Erasmus University Rotterdam

Erasmus School of Economics

Master Thesis Econometrics and Management Science

Bayesian Lasso for Retail Demand Forecasting

Abstract

This research examines the use of Bayesian shrinkage priors for demand forecasting of retail stores. The data consists of the daily demand drivers sales, transactions and footfall of ten jewellery stores. The first model utilises seasonality factors and a time trend in combination with a Bayesian lasso prior to the coefficients. The second model exploits a hierarchical setting between the stores and the third model uses Bayesian autoregression between the demand drivers. The models are compared with conventional forecasting methods such as XGBoost, Random Forest and ARMA. The Bayesian methods outperform these methods by having a smaller Mean Arctangent Absolute Percentage Error.

Company Supervisor:
Ties van den Ende
Authors:
Lennard van der Plas (470328)

Supervisor:
dr. Annika M. Camehl
Second assessor:
dr. Wendun Wang

Date final version: DATE, 2021

Contents

1 Introduction

Accurate demand forecasting is one of the main aspects of successfully managing a retail store because many strategic decisions are based on expectations of the future. Demand forecasting is hard because demand is dependent on a lot of factors. Currently, most forecasts made by companies are based on human experiences, such as managerial judgement or a sales force composite which off the forecasts of the field workers are aggregated (Peterson, 1993). Often, human forecasts have lower accuracy than model forecasts (Takenaka and Shimmura, 2011). Peaks of demand can roughly be predicted by managers, such as in the Christmas period and the summer season, but smaller effects are difficult to estimate. For example, the difference in demand between a Monday morning and a Thursday evening is a lot more difficult for a person to capture. This inability leads to forecasts with a lower accuracy which results in an inefficient business with a higher loss and larger amount of expired inventory. Having a method for accurate demand forecasting gives managers a better base to decide their purchasing of inventory and their staff schedule.

To forecast the store demand we make use of seasonality factors such as previously mentioned, the effect of a certain weekday or a certain month. Ehrenthal et al. (2014) have shown that seasonality is one the most important factors for demand. Seasonality plays a major part as demand is the outcome of people's decisions to buy something or not. These decisions are mostly based on patterns of time. People often go to a particular store at a specific moment of the week just as they decide to spend more at the end of the month when they just received their salary. Therefore the most changes in demand can be traced back to these seasonality effects and is it important to take them into account. Next to that, a store can grow or decrease in customer base over time. Therefore it is important to also include a time trend in the model.

Research is already done for data-driven methods, such as linear regression and moving averages models, but mostly only predicting the total amount of product sales (Miller et al., 1991; Takenaka and Shimmura, 2011; Tanizaki et al., 2019). These methods are an improvement to human predictions, but still do not fully solve the problem. To forecast the number of sales gives the manager insights to base his schedule on, but a manager prefers to also have forecasts on other variables. These models lack in the sense that they have not been developed to also forecast other demand drivers. This is important because the work schedule of a store also depends on other demand drivers such as the number of people entering the store and the number of transactions. The number of required cashiers depends on the number of transactions and less on the number of sales. Other job types like front door greeter and security are dependent on the number of people in the store.

To extend the methods of previous research, forecasting of demand can be done for different demand drivers, for example, transactions, sales and the number of people entering the store which we will from now on name footfall. In this research, we focus on forecasting transactions, sales and footfall of a retail store one month in advance, accounting for time trends and seasonalities. This is different from previous research because demand forecasting is normally done in a univariate setting. To do this, we use three Bayesian methods with shrinkage priors. The first model is a univariate

model as a benchmark, the second is a hierarchical model that incorporates multiple stores and the third is a autoregressive model that uses the information of the three demand drivers together. Using hierarchical Bayes, we can set a type of coherence between the seasonality effects of the stores, which gives a mix of separate regression and pooled regression. To estimate the demand, we use variables like time trend and seasonality dummy variables with, for example, the day of the week and month of the year. Further, the Bayesian autoregressive model allows for different distributions of the error term, which is important when the dependent variables of the model have different conditional distributions. This is useful for time series with different types of data within the same model, such as one variable being continuous and the other being count data. Lastly and most importantly, the demand forecasting model contains a lot of coefficients for the many seasonality factors and time trends. With Bayesian priors, we deduct the problem of multicollinearity and makes it possible to incorporate beliefs. These increase the amount of information we have relative to the number of estimated coefficients. It reduces the variance in the model and the possibility of overfitting.

We solve this problem by using shrinkage priors and incorporating multiple time series in two of the Bayesian models. This is important for the work schedule as the number of required available cashiers is dependent on the number of transactions, but the amount of stock clerks is dependent on the number of sales. Having a prediction for all three makes solving an optimal work schedule more efficient.

We practice the estimation of the demand drivers with three models. The first model is a linear seasonality trend model (LST). The model is developed by Posch et al. (2020) and consists of a spline trend function and a seasonality term. It incorporates the Bayesian lasso to regularise a large number of coefficients. The seasonality function captures effects like the day of the week and day of the month where demand is highly dependant. In our model specification, we assume a multiplicative effect of the seasonalities that stays constant over time. To capture the growth or decline of demand over the time period, we use a linear time trend with a changing slope. We use a spline function that divides the time period into different domains between knots, where on every domain the slope of the time trend can change by a small amount. This spline specification has also been used by Gasthaus et al. (2019) and Ugarte et al. (2009) Further, the model allows for examination of the coefficients which could explain the forecasts. This helps the manager with his choice of inventory and staff management. The second model is a hierarchical LST (HLST), which is an extension of the first model. The HLST is a hierarchical random coefficients model that uses a hierarchical setting for the coefficients across stores. This model ensures coherence between stores such that we can use information from one store to estimate the effects of other stores. It can be seen as a combination of separate and pooled regression. The last model is a Bayesian autoregressive model (BARX), it incorporates the lagged values of the time series to capture the relationship between the transactions, sales and footfall. In comparison to the prior two methods, the BARX has the advantage to use the information of all three demand drivers together in one specification. Although most of the demand is driven by seasonality, it is also influenced by the near

past. Likewise, the demand drivers also influence each other, an increase of people who orientate on their purchase which is measured in footfall could result in an increase in sales the week after.

To measure the performance of the three models, data of ten jewellery stores in the US is used with the daily amount of sales, the number of transactions and footfall in the period January 2016 to April 2020. We find that the three Bayesian methods almost always is better than the conventional methods by having a smaller Mean Arctangent Absolute Percentage Error. The HLST model performs best for data sets longer than a year, whereas the BARX performs best for shorter data sets. It becomes evident that including shrinkage priors is the most important. Further, we conclude that incorporating information from multiple stores or multiple demand drivers has a positive effect on forecasting performance. For a longer forecasting window, such as two months, the Bayesian methods perform relatively worse.

2 Literature

There are a number of studies that analyse demand forecasting for retail stores. According to Geurts and Kelly (1986) accurate demand forecasting enables management to select appropriate levels of inventory and have a basis to make a shift schedule for the staff. We discuss methods from the frequentist approach in subsection 2.1 as well as the Bayesian approach in subsection 2.2.

2.1 Frequentist approach

We begin the review of the frequentist literature with Miller et al. (1991). This research compares a couple of simpler models such as moving average and exponential smoothing. Takenaka and Shimmura (2011) use linear regression to predict the daily demand of restaurants in Japan. The variables they took into account were holidays, certain temperature thresholds and days of the week. Their method outperformed most managers in forecasting. Managers also had the chance to learn the value of the underlining effects of customer demand. More recently, Tanizaki et al. (2019) used linear Bayes regression, boosted decision tree regression, decision forest regression and stepwise regression where all methods performed similarly. The forecast improved with the addition of number of reservations a couple of days before. The ARIMA and Holt-Winters model are compared by Da Veiga et al. (2014) in performance for demand forecasting. The Holt-Winters model performed better, having a better adjustment and capturing the linear behaviour of the series. Machine learning methods such as gradient-boosted decision trees and neural networks have been researched by Huber and Stuckenschmidt (2020). They conclude that machine learning methods outperform time series models and linear regularisation models. The Bayesian models in this research also make use of regularisation in the way of Bayesian lasso. Therefore it is interesting to see how the Bayesian methods compare with machine learning methods used as a benchmark in this thesis. A large review of multiple methods is done by Pavlyshenko (2019) on weekly sales of a supermarket including but not limited to linear regression, Bayesian regression, decision forest regression and boosted decision tree regression. They concluded that boosted decision tree regression

performed the best, but a simple normal prior was used for the Bayesian regression. In this research, we also compare the Bayesian regression method with forest-based methods such as random forest regression and XGBoost.

2.2 Bayesian approach

Bayesian methods take a different assumption for the coefficients by interpreting them as stochastic variables. This enables the use of regularisation priors on the coefficients that enable a type of feature selection. While regularization is also possible for the frequentist approach Park and Casella (2008) argue that the Bayesian regularisation is better than Lasso and ridge regression. For different values of the shrinkage parameters, the Bayesian Lasso produces smooth paths of the estimates of the coefficients like ridge regression but also pulls weak parameters fast to zero like Lasso regression. Therefore Bayesian Lasso is a good technique in this experiment. Regularization is important for demand forecasting, as it consists of a lot of seasonalities with many variables, where data is limited relative to the number of parameters.

In Posch et al. (2020) the problem of predicting food transactions is covered using Bayesian general additive models (GAM) with a normal distribution and a negative binomial distribution for the conditional of the demand. The GAM use a spline function for trend and dummy variables for seasonality. Further, their method includes using a Bayesian Lasso by using Laplace priors for coefficients of the trend and seasonality function as suggested by Park and Casella (2008). In this thesis, this model is taken as a basis to develop an easy to use method for demand forecasting. We extend their model by also using it for footfall and sales, which is a continuous variable. Further, Posch et al. (2020) only cover the problem with an univariate approach. We broaden their research by extending their models to a hierarchical specification between stores and an autoregressive specification between demand drivers. The forecasting for products with a low amount of sales is done by Pitkin et al. (2018) and uses Bayesian hierarchical modelling. They emphasise how to deal with a large number of zero observations.

This research will deal with the usage of a Bayesian VAR model in a business type setting, where data has different characteristics compared to finance or macroeconomics data. Where most financial variables of interest follow a random walk, the number of transactions does not. It is much more influenced by seasonalities and less by lagged values. Further, the number of daily transactions of a store can take the value zero whereas the price of a stock does not. This means that the typical Bayesian autoregressive setting has to be adapted to be used in a business setting. Yelland and Dong (2014) use Bayesian methods to forecast the demand of fashion goods but used simulated data. They specify the model in a hierarchical Bayesian setting. With the hierarchical setting, a relation can be defined between the fashion goods that gives a proper mix between separate estimation and pooled estimation of the fashion goods. In this research, the hierarchical Bayesian setting will be used to define the coherence between the different retail stores. Wong et al. (2006) use a Bayesian vector autoregressive model to predict tourism demand of countries. We bring value by to use the autoregressive setting for micro economic variables like demand drivers of the store, such

as transactions, sales and footfall. Estimating the model in a this design enables the relationship between demand drivers instead of estimating the time series separately. It is shown in Todd (1990) that using a autoregressive setting causes great forecasting capabilities. In the research of Mercy and Kihoro (2015) a vector autoregressive model is compared with a univariate SARIMA model, where the former model came out on top showing the added value of an assembled autoregressive model.

A complex method is used by İşlek and Öğüdücü (2015). They have a procedure where they first cluster similar warehouses. Thereafter, A hybrid model using a moving average model and Bayesian Network machine learning algorithm is applied. This method produces great forecasts, but has a longer computation time, is harder to implement and is not understandable for a manager. These are concepts that we take into consideration for the methods of this thesis by having easy to understand coefficients.

3 Data

To test the performance of the demand forecasting methods, data will be used from 10 stores of a jewellery franchise in the US. The data set contains information about the number of transactions, sales and footfall on a daily level. Transactions is the number of payments made. Sales are the sum of the amounts of transactions. The footfall is the number of customers entering a store. The time period of the sales and transactions data runs from January 2016 to April 2020 but varies a couple of months per store. The footfall data runs from April 2019 to April 2020. An example of a time series can be seen in Figure1. Most notably, there is a clear case of seasonality with large peaks at the end of the year due to Christmas. Another reoccurring peak is at the beginning of May related to Mother's Day. It is notable that the time series do not have a clear upward or downward trend and are volatile over time. Days, where the store was closed, are not taken into account for this research.

Figure 1: Timeseries of the footfall, sales and transactions for one of the jewellery stores.

We further investigate this particular store with Figure 2. In Figure 2a we see the average demand over the week. It is clear that the demand is higher at the weekend, especially Saturday. The items that this franchise sells do not have an immediate need and could be planned by the customers to be bought at the weekend. Sunday shows a decline relative to Saturday as there are shorter opening hours on Sundays. 2b describes the average demand per month. It can be reconfirmed that May and December have a peak. For May this peak is due to Mother's day and a yearly promotion which let customers earn credit for spending. In December the peak is due to the Christmas period. 2c shows the average demand for every day in the month. Peak demands such as Mother's Day and Christmas have been taken out as they pollute the data with outliers. It is notable that demand grows during the month. A probable cause for this is that most wages are paid at the end of the month. In the last figure, 2d, we have removed the seasonality of the demand as described by Nau (2014). This is done by determining the average increase of seasonality in relation to the average demand. For example, for the effect of Saturday, we divide the average demand of all Saturdays by the average demand and find that it is an increase of 50%. Then we divide all Saturdays by 1 + 50%. We repeat this for all weekdays, months and month days. This results in the blue line which has smaller peaks than the actual demand shown in Figure 1. The orange line is the moving average with a window size of 30 days. As the line changes over the demand we can see that there is some local time trend that is not explained by the seasonality. Therefore, we assume that a local time trend does exist in these series.

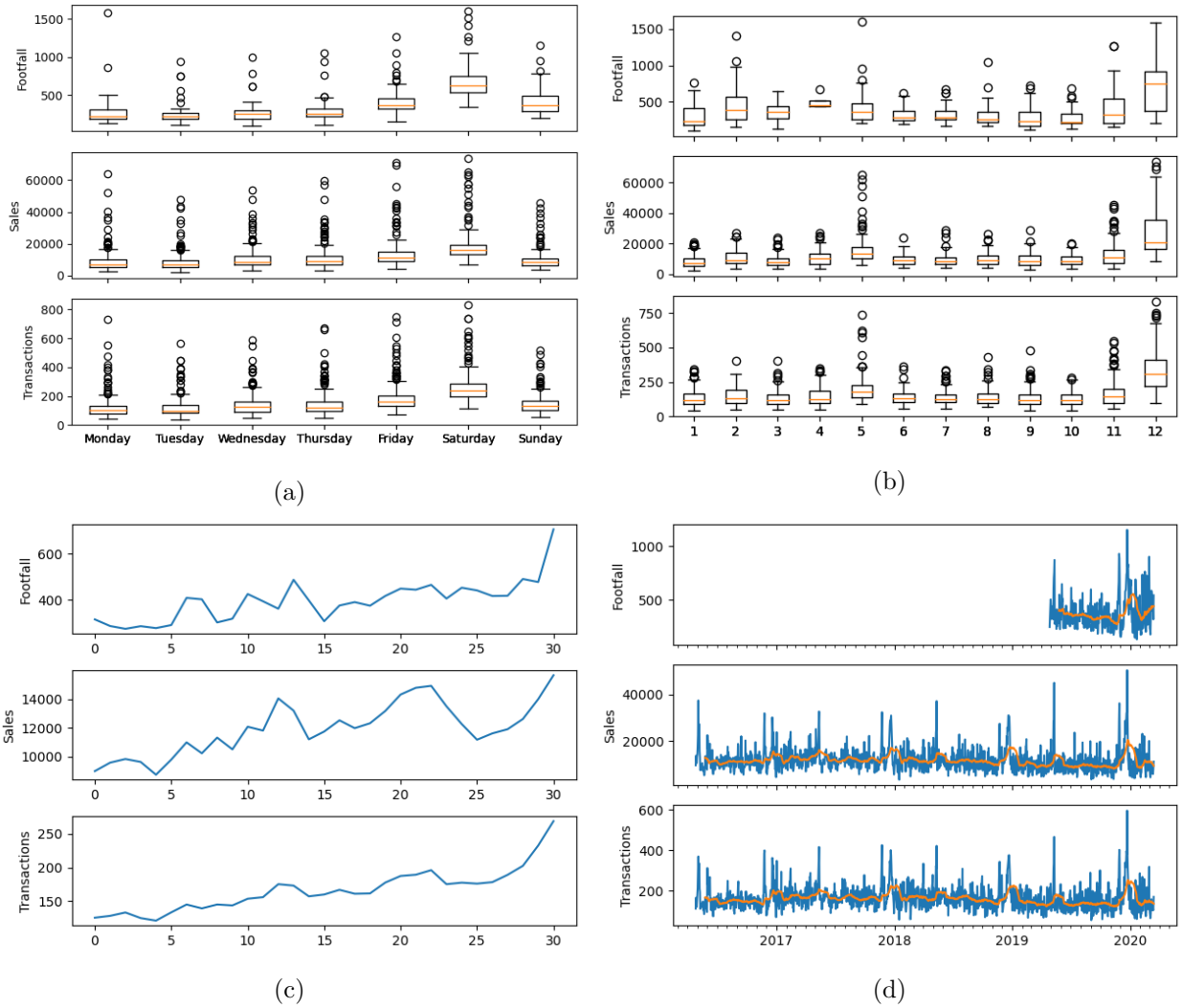

Figure 2: A set of four subfigures: (a) describes the distribution of the demand for every weekday; (b) describes the distribution of the demand for every month; (c) describes the average demand for every day in the month; and, (d) demand with the seasonality removed. The orange line is a moving average with a window size of 30.

With this preliminary research, we conclude this section. We used one store as an example to get more insights into the data. For this research, we use a total of 10 stores of the same jewellery franchise. These stores vary in size, resulting in different scales of demand. The stores however show similar patterns as the store in the figures above. Therefore we assume that there is coherence between the effects of the stores.

4 Methodology

We are using data on a day level to forecast the demand of a retail store. The data for a store consists of the sales, amount of transactions and the footfall which we model differently. We denote

the value of store r of demand driver i at day t as y_{itr}, where $i \in \{sales, transactions, footfall\}$ and $r \in \{1, ..., 10\}$. This describes the observation where the Poisson and negative binomial distributions are used for the count data footfall and transactions. For the sales data, a log-normal distribution is used.

An advantage of using the Bayesian models in this setting is that different distributions can be easily be appointed. Next, the models give more information than a point forecast, because they give a whole distribution, which can be useful to create credible prediction intervals. Most importantly, when estimating a large number of coefficients, Bayesian techniques can reduce the amount of variance in the model by use of priors. Park and Casella, 2008 show that Bayesian Lasso estimates are a compromise between the frequentist Lasso and ridge regression estimates. They argue that the Bayesian Lasso pulls uninformative parameters to zero harder than ridge regression, which helps with the elimination of many parameters. In the next subsection, we start with the specification of the time trend and seasonality function which we use to estimate y_{itr}.

4.1 Time trend and seasonality function

The demand y_{itr} depends on a time trend and seasonality effects. For a time series, the variable t is the number of days since the first date t_0, where $t_0 = 0$. t_T is the last date in the time series.

The demand is mainly dependent on the seasonality factors, but also influenced by local trends. This could be caused by a growing customer base or a temporary promotion. To capture these events, we include a trend function that is dependent on time. A linear trend cannot capture temporary changes in the time trend, therefore we use splines. A spline function is a piecewise polynomial having different functions for different segments of the domain of the variable. For the trend function, this variable will be t, the time in days since the first observation. The separation is declared by knots, which are specified by the user. From the first observation in the time series, there is a knot after every thirty days. We chose thirty days because we want to capture deviance in the level of demand that happen over a month or more. We choose the first-order polynomial because a spline of higher-order has the tendency to pick up noise and peaks of the seasonalities. A first-order spline results in linear functions between the knots. The spline function has one requirement: two polynomials next to each other should connect such that they have the same value at the knot. Because the linear polynomial has two variables and one condition there is one free variable per knot open to change. This will result in a slightly different function between every pair of knots. Take the set of knots $k_1 < ... < k_m$, where m is the amount of knots. To capture the effect of a domain between two knots we use the indicator function

$$(t - k_j)_+^l = \begin{cases} (t - k_j)^l, & t \geq k_j, \\ 0, & \text{else.} \end{cases} \tag{1}$$

Here, j is the index of the knot which is between 1 and m. This function gives the positive distance between day t and the knot k_j. This means that if t is lower than the value of the knot, it

will return zero, but if it is higher, it will return the distance. This distance is then raised to the power l, which is in the case of a linear spline $l = 1$. Using the distances with the indicator function enables us to use the splines specification in an easier setting to estimate the coefficients.

We specify the variables for the time trend with coefficients γ:

$$g(t) = \gamma_1 + \gamma_2 t + \gamma_3 (t - k_1)_+ + ... + \gamma_{m+2} (t - k_m)_+ . \tag{2}$$

At each knot k_i, the function changes which is captured in the coefficients γ_{i+2}. Note the coefficients that influence the global function, γ_1 and γ_2. These values impact the overall behaviour of the linear spline. The coefficients γ_3 to γ_{m+2} influence the slope of the function in the domains of the knots and are named the local coefficients. To elaborate, before the first knot k_1, the slope of the spline function is γ_2, but after just passing k_1 the slope becomes $\gamma_2 + \gamma_3$. Because we assume the local changes to be minimal, we expect small values for γ_3 to γ_{m+2}, which will contribute to the priors we impose on these local coefficients. We expect the coefficients of the global function to be larger, therefore we appoint a less strict prior.

The time series of the demand of a store often has a strong seasonality segment, sales in December are often higher than in June (Chandra and Chaloupka, 2003). To capture this effect we add a set of dummy variables for the *month of the year*, *day of the month* and *day of the week* having values 0 or 1. This leads to a total of 50 variables $x_1, ..., x_{50}$ which gives us the seasonality function:

$$s(t) = \beta_1 x_1(t) + \ldots + \beta_{50} x_{50}(t). \tag{3}$$

To elaborate, β_1 to β_{12} give the effect of the months, β_{13} to β_{43} the effects of the day of the months and β_{44} to β_{50} the effect of the weekday on the demand. Whereas normally one variable is left out of the model with categorical dummy variables, we include all options as we use a form of shrinkage. Leaving one option of every category out would influence the model heavily depending on the variable left out. The time trend and seasonality function are the base of the model. The following subsection will show how we use these to estimate the demand.

4.2 Linear seasonality trend model

The first Bayesian model is a generalised additive model, which assumes an additive effect of the variables on the demand. The model takes the seasonality and time trend into account, thus named the Linear seasonality trend model (LST). The LST is based on the model specification of (Posch et al., 2020). The assumption is that demand is based on a trend in the time and seasonality effects. In the case of count data, namely transactions and footfall, we use a conditional negative binomial distribution for $y_{trans,t,r}$ and $y_{foot,t,r}$. We declare them as:

$$y_{i,t,r} \mid \boldsymbol{\beta}, \boldsymbol{\gamma}, a \sim \text{NegBinom}\left(\exp(g_{ir}(t) + s_{ir}(t)), \frac{1}{a_{ir}^2}\right), \tag{4}$$
$$\text{for } i \in \{transactions, footfall\} \text{ and } r \in \{1, ..., 10\},$$

Potential heteroskedasticity for larger values of demand is counteracted by the exponential transformation. Therefore, we assume that the mean of the demand varies over time, but the variance does not. The functions $g_{ir}(t)$ and $s_{ir}(t)$ are the trend function and seasonality function which have different coefficients for every demand driver and store. The exponent function ensures that the mean of the distribution will always be positive. Next to that, the trend and seasonality have a multiplicative effect on the demand, because we have established an exponential relation between the demand and the independent variables. β_{ir} and γ_{ir} are the coefficients of the trend function and seasonality effects. The parameter of the variance is a_{ir}. As advised by Gelman (2020), the prior for a_{ir} is the half-standard normal prior $a_{ir} \sim N^{+}(0, 1)$. When assigning the half-normal prior directly to the variance term of the Binomial distribution, most of the prior mass would be on over-dispersed models.

We assume a log-normal distribution for the sales data for a couple of reasons. First, it restricts the model from estimating negative values for the number of sales. Second, the sales have a right-skewed distribution. A log transformation lowers large values of sales and takes out heteroscedasticity for the large values. Third, the coefficients represent a multiplicative term to the sales which is scale-invariant, for example, a 20% increase in sales on Wednesday. To simplify the sampling, we take the logarithm of the sales and use the normal distribution.

$$\log(y_{sales,t,r}) \mid \boldsymbol{\beta_{sales,r}}, \boldsymbol{\gamma_{sales,r}}, \sigma_r^2 \sim \mathrm{N}\left(g_{sales,r}(t) + s_{sales,r}(t), \sigma_r^2\right), \quad \text{for } r \in \{1, ..., 10\}, \tag{5}$$

The trend function has a lot of coefficients to estimate, it needs regularisation. As described by Posch et al. (2020), we induce a Bayesian Lasso prior on these coefficients by Park and Casella (2008). The Bayesian Lasso regularises by pulling the coefficients to zero. We make a distinction between the variables that influence the global shape of the spline and the variables that influence the domains between the knots. For the latter, we induce a strict prior as we assume that that the temporary deviance is small and do not want to capture potential noise. Therefore we use the Laplace distribution as a prior, also known as the double exponential distribution, for the spline region-specific coefficients, γ_j for $j \geq 3$. When assuming a normal distribution with variance σ^2 for the likelihood function of $\log(y)$, we get the prior conditional on σ^2.

$$\gamma_{3,sales,r} \left| \sigma_r^2, ..., \gamma_{m+2,sales,r} \right| \sigma_r^2 \underset{i.i.d.}{\sim} \text{Laplace}\left(0, \frac{\sqrt{\sigma_r^2}}{\tau_1}\right) \quad \text{for } r \in \{1, ..., 10\}, \tag{6}$$

The tuning parameter τ_1 has to be specified and influences the sparsity of the estimator of the coefficients. We do not make use of a prior on this tuning parameter as it complicates the sampling and lessens computational efficiency, which is not what we are aiming for with this research. We use predefined values for the hyperparameters which will be discussed in subsection 4.5. A larger value for τ_1 restricts the parameters more around zero, whereas a small value gives them more freedom.

Park and Casella (2008) have shown conditioning on σ^2 is important to have unimodality in the posterior when using a Laplace prior, which otherwise would lead to two optima in the posterior. As we use the mean estimate when we need point estimates, a unimodal posterior will lead to better estimates. For the transactions and footfall, we assumed a negative binomial distribution, which leads to the prior:

$$\gamma_{3,i,r}, ..., \gamma_{m+2,i,r} \underset{i.i.d.}{\sim} \text{Laplace}\left(0, \frac{1}{\tau_1}\right) \tag{7}$$
$$\text{for } i \in \{transactions, footfall\} \text{ and } r \in \{1, ..., 10\},$$

It can be noted that when using the negative binomial distribution, there is no need of including the variance term in the prior as unimodality in the posterior is already guaranteed.

As specified earlier, we assume that there is a global time trend with a small deviation. Therefore we use a strict Laplace prior for the knot variables. Contrary to that, we want the coefficients that determine the global trend to be less restricted. Hence, we do not want to pull them hard to zero, but only regularise them around zero. We do this with Bayesian ridge regression by using a normal distribution prior

$$\gamma_{2,i,r}, \gamma_{3,i,r} \underset{i.i.d.}{\sim} N\left(0, \tau_2^2\right) \tag{8}$$
$$\text{for } i \in \{sales, transactions, footfall\} \text{ and } r \in \{1, ..., 10\},$$

Here the second tuning parameter τ_2 has the same effect as τ_1 influences the freedom around zero. The normal prior does not endanger the unimodal posterior, therefore it is not necessary to include σ^2 in the prior. Lastly, we assign an improper and diffuse prior $p(\gamma_1) \propto 1$ to the intercept of the trend function.

As a prior for the seasonality coefficients, we again apply the Bayesian Lasso by Park and Casella (2008). This gives us the Laplace distribution as a prior where we use tuning parameter τ_3. For the case of a normal conditional distribution with sales we use

$$\beta_{1,sales,r} \left| \sigma_r^2, ..., \beta_{s,sales,r} \right| \sigma_r^2 \underset{i.i.d.}{\sim} \text{Laplace}\left(0, \frac{\sqrt{\sigma_r^2}}{\tau_3}\right) \tag{9}$$
$$\text{for } r \in \{1, ..., 10\}.$$

In the case of sales and footfall, we exclude the variance term.

$$\beta_{1,i,r}, ..., \beta_{s,i,r} \underset{i.i.d.}{\sim} \text{Laplace}\left(0, \frac{1}{\tau_3}\right) \tag{10}$$
$$\text{for } i \in \{transactions, footfall\} \text{ and } r \in \{1, ..., 10\}.$$

As the seasonality function does not have a discrepancy in global and local variables, all variables have the same prior. We assign a diffuse prior for the variance of the normal distribution, $p(\sigma_r^2) \propto \frac{1}{\sigma^2}$.

The specification of the LST model is univariate, meaning we do not establish a relation between the stores or different demand drivers. Therefore we estimate the models separately per store per demand driver.

4.3 Hierarchical semi-pooled model

The data contains information about multiple stores on their three demand variables. Stacking the data sets of the stores results into panel data. We denote the demand i of store r at time t as $y_{i,t,r}$. The previous model specification does not exploit the panel structure but has a model separately for each store. We extend the formulation in subsection 4.2 in the way that we exploit the panel structure in the priors for the coefficients. Therefore we refer to the model by hierarchical linear seasonality trend model (HLST). We assume that there is a relation between the seasonality effect of the stores. This can be seen in the model specification by the use of a $\tilde{\beta}_i$, which captures the overall seasonality effects of all stores. It is plausible that the seasonality effect of December on sales has a similar effect on store r as on store r^-. It is possible that store r is a lot bigger than store r^- with a lot more sales overall, which results in a different scale of the dependent variable. This is countered by the log transformation which ensures that the coefficients have a multiplicative effect so that they are less affected by the overall scale of the particular demand of the store. We do not pool the coefficients of the time trend as the data sets of the stores overlap not perfectly and the placements of the knots is store specific which would lead to unbalanced panel data. This is not an issue for the seasonality effects as we have the seasonality variables for every time series available. We assume that the effect of seasonality is similar but not the same across the stores, therefore we use a random coefficients model. We extend the previous specification in the following way:

$$\gamma_{i,r} = [\gamma_{1,i,r}, ..., \gamma_{m+2,i,r}]' \tag{11}$$

$$\beta_{i,r} = [\beta_{1,i,r}, ..., \beta_{s,i,r}]' \tag{12}$$

$$\beta_{i,r} \sim \text{MVN}(\tilde{\beta}_i, \Sigma_{\tilde{\beta}_i}), \tag{13}$$

$$\tilde{\beta}_i = [\tilde{\beta}_{1,i}, ..., \tilde{\beta}_{s,i}]' \text{ for } i \in \{sales, transactions, footfall\}. \tag{14}$$

$$\tilde{\beta}_{1,sales} \left| \sigma^2, ..., \tilde{\beta}_{s,sales} \right| \sigma^2 \underset{i.i.d.}{\sim} \text{Laplace} \left(0, \frac{\sqrt{\sigma^2}}{\tau_3}\right), \tag{15}$$

$$\tilde{\beta}_{1,i}, ..., \tilde{\beta}_{s,i} \underset{i.i.d.}{\sim} \text{Laplace} \left(0, \frac{1}{\tau_3}\right) \text{ for } i \in \{transactions, footfall\}, \tag{16}$$

$$\Sigma_{\beta_i} \sim \text{InvWish}(I, s+1). \tag{17}$$

As indicated, the seasonality function and time trend are subjective to the store. The number of coefficients of the time trend function is $m + 2$ and for the seasonality function equal to s. We see $\gamma_{i,r}$ as the vector containing the coefficients of the time trend for store r with length $m + 2$. These coefficients have the same normal and Laplace prior as discussed in subsection 4.2 and do not exploit the panel structure as explained before. Further, we have $\tilde{\beta}_i$ as the overall coefficients for

the seasonality function for demand i having length s. It can be regarded as the pooled seasonality effect across all stores, where the elements define the specific effects in a pooled setting. $\Sigma_{\tilde{\beta}_i}$ is the covariance matrix of the multivariate distribution with size $s \times s$ for demand i. The $\beta_{i,r}$ is the store-specific vector of seasonality effects for store r and demand i. It is related to $\tilde{\beta}_i$ as it can be seen as a variant of the pooled effect. As suggested by Lemoine (2019), it has an Inverse Wishart prior which is conjugate for the multivariate distribution. This specification is an extension of the research of Posch et al. (2020) as they did not use a hierarchical specification. We estimate the models separately for every demand driver.

4.4 Bayesian autoregressive model

The third model is a Bayesian autoregressive model (BARX). With this model, lagged values of multiple demand drivers of one store are used to estimate the current value of a demand driver. The demand driver $y_{i,t,r}$ contains the values for $i \in [sales, transactions, footfall]$ at time t of store r. As we assume for the sales a log-normal distribution and the other two a negative binomial distribution, we do not choose to establish a vector model using a multivariate distribution. We induce the previously defined conditionals on the demand drivers. The BARX takes lagged demand of the time series into account. In section 3 we discussed the relation between the values for sales, transactions and footfall. The coherence of these variables makes them a great candidate for this model, where knowledge of one variable is extra information for the other variable. The model uses previous values of the variables to update current values.

To extend the LST described in subsection 4.2 we use the lagged values of demand. This results in the following model:

$$\begin{aligned} y_{i,t,r} &= \theta_{i,r,1} y_{sales,t-1,r} + \theta_{i,r,2} y_{transactions,t-1,r} + \theta_{i,r,3} y_{footfall,t-1,r} + \cdots \\ &\quad + \theta_{i,r,3p-2} y_{sales,t-p,r} + \theta_{i,r,3p-1} y_{transactions,t-p,r} + \theta_{i,r,3p} y_{footfall,t-p,r} \\ &\quad + S_{i,r}(t) + G_{i,r}(t) + \varepsilon_{i,t,r}, \end{aligned} \tag{18}$$

$$S_{i,r}(t) = \beta_{i,r} x(t), \tag{19}$$

$$G_{i,r}(t) = \gamma_{i,r} B(t) \text{ for }, \tag{20}$$

$$\text{for } i \in \{sales, transactions, footfall\} \text{ and } r \in \{1, ..., 10\}.$$

Here, $\theta_{i,r}$ are the coefficients of the autoregressive terms for demand driver i and store r. The order of the lagged variables is denoted by value p. In subsection 4.5 we cover the choice for this value. The time trend, $S_{i,r}(t)$, and seasonality function, $G_{i,r}(t)$, have independent coefficients for demand driver i and store r. $x(t)$ is a row vector with length S containing all the values of the seasonality dummies. The covariates of the time trend are captured in row vector $B(t)$ with length $m+2$. This implies that the time trend and seasonality function are taken into account for the time series $y_{i,t,r}$, but each has its own set of coefficients for these functions.

With a large value for p, the amount of coefficients increases drastically. This is because the amount of coefficients of the autoregressive part is equal to $9p$ meaning, if we would like to add an

additional lag order, we would add nine more coefficients. A larger amount of coefficients increases the variance of the model and is often bad for the forecasting performance. To counter this, we use a ridge normal prior for the coefficients of the autoregressive terms.

$$\theta_{i,r,w} \sim \mathrm{N}\left(0, \frac{1}{\tau_3}\right), \text{for } i \in \{sales, transactions, footfall\}, r \in \{1, ..., 10\} \text{ and } w \in \{1, \cdots, 3p\}.$$

The elements of $\beta_{i,r}$ and $\gamma_{i,r}$ have the same normal and Laplace prior as specified in subsection 4.2.

4.5 Implementation details

The use of a time trend function is discussed in subsection 4.1. In this function, the variable t is used to define the number of days from the first observation. The different time series do not have the same length, so the interval of t differs per time series. To make the priors distributions universal, we min-max standardise the variable t by dividing by the total amount of observations T. As a result variable $t \in [0, 1)]$.

The formulation of the models makes use of hyperparameters that can be predefined or set with a prior. To reduce the sampling time and complexity of the methods, we use predefined values for these parameters. With a couple preliminary experiments we find that the following values for the hyperparameters result in low forecasting errors:

$$\tau_1 = 5, \quad \tau_2 = 0.5, \quad \tau_3 = 6 \quad \tau_4 = 0.5. \tag{21}$$

We continue to use these values for the hyperparameters for the next experiments.

We obtain the forecasts of the models by simulating the forecasts with the sampled coefficients. This leads to a distribution for every forecasted point. To evaluate the forecast on prediction power, we use a point forecast. To get from a forecast distribution to a point forecast, we use the minimum mean square error estimation. This means that we calculate the mean of the forecast distribution and use that point to evaluate the metrics. We use the mean as it has a higher efficiency than the mode or the median.

For the BARX we use a lag length of seven days as it gives information about the current level of demand and includes the same weekday of the week before, which is highly correlated with the demand of the current day. The BARX model uses the three demand drivers of a single store. The length of the footfall data is much shorter than the sales and transactions data, therefore we only use the observations of the demand drivers where all three are known for the BARX model.

Furthermore, the models will be compared with some ready-to-use methods, such as ARMA, exponential smoothing, XGBoost and Random Forest. These models are widely used for demand forecasting and have splendid forecasting performance (Ke et al., 2017; Takenaka and Shimmura, 2011; Tanizaki et al., 2019).

Forecasting is done in a cross-validation setting where the last four months are used for testing and the data before for training. In the first iteration, we train the model on the train data. With

this model, we forecast a one month ahead forecast for the first test month on daily level. Comparing these forecasts with the real daily values of the first test month, we get our metrics. In the second iteration, we include the first test month in the new train data, train the model again and forecast for the second test month. This process is repeated for all four test months in the data set. Metrics are calculated on every run and averaged over the four folds. With ten stores each having three demand driver, we end up with 120 small test sets with a size of one month per model. Because we use we use the last four months of every data set we are biased towards those months. The best performing models perform best for those four months, but we do not know the performance for the other months. We still want to stick with the previously described setup as we are restricted by the computation time of the sampling process of the models.

We also investigate the performance of the models with a longer forecast horizon. In this second experiment we use a window of 2 months. Here we make two folds of the last four months of every data set.

The whole experiment will make use of Python and the PyStan package (Riddell et al., 2021), which is a Python interface for the Stan language. Stan is the conventional way of Bayesian inference (Carpenter et al., 2017). For the Bayesian models, Markov chain Monte Carlo sampling will be used. For all Bayesian estimations, we use 1000 draws, of which 150 are used for the burn-in phase. 350 draws are used to tune the parameters of the Hamiltonian Monte Carlo algorithm which influence the sampling efficiency. These are the default values for the Stan language. The remaining 500 draws are used for sampling and will give us the posterior distributions. In subsection 5.1 we show that these settings suffice and the algorithm is converging.

4.6 Predictive power

An important metric in demand forecasting is the Mean Absolute Percentage Error (MAPE) (De Myttenaere et al., 2016). An advantage of this metric relative to the Mean Squared Error is that the MAPE is scale-invariant because the error is divided by the actual value so that we get the error as a fraction that we are off. This makes sure that we can compare the models for data with different scales. A disadvantage of the MAPE is that we cannot calculate it when the real value is zero, as divisions by zero are going to infinity. To combat this, Kim and Kim (2016) propose the use of the mean arctangent absolute percentage error MAAPE, which by using the arctangent function take away this drawback, because the arctangent function runs to $\frac{\pi}{2}$ for very large values and divisions by zero. Therefore we decide to use this metric to assess the performance of the models. Further, we also measure the Mean Squared Error (MSE). With the Bayesian models, we can also do density forecasting. We will use these to calculate the average log density score (ALDS) to determine the model which has the best posterior distribution for the forecast value. To further elaborate on the metrics we declare y_t on $t \in [1, ..., T]$ as the real values in the prediction interval and $\hat{y_t}$ on $t \in [1, ..., T]$ as the predicted values.

- **MAAPE**: $\frac{1}{T} \sum_{t=1}^{T} \arctan\left(\left|\frac{y_t - \hat{y_t}}{y_t}\right|\right)$

- **MSE**: $\frac{1}{T}\sum_{t=1}^{T}(y_t - \hat{y}_t)^2$
- **ALDS**: $\frac{1}{T}\sum_{t=1}^{T}\log p\left(y_t \mid y, \mathcal{M}\right)$

The appendix also provides the metrics of the Weighted Average Percentage Error (WAPE) which can also deal with zero values just as the MAAPE by summing up the values of all test observations. The Mean Absolute Deviation (MAD) is the average absolute error. The Mean Forecast Error (MFE) denotes the bias of the forecasts. These metrics will be displayed in the Appendix as they are outside the scope of this research.

To test for significance for the best point forecasts we use a Diebold-Mariano test to test all models Diebold and Mariano, 1995. The Bayesian models are also tested with a Weighted Likelihood Ratio test by Amisano and Giacomini (2007) where we compare the density of the forecasts.

5 Results

We first investigate the convergence of the Bayesian models in subsection 5.1 We start by discussing the results of the experiment of the sales data in subsection 5.4. Subsection 5.5 provides insights on the performance of the transactions data. Further, subsection 5.6 covers the forecasting of the footfall. In subsection 5.7 the effects of a longer forecast window are investigated. To conclude this section, we cover remarks in subsection 5.8.

5.1 Convergence diagnosis

When using Bayesian statistical inference with MCMC sampling it is important to investigate if the sampler has converged and if there is a correlation between the samples. In such a case, more samples are needed.

All Bayesian estimations converge and have a potential scale reduction factor (PSRF) between 1 and 1.01, meaning that the scale of the distribution of the sampling does not change. If they would have a larger PSRF that would mean that the variance of the distribution would be smaller if sampling would continue. This would be a sign that the sampled distribution is not representative of the real posterior.

We use the samples of the coefficients and calculate the autocorrelation. We group the autocorrelation by γ, β, σ and θ. Then we use these boxplots to get insights into these autocorrelations. These boxplots can be seen in Figure 3. First of all, we have the blue dotted line indicating the zero axes and the red lines indicating the -0.1 and 0.1. We assume that when the autocorrelation is around zero and between -0.1 and 0.1, there is no serious autocorrelation. Autocorrelations that are 1.5 times the interquartile range larger than the upper quartile or smaller than the lower quartile are classified as outliers. This is done because of the large number of autocorrelations that are represented in the boxplot. We can see that the outer whiskers of the boxplot are almost all between the red lines, The median of the autocorrelations is always between the red lines and around zero. This shows us that there is no concern with the autocorrelation of the samples.

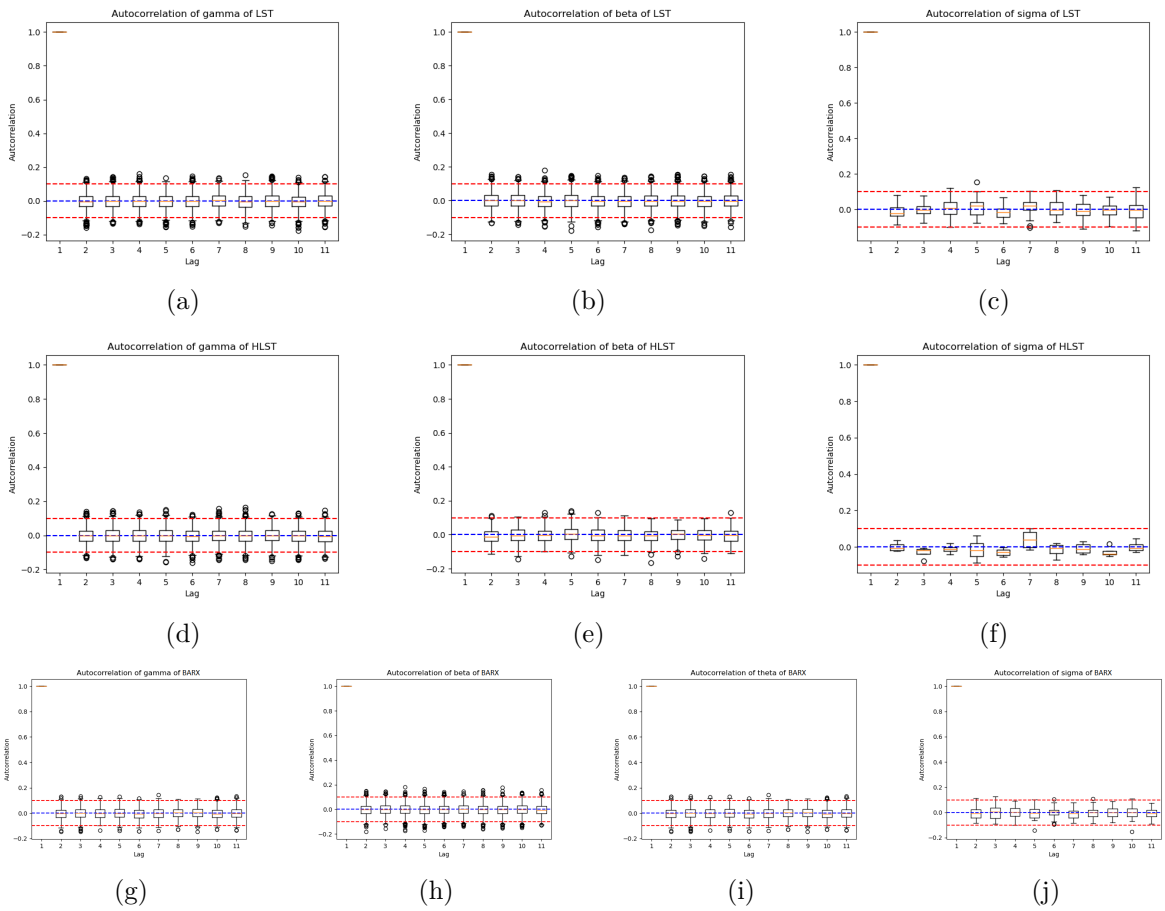

Figure 3: Boxplots of the autocorrelations of the different coefficients. The blue line notes the zero line and the red lines the value -0.1 and 0.1.

5.2 Important predictors

As the LST model is a linear model, the coefficients can also be estimated using regular OLS. This way we do not use the priors, but we can still compare the coefficients. We fit the models on all 10 stores and take the mean of the coefficients to get the average effect. For sales, this is shown in Figure 4. To get point estimates of the coefficients, we used the mean of the draws of the coefficients for the Bayesian models. The coefficients can be interpreted as an increase of $(exp(v) - 1)\%$, where v is the value of the coefficient. This means that values lower than zero have a decreasing effect and above zero have a positive effect on the sales. For example, the coefficient of Saturday for LST is 0.554 which means that the model estimates an increase of 74% relative to the local time trend. What immediately can be noted is that the coefficients of OLS are larger and further from zero indicating that the use of the priors does have a regularisation effect on the estimates. We assume the OLS to be overfitting and will continue with the analysis of these estimates for the Bayesian methods. It can be noted that the three Bayesian methods have very similar estimates suggesting that the model specification is similar. Going over the various seasonalities, we see that mainly the

month of the year and the day of the week play a big part as they have larger coefficients. For months of the year, there is a downward effect for January and a positive effect for May, November and December. For the weekdays, we see that the later in the week the sales are much higher, except on Sunday. We see a high peak on Saturday indicating that it has a large positive effect on the number of sales. The days of the month have smaller effects illustrating that these are less important, but show an increase over the month, telling us that the number of sales is higher later in the month, which is also concluded in Section 3.

Figure 4: Coefficient estimates for the seasonalities of OLS and the Bayesian methods for sales.

For the transactions, we see similar results in Figure 5 although it seems that the OLS estimates are less off from the Bayesian estimates. Further, it can be noted that the day of the month plays a bigger part suggesting that there are more transactions later in the month even relative to the sales. The other effects are akin to the findings we did with the sales estimates.

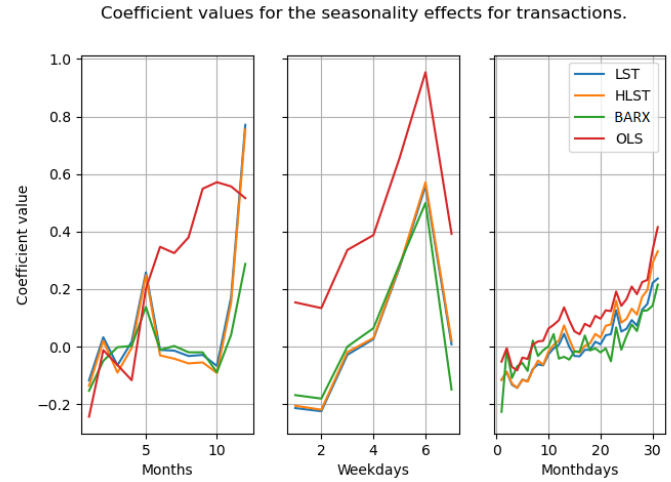

Figure 5: Coefficient estimates for the seasonalities of OLS and the Bayesian methods for transactions.

What strikes the eye in Figure 6 of footfall is that the OLS estimate of the month of the year is much larger than that of the Bayesian methods signalling that the smaller amount of observations of footfall resulted in even more overfitting. Further, for the Bayesian methods, the peak of May is a lot smaller relative to the estimates of sales and transactions. The month of the day plays a small part in the footfall and only has an increase in the last 7 days.

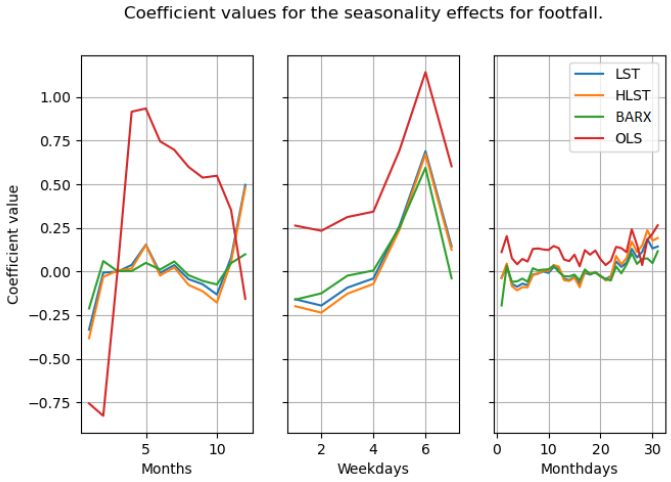

Figure 6: Coefficient estimates for the seasonalities of OLS and the Bayesian methods for footfall.

Next to the seasonalities, we also estimate the time trend. As it is hard to give a comprehensive graph of the time trend for the different stores and demand drivers, we take the sales of a single store as an example. In Figure7 we see the actual number of sales in blue. In orange, we display the time trend twice. It can be noted that the time trend is very stable relative to the actual demand. It is only if we change the scale that we see a changing slope that for the first 150 days is downwards but the slope changes to an upwards trend that increases over the remainder of the days. The time trend function does a great job capturing the overall trend and is not influenced by periodic peaks. The time trend can be seen as a time dynamic mean where the seasonalities cause change around this mean. This is in line with the research by Posch et al. (2020), which found that the effect of the time trend was relatively small to the seasonality effects.

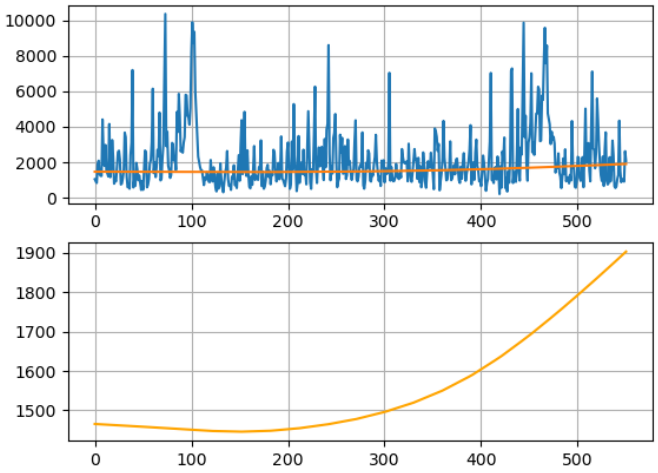

Figure 7: In orange the time trend for the sales of store 1. In blue are the actual sales.

5.3 Prediction metrics

The results of the experiment can be seen in Table 1. The value of the metrics is shown relative to the LST method, where a value larger than 1 shows a better result than the LST method and a value smaller than 1 has a worse result. These results are discussed in the following subsections per demand driver.

Table 1: The performance of the different methods relative to the LST method. A value larger than 1 shows a better score than LST. DM show the p-value of the Diebold-Mariano test where a value smaller than 0.05 shows that the LST outperforms the reference model. DFT show the p-value of the two-sided density forecast test which denotes the probability that the LST model outperforms the reference model regarding the predicted density. * signals a significant result.

		OLS	**XGBoost**	**Ran. For.**	**ARMA**	**Exp. Smth.**	**HLST**	**BARX**
Sales	*MAAPE*	0.580	0.807	0.825	0.667	0.754	1.031	0.874
	MSE	0.000	0.912	0.811	0.597	0.583	1.033	0.845
	DM	0.000	0.000	0.000	0.000	0.000	1.000	0.000
	ALDS						1.012	1.049
	DFT						0.085	0.979*
Trans.	*MAAPE*	0.617	0.890	0.855	0.632	0.763	1.027	0.951
	MSE	0.001	0.901	0.613	0.000	0.457	1.005	0.988
	DM	0.000	0.000	0.000	0.000	0.000	1.000	0.001
	ALDS						1.012	1.000
	DFT						0.00*	0.506
Footfall	*MAAPE*	0.562	1.040	1.077	0.773	0.926	1.037	1.294
	MSE	0.005	0.988	0.985	0.000	0.543	0.967	1.611
	DM	0.000	0.197	0.946	0.000	0.042	0.670	0.283
	ALDS						1.012	1.049
	DFT						0.438	0.415

5.4 Sales

First of all, the *MAAPE* and *MSE* of OLS are much worse than the LST method. For OLS, this can be explained by overfitting the spline trend function, which results in a strong trend upwards or downwards in the forecasting period. Looking at the main focus of accuracy, the MAAPE, we see that all Bayesian methods outperform the benchmark models. The best performing method is the Hierarchical method, which takes the sales data of all stores into account. Further, the Average Log Density Score suggests that the Hierarchical Bayes model and BARX model produce a forecast density that is more in line with the real values in the forecast interval. Other metrics have also been calculated and can be found in A1. Although the HLST does not have the lowest bias, the other metrics verify that the HLST does indeed have the best forecasting performance.

To test the significance of the metrics, we use a one-sided Diebold-Mariano (DM) test to see if the forecasts of HLST are indeed better (Diebold and Mariano, 1995). We use the forecasts of all four folds of all ten stores and compare all methods against each other. To interpret the results in Figure 8, the row of a method where a cell is green means that the method has better forecasts than the method on the horizontal axis. Therefore it can be seen that OLS does not outperform the other methods and HLST has significantly better forecasts than any other method. This confirms the previous observation of having the lowest error metric. Also, it can be seen that

all Bayesian methods surpass the conventional methods in forecasting. This is in line with the research of Posch et al. (2020), where their LST model outperformed the conventional methods on the forecasting of transactions. Pavlyshenko (2019) concluded that the forest-based methods such as random forest and XGBoost outperform the Bayesian methods. Therefore we can note that by using a more sophisticated prior than just the normal prior for the coefficients, we get better performing Bayesian methods.

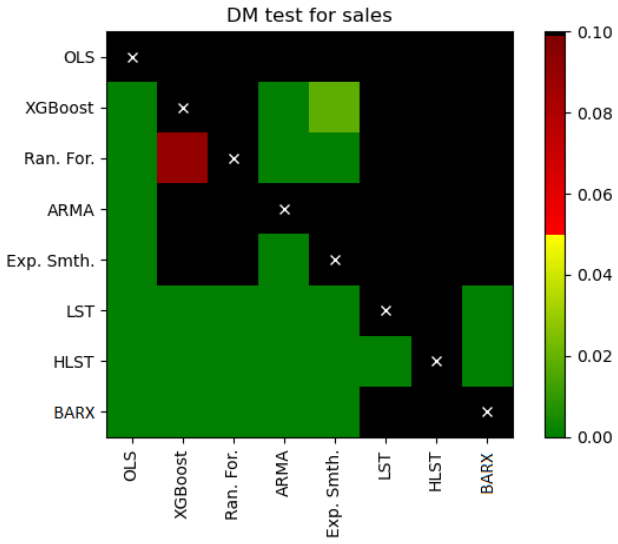

Figure 8: Diebold-Mariano test for the sales of the ten stores with four folds. The colour of the cell denotes the p-value of the null hypothesis of both methods having the same forecasting performance. A green cell indicates that the method on the vertical axis has significantly better forecasts than the method on the horizontal axis.

5.5 Transactions

With the transactions data, we assumed a conditional negative binomial distribution as it is considered as discrete count data. Again, we see that the OLS method performs the worst, possibly due to overfitting. With the ARMA model, having a peak in the lagged values would result in a very large forecast with, as a result, a relatively high error that can be seen in the metrics. The benchmark models are all outperformed by the Bayesian methods. The Hierarchical Bayes model has the best performance overall, but the other two Bayesian methods perform nearly as good. It can be noted that the BARX model has the lowest ALDS, meaning that the forecasts densities are best for this model. Results for the other metrics can be found in A2. There we see that the HLST has the best performance although the LST has a smaller bias.

Again, we use a one-sided Diebold-Mariano (DM) to test the forecasts (Diebold and Mariano, 1995). Results are shown in Figure 9. Again, we see that Bayesian methods significantly outperform conventional methods. We can also see that the HLST method achieved the best forecasts. This means that having information of transactions of multiple stores and assigning a hierarchical

relationship improves the overall forecasting power.

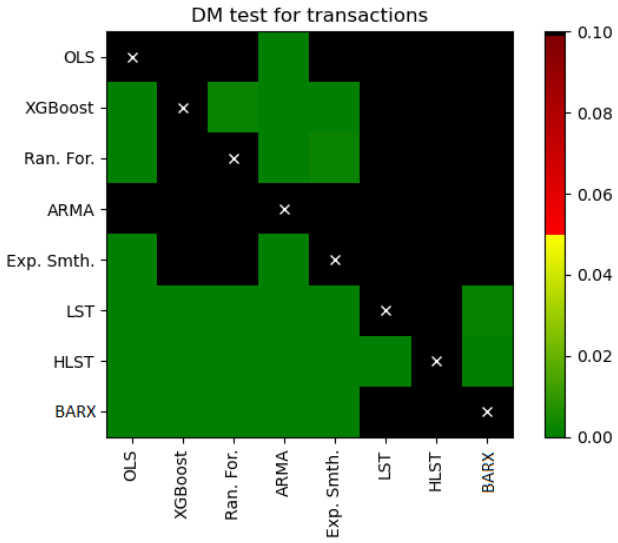

Figure 9: Diabold-Mariano test for the transactions of the ten stores with four folds. The color of the cell denotes the p-value of the null hypothesis of both methods having the same forecasting performance. A green cell indicates that the method on the vertical axis has significantly better forecasts than the method on the horizontal axis.

5.6 Footfall

The footfall data differs from the previous two data sets by being a lot smaller. It can be noted that OLS, ARMA and Exponential Smoothing perform worse than the LST method. Looking at the main focus of accuracy, the MAAPE, we see that the XGBoost and Random Forest perform better than the Bayesian LST model. This could be due to the fact that the number of observations in the data set was too small for this model. This is where the other Bayesian models shine by having multiple data sets included. The BARX model has data on the sales and transactions of a particular store. Thus it can be seen that this addition of information really benefits this method. Looking at the ALDS, we see that the BARX model also has the best density forecasts. A3 show the other metrics for the transactions. These show that the BARX has the best forecasting performance for every metric. The BARX model performs the best for transactions which can also be seen in Figure 10. It does not have the best performance for every fold, but is shows stable scores and does not have large peaks in the error metric.

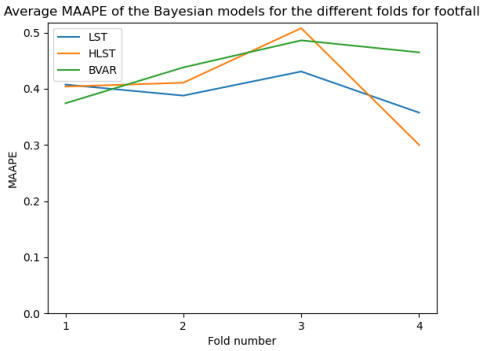

Figure 10: The average MAAPE over the different folds for footfall.

We validate the results with a Diebold-Mariano (DM) to test the forecasts (Diebold and Mariano, 1995). We find the results in Figure 11, which are more inconclusive than the previous two demand drivers. Rejecting p-values smaller than 5% results in not having a single best method for the forecasts. Both the Random Forest, LST and HLST methods have a similar performance in forecasting, which makes it impossible to declare the best method. However, we can clearly see that OLS and ARMA do have worse forecasts than the other methods.

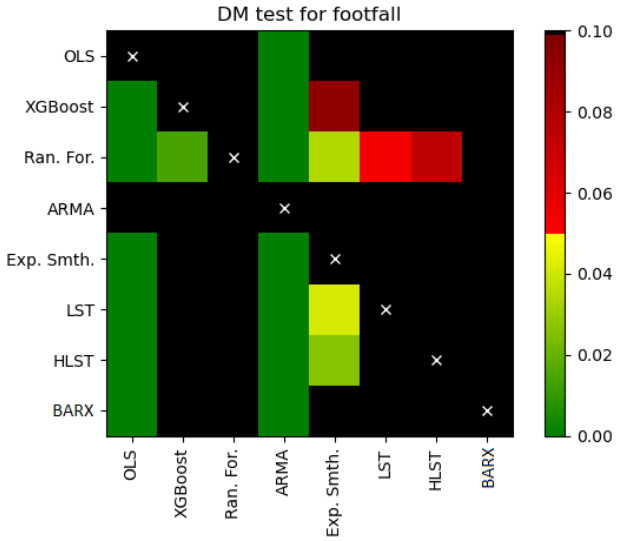

Figure 11: Diebold-Mariano test for the footfall of the ten stores with four folds. The colour of the cell denotes the p-value of the null hypothesis of both methods having the same forecasting performance. A green cell indicates that the method on the vertical axis has significantly better forecasts than the method on the horizontal axis.

5.7 Longer forecast window

To further investigate the performance of the models, we repeat the experiment with a longer forecasting window. Results are shown in Table 2. For the demand driver sales we can now see that the LST and HLST are the best performing models, but the BARX model is outperformed by the

XGBoost and Random Forest methods. For the transactions we see that now the Bayesian methods are not the best performing methods anymore. The XGBoost and Random Forest methods perform best. However with footfall we see that the BARX model has the best accuracy whereafter the XGBoost and Random Forest follow. In general we see that the Bayesian models perform relatively worse for a longer forecasting window. The absolute metrics can be found in Tables A4, A5, A6.

Table 2: The performance of the different methods relative to the LST method. A value larger than 1 shows a better score than LST with a two month forecasting window. DM show the p-value of the Diebold-Mariano test where a value smaller than 0.05 shows that the LST outperforms the reference model. DFT show the p-value of the two sided density forecast test which denotes the probability that the LST model outperforms the reference model regarding the predicted density. * signals a significant result.

		OLS	**XGBoost**	**Ran. For.**	**ARMA**	**Exp. Smth.**	**HLST**	**BARX**
Sales	*MAAPE*	0.615	0.998	0.996	0.677	0.847	1.030	0.877
	MSE	0.000	0.902	1.195	0.483	0.410	1.051	0.556
	DM	0.001*	0.928	0.734	0.000*	0.000*	1.000*	0.000*
	ALDS						1.038	0.796
	DFT						0.247	0.658
Trans.	*MAAPE*	0.629	1.044	1.082	0.749	0.81	1.038	0.869
	MSE	0.001	0.743	1.095	0.428	0.276	1.041	0.627
	DM	0.000*	0.999*	1.000*	0.000*	0.000*	1.000*	0.000*
	ALDS						1.069	0.9354
	DFT						0.129	0.665
Footfall	*MAAPE*	0.705	1.186	1.224	1.076	0.991	1.018	1.309
	MSE	0.387	0.749	0.873	0.892	0.534	1.012	1.426
	DM	0.000*	1.000*	1.000*	1.000*	0.350	0.562	1.000*
	ALDS						1.107	1.361
	DFT						0.142	0.384

5.8 Summarizing remarks

Taking all results into account we can see that the most important factor in modeling the demand is including a shrinkage factor to regularise the large amount of variables. This can be seen in the difference in performance between OLS and LST. Second to that, the using cross-sectional data between stores improves plays a large role in the performance as we can see that the HLST works the best overall. Using autoregressive terms from multiple demand drivers works better for shorter data sets. For longer forecasting windows we see the same results, except that the tree based methods perform on par as the Bayesian methods.

6 Conclusion

The analysis of this thesis covers the performance of Bayesian methods in demand forecasting relative to conventional methods. For this research, data of a jewellery franchise is used with the daily value of demand drivers which are sales, transactions and footfall. The results draw the conclusion that a Bayesian method developed by Park and Casella (2008) using a seasonality and trend function in conjoint with normal and Laplace priors performs better in forecasting than conventional methods such as Random Forests, XGBoost, ARIMA and exponential smoothing. Two variations of this Bayesian model were also tested, the first used a hierarchical setting between the coefficients of stores and the second used a autoregressive framework with the three demand drivers. For every demand driver, the Bayesian methods performed best. Of the three Bayesian methods, the hierarchical method performed the best on sales and transactions. The footfall is best forecast by the Bayesian autoregressive model. It can be concluded that data sets longer than a year are suited for the hierarchical model. Data with fewer observations than a year is better modelled by the autoregressive method. For longer forecasting windows, for example, two months, the performance of the Bayesian models worsen relatively to the other models.

This research is limited by the small number of folds to determine the performance. Having ten or more folds instead of 4 will increase the firmness of the results. In addition, the sampling of the Bayesian methods can be computationally heavy. Estimating the models takes shorter for the conventional methods, therefore a lower forecasting accuracy could weigh more in the choice of the best model.

We do three suggestions for further research. A topic that is not investigated in this thesis is the research of the coefficients of the models. Most machine learning forecasting models do not have interpretable coefficients, but our models do. These can be examined to determine the size of and particular seasonality effect or trend. This helps store managers by giving insight into how a certain prediction was built.

As demand forecasting is among other things used for staff planning, it is interesting to investigate the performance of intraday forecasting. With forecasts on, for example, hourly level, a more efficient schedule could be arranged that takes peaks in the day into account. This way, less staff could be scheduled in between peaks and more staff during peaks.

In this research, we used the normal and Laplace prior for the coefficients. In current literature, more priors have been developed that are more complex but could have better performance. Examples are the horseshoe prior and the spike-and-slab prior (Ishwaran and Rao, 2005). Further, the values of the hyperparameters used in this research could be tweaked based on a grid search and cross-validation.

References

Amisano, G., & Giacomini, R. (2007). Comparing density forecasts via weighted likelihood ratio tests. *Journal of Business & Economic Statistics*, *25*(2), 177–190.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of statistical software*, *76*(1).

Chandra, S., & Chaloupka, F. J. (2003). Seasonality in cigarette sales: Patterns and implications for tobacco control. *Tobacco Control*, *12*(1), 105–107. https://doi.org/10.1136/tc.12.1.105

Da Veiga, C. P., Da Veiga, C. R. P., Catapan, A., Tortato, U., & Da Silva, W. V. (2014). Demand forecasting in food retail: A comparison between the holt-winters and arima models. *WSEAS transactions on business and economics*, *11*(1), 608–614.

De Myttenaere, A., Golden, B., Le Grand, B., & Rossi, F. (2016). Mean absolute percentage error for regression models. *Neurocomputing*, *192*, 38–48.

Diebold, F., & Mariano, R. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, *13*(3), 253–63. https://EconPapers.repec.org/RePEc:bes:jnlbes:v:13:y:1995:i:3:p:253-63

Ehrenthal, J., Honhon, D., & Van Woensel, T. (2014). Demand seasonality in retail inventory management. *European Journal of Operational Research*, *238*(2), 527–539.

Gasthaus, J., Benidis, K., Wang, Y., Rangapuram, S. S., Salinas, D., Flunkert, V., & Januschowski, T. (2019). Probabilistic forecasting with spline quantile function rnns. *The 22nd international conference on artificial intelligence and statistics*, 1901–1910.

Gelman, A. (2020). *Prior choice recommendations*. Retrieved 14-05-2021, from https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations

Geurts, M. D., & Kelly, J. P. (1986). Forecasting retail sales using alternative models. *International Journal of Forecasting*, *2*(3), 261–272.

Huber, J., & Stuckenschmidt, H. (2020). Daily retail demand forecasting using machine learning with emphasis on calendric special days. *International Journal of Forecasting*, *36*(4), 1420–1438.

Ishwaran, H., & Rao, J. S. (2005). Spike and slab variable selection: Frequentist and bayesian strategies. *The Annals of Statistics*, *33*(2), 730–773.

İşlek, İ., & Öğüdücü, Ş. G. (2015). A retail demand forecasting model based on data mining techniques. *2015 IEEE 24th International Symposium on Industrial Electronics (ISIE)*, 55–60.

Ke, J., Zheng, H., Yang, H., & Chen, X. M. (2017). Short-term forecasting of passenger demand under on-demand ride services: A spatio-temporal deep learning approach. *Transportation Research Part C: Emerging Technologies*, *85*, 591–608.

Kim, S., & Kim, H. (2016). A new metric of absolute percentage error for intermittent demand forecasts. *International Journal of Forecasting*, *32*(3), 669–679.

Lemoine, N. P. (2019). Moving beyond noninformative priors: Why and how to choose weakly informative priors in bayesian analyses. *Oikos*, *128*(7), 912–928.

Mercy, C., & Kihoro, J. (2015). Application of vector autoregressive (var) process in modelling reshaped seasonal univariate time series. *Science Journal of Applied Mathematics and Statistics*, *3*(3), 124–135.

Miller, J. J., McCahon, C. S., & Miller, J. L. (1991). Foodservice forecasting using simple mathematical models. *Hospitality Research Journal*, *15*(1), 43–58. https://doi.org/10.1177/109634809101500105

Nau, R. (2014). Principles and risks of forecasting. *Fuqua School of Business, Duke University September*.

Park, T., & Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association*, *103*(482), 681–686.

Pavlyshenko, B. M. (2019). Machine-learning models for sales time series forecasting. *Data*, *4*(1). https://doi.org/10.3390/data4010015

Peterson, R. T. (1993). Forecasting practices in retail industry. *The Journal of Business Forecasting*, *12*(1), 11.

Pitkin, J., Manolopoulou, I., & Ross, G. (2018). Bayesian hierarchical modelling of sparse count processes in retail analytics. *arXiv preprint arXiv:1805.05657*.

Posch, K., Truden, C., Hungerländer, P., & Pilz, J. (2020). A bayesian approach for predicting food and beverage sales in staff canteens and restaurants. *arXiv preprint arXiv:2005.12647*.

Riddell, A., Hartikainen, A., & Carter, M. (2021). Pystan (3.0.0).

Takenaka, T., & Shimmura, T. (2011). Practical and interactive demand forecasting method for retail and restaurant services. *Proc. of International Conference Advances in Production Management Systems*, (3-4), 2.

Tanizaki, T., Hoshino, T., Shimmura, T., & Takenaka, T. (2019). Demand forecasting in restaurants using machine learning and statistical analysis. *Procedia CIRP*, *79*, 679–683.

Todd, R. M. (1990). Improving economic forecasting with bayesian vector autoregression. *Modelling economic series*, 214–34.

Ugarte, M. D., Goicoa, T., Militino, A. F., & Durbán, M. (2009). Spline smoothing in small area trend estimation and forecasting. *Computational statistics & data analysis*, *53*(10), 3616–3629.

Wong, K. K., Song, H., & Chon, K. S. (2006). Bayesian models for tourism demand forecasting. *Tourism Management*, *27*(5), 773–780.

Yelland, P. M., & Dong, X. (2014). Forecasting demand for fashion goods: A hierarchical bayesian approach. *Intelligent fashion forecasting systems: Models and applications* (pp. 71–94). Springer.

7 Appendix

The Appendix is organised as follows. First, Table A1 shows all the average metrics for the sales datasets of the ten stores. Second, the results of the transactions datasets are in A2. To finalize, Table A3 show the average metrics of the methods for the footfall data.

Table A1: The performance of the different methods for sales.

	OLS	**XGBoost**	**Ran. For.**	**ARMA**	**Exp. Smth.**	**LST**	**HLST**	**BARX**
MAAPE	0.67	0.48	0.47	0.58	0.52	0.39	0.38	0.45
MSE	19374882128323.40	33020473.70	37151172.79	50423336.14	51693179.02	30118336.94	29168759.84	35655185.27
WAPE	59.12	0.47	0.47	0.60	0.56	0.38	0.37	0.44
MAD	412093.66	3270.04	3293.12	4156.09	3909.28	2633.01	2600.75	3069.68
MFE	-407165.46	456.76	602.34	-191.84	1507.15	839.30	667.62	1195.73

Table A2: The performance of the different methods for transactions.

	OLS	**XGBoost**	**Ran. For.**	**ARMA**	**Exp. Smth.**	**LST**	**HLST**	**BARX**
MAAPE	0.55	0.38	0.40	0.54	0.45	0.34	0.33	0.36
MSE	5060712.03	3433.25	5046.05	39977452778433200000000000000000000000.00	6759.72	3092.01	3078.10	3130.62
WAPE	3.05	0.35	0.42	1119131226973900.00	0.49	0.31	0.31	0.32
MAD	272.36	31.60	37.48	9997181200542000.00	43.75	28.06	27.71	28.74
MFE	-226.84	4.37	6.58	-99971811999908400.00	20.33	4.42	4.35	0.94

Table A3: The performance of the different methods for footfall.

	OLS	**XGBoost**	**Ran. For.**	**ARMA**	**Exp. Smth.**	**LST**	**HLST**	**BARX**
MAAPE	0.70	0.38	0.36	0.51	0.42	0.39	0.38	0.30
MSE	9159505.99	41778.96	41932.03	18720761613203300000000000000000000.00	75988.31	41288.60	42706.13	25621.57
WAPE	3.35	0.40	0.39	76000128493986.60	0.51	0.40	0.40	0.30
MAD	952.61	114.74	111.82	21619462358638600.00	146.10	115.01	114.13	85.52
MFE	-765.01	30.43	48.45	-21619462357463700.00	75.78	29.86	32.87	26.52

Table A4: The performance of the different methods for sales with a forecasting window of two months.

	OLS	**XGBoost**	**Ran. For.**	**ARMA**	**Exp. Smth.**	**LST**	**HLST**	**BARX**
MAAPE	0.62	0.38	0.38	0.56	0.45	0.38	0.37	0.43
MSE	207224316693171.00	34021356.88	25676330.90	63553534.14	74827855.42	30670388.38	29188876.65	55203112.44
WAPE	200.26	0.41	0.38	0.62	0.59	0.38	0.37	0.49
MAD	1406263.29	2896.32	2651.23	4331.16	4176.77	2652.00	2596.11	3448.55
MFE	-1397062.13	2208.34	1484.60	649.00	3611.34	1018.59	816.88	2148.43

Table A5: The performance of the different methods for transactions with a forecasting window of two months.

	OLS	**XGBoost**	**Ran. For.**	**ARMA**	**Exp. Smth.**	**LST**	**HLST**	**BARX**
MAAPE	0.55	0.33	0.32	0.46	0.42	0.34	0.33	0.40
MSE	3182658.35	4246.64	2884.62	7370.52	11451.57	3157.24	3033.20	5032.39
WAPE	2.85	0.36	0.31	0.49	0.58	0.32	0.31	0.40
MAD	256.76	32.03	28.00	44.60	52.17	28.72	27.81	36.42
MFE	-146.09	21.65	10.83	16.61	45.43	5.46	5.19	10.90

Table A6: The performance of the different methods for footfall with a forecasting window of two months.

	OLS	**XGBoost**	**Ran. For.**	**ARMA**	**Exp. Smth.**	**LST**	**HLST**	**BARX**
MAAPE	0.59	0.35	0.34	0.38	0.42	0.41	0.41	0.32
MSE	145274.39	75024.75	64360.62	62983.44	105127.84	56176.40	55501.71	39404.91
WAPE	0.76	0.48	0.45	0.46	0.59	0.46	0.45	0.36
MAD	218.88	138.39	129.98	131.26	168.49	131.01	130.50	102.65
MFE	161.96	123.52	111.86	92.70	126.07	40.91	21.70	56.71