European Journal of Operational Research 307 (2023) 1177–1191

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Production, Manufacturing, Transportation and Logistics

Use of contextual and model-based information in adjusting promotional forecasts

Anna Sroginis[a,*], Robert Fildes[a], Nikolaos Kourentzes[b]

[a] Lancaster University Management School, Department of Management Science, Lancaster LA1 4YX, UK
[b] Skövde Artificial Intelligence Lab, School of Informatics, University of Skövde, Sweden

A R T I C L E I N F O

Article history:
Received 16 July 2021
Accepted 3 October 2022
Available online 8 October 2022

Keywords:
Forecasting
Judgmental forecasting
Behavioural OR
Decision support systems
Promotions

A B S T R A C T

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 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 which conditions this is beneficial. Using a multi-method approach, we first analyse a UK retailer case study exploring its operations and the forecasting process. The case study provides a contextual setting for the laboratory experiments that simulate a typical supply chain forecasting process. In the experimental study, we provide past sales, statistical forecasts (using baseline and promotional models) and qualitative information about past and future promotional periods. We include contextual information, with and without predictive value, that allows us to investigate whether forecasters can filter such information correctly. 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. 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.

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 understood: it may be used in data collection (data selection), in the model selection stages or, more crucially, directly in making judgmental adjustments/overrides (see extensive overviews of judgmental forecasting literature by Arvan, Fahimnia, Reisi, & Siemsen, 2019; Lawrence, Goodwin, O'Connor, & Önkal, 2006; Perera, Hurley, Fahimnia, & Reisi, 2019). It has been argued that expert adjustments to statistical forecasts can be beneficial for a company (Fildes, Goodwin, Lawrence, & Nikolopoulos, 2009), since adjustments can incorporate contextual information that is not taken into account by the statistical model available (Lawrence et al., 2006). However, it remains unclear how and why experts modify these forecasts and whether the increased availability of the more advanced forecasting methods influences the effectiveness of the process overall.

In this study, 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 contextual information 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 & 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 Support Systems (DSS), and the adjustments demand planners made.

This study contributes to the existing literature in two ways: (i) it presents a case study of the use of a FSS in the demand planning process of a large retailer, providing insights into the

* Corresponding author.
E-mail address: a.sroginis@lancaster.ac.uk (A. Sroginis).

https://doi.org/10.1016/j.ejor.2022.10.005

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

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 for further research into support systems and the interaction between decision-makers and support systems.

This study is organised as follows. Section 2 provides an overview of the literature on judgmental adjustments and promotional modelling and motivates a research question that is common for both the case study and the laboratory experiment. In Section 3, we discuss our multi-method research approach. Section 4 presents a case study of a major UK retail company that motivates our experiments. Section 5 introduces the experiment setup and corresponding hypotheses, and Section 6 provides the analysis of the experimental results, followed by a discussion and concluding remarks in Section 7.

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 & 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, Goodwin, Thomson, Gönül, & Pollock, 2009), and this approach prevails in practice (McCarthy, Davis, Golicic, & Mentzer, 2006). This is also supported by many surveys over the last decade. In Table 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 & 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. Even though the recent study by Khosrowabadi, Hoberg, & Imdahl (2022) reported a smaller percentage of 5.1% of all forecasts at the SKU-store-day level being adjusted by demand planners, this corresponds to 1.5 trillion adjustments per year across 500 stores and 40,000 different SKUs.

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 (Fildes et al., 2009; Franses, 2013; Franses & Legerstee, 2011; Sanders & Graman, 2009; Trapero, Pedregal, Fildes, & Kourentzes, 2013; Van den Broeke, De Baets, Vereecke, Baecke, & Vanderheyden, 2019), focusing on their effectiveness and possible 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 actual/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 & 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 (Fildes & Petropoulos, 2015; Galbraith & Merrill, 1996; Lawrence, O'Connor, & Edmundson, 2000). In addition, various biases and human factors may influence the size and direction of adjustments, such as a sense of ownership (Önkal & 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 Gigerenzer, 1999; Kahneman, 2012; Tversky & Kahneman, 1974).

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) and decision training (Siebert, Kunz, & Rolf, 2021); improved statistical FSSs in order to produce better system forecasts (Fildes et al., 2009) and better support to rely on (Hewage, Perera, & De Baets, 2022); reliable choice and interpretation of error measures (Davydenko & Fildes, 2013; Petropoulos, Kourentzes, & Nikolopoulos, 2016); recording reasons for adjustments (Fildes & Goodwin, 2007a; Goodwin, 2000); improvements in the S&OP process in which potentially useful information is collected as a whole in order to reconcile different sources of information in the forecasting and planning decisions (Fildes, Goodwin, & Lawrence, 2006). This is relevant not only for model-based forecasting, but also for optimisation models (Käi, Kemppainen, & Liesiö, 2019).

Several laboratory experiments have been conducted to analyse the use of contextual information in forecasting. Initially, in one company example Edmundson, Lawrence, & O'Connor (1988) showed that product knowledge could significantly improve forecast accuracy compared to extrapolative forecasting approaches alone. Sanders & Ritzman (1992) conducted a similar experiment between students and practitioners against more sophisticated statistical methods, reporting that 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 a user-friendly interface (e.g., presenting information only in a tabular form and considering experience and domain knowledge as essential attributes of practitioners). And later, in a different experimental study, Webby, O'Connor, & Edmundson (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 the numbers of contextual statements presented.

Recently, Fildes, Goodwin, & Önkal (2019a) 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 the contextual information. Our research expands on this work. More specifically, we attempt to simulate business practice more faithfully by highlighting different types of contextual information 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 predictive value 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 value.

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 to the research question will have implications for the design of organisational forecasting process.

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 & Rosenzweig, 2020). Choi, Cheng, & Zhao (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 their 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 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. (2016)'s argument, the research question discussed in the previous section require 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 FSSs at the heart of the demand planning process, we need to examine how the actors perform their tasks in an archetypical 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.

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 into practice. The evidence on the ubiquity of the forecasting process is provided in many field studies (e.g., Fildes et al., 2009; Franses & Legerstee, 2011; Trapero et al., 2013; Van den Broeke et al., 2019) with further evidence of its use in retail demand forecasting (Fildes, Ma, & Kolassa, 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) interviewing actors and observing the process by which the demand forecasts are produced and the information flows; (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.

3.2. Experimental study

A well-established 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 cannot be fully transferred to an experimental setting. For instance, both De Baets & Harvey (2018) and Fildes et al. (2019a) investigated different aspects of judgmental adjustments of statistical forecasts using laboratory experiments. In particular, De Baets & 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 organisations 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).

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

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.

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 fifty thousand SKUs (Stock Keeping Unit) organised in three supercategories and around twenty lower-level categories. Only twenty 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 (ten types of promotions and each type having many different durations).

Fig. 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 Fig. 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.

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 Fig. 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 Fig. 1). Then these might be judgmentally adjusted by demand planners based on weather, known special events and 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; Fildes et al., 2019b; Syntetos, Babai, Boylan, Kolassa, & Nikolopoulos, 2016) and highlighted that even when statistical models incorporated promotional or special events effects, manual adjustments were frequently used to override model predictions.

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). Although, there was no clear indication how these were operationalised or affected the forecasting process. Nonetheless, the forecasting team checked its performance by manually calculating the Mean Absolute Percentage Error (MAPE), see Eq. A.1 in Appendix A (can be found online in the supplementary materials), 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 the absence of forecast evaluation, which is seen as an important dimension for successful forecasting (Moon, Mentzer, & Smith, 2003) and a common limitation of systems in practice (Ord, Fildes, & Kourentzes, 2017, Chapter 13). According to Petropoulos, Goodwin, & Fildes (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 (Kourentzes, Trapero, & Barrow, 2020; Sanders & Graman, 2009). 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 5.

4.2. Interface

The SAP F&R system uses a visual interface that plots sales and forecasts stored in the system. The interface provides markers for

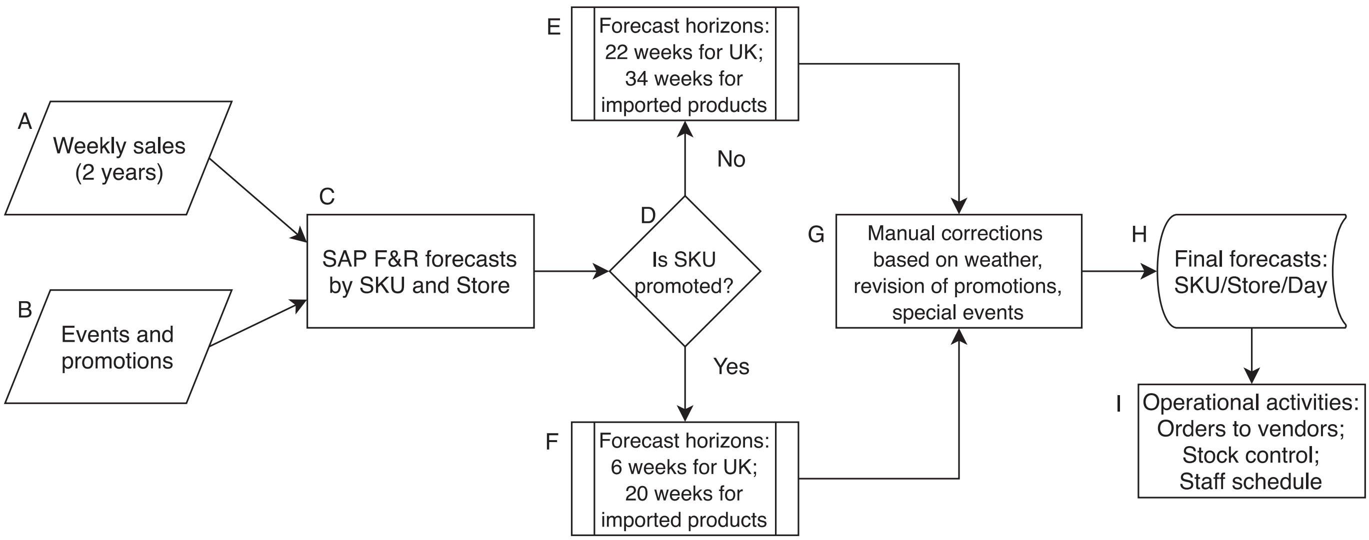

Fig. 1. Forecasting process of the case company.

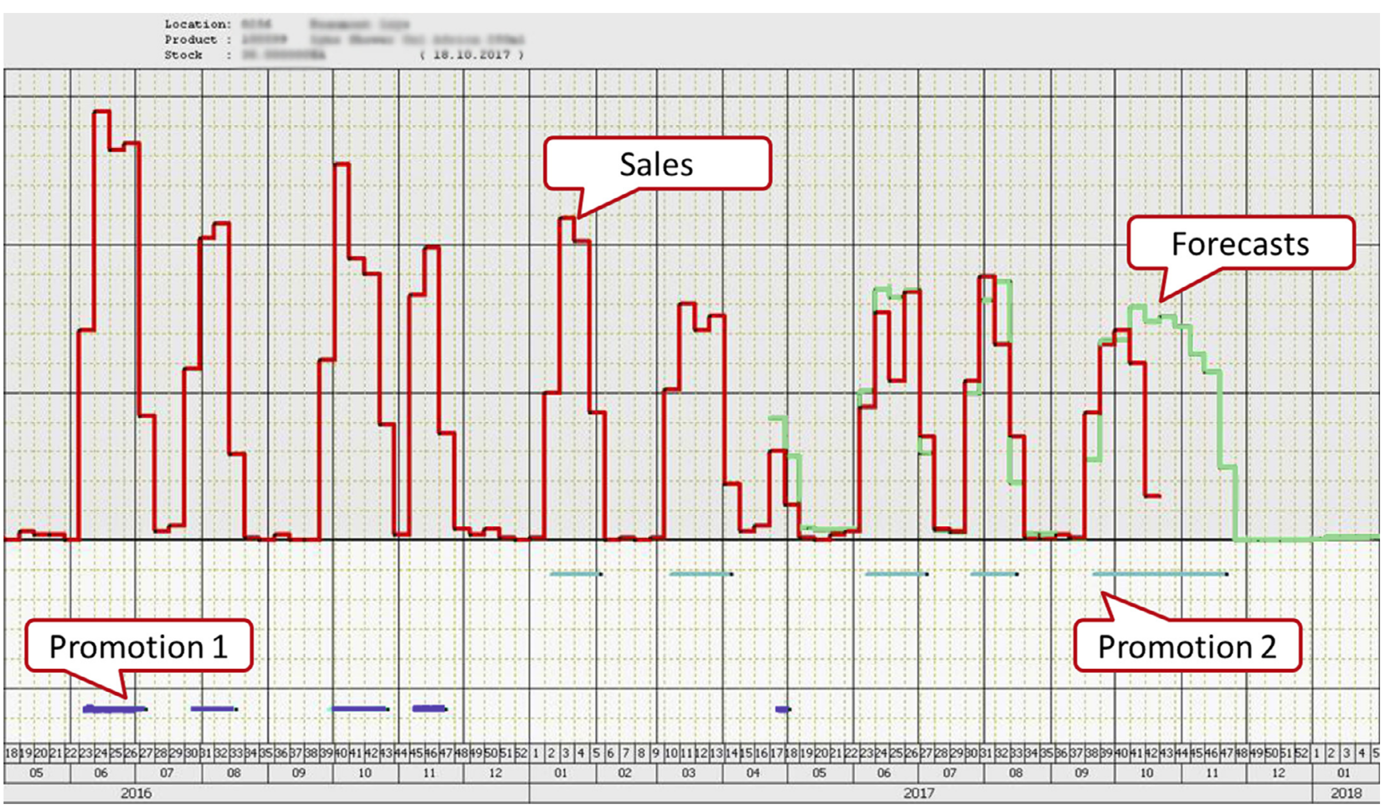

Fig. 2. Screenshot of the software used by the case company.

promotional and seasonal events, which help demand planners visually assess the quality of system forecasts.

Fig. 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.

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 latter 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

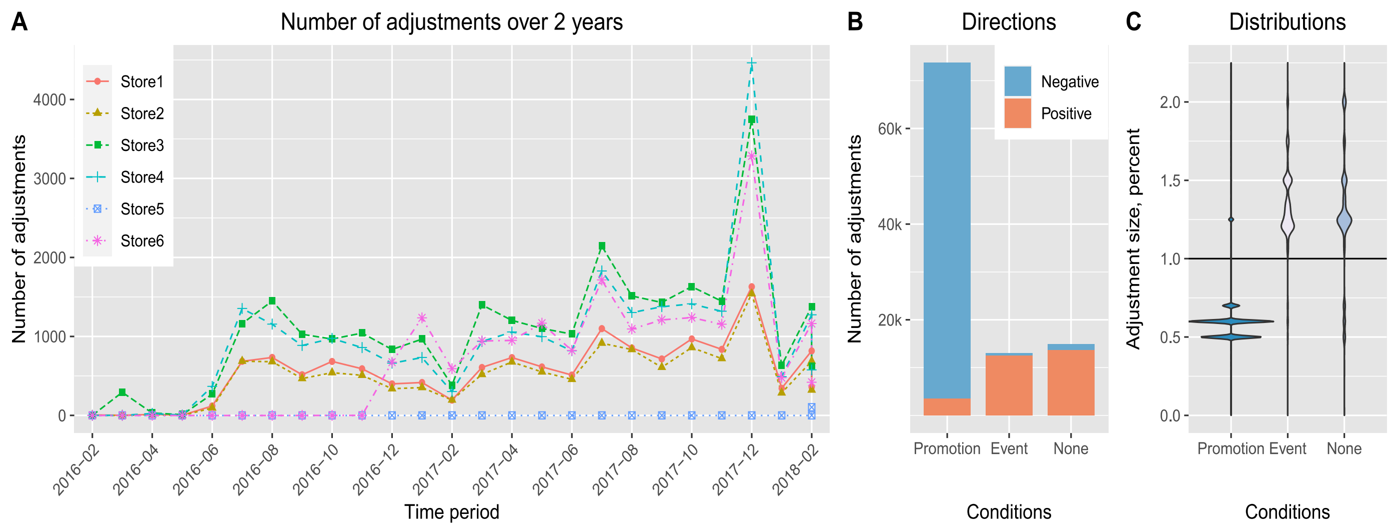

Fig. 3. Total number of manual adjustments over the 2 years (A); Ratio of positive and negative adjustments (2 years) (B); Distribution of percentage adjustments (2 years) (C).

a range between twenty and fifty 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; Fildes et al., 2019a; Franses & Legerstee, 2009), particularly for promotions (Trapero, Kourentzes, & Fildes, 2015; Trapero et al., 2013).

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, Nikolopoulos, Boylan, Fildes, & Goodwin (2009) showed, such adjustments could be effective in intermittent demand cases.

Fig. 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 period with well-known effects on sales. This spike highlights a perceived problem with the statistical forecasts 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 Fig. 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. This suggests that the system forecast could be over-forecasting in those cases. Interestingly, Fig. 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 and 30%. This brief analysis highlights tendency to adjust down often for promotional periods, but also to increase forecasts by a small amount in other cases. In conversation with the case company, one reason given for the number and sizes of the negative adjustments is the timing and length of promotions. System forecasts are weekly (Box C in Fig. 1) and are then disaggregated to daily forecasts (Box H in Fig. 1). Since planned promotions can last less than a week, experts adjust these forecasts negatively to account for this.

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 suggests that these adjustments are not always beneficial in terms of accuracy (Fildes & Goodwin, 2007b; Fildes et al., 2009; Franses & 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.

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
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)

Table 2 shows how many adjustments have been done over 24 weeks in these six stores, highlighting the number of observations contributed 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 Equations A.4 and A.5 in Appendix A). More details on negative adjustments and their accuracy could be found in Table 1 in Appendix 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 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 & 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 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 substantially (see the first row in Table 2) 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 contextual information is available. The effect of human adjustments of statistical forecasts during promotional periods that are based on contextual information of unknown predictive value is mainly positive. Still, incorrect judgments lead to substantial errors and diminish all improvements. As for contextual information, it is collected outside the FSS. 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 predictive value. 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.

5. Experimental study

5.1. Hypotheses

In the experimental part, we aim to investigate adjustments when contextual information from different sources is available to forecasters. As we observed in the case study, demand planners use several types of information to make their decisions: (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). Usually the predictive value of the qualitative information is unknown. In our experiment, we include all these types of information. All qualitative pieces have unknown value for the 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 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. When a promotional forecasting model is provided,

(a) judgmental forecast adjustments are close to zero despite the availability of promotional contextual information of unknown predictive value, and
(b) this model yields better forecast accuracy than any adjustment process.

Conversely,

Hypothesis 2. When the statistical forecast does not contain promotional information, judgmental forecast adjustments are based on contextual statements with unknown predictive value following its direction and perceived value.

In the absence of a trustworthy statistical forecast, the forecaster gives weight to additional statements, leading to adjustments in the same direction as salient contextual reasons, but the expectation is that the forecaster is unable to achieve parity with forecasts produced solely by the promotional statistical model.

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 contextual information and incorporate it in their adjustments, addressing the first aim from Section 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 FSSs along with product codes. We include two statistical models where the baseline model allows the forecaster to check how the SKU's performance changes when only a flat line forecast is presented and the promotional model increase 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 contextual information 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 unknown predictive value 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 had predictive value. 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 C. The first two groups had an equal number of positive/negative reasons. The reasons have been previously validated by experts as an illustration of qualitative information shared in demand forecasting. We prescribed the number of reasons and their directions (positive/negative), but they were otherwise drawn randomly for each promotional period. The timing of promotions was specified with a slight variation between time series.

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 6.4.

The differences between the case study F&R system (Section 4) and our experimental design are explored in Table 3. It shows the main assumptions and simplifications for the experiment.

5.3. Participants and incentives

There were two groups of participants: (1) students, and (2) users of a crowd-sourcing platform for conducting research experiments. The first group had completed a "Business Forecasting" module with one lecture about judgmental forecasting. The experiment was introduced as a voluntary exercise in a controlled setting, with a maximum of 45 minutes to complete it. We used different methods of motivation: individual money tokens for participation, one/two prizes for the best performance and only verbal encouragement. All student 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. We found no significant difference in performance between groups with monetary incentives and a group with only verbal encouragement, which aligns with findings from other studies (e.g., Camerer & Hogarth, 1999; Katok, 2018). The second group of participants was recruited through an online crowd-sourcing platform "Prolific.co", which uses a fair reward policy of 7.5 per hour. We enlisted 90 participants of different backgrounds without prior experience in forecasting or demand planning. The average finishing time of the experiment was around 22 minutes, so the reward was set at 2.75.

5.4. Data generating process

The data for the experiment was generated using:

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

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 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 mimiced 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.

Table 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

Table 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	46	49	41	42
Mean adjustment size (%)	17.1	6	25.5	1
# Adjustments	366	376	341	321
# Negative / positive adjustments	14% / 86%	35% / 65%	11% / 89%	46% / 55%
Positive adjustment size (%)	27.5	19.1	35.8	16.9
Negative adjustment size (%)	−14.8	−13.4	−12.5	−17.4
Adjusted cases	80%	77%	83%	76%
Statistical bias (scaled by mean sales)	0.331	−0.068	0.463	−0.081
Adjustment bias (scaled by mean sales)	0.174	−0.151	0.225	−0.079
Directional bias	0.7	0.3	0.8	0.1
FVA	36.8%	−39.2%	43%	−25.3%

There were four experimental settings (see Table 4). Each treatment was assigned randomly when a participant agreed to contribute in this experiment. However, not all participants finished their attempt, hence, the number of participants across treatments is not uniform. The first two treatments are used as control, while the last two treatments tested “Enhanced” sales, where the effect of promotions was increased by a further 25% when a positive promotional statement is provided. Thus, “Promotional Campaign” statements 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.

Building on the results of Fildes et al. (2019a), we explore the forecasting process not only as a null experiment with no qualitative information affecting the promotion 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 1a where model-based information is stronger than any contextual statements and it is more accurate, Hypothesis 1b); (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.

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 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 system inspired our interface, but some features have been simplified and magnified. The result can be seen in Fig. 4.

The interface is implemented using R Shiny Apps environment 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 3 in Appendix E.

6. Analysis and findings

6.1. Sample and descriptive analysis

In total, 256 participants started the experiment with 88 complete runs for students and 90 for prolific users. According to Cook’s distance, there were 3 outliers in the student dataset 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. No outliers were identified in the prolific sample.

Descriptive statistics of the treatments are provided in Table 4. The table shows the number of participants, the average adjustment size, total adjustments, negative/positive adjustment ratio, adjusted cases, various biases and Forecast Value Added (FVA).

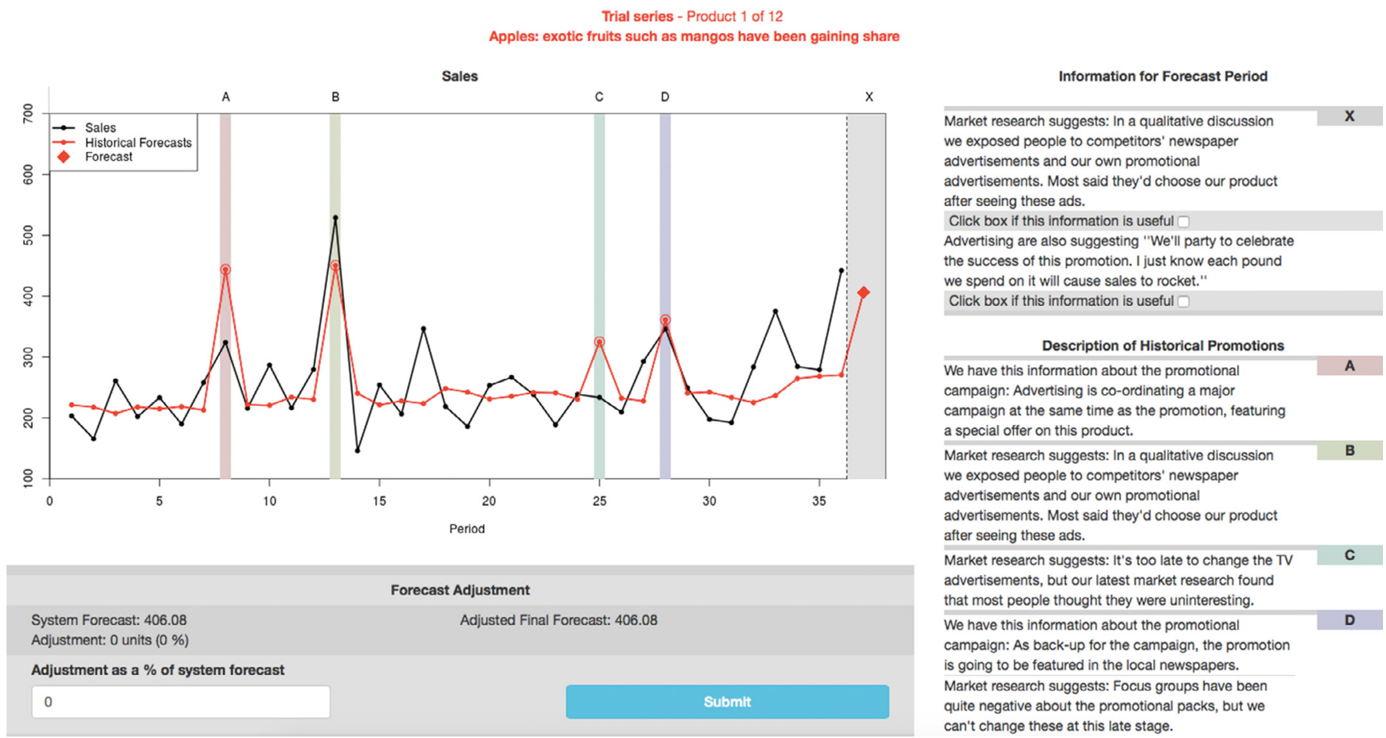

Fig. 4. Experimental FSS. Screenshot example.

The number of participants ranges from 41 to 49 per treatment due to the treatment assignment described in the Section 5.4. The average adjustments for the baseline model (treatments 1 and 3) are much less than the anchor of 50%. However, in the case of the enhanced time series, the adjustments are higher for treatment 3 compared to treatment 1, as the promotional uplifts are more pronounced visually. Nonetheless, the average adjustment still falls short of the expected size, especially once adjusted for the enhanced effect. In treatments 2 and 4, with the promotional model, the size of adjustments is minimal, and it is insignificantly different from zero for the treatment 4 (p-values for two-sided nonparametric test: 0.0000 and 0.3704, respectively). However, considering the percentage of adjusted cases, these time series were adjusted as frequently as in other treatments. For treatment 2, the participants from the online crowd-sourcing platform performed differently from the student sample (Table 2 in Appendix D). These participants adjusted forecasts positively even when there was a statistical model with promotional effects in place. The remaining treatments are similar across the groups.

Comparing between baseline and promotional models, while the total number of adjustments varies slightly, the frequency of negative to positive adjustments changes substantially (from only 11–14% to 35–46% of negative adjustments). In treatments 2 and 4, these positive/negative adjustments indicate "tweaking" the forecasts rather than accepting the statistical forecasts. The frequency of checked promotional, marketing and hype statements can be found in Table 2 in Appendix D, where 60–71% of all promotional and marketing reasons, and only 11% of overly-positive hype statements, are marked by participants as "Useful".

The mean error (bias) typically detects systematic bias in a forecast (see Eq. A.2 in Appendix A that can be found in the online supplement): 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 under-forecasting, but increase the negative bias for the promotional model - over-forecasting for these cases (except treatment 4). The mean directional bias (Eq. A.3) 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 Eq. A.4 in Appendix A) suggest accuracy improvements for the baseline model settings, in contrast to the other two. This suggests that participants corrected forecasts due to either not trusting the statistical forecasts or alternatively falsely relying on the additional contextual information. These results provide evidence to reject Hypothesis 1b. 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 FSS.

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 accounting for differences in personal characteristics and background) 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, Mächler, Bolker, & Walker, 2015) for R.

Following Davydenko & 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 observation, j, the time series and k, participant, denoting the first, second and

Table 5
Linear mixed effects model output for logarithmic relative adjustment size.

Fixed effects			Confidence intervals	
	Estimate	Std. Error	2.5%	97.5%
Intercept	0.3122	0.4255	−0.5013	1.1299
Treatment 2	0.0537	0.0287	−0.0032	0.1090
Treatment 3	0.0707	0.0272	0.0176	0.1237
Treatment 4	0.0205	0.0304	−0.0398	0.0789
Promotional reasons useful	−0.0472	0.0140	−0.0736	−0.0181
Promo reasons useful × Treatment 1&3	0.0491	0.0150	0.0198	0.0785
Market reasons useful	−0.0343	0.0122	−0.0578	−0.0105
Hype useful	0.0262	0.0191	−0.0113	0.0636
Hype useful × Treatment 1&3	0.0562	0.0257	0.0061	0.1067
Low noise × Treatment 1&3	0.0856	0.0151	0.0562	0.1148
Promo reasons useful, positive	0.0851	0.0159	0.0505	0.1151
Market reasons useful, positive	0.0849	0.0175	0.0511	0.1191
Current forecast	−0.2044	0.0330	−0.2672	−0.1381
Last promotional uplift	0.1477	0.0759	0.0035	0.2928
Random effects				
	St. Dev.			
Participant (intercept)	0.1186			
Time series (intercept)	0.0580			
Residual	0.1407			

third levels of a multilevel model with two random intercepts for time series and participants' effects respectively.

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

$$y_{ijk} = \underbrace{\beta_0 + \sum_{a=1}^{n} \beta_a X_{aijk}}_{\text{fixed part}} + \underbrace{v_k + u_{jk}}_{\text{random part}} + \varepsilon_{ijk}, \tag{2}$$

where y_{ijk} is the 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 3 in Appendix E 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. We checked for the inclusion of variables for the number of statements presented, the number of positive/negative reasons or their ratio, time taken for each time series, and none were used. Table 5 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.

6.2.1. *Qualitative information about past promotions*

The results show 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 sparse. The experiment simulates a realistic business environment such as the case study, where there is an excessive amount of information with unknown predictive value, and therefore participants may have been overloaded with information. We will return to this argument in Section 7.

6.2.2. *Contextual information for the forecasted promotional period*

Table 5 provides the final model with the smallest AIC value together with confidence intervals for each coefficient. The contextual reasons included in the model were the statements that had been marked as useful by participants. Looking at the model coefficients, we observe that the additional information was retained, yet misused. All positive promotional and marketing statements increased the relative adjustments. In the case of the treatments with promotional model forecasts, we observe that the promotional, marketing, and hype information have either negative or close to zero coefficients, supporting Hypothesis 1a.

Interaction variables for contextual information and the baseline model ("Promo reasons useful × Treatment 1&3" and "Hype useful × Treatment 1&3") shows that there is a positive effect on adjustments which is as expected, since the baseline model requires corrections. Note that a similar interaction effect with marketing information was excluded from the model. These results partially support Hypothesis 2. In the enhanced promotions treatments (3 and 4) we do not find an interaction between the usefulness of promotions and the specific treatments (Promo reasons useful × Treatment 3&4). Nonetheless, as the promotions are included in the model, we interpret this as the participants being able to identify those promotional statements as predictively useful, but the enhanced effect is not identified and did not increase their reliance on promotional statements.

"Hype" information led participants to forecast a higher uplift, while any "Promotional" and "Market Research" 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 useful information. At the same time, positive promotional and marketing statements have strong positive effects, which supports Hypothesis 2 again regardless of the initial statistical model.

Summarising these findings, participants were confused by overly-positive information (e.g., "Hype") and were not able to identify the effect of helpful statements in enhanced treatments. Nonetheless, the highest coefficients among the contextual information belong to promotional statements. Therefore, we conclude that there is a potential for participants to recognise valuable information. We see that (i) all positive contextual statements have a positive effect on adjustments for both statistical forecasting models, but in the case of the promotional model, this effect is small (supporting Hypothesis 1a); (ii) there is a strong impact of contextual information on adjustments when the baseline statistical forecast is provided (supporting Hypothesis 2). In particular, we observe positive coefficients for both promotional and hype statements.

6.2.3. *Other variables*

The mixed effects model shows that several other variables are highly significant. Table 5 suggests that all variables connected with algorithmic information, such as the current forecast and last

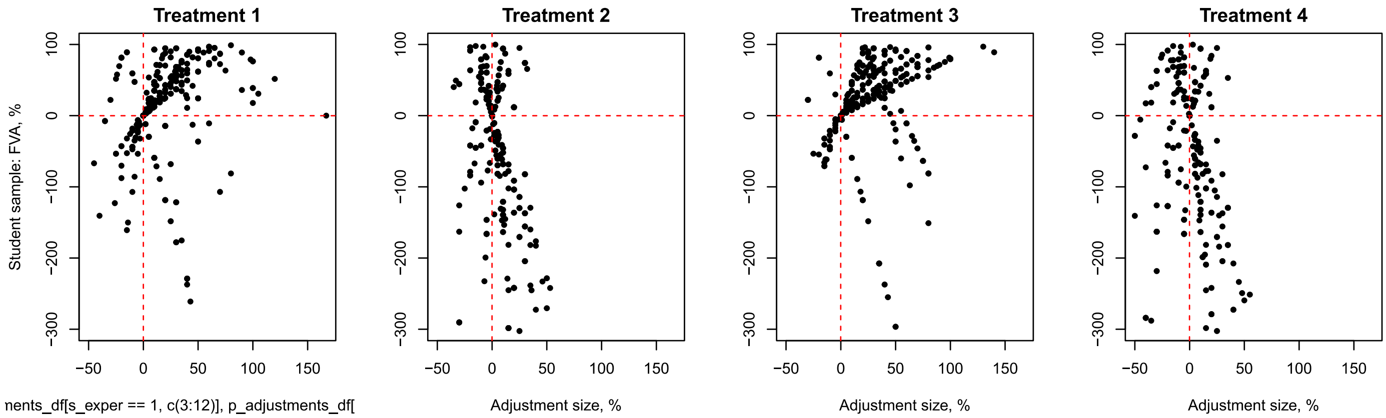

Fig. 5. Adjustment size impact on accuracy (based on FVA values).

promotional uplift 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. (2019a). 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 dampen it more. We observe that 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 the Hypothesis 1b.

6.3. Forecasting accuracy assessment

The improvement in terms of Relative Mean Absolute Errors (RelMAE) has been used to gauge changes in forecast accuracy. The FVA is calculated using Equation A.5 (Appendix A) as measured by MAE, where positive values correspond to improvements in accuracy following interventions.

Fig. 5 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, Fildes, & Davydenko (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 experts: many adjustments were too high, which damaged forecast accuracy. Participants adjusted somewhat randomly, reacting to various pieces of information with unknown value. 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).

A mixed effects model of FVA, with the same variables as in Table 3 in Appendix E, indicated that the reaction to contextual information for the upcoming promotional period changed the overall accuracy differently depending on the information contained in the forecasts. In treatments 1 and 3 promotional statements led to better forecasts, while the opposite was found for treatments 2 and 4. The reaction to hype information caused a dramatic decrease in accuracy for all treatments. We also observed that there was no strong linear correlation between adjustment size and FVA. Therefore, we argue that additional qualitative information with unknown predictive value can be misused and decrease final forecast accuracy (supporting Hypothesis 1b, Section 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 Appendix F show that on average participants had little confidence about their knowledge of sales forecasting (mainly driven by a prolific group). 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 2.7 out of 5 (where 5 is "Very accurate"). The general public from the online sample indicated lower confidence and knowledge of forecasting than the students. 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.

7. Discussion and conclusions

The case study demonstrated that adjusting forecasts remains a core task in demand planning and it follows a complex process. We observed that even when an automatic forecasting model for promotional periods with multiple explanatory factors is available, demand planners still adjust based on additional information available outside of the FSS. We identified that (i) all additional information (e.g., promotional/marketing information, weather forecasts) is handled outside of the FSS; (ii) the adjustments for promotional periods are typically negative, while for special events they are positive; and (iii) forecast evaluation of judgmental interventions is new at this company, mirroring our experience with other companies. The case study provides a detailed mapping of a process that is often left assumed in the literature. Furthermore, the case study highlights problems and issues that can be simulated and analysed in a controlled research environment. Therefore, it acts as an inspiration for the design of our experiment. We argue that this type of research should mirror organisational practices as closely as possible; hence it is crucial to investigate processes in companies prior to any experiments.

In the case study, expert adjustments due to contextual information of unknown predictive value were found to have a positive effect in only half of the cases. We observed negative adjustments for promotional periods that were beneficial (contrary to Fildes & Goodwin, 2007a; Trapero et al., 2013). This suggests that the system forecast could be over-forecasting in those cases. For special events (e.g., weather changes, some marketing activities) the forecasts were adjusted upwards with at times substantial losses in forecast accuracy. This aligns with the previous finding in the literature (e.g., Fildes et al., 2009).

Additionally, the FFS interface in the case organisation is overly complex and thus not user-friendly. As a result, the whole forecasting process and evaluation become extremely obscure 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 FSSs is hardly discussed or considered by many behavioural papers. The key issues can be highlighted: (i) the reliability of the promotional model; (ii) the value added by forecasters; (ii) how experts filter and use the circumstantial sources of information. This opens many directions for further research on the expert judgment in FSSs. These issues persist even when using sophisticated Machine Learning or Artificial Intelligence algorithms as evidenced by Khosrowabadi et al. (2022). In the latter case, research suggests that there may be different trust dynamics between users and models (Glikson & Woolley, 2020), with users anthropomorphising AI and therefore having different expectations of AI. Together with the increasing prevalence of AI, this motivates further research in judgmental adjustments of AI-generated forecasts and more generally in model trust in forecasting (Spavound, Kourentzes et al., 2022).

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, 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 predictive value is available.

One interesting finding is a forecaster's reaction to available information. Apart from Fildes et al. (2019a); Goodwin, Onkal Atay, Thomson, Pollock, & Macaulay (2004); Sanders & Ritzman (1992); Webby et al. (2005), 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, Goodwin, Fildes, Nikolopoulos, & Lawrence (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 the 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, strongly positive statements for the *upcoming* period shifted the adjustments upwards, and, surprisingly, participants did not anchor more to the overly-positive "Hype" statements than any other positive information. Even though participants were influenced by this misleading information, they relied more on promotional statements of unknown predictive value. This is an important result showing that users can filter relevant pieces of information even in such complex conditions. We suggest that pre-processing the contextual information to present it better to forecasters can be beneficial, but how to best do this remains an open research question.

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

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 & Legerstee, 2011). However, both Fildes et al. (2019a) and De Baets & 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. We recognise that other unidentified biases may be relevant here.

A limitation of this study is that we rely on student and general public (via an online platform) participants, although ideally we would like to have experienced forecasters as participants. However, collecting a sufficient sample of practitioners can be very challenging. Therefore, we have relied on participants with limited or no demand planning experience. To mitigate this, we use a relatable example (i.e., a supermarket setting) and introduce two training runs with performance feedback. Although previous studies have indicated that the qualitative findings between using experts and non-experts are similar (e.g., Petropoulos, Kourentzes, Nikolopoulos, & Siemsen, 2018), additional experimentation with practitioners would strengthen our findings. This also connects to our prior point about trust, as it remains an open question whether trust in models differs between expert and non-expert forecasters. Furthermore, 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. Those assumptions were crucial due to insufficient data, which is a typical problem for real company examples. This indicates the need for more multi-method methodology studies in judgmental forecasting, which would provide both findings with high relevance to practice and importantly more examples of archetypical organisations to motivate research and reduce the need for such assumptions. Third, both the case study and our laboratory experiment are very context-, data- and case-specific. Hence, we are not able to generalise these findings widely: despite the ecological validity of our assumptions there may be situations when our conclusions do not hold.

This study emphasises the role of forecasting models, highlighting how the behaviour of forecasters varies depending on the model provided. It suggests potential trust and acceptance challenges (see, Alvarado-Valencia, Barrero, Önkal, & Dennerlein, 2017; Dietvorst, Simmons, & Massey, 2015; 2018) that are not explicitly explored here. We also do not consider other dimensions, such as product categories, season, etc. that can be relevant. While in laboratory settings, we can design an optimal statistical model for generated time series, there is also 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.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ejor.2022.10.005.

References

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. https://doi.org/10.1016/j.ijforecast.2015.12.010.

Arvan, M., Fahimnia, B., Reisi, M., & Siemsen, E. (2019). Integrating human judgement into quantitative forecasting methods: A review. *Omega, 86*, 237–252. https://doi.org/10.1016/j.omega.2018.07.012.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01.

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. https://doi.org/10.1023/a:1007850605129.

Choi, T.-M., Cheng, T. C. E., & Zhao, X. (2016). Multi-methodological research in operations management. *Production and Operations Management, 25*(3), 379–389. https://doi.org/10.1111/poms.12534.

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. https://doi.org/10.1016/j.ijforecast.2012.09.002.

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. https://doi.org/10.1016/j.ijforecast.2017.09.007.

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. https://doi.org/10.1037/xge0000033.

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. https://doi.org/10.1287/mnsc.2016.2643.

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. https://doi.org/10.1002/for.3980070305.

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. https://doi.org/10.1287/inte.1070.0309.

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., & Lawrence, M. (2006). The design features of forecasting support systems and their effectiveness. *Decision Support Systems, 42*, 351–361. https://doi.org/10.1016/j.dss.2005.01.003.

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. https://doi.org/10.1016/j.ijforecast.2008.11.010.

Fildes, R., Goodwin, P., & Önkal, D. (2019). Use and misuse of information in supply chain forecasting of promotion effects. *International Journal of Forecasting, 35*(1), 144–156. https://doi.org/10.1016/j.ijforecast.2017.12.006.

Fildes, R., Ma, S., & Kolassa, S. (2019). Retail forecasting: Research and practice. *International Journal of Forecasting*. https://doi.org/10.1016/j.ijforecast.2019.06.004.

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

Franses, P. H. (2013). Improving judgmental adjustment of model-based forecasts. *Mathematics and Computers in Simulation, 93*, 1–8. https://doi.org/10.1016/j.matcom.2012.11.007.

Franses, P. H., & Legerstee, R. (2009). Properties of expert adjustments on model-based SKU-level forecasts. *International Journal of Forecasting, 25*(1), 35–47. https://doi.org/10.1016/j.ijforecast.2008.11.009.

Franses, P. H., & Legerstee, R. (2011). Combining SKU-level sales forecasts from models and experts. *Expert Systems with Applications, 38*(3), 2365–2370. https://doi.org/10.1016/j.eswa.2010.08.024.

Galbraith, C. S., & Merrill, G. B. (1996). The politics of forecasting: Managing the truth. *California Management Review, 38*(2), 29–43. https://doi.org/10.2307/41165831.

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

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals, 14*(2), 627–660. https://doi.org/10.5465/annals.2018.0057.

Goldstein, H. (2011). Multilevel statistical models. *Wiley series in probability and statistics* (4th ed.). Hoboken, N.J.: Wiley.

Goodwin, P. (2000). Improving the voluntary integration of statistical forecasts and judgment. *International Journal of Forecasting, 16*(1), 85–99. https://doi.org/10.1016/s0169-2070(99)00026-6.

Goodwin, P., Onkal Atay, D., Thomson, M. E., Pollock, A. C., & Macaulay, A. (2004). Feedback-labelling synergies in judgmental stock price forecasting. *Decision Support Systems, 37*(1), 175–186. https://doi.org/10.1016/S0167-9236(03)00002-2.

Harvey, N. (2007). Use of heuristics: Insights from forecasting research. *Thinking and Reasoning, 13*(1), 5–24. https://doi.org/10.1080/13546780600872502.

Hewage, H. C., Perera, H. N., & De Baets, S. (2022). Forecast adjustments during post-promotional periods. *European Journal of Operational Research, 300*(2), 461–472. https://doi.org/10.1016/j.ejor.2021.07.057.

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

Kahneman, D. (2012). *Thinking, fast and slow*. London: Penguin.

Katok, E. (2018). *Designing and conducting laboratory experiments* (pp. 1–33)). John Wiley & Sons, Ltd.

Khosrowabadi, N., Hoberg, K., & Imdahl, C. (2022). Evaluating human behaviour in response to AIrecommendations for judgemental forecasting. *European Journal of Operational Research, 303*(3), 1151–1167. https://doi.org/10.1016/j.ejor.2022.03.017.

Kourentzes, N., Trapero, J. R., & Barrow, D. K. (2020). Optimising forecasting models for inventory planning. *International Journal of Production Economics, 225*, 107597. https://doi.org/10.1016/j.ijpe.2019.107597.

Käi, A., Kemppainen, K., & Liesiö, J. (2019). What to do when decision-makers deviate from model recommendations? empirical evidence from hydropower industry. *European Journal of Operational Research, 278*(3), 869–882. https://doi.org/10.1016/j.ejor.2019.04.021.

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. https://doi.org/10.1016/j.ijforecast.2006.03.007.

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. https://doi.org/10.1016/s0377-2217(99)00085-5.

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. https://doi.org/10.1016/j.ijforecast.2007.02.006.

McCarthy, T. M., Davis, D. F., Golicic, S. L., & Mentzer, J. T. (2006). The evolution of sales forecasting management: A20-year longitudinal study of forecasting practices. *Journal of Forecasting, 25*(5), 303–324. https://doi.org/10.1002/for.989.

Moon, M. A., Mentzer, J. T., & Smith, C. D. (2003). Conducting a sales forecasting audit. *International Journal of Forecasting, 19*(1), 5–25. https://doi.org/10.1016/s0169-2070(02)00032-8.

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. https://doi.org/10.1111/j.1937-5956.2009.01003.x.

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. https://doi.org/10.1016/j.jom.2010.11.012.

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. https://doi.org/10.1002/bdm.637.

Ord, K., Fildes, R., & Kourentzes, N. (2017). *Principles of business forecasting* (2nd ed.). New York: Wessex Press Publishing Co.

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. https://doi.org/10.1016/j.ejor.2018.10.028.

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. https://doi.org/10.1016/j.ijforecast.2015.12.006.

Petropoulos, F., Kourentzes, N., & Nikolopoulos, K. (2016). Another look at estimators for intermittent demand. *International Journal of Production Economics, 181*(3), 154–161. https://doi.org/10.1016/j.ijpe.2016.04.017.

Petropoulos, F., Kourentzes, N., Nikolopoulos, K., & Siemsen, E. (2018). Judgmental selection of forecasting models. *Journal of Operations Management, 60*(1), 34–46. https://doi.org/10.1016/j.jom.2018.05.005.

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. https://doi.org/10.1287/msom.2019.0829.

Sanders, N. R., & Graman, G. A. (2009). Quantifying costs of forecast errors: A case study of the warehouse environment. *Omega, 37*(1), 116–125. https://doi.org/10.1016/j.omega.2006.10.004.

Sanders, N. R., & Manrodt, K. B. (2003). The efficacy of using judgmental versus

quantitative forecasting methods in practice. *Omega, 31*(6), 511–522. https://doi.org/10.1016/j.omega.2003.08.007.
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. https://doi.org/10.1016/j.ijforecast.2018.03.001.
Siebert, J. U., Kunz, R. E., & Rolf, P. (2021). Effects of decision training on individuals decision-making proactivity. *European Journal of Operational Research, 294*(1), 264–282. https://doi.org/10.1016/j.ejor.2021.01.010.
Spavound, S., Kourentzes, N., et al., (2022). Making forecasts more trustworthy. *Foresight: The International Journal of Applied Forecasting*, (66), 21–25.
Stahl, R. (2010). Executive S&OP: Managing to achieve consensus. *Foresight: The International Journal of Applied Forecasting, 19*, 34–38.
Syntetos, A. A., Babai, Z., Boylan, J. E., Kolassa, S., & Nikolopoulos, K. (2016). Supply chain forecasting: Theory, practice, their gap and the future. *European Journal of Operational Research, 252*(1), 1–26. https://doi.org/10.1016/j.ejor.2015.11.010.
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. https://doi.org/10.1016/j.ijpe.2008.08.011.
Trapero, J. R., Fildes, R., & Davydenko, A. (2011). Nonlinear identification of judgmental forecasts effects at SKU level. *Journal of Forecasting, 30*(5), 490–508. https://doi.org/10.1002/for.1184.
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. https://doi.org/10.1057/jors.2013.174.
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. https://doi.org/10.1016/j.ijforecast.2012.10.002.
Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124.
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. https://doi.org/10.1016/j.omega.2018.09.008.
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. https://doi.org/10.1016/j.ijforecast.2004.10.005.
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/.