Dharmawardane, Chethana; Sillanpää, Ville; Holmström, Jan

High-frequency forecasting for grocery point-of-sales

Published in:
OPERATIONS MANAGEMENT RESEARCH

DOI:
10.1007/s12063-020-00176-7

Published: 01/06/2021

Document Version
Peer-reviewed accepted author manuscript, also known as Final accepted manuscript or Post-print

Please cite the original version:
Dharmawardane, C., Sillanpää, V., & Holmström, J. (2021). High-frequency forecasting for grocery point-of-sales: intervention in practice and theoretical implications for operational design. *OPERATIONS MANAGEMENT RESEARCH*, *14*(1-2), 38-60. https://doi.org/10.1007/s12063-020-00176-7

High-frequency forecasting for grocery point-of-sales: Intervention in practice and theoretical implications for operational design

(author details)

Abstract

Food waste in grocery supply chains may exceed one third of the total volume, depending on the category. To address this problem effectively, grocery retailers are introducing automated systems for more efficient store replenishment and dynamic pricing. The stock keeping unit (SKU) and store level forecast is pivotal in these operations management solutions, but operationally challenging. Large grocery retailers have millions of SKU-store combinations that depending on the operational application would need to be forecasted on a weekly, daily, hourly, or even 15-minute frequency. However, in grocery it is challenging to account for demand variation at high frequencies without introducing manual decisions into the process of forecast model configuration. To investigate the limits of current practice and explore opportunities of technology-enabled change, we explore how an advanced forecasting method for electricity demand, called TBATS, can automate daily forecasting for grocery store replenishment. Adopting an interventionist approach, we explore the implications for the design of the operational process in the operational setting provided by the case company. We find that TBATS can produce high frequency base forecasts for the SKU-store level accurately for a period exceeding 3 months. This finding points to the opportunity of shifting operational focus from recalculating forecasts to monitoring forecast errors. Introducing variable, even indefinite re-training frequencies for forecasting models is a significant change of the forecasting process for situations where monitoring requires less computation than retraining, potentially reducing the time and cost associated with increasing the forecast frequency.

Key words: Grocery retail operations, automatic store replenishment, high frequency forecasting, retraining frequency, error monitoring

1. Introduction

The definition of high frequency forecasting depends mainly on the established practices for a specific context, as well as available techniques and computational resources (Kourentzes and Crone 2008). In grocery retail, forecasts are generally produced on a weekly basis (Fildes et al. 2019), therefore, forecasting at smaller intervals – for e.g. daily – can be considered as high frequency. High frequency forecasting can be used to identify granular levels of seasonality in demand (Fildes et al. 2019), which helps grocery retailers improve product freshness and availability while reducing spoilage, overstocking and inventory costs. The grocery industry operates on intense price competition and low margins (Kotzab and Teller 2003), which makes improved availability without spoilage a powerful lever to improve customer service and profitability (Corsten and Gruen 2003). In addition, reducing wastage can reduce a retailer's environmental footprint and promote sustainability in their public relations, since retailers are gatekeepers of the wider food system influencing the amount of food waste generated (Gruber et al. 2016).

There is a strong incentive to automate the forecasting process (fully or partially) when calculating frequent forecasts (Taylor 2007). The need for automated replenishment systems based on information from the point of sales is well documented in literature (Ali et al. 2009; De Toni and Zamolo 2005; Whiteoak 2004). Kiil et al. (2018) finds that automatic forecasting and replenishment reduces food waste by as much as one fifth. In addition,

the sheer number of forecasts required often makes it imperative to automate the forecasting process. Seaman (2018) explains the scale of the problem by using a typical Walmart store, which can hold more than 200,000 different products. Since there are approximately 5000 Walmart stores in the USA, this amounts to almost 1 billion unique forecasts. Even for a mid-sized grocery store chain, daily forecasts for thousands of products in hundreds of stores are needed. This type of high frequency forecasting – the need for many forecasts at frequent intervals – pushes the frontiers of big data analytics and machine learning (Choi et al. 2017) with major implications for digitalization of operations management (Holmström et al. 2019).

In the comprehensive review by Fildes et al. (2019), we can find examples of recent research on weekly forecasting on the level of individual products, or stock keeping units (SKUs), across multiple supermarkets. Despite its high practical relevance, daily forecasting on the level of product and store has not been examined in research literature since Taylor's (2007) pioneering experimentation to forecast the daily demand for a sample of products in one grocery store. The current usage of daily forecasting is limited at best in grocery[1]. Many practical and technological challenges constrain the shift to daily, product and supermarket specific forecasting to support the replenishment systems used by grocery retailers (Ehrentahal et al. 2014).

To investigate the limits of current practice and explore opportunities of technology enabled change in automated grocery forecasting and replenishment, we are working with a leading provider of retail planning solutions (hereafter referred to as the 'case company'). The case company is based in the Nordics, with customers on all continents and offices across Europe, North America and Asia. Due to their proprietary in-memory computing and early adoption of a 'software as a service' business model, they are recognized as a technology leader in retail planning solutions. At the time of this research, they provided weekly product and store specific forecasts using the Holt-Winters method (Winters 1960), considering annual seasonality effects. To manage the within-week variability in daily forecasting, they used predefined 'week profiles', i.e., a set of rules. This rule-based approach requires significant manual input and monitoring to manage bias and precision for daily, store level forecasts. Therefore, the case company was interested in investigating alternative forecasting methods or practices which could reduce the amount of manual input required.

Our research problem – reducing reliance on manual input for producing daily base forecasts at product and store level – stems from the challenge faced by the case company. However, automating a forecasting system while increasing the accuracy and frequency of forecasts is a challenge faced by almost every grocery retailer and retail solutions provider. Therefore, this research has the potential to provide managerially relevant outcomes as well as address the research gap in high frequency forecasting for grocery. After comparing some existing methods of high frequency forecasting, we decided to utilize TBATS (De Livera et al. 2011) in this research. Thus, we explore two research questions in this paper: How can an alternative method – TBATS – which is developed for high frequency utility demand forecasting (refer Lago et al. 2018, and Karabiber and Xydis 2019), be introduced for automated and daily forecasting of grocery demand? What are the implications of such an introduction for the design of the operational process?

[1] On the retail supply chain level, van Donselaar et al. (2006) investigated daily replenishment, but not forecasting, of perishables in six supermarkets. Daily forecasting of aggregate replenishment orders placed to the distribution center using different types of demand signals (Narayanan et al. 2019) and different types of demand representations (Sillanpää and Liesiö 2018) was recently investigated. Note that daily forecasting of aggregate replenishment orders is different by an order of magnitude from daily forecasting in individual supermarkets.

Adopting an interventionist approach (Oliva 2019), we investigate in an operational setting provided by the case company, the transition to a unified forecasting process; i.e. a process without the current use of week profiles. For perishables, we find that automation can produce baseline forecasts with improved accuracy and bias in comparison to the case company's current solution. For highly intermittent demand, automated high frequency forecasting produces comparable results to the current solution, providing potential benefits for operations management through automation. In addition, we find that forecast bias can be significantly reduced when using automatic daily forecasting, which for some examined product groups exceeded 50 percent. Examining how the forecast horizon influences the performance and ranking of forecasting models, we question the widely used procedure – in both research and practice – of training forecasting models at fixed intervals. We propose a method of monitoring forecast errors, and training forecasting models only when required. We believe this method can significantly reduce forecasting costs for grocery retailers who stock a substantial number of perishable products.

In the next section, we review literature to outline our theoretical positioning. Then, we introduce our case in more detail, and describe the problem situation at hand and the intervention design to explore it in the case setting. We report our and the case company expectations and compare those with the results obtained. To conclude, we elaborate on the insights gained from the intervention and discuss their practical implications.

2. Literature review

In this section, we review literature to develop a theoretical framework for the intervention design and expected outcomes. First, we review forecasting and its operational uses in grocery. Next, we examine three approaches for high frequency forecasting that are potentially relevant for the problem at hand. Then, we select one of these approaches for further examination and testing in daily forecasting on product and store levels. To conclude, we examine the operationally important consideration of forecasting horizon and retraining frequency in introducing daily forecasting for grocery.

2.1 Forecasting and its operational uses in grocery retail

Fildes et al. (2019) provide a succinct review of retail forecasting, discussing common forecasting problems faced by retailers and the reasoning for forecasting at different levels (strategic, tactical and operational). At the operational level, frequent forecasting of demand at SKU-store level have many benefits, such as increased revenue and improved customer satisfaction. Taylor's (2007) pioneering research, was the first to forecast individual SKUs in the store at a daily level. Only grocery products showing fast moving demand were chosen for this research, and the forecasting horizon ranged from 1-14 days. This research was limited to SKUs from one store. Most of the literature consider weekly forecasting of grocery demand. A recent example is Ramos and Fildes (2017), who use data from one store and exclude SKUs showing intermittent demand. They calculate 1-step ahead rolling forecasts.

Grocery retailers depend on timely replenishment to keep their businesses running smoothly. Most grocery retailers stock large numbers of SKUs in wide assortments in their stores (Amine and Cadenat 2003), which makes manual forecasting difficult and time consuming. Demand forecasting alongside timely replenishment can contribute towards reducing stock outs, which directly increases customer satisfaction and loyalty (Corsten and Gruen 2003; Anderson Consulting 1996). Fernie and Grant (2008) discuss the repercussions of stock outs and observe that most customers will go to another store, or not make a purchase at all, when there is no on-shelf availability for a desired product. Corsten and Gruen (2003) found that some customers might even permanently

switch stores due to out of stock situations. Therefore, stock outs can lead to loss of revenue for grocery retailers in multiple ways.

Corsten and Gruen (2003) also found that retail stock out rates are not random. Instead, they generally vary by the day of the week in a distinctive manner. For e.g., Sundays and Mondays have the largest average levels of stock outs. This could be a result of inefficient delivery schedules, but the main reason is that shopping is generally highest during the weekend. Demand levels usually differ during the week (based on the day of the week, sometimes even the time of the day): it is easily observable that there are more people shopping on Friday evenings in a supermarket, when compared to Tuesdays. Therefore, clearly there is within-weekly seasonality present in grocery demand, and grocery store replenishment generally requires forecasts with daily granularity (Fildes et al. 2019) to avoid stock outs within the week. However, the week level is still the preferred time granularity for grocery forecasting, which is evident when investigating the related literature (refer the review by Fildes et al. 2019). As mentioned, this is partly due to the inability to perform daily forecasting in many automated replenishment systems used by retailers (Ehrenthal et al. 2014). This limitation is especially a cause for concern for grocery retailers as they routinely deal with perishables such as fresh produce. Excess will lead to wastage since these products perish quickly, usually within a few days. On the other hand, customer retention levels will be negatively affected in the case of stock outs, as perishables act as the main driver in customers deciding where to shop (Peterson 2014). It is evident that accurate daily forecasts are very useful, especially for perishables, for grocery retailers to improve their competitive position.

2.2. High frequency forecasting: Three approaches

High frequency forecasting is not a new phenomenon. However, there is a limited use of high frequency forecasting in grocery. In response, we broaden our search to include also other industries where high frequency forecasting is more common. We identify three forecasting models that are of potential interest. Two of these (ARIMA and Holt-Winters) are established models which were modified later to suit multiple seasonality. The other (TBATS) was developed especially with multiple seasonality in mind.

ARIMA (Box et al. 2015) is a popular forecasting model used by both practitioners and researchers. It stands for Autoregressive Integrated Moving Average. The parameters p, d and q are used to classify an ARIMA model, and they depict the order of the autoregressive, integrated and moving average parts of the model respectively (Christodoulos et al. 2010). Considered as a benchmark in research literature, ARIMA has remained popular especially for short term forecasting in many different practical contexts (Taylor 2003).

The Holt-Winters model is widely known and extensively used in grocery retailing. Practitioners have favored it over the years due to its general robustness, simplicity and reliability (Gardner 2006). It is an extension of exponential smoothing, and has three equations for trend, level and seasonality and a smoothing parameter for each (Winters 1960). Another major benefit is that it is quite simple when compared to ARIMA. However, a drawback is that it can be extra sensitive to unusual events and outliers. Even at present, researchers are trying to find ways to improve it, for e.g. by widening its prediction intervals in order to capture the uncertainty of forecasts (Goodwin 2010).

We can consider either ARIMA or Holt-Winters as a valid choice for daily and SKU level forecasting, but they both share a common problem: their original versions are limited to one seasonal pattern. We expect daily grocery demand data to have at least two seasonal patterns: annual and weekly. One solution is to use adapted ARIMA and Holt-Winters (Taylor 2003; Taylor et al. 2006; Taylor 2010). For e.g., Taylor (2003) used

multiplicative double seasonal ARIMA and double seasonal Holt-Winters to forecast short-term electricity demand successfully. For this, he introduced an additional seasonal index with an extra smoothing equation to the forecasting models.

Another solution is to use a forecasting model custom made for multiple seasonality, instead of an adapted forecasting model. TBATS is one such option (De Livera et al. 2011) which is slowly gaining popularity, especially in utility demand forecasting. It uses trigonometric functions that also allow it to handle complex demand patterns with non-integer and high frequency seasonality (Grmanová et al. 2016). In utility demand forecasting, external variables such as weather patterns tend to affect forecasts. Lago et al. (2018) analyzed the accuracy of 27 forecasting models and found that TBATS shows a high level of accuracy which makes it a good choice (even though at present it cannot be used with external variables). This was the case even when TBATS was compared to forecasting models that could use external variables. However, TBATS has not yet been used in grocery demand forecasting.

2.3 What is TBATS exactly, and why is it interesting for high frequency grocery forecasting?

The TBATS model is an extension of the BATS model. In Appendix 1, we outline the reasons for modifying BATS in order to obtain TBATS and discuss both models in detail. We also explain the usage of the Box-Cox transform, ARMA errors and dampened trend (Appendix 1, Table 1) and summarize the comparison between the adapted ARIMA, adapted Holt-Winters and TBATS models (Appendix 1, Table 2). We have also briefly discussed Bayesian Regression Modeling (Appendix 1, Table 2) since at the time of writing this paper, the case company is introducing this method in their forecasting solution. However, we did not include this method in our comparison, since our specific *research problem* – reducing the reliance on manual input for transitioning to daily SKU-store level base forecasts – does not call for a method with its particular characteristics (such as the ability to use additional components in the forecasting model)[2].

When an ARIMA model is dominated by seasonality and trend variations, its forecasting performance decreases. It also lacks transparency when compared to other models due to added seasonal factors (Gould et al. 2008). After comparing this with double exponential smoothing Holt-Winters model, Taylor (2003) concluded that even though it performs reasonably well, the Holt-Winters model outperforms it considerably in short-term electricity demand forecasting. However, there are many parameters to be estimated in the adapted Holt-Winters model, and this can be difficult when there are many seasonal components (De Livera et al. 2011).

As mentioned by De Livera et al. (2011), difficulties arise when using traditional forecasting models in the presence of complex seasonal patterns. For e.g. when forecasting is done once every week and when the time series (used to train the forecasting model) is longer than a year, annual seasonality effects must be factored into the model. Some years have 52 weeks while some have 53, therefore a frequency between the two would be logically more accurate for calculation purposes. In other words, the concept of leap year would be considered and the number of days in a year would be identified as 365.25 (average of 366 + 365*3), and the frequency as 52.18 (365.25/7). However, in practice, the frequency used is either 52 or 53 when models cannot handle non-integer frequencies (Kourentzes et al. 2014), which is usually the case. One of the biggest advantages of TBATS is that it allows the usage of such non-integer frequencies; in fact, it has been designed to handle such complex seasonal patterns (De Livera et al. 2011).

[2] See Pulkka (2020) for a further description of the components and functionality of the Bayesian Regression Model applied in the context of grocery retailing.

TBATS also considers the fact that errors can be correlated and tries to model it to make the forecasts more accurate (De Livera et al. 2011). Errors correlate when some effects such as seasonal patterns are not captured adequately by model parameters (for e.g. when limited sample sizes are used to train the models). This was also discovered by Taylor (2003), who found that using an AR(1) process to model the errors after using the double seasonal Holt-Winters model provided a significant boost to forecast accuracy. By considering the correlation in errors automatically, TBATS makes the process of forecasting more efficient and easier.

A special note must be made about what TBATS has been used for since it was introduced in 2011, and its limitations for using it in the grocery setting. TBATS has been used in empirical studies across a limited number of industries. For e.g. De Livera et al. (2011), with its introduction, displayed its usage in three case studies including US gasoline data. Additionally, Pereira (2016) used it to forecast hotel occupancies. However, the majority of TBATS studies are done for electricity load forecasting (refer Brozyna et al. 2018; De Livera et al. 2011; Grmanová et al. 2016; Lago et al. 2018 for examples). Usually one or very few forecasts are produced in these contexts, and multiple seasonality patterns are considered (generally at annual, weekly and daily levels).

When Pereira (2016) used TBATS to forecast daily hotel occupancies, he also conducted an empirical study that compared the performance of different models with that of TBATS. He found that TBATS generally outperformed other models due to its ability to tackle complex seasonal patterns. Another aspect pointed out by Pereira (2016) was that most forecasting models cannot handle large frequencies, thus they ignore the annual frequency (which is 365.25). TBATS is suitable even when large data sets (containing more than a year's worth of data) are used to train the forecasting model. Veit et al. (2014), who tested some state-of-the-art methods for forecasting electricity demand, found that TBATS performed in a robust manner even when forecasting horizons were increased. This is useful since robustness is a requirement for grocery demand forecasting.

2.4 The impact of the forecast horizon and retraining frequency

When calculating forecasts, another element to consider is the forecast horizon. Fildes et al. (1998) found that certain methods such as 'Robust Trend' and 'Holt' showed considerably improved performance with longer forecasting horizons, whereas methods such as 'Damped Trend' showed superior performance at shorter horizons. Koehler and Murphree (1988) found that the 'Single Exponential Smoothing' method performed best at shorter forecast horizons, while a 'State Space' method outperformed it at longer forecast horizons. These examples show that the performance of a forecasting model depends on the forecasting horizon in question.

It is well accepted both in theory and practice that forecasts into the distant future are less reliable than forecasts that are closer to the present (Chand et al. 2002). Forecasting models are retrained (model parameters are re-estimated) at regular intervals based on newest data in order to improve forecasting accuracy. The retraining frequency usually depends on the forecast horizon, which is generally based on ordering and planning periods of the retailer (Ma et al. 2016). Sometimes retraining is done as frequently as one-step ahead (with each new forecasted value). One example is Ramos and Fildes (2017), who forecast weekly grocery demand. They produce only one forecast at a time for each SKU. After each forecast, they add the actual sales data (from the test set) to retrain the forecasting model before creating the next forecast. Ma et al. (2016), who also forecast at a weekly level, produce forecasts both one and four-steps ahead at a time.

There are also instances where researchers have preferred not to calculate rolling forecasts, but instead use a fixed forecasting scheme. For e.g. Ma et al. (2016), in addition to rolling forecasts as mentioned above, also calculate forecasts for the entire test set using the same forecasting model. There is no need to retrain the model

in this case. As the authors mention, this is not a likely approach in practice. However, this can be used to evaluate a forecasting model, particularly its robustness over longer forecast horizons.

It is interesting to inspect how the retraining frequency of a forecasting model influences its performance. Retraining at smaller intervals might increase forecasting performance, but it also increases costs (in terms of computational time). This cost can be substantial for mid-size or large grocery retailers with thousands of SKUs to forecast. Therefore, in digitalizing their operations it is imperative that retailers should seek to develop approaches to avoid retraining models and calculating forecasts unnecessarily.

3. Methodology

Our research methodology is design science, seeking theoretical insight through an intervention and evaluation in practice (Holmström et al. 2009). The goal of design science is to frame design propositions in order to transcend a case specific situation to a broad context (Oliva, 2019). In design science in operations management, CIMO is a common method to generalize insights from a specific solution and context. CIMO stands for Context, Intervention, Mechanism and Outcome, and the framed design propositions follow CIMO logic as follows: *"For this problem-in-Context it is useful to use this Intervention, which will produce, through these Mechanisms, this Outcome"* (Denyer et al. 2008). The investigation of mechanisms provides the basis for conceptualization and theory development. As observed by Oliva (2019), for an intervention in practice to inform theory, three elements of the research design are key: Framing the *problem situation* that a practitioner seeks to resolve or improve; Positioning the problem in a *theoretical framework* that can be illuminated through intervention; Designing the *intervention* to inform theory. Our research is conducted following this interventionist research approach. First, we briefly describe our motivation for selecting the case, and then we discuss the problem situation, theoretical positioning and intervention design.

3.1 Motivation for case selection

The case company, a retail solutions provider, was interested in exploring the possibility of producing more automated and granular forecasts than were attainable through their solution in use (at the time). This case introduced us to the cutting edge of practice, and the challenge of high frequency and big data forecasting. This challenge, as it was presented to us, was a primary motivation to start the research collaboration. Another motivation, as we soon found out, was the research gap in terms of daily forecasting in grocery. The case company provided the necessary data and evaluated our alternative design against their current solution. Even though the target was to come up with an alternative design, our intervention also provided other interesting insights, for e.g. regarding forecasting model retraining frequencies. These insights allowed us to move beyond this specific intervention to more general and relevant outcomes.

3.2 Problem situation

The case company initially provided logistics planning and control services to a wide variety of retailers in industries such as agriculture, books and clothes. In 2012, they also started working with grocery retailers. Grocery demand forecasting and replenishment is more challenging than many other types of retailing due to many products and stores, and intensive competition. There is a large variety of products and aggressive competition based on price (due to small margins), leading to variable and unpredictable sales (Alftan et al. 2015). However, retailers need forecasts to be accurate and reliable in order to efficiently plan the supply chain and replenish stores in a timely manner.

The case company has been meeting a clear need for simple, yet effective forecasting and replenishment systems in grocery. However, as the case company is growing and expanding internationally, it has recognized that their current forecasting solution has weaknesses from an operations perspective. At the time this research was conducted, the case company was exploring alternative ways to address these weaknesses. One major drawback was that while their approach to weekly forecasting is simple and effective, the solution for daily forecasting requires an additional process step and is inaccurate without significant manual input.

The case company uses the Holt-Winters model to estimate annual seasonality effects and forecast weekly demand. The weekly forecast values are then split into daily levels using a set of rules (called 'week profiles'). In essence, these algorithms take into account the past demand for each day of the week and divides the weekly demand pro rata. In this step, the end user must select whether week profiles are estimated based on the product's sales history, or whether an aggregate level (for e.g. product group) is used to derive forecasts for each day.

Using this two-step method, base forecasts are produced. In case of special events, such as promotions, additional calculations are needed on top of these base forecasts, which the case company does in a third step – benefiting and requiring situational knowledge – in their forecasting process. Afterwards, they use these forecasts in their replenishment process to provide improved availability and reduced spoilage. In this research, our goal was to explore the potential to automate the two first steps in the forecasting process. In Figure 1, we identify these as the 'base forecasting process'.

A benefit of using the two-step method for base forecast generation is that it is possible to consider the effects of annual and weekly seasonality separately. This allows the case company to use a computationally economical model such as Holt-Winters for weekly demand forecasting, which is important in the grocery problem setting with the large number of SKU-store combinations to be forecasted. Many forecasting models that work well for quickly calculating a large number of forecasts effectively deals with one seasonal pattern but cannot be used in the presence of multiple seasonality effects (De Livera et al. 2011; Godfrey and Powell 2000). Hence, the two-step solution of the case company is used to obtain daily forecasts.

The case company has some years earlier introduced a significantly more effective database, a proprietary technology, which lies behind the case company's expansion in grocery retailing. This technological capability

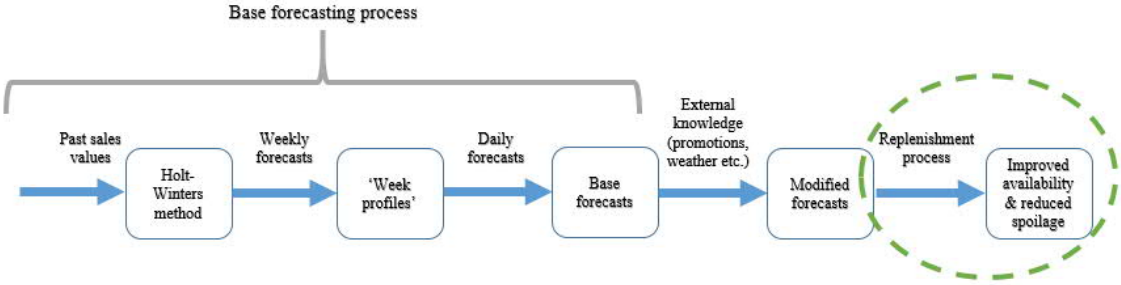

Figure 1: Current solution of the case company

additionally allows the case company to explore opportunities to change the current process. The case company have sought to explore alternative forecasting procedures (in comparison to the current two-step process) and eliminate end-user involvement in their existing base forecasting process. Our intervention constitutes one of these explorations and the practical problem setting of our explorative design research. Through the design of an alternative solution for automatic daily forecasting, we seek implications for operations management theory.

3.3 Theoretical positioning

We have given a broad outline of the theoretical positioning for our intervention in our literature review. To recap, our review covered the use of daily forecasts in grocery operations, high frequency forecasting models and the impact of the forecast horizon and retraining frequency on forecasting model performance.

The current approach of the case company has been discussed and utilized to an extent in literature. For e.g., Taylor (2007) describes the method currently used by their case company (a grocery retailer in the UK), which involves smoothing the weekly sales values using a constant smoothing parameter (α, which is always equal to 0.7). Next, they divide the weekly sales for each day of the week considering past sales and a splitting parameter (γ, which is always equal to 0.1). This is like the two-step process used by the case company, except for the parameter values. Taylor (2011) calls this method *'total and split' exponential smoothing* and uses it to forecast monthly sales for a publishing company (by smoothing the total yearly sales and splitting the total for the months of the year). However, they optimize the smoothing parameters (instead of always using α=0.7 and γ=0.1) to obtain better results. Hence, there is a theoretical foundation behind the case company's current solution.

Accurate daily forecasts are essential, especially for perishables with short shelf lives. Using daily forecasts, the case company optimizes replenishment and allocation processes for their clients. Other functions such as space and workforce planning in grocery retail stores are also affected by forecasts. Daily forecasts are an important component of their retail planning solution, helping them achieve their goals of improved shelf availability, reduced food waste etc. At present, forecasts are produced every four weeks using their current solution. This is done in a rolling manner, where the forecasting model is also retrained with the most current data at the end of each four-week period. However, if a retailer requires it, forecasts can be produced at a different frequency (for e.g. every two weeks). In research, different kinds of rolling forecast horizons are used when forecasting grocery demand. Some are extremely short (for e.g. one-step ahead), while others are longer (for e.g. 16-weeks ahead). Fildes et al. (2019) provide examples in their comprehensive review.

When using shorter forecasting horizons on a rolling basis, it is necessary to retrain forecasting models more frequently. Poler and Mula (2011) point out that forecasting accuracy is calculated using one-step ahead forecasting error in most research, but that in practice, companies are generally interested in the accuracy of forecasts over a longer time horizon. We could not find any literature that explicitly discusses or provides any guidelines on how to choose the rolling forecast horizon or the retraining frequency for a forecasting model. It seems that in literature and practice both, the forecasting models are retrained at fixed time intervals, be it at one-step, four-steps, or even twelve-steps ahead, though the shorter intervals are more common. Here, increasing demands for computational resources to improve forecast accuracy implies an *operational performance frontier* (Schmenner and Swink, 1998), where technology barriers limit the performance of the forecasting process. However, since the new database – a superior technology – was introduced, the previous restrictions have significantly reduced. A question for theory and practice is whether the performance frontier could now potentially, in specific situations, be shifted through operational process redesigns that approach forecasting horizon and retraining frequency as distinct design variables.

3.4 Intervention design

An intervention design is needed for intervening in practice and contribute to the theoretical framework of interest. Our proposed intervention compares two different approaches for daily forecasting, one established in practice, and one novel for the grocery context. Since the main identified drawback in the current solution was

that base forecasting was done in two steps, we decided to investigate forecasting models that will enable us to automatically create daily forecasts in a single step. We introduce TBATS as an alternative forecasting model to the case company setting. The intervention design results in a unified solution that makes the entire base forecasting process automatic, thus requiring no end-user input.

From the existing multiple seasonality forecasting models, we selected TBATS for the intervention design based on our review presented in 2.3. Pereira (2016) mentions that further empirical research using real life data sets and multiple forecasting horizons is needed to understand the applicability and accuracy of TBATS. To our knowledge, there has been no research done that utilized TBATS to forecast demand in the grocery retail industry, at a daily or other level. TBATS seems like a promising method for this purpose, especially with benefits such as the ability to handle multiple seasonality patterns, non-integer frequencies and large frequencies. It can also be utilized in an empirical research using the *forecast* package for R (Hyndman et al. 2018), where the process of constructing the model is fully automated.

We compare the current solution of the case company to a TBATS-based solution. The daily sales forecast is developed form historical daily sales. As each time series is roughly 2.5 years, annual seasonality is also possible to detect from the historical sales. We did not pilot our proposed method for a grocery store in real-life operations. However, we used operational data to test the intervention, and two company representatives joined in the evaluation of the alternative method compared to the case company's current solution. One of the representatives was a Product Owner during this time and had worked at the case company for more than 6 years. He managed the development of automated forecasting solutions for retail, including grocery. The other representative was a Solution Expert and had been at the case company for more than 3 years. He was extremely knowledgeable in all aspects of their solution and its different functionalities.

Our context for using TBATS is different from those where it has been previously applied. Moving TBATS from electricity load forecasting to grocery demand forecasting creates a risk of underperformance, mostly due to its inherent complexity. For e.g., the usage of many parameters in TBATS might make the forecasting model over-fit to the training data, thereby producing sub-standard forecasts. Another obvious difference is that we produce a large number of forecasts in our intervention. Testing it on multiple SKU-stores is necessary because we are working in the grocery industry, where there are many SKU-store combinations to be forecasted at regular intervals. In addition, we only consider two seasonality patterns: annual and weekly, since this is generally sufficient for grocery products as of now. In a grocery retailing setting, retailers not collecting in-store sales data on an hourly basis for their SKU-stores, preventing an investigation of daily seasonality, also limit us.

3.5 Evaluation criteria

To conclude the methodology section, we present the evaluation criteria for our intervention. We use these criteria to compare the performance of our intervention (both rolling and long forecasts) with that of the current solution. Since the new solution in question is a method for forecasting, forecast accuracy is generally considered as the main performance measure (Yokuma and Armstrong 1995). However, Armstrong (2011) states that there are other aspects that can also be considered as vital measures of forecasting model performance, especially implementation related ones such as ease of usage and transparency. We consider that grocery retailers do demand forecasting for large amounts of SKUs at frequent intervals and choose the following performance measures: forecast accuracy and forecast bias, level of automation and computational efficiency, and transparency and ease of usage. We describe the evaluation criteria in detail in Appendix 2 (Table 3 and Table 4).

4. Intervention and outcome

In this section, we present our intervention in detail, starting with a description of the empirical data, our preliminary analyses and the creation of training and testing sets, after which we present the outcomes.

4.1 Data and modeling

The data for this research spans 136 weeks of sales and 18 weeks of forecasts for 1164 SKU-store combinations. The sales values are for the date range 18.10.2015 to 28.05.2018, for all 7 days of the week (954 data points per SKU-store combination). The forecasted values are for the date range 17.01.2018 to 28.05.2018, for all 7 days of the week (132 data points per SKU-store combination). Overall, there are 4 product categories, 61 product sub-categories and 20 stores (summarized in Table 5).

Table 5: Summary of case company data

Category	SKUs	Sub-categories	Covered stores
Category 1	145	8	20
Category 2	391	18	20
Category 3	291	14	20
Category 4	337	21	20

Due to confidentiality, the case company modified the data. The sales volumes were scaled so that they are between 0 and 1. The dates were shifted, but the shifting was done so that seasonality effects remained the same. Since we are not focusing on promotional effects in this research, the data gathered during promotional periods were removed from the data set, and data interpolation was used to estimate values in their place (this was done by the case company). We did not remove outliers from the final data set. It is quite possible that they are there due to holidays, special events or even as part of some cyclic trends (and would affect the model parameters).

The product category and sub-category names are confidential. However, we have provided a brief description of each product category, especially about its shelf life, for readers to obtain an idea about the type of products included. Categories 1 and 2 can be classified as high volume whereas categories 3 and 4 are comparatively low volume. None of the sales data was missing; however, there were instances of zero sales for all product categories. SKUs belonging to product category 3 and 4 generally showed a high number of zero sales, which is characteristic of highly intermittent demand. In comparison, categories 1 and 2 had much less zero sales, which means that these SKUs can be classified as fast moving. We summarize this information in Table 6.

Table 6: Characteristics of case company data

Category	SKUs	Characteristics	Total Volume	Volume per SKU	Zero sales days	Demand type
Category 1	145	Perishable Shelf life: ≈ 1 week	288.5	1.99	8.2%	Fast moving/ high volume
Category 2	391	Perishable Shelf life: ≈ 5 days	2353	6.02	7.7%	Fast moving/ high volume
Category 3	291	Non-perishable Shelf life: ≈ 10 months	55.05	0.19	63.8%	Highly intermittent/ low volume
Category 4	337	Non-perishable Shelf life: ≈ 5 years	34	0.1	77.1%	Highly intermittent/ low volume

In order to analyze the seasonal patterns present in the case company data, we drew monthly and daily box plots for each product category. We observed annual seasonal patterns for all products through the monthly box

plots, which shows that demand changes based on the time of the year. The mean seems to be a bit larger than the median in most cases. This means that the distribution of sales values for most months is slightly skewed to the right; i.e. there are a few exceptional large sales values occurring during most months. The box plots are given in Appendix 3 (Figures 3 – 6).

We specify the ratios of highest mean demand to lowest mean demand in Table 7. This value is somewhat low for categories 1 and 2 when compared to categories 3 and 4. Therefore, we can posit that annual seasonality effects are slightly weak for categories 1 and 2. In comparison, demand throughout the year is more varied for products belonging to categories 3 and 4.

Table 7: Highest mean demand to lowest mean demand ratios (from monthly box plots)

Category	1	2	3	4
Ratio	1.28	1.11	1.62	1.71

In the daily box plots, the mean and median seem to be mostly similar, which means that the sales for each day of the week generally follow a normal distribution. The box plots are given in Appendix 3 (Figures 7 – 10).

The highest mean demand to lowest mean demand ratios are specified in Table 8. This value exceeds 2 in all categories except category 2. Categories 3 and 4 show a large amount of variation in sales values within the week when compared to categories 1 and 2. In other words, products belonging to category 4 have a more varied demand within the week than products belonging to category 2.

Table 8: Highest mean demand to lowest mean demand ratios (from daily box plots)

Category	1	2	3	4
Ratio	2.1	1.8	2.4	2.63

We consider the different demand patterns that occur in the grocery industry (fast moving and intermittent) to understand the best ways in which to apply a daily demand forecasting method. Also of importance are the different contexts in which daily demand forecasting provides significant benefits to retailers, which will greatly affect their decision to switch to a new system. The overarching objective is to evaluate whether a solution that utilizes automatic daily forecasting performs better than the case company's current solution. Through this, we hope to gain insights on how the case company can modify their retail planning solution to serve their customers (grocery retailers) better.

In order to investigate the effect of the forecast horizon, we consider two types of forecast horizons in our intervention. We refer to the first one as a *rolling forecast*, where forecasts are calculated for 4-week periods at a time. The second, called the *long forecast*, has forecasts calculated for the entire test period (18 weeks) in one go. We expect the rolling forecasts to perform generally better than the long forecasts for the same forecasting model. We will compare both sets of forecasts with the forecasts provided by the case company (which have been calculated on a rolling 4-week basis).

Training and test sets were created for each time series using the sales values provided by the case company. In order to convert data into time series, the frequencies of 7 and 365.25 were used. When calculating long forecasts, the last 126 data points of each time series was used as its test set, and the remaining 828 data points were used to train the TBATS model. The TBATS model was then used to forecast the next 126 instances.

When calculating rolling forecasts, the last 126 data points of each time series were used as its test set. An iterative process was then used to create the forecasts. In the beginning, the first 828 data points were used to train the TBATS model, and 28 instances were forecasted using this model. Afterwards, the actual sales values of these

28 instances were also included in the training set and the TBATS model was re-trained. Likewise, the iterative process was run until all 126 data points were forecasted. We describe additional details (practicalities) of the intervention in Appendix 4.

4.2 Outcomes

In this section, we describe the results from the intervention based on the three evaluation criteria: forecast accuracy and bias, level of automation, and transparency and ease of usage.

4.2.1. Forecast accuracy and bias

To increase readability, we separate the results in the forecast accuracy and bias section into the two demand contexts observed in the product categories: fast moving/ high volume and highly intermittent/ low volume. Except for the number of SKU-stores with extremely high bias, the results are small numbers since the case company scaled the sales values to be in the range of 0-1. To make the results easier to present, we have taken the current solution as the benchmark (results equal to 1) and calculated the results of the unified model relative to it. This makes the results easier to interpret as well. For e.g., if the results of the unified model are lesser than 1, this means that the unified model is more accurate compared to the current solution. This process is also used by Pereira (2016) and recommended by Kouprioucchina (2014), and Hyndman and Koehler (2006). The best results in each instance are given in italics. Table 9 contains the results for fast moving/ high volume demand and Table 10 the results for highly intermittent/ low volume demand.

Fast moving/ high volume demand

Table 9: Results for fast moving/ high volume demand for rolling and long forecast horizons

Forecast horizon	Category	Forecast model	RelRMSE	RelMAE	RelME	SKU-stores with extreme biases
Rolling (4 weeks)	Category 1	Current	1	1	1	33.7%
		Unified	*0.9351*	*0.9200*	*0.1714*	*4.8%*
	Category 2	Current	1	1	1	61%
		Unified	*0.7402*	*0.7041*	*0.0046*	*4%*
Long (18 weeks)	Category 1	Current	1	1	1	33.7%
		Unified	*0.9722*	*0.9770*	*0.4857*	*22%*
	Category 2	Current	1	1	1	61%
		Unified	*0.8212*	*0.7891*	*0.1055*	*31.4%*

RelRMSE, RelMAE and RelME stand for Relative Root Mean Square Error, Relative Mean Absolute Error and Relative Mean Error, respectively. These are calculated for the unified solution by taking the results relative to the current solution (=1)

It was encouraging to observe that the new solution performed extremely well both in terms of forecast accuracy and bias. At 4 weeks, the unified model outperformed the current solution in every performance measure we considered. The reduction in forecast bias with the new solution was especially significant. It was at least 28% better than the current solution in every situation, with the unified model showing the best performance for category 2 with a bias reduction of 57%.

At 18 weeks, the unified model outperformed the current solution in every performance measure. This is especially surprising since the forecasts by the current solution were produced at 4-week (rolling) intervals. In other words, the forecasts produced every 18 weeks by the new solution (the unified model) were more accurate and less biased than the forecasts produced every 4 weeks by the current solution. This implies the robustness of

the new solution. Since we are mostly interested in the RelRMSE values and the number of extremely high biases, these results are highlighted in Figure 2.

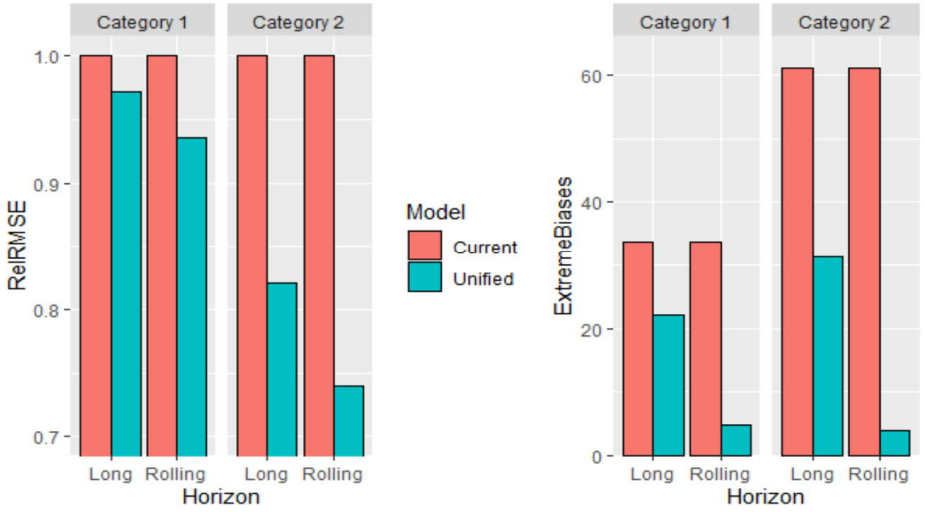

Figure 2: RelRMSE and extreme bias values for fast moving/ high volume demand

Highly intermittent/ low volume demand

Table 10: Results for highly intermittent/ low volume demand for rolling and long forecast horizons

Forecast horizon	Category	Forecast model	RelRMSE	RelMAE	RelME	SKU-stores with extreme biases
Rolling	Category 3	Current	*1*	*1*	*1*	35%
		Unified	1.0588	1.0400	2.000	*22%*
	Category 4	Current	*1*	*1*	1	76.8%
		Unified	1.0333	1.1765	*0.1429*	*35.9%*
Long	Category 3	Current	*1*	*1*	1	*35%*
		Unified	1.1176	1.1200	*1*	47%
	Category 4	Current	*1*	*1*	1	76.8%
		Unified	1.1333	1.3530	*0.2857*	*47.7%*

RelRMSE, RelMAE and RelME stand for Relative Root Mean Square Error, Relative Mean Absolute Error and Relative Mean Error, respectively. These are calculated for the unified solution by taking the results relative to the current solution (=1)

At 4 weeks, the current solution showed better overall performance than the new solution. However, in terms of extreme bias reduction, the new solution showed better performance. Therefore, the results were mixed in this case. For the long horizon, the current solution seems to have an edge over the new solution, at least in terms of forecast accuracy in the highly intermittent demand context. However, the differences seem to be negligible for operational performance, and we can consider the ability to fully automate our solution as a trade-off.

4.2.2 Level of automation

We can fully automate high frequency forecasting through the new solution. Other than point forecasts (which have been used in this research), forecasts can also be created for specific prediction intervals (for e.g. 95% and 80%) if needed. In comparison, the current solution is not fully automated, and needs two steps to go from past sales values to daily forecasted values. From an automation point of view, the new solution is more beneficial for end users.

We calculated the average computation times for each horizon and category combination for the unified model. The computation time includes the time taken to create the model and produce forecasts. We used a system with the following characteristics: Intel® Core™ i5-4300U CPU @ 1.90GHz, 8.00 GB RAM and x64 based processor. In order to perform a more direct comparison, we used the same system to also create weekly forecasts using a standard Holt-Winters function in RStudio (*HoltWinters* function in the *stats* package). We then used another code to split these forecasts pro-rata (to obtain daily forecasts). We followed this process since we were unable to test the case company's actual solution on our system. For ease of comparison, we have provided the results of the unified model relative to the (assumed) current solution (in Table 11). There were no significant differences in timing when it came to different product categories and forecast horizons for both models.

Table 11: Average relative computation times of the unified model

Forecast horizon	Category	Relative computation time of the unified model
Rolling	Category 1	10.89
	Category 2	10.77
	Category 3	9.20
	Category 4	10.94
Long	Category 1	10.88
	Category 2	10.45
	Category 3	9.01
	Category 4	9.55

4.2.2. Transparency and ease of usage

Our proposed solution employs TBATS, which is mostly unknown in the grocery industry. Holt-Winters (used in the current solution) is well established. Due to the addition of components such as ARMA and Box Cox, TBATS is more complex than the current solution. It is also less intuitive, as the aforementioned elements are not as easy to explain as trend, level and seasonality (the basic concepts of Holt-Winters). Increased complexity (which leads to decreased transparency) could create problems in practice, as grocery retailers might show reluctance to employ these methods.

After concluding the exploration of alternative approaches, of which our research was one exploration, the case company decided to introduce a new solution that includes some components similar to those in TBATS. We can assume that this is more or less as complex as TBATS, since any forecasting system utilizing such components will have a decreased level of transparency compared to the current solution. In other words, when obtaining a more efficient, automated forecasting model, a decrease in transparency is expected. Clearly, we can obtain superior results using TBATS as a forecasting model, which is a tradeoff for its increased complexity. As for ease of usage, the ability to automate fully is highly beneficial.

However, we also must consider the next operational steps in the forecasting process. For e.g., special events which can cause spikes in demand, such as promotions and weather, need to be factored in to base forecasts. As of now, it is not possible to add external variables to TBATS models. Therefore, TBATS does not facilitate the automation of the entire forecasting process. A single model which calculates daily forecasts taking also special events or 'exceptions' into account would be highly preferable.

5. Outcome evaluation and theoretical analysis

The outcome evaluation by the case company representatives found the unified model with rolling forecasts to be a good solution for fast moving SKUs. Even in the case of intermittent demand, they do not believe that this

approach is worse than the current solution. The reasoning is that the number of extreme biases has reduced significantly with the unified model, and the differences in other measures such as the RMSE can be considered as negligible. Overall, it is their belief that this approach adds value. The case company representatives also estimated that the current solution should be able to compute 10-100 times faster than the proposed TBATS alternative. This estimation is largely supported by the computation times we obtained. The current solution is more computationally economical, which is expected since it has a much simpler structure.

The insights through our intervention contribute to theory of digitalization of operations. An unexpected aspect we deliberated on was the forecasting model re-training frequency. Both in existing literature and practice, forecasting models are retrained at fixed time intervals, and this can be as frequently as one-step ahead (Ramos and Fildes 2017; Ma et al. 2016). The prevalent process is automatically retraining forecasting models as soon as enough new data is gathered (for e.g. every two weeks), irrespective of the necessity to do so. Per our results, in the fast moving/ high volume demand context, the new solution performs better than the current solution even when forecasts are calculated for 18 weeks. This was surprising for us, since we expected that the current solution would perform better in this case (since the forecasts we received from the case company were calculated every 4 weeks). Switching to a long horizon base forecast using a method like TBATS would save a considerable amount of time, processing power and memory if the forecasts would not have to be calculated frequently as a rolling forecast. Thus, we question the practice of fixed re-training frequency for forecasting models.

The outcome of the intervention is to demonstrate that base forecasts can be calculated accurately for a period exceeding 3 months. It is possible that the forecasting horizon can be even longer. The opportunity for long horizon base forecasts was apparent in particular for fast moving SKUs. Assuming there is a mechanism to account for extraordinary events that significantly affect demand, some SKU-store combinations may never need retraining. This points to the possibility of monitoring the forecast errors continuously, and only re-training the model when errors cross some threshold. Introducing variable, even indefinite re-training frequencies for forecasting models will mean a significant change to the forecasting process. Investigating the change may be worthwhile if monitoring requires significantly less computation than retraining. It would then have the potential to greatly reduce the time and cost associated with high-frequency forecasts. We believe this will be an interesting topic for further research. We propose that switching to a unified approach for daily base forecasts, while reducing retraining, is a potential method of shifting the performance frontier of forecasting operations. Under the right circumstances, such as the fast movers in the intervention, it would enable improved forecasting accuracy without increasing computational cost. What is necessary, then, is being able to detect the need for change and retraining.

As mentioned in section 3, after their research into alternative forecasting models, the case company decided to proceed with *Bayesian regression modeling* for their new solution. According to a company representative, the ability to add an unlimited number of additional components (for exceptions) to the model is a great advantage. Another benefit is that the entire forecasting process can be performed using one model, though the additional components first need to be optimized for different product groups. This requires a deep understanding of the underlying functionality. After the optimization, the forecasting process is automatic. However, its complexity and increased computational time are some identified drawbacks.

Bayesian regression modeling incorporates some components included in TBATS. For e.g., Fourier terms are used to estimate seasonality in this solution (using dummy variables is also possible). The insight is that a key element of our intervention – low frequency retraining – is practically useful when there are no major changes in

demand or supply for the SKU-store combinations in question. When there are changes within the forecasting period, the forecasting model will have to be re-trained using the latest sales data, and relevant external factors that caused the said changes (for e.g. weather, promotions). Here, Bayesian regression modeling is useful since the forecasting model already includes components to handle exceptional events, and if not, these components can be added to the model. The SKUs with changes can be singled out for retraining. Shifting efforts from retraining to monitoring creates a possibility to retrain forecasts responsively to SKUs affected by unanticipated and external factors, while reducing forecasting costs for SKUs unaffected by such change factors. This insight will be beneficial for forecasting grocery retail operations that stock a substantial number of SKUs in very many stores. This is also significant when we consider the future of the grocery industry and the attempts to reduce food waste. This development is leading grocery retailers to implement concepts such as 'grocery happy hours', as described by Segal (2019). In order to have automatic responsive pricing of perishables, retailers will need forecasts hourly or even more frequently. Here our results indicate an opportunity to reduce the massive need for computational resources by adopting a monitoring method for forecast retraining.

Furthermore, we were surprised to find that the unified solution works extraordinarily well for extreme bias reduction. Our results show that the current solution tends to highly over- or under-forecast often in both fast moving and highly intermittent demand contexts. This is worrisome as extreme biases can lead to a variety of operational issues for grocery retailers. The unified solution has the potential to solve this problem by reducing extreme biases, sometimes by more than 50% when compared to the current solution. On the long term, reducing extreme biases has the possibility of producing major savings in the grocery industry through the reduction of out of stocks, as well as wastage due to over-stocking.

An important insight on its own is that it is feasible to use TBATS in the grocery demand context. In other words, it is possible to use a complex, non-conventional model such as TBATS in a comparatively simpler context, albeit one where we produce many more forecasts. This is in contrast to most previous research done using TBATS where three or more seasonality levels were studied (refer Brozyna et al. 2018; De Livera et al. 2011; Grmanová et al. 2016; Lago et al. 2018), and we can do this in a way that gives us superior results in an automatic manner, reducing manual steps. This also means that it should be possible to use TBATS in other similar settings more widely.

6. Conclusion

We conducted an intervention in practice, examining daily forecasting on the level of SKU-store using TBATS as our forecasting model. When moving to high frequency forecasting, our intervention exposes a mechanism to shift the performance trade-off between more frequent forecasts and more computing resources. More frequent forecasts without massive increase in computing can be achieved by retraining less frequently. Thus, based on our study we question the need to retrain forecasting models in a rigid manner (at regular and usually short periods), which is currently done in both literature and practice. We posit that forecasting models should be retrained only when necessary, especially since models such as TBATS can be used to obtain superior results even at long forecasting horizons. In practice, such an approach has the potential to provide significant benefits for retailers who stock large amounts of perishables with short replenishment cycles. It also provides a starting point for applications that utilize automatic daily forecasting in retail, with many opportunities for adaptations and development.

The finding that with TBATS, frequent rolling forecasts are not necessarily beneficial implies a novel contribution to digitalization of operations management: design the operational process so that reliance on high frequency heavy computing is replaced by lighter monitoring and postponed computing. In the examined setting, such monitoring would challenge the current practice of training forecasting models at fixed intervals.

References

Alftan A, Kaipia R, Loikkanen L, Spens K (2015) Centralised grocery supply chain planning: improved exception management. International Journal of Physical Distribution & Logistics Management 45(3):237-259

Ali ÖG, Sayın S, Van Woensel T, Fransoo J (2009) SKU demand forecasting in the presence of promotions. Expert Systems with Applications 36(10):12340-12348

Amine A, Cadenat S (2003) Efficient retailer assortment: a consumer choice evaluation perspective. International Journal of Retail & Distribution Management 31(10):486-497

Anderson Consulting (1996) Where to look for incremental sales gains: The retail problem of out-of-stock merchandise. http://www.ccrrc.org/wp-content/uploads/sites/24/2014/02/Where_to_Look_for_Incremental_Sales_Gains_The_Retail_Problem_of_Out-of-Stock_Merchandise_1996.pdf. Accessed 20 October 2019

Andrawis RR, Atiya AF, El-Shishiny H (2011) Combination of long term and short term forecasts, with application to tourism demand forecasting. International Journal of Forecasting 27(3):870-886

Armstrong JS (2001) Evaluating forecasting methods. In: Armstrong, J.S. (ed) Principles of forecasting: a handbook for researchers and practitioners. Springer Science & Business Media, pp 441-472

Belt T (2017) When is forecast accuracy important in the retail industry? Effect of key product parameters. M.Sc. thesis, Aalto University, Finland

Bishop CM, Tipping ME (2003) Bayesian regression and classification. Nato Science Series sub Series III Computer and Systems Sciences 190:267-288

Box GE, Jenkins GM, Reinsel GC, Ljung GM (2015) Time series analysis: forecasting and control. John Wiley & Sons

Brozyna J, Mentel G, Szetela B, Strielkowski W (2018) Multi-seasonality in the TBATS model using demand for electric energy as a case study. Economic Computation & Economic Cybernetics Studies & Research 52(1)

Bunn DW (2000) Forecasting loads and prices in competitive power markets. Proceedings of the IEEE 2000 88(2):163-169

Chai T, Draxler RR (2014) Root mean square error (RMSE) or mean absolute error (MAE)?– Arguments against avoiding RMSE in the literature. Geoscientific model development 7(3):1247-1250

Chand S, Hsu VN, Sethi S (2002) Forecast, solution, and rolling horizons in operations management problems: A classified bibliography. Manufacturing & Service Operations Management 4(1):25-43

Choi TM, Wallace SW, Wang Y (2018) Big data analytics in operations management. Production and Operations Management 27(10):1868-1883

Christodoulos C, Michalakelis C, Varoutas D (2010) Forecasting with limited data: Combining ARIMA and diffusion models. Technological forecasting and social change 77(4):558-565

Corsten D, Gruen T (2003) Desperately seeking shelf availability: an examination of the extent, the causes, and the efforts to address retail out-of-stocks. International Journal of Retail & Distribution Management 31(12):605-617

Crone SF, Kourentzes N (2011) Segmenting electrical load time series for forecasting? an empirical evaluation of daily UK load patterns. In Proceedings of the 2011 International Joint Conference on Neural Networks

Darbellay GA, Slama M (2000) Forecasting the short-term demand for electricity: Do neural networks stand a better chance? International Journal of Forecasting 16(1):71-83

De Livera AM, Hyndman RJ, Snyder RD (2011) Forecasting time series with complex seasonal patterns using exponential smoothing. Journal of the American Statistical Association 106(496):1513-1527

Denyer D, Tranfield D, van Aken JE (2008) Developing design propositions through research synthesis. Organization Studies 29: 393-413

De Toni AF, Zamolo E (2005) From a traditional replenishment system to vendor-managed inventory: A case study from the household electrical appliances sector. International Journal of Production Economics 96(1):63-79

Ehrenthal JCF, Honhon D, Van Woensel T (2014) Demand seasonality in retail inventory management. European Journal of Operational Research 238(2):527-539

Fernie J, Grant DB (2008) On-shelf availability: The case of a UK grocery retailer. The International Journal of Logistics Management 19(3):293-308

Fildes R, Hibon M, Makridakis S, Meade N (1998) Generalising about univariate forecasting methods: further empirical evidence. International Journal of Forecasting 14(3):339-358

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

García-Martos C, Rodríguez J, Sanchez MJ (2007) Mixed models for short-run forecasting of electricity prices: application for the Spanish market. IEEE Transactions on Power Systems 22(2):544-552

Gardner ES (2006) Exponential smoothing: The state of the art, Part II. International Journal of Forecasting 22(4):637–666

Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB (2013) Bayesian data analysis. Chapman and Hall/CRC

Godfrey GA, Powell WB (2000) Adaptive estimation of daily demands with complex calendar effects for freight transportation. Transportation Research Part B: Methodological 34(6):451-469

Gould PG, Koehler AB, Ord JK, Snyder RD, Hyndman RJ, Vahid-Araghi F (2008) Forecasting time series with multiple seasonal patterns. European Journal of Operational Research 191(1):207-222

Grmanová G, Laurinec P, Rozinajová V, Ezzeddine AB, Lucká M, Lacko P, Návrat P (2016) Incremental ensemble learning for electricity load forecasting. Acta Polytechnica Hungarica 13(2):97-117

Gruber V, Holweg C, Teller C (2016) What a waste! Exploring the human reality of food waste from the store manager's perspective. Journal of Public Policy & Marketing 35(1):3-25

Hansen JV, Nelson RD (2003) Forecasting and recombining time-series components by using neural networks. Journal of the Operational Research Society 54(3):307-317

Holmström J, Ketokivi M, Hameri AP (2009) Bridging practice and theory: A design science approach. Decision Sciences 40(1):65-87

Hyndman RJ, Koehler AB (2009) Another look at measures of forecast accuracy. International Journal of Forecasting 22(4):679-688

Karabiber OA, Xydis G (2019) Electricity Price Forecasting in the Danish Day-Ahead Market Using the TBATS, ANN and ARIMA Methods. Energies 12(5):928

Khotanzad A (2016) Short-term load and price forecasting with artificial neural networks. In: Grigsby L.L. (ed) Power Systems, 3rd edn. CRC Press, pp 255-265

Kiil K, Dreyer HC, Hvolby HH, Chabada L (2018) Sustainable food supply chains: the impact of automatic replenishment in grocery stores. Production Planning & Control 29(2):106-116

Koehler AB, Murphree ES (1988) A comparison of results from state space forecasting with forecasts from the Makridakis competition. International Journal of Forecasting 4(1):45-55

Kotzab H, Teller C (2003) Value-adding partnerships and co-opetition models in the grocery industry. International journal of physical distribution & logistics management 33(3):268-281

Koupriouchina L, van der Rest JP, Schwartz Z (2014) On revenue management and the use of occupancy forecasting error measures. International Journal of Hospitality Management 41:104-114

Kourentzes N, Crone SF (2008) Forecasting high-frequency time series with neural networks - an analysis of modelling challenges from increasing data frequency. The 4th International Conference on Data Mining

Kourentzes N, Petropoulos F, Trapero JR (2014) Improving forecasting by estimating time series structural components across multiple frequencies. International Journal of Forecasting 30(2):291-302

Kourentzes N, Trapero JR, Svetunkov I (2018) Measuring the behaviour of experts on demand forecasting: a complex task http://kourentzes.com/forecasting/wp-content/uploads/2014/12/Kourentzes_Complex-bias.pdf. Accessed 1 October 2019

Lago J, De Ridder F, De Schutter B (2018) Forecasting spot electricity prices: Deep learning approaches and empirical comparison of traditional algorithms. Applied Energy 221:386-405

Lamedica R, Prudenzi A, Sforna M, Caciotta M, Cencellli VO (1996) A neural network based technique for short-term forecasting of anomalous load periods. IEEE Transactions on Power Systems11(4):1749-1756

Ma S, Fildes R, Huang T (2016) Demand forecasting with high dimensional data: The case of SKU retail sales forecasting with intra-and inter-category promotional information. European Journal of Operational Research 249(1):245-257

Narayanan A, Sahin F, Robinson EP (2019) Demand and order-fulfillment planning: The impact of point-of-sale data, retailer orders and distribution center orders on forecast accuracy. Journal of Operations Management 65(5):468-486

Oliva R (2019) Intervention as a research strategy. Journal of Operations Management 65(7): 710-724

Peffers K, Tuunanen T, Rothenberger MA, Chatterjee S (2007) A design science research methodology for information systems research. Journal of management information systems 24(3):45-77

Peffers K, Tuunanen T, Niehaves B (2018) Design science research genres: introduction to the special issue on exemplars and criteria for applicable design science research 129-139

Pereira LN (2016) An introduction to helpful forecasting methods for hotel revenue management. International Journal of Hospitality Management 58:13-23

Peterson H (2018) 4 Ways American Grocery Shopping Is Changing Forever https://www.businessinsider.com/trends-that-are-changing-grocery-stores-2014-4?r=US&IR=T&IR=T. Accessed 4 October 2019

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

Poler R, Mula J (2011) Forecasting model selection through out-of-sample rolling horizon weighted errors. Expert Systems with Applications 38(12):14778-14785

Pulkka A (2020) Applying Bayesian regression to forecast retail demand in the Christmas season. M.Sc. thesis, Aalto University, Finland

Ramos P, Fildes R (2017) Characterizing retail demand with promotional effects. In International Symposium on Forecasting. International Institute of Forecasters Cairns, Australia

Schmenner RW, Swink ML (1998) On theory in operations management. Journal of Operations Management 17(1):97-113

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

Segal D (2019) The world wastes tons of food. A grocery 'happy hour' in one answer. The New York Times. https://www.nytimes.com/2019/09/08/business/food-waste-climate-change.html. Accessed 3 February 2020

Sillanpää V, Liesiö J (2018) Forecasting replenishment orders in retail: value of modelling low and intermittent consumer demand with distributions. International Journal of Production Research 56(12):4168-4185

Snyder RD, Ord JK, Beaumont A (2012) Forecasting the intermittent demand for slow-moving inventories: A modelling approach. International Journal of Forecasting 28(2):485-496

Syntetos AA, Boylan JE (2001) On the bias of intermittent demand estimates. International Journal of Production Economics 71(1-3):457-466

Taylor JW (2003) Short-term electricity demand forecasting using double seasonal exponential smoothing. Journal of the Operational Research Society 54(8):799-805

Taylor JW, De Menezes LM, McSharry PE (2006) A comparison of univariate methods for forecasting electricity demand up to a day ahead. International Journal of Forecasting 22(1):1-16

Taylor JW (2007) Forecasting daily supermarket sales using exponentially weighted quantile regression. European Journal of Operational Research 178(1):154-167

Taylor JW (2010) Triple seasonal methods for short-term electricity demand forecasting. European Journal of Operational Research 204(1):139-152

Taylor JW (2011) Multi-item sales forecasting with total and split exponential smoothing. Journal of the Operational Research Society 62(3):555-563

Teunter R, Sani B (2009) On the bias of Croston's forecasting method. European Journal of Operational Research 194(1):177-183

Van Donselaar K, van Woensel T, Broekmeulen RACM, Fransoo J (2006) Inventory control of perishables in supermarkets. International Journal of Production Economics 104(2):462-472

Veit A, Goebel C, Tidke R, Doblander C, Jacobsen HA (2014) Household electricity demand forecasting: benchmarking state-of-the-art methods. In Proceedings of the 5th international conference on Future energy systems pp 233-234

Whiteoak P (2004) Rethinking efficient replenishment in the grocery sector. In: Fernie, J. & Sparks, L. (ed) Logistics and Retail Management, 2nd edn. London & Sterling, pp 138-163

Willemain TR, Smart CN, Schwarz HF (2004) A new approach to forecasting intermittent demand for service parts inventories. International Journal of forecasting 20(3):375-387

Winters PR (1960) Forecasting sales by exponentially weighted moving averages. Management Science 6(3):324-342

Yokuma JT, Armstrong JS (1995) Beyond accuracy: Comparison of criteria used to select forecasting methods. International Journal of Forecasting 11(4):591-597

Zhou SL, McMahon TA, Walton A, Lewis J (2000) Forecasting daily urban water demand: a case study of Melbourne. Journal of hydrology 236(3-4):153-164

Appendix 1

BATS and TBATS models

De Livera et al. (2011) introduced both BATS and TBATS models. BATS cannot handle non-integer seasonality, which is why TBATS was created. TBATS uses Fourier terms to represent seasonal components in a trigonometric manner, thus allowing for non-integer seasonality. The components of both models are explained in Table 1.

Table 1: Description BATS and TBATS model components

	BATS	TBATS
T (Trigonometric)	Not included	• Uses Fourier terms to represent seasonal components, like methods such as dynamic harmonic regression (Hyndman and Athanasopoulos, n.d.) • Beneficial when seasonal period is large since it reduces the number of parameters, especially with daily data (annual seasonality = 365.25) • Seasonal patterns not fixed; they can change over time (unlike with dynamic harmonic regression) $$\quad (1)$$ $$\quad (2)$$ $$\quad (3)$$
B (Box-Cox transform)	• Allows for non-linearity and transforms model to make it more normally distributed (De Livera et al. 2011). Helpful when dealing with certain weaknesses, for e.g. instability of the model due to infinite variances beyond some forecasting horizons • Value can range from 0 to 1, where 1 means that a strong transformation is not needed to handle non-linearity (if any) • Limits TBATS models to positive time series, but most time series in practice have only positive values (and zeros). This is especially not a limitation in grocery $$\quad (4)$$	
A (ARMA) – Autoregressive Moving Average	• Used to increase forecast accuracy by modeling error explicitly (refer Taylor 2003) • Most suitable orders for autoregressive and moving average parts selected using the Akaike Information Criterion (estimator of the quality of a model for a given dataset) • When both AR and MA orders = 0, error is taken to be random $$= \sum \quad + \sum \quad + \quad (5)$$	
T (Trend)	• Trend dampened to avoid overly optimistic/ pessimistic forecasts (method by Gardner and McKenzie 1985 is used) • Long run trend (b) calculated so that future forecasted values will converge to this trend (otherwise due to the damped trend future values can converge to zero) • Damping factor ranges from 0 to 1, where 1 = no damping effect needed $$= \quad + \quad + \quad (6)$$ $$= (1 - \quad) + \quad + \quad (7)$$	
S (Seasonality)	• Allows for multiple, non-nested seasonal patterns $$\quad (8)$$	Uses the trigonometric seasonal formulation specified in the 'T (Trigonometric)' row
Other details	• Can accommodate multiple seasonal patterns, but is unable to function with non-integer seasonality • Many initial seasonal states must also be estimated when seasonal period is large, reducing computational efficiency and increasing required memory	• Has to estimate $2(\quad)$ initial seasonal states (likely to need less estimation than a BATS model) • Less harmonics will be required by most seasonal components, which reduces the number of parameters (faster and more efficient in terms of memory in comparison to BATS)

	• Supplemented with arguments („„,1, 2,…..,). For e.g., the Holt-Winters additive single seasonal method is (1,1,0,0, 1), with no Box-Cox transform, damping effect or ARMA error, and one seasonal period (seasonal cycle = 1) $^{(\ )} = {}_{-1} + {}_{-1} + \sum_{=1} {}^{(\ )}_{-} +$ (9)	• Supplemented with arguments (, , , ,{ 1, 1},{ 2, 2},.,{ , }) with the number of Fourier coefficients specified for each seasonal period (refer De Livera et al., 2011 for the estimation process for initial states, methods of model selection etc.) $^{(\ )} = {}_{-1} + {}_{-1} + \sum_{=1} {}^{(\ )}_{-1} +$ (10)

The forecast or measurement is given by . 1 … … denote all the seasonal periods, is the long run trend, and are the short run trend and level at time respectively, is the ARMA process whose orders and parameters are signified by and (for AR) and and (for MA) respectively, and is the random error. The smoothing parameters are given by , and for $= 1, \ldots\ldots,$ *, and the damping factor is given by . denotes the number of harmonics required for the* $^{\bar{A}h}$ *seasonal component, the smoothing parameters are given by* $_1$ *and* $_2$*, and* $= \frac{2\bar{A}\bar{A}}{}$*. The stochastic level of the* $^{\bar{A}h}$ *seasonal component and the change in the seasonal component over time (the stochastic growth in the level of the* $^{\bar{A}h}$ *seasonal component) are obtained using* $^{(\ )}_{,}$ *and* $^{*(\ )}_{,}$*, respectively.*

Comparison of approaches for high frequency forecasting

Table 2: Summary of the comparison between selected high frequency forecasting models

	Accuracy	Complexity	Non-integer/ large frequencies	References
ARIMA / ARMA with multiple seasonality	Useful when dominated by short term correlation but not when dominated by seasonality and trend	Complex theory, not easy to understand by laymen	Not possible practically, designed to accommodate smaller integer frequencies	Gould et al., 2008; Taylor, 2003
Double exponential smoothing Holt-Winters	Performs better than ARIMA. Using heuristics to estimate seed states decreases accuracy	Easiest to understand	Not possible, generally designed to accommodate smaller integer frequencies	De Livera et al., 2011; Gardner, 2006; Hyndman et al., 2008; Taylor, 2003; Taylor et al., 2006; Taylor, 2010
TBATS	Good performance for complex seasonality. Robust in different forecasting horizons. Using least squares to estimate seed states increases accuracy. Able to model errors automatically	Rather complex theory, not easy to understand by laymen	Possible, designed especially to handle complex seasonality	De Livera et al., 2011; Grmanová et al., 2016; Pereira, 2016; Veit, 2014
Bayesian Regression Modeling	Possible to reduce overfitting. Additional components can increase accuracy	Complex theory, is intensive to compute estimators	Possible if user chooses to utilize Fourier terms	Bishop, C. M., and Tipping, M. E., 2003; Gelman et al., 2013; Pulkka, 2020

Appendix 2

Evaluation criteria

The case company noted the importance of highlighting measures that show that a forecasting model(s) has reduced large forecast errors/ biases significantly when compared to another model, as this is crucial in practice. Due to the occurrence of zero sales, we disregard scale independent measures as they become undefined in the presence of zeros. We also calculate the forecast bias separately because even though some amount of bias is permissible, it is necessary to escalate extreme biases to management to be dealt with manually. We use four measures overall (explained in detail in Table 3):

- Root Mean Square Error (RMSE) and Mean Absolute Error (MAE): for forecast accuracy
- Mean Error (ME) and bias coefficient: for forecast bias

RMSE, MAE and ME are well established in literature while the bias coefficient was proposed recently by Kourentzes et al. (2014). We use the bias coefficient to calculate the number of SKU-stores showing extremely high biases (the value of 'extreme bias' can be decided based on the industry and products in question). Since they most highlight/ penalize large errors, we pay special consideration to the *RMSE* as the primary measure of forecast accuracy and *extremely high bias* as the primary measure of forecast bias.

Table 3: Measures of forecast accuracy and bias

Criteria	Measure	Formulae	Advantages	Disadvantages	References
Forecast accuracy	RMSE	$= \sqrt{\frac{\sum_{\tilde{A}=\exists}(\quad -\quad)^2}{}}$ (11)	• Historically popular/ accepted in statistical modeling • Same scale as data • Useful when large errors are undesirable	• Hard to explain in layman's terms • More sensitive to outliers than MAE	Chai and Draxler, 2014; Hyndman and Koehler, 2006; Kouprioucchina et al., 2014
	MAE	$= \frac{\sum_{\tilde{A}=\exists}\lvert \quad - \quad \rvert}{}$ (12)	• Easier to interpret than RMSE • Less sensitive to outliers	• Gives equal weight to all errors • Absolute values undesirable in calculations	Chai and Draxler, 2014; Hyndman and Koehler, 2006; Kouprioucchina et al., 2014
Forecast bias	ME	$= \frac{\sum_{\tilde{A}=\exists}(\quad -\quad)}{}$ (13)	• Useful to indicate systematic over- or under-forecasting	• Cancels out positive and negative errors • Unbounded, not easy to interpret • Size of errors can be lost	Kouprioucchina et al., 2014
	Bias coefficient	All formulae found in Kourentzes et al., 2014 When bias coefficient >	• Easy to compare, interpret and present results	Not established in academia	Kourentzes et al., 2014

		0.5 => 'extremely negative bias' When bias coefficient < -0.5 => 'extremely positive bias'	due to range of -1 to 1 • Unit free • Retains size of errors		

RMSE, MAE and ME stand for Root Mean Square Error, Mean Absolute Error and Mean Error, respectively, and f = forecast, y = actual amount, n = the number of forecasts

Aggregation and presentation of results

It is necessary to aggregate forecast errors for ease of evaluation and comparison between solutions. We chose the entire forecast period (126 days) and product category as aggregation levels for all four measures.

Some additional steps are also needed for the RMSE, MAE and ME. Since different products can have vastly different sales volumes even within the same product category, these forecast errors should be weighted based on product volume. We also present these results relative to the current solution for increased readability (since the original numbers are quite small). This process is summarized in Table 4.

Table 4: Methods of aggregation for forecast measures

	Measure	Used relative to current solution results	Volume weighted	Aggregated across total forecast period	Aggregated across product category
Forecast accuracy	RMSE	Yes	Yes	Yes	Yes
	MAE	Yes	Yes	Yes	Yes
Forecast bias	ME	Yes	Yes	Yes	Yes
	Number of extremely high bias (either positive or negative)	No	No	Yes	Yes

RMSE, MAE and ME stand for Root Mean Square Error, Mean Absolute Error and Mean Error, respectively

Level of automation

Automation eliminates user involvement and thus reduces effort. However, computational efficiency is likely to decrease when complex methods are fully automated. Therefore, the feasibility of automating such methods and whether other aspects such as increased accuracy can be used as tradeoffs have to be discussed before they are used practically. When measuring the performance of the new solution, we calculated the time taken to create the forecasting model(s) and forecast as a measure of computational efficiency. These numbers were discussed with case company representatives to understand their practical relevance. We also calculated the time to forecast using an assumed model (in place of the case company's current solution) in order to perform a more direct comparison.

Transparency and ease of usage

Forecasting models are not used by researchers or statisticians in business contexts, but usually by nonprofessionals in the fields of statistics and numerical modeling. Therefore, the new solution must ideally be intuitive and easy to use. We discussed and compared the new solution's complexity with the current solution, with feedback from the case company. It was also necessary to consider if there are justifiable tradeoffs for decreased transparency, which we reasonably expect with a more intricate forecasting model such as TBATS.

Appendix 3

Monthly box plots

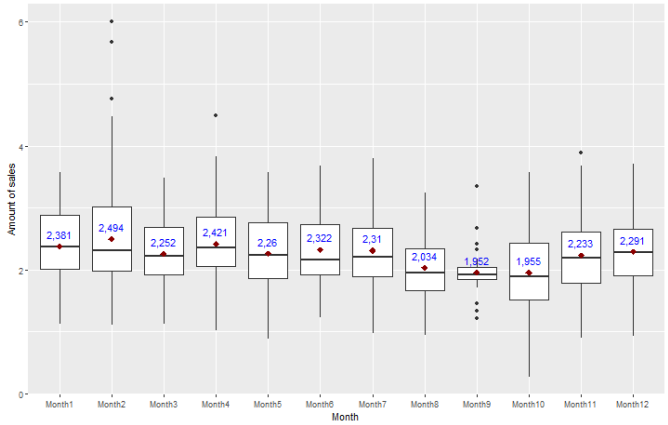

Figure 3: Category 1 product sales per month

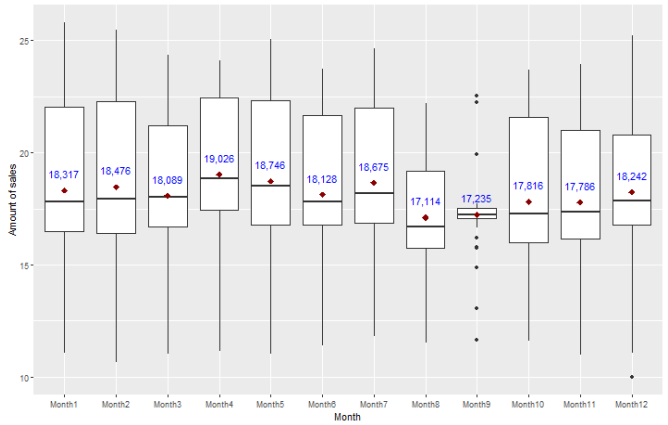

Figure 4: Category 2 product sales per month

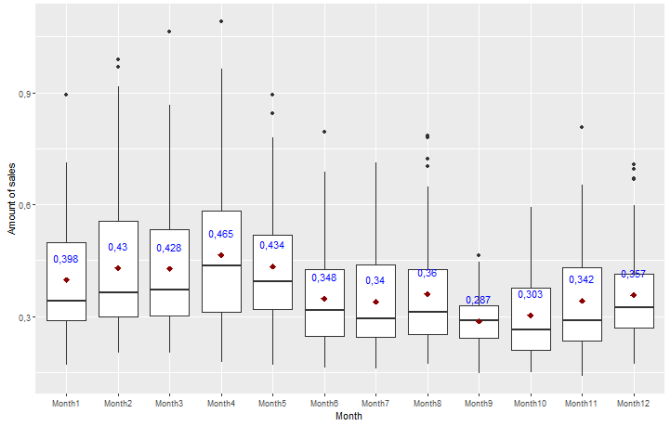

Figure 5: Category 3 product sales per month

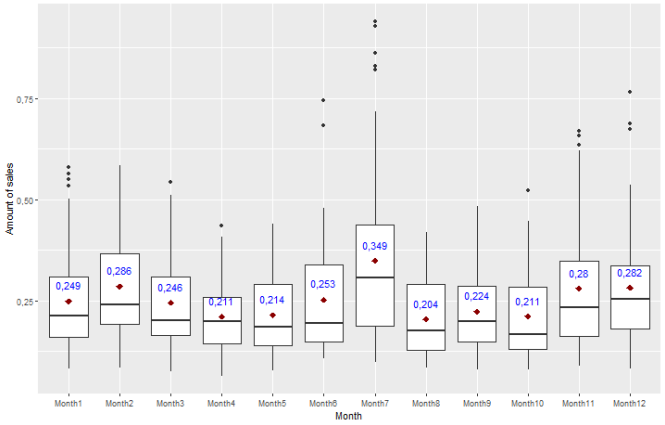

Figure 6: Category 4 product sales per month

Daily box plots

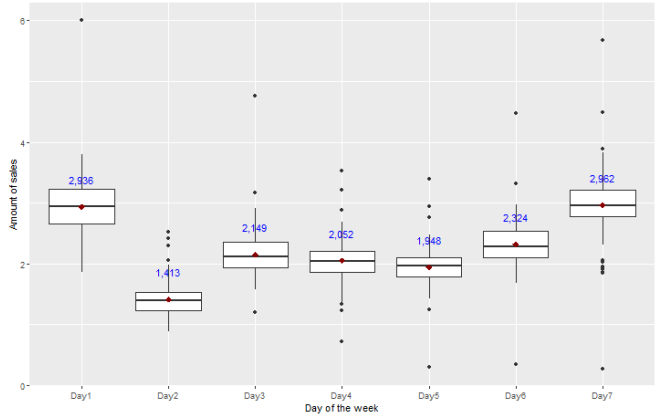

Figure 7: Category 1 product sales per day

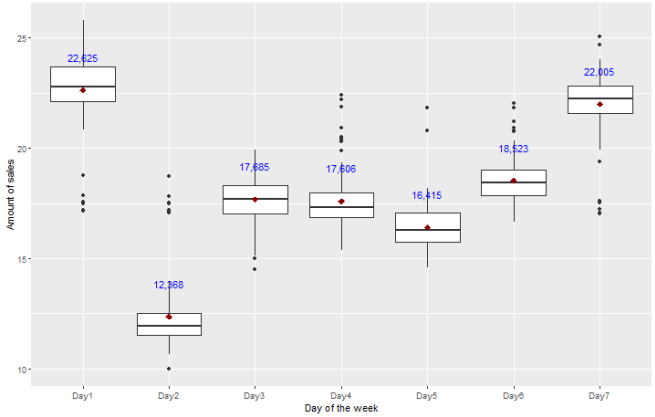

Figure 8: Category 2 product sales per day

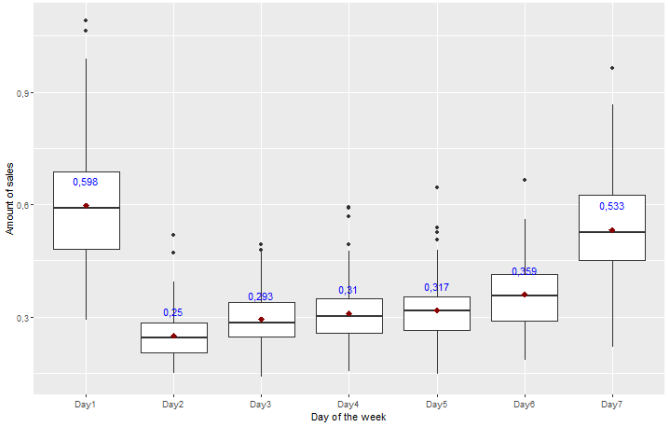

Figure 9: Category 3 product sales per day

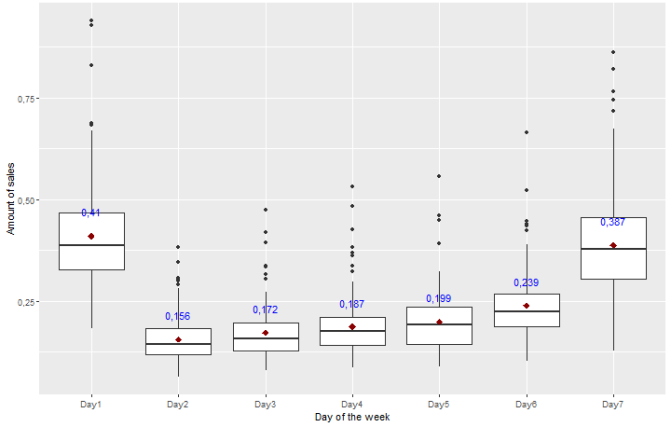

Figure 10: Category 4 product sales per day

Appendix 4

Practicalities in using the intervention

Since the new solution is fully automatic, a user simply has to have the sales data in a format that can be read by RStudio. Data can be imported into RStudio in a variety of formats including csv, xls and xlsx. This solution currently cannot be used for new product forecasting; therefore, it is recommended to have at least a few weeks of sales data to train the TBATS model.

The user can change certain parameters easily, such as the frequencies considered when training the TBATS model(s), training period for the TBATS model(s), the number of forecasts produced, whether a single or multiple SKUs are forecasted etc. Currently, the unified and segmented approaches are run separately (the user can choose which approach to use). However, this can easily be modified so that the approach is chosen automatically based on some criteria (for e.g. the product category). Rolling forecasts and long forecasts are also separately calculated based on the user's choice, but this too can be modified in a similar manner.

It is also possible to modify the technical parameters of the TBATS model, such as:

- Whether or not to use the Box-Cox transform, trend, damping parameter for the trend and ARMA errors
- The lower and upper limits for the Box-Cox transform
- The values (p and q) for ARMA errors

However, this is generally unnecessary since all such possible scenarios are automatically tried out and the best options are chosen to train the TBATS model. Modifying these parameters could be useful, for e.g. if it is clear that the time series in question does not need any transform because it is linear (Box-Cox transform is then set to NULL). Setting unnecessary parameters to NULL makes the forecasting process faster, however this requires the user to have expertise in this area.

The produced forecasts are automatically exported to an xlsx file (this can be changed if the user prefers some other format). This xlsx file shows both the dates and corresponding forecasts for the chosen forecast horizon. If multiple SKUs are forecasted at the same time, these results will be displayed in the same file (in separate columns or sheets, as can be chosen by the user).