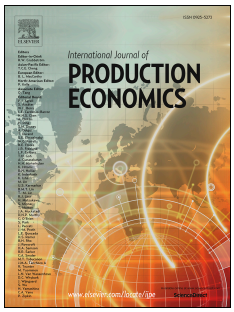
Accepted Manuscript



Investigating the added value of integrating human judgement into statistical demand forecasting systems

Philippe Baecke, Shari De Baets, Karlien Vanderheyden

PII: S0925-5273(17)30165-2

DOI: 10.1016/j.ijpe.2017.05.016

Reference: PROECO 6726

To appear in: *International Journal of Production Economics*

Received Date: 25 July 2016

Revised Date: 29 May 2017

Accepted Date: 31 May 2017

Please cite this article as: Baecke, P., De Baets, S., Vanderheyden, K., Investigating the added value of integrating human judgement into statistical demand forecasting systems, *International Journal of Production Economics* (2017), doi: 10.1016/j.ijpe.2017.05.016.

Investigating the added value of integrating human judgement into statistical demand forecasting systems

Philippe Baecke [a], Shari De Baets [a, b] & Karlien Vanderheyden [a]

[a] *Vlerick Business School, Belgium*

[b] *Ghent University, Belgium*

PHILIPPE BAECKE (first author): philippe.baecke@vlerick.com

Vlerick Business School
Department Marketing
Reep 1
9000 Gent
Belgium
Phone number: 0032 9 210 92 28
Fax number: 0032 9 210 97 00

SHARI DE BAETS (corresponding author): shari.debaets@vlerick.com

Vlerick Business School
Department People and Organisation
Reep 1
9000 Gent
Belgium
Phone number: 0032 16 24 88 26
Fax number: 0032 16 24 88 00

KARLIEN VANDERHEYDEN: karlien.vanderheyden@vlerick.com

Vlerick Business School
Department People and Organisation
Reep 1
9000 Gent
Belgium
Phone number: 0032 9 210 97 67
Fax number: 0032 9 210 97 00

Investigating the added value of integrating human judgement into statistical demand forecasting systems

Abstract

Whilst the research literature points towards the benefits of a statistical approach, business practice continues in many cases to rely on judgmental approaches for demand forecasting. In today's dynamic environment, it is especially relevant to consider a combination of both approaches. However, the question remains as to how this combination should occur. This study compares two different ways of combining statistical and judgmental forecasting, employing real-life data from an international publishing company that produces weekly forecasts on regular and exceptional products. Two forecasting methodologies that are able to include human judgment are compared. In a 'restrictive judgement' model, expert predictions are incorporated as restrictions on the forecasting model. In an 'integrative judgment' model, this information is taken into account as a predictive variable in the demand forecasting process. The proposed models are compared on error metrics and analysed with regard to the properties of the adjustments (direction, size) and of the forecast itself (volatility, periodicity). The integrative approach has a positive effect on accuracy in all scenarios. However, in those cases where the restrictive approach proved to be beneficial, the integrative approach limited these beneficial effects. The study links with demand planning by using the forecasts as input for an optimization model to determine the ideal number of SKUs per Point of Sale (PoS), making a distinction between SKU forecasts and SKU per PoS forecasts. Importantly, this enables performance to be expressed as a measure of profitability, which proves to be higher for the integrative approach than for the restrictive approach.

Keywords: Demand forecasting, judgmental forecasting, human judgment

1. Introduction

Accurate demand forecasting is a first vital step for supply chain management (Fildes et al., 2006). Such forecasts have consequences for decisions within the organisation (e.g., manufacturing, marketing, logistics) and within the larger supply chain (e.g., suppliers, retailers). Forecasting however, is not an easy task and errors can have potential negative effects (e.g., Worthen, 2003). While the optimization of the forecasting process can yield significant advantages such as increased profitability and increased customer service levels (Moon et al., 2003), empirical research remains fairly limited. Specifically, few studies have focussed on real company data (Sanders, 2009) and until recently (Fildes et al., 2009, Franses and Legerstee, 2011, 2013, Syntetos et al., 2016, Trapero et al., 2013), the topic of judgmental adjustments in the operational domain has been overlooked. While experimental research has provided a solid basis for investigating judgmental forecasting, organisation-based research is needed to further understand how forecasts are made (Franses and Legerstee, 2013, Sanders, 2009). Recent examples are the papers from Franses and Legerstee (2009, 2011, 2013), who work with data from a pharmaceutical company; Trapero et al. (2013) with manufacturing company data, and Fildes et al. (2009), who compare four UK-based companies on the efficacy of their judgmental adjustments. We build further on their work by investigating the effect of judgmental forecasting within the context of a publishing company.

We extend their studies in four ways: first, we investigate whether the approach of formally including expert knowledge as an additional explanatory variable as proposed by Franses and Legerstee (2013), although unsuccessful in the pharmacy industry, is more successful in an industry which handles both normal and exceptional products. Incorporating the expectations of experts has the possibility of improving the predictive

performance of these models significantly, especially in the case of promotions (Goodwin, 2002, Goodwin and Fildes, 1999, Trapero et al., 2013, Fildes et al., 2016). This is because the knowledge of expert forecasters represents a previously unmodelled component (Lawrence et al., 2006). Although forecasting models are generally superior to human judgement based on the same information, experts can still add value because they are better able to recognize when predictions should be adapted based on additional information or the existence of exceptional events (Goodwin and Fildes, 1999, Jones and Brown, 2002). A study by Sinha and Zhao (2008) in the field of data-mining demonstrated the added value of expert knowledge on a wide set of classifiers. The current study explicitly compares two different approaches: a restrictive approach and an integrative approach. In the former, predictions are restrictions of the demand forecasting model: i.e., the traditional case of judgmental adjustment of a statistical forecast. In the integrative approach, this information is taken into account as a predictive variable in the demand forecasting model itself. The advantage of this approach is two-fold: the forecasters retain their input and feeling of ownership, while the damaging effects of unnecessary adjustments are mitigated. In addition, this method should motivate forecasters to make the correct adjustment. The more accurate the adjustments of the forecaster were in the past, the more likely this variable will be picked up as a significant predictor for future demand.

Second, we integrate the papers of Fildes et al. (2009) and Franses and Legerstee (2013), by providing an in-depth analysis of the characteristics of the judgmental adjustments (size and direction), and of the data (volatility and periodicity). Regarding volatility, human judgment has been said to be especially relevant in the context of high volatility products due to special events such as promotions (Sanders and Ritzman, 1992). Our dataset gives a unique insight into volatility defined in two ways: as the variation in the data series (Sanders and Ritzman, 1992), and volatility defined as exceptional products. Additionally, this study looks at periodicity of the product as a

proxy for familiarity with the product. To the best of our knowledge, this has not yet been studied in empirical judgmental forecasting literature.

Third, this study distinguishes between two levels of granularity in the forecasting process: SKU and SKU per PoS. Kremer et al. (2015) have pointed out that judgmental forecasting research has not yet studied the effects on hierarchical forecasting. The accuracy of judgmental forecasts on the top-level and the bottom-level are generally not the same (Kremer et al., 2015). The integration of human judgement in forecasting models on SKU per PoS level may create an additional advantage, in that the algorithm can determine in which PoS the factors taken into account by the expert are most influential.

Fourth, to the best of our knowledge, this is the first study that includes data on both sales numbers and profit margins. Steenburgh et al. (2003) already showed that even small improvements in predictive performance can have a serious impact on a firms profitability. In order to investigate both aspects, this study makes a distinction between the forecasting system and the optimization system, as recommended in previous research (Fildes et al., 2009). The forecasting system aims to predict the demand of a product, whereas the goal of the optimization model is to maximize profit, taking price, production cost, delivery cost, recollection cost and expected revenue into account. Whereas the forecasting system enables expressions of accuracy with measures such as MAPE or MdAPE, the optimization model indicates the results via a profitability metric. The literature associated with these gaps is summarized in Table 1.

Table 1: Gaps - overview of core papers

	Empirical Research	Properties of the forecast and of adjustments	Granularity of the forecast	Effects on profitability
Fildes, Goodwin, Lawrence, & Nikolopoulos, 2009	X	X		
Franses & Legerstee, 2009	X	X		
Franses & Legerstee, 2013	X	X		
Hyndman, Ahmed, Athanasopoulos, & Shang, 2011	X		X	
Kremer, Siemsen, & Thomas, 2015			X	
Syntetos, Nikolopoulos, Boylan, Fildes, & Goodwin, 2009	X	X		
Syntetos, Kholidasari, & Naim, 2016	X	X		
Trapero, Pedregal, Fildes, & Kourentzes, 2013	X	X		
Zotteri & Kalchschmidt, 2007			X	
This study	X	X	X	X

2. Background Literature

The potential for accurate demand forecasting has only increased because of the exponential increase in computational power, the rise of the internet and the decrease in data warehousing costs. However, due to accelerated product lifecycles and unpredictable customer demand, forecasting remains challenging (Merzifonluoglu, 2015). Data analytics and human judgment need to be combined to achieve better results, rather than being seen as a dichotomy (Ransbotham et al., 2016). However, it has proven to be a significant challenge to successfully combine the analytical power of computers with human judgment of organisational forecasters. Forecasters usually have a choice between leaving the prediction of the statistical model as it is, or adjusting it in a number of ways. Judgmental adjustment takes the statistical model as starting point, and allows for an adjustment based on human judgment. Given its cost efficiency and it being easy-to-use, it is not surprising that this method is common practice (Turner, 1990). The downside is the possibility of biases and unnecessary adjustments (e.g., Eroglu and Croxton, 2010, Lawrence et al., 2006, Webby and O'Connor, 1996). The latter is hypothesized to be caused by the illusion of control effect, where people have increased confidence in forecasts they have adjusted (Kotte-

man et al., 1994), and the tendency of humans to see patterns in noise (Harvey, 1995). On the other hand, forecasting models have difficulties incorporating the effect of exceptional events such as promotions or new product launches (Goodwin, 2002, Goodwin and Fildes, 1999, Scarpel, 2015). As a result, expert judgment is complementary with the process of mechanically analysing large amounts of data (Alvarado-Valencia et al., n.d., Blattberg and Hoch, 1990, Goodwin, 2002, Goodwin et al., 2011). The question remains how to optimally integrate expert judgment with forecasting models. Recently, Alvarado-Valencia et al. (n.d.) compared three integration methods: judgmental adjustment, 50/50 combination and divide-and-conquer (restricting information access). Judgmental adjustment performed best, possibly because the other techniques restricted access to information (Alvarado-Valencia et al., n.d.). They conclude that bias reduction through information restriction does not work. Other research with decision support systems has illustrated that pointing the forecasters towards the damaging effect (in terms of forecasting accuracy) of their adjustments is not the solution. The tendency to adjust persists despite warning (Lim and O'Connor, 1995). This tendency to discount advice is especially troublesome, given that it implies that conscious efforts toward debiasing judgmental adjustments may be in vain.

Goodwin et al. (2011) focussed on two possibilities to counter harmful judgmental adjustments: restrictiveness (limiting options) and guidance (providing information and explanation). Similar to previous studies in advice taking (e.g., Lim and O'Connor, 1995), the provision of guidance had no significant effect. Restricting user's options to make adjustments to the forecast even led to a significant reduction in forecasting accuracy. An additional problem poses itself in the case of restricting judgmental adjustment: the acceptability of the system in the eyes of the forecaster (Goodwin et al., 2011). Reducing harmful adjustments can come at the price of reduced acceptance (Goodwin et al., 2011), feelings of loss of control and ownership (Goodwin, 2002) and decreased trust of the final forecast (Önkal, Goodwin, Thomson, Gönul and

Pollock, 2009).

While some researchers work on debiasing the judgmental component, others focus on improving predictive performance by applying more advanced statistical techniques (e.g., Kourentzes and Petropoulos, 2016). However, surveys have indicated that judgment persists in business forecasting (Fildes and Goodwin, 2007) and simultaneously, that forecasting accuracy in business practice is not improving (Armstrong et al., 2013). We therefore propose to re-visit the possibility of incorporating human judgment factors in the model itself (Franses and Legerstee, 2013), in order to improve the accuracy of predictions. This way of working has the benefit of potentially mitigating the harmful effects of judgment in two ways: first, by discounting those judgments that are biased or with great error, based on the historic performance. Second, it takes the widespread practice of incorporating judgment in forecasting into account. Forecasters will still be able to submit their judgment and show that they are attending to the task (Fildes et al., 2009). Importantly, forecasters will not be put off by the system (Silver, 1991) and retain their sense of ownership. The model will incorporate human judgment as a predictor if these adjustments have proved to add value in the past. In sum, while previous attempts at optimizing the judgment-statistics combination via a forecast support system have demonstrated potential damaging effects on forecaster performance and forecasting accuracy, the integrative model collects input from both sources, and should therefore mitigate potential harmful effects of human judgment, while recognizing the role of acceptance of the forecaster/user. In the dataset of Franses and Legerstee (2013), the judgmental component proved only beneficial if the statistical model was not performing well. Overall, the original statistical model does not seem to improve by including judgment. However, ignoring the value of expert judgment can be especially dangerous in time series with disturbances, such as promotions and other exceptional events. We therefore propose an integrative approach that can be of significant value for industries dealing with exceptional events, such as

promotions.

3. Integrative judgment forecasting model

In order to gain a more in-depth understanding of the effects of integrative judgment versus restrictive judgment, we investigate adjustment sizes, direction of the adjustment, the periodicity of the product and the volatility of the data series. Previous research has shown that downward adjustments are more likely to be beneficial than upward adjustments (Fildes et al., 2009, Franses and Legerstee, 2009, Syntetos et al., 2009). A possible explanation here is that downward adjustments are only made in the presence of evidence that a downturn may arise, while upward adjustments are mostly a reflection of over-optimism and wishful thinking of the forecaster (Fildes et al., 2009). An integrative approach should counter the negative effects of positive adjustments as follows: In the integrative model, the historical human judgment forecast is added as an additional predictive variable. The model determines whether or not to take this variable into account based on past performance. Only if adjustments have been historically sufficiently accurate and have added value, the model will take them into account. Goodwin et al. (2011) state that a forecast support system which integrates judgment and statistics, should support two stages of this task: first, the decision whether or not the adjustment adds value and second, the determination of how large the adjustment should be. The integrative approach takes this into account, by looking at the significance and weight of the human judgment predictor.

In addition to the direction of the adjustment, previous research suggests a relationship between the size of the adjustment and forecasting accuracy, such that mostly big adjustments are beneficial and small adjustments should be avoided as to not harm forecasting accuracy (Diamantopolous and Mathews, 1989, Fildes et al., 2009). This can be explained by the 'tinkering with data' effect, where forecasters make small adjustments to show that they are working on the task and feel responsible and in control

of the forecasting process (Fildes et al., 2009). Large adjustments on the other hand, are an indicator of knowledge available to the forecaster that is not yet incorporated in the system. However, Fildes et al. (2009) find a tipping point in the case of negative adjustments, such that very large adjustments are equally damaging for forecasting accuracy. Indeed, the effect of adjustment size on forecasting accuracy is curvilineair (inverted U-shape), such that both small and very large adjustments are damaging for forecasting accuracy. Similarly, Trapero et al. (2013) found adjusted forecasts of promotional periods to be potentially beneficial, but not when they were overly large. In other words, the integrative approach should mitigate the damaging effects of both too small and too large adjustments. A third influencing factor is volatility. Forecasting models typically have problems dealing with exceptional events (Goodwin and Fildes, 1999). These exceptional events can occur in several ways, for example the occurrence of the Olympic games when predicting aviation traffic in a country or predicting the sales of a special issue of a magazine. Since little past information about these events is available, it is difficult for computerized models to take this effect into account. Especially in these situations, the incorporation of human judgement can add significant value and could enhance the accuracy of the models (Goodwin and Fildes, 1999). In other words, expert judgment can prove beneficial in volatile data series. However, a distinction must be made between volatile data series because of special events or because of noise. In the former case, human judgment is expected to outperform models (Sanders and Ritzman, 1992) while in the latter case, models outperform judgment (O'Connor et al., 1993). In this study, volatility is determined by the presence of promotions. In concordance with previous literature, we therefore expect that human judgment will have added value over the statistical model if promotions are present (high volatility). In contrast, in low volatility series (periods without promotions), judgment is expected to damage forecasting accuracy (Sanders and Ritzman, 1992). The damaging effect of judgment in low volatility series should be mitigated by the integrative approach, compared to the

restrictive approach.

A fourth factor that may play a role in forecast accuracy, is the periodicity of the product. In this study, a distinction is made between high (weekly products) and low (monthly products) periodicity. Products with weekly forecasts should increase familiarity with the product. Edmundson et al. (1988) tested three groups of participants with varying degrees of non-time series related knowledge: no contextual knowledge, industry knowledge, and industry plus product knowledge. They found that product familiarity was the most significant in improving forecast accuracy. Thus, we expect forecasts of weekly products to increase product familiarity, and therefore, to be more accurate than monthly products.

In addition to the above mentioned qualities of the data, this study decomposes forecasting models on different levels of granularity. Kremer et al. (2015) indicated the lack of cross-over between judgmental forecasting and hierarchical forecasting research. Yet, in many business situations, the forecasted SKUs are distributed over multiple point of sales (PoS). Hence, an organization has the choice between generating forecasts on SKU level and rolling this down to SKU per PoS level by using business rules, or making the forecasts on PoS level directly per SKU. For statistical forecasts, both options are easily implemented. However, in the latter situation, integrating human judgement is challenging since the number of SKU and PoS combinations is typically very high. This makes it impossible to apply expert adjustments on the lowest level of granularity. As an alternative, in a traditional restrictive approach, the forecasts on SKU per PoS level could be rolled up again on SKU level, which allows adjustments based on human judgement expertise. Next, these corrections would be applied over all PoS equally. However, this equal correction over all PoS is not always optimal. For example, if the expert systematically takes weather into account to make adjustments to the statistical forecast, an equal division would not incorporate geographical location effects per PoS: the influence of weather might be greater in PoS in touristic

and commercial areas than in other rural areas. Thus, a trade-off exists between level of detail (granularity) and potential error (Zotteri and Kalchschmidt, 2007, Hyndman et al., 2011). In this situation, the use of an integrative judgement model adds value, in that not all PoS will be treated equally. Since expert adjustments are taken into account as an explanatory variable in the forecasting models on SKU per PoS level, it is able to distinguish the PoS that are typically more influenced by the expert related factors than others.

While accuracy is an important indicator of forecasting performance, this indicator remains in the theoretical realm. Several researchers highlight the need to link forecast accuracy to measures linked to other business performance (Mahmoud et al., 1992, Mentzer et al., 1999, Moon et al., 2003, Kerkkänen et al., 2009). Previous studies have indicated links between improved demand forecasting accuracy and inventory management (e.g., Clarke, 2006, Oliva and Watson, 2009, Syntetos et al., 2010) and delivery times (Shan et al., 2009). This study focusses specifically on the practical consequences of demand forecasting accuracy on financial results. By doing this, we are able to express the effect of increased forecasting accuracy on the profitability of the organisation. The forecasts are used as an input for an optimization algorithm that determines the ideal number of SKUs per point of sales, taking revenue and costs into account. Hence, the focus of this optimization algorithm is to maximize profitability and not accuracy. Although the optimization algorithm on itself is out of the scope of this research, this enables us to express the results of this study in a more business relevant profit metric. To the best of our knowledge, this is the first study that takes this effect into account.

In sum, this study looks at the potential beneficial effects of integrating expert judgment in the statistical forecasting model. In contrast to Franses and Legerstee (2013), our time series are disturbed by exceptional events (promotions), which should increase the potential for an added value of judgmental intervention. We analyse our data-set

in-depth according to the paper of Fildes et al. (2009) to provide a basis for comparison of the dataset's characteristics. Additionally, we follow the recommendation of Kremer et al. (2015) to distinguish between the different hierarchical levels of forecasting. Lastly, we believe this to be the first study that links changes in forecast accuracy to a measure of direct profitability.

4. Materials and Methods

4.1. Data

Data from a European publishing company were collected. The company distributes weekly and monthly magazines to 5902 points of sales. The data included statistical system demand forecasts, judgemental forecasting adjustments, recommendations for optimal distribution for maximizing profit and corresponding actual outcomes. The products are divided into two categories: regular products and exceptional products. An exceptional product is defined as a regular product with an extra: a magazine that includes another magazine, a collectible, dvd, cd, etcetera. In total 3 575 263 data points were collected over a time period of 16 months. These data points are containing forecasting data on SKU and PoS level. This data can be rolled up to 1312 aggregated forecasts on SKU levels. Data cleaning led to the deletion of cases with missing information, magazines (SKU) with a sale less than 100 (in comparison, mean sales per SKU = 30 604) (Fildes et al., 2009), and two cases in which the system forecast deviated extremely from the final sales (APE larger than 2000%), resulting in a final of 1223 aggregated forecasts. Out of those 1223, 850 were classified as regular products (69.50%) and 373 as exceptional products (30.50%) by the company. The forecasting process was discussed with the company and follows a fixed procedure discussed below.

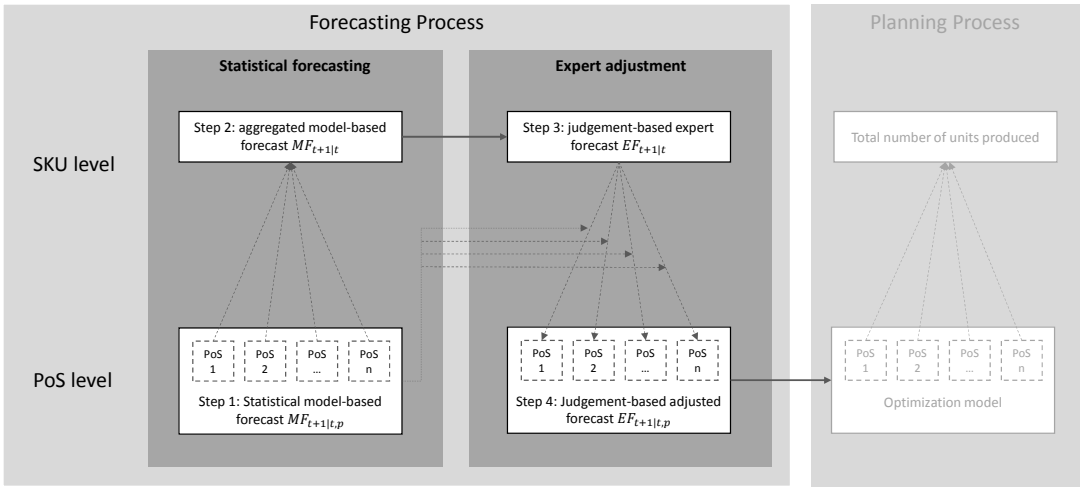


Figure 1: Restrictive judgment process

4.2. Methodology

Figure 1 shows the original forecasting process of the company. This process takes place on two levels of granularity: SKU per Point of Sales (PoS) and SKU level (the aggregated demand forecast per magazine).

In a first step, the demand per PoS is estimated using a statistical forecasting support system. This forecast is denoted as $MF_{t+1|t,p}$, which represents the statistical forecast for Point of Sales p for time t+1 at origin t. This forecast is generated by several time series based forecasting techniques such as moving average, exponential smoothing

or ARIMA models. Each of these techniques is based on lagged sales data. We will illustrate this using an autoregressive model of order two, similar to the one used in Franses and Legerstee (Franses and Legerstee, 2013) :

$$D_{t,p} = \alpha_{1,p} D_{t-1,p} + \alpha_{2,p} D_{t-2,p} + u_{t,p} \tag{1}$$

Where $D_{t,p}$ represents the Demand at time t at Point of Sales p and u_t is an unobserved error term. Based on this forecasting support system the statistical model-based forecast for each PoS can be obtained by:

$$MF_{t+1|t,p} = \alpha_{1,p} D_{t,p} + \alpha_{2,p} D_{t-1,p} \tag{2}$$

Since the number of PoS is typically very large (i.e., 5902) it is impossible to evaluate each forecast using expert judgment. Therefore, in a second step, this data is rolled up to a higher level using the following equation:

$$MF_{t+1|t} = \sum_{d=1}^{n} MF_{t+1|t,p} \tag{3}$$

Where n represents the total number of PoS. Thus, bottom-up forecasting is used, deemed appropriate as the goal is to forecast as accurately as possible on the lowest level (Hyndman et al., 2011). In a third step, this aggregated forecast is send to an expert, who can adjust the forecast. This judgement-based expert forecast can be denoted as $EF_{t+1|t}$, which represents the expert forecast on an aggregated SKU level for time t+1 at origin t. This expert forecast can be formulated by the following equation (Franses and Legerstee, 2013) :

$$EF_{t+1|t} = \lambda MF_{t+1|t} + \beta_t X_{t+1|t} \tag{4}$$

This assumes that the judgement-based expert forecast can be decomposed out of

a weighted average of the statistically-based model forecast and own expert knowledge. This weight is represented by λ , which can range between 0, when the expert ignores the model forecast, and 1, when the expert completely accepts the model forecast. It is important to mention that in practice, a company often does not know how much the expert relies on the model forecast. In addition, equally frequently unknown are the factors that are taken into consideration by the expert to adjust the forecast, represented by $X_{\text{t-1|t}}$ with weight β in equation (4) (Fildes and Goodwin, 2007). Forecasters may have knowledge that is not included in the model. For instance, technical knowledge (knowledge about the data analysis and forecasting procedures), causal knowledge (an understanding of the cause-and-effect relationships involved), or product knowledge (Sanders and Ritzman, 1992, Webby and O'Connor, 1996)). Experts may have knowledge of recent events that have not yet been included in the time series; unusual events that have occurred in the past, but not expected to occur again in the future; or unusual events that have not yet occurred, but expected to occur in the future (Armstrong and Collopy, 1998, Sanders and Ritzman, 2004). In a fourth step, these adjusted forecasts are drilled down again over all PoS equally as followed:

$$EF_{t+1|t,p} = (1 + \delta_{\text{t}})MF_{t+1|t,p} \text{ defining } \delta_{\text{t}} = \frac{EF_{t+1|t} - MF_{t+1|t}}{MF_{t+1|t}} \tag{5}$$

This forecast on PoS level can be passed on to planning and used as input for an optimization model that maximises the expected profit. This process is further referred to as restrictive judgment, since the judgmental adjustment of the forecaster serves as a restriction on the model output (see Figure 1).

In the integrative model (Figure 2) the beginning of the process is the same as currently used in the company and has been described above: forecasting support systems based on past demand provide a statistically-based forecast in each PoS. This is aggregated to SKU level and presented to the expert for evaluation and adjustment.

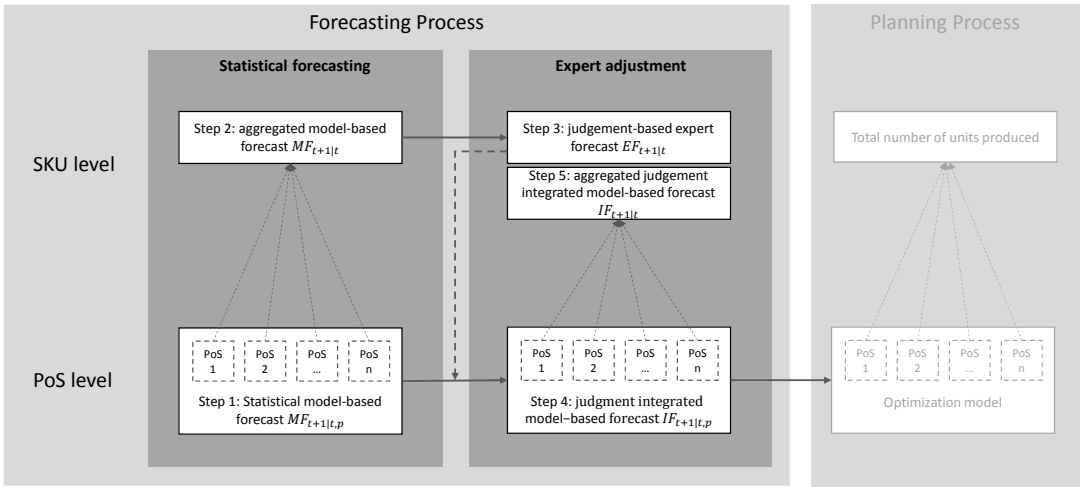


Figure 2: Integrative judgment process

However, in contrast with the previous method, the adjusted forecast is not directly drilled down to PoS level afterwards. Rather, the judgment-based expert adjustments now serve as a predictor variable for the statistically-based forecasting model on PoS level, which can be formulated as (Franses and Legerstee, 2013):

$$D_{t,p} = \alpha_{1,p} D_{t-1,p} + \alpha_{2,p} D_{t-2,p} + \tau_p \beta_t X_{t|t-1} + u_{t,p} \tag{6}$$

Based on equation (4) this can be rewritten as

$$D_{t,p} = \alpha_{1,p} D_{t-1,p} + \alpha_{2,p} D_{t-2,p} + \tau_p (EF_{t|t-1} - \lambda MF_{t|t-1}) + u_{t,p} \quad (7)$$

Defining $\tau_{\mathrm{p}} = \beta_{1,\mathrm{p}} + \beta_{2,\mathrm{p}}$ and $\tau_{\mathrm{p}}\lambda = \beta_{2,\mathrm{p}}$, equation (7) can be rewritten as

$$D_{t,p} = \alpha_{1,p} D_{t-1,p} + \alpha_{2,p} D_{t-2,p} + \beta_{1,p} EF_{t|t-1} + \beta_{2,p} (EF_{t|t-1} - \lambda MF_{t|t-1}) + u_{t,p} \quad (8)$$

$$IF_{t+1|t,p} = \alpha_{1,p} D_{t,p} + \alpha_{2,p} D_{t-1,p} + \beta_{1,p} EF_{t+1|t} + \beta_{2,p} (EF_{t+1|t} - \lambda MF_{t+1|t}) \quad (9)$$

This process is termed integrative judgment (see Figure 2), since the expert adjustment is integrated in the forecast equation. The parameters estimated in the equation are refitted each time using the information of the most recent week or month. Thus, each SKU is forecasted independently, for each time t, taking the previous adjustment performance for the product into account. Equation (8) shows that this statistically-based forecasting model includes the judgement-based expert forecast on SKU level. The more systematically accurate the expert forecast $EF_{\mathrm{t|t\text{-}1}}$ was in the past, the higher the relevance of this variable will be in the forecasting support system. The model allows parameters $\beta_{1,\mathrm{p}}$ and $\beta_{2,\mathrm{p}}$ to vary depending on the PoS. By this the forecasting support system can distinguish between PoS that are systematically more or less influenced by the unobserved factors taken into account by the expert. In other words, the adjustments can be ignored for some PoS, but taken into account for others. The impact of the adjustment can even be increased due to the weights assigned to them. A possible reason for applying different weights to different PoS would be a location effect: a warm month may increase tourism at the seaside and hence, sales in that lo-

cation. An increase of weights could be necessary to counter for an under-forecasting bias: the expert adjustments are in the right direction but always insufficient. Note that in a situation in which no aggregated model forecast is proved (i.e. step 1 until 3 is skipped in the forecasting process), as a result of a lack of time or standardized forecasting process, or when the expert totally ignores this input, a judgment integrated model-based forecast can still be generated. In this specific situation the weight $\lambda = 0$, which reduces equation (9) to:

$$IF_{t+1|t,p} = \alpha_{1,p} D_{t,p} + \alpha_{2,p} D_{t-1,p} + \tau_{1,p} EF_{t+1|t} \tag{10}$$

Finally, in a fifth step, these final integrative forecasts on PoS level could still be rolled up to SKU level, defined by $IF_{t+1|t}$, to determine the aggregated demand per SKU:

$$IF_{t+1|t} = \sum_{p=1}^{n} IF_{t+1|t,p} \tag{11}$$

Although the forecast on PoS level expressed in equation (10) will be used for planning, this aggregated forecast is still interesting for comparing the integrative approach with the traditional restrictive approach defined in equation (4).

Since the goal of the forecasting models is to maximize prediction accuracy, this comparison can be expressed in metrics such as Mean Absolute Percentage Error (MAPE) and Median Absolute Percentage Error (MdAPE). However, the forecasting evaluation in business practice goes beyond accuracy, as it serves as a starting point for planning (Mahmoud et al., 1992, Mentzer et al., 1999, Moon et al., 2003, Kremer et al., 2015). In order to plan the optimal number of units in each PoS, an optimization model is used with the forecasted demand on PoS level as input (see right-hand side on Figures 1 and 2). This algorithm has the goal to maximize profitability taking potential costs into account. More specifically, profitability can be expressed by the following

equation:

$$P_{t,p} = R * D_{t,p} - I * Q_{t,p} - O * (Q_{t,p} - D_{t,p}) \tag{12}$$

With $P_{t,p}$ being the operational profit at time t in PoS p (excluding all overhead and fixed costs), R being the revenue earned from selling one magazine, $D_{t,p}$ being the demand of magazines at time t in PoS p, I the cost to put a magazine in the market (printing + distribution cost), $Q_{t,p}$ the number of magazines distributed to PoS p at time t and O being the cost to get an unsold magazine out of the market.

The outcome of the optimization model is the optimal quantity $Q_{t,p}$ to be delivered at time t per PoS p, that maximises the expected profit taking the expected demand of the forecasting model, the inventory cost and opportunity cost into account. This can be aggregated to know the total number of units to be produced. Note that in all cases I $>$ O and R is at least 2(I+O), hence the optimal quantity that results from the optimisation process is at least the predicted demand or higher. The focus of this study is on the forecasting process by comparing a restrictive judgment and integrative judgment approach; however, the effect on the planning process will be discussed as well, by expressing the results in a profitability metric.

5. Results

5.1. Descriptive statistics

Judgemental Adjustment, defined as JA, can be measured as followed:

For a restrictive approach (judgment as a restriction on the model forecast) :

$$JA = 100 * \frac{EF_{t+1|t} - MF_{t+1|t}}{MF_{t+1|t}} \tag{13}$$

For an integrative approach (judgment incorporated as a predictive variable in the

model forecast) :

$$JA = 100 * \frac{IF_{t+1|t} - MF_{t+1|t}}{MF_{t+1|t}} \tag{14}$$

Table 2 indicates the relative adjustment sizes for the restrictive and integrative judgment model respectively, with the number of adjustments distributed evenly over four quantiles. The adjustment sizes of the integrative model are noticeably smaller than those of the restrictive model, which shows that the integrative approach is more careful in adapting based on judgemental expertise than the restrictive approach.

Table 2: Quartiles, Mean and Median of percentage adjustments

	Restrictive Judgment	Integrative Judgment
25%quantile	4.93	0.00
Mean	24.51	4.74
Median	11.94	.67
75%quantile	31.01	3.97

N = 1223

The data was checked against outliers above 250%, similar to Fildes et al. (2009). No cases needed to be deleted: the largest adjustment made in both the restrictive and the integrative model was 149,55%. Table 3 shows the mean and median sizes of the relative adjustments according to the direction of the adjustment (positive indicating an adjusted forecast that is larger than the system forecast).

Table 3: Mean and median of adjustments, ordered by direction of adjustment

	Restrictive Judgment			Integrative Judgment		
	N	Mean	Median	N	Mean	Median
Downward	587 (48%)	-17.70%	-12.31%	464 (37.94%)	-5.71%	-1.93%
No adjustment	28 (2.29%)	-	-	375 (30.66%)	-	-
Upward	608 (49.71%)	31.57%	12.12%	384 (31.40%)	8.21%	2.48%

As the table indicates, for the restrictive judgment model, adjustments were made in 1195 or 97.71% of the cases, of which 587 or 48,00% were downward (Mean adjustment size = -17.70%, Median adjustment size = -12.31%) and 49.71% upward (Mean adjustment size = 31.57%, Median adjustment size = 12.12%). Comparatively, in an

integrative judgment model, the prediction of the basic model was deemed already optimal in 375 or 30.66% of the cases, resulting in 464 or 37.94% downward adjustments (Mean adjustment size = -5.71%, Median adjustment size = -1.93%) and 384 or 31.40% upward adjustments (Mean adjustment size = 8.21%, Median adjustment size = 2.48%) .

Figure 3 and Figure 4 show the distribution of the adjustment sizes (percentages expressed as decimals) for the restrictive model and the integrative model respectively. The histograms show a clear difference between the integrative and restrictive approach. The adjustments within the restrictive approach are clearly more frequent and larger than in the integrative approach. In addition, the restrictive approach is somewhat skewed towards positive adjustment, which cannot be observed in the histogram of the integrative approach.

Table 4 details the number of times the adjustment was discarded (not significant), dampened, enforced or flipped (a downward adjustment translated into an upward one or vice versa, in decreased or increased size). Expert judgment was discarded more often in the case of upward adjustments. This is similar to Fildes et al. (2009), who found upward adjustments to be damaging. This damaging effect is recognized and compensated for by the integrative judgment approach. When adjustments were accepted, it appears that they were more often dampened than enforced. Consistent overly large adjustments are made by the forecaster and dampened by the integrative judgment model. In some cases the integrative model flipped the expert-based adjustment in direction of adjustment. This could, for example, be the case if good weather would increase the seaside-sales, but decrease sales inland due to the move of potential buyers.

Table 4: Discarding and acceptance of expert judgments by the integrative judgment model

	Discarded	Dampened	Enforced	Flipped
Downward adjustments	69.69%	14.53%	8.82%	6.95%
Upward adjustments	75.02%	9.87%	8.67%	6.44%

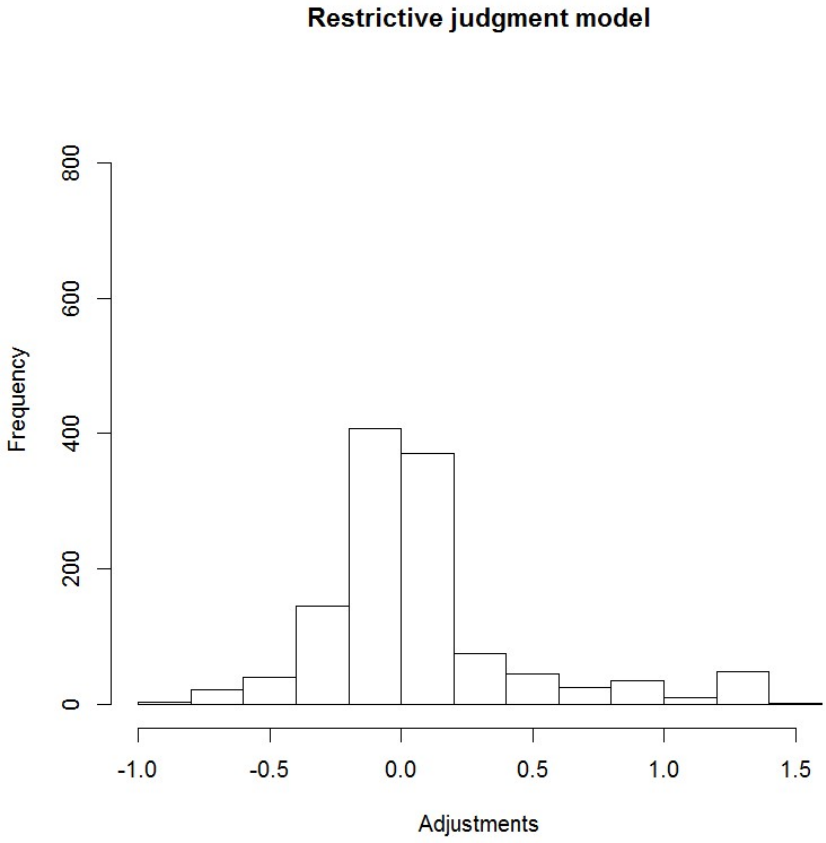


Figure 3: Histogram restrictive judgment

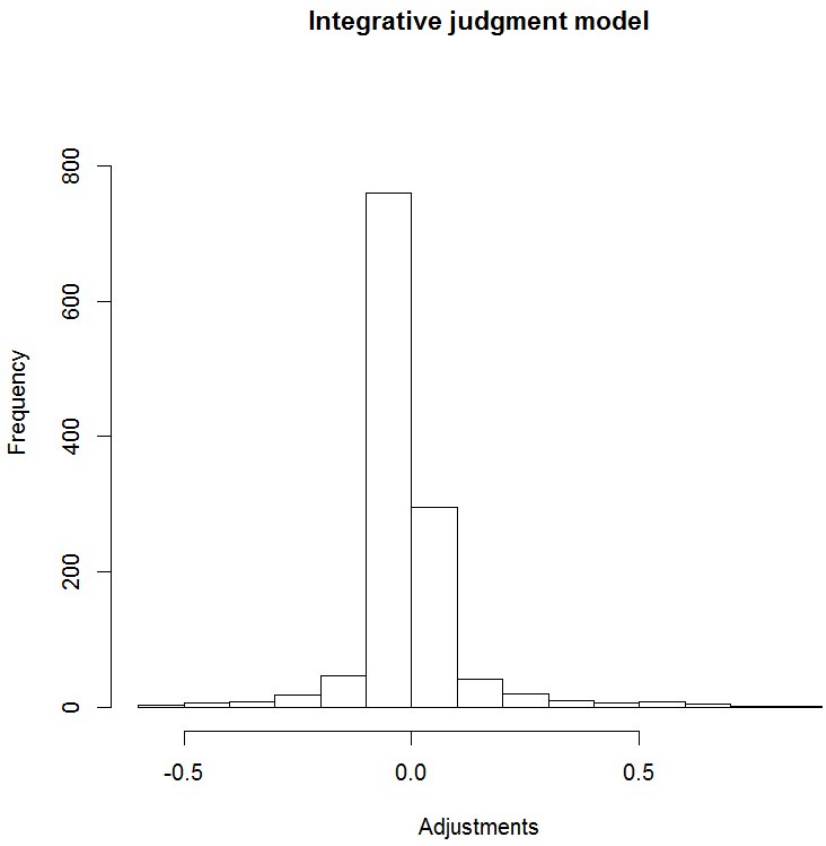


Figure 4: Histogram integrative judgment

5.2. Performance

5.2.1. Error metrics

Table 5 indicates the relative performance of the models. Performance is measured as the forecasted demand, compared to the actual sales. To ensure comparibility with previous studies (e.g., Fildes et al., 2009), we employ MAPE and MdAPE as accuracy measures. The results of MAPE and MdAPE differ such that the median scores are lower than the mean scores, indicating that the data is skewed to the right (similar to Fildes et al., 2009). We therefore report both as measures of forecasting accuracy. To compare the performance of the restrictive model and the integrative model with the basic statistical model, we use FCIMP (forecast improvement) according to the formula:

For a restrictive approach, FCIMP will be defined as:

$$FCIMP = 100 * \frac{|S_{t+1} - MF_{t+1|t}| - |S_{t+1} - EF_{t+1|t}|}{S_{t+1}} \tag{15}$$

For an integrative approach, FCIMP will be defined as:

$$FCIMP = 100 * \frac{|S_{t+1} - MF_{t+1|t}| - |S_{t+1} - IF_{t+1|t}|}{S_{t+1}} \tag{16}$$

This indicates the difference between the APE from the system forecast and the APE of the adjusted forecast (Fildes et al., 2009), based on a restrictive or on an integrative approach. Regarding analysis, for those comparisons where all the observations were correlated (general comparison, volatility (category) and periodicity), we used a paired t-test bootstrapped based on 1000 samples for comparing the MAPE, FCIMP and profit. In order to compare the MdAPE, we compared the medians with a Wilcoxon signed rank test. The comparison between an integrative and restrictive approach for adjustment direction, size and standard deviation contains partially correlated observations (i.e. not all forecasts belong to the same group if restrictive and integrative

approaches are compared). For these comparisions we used a bootstrapped weighed t-test (with 1000 samples) as proposed by Samawi and Vogel (2013)). In a similar way, a Wilcoxon Rank Sum Test is combined with a Wilcoxin Signed Rank Test to compare the MdAPE's of partially correlated samples.

Table 5: Model performance expressed as MAPE, MdAPE, and FCIMP (%)

	Basic Statistical Model	Restrictive Judgment	Integrative Judgment
MAPE	25.03%	26.18%$^{\text{BSM: n.s.}}$	21.88%$^{\text{BSM: *** / RM: **}}$
MdAPE	9.95%	10.83%$^{\text{BSM: n.s.}}$	8.24%$^{\text{BSM: *** / RM: ***}}$
FCIMP	-	-1.15%	3.15%$^{\text{RM: **}}$

Note. Proposed models are compared with the basic statistical model (BSM) and each other (RM).

Note. Significance levels: * $p < .05$, ** $p < .01$, *** $p < .001$, $^{\text{n.s.}}$ not significant

Integrative judgment outperforms restrictive judgment and the basic statistical model in terms of the mean and median absolute percentage error (Table 5). While restrictive judgment harms forecasting accuracy compared to the basic model (-1.15 percentage points (*p.p.*), *n.s.*), the integrative judgment proves beneficial (+2.49 *p.p.*, $p < .01$). In other words, the integrative model (MAPE = 22.54%) has the lowest error and outperforms both the basic statistical model (MAPE = 25.03%, $p < .001$) and the restrictive model (MAPE = 26.18%, $p < .01$).

The integrative judgment approach was further compared with a shrinkage method. The average size of the adjustments in the integrative model is 4.74. When compared to the average size of the adjustments in the restrictive model (mean = 24.51), it can be derived that on average the adjustments of the integrative approach are only 19.34% the size of the restrictive approach. The question is whether the integrative judgment approach has any added value over a simple shrinkage method that reduces all adjustments in the restrictive approach by 19.34%. This shrunk restrictive model (MAPE = 23.80%, MdAPE = 11.34%, FCIMP = 1.23%) performs significantly better ($p_{\text{MAPE}} = .056$, $p_{\text{MdAPE}} < .001$, $p_{\text{FCIMP}} < .056$) than the original restrictive model. However,

the integrative judgment model still significantly outperforms the shrinkage method ($p_{MAPE} < .001$, $p_{MdAPE} < .001$, $p_{FCIMP} < .001$). This indicates that the improvement found in the integrative judgment model is not solely the result of more conservative behaviour towards judgmental adjustment. Rather, the model is selective in when to incorporate adjustments and in what manner.

Digging deeper into the data, we compare both models with regard to the direction of the adjustments, their size, the volatility of the data, and the periodicity of the product.

Table 6: MAPE, MdAPE , and FCIMP, ordered by direction of adjustment

	Restrictive Judgment				Integrative Judgment			
	N	MAPE	MdAPE	FCIMP	N	MAPE	MdAPE	FCIMP
Downward	587 (48%)	16.79%	9.05%	18.24%	464 (37.94%)	24.67%***	8.98%$^{n.s.}$	5.01%***
No adjustment	28 (2.29%)	53.84%	32.76%	-	375 (30.66%)	19.97%	6.51%	-
Upward	608 (49.71%)	33.97%	12.63%	-19.91%	384 (31.40%)	20.38%*	9.02%***	1.97%***

Note. Integrative judgment model is compared with the restrictive judgment model.

Note. Significance levels: * $p < .05$, ** $p < .01$, *** $p < .001$, $^{n.s.}$ not significant

With regard to the direction of the adjustments (Table 6), our dataset confirms previous literature, which indicates upward adjustments as damaging to forecasting accuracy (FCIMP = -19.91%, compared to the basic model), while downward adjustments are beneficial (FCIMP = +18.24%, compared to the basic model) in a traditional restrictive approach. However, when judgment is included in an integrative way, both upward (FCIMP = +1.97%, $p < .001$) and downward (FCIMP = +5.01%, $p < .001$) adjustments are beneficial. The integrative approach thus mitigates the severely damaging effects of restrictive judgmental adjustment. However, while the damaging effect of upward adjustments is neutralized and even translated into a small beneficial effect, the beneficial effects of downward adjustments are tempered (from 18.24% improved accuracy with restrictive judgment, to 5.01% improved accuracy in the case of integra-

tive judgment). These results show that, on average, the integrative approach is able to ignore the adjustments that are consistently over-optimistic and mainly incorporates only useful positive adjustments.

With regard to adjustment size for the restrictive model (Table 7), we find a curvilinear relationship similar to Fildes et al. (2009) between adjustment size (expressed in four quantiles) and forecasting accuracy, such that small adjustments (Q1= -0.3% and Q2= -1.47%) and overly large adjustments (Q4= -8.89%) are damaging. Only medium adjustments proved to be beneficial (Q3 = 5.99%). However, when judgment is included in an integrative way, the damaging effects of Q1, Q2 and Q4 are translated into a beneficial effect, showing a concave relationship.

Table 7: Adjustments ordered by size in four quantiles: MAPE, MdAPE, and FCIMP

	Restrictive Judgment				Integrative Judgment			
	N	MAPE	MdAPE	FCIMP	N	MAPE	MdAPE	FCIMP
No adjustment	28	53.85%	32.76%	-	375	19.97%	6.51%	-
Size Q1	298	12.49%	5.86%	-0.03%	212	28.44%*	8.67%***	0.06%^{*1}
Size Q2	298	13.66%	8.49%	-1.47%	212	18.11%***	7.62%$^{n.s.}$	0.44%$^{n.s.}$
Size Q3	298	18.55%	12.51%	5.99%	212	21.25%$^{n.s.}$	8.63%*	1.57%***
Size Q4	298	57.37%	39.29%	-8.89%	212	23.12%**	12.82%***	16.11%***

Note. Integrative judgment model is compared with the restrictive judgment model.

Note. [1] the value was marginally significant at .051.

Note. Significance levels: * $p < .05$, ** $p < .01$, *** $p < .001$, $^{n.s.}$ not significant

While adjustments are always beneficial in the case of integrative judgment, the large beneficial effect of Q3 with restrictive judgment is limited (a decline in beneficial effect from 5.99% to 1.57%, $p < .001$). However, the severely damaging effect of overly large adjustments (-8.89%) is translated into a highly beneficial effect (16.11%, $p < .001$).

To calculate volatility, we employ two different procedures. First, we calculate volatility as the coefficient of variation of the system forecast absolute error (Fildes et al., 2009). The resulting volatility scores are divided into four quantiles, ranging

from low to high (Table 8).

Table 8: Performance ordered by volatility according to SD: MAPE, MdAPE, and FCIMP

	Restrictive Judgment				Integrative Judgment			
	N	MAPE	MdAPE	FCIMP	N	MAPE	MdAPE	FCIMP
SD Q1	305	54.25%	31.91%	-28.84%	305	24.81%***	12.95%***	0.60%***
SD Q2	306	26.20%	14.68%	6.40%	306	28.95%$^{n.s.}$	15.46%$^{n.s.}$	3.66%$^{n.s.}$
SD Q3	306	13.16%	7.53%	10.87%	306	18.27%$^{n.s.}$	6.42%$^{n.s.}$	5.76%$^{n.s.}$
SD Q4	306	11.19%	7.31%	6.89%	306	15.51%$^{n.s.}$	4.90%***	2.57%$^{n.s.}$

Note. Integrative judgment model is compared with the restrictive judgment model.

Note. Significance levels: * p $<$ .05, ** p $<$ .01, *** p $<$ .001, $^{n.s.}$ not significant

In the case of restrictive judgment, judgmental adjustments are especially troublesome in the lowest quantile, with a reduction of -28.84% in forecasting accuracy. The other quantiles with medium to high volatility on the other hand, show a forecasting improvement 6.40% (Q2), 10.87% (Q3) and 6.89% (Q4). In the integrative model, there is a slight forecasting improvement in all quantiles and ranges of volatility, displaying a curvilinear effect with Q1 = .6% ($p < .001$), Q2 = 3.66% (*n.s.*), Q3 = 5.76% (*n.s.*) and Q4 = 2.57% (*n.s.*). However, the improvement of the integrative approach is significantly better than the restrictive approach for the first quantile only.

Second, we follow the company's categorization in regular products (considered easier to predict) and exceptional products (considered more difficult to predict) as an analogy for low and high volatility (Table 9). The correlation with the previous metric equals r = 0.3468, suggesting a relationship but not a full overlap. For instance, within the low volatility category, there is still an effect of volatility as variation of the system forecast absolute error.

Table 9: Performance ordered by volatility according to category: MAPE, MdAPE,and FCIMP

	Restrictive Judgment				Integrative Judgment			
	N	MAPE	MdAPE	FCIMP	N	MAPE	MdAPE	FCIMP
Regular Products	850	26.26%	9.25%	-12.36%	850	10.99%***	5.33%***	2.90%***
Exceptional Products	373	26.00%	14.58%	24.42%	373	46.70%***	21.56%***	3.72%***

Note. Integrative judgment model is compared with the restrictive judgment model.

Note. Significance levels: * p $<$ .05, ** p $<$ .01, *** p $<$ .001, $^{n.s.}$ not significant

The results of the restrictive model confirm that judgmental adjustment in case of low volatility (regular products) is damaging (FCIMP = -12.36%) and beneficial in the case of high volatility (FCIMP = 24.42%). In the integrative model, the damaging effect of judgment in case of regular products is eliminated and translated into a positive improvement in forecasting accuracy (FCIMP = 2.90%, $p < .001$). In case of high volatility, the forecast improvement is 3.72% ($p < .001$). Logically, for non-volatile products, the statistical method is quite adept at predicting demand. Judgmental adjustment will on average be more harmful (e.g., due to tinkering with the data) than beneficial (e.g., due to having additional information). In situations with high volatility, expert adjustment is beneficial. However, the integrative approach limits the effect.

The effects of product periodicity for restrictive judgment are as expected: weekly products are forecast more accurately than monthly products (Table 10). Restrictive judgment performs worse than the basic statistical model for monthly products (FCIMP: -14.27%). This harmful effect is eliminated by the integrative judgment model and translated into a slight improvement (FCIMP: 1.13%, ($p < .001$). However, where weekly products are concerned, the integrative judgment model shows no significant improvement on the restrictive judgment model. The forecaster is quite adept at forecasting those products that are more familiar; the integrative model cannot provide any added value.

Table 10: Performance ordered by periodicity according to category: MAPE, MdAPE,and FCIMP

	Restrictive Judgment				Integrative Judgment			
	N	MAPE	MdAPE	FCIMP	N	MAPE	MdAPE	FCIMP
High (weekly)	977	22.45%	8.84%	2.16%	977	20.96%$^{\text{n.s.}}$	6.81%***	3.66%$^{\text{n.s.}}$
Low (monthly)	246	40.95%	28.87%	-14.27%	246	25.55%***	17.60%***	1.13%***

Note. Integrative judgment model is compared with the restrictive judgment model.

Note. Significance levels: * $p < .05$, ** $p < .01$, *** $p < .001$, $^{\text{n.s.}}$ not significant

Next to comparing the difference in forecasting improvement between a restrictive and integrative model, the effect of the characteristics of the forecast and judgemental adjustments are studied in detail. Previous results indicate that for both models, the forecasting improvement resulting from expert adjustments is relatively higher when dealing with downward adjustments, more volatile data and products with a high periodicity. Table 11 investigates these effects in more detail. A multivariate regression analysis is executed to investigate the effect of adjustment size, direction, volatility and periodicity on FCIMP. Given that Table 7 and 8 indicate a non-linear relationship for adjustment size and volatility measured by the standard deviation, the squares of these variables are included in the regression analysis.

Table 11: Investigating the effect of volitality (SD) and periodicity on FCIMP, controlled for the effects of adjustment size and direction

	Restrictive judgment model			Integrative judgment model		
	B	SE B	Std Beta	B	SE B	Std Beta
Intercept	-0.423	0.086	0^{***}	-0.017	0.010	0
Adjustment size	1.362	0.130	0.803^{***}	1.018	0.071	0.820^{***}
Adjustment size sq.	-1.470	0.113	-1.019^{***}	-0.527	0.140	-0.206^{***}
Downward adj	-0.036	0.081	-0.034	0.012	0.007	0.049
Upward adj	-0.213	0.082	-0.203^{***}	-0.052	0.008	-0.196^{***}
Volatility (SD)	0	0	0.686^{***}	0	0	0.208^{*}
Volatility (SD) sq	0	0	-0.473^{***}	0	0	-0.198^{*}
Volatility (cat.)	0.254	0.033	0.223^{***}	0.006	0.007	0.021
Periodicity	0.028	0.035	0.021	-0.001	0.008	-0.003

Note. Significance levels: * p $<$ .05, ** p $<$ .01, *** p $<$ .001

Note. R^2_{Restr} = .418, R^2_{Integr} = .405

This table shows that the volatility of a product has a larger impact for the restrictive model than the integrative model. Interestingly, the impact of volatility expressed as exceptional products is not significant for the integrative model. This can be explained by the fact that every special item has its unique characteristics (e.g., a CD of a local band will have a different effect than a CD of an international band), and therefore, this effect is hard to capture by the integrative model. In the restrictive judgment model, however, the expert judge is better equipped to account for these exceptional products. The effect of the periodicity of the product on forecasting improvement is not significant in both regression analyses. Although a positive effect is visible in Table 10 for both models, where weekly products show a higher improvement than monthly products, this effect disappears once controlled for the other variables. This improvement is entirely due to a higher volatility of weekly products and a decreased error in direction and size of the adjustment. Presumably, the forecaster is more familiar with weekly products than with monthly products. This increased familiarity allows the forecaster to be better calibrated in their forecasting.

In sum, the integrative model consistently outperforms the basic statistical model.

However, in those cases where the restrictive model has a beneficial effect compared to the basic model, this improvement is generally larger than with the integrative model.

5.2.2. Profitability

The output of the prediction model serves as input for an optimization model, which calculates the optimal number of magazines for each PoS. The optimization enables an expression of model performance in profatibility. To guarantee anonymity of the company and protect confidential data, we report the measure of profit expressed as a percentage of the hypothetical maximum profit. The hypothetical maximum profit is reached if every PoS would sell the exact amount of predicted magazines. In this situation, no product would have to be returned or none would have to be delivered additionally, maximizing profit. The results of the optimization model for restrictive judgment and integrative judgment respectively, are compared to this perfect situation.

Table 12: Model performance expressed as MAPE, MdAPE, FCIMP and profit (%)

	Basic Statistical Model	Restrictive Judgment	Integrative Judgment
MAPE	25.03%	26.18%$^{\text{BSM: n.s.}}$	22.54%$^{\text{BSM: *** / RM: **}}$
MdAPE	9.95%	10.83%$^{\text{BSM: n.s.}}$	8.24%$^{\text{BSM: *** / RM: ***}}$
FCIMP	-	-1.15%	2.49%$^{\text{RM: **}}$
Profit	82.00%	77.17%$^{\text{BSM: ***}}$	82.80%$^{\text{BSM: *** / RM: ***}}$

Note. Proposed models are compared with the basic statistical model (BSM) and each other (RM).

Note. Significance levels: * $p < .05$, ** $p < .01$, *** $p < .001$, $^{\text{n.s.}}$ not significant

The result is a percentage which indicates how close the optimization model comes to the ideal situation. This amounts to a profit percentage of 82,00% for the basic statistical model (see Table 12). This implies that there is still a potential to improve profitability with 18% if the forecast in each PoS would be perfectly accurate. The restrictive model demonstrates a damaging effect on profit, reducing the profit percentage to 77,17% (a decline of 4.83 *p.p.* compared to the statistical model, $p < .001$). The integrative model however, displays a positive effect on profit compared to the statistical model (an improvement of .80 *p.p.*, $p < .001$), with a profit percentage of 82.80%.

This difference is significant at the 0.001 level, presumably due to the better allocation of magazines on the PoS level. This shows that by simply adapting the forecasting process, significant profits can be obtained by the organization.

6. Discussion

6.1. Summary of findings

This study builds further on previous judgmental forecasting research, which indicates that restrictive human judgment proves to be valuable only in specific cases. The method of incorporating human judgement predictions in an integrative way suggests a way to counter the harmful effects of upward adjustments, small or overly large adjustments and adjustments made in the case of low periodicity or volatility. Previous research has extensively proven that people have a tendency to adjust forecasts, even if they have no additional information (Lawrence et al., 2006). Forecasters tend to have increased confidence in forecasts they have adjusted (Kotteman et al., 1994) and tinker unnecessarily with the outcomes of the statistical model, simply to show they are paying attention to the task (Fildes et al., 2009). These unnecessary adjustments lead to a general decline in forecasting accuracy. The integrative approach counters the harmful effects of these adjustments. Indeed, the methodology has a positive effect on accuracy in all scenarios. However, in those cases where the restrictive approach proved to be beneficial (medium sized and big adjustments, downward adjustments and adjustments in case of high volatility), the integrative approach limited these beneficial effects.

6.2. Managerial contributions

The integrative approach can prove beneficial for business practice on four different aspects: accuracy, process, profit and people management. First, the integrative approach has proven to be able to counter any damaging effects that existed in the forecasts because of judgmental factors. While the traditional restrictive approach demonstrated the same pitfalls (damaging small adjustments, positive adjustments) as found

in previous literature (e.g., Fildes et al., 2009), the integrative approach cancelled out all effects that were detrimental for forecasting accuracy. Classic forecast accuracy metrics such as MAPE and MdAPE show a reduction in error rates and thus an improvement in general forecasting accuracy.

Second, the integrative approach can lead to process improvement. The model enables a tailored drill-down of judgmental adjustment to the different PoS. Manual adjustment per PoS would require a labour intensive way of adjusting forecasts. Using business rules may not always prove accurate. By using the integrative approach, the model parameters automatically indicate the importance of judgmental adjustment for each PoS separately.

Third, this dataset provided the opportunity to look beyond established theoretical measures of accuracy and provided a picture of profitability by distinguishing between the forecasting system and the optimization system. The integrative approach thus has an even more tangible positive consequence, in that it heightens not only accuracy measures but also heightens profit. While the classic, restrictive approach damaged profit compared to the basic statistical forecasting model, the integrative approach not only mitigated this damaging effect but caused an increase in gained profit on top of the statistical forecasting model. While the percentages expressed in the results section may be seen as small numbers, it is important to note that these correspond to large differences in absolute profit numbers. The translation from error measures to profit margins provides a unique opportunity for a better communication between researchers and practitioners on the importance of improved forecasting accuracy.

Fourth, the integrative approach has a high chance of being seen as acceptable by the forecasters, since it does not take away their input and thus sense of ownership. Previous efforts aimed at reducing error due to faulty judgment in forecasting have focussed on biases associated with judgment, and on debiasing approaches. A popular technique has been debiasing via consciously attempting to correct judgment via ad-

vice or explicitly pointing towards the declining accuracy. While this technique has the potential to be applicable to biases that can vary over time, judgmental forecasters however, seem to persist in their damaging adjustments (Lim and O'Connor, 1995). Önkal, Goodwin, Thomson, Gönul and Pollock (2009) interviewed forecasters and found that they not only adjust to integrate their knowledge, but also to own the forecast, contribute to the forecast and gain a sense of control over them. Consequently, when forecasters are advised to leave the forecast alone, this sense of ownership is taken away from them, possibly leading to resistance to this way of working. With an integrative approach on the other hand, the possibility remains for the forecasters to provide their input, while simultaneously correcting for possible damaging adjustments. Indeed, forecasters continue to be asked for their input. Additionally, the integrative model takes the judgment variable into account according to its predictive power in past forecasts. Increased judgmental accuracy in the past will therefore lead to increased impact of the judgment of the forecasters in the present. If the forecasters performs badly, e.g. consistent under-forecasting by a sales forecaster to easily achieve its target, judgment will no longer have a significant influence on the forecast result. This method would thus motivate forecasters to perform accurate and more objectively. Additionally, forecasters can be informed on when to rely on their own judgment (restrictive) and when to rely on the integrative model for better results.

6.3. Limitations and further research

The study has some limitations. First, the dataset contains information about forecasts, adjustments, sales numbers and profit from a single company. However, our analysis finds the same pattern as previous studies (e.g., Fildes et al., 2009), providing an indication that the data is comparable. Similar to previous literature (Fildes et al., 2009), negative adjustments were more profitable than positive adjustments. Similar to Sanders and Ritzman (1992), judgmental adjustment was found to be beneficial in high volatility series, measured both by the forecast error and the classification by the

company. In low volatility series, restrictive judgment damaged accuracy. Given the large similarities in data patterns, results should hold in further research with the integrative model in other companies, to further test the robustness of this solution.

Additionally, qualitative inquiry is needed to test the acceptance of this new method by the forecast users. Previous research has indicated trust issues with forecast support systems that restrict judgment (Goodwin et al., 2011). The proposed method should counter the negative effects of judgmental adjustments, while retaining the positive feelings of the forecaster concerning their input, value and ownership. An interesting avenue for research would be to apply the integrative model on forecaster level, rather than on product level. In this dataset, no information was available on the past performance of the individual forecaster. Future research could investigate methodologies that incorporate the forecaster's performance across products, in order to determine even more accurately the weight of the expert forecast in the integrative model. In the organization involved in this study, the same forecaster was typically responsible for the same products. This could be different in other organisations. Especially in this situation, this might play an important role. If forecaster X has a good performance history on product Y, but is sometimes replaced with the less effective forecaster Z, the integrative model might wrongfully include forecaster Z's adaptations. Notably, when moving to individual level data and inclusion based on past performance, motivation will play an important factor.

Next, in this study, including judgment in an integrative way improves performance consistently. Unfortunately, while the integrative judgment was able to translate all damaging judgmental adjustments into beneficial adjustments, it also tempered the magnitude of the beneficial judgmental adjustments. Future research should further look into optimizing the effects of integrative judgment, such that the beneficial effects of judgment remain equally large. A possible avenue is the establishment of application rules with regard to the use of restrictive judgment or integrative judgment.

Additionally, future research could investigate the impact of the reconsiliation strategy of Hyndman et al. ((Hyndman et al., 2011)) on judgmental forecasting. While their paper provides on statistical forecasting, an application on combined judgmental and statistical forecasting could provide valuable insights.

7. Conclusion

This study looked into a forecasting model which integrated expert adjustment into the model itself. The results show that there is a beneficial effect of integrative judgment, compared to the restrictive approach and the basic statistical model. The dataset was tested extensively according to size of the adjustments, their direction, and the volatility of the data. The latter was tested in two ways: by using the Standard Deviation (Fildes et al., 2009) and by using a category as defined by the company. Results showed similar patterns to the extensive dataset of Fildes et al. (2009), indicating a level of comparability between our dataset and others. Additionally, this study takes into account the call for judgmental forecasting researchers to pay attention to hierarchical levels of forecasters (Kremer et al., 2015). The analyses show the added value of the integrated approach on both levels of granularity. Lastly, this study was able to provide an insight into the direct financial consequences of improved forecasting accuracy. The relationship with direct financial gain is close to non-existent in judgmental forecasting literature, while it plays a pivotal role for the practitioner. Improving forecasting accuracy in practice is an important task for researchers in our field (Sanders and Manrodt, 2003). This study thus responds to a call for more research with company data (Sanders, 2009). The integrative approach can debias judgmental forecasting without negatively affecting feelings of ownership from the forecaster, while improving forecasting accuracy.

8. Acknowledgements

The authors would like to thank the Flemish Research Council (the ICM-FWO fellowship for author 2) and the Academic Research Fund of the Vlerick Business School for financial support; and Provideor for sharing their data and further cooperation.

9. References

Alvarado-Valencia, J., Barrero, L., Önkal, D. and Dennerlein, J. (n.d.), 'Expertise, credibility of system forecasts and integration methods in judgmental demand forecasting', *International Journal of Forecasting* **33**.

Armstrong, J. and Collopy, F. (1998), *Integration of statistical methods and judgment for time series forecasting: principles from empirical research*, John Wiley and Sons, New York, pp. 269 – 293.

Armstrong, J., Green, K. and Graefe, A. (2013), Golden rule of forecasting: Be conservative.

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

Clarke, S. (2006), 'Transformation lessons from coca-cola enterprises inc.: Managing the introduction of a structured forecast process', *Foresight: The International Journal of Applied Forecasting* (4), 21 – 25.

Diamantopolous, A. and Mathews, B. (1989), 'Factors affecting the nature and effectiveness of subjective revision in sales forecasting: An empirical study', *Managerial and Decision Economics* **10**, 51 – 59.

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

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

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

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

Fildes, R., Goodwin, P., Lawrence, M. and Nikolopoulos, K. (2009), 'Effective forecasting and judgmental adjustments: an empirical evaluation and strategies for improvement in supply-chain planning', *International Journal of Forecasting* **25**(1), 3 – 23.

Fildes, R., Goodwin, P. and Onkal, D. (2016), Information use in supply chain planning.

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

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

Franses, P. H. and Legerstee, R. (2013), 'Do statistical forecasting models for sku-level data benefit from including past expert knowledge?', *International Journal of Forecasting* **29**(1), 80 – 87.

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

Goodwin, P. and Fildes, R. (1999), 'Judgmental forecasts of time series affected by special events: does providing a statistical forecast improve accuracy?', *Journal of Behavioral Decision Making* **12**(1), 37 – 23.

Goodwin, P., Fildes, R., Lawrence, M. and Stephens, G. (2011), 'Restrictiveness and guidance in support systems', *Omega : The International Journal of Management Science* **39**(3), 242 – 253.

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

Hyndman, R., Ahmed, R., Athanasopoulos, G. and Shang, H. (2011), 'Optimal combination forecasts for hierarchical time series', *Computational Statistics and Data Analysis* **55**, 2579–2589.

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

Kerkkänen, A., Korpela, J. and Huiskonen, J. (2009), 'Demand forecasting errors in industrial context: Measurement and impacts', *International Journal of Production Economics* **118**(1), 43 – 48.

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

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

Kremer, M., Siemsen, E. and Thomas, D. (2015), 'The sum and its parts: Judgmental hierarchical forecasting', *Management Science* .

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

Lim, J. and O'Connor, M. (1995), 'Judgmental adjustment of initial forecasts: its effectiveness and biases', *Journal of Behavioral Decision Making* **8**, 149 – 168.

Mahmoud, E., DeRoeck, R., Brown, R. and Rice, G. (1992), 'Bridging the gap between theory and practice in forecasting', *International Journal of Forecasting* **8**(2), 251 – 267.

Mentzer, J., Bienstock, C. and Kahn, K. (1999), 'Benchmarking sales forecasting management', *Business Horizons* **42**(3), 48–56.

Merzifonluoglu, Y. (2015), 'Risk averse supply portfolio selection with supply, demand and spot market volatility', *Omega* **57A**, 40 – 53.

Moon, M., Mentzer, J. and Smith, C. (2003), 'Conducting a sales forecasting audit', *International Journal of Forecasting* **19**, 5 – 25.

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

Oliva, R. and 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.

■Önkal et al.

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

Ransbotham, S., Kiron, D. and Prentice, P. (2016), 'Beyond the hype: The hard work behind analytics success', *Mit Sloan Management Review* **March**.

Samawi, H. and Vogel, R. (2013), 'Notes on two sample tests for partially correlated (paired) data', *Journal of Applied Statistics* **41**(1), 109 – 117.

Sanders, N. (2009), 'Comments on "effective forecasting and judgmental adjustments: an empirical evaluation and strategies for improvement in supply-chain planning"', *International Journal of Forecasting* **25**, 24 – 26.

Sanders, N. and Manrodt, K. (2003), 'The efficacy of using judgmental versus quantiative forecasting methods in practice', *Omega* **31**, 511 – 522.

Sanders, N. R. and 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.

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

Scarpel, R. (2015), 'An integrated mixture of local experts model for demand forecasting', *International Journal of Production Economics* **164**, 35 – 42.

Shan, J., Ward, J., Jain, S., Beltran, J., Amirjalayer, F. and Kim, Y. (2009), 'Spare-parts forecasting: A case study at hewlett-packard', *Foresight: The International Journal of Applied Forecasting* **14**, 40–47.

Silver, M. (1991), 'Decisional guidance for computer-based decision support', *MIS Quarterly* **15**(1), 105 – 122.

Sinha, A. and Zhao, H. (2008), 'Incorporating domain knowledge into data mining classifiers: An application in indirect lending', *Decision Support Systems* **46**, 287 – 299.

Steenburgh, T. J., Ainslie, A. and Engebretson, P. H. (2003), 'Massively categorical variables: Revealing the information in zip codes', *Marketing Science* **22**(1), 40 – 57.

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

Syntetos, A., Nikolopoulos, K. and Boylan, J. (2010), 'Judging the judges through accuracy-implication metrics: The case of inventory forecasting', *International Journal of Forecasting* **26**(1), 134 – 143.

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

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

Turner, D. (1990), 'The role of judgement in mactroeconomic forecasting', *Journal of Forecasting* **9**, 315 – 346.

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

Worthen, B. (2003), 'Future results not guaranteed; contrary to what vendors tell you, computer systems alone are incapable of producing accurate forecasts', *CIO* **16**(19), 1 – 4.

Zotteri, G. and Kalchschmidt, M. (2007), 'A model for selecting the appropriate level of aggregation in forecasting processes', *International Journal of Production Economics* **108**(1-2), 74 – 83.