JBA

JOURNAL OF BUSINESS ANALYTICS

Dursun Delen
Sudha Ram
Richard Vidgen

Journal of Business Analytics

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tjba20

When and how to adjust statistical forecasts in supply chains? Insight from causal machine learning

Budhi S. Wibowo

To cite this article: Budhi S. Wibowo (17 Aug 2023): When and how to adjust statistical forecasts in supply chains? Insight from causal machine learning, Journal of Business Analytics, DOI: 10.1080/2573234X.2023.2248203

To link to this article: https://doi.org/10.1080/2573234X.2023.2248203

Published online: 17 Aug 2023.

JOURNAL OF BUSINESS ANALYTICS
https://doi.org/10.1080/2573234X.2023.2248203

RESEARCH ARTICLE

When and how to adjust statistical forecasts in supply chains? Insight from causal machine learning

Budhi S. Wibowo

Mechanical and Industrial Engineering Department, Universitas Gadjah Mada, Yogyakarta, Indonesia

ABSTRACT

Empirical studies have discovered that most statistical forecasts in supply chains are subjected to judgemental adjustments during a forecast review. Although such a practice requires significant management effort and frequently reduces forecast accuracy, many organisations prefer this approach as part of their Sales & Operations Planning process. This study aims to identify the optimal policy to achieve significant accuracy improvement from forecast review. We focus on a practical situation where managers periodically review forecasts from the statistical software and compare them with judgemental forecasts from the sales and marketing functions. Managers must decide whether to disregard the judgement and continue with the existing forecast or revise the statistical forecast based on the judgement. To determine the best course of action, we conducted a numerical experiment using data from five supply-chain companies wit more than 12,000-point forecasts. The experiment considered three alternative actions: "do-nothing", "follow the judgement", and "simple-average". Using a causal machine learning method, namely a policy tree, we develop a set of decision rules that maximise the expected accuracy gains given the variation in forecasting features. The result proposes a simple yet effective policy to recommend suitable actions based on two identified key features: "judgment direction" and the "accuracy of statistical forecasts". The policy was tested against real-world data and achieved remarkable accuracy with roughly a 3–11 percentage points improvement over the baseline. Our findings offer valuable insights for managers to customise their forecast review policy based on their unique environment.

ARTICLE HISTORY
Received 28 March 2023
Accepted 8 August 2023

KEYWORDS
Statistical forecasting; judgemental adjustment; policy learning; supply chain management; sales and operations planning

1. Introduction

Effective demand planning is critical to the success of supply chain management. By anticipating future demand, companies can make timely and informed decisions about inventory levels, production capacity, and distributions, ultimately leading to optimised operations. Effective demand planning also helps reduce logistics costs and minimises the risk of obsolete inventory. Furthermore, demand planning helps companies proactively identify and address supply chain issues, minimising disruptions and improving overall business performance. Companies that excel in demand planning can gain a competitive advantage by quickly responding to changing market conditions and delivering high-quality products to customers on time. As a result, many companies invest in improving their demand planning processes.

The typical approach to forecasting demand in companies involves computer software employing univariate and multivariate statistical methods to generate baseline forecasts from historical data (hereafter referred to as "statistical forecast"). These statistical forecasts are then reviewed and sometimes adjusted by the demand planners to correct perceived flaws in the initial forecasts or to incorporate additional relevant information (Fildes et al., 2009). The additional information is typically obtained from the sales and marketing function, which has a

The additional information is typically obtained from the sales and marketing function, which has a closer relationship with customers and frequently develops intuition about their future intentions. The sales and marketing function often uses internal and external information (e.g., company objectives, market research, promotion plans, competitive landscape, and special events) to envisage future demand for a longer forecast horizon (hereafter referred to as "judgemental forecasts"). Although highly subjective, independent judgemental forecasts often include helpful information, which is a valuable predictor, especially in volatile environments (Remus et al., 1995; Tan, 2008; Wibowo et al., 2021). Thus, by combining statistical and judgemental forecasts, companies can gain higher accuracy in demand forecasting.

In many companies, the combination of statistical and judgemental forecasts is often formalised as part of the Sales & Operations Planning (S&OP) process in which various business functions collaborate to produce so-called "final forecasts". During this meeting, the statistical forecasts are reviewed and sometimes

CONTACT Budhi S. Wibowo budhi.sholehwibowo@ugm.ac.id Mechanical and Industrial Engineering Department, Universitas Gadjah Mada, Jl. Grafika 2, Sleman, Yogyakarta 55284, Indonesia

compared with judgemental forecasts from the sales and marketing function. Using their expertise, intuition, and available data, the managers must determine whether to adjust the statistical forecasts, possibly taking into account the judgemental forecast or leaving them unaltered.

Academic literature has documented support for judgemental adjustment of the statistical forecast. Experimental and empirical studies showed that such an adjustment could add value to forecast accuracy (see Perera et al. (2019) for a comprehensive review). However, specific principles must be followed for the judgemental adjustment to be effective. Sanders and Ritzman (2001) suggested that intervening statistical forecasts should only be applied when domain knowledge is present and environmental changes are known. Moreover, the adjustment's size and direction should be carefully evaluated, given that human judgements often exhibit systematic biases (Eroglu & Croxton, 2010; Fildes et al., 2009). Thus, the detrimental effect might be reduced by making the forecasters aware of the potential judgement biases.

Unfortunately, such principles are often disregarded in practices, leading to reduced forecast accuracy and unintended biases (Fildes & Goodwin, 2021; Franses & Legerstee, 2010). Based on data from four supply-chain companies, Fildes et al. (2009) observed that up to 90% of the statistical forecasts were judgmentally adjusted. Many of those adjustments were trivial and not based on reliable information. Besides, the adjustments were frequently made in the wrong direction and tended to overshoot, indicating poor judgement quality (Fildes et al., 2009; Khosrowabadi et al., 2022; Petropoulos et al., 2016). Worse, forecasters often persistently choose to intervene in the statistical forecasts, despite feedback affirming that it reduces the accuracy (Lim & O'Connor, 1995; Petropoulos et al., 2016). Such persistence can be motivated by "algorithm aversion" and "illusion of control", where people believe that models are imperfect while overestimating their capability to correct them (Dietvorst et al., 2018; Kottemann et al., 1994).

As a response, several strategies have been proposed to minimise the negative impact of judgemental adjustments in demand forecasting, such as training programs for forecasters (Alvarado-Valencia et al., 2017; Eroglu & Croxton, 2010; Eroglu & Knemeyer, 2010; Sanders & Ritzman, 1992), restricting the opportunity to intervene the statistical forecasts (Fildes et al., 2009; Goodwin, 2000b; Petropoulos et al., 2016; Sanders & Ritzman, 2001), and mechanically combining judgement and statistical forecasts (Petropoulos et al., 2016; Sanders & Ritzman, 2001; Wibowo et al., 2021). Interestingly, although the judgemental adjustment of statistical forecasts is a widespread practice in the industry, the decision rules to recognise "when" and "how" to adjust statistical forecasts effectively remain underexplored by researchers.

Recently, policy learning techniques based on causal machine learning (CML) models have gained popularity in various fields, including operation management studies (for example, Huber et al., 2022; Leoni & Nilsson, 2021; Luo et al., 2019; Truong, 2021). The methods allow researchers to simulate experiments and estimate causal effects from observational data (Athey & Wager, 2021; Jacob, 2021). Given the vast amount of forecast data in supply chains, such techniques enable empirical observations to be used to learn the optimal forecast review policy. Our study is among the first that utilised machine learning techniques to investigate the best policy from a large set of supply chain data. The study complements existing approaches in demand planning literature that are typically developed based on conventionally designed experiments and hypothesis testing.

The primary objective of our research is to determine the best policy for reviewing statistical forecasts in supply chains, considering the diverse forecasting conditions and demand characteristics. Our study demonstrates that the policy learning method can offer valuable recommendations to demand planners for enhancing the efficiency of the forecast review process. To achieve this goal, we evaluated three alternative actions, namely, "do-nothing", "follow the judgement", and "simple-average". By leveraging the "policy tree" method, we identify the best policy using over 12,000-point forecast data from five Asia-Pacific-based companies. We then examine the policy's performance in improving forecast accuracy and estimate how the proposed actions deviate from the true optimal. Finally, we discuss the results and provide concluding remarks.

Our study makes the following contributions:

- It showcases the application of a CML framework in evaluating alternative policies designed to improve forecast accuracy during the forecast review process. Companies can leverage this methodology to create customised policies based on their unique forecasting environment.
- It provides an effective policy for demand planners to aid the review of statistical forecasts in the S&OP process, which can be integrated into forecasting support systems (FSS).
- It explores strategies to limit the negative impact of judgemental adjustments on statistical forecasts, which remains scarce in the literature.

The following sections review the literature on judgemental adjustments of statistical forecasting and the applications of CML. Section 3 describes the data and methods used in the study, focusing on the practical application of CML. Section 5 presents the results. Section 6 discusses the findings and managerial implications. Section 6 summarises the main findings and study limitations.

2. Literature review

2.1. Judgemental adjustment of statistical forecasts

Despite the reliance of supply-chain forecasting on computer software, the human factor still plays a crucial role in the actual forecasting process (Arvan et al., 2019; Perera et al., 2019). Several surveys have revealed that the most prevalent forecasting practice in many industries involves experts/humans reviewing and modifying statistically generated forecasts based on their intuition, expertise, or other contextual factors (Fildes & Goodwin, 2007; Sanders & Manrodt, 2003). The process of adjusting forecasts typically comprises two stages. First, deciding whether the statistical forecasts require revision ("when"). Second, determining the size and direction of the adjustments ("how"). The primary motivation for this approach was to correct perceived flaws and consider new information not included in statistical forecasts (Sanders & Ritzman, 2001, 2004).

The literature provides mixed results on whether judgemental adjustments to statistical forecasts improve accuracy. In the beginning, many scholars discouraged using judgement to revise statistical forecasts because it often resulted in worse performance (Armstrong, 1986; Lawrence & O'Connor, 1992, 1996; Sanders, 1992). One reason to be wary of this practice is that judgemental adjustment forecasts are prone to large errors due to biases (Bunn & Wright, 1991). Such biases are often driven by the "illusion of control" and "overconfidence", where forecasters tend to believe that they are more capable of predicting and influencing the outcomes (de Baets & Harvey, 2018; Kottemann et al., 1994). In addition, some studies suggest that human judgements are prone to "loss aversion", which is a tendency for people to treat losses and gains asymmetrically. Tong et al. (2018) reported that the risk-averse forecaster may intentionally choose to over-forecast to avoid stockouts, resulting in "optimism bias". Moreover, organisational politics and performance targets are important in shaping judgemental adjustments to forecasts (Harvey & Reimers, 2013; Lawrence et al., 2006; Webby & O'Connor, 1996). Judgemental forecasts were also characterised by other biases, such as a tendency to see systems in random patterns, dampening downward trends, and a narrow focus on recent observations or recency bias (Fildes & Goodwin, 2021; Lawrence et al., 2006).

Nonetheless, further studies have increasingly recognised the value-added of human judgement in demand forecasting (Baecke et al., 2017). The main recommendation in the literature is that judgemental adjustments should only be made when specific contextual information is available or when using statistical models becomes less reasonable, such as in highly uncertain situations (Alvarado-Valencia et al., 2017; Fildes & Goodwin, 2007; Sanders & Ritzman, 2001; Webby & O'Connor, 1996). Further studies found that humans better recognise negative events than positive ones. Based on empirical data from companies, Fildes et al. (2009) and Syntetos et al. (2009) observed that positive adjustments often lead to larger errors, while negative adjustments improve forecasting accuracy.

Several efforts have been made to assist forecasters in better judging statistical forecasts. Some proposed training programs to improve forecasters' domain and technical knowledge (Alvarado-Valencia et al., 2017; Sanders & Ritzman, 1992). Alvarado-Valencia et al. (2017) suggested adjusting the forecasts only when the forecasters' expertise is high and the credibility of statistical models is low. Other studies suggested training to make the forecasters aware of the potential biases in judgemental adjustment (Eroglu & Croxton, 2010; Eroglu & Knemeyer, 2010).

Another strategy is to restrict unnecessary changes to statistical forecasts. One method is for managers to write down and document the reason for the change before acting (Goodwin, 2000b; Sanders & Ritzman, 2001). The restriction can also be held by preventing minor alterations and large positive adjustments to statistical forecasts (Fildes et al., 2009; Syntetos et al., 2009). To smooth out the adjustment, managers can also consider taking a 50% judgement and 50% statistical forecast (Blattberg & Hoch, 1990). Despite its simplicity, such a method effectively improves accuracy and has been seen as an alternative way to avoid the negative impact of judgemental adjustments (Franses, 2011; Sanders & Ritzman, 2001; Wibowo et al., 2021).

While judgemental adjustment has been a topic of interest in academic literature for the past two decades, there is still a lack of conclusive evidence and significant room for empirical studies and methodological contributions concerning judgemental adjustments in supply-chain forecasting. Our research contributes to the literature by presenting a framework that leverages advanced technologies, such as machine learning to recommend optimal strategies for enhancing the efficiency and effectiveness of judgemental adjustments in forecasting practices.

2.2. Causal machine learning applications in operations management

Over the last few years, CML has garnered attention from researchers to help them to learn optimal policy and maximise operational goals. CML techniques combine advances in machine learning with causal inference theory to imitate randomised experiments and estimate causal effects from treatments (Athey & Wager, 2021; Jacob, 2021). Such methods are handy when researchers do not have access to a well-designed experiment but have some observational data (Hair & Sarstedt, 2021; Kreif & DiazOrdaz, 2019).

Several studies have shown that using CML techniques can help businesses to achieve their operational goals more effectively. For instance, Luo et al. (2019) utilised a CML technique called causal forest optimisation to improve e-commerce cart targeting (ECT) and reduce cart abandonment in e-commerce platforms. Carbo-Valverde et al. (2020) examined features responsible for digitalising bank customers using the causal random forests method. Gubela et al. (2020) studied the effects of a marketing campaign on customers using a technique called causal uplift modelling. They applied the model in a German online marketing agency and discovered that it could help allocate marketing budgets more effectively and target marketing efforts based on customers' behaviour. Truong (2021) employed the CML framework to predict the probability of flight delays. The study highlights the benefit of CML methods in exploring how the variables interact and how these interactions lead to delayed occurrences. In a recent study, Huber et al. (2022) surveyed customers who bought supersaver train tickets to explore how discounts affect demand. Using CML methods, they found that increasing the discount rate can increase the number of rescheduled trips.

Despite the recent advancements in CML techniques, none of the existing studies have explored its potential to support policy in the S&OP context. CML can comprehend intricate interrelationships from vast data pools within supply chains, enabling it to suggest optimal strategies for enhancing forecasting accuracy. Our research aims to contribute to the literature by showcasing the application of state-of-the-art machine learning techniques to identify custom policies in the forecast review process, using extensive archival data from supply chains.

To our knowledge, only Baker (2021) has utilised a machine learning model to address a similar problem to our study. The study uses a classification tree to identify situations where overriding the statistical forecast can give value-added to the accuracy. It works by treating the problem as a prediction task, where the goal is to assign predefined class labels to input data instances based on their features. However, a limitation of using classification trees is that they mainly establish associations in the data without estimating cause-and-effect relationships. In contrast, our study utilises a policy tree algorithm specially designed to estimate cause-and-effect relationships from various alternative decisions. By identifying policies or rules that optimise desired outcomes, this approach goes beyond mere association and transforms the problem into a prescription task. Consequently, we gain a deeper understanding of the underlying causal relationships between actions and their estimated impacts, which enables us to make effective policies based on reliable causal inference.

3. Methods

3.1. Problem formulation

Our study utilises archival data from five Asian-Pacific companies specialising in producing and distributing fast-moving consumer goods for global markets, as outlined in Table 1. To facilitate demand planning, advanced planning systems are employed to manage supply chains, incorporating forecasting software that generates forecasts based on historical data using both univariate and multivariate models. To supplement statistical forecasts, companies acquire judgemental forecasts from the sales and marketing function, which typically depend on qualitative market research and promotion plans.

Subsequently, managers review statistical and judgemental forecasts to create a singular "final forecast". They can adjust the statistical forecasts if they believe some changes are necessary due to external factors (e.g., promotion, special events, competition landscape). The size and direction of the adjustments are made judgmentally based on the expertise and

Table 1. Summary of forecast data from five companies.

	Company				
Summary	A	B	C	D	E
Forecast units (SKUs)	55	57	124	265	52
Forecast points	1,701	1,813	3,950	8,787	1,756
Accuracy of statistical forecasts ($MAPE_S$, %)	79.31	231.22	73.28	24.91	23.70
Accuracy of judgmental forecasts ($MAPE_J$, %)	108.63	212.09	129.78	32.94	28.72
Demand variability (CV_A)	0.44	0.60	0.41	0.20	0.20

information available to the managers. In real situations, managers have the flexibility to make various adjustments as they see fit. However, to simplify the problem, we only consider three possible actions in this study:

- "Do-nothing": ignores the proposed adjustment and leaves the statistical forecast untreated.
- "Follow the judgement": replaces the statistical forecast with the judgemental forecast.
- "Simple-average": takes the average of statistical and judgemental forecasts as the final forecast (Blattberg – Hoch method).

The objective of the problem is to assign actions to the baseline forecasts that provide the highest accuracy improvement.

We formulate the instance as a policy learning problem where the objective is to develop a policy that prescribes actions that result in the best outcomes given a set of observed features in the data (see Figure 1). Suppose we have independent and identically distributed data $\{(\mathbf{x}_i, y_i, z_i)\}_{i=1}^{n}$ comprising n observations of the forecasting task. Each observation i comprises a vector of features $\mathbf{x}_\mathrm{i} \in \mathbb{R}$, an outcome $y_i \in \mathbb{R}$, and a set of available actions $z_i \in (1, \ldots, m)$. We attain a reward Q_{iz} for every observation i under a prescribed action k. In this problem, the reward Q_{iz} is measured in *absolute percentage error* (APE), such that a lower reward corresponds to better action. Therefore, we formulate the problem in a minimisation mode:

$$\min_{\pi(.)} \sum_{i=1}^{n} \sum_{z=1}^{m} 1\{\pi(\mathbf{x}_i) = k\} \cdot Q_{ik} \tag{1}$$

where $\pi(\mathbf{x})$ is a policy that recommends actions towards observations based on observed features $\mathbf{x}$, and $1\{.\}$ is the indicator function that takes value one if the argument is valid and zero otherwise.

To obtain a suitable dataset for policy learning, we conducted a numerical experiment based on empirical data, which allowed us to directly observe the "true outcome" for any observation under each prescription. The dataset and the details of the feature engineering process are discussed in the following subsection.

3.2. Feature engineering

The data for the study is based on the archival data of three interrelated time series: actual demand **A**, statistical forecasts **S**, and judgemental forecasts **J**. Each set comprised 18,007 points, representing the monthly observations of 553 forecast units within three years. The forecasting features **x** corresponds to the situational characteristics of the forecasting task at time t. We assume that demand planners only consider the characteristics of the last six observations in a forecasting task. This assumption was made considering the natural bias commonly observed in forecasters focusing primarily on recent observations rather than the longer term (Fildes & Goodwin, 2021).

Based on the assumption, each feature in vector **x** is measured based on the last six observations in the triples (refer to Table 2 for details). To measure the features, we first reshape the time series into a collection of individual forecasting tasks, where each instance comprises the last six observations of the triples. Note that forecasts between consecutive periods might be correlated. Nevertheless, assuming independence between consecutive forecasts is reasonable since forecasters typically perceive problems as distinct and separate instances (Trapero et al., 2013). Further, we excluded instances with incomplete triples or zero actuals in over half of the series. We normalised the data using L1 regularisation to eliminate scale variations while preserving the covariance relationships between triples. Finally, we established ten features using the information obtained from the triples. The final dataset for the study consists of 12,023 observations.

3.3. The policy tree learning method

To address the challenge of policy learning, we employed a novel method known as *the policy tree* (Amram et al., 2022; Athey & Wager, 2021), a variant of recent CML techniques. The policy tree

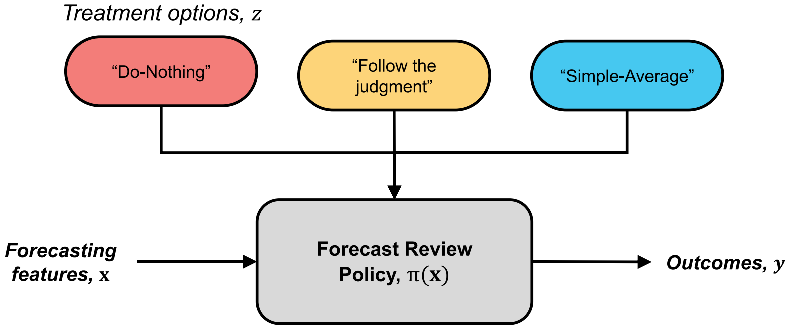

Figure 1. Policy learning model.

Table 2. Features of a forecasting task.

Group	Feature	Definition	Supporting References
Demand characteristics	CV_A	Coefficient of the variation of detrended actuals in the last six observations	(Abolghasemi et al., 2020; Sanders & Ritzman, 1992; Wibowo et al., 2021)
	β_A	The trend of actuals in the last six observations	
Past performance of judgmental forecasts	$MAPE_J$	MAPE of judgmental forecasts based on the last six observations; a higher value indicates lower forecast accuracy	(Eroglu & Croxton, 2010; Sanders & Ritzman, 2001)
	MPE_J	Mean percentage error from judgmental forecasts based on the last six observations; a positive value indicates a tendency to over-forecast (positive bias), whereas a negative value indicates a tendency to under-forecast (negative bias).	
Past performance of statistical forecasts	$MAPE_S$	MAPE of statistical forecasts based on the last six observations; a higher value indicates lower forecast accuracy.	(Alvarado-Valencia et al., 2017; Sanders & Ritzman, 2001)
	$MAPE_S$	Mean percentage error from statistical forecast based on the last six observations; a positive value indicates a tendency to over-forecast (positive bias), whereas a negative value indicates a tendency to under-forecast (negative bias).	
Error correlation between judgmental and statistical forecast	$\lvert r^{\varepsilon}_{JS} \rvert$	The relationship strength between judgmental and statistical forecast errors in the last six observations; a higher value indicates a stronger relationship;	(Armstrong, 2001; Blanc & Setzer, 2016; Goodwin, 2000a)
	$\overrightarrow{r^{\varepsilon}_{JS}}$	The relationship direction between judgmental and statistical forecast errors in the last six observations; $\overrightarrow{r^{\varepsilon}_{JS}}$ is "1" if the correlation is positive, and "−1" otherwise.	
Difference between judgmental and statistical forecast	$\lvert \delta_{JS} \rvert$	The absolute difference between judgmental and statistical forecasts $(t+1)$.	(Fildes et al., 2009; Syntetos et al., 2009)
	$\overrightarrow{\delta_{JS}}$	The discrepancy of judgment $(t+1)$ to the statistical forecast $(t+1)$. The value of $\overrightarrow{\delta_{JS}}$ is "1" if the judgmental forecast is greater than or equal to the statistical forecast and "−1" otherwise.	

leverages the well-established decision tree learning algorithm to recommend actions that can lead to the desired outcome, given the observed data features. The decision tree algorithm is appealing for policy learning as it offers transparent decision logic represented in a simple binary tree structure. This interpretability allows domain experts to inspect and comprehend the decision-making process and make corrections when necessary. This is particularly important in prescriptive analytics, where transparency and interpretability of models are vital, as the recommended actions can have significant and direct consequences. We provided an example of a policy tree in Figure 2.

In brief, the policy tree algorithm operates by recursively dividing observations into subgroups or leaves. This is achieved by conditioning feature values to assign each observation to a specific leaf. All the observations belonging to a given leaf receive a similar course of action, enabling the observation of the reward or treatment effects. This reward is computed by calculating the difference in average outcomes between treated and untreated observations

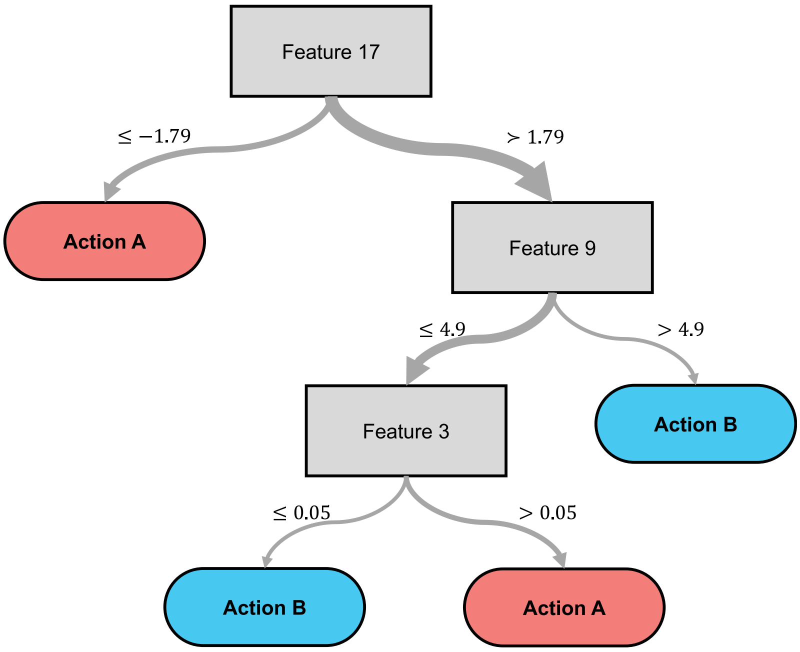

Figure 2. Example of a policy tree prescribing binary treatments based on feature values.

within the sample. The relative cost of alternative actions is evaluated based on this reward. The main objective of the policy tree is to optimise the total value of the policy by identifying the most distinct splits where a treatment creates the largest difference. To this end, the policy tree uses the relative values of the rewards for each observation rather than focusing solely on actions that yield the best reward.

Specifically, let us denote the group of observations in a leaf as v and the action prescribed for each leaf ℓ as $z_\ell \in (1, \ldots, m)$. Thus, the objective of the policy tree can be formulated as follow:

$$\min_{v(.),z} \sum_{i=1}^{n} \sum_{\ell} 1\{\mathrm{v}(\boldsymbol{x}_i) = \ell\} \cdot Q_{iz_\ell}, \tag{2}$$

such that for each observation i, we identify the leaf $\ell = \mathrm{v}(\mathbf{x}_i)$ containing this observation and use the reward Q_{iz_ℓ} corresponding to the prescription in this leaf. Therefore, given a tree structure $\mathrm{v}(\mathbf{x}_i)$, we can find the best action for each leaf ℓ by enumerating all the alternative actions as follows:

$$z_\ell = \arg\min_{k} \sum_{i:\mathrm{v}(\mathbf{x}_i)=\ell} Q_{ik} \tag{3}$$

In all our experiments, we utilise *policytree* package for *R* (Sverdrup et al., 2020) to train the decision tree. Given that finding an optimal tree with arbitrary depth is an NP-hard problem, we only train a shallow tree (with a depth of less than five) using a hybrid greedy search. We also implemented *the honesty* principle to obtain consistent reward estimates, which divided the dataset into two subsamples randomly. Half of the data was used for constructing the trees, and the other half was used to estimate the reward. The *honest* estimation guarantees consistency and asymptotic normality in the estimator, providing reliable confidence intervals of the parameter estimated (Athey & Imbens, 2016).

3.4. Evaluation procedure

We used the following procedure to train and assess the effectiveness of the prescribed policy:

(1) Given the dataset $(\mathbf{x}_i, y_i, z_i)$, split the observations randomly into two subsets: 80% training set and 20% test set. The training set is used to construct the policy and evaluate the expected rewards, whereas the test set is used to validate the policy performance.

(2) Train the policy tree using the features $\mathbf{x}_i$ and the rewards Q_{it} from the test set to yield a prescribed policy.

(3) Using known observations in the training set:

a. Evaluate the *mean reward* of the prescribed policy by averaging the reward from the prescribed action. This measure corresponds to the MAPE of final forecasts. The use of MAPE as the main performance measure is due to its widespread adoption in companies. Additionally, MAPE aligns well with the policy tree algorithm, which employs the arithmetic mean to evaluate rewards for each action.

b. Evaluate the *mean difference* by taking the difference between the outcome of the prescribed policy and the outcome of the "do-nothing" policy, averaged across all the points in the set. The objective is to evaluate how far the prescribed policy can improve the baseline of forecasting performance.

c. Evaluate the *mean regret* of the prescribed policy. This step evaluates how far the prescribed policy is from the "true optimal". The "optimal" policy corresponds to the set of optimal actions based on the true outcomes.

d. Evaluate the average relative mean absolute error (AvgRelMAE) of the prescribed policy. AvgRelMAE is a new forecast accuracy metric proposed by Davydenko and Fildes (2013) to measure the improvement gained from a forecasting technique that often has a skewed distribution. This new measure accompanies the MAPE as the main metric as it has some known drawbacks (see Koutsandreas et al. (2022) for details). The AvgRelMAE takes the geometric mean of relative absolute errors between the proposed and baseline forecasts. Let MAE_i^f denote the mean absolute errors of forecast series i, and $R_i = \frac{\mathrm{MAE}_i^f}{\mathrm{MAE}_i^b}$ denotes the relative measure between proposed forecasts and the baseline. The AvgRelMAE is formulated as follows: The value of AvgRelMAE shows the relative efficiency between the two forecasts. When AvgRelMAE < 1, the proposed method is better than the baseline method. In contrast, AvgRelMAE > 1 indicates the opposite.

$$\mathrm{AvgRelMAE} = \exp\left[\frac{1}{\sum_{i=1}^{m} n_i} \sum_{i=1}^{m} n_i \ln(R_i)\right] \tag{4}$$

(4) Use the test set to perform an out-of-sample evaluation to evaluate the mean reward, improvement, and regret. This step corresponds to the unbiased performance evaluation of the prescribed policy.

The performance of the prescribed policy is also validated by comparing the outcomes under the

Table 3. Descriptive summary of covariates.

Feature	Mean	Std. Deviation	25th percentile	50th percentile	75th percentile	VIF
CV_A	0.285	0.267	.121	.183	.346	3.05
β_A	0.135	1.965	−.691	−.041	.705	1.21
$MAPE_J$ (%)	55.1	132.1	14.4	24.4	48.2	1.91
$MPE_J(\%)$	−228.2	127.93	−272.2	45.3	135.5	1.36
$MAPE_S$ (%)	38.5	89.7	1.7	18.1	33.6	1.85
MPE_S (%)	−119.4	69.49	−144.8	16.3	77.01	1.35
$\left\|r^{\varepsilon}_{JS}\right\|$	0.614	0.282	.398	.675	.857	3.91
$\overrightarrow{r^{\varepsilon}_{JS}}$	0.764	0.645	1.000	1.000	1.000	2.83
$\|\delta_{JS}\|$	0.344	1.152	.076	.176	.364	1.10
$\overrightarrow{\delta_{JS}}$	−0.031	1.000	−1.000	−1.000	1.000	1.13

prescribed policy with the outcome of the "constant treatment" policy, i.e., "simple-average" and "follow the judgement" as a benchmark.

4. Results

4.1. Descriptive analysis

Table 3 provides a descriptive summary of the covariates based on data from all five companies. The summary reveals that the demand series have low variability, as shown by their low mean of CV_A. They also have a relatively balanced distribution of upward and downward trends based on the distribution of β_A. We also observe that the statistical forecasts outperform judgemental forecasts,x with the MAPE distribution indicating lower error rates for statistical forecasts. Additionally, judgemental forecasts show higher mean biases than statistical forecasts, with most biases skewing right, indicating a tendency towards positive biases.

Table 4 presents the correlations between covariates, which indicate interesting relationships. The correlations demonstrate that both statistical and judgemental forecast performance moderately correlates with demand volatility. In other words, the more volatile the demand, the poorer the forecast performance. We also observe that the gap between statistical and judgemental forecasts is associated with the forecast performance of both types of forecasts. This finding suggests that the more unpredictable the demand, the more likely the judgemental and statistical forecasts differ.

Table 4 also reveals that judgemental biases (MPE_J) have a negative correlation with the discrepancy between judgemental forecasts to the statistical forecasts $(|\delta_{JS}|)$. This result indicates that systematic biases are mostly observed when the judgements are higher than the statistical forecasts, suggesting "optimism" biases. Interestingly, the biases of statistical forecasts (MPE_S) positively correlate with demand trends (β_A). This finding suggests that the statistical models tend to dampen strong trends and follow the mean of historical demand.

Regarding the relationships between the covariates, Table 3 demonstrates that they exhibit low correlations, implying that their individual contributions to the model can be distinctly identified. This low correlation indicates the presence of low multicollinearity, which is further supported by the low scores of the variance inflation factors (VIF). The VIF is a measure that quantifies how much the estimated regression coefficients' variance is inflated due to multicollinearity. To calculate the VIF for a predictor variable, the reciprocal of the coefficient of determination obtained from a regression model involving all other predictor variables is taken. Typically, a VIF value lower than five indicates low multicollinearity.

Further, we delved deeper into the quality of human judgements as they are the primary benchmark during forecast review. To assess the judgements, we categorised them based on their size and direction of adjustment (see Figure 3). A judgement was considered incorrect if it moved the final forecast further from the actuals. If the judgement had the correct direction but recommended a 25% higher adjustment

Table 4. Kendall's correlations of covariates.

	CV_A	β_A	$MAPE_J$	MPE_J	$MAPE_S$	MPE_S	$\left\|r^{\varepsilon}_{JS}\right\|$	$\overrightarrow{r^{\varepsilon}_{JS}}$	$\|\delta_{JS}\|$	$\overrightarrow{\delta_{JS}}$
CV_A	1									
β_A	0.037	1								
$MAPE_J$	0.455	0.042	1							
MPE_J	−0.083	0.034	−0.224	1						
$MAPE_S$	0.524	−0.016	0.553	−0.104	1					
MPE_S	−0.145	0.208	−0.150	−0.221	−0.216	1				
$\left\|r^{\varepsilon}_{JS}\right\|$	0.178	−0.006	0.068	0.070	0.139	−0.033	1			
$\overrightarrow{r^{\varepsilon}_{JS}}$	0.065	−0.029	−0.019	0.090	0.055	−0.006	0.296	1		
$\|\delta_{JS}\|$	0.254	0.035	0.338	−0.077	0.287	−0.073	0.008	−0.046	1	
$\overrightarrow{\delta_{JS}}$	−0.040	−0.059	0.053	−0.251	−0.018	−0.049	−0.059	−0.056	0.045	1

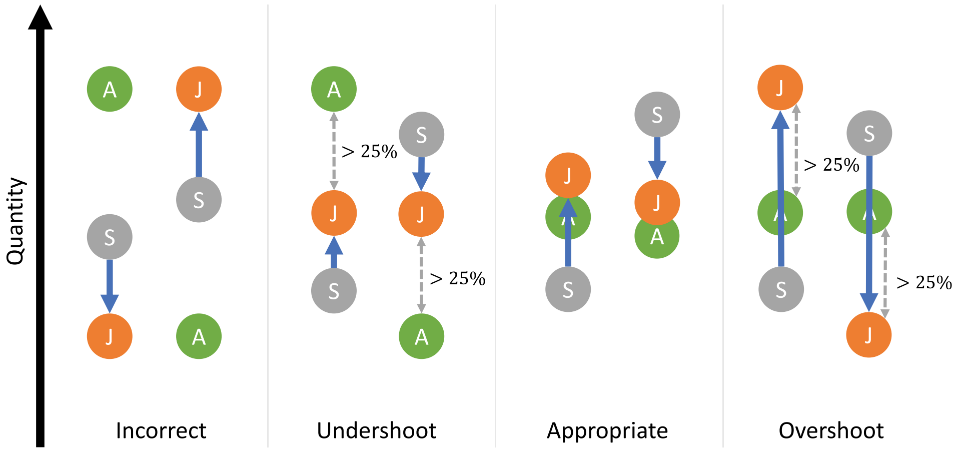

Figure 3. The quality of human judgments based on the size and direction of adjustment (A = actual, S = statistical forecast, J = human judgment).

Table 5. Quality of positive and negative judgements measured in proportion (%) of the total column.

	Positive judgments	Negative judgments	Total
Wrong Direction	2,840 (47%)	2,093 (35%)	4,933 (41%)
Undershoot	725 (12%)	1,077 (18%)	1,802 (15%)
Appropriate	483 (8%)	837 (14%)	1,320 (11%)
Overshoot	1,994 (33%)	1,974 (33%)	3,968 (33%)
Total	6,042 (100%)	5,981(100%)	12,023 (100%)

than required, it was considered an “overshoot”. Conversely, if the proposed adjustment was 25% smaller than required, it was considered an “undershoot”. If the proposed adjustment was within ± 25% of the actuals, it was considered “appropriate”. We present our results in Table 5.

The result showed that 41% of judgements were in the wrong direction (incorrect). For cases with correct directions, 81% of judgements are either overshoot or undershoot, with only a small fraction (11%) having the appropriate size and correct direction. Additionally, we found that positive judgements often performed worse than negative judgements. Positive judgements tended to overshoot (33%) and signal false adjustment direction (47%) compared to negative judgements.

4.2. Optimal policy tree

We implemented the *policy tree* method to learn the optimal policy based on 9,618 samples in the training set (80% of the dataset). To overcome overfitting and improve the generalisation of the resulting policy, we applied the *honesty* principle in the analysis, where we split the training set into two subsamples: 50% for splitting and 50% for estimation purposes. The final tree is shown in Figure 4, where it takes a simple structure with only two levels of depth, two decision nodes, and three terminal leaves. Among the available covariates, only two out of ten covariates were included in the final tree, i.e., the direction of judgement ($\overrightarrow{\delta_{JS}}$) and the accuracy of statistical forecasts (MAPE$_s$).

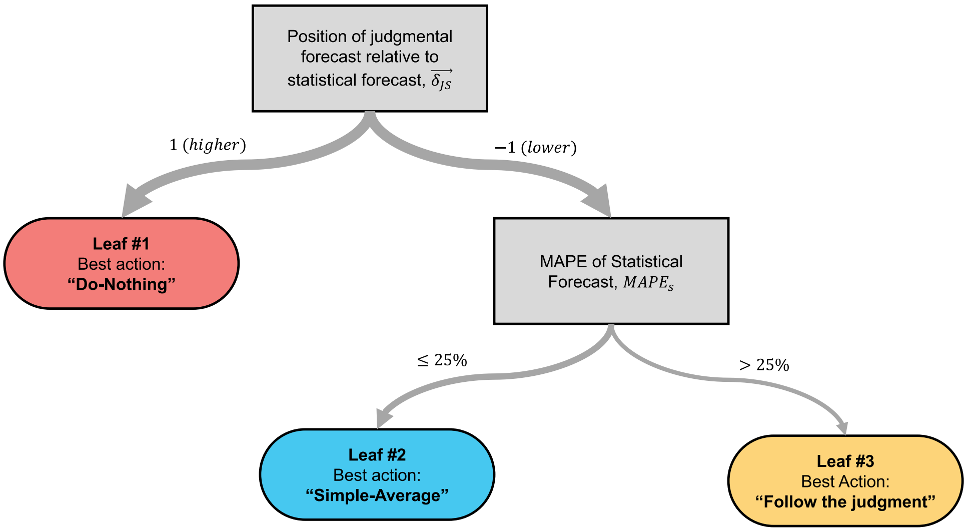

Figure 4. Policy tree inferred from the data.

Table 6. Mean rewards of performing an action on each leaf within the policy tree. Lower values indicate a better course of action.

	Action: "Do-Nothing"		Action: "Follow the judgment"		Action: "Simple-Average"	
Leaf	MAPE, %	Std. Error	MAPE, %	Std. Error	MAPE, %	Std. Error
Leaf #1	32.04	1.4	67.29	3.5	46.75	2.3
Leaf #2	23.21	1.5	22.09	0.8	19.90	1.0
Leaf #3	119.69	10.2	61.50	3.3	83.90	6.4

Table 7. Mean improvement relative to "do-nothing". A positive value indicates a better course of action. The statistical significance of the median difference is tested using the Wilcoxon signed-rank test.

		Action: "Follow the judgment"				Action: "Simple-Average"			
Leaf	Estimating subsample	Mean diff.	Median diff.	Std. Error	p-value	Mean diff.	Median diff.	Std. Error	p-value
Leaf #1	2,329 (48%)	−35.30	−12.00	2.60	<0.01	−14.71	−4.00	1.28	<0.01
Leaf #2	1,515 (32%)	1.10	−2.00	1.00	0.278	3.31	1.01	0.59	<0.01
Leaf #3	965 (20%)	58.20	11.52	8.10	<0.01	35.50	12.01	4.07	<0.01

Table 8. Feature summary showing the mean and standard deviation (in brackets) on each leaf in the policy tree.

Feature	Leaf #1 "Do-Nothing"	Leaf #2 "Simple-Average"	Leaf #3 "Follow the Judgment"
CV_A	0.265 (0.236)	.180 (.162)	0.518 (0.339)
β_A	0.023 (1.799)	.248 (1.380)	0.277 (2.957)
$MAPE_J$	64% (163.9%)	21.7% (3.4%)	89.4% (150.6%)
MPE_J	4.215 (16.320)	−.389 (4.450)	2.045 (13.063)
$MAPE_S$	35% (69%)	13.1% (5.8%)	91.2% (167.9%)
MPE_S	1.086 (5.119)	−.183 (1.918)	4.079 (13.541)
$\lvert r^{\varepsilon}_{JS}\rvert$	0.591 (0.284)	.595 (.274)	0.699 (0.274)
$\overrightarrow{r^{\varepsilon}_{JS}}$	0.720 (0.693)	.774 (.633)	0.839 (0.544)
$\lvert\delta_{JS}\rvert$	0.442 (1.512)	.172 (.174)	0.355 (0.256)
$\overrightarrow{\delta_{JS}}$	1 (0)	−1 (0)	−1 (0)

The results of the rewards and improvements from each action on each leaf are presented in Tables 6–8. Table 6 displays the rewards of three alternative actions for each leaf. It is important to note that the action reward is measured in a percentage-based metric (MAPE) that can exceed 100%. This can occur when the actual demand is very low, and the forecasted demand is positive. The optimal action for each leaf was determined based on the highest reward or lowest MAPE. For instance, the recommended action in Leaf #1 is to "do-nothing", i.e., follow the statistical forecasts. Table 7 demonstrates that the median difference between the rewards on each leaf is statistically significant ($p < 0.01$). Due to the non-parametric nature of the decision tree algorithm, medians were employed in hypothesis testing. Finally, Table 8 summarises each leaf's features to describe the unique characteristics observed across the leaves.

The final tree model proposes a sequential decision-making process to enhance the accuracy of forecast reviews. The model first considers the discrepancy between the judgemental and statistical forecasts, $\overrightarrow{\delta_{JS}}$. If the discrepancy is positive ($\overrightarrow{\delta_{JS}} = 1$), the tree recommends a "do-nothing" action, where the statistical forecast is left unaltered. This action yields a reward improvement of 32.04 compared to alternative actions such as "follow the judgement" and "simple-average", resulting in an average decrease of 67.29 and 46.75 points from the baseline, respectively.

In contrast, if the judgemental forecast has a negative discrepancy ($\overrightarrow{\delta_{JS}} = -1$), the tree recommends a negative adjustment to the statistical forecast. However, the size of this adjustment depends on the historical performance of the statistical forecast. When the statistical forecast performs well, the tree recommends a smaller adjustment by averaging the statistical and judgemental forecasts through a "simple-average" approach. This results in a modest increase of 3.31 percentage points on average over the baseline. Conversely, when the

Table 9. The performance of the optimal policy tree on the training set, compared to constant treatment policies.

Policy	Mean Reward	Mean Improvement	Mean Regret	AvgRelMAE
"Do-nothing"	46.12	0	19.75	1
"Simple-average"	45.10	1.02	18.73	1.045
"Follow the judgment".	51.13	−5.01	24.76	1.321
Policy tree	33.46	12.66	7.09	0.950

statistical forecast performs poorly, the tree prescribes a “follow the judgement” action, which yields a significantly higher reward with 57.70 improvement points on average over the baseline. This significant effect is due to the high demand variability and poor accuracy of the statistical forecasts in this subgroup (refer to Leaf #3 in Table 8).

Collectively, the optimal policy derived from the tree structure produces a promising outcome, demonstrating a mean improvement of 12.66 percentage points compared to the statistical forecasts, as presented in Table 9. The AvgRelMAE scores highlight that our policy was the sole solution that effectively enhanced the baseline. Conversely, judgemental forecasts were, on average, 32 percentage points lower than the baseline, and the “simple-average” outcomes did not yield significant improvements. Furthermore, our policy displayed an excellent mean regret performance, with a marginal difference of only 7.20 points from the “true optimal” outcomes.

4.3. Policy evaluation

We evaluated the performance of our policy on 2,406 new observations in the test set, which accounts for 20% of the dataset. We used mean reward, improvement, regret, and AvgRelMAE to evaluate its performance, and the results are presented in Table 10. The optimal policy tree yielded the largest mean improvement, outperforming statistical forecasts by 9.34 percentage points. This improvement is almost identical to the training result, indicating that the model did not overfit and that the recommended policy can generalise well to new observations. We also observed that the reward from the policy deviates only 7.52 percentage points from the “true optimal”, which is much closer than other constant treatments, demonstrating its effectiveness in enhancing forecast accuracy.

However, it is worth noting that the policy tree model is based on regression tree learning, which is non-robust and can lead to large changes in the tree structure with minor changes in the data. To evaluate the robustness of our proposed policy, we performed a numerical simulation by varying the number of training samples between 10% to 80% of all observations while keeping the test set fixed. We conducted ten replications of the sampling process and took the average to evaluate the outcomes, as shown in Figure 5. The results indicate that the model quickly converges to the current solution with only 30% of

Table 10. “Out-of-sample” performance of the optimal policy tree on the test set, compared to constant treatments.

Policy	Mean Reward	Mean Improvement	Mean Regret	AvgRelMAE
“Do-nothing”	44.34	0	16.86	1
“Simple-average”	47.83	−3.48	20.35	1.033
“Follow the judgment”	58.56	−14.21	31.08	1.383
Policy Tree	35.00	9.34	7.52	0.922

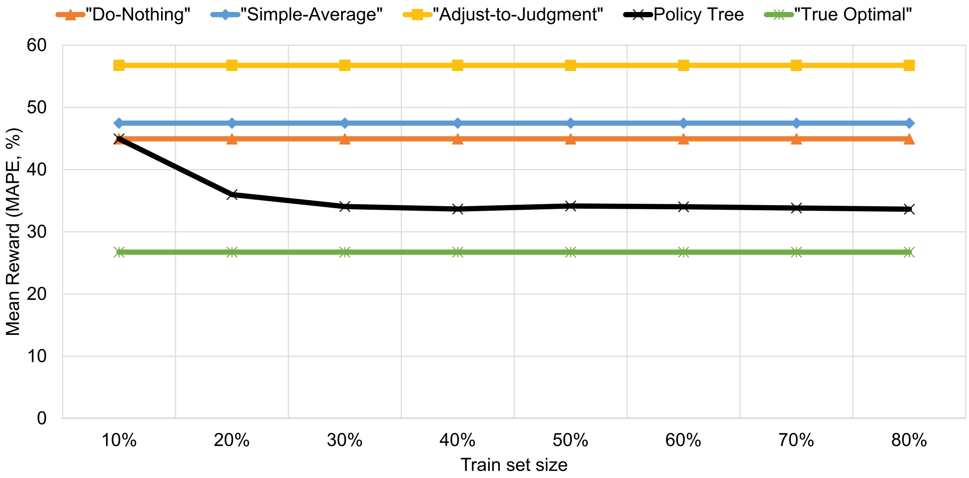

Figure 5. “Out-of-sample” performance from different amounts of training data.

Table 11. Performance of the optimal policy tree on five company data.

Company	Baseline (*$MAPE_5$*)	Mean Reward	Mean Improvement	Mean Regret	AvgRelMAE
A	79.81	57.24	22.56	9.71	0.912
B	231.22	154.68	76.54	13.35	0.884
C	73.28	56.28	16.41	10.33	0.937
D	24.91	20.95	3.96	5.73	0.953
E	23.70	21.12	2.58	6.51	0.970

observations, suggesting that the proposed solution is relatively robust and that the tree structure is unlikely to differ much with additional observations.

We also tested the effectiveness of our prescribed policy on data from five different companies to assess its heterogeneous effect on each organisation (refer to Table 11 for results). The outcome indicated that the policy significantly improves the forecasting accuracy of all five companies, as demonstrated by the mean improvement. In addition, the AvgRelMAE scores showed that our policy outperforms the baseline by 3–11 percentage points. Interestingly, companies with relatively poor forecasting performance (A, B, and C) benefit the most from the policy, gaining an average two-digit accuracy improvement. Conversely, companies with relatively good forecasting performance (D and E) only receive modest improvements with an average of 2–3 percentage points. Despite the small magnitude, these improvements were close to the "true optimal" with only a 5–6 percentage point gap.

4.4. Policy comparison

In the literature, several policies exist that propose ways to improve the effectiveness of judgemental adjustment of statistical forecasts in supply chains. Therefore, we compared our policy with four alternative policies to evaluate its efficiency. Here are the policies we evaluated:

(i) The first policy, based on Fildes et al. (2009), recommends taking a "simple-average" when the judgement is positive ($\overrightarrow{\delta_{JS}} = 1$) and "following the judgement" when the judgement is negative ($\overrightarrow{\delta_{JS}} = -1$).

(ii) The second policy, based on Syntetos et al. (2009), suggests restricting positive adjustments and only accepting large negative adjustments during forecast review. We consider an absolute adjustment as large when it is 10% greater than the statistical forecast ($|\delta_{JS}| > 0.1$).

(iii) The third policy, based on Alvarado-Valencia et al. (2017), advises "follow the judgement" only when the credibility of a statistical model is low and the forecaster's expertise is high. In this study, we measured the credibility of statistical forecasts based on their historical accuracy ($MAPE_S$ >25%) and the forecasters' expertise derived from their forecasting performance ($MAPE_J$ <25%).

(iv) Lastly, the fourth policy, based on Baker (2021), suggests "follow the judgement" only when the accuracy of the statistical forecast is low and the judgement is large and negative. In this study, we interpreted Baker's policy as follows: "follow the judgement" when $MAPE_S$ >25%, $|\delta_{JS}| > 0.1$, and $\overrightarrow{\delta_{JS}} = -1$.

We summarised the comparison results in Table 12. The results indicate that three of the four policies are effective on the test set, with the policy tree showing the largest effect. Specifically, we observed that Fildes et al'.s policy negatively affected both MAPE and AvgRelMAE. Note that Fildes et al.‘s policy ended up with 100% adjustments since the policy recommends either following the judgement or taking a simple average. The outcomes of Syntetos et al.‘s policy had mixed results, with the mean improvement showing a significant positive effect but the AvgRelMAE suggesting a modest negative effect. However, given that the mean improvement is based on the difference between MAPEs, the scores may be biased due to the extreme values in error distribution. The AvgRelMAE is the more reliable measure for forecast comparison as it considers a geometric mean of relative errors instead of an arithmetic mean of percentage errors. Hence, we concluded that Syntetos et al.‘s policy had a small negative effect on our dataset. Additionally, we found that Alvarado-Valencia et al.‘s policy had a moderate effect, with less than one percentage point improvement. Last, when implemented to our dataset, Baker's policy performs worst based on MAPE and AvgRelMAE. Overall, our policy outperformed the alternative policies, with the policy tree showing the largest effect on forecasting accuracy.

5. Discussion

Our dataset is based on five supply-chain companies' archival data in Asia-Pacific. Although it may not represent the globe, the descriptive analysis shows that our dataset has typical data characteristics in the literature. First, we found that on our dataset, the statistical forecast, on average, performs better than judgemental forecasts (Sanders & Manrodt, 2003;

Table 12. Performance comparison of alternative policies on the test set.

Policy	Frequency of adjustments	Mean Reward	Mean Improvement	Mean Regret	AvgRelMAE
Policy tree	52%	35.00	9.34	7.52	0.922
(Fildes et al., 2009)	100%	45.50	–1.15	18.01	1.130
(Syntetos et al., 2009)	35%	35.71	8.64	8.22	1.012
(Alvarado-Valencia et al., 2017)	5%	43.68	.67	16.19	0.992
(Baker, 2021)	19%	45.72	–1.38	18.97	1.151

Webby & O'Connor, 1996). Second, judgemental forecasts tend to overshoot and often signal wrong adjustment directions (Fildes et al., 2009; Petropoulos et al., 2016; Trapero et al., 2013). Third, the judgemental forecasts tend to exhibit positive biases, suggesting a bias towards "optimism" (Eroglu & Croxton, 2010; Fildes et al., 2009; Tong et al., 2018). This trait similarity suggests that our dataset is typical and can be a good basis for analysing typical forecasting practices in supply chains.

We compared three alternative actions during the forecast review and found that constantly following judgemental forecasts from the marketing function can hurt the forecast accuracy by roughly 5–14 percentage points lower than the baseline (refer to Tables 9 and 10). Meanwhile, continuously taking a "simple average" between the two forecasts gives an insignificant difference in accuracy. These results suggest that adjusting statistical forecasts may not have a positive effect on forecasting accuracy.

However, our detailed analysis using a policy tree approach revealed that improving forecast accuracy through adjustments is possible by considering the individual characteristics of the forecasting task. One major benefit of using the policy tree is that it is simple to understand and interpret (white-box model). Specifically, our proposed policy can be summarised into three simple rules:

(i) Restrict positive adjustments of statistical forecasts.
(ii) In a highly uncertain environment, allow only negative adjustments of statistical forecasts.
(iii) When the statistical models perform well, dampen the size of negative judgements using the "simple-average" method.

Our policy shares some similarities with that of Syntetos et al. (2009), which recommended limiting positive adjustments and only allowing large negative adjustments during the forecast review. Such a recommendation is reasonable because most positive adjustments in our dataset exhibit systematic biases which can be driven by motivational orientation (Eroglu & Knemeyer, 2010), performance targets (Harvey & Reimers, 2013), and risk aversion towards lost sales (Tong et al., 2018). Trapero et al. (2013) also suggest that human judgement tends to overshoot during promotional campaigns, indicating "overconfidence" in positive events. Additionally, a recent study by Sroginis et al. (2023) concluded that when sales forecasts are presented, people tend to misinterpret the information and make excessive adjustments, which negatively impacts accuracy. Given these biases in human judgement, avoiding excessive positive adjustments can benefit accuracy.

Our result in Table 7 also shows that avoiding positive adjustments to statistical forecasts can significantly improve forecast accuracy. We observed an average improvement of nearly 35 percentage points and a median improvement of 12 percentage points. One possible explanation for this large improvement is that some companies in our sample use multivariate forecasting techniques that incorporate promotion plans and special events as input. These methods may be more accurate in estimating the demand uplift than relying solely on human judgements.

Furthermore, our policy aligns with most findings in the literature regarding the effectiveness of negative adjustments on statistical forecasts (Fildes et al., 2009; Syntetos et al., 2009). Our data suggest that, on average, negative adjustments improve forecast accuracy, which implies that humans are generally better at detecting negative changes in market demand than positive ones. Fildes et al. (2019) suggested that negative adjustments are often based on reliable arguments rather than intuition or personal interests, which makes them more accurate. Furthermore, it takes courage to hold a negative opinion of forecasts when positive demand growth is highly valued in most organisational settings (Harvey & Reimers, 2013).

Despite the similarities with the policies proposed by Fildes et al. (2009) and Syntetos et al. (2009), our policy offers more specific guidance on how to carry out negative adjustments. Rather than solely relying on the size of adjustments, our policy suggests basing adjustments on the performance of statistical forecasts. Prior studies have also recognised the credibility of statistical forecasts as a critical factor in deciding whether to adjust them. The common conclusions were that the adjustment should only be made when the statistical forecasts have a poor performance (Alvarado-Valencia et al., 2017; Baker, 2021; Sanders & Ritzman, 2001).

When the statistical models perform poorly, our policy suggests following human judgement and proceeding with a negative adjustment. Our analysis suggests that human judgements are more valuable within uncertain situations, especially where statistical models fail to capture demand patterns. In such situations, expertise and contextual information from humans play a bigger role than statistical models for creating accurate forecasts (Alvarado-Valencia et al., 2017; Sanders & Ritzman, 1992, 2001).

In contrast, when the statistical models are reliable, our policy suggests dampening the size of negative adjustment using a "simple-average". Although such action only yields a modest improvement, several studies have reported similar effects from simple-average forecasts, ranging from 1–5 percentage points (Sanders, 1992; Sanders & Ritzman, 2004; Wibowo et al., 2021). The modest improvement is reasonable since the statistical models were already performed well. Therefore, huge improvements were less likely to occur.

This non-linear effect of judgemental adjustment has also been observed in a previous study by Trapero et al. (2011). They employed a non-linear model to correct the bias resulting from judgemental adjustments. The study revealed a preference for human judgements when the adjustment was negative. However, when adjustments were positive, non-linear effects emerged based on the magnitude of the adjustment. Our study differs from theirs in terms of the factor influencing the non-linear effect of negative adjustment. While their research suggests that the adjustment size is the influencing factor, we propose that $MAPE_S$ is a more reliable indicator for determining whether to follow a negative judgement or employ a simple average.

We conducted a comparative analysis of our policy against various existing policies documented in the literature. The assessment outcomes indicate that the benchmark policies display mixed results on our dataset, as shown in Table 12. This outcome variation could be due to dissimilarities in the datasets used in prior studies. Thus, the outcomes may differ depending on the characteristics of the dataset. In their review, Syntetos et al. (2016) observed limited empirical studies in judgemental adjustment research, leading to inconclusive findings. Most of the recommended policies in the literature were derived from samples from specific companies. Hence, they may not be effective as general rules. Therefore, further empirical research is necessary to test and validate proposed solutions in this field, and more data should be made more accessible to facilitate this research.

5.1. Managerial consideration

It is also important to note that the prescribed policy from the decision tree model is prone to change with different forecast measures and data characteristics. Given the variety of individuals and tasks involved in supply chain forecasting, companies may benefit from seeking a personalised policy based on their unique environments. To achieve this, our study proposes a data-driven framework that allows businesses to learn an optimal policy based on observed forecasting features. This framework can help businesses identify actions that increase accuracy and understand key factors that influence the efficacy of statistical forecast adjustment. Our model can also be incorporated into an FSS to support managers in making effective decisions on when and how to adjust baseline forecasts in supply chains.

However, implementing an automated policy to correct bias may face significant challenges. If forecasters know their forecasts will be "corrected", they may perceive their judgement as futile and alter their approach. This could gradually erode their sense of ownership and diminish their motivation to produce high-quality forecasts, knowing their efforts will ultimately be modified (Belton & Goodwin, 1996). One potential solution to address this issue is to separate the users of judgemental forecasts from the review process, preventing them from knowing whether the final forecasts have been adjusted (Sanders & Ritzman, 2004). However, this approach may present practical challenges. It may not be sustainable in the long term since the S&OP process aims to foster information exchange and alignment across departments, facilitating collaborative decision-making. In this context, the final forecast from previous periods may emerge and disrupt the information flow. An alternative approach is to frame the adjustment policy as a "nudge" to influence the behaviour of forecasters (Baker, 2021). By adopting this option, it might be possible to preserve the forecasters' sense of ownership while guiding them to minimise any undesirable biases.

6. Conclusions

This study demonstrated the application of CML methods to identify the optimal policy for improving forecast accuracy during the S&OP process. Specifically, we considered three alternative actions – "do-nothing", "follow the judgement", and "simple-average", - and used more than approximately 12,000 forecast data from five supply-chain companies based in Asia-Pacific for the experiment.

Our study found that abruptly adjusting statistical forecasts has a detrimental effect on forecast accuracy. Instead, managers should tailor their approach to each forecasting task. Using the policy learning method, we identified two key features to consider during the forecast review: the "judgement direction" and the "performance of the statistical model". These features led to three simple decision rules for making appropriate adjustments to statistical forecasts. The prescribed rules are also consistent with prior research indicating the efficacy of negative adjustments, especially when human judgements display positive biases.

We also tested the performance of our policy using real-world data and compared it to other policies in the literature. The results suggest the effectiveness and efficiency of our policy in improving forecast accuracy, with an average 3–11 percentage point improvement from baseline. However, since the forecast environments in supply chains vary, the optimal policy obtained from our dataset may not generalise well to other datasets. Therefore, it may be advisable for businesses to develop a customised policy based on their specific environment. Our policy learning method can serve as a framework for learning the best policy from company-specific data and gaining insights into effectively adjusting statistical forecasts in supply chains.

Our study has some limitations that warrant consideration. Firstly, our sample was drawn from five Asian-Pacific companies in the consumer goods industry, which may only partially represent the global population. Nonetheless, it exhibits typical forecast data characteristics in the supply chain literature, indicating that it is representative. Secondly, our policy learning model only accounts for three types of alternative actions, which, while popular in the literature, may only encompass some of the available adjustment methods. Future studies could expand the scope by including more alternatives or incorporating continuous variable adjustments to provide a more comprehensive analysis. Thirdly, the policy tree structure developed in our study was based on MAPE as a performance measure. While MAPE is widely used, it does have known drawbacks as a performance metric in forecasting. Incorporating more robust metrics could yield further insights and enhance the model's accuracy. Lastly, our policy was developed based on single-point, one-month-ahead forecast data, which may not apply to different forecasting environments, such as those involving long-term or intermittent demand forecasts. In such scenarios, human judgement may be more reliable than statistical forecasts. Finally, there might be some managerial challenges to implement this approach in practice.

Despite these limitations, our study provides valuable insights into improving forecast accuracy through data-driven policy. Future research could expand upon our findings by exploring the applicability of our policy in diverse forecasting environments and incorporating additional adjustment methods into the analysis.

Disclosure statement

No potential conflict of interest was reported by the author.

Funding

This research received no specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data availability statement

Due to the nature of the research and legal concerns, supporting data is unavailable.

References

Abolghasemi, M., Beh, E., Tarr, G., & Gerlach, R. (2020). Demand forecasting in supply chain: The impact of demand volatility in the presence of promotion. *Computers & Industrial Engineering*, *142*(July 2019), 106380. https://doi.org/10.1016/j.cie.2020.106380

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

Amram, M., Dunn, J., & Zhuo, Y. D. (2022). Optimal policy trees. *Machine Learning*, *111*(7), 2741–2768. https://doi.org/10.1007/s10994-022-06128-5

Armstrong, J. S. (1986). The Ombudsman: Research on forecasting: A quarter-century review, 1960–1984. *Interfaces*, *16*(1), 89–109. https://doi.org/10.1287/inte.16.1.89

Armstrong, J. S. (2001). *Combining forecasts* (pp. 417–439). https://doi.org/10.1007/978-0-306-47630-3_19

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

Athey, S., & Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(27), 7353–7360. https://doi.org/10.1073/pnas.1510489113

Athey, S., & Wager, S. (2021). Policy learning with observational data. *Econometrica*, *89*(1), 133–161. https://doi.org/10.3982/ecta15732

Baecke, P., de Baets, S., & Vanderheyden, K. (2017). Investigating the added value of integrating human judgement into statistical demand forecasting systems. *International Journal of Production Economics*, *191*, 85–96. https://doi.org/10.1016/j.ijpe.2017.05.016

Baker, J. (2021). Maximizing forecast value added through machine learning and "Nudges. *Foresight: The International Journal of Applied Forecasting*, (60), 8–15.

Belton, V., & Goodwin, P. (1996). Remarks on the application of the analytic hierarchy process to judgmental forecasting. *International Journal of Forecasting*, *12*(1), 155–161. https://doi.org/10.1016/0169-2070(95)00643-5

Blanc, S. M., & Setzer, T. (2016). When to choose the simple average in forecast combination. *Journal of Business Research*, *69*(10), 3951–3962. https://doi.org/10.1016/j.jbusres.2016.05.013

Blattberg, R. C., & Hoch, S. J. (1990). Database models and managerial intuition: 50% model + 50% manager. *Management Science*, *36*(8), 887–899. https://doi.org/10.1287/mnsc.36.8.887

Bunn, D., & Wright, G. (1991). Interaction of judgemental and statistical forecasting methods: Issues & analysis. *Management Science*, *37*(5), 501–518. https://doi.org/10.1287/mnsc.37.5.501

Carbo-Valverde, S., Cuadros-Solas, P., Rodríguez-Fernández, F., & Xin, B. (2020, October). A machine learning approach to the digitalization of bank customers: Evidence from random and causal forests. *PLoS ONE*, *15*(10), e0240362. https://doi.org/10.1371/journal.pone.0240362

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. (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

Eroglu, C., & Croxton, K. L. (2010). Biases in judgmental adjustments of statistical forecasts: The role of individual differences. *International Journal of Forecasting*, *26*(1), 116–133. https://doi.org/10.1016/j.ijforecast.2009.02.005

Eroglu, C., & Knemeyer, A. M. (2010). Exploring the potential effects of forecaster motivational orientation and gender on judgmental adjustments of statistical forecasts. *Journal of Business Logistics*, *31*(1), 179–195. https://doi.org/10.1002/j.2158-1592.2010.tb00134.x

Fildes, R., & Goodwin, P. (2007). Against your better judgment? How organizations can improve their use of management judgment in forecasting. *Interfaces*, *37*(6), 570–576. https://doi.org/10.1287/inte.1070.0309

Fildes, R., & Goodwin, P. (2021). Stability in the inefficient use of forecasting systems: A case study in a supply chain company. *International Journal of Forecasting*, *37*(2), 1031–1046. https://doi.org/10.1016/j.ijforecast.2020.11.004

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

Franses, P. H. (2011). Averaging model forecasts and expert forecasts: Why does it work? *Interfaces*, *41*(2), 177–181. https://doi.org/10.1287/inte.1100.0554

Franses, P. H., & Legerstee, R. (2010). Do experts' adjustments on model-based SKU-level forecasts improve forecast quality? *Journal of Forecasting*, *29*(3), 331–340. https://doi.org/10.1002/for.1129

Goodwin, P. (2000a). Correct or combine? Mechanically integrating judgmental forecasts with statistical methods. *International Journal of Forecasting*, *16*(2), 261–275. https://doi.org/10.1016/S0169-2070(00)00038-8

Goodwin, P. (2000b). 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

Gubela, R. M. R. M., Lessmann, S., & Jaroszewicz, S. (2020). Response transformation and profit decomposition for revenue uplift modeling. *European Journal of Operational Research*, *283*(2), 647–661. https://doi.org/10.1016/j.ejor.2019.11.030

Hair, J. F., & Sarstedt, M. (2021). Data, measurement, and causal inferences in machine learning: Opportunities and challenges for marketing. *Journal of Marketing Theory & Practice*, *29*(1), 65–77. https://doi.org/10.1080/10696679.2020.1860683

Harvey, N., & Reimers, S. (2013). Trend damping: Under-adjustment, experimental artifact, or adaptation to features of the natural environment? *Journal of Experimental Psychology: Learning Memory and Cognition*, *39*(2), 589–607. https://doi.org/10.1037/a0029179

Huber, M., Meier, J., & Wallimann, H. (2022). Business analytics meets artificial intelligence: Assessing the demand effects of discounts on Swiss train tickets. *Transportation Research Part B: Methodological*, *163*, 22–39. https://doi.org/10.1016/j.trb.2022.06.006

Jacob, D. (2021). CATE meets ML. *Digital Finance*, *3*(2), 99–148. https://doi.org/10.1007/s42521-021-00033-7

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

Kottemann, J. E., Davis, F. D., & Remus, W. E. (1994). Computer-assisted decision making: Performance, beliefs, and the illusion of control. *Organizational Behavior and Human Decision Processes*, *57*(1), 26–37. https://doi.org/10.1006/obhd.1994.1002

Koutsandreas, D., Spiliotis, E., Petropoulos, F., & Assimakopoulos, V. (2022). On the selection of forecasting accuracy measures. *Journal of the Operational Research Society*, *73*(5), 937–954. https://doi.org/10.1080/01605682.2021.1892464

Kreif, N., & DiazOrdaz, K. (2019). Machine learning in policy evaluation: New tools for causal inference. *Oxford Research Encyclopedia of Economics and Finance*. https://doi.org/10.1093/acrefore/9780190625979.013.256

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. (1992). Exploring judgemental forecasting. *International Journal of Forecasting*, *8*(1), 15–26. https://doi.org/10.1016/0169-2070(92)90004-S

Lawrence, M., & O'Connor, M. (1996). Judgement or models: The importance of task differences. *Omega*, *24*(3), 245–254. https://doi.org/10.1016/0305-0483(96)00006-0

Leoni, V., & Nilsson, W. (2021). Dynamic pricing and revenues of Airbnb listings: Estimating heterogeneous causal effects. *International Journal of Hospitality Management*, *95*(March), 102914. https://doi.org/10.1016/j.ijhm.2021.102914

Lim, J. S., & O'Connor, M. (1995). Judgemental adjustment of initial forecasts: Its effectiveness and biases. *Journal of Behavioral Decision Making*, *8*(3), 149–168. https://doi.org/10.1002/bdm.3960080302

Luo, X., Lu, X., & Li, J. (2019). When and how to leverage E-commerce cart targeting: The relative and moderated effects of scarcity and price incentives with a two-stage field experiment and causal forest optimization. *Information Systems Research*, *30*(4), 1203–1227. https://doi.org/10.1287/isre.2019.0859

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., Fildes, R., & Goodwin, P. (2016). Do "big losses" in judgmental adjustments to statistical forecasts affect experts' behaviour? *European Journal of Operational Research*, *249*(3), 842–852. https://doi.org/10.1016/j.ejor.2015.06.002

Remus, W., O'Connor, M., & Griggs, K. (1995). Does reliable information improve the accuracy of judgmental forecasts? *International Journal of Forecasting*, *11*(2), 285–293. https://doi.org/10.1016/0169-2070(94)00578-Z

Sanders, N. R. (1992). Accuracy of judgmental forecasts: A comparison. *Omega*, *20*(3), 353–364. https://doi.org/10.1016/0305-0483(92)90040-E

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. https://doi.org/10.1002/bdm.3960050106

Sanders, N. R., & Ritzman, L. P. 2001. *Judgmental adjustment of statistical forecasts*. Issue January. https://doi.org/10.1007/978-0-306-47630-3_18.

Sanders, N. R., & Ritzman, L. P. (2004). Integrating judgmental and quantitative forecasts: Methodologies for pooling marketing and operations information. *International Journal of Operations and Production Management*, *24*(5–6), 514–529. https://doi.org/10.1108/01443570410532560

Sroginis, A., Fildes, R., & Kourentzes, N. (2023). Use of contextual and model-based information in adjusting promotional forecasts. *European Journal of Operational Research*, *307*(3), 1177–1191. https://doi.org/10.1016/j.ejor.2022.10.005

Sverdrup, E., Kanodia, A., Zhou, Z., Athey, S., & Wager, S. (2020). Policytree: Policy learning via doubly robust empirical welfare maximization over trees. *Journal of Open Source Software*, *5*(50), 2232. https://doi.org/10.21105/joss.02232

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

Tan, T. (2008). Using imperfect advance demand information in forecasting. *IMA Journal of Management Mathematics*, *19*(2), 163–173. https://doi.org/10.1093/imaman/dpn002

Tong, J., Feiler, D., & Larrick, R. (2018). A behavioral remedy for the censorship bias. *Production and Operations Management*, *27*(4), 624–643. https://doi.org/10.1111/poms.12823

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., 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

Truong, D. (2021). Using causal machine learning for predicting the risk of flight delays in air transportation. *Journal of Air Transport Management*, *91*(May 2020), 101993. https://doi.org/10.1016/j.jairtraman.2020.101993

Webby, R., & O’Connor, M. (1996). Judgemental and statistical time series forecasting: A review of the literature. *International Journal of Forecasting*, *12*(1), 91–118. https://doi.org/10.1016/0169-2070(95)00644-3

Wibowo, B. S., Prakoso, Y. J., & Masruroh, N. A. (2021). Performance of judgmental–statistical forecast combination strategies under product-market configurations. *International Journal of Management Science and Engineering Management*, *00*(), 1–14. https://doi.org/10.1080/17509653.2021.2015472