The use of contextual information in demand forecasting

Anna Sroginis

Submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy at the Department of Management Science, Lancaster University

September, 2021

Declaration

This thesis is my own work and it has not been submitted in support of an application for another higher degree or qualification elsewhere.

Anna Sroginis

Abstract

Despite the vast advances in quantitative forecasting methods and their prevailing use in practice, expert judgment remains crucial in forecasting; either by specifying and supervising models, or by adjusting model forecasts using expert knowledge and contextual information. The latter is typically done outside any Forecasting Support System (FSS) since, unfortunately, many FSSs neglect judgmental interventions. This can complicate the job of analysts and makes the use of judgment challenging to track and evaluate. This thesis aims to explore how forecasters use contextual information to adjust statistical forecasts, narrowing our focus to demand forecasting.

First, we analyse a UK-retailer case study exploring its operations and forecasting process. Inspired by this case company, we describe laboratory experiments, simulating a demand planning process where both qualitative and quantitative information of unknown quality is presented to experimental participants. In the first study, we find that forecasters mainly focus on model-based anchors, yet they are also prone to react to overly optimistic qualitative statements. The results show the potential for participants to extract useful information even despite this task's complexity and information overload. One option for reducing the cognitive load is by structuring the contextual information. Hence our second study investigates whether the decomposition of qualitative information in the forecasting process with system support could be an effective tool for managing information overload and helping to weight it appropriately. Using a similar setup to the first experiment, we observe better forecasting performance when decomposition is employed. We also find it reduces the number and size of adjustments across all investigated treatments, which could counter some cognitive biases, such as the optimism bias.

While judgmental adjustments are the most common and well-documented form of expert interventions, in the case company we observe an alternative way of incorporating contextual information, which we call judgmental model tuning. Judgmental model tuning represents all changes that analysts make during the model building stage, such as adding

features and changes to the model form. We analyse the use of judgmental model tuning, its effect on forecast accuracy, and contrast it with judgmental forecast adjustments. We find that model tuning is not as effective as we expect since demand planners tend to overuse it. Furthermore, we show that demand planners have the potential to identify and add important variables that are meant to capture special events into a forecasting model, but they do that inconsistently and too frequently. Nonetheless, our results are promising, since judgmental model tuning is cognitively easier, scalable, and can be adopted for multiple products simultaneously. The overall aim of this research is to understand how organisations utilise contextual information and algorithms and to design processes that better integrate these various sources of information to yield more accurate forecasts. This thesis contributes to these areas and identifies new future research questions that are important for effective use of human interventions in forecasts.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Forecasting and planning are critical for many human and business activities. For businesses, more accurate forecasts give a significant competitive advantage of being able to plan their operations efficiently. This allows them to maximise their profits while avoiding waste of resources, both financial and human (Ord et al., 2017, Chapter 1). Forecasting supports many different fields, such as demand planning, supply chain management, energy consumption, weather modelling, financial and economic predictions, etc. However, individuals and organisations can still underestimate the importance of forecasting and the associated planning processes.

There are three principal forecasting methodologies: (i) quantitative methods, such as extrapolative (e.g., exponential smoothing, ARIMA models), causal (regression-based models) and advanced nonlinear methods (Neural Networks and various Machine Learning techniques); (ii) pure judgmental forecasts, prevalent in cases where there is no sufficient quantitative data; (iii) a combination of quantitative methods and judgment. Even though it might seem that these forecasting approaches compete with one another, in fact, they are tailored for specific situations, depending on the availability and quality of both data and human resources (Blattberg and Hoch, 1990; Goodwin and Fildes, 1999; Alvarado-Valencia et al., 2017).

While the use of quantitative methods is frequently adopted and technically convenient for the majority of forecasting tasks, as many forecasting competitions demonstrate (Hyndman, 2020), human judgment is irreplaceable for certain situations. For

instance, expert judgment is an instrument to incorporate contextual information that is not available to the statistical model into the final forecasts. According to Lawrence et al. (2006), this contextual information (sometimes referred to as domain knowledge in the literature) is all pieces of information additional to the time series data history which may be called on to help understand the past and to project the future: this might include "past and future promotional plans, competitor data, manufacturing data and macroeconomic forecast data". Thus, the combination of human judgment based on this additional information and quantitative methods can potentially yield more accurate and prompt results (Zellner et al., 2021).

1.1 Judgmental forecasting research

Before the foundation of The International Institute of Forecasters in 1981 and subsequent major forecasting journals (Journal of Forecasting (JoF) and the International Journal of Forecasting (IJF)), one of the areas of behavioural forecasting in time series was accounting earnings and auditing (Libby, 1975; Libby and Lewis, 1982; Brown, 1993; Nelson and Tan, 2005), where lots of behavioural laboratory experiments were carried out (Swieringa and Weick, 1982). These studies described forecasters' behaviour when analysing and predicting company earnings (Ramnath et al., 2008). Following developments in economic and macroeconomic forecasting (McNees, 1990; Bathcelor and Dua, 1990; Turner, 1990; Lim and O'Connor, 1995; Clements, 1995), behavioural forecasting has changed its course from warning against the use of predictive human judgment due to less accurate estimation, biases and psychological issues (Hogarth and Makridakis, 1981) to acceptance, becoming an essential part of sales forecasting as well (Bunn and Wright, 1991). Researchers realised that experts can still add value either by specifying and supervising the statistical methods or incorporating the information that has not been taken into account (Lawrence et al.,

2006; Arvan et al., 2019; Perera et al., 2019). Hence, we argue that it is important to understand how human judgment complements quantitative methods in the best way possible.

Despite its important role, it is impossible to quickly evaluate the impact of expert judgment in a similar way to the majority of forecasting competitions such as the M-competitions (Makridakis and Hibon, 2000; Makridakis et al., 2020) or Global Energy forecasting competitions (Hong et al., 2014, 2016) where various forecasting techniques are tested on many time series. There was an exception of the M2 competition, where the combination of judgmental and statistical modelling was used on only 29 time series (Makridakis et al., 1993). But it is still unclear how to conduct appropriate forecasting competitions with expert judgment to avoid prospects of forecasters "eyeballing" the past history of the time series (Lawrence, 1993). The second issue arises from organisational practices and their impact on expert judgment. Hence, it is not trivial to estimate and compare Forecast Value Added (FVA) (Gilliland, 2010) for the performance before and after the incorporation of expert judgment across different datasets and experts.

From the early research of Mathews and Diamantopoulos (1986), there has been an increasing interest in the human role in demand forecasting in both academia and practice. In practice, we see a persistent use of expert judgment when producing forecasts (over a third of all forecasts reported) in different companies over the last two decades (Sanders and Manrodt, 2003; McCarthy et al., 2006; Fildes and Goodwin, 2007a; Weller and Crone, 2012; Fildes and Petropoulos, 2015), while the number of published academic papers on human judgment has been increasing over time (Arvan et al., 2019), although not necessarily as a proportion of forecasting publications. Moreover, this trend in research is also evident across different applications: from operations research to other management practices (see special issues of JOM and EJOR, e.g. Zhao et al., 2013; Franco and Hämäläinen, 2016; Durbach and Montibeller, 2018).

This has a twofold benefit. First, more studies show that human decision-making is essential in many business processes. Second, it allows us to explore human judgment, its various known cognitive biases and inefficiencies in detail in the context of time series analysis, and develop tools to aid decision-makers using modern computer systems and advanced behavioural methods. This thesis aims to contribute to this fast-growing body of literature on human decision-making in forecasting, focusing on the effective combination of both judgment and model inputs.

In forecasting, the main body of the research on human judgment has been focusing on its accuracy and a comparison to quantitative methods (Fildes et al., 2009; Franses and Legerstee, 2010, 2011; Trapero et al., 2013; Legerstee and Franses, 2014; Van den Broeke et al., 2019), rarely describing the process and outlining the reasons behind the incorporated judgment, while ignoring the interaction between experts and systems altogether. However, the latter is an important aspect of this process since we know that experts tend to make unnecessary and often harmful adjustments (Goodwin, 2000; Lawrence et al., 2006; Fildes et al., 2009), see false patterns in noise (Harvey, 1995; Goodwin and Fildes, 1999; O'Connor et al., 1993; Reimers and Harvey, 2011), and are prone to many known cognitive biases (Tversky and Kahneman, 1974; Kahneman, 2012) such as anchoring (Harvey, 2007), and over-optimism (Fildes and Goodwin, 2007b; Franses and Legerstee, 2009; Windschitl et al., 2010; Libby and Rennekamp, 2012) and their sources (e.g., a sense of ownership described by Önkal and Gönül (2005)). Neither scholars nor practitioners know yet how to control for these weaknesses effectively (see some attempts and suggestions relevant to forecasting by Goodwin, 2000; Eroglu and Croxton, 2010; Fildes and Goodwin, 2007a; Goodwin et al., 2011).

Despite these known cognitive biases and inefficiencies, human judgment is essential in processing and incorporating contextual information into quantitative methods in a timely manner (Webby and O'Connor, 1996; Goodwin and Fildes, 1999;

Cheikhrouhou et al., 2011; Trapero et al., 2013). Hence, we need to understand how this information can be used, whether forecasters can filter and focus on salient pieces and how they can be guided to better forecasts. Using controlled laboratory experiments, Lawrence et al. (1985), Edmundson et al. (1988) and Sanders and Ritzman (1992) started analysing the use of contextual information in pure judgmental sales forecasting, demonstrating improvements when adopting this additional information. Webby et al. (2005) incorporated contextual information into a Forecasting Support System (FSS), showing its effectiveness for special events (e.g., promotions, strikes). However, all these studies investigated judgmental forecasts exclusively, i.e., without any assistance from quantitative methods. This limitation is not relevant for many companies due to the wide adoption of various computer systems that had enabled algorithmic forecasting.

1.2 The process of demand forecasting

Algorithmic forecasting is a key element in modern sales and operations planning (S&OP) business processes. S&OP is an integrated business framework, which usually provides a sales and operations planning tool not only for production but also for sales, demand forecasting, and resource capacity planning (Thomé et al., 2012; Tuomikangas and Kaipia, 2014). At the heart of any S&OP process is the forecasting and demand planning activities.

Figure 1.1 is a general presentation of observed practices in organisations, where both quantitative and qualitative data are used. While a FSS facilitates algorithmic forecasts, a demand planner has to integrate all additional information coming from sales, marketing, finance and production. Despite the fact that judgmental interventions are common, it is not yet a common practice to support expert adjustments in FSSs (Fildes et al., 2006), track their reasons, evaluate final forecasts and give

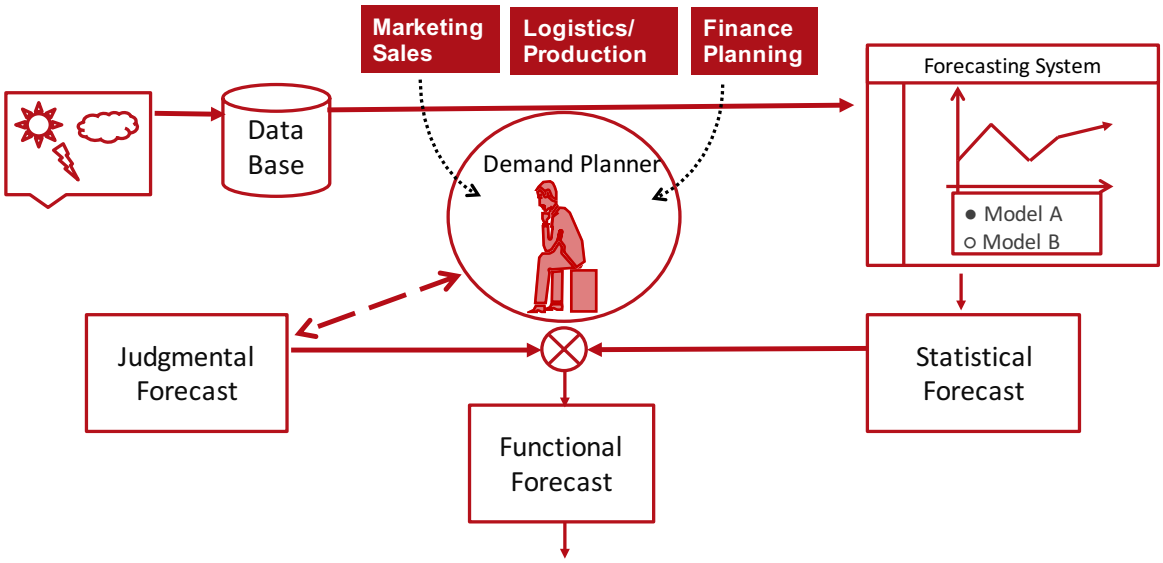

Figure 1.1: Forecasting process in S&OP, adapted from a presentation "Experiments in Operations Management: Information use in supply chain" by Fildes R. in 2018

appropriate feedback (Fildes et al., 2018, 2020). Hence, Fildes et al. (2019a) moved a step further and conducted an experiment where contextual information, graphical time-series sales history and statistical forecasts were presented side by side in a simulated FSS. In that study, all the contextual information did not have predictive power by design. The results showed that participants mis-weighted this qualitative information and tended to underestimate the required adjustments for the promotion. Yet, it is unclear to what extent this contextual information affects human judgment, especially if the mix of relevant and irrelevant pieces of qualitative information is presented, and how it interacts with model-based advice.

1.3 Research question and methodology

This doctoral thesis focuses on the use of contextual information in Forecasting Support Systems in demand planning, looking at the behaviour of forecasters when dealing with such forecasting tasks. In this thesis, we aim to address the following broad research question:

Research question: *How do forecasters use contextual information with unknown predictive value (diagnosticity)?*

To address this wide RQ, we look at the following detailed questions: (1) What is the effect of contextual information of unknown diagnosticity on human adjustments of statistical forecasts during special events and promotional periods? (2) Does decomposition of contextual information help FSS users to identify relevant statements more effectively? (3) Is judgmental model tuning more accurate than forecast adjustments? These questions aim to disclose and analyse some of issues that we observe in practice.

The first two research questions naturally require experimental methodology, which allow the analysis of causal effects in a controlled setting. To make the experimental design and key features more informed, we draw the details from a case study of a major UK-based retailer. This case study describes the company forecasting process, their use of statistical models, software and judgment in demand planning and forecasting. The case study approach is an important tool for gaining deeper and more detailed organisational insights (Baškarada, 2014; Walsham, 2006), yet it is not frequently used in forecasting research (Boone et al., 2019). For instance, Roth and Rosenzweig (2020) reported no case studies published in the Manufacturing and Service Operations Management journal since its inception in 1999, while in the forecasting literature there were only a handful of case-based papers (Fildes, 2017; Fildes and Goodwin, 2021). Hence, we adopt a multi-methodological approach where a case study inspires not only the research questions, but also the key elements in the laboratory experiment itself. While behavioural studies are typically conducted using experiments where subjects complete tasks on a computer either online or in a laboratory (Bendoly et al., 2006), the inclusion of a complementary case study should lead to a stronger connection between research and practice and better triangulation of the results (Croson et al., 2012).

1.4 Contributions

Chapter 2 starts with a case study, highlighting key elements in the forecasting process of a major retailer who combines both qualitative and quantitative information quite frequently. Then we introduce a laboratory experiment with both model-based and contextual information with different predictive power presented to a user. We found that forecasters could be overwhelmed by the amount of information that they need to take into account in their decision-making process. Yet, they exhibit the potential to extract useful information even under such conditions.

The results of this chapter have been presented at the *37th & 38th International Symposium on Forecasting* and at *Operational Research Conference (OR60)*. It is available as a working paper (Sroginis et al., 2019) and has been submitted to the *Production and Operations Management* journal and later rejected with helpful reviews. It is now submitted to the *European Journal of Operational Research*.

Generally, forecasters are required (1) to assess the quality of system inputs and outputs, and then (2) to weight multiple pieces of contextual information to incorporate into the final decision. Hence, the simultaneous use of qualitative and quantitative information of unknown quality requires sufficient cognitive resources to process and advise a final synthesised forecast. The process of integration could introduce an information overload and dramatically increase the task complexity (correlated to the number of contextual statements presented) (Webby et al., 2005). One of the possible ways to reduce cognitive load is by structuring knowledge into schemas (Cook, 2006), which is known as decomposition in the forecasting literature (Goodwin and Wright, 1993). Wolfe and Flores (1990) proposed using the analytical hierarchy process (AHP) for improving judgmental adjustments by weighting various factors structurally. However, Belton and Goodwin (1996) outlined some potential problems, highlighting that this approach allows only relative weighting (where one of the fac-

tors is a baseline) rather than absolute, which might not be useful in judgmental forecasting since absolute weights are needed.

In the second paper of this thesis (Chapter 3) we analyse the effect of the decomposed presentation of contextual information on expert adjustments in forecasting. This study is inspired by the first paper and ultimately aims to investigate whether the decomposition of qualitative information could be an effective tool for managing cognitive load in forecasting tasks. We found that decomposition helps to identify relevant information for both baseline and promotional models. We argue that this is due to reducing the task complexity by guiding people to weight textual information sequentially. We also found that decomposition of qualitative information reduced the size of judgmental adjustments on average across all conditions and this could be a useful design feature for support systems.

Chapter 3 is now a working paper with the aim to be submitted to the *Decision Support Systems* journal. This work has been presented at the *39th International Symposium on Forecasting* and at the *15th Annual Behavioral Operations Conference(BOC).*

Both the first and the second studies employ an experimental methodology to investigate the use of contextual information in a simulated FSS. This support system was based on a case study (presented in Chapter 2) of a large UK retailer, who uses (1) the well-established SAP F&R system for their statistical modelling for both non-promotional and promotional products, and also (2) qualitative information received through various channels, such as formal/informal meetings, emails and phone calls. While judgmental adjustments are the most common approach of incorporating this information into final forecasts, we observed an alternative, previously unrecorded, way of introducing this knowledge.

The third paper (Chapter 4) investigates this new method of injecting human judgment into forecasts and compares it to judgmental adjustments. Instead of es-

timating the effect of contextual information judgmentally, the forecaster can add additional indicator variables into a statistical model. The system then estimates an appropriate modification. We call this process *judgmental model tuning* of an initial statistical model. Judgmental model tuning should be easier than adjusting forecasts since only the location of a change is required rather than both location and value assignments. However, since not every piece of contextual information can be transformed into indicator variables for various reasons (e.g., not repetitive events, sporadic behaviour, too specific), judgmental model tuning is hardly a substitute for judgmental adjustments. The main risk for judgmental model tuning is that it is easier than making adjustments, and experts might saturate the model by adding spurious variables, leading to potential overfitting.

We found that model tuning is ineffective. Some of these added variables were observed to be useful and beneficial, but their effect was diminished by the number of harmful indicators. However, once we remove judgmentally imposed insignificant explanatory variables, it provided accuracy gains, indicating that model tuning can be subject to substantial biases that can harm its performance, yet remains promising as it scales better than judgmental adjustments of statistical forecasts.

Chapter 4 is ready to be submitted to the *International Journal of Forecasting*. It has been presented at the *40th International Symposium on Forecasting* and at *INFORMS Annual Meeting 2020*.

Human judgment and the role of contextual information within it connect all three papers, bringing new insights into the forecasting process, its complexities, possible problems and potential solutions. At the same time, we identify many more questions to explore regarding human trust in models, its impact on their decisions and the potential limitations of the combination.

1.5 Structure of the thesis

The remainder of this thesis is structured as follows. The next chapter presents an experimental study on the effect of contextual information of unknown value on human judgment in forecasting. In Chapter 3, we investigate the impact of decomposition on human cognitive load in forecasting tasks. In Chapter 4, we explore an alternative method to incorporate expertise at different stages of the forecasting process. Finally, Chapter 5 discusses the contributions this thesis makes to the literature and concludes with managerial and software design implications; alongside the limitations, the research proposes ideas for future research.

Chapter 2

Use of contextual and model-based information in behavioural operations

Abstract

Despite improvements in statistical forecasting, human judgment remains fundamental to business forecasting and demand planning. Typically, forecasters do not rely solely on statistical forecasts; they also adjust forecasts according to their knowledge, experience and information that is not available to the statistical models. However, we have limited understanding of the adjustment mechanisms employed, particularly how people use additional information (e.g. special events and promotions, weather, holidays) and under what conditions this is beneficial. Using a multi-method approach, we first analyse a UK-retailer case study exploring its operations and forecasting process. We build on this by laboratory experiments that simulate a typical supply chain forecasting process. We provide past sales, statistical forecasts (using baseline and promotional models) and qualitative information about past and future promotional periods. We find that when adjusting, forecasters tend to focus on model-based anchors, such as the last promotional uplift and the current statistical forecast, ignoring past baseline promotional values and additional information about previous promotions. But still the effect of contextual information on human decisions is significant

in both the case study and the experiments. The impact of contextual statements for the forecasting period depends on the type of statistical predictions provided: when a promotional forecasting model is presented, people tend to misinterpret the provided information and over-adjust, harming accuracy.

2.1 Introduction

Supply chain and inventory decisions directly depend on accurate demand forecasts, especially in retail operations. A typical forecasting process relies on two main components: statistical modelling and human judgment. While the former aspect is well developed and implemented in various forecasting packages, the latter is less well understood: it may be used in data collection (data selection), or in the model selection stages or, more crucially, directly in making judgmental adjustments/overrides (see extensive overviews of judgmental forecasting literature by Lawrence et al., 2006; Arvan et al., 2019; Perera et al., 2019). It has been argued that expert adjustments to statistical forecasts can be beneficial for a company (Fildes et al., 2009), since adjustments can incorporate contextual information[1] that is not taken into account by the statistical model available. However, it remains unclear how and why experts modify these forecasts and whether the increased availability of the more advanced forecasting methods, that have become available, influence the effectiveness of the process overall.

In this chapter, we focus on the human decision-making process in adjusting sales forecasts, when a statistical forecast is given. There is limited research on the use of domain knowledge in forecast adjustment, along with the factors that can influence the accuracy of these decisions, and yet this combination is at the heart of the Sales and Operations Planning (S&OP) process (Oliva and Watson, 2009; Stahl, 2010). We investigate these questions using a multi-method approach combining a case study with a laboratory experiment. The experimental design is motivated by the practices of a major UK retailer established through a case study of their forecasting process, their use of a Forecasting Support System (FSS), a specialised version of Decision

[1]The domain or contextual data is non-time series information. This might include past and future promotional plans, competitor data, manufacturing data and macroeconomic forecast data (Lawrence et al., 2006).

Support Systems (DSS), and the adjustments demand planners made.

We aim here (i) to explore the effects of qualitative information on human adjustments as they depend on *two types of statistical forecasts*: baseline (where the forecasts are purely time series extrapolation), and a promotional model capturing past and future promotional events; (ii) to investigate the factors that influence final forecast accuracy in these conditions; and (iii) to examine potential anchoring effects in demand planners and their acceptance of statistical forecasts with promotional effects. In particular, we want to understand and test if people can identify and correctly use relevant information, while filtering out statements of dubious predictive value, investigating a common situation observed in practice.

This study aims to reveal those factors that influence the interaction of judgment with model-based forecasts and provides a foundation to explore how these adjustments could be improved. In part, accuracy improvements may be achieved through the use of a more accurate forecasting model (e.g., capturing promotional events), but we are interested to see if the results are affected by how the different statistical models are used.

This chapter contributes to the existing literature in two ways: (i) it presents a case study of the use of a Forecasting Support System in the demand planning process of a large retailer, providing insights into the questions raised above and (ii) using controlled laboratory experiments, which are inspired by the case study, it analyses expert adjustments in the presence of promotions, when the forecasters are supplied with model-based and contextual information. Our results are highly relevant to practice, identifying weaknesses in typical forecasting processes and their reliance on correctly interpreting the advice delivered by support systems. The research opens up novel paths of further research into support systems and the interaction between decision makers and support systems.

This chapter is organised as follows: Section 2.2 provides an overview of the lit-

erature on judgmental adjustments and promotional modelling. In Section 2.3, we discuss multi-method approaches and their value in research. Section 2.4 presents a case study of a major UK retail company. Following Section 2.5 introduces the experiment setting, and Section 2.6 provides the analysis of the experimental results. The discussion is presented and conclusions are drawn in Section 2.7.

2.2 Related literature and research questions

Advice on using judgment in forecasting has been changing over time: from warning against its use, due to less accurate predictions, biases and psychological factors (Hogarth and Makridakis, 1981), to the acceptance that judgmentally adjusted forecasts can add value to organisations (Lawrence et al., 2006). In fact, even if the majority of statistical forecasts are more consistent (following a specific statistical method or model) than human predictions, there is evidence that people usually prefer judgmental methods (Önkal et al., 2009), and this approach prevails in practice (McCarthy et al., 2006). This is also supported by many surveys over the last two decades. In Table 2.1 we have summarised the findings of all available surveys, which shows the use of judgmentally adjusted statistical forecasts corresponds to around 40% of responses. Further evidence comes from field studies: Fildes et al. (2009) and Franses and Legerstee (2009) reported a high percentage of judgmental interventions of statistical demand forecasts in their company-based case studies, up to 90% in some cases.

The effects of expert adjustments can be substantial and are currently poorly understood. To this end, a number of studies have attempted to investigate the effect of judgmental adjustments using company data (Sanders and Graman, 2009; Fildes et al., 2009; Franses and Legerstee, 2011; Franses, 2013; Trapero et al., 2013; Syntetos et al., 2016b; Van den Broeke et al., 2019), focusing on their effectiveness and possible

Table 2.1: Survey studies of methods used in practice.

Method	Study[a]				Average weighted by sample size
	A	B	C	D	
Judgment alone	30%	25%	24%	14%	26%
Statistical methods exclusively	29%	25%	32%	30%	28%
Average statistical & judgment	41%	17%	–	19%	17%[b]
Adjusted statistical forecast		33%	44%	37%	36%[b]
Sample size	240	149	59	42	

[a] A: Sanders and Manrodt (2003); B: Fildes and Goodwin (2007a); C: Weller and Crone (2012); D: Fildes and Petropoulos (2015).
[b] Excluding Study A from the calculations.

weaknesses, in particular, their behavioural characteristics. For instance, Fildes et al. (2009) analysed data from four companies (three manufacturers and one retailer) and found that negative adjustments of statistical forecasts on average increased final accuracy, while positive ones did not (especially in the case of the retailer).

In the presence of promotions, Trapero et al. (2013) showed that judgmental adjustments could enhance baseline forecasts during promotions, but not systematically. Even when using an appropriate statistical model with promotional effects, there were cases where humans still added value, pointing to the need for more research on the conditions where human interventions would add value to forecasts when promotional effects took place.

Using real data from four companies, Van den Broeke et al. (2019) analysed expert adjustments over different time horizons. Not only was a high number of adjustments reported, but also the authors showed that the number of adjustments increased closer to the sales point, and these corrections became larger and more positive but not necessarily more accurate. Again, these adjustments were typically based on knowledge of the product, the market, or new information regarding future actions (such as promotions).

All these empirical studies neglect or just briefly touch upon the analysis of ac-

tual/possible reasons for these judgmental adjustments. Hence, they are missing an essential step in describing and evaluating the potential benefits and drawbacks of this process. Only a few studies have explored the reasons for these interventions. According to Fildes and Goodwin (2007a), the main causes of interventions are usually various special events, such as promotions, price changes and holidays. Also, evidence has accumulated that political and budget reasons could motivate experts to intervene (Galbraith and Merrill, 1996; Lawrence et al., 2000; Fildes and Petropoulos, 2015). In addition, various biases and human factors may influence the size and direction of adjustments, such as a sense of ownership (Önkal and Gönül, 2005), anchoring (Harvey, 2007), neglecting base rates and looking for similar patterns in the data, and reliance on readily available information (for general psychological overviews, see seminal books of Tversky and Kahneman, 1974; Gigerenzer, 1999; Kahneman, 2012).

The options that may help to mitigate human biases and inefficiencies in time-series forecasting could be summarised as: forecasters' training (Fildes et al., 2009); improved statistical forecasting support systems in order to produce better system forecasts (Fildes et al., 2009); reliable choice and interpretation of error measures (Davydenko and Fildes, 2013; Petropoulos et al., 2016); recording reasons for adjustments (Fildes and Goodwin, 2007a; Goodwin, 2000); provision of guidance or feedback (Goodwin et al., 2011); improvements in the S&OP process in which diagnostic information[2] is collected as a whole in order to reconcile different sources of information in the forecasting and planning decisions (Fildes et al., 2006).

Several laboratory experiments have been conducted to analyse the use of contextual information in forecasting. Initially, in one company example Edmundson et al. (1988) showed that product knowledge could significantly improve forecast accuracy compared to extrapolative forecasting approaches alone. Sanders and Ritzman (1992) conducted a similar experiment involving students and practitioners, reporting that

[2]The diagnosticity of a piece of information is a measure of its helpfulness and usefulness for making a judgment (or forecast) in empirical studies (Qiu et al., 2012).

judgmental forecasts with contextual information were significantly better than judgmental forecasts without such information. However, both studies looked at pure judgmental forecasts without any support from statistical methods and user-friendly interface (e.g. presenting information only in a tabular form and considering experience and domain knowledge as an essential attribute of practitioners). And later, in a different experimental study, Webby et al. (2005) showed the effectiveness of a FSS in judgmental forecasting (without statistical model support) under special events (e.g., promotions, strikes) with different information loads, ranging numbers of contextual statements presented.

Recently, Fildes et al. (2018) investigated the use of additional qualitative information along with graphical time-series sales history and statistical forecasts assisted by an FSS. Participants were provided with information about the baseline uplift for promotions and could adjust the forthcoming promotional period having all the available information on the same screen. The authors found that participants mis-weighted the provided information and tended to underestimate the required adjustments for the promotion. This study raised a series of important questions about the effect of contextual information on forecast accuracy, its appropriate representation and access to domain knowledge. Our research expands on this work. More specifically, we attempt to simulate business practice more faithfully by highlighting different types of domain knowledge inputs.

Building on the findings of both field- and laboratory-based studies, we aim to answer the following question:

Research question: *What is the effect of contextual information of unknown diagnosticity on human adjustments of statistical forecasts during promotional periods?*

In this context, any verbal/textual information that contains some subjective belief, is motivated or biased, sometimes can be conflicting and has an unknown degree of

reliability. For example, the Marketing department might say: "Based on market research, the Marketing Manager concluded there would be a positive reaction to the promotional campaign" together with "Our spending on this campaign is only 50% of our normal costs". What information do people use when making their adjustments and how do they weight it? This additional information might be either quantitative or qualitative, originating from different sources (e.g. marketing, production or supply chain departments) and with unknown predictive diagnosticity.

In summary, there is limited research on the subject of the use of FSSs and contextual information that naturally accrues from various aspects of S&OP processes. Judgmental adjustments are commonplace in practice, and therefore it is important to understand the underlying mechanics better, particularly as they are affected by the forecasting advice given by models of different complexity and completeness. Answering the research question will have implications for the design of organizational forecasting processes.

2.3 Methodology

Operations Management (OM) in general, and forecasting in particular, has seen little embedded research where the models have been developed in a specific context embracing the operational constraints organisations face. There has been little organisationally based research in the OM literature (Roth and Rosenzweig, 2020). Choi et al. (2016) by analysing a wide range of OM journals found that while there has been a growth in case-based research, the analytical research methodology dominates both quantitative empirical and case-based studies. This leads to the criticism that without adopting a multi-method approach, relevance can be unnecessarily sacrificed, instead favouring uni-disciplinary rigour.

Similarly, the primary emphasis in forecasting research has been on new modelling

methods and their comparative effectiveness (Fildes, 2017). Nonetheless, there has also been a trickle of studies that have focused on what we have demonstrated through Table 2.1 in the literature review: the common practice of judgmental interaction with models and forecasts. However, the results from the various studies are often conflicting, arguably due to the very different situations in which data have been gathered and the epistemological perspective of the researchers. Following on from Choi et al.'s (2016) argument, the research question discussed in the previous section requires a multi-method methodology, if the results are going to have any validity and practical application in the research context of supply chain forecasting. Since one of our aims is to make recommendations as to how demand planners should use the forecasting support systems at the heart of the demand planning process, we need to examine how the actors perform their tasks in an archetypal organisation. As a consequence, we adopt a multi-method approach to develop guidelines for research directly relevant to practice in the design of forecasting processes and the use of FSSs. The methods we use and their justification are discussed below.

2.3.1 Case analysis

Much of the research on judgment in forecasting is carried out in abstract (Lawrence et al., 2006). Yet, the organisational context in which forecasting occurs is rich in features that have the potential to influence the results of any field studies and also impact the validity of interpreting experimental results in practice. The evidence on the ubiquity of the forecasting process is provided in many field studies (e.g., Fildes et al., 2009; Trapero et al., 2013; Franses and Legerstee, 2011; Van den Broeke et al., 2019) with further evidence of its use in retail demand forecasting (Fildes et al., 2019b). Thus, it is important to ground any experimental (and analytical) work in these organisational realities. These features include: (i) the nature of the demand process, the core baseline forecasting models, and the more advanced promotional

models increasingly used in companies; (ii) the interventions that demand forecasters make; (iii) the events such as promotions that provoke interventions, and other relevant indicators, such as calendar events; (iv) the information (and organisational processes) used underpinning these interventions. All this richness is typically either ignored or not reported in many academic papers.

We have spent considerable time with the case company, a major UK non-grocery retailer (i) semi-structurally interviewing actors and observing the process by which the demand forecasts are produced and the information flows (unstructured/naturalistic observation method, where demand planners show how they are dealing with some forecasts); (ii) collecting data on the model-based forecasting routines, initially exponential smoothing and latterly a regression-based method; and (iii) the evaluation of their forecasts to provide evidence of the efficiency of these interventions. This company is a typical example of a large enterprise that shows how complex the whole planning process is. However, the results of this case study can not be easily generalised since they depend on a combination of technological and human factors.

2.3.2 Experimental study

An increasingly popular methodology for analysing judgmental adjustments is laboratory-based experiments either with students or practitioners. This gives control of the sample, conditions and experiment setting. However, such experiments are typically focused on a simplified problem, as the organisational complexities can not be fully transferred to an experimental setting. For instance, both De Baets and Harvey (2018) and Fildes et al. (2018) investigated different aspects of judgmental adjustments of statistical forecasts using laboratory experiments.

In particular, De Baets and Harvey (2018) studied adjustments under various levels of forecast support (from none to using an exponential smoothing model with

an integrated promotional effect) on data with promotions. The authors showed that providing statistical forecasts was beneficial in every case and was even more useful when statistical forecasts incorporated promotional effects.

Our research question focuses on the interpretation of information in the light of two different forecasting models (with and without integrated promotional effects) employed. By controlling for the various factors we have ascertained as potentially affecting an organisation's forecasts, we have aimed to identify those that are important both in practice and in the design of FSSs.

Relying on our two-pronged approach, we validate our experimental study by reflecting on our analysis of the field data.

In the next section, we discuss a case study from a UK retailer to provide background information on the various types of data and how these are presented and interpreted in practice, as well as the forecasting process in the organisation, so as to motivate the design of our laboratory experiments. This ensures that the findings are directly relevant to retail practice (Fildes et al., 2019b).

2.4 Case study: adjustments in practice

Earlier research by Fildes et al. (2009) had investigated the role of judgment in demand planning activities in four companies, all of which used an exponential smoothing forecasting algorithm. One of them, a retailer, continued to work with Robert Fildes, on promotional modelling amongst other topics. With the move to SAP software, most recently to the F&R module (Forecasting and Replenishment), the company offered an ideal base for understanding how this new software was used and the new role of the judgments made by the demand planners. Thus, we present a case study of a UK-based major retail company that provides an interesting example of the current processes employed in companies and shows many similarities with other retailers

(Seaman, 2018; Fildes et al., 2019b).

We have collected and analysed data between 2017 and 2019. Further evidence was collected from interviews with demand planners, forecasting managers and also from an analysis of the software manuals as well as an extensive analysis of the company's data.

2.4.1 The forecasting process

The case study is based on a retail company that focuses primarily on household and fast-moving consumer goods (FMCG). There are around 50 thousand SKUs (Stock Keeping Units) organised in 3 supercategories and around 20 lower-level categories. Only 20 thousand of these products are regular items, and all others are either new or seasonal. The data are highly intermittent at store level: only about 10% of all SKUs have fewer than 20% of zero-demand periods (based on approximately two years of weekly data). Moreover, items are heavily promoted, especially in the supercategory of FMCG (10 types of promotions and each type having many different durations).

Figure 2.1 presents a flowchart of the forecasting process implemented in the case company. Initially, there were several inputs: a rolling history of 120 weeks of sales as a base for its weekly forecasts; yearly promotional plans and calendar events (A and B input nodes in Figure 2.1). The company used the SAP F&R module, where the implemented algorithm switches forecasting models between non-promotional and promotional periods. The exact forecasting methods and heuristics are not publicly available. The company's forecasting process manual stated that a weighted mean was used for non-promotional periods, while a regression-based model with dummy variables was used during promotional periods, holidays and other events. Since the company had promotionally intensive products, demand planners could adjust the F&R forecasts based on any qualitative and quantitative information available to them.

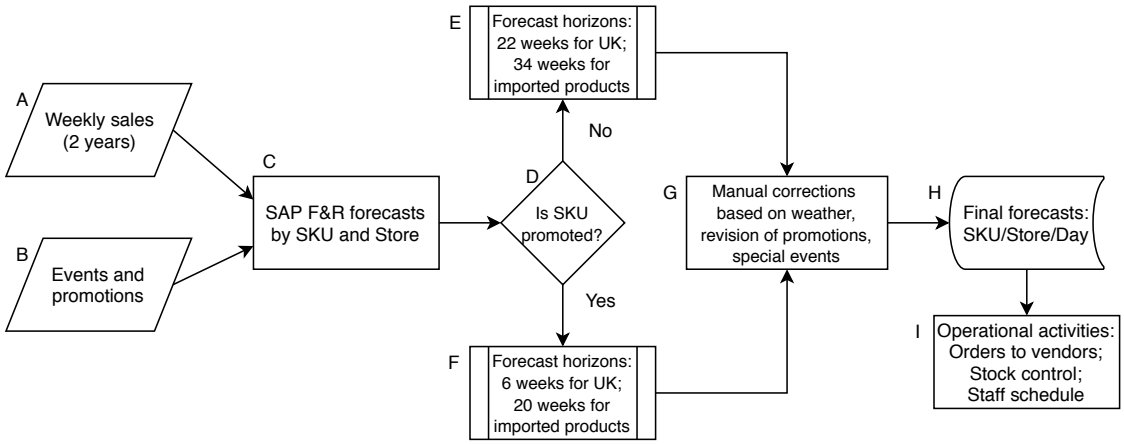

Figure 2.1: Forecasting process of the case company.

In general, there were several operational forecast horizons that depended on the type of a product (locally sourced/imported) and whether this product was on promotion or not. These horizons can be seen in boxes E and F in Figure 2.1. Surprisingly, only one week (one step) ahead final forecasts for the past six months were easily extractable from the system. Thus, neither we nor the company could evaluate the accuracy for the operational horizons. This also meant that the company was not able to assess its final forecasts for the longer lead times, as they did not store past forecasts externally.

Statistical forecasts were produced at a store level for each SKU, weekly, using the SAP F&R system (Box C in Figure 2.1). Then these might be judgmentally adjusted by demand planners based on weather, known special events or revisions of promotions (Box G). Finally, the resulting values were disaggregated into days using weekly SKU profiles (which were calculated internally in the system and updated every month) and stored for the following operational activities: orders for distribution centres, stock control and staff scheduling (Boxes H and I respectively). This added further supportive evidence to other studies which described the forecasting process in the retail context (Fildes et al., 2009; Syntetos et al., 2016a; Fildes et al., 2019b) and highlighted that even when statistical models incorporated promotional or special

events effects, manual adjustments were frequently used to override model predictions.

Figure 2.2: Screenshot of an example software used in practice.

The company had several non-forecasting focused performance indicators for evaluating the accuracy, such as working capital, supplier service and availability (number of stock outs). Nonetheless, the forecasting team checked its performance by manually calculating the Mean Absolute Percentage Error (MAPE), see Formula 2.3 in Appendix 2.B, across stores for one step ahead forecast horizon. The internal guideline was that this error should not exceed 20% on average. However, this evaluation of forecasts had only been introduced shortly before our study, and, therefore, there was no historical data of past forecast performance, nor does the measure relate to the operational horizons. Thus, a weakness of this F&R system implementation was absence of forecast evaluation, which is seen as an important dimension for successful forecasting (Moon et al., 2003) and a common limitation of systems in practice (Ord et al., 2017, Chapter 13). According to Petropoulos et al. (2017), experts revise predictions better when provided with forecast bias feedback, which is also absent in the current process. Given that the forecasts support stock control decisions, the forecast bias is particularly relevant for the inventory performance (Sanders and Graman, 2009; Kourentzes et al., 2020). Moreover, the reliability of the statistical forecast,

prior to any adjustments, is also unknown to the users. This can influence their trust in the statistical forecast and therefore their propensity to adjust, making the benchmarking of its performance important. These observations influenced the design of our experiment, as discussed in Section 2.5.

2.4.2 Interface

The SAP F&R system uses a visual interface that plots sales and forecasts stored in the system. The interface provides indicators for promotional and seasonal events, which help demand planners visually assess the quality of system forecasts. Figure 2.2 provides a screenshot, where the dark (red) line corresponds to actual sales and the light (green) one to the company's forecasts. The horizontal lines under the graph represent different types of promotions. Various events, including seasonal events and holidays, are also colour coded and presented in the same box as promotional markers. There is no qualitative information in the system (e.g., marketing information or weather predictions) and demand planners manually track possible reasons for adjustments that are not provided internally in the F&R interface.

2.4.3 Expert Adjustments

There are three types of possible forecast adjustments in the F&R: (i) overwriting the statistical forecasts by inputting a new value (in our dataset this is used only for new products); (ii) to increase/decrease the forecasted value by a given value (no examples in the dataset); (iii) a percentage increase/decrease of the system forecast from the initial (the majority of cases). We focus on the effects of the last option, because this is the most frequent and relevant strategy used in the case company.

The dataset consists of six stores with different turnover (low, medium and high) and is 2% of the stores in the company. Having a range between 20 and 50 thousand SKUs per store, about 25% of all SKUs were adjusted at the store level (around 7000

time series) over the two years of sales history available. It translates to 2 corrections per SKU on average (the maximum observed adjustments per SKU is 14 over 100 observations). Other field studies based on supply-chain and retail companies have shown that the frequency of expert adjustments tends to be similarly very high (Fildes et al., 2009; Franses and Legerstee, 2009), particularly for promotions (Trapero et al., 2013, 2015).

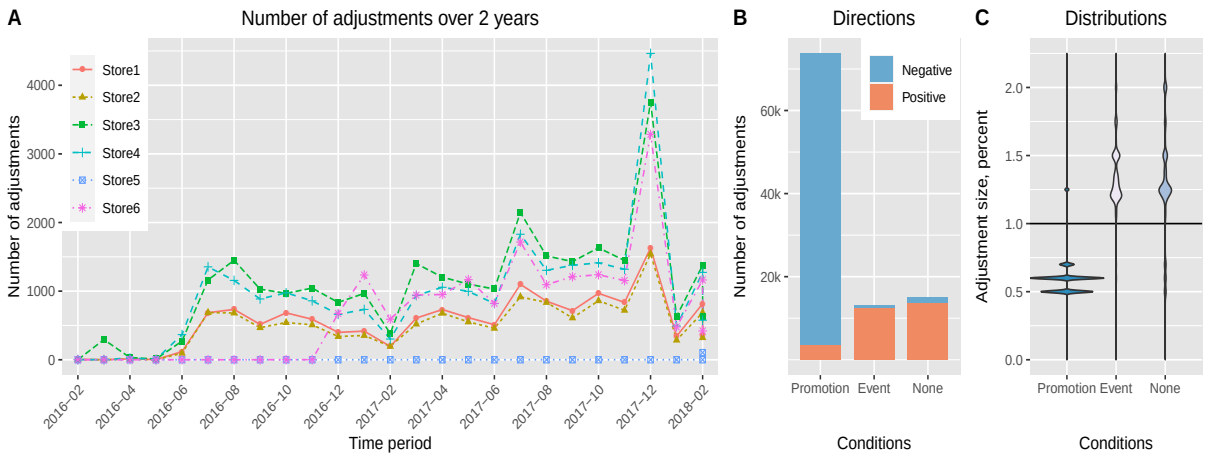

Figure 2.3: Total number of manual adjustments over the 2 years (A); Ratio of positive and negative adjustments (2 years) (B); Distribution of percentage adjustments (2 years) (C).

Around 70% of all adjustments were within the FMCG category of fast-moving products. Interestingly, only 18% of all adjustments were performed on time series with continuous demand across SKUs, suggesting that there was a perception that intermittent time series were not forecasted accurately. As the empirical study of Syntetos et al. (2009) showed, such adjustments could be effective in intermittent demand cases.

Figure 2.3A plots the number of adjustments for the different stores across time. On average, there were around 600 adjustments per month at each individual store. The number of adjustments increased dramatically over the Christmas period (observe the spike in December), which is, in general, a diagnostic period with well-known effects on sales. This spike highlights a perceived problem with the statistical forecast

during the Christmas period, even though the expectation would be that statistical models with explanatory variables for promotions and seasonal effects would provide reliable results.

We identified that there were two main groups of effects in this dataset: promotions and special events (e.g., weather, calendar events). Regarding the direction of these adjustments (see Figure 2.3B), around 89% of all adjustments at the individual store level were negative (i.e., smaller than 1, where 1 is a non-adjusted system forecasts). In particular, corrections for promotional periods were prevailing and were mainly negative, while adjustments for special events and the rest were positive. Interestingly, Figure 2.3C shows truncated distributions of these adjustments: during promotions they were mainly around 50-70% decrease, while for other periods they were typically lower, between 20-30%. This brief analysis highlights a tendency to adjust down often for promotional periods, but also to increase forecasts by a small amount in other cases.

The company did not have any strict procedure for recording and justifying any interventions made, complicating the categorisation of the adjustment reasons. Nonetheless, the recorded descriptions stored in F&R system revealed common keywords and phrases. We were able to extract the keywords: "Adjustments for Promo", "Xmas", "Clearance", "Promo Correct", "Increase" and some common numeric codes. These keywords appeared in descriptions that corresponded to 90% of adjustments and were connected with either promotions or special events. Other reasons for adjustments included weather forecasts and marketing information about competitors' events. Although collecting information for these events was hard and required a substantial time investment, they were usually considered in demand planning meetings. This is in line with findings by Fildes et al. (2019b).

Field studies suggest that these adjustments are not always beneficial in terms of accuracy (Fildes and Goodwin, 2007b; Fildes et al., 2009; Franses and Legerstee, 2009;

Trapero et al., 2013). We proceed to evaluate the added value of the adjustments for the case company, so as to provide a sense of the impact all these adjustments have on the accuracy of the forecasting process.

2.4.4 Accuracy of the adjustments

To evaluate expert adjustments, we ideally need data triplets of actual sales, system, and final adjusted forecasts. In our case, the system forecasts were unavailable, and only 24 weeks of final (adjusted) one-week ahead forecasts and the corresponding actual sales were provided. Sales time series were updated on a weekly level, while the demand planners made adjustments and flagged promotions at a daily level. In principle, the system disaggregates the weekly data using daily weights and introduces the adjustments to the final forecasts. However, these weights were not extractable from the system, and therefore, without these weights, we cannot reproduce the system forecasts. Instead, we relied on a table with differences between internal baseline system forecasts (i.e., without any promotional or other events) and final adjusted forecasts, recorded at a weekly level. Using that, we reconstructed initial system forecasts for the adjusted periods, assuming that even if only one day in a week was adjusted, the effect was propagated to the whole week.

Table 2.2 shows how many adjustments have been done over 24 weeks in these six stores, highlighting the number of observations contributing to better accuracy compared to our reconstructed forecasts. The majority of negative adjustments (i.e. a decrease by 50%, 60% and 70%) lead to positive Forecast Value Added (FVA) calculated using the AvgRelMAE (see Formulas 2.6 and 2.7 in Appendix 2.B). Positive adjustments are split into four groups with a step of 25%, and the last group includes all corrections of more than 100%. There are several interesting observations here. In almost all groups, the majority of adjustments decrease forecast error of the system forecasts. However, taking into account statistics of RelMAE, we conclude that the

Table 2.2: Case study: accuracy of adjustments, 6 months and December (in brackets) evaluation.

	Negative	Positive adjustments					
		Overall	5-25%	26-50%	51-75%	76-100%	101-1500%
Adjustments	3469 (1133)	3436 (2839)	1837 (1509)	1175 (1037)	192 (160)	103 (81)	129 (52)
Negative FVA	1073 (316)	1544 (1307)	815 (677)	529 (474)	86 (78)	63 (53)	51 (25)
Positive FVA	2396 (817)	1892 (1532)	1022 (823)	646 (563)	106 (82)	40 (28)	78 (27)
RelMAE							
Mean	2.00 (1.64)	4.10 (4.08)	5.92 (5.81)	2.02 (2.13)	1.69 (1.60)	3.27 (3.69)	1.43 (1.20)
Median	0.75 (0.73)	0.96 (0.98)	0.96 (0.97)	0.96 (0.97)	0.97 (1.00)	1.06 (1.06)	0.68 (0.97)
St. deviation	16.00 (12.76)	58.76 (60.10)	79.94 (81.97)	9.13 (9.69)	4.05 (3.10)	8.96 (10.01)	4.15 (1.57)

distribution of errors is skewed to the right due to many large outliers. Hence, on average, these adjustments are beneficial, but if they miss, the negative effect is substantial. We can observe several interesting things: first, the smallest positive corrections have the largest variation (i.e., the percentage increase from 5% to 25%); second, adjustments between 76% and 100% have the worst accuracy amongst all groups, while the largest corrections have the best; third, the negative adjustments also prove to be quite accurate. This latter result aligns with Fildes and Goodwin (2007b), where negative adjustments proved more beneficial than positive.

One of the main periods for adjustments was Christmas. About 58% of all changes during these six months happened during December. But only one-third of these adjustments occurred together with promotions, even though it was the busiest month for promotions and other events. The distribution of positive/negative adjustments and its accuracy can also be observed in Table 2.2 in brackets. The number of corrections during this period indicated that the F&R system forecast was not perceived as performing well: this was further demonstrated in a separate benchmarking exercise (against a reconstructed regression-based system) that provided evidence for this case study. So increasing forecasts (see the first row in Table 2.2) substantially could be motivated by reliable information about some future events (e.g. having additional stock, expecting demand for slow-moving products). These overall accuracy results are specific to this company, yet they give a general idea about the effect of

adjustments by size and direction.

As for the main research question, the case study demonstrated once again that judgmental adjustments are common in practice, in particular when potentially relevant domain knowledge is available. The effect of human adjustments of statistical forecasts during promotional periods that are based on contextual information of unknown diagnosticity is mainly positive. Still, incorrect judgments lead to substantial errors and diminish all improvements. As for contextual information, it is collected outside the forecasting support system. And given the lack of detailed recording of historical adjustments, it is impossible to evaluate the information content and usefulness of it, solely by using the case study evidence. More crucially, it is infeasible to reveal how experts distinguish and use different sources of information with unknown diagnosticity. This motivates our experimental stage, which mimics many of the observed features of the forecasting process and F&R system of the case company, so as to provide a realistic test environment to investigate our research questions and related hypotheses.

2.5 Experimental study

2.5.1 Hypotheses

In the experimental part, we aim to investigate adjustments when contextual information from different sources is available to forecasters. As we have observed in the case study, demand planners use different sources of information to make their decisions. Essentially, there are several types of information that could be used for time-series forecasting: (i) model-based (e.g. statistical model fit and forecasts); (ii) qualitative (e.g. contextual statements and explanations); (iii) historical (e.g. past sales and events history). Sometimes, the qualitative information has little or no predictive value, but usually, the value is unknown so that we will characterise this situation

as "unknown diagnosticity". All these pieces of information are non-diagnostic for a user since no background or feedback is provided. Importantly, we analyse expert adjustment given two types of statistical forecasts: *baseline* (a simple extrapolation based on data with promotional observations removed) and a *promotional* model with integrated diagnostic promotional effects. This permits our analysis to be conditioned with respect to the richness of the statistical model. We formulate the following hypotheses for our laboratory experiments:

Hypothesis 1: *Promotional forecasting models are valued more than contextual information of unknown diagnosticity related to promotions.*

The model with integrated promotional effects is perceived to be more valuable than, for example, the statements provided by other functional sources to the forecaster: this leads to more accurate forecasting.

Hypothesis 2: *When the statistical forecast does not contain diagnostic promotional information, the weight of statements with unknown diagnosticity increases.*

In the absence of a trustworthy statistical forecast, the forecaster gives additional weight to both the diagnostic and non-diagnostic statements but is unable to achieve parity with forecasts produced by the diagnostic statistical model.

Hypothesis 3: *Judgmental adjustments based on contextual information with unknown diagnosticity lead to less accurate forecasts than promotional model forecasts.*

Information of unknown diagnosticity potentially induces unnecessary and harmful adjustments, whatever the statistical model is.

The first hypothesis focuses on the use of contextual information where the forecasts provided aim to include promotional effects, which is the situation in more advanced FSSs. The second and third hypotheses are about how contextual information,

arising in an S&OP process, damage or add value to the statistical model forecasting accuracy. More specifically, we aim to analyse how these adjustments change depending on model-based information checking how the reaction to statements of unknown diagnosticity changes depending on the level of diagnostic information included in the model. The forecast accuracy of all scenarios is investigated in the last hypothesis.

2.5.2 Experimental procedure

Building on the insights and information we have gained from the case study, we designed a behavioural experiment that aimed to illuminate how experts perceive the usefulness of domain information and incorporate it in their adjustments, addressing the first aim from Section 2.1. The research question focuses on the forecasters' use of both model-based and contextual information and its relationship to the completeness of the statistical model. Therefore, to identify the impact of qualitative information on human judgments of statistical forecasts, for each of 12 time series the participants were provided with:

1. historical data (a graph of sales for the last 36 periods);
2. one-step ahead statistical forecasts for the same 36 periods (each graph shows the effect on sales of four earlier promotions) and a forecast for the future 37th period;
3. a description of the product considered;
4. an indication that the mean promotion sales uplift (base rate) was 50%;
5. additional information that is not incorporated into statistical forecasts such as reasons for the success (or failure) of past and future promotions. This information has no predictive value except in one case situation that is discussed subsequently.

The first three features are already widely incorporated in major forecasting support systems along with product codes. We include two statistical models (i) the baseline model which allows the forecaster to check how the SKU's performance changes when only a flat line forecast is presented and (ii) the promotional model which increases reliability of the model-based information and creates a potentially stronger anchor towards it. In both cases, participants were clearly informed as to which model is presented. This design allows us to explain how forecasters respond to two situations observed in practice: where just a baseline statistical forecast is provided (e.g. SAP APO system) or, as in our case study, where promotional events are included (e.g SAP F&R system).

At the beginning of the experiment, participants were asked their predictions for the average promotional uplift in a retail company selling groceries, which we interpreted as their prior expert estimates. Even if one can argue that students probably do not have any reliable priors for promotional modelling, we use these predictions to identify whether people tend to follow their prior forecasts or switch to the provided anchors.

In the main part of the experiment, the participants were asked to forecast one-step ahead for twelve time series during a promotion planned for the 37th period. There were two trial runs for subjects to familiarise themselves with the system and its structure; they were excluded from the analysis.

Additional information, as in our case study, would be received through various channels, such as formal/informal meetings, emails and phone calls. To reflect that, we simulated different statements that are realistic in a business environment including overly optimistic ones. The domain knowledge provided for each product was of the form:

1. positive/negative reasons and explanations of success/failure for the past four promotions (e.g. "Our spending on this campaign is only 50% of our normal

cost") and/or market research (e.g. "Based on market research, the Marketing Manager concluded there would be a positive reaction to the promotional campaign'). There was a maximum of eight statements for the four promotions in total. All of these statements had no diagnosticity from data assumptions (except in one setting that is discussed later).

2. positive/negative arguments for the upcoming promotion (in the 37th period) related again to the promotional campaign and/or market research. In some experimental treatments, positive promotional information was diagnostic. This effect will be explained shortly.

3. additional positive statements that we call "Hype" which reflected personal feelings and perception (e.g."Sales already told the guys from the top to expect a huge boost to sales following this promotion") and had no predictive value.

A full list of reasons, which contained 22 statements for each "Promotional Campaign" and "Market Research" topics and 15 for "Hype", is available in Appendix 2.A. The first two groups had an equal number of positive/negative reasons. The reasons had been previously validated by experts as an illustration of qualitative information shared in demand forecasting.

All these expressions were displayed randomly for each promotional period, both with regard to the number and order of the presentation. Participants indicated whether any of the displayed qualitative statements, about the upcoming promotional period was regarded as useful for their judgmental decision. No feedback was provided during the experiment, except the overall average Mean Absolute Percentage Error (MAPE) result at the end of the experiment. Finally, participants were asked to complete a short post-experimental questionnaire that is discussed further in Section 2.6.4.

The differences between the case study F&R system (Section 2.4) and our ex-

perimental design are explored in Table 2.3. It shows the main assumptions and simplifications for the experiment.

Table 2.3: Comparison of the designs.

	Case study, F&R system	Our study
Statistical model	Promotional	Baseline & Promotional
Contextual information	Included in the process, but not in the system	Included and presented together in the system
Time buckets	Weekly	Not specified
Duration of promotions	Various	1 period
Average promotional uplift	Not specified	50%
Forecast horizon	From 6 to 34 weeks	1 period ahead
Promotional markers	Under the graph	Colour highlights

2.5.3 Participants and incentives

Participants were students who had completed a "Business Forecasting" module with one lecture about judgmental forecasting. The experiment was introduced as a voluntary exercise for judgmental forecasting with different methods of motivation: individual money tokens for participation, one/two prizes for the best performance and only verbal encouragement. All groups' performances were compared (using non-parametric tests due to different size of groups) prior to conducting the detailed analysis, and no evidence of difference was identified. Interestingly, we found no benefit from monetary incentives in any form, which aligns with some findings from other studies (e.g. Camerer and Hogarth, 1999; Katok, 2018).

2.5.4 Data generating process

The data for the experiment was generated using:

$$\text{Sales}_t = \text{PromoEffect}_t^{c_t}(\alpha\text{Sales}_{t-1} + (1-\alpha)\text{BaseSales}_{t-1})\varepsilon_t. \tag{2.1}$$

A baseline statistical forecast was produced using Simple Exponential Smoothing with a fixed smoothing parameter of 0.2 and initial level of 200 units. The promotional effect was 50% (with 10% variation, as showed in Table 2.4) of the non-promotional historical sales where c_t is 1 for promotional periods and 0 otherwise, and $\varepsilon_t \sim \log \mathcal{N}(0, \sigma^2)$, where σ^2 is variance of the noise. There were two levels of noise: low and high (with values of 0.1 and 0.2, respectively), which implied different variation in the time series. The multiplicative promotional interaction mimicked a pattern typically found in promotional sales data.

In total, there were five promotional periods: four were randomly allocated in the historical sample, a promotional event in each of four equal time periods, while the last one was always on the 37th period where the forecast was needed. Each promotion lasted for only one period without any lag or lead effects.

There were four experimental settings (see Table 2.4). For the last two treatments we used "Enhanced" sales, where the effect of promotions was increased by a further 25% when a positive promotional statement is provided. Thus, "Promotional Campaign" statements were diagnostic and had predictive value for historical sales and the forecast period. Enhanced time series were introduced to investigate how forecasters would react to more extreme special events.

To visualise the difference, one example time series in four experimental setting is provided on Figure 2.4. Historical sales and statistical forecasts are black and red (grey) lines respectively. The differences during some promotional periods (by 25%) in treatments 3 and 4 can be observed, which in these cases reach 625 units compared to 500 as in treatments 1 and 2. In simulating the "Enhanced" setting, we are able to investigate if this increase of promotional effects, making them more obviously salient, influences the overall impact of the soft information.

Building on the results of Fildes et al. (2018), we explore the forecasting process not only as a null experiment with no qualitative information affecting the promotion

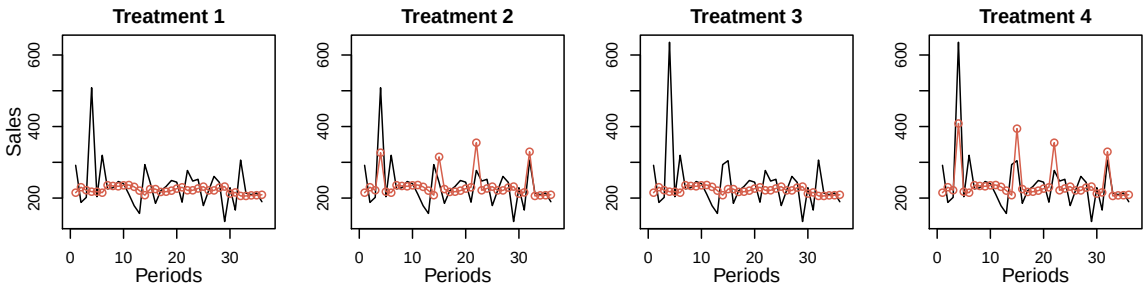

Figure 2.4: Implementation of the experiment: one time series across settings.

but also connecting it to practice by including additional qualitative information with predictive power. Thus, the two settings give a broader perspective for understanding the underlying mechanics of adjustments in the context of promotions. We anticipate different interactions with the information provided in all four treatments, which have not been examined before in the literature. The ideal strategy is: (1) to accept or adjust insignificantly statistical forecasts with promotional effects for treatments 2 and 4 (answering Hypothesis 1 where model-based information is stronger than any contextual statements); (2) to adjust by 50%, on average, in other cases (taking into account the forthcoming promotional period and the additional information available about it, checking Hypothesis 2). Moreover, the adjustments are expected to be higher in the experimental settings with enhanced time series, since past promotional uplifts are more apparent compared with non-enhanced treatments.

2.5.5 Interface and tracked data

The interface design, as well as the type of information provided, was based on the forecasting system of the case company (see Section 2.4 for details), where a software screen provided actual sales, information about promotional type and some special events such as Christmas, Easter, etc. The SAP F&R screenshot (Figure 2.2) inspired our interface, but some features have been simplified and magnified. The result can be seen in Figure 2.5.

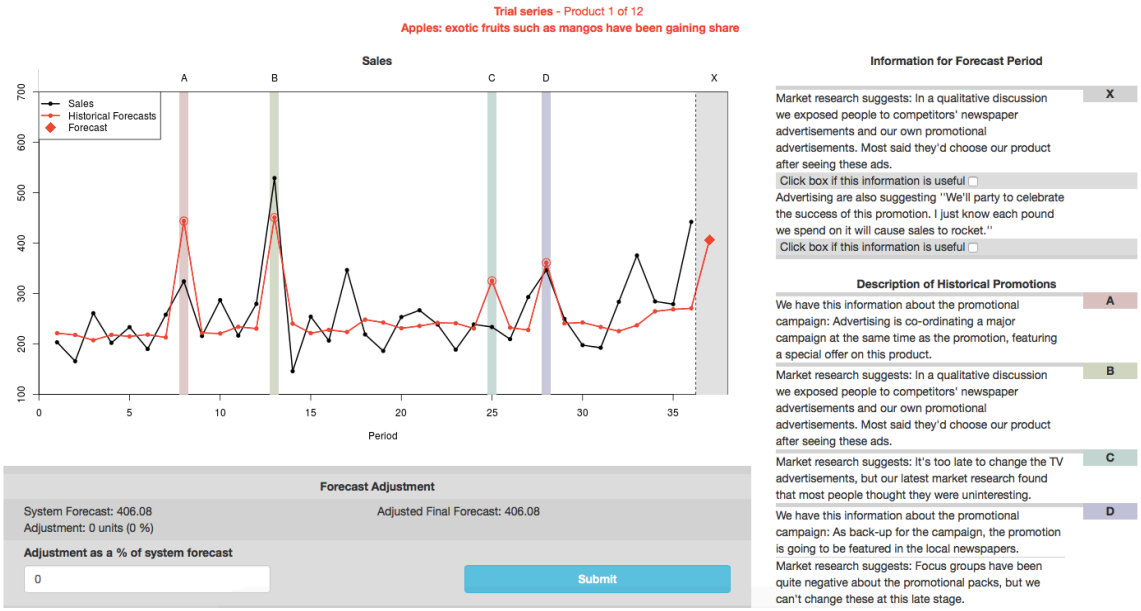

Figure 2.5: Experimental Forecasting Support System. Screenshot example.

The interface is implemented using R (R Core Team, 2017) Shiny Apps environment (Chang et al., 2017) and hosted online at http://www.shinyapps.io. The presentation of the adjustments was interactive: if a participant inputs a number in the adjustment box, the forecast on the graph changes accordingly. The right-hand side of the screen was devoted to the contextual information. Participants were able to indicate whether the piece of information was useful with checkboxes. Evident colour highlights separated different promotional periods on the graph and in a text field, marked by the letters: A, B, C, D and X.

During the experiment we tracked detailed time series information, including the timing of promotions, adjustments and calculated accuracy (MAPE). This data was then transformed to model variables that are defined in Table 2.5 in the next section.

2.6 Analysis and Findings

2.6.1 Sample and descriptive analysis

In total, 128 participants started the experiment with 88 complete runs. As we stated before, this experiment was a voluntary exercise for all participants, hence, we think that the response rate of almost 69% is appropriate for such task.

According to Cook's distance, there were 3 outliers in the data that result in the long tails observed in the distributions, primarily in treatments 2 and 4. These outliers were due to participants who overall had average performance as measured by MAPE. Hence, we retained these, as they were not likely to be associated with any mis-understanding by these individuals.

Descriptive statistics of the treatments are provided in Table 2.4. The table shows the number of participants, the average adjustment size, adjusted cases, various biases and Forecast Value Added (FVA). The participants were assigned to the treatments randomly as they enter the experiment, so that the treatment samples are of the same size.

The average adjustments for the baseline model (treatments 1 and 3) are much less than the anchor of 50%. However, in the case of enhanced time series the adjustments are almost 30% which is 1.8 times higher than for non-enhanced meaning that the promotional uplifts in treatments 3 and 4 are more pronounced and salient graphically. Nonetheless, the average adjustment still falls short of the expected size, especially once adjusted for the enhanced effect. In treatments with the promotional model, the size of the adjustments is minimal, while it is insignificantly different from zero for the the second one (p-values for one-sided t-tests: 0.150 and 0.003, respectively). However, considering the percentage of adjusted cases, these time series were adjusted as frequently as in other treatments.

Table 2.4: Experiment settings and descriptive statistics.

	Treatment			
	1	2	3	4
Time Series	Not Enhanced	Not Enhanced	Enhanced	Enhanced
Statistical forecast	Baseline	Promotional	Baseline	Promotional
Expected Adjustment in population	50%	0%	50%	0%
Expected Adjustment in sample	40%	0%	60%	0%
Participants	17	24	26	21
Mean adjustment size	16.1%	1.4%	29.0%	2.9%
Adjusted cases	81%	70%	82%	83%
Statistical bias (scaled by mean sales)	0.400	-0.024	0.536	-0.035
Adjustment bias (scaled by mean sales)	0.253	-0.041	0.270	-0.062
Directional bias	0.6	0.1	0.8	0.2
FVA	31%	-25%	47%	-8%

The mean error (bias) typically detects systematic bias in a forecast (see Formula 2.4 in 2.B for calculations): when the actual value is consistently greater (less) than the forecast, then the bias is positive (negative). Comparing our statistical and adjustment bias, we see that participants substantially reduce the positive bias for the baseline model, resulting in underforecasting, but increase the negative bias for the promotional model - overforecasting for these cases. The mean directional bias (Formula 2.5) shows a ratio of positive adjustments to negative ones, providing evidence that all treatments have a predominant number of positive adjustments (i.e. positive directional bias).

The FVA figures (that are calculated using the AvgRelMAE, see Formula 2.7 in 2.B) suggest accuracy improvements for the baseline model settings, in contrast to the promotional model. This suggests that participants corrected promotional forecasts due to either not trusting the statistical forecasts or alternatively falsely relying on the additional non-diagnostic contextual information. These results provide evidence

to reject Hypothesis 1. However, FVA for the fourth treatment is negative, but not significantly different from zero (t-test p-value is 0.161). Both third and fourth settings with enhanced time series, where positive promotional statements influence forecasts, perform better based on FVA compared to the non-enhanced.

To further explore participants' adjustment processes and investigate the hypotheses, we built a mixed effects model relating the size of the adjustments with the model-based and soft information provided by the forecasting support system.

2.6.2 Mixed effects models for adjustment size

To account for variability in the time series and participants' performance, we need to include random effects. These random effects help to estimate participant specific intercepts (removing bias) and allow for the changing variance between settings. Also, we explored the inclusion of the effects both as intercepts and slopes for variables. The alternative specifications were assessed using Akaike's Information Criterion (AIC) and the models were fitted using the *lme*4 package (Bates et al., 2015) for R.

Following Davydenko and Fildes (2013), we transformed the adjustments to relative adjustments, using $\log(1 + \text{Adjustment})$ as this leads to a more symmetric distribution for modelling purposes. The new target variable is y_{ijk}, with i, the participant; j, the time series and k, participant group, denoting the first, second and third levels of a multilevel model with two random intercepts for time series and participants' groups respectively.

We also logarithmically transformed all continuous variables. The mixed effects model has this general form with three levels (Goldstein, 2011, p.86):

$$y_{ijk} = \underbrace{\beta_0 + \sum_{a=1}^{n} \beta_a X_{aijk}}_{\text{fixed part}} + \underbrace{v_k + u_{jk}}_{\text{random part}} + \varepsilon_{ijk}, \tag{2.2}$$

where y_{ijk} is a response variable, v_k and u_{jk} are the model's errors from the random

effects; and $\sum_{a=1}^{n} \beta_a X_{aijk}$ are fixed effects for all variables in Table 2.5 where a brief description of the variables is also provided.

Time, order and a number of observations were considered, but were rejected by using the AIC. The same procedure was applied for all "Excluded" variables. Table 2.6 provides the estimated coefficients for the final model. The random effect intercepts for participants and time series show small variance, but according to the AIC, it is useful for the model fit. We proceed with a discussion of the variables associated with past promotions and the current forecast period.

Qualitative information about past promotions

Observe that none of the variables corresponding to additional information for past promotions improved the model fit. Evidently, participants tended to ignore these contextual statements or their effect is too limited. The experiment simulates a realistic business environment such as the case study, where there is an excessive amount of information with unknown diagnosticity, and therefore participants may have been overloaded with information. We will return to this argument in Section 2.7.

Domain knowledge for the forecasted promotional period

Table 2.6 provides the final model with the smallest AIC value together with p-values for each coefficient. All contextual reasons included in the model were those statements that had been ticked/checked as useful by participants. Looking at the model coefficients, we see that in the period of interest the additional information was retained, yet misused. In particular, all positive promotional and marketing statements increased the relative adjustments. In the case of the treatments with promotional model forecasts, we can observe that the promotional, marketing and hype information have either negative or close to zero coefficients, supporting Hypothesis 1 (i.e. the effect of positive reasons checked for treatments 3 and 4 is -0.0592+0.0606).

Table 2.5: List of variables used as fixed effects.

Variables	Description
Included	
Treatment	Factor variable, where Treatment 1 is the baseline
Promo reasons	Promotional reasons useful (checked)
Promo reasons × Treatment 1&3	Interaction variable of the checked "Promotional Campaign" reasons for the upcoming period and Treatments 1&3
Promo reasons × Treatment 3&4	Checked "Promotional Campaign" reasons for enhanced settings of experiment settings 3&4
Hype	"Hype" reasons useful (checked)
Hype × Treatment 1&3	Interaction variable of the checked "Hype" reasons and Treatments 1&3
Promo reasons, positive	Checked positive "Promo" reason
Market reasons, positive	Checked positive "Market" reason
Expert prior	Forecaster's prior estimate for the promotional uplift
Last promotional uplift	Last promotional uplift
Current forecast	Current statistical forecast
Low noise × Treatment 1&3	Low/high noise for Treatments 1&3
Excluded	
Order	Order of time series excluding trial runs
Promo reasons, past events	4 dummies, one for each past promotion
Market reasons, past events	4 dummies, one for each past promotion
Market reasons	Marketing reasons useful (checked)
Marketing reasons × Treatments 1&3	Interaction variable of the checked "Market Research" reasons and Treatments 1&3
Average promo uplift	Average of past promotional uplifts
Last actual	Last actual sales before the forecasting period
Low noise	Dummy for low/high noise
Past reasons	Number of past statements presented (as a set of dummies)
Current reasons	Number of current statements presented (as a set of dummies)

Interaction variables for domain information and the baseline model ("Promo reasons useful × Treatment 1&3" and "Hype useful × Treatment 1&3") show that there is a positive effect on adjustments which is as expected, since the baseline model re-

Table 2.6: Linear mixed effects model output for logarithmic relative adjustment size.

Fixed effects			
	Estimate	Standard Error	P-value
Intercept	-0.0048	0.4303	0.9913
Treatment 2	0.0300	0.0249	0.2296
Treatment 3	0.0695	0.0201	0.0006
Treatment 4	0.0391	0.0299	0.1915
Promotional reasons useful	-0.0592	0.0226	0.0090
Promo reasons useful × Treatment 1&3	0.0449	0.0224	0.0450
Promo reasons useful × Treatment 3&4	0.0606	0.0223	0.0067
Hype useful	0.0046	0.0260	0.8608
Hype useful × Treatment 1&3	0.0822	0.0359	0.0222
Promo reasons useful, positive	0.0712	0.0240	0.0033
Market reasons useful, positive	0.0755	0.0211	0.0004
Low noise × Treatment 1&3	0.0949	0.0240	0.0000
Expert prior	0.0187	0.0064	0.0035
Last promotional uplift	0.1896	0.0820	0.0418
Current forecast	-0.2080	0.0514	0.0000
Random effects			
	St. Dev.		
Participant group (intercept)	0.0096		
Time series (intercept)	0.0565		
Residual	0.1613		

quires corrections. Note that marketing information was excluded from the model. These results partially support Hypothesis 2.

According to the positive statistically significant coefficient for the interaction effect of diagnostic promotional statements in the enhanced treatments (3 and 4), we conclude that participants were able to weight this information correctly. This is an average effect for both settings, and if we split this variable into two independent ones, then the result is higher for treatment 3 (estimate of 0.0987 with st. error of 0.0321) than for the fourth one (0.0254 with st. error of 0.0309), which is as expected due to the different forecasting model settings.

"Hype" information led participants to forecast a higher uplift, while any "Promotional" statements led to a decrease in relative adjustments, on average, across all

treatments. This implies that overly optimistic statement mislead subjects, obfuscating their judgment with regard to diagnostic information.

Summarising these findings with respect to domain knowledge applied to the forecast period, participants were confused by non-diagnostic information (e.g. "Hype") but correctly identified the effect of helpful statements in enhanced treatments (a positive coefficient for promotional information for treatments 3 and 4). We see that (i) all positive contextual statements have a positive effect on adjustments given both statistical forecasting models, but in the case of the promotional model, this effect is not substantial (supporting Hypothesis 1); (ii) there is a strong impact of domain knowledge on adjustments when only the baseline statistical forecast is provided (supporting Hypothesis 2), in particular, we observe positive coefficients for promotional and hype statements; (iii) participants were able to identify diagnosticity of additional promotional reasons for enhanced treatments, even though the size of the coefficient is not substantial (about 6% as an average effect against initial 25% by design). Market reasons with no diagnostic power were misinterpreted and weighted more than the diagnostic promotional reasons.

Other variables

The mixed effects model shows that several other variables are highly significant. Table 2.6 suggests that all variables connected with algorithmic information, such as the current forecast, last promotional uplift and expert priors have stronger effects on relative adjustments compared to any qualitative statements. The last promotional uplift has a positive effect suggesting that it is one of the major anchors for decision makers. This confirms Fildes et al. (2018). However, note that participants ignored any qualitative information about the last historical promotion indicating that it was the size of the past promotional uplift rather than the reasoning behind the observed effect that was important for them.

The significant negative coefficient for the statistical baseline forecast ("Current forecast") implies that participants lowered their adjustments for high forecasts or dampened it more.

The results indicate that there is no evidence of a significant effect of qualitative information (with unknown diagnosticity) for past promotional periods. Instead, participants focused on the current forecasting period, taking into account all major anchors such as last promotional uplift, current statistical forecast and contextual reasons for possible success or failure. In the next section, we explore the effects on forecast accuracy to test Hypothesis 3.

2.6.3 Forecasting accuracy assessment

The improvement in terms of Relative Mean Absolute Errors (RelMAE) has been used to gauge changes in forecast accuracy and investigate whether the previous findings for the relative adjustment sizes as shown in Figure 5 hold unconditionally. The FVA is calculated using Equation 2.7 as measured by MAE, where positive values correspond to improvements in accuracy following interventions.

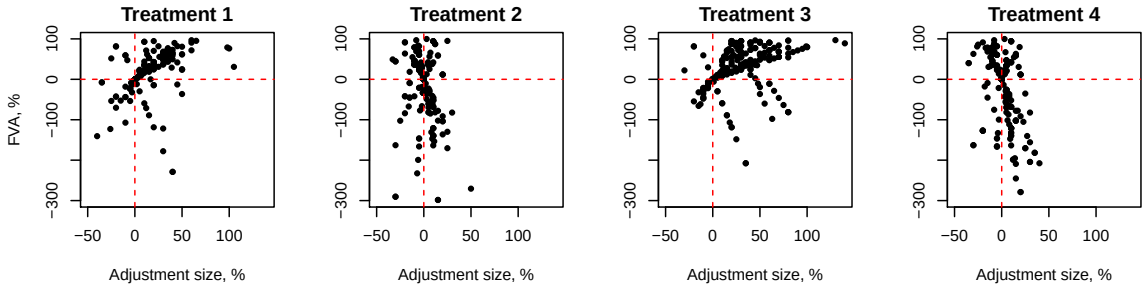

Figure 2.6: Adjustment size impact on accuracy (based on FVA values).

Figure 2.6 provides scatter plots of FVA for all four settings, where there is a clear difference between the baseline and promotional models (pairs of 1&3 and 2&4). On the vertical axis, positive values refer to accuracy improvements, while on the horizontal axis, positive values correspond to positive adjustments. The majority of

adjustments for the first and third settings are located in the top right corner and are positive, as expected. The observed pattern in the scatter plot is similar to the MAE of normalised positive adjustments that was reported by Trapero et al. (2011) in a separate case study. This validates the setup showing that participants adjust as expected (in particular, for the baseline model). Since the optimal decision for the promotional model is to accept or adjust insignificantly, the results for these treatments are indicative of erratic behaviour of participants: many adjustments were too high, which damaged forecast accuracy. Participants adjusted somewhat randomly, reacting to various pieces of non-diagnostic information. Observe that the magnitude of the adjustments is smaller, closer to tweaking, which has been reported in the literature to be damaging (Fildes et al., 2009).

Building a mixed effects model of FVA, with the same structure as in Table 2.5, indicates apparent effects such as the reaction to contextual information for the upcoming promotional period changes overall accuracy asymmetrically: in treatments 1 and 3 promotional statements led to better forecasts, while it displayed the opposite for treatments 2 and 4. Essentially, the value of the contextual information depended on it being additional to that included in the model - a correct appraisal of the two different forecasting models. The reaction to hype information caused a dramatic decrease in accuracy for the second and fourth experiment types, and a smaller negative effect for treatments 1 and 3. A possible explanation for this might be that in the baseline setting participants needed to adjust, so they assigned some weight to this piece of information, but it was still underweighted. We also see that there was no strong linear correlation between adjustment size and FVA. Thus, we can claim that additional qualitative information with unknown diagnosticity is typically misused and reduced final forecast accuracy (supporting Hypothesis 3 from Section 2.2).

2.6.4 Post-experiment questionnaire: user expectations

A post-experiment questionnaire was designed to assess how motivated and reliable participants were. The results in 2.C show that on average participants had medium confidence about their knowledge of sales forecasting. They found the interface (e.g. time series graphs and statistical models) to be useful. They were generally motivated and had experience of purchasing products on promotions. Participants felt that the reasons provided had a direct influence on their forecasts. The overall expectations of making accurate forecasts were 2.9 out of 5 (where 5 is "Very accurate"). In general, the results of the questionnaire showed that participants were motivated and understood the design and interface, although they were conservative in their performance expectations.

2.7 Discussion and conclusions

Forecast adjustment is a core task in demand planning as the case study demonstrated. In our experiment, we asked forecasters to adjust, if needed, for a promotional period under various conditions in order to investigate how they reacted to a mix of quantitative and qualitative information presented on the same screen, the setting that is regularly desirous in organisations, and how accurate their predictions were. These key conditions were: (i) baseline model or promotional model forecasts; (ii) different sizes of promotional effects paired with qualitative information (e.g. enhanced settings); (iii) different sets of positive and negative contextual statements from several sources. Based on the case study, we simulated the company forecaster's reality as closely as possible to understand the mechanics behind adjustments in a promotional context when model-based and contextual information of unknown diagnosticity is available.

One interesting finding is a forecaster's reaction to available information. Apart

from Sanders and Ritzman (1992); Webby et al. (2005); Fildes et al. (2018), there is neither a laboratory experiment nor an empirical study on the use of contextual information in the forecasting process. Our study provides some important insights into the decision-making process for making judgmental adjustments to statistical forecasts and offers answers to our main research question.

First, we found that there was no significant effect of additional qualitative reasons or explanations of success/failure for *past* promotional periods on adjustment size. In a prototype FSS, Lee et al. (2007) succeeded in providing valuable summary information summarising past promotions. Given that we tried to simulate a realistic system, capturing major elements of the S&OP process, but through the FSS, the forecasters might have been experiencing an information overload which led to under-weighing of historical information. In practice, information accrues sequentially, usually via emails or meetings. We postulate that information overload for our experimental forecasters was an issue and therefore further research should investigate it (e.g., methods based on Lee et al. (2007) or decomposition of the available information might well prove valuable here).

Second, the same information for the *upcoming* period showed that strongly positive statements shifted the adjustments upwards and, surprisingly, participants were more sensitive to the over-positive non-diagnostic "Hype" statements rather than the diagnostic promotional information. Event though participants were confused by this non-diagnostic information, yet they were able to correctly identify the effect of helpful statements in Enhanced treatments, where those effects were more pronounced. This is an important result showing that users can extract relevant pieces of information even in such complex conditions, but we need to pre-process the contextual information as much as possible before providing it to forecasters to avoid any evident cognitive biases.

For both baseline and promotional models, we also observed that people tend to

underestimate adjustments for promotional periods, which is unexpected since there is a well-known notion of "over-optimism" bias in human judgment. In forecasting, this bias has been observed in several field studies (e.g. Fildes et al., 2009; Franses and Legerstee, 2011). However, both Fildes et al. (2018) and De Baets and Harvey (2018) observed the same effect. We argue that participants may be somewhat more conservative in their adjustments than practitioners enmeshed in a real company. In fact, this was somewhat indicated in the post-experimental questionnaire.

The case study has played a crucial role in exploring how much expert judgment is used in practice, providing insights on its operations and resultant forecast accuracy. According to the forecasting process observed in the company, even having an automatic multivariate model to forecast promotional periods with many explanatory factors taken into account, demand planners still often adjust based on the additional information available to them outside of the forecasting support system. There are various reasons for adjusting statistical forecasts, such as revisions of promotions, weather and other events, which could lead to the different behaviour of forecasters.

We found that the effect of expert adjustments that were based on contextual information of unknown diagnosticity was positive in a half of the cases, but in the other half, it led to substantial errors and diminished all improvements. This finding is highly dependent on the statistical model implemented and highlights the importance of FSS and its lack of appropriate feedback features.

Additionally, the interface we examined in the case organisation, is overly complex and thus not user-friendly. As a result, the whole forecasting process and evaluation becomes extremely complex and difficult to track. All highlighted issues make a forecaster's job even harder, given the number of time series that they need to be handled efficiently. Surprisingly, the importance of soft information integration and its effective provision (as an interface feature) in forecasting support systems is hardly discussed or considered by many behavioural papers. The key issues can be highlighted: (i) the

reliability of the promotional model; (ii) value added by forecasters; (ii) how experts filter and use the circumstantial sources of information. This opens many directions for further research on expert judgment in Forecasting Support Systems.

The main limitation of this study is that we managed to employ only novice (student) participants, although initially, we aimed to engage our company contacts as a separate group of participants. Therefore, there is an open question about the difference between student participants and demand planners performances, because they are likely to have different levels of involvement and knowledge in such forecasting exercises. Second, we adopted several essential assumptions in the data analysis of the case study, e.g., a transformation of time buckets of sales, promotions and other factors. Third, both case study and our laboratory experiment are very context, data and case specific. Hence, we are confined to such details and are not able to generalise these findings appropriately.

The failure of the experimental forecasters to distinguish between the diagnostic and non-diagnostic pieces of information suggests that there is a need to carefully consider how to design FSS in such way that only diagnostic qualitative information is presented. This means that companies with similar forecasting procedures need to take care in filtering upcoming information more thoroughly, potentially a problem in designing a forecasting process (Oliva and Watson, 2010).

Essentially this study emphasises the role of forecasting models, highlighting how forecaster behaviour varies depending on the model provided. This outcome is new and has substantial implications for existing and future research. While in laboratory settings we can design an optimal diagnostic statistical model for generated time series, there is a question about model uncertainty in practice. Are company forecasters confident in dealing with optimal statistical forecasts? How do experts interpret both model-based and soft information in behavioural operations? These questions open completely new directions in behavioural operations research.

Appendix 2.A Contextual statements

2.A.1 Promotional statements

- Our main competitor has just started a directly competitive promotion and they've put a lot of resources behind it.
- We were hoping for a celebrity endorsement of our product as part of the campaign, but negotiations have not been successful and, unfortunately, we will have to run the campaign without this endorsement.
- It has been decided not to feature the promotion in local newspapers. In retrospect this may be a mistake.
- No advertising is planned beyond the regular TV bursts.
- Sales staff have been in discussions with the supermarkets. Unfortunately there was no agreement display the product prominently during the campaign.
- No in store support is made available to overcome consumer resistance.
- Our spending on this campaign is only 20% of our normal spend.
- Because of a budget constraint we have had to economise so our plan to recruit sales staff to promote the product in stores has been cancelled.
- The campaign is a buy-one-get-one-free promotion. For us these have proved disappointing compared to the normal uplift we would expect.
- Only the largest campaigns have proved effective and this one is under-par.
- Our promotion spend is only half that of our competitors.
- Advertising is co-ordinating a major campaign at the same time as the promotion, featuring a special offer on this product.

- Our campaign will be fronted by an A-list celebrity. In previous campaigns this has proved to be an effective way of significantly boosting sales.
- Sales staff have been in discussions with the supermarkets. The good news is that they have agreed to display the product prominently during the campaign.
- As back-up for the campaign, the promotion is going to be featured in the local newspapers.
- From their contacts, sales staff believe the main competing product will not be promoted during this period.
- The promotion is supported by a complimentary trial offer in store to overcome consumer resistance.
- From their contacts, sales staff believe the main competing product will not be promoted during this period.
- We have spent a considerable amount on recruiting sales representatives who will promote the product in stores.
- The campaign is a buy-one-get-one-free promotion. These always seem to have a substantial impact above the normal uplift.
- These large campaigns have proved to be effective in the past.
- The campaign is a 3 for the price of 2. This has proved very effective for this product category.

2.A.2 Market Research statements

- It's too late to change the TV advertisements, but our latest market research found that most people thought they were uninteresting.

- In a simulated shopping experiment with potential customers, relatively few chose the product in its new promotional pack.
- In a consumer choice experiment that Marketing carried out, consumers unattracted to the promotional offer in comparison to the regular product offer.
- Market research results relating to the promotion campaign have not been encouraging according to the Marketing Manager.
- It's too late to cancel the TV advertisements, but our latest market research found that most people thought they were uninteresting.
- Focus groups have been quite negative about the promotional packs, but we can't change these at this late stage.
- In a qualitative study we exposed people to competitors' newspaper advertisements and our own promotional advertisements. Few people said they'd choose our product after seeing these ads.
- Market research has shown that consumers are not responding to the proposed campaign. Focus groups have been quite negative about the promotional packs, but we can't change these at this late stage.
- In a qualitative discussion we exposed people to competitors' newspaper advertisements and our own promotional advertisements. Few people said they'd choose our product after seeing these ads.
- Focus groups were unimpressed by the benefits of the promotion.
- In a simulated shopping experiment with potential customers, the product proved very attractive in its new promotional pack.
- Our new advertising to support the promotion was very well received when market tested.

- Qualitative research has found that the newspaper advertisements we will run were highly likely to persuade people to choose our product.
- In a consumer choice experiment that Marketing carried out, the promoted product was chosen by almost double the number of potential users compared to the regular product.
- Based on market research, the Marketing Manager suggests there will be a positive reaction to the promotion campaign.
- Our television advertisements for the product were very well received by potential customers according to our market research.
- Focus groups discussing the proposed promotional pack thought the design excellent.
- Based on market research, the Marketing Manager concluded there would be a positive reaction to the promotion campaign.
- Focus groups attracted by the store displays.
- In a qualitative discussion we exposed people to competitors' newspaper advertisements and our own promotional advertisements. Most said they'd choose our product after seeing these ads.
- Focus groups were excited by the benefits of the promotion.

2.A.3 Hype (over-positive) statements

- "With this promotion we just can't go wrong. I can smell success already and it's sweet." This came in from the Marketing Director
- The accounts manager has added "Mark my words. This campaign is sure to be a success."

- "The marketing guys we've got here are unbeatable. They're the best in the business."
- Advertising are also suggesting "We'll party to celebrate the success of this promotion. I just know each pound we spend on it will cause sales to rocket."
- "The timeline has shown we've had meeting after meeting to get this promotion right. Now everyone's on board. We firmly believe we'll meet our target of at least 80% lift and get way beyond it."
- The team just knows this promotion is going to give our sales a huge uplift.
- "Our team all left the promotion planning meeting on a high. Everything about the promotion feels great. Let's look forward to a great boost in sales."
- "Sales' gut feel says that we are going to be toasting success when we look back at the sales generated by this promotion."
- "This promotion is going to be historic! In the early planning meeting the belief has been these sales would rocket."
- "Our view in planning is this campaign is it will knock out the competition. It's just got that feel of being a winner."
- The MD has said he's certain we'll soon be celebrating this campaign.
- "Sales already told the guys at the top to expect a huge boost to sales following this promotion."
- "The planning team's view is this promotion campaign has got success written all over it."
- "Logistics need to know. We should tell the transport people they'll soon be very busy. This campaign looks brilliant"

- "Sales has no doubts at all. We've got a winning promotion formula here."

Appendix 2.B Error measures

The Mean Absolute Percentage Error (MAPE) is:

$$\text{MAPE} = \frac{1}{n}\sum_{j=1}^{n}\frac{|Actual_t - Forecast_t|}{Actual_t}. \tag{2.3}$$

MAPE is appropriate only when there is a meaningful zero for $Actual_t$. Naturally the denominator should be non-negative.

Bias is:

$$\text{Bias} = \frac{1}{n}\sum_{j=1}^{n}\left(Actual_t - Forecast_t\right), \tag{2.4}$$

where $Forecast_t$ is either a pure system forecast or adjusted by a forecaster.

The Mean Directional Bias (MDB) is equal to:

$$\text{MDB} = \frac{(n_{pos} - n_{neg})}{n}, \tag{2.5}$$

where n_{pos} corresponds to the number of positive errors and n_{neg} corresponds to the number of negative errors. It varies from -1 to 1, resulting in zero only when the number of positive errors in the sample is equal to the number of negative errors. MDB is scale independent and robust to the size of errors.

We use the Average Relative Mean Absolute Error (AvgRelMAE: Davydenko and Fildes, 2013) on non-zero forecast periods (in order to reduce overall bias towards zero forecasts) to calculate the performance of expert adjustments. AvgRelMAE is calculated as:

$$\text{AvgRelMAE} = \left(\prod_{j=1}^{k}\frac{\text{MAE}_{adj}}{\text{MAE}_{stat}}\right)^{1/k} \tag{2.6}$$

where MAE_{stat} is the Mean Absolute Error (MAE) calculated for the reconstructed

statistical forecasts over all periods, MAE_{adj} is MAE for the final forecasts, k is a number of time series. Any values below one suggest that adjustments decrease the forecast error. Alternatively, we can calculate the forecast value added (FVA) due to the adjustments:

$$\text{FVA} = (1 - \text{AvgRelMAE}) \cdot 100\% \tag{2.7}$$

In this case, negative values suggest that adjustments destroy value.

Appendix 2.C Post-questionnaire responses

Table 2.7 provides descriptive statistic of the responses, ordered from none/low (1) to high (5).

Table 2.7: Mean, median, and standard deviation of post-questionnaire responses.

Question	Mean	Median	Std. Dev
Overall knowledge of sales forecasting	2.625	3	0.821
The time-series graphs' usefulness	3.920	4	1.106
Expected accuracy of adjusted forecasts	2.898	3	0.858
The statistical forecasts' usefulness	3.693	4	0.998
Expectation for the accuracy of the statistical forecasts	2.989	3	0.780
The provided reasons were very easy to understand and use	3.239	3	1.028
The reasons provided had a direct influence on my forecasts	3.466	4	1.005
The occasional highly positive messages had a direct influence on my forecasts	3.136	3	1.063
Motivation to estimate the promotional effects accurately	3.591	4	0.918
Overall customer experience with promotional products in supermarkets	3.693	4	1.054

Chapter 3

Managing cognitive load in Forecasting Support Systems

Abstract

In generating forecasts, taking into account both model-based (quantitative) and contextual (qualitative) information can benefit a decision maker. While analysts use both information sources, most Forecasting Support Systems (FSSs) do not record or structure qualitative statements. This confounds the efforts of analysts, and the forecasting process becomes challenging to track and evaluate. We focus on how forecasters use textual information to adjust statistical sales forecasts. We investigate whether structuring this information by decomposition can assist forecasters in identifying predictively useful contextual information and weight it appropriately. By using laboratory experiments designed to match key features of a real FSS in companies, we found that decomposition helps to determine relevant information for both baseline (sub-optimal) and promotional (optimal) forecasting models. We argue that this is due to reducing this task's cognitive load by guiding people to weight these pieces of information sequentially. We also observed that decomposition of qualitative information reduces judgmental adjustments in all treatments as a natural restrictive feature which can be a useful design tool in many support systems.

3.1 Introduction

Human judgment has an important role in forecasting and demand planning. Algorithmic approaches are increasingly prominent in forecasting, due to rapid developments in different statistical and machine learning methods, as well as increased availability of data. Nonetheless, humans remain an integral part of the forecasting process, both by supervising the process and by embedding contextual information. Given various cognitive biases and heuristics, unclear organisational practices, imperfect information sharing, insufficient statistical knowledge, and sub-optimal decision support systems, the expert's job is not trivial.

The most common way of applying judgment in forecasting is through the adjustment of statistical forecasts (Sanders and Manrodt, 2003; Fildes and Goodwin, 2007a; Weller and Crone, 2012; Fildes and Petropoulos, 2015). These aim to account for the rapidly changing business environment and new timely qualitative information that is not taken into account in statistical modelling. The main trends and findings of this line of research are discussed in the recent literature reviews by Perera et al. (2019) and Arvan et al. (2019), highlighting some conditions where these adjustments can be beneficial (e.g., Fildes et al., 2009; Trapero et al., 2013). Even though there is a broad understanding of how adjustments are implemented, there are no studies analysing the process of applying adjustments nor examining the underlying reasons and their efficiency in practice.

Many existing Forecasting Support Systems (FSSs), a special type of Decision Support System (DSS), lack interface features that can support manual corrections (Fildes et al., 2006), track adjustment reasons, evaluate final forecasts and give appropriate feedback (see last forecasting software surveys by Fildes et al., 2018, 2020). In the meantime, a user is potentially faced with an overwhelming amount of information, generated from the FSS and the demand planning review required before the

forecasts are signed off. In fact, when using sophisticated systems, a user has to (1) assess the quality of system inputs and outputs, and (2) weight other external factors (e.g., contextual information) to be able to include their effects. The complexity of this task arise not only due to the uncertainties on all levels of this task, but because the notion of optimality can often be elusive in a demand planning setting (Kourentzes et al., 2020). Ideally, we would like to have (i) a model, that follows the structure of the underlying data generating process (DGP), and (ii) only relevant additional external information that has predictive value and could be used to elicit accurate decisions based on it. However, this is never the case in practice (e.g., see a case study analysis in Chapter 2.4). Hence, these factors contribute to the demand planning expert having to consider multiple complex factors. One of the possible ways to reduce cognitive load is to rely on decomposition of the information, which is a strategy to split the task into smaller parts or sub-problems. There are no known studies investigating its application to qualitative information in judgmental forecasting.

In this chapter, we aim to analyse the effect of the decomposed presentation of contextual information on expert adjustments comparing to a control group where both model-based and domain data are presented simultaneously. The control group is motivated by a case study that was described in Chapter 2.4. The most obvious setting for such a study is promotional planning, since it is the primary reason to adjust system forecasts according to a survey conducted by Fildes and Goodwin (2007a).

We define contextual information (or domain knowledge) as any qualitative statements available to the expert, typically exchanged via telephone calls, meetings or emails and not included in statistical modelling. Since not all information transmitted in the company is reliable, we consider both *diagnostic* (i.e., helpful and useful for making a judgment) and *non-diagnostic* information (essentially noise). This dimension of complexity provides a better approximation of real organisational processes and is able to reveal whether humans can sift the relevant from the irrelevant infor-

mation.

The second dimension of complexity is connected with the system predictions that forecasters have at their disposal: (i) baseline forecasts, which are not optimal for time series with promotional or other events and (ii) promotional models that capture these promotional effects. These statistical forecasts provide different levels of support to make accurate decisions about future sales.

We use a laboratory experiment, where the forecasting process is fully defined, and controlled conditions are prescribed. The implications of this study are cross-disciplinary being highly relevant (i) to practice in terms of interface design features to support human judgment in forecasting support systems (Lee et al., 2018); and (ii) to decision-making theory where both the available information and its presentation matters.

This chapter is organised as follows: Section 3.2 of this chapter provides an overview of the literature on cognitive issues in forecasting and decision support systems together with our hypotheses. Section 3.3 introduces the design of the experiment, and Section 3.4 provides the analysis of the experimental results. Section 3.5 discusses the results and the main conclusions for this chapter are in Section 3.6.

3.2 Literature review and hypotheses

While there has been a rapid development of computer decision support systems with sophisticated statistical algorithms, an essential element of the process has been ignored: the role of the decision maker. Starting from selecting model-based algorithms to delivering the final decision, human input is still crucial even when considering the rise of machine learning and artificial intelligence. The development of Behavioural Operations Research (BOR) is motivated by behavioural issues that arise in various OR problems. In particular, BOR aims to explore the behaviour of managers, an-

alysts, and customers with regards to operational decisions and outcomes (see the latest overview by Donohue et al., 2020). With some parallels, Decision Support Systems (DSS) research is dedicated to providing managers with tools that can support and improve their decision-making process. Our study combines both fields aiming to understand how the demand planners use their judgment in forecasting and to assist them in this process through a Forecasting Support System.

In the "big data" era, extensive data analysis can guide business processes, given that it is supported by technologies, technical skills and well-organised information sharing (Fosso Wamba et al., 2015; Wang et al., 2016). One of the natural implications of the impact of big data is a considerable interest in Machine Learning (ML) and various Artificial Intellegence (AI) techniques which are able to take advantage of vast volumes of data (e.g., see case studies by Kraus et al., 2020). Even though these sophisticated techniques can be promising for businesses, the major challenge of model interpretation and transparency puts a barrier to the adoption of this technology (Ransbotham et al., 2017). The complex nature of these methods, together with the lack of technical and statistical skills in firms often limits their implementation. We observe the same issue with statistical models, which, even though more transparent, remain challengingly complex. This amplifies the role of humans in these processes. As a consequence, we have to consider many cognitive restrictions that can prevent the efficient use of technologies and data. In the following subsection, we introduce the relevant cognitive problems that arise during the demand planning process.

3.2.1 Cognitive restrictions

Many cognitive biases and inefficiencies are known to affect the decision-making process and perception of qualitative information in particular. Humans have only limited cognitive resources to weight information effectively, especially in analytical tasks. A considerable amount of literature has been published on judgment; the most promi-

nent are Tversky and Kahneman (1974) and Gigerenzer (1999), which explore many biases and heuristics of the human mind. According to Tversky and Kahneman (1974), humans have propensities (i) to anchor to our own experience or beliefs; (ii) to rely on similarity to the specific evidence with unknown validity ("Representativeness bias") and (iii) to weight available knowledge more ("Availability bias"). In forecasting, we have evidence of all of these (Harvey, 2007) together with a sense of ownership, overconfidence (Bovi, 2009; Eroglu and Croxton, 2010; Libby and Rennekamp, 2012), and organisational and political biases (Oliva and Watson, 2009, 2010; Pennings et al., 2019). However, there are no studies analysing the influence of Forecasting Support Systems design on the decision-making process and its impact on cognitive limitations.

The simultaneous incorporation of both qualitative and quantitative information is common in many types of forecasting including weather, macroeconomic forecasting and the context studied here, demand planning (Franses, 2014). This combination is typically handled by demand planners outside of the systems available to them, which means that there is no opportunity to track the effectiveness and accuracy of these decisions. This also limits the potential for training and improving their adjustment skills organically while performing their forecasting task. Their intervention typically amounts to a series of judgmental adjustments of the model outputs (Fildes and Goodwin, 2007a). This issue is crucial since human adjustments are frequently used according to many surveys of practitioners (Sanders and Manrodt, 2003; Fildes and Goodwin, 2007a; Weller and Crone, 2012; Fildes and Petropoulos, 2015). In order to be able to track, evaluate and improve decisions, we need to include both qualitative and quantitative information in the FSS. However, this could potentially introduce an information overload and dramatically increase the task complexity. In educational and learning psychology, this phenomenon is called cognitive load and explained in Cognitive Load Theory (CLT), which provides an understanding of human cognition based on the working memory principle to generate instructional procedures (Sweller,

2005).

According to the literature on CLT, there are two main ways to reduce cognitive load: (i) increasing the working memory capacity by using both visual and verbal information; or (ii) structuring knowledge into schemas to store it in the long-term memory (Cook, 2006). While the former is not practical since it would be difficult to implement it in the forecasting support systems (e.g., narrative instructions for forecasts), the latter helps to manage information by splitting it into chunks that are meant to reduce cognitive load and eliminate the need for simultaneous processing of all elements. According to Cook (2006), this is especially helpful when these elements are highly correlated.

In the forecasting literature, there is a similar notion of decomposition which is defined as "the strategy of breaking a problem into subproblems, solving them, and then combining the solutions to the subproblems to get the overall solution" (Armstrong et al., 1975; Edmundson, 1990). Goodwin and Wright (1993) highlighted several forms of decomposition, where the main difference was how the task is presented to a user. Webby et al. (2005) reported that the presentation of both time series and non-time series information selectively and sequentially in a FSS was found to improve accuracy, except in a case of high information load. Fildes et al. (2019a) analysed the use of a combination of qualitative information, graphical time-series history and statistical forecasts simulated in an FSS. All the available data was presented on the same screen. The participants mis-weighted the provided information and tended to underestimate the adjustments, which highlights the problem of information overload and the high complexity of such tasks. According to Harvey et al. (2000), people are better at assessing the quality of additional information than at using it, since taking all available information into account at once imposes a heavier cognitive load than assessing its relevance step by step. These findings suggest filtering of available cues to identify and present this information structurally, and leads to the main research

question in this study:

Research question: *Are there benefits from a visually structured presentation of contextual information on the accuracy of judgmental adjustments to statistical forecasts?*

Subsequently, our first hypothesis connects cognitive load and ability to assess information sequentially:

Hypothesis 1: *Decomposition of contextual information reduces cognitive load and, as a result, helps users to identify relevant statements more effectively than a non-structured presentation.*

3.2.2 Information quality assessment

Beyond the information overload that users may experience, the complexity of the forecasting task is also directly linked to the quality of the information provided. There is a question of information quality assessment, which is a crucial cross-disciplinary question on its own (see the extensive review and several books on data quality by Redman and Godfrey, 1997; Redman, 2001; Batini et al., 2009; Batini and Scannapieco, 2016).

According to this literature, there are several dimensions of information quality, such as relevancy, value-added, quantity, reliability, accessibility, and reputation of the data (Hazen et al., 2014, p.73). Some of them are defined by the attributes that are native to the data (intrinsic) while others depended "on the context in which the data are observed or used" (contextual). The typical methodology for measuring information quality remains self-reported user questionnaires, which can introduce many biases depending on the protocols. We use both regression and questionnaire analysis to assess the perceived quality of contextual information through the weights assigned in the experimental forecasting tasks due to its perceived quality. It is

important to note here that allocating different weights to information is cognitively much more demanding than identifying relevant cues (that is a binary classification, referring to the first hypothesis). This is supported by the link to cognitive load, implying that the amount of cognitive resources available is positively correlated to "the likelihood that a particular individual will undertake the additional cognitive effort to systematically process the newly received information, as opposed to relying on heuristics" (Watts et al., 2009). This leads to us to one of the arguments for decomposition of contextual information, ultimately aiming to persuade humans to evaluate any information systematically.

In forecasting, we see some evidence that humans can assess time-series and non-time series information correctly. In the case of time-series information, Petropoulos et al. (2018) showed that judgmental selection of a forecasting model visually could perform similarly to an automatic algorithmic counterpart, which means that people are able to identify the best model quite frequently using only graphical time-series presentation. Regarding non-time series information assessment, a post-processing analysis of company data and their judgmental adjustments showed that human judgment could supplement statistical methods well (Trapero et al., 2013). The results of Fildes et al. (2019a) and Chapter 2.4 suggest that humans tend to focus on qualitative statements as much as on the time-series information, but they are not proficient in identifying and weighting relevant ones due to the information overload. In this study, we assume that the decomposition reduces cognitive load, hence humans will be motivated to systematically process any given information, which would lead to more appropriate weighting. Therefore, we can hypothesise the following:

Hypothesis 2: *Decomposition of compound information helps a forecaster weight (i) diagnostic or (ii) non-diagnostic qualitative information appropriately.*

3.2.3 Acceptance of statistical model recommendations

Another important feature of any support system is the level of statistical support provided in a DSS. It is well accepted that we need to provide model-based support to decision-makers, but to what extent they follow these recommendations is still an open question. There are multiple dimensions of interest for research, such as statistical modelling adoption, transparency, acceptance and trust in a DSS (e.g., Shibl et al., 2013). For instance, Önkal et al. (2009) showed that people tend to take advice from human experts a bit more willingly than from statistical methods. Moreover, there is the popular notion of "algorithm aversion" which generally suggests that humans are resistant to use model recommendations provided by DSS, especially after seeing these models err (Dietvorst et al., 2015). This phenomenon could appear due to several reasons, such as lack of trust and credibility of the system (Prahl and Van Swol, 2017), over-restrictiveness of the systems (Dietvorst et al., 2018), or a preference of human recommendations over the algorithms (Yeomans et al., 2019). In fact, Dietvorst et al. (2018) found that giving some degree of control over the final decisions (allowing adjustments and corrections) increased the subsequent use of model forecasts, and even a limited control might be enough to satisfy users and motivate them to trust even imperfect algorithms. In our study, we are not aiming to dive into the question of acceptance of model over human recommendations, rather we want to contrast the use of qualitative and quantitative information depending on the level of DSS algorithmic support. In this case, we contrast the baseline and promotional models to explore the effect of statistical support provided in the software system on the weighting of qualitative information. Hence, we arrive at the following hypothesis:

Hypothesis 3: *Decomposition increases acceptance of promotional model forecasts.*

Since a baseline model requires judgmental adjustments to incorporate the missing information, we expect a positive effect of the decomposition in this case. On the other

hand, where the promotional model provides a good fit already, a human may tend to accept these results or adjust insignificantly. Hence, the case of the promotional model may lead to an imperceptible effect, where positive and negative changes will balance each other out.

Using decomposition, we aim at reducing cognitive load and subsequently simplifying identification of relevant cues, and ultimately, making weighting of all available information more effective. Changes in weighting will affect the accuracy of the final forecast.

3.3 Experimental study

3.3.1 Design

We have conducted a behavioural experiment simulating a FSS for producing forecasts in a retailing context with two by two between-subjects factorial design. One dimension of this design is related to the use of information decomposition when making judgmental adjustments to retail forecasts. This allows the analysis of whether it is possible to reduce cognitive load during this decision making process. We included both time-series (e.g., sales history, statistical model fit and forecasts) and non-time series data (e.g., contextual information about promotions or other events) with the focus being on the handling of the latter. The second dimension of the design is concerned with the statistical modelling element in the process. Since judgmental adjustments depend on the quality of the statistical model (De Baets and Harvey, 2020), we included two types: a baseline model (a simple extrapolation method for the data without information about promotional periods) and a promotional model which integrates promotional periods and is optimal with regards to the promotional effects. Naturally, the behaviour of forecasters should be different between the two setups: for the baseline case, the adjustment should take into account not only additional

information but also the promotional effect itself, while for a promotional model, this effect is already incorporated and the judgment should be based only on the contextual information presented. The decomposition is expected to help users to weight this additional information more efficiently than a non-structured presentation.

3.3.2 Procedure

The main task was to forecast grocery products for a local supermarket by taking into account all available information. The experiment started with a question "Given your groceries shopping experience, what would you estimate is the size of the typical promotional uplift for supermarket products (as a % of sales above the sales without promotions)?". We considered the responses as a prior estimate of a promotional uplift as one of the explanatory variables in our analysis.

Each participant was provided with an interface of a Forecasting Support System. The interface was implemented using R (R Core Team, 2017) Shiny Apps environment (Chang et al., 2017) and hosted on http://www.shinyapps.io and was available online using a web browser. There were ten different products to forecast (e.g., shampoo, apples, vitamin C, bottled water). The first two products were used as trials, so the participants could familiarise themselves with the interface (these are excluded from the analysis). For the trial runs we included performance feedback as a learning tool for participants which has proved to be an effective tool for improving accuracy (Petropoulos et al., 2017) and which has rarely been implemented in the currently available forecasting software (Fildes et al., 2020).

The screen of the FSS was divided into two parts. The left half was designated to present time series data graphically, and the right half was for the contextual information (see Figures 3.1 and 3.2). On the left side, we provided sales, historical predictions and the forecast graphically and a box for adjustments. The adjustment as a per cent of the system forecast was a manipulated (dependent) variable, a change

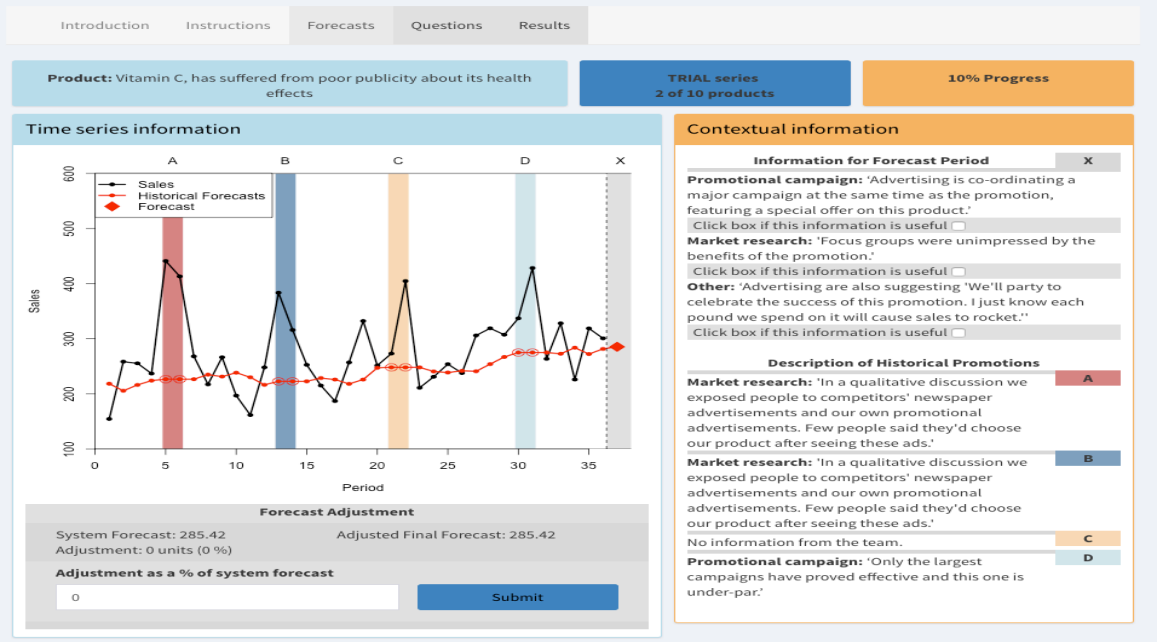

Figure 3.1: Screenshot of a control design (with a baseline model).

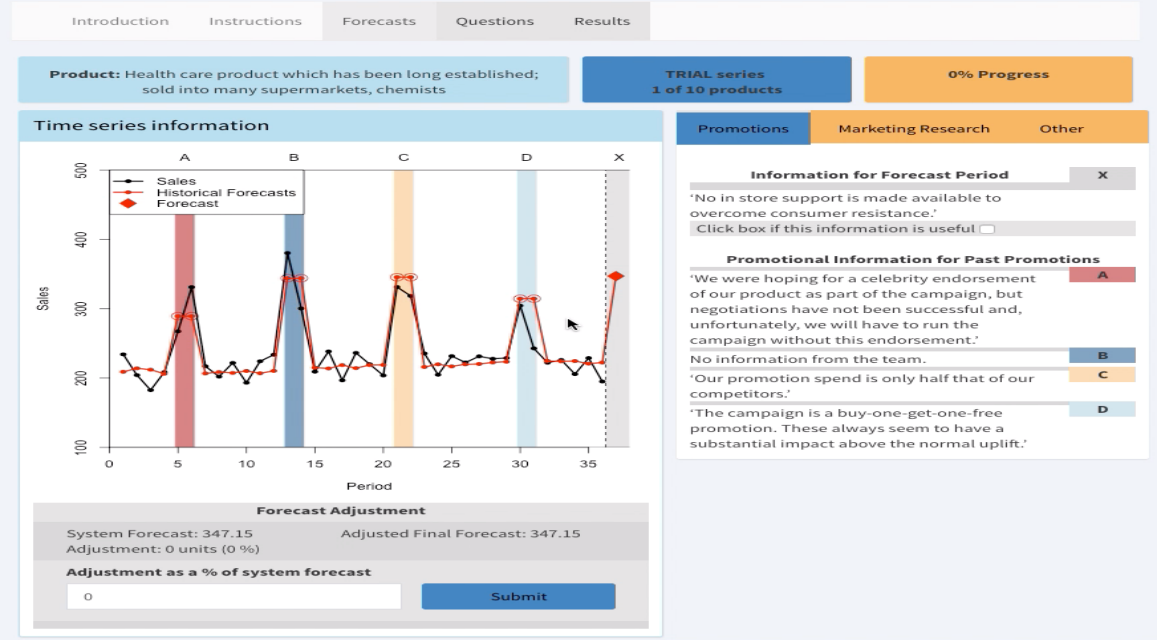

Figure 3.2: Screenshot of a decomposition design (with a promotional model).

of which initiated a recalculation of the system along with a shift in the final forecast on the graph. This feature introduced interactivity between users and the system, allowing them to assess different values and to find the most suitable visually.

The right side was designated for contextual information. Figures 3.1 and 3.2 display the difference between structured and unstructured presentation of the qualitative statements (which are drawn randomly out of the statement pool). In particular, in the first picture there is a full block of text on the right-hand side of the screen ("No decomposition" treatment), while in the second, this block is split into three categories: "Promotions", "Marketing Research" and "Other" (we will call it "Decomposition" treatment), and the participants were needed to click through these buttons in order to access all contextual information. This provided a clear structure of the information provided, where the presentation order of these categories is fixed. In both cases, there were checkboxes for the participant to indicate the contextual statements that were useful. The promotion type was unknown, but each lasted for two time periods and was marked clearly with colour highlights and the letters: A, B, C, D and X.

During the experiment, we tracked detailed time-series information, including the timing of promotions, adjustments and calculated the forecast accuracy (MAPE). After the main part of the experiment, a short closed-ended questionnaire followed, focusing on the participants experience and expectations. This is detailed in Section 3.4.3.

3.3.3 Data

We provided the participants with the following information:

1. a graph of historical sales (36 time periods);

2. a line of statistical forecasts (36 points of history and 1 point for the upcoming

period);

3. a description of the product;

4. a note that average promotion sales uplift (base rate) is 50%;

5. additional qualitative information including a general understanding of each product, important events and possible problems for the forecast.

The graphical information is commonly used in many current forecasting support systems in either graphical or tabular formats, while the qualitative information is typically handled outside the system based on meetings, emails and phone discussions.

Time series

We used the same data generating process as in Chapter 2, where the time series are generated as:

$$\text{Sales}_t = \text{PromoEffect}_t^{c_t}(\alpha\text{Sales}_{t-1} + (1-\alpha)\text{BaseSales}_{t-1})\varepsilon_t. \tag{3.1}$$

A Simple Exponential Smoothing model with a fixed smoothing parameter of 0.2 gave the baseline statistical forecast. The initial level is set to 200 units. The promotional effect was on average 50% (with 10% variation) of the non-promotional historical sales, where c_t is 1 for promotional periods and 0 otherwise, and $\varepsilon_t \sim \log\mathcal{N}(0, \sigma^2)$, where σ^2 is the variance of the noise ε_t. We varied a noise parameter to have different levels: low and high variance (with values of 0.1 and 0.2, respectively). We opted to use 2-period promotions, which was a generalisation of case studies (Sroginis et al., 2019; Trapero et al., 2013, where retailers typically run a half month promotions).

In total, there were five promotional periods: four (two data points each) were randomly allocated in the historical sample, while the last one was always on the last

37th period. For simplification purposes, these promotions did not have any lag or lead effects.

Contextual information

In the experiments, we included contextual statements that are realistic in a business environment, including overly optimistic ones. We categorise the domain knowledge into several groups:

- information about past promotions (a maximum of eight statements for the four promotions in total):
  - positive (negative) reasons and explanations of success (failure) for the past four promotions (e.g., "Our spending on this campaign is only 50% of our normal cost");
  - market research (e.g., "Based on market research, the Marketing Manager concluded there would be a positive reaction to the promotional campaign').
- information about the upcoming promotion (in the 37th period):
  - positive (negative) statements about the promotional campaign, which are diagnostic and directly influence sales (on average increasing (decreasing) sales by 20 units, around 10% of average sales level);
  - positive (negative) statements based on the market research (with no predictive value);
  - overly positive statements that we call "Other" which reflects personal feelings and perception (e.g.,"Sales already told the guys from the top to expect a huge boost to sales following this promotion") and has no predictive value.

These statements were displayed randomly for each promotional period with a static note "No information is provided by the team" when no qualitative information is provided. A full list of reasons with 22 statements for each "Promotional Campaign" and "Market Research" topics and 15 for "Other" is available upon request from authors.

3.3.4 Participants

We recruited 197 participants, including 19 students from a UK university. The student group was comprised of undergraduate students who attended a Business Forecasting module during their program. At the same time, the rest of the participants were solicited via a crowd-sourcing platform for research experiments "Prolific Academic", https://www.prolific.co/. Subjects were randomly assigned to one of the four treatments. Students participation was completely voluntary and was a part of an additional exercise for the course. The Prolific system has a "fair pay" policy, so each participant was paid a fixed fee of two pounds for a twenty-minutes session.

Both groups of participants were checked for outliers in terms of time completion and performance metrics. Nine participants were omitted from the sample due to the rapid completion time that could be either a problem of understanding the task or lack of involvement. Both groups were compared in terms of compatibility of distributions and key metrics. No significant difference was found. In total the sample has 188 observations.

3.4 Results

3.4.1 Descriptive analysis

Table 3.1 shows the descriptive statistics for each treatment. Each participant finished ten time series, and only eight were considered for the analysis since the first two

were explicitly marked as trial runs. Hence, all averages were obtained across all eight time series and participants in the treatments. We can see that the average adjustment sizes (in the fifth row) are significantly different (all t-test $p < 0.000$) from the expected values (in the fourth row). They fall dramatically lower for the baseline cases, where the system forecasts do not account for the promotional periods. For this case, surprisingly, the decomposition decreases the average size of adjustments even further (standard deviations are 33.6% and 25.6% respectively). We will return to this observation in the regression and discussion sections. For the promotional model, the average adjustments are positive, meaning that participants were correcting the forecasts even though they were optimal, a fact that was explicitly stated in the instructions.

Table 3.1: Experiment settings and descriptive statistics.

	Treatment			
	1	2	3	4
Statistical forecast	Baseline	Promotional	Baseline	Promotional
Decomposition	No	No	Yes	Yes
Participants	47	47	52	42
Expected adjustment[a]	50% (53%)	0% (1%)	50% (49%)	0% (-1%)
Observed adjustment	31.9%	10.6%	21.7%	2.9%
Adjusted cases	95%	90%	88%	83%
Correctly identified cues[b]	66%	62%	53%	56%
Incorrectly identified cues[c]	56%	48%	34%	38%
Biases				
Statistical[d]	0.34	0.01	0.33	-0.01
Adjusted[d]	0.13	-0.09	0.18	-0.06
FVA	31%	-95%	26%	-56%

[a] Estimated in population and in sample (in brackets) using the same level of 200 units, the variation is explained by added noise.
[b] Based on clicked promotional statements.
[c] Based on clicked marketing and other statements.
[d] Scaled by average sales history.

The proportion of adjusted cases changes with regards to the decomposition treatments only between treatments 1 and 3 (t-test: $p = 0.026$), which indicates that there is an effect only for the baseline model, while there is no significant difference for the promotional models cases (t-test: $p = 0.119$).

Since there were checkboxes for the participant to indicate whether the contextual information was useful, we calculated a percentage of correctly (based on promotional statements only) and incorrectly (counting both "Marketing research" and "Other" statements) identified cues. The notion of usefulness was not defined in the experiment, so we use these values as an approximation for user attention or interest in particular statements. The results in Table 3.1 show that promotional contextual information was identified by users as useful in more than a half of cases, from 53% to 66% on average across treatments. However, the difference between decomposition and non-decomposition treatments is not statistically significant at 5% significance level ($p = 0.062$ for treatments 1 and 3 and $p = 0.710$ for treatments 2 and 4). At the same time, the impact of decomposition is significant for the incorrectly identified cues (both $p < 0.005$), and the average impact is higher for the baseline model in comparison to the promotional one. This could indicate that the decomposition may be helping people to be more careful and thoughtful with regards to any statements, especially "Marketing research" and "Other" statements. This provides some evidence to support the first hypothesis.

In order to understand whether participants adjusted correctly, we analyse the bias of the statistical and adjusted forecasts. The bias is calculated as:

$$\text{Bias} = \frac{1}{n} \sum_{j=1}^{n} \left(Actual_t - Forecast_t \right), \tag{3.2}$$

where $Forecast_t$ is either the model forecast or adjusted by a forecaster. The bias should be close to zero. It has a positive value, when the actual value is on average

greater than the forecast and vice versa for negative values. This means that the sales value is underforecasted or overforecasted respectively. The adjustments for the baseline model decrease bias by more than 62% for the non-decomposition case, and by 54% for the decomposition one (in percentage point terms 21 and 15). The opposite is true for the promotional model, where the decomposition does not aggravate the problem of overforecasting as much as in non-decomposition case. These descriptive results indicate that there is an impact of decomposition on forecasts, and arguably, on the cognitive load in this decision-making process.

Accuracy is one of the key measures in forecasting that allows to compare different models and processes. In this case, we calculate the Forecast Value Added (FVA), Formula 3.4, of the adjustments based on the Average Relative Mean Absolute Error (AvgRelMAE: Davydenko and Fildes, 2013) of non-zero forecast periods in order to reduce overall bias towards zero forecasts. AvgRelMAE is calculated as:

$$\text{AvgRelMAE} = \left(\prod_{j=1}^{k} \frac{\text{MAE}_{adj}}{\text{MAE}_{stat}} \right)^{1/k}, \tag{3.3}$$

where MAE_{stat} is the Mean Absolute Error (MAE) calculated for the statistical forecasts over all periods, MAE_{adj} is MAE for the adjusted forecasts, k is the number of time series. Any values below one suggest that adjustments decrease the forecast error.

If FVA is positive, the adjustments improve overall accuracy of the forecasts:

$$\text{FVA} = (1 - \text{AvgRelMAE}) \cdot 100\%. \tag{3.4}$$

The last row of Table 3.1 displays FVA for each treatment. The difference between averages of value added in Treatment 2 and 4 is striking, yet not statistically significant for means ($p = 0.251$ and $p = 0.139$ for treatments 1&3 and 2&4 respectively). The high variance between treatments (Figure 3.3) highlights some inconsistency in

decision making between participants, but reveals the effect of the decomposition scenario (a F-test on the difference between the variances of FVA show a significant difference between treatments 1&3 and 2&4, $p = 0.006$ and $p = 0.000$, respectively).

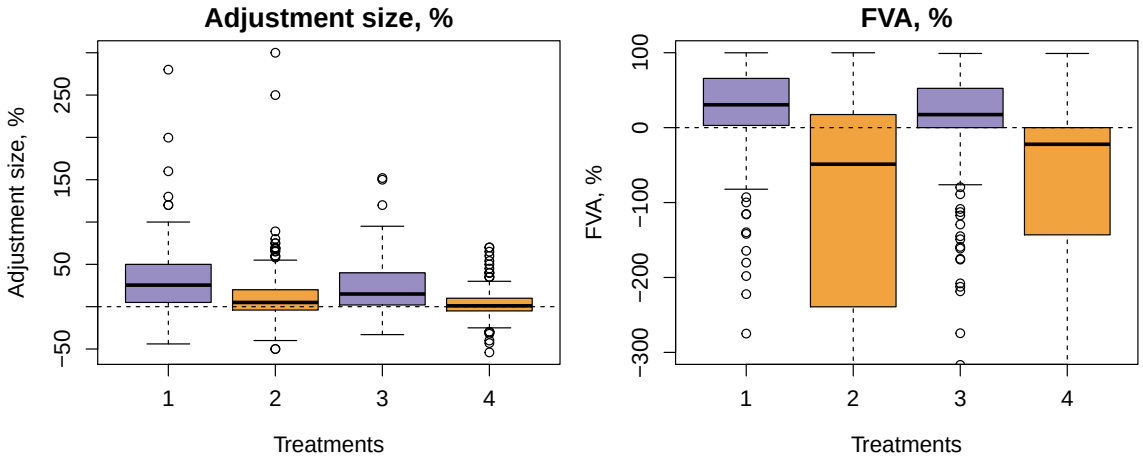

Figure 3.3: Adjustment size (left) and trimmed accuracy (right) boxplots across treatments.

3.4.2 Regression Modelling

The descriptive analysis has already shown some interesting features across treatments, but it is an insufficient tool to investigate the hypotheses and identify possible interactions between variables. Hence, we report the results of regression modelling, where the relative adjustments of $\log(1 + \text{Adjustment})$ is the dependent variable y_i. Davydenko and Fildes (2013) argue that this transformation of the adjustment variable gives a more symmetric distribution.

To test the first two hypotheses, we need to investigate into how users weight different types of information in these scenarios. Hence, we use explanatory variables that (1) correspond to the experimental design cases: baseline/promotional models and no/decomposition; (2) are connected to the system forecast, such as current system forecast, last promotional observation, etc.; and (3) are related to the contextual information, for example reasons for past/future promotions. The full list without

Table 3.2: List of independent variables (without any interaction effects).

Variables	Description
Treatments	
Treatment	Factor variable, where Treatment 1 is the baseline
Time series related	
Last promotional uplift	Last promotional uplift
Current forecast	Current statistical forecast
Average promo uplift	Average of past promotional uplifts
Last actual	Last actual sales before the forecasting period
Contextual information related	
"Promotions"	Promotional reasons (positive/negative)
"Marketing Research"	Marketing reasons (positive/negative)
"Other"	"Other" (over-positive) reasons
Checked "Promotions"	Promotional reasons useful (checked)
Checked "Marketing Research"	Marketing reasons useful (checked)
Checked "Other"	"Other" (over-optimistic) reasons useful (checked)
Promo reasons, past events	4 dummies, one for each past promotion
Market reasons, past events	4 dummies, one for each past promotion
Past reasons	Number of past statements presented (as a set of dummies)
Current reasons	Number of current statements presented (as a set of dummies)
Misc	
Low noise	Dummy for low/high noise
Order	Order of time series excluding trial runs
Expert prior	Forecaster's prior estimate for the promotional uplift

any interaction effects can be seen in Table 3.2. Since the time series were generated automatically for each participant separately, there is no need to account for the time series variability.

Table 3.3 shows the results for Ordinary Least Squares (OLS) regression with variable selection using the Akaike's Information Criterion (AIC) in a step-wise manner. We added three sets of interaction effects for "Promotions", "Marketing" and "Other" statements with corresponding treatments, since we test whether the decomposition has any effect.

Table 3.3: OLS Regression results for Model 1, dependent variable is log(1 + Adjustment).

	Model 1
(Intercept)	0.1768 (0.0375)***
Treatments	
Treatment 2	−0.1252 (0.0294)***
Treatment 3	−0.1155 (0.0246)***
Treatment 4	−0.2040 (0.0303)***
Contextual information related	
"Promotions" (negative)	0.0010 (0.0258)
"Promotions" (positive)	−0.0068 (0.0243)
"Promotions" (negative)* Treatment 2	−0.0100 (0.0366)
"Promotions" (positive)* Treatment 2	0.0315 (0.0340)
"Promotions" (negative)* Treatment 3	0.0406 (0.0357)
"Promotions" (positive)* Treatment 3	0.0762 (0.0338)*
"Promotions" (negative)* Treatment 4	0.0137 (0.0372)
"Promotions" (positive)* Treatment 4	0.0751 (0.0357)*
Checked "Other"	0.0711 (0.0161)***
Time series related	
Average promotional uplift	0.0006 (0.0002)***
Current forecast	−0.0006 (0.0002)***
Other	
Expert prior	−0.0001 (0.0001)**
R^2	0.1931
Adj. R^2	0.1850
Num. obs.	1504

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$. Figures in parentheses are standard errors.

H1: Decomposition helps users identify relevant statements more effectively. Table 3.3 shows that only promotional interaction effects stayed in the final model with significant coefficients for positive promotional statements in treatments 3 and 4 ($p = 0.024$ and $p = 0.036$ respectively), which suggests that the decomposition helps to identify relevant contextual cues, which are any promotional statements in this case. This finding supports H1.

H2: Decomposition helps to weight qualitative information appropriately. While the results for the first hypothesis are promising, indicating that the decomposition

is able to assist the filtering of contextual information, the appropriate weighting of this information is a harder task. There are two main components in this task (i) to determine the correct direction of the adjustment, and (ii) to assign an appropriate size to a particular piece of contextual information. We split H2 into two parts: (i) for diagnostic and (ii) non-diagnostic qualitative information where the by-design weights should be 10%/-10% for the positive/negative promotional information and zero for the rest.

According to the final model, positive promotional statements were weighted slightly higher than "Other" for both baseline and promotional models in the decomposition case (having coefficients of 0.076 and 0.075 for "Promotions" in treatments 3 and 4 compared with 0.071 for "Other"). In the non-decomposition cases, the most substantial effect was attributed to overly optimistic statements that had by design no predictive power.

To find the net effect for diagnostic and non-diagnostic statements, first we calculated the base effect for each treatment (intercept + treatment coefficient) and then added the effect of the promotional/other dummy coefficients. The results are provided in Table 3.4.

We observe that the adjustments made for negative reasons in treatments 1 to 3 have the wrong sign leading to an inappropriate uplift in the adjustment. Comparing between the decomposition and no decomposition cases, we see that the net effects for the decomposition are closer to the expected -10%. On the contrary, the net effects for positive diagnostic cues are mostly positive and are close to the true value. When comparing the use of decomposition or not, we observe that for the baseline model the net promotional effect for decomposition is closer to the expected coefficient. However, this is not the case when a promotional model is available. Moreover, the comparison of these parameters, across all baseline and promotional models, shows that decomposition decreases these parameter estimates. Overall, there is some support for H2

Table 3.4: The net effects for promotional information (H2).

	Treatment			
	1	2	3	4
Statistical forecast	Baseline	Promotional	Baseline	Promotional
Decomposition	No	No	Yes	Yes
Base effect	0.1768	0.0516	0.0613	-0.0272
Net effect				
"Promotions" negative	**0.1778***	**0.0416***	**0.1019***	-0.0135
	(0.0000)	(0.0023)	(0.0000)	(0.0603)
"Promotions" positive	0.1700	0.0831	0.1375	0.0479
	(0.0614)	(0.7094)	(0.3956)	(0.2512)
"Other"	**0.2479***	**0.1227***	**0.1324***	0.0439
	(0.0000)	(0.0024)	(0.0001)	(0.2805)

T-test for net effects and "true" values (0.1 for "Promotions" and 0 for "Other").
P-values are in parentheses: $^*p < 0.05$.

for diagnostic information.

H3: Decomposition increases acceptance of model forecasts. The coefficients for treatment dummies in the final model indicate the effect of decomposition on model forecast acceptance. A systematic decrease in adjustments may imply either increase of model acceptance or a restrictive feature of the decomposition. This was observed in the descriptive analysis, where the average size of adjustments decreased substantially in the decomposition case. Table 3.5 provides the direction of adjustments; we notice further differences between non-decomposed and decomposed cases. The number of negative adjustments increases from treatment 1 and 3 to 2 and 4, and the highest increase is observed in the "No adjustments" category, which indicates a change in forecasters' behaviour. Surprisingly, this is observed for both baseline and promotional model forecasts, whereas we expected a positive effect from decomposition only on model forecast acceptance, as no adjustment is needed in this case. Hence, participants either exercise extra caution when making any adjustments or just accept models more when encouraged by the design to evaluate information step

by step. Therefore, there is some support towards H3.

Table 3.5: Direction of adjustments across treatments.

	Treatment			
	1	2	3	4
Statistical forecast	Baseline	Promotional	Baseline	Promotional
Decomposition	No	No	Yes	Yes
Positive adjustment	87%	61%	79%	52%
Negative adjustment	8%	29%	9%	31%
No adjustment	5%	10%	**12%**	**17%**

To summarise, decomposition of contextual information performs better than a non-structured presentation in the context of forecasting when combining both qualitative and quantitative information in a DSS. The results show that the decomposition helps to reduce the cognitive load of this complex task, which consequently helps to focus on identifying relevant cues and to weight them efficiently.

3.4.3 Post-experiment questionnaire analysis

The results of the post-experiment questionnaire are presented in Appendix 3.A. This was aimed at assessing the general motivation and involvement of participants after the experiment. Despite participants responding with low confidence about their knowledge of sales forecasting, we were able to see the effect of the structured presentation of qualitative information on their performance (in the previous section). This indicates the effectiveness of such a feature even for novice forecasters. Participants evaluated their perception of the usefulness of statistical models higher than the usefulness of time-series graphs, revealing a primary focus of attention on a number rather than on graph. The results of this questionnaire also affirm the general motivation to estimate the promotional effects accurately, noting that the reasons provided had influenced their decisions. In general, participants understood the design and

interface of the system. We also noticed that they were generally engaged with the experiment, although they were conservative in their performance expectations.

3.5 Discussion

While statistical modelling is undoubtedly crucial in the current supply chains and planning departments, not all information can be incorporated into these statistical methods. For this reason, we need to focus on helping and assisting humans in incorporating such contextual information using available support systems. As Fildes et al. (2006) summarised, a well-designed FSS should be reliable, user-friendly, flexible and commercially attractive for both users and developers. Yet, more than a decade later, we still observe that the integration of both qualitative and quantitative information from multiple sources is typically ignored, despite the fact that it is widely used in practice (Fildes et al., 2020).

3.5.1 Restrictiveness of decomposition

The current study found that decomposition is able to reduce cognitive load, which is caused by the amount of information that is needed to be taken into account when making any decisions. The descriptive analysis pointed out that decomposition decreases the size of adjustments on average for both the baseline and promotional models. To explore further, we have built a regression with two binary variables "Promotional model" and "Decomposition" (Model 2 in Table 3.6). We observe that the intercept between all four treatments changes dramatically: (1) decreasing from 16% to 8% when switching from "no decomposition & baseline" to "decomposition & baseline"; and (2) changing direction from +4% for "no decomposition & promotional model" to -3% for "decomposition & promotional model". The latter switch in behaviour is particularly interesting since it indicates that the structured presentation

Table 3.6: OLS Regression results for Model 2, dependent variable is log(1 + Adjustment).

	Model 2
(Intercept)	0.1477 (0.0357)***
Treatments	
Promotional model	−0.1102 (0.0186)***
Decomposition	−0.0667 (0.0102)***
Contextual information related	
"Promotions" (negative)	0.0138 (0.0129)
"Promotions" (positive)	0.0385 (0.0122)**
Checked "Other"	0.0707 (0.0161)***
"Promotions" 3 (negative)	0.0221 (0.0106)*
"Promotions" 4 (positive)	0.0154 (0.0108)
"Marketing" 4 (negative)	−0.0158 (0.0108)
Time series related	
Average promotional uplift	0.0006 (0.0002)***
Current forecast	−0.0006 (0.0002)***
Other	
Expert prior	−0.0001 (0.0001)**
R^2	0.1929
Adj. R^2	0.1869
Num. obs.	1504

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$. Figures in parentheses are standard errors.

of qualitative information reduces adjustments in all treatments (see the first plot in Figure 3.3 for visual assessment of adjustment sizes between pairs of treatments). This could be a tool to restrict adjustments in the system by providing qualitative infromation in clusters. Nevertheless, these restrictive effects correlate with model acceptance (H3), which is again confirmed by a negative coefficient for the promotional model in Table 3.6. We argue that restrictiveness is a more reasonable explanation, as model acceptance would seem to be occurring even when the model is apparently erroneous (e.g., baseline model).

3.5.2 Accuracy

We discovered that the forecast accuracy improves the most between treatments 2 and 4, while for the baseline model case the effect is negligible. Looking at the second graph in Figure 3.3, we see that the decomposition has significantly reduced variability of FVA between treatments 2 to 4 (F-test p-value is close to zero), and as a result, slightly improving the performance. However, there is no statistically significant difference in the central location of the distributions of FVA for both pairs of treatments (all $p > 0.05$). The reduced variability again supports the proposition that decomposition helps to restrict unnecessary adjustments, with benefits for accuracy.

There is a strong positive correlation between adjustment sizes and FVA for the baseline model (higher adjustments lead to better accuracy). Naturally this is not the case for the promotional model. The comparison of adjustment sizes (the first graph in Figure 3.3) between pairs of treatment 1 and 3 and 2 and 4 shows that the results for non-decomposed and decomposed cases are statistically different (p-values for both t-test and F-tests are close to zero).

3.5.3 Optimal forecasting models are not realistic

In this experiment, we compared (i) baseline and (ii) promotional model forecasts. The baseline model form, although common in practice, does not take the promotional setting into account. The second model assumes full knowledge of the data generating process, which is unrealistic. Good models will reasonably approximate the observed demand, but not perfectly due to the complexity of real demand processes and limitations of the sample. This suggests that contextual information will remain useful to enrich forecasts. Decomposition demonstrates gains for both extremes, reducing adjustments, and allowing participants to better identify diagnostic information. We expect this to be the case for the more realistic condition, where the forecasting model

is neither optimal, nor fully uninformed.

3.6 Conclusions

In this chapter, we have proposed a decomposition design feature of a FSS to handle qualitative information expecting that it will help users to filter and weight contextual information more efficiently. We analysed the effect of the decomposed presentation of contextual information on forecasters' adjustments in a demand-planning setting with promotional events. This presentation was contrasted with a simultaneous display of both model-based and domain data. Moreover, this contextual information can be relevant or irrelevant, which raises the question of whether users are able to filter one from another. We also added a second dimension of complexity, varying baseline (sub-optimal) and promotional (optimal) models provided to users.

We found that the decomposed representation of qualitative statements (1) helps users to filter relevant information for both baseline and promotional models, even though participants remained prone to take into account over-positive statements ("Other"); (2) reduces adjustment size naturally, lessening their variation and therefore imposing organic restrictiveness to judgments and, arguably, increasing model acceptance. The former means that humans have the potential to filter and incorporate only useful information when using this feature in an FSS. The latter suggests that this information is weighed better. It has a similar effect for both the baseline and promotional models.

In general, this design feature could be beneficial in the forecasting process because it imposes lower cognitive load due to a step by step evaluation of quantitative and qualitative information. At the same time, we would claim that it should be relatively easy to implement it in an FSS as long as there is a database of textual information. We show that this design element can make the decision-making process

more consistent and predictable.

Another important implication of this feature is that it performed well as an efficient restrictive tool to manage judgmental interventions to system forecasts. It naturally forces users to evaluate these pieces of information step by step, discourages them from making unnecessary adjustments or at least reducing their sizes. Basically, we transform a multiplicative process into an additive one, which is easier to deal with. Goodwin et al. (2011) showed that neither direct restrictiveness nor guidance were efficient in guiding forecasters during their decision-making process; they either made overly large adjustments or ignored the system's recommendations. On the contrary, we found that decomposition could be efficient for both baseline and promotional models, even though it did not lead to optimal accuracy gains achievable with a prefect interpretation of the cues: it still has the potential of being a useful element in the design of forecasting support systems. Both findings are novel in the forecasting literature, indicating ways to improve not only the process itself but also more generally decision support systems where information can be decomposed.

Decomposition could be implemented in Forecasting Support Systems by introducing information clusters that are easier to process cognitively. However, there are two limitations here. First, while the effective design of FSS is key to successful adoption and use of such systems, we can observe general reluctance of software vendors to merge the judgmental and model-based sides of the forecasting process. This may be due to (1) the risk of possible errors from forecasters' interventions to otherwise appropriate model forecast, but (2) also a lack of design and DSS expertise from software vendors. Second, forecasters are prone to many cognitive biases and possible algorithm aversion when dealing with such tasks. The efficient decision-making process, in this case, requires both system acceptance from the user's side along with the implementation of useful features from the software side, and this balance is still to be established in practice. More research is needed in both directions, given the promi-

nence of both statistical forecasting and judgmental adjustments in organisations, reinforced by an acknowledgement of the importance of integration.

Appendix 3.A Post-questionnaire responses

Table 3.7 provides descriptive statistic of the responses, ordered from none/low (1) to high (5).

Table 3.7: Mean, median, and standard deviation of post-questionnaire responses. Scale is from 1 to 5 where 5 is the highest mark.

Question	Mean	Median	Std. Dev
Overall knowledge of sales forecasting	2.02	2.00	1.01
The time-series graphs' usefulness	2.63	3.00	0.95
Expected accuracy of adjusted forecasts	2.68	3.00	0.91
The statistical forecasts' usefulness	3.73	4.00	1.03
Expectation for the accuracy of the statistical forecasts	3.46	4.00	1.02
The provided reasons were very easy to understand and use	3.53	4.00	1.16
The reasons provided had a direct influence on my forecasts	3.80	4.00	0.92
The occasional highly positive messages had a direct influence on my forecasts	3.42	4.00	1.14
Motivation to estimate the promotional effects accurately	3.85	4.00	0.99
Overall customer experience with promotional products in supermarkets	3.63	4.00	1.12

Chapter 4

Judgmental model tuning versus adjustments

Abstract

Expert judgment is the primary tool for incorporating additional knowledge quickly and effectively into statistical models. Despite being frequently used, there is little understanding of the conditions when these expert interventions are useful and done effectively. Moreover, there is almost no research investigating the use of judgment during the modelling process, even though it is one of the critical stages where expert knowledge can be incorporated into models. In this chapter, we explore the use of judgmental model tuning and its effect on forecast accuracy, and compare it with judgmental adjustments of statistical forecasts. Using data from a case study, we find that model tuning is ineffective. However, once we remove judgmentally imposed events that are spurious, it provides accuracy gains, indicating that model tuning can be subject to substantial biases that can harm its performance, yet promising as it scales better than judgmental adjustments of statistical forecasts.

4.1 Introduction

Human judgment plays an important role in operational research, including demand forecasting and supply chain modelling. Expert input is typically required at all stages of the forecasting process (Ord et al., 2017, chapter 1): (1) before implementing any models: setting aims and purposes for forecasts, obtaining required dataset, understanding organisational performance indicators, accuracy and bias measures; (2) during the modelling process: model pooling, selection, tuning, benchmarking; (3) after: assessing accuracy, adjusting forecasts to take into account additional information (e.g., market intelligence, promotions, macro- and micro-economic changes).

The judgmental forecasting literature is mostly focused on the last category: investigating judgmentally adjusted statistical outputs or pure judgmental methods when statistical models are unavailable (see extensive literature reviews by Lawrence et al., 2006; Arvan et al., 2019; Perera et al., 2019). This focus is also motivated by several survey studies where the prevalence of adjusted statistical forecasts in practice is apparent (Sanders and Manrodt, 2003; Fildes and Goodwin, 2007a; Boulaksil and Franses, 2009; Weller and Crone, 2012; Fildes and Petropoulos, 2015). However, there is almost no literature looking at the use of judgment during the modelling process, even though it is one of the most important stages where expert knowledge can be incorporated into forecasting models. This could be implemented as judgmental model pooling, judgmental model selection, manual model parameterisation or model tuning. The human impact on this stage is arguably harder to track and evaluate since the manipulation is taking place before there is anything to evaluate. The literature typically reports results conditional on a pool of alternative forecasting models or methods, and reasonably not all options are included (Kourentzes et al., 2019). Hence, other analysis methods such as case studies, laboratory experiments or observations need to be carried out to fill this substantial gap in the literature.

Petropoulos et al. (2018) used a behavioural experiment to compare judgmental model selection against a statistical model selection procedure that is dominant in practice. The authors show that judgmental selection is competitive and avoids the worst model choice more consistently than statistical methods. Also, the authors indicate the research potential for judgmental interventions during the modelling stage and highlight the current limitations of many forecasting support systems. To the best of our knowledge, there are no other studies investigating this area. Furthermore, there is no work that explores how these model interventions interact, if at all, with typical judgmental adjustments of statistical forecasts.

In this research, under *judgmental interventions*, we include all human inputs/adjustments that are added either before, during or after the statistical modelling. While *judgmental adjustments* are only those revisions that take place to statistically generated forecasts using some contextual information (Lawrence et al., 2006), we define *judgmental model tuning* as all changes that analysts make during the model building stage, such as inclusion of features and changes in the model form. Crucially, due to the structure of these interventions, judgmental model tuning precedes all possible judgmental adjustments to statistical forecasts, and hence, we can expect that errors from the model adjustments can propagate to the following revisions.

Using data from a case company, we explore the use of a two-level judgmental intervention process of statistical forecasts and inspect issues connected to that. First, demand planners can incorporate their knowledge during the model building stage, manually adding indicator variables to models to identify specific events/conditions, implementing judgmental model tuning as we defined above. Second, they can manually adjust the system forecasts at the final stage of the forecasting process. Both options could be implemented by the same group of experts, which can potentially introduce additional biases and overconfidence. We investigate these interventions in

a company case from a major retailer in the UK.

The aims of this chapter are (1) to analyse judgmental model tuning accuracy; (2) to explore the interaction between model tuning and judgmental adjustments. This chapter contributes to the existing literature on the use of human judgment in forecasting by investigating alternative ways to incorporate expertise at different stages of the forecasting process. Both have been largely overlooked in the literature.

This chapter is structured as follows. The next section provides an overview of the literature on judgmental adjustments in promotional modelling, concentrating on features that are relevant for our case, but also the leading motivation for adjusting forecasts and current known case studies. Section 4.3 presents a case study of a major UK retail company and Section 4.4 provides an exploratory analysis of the interventions. Section 4.5 analyses human judgment from a modelling perspective, while in Section 4.6 we propose how to reduce the number of judgmentally imposed explanatory variables. We explain the results and draw implications of this case study in Section 4.7.

4.2 Literature review

The forecasting task typically combines time series modelling and expert knowledge to capture the appropriate time series components using statistical methods, while being able to react to new information using managerial judgment. Forecasting Support Systems (FSS) play a major role in the suitable combination of these two methods. Petropoulos et al. (2018) listed all stages of the forecasting process where practitioners can potentially apply judgment: (1) definition of a set of candidate models, (2) selection of a model,Mparameterisation of models, (4) production of forecasts, and (5) forecast revisions/adjustments. Modern FSSs typically store and present all relevant quantitative information, define and provide automatically specified statistical models

to use, and evaluate the accuracy of these outcomes when new data is available. In an ideal situation, all steps to produce good-quality forecasts can be done within a FSS. However, there can be non-times-series information (e.g., marketing plans, actions of competitors, extreme weather events as in Webby et al., 2005; Lawrence et al., 2006; Fildes and Goodwin, 2013) that might motivate demand planners to make adjustments to both the model building process and the forecasts. In that case, the most important role of the FSS is to support and assess the use of managerial judgment.

4.2.1 Judgmental adjustments

Our knowledge about judgmental adjustments is primarily based on empirical studies that investigate how these are done, in what conditions and the impact on the final forecast accuracy (e.g., Mathews and Diamantopoulos, 1986; Fildes et al., 2009; Franses and Legerstee, 2009; Trapero et al., 2013; Franses, 2014). Based on the analysis of four different companies by Fildes and Goodwin (2007b) and of a pharmaceutical dataset for 37 countries by Franses and Legerstee (2009), around 75% and 90% of all forecasts respectively had been adjusted in these cases. And similar results have been reported not only for one-step-ahead forecasts. For instance, Mathews and Diamantopoulos (1989) showed that a number of adjustments fluctuated between one third and one-half of all products (around 900 products) across six time periods. Surprisingly, the recent evidence by Van den Broeke et al. (2019) and Kourentzes and Fildes (2021) showed that the number of adjustments has not decreased since Mathews and Diamantopoulos (1989): in the first case it ranged from 10% to 99% in four companies; in the second case, more than 90% of all forecasts were adjusted across horizons. This again highlights the importance of judgement in forecasting and the ineffective use of information in FSSs.

As for the accuracy of these adjustments, positive judgmental adjustments (increasing statistical baseline forecasts) were far less effective than negative (decreasing

system output, respectively). In contrast, small adjustments harmed forecast accuracy and should be avoided (Fildes et al., 2009). Franses and Legerstee (2009) also showed that the experts make frequent adjustments aiming upwards, rather than downwards, which can be predicted to some extent using a regression model. Trapero et al. (2013) analysed the judgmental adjustments connected to promotions and found that there was a positive impact of these corrections on forecast accuracy, but it was inconsistent. Modelling these adjustments statistically proved a promising alternative, but there were still cases when humans added value.

Despite the frequency and persistence of expert adjustments in practice, the challenge remains that their quality depends on the individuals or groups making the interventions and the available information to them (Davydenko and Fildes, 2013; Fildes et al., 2019a). Overall, the literature demonstrates both beneficial and harmful judgmental adjustments (for examples, see Fildes et al., 2009; Oliva and Watson, 2009; Franses and Legerstee, 2011; Syntetos et al., 2009) in different contexts. This lack of consistent findings is a limitation of a field study methodology. To address this, there has been an increasing number of laboratory experiments that try to understand such adjustments in controlled conditions and highlight when these are useful (Fildes et al., 2019a; De Baets and Harvey, 2018; Sroginis et al., 2019; De Baets and Harvey, 2020).

Even though judgmental forecasting is a prominent area of research in demand planning and forecasting, there are still open questions about the use of judgment in this field, appropriate system support and possible implications of inefficient combination of humans and systems. The consistent observed biases and inefficiencies show that there is still much room for improvement.

4.2.2 Judgmental model tuning

Judgmental adjustments require humans to weight the additional information that is not available to the statistical model. We know that these decisions are prone to

not only many cognitive biases and heuristics (see an overview of Tversky and Kahneman (1974)'s heuristics in forecasting by Harvey, 2007), but also to seeing false patterns in noise (Harvey, 1995; Goodwin and Fildes, 1999). Despite all these and other cognitive challenges, people still can add value to statistical forecasts by considering unaccounted qualitative information (e.g., Fildes et al., 2009; Trapero et al., 2013). An alternative approach, instead of estimating this effect judgmentally, is to add additional variables into a statistical model, which is a partial judgmental parametrization of an initial statistical model. This can be considered as a part of the modelling process. At the same time, we note that many FSSs offer either automatic or semi-automatic modelling algorithms (Fry and Mehrotra, 2016; Fildes et al., 2020) limiting direct access to humans. Thus, they artificially restrict humans, and potentially exacerbate the problem of efficient combination of humans and computer systems.

Additional qualitative information is typically handled outside the system, which means that the whole planning process becomes even harder to track and evaluate. Hence, transforming some qualitative information into additional indicator variables (e.g., Trapero et al., 2013) might be an effective way to optimise the time and effort committed to forecast revisions. The resulting use of judgmental model tuning and judgmental adjustments can accelerate the forecasting process since demand planners can then focus only on low-level adjustments for a specific product while model adjustments save some time for more general information.

We assume that judgmental model tuning can potentially be (1) easily scalable across different stores/products; (2) time-efficient; (3) easier to implement in the system; (4) possible to track, evaluate the accuracy, and provide feedback; (5) free of direct cognitive biases, and therefore more accurate. In particular, model tuning is supposed to be cognitively easier since only the location of a change is required rather than both location and value assignments. However, there is a risk for experts

attempting to capture too much complexity by adding too many variables or to omit some important variables. If both model tuning and adjustments are used, a double-counting bias might be an additional issue (Bunn and Salo, 1996).

This chapter aims to address this gap in the literature by analysing the accuracy of judgmental model tuning in comparison to adjustments. Even though this analysis is motivated by a specific case study, the ultimate aim is to investigate whether judgmental model tuning is a promising development for more general use in Forecasting Support Systems as an alternative (or extension) to judgmental adjustments.

4.3 Case study data

The data was collected from a retail company specialising in household and fast-moving consumer goods (FMCG) products. It was comprised of (1) sales in units; (2) one-step-ahead final forecasts; (3) judgmental adjustment effect of system forecasts (percentage increase/decrease); (4) binary indicator variables having a value of one if there is a promotion and 0 otherwise and (5) binary indicator variables for indicating various special events. This last element embodies judgmental model tuning that is the main focus of this study.

The dataset contains a subset of stores with different volume of sales (we class them as low, medium and high), having a different number of Stock Keeping Units (SKUs), ranging from 22 to 48 thousands per store. The collected sales and forecasts data was at a SKU/store level, sampled at a weekly frequency, while all promotional markers and adjustments were made for specific days. In the case company the one-step-ahead system forecasts were not saved separately in the system and cannot be reconstructed fully. Hence, there were no complete triplets (actuals, system and final forecasts) available. Note that the data was highly intermittent: around 77% of all observations were zero-sales periods, while forecasts predicted around 57% of zero-demand periods.

A detailed description of the forecasting process in the case company is provided in Section 2.4.

There were two levels of human interventions to the system forecasts: (1) inclusion of binary variables for events as judgemental model tuning; (2) percentage adjustments on top of the statistical forecast outcome. The latter was either local, applied for a particular SKU in a particular store, or global, applied for a particular SKU in all stores in the company. To visualise these categories, we adapted a graph from Petropoulos (2019) to create Figure 4.1, highlighting the two steps that corresponded to our types of interventions with the triggers that affected them.

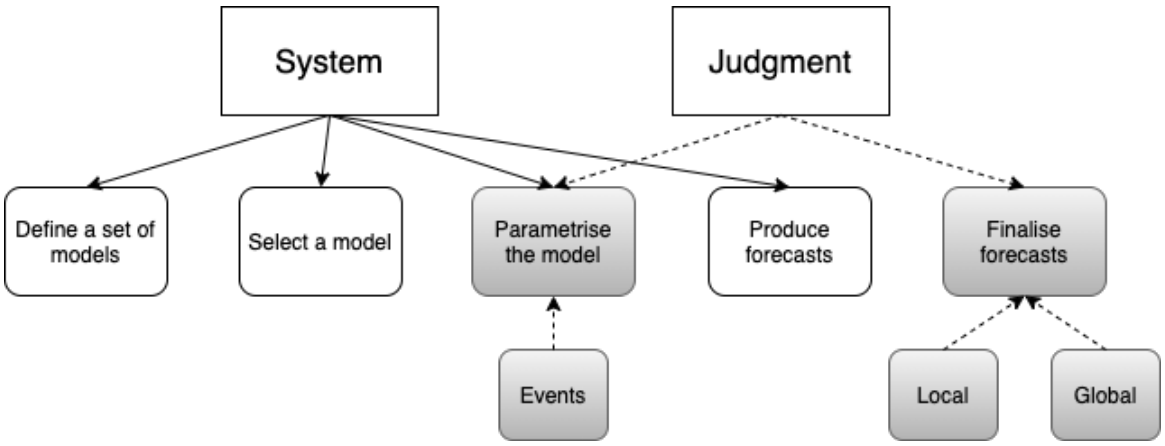

Figure 4.1: The role of judgment in the FSS, adapted from Petropoulos (2019).

The promotional periods were frequent and crucial for the company. They were planned in advance and included in the FSS. The different types of promotions (10 in all) were hard-coded at the local level as indicator variables, and their effect estimated by the system automatically, using regression modelling. We considered this to be a part of the system forecast.

The system allowed demand planners to manually introduce binary indicator variables, based on their knowledge and expertise, for various events, such as "Exam period in May" or "Back to college". These special events were implemented at the global level, across all stores. They were processed in the FSS and represented the

judgmental model tuning.

There were several ways to make judgmental adjustments of statistical forecasts: (1) overwriting the forecasted value; (2) adding extra units of sales to the system forecast; and (3) a percentage change applied to the statistical forecast provided by the system. The majority of adjustments were done as percentage changes. Adjustments were made for event periods, suggesting a double expert intervention: at the modelling and then at the post-processing stage, as a combination of judgmental model tuning and adjustments. Since there were double adjustments, we also attempted to answer (1) how these type of adjustments interacted, and more crucially, (2) what was the role of judgmental adjustments when model tuning is used.

Figure 4.2 provides a screenshot of the FSS interface, showing how the information is reported to the experts. Observe the promotional periods indicated at the lower part of the screen, where horizontal lines of different colours represent event markers.

Figure 4.2: Screenshot of an example time series in the FSS.

4.4 Descriptive analysis of judgmental interventions

To assess the scale of these interventions, we have examined all observations for the 24 weeks (6 months) of the collected data, where all types of data (sales, forecasts, adjustments) were available. Final forecasts were saved only for this period of time. We split the data into several major categories: (i) promotions; (ii) judgmental model tuning (events); (iii) local and (iv) global adjustments. The adjustments were measured as a percentage of the model-based forecast.

While the total number of SKUs across stores is 197,007, 92% of them have at least one event, emphasising the range and wide use of these indicator variables. Table 4.1 provides other percentages of product that have been at least once under either promotions (28%) or local/global adjustments (21% and 64% of time series, respectively). Since one product might have seen several of these conditions at once, these percentages are not mutually exclusive.

Table 4.1: Number of SKUs that have been at least once under promotions, events or local/global adjustments.

	Stock Keeping Units				
	Overall	Promotions	Events	Local Adjustments	Global Adjustments
# SKU	197,007	28%	92%	21%	64%

In Table 4.2 we see that only around 7% of all observations were under promotion, while additional events were added for more than a third of observations during the investigated period. The number of adjustments varied from store to store, ranging from 0.5% to 3% over the sample period. Around 40% of all modifications were made on zero forecasts that indicated either a restocking motivation or additional managerial knowledge. In just 1% of all observations both promotions and events occurred, and just a few observations have been adjusted both at local and global

levels; hence we disregarded these observations.

Table 4.2: Descriptive statistics for the last 24 weeks (% of all observations): Periods across all SKUs where Promotions, Events, Local and Global adjustments are made.

	Observations				
	Overall	Promotion	Event	Local Adjustment	Global Adjustment
Store 1	581,496	38,415 (7%)	214,421 (37%)	6,464 (1%)	14,930 (3%)
Store 2	552,936	34,056 (6%)	205,874 (37%)	5,919 (1%)	14,060 (3%)
Store 6	513,696	61,858 (12%)	186,496 (36%)	10,616 (2%)	17,540 (3%)
Store 4	828,840	65,280 (8%)	302,708 (37%)	12,960 (2%)	19,401 (2%)
Store 3	1,100,232	67,184 (6%)	413,848 (38%)	13,265 (1%)	21,423 (2%)
Store 5	1,150,968	46,498 (4%)	435,040 (38%)	2,191 (0%)	21,743 (2%)

Demand planners in the company reported that the main reason for local adjustments was revisions of forecasts during promotional periods. Further descriptive analysis of interactions between these conditions is presented in Figure 4.3. Judgmental model tuning (events) is scalable, taking up to 70% of all judgmental interventions in the dataset. Furthermore, it was subject to many local and global adjustments, with more than 40% of manual adjustments overlapping with model tuning. Hence, double expert judgment amounted to a significant part of all interventions.

Global adjustments were implemented more frequently than local ones, both on promotional and events periods. While experts exhibited careful behaviour with local adjustments (e.g., no corrections for observations under promotions and events simultaneously), there were global adjustments applied to many products simultaneously (again, almost two-thirds of all SKUs, while only 21% of SKUs displayed local corrections).

The first line of plots in Figure 4.4 shows how many weeks out of 111 total weeks, as a percentage, have been corrected using either model tuning (events) or local/global adjustments. It corresponds to the duration of all corrections within the history at the product level. For example, the first graph displays that around forty thousand SKUs were tuned in 55-60% of the time periods (approximately 61-66 weeks). The

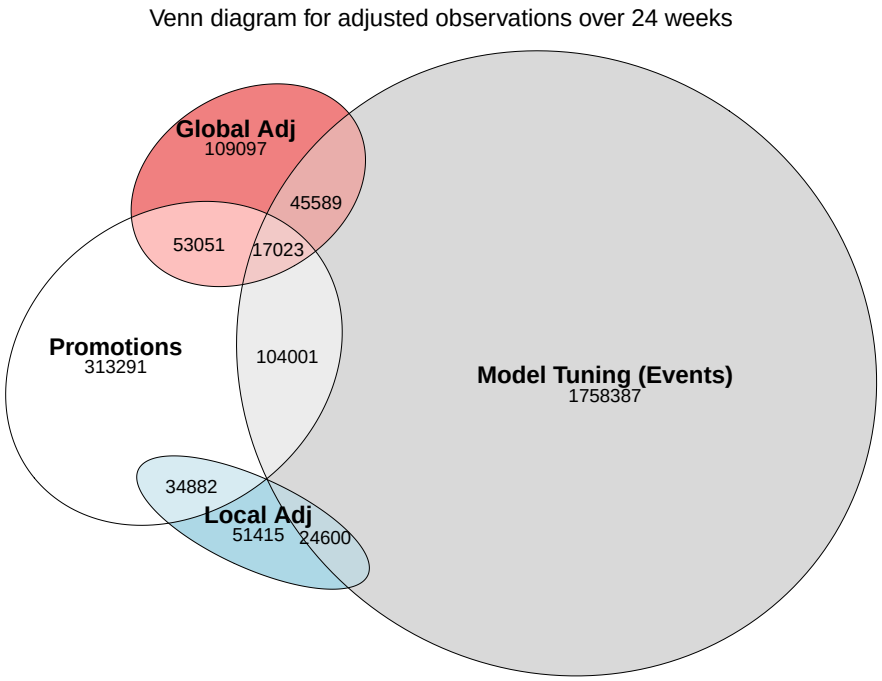

Figure 4.3: Venn diagram for adjusted observations for 24 weeks, across all stores, where a number under a category is the total number of Promotions, Events, Local and Global Adjustments applied (corresponds to the column sum in Table 4.2).

second provides the frequency of unique interventions within a product (SKU). While judgmental model tuning exhibits multi-modal behaviour in both histograms, indicating that for some products more than a half of all observations have been adjusted using indicator variables with a maximum of 31 indicators added to one time series. This result pinpoints a possible problem of superfluous variables in such cases with little history to estimate these variables. Figure 4.5 visualises this problem of the plethora of event indicators together with promotional dummies. This is not a rare case as we evidenced above.

As for local and global adjustments, we once again observe sporadic manual adjustments at the local level (per SKU/store) with a maximum of 5 different values in just a few products. Global adjustments lasted longer and occurred twice as often as local.

The summary statistics of values for local and global adjustments are provided

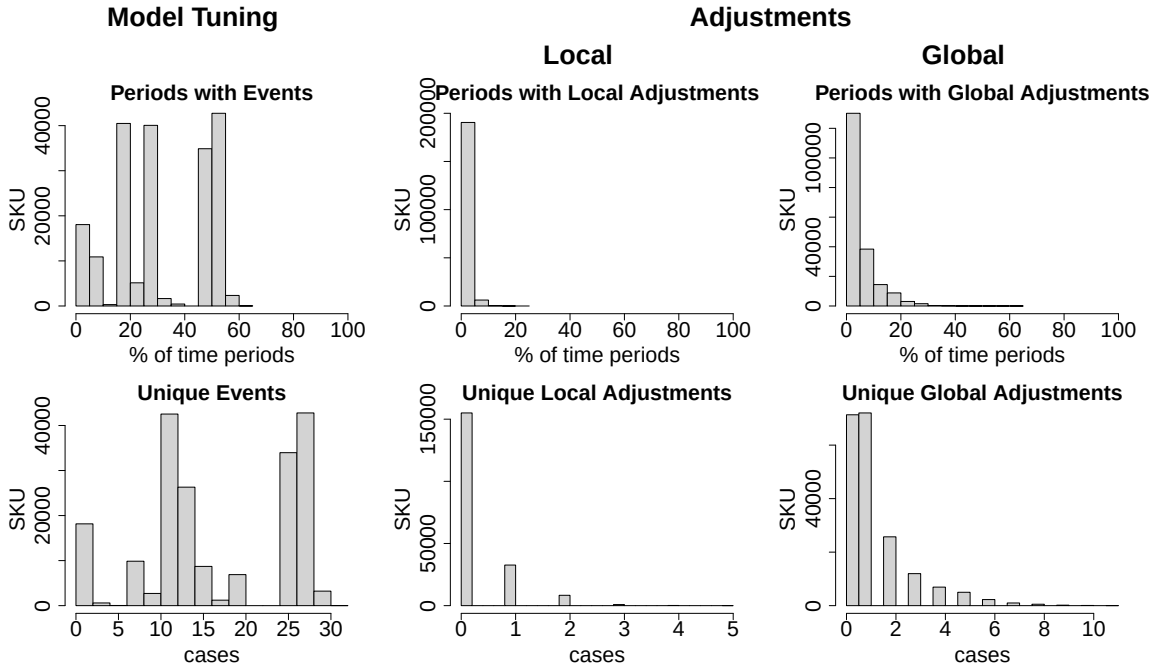

Figure 4.4: Histograms for number of periods adjusted and number of unique corrections per SKU/store.

Table 4.3: Summary statistics for Local and Global Adjustments.

	Min.	1st Quartile	Median	Mean	3rd Quartile	Max.
Local Adjustments	-100%	-50%	-40%	3%	-40%	29400%
Global Adjustments	-100%	10%	10%	37%	25%	42844%

in Table 4.3. Both distributions have many outliers with medians of -40% and 10% respectively. In particular, negative adjustments (e.g., -100% correspond to zero outcome) decrease the number of units by x per cent, are frequent for local adjustments. In contrast, global adjustments tend to primarily increase the final forecasts, with a median of 10% and a mean of 37% increase. Note that both maximums are extremely high, with 29,400% (and 42,844%) sales corrections in some products for the local (global) adjustments respectively.

Even though local adjustments were made less often than global ones, they are typically time-consuming corrections since they are done at the SKU/store level. As for global adjustments and events that are more scalable and easier to implement,

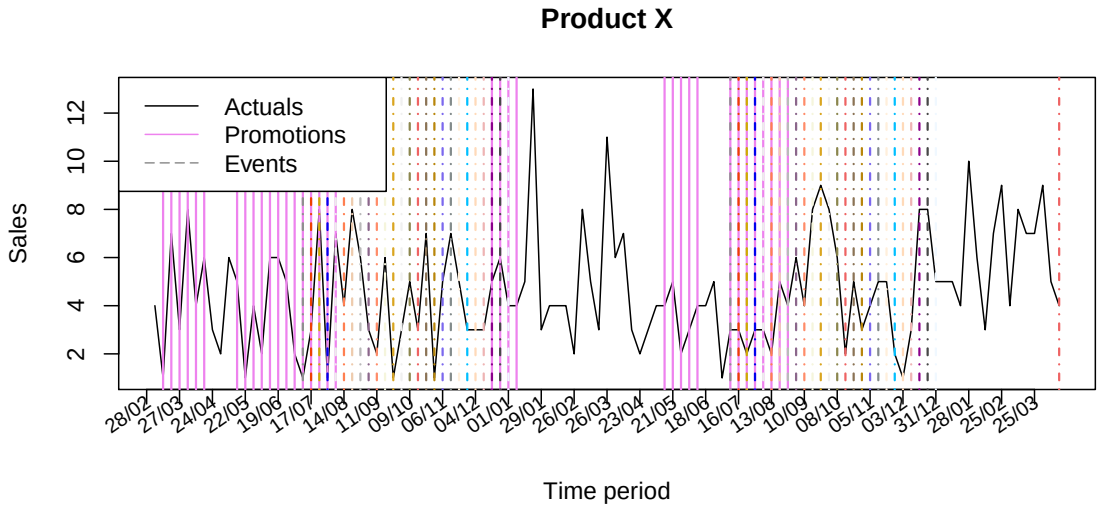

Figure 4.5: Example time series with events and promotional dummies.

there is a chance of misusing it, and as a result, overfitting the data. For instance, these interventions can potentially be useful only for reoccurring events (observed 2 or more times) or for small categories of products that exhibit similar purchase behaviour.

These results lead to several questions: (1) are these events beneficial for the final forecast accuracy? (2) is there a problem with the system forecasts that so many adjustments are being made during promotions? (3) how accurate are the forecasts that have been adjusted twice? We examine these questions later in the chapter.

4.4.1 Accuracy of adjustments

Since pure system forecasts were not available (neither for us nor for the company), we used two benchmarks for forecast error evaluation: (1) simple exponential smoothing method with indicators for promotions (ETS+Promo), (2) simple exponential smoothing method with indicators for promotions and events (ETS+Promo+Events). The first model generated one-step-ahead rolling origin forecasts for a period of 24 weeks using only promotional dummies when necessary. The second model expanded the

list of indicators to include judgmentally added ones for various events (judgmental model tuning). Exponential smoothing has been shown to have relatively good performance and ease of use (Gardner, 2006; Ord et al., 2017), even in the context of retail data (Kolassa, 2019). These models were baseline alternatives to the company's algorithms that were fully controlled, so that we could evaluate the effect of these variables thoroughly.

Using Hyndman et al. (2002) state space framework that is implemented in the `adam` function from "smooth" package v3.1.1. (Svetunkov, 2021b) in R (R Core Team, 2017), we produced one-step-ahead forecasts with initial parameters being produced using a backcasting procedure (see details on the implementation in Svetunkov, 2021a) since it is computationally much faster.

Initially, there were forty events and ten types of promotions in the dataset. Even if some indicator variables were well-known calendar events, such as Christmas or Easter, a simple exponential smoothing model should be able to estimate their parameters efficiently if there are enough observations. Interestingly, almost half of all judgmental model tuning was applied to products with a very short sale history (3 data points or fewer), meaning that it was impossible to evaluate almost half of judgmental model tuning cases - neither local regression not exponential smoothing with exogenous variables could be built for such short time series. We excluded these cases. This finding suggests that these indicators were employed automatically for categories of products, potentially leading to redundant variables.

Since both judgmental adjustments and model tuning played an important role in this case, we split the data into the following mutually exclusive categories: (1) baseline forecasts (with or without promotions); (2) judgmental model tuning; (3) judgmental adjustments (both local and global levels); (4) the double judgment category, which includes both model tuning and judgmental adjustments. The last category was comprised of nested adjustments since model tuning was accounted for in

Table 4.4: The scaled Root Mean Squared Error (sRMSE) across all stores.

	# Observation	Company Final Forecast	ETS +Promo	ETS+Promo +Events
Baseline forecasts	1,799,640	**95.93**	124.60	124.60
Baseline with Promotions	141,309	20.05	**13.47**	14.02
Judgmental model tuning	1,162,224	121.80	**41.75**	**41.75**
Local Adjustments	2,888	**83.76**	172.91	173.06
positive	2,388	**91.47**	189.85	190.01
negative	500	23.68	23.44	**23.43**
Local Adjustments for Promotions	18,021	**13.76**	18.42	19.75
positive	248	**28.70**	31.27	31.29
negative	17,773	**13.44**	18.18	19.54
Global Adjustments	22,750	89.48	19.07	**18.52**
positive	17,049	102.63	21.01	**20.34**
negative	5,701	21.22	**11.45**	11.47
Judgmental model tuning & Global Adj	23,840	267.07	62.32	**62.08**
positive	20,473	286.06	66.07	**65.81**
negative	3,367	86.30	30.88	**30.85**

the system forecast first, and then some products were adjusted further manually. We used an error measure of Root Mean Squared Error (Fildes, 1992) that has been scaled by the in-sample sales. This error measure can be used on intermittent time series as suggested by Petropoulos and Kourentzes (2015). All products, where all sales or forecasts were zeroes, were excluded from this evaluation. The formulas can be found in Appendix 4.A.

Table 4.4 presents the accuracy results across these categories; a horizontal line separates each category and the boldface values indicate the most accurate method within each category. The number of observations was smaller than in both Table 4.2 and Figure 4.3 due to additional filters: (i) mutual exclusivity of categories; (ii) exclusion of infinity and impossible values (division by zero). We observe that the number of observations (the first column) in some categories is striking. For instance, more than a million observations have been adjusted by events (model tuning), which shows how scalable, easy to implement for many products, this method is. As expected, the

smallest category is local adjustments that are applied to a particular SKU at a specific store. However, the same category of adjustments in promotional periods is six times bigger and would require significant effort and time from the experts.

The first line provides the results for baseline forecasts for periods without any promotions, events or otherwise adjusted. The final one-step-ahead predictions are better than both of our baselines, which can be attributed to corrections for out-of-stock that are implemented in the system. Note that the situation changes for promotional periods: simple exponential smoothing with exogenous variables performs better than the regression-based method used in the company. The accuracy of exponential smoothing with promotions over regression has been reported in the literature before (Kourentzes and Petropoulos, 2016).

We conclude that (1) local adjustments increase forecast accuracy for both promotional and non-promotional modelling; (2) neither global adjustments, judgmental model tuning, nor double adjustments perform well compared to simple statistical baselines. These preliminary results show that neither judgmental model tuning nor a combination of adjustments achieves their goals. The worst performance can be seen in the double judgment category, which is somewhat expected since both judgmental model tuning and global adjustments decrease the forecast accuracy separately.

4.4.2 Directions of adjustments

Table 4.4 also reports the errors by the direction of adjustments (positive or negative). For non-promotional modelling, the positive changes are more frequent than negative, while during promotions, only 1% of all adjustments were positive. However, the accuracy of these adjustments is consistent across subcategories: negative adjustments perform much better than positive ones. This result is consistent with previous case studies (Fildes et al., 2009; Trapero et al., 2011, 2013; Van den Broeke et al., 2019). Global adjustments tend to be overall positive and harm the final accuracy. While

Table 4.5: The scaled Error (sE) across all stores.

	Company Final Forecast	ETS +Promo	ETS+Promo +Events
Baseline forecasts	-0.86	**-0.59**	-0.60
Baseline with Promotions	-0.41	**-0.01**	**0.01**
Judgmental model tuning	-3.34	**0.21**	0.25
Local Adjustments for non-promotions	**12.59**	-47.79	-47.89
positive	**14.67**	-58.43	-58.54
negative	**2.69**	2.99	2.97
Local Adjustments for promotions	**0.16**	-1.31	-1.27
positive	-2.24	-1.55	**-1.22**
negative	**0.19**	-1.31	-1.28
Global Adjustments	-2.05	-0.31	**-0.26**
positive	-2.49	-0.52	**-0.45**
negative	-0.74	**0.30**	0.33
Double judgment	-12.82	**1.00**	1.35
positive	-14.60	**1.01**	1.40
negative	-2.01	**0.92**	1.04

positive local adjustments are beneficial for both promotional and non-promotional settings, global adjustments seem unnecessary as even a simple baseline provides better forecasts. Finally, worse performance can be seen in the positive double judgement category, highlighting the harm done by introducing both additional indicator variables and manual interventions.

4.4.3 Bias of adjustments

To analyse the forecast performance thoroughly, we need to measure the bias of the final forecasts (Formula 4.2 in Appendix 4.A), which indicates whether there is a tendency to over-forecast or under-forecast, on average. The scaled error should be close to zero to claim that the forecasts are unbiased. The positive results indicate that the actual values are higher than the final estimates and vice versa.

Table 4.5 provides the scaled errors for three methods for all SKUs. The closest

absolute values to zero in each category are highlighted in boldface. It shows that the final company forecasts' bias is almost always negative, which indicates over-forecasting, especially in the case of double judgment (the last row). These results are consistent with previously reported case studies (Fildes et al., 2009; Trapero et al., 2011, 2013). Even baseline forecasts tend to over-forecast, raising questions about the accuracy of the statistical model that could motivate demand planners to use judgmental adjustments and model tuning to correct such forecasts.

Looking at the errors of both baseline models, the worst performance is in the local adjustments category, producing highly over-forecasted sales, suggesting potential usefulness of expert interventions. We revisit this subset of products in Section 4.5. Our baseline models are less biased than the company's final model forecasts in all other categories. In summary, we observe optimism bias (over-forecasting across products) by the demand planners, especially for periods when special events are introduced. In the next section, we explore the consistency and impact of judgmental model tuning in detail.

4.5 Analysis of judgmental interventions

While judgmental model tuning is potentially more efficient in terms of time and effort required to use it, there is a danger of overusing it. Enabling a statistical model to estimate these indicator variables could be misleading in two ways: (1) accounting for spurious events may result in overfitting rather than improving the baseline model; (2) model parameters estimated on small samples can potentially be biased and inefficient. The redundant variables in a model together with a small sample can lead to fitting to noise and as a result, to poor forecasting performance. However, if the model form is correct, the parameter estimates will be unbiased and efficient, given that the sample set is big enough.

To analyse the effects of both judgmental model tuning and manual adjustments, we build a pooled regression model of forecast errors and all included interventions to find their impact as:

$$
\begin{aligned}
scaledAE_t = & \underbrace{\beta_0 + \sum_{j=1}^{10} \beta_j Promo_{j,t}}_{\text{baseline model part}} + \underbrace{\sum_{j=1}^{23} \beta_{j+10} Events_{j,t}}_{\text{model tuning part}} \\
& + \underbrace{\beta_{34} LocalAdj + \beta_{35} GlobalAdj + \sum_{j=1}^{10} \beta_{j+35} LocalAdj_t \times Promo_{j,t}}_{\text{adjustment part}} \\
& + \underbrace{\sum_{j=1}^{23} \beta_{j+45} LocalAdj_t \times Events_{j,t} + \sum_{j=1}^{23} \beta_{j+68} GlobalAdj_t \times Events_{j,t}}_{\text{double judgment part}} + \epsilon_t
\end{aligned}
\tag{4.1}
$$

where $scaledAE_t$ is a response variable of scaled absolute errors of the company's final forecasts, $Promo$ and $Events$ are indicator variables for the corresponding periods; and $LocalAdj$ and $GlobalAdj$ are logarithmically transformed percentage adjustments, where initially values under 1 correspond to decrease of model forecasts, and all values above 1 indicate a percentage increase (e.g., a value of 0.5 implies 50% decrease of the initial model forecast while 1.5 is 50% increase from the baseline). The logarithmic transformation of these provides a better distribution than the summary statistics in Table 4.3 represent.

The rest of the variables in Formula 4.1 are interaction terms, except local adjustments during promotional periods, that account for double judgments. This model comprises four parts: (1) a baseline model with promotional effects; (2) model tuning part; (3) adjustment part (both local and global adjustments for non-promotional and promotional periods); (4) double judgment, nested adjustments of local/global adjustments during events.

Table 4.6 provides the output of the regression model. Out of 197,007 SKUs, 124,112 are suitable for regression due to missing or infinite values (e.g., short time

series). However, there are 24 periods for each SKU to include, so in total, we would have a regression with almost 3 million observations and 91 explanatory variables. Due to computational restrictions, we limit our set of SKUs to only those with either local or global events and then randomly sample the maximum possible observations (in this case, 900,000 observations). The results for different samples are similar as well as the average coefficients across several samples, and therefore, we only present one output. Note that R^2 is small due to the variability in the huge number of observations.

4.5.1 Judgmental model tuning

The first set of variables correspond to judgmental model tuning, which is comprised of 23 events. Note that this model is evaluated on 24 weeks, meaning that there is an event for almost every week in this sample. Once more, this provides evidence that demand planners add too many indicators. Estimating a regression on only two years of weekly data (111 observations, in total) with variables for almost a half of the data can result in statistically inefficient parameter estimates, which might lead to inaccurate forecasts. Moreover, 9 out of 23 events are significant at 5% level and positive, meaning that these events increase forecast errors systematically. These significant events can be grouped around Christmas, revealing how experts perceive this busy period, trying to correct baseline forecasts. At the same time, the effect of other model tuning dummies is indistinguishable from zero. This finding is consistent across different samples. Overall, these results indicate that (1) redundant variables are added into the system, more than a half of which are either impossible to estimate (short time series) or irrelevant; (2) significant dummies decrease forecast accuracy consistently, meaning that the baseline model performs reasonably well despite demand planners' perception.

Table 4.6: OLS Regression results: scaled AE as a dependent variable, parameters and std. errors in brackets.

(Intercept)	2.94 (0.13)***
Event1	−0.55 (0.81)
Event2	−2.59 (4.00)
Event3	−0.51 (3.66)
Event4	−1.03 (7.33)
Event5	−1.06 (4.33)
Event6	−1.29 (1.68)
Event7	−1.38 (10.78)
Event8	−0.69 (1.11)
Event9	−1.71 (15.05)
Event10	1.48 (0.57)**
Event11	−1.24 (4.66)
Event12	3.73 (20.76)
Event13	−0.17 (0.82)
Event14	1.81 (0.64)**
Event15	1.96 (0.56)***
Event16	3.84 (0.56)***
Event17	7.78 (0.56)***
Event18	9.09 (0.56)***
Event19	1.60 (0.56)**
Event20	1.36 (0.57)*
Event21	1.99 (0.56)***
Event22	0.05 (1.43)
Event23	−1.25 (0.96)
Promo1	−0.53 (2.09)
Promo2	9.60 (3.22)**
Promo3	9.44 (1.91)***
Promo4	7.51 (1.91)***
Promo5	3.92 (0.83)***
Promo6	−0.36 (1.09)
Promo7	0.52 (0.61)
Promo8	−0.21 (0.47)
log(Local_Adj)	101.27 (3.63)***
log(Global_Adj)	3.33 (1.29)**
Event1:log(Local_Adj)	−79.34 (10.84)***
Event2:log(Local_Adj)	−27.12 (26.12)
Event3:log(Local_Adj)	10.38 (108.08)
Event4:log(Local_Adj)	17.15 (44.82)
Event5:log(Local_Adj)	−112.19 (463.43)
Event6:log(Local_Adj)	−45.17 (15.26)**
Event8:log(Local_Adj)	9.59 (26.46)
Event10:log(Local_Adj)	49.14 (8.15)***
Event11:log(Local_Adj)	21.29 (44.05)
Event12:log(Local_Adj)	−5.79 (51.21)
Event13:log(Local_Adj)	−18.16 (10.15)
Event14:log(Local_Adj)	19.83 (18.57)
Event15:log(Local_Adj)	5.40 (10.66)
Event16:log(Local_Adj)	72.14 (7.17)***
Event17:log(Local_Adj)	235.01 (7.52)***
Event18:log(Local_Adj)	303.01 (6.04)***
Event19:log(Local_Adj)	13.45 (9.80)
Event20:log(Local_Adj)	1.39 (8.98)
Event21:log(Local_Adj)	−39.43 (10.96)***
Event22:log(Local_Adj)	5.84 (6.51)
Event23:log(Local_Adj)	1.01 (8.39)
Promo1:log(Local_Adj)	−99.19 (7.78)***
Promo2:log(Local_Adj)	−96.11 (12.66)***
Promo3:log(Local_Adj)	−92.19 (9.01)***
Promo4:log(Local_Adj)	−96.47 (9.28)***
Promo5:log(Local_Adj)	−87.41 (5.01)***
Promo6:log(Local_Adj)	−120.38 (7.32)***
Promo7:log(Local_Adj)	−127.41 (4.86)***
Promo8:log(Local_Adj)	−121.01 (4.12)***
Event1:log(Global_Adj)	10.51 (12.66)
Event2:log(Global_Adj)	−3.81 (20.15)
Event3:log(Global_Adj)	1.15 (16.72)
Event4:log(Global_Adj)	−4.64 (66.95)
Event5:log(Global_Adj)	2.74 (39.72)
Event6:log(Global_Adj)	−2.82 (8.21)
Event7:log(Global_Adj)	−9.56 (487.54)
Event8:log(Global_Adj)	−3.25 (18.51)
Event10:log(Global_Adj)	14.26 (3.64)***
Event11:log(Global_Adj)	−3.55 (32.74)
Event13:log(Global_Adj)	4.25 (7.38)
Event14:log(Global_Adj)	0.99 (8.17)
Event15:log(Global_Adj)	32.58 (4.37)***
Event16:log(Global_Adj)	71.27 (4.23)***
Event17:log(Global_Adj)	142.33 (3.61)***
Event18:log(Global_Adj)	153.66 (3.34)***
Event19:log(Global_Adj)	4.01 (3.31)
Event20:log(Global_Adj)	7.65 (5.60)
Event21:log(Global_Adj)	28.79 (5.71)***
Event22:log(Global_Adj)	−3.15 (12.70)
Event23:log(Global_Adj)	0.60 (5.52)
R^2	0.01
Adj. R^2	0.01
Num. obs.	900000

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

4.5.2 Judgmental adjustments

Since the distributions of local and global adjustments are strongly asymmetric (see the summary statistics in Table 4.3), we transform these variables by taking log-

arithms. This allows the interpretation of these parameters as 1% increase of local/global adjustments changes absolute errors by $\beta_{34}/100$ and $\beta_{35}/100$ units respectively. First, both parameters are positive, signalling a positive correlation between adjustments and their accuracy. Second, local adjustments have higher positive parameters than global ones, which is counter-intuitive given our previous descriptive statistics (e.g., Table 4.2), where local adjustments produced better accuracy on average than our baseline models. This is an interesting result that on average local adjustments lead to a decrease of accuracy. They perform better than the baseline exponential smoothing model, giving a marginal advantage on average across all products. For instance, all local adjustments during promotional periods are statistically significant and beneficial for accuracy, improving the final model substantially. Third, global adjustments are harmful in all cases, making the final forecasts worse. Therefore, based on these results, we conclude that global adjustments are dangerous and unnecessary, even though they are much easier to implement. This shows how important it is to understand where the forecast value added could be introduced by demand planners. In this case, generalisation of adjustments across different stores for a product does not improve the forecasts; on the contrary, it complicates the model evaluation process and harms the overall forecast accuracy.

4.5.3 Double judgments

The last category of judgmental interventions is double judgments which are nested adjustments, where model tuning is implemented first followed by adjustments to the resulting statistical forecasts. There are two sets of such interaction terms in the model for local and global adjustments. We observe that almost all significant interaction effects increase forecast errors, while only a few provide a modest benefit. However, the overall effect increase errors, highlighting the damage done by using double judgments. Based on this result, double adjustments are unreliable, harmful

and should be avoided.

4.6 Removing insignificant explanatory variables

In the previous section, we reported a problem of superfluous variables added to the forecasting model. For this reason, the next question is whether we can decrease the number of explanatory variables automatically using algorithms rather than human judgment only, checking their significance or predictive power (for example, using information criteria). This process is computationally expensive since we obtain model outcomes twice: (1) a simple exponential smoothing model with all variables; (2) with a subset of significant ones. The first step is required to identify the events' effects and extract corresponding parameters and standard errors to evaluate them. We use all 111 observations to fit and estimate the initial model, and then we extract coefficients and their significance based on p-values. For the final forecasts, we rerun $ETS + Promo + Events$ model from Table 4.4 to only include a subset of significant explanatory variables for promotions and events.

Table 4.7 provides the final results for a comparison between the company's final forecasts and ours. We find that this final model performs much better except for the local adjustments cases than both the company's model or any of our initial baseline models. However, the demand planners' negative local adjustments for promotional periods perform exceptionally well, and demonstrate the experience and domain knowledge of demand planners working with these products. this result reinforces our previous findings from the regression model, indicating a strong value added in these particular periods.

We find that only 11%, on average, of all event dummies across all products are significant at a 5% level. This is an important finding because it shows that we improve forecast accuracy on average once we remove insignificant judgmentally

Table 4.7: Comparison of four different methods, sRMSE.

	Company Forecast	ETS: +Promo	ETS+Promo: +Events	ETS+Promo: +SubsetEvents
Baseline	**95.93**	124.60	124.60	118.14
Baseline with Promotions	20.05	13.47	14.02	**13.38**
Local Adjustments	**83.76**	172.91	173.06	85.75
positive	**91.47**	189.85	190.01	93.25
negative	23.68	23.44	**23.43**	30.74
Local Adjustments for Promotions	**13.76**	18.42	19.75	18.19
positive	28.70	31.27	31.29	**26.29**
negative	**13.44**	18.18	19.54	18.06
Global Adjustments	89.48	19.07	18.52	**15.21**
positive	102.63	21.01	20.34	**16.29**
negative	21.22	11.45	11.47	**11.36**
Judgmental model tuning	121.80	41.75	41.75	**37.74**
Double judgment	267.07	62.32	62.08	**60.02**
positive	286.06	66.07	65.81	**63.56**
negative	86.30	30.88	30.85	**30.65**

imposed events. Therefore, it indicates that model tuning is subject to substantial biases that can harm its performance, yet promising as it scales better than judgmental adjustments of forecasts.

Interestingly, out of 40 various events, one would expect that some of them would be left out after optimising the model, but on the contrary, the results show that almost all events are significant for some products, indicating that demand planners add valuable information, but not precisely enough.

Figure 4.6 (left subplot) shows the percent of judgmentally modified time periods (out of 111 in total) for two models: (1) with all judgmental model tuning (light purple); (2) with the subset of significant ones (light pink), where a darker purple colour corresponds to an intersection between these two. We see that after selecting indicators there are still time series where around 60 out 111 time periods are judgmentally tuned, but the number of SKUs is significantly lower (from 40 to 18 thousands of

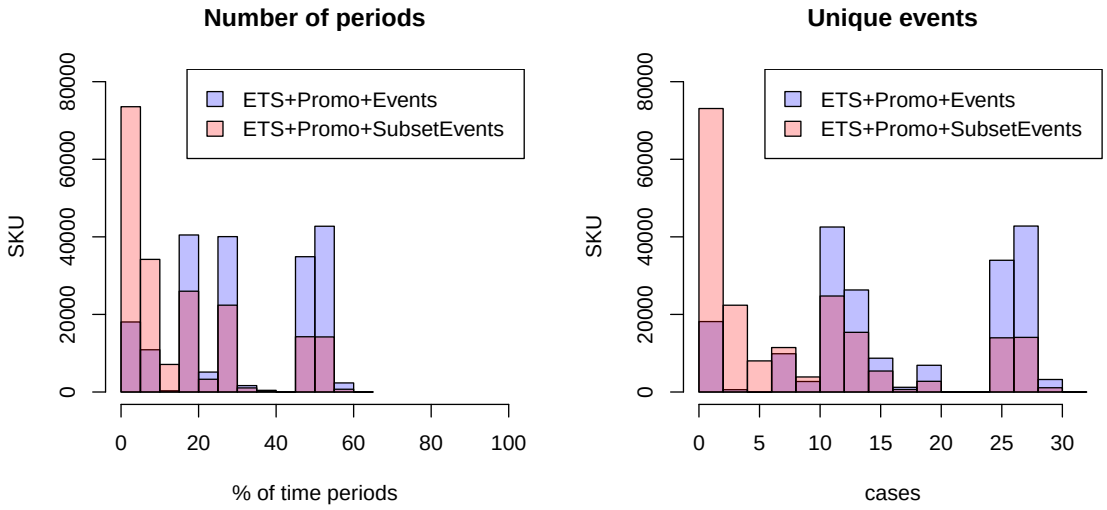

Figure 4.6: Histograms for number of periods adjusted and number of unique corrections per SKU/store by two methods: exponential model with all explanatory events or its subset.

SKUs). This means that experts are able to impose explanatory variables correctly, but for a smaller subset of products. There are no events that are not found significant at least for one time series. A similar picture can be observed on the right subplot, indicating that the variability indeed matters, but not for all products. For instance, even with the subset of significant events, 20 thousand of SKUs out of the 197 thousands still benefit from 25-30 unique events over two years of weekly data. Thus, demand planners are capable to extract and incorporate valuable information into a statistical model as model tuning for some products in particular. However, they tend to overuse it and pass this information to other products as well, diminishing the overall accuracy gains.

Overall, we find that demand planners have the potential to add important variables into the model. However, they do that inconsistently, for too many products and ignoring modelling constraints (a ratio of number of variables to a number of observations). Nonetheless, these issues are likely to be solvable when using both humans and computer systems as complementary agents in the demand planning process.

4.7 Discussion and conclusions

This chapter explored and analysed two different types of judgmental interventions in forecasting, where the first was model tuning and the second was adjustments of forecasts. While the latter is generally well documented and widely investigated in the literature, the former has never been investigated in detail before. In theory, judgmental model tuning should perform better than adjustments because it needs only the location of the events as an input. Then the effect is estimated statistically, making these interventions more scalable across different stores and products as well as unbiased in terms the estimation. Crucially, humans would only indicate periods when an event is happening, giving enough flexibility to a model to estimate this effect efficiently. In practice, we observed that demand planners in the case organisation overused this tool, resulting in overwhelming the model with redundant variables and too many parameters to estimate. It means that even though some of these judgmentally added variables were statistically significant and enhanced the baseline model, on average, their effects were undermined due to the number of these interventions. Nonetheless, we believe that the key to success lies in the organic combination of both humans and computer systems, trying to capitalise on the strengths of both. We note that model tuning is not a replacement to judgmental adjustments, though in many circumstances it has the potential to be more effective. Consider that model tuning estimates the size of a presumed effect on historical observations, while adjustments can incorporate unobserved ill-defined effects.

Considering that experts can extract and incorporate potentially valuable contextual information via judgmental model tuning, the next step in capitalising from this would be to enhance this process to make these interventions effective. For instance, a simple question such as "Do you expect this information of an event or a condition to influence your forecasts just once?" can recommend using either an adjustment for

one-off events or model tuning for reoccurring conditions. We can incorporate similar guidance to highlight the issue of overfitting. This would remind users to check the number of indicator variables in comparison to the sample size of the time series (can be one of the automatic features in support systems), aiming to reduce the number of redundant variables in the sample. Another possibility that we just touched upon earlier is grouping/classifying both time series and indicators, which might help identify the most impactful categories for experts to focus on.

Since judgmental model tuning and adjustments are sequential, the second question in this research was about the interaction between these interventions and their accuracy. We observed that the errors from the first could propagate to the second since a significant number of adjustments were made on previously judgmentally tuned observations. And as a result, the use of double judgment is particularly harmful to forecast accuracy since it exacerbates possible positive biases and overwhelms both the model and the forecaster. Especially, if the same person carries out each task, otherwise one can countermand the other. Based on the results from this case study, we would strongly recommend avoiding double adjustments, at least until model tuning interventions perform consistently well. At this point we also raise the question of what information is used in each of these interventions. In our study we did not have these details, yet it is reasonable to assume that for both types of interventions to be beneficial they need to introduce different and new information.

This study has confirmed some key findings from the previous case studies by Fildes et al. (2009) and Trapero et al. (2013). For instance, we observe that judgmental adjustments add value and increase forecast accuracy on average, but somewhat inconsistently. While local adjustments perform exceptionally well (especially negative ones for promotional periods), global adjustments (per SKU across all stores) do not achieve the expected result, emphasising the value of the low-level local corrections, where the most specific, hence useful, information can be obtained and incorpo-

rated into the forecasts. The findings also demonstrate that (1) negative adjustments perform better than positive ones; (2) large/substantive adjustments have more benefit than small ones; (3) the frequency of adjustments has not decreased over time comparing to previous case studies; (4) the perceived quality of the baseline system model can lead to a high number of manual adjustments. The latter is a current direction of the research in many disciplines exploring decision-making in practice, such as behavioural operational research, computer science, behavioural economics and psychology (Kunc, 2019).

Surprisingly, local adjustments perform better than the baseline regression, suggesting that either experts have essential information or the regression outputs are consistently biased. Notably, local adjustments outperform model tuning for this case study: to understand why it would require qualitative investigation, interviewing demand planners on the process of making these judgmental interventions, the qualitative information and cues behind them, and possibly tracking their performance over time.

Our research is not without its limitations. First, the company's system forecasts are unavailable. We made assumptions to obtain our alternative models for the baseline forecasts. As far as we know, demand planners cannot access the historical system forecasts either. Second, all judgmental interventions are done at a daily level, while all sales and forecasts are stored and generated at a weekly level. This creates a coherency problem between these levels. This is managed by the system internally using daily profiles for each individual product. Once again, these weights were not available, so we had to work at a weekly level.

This chapter makes several contributions; first, it shows how complex judgmental interventions are done in practice, identifying a hierarchy of adjustments with several variations within each of them. Second, despite the many theoretical advantages of judgmental model tuning, it performs relatively poorly due to its misuse. Third,

we show that we can increase the final forecast accuracy by removing insignificant judgmentally imposed events. This suggests that humans can still add value and identify important additional events.

Appendix 4.A Bias and accuracy metrics

The scaled Error (sE) measure for the series i and horizon h is calculated as:

$$\text{sE}_{i,h} = \frac{Actual_{N+h} - Forecast_h}{\frac{1}{N}\sum_{t=1}^{N} Actual_t}, \tag{4.2}$$

where N is the number of in-sample observations, $Actual_{N+h}$ is the actual value of the hth out-of-sample period and $Forecast_h$ is the h-steps-ahead forecast (Petropoulos and Kourentzes, 2015).

The scaled Absolute Error (sAE) measure for the series i and horizon h is determined as:

$$\text{sAE}_{i,h} = \frac{|Actual_{N+h} - Forecast_h|}{\frac{1}{N}\sum_{t=1}^{N} Actual_t}. \tag{4.3}$$

Similarly, the scaled Squared Error (sSE) is defined as:

$$\text{sSE}_{i,h} = \left(\frac{Actual_{N+h} - Forecast_h}{\frac{1}{N}\sum_{t=1}^{N} Actual_t}\right)^2, \tag{4.4}$$

while the scaled Mean Squared Error (sMSE) is calculated as the average of sSE across all series and horizons. Based on that, we can calculate sRMSE that brings the error to the origin unit scale.

Chapter 5

Discussion and conclusions

This doctoral thesis has investigated the effect and modes of presentation of contextual information in forecasting. In particular, we focused on the combination of judgment and model inputs, trying to identify the effective use of those, taking into account advances in computer systems, statistical modelling and understanding of human behaviour. Knowing both strengths and weaknesses of human decision-making, we attempted to investigate the conditions under which forecasters add value despite many well-known cognitive biases and heuristics (Kahneman, 2012; Tversky and Kahneman, 1974).

First, we set out to conduct an online experiment to integrate contextual information of varying predictive value into a Forecasting Support System. This idea was inspired by Fildes et al. (2019a), the first study that used both time series and qualitative information simultaneously on the same screen. However, the authors implemented a simplified design, providing an initial picture of how forecasters might manage such combination of information in a controlled experimental environment. They found that contextual information of unknown diagnosticity was detrimental to forecast accuracy. Fildes et al. (2019a) emphasised the value of filtering and providing only relevant qualitative information, a process that needs to be a part of a forecasting system. We pursued this idea and designed an experiment with a mix of contextual statements with and without any predictive value, expecting forecasters to identify and use only the relevant pieces.

This mix of quantitative and qualitative information of different quality depicts imperfect information sharing in organisations that is ingrained in the Sales and Operations Planning (S&OP) process (Chen, 2003; Cui et al., 2015). One can argue that a trivial concept of "garbage in, garbage out" would easily predict an outcome for such information. Nevertheless, this is precisely what we observe in practice - various pieces of information of unknown predictive value are collected from multiple sources such as calls, discussions, reports, the Internet, etc., and then considered in the forecasting and decision-making processes. Hence, we argue that such experiments can help to identify the conditions where forecasters are prone to make mistakes or, on the contrary, can add value.

The design of this experiment was largely inspired by a case study that we conducted shortly before the first experimental blueprint. This acquaintance with the case company's processes persuaded us to use (1) promotional modelling as the main motivation for the judgmental adjustments of statistical forecasts (one of the main reasons for interventions according to Fildes and Goodwin, 2007a); (2) both baseline and promotional models to produce system forecasts; (3) both optimistic and pessimistic contextual information from different contexts. This also matches previous surveys on forecast adjustment causes (Sanders and Manrodt, 2003; Fildes and Goodwin, 2007a; Boulaksil and Franses, 2009; Weller and Crone, 2012; Fildes and Petropoulos, 2015). We observed that demand planners were prone to incorporate their personal beliefs into the final forecasts, which motivated us to include a novel category of qualitative information - overly optimistic statements (unknown to the subjects, these were without predictive value). Despite the apparent distraction by the latter, we noticed that forecasters could determine valuable pieces of qualitative information in treatments with stronger promotional effects. This finding is important indicating that forecasters (non-experts) could potentially process quantitative and qualitative information of different quality simultaneously in this overwhelming

yet realistic setting. This opens new avenues for cross-disciplinary studies on the effective presentation of these types of information to the user. For instance, we need to identify ways to pre-process qualitative information first, trying to avoid overly emotional statements or to carefully classify the statements into groups to reduce the task complexity.

In our first experiment (Chapter 2), we presented all qualitative statements simultaneously, and found signs of information overload that may have resulted in poorer overall performance and user experience. Hence, in Chapter 3, we altered how textual information was presented, adding structure by decomposing these statements into subgroups and independent tabs on the screen. Using the initial design as a control group, we questioned whether decomposition could reduce the information load and, as a result, could help users to identify contextual information with predictive value and weight it accordingly. The results showed that the decomposition reduced the number and the size of forecast adjustments in all treatments. The latter result was unexpected and could be explained in two ways: the decomposition restricted users naturally, or alternatively the users accepted models more often under this condition. We argued that it was the former since this effect was observed even when the model was clearly erroneous. We concluded that the decomposed presentation of qualitative statements performed better than the control setting, even though subjects remained prone to weight overly positive statements despite their unknown predictive value. In other elements, we retained the same experimental setup that was representative of the case company reported in Chapter 2, and we only varied how qualitative information was presented. This design follows the representative design concept proposed by Brunswik (1955) that is widely adopted for empirical and experimental studies with human judgment.

Both Chapters 2 and 3 highlight a problem of little knowledge of how forecasters use contextual information in their decision-making process. Only a few studies have

analysed the use of textual information employing either empirical or experimental methods (Sanders and Ritzman, 1992; Webby et al., 2005; Fildes et al., 2018). Our studies open up the research avenue of investigating the value of such information, its use and possible pitfalls from software or company sides. For instance, there are two main issues with the current processes: (1) insufficient support of judgmental adjustments in FSSs; (2) unstructured utilisation of available contextual information from different sources in organisations, which is itself unstructured. The possible implications will be discussed in the following subsection of practical and research implications.

While conducting the case study described in Chapter 2, we noticed a new, unreported way of incorporating expert experience and additional knowledge into statistical forecasts by making changes at the model building stage that we defined as *judgmental model tuning*. In the company case study, we observed the addition of indicator variables for repeating events, which were not accounted for by the statistical model in the demand planners' opinion. In Chapter 4, we explored and compared two types of judgmental interventions: model tuning and forecast adjustments. We expected that judgmental model tuning would be a good alternative to forecast adjustments and would perform better since it requires inputting only the location of the events whose impacts were later estimated statistically by the enhanced model. However, the results showed that demand planners overused this instrument, saturating the model with redundant variables. Some indicator variables were found to be useful and beneficial, but their effect was diminished by the number of harmful indicators. We also found that negative adjustments perform better than positive ones, and more substantive adjustments have more benefit than minor "tweaking" ones, confirming findings by Fildes et al. (2009) and Trapero et al. (2013). While judgmental model tuning might appear to be a good alternative to judgmental adjustments, it is not a replacement for the latter since not every contextual piece of information can be

translated into an indicator variable.

5.1 Practical and research implications

In an age of advanced computer systems and artificial intelligence approaching from different corners, individuals and organisations are keen to improve their operational processes, and, ultimately, to build a better, more sustainable future. This research recognises that judgment can bring valuable inputs into the forecasting processes. Our research provides a new perspective on the use of human judgment in forecasting by investigating different ways of presenting and incorporating textual information into statistical forecasts. This direction involves several types of stakeholders, namely, demand planners and managers, software vendors and researchers. We discuss implications and practicalities for each in turn.

5.1.1 Demand planners and managers

The case study analysis revealed a high frequency of judgmental interventions in the company (supporting other similar findings by (Fildes et al., 2009; Syntetos et al., 2016a; Fildes et al., 2019b) that are based on unstructured, often subjective contextual information. The results of these interventions are rarely tracked and evaluated, and a big challenge remains that many FSSs do not support this process, as was the case for the case company. This indicates a problem of sub-optimal planning processes, where any judgmental interventions require more time and effort that leads to labour intensive adjustments of thousands of products. Hence, by analysing and finding ways to consolidate both judgment and quantitative methods, we can counterbalance their independent shortcomings and increase forecast accuracy and reliability, ultimately, bringing about better practices (Zellner et al., 2021).

Nevertheless, the first step could be collecting and recording all contextual state-

ments used in the decision-making process, aiming to analyse their practicality. The second step would be to structure the process as much as possible, so one would be able to trace back all implemented changes both during the modelling process and at the later stages. Then these inputs (all contextual information and the steps) could be incorporated into the Forecasting Support Systems by software vendors. Finally, there is overwhelming evidence that users can benefit from software that is attentive to interface design issues.

5.1.2 Software vendors

Given that statistical forecasts can capture the underlying components of a time series such as level, trend and seasonality, forecasters can focus on either the most valuable or unusual products or embody any other information (e.g., weather, competitors, macroeconomic factors) that might impact product sales (Fildes and Goodwin, 2007a; Boulaksil and Franses, 2009). Asimakopoulos et al. (2011) showed how unstructured and non-linear the forecasting process is, as well as how many stages/steps it includes, especially connected to additional qualitative information. Fildes and Goodwin (2021) explored the FSS impact and use in the business environment over fifteen years, showing two extremes on the same case example: forecasters can abuse the system, and the system can over-restrict expert interventions. This situation suggests that we still are far away from the efficient use of both support systems and the human judgments made within them.

Despite many known cognitive issues that might arise when using human judgment, we argue that FSSs need to be redesigned to take into account judgmental interventions. The inclusion of judgment in FSSs aims not only to acknowledge the role of forecasters, but also to facilitate its effective use, allowing for tracking, evaluating and aiding forecasters in this process. The reports of the latest surveys about FSSs (Fildes et al., 2018, 2020) show that not much has been improved since Fildes

et al. (2006), where the authors extensively discussed what features are essential in a good FSS. For instance, our own example of a screenshot of an FSS in the case study (see Figure 2 in Chapter 2) shows that the FSS interface is not informative, sometimes either inexplicably complex or, on the contrary, overly simplified. Therefore, it is practically impossible to track and evaluate such judgmental interventions without appropriate technical support.

Even though demand planners are motivated to assess the final forecast accuracy, we observe that they are reluctant to do so since it makes their job even harder (e.g., extra steps to transfer all the data outside the system to calculate any forecast errors or other KPIs). Surprisingly, neither researchers nor software vendors discuss these issues in forecasting. Various interface designs and the effective use of "soft" information in it has the potential to improve such situations tremendously. In Chapter 2 and 3, we highlight some evident issues: (1) a presentation of contextual information; (2) its classification; (3) the reliability of the statistical (system) model; (4) estimation of forecast value added (FVA) and providing feedback (especially given that forecasters might change a model form as we explore in Chapter 4). These questions require more cross-disciplinary studies (e.g., with insights drawn from computer science, human-computer interactions, human trust in algorithms and possible explainable artificial intelligence) and involvement of software vendors for optimal results in practice.

5.1.3 Researchers

Both Chapter 2 and 3 highlight the importance of a multi-methodological approach in research (Choi et al., 2016). First, we interviewed demand planners in a retailing company, observed the forecasting process, collected and analysed the data with regards to the overall accuracy and judgmental interventions. This case study provided many insights on how complex this process is, emphasising the role of forecasters in it. Then, we developed a controlled experiment to assess the use of contextual infor-

mation simulating the FSS used in the company but adding salient features to test our hypotheses. Such a mix of methodologies facilitates not only multiple points of view on a problem, but also provide grounded assumptions in further experimental studies (Singhal and Singhal, 2011a,b). It is especially essential in behavioural operations where there is a tendency for controlled experiments to be used for many contexts (Bendoly et al., 2006), with slow adoption of more diverse methodologies (Croson et al., 2012). The barriers are substantial: as Choi et al. (2016) outlined, a multi-methodological approach is more demanding, time-consuming, expensive and inessential for publishing. Yet, the authors are optimistic about its adoption in a wider scientific society soon.

Also, scholars need to explore non-conventional ways to incorporate "soft" unstructured information into algorithmic methods. For instance, there is evidence that Artificial Intelligence (AI) can process non-structured contextual information using pattern tracking, image and natural language processing methods (Ittoo et al., 2016; Kreimeyer et al., 2017; Zhang et al., 2020). This could be a pre-processing step to embody contextual information into the final forecasts. However, the collection and preparation of such information would need additional tools and time. Yet, it is a promising for supply-chain management as a whole (Beheshti-Kashi et al., 2019).

5.2 Limitations and future work

This thesis has attempted to look at the use of contextual information in detail, trying to capture some of the complexities that arise when forecasters use modern computer support systems. By adopting case study, analytical and experimental methodologies, we have investigated this question from different perspectives. Nevertheless, there are several important limitations to this research. First, we employed only novice (student and non-student) participants (Chapter 2 and 3), although initially, we aimed to utilise

our company contacts to include experts as a separate group of participants. Due to the lack of involvement and resources, we were not able to collect an appropriate sample. Hence, there is an open question about the difference in involvement between participants and demand planners, that needs to be checked either using extended experiments or field studies (Siemsen, 2011; Fildes et al., 2019a). Second, we made several essential assumptions in the data analysis of the case study (Chapter 2 and 4), such as the initial baseline model form (trying to simulate historical system forecasts), a choice of common time buckets and the corresponding transformations. Third, both case study and laboratory experiments are very context, data and case specific. We need more evidence for better understanding of the conditions. For instance, we moved closer to reality, but still looked at a narrow case of context - judgmental forecast adjustments in retailing. This made it easier for participants to connect to, but also ignored very specific dynamics connected to this sector (e.g., organisational politics).

Our research opens multiple avenues for various multi-disciplinary research questions. For instance, there is a popular notion of "algorithm aversion", a situation where humans show their reluctance to take advice generated by a machine algorithm or quantitative model, introduced by Dietvorst et al. (2015), which has been challenged by the "algorithm appreciation" idea (Logg et al., 2019). Basically, both ideas refer to a bigger question of human trust in algorithmically derived forecasts. And there are many factors that may dominate a final decision (Dietvorst et al., 2018; Kaufmann and Budescu, 2019; Yeomans et al., 2019). This issue of trust is especially crucial with "black-box" algorithms such as machine learning (ML) and artificial intelligence (AI) (Glikson and Woolley, 2020). We indirectly touched upon model acceptance and trust in algorithms in Chapter 3 (Alvarado-Valencia and Barrero, 2014), but more studies are needed to draw any implications of human trust in models in forecasting.

This research shows how vital support systems are, and we believe that it is possible to extract and combine the best of both judgmental and quantitative methods in forecasting and demand planning. Identifying and analysing the best practices of utilising contextual information and trustworthy algorithms in a reliable FSS is the ultimate goal of this direction (Fildes, 2017). This thesis is just a starting point in this journey, where many more questions need to be investigated. For instance, further studies on the use of judgment in forecasting might focus not only on simple, well-established statistical models but also include advanced ML methods that are able to model textual information as well. Together with insights from computer systems, there are strong prospects to develop reliable and sustainable organisational practices for both forecasters and support systems. Success would lead to major efficiency saving in the many areas where judgment is combined with the formal model-based forecasts.

Bibliography

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.

Alvarado-Valencia, J. A., Barrero, L. H., 2014. Reliance, trust and heuristics in judgmental forecasting. Computers in Human Behavior 36, 102–113.

Armstrong, J., Denniston, W. B., Gordon, M. M., 1975. The use of the decomposition principle in making judgments. Organizational Behavior and Human Performance 14 (2), 257–263.

Arvan, M., Fahimnia, B., Reisi, M., Siemsen, E., 2019. Integrating human judgement into quantitative forecasting methods: A review. Omega 86, 237–252.

Asimakopoulos, S., Dix, A., Fildes, R., 2011. Using hierarchical task decomposition as a grammar to map actions in context: Application to forecasting systems in supply chain planning. International Journal of Human-Computer Studies 69 (4), 234–250.

Baškarada, S., 2014. Qualitative case study guidelines. The Qualitative Report 19 (40), 1–18.

Bates, D., Mächler, M., Bolker, B., Walker, S., 2015. Fitting linear mixed-effects models using lme4. Journal of Statistical Software 67 (1), 1–48.

Bathcelor, R., Dua, P., 1990. Forecaster ideology, forecasting technique, and the accuracy of economic forecasts. International Journal of Forecasting 6 (1), 3–10.

Batini, C., Cappiello, C., Francalanci, C., Maurino, A., 2009. Methodologies for data quality assessment and improvement. ACM Computing Surveys 41 (3), 1–52.

Batini, C., Scannapieco, M., 2016. Data and Information Quality. Springer International Publishing.

Beheshti-Kashi, S., Pannek, J., Kinra, A., 2019. Complementing decision support and forecasting risk in supply chain with unstructured data. IFAC-PapersOnLine 52 (13), 1721–1726.

Belton, V., Goodwin, P., 1996. Remarks on the application of the analytic hierarchy process to judgmental forecasting. International Journal of Forecasting 12 (1), 155–161.

Bendoly, E., Donohue, K., Schultz, K. L., 2006. Behavior in operations management: Assessing recent findings and revisiting old assumptions. Journal of Operations Management 24 (6), 737–752.

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

Boone, T., Boylan, J. E., Fildes, R., Ganeshan, R., Sanders, N., 2019. Perspectives on supply chain forecasting. International Journal of Forecasting 35 (1), 121–127.

Boulaksil, Y., Franses, P. H., 2009. Experts' stated behavior. Interfaces 39 (2), 168–171.

Bovi, M., 2009. Economic versus psychological forecasting. Evidence from consumer confidence surveys. Journal of Economic Psychology 30 (4), 563–574.

Brown, L. D., 1993. Earnings forecasting research: its implications for capital markets research. International Journal of Forecasting 9 (3), 295–320.

Brunswik, E., 1955. Representative design and probabilistic theory in a functional psychology. Psychological Review 62 (3), 193–217.

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

Bunn, D. W., Salo, A. A., 1996. Adjustment of forecasts with model consistent expectations. International Journal of Forecasting 12 (1), 163–170.

Camerer, C. F., Hogarth, R. M., 1999. The effects of financial incentives in experiments: A review and capital-labor-production framework. Journal of Risk and Uncertainty 19 (1), 7–42.

Chang, W., Cheng, J., Allaire, J., Xie, Y., McPherson, J., 2017. shiny: Web Application Framework for R. R package version 1.0.5.
URL https://CRAN.R-project.org/package=shiny

Cheikhrouhou, N., Marmier, F., Ayadi, O., Wieser, P., 2011. A collaborative demand forecasting process with event-based fuzzy judgements. Computers & Industrial Engineering 61 (2), 409–421.

Chen, F., 2003. Information sharing and supply chain coordination. In: Supply Chain Management: Design, Coordination and Operation. Elsevier, pp. 341–421.

Choi, T.-M., Cheng, T. C. E., Zhao, X., 2016. Multi-methodological research in operations management. Production and Operations Management 25 (3), 379–389.

Clements, M. P., 1995. Rationality and the role of judgement in macroeconomic forecasting. The Economic Journal 105 (429), 410.

Cook, M. P., 2006. Visual representations in science education: The influence of prior knowledge and cognitive load theory on instructional design principles. Science Education 90 (6), 1073–1091.

Croson, R., Schultz, K., Siemsen, E., Yeo, M., 2012. Behavioral operations: The state of the field. Journal of Operations Management 31 (1-2), 1–5.

Cui, R., Allon, G., Bassamboo, A., Mieghem, J. A. V., 2015. Information sharing in supply chains: An empirical and theoretical valuation. Management Science 61 (11), 2803–2824.

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.

De Baets, S., Harvey, N., 2018. Forecasting from time series subject to sporadic perturbations: Effectiveness of different types of forecasting support. International Journal of Forecasting 34 (2), 163–180.

De Baets, S., Harvey, N., 2020. Using judgment to select and adjust forecasts from statistical models. European Journal of Operational Research 284 (3), 882–895.

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), 114–126.

Dietvorst, B. J., Simmons, J. P., Massey, C., 2018. Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. Management Science 64 (3), 1155–1170.

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

Durbach, I. N., Montibeller, G., 2018. Behavioural analytics: Exploring judgments and choices in large data sets. Journal of the Operational Research Society 70 (2), 255–268.

Edmundson, B., Lawrence, M., O'Connor, M., 1988. The use of non-time series information in sales forecasting: A case study. Journal of Forecasting 7 (3), 201–211.

Edmundson, R. H., 1990. Decomposition: a strategy for judgemental forecasting. Journal of Forecasting 9 (4), 305–314.

Eroglu, C., Croxton, K. L., 2010. Biases in judgmental adjustments of statistical forecasts: The role of individual differences. International Journal of Forecasting 26 (1), 116–133.

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

Fildes, R., 2017. Research into forecasting practice. Foresight: The International Journal of Applied Forecasting 44, 39–46.

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

Fildes, R., Goodwin, P., 2007b. Good and bad judgment in forecasting: lessons from four companies. Foresight: The International Journal of Applied Forecasting 8 (8), 5–10.

Fildes, R., Goodwin, P., 2013. Forecasting support systems: What we know, what we need to know. International Journal of Forecasting 29 (2), 290–294.

Fildes, R., Goodwin, P., 2021. Stability in the inefficient use of forecasting systems: A case study in a supply chain company. International Journal of Forecasting 37 (2), 1031–1046.

Fildes, R., Goodwin, P., Lawrence, M., 2006. The design features of forecasting support systems and their effectiveness. Decision Support Systems 42, 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., 2019a. Use and misuse of information in supply chain forecasting of promotion effects. International Journal of Forecasting 35 (1), 144–156.

Fildes, R., Ma, S., Kolassa, S., 2019b. Retail forecasting: Research and practice. International Journal of Forecasting.

Fildes, R., Petropoulos, F., 2015. How to improve forecast quality: A new survey. Foresight: The International Journal of Applied Forecasting 36, 5–12.

Fildes, R., Schaer, O., Svetunkov, I., 2018. Forecasting 2018. OR/MS Today 46 (3), 44–46.

Fildes, R., Schaer, O., Svetunkov, I., Yusupova, A., 2020. Survey: What's new in forecasting software? OR/MS Today 47 (4), 1–17.

Fosso Wamba, S., Akter, S., Edwards, A., Chopin, G., Gnanzou, D., 2015. How 'big data' can make big impact: Findings from a systematic review and a longitudinal case study. International Journal of Production Economics 165, 234 – 246.

Franco, L. A., Hämäläinen, R. P., 2016. Behavioural operational research: Returning to the roots of the OR profession. European Journal of Operational Research 249 (3), 791–795.

Franses, P. H., 2013. Improving judgmental adjustment of model-based forecasts. Mathematics and Computers in Simulation 93, 1–8.

Franses, P. H., 2014. Expert Adjustments of Model Forecasts : Theory, Practice and Strategies for Improvement. Cambridge: Cambridge University Press.

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., 2010. Do experts' adjustments on model-based SKU-level forecasts improve forecast quality? Journal of Forecasting 29, 331–340.

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

Fry, C., Mehrotra, V., 2016. Forecasting 2016 (software survey). OR/MS Today 43 (3), 44–54.

Galbraith, C. S., Merrill, G. B., 1996. The politics of forecasting: Managing the truth. California Management Review 38 (2), 29–43.

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

Gigerenzer, G., 1999. Simple heuristics that make us smart. Evolution and cognition. Oxford University Press, New York.

Gilliland, M., 2010. The business forecasting deal: exposing myths, eliminating bad practices, providing practical solutions. Vol. 27 of Wiley and SAS Business Series. Wiley.

Glikson, E., Woolley, A. W., 2020. Human trust in artificial intelligence: Review of empirical research. Academy of Management Annals 14 (2), 627–660.

Goldstein, H., 2011. Multilevel statistical models, 4th Edition. Wiley series in probability and statistics. Wiley, Hoboken, N.J.

Goodwin, P., 2000. Improving the voluntary integration of statistical forecasts and judgment. International Journal of Forecasting 16 (1), 85–99.

Goodwin, P., Fildes, R., 1999. Judgmental forecasts of time series affected by special events: does providing a statistical forecast improve accuracy? Journal of Behavioral Decision Making 12 (1), 37–53.

Goodwin, P., Fildes, R., Lawrence, M., Stephens, G., 2011. Restrictiveness and guidance in support systems. Omega 39 (3), 242–253.

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.

Harvey, N., 1995. Why are judgments less consistent in less predictable task situations? Organizational behaviours and human decision processes 63 (3), 247–263.

Harvey, N., 2007. Use of heuristics: Insights from forecasting research. Thinking & Reasoning 13 (1), 5–24.

Harvey, N., Harries, C., Fischer, I., 2000. Using advice and assessing its quality. Organizational Behavior and Human Decision Processes 81 (2), 252–273.

Hazen, B. T., Boone, C. A., Ezell, J. D., Jones-Farmer, L. A., 2014. Data quality for data science, predictive analytics, and big data in supply chain management: An introduction to the problem and suggestions for research and applications. International Journal of Production Economics 154, 72–80.

Hogarth, R. M., Makridakis, S., 1981. Forecasting and planning: An evaluation. Management Science 27 (2), 115.

Hong, T., Pinson, P., Fan, S., 2014. Global energy forecasting competition 2012. International Journal of Forecasting 30 (2), 357–363.

Hong, T., Pinson, P., Fan, S., Zareipour, H., Troccoli, A., Hyndman, R. J., 2016.

Probabilistic energy forecasting: Global energy forecasting competition 2014 and beyond. International Journal of Forecasting 32 (3), 896–913.

Hyndman, R. J., 2020. A brief history of forecasting competitions. International Journal of Forecasting 36 (1), 7–14.

Hyndman, R. J., Koehler, A. B., Snyder, R. D., Grose, S., 2002. A state space framework for automatic forecasting using exponential smoothing methods. International Journal of Forecasting 18 (3), 439–454.

Ittoo, A., Nguyen, L. M., van den Bosch, A., 2016. Text analytics in industry: Challenges, desiderata and trends. Computers in Industry 78, 96–107.

Kahneman, D., 2012. Thinking, fast and slow. Penguin, London.

Katok, E., 2018. Designing and Conducting Laboratory Experiments. John Wiley & Sons, Ltd, Ch. 1, pp. 1–33.

Kaufmann, E., Budescu, D. V., 2019. Do teachers consider advice? On the acceptance of computerized expert models. Journal of Educational Measurement 57 (2), 311–342.

Kolassa, S., 2019. Forecasting the future of retail forecasting. Foresight: The International Journal of Applied Forecasting (52), 11–19.

Kourentzes, N., Barrow, D., Petropoulos, F., 2019. Another look at forecast selection and combination: Evidence from forecast pooling. International Journal of Production Economics 209, 226–235.

Kourentzes, N., Fildes, R. A., 2021. The dynamics of judgemental adjustments in demand planning, available at SSRN: https://ssrn.com/abstract=3784616.

Kourentzes, N., Petropoulos, F., 2016. Forecasting with multivariate temporal aggregation: The case of promotional modelling. International Journal of Production Economics 181, 145–153.

Kourentzes, N., Trapero, J. R., Barrow, D. K., 2020. Optimising forecasting models for inventory planning. International Journal of Production Economics 225, 107597.

Kraus, M., Feuerriegel, S., Oztekin, A., 2020. Deep learning in business analytics and operations research: Models, applications and managerial implications. European Journal of Operational Research 281 (3), 628–641.

Kreimeyer, K., Foster, M., Pandey, A., Arya, N., Halford, G., Jones, S. F., Forshee, R., Walderhaug, M., Botsis, T., 2017. Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review. Journal of Biomedical Informatics 73, 14–29.

Kunc, M., 2019. Behavioral operations and behavioral operational research: Similarities and differences in competences and capabilities. In: Behavioral Operational Research. Springer International Publishing, pp. 3–22.

Lawrence, M., 1993. The M2-competition: Some personal views. International Journal of Forecasting 9 (1), 25–26.

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

Lawrence, M., O'Connor, M., Edmundson, B., 2000. Field study of sales forecasting accuracy and processes. European Journal of Operational Research 122 (1), 151–160.

Lawrence, M. J., Edmundson, R. H., O'Connor, M. J., 1985. An examination of the accuracy of judgmental extrapolation of time series. International Journal of Forecasting 1 (1), 25–35.

Lee, W. Y., Goodwin, P., Fildes, R., Nikolopoulos, K., Lawrence, M., 2007. Providing support for the use of analogies in demand forecasting tasks. International Journal of Forecasting 23 (3), 377–390.

Lee, Y. S., Seo, Y. W., Siemsen, E., 2018. Running behavioral operations experiments using Amazon's mechanical turk. Production and Operations Management 27 (5), 973–989.

Legerstee, R., Franses, P. H., 2014. Do experts' SKU forecasts improve after feedback? Journal of Forecasting 33 (1), 69–79.

Libby, R., 1975. Accounting ratios and the prediction of failure: Some behavioral evidence. Journal of Accounting Research 13 (1), 150.

Libby, R., Lewis, B. L., 1982. Human information processing research in accounting: The state of the art in 1982. Accounting, Organizations and Society 7 (3), 231–285.

Libby, R., Rennekamp, K., 2012. Self-serving attribution bias, overconfidence, and the issuance of management forecasts. Journal of Accounting Research 50 (1), 197–231.

Lim, J. S., O'Connor, M., 1995. Judgemental adjustment of initial forecasts: Its effectiveness and biases. Journal of Behavioral Decision Making 8 (3), 149–168.

Logg, J. M., Minson, J. A., Moore, D. A., 2019. Algorithm appreciation: People prefer algorithmic to human judgment. Organizational Behavior and Human Decision Processes 151, 90–103.

Makridakis, S., Chatfield, C., Hibon, M., Lawrence, M., Mills, T., Ord, K., Simmons,

L. F., 1993. The M2-competition: A real-time judgmentally based forecasting study. International Journal of Forecasting 9 (1), 5–22.

Makridakis, S., Hibon, M., 2000. The M3-competition: results, conclusions and implications. International Journal of Forecasting 16 (4), 451–476.

Makridakis, S., Spiliotis, E., Assimakopoulos, V., 2020. The M4 competition: 100,000 time series and 61 forecasting methods. International Journal of Forecasting 36 (1), 54–74, m4 Competition.

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., Diamantopoulos, A., 1989. Judgemental revision of sales forecasts: A longitudinal extension. Journal of Forecasting 8 (2), 129–140.

McCarthy, T. M., Davis, D. F., Golicic, S. L., Mentzer, J. T., 2006. The evolution of sales forecasting management: a 20-year longitudinal study of forecasting practices. Journal of Forecasting 25 (5), 303–324.

McNees, S. K., 1990. The role of judgment in macroeconomic forecasting accuracy. International Journal of Forecasting 6 (3), 287–299.

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

Nelson, M., Tan, H.-T., 2005. Judgment and decision making research in auditing: A task, person, and interpersonal interaction perspective. AUDITING: A Journal of Practice & Theory 24 (s-1), 41–71.

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

Oliva, R., Watson, N., 2009. Managing functional biases in organizational forecasts: A case study of consensus forecasting in supply chain planning. Production and Operations Management 18 (2), 138–151.

Oliva, R., Watson, N., 2010. Cross-functional alignment in supply chain planning: A case study of sales and operations planning. Journal of Operations Management 29 (5), 434–448.

Önkal, D., Gönül, S., 2005. Judgmental adjustment: A challenge for providers and users of forecasts. Foresight: The International Journal of Applied Forecasting 1 (1), 13–17.

Önkal, D., Goodwin, P., Thomson, M., Gönül, S., Pollock, A., 2009. The relative influence of advice from human experts and statistical methods on forecast adjustments. Journal of Behavioral Decision Making 22 (4), 390–409.

Ord, K., Fildes, R., Kourentzes, N., 2017. Principles of Business Forecasting, 2nd Edition. Wessex Press Publishing Co, New York.

Pennings, C. L., van Dalen, J., Rook, L., 2019. Coordinating judgmental forecasting: Coping with intentional biases. Omega 87, 46–56.

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., 2019. Judgmental model selection. Foresight: the International Journal of Applied Forecasting 2019 (54), 4–10.

Petropoulos, F., Goodwin, P., Fildes, R., 2017. Using a rolling training approach to improve judgmental extrapolations elicited from forecasters with technical knowledge. International Journal of Forecasting 33 (1), 314–324.

Petropoulos, F., Kourentzes, N., 2015. Forecast combinations for intermittent demand. Journal of the Operational Research Society 66 (6), 914–924.

Petropoulos, F., Kourentzes, N., Nikolopoulos, K., 2016. Another look at estimators for intermittent demand. International Journal of Production Economics 181 (3), 154–161.

Petropoulos, F., Kourentzes, N., Nikolopoulos, K., Siemsen, E., 2018. Judgmental selection of forecasting models. Journal of Operations Management 60 (1), 34–46.

Prahl, A., Van Swol, L., 2017. Understanding algorithm aversion: When is advice from automation discounted? Journal of Forecasting 36 (6), 691–702.

Qiu, L., Pang, J., Lim, K. H., 2012. Effects of conflicting aggregated rating on eWOM review credibility and diagnosticity: The moderating role of review valence. Decision Support Systems 54 (1), 631–643.

R Core Team, 2017. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
URL https://www.R-project.org/

Ramnath, S., Rock, S., Shane, P., 2008. The financial analyst forecasting literature: A taxonomy with suggestions for further research. International Journal of Forecasting 24 (1), 34–75.

Ransbotham, S., Kiron, D., Gerbert, P., Reeves, M., 2017. Reshaping business with artificial intelligence: Closing the gap between ambition and action. MIT Sloan Management Review 59 (1).

Redman, T., 2001. Data Quality: The Field Guide. Data management series. Elsevier Science.

Redman, T. C., Godfrey, A. B., 1997. Data Quality for the Information Age, 1st Edition. Artech House, Inc., USA.

Reimers, S., Harvey, N., 2011. Sensitivity to autocorrelation in judgmental time series forecasting. International Journal of Forecasting 27 (4), 1196–1214.

Roth, A., Rosenzweig, E., 2020. Advancing empirical science in operations management research: A clarion call to action. Manufacturing and Service Operations Management 22 (1), 179–190.

Sanders, N. R., Graman, G. A., 2009. Quantifying costs of forecast errors: A case study of the warehouse environment. Omega 37 (1), 116–125.

Sanders, N. R., Manrodt, K. B., 2003. The efficacy of using judgmental versus quantitative forecasting methods in practice. Omega 31 (6), 511–522.

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.

Seaman, B., 2018. Considerations of a retail forecasting practitioner. International Journal of Forecasting 34 (4), 822–829.

Shibl, R., Lawley, M., Debuse, J., 2013. Factors influencing decision support system acceptance. Decision Support Systems 54 (2), 953–961.

Siemsen, E., 2011. The usefulness of behavioral laboratory experiments in supply chain management research. Journal of Supply Chain Management 47 (3), 17–18.

Singhal, K., Singhal, J., 2011a. Imperatives of the science of operations and supply-chain management. Journal of Operations Management 30 (3), 237–244.

Singhal, K., Singhal, J., 2011b. Opportunities for developing the science of operations and supply-chain management. Journal of Operations Management 30 (3), 245–252.

Sroginis, A., Fildes, R. A., Kourentzes, N., 2019. Use of contextual and model-based information in behavioural operations, available at SSRN: https://ssrn.com/abstract=3466929.

Stahl, R., 2010. Executive S&OP: Managing to achieve consensus. Foresight: The International Journal of Applied Forecasting 19, 34–38.

Svetunkov, I., 2021a. Forecasting and Analytics with ADAM. (version: 2021-04-12). URL https://openforecast.org/adam/

Svetunkov, I., 2021b. smooth: Forecasting using state space models. R package version 3.1.1.41013.
URL https://github.com/config-i1/smooth

Sweller, J., 2005. Implications of cognitive load theory for multimedia learning. In: The Cambridge Handbook of Multimedia Learning. Cambridge University Press, pp. 19–30.

Swieringa, R. J., Weick, K. E., 1982. An assessment of laboratory experiments in accounting. Journal of Accounting Research 20, 56.

Syntetos, A. A., Babai, Z., Boylan, J. E., Kolassa, S., Nikolopoulos, K., 2016a. Supply chain forecasting: Theory, practice, their gap and the future. European Journal of Operational Research 252 (1), 1–26.

Syntetos, A. A., Kholidasari, I., Naim, M. M., 2016b. 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.

Thomé, A. M. T., Scavarda, L. F., Fernandez, N. S., Scavarda, A. J., 2012. Sales and operations planning: A research synthesis. International Journal of Production Economics 138 (1), 1–13.

Trapero, J. R., Fildes, R., Davydenko, A., 2011. Nonlinear identification of judgmental forecasts effects at SKU level. Journal of Forecasting 30 (5), 490–508.

Trapero, J. R., Kourentzes, N., Fildes, R., 2015. On the identification of sales forecasting models in the presence of promotions. Journal of the Operational Research Society 66, 299–307.

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.

Tuomikangas, N., Kaipia, R., 2014. A coordination framework for sales and operations planning (S&OP): Synthesis from the literature. International Journal of Production Economics 154, 243–262.

Turner, D. S., 1990. The role of judgement in macroeconomic forecasting. Journal of Forecasting 9 (4), 315–345.

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.

Walsham, G., 2006. Doing interpretive research. European Journal of Information Systems 15 (3), 320–330.

Wang, G., Gunasekaran, A., Ngai, E. W., Papadopoulos, T., 2016. Big data analytics

in logistics and supply chain management: Certain investigations for research and applications. International Journal of Production Economics 176, 98–110.

Watts, S., Shankaranarayanan, G., Even, A., 2009. Data quality assessment in context: A cognitive perspective. Decision Support Systems 48 (1), 202–211.

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., Edmundson, B., 2005. Forecasting support systems for the incorporation of event information: An empirical investigation. International Journal of Forecasting 21 (3), 411–423.

Weller, M., Crone, S., 2012. Supply chain forecasting: Best practices - benchmarking study, available at Lancaster University https://eprints.lancs.ac.uk/id/eprint/135958/1/Weller_Crone_Techical_Report_Supply_Chain_Forecasting_Best_Practices_and_Benchmarking_Study.pdf.

Windschitl, P. D., Smith, A. R., Rose, J. P., Krizan, Z., 2010. The desirability bias in predictions: Going optimistic without leaving realism. Organizational Behavior and Human Decision Processes 111 (1), 33–47.

Wolfe, C., Flores, B., 1990. Judgmental adjustment of earnings forecasts. Journal of Forecasting 9 (4), 389–405.

Yeomans, M., Shah, A., Mullainathan, S., Kleinberg, J., 2019. Making sense of recommendations. Journal of Behavioral Decision Making 32 (4), 403–414.

Zellner, M., Abbas, A. E., Budescu, D. V., Galstyan, A., 2021. A survey of human judgement and quantitative forecasting methods. Royal Society Open Science 8 (2).

Zhang, D., Yin, C., Zeng, J., Yuan, X., Zhang, P., 2020. Combining structured and

unstructured data for predictive models: a deep learning approach. BMC Medical Informatics and Decision Making 20 (1).

Zhao, X., Zhao, X., Wu, Y., 2013. Opportunities for research in behavioral operations management. International Journal of Production Economics 142 (1), 1–2.