Topics in Retail Forecasting

Daniel Waller

Submitted for the degree of Doctor of Philosophy at Lancaster University.

September 2019

Abstract

Retail forecasting is a diverse and dynamic research area encompassing a variety of different topics. The advent of online channels, the increasing complexity of product ranges, and the shortening lifespan of many items are as examples of some of the new challenges that maintain the importance of improving forecasting in this domain. This thesis aims to address questions in retail forecasting that are closely linked with relevant problems faced in the industry. As such, the problems have been identified through a combination of reviewing the academic literature, discussion, and engagement with practitioners.

This thesis starts by considering the situation where demand series are influenced by multiple seasonal and calendar effects. This is a challenge which is widespread due to high frequency sampling and decision making in retailing. We develop a new model to accommodate flexibility in modelling complex seasonal patterns, which also aids with mitigating the effect of short demand histories on forecasting performance. The new model is embedded in an innovations state-space formulation and it is demonstrated empirically using wholesale food data to provide competitive forecasting accuracy to established benchmarks.

Next, the dual problems of SKU-level model parameter estimation and forecasting are considered. For retailers experiencing frequent promotional activities, this is a principal issue. The parameter estimates provide insights about the elasticity of different factors on demand for the SKU, and therefore inform marketing planning. Accurate forecasts, for both promotional and baseline periods, support other functions such as replenishment and inventory management. First, a geometric parameter inheritance procedure is proposed, which uses aggregate information within a product hierarchy to improve parameter estimates under certain assumptions. At brand level, it is typically easier to better estimate elasticity effects, making this strategy preferable. Second, a debiasing approximation is derived for the forecasting procedure, which is demonstrated to reduce bias, whilst remaining competitive in terms of forecast accuracy, as shown in a simulation study. The debiasing approximation is then evaluated with an inventory simulation study, which examines the conditions under which improvements in inventory performance can be gained. The conclusions give useful insights for inventory managers, and demonstrate that bias is a significant factor in inventory performance.

Acknowledgements

Firstly, I gratefully acknowledge the financial support of the EPSRC, through the STOR-i Centre for Doctoral Training, and of Aimia, in co-sponsoring this work. I'd like to thank all those at Aimia that have helped shape this research.

I'd like to thank the directors of STOR-i, and the administrative staff, along with anyone else who has been involved in running the centre, for their tireless efforts behind the scenes in making STOR-i what it is. It's hard to imagine a better place to do a PhD. I would also like to thank everyone in the community of STOR-i students that has come and gone during my time here for your friendship. The people at STOR-i really are the best there are.

I would like to give great thanks to both of my supervisors, John Boylan and Nikolaos Kourentzes, for all the time and effort they have spent to guide me in this project. It would have been simply impossible without their consistent, invaluable advice, patience and encouragement. I also thank the members, past and present, of the Centre for Forecasting, who have always been generous in sharing their knowledge.

Last but not least, I would like to thank my family, and my girlfriend Anne, for their love and support throughout this time.

Declaration

I declare that the work in this thesis has been done by myself and has not been submitted elsewhere for the award of any other degree.

Daniel Waller

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The quality of retail forecasting has been widely demonstrated to have a significant impact on retailers' performance. Both academic researchers and forecasting practitioners have directed a lot of attention to improving and developing forecasting models for many years (Fildes et al., 2019); it continues to be a crucial research focus, considering the changing environment in the industry today. In a recent article, Seaman (2018) presents some of the considerations that retail forecasters should take into account and how forecasts can be used differently, depending on the objectives. One of the important challenges touched upon is forecasting for highly seasonal and promotional items, particularly around the holiday period. Commenting on this article, Boylan (2018) provides more background on recent trends. A surge in competition, as well as changes in the mix of store types and a shift towards offering more diverse product ranges have all pushed retailers towards data analytics as a means to gain an advantage over rival firms. Boylan (2018) also cites how shorter supplier lead times have led to retail data being captured in shorter time buckets than ever before.

This increased granularity of data has revealed seasonal patterns that were previously unobservable due to their high frequency, leading directly to a growing importance of modelling complex seasonality. Furthermore, a proliferation of data, partly influenced by the rise of loyalty programmes, has placed an emphasis back on extracting signals from noisy, disaggregate information. The dual problems of parameter estimation and forecasting are important challenges for this data type, particularly in the context of promotional modelling; they influence many retail functions, including inventory replenishment and promotional strategy.

In this thesis, we are motivated to provide solutions to real issues in retail forecasting that are faced by practitioners. In particular, part of the motivation for the thesis is from co-sponsorship by Aimia, a loyalty analytics company. In dealing with retail data from a wide range of different sources, the company is well-placed to recognise the problems that are faced within the sector, and many of the problems that motivate this research have come from discussions with them and other practitioners in the area. Additionally, real-world data that has been shared by the company have been invaluable in helping to provide both intuition and motivation for the work. Chapter 2 sees the most direct example of this, where various methods for multi-seasonal forecasting are applied on one of these datasets in an empirical study; nonetheless, insights gained from the real datasets have influenced all chapters. Meanwhile, this work has also aimed to contribute to various different areas in the academic literature. As the contents and structure of the thesis is laid out in the following paragraphs, we try to describe the process by which a practical problem was taken as motivation and then abstracted into a research question that addresses gaps in the academic literature in each case.

In Chapter 2, the motivation to study the issue of multiple seasonality was formed through both an appraisal of demand series from real retail datasets from varying sources, and also through conversations with forecasting practitioners, who identified the interaction between special events and other more regular calendar-based seasonalities as a difficult phenomenon to forecast. To examine this issue in an academic setting, the general methodology of dealing with seasonalities in time series forecasting was reviewed. It was found that, whilst recent work (eg. Taylor and Snyder, 2012; De Livera et al., 2011) proposed methods to reduce the number of parameters needing to be estimated, no approach specifically accounted for the short demand histories typical of retail, whilst still accommodating flexible modelling of seasonalities and the interaction between them. A new model is developed, the mixed parsimonious model, which addresses this problem, and an empirical evaluation, carried out on real-world data from a food wholesaler, yields promising results for the proposed model.

Having overcome modelling challenges associated with multiple seasonal periodicities in retail data, the focus turns to modelling promotional events, which is another major complication for forecasting in the sector. In Chapter 3, the initial problem identified is the estimation of the typically-seen-in-practice loglinear regression-model parameters, for forecasting at the stock keeping unit (SKU) level. This is a problem for practitioners because SKU-level data series often have short histories and are very noisy, resulting in parameter estimates that are inaccurate, which has knock-on effects for both forecasting and other analytics carried out by the retailer; alternatively, insights into products can be gained by clustering SKUs within a subcategory eg. brand or target market. By contrast, parameter estimation and forecasting at higher levels of the retail hierarchy is much easier; thus, we are motivated to explore using hierarchical information at higher levels of aggregation to alleviate problems at lower levels. We found that potential gaps existed to exploit the abundance of SKU-level data available in recent times to better use hierarchical information, with much research taking place before such data was available (eg. Christen et al., 1997). Furthermore, the bias in the forecasting procedure had not been fully explored eg. Miller (1984). Two methodological developments were made: (i) a geometric parameter inheritance scheme to estimate a common parameter at an aggregate level without incurring bias, and (ii) a forecast debiasing approximation which corrects for bias incurred in the parameter estimation process. The performance of both of these developments, along with the combination of the two, is examined through a simulation study.

In Chapter 4, the impact of the forecast approximation developed in the previous chapter on inventory management is examined. Linking improvements in forecasting procedure through to their consequences in stock and service levels is important for practitioners (Gardner, 1990); inventory performance metrics are often much closer to the real factors that influence operational decisions than more abstract forecast accuracy measures (Kourentzes, 2013). We found this relationship to be important, and underdiscussed in the academic literature despite the earlier references. Furthermore, issues around bridging promotional modelling and inventory management, such as calculating safety stock for promotional periods, are also gaps needing further research. This chapter contains details of a simulation study, which demonstrates the performance of our approach under a variety of conditions. We also examine the relationship between different properties of the forecasts, such as bias, accuracy and variance, and dimensions of inventory performance such as average on-hand inventory and service level, to further understand the drivers of inventory performance.

We summarise the findings of the thesis in Chapter 5, outlining the main contributions that have been made. The implications for practitioners are discussed, along with comments on the limitations of the work, and then we expand on some possible directions for future research.

Before presenting the main body of the thesis, we summarise the contributions. The first contribution is a new method for modelling complex seasonal patterns, embedded in an innovations state-space model. The new method is found through an empirical study to improve forecasting accuracy over existing methods in all periods, including where calendar effects are present, and for cumulative forecast horizons. The second contribution is a geometric parameter inheritance scheme for estimating SKU-level parameters, which avoids bias normally present when estimating parameters at a higher aggregation level. The procedure is examined through simulations and found to improve the accuracy of parameter estimates over individual estimation, under the assumption that the SKUs share a parameter in common. Third, an approximation for debiasing forecasts is derived and found through simulations to significantly reduce bias, whilst yielding similar accuracy compared with existing forecast approximations. A combined method implementing geometric parameter inheritance and the debiasing approximation is also demonstrated to reduce bias in the forecasts in return for a modest increase in forecast mean squared error. Lastly, the debiasing approximation is shown through simulations to reduce holding costs, albeit at expense of a slightly reduced service level, in an inventory management system. Examining the forecast properties in terms of bias, accuracy and variance, the situations in which stock holding costs are decreased the most are found to be somewhat associated with those where the bias is reduced most.

Chapter 2

Multiple seasonality in retail

Abstract

In retailing, there are often time series where the value in a period is influenced by more than one seasonal effect. A typical example is that of a daily time series of sales, fluctuating due to both the weekly and annual patterns, in addition to any calendar effects. These complex seasonal patterns are more common with the increasing granularity of data, and present a challenge to forecast. In this paper, we examine how multiple seasonal forecasting methods mainly originating from the short-term energy forecasting literature, can be implemented in the retail domain. The features of retail data are discussed, and short data histories are found to be particularly problematic for existing forecasting methods. To address this issue, a novel approach is developed for modelling multiple seasonality. The new method, embedded in an innovations state-space framework, is based on mixing two alternative representations of seasonality to make best use of the limited data. It is tested against current methods in an empirical study, using data from the food wholesale sector, and the results show that our approach outperforms other methods in most cases. We discuss reasons behind the observed improvements, and suggest how the logic behind our model might be extended to a more general setting.

2.1 Introduction

Seasonal variation in time series is a common and wide-ranging phenomenon which has been extensively studied in many guises. Exponential smoothing forecasting models have treated times series as a composition of three components: trend, seasonality and error (Hyndman et al., 2008); these models are studied and adapted to cope with new challenges in measuring seasonal demand in retail by using aggregate sales information to estimate seasonal indices (Dekker et al., 2004) or by using regressors for calendar effects to facilitate temporal aggregation (Kourentzes and Petropoulos, 2016). ARIMA models are built by considering autocorrelations, where the seasonal variant contains additional components of this type associated with seasonal differences (Box et al., 2015). Multivariate time series methods, such as regression, often encode seasonality via a set of dummy variables, while machine learning approaches typically follow one of these approaches (Barrow and Kourentzes, 2018). These approaches are widespread in both research and practice.

The seasonality of a series is often defined as a predictable, periodic variation in the series mean (Ord et al., 2017). Implicit in this definition is the idea of a single seasonality, repeating over time. The development of models able to account for multiple seasonalities is much newer (Taylor, 2003).

With the availability of more granular sales data, modelling multiple seasonality and calendar effects is a problem which is becoming more important for the retail sector. Many product sales vary by weekly and annual patterns of seasonality; calendar effects such as Christmas also have a strong influence. Understanding these influences is vital for producing accurate forecasts, and subsequently for making inventory management decisions. (Ramos et al., 2015)

So far in the academic literature, there is a significant gap in the application of multiple seasonal forecasting methods in retailing contexts. The multiple seasonality literature is motivated by applications in sectors such as short-term energy/utility forecasting and call centre forecasting, among others. One of the contributions of this paper is to consider this body of methods in the context of retail, and demonstrate their benefits and drawbacks.

It is obvious that retail forecasting throws up other hurdles that are not present in other domains. One significant issue is that data histories in retail are typically much shorter, often lasting two to three years, or even less; the unavailability of long histories can either be down to a lack of storage capability or the lack of a process to store data for that length of time. This places limitations on the complexity of the model that can be estimated, which in turn limits the applicability of many existing methods. Another issue is the interaction of different seasonal effects; for example, the pattern of spending around Christmas differs according to which day of the week Christmas Day itself falls on. The contribution of this paper is a new model which is an adaptation of previous models to address theses hurdles.

The rest of this paper is set out as follows. Section 2 comprises a literature review on multi-seasonal forecasting, examining methods from a retail perspective. Section 3 describes the proposed model, which we term the *mixed parsimonious* model. Section 4 sets out two empirical studies, in different areas of retail forecasting, that assess how the methods can be implemented in this setting and assess their performance. Section 5 concludes and hints at future research.

2.2 Literature review

2.2.1 Types of seasonality

We clarify a few definitions related to seasonality for ease of future reading. First, seasonality may be thought of as being either deterministic or stochastic. Deterministic seasonality is defined as behaviour where the unconditional mean of the process varies throughout the period, but the seasonal profile is stable over time (Ghysels and Osborn, 2001). For instance, a deterministic representation of a series y_t might look like:

$$y_t = \sum_{s=1}^{S} z_s \delta_{st} + \varepsilon_t \quad , \tag{2.2.1}$$

where z_s is the conditional mean for season s, δ_{st} are dummy variables and ε_t is a weakly stationary, zero-mean, IID stochastic process. Stochastic seasonality describes a process where the seasonal shape depends on previous disturbance values, therefore allowing the seasonal profile to vary over time. For example, in a stochastic seasonal AR(1) process, the seasonal indices are defined as:

$$z_{s,t} = \theta z_{s,t-1} + \varepsilon_{s,t} \quad , \tag{2.2.2}$$

where t is the current season. Note that the stochastic seasonal process requires s initial seasonal values $z_{s,0}$, where s is the length of the seasonality.

The second dichotomy we present is in the representation of deterministic seasonality; both a dummy variable representation and trigonometric representation are possible. The dummy variable representation is shown in (2.2.1); the trigonometric one is

$$y_t = \mu + \sum_{k=1}^{S/2} \left[\alpha_k \cos\left(\frac{2\pi kt}{S}\right) + \beta_k \sin\left(\frac{2\pi kt}{S}\right)\right] + \varepsilon_t \quad , \tag{2.2.3}$$

where α_k and β_k are coefficients. The two representations can be used interchangeably (Ghysels and Osborn, 2001).

2.2.2 Multiple seasonality

Approaches to forecasting with multiple seasonalities can broadly be classified into four approaches: exponential smoothing; seasonal ARIMA; regression; and machine learning. All these approaches share many common features and typically rely on one, or more, of the definitions provided in the above section. In this paper we focus on exponential smoothing based approaches, but we first review the four approaches in the literature and compare benefits and drawbacks.

Starting with machine learning, we find that the most common approach by far to forecasting multi-seasonal data has been with neural networks. We assess that the general performance of these in past studies has been extremely mixed, but there is evidence that, if best practices are adopted, they can produce highly accurate forecasts. Crone and Kourentzes (2010) and Kourentzes et al. (2014) discuss various issues in the training and specification of neural networks, as well as remedies for common problems. The poor neural network performance reported by Taylor (2010b) and Taylor and Snyder (2012) can be partly attributed to not following many of the practices outlined there. This can explain the stark contrast to the results of the literature review by Hippert et al. (2001) that finds neural networks to be particularly suited to electricity load forecasting. Note that in the electricity load forecasting literature it is very common to separate a multiple seasonal time series into multiple time series, to reduce the number of seasonalities. For example, one could construct seven separate time series, one for each day of the week and model only the annual seasonality, instead of modelling simultaneously both seasonalities, as the methods reviewed here do. Hippert et al. (2001) report that this is very common practice. However, Crone and Kourentzes (2011) evaluate this practice and find it to be always inferior to modelling the original time series directly.

Barrow and Kourentzes (2018) evaluate multi-seasonal exponential smoothing, ARIMA, and neural networks in forecasting call centre demand. Single seasonal versions of the same models, as well as other statistical models such as seasonal moving average are also considered. Interestingly, the authors find that ARIMA models that focus only on the single longer seasonal cycle are substantially more accurate than double seasonal ARIMA models, and comparable to double seasonal exponential smoothing models; this is attributed to the ease of specifying single seasonal ARIMA models. Other studies have reported the superiority of double seasonal Holt-Winters (DSHW) over ARIMA on high frequency datasets; Taylor and McSharry (2007) forecast half-hourly electricity data, and Taylor (2008) examines minute-by-minute observations;

in both cases DSHW is shown to outperform double-seasonal ARMA models specified by following the Box-Jenkins methodology. Notably, the simplistic seasonal moving average method, when tuned to capture the longest seasonal cycle, is not substantially worse than the more complex ARIMA and exponential smoothing models, echoing the results by Barrow (2016). Finally, Barrow and Kourentzes (2018) find neural networks to outperform all statistical contenders. They use a very parsimonious trigonometric encoding of seasonality that is feasible only due to the nonlinear nature of neural networks. Nonetheless, neural networks require a substantial training sample that is often not available for retail time series. Moreover, the computational cost of neural networks can make them prohibitively slow (or equivalently expensive) to use in retailing, due to the large number of forecasts required. Furthermore, quantities such as price elasticities and promotional uplifts are often important for retailers to know, alongside the forecasts; neural networks cannot provide any interpretable information on these quantities.

Another family of techniques that lend themselves to modelling seasonality is wavelets. For instance, Pindoriya et al. (2008) uses an adaptive wavelet neural network for short term price forecasting in the electricity market, and wavelet methods have also been applied to analysing periodic behvaiour of radon concentration within soil (Siino et al., 2019). Much remains to be explored in this area.

A recent example of regression for multi-seasonal data is Trapero et al. (2015), where the objective is to forecast solar irradiance time series of hourly granularity for horizons of up to a day ahead. The authors evaluate dynamic harmonic regression, which is harmonic regression with time-varying coefficients. The motivation for using this model was that the shape of daily solar irradiance varies across the year, dependent on daylight hours. This is not a principal problem for retailing; therefore, we do not consider this model further. In the retail domain, Arunraj and Ahrens (2015) develop a seasonal ARIMA with explanatory variables (SARIMAX) model for daily food sales, where covariates were used to incorporate additional seasonal effects on top of the day-of-week effect, such as month of year and calendar effects. The addition of the seasonally-related explanatory variables was found to reduce out-of-sample forecast errors from those of the single-seasonal SARIMA model.

2.2.3 Exponential smoothing based methods

Multi-seasonal Holt-Winters

Taylor (2003) proposed the *double seasonal Holt-Winters* (DSHW) model. This was motivated by the desire to capture information from both intra-week and intraday seasonal patterns in half-hourly electricity demand data. The method extends the single seasonal Holt-Winters method by introducing a second seasonal vector and a corresponding smoothing equation to capture two separate stochastic seasonal processes simultaneously, along with stochastic level and trend components. The assumption is that both seasonal processes are regular and periodic. Both additive and multiplicative representations of the seasonality can be adopted; the additive seasonal representation (without trend) is presented below:

$$\hat{y}_{t+h} = l_t + s^1_{t+h-m_1} + s^2_{t+h-m_2} + \phi^h e_t$$

$$e_t = y_t - (l_{t-1} + s^1_{t+h-m_1} + s^2_{t+h-m_2})$$

$$l_t = \alpha(y_t - s_{t-m}) + (1-\alpha)l_{t-1} \tag{2.2.4}$$

$$s^1_t = \gamma(y_t - l_{t-1} - s^2_{t-m_2}) + (1-\gamma)s^1_{t-m_1} \tag{2.2.5}$$

$$s^2_t = \omega(y_t - l_{t-1} - s^1_{t-m_2}) + (1-\omega)s^2_{t-m_1} \tag{2.2.6}$$

Here, $\hat{y}_{t+h}$ is the h-step ahead forecast made at the current time t; l_t is the current level; s^1_t is the seasonal index for the first seasonal pattern; s^2_t is the seasonal index for the second seasonal pattern; m_1 and m_2 are the seasonal periods for the first and second patterns; and α, γ and ω are smoothing parameters, bounded between 0 and 1. ϕ is a parameter representing an adjustment due to first order autocorrelation of the residuals, a well-documented phenomenon (see eg. Chatfield, 1978). ϕ is bounded between -1 and 1.

The method has been extended to the triple seasonal case by Taylor (2010b) to capture intrayear seasonality in electricity demand, and models underpinning the processes have been shown to fit within the framework of an innovations state space model (Hyndman et al., 2008). Theoretically, it is possible to capture any number of seasonal patterns by extension in the same way. Applicability of the model is not restricted to intraday, intraweek and intrayear seasonality; seasonal periods of any length may be included, even if the periods do not nest within each other.

An important limitation of the DSHW model is the number of parameters and initial terms (for the level, and seasonal components) that require estimation, which is

m_1+m_2+5. Taylor (2003) uses 385 parameters and initial terms to model the British electricity data series. For this application this issue is mitigated by the relatively long training set of 8 weeks, with the weekly seasonality being the longer one. We see this as a significant issue for retail issues, where even optimistically only 2 to 3 years worth of historical data might be available. It may not even be possible to distinguish between stochastic and deterministic seasonality in this case. Note that as the additive seasonal exponential smoothing model has an ARIMA equivalent, DSHW is closely connected to ARIMA. However, the latter is often much more difficult to specify than DSHW; this leads to DSHW often being found to perform better (eg. Taylor, 2008).

Intraday cycle exponential smoothing

Gould et al. (2008) proposed relaxing the assumption of DSHW that the intraday seasonal pattern would have the same components for each day of the week. This was achieved by dropping the intraweek seasonal vector and allowing different intraday seasonal patterns for each day of the week. Days that exhibited similar patterns were permitted to share the same intraday component. This model, termed Intra-day Cycle Exponential Smoothing (ICES), takes the following innovations state space form:

$$
\begin{aligned}
y_t &= l_{t-1} + b_{t-1} + \sum_{i=1}^{r} x_{it} s_{i,t-m} + \varepsilon_t \\
l_t &= l_{t-1} + b_{t-1} + \alpha \varepsilon_t \\
b_t &= b_{t-1} + \beta \varepsilon_t \\
s_{it} &= s_{i,t-m} + \Big(\sum_{j=1}^{r} \gamma_{ij} x_{jt} \Big) \varepsilon_t \\
x_{jt} &= \begin{cases} 1 & \text{if time period } t \text{ occurs during a day of type } j \\ 0 & \text{otherwise} \end{cases}
\end{aligned} \tag{2.2.7}
$$

where l_t is the current level at time t; b_t is the current trend; $s_{i,t}$ the current seasonal index for day type i; m the intraday seasonal period; α, β and γ_{ij} are smoothing parameters, bounded between 0 and 1; and $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$ is the innovations term. Taylor and Snyder (2012) recommend also retaining an adjustment for first order residual autocorrelation.

ICES is more parsimonious than DSHW, as clustering allows for a reduction in the number of seasonal components to be estimated. The method can be extended to more general situations. When no two days are considered to exhibit the same pattern, the method reverts to the DSHW method, which is a special case.

BATS and TBATS

De Livera et al. (2011) consider a different generalisation of the DSHW method. They propose supplementing the multiple seasonal components with two additional features, namely an ARMA error structure and a Box-Cox transformation of the data. The resulting model is termed BATS (*B*ox-Cox, *A*RMA errors, *T*rend, *S*easonal). BATS takes arguments ω (the Box-Cox parameter), α, β and γ_i (smoothing parameters bounded by 0 and 1), ϕ (the damping parameter, between 0 and 1), p and q (the orders of the ARMA errors) and m_1,..., m_T (the periods of the T seasonal patterns). The BATS methodology begins with a possible Box-Cox transform:

$$y_t^{(\omega)} = \begin{cases} \frac{y_t^{\omega}-1}{\omega}, & \omega \neq 0 \\ \\ \log y_t, & \omega = 0 \end{cases}$$

and the model then takes the following form:

$$\begin{aligned} y_t^{(\omega)} &= l_{t-1} + \phi b_{t-1} + \sum_{i=1}^{T} s_{t-m_i}^{(i)} + d_t \\ l_t &= l_{t-1} + \phi b_{t-1} + \alpha d_t \\ b_t &= (1-\phi)b + \phi b_{t-1} + \beta d_t \\ s_t^{(i)} &= s_{t-m_i}^{(i)} + \gamma_i d_t \\ d_t &= \sum_{i=1}^{p} \Phi_i d_{t-i} + \sum_{i=1}^{q} \theta_i \epsilon_{t-i} + \varepsilon_t \end{aligned}$$

where l_t is the current level at time t; b_t is the current trend; s_t^i is the i-th seasonal component; d_t represents an ARMA process and $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$ is the innovation term. We note that DSHW is represented by the BATS(1,1,1,0,m_1,m_2) model.

As a generalisation of DSHW, the BATS model also suffers from heavy parameterisation. De Livera et al. (2011) try to mitigate this by replacing the seasonal vectors with a trigonometric representation of seasonality, shown by the following sum of harmonic terms:

$$s_t^{(i)} = \sum_{j=1}^{k_i} s_{j,t}^{(i)} \tag{2.2.8}$$

$$s_{j,t}^{(i)} = s_{j,t-1}^{(i)} \cos \lambda_j^{(i)} + s_{j,t-1}^{*(i)} \sin \lambda_j^{(i)} + \gamma_1^{(i)} d_t \tag{2.2.9}$$

$$s_{j,t}^{*(i)} = -s_{j,t-1}^{(i)} \sin \lambda_j^{(i)} + s_{j,t-1}^{*(i)} \cos \lambda_j^{(i)} + \gamma_2^{(i)} d_t \tag{2.2.10}$$

$$\tag{2.2.11}$$

Here, $\gamma_1^{(i)}$ and $\gamma_2^{(i)}$ are smoothing parameters, $\lambda_j^{(i)} = \frac{2\pi j}{m_i}$ represent the different frequencies, $s_{j,t}^{(i)}$ represents the stochastic level of the i-th seasonal component, $s_{j,t}^{(i)}$ represents the stochastic growth of this level, and $s_t^{(i)}$ represents the i-th seasonal component itself. k_i represents the number of harmonics that is required for the i-th seasonal component.

The BATS model with this trigonometric seasonal representation is known as TBATS. Although the trigonometric representation of seasonality is used, it is considered as stochastic, allowing for the shape of the seasonality to evolve over time. The number of harmonic terms is selected via a heuristic that starts with none and gradually adds in additional frequencies. The heuristic considers one seasonal component at a time, keeping the others fixed. Significance testing is used to determine whether the additional harmonic term is kept or discarded at each stage. The trigonometric representation also allows handling some special cases, such as seasons of fractional length.

The TBATS model is a significant improvement on BATS in terms of parsimony. A limitation of the method is its computational speed when the seasonal lengths $m_1, \ldots, m_T$ are not specified, as the optimisation routine used to determine these is slow. Pre-specifying these parameters speeds computation up considerably.

The application studies in De Livera et al. (2011) compare BATS and TBATS only; the latter is shown to have better performance, with the argument made that the BATS approach encompasses all traditional exponential smoothing models. A further comparison with DSHW would have been of interest; one point it might have illustrated would be if the extra complexity involved specifying the BATS model was worth it in terms of more accurate forecasts.

Parsimonious exponential smoothing

The first allusion to a 'parsimonious' seasonal exponential smoothing model comes from Hyndman et al. (2008), p.49-50. They consider a simple hypothetical example involving sales which are similar in all months, except December when they peak. In this case, it may not be necessary to rigorously define different seasonal states for every period in a season. If certain periods in a season can be assumed to follow the same generating process, then they should take the same seasonal component. Hence, in this example, the use of just two 'seasons' is appropriate, with all but December being classed as periods in season 1, and December observations being classed as season 2.

The logic can easily be extended to multiple seasonalities. These may have differing strength; for example, a daily time series might exhibit a very strong day-of-week pattern, with a weaker week-of-year effect showing prominently in only a few special weeks. The parsimonious approach allows us to model only the parts of each seasonality that are pronounced enough to merit it, thus achieving parsimony.

As the name suggests, Parsimonious Exponential Smoothing (PES) is focussed on reducing the number of seasonal terms needed to model the data, so as to achieve a balance between model simplicity and complexity. PES was introduced by Taylor and Snyder (2012) and fully extends the idea of Gould et al. (2008) by considering not only that days can be clustered into similar profiles, but also that different periods from different days can be clustered as well. In this way, PES encompasses ICES, and allows for the unconstrained clustering of periods into groups, which are considered *seasons* in this model. However, the encoding of seasonality is fundamentally different. PES completely removes the assumption that seasons occur at regular, periodic intervals and allows them to occur at any time, at the discretion of the modeller.

We present below the general form of the model. The authors consider various refinements that are specific to intraday/intraweek seasonalities. Although this is given in fully additive form, a fully multiplicative version is easily obtainable; the formulation for this is provided in A.2.

$$\begin{aligned}
y_t &= \sum_{i=1}^{M} I_{it} s_{i,t-1} + \phi e_{t-1} + \varepsilon_t \\
e_t &= y_t - \sum_{i=1}^{M} I_{it} s_{i,t-1} \\
s_{it} &= s_{i,t-1} + (\alpha + \omega I_{it}) e_t \qquad i = 1, 2, \ldots, M \qquad . \\
I_{it} &= \begin{cases} 1, & \text{if period t occurs in season i} \\ 0, & \text{otherwise} \end{cases}
\end{aligned}$$

Here y_t is the value of the series at time t, M is the total number of distinct seasons chosen in the model, s_{it} is the seasonal state of season i at time t. $\epsilon_t \sim N(0, \sigma^2)$ is the independently distributed error process, whilst α and ω are smoothing parameters taking values between 0 and 1; ϕ is the parameter of a residual autoregressive term.

Unlike the previous models, the level is absorbed into the seasonal vector $\mathbf{s_t}$. This more concise formulation also makes for a simpler initialisation of the states. It is also possible to include a trend component, but this is omitted here for clarity.

We note that the entire vector of seasonal components is updated at each step by αe_t, except the component corresponding to the current season, which is adjusted by $(\alpha + \omega)e_t$ instead. This allows for updating of seasonal components outside of the periods for which they occur. We also note again the presence of an autoregressive parameter, capturing the first-order autocorrelation in the residuals.

The parsimonious approach complicates model selection; a particular configuration of seasons must be chosen for each application of the model. Taylor and Snyder (2012) consider judgemental and statistical approaches. The judgemental approach is to produce average plots of the smaller seasonal cycles (eg. the intra-day cycles), taking note of where they appear to overlap and where they diverge. This approach gives clusters which are interpretable. However, a major limitation is that it is not an automatic procedure and cannot be used for large numbers of series. The authors note that attempts to use basic statistical clustering techniques, such as hierarchical and k-means were not successful. We see automation of model selection as an open question which is quite relevant to retailing, owing to a common need for forecasts for a large number of series. A criticism of PES is that, despite its attempted parsimony, the number of seasonal terms can still be large (Dudek, 2016), making initialisation a potential problem. On the other hand, PES provides an framework where the user can actively control the level of parsimony.

Alternative methods

Taylor (2010a) present some alternative approaches for modelling intraday/intraweek seasonalities, such as the double-seasonal total and split exponential smoothing, which extends a model presented in Taylor (2011) to multiple seasonalities. This smooths both the weekly total sales and the proportional split of sales between each period in both the week and day are smoothed. Since this method (and others described in the paper) are more case-specific, we choose not to focus on them in this paper.

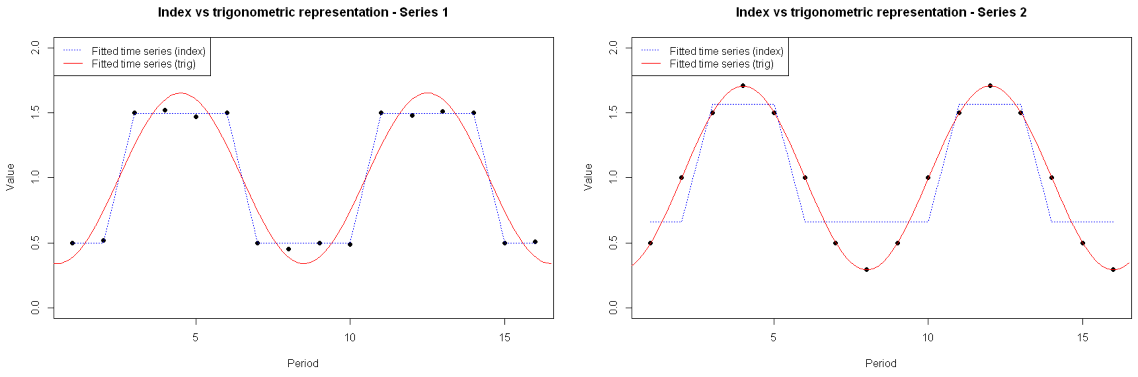
Conclusions

Concluding the examination at exponential smoothing-based methods, we draw together the criticisms of the methods we have seen. Firstly, it is opined that most of the above methods suffer from the need to estimate a large number of parameters, a limitation acceptable in the electrical load forecasting domain where they are applied, but not in retailing, where short demand histories are a defining characteristic. Additionally, whilst the TBATS and PES methods do make some attempt to reduce the number of parameters, the TBATS method is slow computationally, whilst PES requires manual model selection. Neither method allows for both dummy variable and trigonometric representations of seasonality to be used simultaneously. Additionally, there is no empirical evidence using data with short histories, which is vital for demonstrating applicability of multiple seasonal methods in this new area.

2.3 A mixed-representation parsimonious seasonal model

Drawing on our conclusions from the literature review, we propose a new model for multiple seasonal forecasting, gearing our approach specifically to the case of daily data with weekly and annual seasonal effects. The model is novel in that it mixes a trigonometric representation of a seasonal component for day-of-year seasonality with a seasonal index representation for the day-of-week seasonality. Both seasonal representations allow for parsimony. The justification for this is that the day-of-year seasonality gradually changes over the year, particularly in the retailing context, which seems more akin to a sinusoidal function than a piecewise constant function. The day-of-week seasonality, by contrast, is more changeable over the course of its period, and it seems less natural to model this using harmonic terms than simply seasonal indices, when parsimony is the objective. Since we anticipate being unable to distinguish between deterministic and stochastic seasonality for a yearly pattern due to limited sample size, the trigonometric seasonal component of the model is deterministic, while the seasonal index component is kept stochastic as there are plenty of examples of the higher frequency seasonality.

Although using seasonal dummies to represent seasonality is equivalent to using a sum of harmonic terms, this does not hold once terms are removed, since the representations become sparse in different ways. Figure 2.3.1 illustrates how the different representations might be more appropriate in different situations. Both panels display a seasonal data series of length 16 with seasonal period 8. For both series, an attempt



(a) Index is a better fit

(b) Trigonometric is a better fit.

Figure 2.3.1: An illustration that both the trigonometric and index representations of seasonality can fit better to different series.

has been made to model the series using just 2 seasonal parameters, trying both representations. It can be clearly seen that, on the left hand side, the index approach (dotted line) is more suitable, whereas on the right hand side the trigonometric (solid line) is superior.

Given our reasoning, for the lower frequency annual seasonal component we estimate a harmonic regression of the form:

$$y_{s,t} = \sum_{k=0}^{\frac{S}{2}} \left[a_k \cos\left(\frac{2\pi kt}{S}\right) + b_k \sin\left(\frac{2\pi kt}{S}\right) \right] + \varepsilon_t \tag{2.3.1}$$

In this equation, a_k and b_k are the coefficients of the trigonometric functions, and k is the frequency of the harmonic.

The parsimony here will be introduced by a number of the α_k and β_k terms being set equal to 0. A good analogy is the decomposition of a signal via a Fourier transform into its component frequencies. Infinitely many components of different frequencies may be sequentially added to build up a closer and closer approximation of a continuous signal. However, diminishing returns occur with the addition of each one. At some point we stop, as the accuracy gained from adding another term is disproportionate to the cost in complexity. The same principle applies in our case.

Our idea is to estimate the full regression with all harmonic coefficients at first, and then gradually eliminate terms using the Akaike Information Criteria (AIC) in a backward fashion, so as to strike a good balance between parsimony and model fit.

For the high frequency intraweek seasonality, we use the conventional binary dummy-based seasonal representation. We do that for two reasons: i) our data suggests that this seasonality is more discontinuous, thus not lending itself towards increased parsimony by trigonometric encoding; and ii) we want to retain the advantages of PES, that is to capture parsimoniously non-regular seasonal elements with ease.

2.3.1 Innovations state-space model

We propose an innovations state-space model to produce the required forecasts. This model framework has been advocated as a way to underpin forecasting methods in recent times and is commonly referred to as single-source-of-error (SSOE) models, due to all the error sources being perfectly correlated. The main alternative to the SSOE formulation is a multiple-source-of-error (MSOE) formulation, where the error sources are independent. Both formulations have their strengths, but we choose the SSOE form primarily to facilitate easier comparison between our model and the others discussed in the previous section, due to SSOE being the choice of model form there throughout. Further discussion of SSOE vs. MSOE can be found elsewhere (see eg.

Hyndman et al., 2008).

Another advantage in employing a state-space model to underpin our method is that theoretical expressions of variance are possible to obtain, allowing the provision of probabilistic forecasts. We do not investigate this here but instead include it as further research in Chapter 2.5.

The general form for a linear innovations state space model is given by:

$$y_t = \mathbf{w}^{\mathbf{T}}\mathbf{x}_{\mathbf{t-1}} + \varepsilon_t \tag{2.3.2}$$

$$\mathbf{x}_{\mathbf{t}} = \mathbf{F}\mathbf{x}_{\mathbf{t-1}} + \mathbf{g}\varepsilon_t \tag{2.3.3}$$

The first equation is known as the *measurement* equation, where the observation y_t is described as the sum of the states $\mathbf{x}_{\mathbf{t-1}}$, multiplied by coefficients $\mathbf{w}$, plus the innovations term $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$. The second equation is the *transition* equation, where the states are updated. The state vector $\mathbf{x}_{\mathbf{t-1}}$ is multiplied by a transition matrix F, and the error term is included in places via the vector of coefficients $\mathbf{g}$.

We set $\epsilon_t = e_t = y_t - \hat{y}_{t|t-1}$ and set out our list of states: stochastic seasonal indices $s_1, \dots, s_M$, deterministic harmonic seasonal components $w_1, \dots, w_S$ and the autoregressive state e_t. Substituting ε_t into our measurement equation, we obtain

$$y_t = \sum_{i=1}^{M} I_{it}s_{i,t-1} + w_{t-S} + \phi e_{t-1} + \varepsilon_t \tag{2.3.4}$$

and the vector of coefficients $\mathbf{w}$ in our measurement equation takes the form

$$\mathbf{w}_{\mathbf{t}} = (\mathbf{I}_{\star\mathbf{t}}^{\mathbf{T}}, 1, \mathbf{0}_{S-1}, \phi) \tag{2.3.5}$$

We notice that $\mathbf{w}_{\mathbf{t}}$ is time-varying, since it depends on a different row of the matrix I at each time point.

We move now to deal with the transition equations. Starting with the autoregressive term e_t, we use the measurement equation just described to see that

$$e_t = y_t - \sum_{i=1}^{M} I_{it} s_{i,t-1} - w_{t-S} \tag{2.3.6}$$

$$= (\sum_{i=1}^{M} I_{it} s_{i,t-1} + w_{t-S} + \phi e_{t-1} + \epsilon_t) - \sum_{i=1}^{M} I_{it} s_{i,t-1} - w_{t-S} \tag{2.3.7}$$

$$= \phi e_{t-1} + \epsilon_t \tag{2.3.8}$$

Our deterministic seasonal values $w_{t-S}, \ldots, w_{t-1}$ do not change, but we do rotate them by 1 step to get the correct values in place for the next time period. That is, the value at w_{t-1} moves to w_{t-2}, and so on, with the value from w_{t-S} which was used in the last measurement equation moving to w_{t-1}.

For the stochastic seasonal indices, we use the equation for the autoregressive term to see that

$$s_{it} = s_{i,t-1} + (\alpha + \omega I_{it}) e_t \tag{2.3.9}$$

$$= s_{i,t-1} + (\alpha + \omega I_{it})(\phi e_{t-1} + \epsilon_t) \tag{2.3.10}$$

$$= s_{i,t-1} + (\alpha + \omega I_{it})\phi e_{t-1} + (\alpha + \omega I_{it})\epsilon_t \tag{2.3.11}$$

Hence we can see that our transition matrix $\mathbf{F}$ takes form

$$\mathbf{F} = \begin{bmatrix} \mathbf{1}_{MxM} & \mathbf{0}_S & (\alpha + \omega \mathbf{I_{\star t}})\phi \\ \mathbf{0}_M & 1(c) & 0 \\ \mathbf{0}_M & \mathbf{0}_S & \phi \end{bmatrix} , \tag{2.3.12}$$

and the coefficients vector $\mathbf{g}$ takes form

$$\mathbf{g} = ((\alpha + \omega \mathbf{I_{\star t}}), \mathbf{0}_S, 1) \tag{2.3.13}$$

Putting all the above together, we obtain a full state-space formulation of our model.

2.3.2 Estimation

Maximum likelihood estimation

We use maximum likelihood estimation to parameterise w and x_0, the vector of initial states. This is undertaken in the time domain, as the procedure is already worked out and fairly straightforward; however, with harmonic terms prominent in the model, estimation in the frequency domain would have been a sensible alternative. For the proposed additive state-space model, the likelihood function can be reduced to (Hyndman et al., 2008):

$$\mathcal{L}(\theta, \mathbf{x_0}, \sigma^2|\{y_1, \ldots, y_T\}) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \cdot \exp\Big(- \frac{1}{2\sigma^2} \sum_{t=1}^{T} \epsilon_t^2 \Big) \tag{2.3.14}$$

The variance parameter σ^2 is concentrated out by substituting in its maximum likelihood estimator

$$\hat{\sigma^2} = \frac{1}{T} \sum_{t=1}^{T} \epsilon_t^2 \tag{2.3.15}$$

a result achieved by partial differentiation of the previous equation. Once this is done, then it follows that the MLEs of the other parameters are achieved by simply minimising $\sum_{t=1}^{T} \epsilon_t^2$, the sum of squared errors.

Initial seed vector

Our method contains 3 smoothing parameters, M stochastic seasonal indices and S deterministic seasonal values, which is a large number of parameters to be optimised simultaneously. Accordingly, when starting the optimisation from a randomised initial seed vector, it was found that the outcome was heavily dependent upon the starting conditions and often converged to local minima, giving poor out-of-sample accuracy.

As a result, we facilitate the optimisation process by starting from a reasonable initial x_0 computed via the following heuristic, which we have found to work well in practice. We start by estimating the deterministic part of the seasonality. A moving average is used to smooth the data, in order to fit the harmonic regression. In our case, a good length for the moving average was a couple of times the length of the shorter seasonal period we are trying to smooth out, for example two weeks. We then run the harmonic regression to obtain initial states for the deterministic seasonal component. Then we subtract the estimated seasonality and run the usual estimation procedure for PES on the resulting series. This produces a full initial x_0 from which the optimisation starts.

2.3.3 Model specification

To select the number of harmonics for the model, we propose to use a backwards regression procedure. We start with a model with all harmonic terms. As noted, estimating this model would be equivalent to estimating a seasonal-index based model, as the number of degrees of freedom is equal to the number of points. Using the AIC values we evaluate models with one less term, until we cannot improve any further. As the search is one-directional, it is relatively fast.

For the stochastic seasonality, we rely on the conventional PES procedure, for which A.1 provides guidelines; this cannot be fully automated without a heuristic, since there is a huge number of possible models and no clear order of progress, but we can use AIC to compare potential alternatives. Moreover, the use of AIC improves upon the purely judgemental model selection procedure of Taylor and Snyder (2012), exploiting a key advantage of the state-space model formulation.

2.4 Empirical evaluation

2.4.1 Dataset

We undertake an empirical study to compare the performance of our new model against the existing methods discussed in the literature review. The data comes from a food wholesaler and has 12 daily series describing sales of assorted food products to clients within a particular sector over 3 years from 1/1/14 to 31/12/16. The 12 series represent aggregated sales over clients from different sectors of the economy; examples include the education sector (schools, colleges, universities), pubs, fast food restaurants and hotels. The wholesaler experiences distinct complex seasonal patterns of incoming orders from each sector; this characteristic in the data is one primary motivation behind choosing it for this evaluation. Individual product level sales were not available, and so the sales are aggregated by transaction value. The sales series are used as proxy for the total demand that exists; no information is available on unobservable lost sales, but we do not expect that the data characteristics would be different in nature if those lost sales were factored in due to our expectation that the wholesaler is able to meet the vast majority of demand. We use the first 2 years for estimation and the remaining data as a test set for forecast evaluation. We examine a range of forecast horizons, from 1 day through to 28 days, to assess both short-term and medium-term forecasting accuracy. This reflects a realistic range of supplier lead times that are encountered in practice. For inventory decisions, cumulative forecasts are required, thus we also examine the cumulative forecast horizons of 7, 14 and 28.

2.4.2 Methods

The performance of the proposed method is evaluated against a set of simple benchmarks and established methods from the literature, listed in Table 1. The benchmarks are basic statistical methods which capture only the highest frequency seasonality. These are as follows: i) the seasonal naive $y_t = y_{t-s} + \varepsilon_t$, where s is the seasonal period, simply sets the forecast equal to the observed value last period (here, a week). ii) the well-known Holt-Winters, also known as single-seasonal exponential smoothing (ES), embedded in a state-space framework Hyndman et al. (2008) is fitted on the weekly seasonality, allowing us to assess the need for double seasonality. iii) a simple regression model with binary dummies $y_t = \alpha + \sum_i \beta_i d_{i,t} + \varepsilon_t$ provides a deterministic seasonality benchmark. The established models are as described in Section 2; for the PES models, we specify seasonality as in A.1. Both the additive version Taylor and Snyder (2012) and a multiplicative variant described in A.2 are implemented. We evaluate existing forecasting approaches on the retailing case to demonstrate their relative merits and highlight the benefits of the proposed model, and the inclusion of simple models allows us to assess any gains due to the additional complexity of multiple seasonality.

Custom implementations of the DSHW method, both PES methods, and our new model were created in R Core Team (2013) for the evaluation. For TBATS, we used the `tbats` function found in the `forecast` package for R (Hyndman et al. (2007)), whilst for single-seasonal ES we used the `es` function from the `smooth` package for R,

Table 2.4.1: Benchmark methods

Method	Seasonality
Seasonal naive	7
Single-seasonal ES	7
Regression with seasonal dummies	7, special days
TBATS	7, 365
DSHW	7, 365
PES(additive)	7, special days
PES(multiplicative)	7, special days

(Svetunkov (2017)).

2.4.3 Error measures

We use two different scale-independent error metrics to assess forecast accuracy: Mean Absolute Percentage Error (MAPE) and Average Relative Mean Absolute Error (AvgRelMAE), introduced by Davydenko and Fildes (2013). We use MAPE as it is a common metric in industrial practice, whilst AvgRelMAE has favourable statistical properties and is easy to interpret. An AvgRelMAE of less than 1 indicates the evaluated approach outperforms the benchmark, and vice versa. Gains can be expressed as a percentage by calculating $(1 - AvgRelMAE) \times 100\%$.

Letting $e_t = y_t - \hat{y}_t$ be the forecast error as the difference between the observed value y_t at period t and the forecast $\hat{y}_t$, the formulae for the error metrics are given below:

$$MAPE = \frac{1}{T}\sum_{t=1}^{T}\left|100\frac{e_t}{y_t}\right|, \tag{2.4.1}$$

$$MAE = \frac{1}{T}\sum_{t=1}^{T}|e_t|, \tag{2.4.2}$$

$$AvgRelMAE = \Big(\prod_{t=1}^{T} r_t\Big)^{\frac{1}{T}}, \qquad r_t = \frac{MAE_{\hat{y}}}{MAE_b} \tag{2.4.3}$$

where $MAE_{\hat{y}}$ is the Mean Absolute Error (MAE) of the forecasting method being evaluated, and MAE_b is the MAE of a benchmark method. For this evaluation, the seasonal naive method is used as the benchmark. Note that there is no reference to horizon in the accuracy notation, since both metrics may be applied in any specific case.

2.4.4 Results

We start by examining forecast accuracy for individual horizons. Table 2.4.2 presents the AvgRelMAE results for each method/horizon combination, averaged across the 12 series. The best forecast in each column is highlighted in bold. We focus on AvgRelMAE for the discussion; the MAPE results can be found in A.3, and lead to similar conclusions.

The simple benchmark methods, on average, do not perform the best. Overall, the single seasonal ES performs the best of the three and is competitive with more sophisticated methods. We also note that the stochastic seasonality of the seasonal

Table 2.4.2: AvgRelMAE figures, individual horizons.

Forecast	Horizon			
	1	7	14	28
Seasonal naive (7)	1.000	1.000	1.000	1.000
Single-seasonal ES (7)	0.985	**0.836**	**0.821**	0.864
Regression	1.644	1.311	1.333	1.318
TBATS (7,365.25)	1.284	1.070	1.050	1.050
DSHW (7,365)	1.020	0.979	1.050	1.048
PES (additive)	0.956	0.913	0.931	0.953
PES (multiplicative)	0.966	0.924	0.978	1.020
Mixed parsimonious	**0.932**	0.848	0.856	**0.859**

naive is better than the deterministic regression. Among the multiple seasonal methods, the parsimonious methods are proving to be most accurate, and TBATS and DHSW are performing rather poorly. This is in line with our expectations. It was noted in Section 2 that DSHW may suffer from being over-parameterised when short data histories are in effect; this seems to be the case here. We justify the TBATS results due to i) the TBATS model is difficult to estimate with a limited sample due to its complexity, and ii) TBATS lacks the flexibility of the PES methods in modelling seasonality that does not fit the regular patterns. The good performance of the PES methods is somewhat expected given that parsimony is a significant benefit when dealing with short data histories. The additive version is overall marginally better than the multiplicative; this is due to a lack of strong trend in the data, which would make multiplicative seasonality more prominent. However, it is interesting to note that at horizons longer than 1 step ahead, the much simpler single-seasonal ES performs better than all the multi-seasonal methods. This can be attributed to limited history available in our data.

The new mixed parsimonious method is the most accurate method for 1-step and 28-steps ahead forecasts, and quite competitive for 7-steps and 14-steps ahead, where it comes second to the single-seasonal ES. The consistent performance meets our expectations due to the benefits of parsimony, and justifies our modelling decision to represent seasonality in different ways.

We also consider cumulative accuracy figures for three different horizons: 7, 14 and 28 (ie. 1, 2 and 4 weeks ahead). The cumulative forecasts represent the sum of all demand from the origin up until the horizon. Their accuracy is important for demand

Table 2.4.3: AvgRelMAE figures, cumulative horizons.

Forecast	Horizon		
	1-7	1-14	1-28
Seasonal naive (7)	1.000	1.000	1.000
Single-seasonal ES (7)	1.324	1.348	1.357
Regression	0.952	0.996	1.023
TBATS (7,365.25)	1.343	1.401	1.452
DSHW (7,365)	0.932	0.921	0.941
PES (additive)	0.911	0.921	0.922
PES (multiplicative)	0.962	0.965	0.961
Mixed parsimonious	**0.902**	**0.906**	**0.911**

planners, as they may make replenishment orders based on covering demand over relevant lead times, rather than a single period. Table 2.4.3 presents the AvgRelMAE results for the three cumulative horizons previously mentioned. We observe that, in general, the methods that see the lowest drop-off in individual forecast accuracy as the horizon increases also see the biggest improvements (or smallest declines) in accuracy as the cumulative forecast horizon is extended.

The mixed parsimonious model is consistently the best performer for all horizons, with substantial gains overall. Comparing Table 2.4.3 with the previous Table 2.4.2, we can observe that the multi-seasonal methods perform relatively better, while that is no longer the case for the single seasonal benchmarks. The difference is most striking at the longest cumulation, 1-28 days where with the exception of TBATS, all

DSHW and PES variants perform only second to the proposed model. Although on single periods, the gains provided by better overall tracking of the time series shape through the second seasonality is not evident, with cumulative errors the benefits become clear.

Lastly, we want to be confident that our model will work both for intervals containing 'special days' and otherwise, as special events are very frequent in retailing; for a definition of what we consider special days, see A.1. This is a vital consideration for demand planners who are seeking a robust forecasting system that will produce adequate results under all circumstances. To assess this, we look again at the cumulative forecast accuracy, assessing periods which contain at least 1 special day separately to those which only contain 'usual' days. Table 2.4.4 shows the AvgRelMAE figures for the 1-7 day case; the results for the other horizons are generally similar.

Again, we conclude that the mixed parsimonious method compares favourably with all other methods. The improvement shown over other methods seems strongest for those intervals which contain at least 1 special day. We note particularly the substantial difference in the performance of single-seasonal ES, which performs relatively well on special days but not on normal ones. This explains further the cumulative results in Table 2.4.3; it also agrees with findings in Barrow and Kourentzes (2018), where this effect is also observed. Since we report AvgRelMAE, the errors when at least 1 special day occurs are higher as absolute values, but appear lower in Table 2.4.4 since they are expressed relative to the seasonal naive. Our results clearly demonstrate the value in representing different types of seasonality in different ways and in the focus on model parsimony, since the parsimonious methods perform the best.

Table 2.4.4: AvgRelMAE figures, 1-7 day cumulative horizon, special days vs. no special days.

Horizon	No special days	At least 1 special day
Seasonal naive (7)	1.000	1.000
Single-seasonal ES (7)	1.327	0.924
Regression	0.953	0.944
TBATS (7,365.25)	1.361	1.050
DSHW (7,365)	0.945	0.921
PES (additive)	0.912	0.871
PES (multiplicative)	0.964	0.890
Mixed parsimonious	**0.909**	**0.844**

2.5 Conclusions

The use of multiple seasonal forecasting methods for forecasting in retail has been investigated in this paper. The results show that these methods, many of which have hitherto been applied mainly in the short-term energy forecasting domain, are not directly applicable in the retailing context, due to the particular properties exhibited by those time series. It is noted that price and promotion influences are often crucial for forecasting SKU-level items; however, we do not address this here, as these are additional to the seasonal effects. In addition, the model is not fully automated in its current form, and as such the issue of forecasting thousands of time series is not addressed. However, the model does allow great flexibility for the user to model seasonality manually; and recent research (Petropoulos et al., 2018) has shown judgmental model building to perform strongly when compared directly to automated model selection.

We introduced a new approach for complex seasonal forecasting, the mixed parsimonious method. The novel feature of this approach is the mixture of trigonometric and seasonal dummy parsimonious representations of seasonality. It was shown that the two representations are sparse in different ways; this logic, in combination with retail-specific demands such as flexibility and short data history, was used as the justification for the construction of the mixed model.

The empirical studies showed that the mixed parsimonious method generally outperformed benchmark methods and other methods dealing with complex seasonal forecasting. Importantly for the demand planner, the model was consistently best in the situation where cumulative horizons of differing lengths were considered, which is directly relevant to inventory related decisions, a critical function in retailing. The improvement in accuracy also held when special days were introduced into the forecast period, demonstrating that the model is robust to such situations which occur frequently in practice. Note that one key advantage of the proposed model is the ability to calculate expressions for the variance through the state-space framework, which is also important to support these decisions. This is a viable alternative to empirical methods, which are often problematic due to limited sample sizes.

The process of modelling seasonality using mixed parsimonious representations was moulded very specifically to the case of daily aggregate data in this research. A logical next step would be to generalise this approach to all types of multiple seasonal series, with one possible approach being to establish a decision rule which would dictate which combination of seasonal representations should be used in a given situation. The current framework provides a natural route for such extensions.

One of the advantages of underpinning our new forecasting method with a state-space model that has not been touched upon much is the ability to compute theoretical variance expressions, which would allow the provision of prediction intervals alongside point forecasts. Probabilistic forecasts of this form have received attention in related areas (see eg. Hong et al., 2016) and there is a natural benefit to demand planners in the context of inventory management, where safety stock calculations rely on quantiles of expected demand as inputs. Conducting further research in this direction would thus be useful to practitioners and topical from an academic standpoint.

Additionally, the problem of automatic model selection was touched upon. Although we do not provide a fully automated specification for PES (and by extension, the proposed model) the ability to calculate AIC values for alternative representations makes the process much simpler and more quantitative. Further research in this area could look at efficient heuristics for searching within the model space.

Chapter 3

Sources of bias in loglinear models for retail

Abstract

All retailers face the essential task of producing forecasts of sales at the individual item or stock keeping unit (SKU) level, which facilitate a variety of inventory management and supply chain decisions. Log-linear regression models are often used for this task since they yield parameters that take a useful interpretation, such as price elasticity of demand and promotional uplift. However, both the parameter estimation and forecast generation processes in these models can be subject to bias, affecting performance. This paper addresses both facets of this problem. Firstly, we layout a straightforward procedure for improving parameter accuracy at the disaggregate level through the inheritance of estimates from a higher level of aggregation, avoiding aggregation-related bias. Secondly, we investigate the forecast bias theoretically, and propose an

approximation to de-bias the forecasts by accounting for parameter uncertainty as well as model bias. Through simulations, we demonstrate the performance in parameter accuracy and forecast bias in a range of scenarios.

3.1 Introduction

A vital component of any retailer's inventory management system are forecasts at the individual product/stock keeping unit (SKU) level, for each store location. The number of series where such forecasts are required has been increasing in recent years and can amount to hundreds of thousands for some large retailers, or even more for the very largest; for instance, Seaman (2018) puts the number of SKU x store combinations required for Walmart at over 1 billion. Additionally, the ability to record and store sales data on a transactional basis has led to the possibility of forming sales series from smaller time windows, such as daily totals. Demand series can be sparse or even intermittent at these timeframes, introducing the challenge of aggregating data at the correct temporal scale Nikolopoulos et al. (2011). More granular time series naturally exhibit a lower signal-to-noise ratio and hence are often the most difficult to forecast; demand uncertainty is already among the most important challenges for retailers (Chen and Blue, 2010). In addition, parameter estimates are frequently extremely variable and this error translates through into high forecast errors.

The increased level of detail exposed by these trends poses challenges for forecast practitioners operating on multiple different levels of the business (see eg. Fildes et al., 2019). Inventory managers wish to be able to produce forecasts for smaller timeframes to avert stockouts with less notice. Marketers need to devise promotional strategies in situations where the number of SKUs is increasing, particularly with greater variation in pack sizes and other attributes of the same product. The most common model form used to forecast such series is a loglinear model of sales, which depends multiplicatively on variables such as price and promotional activity, among others. The adaptation of these traditional forecasting techniques, both to overcome these obstacles and to take best advantage of the new possibilities granted, is of paramount importance for retailers.

This paper makes two main contributions. The first is to examine the idea of using more aggregate series within the hierarchy to stabilise parameter estimates at the disaggregate level. We outline a simple procedure which is only possible with an increased level of data granularity, and examine how it provides improvement in a wide range of practical situations. The procedure also removes a source of bias that is omnipresent in other aggregation schemes. The second contribution is towards quantifying the forecast bias that results in estimating these regression models in the log domain. We quantify this bias and provide an approximation equation to alleviate it, evaluating its performance under a range of conditions. Overall, the paper presents two ways in which practitioners can alleviate parameter and forecast bias with simple techniques, along with guidance as to where these techniques will provide the most benefit, both in the estimation itself and in improved forecast performance.

The rest of this paper is as follows. In Section 3.2 we briefly review the literature on sales response models before a review of aggregation in retail forecasting models and bias in loglinear models. In Section 3.3 we outline our proposed scheme to improve disaggregate parameter estimates using aggregate information. Section 3.4 reports our proposed approximation to reduce bias in parameter estimation from loglinear models. The two individual ideas are then combined in Section 3.5 and simulations are provided to demonstrate the improved performance. Section 3.6 concludes with a discussion of the results and areas for possible future research.

3.2 Literature review

3.2.1 Sales response models

A number of different approaches to modelling sales occur in the literature, underpinned by different assumptions about what drives demand in retail. Regression type models are by far the most common, both in the literature and in practice. Regression models assume that demand for a SKU is dependent upon a combination of certain variables, including the price and promotional activity. Furthermore, a loglinear or multiplicative form of regression is the most popular choice of sales response function (Hanssens et al., 2003). One of the main reasons for this is that, under this form, the parameters take useful interpretations; for example, the coefficient of price is interpreted as price elasticity of demand under this setting. Another reason is that uplift, for instance caused by a promotion taking effect, is likely to scale as the baseline scales; an additive promotional effect which is independent of the series' current baseline does not make as much sense.

There are many examples of case studies in the literature where the models formulated follow a loglinear/multiplicative regression form. Cooper et al. (1999) present the PromoCast system, where the natural log of sales is regressed against the log of price and other unit related variables, along with advertising and display combination variables. Divakar et al. (2005) implemented both linear and loglinear forms of their regression-based forecasting model CHAN4CAST, and found that the loglinear model performed best, although the difference between the two was slight. The SCAN*PRO model of Wittink et al. (1988) presents a multiplicative equation with the product of relative price, promotional and seasonal factors used to represent weekly sales, relative to the baseline; this model form has facilitated a number of natural extensions to capture different retail phenomena (Van Heerde et al., 2002). Given the popularity of loglinear regression models, we regard them as a worthy candidate for specific study and regard alternative, plausible modelling options as outside of the scope of the paper.

3.2.2 Aggregation in retail modelling

The topic of aggregation is long-standing. Grunfeld and Griliches (1960) presented one of the first works to consider whether an aggregate level model or the composite of a number of disaggregate models could better explain an aggregate level dataset, here in the context of regression. Their conclusion, based on empirical studies, was that aggregation gain may be possible, under certain conditions. This conclusion has been reached a number of times in different analyses of aggregation since, as discussed below.

Passing parameters between aggregation levels has been studied in the context of seasonality. Withycombe (1989) proposed a method for estimating the seasonal indices for a group of items by estimating the index on the aggregate series and passing it down. By way of an empirical study, Bunn and Vassilopoulos (1993) found that aggregating items for estimation in this way yielded a reduction in out-of-sample forecast error over an estimation method of Dalhart (1974), where estimation takes place at the disaggregate level and is then averaged. A theoretical investigation into these methods was conducted by Chen and Boylan (2007), where a set of decision rules were discovered, explaining the circumstances under which gains from aggregation are possible. The key quantity of importance was found to be the coefficient of variation, defined as the ratio of the standard deviation to the mean; if this quantity was minimised on the aggregate series, aggregation would yield gains, otherwise a disaggregate method should be preferred. Chen and Boylan (2008) conducted an empirical study, where use of the decision rules were found to produce more accurate forecasts than other methods. Aggregation gains in a seasonal forecasting context have also been demonstrated by Dekker et al. (2004), who exploited the use of a multiplicative Holt-Winters forecasting model form to propose a procedure where aggregate-level seasonal estimates are transferred to a group of disaggregate level series. Applying the method to retail data, they found significant improvements in forecast accuracy at the SKU level. Many other demand characteristics have been studied in the context of aggregation; for example Widiarta et al. (2007) look at the effectiveness of top-down and bottom-up strategies when demand follows a first-order autoregressive process.

Aggregation across time has also been considered. Temporal aggregation in the context of a relatively simple demand process, with forecasts produced via simple exponential smoothing was considered by Rostami-Tabar et al. (2013), and showed that forecast improvements could be obtained as a result of aggregating demand into lower frequency time buckets. Additionally, Kourentzes and Petropoulos (2016) used exponential smoothing to forecast demand via MAPAx (Multiple Aggregation Prediction Algorithm with exogenous variables). In addition to aggregating across different temporal frequencies, the authors also aggregated promotions across retailers to construct promotions for the manufacturer. This approach was found to outperform other methods in terms of both forecast bias and accuracy; by combining promotional information at different levels of temporal aggregation, the MAPAx algorithm provided forecasts which were robust to model misspecification at any individual level.

Passing parameters between sales regression models at different levels has been studied, but runs into a problem with bias. Simply put, multiplicative regression models are not appropriate for category-level sales, since the linear aggregation of such models from the SKU-level is not well-specified by a multiplicative regression model of the same form. Thus, parameters estimated at the aggregate level will be biased estimates of the corresponding disaggregate parameters. The problem is therefore one of alleviating the bias or discovering a way around it.

Lewbel (1992) analyses this situation from a theoretical standpoint, concluding that aggregation bias will exist unless the variables satisfy a property known as *mean scaling*. Intuitively, this property requires that if the mean of the variable across disaggregate components changes, the relative distribution of the values remains unchanged. In a retail context, this property will not hold; consider the simple example of a binary promotional variable, which takes value 1 to indicate the presence of a promotion, and 0 to indicate its absence. Starting from a period in which no SKUs are on this promotion, when this promotion is run on a few SKUs within a category, the mean will change from 0, but the variance of the relative distribution will be increased. Following this logic, the only special case where mean scalability is present is if the promotional activity is homogeneous, ie. where either all values of a variable are 0 or 1. The same conclusion is reached by Link (1995), who, whilst omitting a rigorous theoretical justification, supplies a fuller list of variables commonly used in sale regressions which exacerbate aggregation bias. The theoretical situation is further analysed by Van Garderen et al. (2000), who set out conditions under which the aggregate model (in the face of bias) or the disaggregate model is best able to forecast at the aggregate level.

Until relatively recently, the typical situation has been that only linearly aggregate scanner data is stored by retailers, and so alleviating the aggregation bias has attracted research attention. One of the first detailed investigations into the problem in a retailing context is described in Wittink et al. (1993), where the authors quantify the bias present when promotional variables are distributed in a range of ways and show via empirical analysis that aggregate level estimates of promotional uplift frequently show significant positive bias. Their recommendation was that linearly aggregated data should not be used to estimate promotional effect at the SKU level, and that alternative methods of aggregation should be explored. Christen et al. (1997) corroborate these findings, and offer a solution for practitioners via a large scale simulation. They construct a large table of debiasing coefficients which can be used in a wide range of situations to improve market-level estimates of disaggregate parameters. Although an empirical study is provided showing improvement in the accuracy of parameter estimates once the debiasing is applied, we find the suggested solution quite unwieldy, and it is unclear that the improvements will generalise for other product domains. A different solution is proposed by Bemmaor and Wagner (2002), who use the arithmetic mean and standard deviation of sales and prices to predict the geometric means of those quantities via the method of moments. Using these quantities in an aggregate model, improvements were found in parameter estimation. This solution seems quite unnatural, and the distributional assumption seems unlikely to be valid, especially for the price variable.

Alternative schemes of aggregation are not considered much in the literature, although Wittink et al. (1993) and later Christen et al. (1997) do mention that geometric aggregation might provide a way around the issue of bias. The principal reasons why this has not attracted more interest can be imagined: (i) geometric aggregates of sales are not stored and the granular level of data needed to compute these are only recently becoming available (ii) the geometric aggregate quantity itself is not of interest to forecast, and a number of papers were either motivated by forecasting at both disaggregate and aggregate levels, or just the latter. Since we focus on forecasting at the SKU level and make the assumption that enough data is stored to compute these geometric aggregates, our situation allows us to explore this signficant gap in more depth.

3.2.3 Bias in loglinear models

A further bias-related problem stems from the estimation procedure of loglinear models. This estimation is typically carried out via ordinary least squares (OLS) on the log-transformed equation, with those parameters then being translated back to the original model via taking exponents (where necessary). However, whilst the statistical properties of the OLS procedure guarantee the optimality of these parameter estimates in the log-domain, they introduce bias in the original domain since the transformed estimates are representative of the median, rather than the mean.

This problem has been identified in the statistical literature. One of the first researchers to examine lognormally distributed quantities was Finney (1941), who derived expressions for the moments of the distribution and formulae for efficient sample estimates, noting indirectly the relationship between the means of normally and lognormally distributed variables. Neyman and Scott (1960) examine the issue of bias by looking at the general case where a transformation of variables takes place. Van Garderen (2001) derives an exact optimal predictor and variance for the dependent variable of a loglinear equation using a Laplace inversion method. However, the expressions require the evaluation of hypergeometric functions and are too complicated for practical situations; moreover, strict assumptions are made about the behaviour of the covariates and parameters which are unlikely to hold.

In the marketing literature, this issue seems to have attracted relatively less attention. One exception is the work of Miller (1984), who focusses on solutions to bias problems from the perspective of practitioners. The research found that a simple bias correction term, where half of the mean squared error is added to the prediction in the log domain before it is exponentiated, can reduce the bias; the analysis was, however, limited in this paper to being theoretical, rather than empirical. Cooper et al. (1999) apply this correction when estimating their Promocast model, noting that it is analogous to a procedure detailed in Wittink et al. (1988) for the estimation of the original SCAN*PRO model; however lack of easy access to the latter paper may mean that many practitioners are unaware of its use. In fact, the inclusion of a correction seems to be overlooked in some academic papers; for example, Andrews et al. (2008) note only that, when fitting the SCAN*PRO model, the parameters of the model 'can be estimated with OLS after a log transformation'.

The idea of an approximate bias correction term is practically appealing. We take Miller (1984) as a starting point for a more thorough theoretical investigation. By testing a range of alternative approximations through simulations, we address another gap in the literature.

3.3 Geometric parameter inheritance

We consider the simplified loglinear regression model for sales which varies solely due to price:

$$S_{j,t} = \alpha_j \tilde{P}_{j,t}^{\beta_j} \varepsilon_{j,t} \quad , \tag{3.3.1}$$

where $S_{j,t}$ represents the sales of SKU j at time period t, α_j represents baseline sales, $\tilde{P}_{j,t}$ represents price relative to the baseline price, β_j represents price elasticity of demand and $\varepsilon_{j,t}$ are lognormally distributed errors $ln\varepsilon_{j,t} \sim \mathcal{N}(0, \sigma^2)$, independent and identically distributed (IID) across both the time and item dimensions. We also assume that $\varepsilon_{j,t}$ is independent from $\tilde{P}_{j,t}$. The analysis can be naturally extended to models which include more covariates, such as promotion-type binaries; we use a simplified model here to focus more clearly on the aggregation issues.

We use the term *parameter inheritance* to describe a procedure where parameter estimates are produced at an aggregate level and then passed down to be used at the disaggregate level. For this to make sense, we need to make the assumption that the individual SKUs share a common parameter, since there will only be one estimate at the aggregate level to be inherited. Thus, we specify the following models:

$$\begin{array}{cc} \text{Disaggregate model} & \text{Aggregate model} \\ S_{j,t} = \alpha_j \tilde{P}_{j,t}^{\beta} \varepsilon_{j,t} & S_t = \alpha \tilde{P}_t^{\beta} \varepsilon_t \end{array} \tag{3.3.2}$$

where β, the common elasticity of demand for all SKUs, appears in both the aggregate and disaggregate models, whereas α, which represents the intercept of the aggregate model, is only seen on that side and is not interpreted in terms of the individual α_j. ε_t, the errors for the aggregate model, are again lognormally distributed and IID.

Much of the research in the previous section looked primarily at alleviating the bias present when using an *arithmetic parameter inheritance* (API) scheme.

1. Form group of items S where $\beta_j = \beta \quad \forall j \in S$.

2. Estimate β from the aggregate level equation:

$$\frac{1}{|S|} \sum_{j \in S} S_{j,t} = \Big(\frac{1}{|S|} \sum_{j \in S} \alpha_j \Big) \Big(\frac{1}{|S|} \sum_{j \in S} P_{j,t}^{\beta_j} \Big) \eta_t \qquad , \tag{3.3.3}$$

   where η_t is another (assumed separate) lognormal IID error process.

3. Estimate all α_j at the disaggregate level.

4. Disaggregate items inherit common β parameter from group.

5. Forecast.

However, we can see that linear aggregation of non-linear equations is not consistent with a linear aggregate equation:

$$\frac{1}{J}\sum_j \left(\alpha_j P_j^{\beta}\right) \neq \left(\frac{1}{J}\sum_j \alpha_j\right)\left(\frac{1}{J}\sum_j P_j\right)^{\beta}$$

In the past, much sales data has only been reported at the aggregate level, meaning that only arithmetic means could be used in practice. However, many businesses have now reached a stage where data is recorded on a more granular level, and hence we can consider geometric mean as an alternative. We can see that taking geometric means will result in estimates of β that are unbiased estimators, since:

$$\left(\prod_j \alpha_j P_j^{\beta}\right)^{\frac{1}{J}} = \left(\prod_j \alpha_j\right)^{\frac{1}{J}}\left(\prod_j P_j\right)^{\frac{\beta}{J}}$$

Hence we propose a *geometric parameter inheritance* (GPI) scheme, laid out as follows:

1. Form group of items S where parameter β_j is assumed to be common to all items ie. $\beta_j = \beta, \quad \forall j \in S$.

2. Estimate β from the aggregate level equation

$$\left(\prod_{j\in S} S_{j,t}\right)^{\frac{1}{|S|}} = \left(\prod_j \alpha_j\right)^{\frac{1}{|S|}}\left(\prod_j P_j\right)^{\frac{\beta}{|S|}}\eta_t \qquad (3.3.4)$$

3. Estimate other parameters at disaggregate level

4. Disaggregate items inherit the common β parameter from the group level.

5. Forecast.

Thus, the GPI scheme results in parameter estimates which are unbiased, and therefore preferable to those yielded through API, whilst also attempting to improve accuracy. This will be demonstrated in Section 3.5.

3.4 Bias in loglinear models

3.4.1 Bias quantification

Again, the start point is the simplified regression model considered in Equation (3.3.1). Let $1, \ldots, T$ be periods for which we have observed sales and price, and $T + 1$ be the period which we wish to model. Using Equation 3.3.1 and taking expectations of both sides we get:

$$\begin{aligned}\mathbb{E}[S_{j,T+1}|S_{j,1}, \ldots, S_{j,T},] &= \mathbb{E}[\alpha_j P_{j,T+1}^{\beta_j} \varepsilon_{j,T+1}] \\ &= \mathbb{E}[\alpha_j P_{j,T+1}^{\beta_j}]\mathbb{E}[\varepsilon_{j,T+1}]\end{aligned}$$

by independence of the $\varepsilon_{j,T+1}$ random variable. We know already that:

$$\mathbb{E}[\varepsilon_{j,T+1}] = \exp\left\{\frac{\sigma^2}{2}\right\} \qquad ,$$

and hence:

$$\mathbb{E}[S_{j,T+1}] = \mathbb{E}[\alpha_j P_{j,T+1}^{\beta_j}] \exp\left\{\frac{\sigma^2}{2}\right\} \qquad . \tag{3.4.1}$$

In the case of our *model*, in which it is assumed that α_j and β_j are known, we have $E[\alpha_j P_{j,T+1}^{\beta_j}] = \alpha_j P_{j,T+1}^{\beta_j}$. Therefore, we have the full expectation being simply:

$$\mathbb{E}[S_{j,T+1}] = \alpha_j P_{j,T+1}^{\beta_j} \exp\left\{\frac{\sigma^2}{2}\right\} \tag{3.4.2}$$

as detailed by Miller (1984). The equation shows that, in the case that our parameters are known, the bias can be quantified and the forecast could be debiased completely. However, in reality we must estimate the parameters α_j, β_j and σ^2, and this estimation introduces further bias into the forecast.

To quantify this additional bias, we need to examine the forecast function (FF):

$$\hat{S}_{j,T+1} = \hat{\alpha}_j P_{j,T+1}^{\hat{\beta}_j} \tag{3.4.3}$$

Again, we start by taking expectations, leading to:

$$\begin{aligned} \mathbb{E}[\hat{S}_{j,T+1}] &= \mathbb{E}[\hat{\alpha}_j P_{j,T+1}^{\hat{\beta}_j}] && (3.4.4) \\ &= \mathbb{E}[\hat{\alpha}_j]\mathbb{E}[P_{j,T+1}^{\hat{\beta}_j}] \quad . && (3.4.5) \end{aligned}$$

We make the assumption here that the expectation of the baseline and price elasticity parameters are independent. This is not strictly valid, as the value of $\mathbb{E}[\hat{\beta}_j]$ will vary as $\hat{\alpha}_j$ varies. However, we can now derive approximate expressions for debiasing the forecasts from Equation (3.3.1). To find expressions for the two terms on the right-hand side of this equation, we consider the estimation procedure. Ordinary least squares (OLS) estimation takes place after a log transform of the original model in Equation (3.3.1), yielding:

$$\log S_{j,T+1} = \tilde{\alpha}_j + \beta_j \tilde{P}_{j,T+1} + \tilde{\varepsilon}_{j,T+1}$$

where, for clarity of notation later on, we introduce the tilde symbol $\sim$ to denote a log transformation eg. $\tilde{\alpha}_j = \log \alpha_j$.

We note that $\tilde{\varepsilon}_{j,T+1}$ is Normally distributed with mean 0. Hence, we see that $\mathbb{E}[\hat{\beta}_j] = \beta_j$, that is that the OLS estimate $\hat{\beta}_j$ is an unbiased estimator of β_j. Hence $\hat{\beta}_j \sim \mathcal{N}(\beta_j, \sigma^2_{\hat{\beta}_j})$, where $\sigma^2_{\hat{\beta}_j}$ is the variance of the estimate $\hat{\beta}_j$. We now use a result from regression theory that states that, under matrix notation where $y = \beta\mathbf{X} + \varepsilon$:

$$Var[\hat{\beta}|X] = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1} \tag{3.4.6}$$

In our case, $\mathbf{X} = \mathbf{P_j} = (P_{j1}, \ldots, P_{j,T})$, and so we get the expression:

$$Var[\hat{\beta}_j] = \sigma^2_{\hat{\beta}_j} = \frac{\sigma^2}{\sum_{t=1}^{T} P^2_{j,t}} \tag{3.4.7}$$

Considering the expectation $\mathbb{E}[P^{\hat{\beta}_j}_{j,t}]$, we see that

$$\log\,\{P^{\hat{\beta}}_{j,t}\} = \hat{\beta}\,\log\,P_{j,t} \sim \mathcal{N}(\beta\,\log\,P_{j,t}, \sigma^2_{\hat{\beta}}\,\log\,{}^2P_{j,t}) \tag{3.4.8}$$

Hence, our original variable $P^{\hat{\beta}}_{j,t}$ must have a lognormal distribution with the same parameters. This leads to:

$$\mathbb{E}[P^{\hat{\beta}_j}_{j,t}] = \exp\left\{\beta_j\,\log\,P_{j,t} + \frac{\sigma^2_{\hat{\beta}_j}\,\log\,{}^2P_{j,t}}{2}\right\} \tag{3.4.9}$$

and since $Var[\hat{\beta}_j] = \sigma^2_{\hat{\beta}}$, we can simplify:

$$\mathbb{E}[P^{\hat{\beta}_j}_{j,t}] = P^{\beta_j}_{j,t}\exp\left\{\frac{\sigma^2\log^2 P_{j,t}}{2\sum_{i=1}^{T} P^2_{j,i}}\right\} \tag{3.4.10}$$

We also need to investigate $\mathbb{E}[\hat{\alpha}_j]$. Since $\mathbb{E}[\hat{\tilde{\alpha}}_j] = \tilde{\alpha}_j$, we have:

$$\mathbb{E}[\hat{\alpha}_j] = \mathbb{E}[\exp\{\hat{\tilde{\alpha}}_j\}] = \exp\left\{\tilde{\alpha}_j + \frac{\sigma^2_{\log\,\alpha_j}}{2}\right\} = \exp\left\{\,\log\,\alpha_j + \frac{\sigma^2_{\log\,\alpha_j}}{2}\right\} \tag{3.4.11}$$

$$= \alpha_j\exp\left\{\frac{\sigma^2_{\log\,\alpha_j}}{2}\right\} \tag{3.4.12}$$

where $\sigma^2_{\log\ \alpha_j}$ is the variance of $\log\ \hat{\alpha}_j$.

Bringing the analysis together, we have the following expressions:

$$\mathbb{E}[S_{j,t}] = \alpha_j P_{j,t}^{\beta_j} \exp\left\{\frac{\sigma^2}{2}\right\} \tag{3.4.13}$$

$$\mathbb{E}[\hat{S}_{j,t}] = \alpha_j \exp\left\{\frac{\sigma^2_{\log\ \alpha_j}}{2}\right\} P_{j,t}^{\beta_j} \exp\left\{\frac{\sigma^2 \log^2 P_{j,t}}{2\sum_{i=1}^{T} P_{j,i}^2}\right\} \tag{3.4.14}$$

This leads to the relationship between the expected value of the model and that of the forecast function being:

$$\mathbb{E}[S_{j,t}] = \mathbb{E}[\hat{S}_{j,t}] \exp\left\{\frac{\sigma^2}{2} - \frac{\sigma^2_{\log \alpha_j}}{2} - \frac{\sigma^2 \log^2 P_{j,t}}{2\sum_{i=1}^{T} P_{j,i}^2}\right\} \tag{3.4.15}$$

Thus, the forecast function in Equation 3.4.3 is biased, needing multiplication by the expression inside the exponent on the right-hand-side to be become unbiased. We see that there are 3 elements to this bias expression.

The first depends only on σ^2, with a higher value of that parameter causing a downwards bias in the forecast function.

The second, dependent on the parameter $\sigma^2_{\log \alpha_j}$, relates to the variability in the estimate of the logged baseline sales. Here, higher values of this parameter, relating to a more uncertain estimate, cause an upwards bias in the forecast function.

The third term is dependent on σ^2, again, but the element of real note here is the sum in the denominator. This sum gets higher as the data history gets longer, indicating that the longer the data history is, the less effect this term will have. We can see that as the data history becomes infinite, the term will reduce to zero.

Overall, a positive value inside the exponent will indicate that the forecast function is biased below the expectation from the data, and a negative value will indicate it is biased above.

3.4.2 Correction approximations

Based on the above analysis, we try to construct an approximate bias correction to the forecast function to create an improved forecasting method. We present three alternative correction ideas, with each idea becoming more detailed than the last by attempting to estimate more elements of the bias expression. The alternatives are:

- Approximation 1:

$$\hat{S}^{(1)}_{j,t} = \hat{\alpha}_j P^{\hat{\beta}_j}_{j,t} \exp\{\frac{s^2}{2}\} \tag{3.4.16}$$

  where s^2 is the sample variance. It assumes that both the variation in the logged baseline sales estimate is negligible, and that the data history is long enough to dominate the numerator in the second element of the bias.

- Approximation 2:

$$\hat{S}^{(2)}_{j,t} = \hat{\alpha}_j P^{\hat{\beta}_j}_{j,t} \exp\{\frac{s^2}{2}[1 - \frac{\log^2 P_{j,t}}{\sum_{i=1}^{T} P^2_{j,i}}]\} \tag{3.4.17}$$

  With this approximation, the data history is reintroduced into the approximation, but the logged baseline variance is still considered negligible.

- Approximation 3:

$$\hat{S}^{(3)}_{j,t} = \hat{\alpha}_j P^{\hat{\beta}_j}_{j,t} \exp\{\frac{s^2}{2}[1 - \frac{\log^2 P_{j,t}}{\sum_{i=1}^{T} P^2_{j,i}}]\} \exp\{\frac{-s^2_{\log_{\alpha_j}}}{2}\} \tag{3.4.18}$$

  where $s^2_{\log_{\alpha_j}}$ is the variance of the estimate $\hat{\alpha}_j$. With this, it is assumed that all 3 quantities are non-negligible and important to alleviating the bias.

Considering the 3 approximations, we see that Approximation 1 is equal to that of Miller (1984) and deals with the fundamental forecasting bias occurring from log-normal error structure. Approximations 2 and 3 are new contributions in this paper which extend the previous work by also accounting for the bias occurring from estimation in the log-domain, where after transformation the variance around the parameter estimates is not symmetric in the original units. Thus, all 3 estimates are consistent as the number of observations tends to infinity; however, for the relatively small data histories encountered in practice, we expect Approximations 2 and 3 to provide increasing value. It is anticipated that, overall, Approximation 3 is the least biased and hence it is our proposed procedure, which will be tested in Section 5.

3.4.3 Combination of parameter inheritance and bias approximations

Examining the conclusions of the previous section and this one, it is suggested that a geometric parameter inheritance (GPI) scheme and a bias correction approximation lead to improved parameter accuracy and forecast accuracy, respectively. A natural extension therefore is to consider consolidating the two procedures into one combined approach.

To go about this, we can look at the expectation:

$$\mathbb{E}[\hat{S}_{j,T+1}] = \mathbb{E}[\hat{\alpha}_j \cdot P_{j,T+1}^{\hat{\beta}}] \tag{3.4.19}$$

where the only change is dropping the subscript on the β to reflect that the GPI procedure is taking place. Using similar logic to that in Section 4, we can see that:

$$\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2_{\hat{\beta}}) \tag{3.4.20}$$

and $Var[\hat{\beta}|X] = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$ as before. However, since the estimation of $\hat{\beta}$ is taking place on the aggregate equation, the σ in this equation is not equal to the σ in the individual SKU equation. To avoid confusing the two, we relabel the aggregate-level variance σ_a and its estimate s_a. Similarly, we can see that $\mathbf{X}$ here is not the same; as the estimation of β is on the aggregate level, $\mathbf{X} = \Big(\big(\prod_j P_{j,1} \big)^{\frac{1}{J}}, \ldots, \big(\prod_j P_{j,t} \big)^{\frac{1}{J}} \Big)$. Substituting these two quantities into Equation 3.4.19 gives us the resulting procedure which combines GPI with the approximation.

3.5 Simulations

We now undertake simulations to investigate the performance of our proposed procedures. The simulations are run in three stages: firstly, the GPI procedure is examined, and then the forecast approximations from Section 3.4. The third stage illustrates where gains in bias and accuracy are made when the two methods are combined.

3.5.1 GPI simulations

We first undertake simulations to investigate the performance of the GPI procedure over simple SKU-level estimation. In each set of simulations, we generate a number of sales series from given parameters, and then compare the two methods in terms of how well they estimate the price elasticity parameter. The accuracy of the parameter estimates is assessed in Mean Absolute Error (MAE) to give an accurate measure of how each method does on average.

Firstly, in order to understand the responsiveness of the two procedures to some important quantities, we present a section of simulations where three quantities are varied individually. The data was generated according to Equation (3.3.1); thus, for these simulations we assume the true model. The quantities varied were: (i) the variance of the error process, varied over the values 0.5, 1 and 1.5 ; (ii) the proportion of periods experiencing a price cut, varied over $\frac{1}{20}$, $\frac{1}{10}$ and $\frac{1}{5}$; and the number of items grouped together for the GPI method, varied between 10, 30 and 50. When one of the other parameters was being varied, the default parameter values were 1, $\frac{1}{5}$ and 50 respectively. 200 simulations were made under each set of conditions; a common price elasticity β is used for each simulated SKU in the group. Here, $\beta = -1$; although other values were also experimented with, there was no significant qualitative difference in the results. 52 observations are generated for each SKU, a quantity chosen to reflect the recording of a year's worth of weekly sales.

We report MAE figures in Table 3.5.1 for the usual approach of individual estimation, referred to here as bottom-up (BU), and for the GPI method under different combinations of varying data variance, promotional frequency and the number of items. The first pair of tables shows that the GPI method has a lower average MAE than BU in all cases, and the difference is slightly larger in the directions of higher variance, whilst similar across promotional frequencies. This is as expected; both of these increase the difficulty of individual estimation of parameters, allowing for greater potential gain via aggregation. The second pair shows that as the number

BU		Standard deviation			GPI		Standard deviation		
		0.2	0.6	1			0.2	0.6	1
Promotional	1/20	0.125	0.375	0.625	Promotional	1/20	**0.118**	**0.354**	**0.591**
frequency	1/10	0.096	0.289	0.482	frequency	1/10	**0.090**	**0.269**	**0.449**
	1/5	0.079	0.236	0.393		1/5	**0.070**	**0.209**	**0.350**

BU		Promotional frequency			GPI		Promotional frequency		
		1/20	1/10	1/5			1/20	1/10	1/5
Number of	10	0.375	0.289	0.236	Number of	10	**0.354**	**0.270**	**0.209**
items	20	0.366	**0.284**	**0.230**	items	20	**0.359**	0.305	0.240
	30	0.368	**0.285**	**0.232**		30	**0.352**	0.307	0.244

BU		Standard deviation			GPI		Standard deviation		
		0.2	0.6	1			0.2	0.6	1
Number of	10	0.079	0.236	0.393	Number of	10	**0.070**	**0.209**	**0.350**
items	20	**0.077**	**0.230**	**0.384**	items	20	0.080	0.240	0.399
	30	**0.077**	**0.232**	**0.387**		30	0.081	0.244	0.406

Table 3.5.1: Parameter MAE figures for the BU and GPI methods under varying standard deviation, price cut frequency and number of items in the group.

of items increases, the GPI parameter estimates generally deteriorate. This can be explained since grouping a larger number of SKUs increases the chance that one SKU will be difficult to estimate, influencing the grouped estimate. The BU parameter estimates are also less accurate at lower promotional frequencies; this is because there are fewer promotional periods in the estimation sample. In the third pair of tables, we see again that increasing the number of items leads to worse performance for the GPI method, although the difference is quite slight. There doesn't seem to be any significant interaction effect between the standard deviation and number of items variables.

Whilst the GPI method is clearly an improvement on BU for a range of scenarios, the difference in MAE between the two methods is not as large as might be expected. This is due partly to GPI also showing a larger variability in accuracy than the BU method, a somewhat surprising result which is a consequence of the strong effect on the accuracy that outliers in a group have on the GPI method. Whilst in the BU method, one outlying estimate often hides among many other good estimates, in the GPI method an outlier biases the estimation of the grouped parameter, affecting all SKUs. This is due to the nature of taking the geometric mean, for instance of the sales observations; the product of all sales in a particular period is calculated, and one abnormally high value can cause that product to be multiplied by several times the value yielded in its absence. This effect is keenly felt even after taking the n-th root.

The results here have important implications for practice. In scenarios where the important factor is to gain more accurate parameter estimates overall, with a few

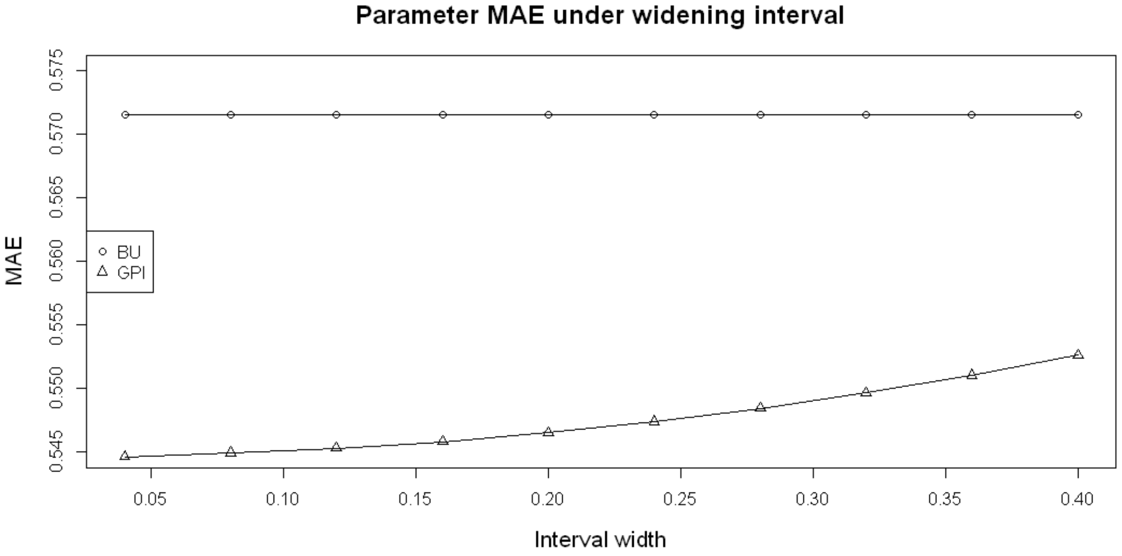


Figure 3.5.1: Lineplot showing the MAE of parameter estimates as the sampling interval widens.

outlying values being unimportant, the GPI method can be preferred. However, in scenarios where a few inaccurate parameter estimates lead to very costly consequences, it may be worth carefully considering their overall impact, as outliers can generally be expected from any data sample.

To explore the performance of GPI further, a second set of simulations was undertaken where the assumption of a common elasticity parameter across all SKUs was broken. Instead, elasticities for the items were drawn from a Uniform distribution within a specified interval, centered on the value -1; starting with an interval of [-1.02,-0.98], both boundaries of the interval were pushed outwards. The aim was to reflect the real world situation, in which elasticities are not known to the retailer without uncertainty, and that considering two separate SKUs to have exactly the same elasticity is somewhat unrealistic.

Figure 3.5.1 shows the effect of on parameter accuracy of widening the interval in which elasticities may be sampled from. It is apparent that the widening of the interval has a small effect on the accuracy of the GPI method, but the MAE figure for GPI lies below BU for all interval widths considered. It should be noted that the width of the interval is increased up to a value of 0.4, which can be considered fairly wide when, for example, very similar products which can be expected to behave in a similar way are grouped. This indicates that the GPI method seems to be robust to a violation of the assumptions and can thus be used to obtain more accurate parameter values over groups of similarly behaved SKUs with some confidence.

3.5.2 Forecast approximation simulations

We now evaluate the performance of the 3 approximations, along with the forecast function given in Equation 3.4.3, in simulations. By varying the conditions of the simulations in terms of the level of price cut in the forecast period, the variance of the data and the length of the data history, we compare Approximation 3 to the other 3 competitors in terms of forecast accuracy.

To assess forecast accuracy, a grid representing each different combination of variance and the proportion of regular price was created. The number of items, considered in the previous section, was not a relevant variable here since grouping does not occur. Variance was varied from 0.02 to 1, with a step-size of 0.02; proportion of regular price was varied from 0.3 to 1, also with a step-size of 0.02. At each grid point, 1000 simulations were undertaken, simulating from the model with corresponding variance with a data history length of 26 observations. Forecasts were calculated for periods with the corresponding price cut using each of the 3 approximations discussed and additionally the forecast function in Equation 3.4.3, and the results were used to create surface plots of the MAE and mean error (ME) incurred.

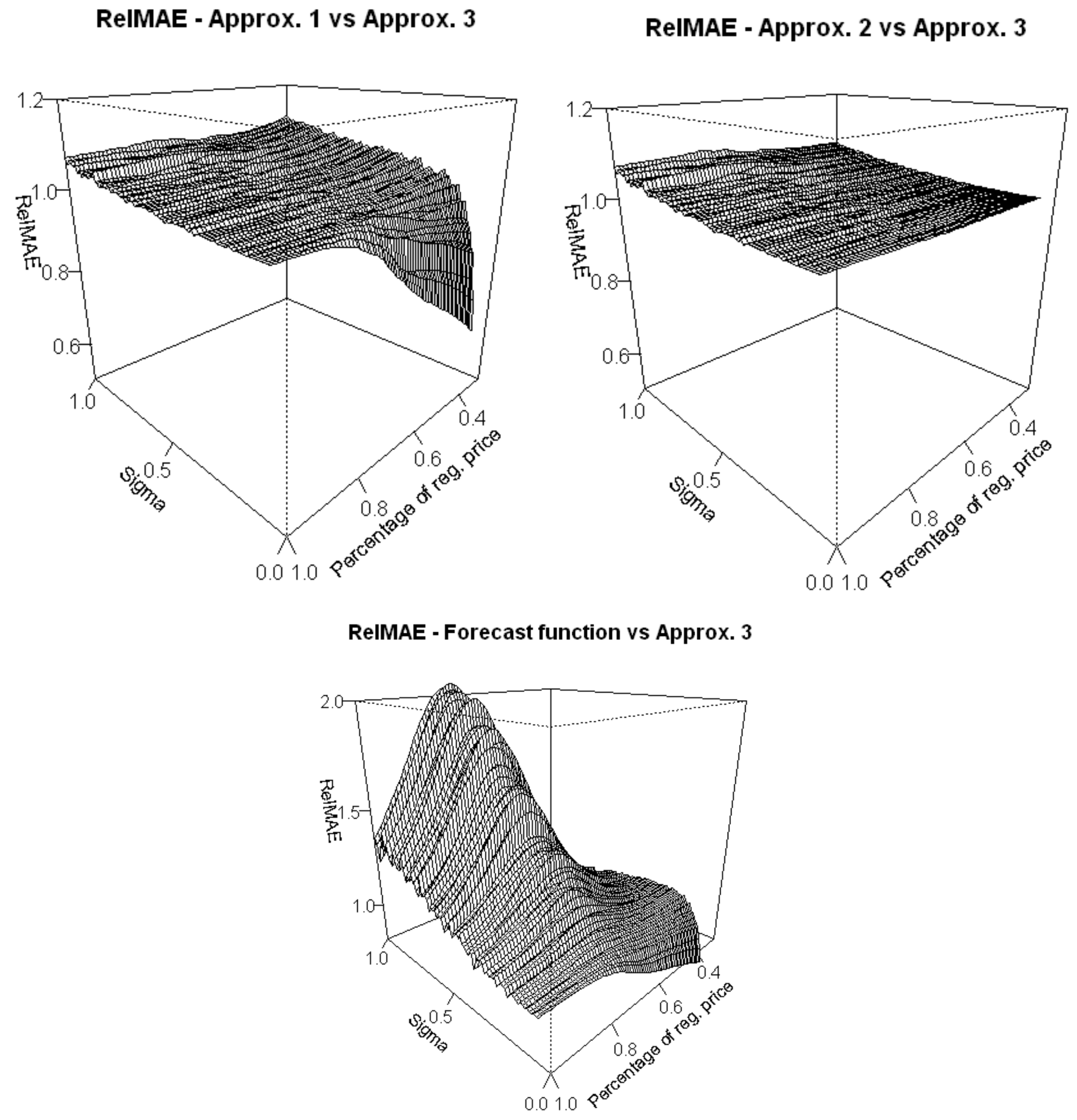
Figure 3.5.2 shows Relative Mean Absolute Error (RelMAE) figures for the forecast accuracy. The RelMAE of a method is defined:

$$\text{RelMAE} = \frac{\text{MAE}_{\text{method}}}{\text{MAE}_{\text{benchmark}}} \tag{3.5.1}$$

For ease of comparison, Approximation 3 (our proposed approximation) has been used to benchmark the other three methods. The 3 surface plots displayed represent the ratio of MAE figures of each of those other 3 methods to that of Approximation 3. Thus, a value of greater than 1 indicates that Approximation 3 has a lower MAE than the method compared, and vice versa. From the first two plots, it can be seen that Approximation 3 outperforms both Approximations 1 and 2 in most cases, with Approximation 1 performing slightly better in the unlikely scenario of very low variance and low proportion of regular price. In addition, the geometric mean of both surfaces is greater than 1, indicating an overall superior performance. This greater accuracy can be explained by the improvement in bias of the forecasts. The 3rd surface plot in this set demonstrates that Approximation 3 is far superior to the base forecast function in the majority of cases, with the closest results coming again when variance is extremely low. This reflects the fact that with a low variance, the bias in parameter estimation in the forecast function is amplified less.

Figure 3.5.3 show Relative Absolute Mean Error (RelAbsME) figures, defined by:

$$\text{RelAbsME} = \frac{|\text{ME}_{\text{method}}|}{|\text{ME}_{\text{benchmark}}|} \tag{3.5.2}$$



RelMAE - Approx. 1 vs Approx. 3

RelMAE - Approx. 2 vs Approx. 3

RelMAE - Forecast function vs Approx. 3

Figure 3.5.2: Surface plots showing the ratio of MAE between different forecasting functions, using Approximation 3 as the benchmark, varying over price cuts and variance. From top (i) Approximation 1 (ii) Approximation 2 (iii) Base forecast function. The geometric means of the RelMAE surfaces are: (i) 1.019 (ii) 1.018 (iii) 1.198 respectively.

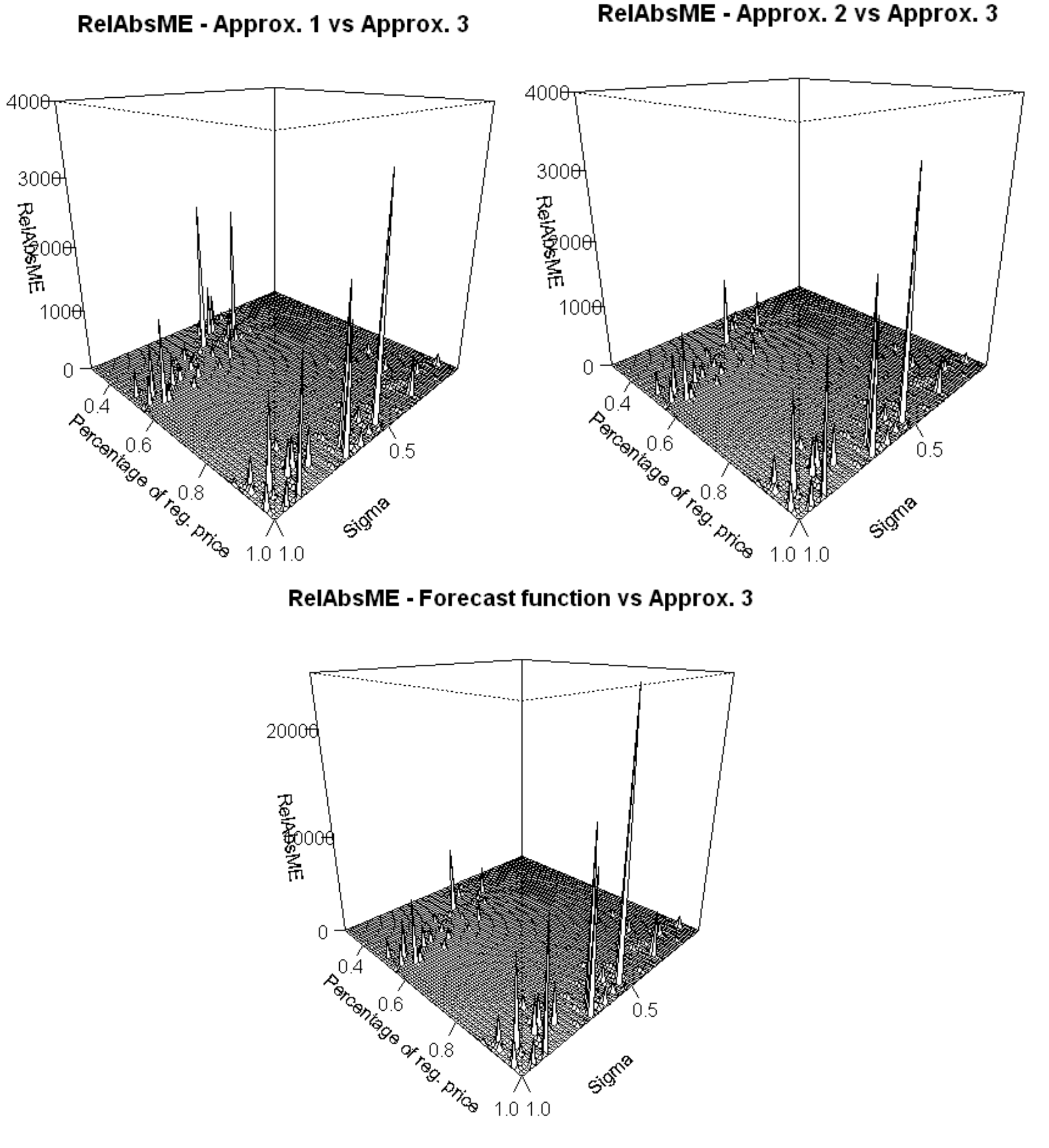


Figure 3.5.3: Surface plots showing RelAbsME between different forecasting functions, using Approximation 3 as the benchmark, varying over price cuts and variance. From top (i) Approximation 1 (ii) Approximation 2 (iii) Base forecast function. The geometric means of the surfaces are (i) 1.039 (ii) 1.042 (iii) 7.534 respectively.

Using the geometric means of the surfaces, we can see that the bias is lower for Approximation 3 than for the other methods, and that the improvement is mostly uniform across the surface; in other words, there is no strong interaction effect between these two variables in terms of bias. The spikiness of the plots shows that there are occasional cases where the bias for Approximation 3 is very low; these points are few enough that they can be considered as points of coincidence.

Simulations to assess the impact of data history length on forecast accuracy and bias were also conducted. To examine these simultaneously, we move from using absolute errors (ie. MAE) to squared errors (ie. mean squared error (MSE)). The change in error metric is to allow us to look at the bias-variance decomposition; it is well known that the MSE of a prediction algorithm can be broken down into bias (in terms of squared mean error), variance (of the predictions) and irreducible error (eg. Hastie et al. (2009)). Figure 3.5.4 shows the distribution of the forecast squared error (SE) of the four methods under different history lengths. The boxplots show that, as history length is increased, the SE of the three approximations all decrease on average, but that the distribution of squared errors at any one history length are broadly similar. The forecast function also performs competitively for the shortest data history considered, but falls away as the history is increased. Broadly, there seems little difference between the three approximations in terms of squared error; this can be contrasted with the surface plots, which showed Approximation 3 to be broadly superior in RelMAE. This finding is because squared errors penalise larger errors more harshly than absolute errors; as such, the absolute errors from Approximation 3 have a longer tail than those from the forecast function.

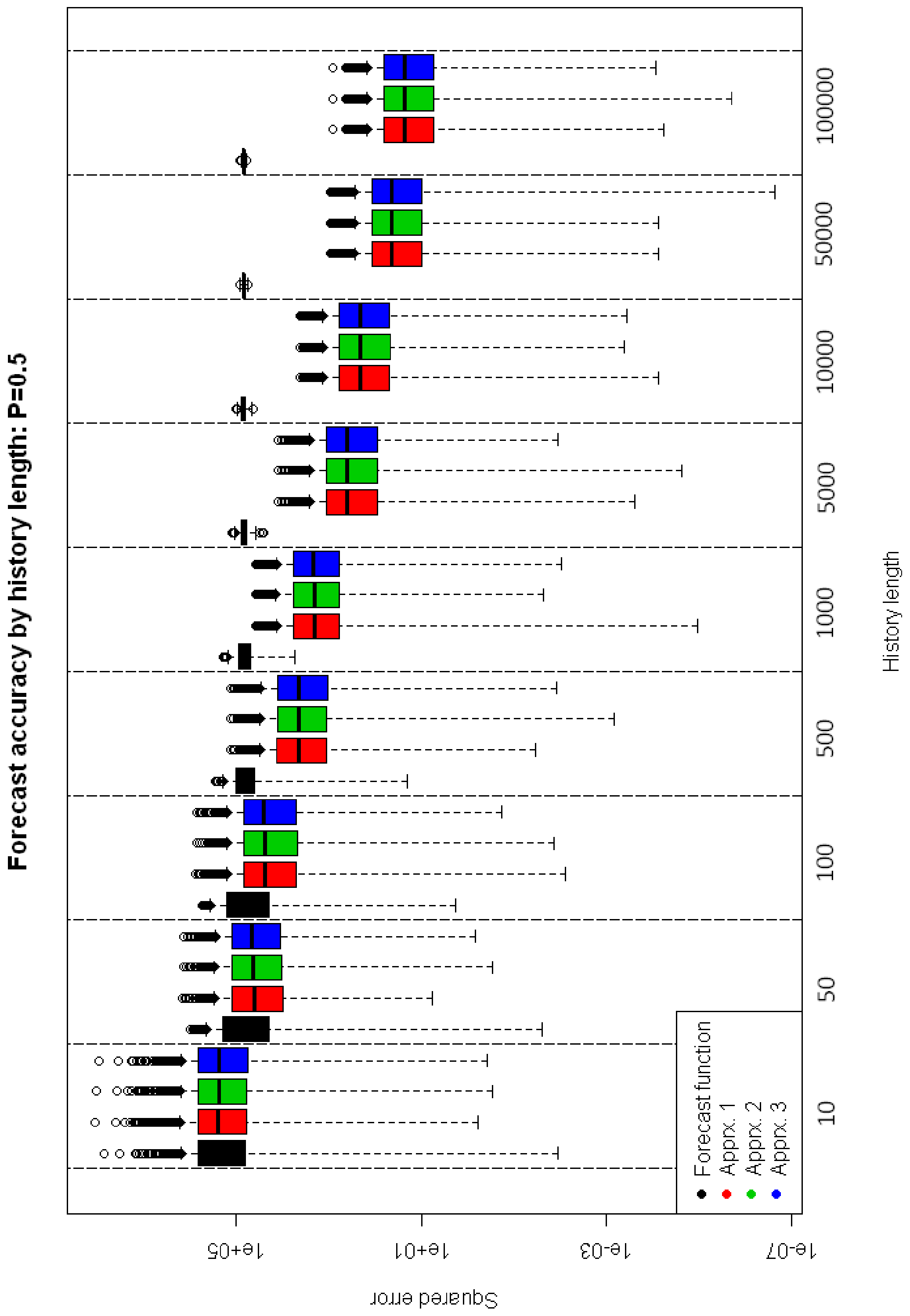


Figure 3.5.4: Boxplots showing the distribution of squared forecast errors when P = 0.5 for varying history lengths.

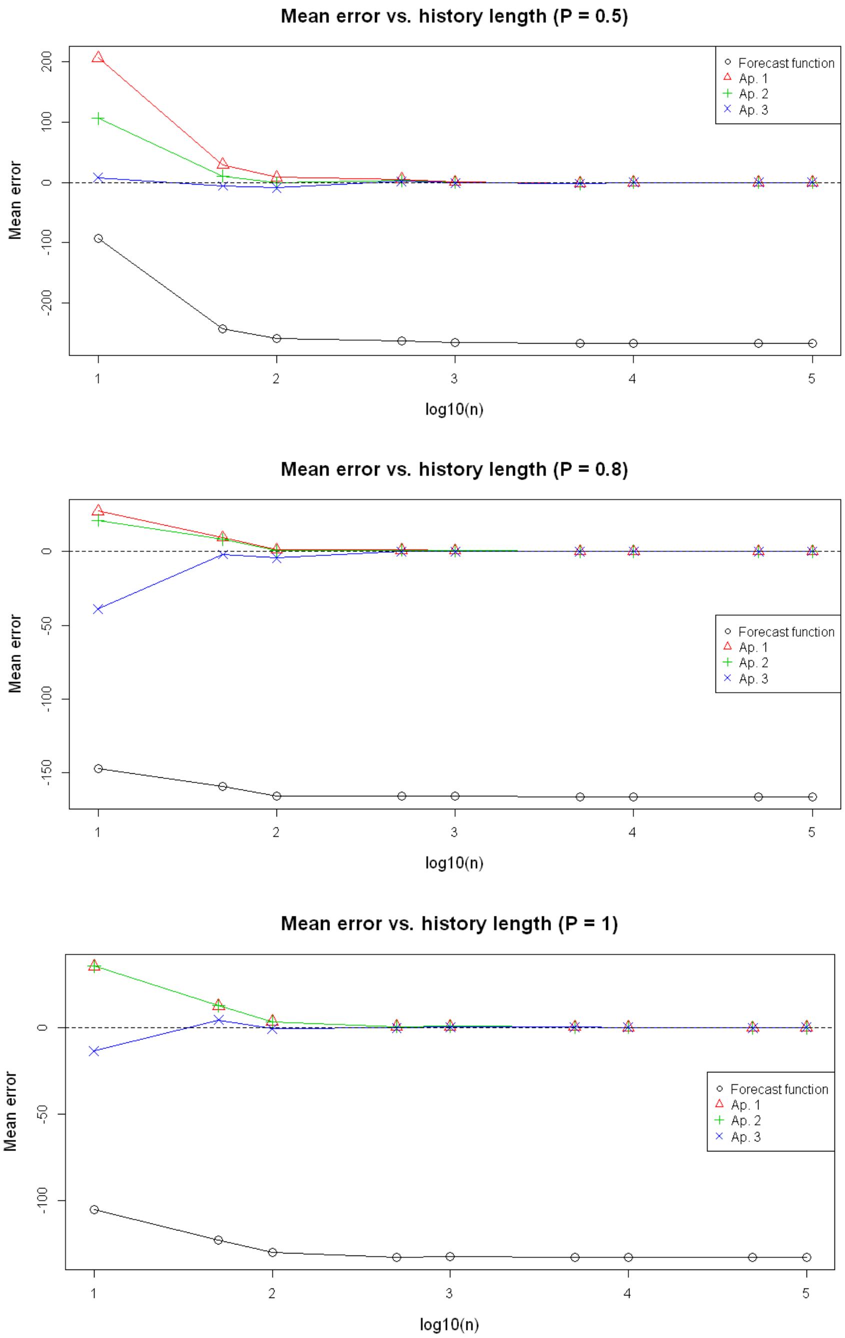


Figure 3.5.5: Line plots showing the mean error of each of the three approximations and the forecast function, compared to the theoretical average, varying over length of data history. From top (i) Price $\frac{1}{2}$ of regular price (ii) Price $\frac{4}{5}$ of regular price (iii) Full regular price.

Considering bias, the line charts in Figure 3.5.5 show the mean error (ME) over 1000 simulations for a fixed variance of 0.5 as the data history increases. Firstly, for each of the three approximations, the bias converges to zero as the data history becomes very large; this is as expected. Secondly, the average forecast from the forecast function converges to a negative value. This is because the forecast function is inherently biased, due to the multiplicative error structure. The differences between the predictions from the 3 approximations become clear when the proportion of regular price is lowered, and when the data history is short. We can see that Approximation 3 is clearly the least biased of the three methods below a history length of 100 when P = 0.5 or 1. At P = 0.8, it seems to be quite biased downwards for the lowest history length of 10, before becoming more unbiased once the history length is increased to 50. Approximation 2 appears superior in this special case, slightly outperforming Approximation 1. For all values of P, Approximations 1 and 2 are positively biased for history lengths of 100 and less. Although not perfect, it can be seen that the method least afflicted by bias is Approximation 3. We suggest this is because it takes into account not only the contribution of the history of price promotions in estimating the price elasticity β, but also uncertainty in estimating the baseline α and the asymmetric way in which that translates into the forecast via the log transformation. This is important, since data lengths of less than 100 are by far the most common situation encountered in practice.

Considering the line plots and boxplots together, we see that whilst Approximation 3 does not reduce the squared error of the forecasts measurably when compared with the other approximations, it does reduce the bias in terms of mean error, especially where the history length is less than 100, the most common situation in practice. Thus, we conclude that Approximation 3 strikes a good balance between bias reduction and an increase in forecast variance; it is especially a large improvement over the forecast function.

3.5.3 Combined simulations

Lastly, we present simulations examining whether combining the two procedures gives any benefit. We examine all four combinations possible by selecting pairs of either the base method of parameter estimation/forecast generation, or the procedure that was found to give the largest individual improvement in performance; those are as follows: (i) individual or bottom up estimation/ forecast function (BU/FF) (ii) GPI estimation/forecast function (GPI/FF) (iii) BU/forecast approximation 3 (BU/Ap3) (iv) the "combined" approach of GPI estimation/forecast approximation 3 (GPI/Ap3).

It has already been seen that the GPI method leads to greater parameter accuracy, so we focus on forecasts. In these simulations, a wide range of different parameter values were examined, covering (i) 5 values of σ, the standard deviation (ii) 3 history lengths, in weeks (iii) 3 values of the price elasticity β (iv) 3 frequencies of price cuts and (v) 3 different hierarchy sizes. Of these, the first two were found to be the biggest drivers of forecast performance and we report these results below. Interaction effects between pairs of variables were also considered, but no strong effects were found. For clarity, we also limit our reporting to forecasting periods where there is a price cut in effect (the specific case is P=0.5). For periods without a price cut, we found the combined GPI/Ap3 approach to outperform the other options, although forecast

(History length = 52)	BU/FF vs. GPI/Ap3			GPI/FF vs. GPI/Ap3			BU/Ap3 vs. GPI/Ap3		
Standard deviation	MSE	Var.	SME	MSE	Var.	SME	MSE	Var.	SME
0.2‘	0.970	0.733	2.193	0.990	0.962	1.190	0.976	0.763	0.364
0.4	0.941	0.637	2.907	0.965	0.853	1.946	0.962	0.748	0.445
0.6	0.903	0.504	2.941	0.929	0.698	2.114	0.943	0.724	0.478
0.8	0.859	0.361	2.667	0.887	0.529	2.001	0.923	0.692	0.493
1	0.814	0.234	2.252	0.839	0.365	1.742	0.899	0.651	0.498

Table 3.5.2: Ratios of MSE and components as standard deviation varies.

errors were relatively very small in all cases compared with the case of a price cut. The values of the other 3 quantities mentioned in the slice of reported results are: a price elasticity of -4, a 1 in 20 frequency of price cuts and a hierarchy size of 30 SKUs.

In each case, the bias variance tradeoff was used to assess performance: mean squared error (MSE), forecast variance (Var.) and the squared mean error (SME) of the forecasts are all reported. The results are presented in terms of comparison against the combined approach GPI/Ap3, by dividing the figures for each of the other 3 methods by those for GPI/Ap3. Thus, a value of greater than 1 indicates the combined approach is superior in a given quantity than the comparison method, and a value of less than 1 indicates it is inferior.

Tables 3.5.2 and 3.5.3 show the results for varying standard deviation and varying history length respectively. Looking at both tables, we first find that the combined approach is the approach with the highest variance in all cases. This is somewhat expected since the combined approach includes both Approximation 3, which includes the most terms in the forecast function, and the GPI parameter estimates, which were

($\sigma = 0.4$)	BU/FF vs. GPI/Ap3			GPI/FF vs. GPI/Ap3			BU/Ap3 vs. GPI/Ap3		
History length	MSE	Var	SME	MSE	Var	SME	MSE	Var	SME
52	0.941	0.637	2.907	0.965	0.853	1.946	0.962	0.748	0.445
26	0.942	0.719	3.012	0.962	0.857	3.257	0.977	0.842	1.168
13	0.806	0.543	1.577	0.928	0.853	4.255	0.855	0.628	3.425

Table 3.5.3: Ratios of MSE and components as history length varies.

found to be slightly more variable. The forecast variance of the combined method increases, relative to the other methods, as the data becomes more variable. Considering the bias next, we see that the combined approach is very much superior on this metric to the two methods which use the base forecast function. In the case of the BU/Ap3 method, the combined approach is superior in the cases where the history length is half a year or less. The biases of both of these methods compared with the base forecast function is very low, although the failure of the improved parameter accuracy of the combined approach to translate into lower bias can be attributed to the fact that the errors in the data are non-Gaussian. Finally, in terms of MSE we see that the combined approach is slightly inferior in all cases, indicating that the variance is a bigger constituent part than the bias. However, especially for the smaller magnitudes of σ, the MSE of the combined approach is not drastically worse. If reducing the bias is a prime objective for the retailer, it is possible to recommend the combined approach. The relative MSE of the combined approach does worsen with increased variance and shorter history length; this is because a more complex procedure is more difficult to undertake with less stable data. For the higher values of σ considered in these simulations, a visual inspection of the data series indicates that the signal-to-noise is very low and the price cut periods are approaching being indistinguishable from regular periods.

3.6 Conclusions

SKU-level forecasting and the estimation of SKU-level parameters is a vital issue for all retailers. In this paper, we have identified two major sources of bias that can occur in loglinear sales modelling: parameter estimation through linearly aggregated sales and the estimation of the parameters in the log domain. To combat these issues, a geometric aggregation scheme for parameter estimation and an adjustment that approximately de-biases forecasts produced from the model have been proposed. Through simulations, we have shown that the geometric aggregation scheme produces more accurate group parameter estimates, including situations where the true individual parameter values are different within a modest range. The forecast approximation eliminates substantially more bias from the final forecasts than previously used approximations by accounting for asymmetric uncertainty in estimating the model parameters in the log domain and translating them back, and is competitive in terms of squared forecast error. Both of these procedures can be used individually to either improve parameter accuracy or to de-bias forecasts, respectively. A natural combination of the two proposed procedures was also derived, which under certain circumstances can provide significant bias reduction in return for a modest increase in forecast mean squared error.

This work focusses on theoretical and simulatory evidence for the methods proposed, and naturally an empirical study would be the next step. We believe that an inventory simulation would be a promising addition to this study, since in practice these types of improvements would be utilised by inventory managers to determine stock levels. Such a study would focus on whether the reduced bias in our methods would compensate for increased variance in terms of stock performance. There is evidence in the literature that forecast bias can have a greater impact on organisational cost than forecast variance (eg. Sanders and Graman (2009)), although this depends on the existing magnitude of bias and its ratio with variance. Given that the suggested approach reduces bias by a considerable factor, it could provide value especially for items with a short history length.

The analysis in this paper also calls into question the mechanisms for hierarchy construction within retail forecast systems themselves. The GPI procedure, for instance, assumes a common value of price elasticity of demand for all items within a group, but retailers more often construct hierarchies of SKUs that are contingent on other factors, such as characteristics of the product. A holistic study looking at the costs and benefits of restructuring retail hierarchies in line with forecasting procedures would be of interest.

Chapter 4

The inventory performance of bias-corrected sales forecasts

Abstract

In this paper, the inventory performance of a new approximation for debiasing forecasts from loglinear promotional sales models is investigated. Through a simulation study, the new approximation is demonstrated to reduce inventory holding costs at the expense of achieved service level, at a rate which is favourable under some circumstances, including when demand histories are short. Furthermore, the properties of the forecasts themselves are linked through to the dimensions of inventory performance. We find that the biggest improvements in holding costs occur where bias reduction is greatest, despite no large difference in either forecast variance or accuracy. The under-discussed issue of calculating safety stock for promotional periods with few historical examples is addressed as a side-issue, with a simple heuristic to pool forecast

errors, combined with a recently proposed density-estimation routine demonstrating improved performance over the standard approach.

4.1 Introduction

Inventory management is a vital operation for any retailer, and the central task in any inventory management system is forecasting demand for the various products that are offered. Short-term forecasts on the stock keeping unit (SKU) level are required to inform decisions on when to reorder stock and how much of each SKU to order. Underestimating how much stock is needed can lead to stockouts which, in addition to the obvious lost revenue opportunities, can have a substantial impact on customer perceptions of the retailer, and more. Equally, overestimating stock requirements can be costly, as holding stock ties up capital and typically requires both administration and physical space.

Clearly, more accurate forecasts of demand are generally desirable, and forecast accuracy has a profound impact on all aspects of inventory management. As a consequence, forecasting for inventory planning is long-studied. Gardner (1990) analysed how forecasting impacted inventory decisions in a distribution system, concluding that models with different forecast accuracy was able to be tracked to substantial differences in terms relevant to the inventory managers, and hence had an impact on the amount of investment put in to achieve customer service level targets. Syntetos et al. (2009) provide a comprehensive review of advancements spanning 50 years of research in this area, concluding that there are many opportunities for further research, including more studies that bridge together the areas between forecasting and inventory planning more tightly. Ali et al. (2012) undertake a theoretical analysis of a particular demand processes which analyses the association between forecast accuracy and inventory holdings from the perspective of information sharing in the supply chain, demonstrating that their results are valid with an empirical evaluation. Syntetos et al. (2010) also argue that when forecasting is done in the context of inventory management, the evaluation should always be done with respect to the implications for stock control, since the linkages between forecast accuracy and inventory performance metrics can be complex. Of further interest is the relationship between inventory performance and the properties of the forecast themselves. Sanders and Graman (2009) studied forecast bias and variance in a warehouse scenario, and found that overall, forecast bias had a significantly greater impact on organisational costs than forecast variance. Additional research to build on their results would be of great value in providing insight into the link between forecasts and inventory performance.

Kourentzes (2013) examined the performance of neural networks for intermittent demand forecasting by tracking both forecast accuracy and inventory performance metrics. Whilst performing poorly in the former, the opposite was true for the latter; the conclusion was that both should play a part in these types of studies. A similar evaluation approach was used by Kourentzes (2014) in assessing the performance of new metrics for parameter optimisation in intermittent demand models, and in Teunter and Duncan (2009), where a large empirical evaluation on intermittent demand forecasting methods is carried out. However, most studies which describe SKU-level retail forecasting approaches in the retail domain evaluate performance solely with regard to conventional accuracy metrics (see eg. Ali et al., 2009; Ma et al., 2016). Given that inventory metrics are more closely related to the actual operations that take place within an organisation, more focus in this area is also needed.

This paper aims to examine the inventory performance of sales forecasts that have been debiased through a new corrective method, developed in previous work. Particular attention is given to the case of promotional forecasts, research on which has been relatively light. A particular promotional sales model in the form of a loglinear regression model is the focus, since it is a model which is often used in practice and important to practitioners. The new approximation is compared with previous approaches through inventory simulations, and it is established that the new approximation generally reduces holding costs at the expense of reducing the achieved service level. Moreover, this tradeoff is most favourable when histories are short, among other circumstances. We demonstrate that in some situations, the tradeoff frontier possible with the new approximation dominates that of other forecast methods.

The results also show that the situations in which the forecast debiasing afforded by the new approximation is proportionally greatest, generally correspond with those situations where the holding costs are reduced most, relative to the size of decrease in achieved service level. Furthermore, this improvement comes even whilst forecast variance and accuracy remain relatively stable. The conclusion for inventory managers is that the reduction in forecast bias is the key driver of possible inventory savings, and that a greater reduction in bias can result in greater possible savings.

The structure of this paper is as follows. In Section 2, we overview the simplified sales model and provide the forecast modifications to be examined. In Section 3, we look at the setup of the inventory simulation and the assumptions that we make in doing so. Section 4 is the main body of the paper and deals with the results from the simulation study. Finally, Section 5 concludes the work and offers future areas of study.

4.2 Methods

4.2.1 Sales model

A simplified loglinear model is chosen in order to isolate the relationship between forecast and inventory performance and put it in clear focus. The model chosen considers sales as being dependent on a time-independent baseline for non-promotional sales, multiplied by a price discounting effect, which is also dependent on the price elasticity of demand for the SKU in question. The model equation is:

$$S_t = \alpha \tilde{P}_t^{\beta} \varepsilon_t \quad , \tag{4.2.1}$$

where S_t denotes the sales in period t, α represents baseline sales, $\tilde{P}_t$ represents price relative to the baseline price, β represents price elasticity of demand and ε_t are log-normally distributed errors $\ln\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$, independent and identically distributed (IID) across both the time and item dimensions. We assume that α and β are estimated separately for each series, and also assume that ε_t is independent from $\tilde{P}_t$.

4.2.2 Forecasting methods

Three forecasting methods are considered for the model. We use a benchmark method and then consider two approximations; each represents a refinement upon the previous approach. Those methods are discussed below.

Forecast function

The forecast function for the sales model in Equation 4.2.1 is:

$$\hat{S}_{t+h} = \hat{\alpha} P_{T+h}^{\hat{\beta}} \quad , \tag{4.2.2}$$

where $\hat{S}_{T+h}$ denotes the h-step ahead sales forecast for time period $T+h$ (conditioned on the known demand up to period T), $\hat{\alpha}$ the estimate of the baseline sales, $\hat{\beta}$ the estimate of price elasticity, and P_{T+h} the price as a proportion of the regular non-promotional price, for the forecast horizon $T+h$.

The forecast function, whilst intuitive, and representative of current practice in some retailing applications, yields biased sales forecasts for two reasons: (i) the expected value of the forecast function is not equal to the expected value of the sales, given that the error distribution of the sales model is asymmetric, and (ii) the parameter estimates $\hat{\alpha}$ and $\hat{\beta}$ are computed in the log-domain, where the estimates are unbiased; that property is not retained once the parameters are transferred back into the original units. These two arguments are expanded upon in Waller et al. (2019). We use it here to benchmark the inventory performance of the following two modifications.

Miller approximation

Based on the forecast function from the previous section, the approximation of Miller (1984) introduces a correction to reduce the bias of the forecasts, accounting for the lognormal error distribution in the sales model. The equation for the Miller approximation is:

$$\hat{S}_{T+h} = \hat{\alpha} P_{T+h}^{\hat{\beta}} \exp\{\frac{s^2}{2}\} \quad , \tag{4.2.3}$$

where $\frac{s^2}{2}$ represents the sample estimate of the variance parameter in the error distribution of the loglinear sales model. This approximation is also representative of some current practices and has been used in some loglinear sales models in the literature (Wittink et al., 1988), (Cooper et al., 1999) and yields forecasts with reduced bias compared with the forecast function, at the expense of forecast variance.

Estimation bias-correction (EBC) approximation

The third approximation introduces further correction terms to account for bias in the estimation of parameters due to estimation in the log domain.

$$\hat{S}_{T+h} = \hat{\alpha} P_{T+h}^{\hat{\beta}} \exp\left\{\frac{s^2}{2}\Big[1 - \frac{\log^2 P_{T+h}}{\sum_{i=1}^{T} P_i^2}\Big]\right\} \exp\left\{\frac{-s^2_{\log_\alpha}}{2}\right\} \tag{4.2.4}$$

where $s^2_{\log_\alpha}$ is the variance of the estimate $\hat{\alpha}$. It can be seen that, in addition to the previous terms, the adjusted forecast is dependent upon the observed history of price cuts, the price cut for the current period and the standard error in estimation of the baseline sales. The derivation can again be found in Waller et al. (2019). This approximation further reduces the bias in the forecast, and is most effective at doing so when the data history is short.

4.3 Inventory setup

In this section, the choices made in the inventory setup are described, along with the rationale for those decisions. The key assumptions are summarised in Table 4.3.1.

4.3.1 Type of inventory process

The inventory process used in the simulations is an (R, S) periodic-review, order-up-to system (see eg. Silver et al., 2017). We set the periodic review time R to 1 period, making a replenishment decision at every period to reflect the daily inventory checks that take place in a large number of retailers with relatively fast-moving SKUs such as fresh food retailing (Minner and Transchel, 2010). Furthermore, we assume that stock once delivered is immediately available to consumers, with no delay in moving stock from an intermediate storage facility to the retail floor.

In dealing with lost sales, the approach adopted in these results was to treat all demand that cannot be immediately fulfilled from stock as being completely lost, with no portion of that demand deferred to later periods. This is closer to reality than complete backordering for retail applications (Gruen et al., 2002), and as such is a common assumption made in research (eg. Van Donselaar et al., 1996; Kapalka et al., 1999). Choosing lost sales rather than backordering also allows for the primary effect on inventories of the forecasting procedure to be isolated more clearly.

4.3.2 Service target and calculation of safety stock

To calculate the order-up-to level S, the two most common service targets used in practice are the cycle-service level (CSL), which is defined as the proportion of periods in which all demand is satisfied, and the fill rate, which is the proportion of overall demand that is satisfied. These are the two standard measures Syntetos et al. (2010), and both of these have their strengths and weaknesses. However, the inventory mechanisms to turn a target fill rate into a replenishment ordering decision system are much more complex, and there is a wariness in the literature of taking it on. Zipkin (2008) warns that the study of discrete-time lost sales systems with constant lead times and stochastic demand is difficult. Moreover, it seems that many important questions are not settled. In calculating safety stock for such inventory systems, Silver and Peterson (1985) argue that target fill rates for complete backorder systems are adequate stand-ins for lost sales systems, but more recent research (van Donselaar and Broekmeulen, 2013) disagrees, arguing further approximations are needed.

Experimentation with using the fill rate yielded significant discrepancies between the target and achieved service levels in many cases, rendering it unappealing. In response to this, the CSL target measure is chosen as a more practical and reliable choice; it is by far the most often used of the two, and the methodology here can more easily accommodate non-Gaussian forecast error distributions (to be explained in the next paragraph).

When using CSL as a target service measure, there are a number of ways in which the resulting safety stock can be calculated. However, some of these assume that the distribution of forecast errors is Gaussian, such as the method detailed by Silver et al. (2017). The sales model here on which forecasts are based assumes a lognormal error process; accordingly, we decide to examine two alternative non-parametric approaches to calculating safety stock, described in Trapero et al. (2019). These two alternatives are:

1. *Empirical percentiles*: Here, the safety stock is simply linked to quantiles of the forecast error distribution.

   Using a CSL of c, we have that

   $$P(E_t < SS_t) = c \tag{4.3.1}$$

   where $E_t = S_t - \hat{S}_t$ is the forecast error (the sales S_t in period t minus the forecasted sales $\hat{S}_t$ for that period, and SS_t is the safety stock. In other words, we find the c-th quantile of the set of forecast errors, and order enough safety stock to cover an error of this size. For instance, with a CSL of 0.95, we order safety stock to cover 95% of forecast errors in the sample.

2. *Kernel density estimation*: The approach followed is the same as the proposed method in Trapero et al. (2019); namely, we use the Epanechnikov kernel and choose the optimal bandwidth:

   $$h_{opt} = 0.9AN^{-0.2} \quad , \tag{4.3.2}$$

   where N is the sample size, and:

   $$A = \min(\text{Standard deviation}, \text{Interquartile range}/1.34) \quad , \tag{4.3.3}$$

is a measure of the spread of the data.

One inherent limitation in using a non-parametric approaches is that small sample sizes can lead to variable outputs. This is, however, exacerbated further in the presence of promotions, since promotional periods are harder to forecast and the errors are larger than in non-promotional periods. In order to apply the safety stock calculations, therefore, the set of forecast errors must be subdivided into non-promotional and promotional errors, with only the relevant set used to determine empirical quantiles. For example, for a SKU with a history of 20 observations and a promotional frequency of 1 in 10, we have 18 forecast errors available to calculate safety stock for non-promotional periods, and just 2 for promotional periods. This sometimes tiny number of available errors for promotional periods is clearly a source of concern.

To the best of our knowledge, this issue has not been discussed in the literature. Therefore, a simple new heuristic is proposed to merge the sets of promotional and non-promotional errors. If non-promotional forecasts stem from the basic model:

$$S_T = \alpha \varepsilon_T \quad , \tag{4.3.4}$$

whereas promotional forecasts stem from the model

$$S_T = \alpha P_T^{\beta} \varepsilon_T \tag{4.3.5}$$

then the difference between these two model forms serves to multiply the set of non-promotional forecast errors up to the level of the promotional errors, ie. when calculating safety stock during a promotional period, we multiply all non-promotional forecast errors in the history by the quantity $P_T^{\hat{\beta}}$, where $\hat{\beta}$ is the current estimate of

Component	Assumption chosen
Inventory system type	(R,S) periodic review, order-up-to (R = 1)
Lost sales or backorders	Fully lost sales
Service target	Cycle service level (CSL)
Safety stock calculation	Forecast error empirical percentiles
Stock availability on arrival	Stock immediately available

Table 4.3.1: Choices for the components of the inventory system

price elasticity, and then merge this set with the promotional errors to form a larger set of errors. Similarly, for a non-promotional period we first divide the promotional forecast errors by the same quantity, before merging the two sets of forecast errors. This heuristic depends on the form of the model with regards to how the promotional variable is incorporated, but remains independent of the error process, which cancels out.

Pooling these sets of errors allows for the maximum possible number of forecast errors to contribute to the calculation of safety stock, and thus results in less variable outcomes. The impact of this heuristic on inventory performance is contrasted with the non-heuristic approach in Section 4.5.3.

4.4 Simulation setup and results

A simulation study is now carried out to investigate the performance of the EBC forecast approximation against both the Miller approximation and the benchmark.

The aims of the simulation study are the following:

1. To identify how the EBC approximation performs compared with the Miller approximation, and the benchmark, in terms of inventory performance metrics that are relevant to practitioners.

2. To enumerate different conditions under which the difference between the two approaches may be greater, or smaller.

3. To examine whether using a heuristic to pool forecast errors together improves inventory performance in the presence of price cuts and promotions.

4. To explore whether there is a possible link between the bias, variance and accuracy properties of the forecasts, and the resulting inventory performance.

500 data series are generated from the model in Equation 4.2.1 for each simulation run. The quantities we fix in these results are: (i) the lead time between a replenishment order and its arrival, set to 1 period, (ii) the review period, also set to 1 period, and (iii) the baseline sales α, set to 100. Also fixed by implication is the forecast horizon over which the forecasts are assessed, at 2 periods (review + lead time). Varying these conditions did not yield figures that alter the narrative of the results, and so they are omitted for concision.

The quantities that are varied are: (i) the price elasticity of demand, taking values -1,-2, and -4 (ii) the ‘noise’ parameter σ in the sales model, taking values 0.2,0.5 and 0.8 (iii) the proportion of promoted periods, which varies between 1 promotion every 5 periods (0.2), 1 in 10 (0.1), and 1 in 20 (0.05), and (iv) the history length, which

Origin	$\beta = -2$, $\sigma = 0.5$, Promo. prop. = 0.1, Hist. length = 20
Dimension	**Alternative values**
Elasticity	-1, -4
Noise	0.2, 0.8
Promo. proportion	0.05, 0.2
History length	10, 30

Table 4.4.1: Showing the four different dimensions across which we vary parameters in the simulation study. The origin represents the parameter combination sitting in the middle of the range in each direction, whilst the other rows show the higher and lower alternatives for each dimension

is varied between 10 periods, 20, and 30. In addition, for each set of parameters we average the results obtained across 3 different target service levels: 90%, 95% and 99%. Table 4.4.1 displays the selected parameter values.

The choices of values of price elasticities, noise, and promotional frequencies are intended to represent the range of these conditions that may be experienced in practice. For example, the promotional frequencies selected range from roughly once a month to roughly twice/three times a year, if the sales frequency is taken to be weekly. The noise parameter values were chosen by visual assessment of the series generated, whilst the elasticities were chosen partly by reviewing literature and real-world data. The history lengths chosen are on the shorter end of what is typically encountered in practice. For longer histories of 50 or more, the difference between the two approaches is expected to narrow considerably, as sufficient data to estimate the parameters more accurately means that there is much less need for any approximation. Therefore, it is of most interest in this situation to analyse the situations where the most difference may occur. With regard to the service level targets, we choose high values of 90% or more, since higher service targets of this level are generally the most desirable in retailing.

The main inventory metrics utilised to judge performance are:

1. Average on-hand inventory (OHI), defined as:

$$\text{Av. OHI} = \frac{1}{T}\sum_{t=1}^{T}\frac{\text{Stock at beginning of period } t + \text{Stock at end of period } t}{2} \tag{4.4.1}$$

   where T is the total number of periods in the test set. The assumption here is that demand comes at a constant rate throughout the period, and therefore the average inventory held during that period is the midpoint between the starting and finishing points.

   Since average OHI is somewhat dependent on the level of the series and hence the choice of parameters, we introduce the following relative measure, termed the *relative average on-hand inventory* (RelAvOHI):

$$\text{RelAvOHI}^{f_i} = \Big[\prod_{t=1}^{T}\Big(\frac{\text{Av. OHI}_t^{f_i}}{\text{Av. OHI}_t^{\text{b}}}\Big)\Big]^{\frac{1}{T}} \quad , \tag{4.4.2}$$

   where f_i is a specified forecast method (here, either the Miller or EBC approximation) and Av. OHI_t^b is the average OHI for the benchmark in period t. We use the geometric mean rather than the arithmetic since it is known to treat ratios of greater and smaller than 1 more symmetrically (Davydenko and Fildes, 2013). Whilst other references, such as the paper just mentioned, apply geometric averaging for evaluation purposes, the application to on-hand inventory is a new one to our knowledge.

2. Service level difference, defined as:

$$\text{SL diff.} = \text{CSL (achieved)} - \text{CSL (target)} \quad , \tag{4.4.3}$$

Ideally, the service level difference will be close to zero, although it is common that for high service level targets, the achieved service level rarely reaches the target level. The measure is not always symmetric; from an organisation's perspective, overperformance may be preferred to underperformance (this is more common than vice-versa). This caveat should be noted whilst, for simplicity, in these results values closer to zero will be indicated as superior.

The forecast performance metrics, to be used later on in the results, are:

1. Forecast *bias*, expressed as the mean error (ME):

$$\text{ME} = \frac{1}{T}\sum_{t=1}^{T} E_t \tag{4.4.4}$$

2. Forecast *accuracy*, expressed as the root mean squared error (RMSE):

$$\text{RMSE} = \sqrt{\frac{1}{T}\sum_{t=1}^{T} E_t^2} \tag{4.4.5}$$

3. Forecast *variance*, the variance of the forecasts themselves:

$$\text{Var} = \frac{\sum_{t=1}^{T}(F_t - \bar{F})^2}{T-1} \tag{4.4.6}$$

The errors E_t used in these metrics are cumulative with respect to the forecast horizon (2 periods) mentioned earlier in the section. We note that whilst forecast accuracy and bias are commonly used to evaluate forecast performance, the use of forecast variance is more unusual. In this setting, however, forecast variance possibly feeds into inventory performance via its influence on calculating the order-up-to level S. More variable forecasts may lead to S fluctuating more rapidly, with consequences for stock levels.

4.5 Results

4.5.1 Safety stock calculation method

To facilitate the presentation of the results, the overall performance of the three different safety stock calculation methods is presented to demonstrate which is the most promising. Figure 4.5.1 shows violin plots of the distribution of the service level difference across all parameter combinations, for each of the three forecast methods combined with each of three methods for estimating safety stock: the non-heuristic method (N), heuristic method (H) and the kernel density estimation with heuristic (KDE) method. We can see that, for each forecast method, the KDE approach achieves a far smaller service level difference than both the non-heuristic and heuristic-only approach. Whilst the heuristic approach improves significantly on non-heuristic, the subsequent improvement from adding in the KDE is even more significant. We attribute the relatively superior performance of the KDE methods to the greater detail in estimating the distribution of forecast errors, compared with linear interpolation between points used in the empirical quantiles approach.

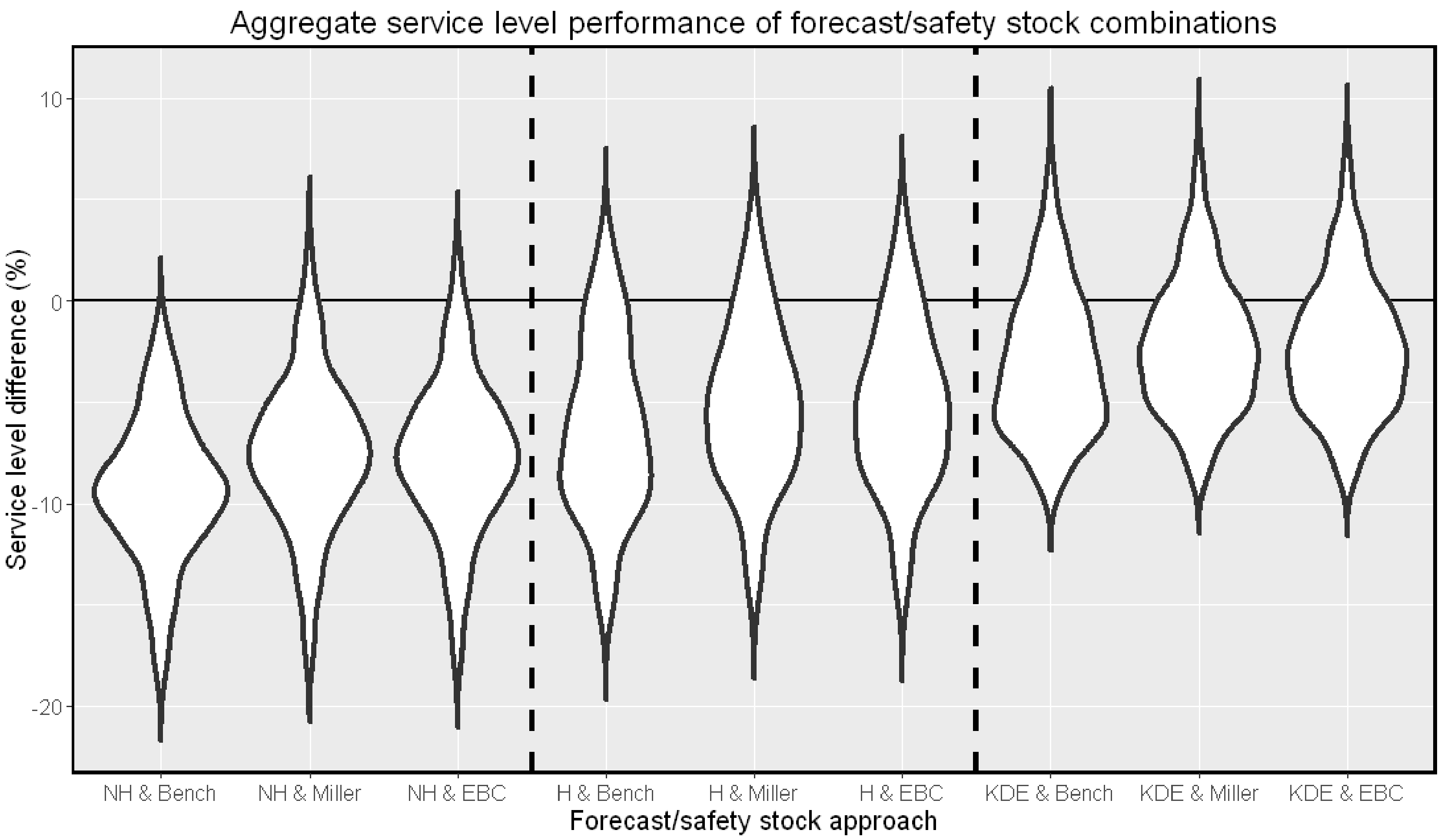


Figure 4.5.1: Aggregate performance over all 243 parameter sets of the 9 forecast method/safety stock calculation method combination.

In view of these results, the KDE approach is the only safety stock calculation method that is considered going forward.

4.5.2 Varying parameter combinations

The service level difference and relative average OHI figures are now presented. Due to the large number of possible parameter combinations, only a section of these are presented in the main body of the paper. To facilitate the description of which combinations are selected, Table 4.4.1 is referenced. The combination of the four parameter values representing the middle in each category is taken as the 'origin' of the parameter space, which can be imagined as having four dimensions with the alternative values given. In order to fully explore interaction effects between dimensions whilst keeping the results concise, results are presented for the 33 parameter combinations where we have moved away from the origin in at most 2 dimensions; these results are presented below in Table 4.5.1. Results for the remaining 48 combinations can be found in Appendix B.

Table 4.5.1: Service level difference and relative average OHI simulation results for Miller and EBC under varying conditions.

Parameter combinations				Service level difference		RelAvOHI	
Elasticity	Sigma	Promo. freq.	History	Miller	EBC	Miller	EBC
-1	0.2	0.1	20	**-5.30**	-5.35	1.021	**1.019**
	0.5	0.05	20	**-4.51**	-4.58	1.070	**1.064**

Continued on next page

Continued from previous page

Parameter combinations				Service level difference		RelAvOHI	
Elasticity	Sigma	Promo. freq.	History	Miller	EBC	Miller	EBC
		0.1	10	**-6.16**	-6.35	1.078	**1.065**
			20	**-4.46**	-4.56	1.070	**1.064**
			30	**-3.71**	-3.77	1.066	**1.063**
		0.2	20	**-4.57**	-4.64	1.064	**1.058**
	0.8	0.1	20	**-2.82**	-2.93	1.109	**1.101**
-2	0.2	0.05	20	**-5.44**	-5.48	1.018	**1.017**
		0.1	10	**-7.82**	-7.92	1.017	**1.014**
			20	**-5.47**	-5.52	1.017	**1.015**
			30	**-5.21**	-5.24	1.016	**1.015**
		0.2	20	**-6.49**	-6.52	1.015	**1.013**
	0.5	0.05	10	-	-	-	-
			20	**-3.45**	-3.53	1.050	**1.047**
			30	**-3.13**	-3.18	1.047	**1.044**
		0.1	10	**-3.99**	-4.16	1.062	**1.052**
			20	**-2.92**	-2.99	1.056	**1.048**
			30	**-2.10**	-2.15	1.046	**1.043**
		0.2	10	**-3.67**	-3.83	1.054	**1.042**
			20	**-2.31**	-2.41	1.044	**1.038**
			30	**-1.75**	-1.81	1.038	**1.035**
	0.8	0.05	20	**-0.09**	-0.10	1.069	**1.056**

Continued on next page

Continued from previous page

Parameter combinations				Service level difference		RelAvOHI	
Elasticity	Sigma	Promo. freq.	History	Miller	EBC	Miller	EBC
		0.1	10	0.01	**0.00**	1.091	**1.083**
			20	**-0.73**	-0.85	1.067	**1.054**
			30	**-0.93**	-0.99	1.068	**1.061**
		0.2	20	**-0.48**	-0.62	1.077	**1.067**
-4	0.2	0.1	20	**-0.63**	-0.67	1.007	**1.006**
	0.5	0.05	20	1.41	**1.35**	1.027	**1.020**
		0.1	10	1.84	**1.77**	1.041	**1.032**
			20	1.43	**1.40**	1.025	**1.024**
			30	1.72	**1.71**	1.017	**1.015**
		0.2	20	1.10	**1.06**	1.019	**1.016**
	0.8	0.1	20	3.23	**3.16**	1.036	**1.029**

The first observation is that, for most cases considered, the service level difference is negative, meaning that the achieved CSL falls short of the target CSL. This is largely because the target CSL values selected are high, and consequently difficulties in estimating the upper tail of the forecast error distribution with smaller than desired sets of errors becomes more visible. In addition, the RelAvOHI figures for both methods are greater than 1, meaning that both the Miller and EBC methods result in more stock being held than in the benchmark case. However, this is entirely as expected, given that the benchmark is biased towards underforecasting demand.

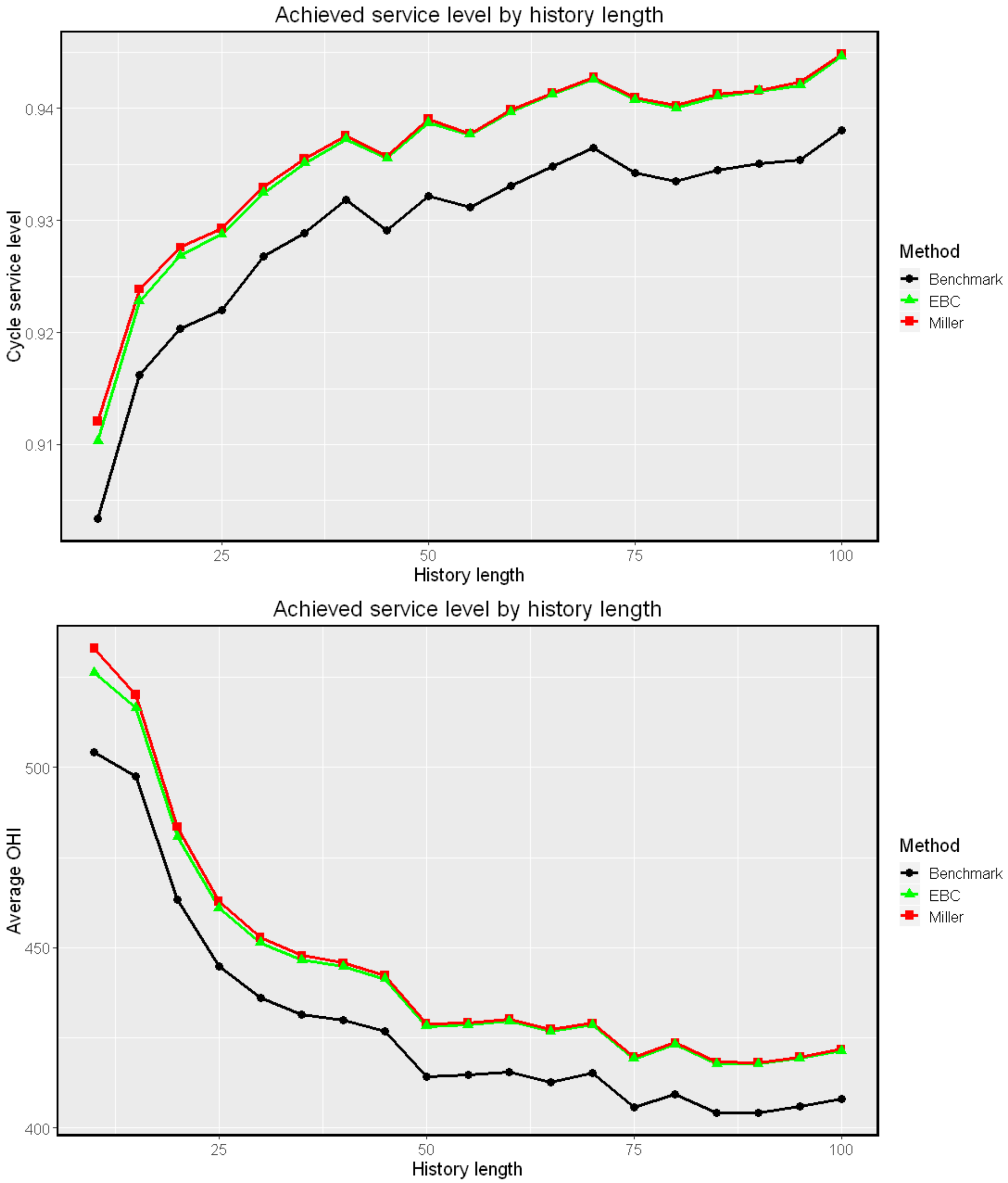


Figure 4.5.2: Lineplots illustrating the performance in terms of (i) CSL, and (ii) Average OHI, as history length varies of the forecasting methods. Results here are for a 95% target CSL.

Comparing the performance of the two approximations, it is clear that the RelAvOHI figures for EBC are superior to those for Miller across the board, meaning that EBC always holds less inventory, compared with Miller. The differences are generally small, but consistent. Given the form of the EBC correction and the fact that it adjusts the forecasts of Miller downwards in all normal situations, this is as expected. The situation is reversed when the figures for service level difference are viewed, with the Miller approximation covering more periods of demand than EBC. However, for most of the situations with the most elastic sales coefficient (-4), the service level difference is actually positive. The reason for this is the large size of the promotional uplift here, which heavily affects stock levels in regular periods. On occasion, the inventory mechanism orders more safety stock than is required; the result is that stock levels rise dramatically and only decay slowly over the following regular periods, boosting the achieved service level. In these cases, the EBC service level ends up closer to the target than the Miller service level.

Considering the effect of various parameter combinations on the relative performance of the methods, there is a general tendency for EBC to yield RelAvOHI figures that are relatively lower than those for Miller when the history length is shorter. This is in line with the expectation given the EBC formula, the denominator of part of which increases with the size of the history, reducing the magnitude of the forecast adjustment. Figure 4.5.2 shows how the EBC and Miller approximations converge in their performance as history length increases, both in service level difference and Av. OHI. Further, the graphs show how the performance flattens considerably as history length increases past 50.

In addition, the EBC also does best when the promotional frequency is highest, the value of σ is higher, and under higher elasticities. This is because the EBC provides benefit in promotional periods; when they occur in the evaluation sample with greater frequency, it is to the favour of EBC. EBC also provides a greater forecasting adjustment when estimation is less stable through the $s^2_{\log_\alpha}$ term, theoretically providing more benefit in those cases, whilst the observed effect from elasticity can be suggested as occurring due to the earlier-mentioned slowly decaying surplus stock levels from promotional periods; starting with lower safety stocks has cumulative benefits in such a scenario.

Summarising the above, we can see that there is something of a tradeoff between service level difference and RelAvOHI; EBC improves on the latter, but generally does worse on the former. There are conditions at which the tradeoff becomes more favourable or less favourable. Of interest is whether we can make a more definitive conclusion about this tradeoff.

4.5.3 Tradeoff between service level difference and average OHI

Tradeoff curves can be examined to discover the frontier of achievable service level difference and on-hand inventory duos by running simulations for the same parameter combination and varying the target CSL incrementally each time. Figure 4.5.3 shows one such tradeoff curve; that for the 'origin' case in our simulations, ie. a promotional frequency of 1 in 10, elasticity of -2, noise parameter of 0.5 and history length 20. The points plotted represent the results for simulations run with target CSLs increasing incrementally by 0.5%, starting from 80% and increasing up to a 99.5% target CSL. The y-axis represents the total % of periods where demand is not fully covered ie. 100% minus the achieved service level, and the x-axis represents the corresponding average OHI. Running from top-left to bottom-right, the points run from low to high service level.

From the plot, it can be seen that both approximations outperform the benchmark for

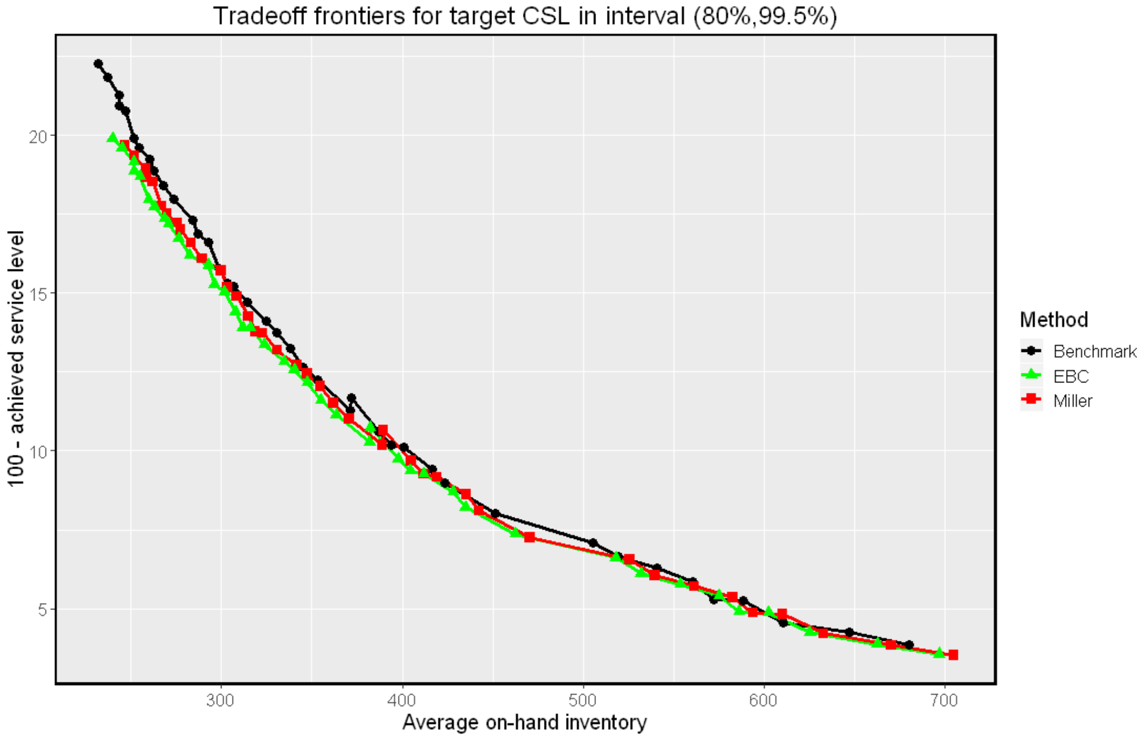


Figure 4.5.3: Tradeoff between achieved service level and average OHI. The parameter set is $\beta = -2$, $\sigma = 0.5$, promotional frequency $= 0.1$ and history length $= 20$.

the majority of target CSLs; as expected, using an approximation to gain better forecasts results in better inventory performance. It can also be seen that the EBC curve dominates the Miller curve in most places, particularly for lower target CSLs. Additionally, since these cases are those that require the least inventory, the savings represents a greater percentage of the total average OHI. For the highest CSLs, corresponding to the target service levels of 95% and above, the lines overlap more and there is not any significant difference. The lack of a significant difference at this upper end may relate to the increased difficulty in estimating the tails of the forecast error distribution for safety stock calculation. Nevertheless, the EBC tradeoff frontier is generally the superior one.

Similar tradeoff curves can be obtained for each of the possible parameter combinations,

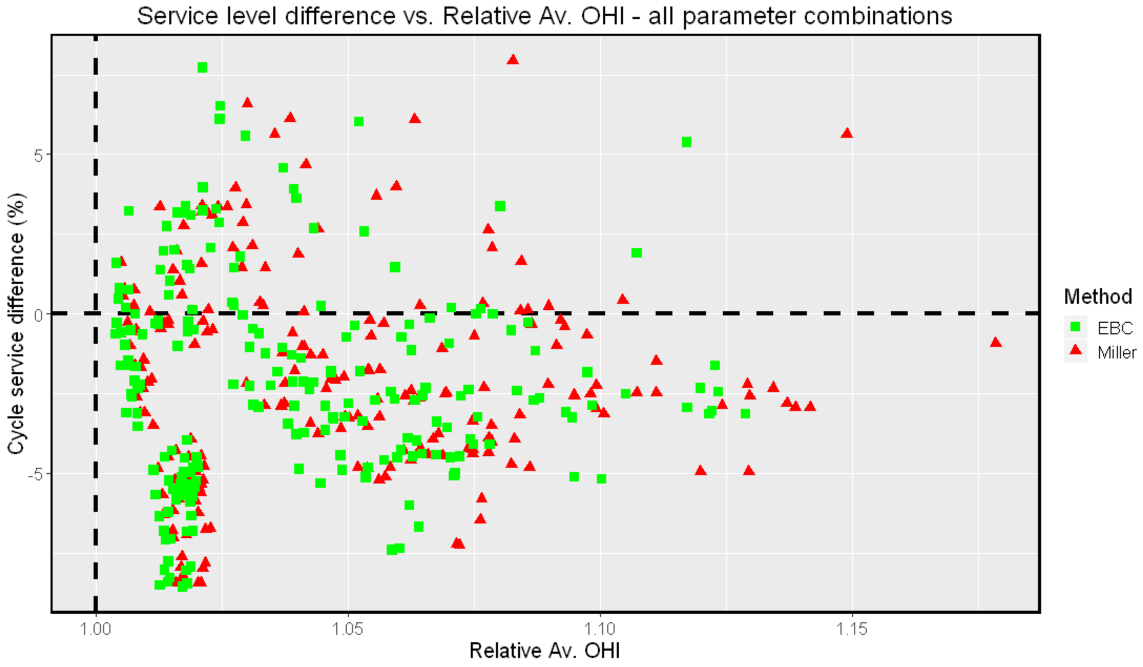


Figure 4.5.4: Scatterplot displaying relative av. OHI vs. service level difference for all parameter combinations.

and the results hold generally. These curves are too much to show in detail here. However, to demonstrate the results in the aggregate, a scatterplot of RelAvOHI vs. service level difference is shown in Figure 4.5.4. All parameter combinations are represented in this scatterplot; furthermore, results for 90%, 95% and 99% CSLs are disaggregated. Results for the higher target CSLs generally experienced a more negative service level difference, due to the difficulty in reaching such high targets. In terms of RelAvOHI, they were not too dissimilar. Visually, it can be seen from the plot that whilst neither method achieves a clustering of results significantly closer to the 0% service level difference line, the EBC cluster is as a whole closer to the left-hand side of the graph, indicating superior RelAvOHI.

For inventory managers, the implication of these results is largely dependent upon the particular utility of reducing inventories, relative to the utility of upholding service levels. However, these results have two generally applicable implications: (i) the EBC approach can be utilised to achieve superior inventory performance, and (ii) the possible gains are greatest for SKUs with a short history, greater variability, more frequent promotions, and a more elastic response to price. It should be noted that the negative service level difference can be a serious practical issue (which affects all promotional forecasting and other difficult-to-forecast situations) if achieved CSL above 90% is required; other target CSLs are achievable through setting the target CSL in the inventory system to be higher than required.

4.5.4 Connection with forecast bias/variance

The question of whether properties of the forecasts can be linked to improvements in inventory performance is now considered, in order to gain insight as to which of these properties may be more important. To assess this, the ratios of forecast bias, variance and accuracy under the possible conditions is calculated and tabulated, along with the percentage inventory savings and percentage service level decrease, are analysed.

We present these results in the form of direct comparison between the Miller and EBC approximations. For the bias, accuracy and variance ratios, we divided the relevant figures in each case for the EBC approximation by the figure for the Miller approximation to obtain ratios of these metrics. For example:

$$\text{ME ratio} = \frac{|\text{ME (EBC)}|}{|\text{ME (Miller)}|} \quad , \tag{4.5.1}$$

is the ratio of mean error. A value of less than 1 indicates superior performance of EBC on this metric; higher than 1 indicates superiority of Miller. Similar ratios are used for forecast variance and root mean squared error (RMSE).

For the inventory savings, we present the percentage decrease in RelAvOHI yielded by EBC with regard to Miller. Since EBC results in a lesser average OHI than Miller, the inventory saving percentage is the savings gained by using EBC instead of Miller. Similarly, for service level decrease, we present the percentage lost by using EBC over Miller.

The results are presented in Table 4.5.2; for concision, only a representative sample of the combinations are displayed. Examining the table, we can see that the greatest bias reduction happens generally at lower history lengths, as expected, as well as at high promotional frequencies and variability . We also see that the EBC approximation forecasts are less variable than those from Miller, with the difference between the two growing to a smaller extent with promotional frequency/variability, as well as shrinking with history length. The accuracy figures follow the same pattern, with the smallest magnitude of improvement of the three. Meanwhile, elasticity seems to play little part in the differences in variance and accuracy figures; however, it does seem that the bias ratios are better for the less elastic cases than for the more elastic ones.

Considering how the inventory savings and service level figures combine with this, it can be seen that the inventory savings figures improve in the same directions as the bias, variance, and accuracy ratios. They are also slightly better for the less elastic cases, which indicates an association with the improved bias figures, since it is only really the bias that improves markedly in this direction. The fact that less elastic cases seem to be better for EBC here seems to contradict the observation from earlier; however, here we are looking at percentage inventory savings from Miller to EBC, whereas earlier, the benchmark was the point of reference, so the measurements are slightly different. Considering the service level decreases as well, the decreases are much more modest for the most elastic cases, so the tradeoff may still be more favourable in that scenario.

Parameter combinations				Bias (ME)	Variance	Accuracy (RMSE)	Inventory	Service level
β	σ	Promo prop.	History	ratio	ratio	ratio	saving (%)	decrease (%)
-1	0.2	0.05	10	-	-	-	-	-
			30	0.767	0.997	0.999	0.07	0.03
	0.5	0.1	10	0.581	0.951	0.991	1.16	0.19
			30	0.575	0.983	0.999	0.29	0.06
	0.8	0.2	10	0.462	0.848	0.973	3.44	0.41
			30	0.459	0.952	0.997	0.57	0.08
-2	0.2	0.05	10	-	-	-	-	-
			30	0.782	0.998	0.999	0.07	0.02
	0.5	0.1	10	0.672	0.955	0.986	0.95	0.17
			30	0.721	0.985	0.998	0.30	0.05
	0.8	0.2	10	0.536	0.863	0.966	2.44	0.18
			30	0.617	0.957	0.995	0.45	0.07

Continued on next page

Continued from previous page

Parameter combinations				Bias (ME)	Variance	Accuracy (RMSE)	Inventory	Service level
β	σ	Promo prop.	History	ratio	ratio	ratio	saving (%)	decrease (%)
-4	0.2	0.05	10	-	-	-	-	-
			30	0.908	0.998	0.999	0.09	0.00
	0.5	0.1	10	0.782	0.956	0.982	0.97	0.07
			30	0.705	0.986	0.997	0.23	0.02
	0.8	0.2	10	0.581	0.870	0.963	2.50	0.11
			30	0.672	0.959	0.995	0.39	0.01

Table 4.5.2: Bias/RMSE/variance ratio of EBC approx. to Miller approx., alongside percentage inventory saved and service level lost.

The conclusions overall here are difficult to pick out, so to visualise the results, scatterplots showing the relationship between forecast bias and variance and the two inventory performance metrics are used. Figure 4.5.5 shows how the forecast bias relates to the RelAvOHI and the service level difference respectively. It can be seen from the first panel that the EBC cluster as a whole has generally a lower bias and a lower RelAvOHI, suggesting that the two are somewhat associated. Additionally, the scenarios where the bias is lower in general are those with the best RelAvOHI; this pattern seems to hold for both methods. The second panel shows a slightly less clear relationship, with the less biased EBC cluster definitely worse on average in terms of service level difference, with the catch that it's not necessarily further from the target in all cases.

Figure 4.5.6 provides the scatterplots for the forecast variance's relationship with RelAvOHI and service level difference. The plots here show a much different relationship to the previous example. In the top panel, we see that although the EBC cluster has lower RelAvOHI figures, it doesn't correspond with the lower variance figures as well. In general, the two clusters have fairly similar variance performance. In the bottom panel, it can be seen that there is very little difference between the EBC and Miller clusters.

We can conclude that if the improvement in performance of EBC over Miller is due to the improved properties of the forecasts, it is likely due to the significant decrease in bias, since there is an association between the two. The scatterplots demonstrate visually that the two clusters are separated most in the direction of bias and in a positive direction in terms of inventory performance overall. This is in line with our expectations, and indicates that, since EBC can lead to significant bias reduction, it can also lead to improved inventory performance in those situations. Additionally, it adds weight to the literature that has previously shown forecast bias to be an important factor in determining inventory performance.

The variance figures are all very close to 1, with very little variation in any direction. On the one hand, the fact that variance is relatively controlled makes the association between the bias and inventory performance more obvious. On the other hand, there is not enough evidence to assess how significant for the inventory reducing the variance would be, and so this remains for further work to determine.

Finally, the scatterplots for forecast accuracy versus RelAvOHI and service level difference are shown in Figure 4.5.7. We see that the two clusters in this case are very similar, which is not surprising given the closeness of the RMSE ratio to 1 in most cases considered. Due to this, there is not much to say about the relationship between the difference in accuracy between the two methods and the difference in inventory performance. A clear relationship is noticed where the service level difference is generally lower in the situations where RMSE is lower, with the most accurate forecasts leading to service levels which fall below the target by up to 9%; this is most likely due to the very high target service levels used, which are known to be hard to achieve in practice, whilst situations where the forecasts are less accurate managing to achieve higher service levels is likely because those sets of forecasts also exhibit higher variance in our case.

4.6 Conclusions

In this paper, the impact on inventory performance of a new forecast approximation (termed the EBC approximation), developed for reducing forecast bias when loglinear sales models are assumed, was examined. Through the use of inventory simulations, it was demonstrated that (i) EBC generally reduces average OHI at the expense of achieved service level, when

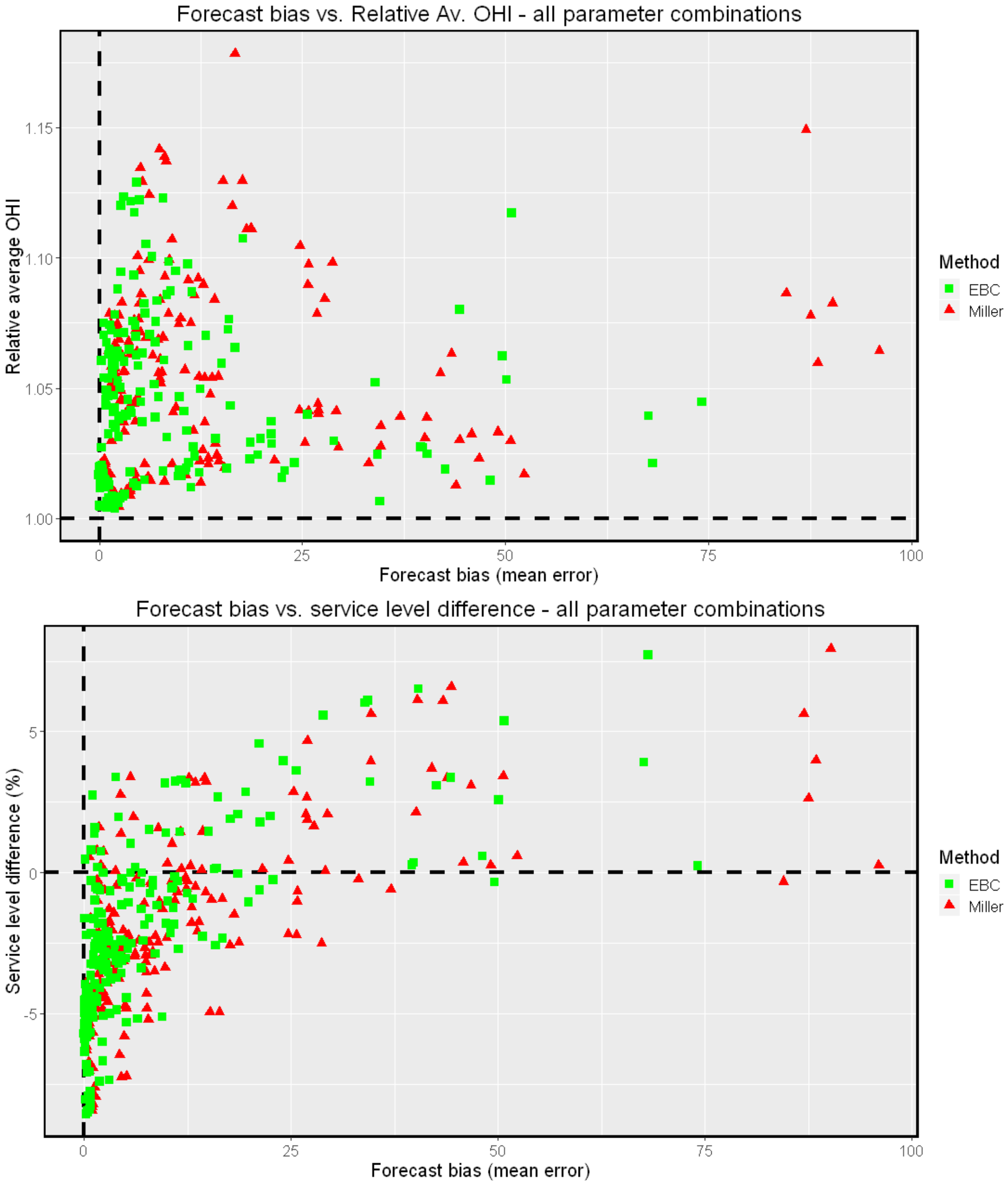


Figure 4.5.5: Forecast bias vs. (i) RelAvOHI (ii) Service level difference, for all parameter combinations.

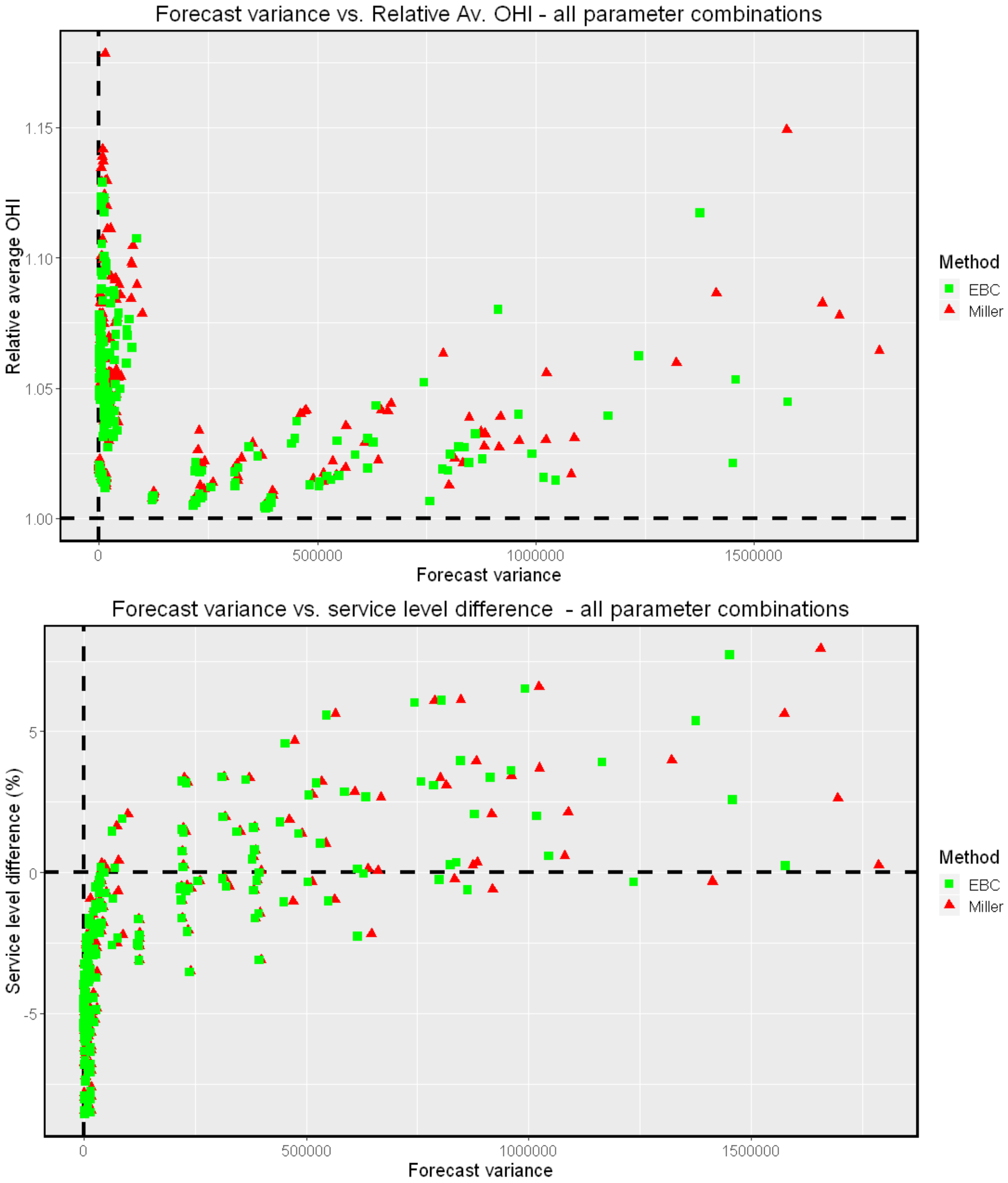


Figure 4.5.6: Forecast variance vs. (i) RelAvOHI (ii) Service level difference, for all parameter combinations.

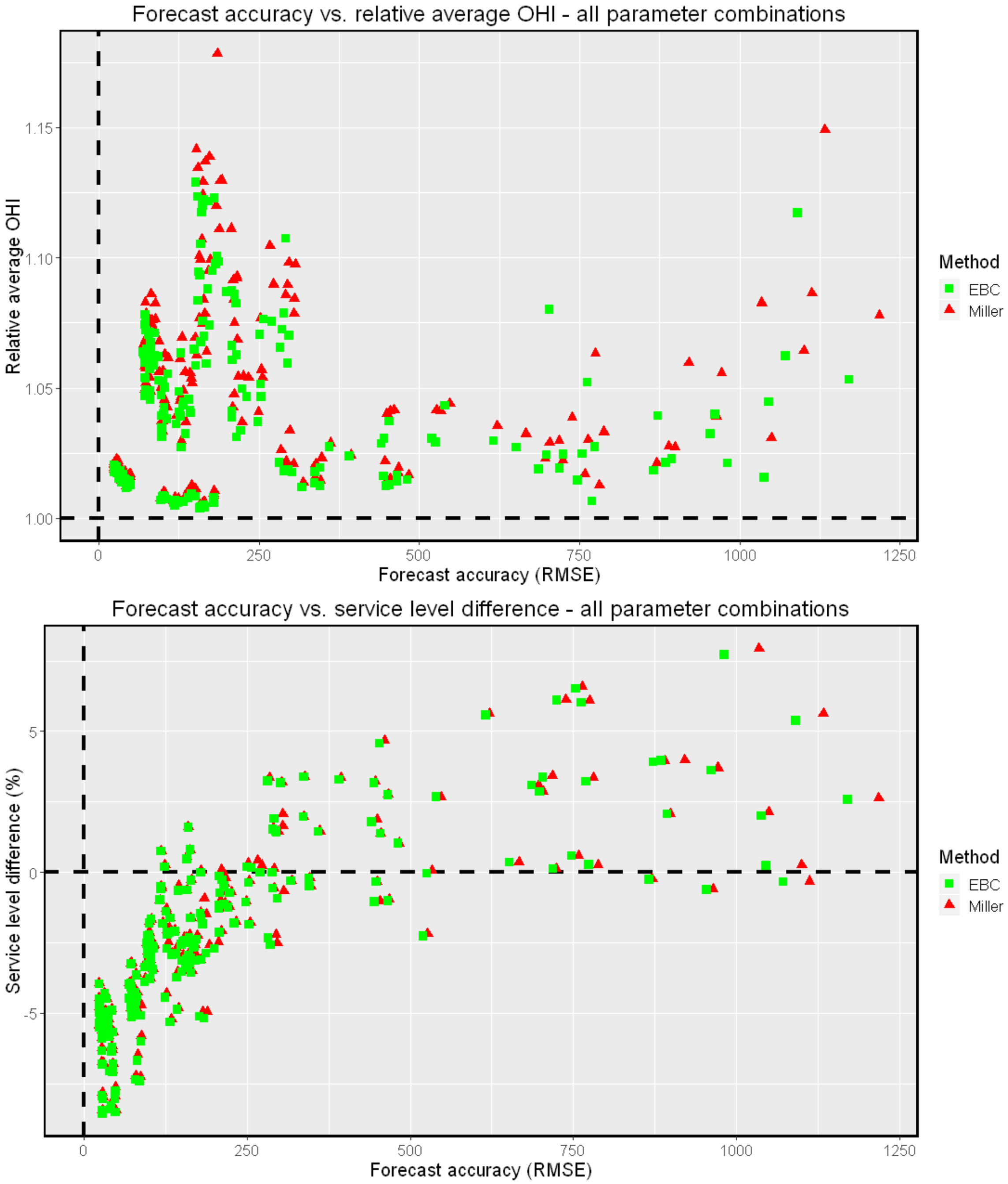


Figure 4.5.7: Forecast accuracy vs. (i) RelAvOHI (ii) Service level difference, for all parameter combinations.

compared with a previously used forecast approximation, and (ii) the size of the improvement in average OHI is greatest at shorter history levels, along with noisier demand, greater promotional frequency, and more elastic response to price. The use of a heuristic to pool forecast errors for the estimation of safety stock when sample sizes are small was demonstrated to be effective at improving service level difference in promotional periods overall. Tradeoff curves between inventory and service level were found to be promising for EBC, and a scatterplot showing the dual performance for all scenarios confirmed that. Furthermore, the forecast properties of the methods themselves were examined and linked to the inventory performance. The evidence showed that scenarios where the reduction of bias was greatest corresponded with scenarios where inventory savings were most favourable for EBC, despite worsening variance. The conclusion is that the reduction in forecast bias yielded by EBC may be associated with the improved inventory performance, and that the situations in which that reduction is greatest are likely to be the most promising for inventory gains. Due to the relatively constant variance and accuracy figures, it was not possible to come to a conclusion on how significant improving either of those forecast properties would be for inventory performance.

For inventory managers, the message that bias reduction in forecasts is important for inventories is one that can be useful in general. The use of a heuristic to improve safety stock calculations in promotional periods seems under-discussed, with short histories not focussed on much in previous work, where the assumption of Gaussian forecast errors is quite common. Certainly, pooling all forecast errors without adjusting non-promotional forecast errors does not help; these types of adjustments are perhaps something that managers can investigate in practice. The refinement of heuristics, such as the one presented here, to improve the service level in promotional periods further may be a fruitful direction for more research, and could focus on whether simply matching the means of the two samples is sufficient, or whether more specific, model-based heuristics give better performance. Another direction of further work would be to expand the scope of the approach here to other situations; for example, having a group of SKUs with a covariance structure between their demand patterns, or expanding the parameters of the existing work, for example looking at fill rate, or lengthening the lead and review times.

Chapter 5

Outcomes and further work

5.1 Thesis summary

In this thesis, we have attempted to address a selection of important practical issues for retail forecasters whilst contributing more generally to gaps in the academic research areas. Chapter 1 expanded upon the motivation for the thesis, giving a general overview of the contributions made. In Chapter 2, we were motivated to study the situation where demand is influenced by complex seasonal patterns, often consisting of multiple repeating seasonalities with calendar effects. A new model was developed, the mixed parsimonious model, which was embedded within an innovations state-space framework, and allowed for the mixing of trigonometric and index representations of seasonality. In an empirical evaluation, transactional data from food wholesaling was aggregated into daily buckets, resulting in series which exhibit both an intraweek and intrayear seasonality, as well as several calendar effects. When applying the mixed parsimonious model, we modelled the intrayear seasonality with a trigonometric representation due to the short history and the gently curving nature of the seasonality, and the intraweek seasonality with an index representation, as the data was sufficient to allow for smoothing to take place. The results demonstrated that the new model was superior in most cases, justifying the modelling approach.

Another topic which is of import to retail organisations is the dual issue of forecasting and parameter estimation at the SKU-level, where the signal-to-noise ratio exhibited in the data is at the lowest in the hierarchy. The use of hierarchical information in alleviating these two problems was examined in Chapter 3, which began with a review of the existing literature on sales response models, informing the choice to focus on loglinear regression models, a common approach in practice. A gap that was identified in the literature was that whilst procedures for using aggregate information to estimate SKU-level parameters were proposed, there were none that do so in a way avoids bias. To address this, we proposed a geometric parameter inheritance scheme which can be implemented in the scenarios where SKUs can be assumed to share a common parameter value. Secondly, the bias that occurs in the forecast was analysed, and a new approximation was proposed that alleviates the bias further compared with other approximations. The combination of these two improvements was also laid out. A simulation study was then carried out, which demonstrated that the geometric parameter inheritance scheme results in more accurate parameter estimates, and that the forecast approximation was successful in significantly reducing bias in forecasts, whilst remaining competitive in accuracy in terms of squared error.

The forecast approximation was carried forward in Chapter 4, where the impact of the debiasing approximation on inventory performance was examined. The inventory setup chosen was a commonly used one in practice, where results showed that a recently proposed approach to calculating safety stock by implementing a kernel density estimator was superior to simply taking quantiles of the empirical forecast error distribution. It was also demonstrated that a simple heuristic to enable the pooling of forecast error sets resulted in better performance, and provided a solution to the problem of having small forecast error sets from promotional periods, which is an issue needing more discussion in the literature. In a simulation study, we varied parameters of the data including price elasticity of demand, noise, promotional frequency and history length, and contrasted the performance of the new approximation with a benchmark forecast function and a previous approximation representing current practice. The results demonstrated that the new approximation generally reduces stock holding costs whilst experiencing a slight decrease in service level. Additionally, the gains are greater under shorter histories, greater promotional frequency, noisier series and for more elastic SKUs. Finally, we were also able to shed some light on the link between forecast properties and inventory metrics, finding some evidence that the reduction in bias was a driver of the improved performance, whilst less was able to be concluded about the effect of the forecast variance and accuracy.

5.2 Implications for practitioners

The work in this thesis strikes a balance between academic research and practice. We have addressed areas that are important considerations for practitioners as in Seaman (2018) which are also trending areas in academic research (Fildes et al., 2019). As such there are immediate points that practitioners can take away as improvements. We summarise some of the main ones in Table 5.2.1. To give some direction in how these points can be translated into practice, we discuss the points raised in the table.

For the manual definition of seasons in the mixed parsimonious model, some guidance is given in Appendix A.1; it may require a little bit of experimentation to get a feel for. The salient point is to create clusters that are thought to be similarly influenced by the seasonality and calendar effects under consideration. It should also be noted that the clusters must be well defined, so that periods in the forecast sample can be attributed to existing clusters before the data for that period is observed.

In using the geometric parameter inheritance, there are a number of ways and situations in which SKUs with common parameters can be identified. This could take the form of conducting pre-analysis ie. estimating parameters individually first and then grouping products which turn out to be similar; or it could be contextual, for example grouping SKUs which are different varieties of the same brand, such as flavours of crisps. Most likely, a combination of the two would be ideal, as SKUs should not be grouped without an underlying logic, but logic alone may lead to some products being clustered incorrectly.

Finally, when seeking to use the debiasing approximation in an inventory context, consideration should be given to defining exactly what terms such as short histories and elastic SKUs are. In our simulations, the parameter values used can be found in Table 4.4.1; perhaps most important to note is the choice of history lengths, which are quite short and reflect that we expect most benefits to come when the history is 30 periods or less. The other values chosen are somewhere close to the middle of the range of values that has been observed in real data. In general, drawing boundaries in the parameter space to indicate where the approximation should be used needs to be considered with respect to the inventory cost function of each organisation; a firm with a higher ratio of holding costs to stockout penalties may find the approximation beneficial in a wider range of scenarios. A good idea is to experiment first by running the system with the approximation alongside the current system and assessing where benefits would have occurred.

Chapter	**Takeaway points**
Multiple seasonality in retail	New mixed parsimonious model flexibly models interactions of seasonality and calendar effects. Requires some manual definition of seasons.
Sources of bias in loglinear models for retail	Geometric parameter inheritance can be considered to improve SKU-level parameter estimates when (i) SKU level data is available, and (ii) the SKUs share a common parameter. New approximation to debias forecasts significantly reduces bias, performs similarly in accuracy terms.
The inventory performance of bias-corrected sales forecasts	EBC approximation can be considered to improve inventory performance, particularly if holding costs are large. Debiasing should be considered particularly when histories short, SKU is elastic, promotional frequency is high and/or noise is high.

Table 5.2.1: Summary of managerial implications.

5.3 Future directions

Whilst the work in this thesis constitutes a contribution towards some of the open questions in retail forecasting, there are limitations to the scope of the research and qualifiers to the conclusions that can be drawn. Attention has been drawn to these at various points throughout; we summarise here some further work that could be done.

Chapter 2 saw the development of a new model for multiple seasonal forecasting, the mixed parsimonious model. As mentioned, there is scope for a couple of extensions to the model. Firstly, there is room to explore the modelling process, in terms of how to choose when to use a trigonometric seasonal representation, and when to use an index one. It was relatively clear how to best represent the intraweek and intrayear seasonalities seen in Chapter 2, but this may not always be the case.

Secondly, a fully automated model selection process for the mixed parsimonious model would be a useful innovation, as it would facilitate the application of the model when the number of series is very large. An automatic determination of the trigonometric terms required is already included in the model, so the central component of such an approach would be an unsupervised season-clustering approach for the index representation. We can envisage a procedure, akin to hierarchical clustering, that starts with fully 'disaggregate' clusters and proposes mergers based on p-values from a statistical test (for example, the Nemenyi paired test). We judge it to be unlikely that any modelling approach would outperform human modelling under mildly challenging conditions; for example, where the dataset has histories too short to permit the use of a validation set. Nevertheless, a comparison of automatic and human modelling which demonstrated an algorithm getting sufficiently close to the manual approach would instill a degree of confidence in using the automated method. Computational cost is also likely to be an issue which may need addressing. Finally, the mixed seasonality encoding may be beneficial in other modelling paradigms where seasonal variables need to be incorporated, such as machine learning, and this may merit further investigation.

In Chapter 3, we saw a method for improving SKU-level parameter estimates using parameter inheritance which applies under the assumption that the parameter is common among all the SKUs in the group. Naturally, the broader case is what can be done in the general case where the parameters are different. A possible line of research is whether forecasts and parameters can be simultaneously adjusted. The optimal combination forecast reconciliation method Hyndman et al. (2011) outlines a method which uses forecasts from all levels of a hierarchy in an adjustment process which ensures that forecasts are aggregate consistent. The idea is to adjust the forecasts at all levels by the minimum amount needed to ensure that the sums of forecasts at lower levels of the hierarchy equal the group forecasts at the higher levels, in a similar manner to the way a line-of-best-fit is fitted in a regression model. We investigated the possibility of adjusting SKU-level parameters in sync with the forecast adjustment procedure, in an attempt to achieve two goals: (i) improved SKU-level parameter estimates, and (ii) ensuring that the parameter estimates and adjusted forecasts are consistent. Such a method would also expand upon the geometric parameter inheritance scheme proposed in Chapter 3 by enabling users to deal with the more general situation of independent parameter values across SKUs. Initial attempts at formulating such a procedure did not lead to improved estimates; however, we do believe there is something achievable here by further experimentation with the form of the regression model assumed and the form of the covariance matrix used in the reconciliation procedure.

Another possible area of future research stemming from Chapter 3 relates to the derivation of the forecast approximation used to debias forecasts. Although a more significant chunk of bias has been removed, there still remains scope to debias the forecasts further by removing some of the assumptions that are involved in getting to this approximation. For example, we make the assumption that $\mathbb{E}[\hat{\alpha}_j]$ and $\mathbb{E}[P_{j,T+1}^{\hat{\beta}_j}]$ are independent, whilst noting that it is not strictly valid. Relaxing this assumption might lead to a further reduction in bias.

In Chapter 4, the problem of calculating safety stock when the set of available forecast errors is small was encountered, mainly in relation to promotional periods. Forecast errors from non-promotional periods in the history were of a much lower magnitude, and so a simple heuristic was developed to adjust and pool all the errors together, which improved performance. The question of how to construct a heuristic could do with further research, as it seems pertinent for many situations in practice. In general, longer histories will exist in practice, but (i) promotions may still be relatively infrequent, and (ii) we only differentiated here between promotional and non-promotional forecast errors, with one type of promotion; in real world scenarios, the group of forecast errors is likely to be further stratified via different levels and types of promotion, as well as other variables which affect the magnitude of error. Further work could investigate whether simply matching properties of the empirical forecast distributions (mean, standard deviation) is sufficient, or whether other model-related quantities, such as the standard error of the parameter estimates, could better explain the relationship between error sets. The former is attractive in that it is simple and easily generalisable; however, in a situation where one set only contains a small number of errors, more may be needed.

Taking into account the limitations just discussed, we believe that the contributions of this thesis can provide a new path forward for researchers and practitioners, even beyond the retailing domain. The methodological developments made have relevance in a range of areas; multiple seasonal signals and special events are considerations in applications such as energy markets and traffic management, whilst bias reduction and noisy data are ubiquitous problems in statistics and data science. We see these issues becoming more and more relevant, especially as the richness of data, in terms of both the number of variables and the sampling frequencies involved, increases.

Appendix A

Multiple seasonality in retail - appendices

A.1 Model specification for PES

The specification procedure used to obtain the PES results is described in this section. Different applications have varying seasonal influences, and therefore the seasonal structure of each requires specific modelling. We describe a process to specify the seasonal elements for PES. As an example, we use a daily time series of sales from the education sector. The following judgemental procedure is used in determining the seasonal structure from the data:

- All days where the sector follows usual weekly behaviour are identified. Initially, seven seasons are created to represent those days.
- From the remaining days, any that follow uncommon but recurring behaviour that occurs more than once a year are identified. These days ranged from school vacation periods in the Education sector, to summer weeks for the Hotels sector. New seasons

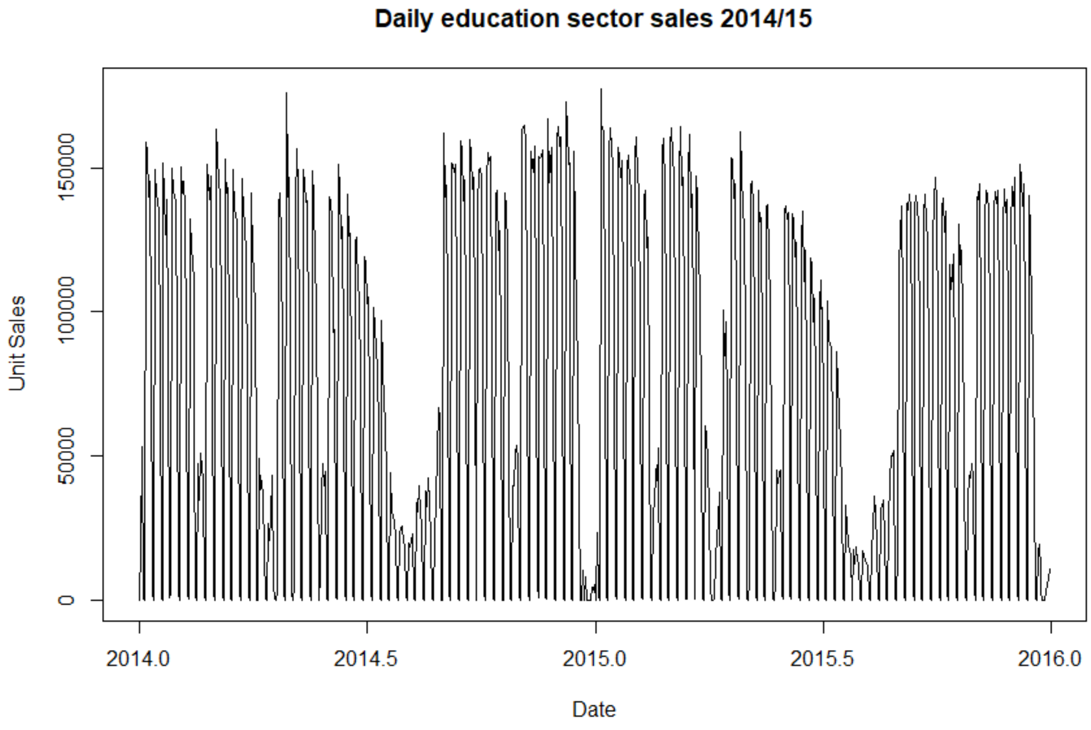


Figure A.1.1: Daily sales to the education sector.

are created for these days accordingly.

- The remaining days are termed 'special days' and correspond with once-a-year phenomena such as the Christmas period or Easter. The idea is that these observations are far more influenced by their position in the year than the weekly seasonality.

In this way, we obtain seasonal structures with a total number of seasons which is the sum of the number of seasons created from these 3 categories.

Figure A.1.1 shows the plot of daily sales to the education sector (including schools, universities and other institutions). We can see that the seasonal pattern is rather complex, with the week of year having a very clear effect. Taking the general approach to defining seasons outlined above, we went through the following process:

- *Normal behaviour*: in this case, there seemed no overarching usual behaviour outside of weekends, which were constant throughout the year (determining the first 2 seasons). However, using contextual information, we split the series up into 6 'term' periods (term dates for primary/seco ndary education), 3 'half-term' periods of a week in length, an Easter holidays period of two weeks, a summer holidays period of 6 weeks, and a Christmas holiday period of two weeks.

- *Recurring behaviour*: we analysed each sector separately to determine the number of seasons. Figure A.1.2 shows a plot of the 6 week summer holiday period, overlaying sales for the period from consecutive years and the average sales over both. From the average shapes (ignoring weekends), we assessed that Week 1 took a unique shape, Weeks 2 and 3 were broadly similar, Weeks 4 and 5 were broadly similar, and Week 6 was unique. This created 20 new seasons. Furthermore, we considered that for Weeks 2 and 3, the level of sales was roughly the same across all days, whereas for Week 6, sales looked to be similar from Wednesday to Friday. Thus, we defined 5 (Week 1) + 1 (Weeks 2/3) + 5 (Weeks 4/5) + 3 (Week 6) = 14 unique seasons for these periods.

  Further analysis led to the following number of seasons being defined:

  – 5 seasons for half-terms,

  – 10 periods for Easter holidays,

  – 10 periods for the spring terms,

  – 15 periods for the summer terms,

  – 15 periods for the autumn terms,

  which led to a total of 2+14+5+10+10+15+15 = 71 seasons so far.

- *Special days*: the final period to be considered was the Christmas holiday period. From

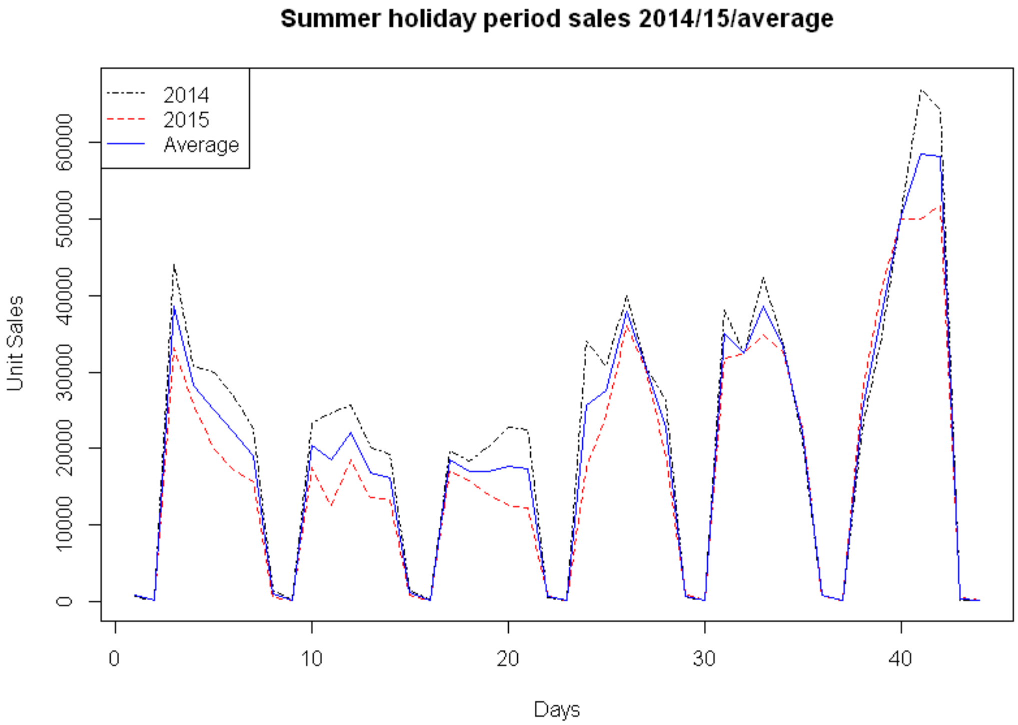


Figure A.1.2: Daily sales to the education sector during the 6 weeks of summer holidays.

the 24th December through to the 2nd January, taking weekly and yearly differences helped to reveal that these seasons were influenced more by the day of the year than the day of week, or even week of year. Thus, we created an additional 10 periods for these days, leading to a final total of 81 days.

A.2 Multiplicative PES

The formulation for multiplicative PES is as follows:

$$y_t = \Big(\prod_{i=1}^{M} s_{i,t-1}^{I_{it}}\Big) e_{t-1}^{\phi} \varepsilon_t \tag{A.2.1}$$

$$e_t = \frac{y_t}{\Big(\prod_{i=1}^{M} s_{i,t-1}^{I_{it}}\Big)} \tag{A.2.2}$$

$$s_{i,t} = (s_{i,t-1})(e_t)^{(\alpha+\omega I_{it})} \quad , \tag{A.2.3}$$

$$I_{it} = \begin{cases} 1, & \text{if period t occurs in season i} \\ 0, & \text{otherwise} \end{cases}$$

where α and ω are smoothing parameters bounded as before, and ϕ is an autoregressive parameter greater than 0. We use this formulation alongside the additive in Section 2.4, since retailing series are commonly influenced by multiplicative influences.

A.3 MAPE results

We present the MAPE results for the individual period (Table A.3.1) and cumulative (Table A.3.2) forecasts. The first table differs from the AvgRelMAE results in that the mixed parsimonious model is shown to be the best at 7 steps ahead, rather than single-seasonal ES, and PES is considered the best at 14 steps ahead. The overall performance of PES is more favourable looking at MAPE than at AvgRelMAE; apart from this, the general conclusions are the same, with the other two multi-seasonal methods failing to outperform the benchmarks. The 1 step-ahead figures for single-seasonal ES and TBATS are unusually large; this is explained at the end of the section.

For the cumulative figures in Table A.3.2, there are no areas where the MAPE results significantly diverge from the AvgRelMAE. The mixed parsimonious method still marginally outperforms the other methods. Note that the MAPE figures are much lower than in Table

Table A.3.1: MAPE figures, individual horizons.

Forecast	Horizon			
	1	7	14	28
Seasonal naive (7)	45.51%	51.49%	49.12%	49.98%
Single-seasonal ES (7)	96.91%	41.76%	48.48%	57.11%
Regression	54.05%	48.44%	48.42%	49.04%
TBATS (7,365.25)	108.24%	71.73%	77.90%	82.44%
Double-seasonal HW (7,365)	52.01%	51.89%	54.56%	59.73%
PES (additive)	40.68%	38.88%	**39.40%**	44.42%
PES (multiplicative)	40.74%	39.00%	40.15%	45.61%
Mixed parsimonious	**40.04%**	**37.66%**	42.33%	**44.10%**

A.3.1, since the individual errors average out to some extent over the interval.

We explore further the magnitude of the large MAPE figures calculated for the one-step ahead single-seasonal ES and TBATS methods. The explanation mainly involves which dates are in the test set at different horizons. The first forecast made when $h = 1$ pertain to the 1st January 2016, but the first forecast when $h = 7$ would pertain to the 7th January. This allows for an additional 6 forecasts to be made for each run when $h = 1$, compared to $h = 7$. Figure A.3.1 shows the distribution of 1-step ahead forecast errors on a particular series for the seasonal naive, single-seasonal ES and TBATS methods. The plot shows that the distributions are relatively similar, except that there are a few more extreme errors in the case of single-seasonal ES and TBATS. These errors (of magnitude up to 2×10^4) are the cause of the much larger values of MAPE for the latter two methods, compared with

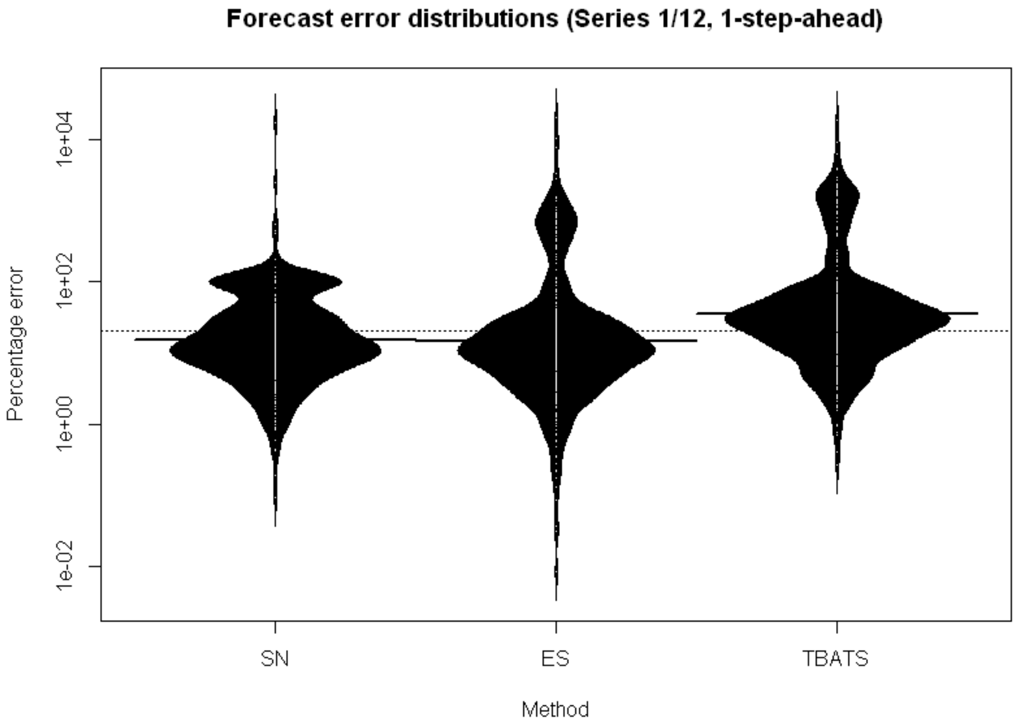


Figure A.3.1: Beanplot showing the one-step-ahead percentage forecast error distribution on the first series, for 3 selected methods.

Table A.3.2: MAPE figures, cumulative horizons.

Forecast	Horizon		
	1-7	1-14	1-28
Seasonal naive (7)	18.75%	16.63%	16.23%
Single-seasonal ES (7)	24.45%	22.78%	25.95%
Regression	13.02%	14.38%	14.71%
TBATS (7,365.25)	25.11%	25.65%	25.48%
Double-seasonal HW (7,365)	15.52%	15.34%	15.30%
PES (additive)	12.80%	12.68%	12.72%
PES (multiplicative)	14.36%	13.87%	13.64%
Mixed parsimonious	**12.69%**	**12.34%**	**12.45%**

the seasonal naive. The figure is representative of other time series in our data.

Both the single-seasonal ES and TBATS methods have large errors in the first part of the series, specifically the first few periods that only occur in the test when $h = 1$ (and not when $h = 7$). These errors are large enough to distort the overall MAPE, due to the actual values being extremely low.

Appendix B

The inventory performance of bias-corrected sales forecasts - appendix

Parameter combinations				Service level difference		RelAvOHI	
Elasticity	Sigma	Promo. freq.	History	Miller	EBC	Miller	EBC
-1	0.2	0.05	10	-	-	-	-
			20	**-5.20**	-5.26	1.020	**1.019**
			30	**-4.58**	-4.61	1.020	**1.019**
		0.1	10	**-7.66**	-7.73	1.022	**1.019**
			30	**-4.74**	-4.77	1.019	**1.018**
		0.2	10	**-7.73**	-7.82	1.021	**1.018**
			20	**-5.82**	-5.90	1.020	**1.019**

Continued on next page

Continued from previous page

Parameter combinations				Service level difference		RelAvOHI	
Elasticity	Sigma	Promo. freq.	History	Miller	EBC	Miller	EBC
			30	**-5.38**	-5.42	1.020	**1.019**
	0.5	0.05	10	-	-	-	-
			30	**-3.85**	-3.93	1.064	**1.061**
		0.2	10	**-5.93**	-6.17	1.077	**1.064**
			30	**-4.06**	-4.15	1.063	**1.059**
	0.8	0.05	10	-	-	-	-
			20	**-2.88**	-2.98	1.098	**1.093**
			30	**-2.75**	-2.83	1.095	**1.090**
		0.1	10	**-2.97**	-3.22	1.114	**1.093**
			30	**-2.66**	-2.75	1.101	**1.094**
		0.2	10	**-2.83**	-3.23	1.146	**1.107**
			20	**-2.89**	-3.03	1.106	**1.088**
			30	**-2.47**	-2.55	1.096	**1.089**
-2	0.2	0.05	10	-	-	-	-
			30	**-4.78**	-4.80	1.018	**1.017**
		0.2	10	**-7.80**	-8.10	1.017	**1.014**
			30	**-5.77**	-5.82	1.014	**1.013**
	0.8	0.05	10	-	-	-	-
			30	**-0.97**	-1.04	1.061	**1.057**
		0.2	10	**-0.51**	-0.69	1.093	**1.067**

Continued on next page

Continued from previous page

Parameter combinations				Service level difference		RelAvOHI	
Elasticity	Sigma	Promo. freq.	History	Miller	EBC	Miller	EBC
			30	**-0.34**	-0.41	1.058	**1.053**
-4	0.2	0.05	10	-	-	-	-
			20	**-2.59**	-2.64	1.009	**1.008**
			30	**-2.21**	**-2.21**	1.009	**1.008**
		0.1	10	**-2.01**	-2.10	1.018	**1.008**
			30	**-0.24**	-0.25	1.007	**1.006**
		0.2	10	**-1.50**	-1.53	1.010	**1.007**
			20	**-0.37**	-0.38	1.006	**1.004**
			30	0.51	**0.48**	1.005	**1.005**
	0.5	0.05	10	-	-	-	-
			30	1.49	**1.46**	1.019	**1.016**
		0.2	10	0.17	**0.12**	1.042	**1.034**
			30	1.26	**1.25**	1.016	**1.013**
	0.8	0.05	10	-	-	-	-
			20	3.30	**3.27**	**1.034**	1.044
			30	3.42	**3.39**	1.023	**1.019**
		0.1	10	4.06	**3.95**	1.069	**1.035**
			30	2.87	**2.85**	1.029	**1.024**
		0.2	10	2.65	**2.53**	1.104	**1.078**
			20	1.74	**1.66**	1.042	**1.029**

Continued on next page

Continued from previous page

Parameter combinations				Service level difference		RelAvOHI	
Elasticity	Sigma	Promo. freq.	History	Miller	EBC	Miller	EBC
			30	1.93	**1.92**	1.025	**1.021**

Bibliography

Ali, M. M., Boylan, J. E., Syntetos, A. A., 2012. Forecast errors and inventory performance under forecast information sharing. International Journal of Forecasting 28 (4), 830–841.

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

Andrews, R. L., Currim, I. S., Leeflang, P., Lim, J., 2008. Estimating the SCAN*PRO model of store sales: HB, FM or just OLS? International Journal of Research in Marketing 25 (1), 22–33.

Arunraj, N. S., Ahrens, D., 2015. A hybrid seasonal autoregressive integrated moving average and quantile regression for daily food sales forecasting. International Journal of Production Economics 170, 321–335.

Barrow, D., Kourentzes, N., 2018. The impact of special days in call arrivals forecasting: A neural network approach to modelling special days. European Journal of Operational Research 264 (3), 967–977.

Barrow, D. K., 2016. Forecasting intraday call arrivals using the seasonal moving average method. Journal of Business Research 69 (12), 6088–6096.

Bemmaor, A. C., Wagner, U., 2002. Estimating market-level multiplicative models of promotion effects with linearly aggregated data: a parametric approach. In: Advances in Econometrics. Emerald Group Publishing Limited, pp. 165–189.

Box, G. E., Jenkins, G. M., Reinsel, G. C., Ljung, G. M., 2015. Time series analysis: forecasting and control, 5th Edition. John Wiley & Sons.

Boylan, J. E., 2018. Commentary on retail forecasting. International Journal of Forecasting 34 (4), 832–834.

Bunn, D. W., Vassilopoulos, A., 1993. Using group seasonal indices in multi-item short-term forecasting. International Journal of Forecasting 9 (4), 517–526.

Chatfield, C., 1978. The Holt-Winters forecasting procedure. Applied Statistics 27 (3), 264–279.

Chen, A., Blue, J., 2010. Performance analysis of demand planning approaches for aggregating, forecasting and disaggregating interrelated demands. International Journal of Production Economics 128 (2), 586 – 602, supply Chain Forecasting Systems.
URL http://www.sciencedirect.com/science/article/pii/S0925527310002318

Chen, H., Boylan, J. E., 2007. Use of individual and group seasonal indices in subaggregate demand forecasting. Journal of the Operational Research Society 58 (12), 1660–1671.

Chen, H., Boylan, J. E., 2008. Empirical evidence on individual, group and shrinkage seasonal indices. International Journal of Forecasting 24 (3), 525–534.

Christen, M., Gupta, S., Porter, J. C., Staelin, R., Wittink, D. R., 1997. Using market-level data to understand promotion effects in a nonlinear model. Journal of Marketing Research, 322–334.

Cooper, L. G., Baron, P., Levy, W., Swisher, M., Gogos, P., 1999. PromoCast: A new forecasting method for promotion planning. Marketing Science 18 (3), 301–316.

Core Team, R., 2013. R: A language and environment for statistical computing:. 201.

Crone, S. F., Kourentzes, N., 2010. Feature selection for time series prediction–a combined filter and wrapper approach for neural networks. Neurocomputing 73 (10-12), 1923–1936.

Crone, S. F., 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. pp. p3285–3292.

Dalhart, G., 1974. Class seasonalitya new approach. Published in Forecasting, 2nd edition. American Production and Inventory Control Society, Washington DC, 11–16.

Davydenko, A., Fildes, R., 2013. Measuring forecasting accuracy: The case of judgmental adjustments to SKU-level demand forecasts. International Journal of Forecasting 29 (3), 510–522.

De Livera, A. M., Hyndman, R. J., Snyder, R. D., 2011. Forecasting time series with complex seasonal patterns using exponential smoothing. Journal of the American Statistical Association 106 (496), 1513–1527.

Dekker, M., Van Donselaar, K., Ouwehand, P., 2004. How to use aggregation and combined forecasting to improve seasonal demand forecasts. International Journal of Production Economics 90 (2), 151–167.

Divakar, S., Ratchford, B. T., Shankar, V., 2005. CHAN4CAST: A multichannel, multiregion sales forecasting model and decision support system for consumer packaged goods. Marketing Science 24 (3), 334–350.

Dudek, G., 2016. Pattern-based local linear regression models for short-term load forecasting. Electric Power Systems Research 130, 139–147.

Fildes, R., Ma, S., Kolassa, S., 2019. Retail forecasting: research and practice. Working paper.

Finney, D., 1941. On the distribution of a variate whose logarithm is normally distributed. Supplement to the Journal of the Royal Statistical Society 7 (2), 155–161.

Gardner, E. S., 1990. Evaluating forecast performance in an inventory control system. Management Science 36 (4), 490–499.

Ghysels, E., Osborn, D. R., 2001. The econometric analysis of seasonal time series. Cambridge University Press.

Gould, P. G., Koehler, A. B., Ord, J. K., Snyder, R. D., Hyndman, R. J., Vahid-Araghi, F., 2008. Forecasting time series with multiple seasonal patterns. European Journal of Operational Research 191 (1), 207–222.

Gruen, T. W., Corsten, D. S., Bharadwaj, S., 2002. Retail out-of-stocks: A worldwide examination of extent, causes and consumer responses. Grocery Manufacturers of America Washington, DC.

Grunfeld, Y., Griliches, Z., 1960. Is aggregation necessarily bad? The Review of Economics and Statistics, 1–13.

Hanssens, D. M., Parsons, L. J., Schultz, R. L., 2003. Market response models: Econometric and time series analysis. Vol. 12. Springer Science & Business Media.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The elements of statistical learning. 2nd edition. Vol. 1, no. 10. Springer series in statistics New York, NY, USA:.

Hippert, H. S., Pedreira, C. E., Souza, R. C., 2001. Neural networks for short-term load forecasting: A review and evaluation. IEEE Transactions on power systems 16 (1), 44–55.

Hong, T., Pinson, P., Fan, S., Zareipour, H., Troccoli, A., Hyndman, R. J., 2016. Probabilistic energy forecasting: Global energy forecasting competition 2014 and beyond.

Hyndman, R., Koehler, A. B., Ord, J. K., Snyder, R. D., 2008. Forecasting with exponential smoothing: the state space approach. Springer Science & Business Media.

Hyndman, R. J., Ahmed, R. A., Athanasopoulos, G., Shang, H. L., 2011. Optimal combination forecasts for hierarchical time series. Computational Statistics & Data Analysis 55 (9), 2579–2589.

Hyndman, R. J., Khandakar, Y., et al., 2007. Automatic time series for forecasting: the forecast package for R. Monash University, Department of Econometrics and Business Statistics.

Kapalka, B. A., Katircioglu, K., Puterman, M. L., 1999. Retail inventory control with lost sales, service constraints, and fractional lead times. Production and operations management 8 (4), 393–408.

Kourentzes, N., 2013. Intermittent demand forecasts with neural networks. International Journal of Production Economics 143 (1), 198–206.

Kourentzes, N., 2014. On intermittent demand model optimisation and selection. International Journal of Production Economics 156, 180–190.

Kourentzes, N., Barrow, D. K., Crone, S. F., 2014. Neural network ensemble operators for time series forecasting. Expert Systems with Applications 41 (9), 4235–4244.

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

Lewbel, A., 1992. Aggregation with log-linear models. The Review of Economic Studies 59 (3), 635–642.

Link, R., 1995. Are aggregate scanner data models biased? Journal of Advertising Research 35 (5), RC8–RC8.

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.

Miller, D. M., 1984. Reducing transformation bias in curve fitting. The American Statistician 38 (2), 124–126.

Minner, S., Transchel, S., 2010. Periodic review inventory-control for perishable products under service-level constraints. OR spectrum 32 (4), 979–996.

Neyman, J., Scott, E. L., 1960. Correction for bias introduced by a transformation of variables. The Annals of Mathematical Statistics 31 (3), 643–655.

Nikolopoulos, K., Syntetos, A. A., Boylan, J. E., Petropoulos, F., Assimakopoulos, V., 2011. An aggregate–disaggregate intermittent demand approach (adida) to forecasting: an empirical proposition and analysis. Journal of the Operational Research Society 62 (3), 544–554.

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

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

Pindoriya, N., Singh, S., Singh, S., 2008. An adaptive wavelet neural network-based energy price forecasting in electricity markets. IEEE Transactions On power systems 23 (3), 1423–1432.

Ramos, P., Santos, N., Rebelo, R., 2015. Performance of state space and arima models for consumer retail sales forecasting. Robotics and computer-integrated manufacturing 34, 151–163.

Rostami-Tabar, B., Babai, M. Z., Syntetos, A., Ducq, Y., 2013. Demand forecasting by temporal aggregation. Naval Research Logistics (NRL) 60 (6), 479–498.

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

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

Siino, M., Scudero, S., Cannelli, V., Piersanti, A., DAlessandro, A., 2019. Multiple seasonality in soil radon time series. Scientific reports 9 (1), 1–13.

Silver, E. A., Peterson, R., 1985. Decision systems for inventory management and production planning. John Wiley & Sons Inc.

Silver, E. A., Pyke, D. F., Peterson, R., Thomas, D. J., 2017. Inventory and production management in supply chains. CRC Press, 4th ed.

Svetunkov, I., 2017. Smooth: Forecasting using smoothing functions. R package version 2.0. 0.

Syntetos, A. A., Boylan, J. E., Disney, S. M., 2009. Forecasting for inventory planning: a 50-year review. Journal of the Operational Research Society 60 (sup1), S149–S160.

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

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

Taylor, J. W., 2008. An evaluation of methods for very short-term load forecasting using minute-by-minute British data. International Journal of Forecasting 24 (4), 645–658.

Taylor, J. W., 2010a. Exponentially weighted methods for forecasting intraday time series with multiple seasonal cycles. International Journal of Forecasting 26 (4), 627–646.

Taylor, J. W., 2010b. Triple seasonal methods for short-term electricity demand forecasting. European Journal of Operational Research 204 (1), 139–152.

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

Taylor, J. W., McSharry, P. E., 2007. Short-term load forecasting methods: An evaluation based on European data. IEEE Transactions on Power Systems 22 (4), 2213–2219.

Taylor, J. W., Snyder, R. D., 2012. Forecasting intraday time series with multiple seasonal cycles using parsimonious seasonal exponential smoothing. Omega 40 (6), 748–757.

Teunter, R. H., Duncan, L., 2009. Forecasting intermittent demand: a comparative study. Journal of the Operational Research Society 60 (3), 321–329.

Trapero, J. R., Cardos, M., Kourentzes, N., 2019. Empirical safety stock estimation based on kernel and GARCH models. Omega 84, 199–211.

Trapero, J. R., Kourentzes, N., Martin, A., 2015. Short-term solar irradiation forecasting based on dynamic harmonic regression. Energy 84 (1), 289–295.

Van Donselaar, K., De Kok, T., Rutten, W., et al., 1996. Two replenishment strategies for the lost sales inventory model: A comparison. International Journal of Production Economics 46, 285–295.

van Donselaar, K. H., Broekmeulen, R. A., 2013. Determination of safety stocks in a lost sales inventory system with periodic review, positive lead-time, lot-sizing and a target fill rate. International Journal of Production Economics 143 (2), 440–448.

Van Garderen, K. J., 2001. Optimal prediction in loglinear models. Journal of Econometrics 104 (1), 119–140.

Van Garderen, K. J., Lee, K., Pesaran, M. H., 2000. Cross-sectional aggregation of non-linear models. Journal of Econometrics 95 (2), 285–331.

Van Heerde, H. J., Leeflang, P. S., Wittink, D. R., 2002. How promotions work: SCAN*PRO-based evolutionary model building. Schmalenbach Business Review 54 (3), 198–220.

Waller, D., Boylan, J., Kourentzes, N., 2019. Sources of bias in loglinear models for retail. Working paper.

Widiarta, H., Viswanathan, S., Piplani, R., 2007. On the effectiveness of top-down strategy for forecasting autoregressive demands. Naval Research Logistics (NRL) 54 (2), 176–188.

Withycombe, R., 1989. Forecasting with combined seasonal indices. International Journal of Forecasting 5 (4), 547–552.

Wittink, D. R., Addona, M. J., Hawkes, W. J., Porter, J. C., 1988. SCAN* PRO: The estimation, validation and use of promotional effects based on scanner data. Internal Paper, Cornell University.

Wittink, D. R., Porter, J. C., Gupta, S., 1993. Dangers in using market-level data for determining promotion effects. MSI, 93–115.

Zipkin, P., 2008. Old and new methods for lost-sales inventory systems. Operations Research 56 (5), 1256–1263.