UNIVERSITY OF TWENTE.

Data-Driven Retail Food Waste Reduction

A comparison of demand forecasting techniques and dynamic pricing strategies

Paula Felix

Master Thesis
August 2018

Study Programmes
MSc Computer Science (CSC)
MSc Business Information Technology (BIT)

Graduation Committee CSC
Dr. N. Sikkel (chairman)
Dr. M. van Keulen

Graduation Committee BIT
Dr. N. Sikkel (chairman)
Dr. A.B.J.M. Wijnhoven

Preface

First of all, I would like to thank the members of my graduation committee, Klaas Sikkel, Fons Wijnhoven and Maurice van Keulen, for guiding me in this 9-month thesis writing process and for all your valuable feedback. I really enjoyed our meetings and always left motivated to continue and improve my thesis. In addition, I would like to thank my Deloitte colleagues for giving me the opportunity to write my thesis there, for their useful suggestions and of course for making my time at the office so enjoyable. Handing in this thesis officially marks the end of my time as a student at the University of Twente and these five years as a student have really flown by. Although I won't be a student anymore, I hope to continue to learn many new things and be challenged on a daily basis in my future career.

Executive Summary

Every year one third of all food that is produced is wasted and one of the UN sustainable development goals is to cut food waste in half by 2030. This thesis focuses on two data-driven strategies to reduce perishable food waste at retailers: demand forecasting and dynamic pricing. Improved demand forecasting techniques can prevent excess inventory by better supporting replenishment decisions, whereas dynamic pricing can reduce excess inventory once it exists by stimulating customers to buy older products at a discount. The performance of both traditional and promising new demand forecasting techniques is compared and implementation guidelines are provided for retailers. In addition, simulations were conducted to investigate the performance of different (dynamic) pricing strategies in terms of revenue, waste and stock-outs. For retailers, reducing perishable food waste results in financial and sustainability benefits.

Demand Forecasting

Demand forecasting techniques that were included in the performance evaluation are: naive (where the forecast equals sales from last period), exponential smoothing (ES), moving average (MA), linear regression (LINREG), auto-regressive integrated moving average (ARIMA), support vector regression (SVR), multi-layer perceptron (MLP), long-short term memory network (LSTM) and adaptive boost (ADA). These techniques were evaluated based on their performance in forecasting sales for 986 perishable food products from an Ecuadorian supermarket. Performance was measured using the relative root mean squared error measure (RelRMSE), which is a robust measure that indicates how well a certain technique performs relative to the naive forecast. Performance was compared across different forecasting scenarios, which differed in terms of their time detail level, location detail level and horizon. In addition, it was investigated what the influence is of using external factors such as the weather in addition to historical sales data to produce forecasts.

Results show that there is no such thing as the ultimate demand forecasting technique and that performance greatly varies across products and forecasting scenarios. By far the best demand forecasting performance overall can be obtained by automatically selecting the best forecasting technique for each individual product, resulting in (depending on the forecasting scenario) a 6% to 32% RelRMSE improvement compared to always using the naive forecast and a 2% to 10% improvement compared to always using the best individual demand forecasting technique. Adding external factors from the weather, promotion, economic and holiday categories has shown to be valuable for forecasting scenarios that have a daily time detail level, enabling an additional 3% to 9% RelRMSE improvement compared to using histor-

ical sales data only. For each forecasting scenario, detailed results and an overview of the top 3 best performing individual techniques are provided in this thesis to help retailers select the appropriate demand forecasting technique(s) in their situation.

In addition, a process is provided with step-by-step guidelines for retailers on how to improve the wider demand forecasting process. It not only considers how to select the right demand forecasting technique(s), which form the quantitative core of the demand forecasting process, but also discusses how to assess forecasting capabilities and which qualitative factors should be taken into account, such as implementation in decision support systems, adoption factors and organizational factors.

Dynamic Pricing

A simulation study was conducted to determine which (dynamic) pricing strategy for perishable products performs best in terms of total revenue, waste and stock-outs. The simulation considers a monopolist grocer selling a single perishable product with a fixed shelf life. The product can be replenished each day based on the demand forecast and a safety factor. Customers have different characteristics: some customers are regular customers that only pay attention to price, whereas others are date-checking customers that also pay attention to remaining shelf life and aim to choose a product that maximizes their value-for-money. Four main pricing strategies for marking down perishable products that approach their expiry date were compared. Strategy 1 applies no price changes at all and serves as a baseline. Strategy 2 applies a fixed discount D at the last day before expiry, while strategy 3 spreads that discount over the last S days before expiry. Strategy 4 dynamically determines discount percentages based on the demand forecast and the remaining inventory.

Almost all strategies that applied a discount resulted in a waste reduction, but they regularly resulted in significantly lower revenues. Multiple experiments were conducted to investigate the effects of varying assumptions for simulation settings. Discounting was most beneficial when product demand was more elastic and when more customers checked expiry dates, because that resulted in higher waste reductions and less negative (or even positive) changes to revenue at the same time. The fixed pricing strategy that most frequently performed best was strategy 2 with a fixed discount of 20% on the last day before the expiration date. Surprisingly, the dynamic pricing strategies did not always perform better than a fixed price strategy, which could be due to the fact that these strategies relied upon imperfect demand forecasts and hence their estimated optimal discount percentage might have been off base. A dynamic pricing strategy already outperformed the best fixed strategy when initial waste levels for a product were high or when a large percentage of customers were regular customers.

Conclusion and Discussion

Both demand forecasting process improvement and dynamic pricing strategies for marking down products that near their expiry dates have a positive impact on waste reduction. Grocers are advised to initially focus on improving the demand forecasting process, since that reduces waste, but does not have the negative impact on revenue that discounting strategies frequently have. Preventing inventory excesses from occurring through improved demand forecasting is better than trying to resolve

such excesses by discounting products that approach their expiration dates. Grocers are advised to follow the demand forecasting improvement process, to use this study as a benchmark for forecasting technique selection, to implement multiple forecasting techniques and to automatically select the best forecasting technique for each product. To resolve any excesses that do occur in stores, grocers are advised to use the pricing strategy that showed best performance in similar situations in the simulations. The best fixed strategy in general is applying a 20% fixed discount on the last day before expiry. A dynamic pricing strategy only outperformed a fixed strategy in a few situations.

This study makes several contributions to both retail practice and the scientific fields of demand forecasting and dynamic pricing. First of all, this study provides a robust and objective comparison of forecasting techniques in different scenarios within a food retail context. To the best of our knowledge, it is the first study that conducts such a comparison. Results show the impact of automatically selecting the best forecasting techniques for each individual product and for including external factors and can be used by retailers as a benchmark for forecasting technique selection. In addition, this study proposes a process for demand forecasting improvement and forecasting technique selection, which can guide retailers in their demand forecasting improvement efforts. This study also provides a robust performance comparison of different pricing strategies that mark down perishable products, giving retailers insight into the impact on revenue, waste and stock-outs. In addition, tools were developed in the form of an algorithm for automatic best DF technique selection (and configuration) and a pricing simulation that can be reused in future work.

Contents

1
Introduction

Each year around one third of all food that is produced is wasted, which equals 1.3 billion tonnes and is worth around $1 trillion [25]. One of the UN sustainable development goals is to cut food waste in half by 2030 [63]. Food retailers such as grocery chains are in a unique position to contribute to this goal. They handle large quantities of food in their distribution centres and stores and are closely connected to both suppliers and consumers.

The food waste hierarchy [48] provides a framework for food waste reduction and shows 5 distinct stages: prevent, reuse, recycle, recover and dispose. Solutions that fall in the prevention stage are most favourable [48]. Two promising directions for reducing food waste in grocery chains using data analytics are enhanced demand forecasting (DF) and dynamic pricing (DP). Both directions fall in the prevention stage of the food waste hierarchy. Better demand forecasting can prevent surplus inventory, whereas dynamic pricing can reduce a surplus once it exists.

1.1 Introduction to Demand Forecasting

The goal of demand forecasting is to predict demand for individual products as accurately as possible to support business decisions such as replenishment. Inaccurate forecasts have been identified as one of the main causes for food waste at the retailer level [43]. So by improving the DF process, replenishment decisions can be better supported and surplus inventory can be prevented.

At the core of the DF process lie the forecasting techniques that are used, which can be qualitative, quantitative or a combination. The quantitative DF techniques are the main focus of this study. Qualitative techniques include expert opinion and the Delphi method. Traditional quantitative techniques include exponential smoothing (ES) and auto-regressive integrated moving average (ARIMA). Advances in predictive analytics have given rise to many new DF techniques over the last years. Newer methods for quantitative demand forecasting include variations of neural networks and ensemble methods. Various external factors can be taken into account as well, ranging from the weather and search trends to more general economic factors. Before selecting a forecasting technique, it is important to get insight into the characteristics of the forecasting problem at hand. Forecasts can differ on multiple dimensions, including the forecast horizon and the level of detail in terms of time, product and location.

It is important to realize that DF is about more than the techniques used, since the goal of DF is to support business decision-making. DF techniques are often implemented as part of a decision support system (DSS). Such a forecasting support system (FSS) can automatically create forecasts for products and aid employees in adjusting these forecasts if necessary to help them decide on replenishment quantities. For DF implementation it is also important to consider organizational factors, such as processes and management practices surrounding demand forecasting, since those can also influence adoption and performance.

1.2 Introduction to Dynamic Pricing

The field of dynamic pricing focuses on finding the optimal product prices over time to maximize profit for the seller. The adoption of dynamic pricing has increased considerably over the last years, mainly due to the increased availability of internal and external demand data, the rise of technologies that make it easier to change prices and the development of decision support systems [20]. A traditional application area of dynamic pricing is the airline industry [13], where there is a fixed capacity (limited number of seats in an aircraft) and ticket prices are dynamically adjusted based on the marginal value of remaining seats.

Even though dynamic pricing traditionally focused on the airline and hospitality industries, it also holds great potential for food retailers since they face fluctuating demand over time and their product offerings include perishable products like fresh foods that lose their value after a certain expiry date. Dynamic pricing strategies can help food retailers to minimize waste when a surplus occurs. Dynamic pricing and demand forecasting are interrelated. Dynamic pricing models use demand forecasts, in addition to other data such as current inventory levels, to determine what price strategy is optimal for an individual product to minimize waste and maximize profit. In turn, the adjusted prices again influence customer demand.

1.3 Benefits for Food Retailers

Retailers operate in a challenging industry with heavy competition, relatively low margins and fluctuating consumer demand. The large variety of products on offer in their stores includes perishables, which decline in quality over time and expire after a certain date, resulting in shrink (lost inventory value). Fresh food items are very important for grocers since they account for up to 40% of their revenue [39] and can be responsible for 80% of the total store shrink [47]. Hence reducing perishable product waste will improve retailer profitability.

One success story comes from Tesco, a large food retailer in the United Kingdom. Their improved DF system saved £100 million a year and a DP system for marking down perishable products further reduced waste by 2% a year and saved an additional £30 million each year [29].

Waste reduction also improves retailer sustainability, which not only has a positive influence on brand image, but can also result in additional financial benefits. Waste reduction contributes to a higher company score on sustainability assessments resulting in a higher chance to be included in sustainability indices, such as the Dow Jones' Sustainability Index, which has a positive impact on market capitalization.

Another benefit of enhanced DF for retailers is improved on-shelf availability, since forecasting inaccuracy is one of the root causes of stock-outs [18]. If a product is out-of-stock, this may result in opportunity cost due to missed sales opportunities. In addition, stock-outs have a negative effect on customer satisfaction and especially when a customer faces stock-outs multiple times, there is a high chance that that customer will go to a different store [18]. So retailers do not just have to prevent surplus inventory, but they also have to prevent stock-outs to keep customers satisfied.

1.4 Problem Statement

With one third of food going to waste each year, it is clearly an issue that has a big impact on our society. Cutting food waste in half by 2030 is one of the UN sustainable development goals. To contribute to achieving this goal, food retailers have to reduce the waste from their distribution centres and stores. For most, it is already part of their sustainability agenda for the coming years.

With the increasing digitization of the industry, opportunities arise for retail companies to implement advanced demand forecasting techniques and dynamic pricing strategies to reduce food waste. Many new forecasting techniques are developed in literature but their adoption in practice lags behind. It is difficult for companies to decide which DF technique to implement because a wide variety of techniques exist and new variations are published frequently. To the best of our knowledge, there is currently no thorough comparison of the performance of demand forecasting techniques for food retailers and practical guidelines for implementation are lacking. The potential of dynamic pricing to reduce waste also remains unclear so far and hence retailers would benefit from a simulation comparing the benefits of different pricing strategies.

1.5 Research Questions

Following from the introduction and the problem statement, the following main research question has been formulated for the thesis:

"How can enhanced demand forecasting and dynamic pricing contribute to reducing food waste at the retailer level?"

Separate sub questions have been formulated for the demand forecasting (DF) and dynamic pricing (DP) parts of the research.

DF1. Which quantitative demand forecasting technique performs best in forecasting perishable product demand?

DF2. What guidelines can be provided to improve the wider forecasting process in practice?

The comparison of demand forecasting techniques for question DF1 considers both commonly used and promising new DF techniques. The selected DF techniques are: naive forecasts, moving average (MA), exponential smoothing (ES), linear regression (LINREG), autoregressive integrated moving average (ARIMA), support vector regression (SVR), neural network variations (MLP and LSTM) and an ensemble technique, ADABoost (ADA). The forecasting problem dimensions will be

varied as well, specifically one- versus multi-step performance, weekly versus daily forecast performance and store- versus country-level performance will be evaluated. In addition, the added value of including external factor data on top of historical sales data will be examined. Since overall forecasting performance depends on more than just the technique used, question DF2 provides guidelines on how to improve the wider forecasting process aspects including decision support systems (DSS) and organizational factors.

DP1. What (dynamic) pricing strategies can be used for the sale of perishable food products in supermarkets?

DP2. What simulation model can be used to simulate perishable food product sales in supermarkets?

DP3. In simulations, which pricing strategy performs best in terms of total revenue, waste and stock-outs?

For question DP1, dynamic pricing strategies will be developed that optimize perishable product prices over time. Fixed pricing strategies will be discussed as well. For question DP2, a simulation model will be developed and assumptions will have to be made about consumer behaviour. The impact of different pricing strategies on revenue, waste and stock-outs will be evaluated using simulation experiments to answer question DP3.

1.6 Thesis Structure

The remainder of the thesis roughly follows the research questions in its structure and is divided into three parts. Part 1 covers demand forecasting, part 2 covers dynamic pricing and part 3 discusses results and draws conclusions. Figure 1.1 gives an overview of the structure of the thesis. Within part 1 and part 2, relevant background literature is covered to provide all readers with a solid background on both topics. Readers in a hurry who already have sufficient background knowledge can decide to skip these chapters without it limiting their ability to understand the rest of the thesis.

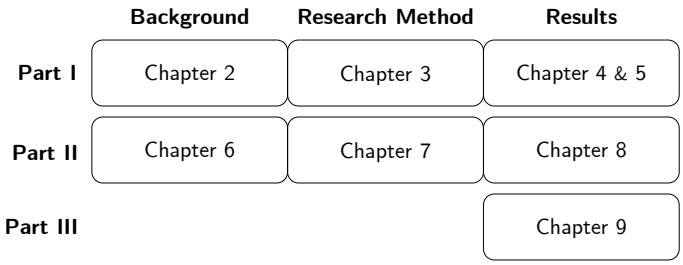

Figure 1.1: Overview of the thesis structure

Part I

Enhancing the Demand Forecasting Process

2
Demand Forecasting Background

The purpose of this chapter is to provide readers with a background on demand forecasting and to introduce the wide variety of forecasting techniques that are described in literature and used in practice. The goal of forecasting is to make predictions about the future value of variables and to do that as accurately as possible to support business decision-making. Section 2.1 describes the different dimensions of forecasting problems.

Figure 2.1 gives an overview of the main classes of forecasting techniques. On the highest level, forecasting techniques can be split into qualitative (section 2.2) and quantitative techniques (section 2.3). Qualitative techniques include expert opinion, the Delphi method, prediction markets, surveys and scenario development. Quantitative techniques can be further split into simple smoothing techniques (section 2.3.2), regression techniques (section 2.3.3), variations of ARIMA models (section 2.3.4), neural networks (section 2.3.5) and ensemble techniques (section 2.3.6). Section 2.3.7 discusses the variety of external factors that could be used as extensions to some of these DF techniques to further improve performance. Section 2.4 discusses forecasting performance measures and evaluation procedures. Section 2.5 provides context of demand forecasting in the grocery sector by reviewing available literature and discussing current practices at one large Dutch grocer.

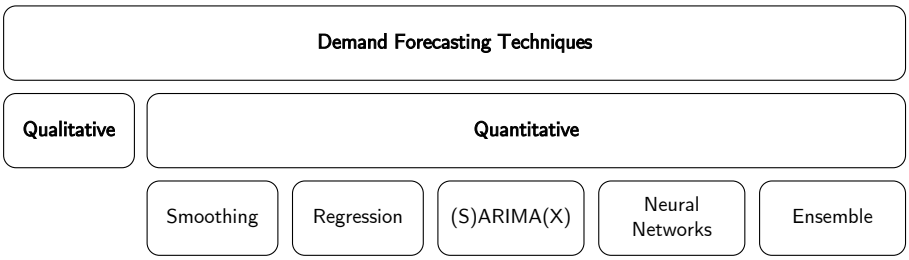

Figure 2.1: Types of forecasting techniques

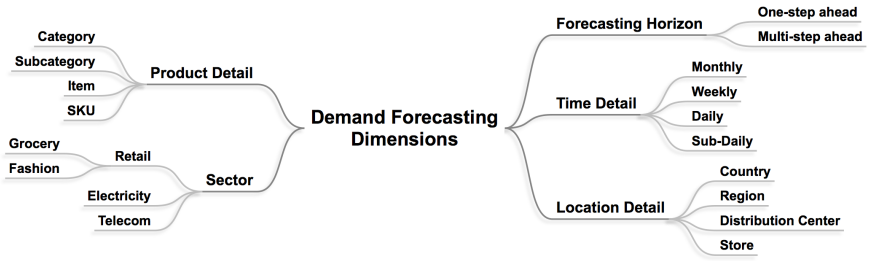

Figure 2.2: Overview of demand forecasting dimensions

2.1 Forecasting Problem Dimensions

Before selecting a forecasting technique, it is important to get insight into the characteristics of the forecasting problem at hand. Forecasts can differ on multiple dimensions, of which figure 2.2 gives an overview.

The forecast horizon dimension and aggregation on the dimensions of time detail, location detail and product detail account for the largest differences between forecasts. Depending on the sector of the company for which the forecast is created and the decision that is supported, the forecasting problem has its own unique characteristics.

One dimension is the forecast horizon. With single-step forecasting a prediction is made for a single future period, while multi-step forecasting predicts values for two or more future periods. Aggregation on three different dimensions also results in different forecasting problems. Aggregation can occur on the time detail, location detail and a product detail dimensions. The time detail of a forecast can for example be monthly, weekly, daily or even sub-daily (e.g. hours or minutes). The appropriate time detail level depends on the forecasting needs. For example in energy sector, a sub-daily demand forecast is crucial, whereas for car sales forecasting a monthly forecast might be enough. On the product detail dimension, forecasts can be created on a category level, subcategory level, individual item level (style) or even at the stock-keeping unit level (e.g. different sizes of each style). On the location detail dimension forecasts can be created country-wide, per region, per distribution centre or per individual store. The lower the aggregation level (so the more detailed the forecast) and the longer the forecasting horizon (more steps ahead), the more complex the forecasting problem is and the harder it is to make accurate forecasts.

2.2 Qualitative Forecasting Techniques

Qualitative forecasting techniques are subjective ways to produce forecasts. Qualitative techniques can be helpful in situations where there is limited availability of historical data, for example with the introduction of new products. With the most basic qualitative forecasting technique, a domain expert decides on a forecast based on his experience, intuition and any information about external factors that he/she knows might influence demand.

One of the more structured judgemental techniques is the Delphi method, which

provides a process for multiple experts to state and combine their forecasts. In multiple rounds, experts create individual forecasts with argumentation and a facilitator anonymously reads all of them. This way, forecasters can adjust their forecast based on other experts' arguments and the forecasts are likely to converge. After the last round the forecasts are combined and through this process the final forecast will likely be less biased than when one forecaster would have made a single forecast.

Prediction markets also aim to combine the knowledge of multiple participants. In a prediction market, forecasters trade contracts with a pay-off that depends on the outcome of uncertain future events [60]. Market prices represent the aggregated expectations of the forecasters about for example demand for a new product.

Another technique that is frequently used is market research, where surveys are sent out to consumers to gauge their demand for certain products as well as to get an idea of what factors influence their demand. Teschner and Weinhardt [60] provide a comparison of prediction markets and surveys, which shows that surveys are less complex, but that prediction markets provide continuous forecasts and enable forecasters to react to events. It is not clear yet whether prediction markets outperform other qualitative techniques, as research so far has produced mixed results [60].

Foresight literature describes several other qualitative techniques for long-term forecasting to support strategic decision making. Foresight is "the systematic process of looking into long-term future of science, technology and innovation" [33]. These techniques could support retailers' strategic decisions such as assortment planning. With scenario development, multiple scenario stories are developed that each depict a different possible future. So these scenarios take into account disruptions that might occur in the market. Roadmapping is another technique for strategic future planning where a visual chart is created that depicts developments from a technological, product and market perspective over time [49]. This can help retailers to assess the impact of market developments and new technologies.

A disadvantage of qualitative techniques is that because the techniques are subjective, the forecasts are susceptible to a number of human biases. It was shown that a forecaster's personality and motivational orientation have a significant effect on forecasting biases [21]. For example, conscientiousness increases overreaction bias and anchoring bias, whereas openness to experience increases optimism bias [21]. With anchoring bias, forecasters rely too heavily on the information that was offered first (the "anchor"). With overreaction bias, forecasters react disproportionally to new information. Optimism bias refers to forecasters consistently overestimating. Mellers et al. [42] also found that forecasting performance can differ between individuals and that the top performing forecasters, so called "superforecasters", maintained high performance over time and for a wide variety of topics (within geopolitical event occurrence probability forecasting). Superforecasters' high performance can be explained by their cognitive abilities and styles (high fluid & crystallized intelligence), their task-specific skills (high scope insensitivity & forecasting granularity), their high motivation & commitment and their enriched environments (they could discuss with other superforecasters) [42].

Most frequently, qualitative and quantitative techniques are used in combination. Forecasting is usually a two-step process [23], where first a quantitative forecast is derived from a model and then adjusted judgementally by an expert. Qualitative and quantitative approaches complement each other and should be combined, with the qualitative aspects' importance increasing as we look further ahead into the future [33].

2.3 Quantitative Forecasting Techniques

This section gives an overview of the main quantitative forecasting techniques. Before diving into individual techniques, section 2.3.1 provides an introduction to time series since that is the main data source for all quantitative techniques. Then section 2.3.2 introduces simple smoothing techniques, section 2.3.3 regression techniques, section 2.3.4 variations of ARIMA, section 2.3.5 neural networks and finally section 2.3.6 introduces ensemble techniques.

2.3.1 Time Series

Time series are "a set of observations y, each being recorded at a specific time t" [8]. The observations are sequential and have equal time intervals. Time series can be either univariate or multivariate, depending on the number of variables observed at each time t. One example of a univariate time series is 'car sales revenue per month', since it provides sequential values of a single variable (revenue from car sales) with a monthly time interval.

A time series can be decomposed into a trend component, seasonal component and residuals. The trend represents a long term growth/decline, whereas the seasonal component describes a recurring pattern over time. The residuals are what remains when the original data is detrended and deseasonalized.

In statistics each observation y_t is viewed as a realization of a certain random variable Y, so one important aspect of time series analysis is selecting a suitable probability model for the data. A general approach for statistical time series modelling consists of these steps [8]: plot the time series and examine the main features of the graph, remove the trend and seasonal components to get stationary residuals, choose a model to fit the residuals, forecast the residuals and invert the transformations done in earlier steps to arrive back at forecast of the original series.

Demand exists when a customer appears who wants to buy a certain product or service. In forecasting, sales are frequently used as a proxy for demand. While this is reasonable in most situations, it is important to realize that when stockouts occur too frequently or for longer periods of time, it is no longer an accurate approximation since there would be customers with demand who can not make a purchase. Throughout the rest of the thesis, the terms sales and demand are used interchangeably unless indicated otherwise. Sales forecasting can be considered a time series forecasting problem, since historical data is available on sales per time period.

2.3.2 Smoothing Techniques

The naive forecasting technique is the most basic way to produce forecasts. This approach takes the observed value from the previous period and returns that as the forecast for the next period(s). This is often used as a baseline to compare the performance of more advanced forecasting techniques.

The moving average (MA) technique is another relatively easy way to do forecasting based on smoothing. MA takes the average of the q previous periods and returns that as the forecast for the next periods. By averaging the q previous periods, this technique smoothes the historical returns by reducing the impact of irregularities. With MA, all previous periods have equal weights.

Table 2.1: Overview of smoothing techniques

Technique	Calculation
NAIVE	$F_t = Y_{t-1}$
MA	$F_t = \sum_{i=1}^{q} Y_{t-i}/q$
ES	$F_t = \alpha Y_{t-1} + (1-\alpha)F_{t-1}$

Another smoothing technique is exponential smoothing (ES), which also takes the average of previous periods but places a higher weight on the observations from the most recent periods. The weights for older periods are decreased gradually using a smoothing parameter α, which lies between 0 and 1. The closer α is to 1, the more emphasis is placed on the most recent observations. Table 2.1 shows how to use these techniques to produce a one-step ahead forecast.

2.3.3 Regression

Regression analysis is a technique that aims to fit a line through observations of a variable Y using explanatory variables X_i. In general a linear regression model looks like this:

$$Y = \alpha + \beta_1 X_1 + ... + \beta_n X_n + \epsilon$$

where α is the intercept, β are the weights for each of the explanatory variables and ϵ is the residual term. Regression analysis produces estimates of α and the regression coefficients β. With linear regression a straight line is fitted through the observations. Polynomial regression is more flexible, since it can fit a curved line through the observations, in this case the explanatory variables in the formula notation above would have increasing powers.

Multiple regression analysis methods exists, including ordinary least squares, ridge and lasso regression. Ordinary least squares selects the regression coefficients in such a way that it minimizes the sum of squares of differences between the true values and the predicted values. Additional variants were developed to increase the predictive accuracy of regression models. With ridge regression, a maximum is imposed on the sum of the absolute regression coefficient values and coefficient values are lowered if necessary. Similar to ridge regression, lasso regression also imposes a maximum value on the sum of the absolute values of the regression coefficients, but when that value is exceeded coefficients are not only lowered, but can also be set to zero. To summarize the difference: ridge regression can reduce the impact that irrelevant explanatory variables have on results, whereas lasso regression can cut out those features completely.

Support vector regression (SVR) may sound similar to normal regression, but is based on the machine learning concept of support vector machines (SVM). SVR is too complex to properly explain in a few sentences, so for a detailed explanation of SVR the reader is referred to the original paper by Drucker et al. [17] or the review paper by Basak et al. [6]. In short, when SVM is used for classification, it fits a hyperplane through the observations such that the observations from the different classes are separated with the highest margin (optimizing the chance of correctly classifying unseen data). To achieve this, SVM can apply (non-linear) transformations to map observations to a higher dimensional space using kernel functions. With a regression task, a margin of tolerance is set (ϵ).

2.3.4 ARIMA and Variations

One of the traditional time series forecasting techniques is autoregressive integrated moving average (ARIMA) [8]. Forecasts are based on an ARIMA(p,d,q) model. Parameter p is used for the model's autoregressive (AR) component, which is a linear regression using the p previous values of the time series itself. Parameter d represents the level of differencing that is applied to the original time series and generally gets the value 0, 1 or 2. Let Y be the regular time series and Y' the differenced series. With $d = 0$ no differencing is applied at all ($Y'_t = Y_t$), with $d = 1$ the first difference of the series ($Y'_t = Y_t - Y_{t-1}$) is used and with $d = 2$ the second difference ($Y'_t = (Y_t - Y_{t-1}) - (Y_{t-1} - Y_{t-2})$). Parameter q is the number of lags for the moving average (MA). The MA component is a linear regression using the previous q random error terms. The ARIMA model can then be represented by the following equation:

$$Y'_t = c + \epsilon_t + \sum_{i=1}^{p} \gamma_i Y'_{t-i} + \sum_{i=1}^{q} \theta_i \epsilon_{t-i}$$

The optimal values for p, d and q can be estimated by examining the auto-correlation function (ACF) & partial auto-correlation function (PACF) plots. The ACF plot shows how correlated the data is with itself at different time lags [8]. At lags greater than q, the ACF plot should not be significantly different from 0. The estimation of parameter p is analogous, but then using the PACF plot instead of the ACF plot. The PACF plot is similar to the ACF plot, but also adjusts for the correlation values in between the two times of the lag [8].

It is important to note that ARIMA (and its variations) requires a stationary time series, which means that the mean as well as the variance of the series should be constant over time [8]. Stationarity can be tested with the Augmented Dickey-Fuller (ADF) test. If this test fails, the time series' stationarity can be improved by employing a transformation technique. These transformation techniques either employ some form of smoothing, like moving average (MA) or exponential smoothing (ES), or some form of differencing. Another constraint is that the ARIMA process assumes the time series has a zero mean [8]. If that is not the case, the time series can be made compliant with this requirement by subtracting the mean from each of the time series values.

Variations of the ARIMA model include SARIMA and (S)ARIMAX. Seasonality in the time series can be taken into account using SARIMA. The SARIMA model has 4 additional parameters: (P, D, Q) describe the seasonality and m represents the number of periods in each season. With (S)ARIMAX additional external factors can be taken into account, by adding a $\beta_i X_i$ for all i external factors to the standard (S)ARIMA model equation. The time series of the external factors have to comply with the same requirements, so it has to be stationary and have a zero mean.

2.3.5 Neural Networks

Artificial Neural Networks (ANNs) originate from the field of computer science and are inspired on the structure of the human nervous system. ANNs are applied to a wide variety of problems and have four characteristics that make them very suitable for forecasting as well [69]. Firstly, ANNs are data-driven and self-adaptive so few assumptions have to be made up front about the underlying models. In addition,

they can generalize and therefore also produce forecasts for previously unseen items. Thirdly, ANNs can approximate any continuous function and their functional forms are more general and flexible than those of other methods. Finally, ANNs are non-linear, which is often the case for real world systems. However, linear models have the advantage that they are less complex and easier to explain. There are several varieties of ANNs, of which the multi-layer perceptron (MLP) is one of the most commonly used. With an MLP, the first layer receives external input and the last layer produces output (a prediction in our case). Between these layers there are multiple hidden layers with processing nodes. These hidden nodes receive a number of input values, apply weights to them and then apply a certain activation function before outputting a new value to the next layer. An MLP is feed-forward, meaning that data is processed through the network in one direction. Some studies implement recurrent neural networks (RNNs) instead, which contain a feedback loop such that information can persist in the network and for example previous outputs could be used as inputs again. One RNN type that is suggested to be suitable for time series forecasting is long short-term memory (LSTM), which has the benefit of being able to learn long term dependencies. To go into some specifics, the main reason why LSTM has this ability is because it solves the vanishing (and exploding) gradient problems. For a detailed explanation of LSTM architecture, readers are referred to the original paper by Hochreiter and Schmidhuber [31].

Another variation of a feed-forward ANN is the extreme learning machine (ELM), which is tailored for bigger datasets. ELM is a new algorithm for training a feed-forward neural network with one hidden layer [67]. It has similar results as traditional ANNs but is much faster (up to 6 orders of magnitude).

ANNs require quite some hyperparameter tuning. There is a lot of flexibility in the number of hidden layers, hidden nodes, activation functions and learning algorithms. Two measures that are frequently used to determine the optimal parameters are the Bayesian Information Criterion (BIC) and the Aikake Information Criterion (AIC) [50]. These information criteria help to select the model with the best fit, while penalizing model complexity. When comparing the AIC or BIC for two models, the model with the lower AIC or BIC is better than the other. Since AIC and BIC are relative measures, their value for just one model does not have any value. Similarly, AIC and BIC can also be used to automate the parameter selection for (S)ARIMA [8]. In addition to parameter tuning, the input parameters for the neural network have to undergo pre-processing, including a transformation to numeric values and scaling to be within a fixed range.

2.3.6 Ensemble Techniques

Ensemble techniques combine multiple (weaker) prediction models to increase forecast accuracy. The main ensemble techniques include bagging, boosting and random forests. With bagging an original training set is uniformly sampled (with replacement) into a number of new training sets. Models are then trained on each of those new training sets and the outputs from all models are averaged. With boosting a sequence of models is generated, where the first model is trained on the training set and the remaining models on the residuals of the previous models. With random forests, multiple regression trees are generated with each a random selection of training variables and this is combined with bagging to improve performance. One popular ensemble technique for boosting is AdaBoost, which was developed by Freund and Schapire [26].

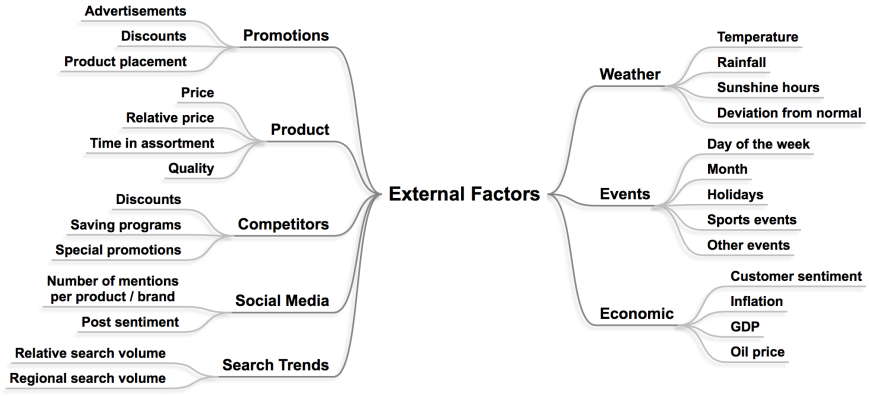

Figure 2.3: Overview of external factor categories with examples

2.3.7 External Factors for Demand Forecasting

When using (S)ARIMAX, regression or neural networks techniques for forecasting several external factors can be taken into account in addition to the historical time series data. A wide variety of external variables can be considered and their relevance could depend on the forecasting problem dimensions. The external factors can be grouped into several categories: weather, events, economic, promotions, product, competitors, social media, search trends. See figure 2.3 for an overview of these external factors. Some of these factors have a causal relationship with demand, for example in case of the weather (more umbrellas will be sold when it's raining). Other factors, such as search trends, may merely correlate with sales (possibly with a lag). These factors are called 'external' because they are not directly about demand. It does not necessarily mean they are external to the retailer, since the promotions and product factors are within the retailer's control (the other factors are not).

Weather-related factors that could be included in the model are: temperature, rainfall, sunshine hours or the deviation of those factors from the normal values at that time of year. It is expected that this factor has a causal relationship with the demand of some products, for example when it's raining there will be more demand for umbrellas and when there's a lot of sunshine BBQ products will be more popular. However, there will also be products for which sales are not influenced by the weather (e.g. toothpaste).

Event factors include all factors that make certain days stand out, including: day of the week, month, holidays (e.g. Christmas), school holidays (e.g. spring break), sports events (e.g. world cup games) and all other events (e.g. carnaval). It is expected that the event factors have a causal relationship with the demand of some products, for example when there's a world cup game sales of beer and chips may rise.

Economic factors include all factors that influence customer buying behaviour on the longer term, including: customer sentiment, inflation, GDP, oil price. It is expected that it is only useful to include these factors in the demand model when the time detail dimension of the forecasting problem is monthly or longer, since the values of these factors are published only on a monthly/quarterly basis and influence customers only on the longer term. These factors are expected to influence luxury products more strongly (e.g. special chocolates), as customers will keep buying

necessity products (e.g. bread).

Promotion factors include factors that increase the attractiveness or visibility of a product. This includes advertisements, discounts and product placement. Product placement can influence sales because customers tend to buy those products that they see at eye-height or that are easy to grab. So for example when there is a special stand with products near the cashier, the products on that stand might have increased sales.

Competitor factors track competitor activities that might influence demand. Such factors can include discounts at the competitor, saving programs (when certain gifts can be earned) and special promotion periods with exceptionally good deals. A great deal at the competitor might directly negatively affect sales for a retailer due to customers switching stores or brands.

Product factors include the price, price relative to similar products, time in assortment and quality. Substitutes and complements of the product could also be included as external factors, since a rise in the demand of one product might cause a decrease in that of another (in case of substitutes) or cause a simultaneous increase in the sales of another product (in case of complements).

With the increasing popularity of social media and search engines, new data sources have become available to use in forecasts. Social media factors that could be considered include: number of mentions for a particular product or brand, number of retailer mentions and sentiment of posts about products/brands. Examples of search factors that could be included are the relative search volume for a product/brand and the difference in those search volumes per region. Social media and search factors might be correlated with sales with a time lag, due to the fact that people search for a product some time before buying it and might evaluate it on social media after buying it. For car sales, it was found that Google search trends and forum data have a significant effect in improving forecast accuracy [27] and that sentiment of social media posts has little further predictive value [66]. It was found that search trends and forum data had the most beneficial effect on forecasts for lower value car brands [27]. Another relevant concept is product involvement, which represents customers' interest in a product and their perceptions regarding its importance and the perceived risk associated with it. Cars are a high-involvement product, so further research could examine the predictive power of these factors for low-involvement products like groceries and movies [27].

In conclusion, this section showed that there is a wide variety of external factors that could be taken into account in demand forecasting. It is likely that the predictive power of these external factors might differ depending on the forecasting problem dimensions and even on the characteristics of the specific products forecasts are generated for.

2.4 Forecasting Performance Evaluation

When evaluating the performance of DF techniques it is important to choose appropriate performance measures (discussed in subsection 2.4.1) and to follow an appropriate evaluation procedure (discussed in subsection 2.4.2).

Table 2.2: Overview of performance measures (adapted from [32])

Name	Short	Type	Calculation
Mean absolute error	MAE	A	$mean(\|e_t\|)$
Mean squared error	MSE	A	$mean(e_t^2)$
Root mean squared error	RMSE	A	$\sqrt{mean(e_t^2)}$
Mean absolute percentage error	MAPE	P	$mean(\|p_t\|)$
Root mean squared percentage error	RMSPE	P	$\sqrt{mean(p_t^2)}$
Mean relative absolute error	MRAE	R	$mean(\|r_t\|)$
Geometric mean relative absolute error	GMRAE	R	$gmean(\|r_t\|)$
Mean absolute scaled error	MASE	S	$mean(\|q_t\|)$

2.4.1 Performance Measures

A wide variety of forecasting performance measures is available and on a high level they can be classified into four main types: absolute, percentage, relative and scaled [32]. Absolute measures are scale-dependent, while percentage, relative and scaled measures are scale-independent. A benefit of scale-independent measures is that they are easier to compare across different forecasts. This is especially helpful when there is a difference in the orders of magnitude of different forecasts, for example when one item is sold much more often than another.

The first step to measure performance is to determine e_t, which is the difference between the observed value (Y_t) and the forecast value (F_t) at time t. Percentage measures use p_t. Relative error measures compare performance to a benchmark e_{tb}, which is generally the naive method. Scaled error measures use q_t which scales errors using the in-sample MAE from the naive method [32] (the calculation provided for q_t is now for a one-step ahead forecast, similar calculations can be made for multi-step ahead forecasts).

$$e_t = Y_t - F_t$$

$$p_t = 100 * e_t / Y_t$$

$$r_t = e_t / e_{tb}$$

$$q_t = \frac{e_t}{\frac{1}{n-1}\sum_{i=2}^{n} |Y_i - Y_{i-1}|}$$

Table 2.2 (adapted from [32]) gives an overview of the most commonly used forecasting performance measures and how they can be calculated.

There is quite some discussion about the suitability of different performance measures. An extensive, critical review of the available forecasting performance measures can be found in the paper by Hyndman [32], where the MASE performance measure was introduced as well. A survey of forecasters in the US showed MAPE is the most popular performance measure in practice [38]. One point of criticism on popular measures like RMSE and MAPE is that outliers can strongly influence the error calculations [24]. A problem with percentage errors is that they are undefined when there are observations with a value of zero. A deficiency of relative error measures is that e_{tb} can be small [32]. Another option would be to use relative measures (e.g. $RelMAE = MAE/MAE_b$) instead of relative errors, however that is not suitable when computing the performance across different series [32]. MASE

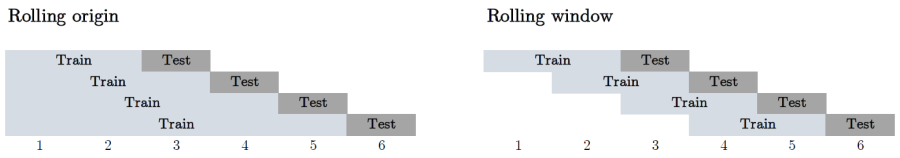

Figure 2.4: Rolling Origin and Rolling Window Procedures

does not have these problems, since it is scale-independent, always defined and symmetric. This discussion indicates that RelRMSE and MASE are two of the most robust performance measures. They are also easy to interpret, since they indicate how well a certain technique performs compared to the naive method. For example, a RelRMSE < 1 means that on average that technique gives smaller errors than the naive method.

2.4.2 Evaluation Procedures

The traditional procedure to evaluate forecasting techniques splits the available time series data into two parts, where the model is trained on the first part and evaluated on the last. However, this procedure does optimally use the available data. By doing cross-validation, a more generalizable impression of model performance could be obtained. With regular, k-fold cross validation, the data is split into k parts. When k $= 5$, the model is trained on 4 parts and evaluated with the remaining part, which is repeated 5 times so all parts have once been the test part. However, regular cross-validation is not a suitable approach in this case, since the data values are not independent. Because sales data is a time series, there is a temporal order in the values. Hence sales data will most likely violate the independence and identically distributed (i.i.d.) assumptions required for cross-validation.

Tashman [58] discusses several evaluation procedures that are tailored to time series forecasting. The forecast origin is the last data point the model uses to compute a forecast. During evaluations, the forecast origin can be fixed, in which case just a single forecast is computed, or rolling. A rolling evaluation procedure can have a rolling origin or rolling window. With a rolling origin the size of the training set differs over time, whereas the size is constant for a rolling window approach. Figure 2.4 illustrates this difference. For the rolling procedure, a choice has to be made between re-calibrating and updating. With re-calibration approach the model is re-trained for each step, whereas this is not the case with updating.

2.5 Forecasting in a Retail Context

There are multiple studies that apply demand forecasting to a case in the retail industry. Table 2.3 gives an overview of demand forecasting literature with a case in different retail sectors. The sectors are indicated with letters, where *G* stands for grocery, *F* fashion, *A* automotive, *CE* consumer electronics and *O* for other. The different studies produce forecasts with different time granularity: daily (D), weekly (W), monthly (M), quarterly (Q) or yearly (Y). Check marks indicate which forecasting techniques are used in each paper.

Table 2.3: Overview of demand forecasting research with retail cases

Paper	Sector	Period	Forecasting Techniques							
			ES	MA	Regression	ARIMA(X)	SARIMA(X)	ANN	Ensemble	Other
[16]	G	D	✓	✓		✓		✓ RBF	✓	
[54]	G	D					✓	✓ MLP		
[11]	G	D						✓ ELM		
[51]	G	W			✓				✓	
[67]	F	M/Q/A				✓		✓ ELM, ENN		
[52]	F	M	✓			✓	✓			
[56]	F	M						✓ ELM		
[53]	F	W				✓		✓ ELM		✓ GM, PPD
[22]	F	H			✓				✓	
[61]	F	Y/W				✓	✓	✓		✓ ANFIS
[27]	A	M			✓			✓ MLP	✓	✓ SVM
[66]	A	M			✓					
[45]	CE	W		✓		✓		✓ MLP, NARX		
[19]	CE	M						✓ MLP		✓ ANFIS
[4]	O	W	✓			✓	✓	✓	✓	✓ ANFIS

The majority of demand forecasting research with a retail case was focused on the fashion industry [67, 52, 56, 53, 22, 61]. For forecasting sales for a variety of fashion items it was shown that the ELM with a harmony search algorithm performed best [67]. In forecasting 5 different types of footwear, Ramos et al. [52] showed that state space models (including exponential smoothing) had a similar performance as ARIMA, so that neither can be said to be best for all circumstances. One paper completely focused on comparing different neural network variations [56] and found that ELM had the best performance & stability compared to two other backpropagation algorithms. Another paper [53] forecast the sales for 6 different fashion items and concluded that their developed pure-panel data (PPD) method performed best, followed by ELM and ARIMA, and that a grey model forecasting technique performed worst. Ferreira et al. [22] focused on forecasting hourly sales at an online fashion outlet called Rue La La and jointly optimizing the sales prices based on the demand model. A study with a large sample of 322 fashion items found that an adaptive network-based fuzzy inference system (ANFIS) had the best performance for forecasting both weekly and yearly sales [61].

Several other studies focused on the grocery sector [16, 54, 11, 51, 59]. For forecasting fresh milk sales, a radial basis function (RBF) neural network with a genetic algorithm for feature selection performed best [16], however since it was only tested on one single product the sample was limited. For forecasting banana sales, it was found that SARIMAX had the best performance [54]. Specifically, the variation that used quantile regression (SARIMAX-QR) instead of the default least squares provided the best prediction intervals and demand factor insights. The sample was again limited, since only banana sales were examined. One study compared neural network variations and showed that Gray ELM had the best performance compared to BPN and MLP for forecasting lunch box sales [11]. Another paper focussed on

regression (ensembles) and showed that multi-variate regression performed best [51]. Their external factors included price, holidays, discounts, inventory and regional factors. Besides demand prediction, the paper also focussed on price optimization. In a comparison of different ES variations for forecasting 256 grocery items, it was found that exponentially weighted quantile regression (EWQR) with just the constant factor performed best [59]. EWQR is equivalent to simple exponential smoothing of the cumulative distribution function.

Other papers had a case in the automotive [66, 27], consumer electronics [45, 19] or furniture sector [4]. The studies with a case in the automotive sector focused mainly on the predictive power of social media data and search trends as external factors and their results have been summarized in section 2.3.7. One of those papers mentions finding similar results for regression, ANN, SVM and ensembles [27], however the results of some techniques (SVM, random forest) were not included in the paper. For forecasting refrigerator component sales, it was found that a Non-linear Autoregressive Network with eXogenous inputs (NARX) had the best performance compared to traditional methods [45]. However, the authors used a limited sample and did no cross-validation. ANFIS was found to have the best performance for predicting sales of cash registers [19], based on a limited sales history and a simulation of additional data. For forecasting sales for 10 furniture items it was found that ensemble methods performed best [4]. Contrary to the most other forecasting literature, Aras et al. [4] did not find a statistically significant difference between the performance of ANN and other techniques such as ES and (S)ARIMA.

From the overview in table 2.3, it becomes clear that even within the retail industry studies have different characteristics. The forecast period differs per sector, with the grocery sector mainly creating forecasts for sales per day, whereas other sectors generally have forecasts with a weekly or monthly time detail level. In addition, a wide variety of forecasting techniques is applied, but there is not yet one technique that stands out in terms of performance for forecasting retail sales. Table 2.3 shows that researchers have a focus on neural networks techniques (included in 11 out of the 15 reviewed studies), that smoothing (4 studies) and regression techniques (4 studies) receive considerably less attention and (S)ARIMA(X) somewhat less (8 studies). It is difficult to do a direct cross-paper comparison of results, since all papers used a different dataset, slightly different variations of forecasting techniques with different external factors, different performance measures and some studies had a limited sample. However, when roughly generalizing the results it seems to point towards neural networks and ensembles as the best performing techniques.

The M3 competition [37] is also worth mentioning here because of its relatively large scale (3003 time series) and the large variety of DF techniques that was evaluated simultaneously. However, it does not focus specifically on retail sales, in fact the majority of time series were of different types, such a demographic time series. Almost no time series were considered with a daily or weekly time detail level, which further limits the usability of these results for retailers.

During the pre-research phase an interview was conducted with a large Dutch grocer to illustrate current practices in the grocery sector. This grocer uses a quite restrictive forecasting support system (FSS) that applies exponential smoothing. Forecasting employees have a lot of freedom to do qualitative forecasting adjustments based on their product knowledge and experience. The forecasting time detail level is now weekly, but the desire is to move towards daily forecasts. In general, the adoption of forecasting techniques from literature is lagging behind in practice.

3
Research Method

The following two sub questions were formulated for the DF part of the research:

DF1. Which quantitative demand forecasting technique performs best in forecasting perishable product demand?

DF2. What guidelines can be provided to improve the wider forecasting process in practice?

This chapter discusses the research method that was followed to answer these questions. Sections 3.1 to 3.3 relate to question DF1 and the literature review process for question DF2 is shortly discussed in this chapter introduction.

For question DF1 the performance of different DF techniques was evaluated across different forecasting scenarios using perishable product sales data from an Ecuadorian food retailer (Favorita Corporacion). Section 3.1 describes the forecasting problem scenarios that were considered. The implemented DF techniques are described in section 3.2. Section 3.2.1 discusses the chosen performance measures and the followed evaluation procedure.

To answer DF2, a structured literature review was conducted. The wider forecasting process aspects that will be investigated include forecasting evaluation frameworks, implementation of DF techniques in decision support systems (DSS) and their adoption, organizational factors influencing forecasting performance and criteria for DF technique selection. The goal of this literature review is to also include 'softer' aspects of the forecasting process and to develop a framework that guides the implementation of the quantitative DF techniques in practice.

3.1 DF Problem Scenarios Considered

To gain insight into which forecasting method is best in which situation, the forecasting problem dimensions will be varied. The following forecasting dimensions will be considered:

- **Forecasting Horizon:** One-step (O) vs. multi-step (M) ahead forecast
- **Time Detail Level:** Weekly (W) vs. daily (D) level
- **Location Detail Level:** Country (C) vs. store (S) level

This results in 8 different scenarios: OWC, OWS, ODC, ODS, MWC, MWS, MDC and MDS. Within the multi-step ahead scenarios, there are again several variations depending on exactly how many steps ahead is forecast. All scenarios will be evaluated with a 2-step ahead forecast (M2). The 3-step (M3) to 7-step (M7) ahead forecasts will be evaluated only for the DC scenario to limit the time needed for the evaluation. Additionally, two variations of input data for the DF techniques will be evaluated:

- **Input Data**: Historical sales only vs. external factors included

These scenarios are common in practice. The literature review and interviews showed that a weekly and daily forecast level is most common in the grocery industry. At least one large Dutch grocer has the desire to move from a weekly to a daily forecast detail level. That same grocer has a multi-step ahead forecast (7-weeks ahead), but the one-step ahead forecast is most important and used for operations. Country-wide forecasts can be used to plan operations in distribution centres, whereas store level forecasts can be used for store-specific replenishment orders.

3.2 DF Techniques Evaluated

Several quantitative DF techniques were selected to be evaluated in this study based on the literature review and interviews. These techniques were selected for comparison:

- Naive, which is used as a baseline for comparison
- Moving average (MA)
- Exponential smoothing (ES)
- Auto-regressive integrated moving average (ARIMA)
- Linear Regression (LINREG)
- Ensemble - ADABoost of Linear Regression (ADA)
- Support Vector Regression (SVR)
- Neural networks (NN)
  - Multi-layer perceptron (MLP)
  - Long short-term memory (LSTM)

This selection represents traditional, commonly used techniques as well as new, increasingly popular techniques. Several of these techniques have been used in previous research, as could be seen from section 2.5, but to the best of our knowledge they were never evaluated simultaneously on a large sample. A survey by McCarthy et al. [38] (from 2006) gives an indication of how familiar forecasters (from the US) are with these techniques: forecasters were most familiar with moving average (84%), exponential smoothing (76%) and regression (73%), whilst they were least familiar with neural networks (17%).

The models were first evaluated with historical data only. In the most basic version, the input for the LINREG, ADA, SVR and NN models consisted of just lag-1 historical data. In a later stage, external factors were added to the LINREG, ADA, SVR and NN models. The reader is referred to appendix A for more

details on how the individual demand forecasting techniques were implemented in Python. Specifically, appendix A.1 gives an overview of which functions / classes from which packages were used as a basis for the implementation of each of the demand forecasting techniques.

Section 3.3.1 describes the dataset that was used and provides some context on the company that provided it. Section 3.3.2 describes the external factors that were considered. Section 3.2.2 describes the set-up of each of the demand forecasting techniques and covers how the parameters for each technique were tuned.

3.2.1 Evaluation Measures and Procedure

Based on the discussion from section 2.4.1, it became clear that RelRMSE is one of the most robust forecasting performance measures. The fact that this is a scaled measure is useful for this study since performance will be compared across many product sales time series which may differ in magnitude. In addition, it allows for an easy comparison of the performance of different techniques compared to the naive forecast baseline. An RelRMSE < 1 means that the particular technique on average performs better than the naive forecast. Percentage errors are unsuitable in this case because there will be days with zero sales for some products, particularly when looking at sales in individual stores.

The raw data was split into a training and a test set for evaluation purposes, where the training set contained 80% of the data available. As discussed in section 2.4.2 the regular k-fold cross-validation approach is unsuitable in this case due to the sequential nature of time series data. From the evaluation procedures that were proposed by Tashman [58] the rolling origin procedure was chosen for this study. The rolling approach makes optimal use of the available data and provides more robust performance results since it creates multiple forecasts. Since recalibration is relatively computationally expensive, the updating variant was chosen.

3.2.2 Hyperparameter Tuning

Several DF techniques require hyperparameter tuning to achieve their best performance. This section describes which hyperparameters were and were not optimized for each technique. Grid searching is one of the methods that is commonly used to optimize hyperparameters. It exhaustively evaluates a hyperparameter set and selects the best-performing one. To make sure a realistic performance result could be provided, cross-validation was performed on the available training data (following the procedures described in section 3.2.1) and the training data was again split into a smaller training set (80% of original training data) and a validation set.

The naive forecast has no hyperparameters and hence does not need hyperparameter tuning. MA uses a window q that was optimized for each SKU individually by doing a grid search for windows within the range [1, 14]. The ES technique uses a smoothing factor α that can be optimized, either directly or indirectly by optimizing the 'span'. From the span, the α can be derived as follows: $\alpha = \frac{2}{span+1}$. A grid search was conducted for spans within the range [1, 14]. For both grid searches, the hyperparameter resulting in the lowest RMSE was selected.

For ARIMA, hyperparameter tuning was more complex. The hyperparameters p, d and q respectively determine the auto-regressive lag, level of differencing and moving average lag. A grid search was performed, but the RMSE was now not used anymore to choose the best hyperparameters since that does not penalize the

number of ARIMA parameters and hence you might end up with an unnecessarily complex model. Instead, the best model was selected based on the BIC, which penalizes the number of parameters (and more heavily so than its alternative AIC). The ranges used during the optimization were (0,3) for p and q and (0,1) for d. Choosing the ARIMA hyperparameters automatically prevents us from having to evaluate all the ACF and PACF plots for all SKUs manually.

LINREG has no hyperparameters, so no optimization was needed. ADA has hyperparameters, but no optimization was performed for them. The number of estimators was fixed at 300 and other hyperparameters were kept default. In addition, ADABoost was only using linear regression models.

For SVR, the hyperparameters that were optimized through grid search were the kernel and penalty parameter C. The kernels that were evaluated during optimization were a linear kernel, a radial basis function kernel and a polynomial kernel. Cs evaluated included 0.25, 0.5, 0.75, 1.0 and 1.25.

For the MLP a single hidden layer was used with rectified linear unit (relu) activation functions. The number of nodes in the hidden layer was optimized for each SKU and were within the range (1, 120) with a step size of 5. For the LSTM no hyperparameter tuning was performed for each individual item since it has a relatively long training time making grid search quite expensive. There was one hidden layer with 3 LSTM neurons. The number of epochs was set to 3, meaning that the data from the training set is passed through the network 3 times to learn from. Although these hyperparameters are not automatically optimized for each SKU, they were definitely not chosen randomly and were selected carefully based on a manual evaluation of a small hyperparameter range for a small set of items. This manual evaluation showed that for those items evaluated, 3 neurons seemed to be the optimal amount both when working with and without external factors. In addition, both MLP and LSTM forecasts were repeated 3 times during the evaluation and their results averaged to minimize the influence of their random start states on the evaluation results.

3.3 Dataset and Preparation

This section describes the data that was used as a baseline for the comparison of DF techniques. The dataset that was used was provided by Favorita Corporacion, a large Ecuadorian retailer. Favorita's revenue in 2016 was $1.824 billion and profit was $135 million, hence their profit margin was roughly 7%. The company aims to improve its demand forecasts and for that purpose they provided a publicly available dataset online[1]. To the best of our knowledge, this is the most extensive and most detailed food sales history data source that is publicly available, which is why it was selected for this study.

3.3.1 Sales History Data

The dataset contains Favorita's daily sales per product per store from January 2013 to August 2017. The dataset covers 54 stores throughout Ecuador and 4100 items in different categories. The raw sales history matches the DS scenario, but this data can be aggregated to a weekly and/or country level to represent all scenarios. Items are identified by a unique item number and their high level category (e.g.

[1]Data source: `https://www.kaggle.com/c/favorita-grocery-sales-forecasting/data`

Table 3.1: Percentage of items per perishable product category in the dataset

Category	% Items
Produce	31,0%
Dairy	24,5%
Bread/Bakery	13,6%
Deli	9,2%
Meats	8,5%
Poultry	5,5%
Eggs	4,2%
Prepared Foods	2,6%
Seafood	0,8%

Figure 3.1: External factors that were considered as part of this study

"dairy") and perishability are provided. For this study, only perishable products will be considered, which leaves a dataset with 986 perishable items. The large size of this dataset makes it a good basis for a robust comparison of different demand forecasting techniques. Table 3.1 gives an overview of the different perishable product categories that are present in the dataset and what percentage of items falls in each of those categories.

3.3.2 External Factors Considered

Figure 3.1 gives an overview of the external factors that are taken into account in this study, which fall in the weather, events, economic and promotions categories. Daily Ecuadorian weather data was obtained from DarkSky[2]. The weather data includes minimum and maximum temperature, cloud cover, humidity, dew point, visibility and wind speeds for the cities where the 54 Favorita stores are located. Data for the other external factor categories was already included in or derived from the Favorita dataset. The events data that is used contains days of the week and national Ecuadorian holidays (regional holidays were excluded). Three types of holidays were distinguished: holidays, events and additional days. Additional days are the days around holidays or events, such as the days before Christmas. The promotions data that is used contains whether a product was on promotion (true or false). The economic data that is used the daily oil prices in Ecuador.

The options for external factor feature selection were evaluated as well, although not automatically for each individual item during the evaluation since this would

[2]DarkSky API: `https://www.darksky.net/dev`

consume too much time. A manual evaluation was performed on a small subset of items where the K (= 5, 10, 15, 20, all) best features were selected based on the mutual information scores. This indicated that using all features resulted in the best performance.

When comparing figure 3.1 with figure 2.3 which provided an overview of all external factor categories that could be used for demand forecasting, it can be seen that some external factor categories were not considered during this study. External factors from the product, competitors, social media and search trends categories were not considered, since such data was unfortunately not available. For example, no detailed product information and no price information was available because Favorita anonymized the publicly available dataset. To find relevant social media messages or search trends, at least brand and preferably product names would have to be available.

3.3.3 Data Preprocessing

No special data preprocessing was necessary for NAIVE, MA and ES. Since ARIMA assumes that the data is stationary and that it has a zero mean, some pre-processing was required to make sure these assumptions are met. The raw time series data was demeaned to make sure it had a zero mean. The augmented dickey-fuller (ADF) test was performed to check whether the time series is stationary before applying the ARIMA model. For the LINREG, ADA, SVR, MLP and LSTM models more pre-processing was required and in particular the data is scaled so all features have the same scale and have values between 0 and 1.

When external factors were included in the evaluation, these were only given as input to the LINREG, ADA, SVR, MLP and LSTM techniques. When these external factors were categorical they were transformed into dummy variables. For example if there is a feature 'Holiday Type' which can be 0 (National), 1 (Regional) or 2 (Additional), some forecasting techniques may misinterpret this and assume there is an order in the data. By transforming this feature to dummy variables, each holiday type will become a separate, binary feature and there is no room anymore for the forecasting techniques to misinterpret the data.

4
Performance Comparison Results

This chapter covers the results from the performance comparison of DF techniques. Section 4.1 discusses results for the one-step ahead forecasting scenarios, section 4.2 for the multi-step ahead forecasting scenarios and section 4.3 covers the results on the influence of external factors.

For each scenario, a table is provided with for each DF technique the average, minimum and maximum RelRMSE score. The average RelRMSE score can be seen as a measure of each technique's performance consistency, since it reflects how well it performs on average across all items. When a retailer wants to use a single forecasting technique for all its items, he most likely wants to choose the DF technique that performs best on average (lowest average RelRMSE). The last row in each table, with 'BEST', reports the RelRMSE score when the best-performing DF technique was chosen for each individual item. The last column in the table reports the %best, which is the percentage of items for which each DF technique was the best-performing one. When the retailer wants to use an algorithm that automatically selects the best DF technique for each item, the best% score reflects how important it is to include each DF technique in the set of techniques the forecasting algorithm can choose from.

Wilcoxon signed-rank tests were conducted for all pairs of DF techniques in each scenario to test whether the distributions of their RelRMSE scores were significantly different. More details about this test and the significance test results can be found in appendix B.

4.1 Results for One-Step Ahead Scenarios

This section first describes the results for each individual one-step ahead scenario and then discusses the results in a cross-scenario comparison.

OWC Scenario results

Table 4.1 contains the RelRMSE scores for the OWC forecasting scenario. ARIMA performed best for the largest amount of items (27.6%), followed closely by LINREG (24.0%) and ADA (18.9%). When looking at performance consistency, it can be

seen that ARIMA and LINREG had the best average RelRMSE scores. Interestingly, all other DF techniques had an average RelRMSE > 1, so on average they performed worse than the naive forecast. Always forecasting using ARIMA or LINREG resulted in respectively a 1.5% and 1.3% error reduction compared to always using the NAIVE forecast. Always automatically selecting the best-performing forecasting technique for each individual item resulted in a 6.4% improvement compared to always using the NAIVE forecast or a 4.9% improvement compared to always using ARIMA.

ODC Scenario results

Table 4.2 contains the RelRMSE scores for the ODC forecasting scenario. The ranking of DF techniques shows that ARIMA by far performed best for the largest amount of items (72.1%), followed at large distance by MA (10.0%) and MLP (6.0%). However, in terms of performance consistency, ARIMA has a RelRMSE > 1, meaning that on average its performance is worse than using NAIVE. The techniques with best performance consistency are MA and ES, followed by MLP. Always using MA or ES resulted in a 12.8% performance improvement compared to always using the NAIVE forecast, always using MLP resulted in a 9% improvement. Always automatically selecting the best-performing forecasting technique for each individual item resulted in a 21.6% improvement compared to always using the NAIVE forecast or an 8.8% improvement compared to always using MA or ES.

OWS Scenario results

Table 4.3 contains the RelRMSE scores for the OWS scenario. ARIMA performed best for the largest amount of items (26.8%), followed by ES (18.7%) and SVR (14.1%). In terms of performance consistency, ES and ARIMA had the best RelRMSE scores, respectively resulting in a 9.3% and 9.1% error reduction compared to always using the NAIVE forecast. Even though SVR was the best technique for many items, it has low performance consistency and on average performs worse than the NAIVE forecast. Always automatically selecting the best-performing forecasting technique for each individual item resulted in a 14.8% improvement compared to always using the NAIVE forecast or an 5.5% improvement compared to always using ES.

ODS Scenario results

Table 4.4 contains the RelRMSE scores for the ODS scenario. ARIMA again performed best for the largest amount of items (37.4%), followed by LSTM (18.0%) and ES (16.4%). ES and LSTM also had the best performance consistency, respectively resulting in a 21.7% and 20.2% improvement in RelRMSE compared to always using NAIVE. Always automatically selecting the best-performing forecasting technique for each individual item resulted in a 24.3% improvement compared to always using the NAIVE forecast or a 2.6% improvement compared to always using ES.

Discussion of results across one-step ahead scenarios

Tables 4.5 and 4.6 give an overview of the results for the one-step ahead scenarios. They respectively contain each scenario's top 3 for the best technique in terms of average RelRMSE (performance consistency) and in terms of the percentage

Table 4.1: OWC results (RelRMSE) for 909 items

	avg	std	min	max	%best
NAIVE	1.000	0.000	1.000	1.000	8.1
MA	1.159	0.519	0.692	8.237	0.3
ES	1.000	0.082	0.721	1.522	8.5
LINREG	**0.987**	0.132	0.699	2.714	**24.0**
ADA	1.004	0.168	0.695	2.973	**18.9**
ARIMA	**0.985**	0.105	0.650	1.963	**27.6**
SVR	1.498	0.673	0.718	6.554	4.7
MLP	1.009	0.188	0.699	3.980	6.4
LSTM	1.374	0.404	0.682	5.289	1.4
BEST	**0.936**	0.051	0.650	1.0	-

Table 4.2: ODC results (RelRMSE) for 986 items

	avg	std	min	max	%best
NAIVE	1.000	0.000	1.000	1.000	1.9
MA	**0.872**	0.108	0.598	1.509	**10.0**
ES	**0.872**	0.094	0.602	1.215	3.9
LINREG	0.912	0.105	0.477	2.102	1.4
ADA	0.931	0.166	0.493	3.015	2.0
ARIMA	1.097	6.009	0.440	142.175	**72.1**
SVR	1.078	0.454	0.538	7.797	1.0
MLP	**0.910**	0.137	0.474	2.260	**6.0**
LSTM	0.929	0.182	0.478	3.064	1.6
BEST	**0.784**	0.095	0.440	1.000	-

Table 4.3: OWS results (RelRMSE) for 4278 store-items

	avg	std	min	max	%best
NAIVE	1.000	0.000	1.000	1.000	4.3
MA	1.123	0.421	0.552	7.429	1.3
ES	**0.907**	0.105	0.641	1.528	**18.7**
LINREG	0.971	0.447	0.598	25.712	8.6
ADA	0.998	0.529	0.334	27.692	11.9
ARIMA	**0.909**	0.475	0.552	30.836	**26.8**
SVR	1.125	0.739	0.118	29.508	**14.1**
MLP	0.981	0.282	0.530	10.136	9.0
LSTM	1.088	0.351	0.179	8.040	5.4
BEST	**0.852**	0.081	0.118	1.000	-

Table 4.4: ODS results (RelRMSE) for 4827 store-items

	avg	std	min	max	%best
NAIVE	1.000	0.000	1.000	1.000	0.3
MA	0.823	0.102	0.380	2.799	2.7
ES	**0.783**	0.060	0.607	1.312	**16.4**
LINREG	0.854	0.586	0.498	36.684	4.6
ADA	0.911	1.007	0.479	58.572	5.4
ARIMA	0.803	0.776	0.430	37.641	**37.4**
SVR	0.850	0.188	0.618	8.628	8.9
MLP	0.845	0.464	0.492	29.428	6.4
LSTM	**0.798**	0.111	0.465	3.915	**18.0**
BEST	**0.757**	0.060	0.380	1.000	-

Table 4.5: Average RelRMSE top 3 per one-step ahead scenario

	OWC	ODC	OWS	ODS
#1	ARIMA (0.985)	MA (0.872)	ES (0.907)	ES (0.783)
#2	LINREG (0.987)	ES (0.872)	ARIMA (0.909)	LSTM (0.798)
#3	-	MLP (0.910)	LINREG (0.971)	ARIMA (0.803)

Table 4.6: Best% top 3 per one-step ahead scenario

	OWC	ODC	OWS	ODS
#1	ARIMA (28%)	ARIMA (72%)	ARIMA (27%)	ARIMA (37%)
#2	LINREG (24%)	MA (10%)	ES (19%)	LSTM (18%)
#3	ADA (20%)	MLP (6%)	SVR (14%)	ES (16%)

of items for which it was the best technique. When looking at the performance of individual forecasting techniques across scenarios, it can be seen that ARIMA performed quite well overall. For all scenarios, it was the model that was 'best' for the largest number of items. However, its performance consistency was not always good. This was particularly visible in the ODC scenario, where ARIMA on average performed worse than NAIVE. The ODC scenario was where MA achieved its best results, where it was one of the techniques with the best average RelRMSE. For the OWC and OWS scenarios, MA on average was worse than NAIVE. ES scored very well in terms of performance consistency, and was in the RelRMSE top 3 for 3 out of 4 scenarios. LINREG showed its best performance in the OWC scenario, where it was almost as good as ARIMA. ADA performed well for a large number of items in the OWC scenario, but its performance consistency was not good. SVR scored well for a large number of items in the OWS scenario, but again had bad performance consistency. MLP was part of the top 3 for the ODC scenario, but also performed quite well in the ODS scenario. LSTM showed its worst performance in the OWC scenario, was still worse than naive for the OWS scenario, performed quite well for the ODC scenario and for the ODS scenario it turned out to be one of the best performing techniques.

Based on these results, it can be concluded that ARIMA performs well for a large number of items in many scenarios but that performance consistency is sometimes an issue, that ES generally has good performance consistency and that the relative performance of neural networks techniques such as MLP and LSTM increase greatly as the forecasting problem becomes more fine-grained.

The best overall performance by far was obtained by automatically choosing the best technique for each individual item(-store combination). It achieves a RelRMSE improvement ranging from 6.4% (OWC scenario) to 24.3% (ODS scenario) compared to NAIVE. So the impact of automatically selecting the best forecasting technique is greatest for forecasting scenarios that are more fine-grained. Automatic technique selection achieves a RelRMSE improvement ranging from 2.6% (ODS scenario) to 8.8% (ODC scenario) compared to always using the best individual technique in each scenario.

Table 4.7: ODS scenario RelRMSE per DF technique per product category

	ADA	ARIMA	BEST	ES	LINREG	LSTM	MA	MLP	SVR
BAKERY	0.98	0.80	0.75	0.77	0.88	0.80	0.81	0.86	0.84
DAIRY	0.88	0.80	0.77	0.79	0.83	0.81	0.85	0.83	0.82
DELI	0.88	0.80	0.76	0.78	0.82	0.79	0.83	0.83	0.82
EGGS	0.89	0.79	0.76	0.80	0.83	0.80	0.83	0.83	0.83
MEATS	0.91	0.87	0.74	0.77	0.83	0.78	0.79	0.83	0.81
POULTRY	0.97	0.79	0.75	0.78	0.90	0.80	0.81	0.89	0.88
PREPARED	0.89	0.81	0.76	0.81	0.86	0.81	0.86	0.85	0.91
PRODUCE	0.91	0.80	0.76	0.79	0.87	0.79	0.82	0.85	0.89
SEAFOOD	0.87	0.77	0.76	0.78	0.84	0.81	0.81	0.84	0.85

Discussion of results across product categories

It was also investigated whether performance for each of the DF techniques differed across item families. The results for this comparison for the ODS scenario can be found in table 4.7. These results show that there are no large differences in terms of RelRMSE performance between product categories for individual DF techniques nor for automatic DF technique selection. This strengthens the generalizability of results, because it indicates that the mix of product categories a specific food retailer offers will not have a large influence on the overall RelRMSE results at that retailer. Similar findings were obtained in all other scenarios in this study and to limit the length of this thesis those results were not included, but they are available for request.

4.2 Results for Multi-Step Ahead Scenarios

There can be several variations of multi-step ahead forecasts, depending on exactly how many steps ahead is forecast. To limit the time needed to run all the evaluations, the two-step ahead forecasts were done for all scenarios and the three- to seven-step ahead forecasts only for the DC (daily, country-level) scenario. The two-step ahead results are discussed in detail per individual scenario, the results for scenarios with more steps ahead are shortly discussed at the end of this section.

M2WC Scenario results

Table 4.8 shows the results for the M2WC scenario. ARIMA is the best performing technique by far, both in terms of the percentage of items for which it is the 'best' technique (68.5%) and in terms of performance consistency. It achieves an average 15.6% improvement in RelRMSE compared to always using NAIVE. The next-best techniques in terms of performance consistency were ES and LINREG, with respectively a 5.9% and 4.6% improvement compared to NAIVE. Always automatically selecting the best technique for each individual item results in a 17.7% improvement compared to always using NAIVE and an additional 2.1% improvement compared to always using ARIMA.

M2DC Scenario results

Table 4.9 shows the results for the M2DC scenario. ARIMA again is the best model for the largest number of items (86.2%). However, it is the second-worst technique in terms of performance consistency since it achieves a 5.2% improvement in average RelRMSE compared to NAIVE. The techniques with best performance consistency were ES, MA and LSTM, which respectively resulted in a 22.3%, 22.2% and 14.9% average RelRMSE improvement compared to always using NAIVE. Automatic selection fo the best technique for each item resulted in a 32.4% improvement in RelRMSE compared to always using NAIVE and 10.1% compared to always using ES.

M2WS Scenario results

Table 4.10 shows the results for the M2WS scenario. ARIMA again performs best for most items (49.5%), followed by ES (15.8%) and SVR (12.9%). ARIMA also has the best performance consistency and achieves a 16% average RelRMSE improvement compared to the NAIVE forecasts. The next-best techniques in terms of performance consistency were ES and LINREG, with respectively a 12.6% and 5.7% improvement compared to NAIVE. Automatic best technique selection results in a 20.4% improvement compared to NAIVE, so that is an additional 4.4% compared to ARIMA.

M2DS Scenario results

Table 4.11 shows the results for the M2DS scenario. ARIMA performs best for most items (52.2%), followed by LSTM (16.4%) and ES (13.2%). ES has the best performance consistency and results in a 24.6% improvement compared to always using NAIVE. Closely behind ES in terms of performance consistency are ARIMA and LSTM, with respectively a 23.9% and 22.9% improvement compared to NAIVE. Automatically selecting the best DF technique for each item results in a 27.9% improvement compared to always using NAIVE, which is 3.3% more compared to always using ES.

Discussion of results across two-step ahead scenarios

Tables 4.12 and 4.13 give an overview of the results for the two-step ahead scenarios. They respectively contain each scenario's top 3 for the best technique in terms of average RelRMSE (performance consistency) and in terms of the percentage of items for which it was the best technique.

When looking at the performance of DF techniques across scenarios, it can be seen that ARIMA performed quite well overall. It is best for the highest percentage of items in all four scenarios. ARIMA is also in the average RelRMSE top 3 for three out of four scenarios (first place for M2WC and M2WS and second place for M2DS). It can be concluded that ARIMA has its lowest performance consistency for scenarios with a daily time detail level.

The technique with best performance consistency overall is ES, as it is in the average RelRMSE top 3 for all scenarios with one first place in scenario M2DS and second places in the other three scenarios.

Although LSTM does not perform well in the M2WC and M2WS scenarios, it does perform well for the M2DC scenario, where it is in the average RelRMSE top

Table 4.8: M2WC results (RelRMSE) for 909 items

	avg	std	min	max	%best
NAIVE	1.000	0.000	1.000	1.000	2.1
MA	1.198	0.702	0.439	10.363	1.3
ES	**0.941**	0.080	0.688	1.407	**6.5**
LINREG	**0.954**	0.147	0.711	2.647	4.1
ADA	0.976	0.199	0.713	3.168	5.1
ARIMA	**0.844**	0.120	0.411	1.867	**68.5**
SVR	1.238	0.573	0.454	5.658	**8.6**
MLP	0.971	0.188	0.711	3.809	1.3
LSTM	1.159	0.337	0.510	3.197	2.5
BEST	**0.823**	0.096	0.411	1.000	-

Table 4.9: M2DC results (RelRMSE) for 986 items

	avg	std	min	max	%best
NAIVE	1.000	0.000	1.000	1.000	0.6
MA	**0.778**	0.102	0.604	1.308	**6.7**
ES	**0.777**	0.074	0.606	1.154	**3.4**
LINREG	0.881	0.157	0.602	2.616	0.3
ADA	0.894	0.196	0.600	2.894	0.2
ARIMA	0.948	5.479	0.377	133.141	**86.2**
SVR	0.962	0.367	0.609	5.847	0.8
MLP	0.877	0.169	0.573	2.616	0.3
LSTM	**0.851**	0.190	0.606	3.480	1.4
BEST	**0.676**	0.090	0.377	1.000	-

Table 4.10: M2WS results (RelRMSE) for 4278 store-items

	avg	std	min	max	%best
NAIVE	1.000	0.000	1.000	1.000	0.9
MA	1.142	0.548	0.198	9.774	2.0
ES	0.874	0.090	0.649	1.374	**15.8**
LINREG	0.943	0.487	0.537	28.762	4.0
ADA	0.980	0.730	0.475	30.993	5.7
ARIMA	0.840	0.473	0.183	30.823	**49.5**
SVR	1.046	0.648	0.119	23.418	**12.9**
MLP	0.961	0.585	0.539	28.646	3.8
LSTM	1.016	0.442	0.152	16.408	5.5
BEST	0.796	0.077	0.119	1.000	-

Table 4.11: M2DS results (RelRMSE) for 4827 store-items

	avg	std	min	max	%best
NAIVE	1.000	0.000	1.000	1.000	0.2
MA	0.792	0.141	0.387	6.556	2.2
ES	**0.754**	0.048	0.645	1.201	**13.2**
LINREG	0.834	0.539	0.610	32.468	2.8
ADA	0.897	1.165	0.602	70.730	3.8
ARIMA	**0.761**	0.764	0.391	39.024	**52.2**
SVR	0.826	0.150	0.613	5.535	7.0
MLP	0.822	0.310	0.610	15.107	2.3
LSTM	**0.771**	0.120	0.596	4.280	**16.4**
BEST	**0.721**	0.053	0.387	1.000	-

Table 4.12: Average RelRMSE top 3 per two-step ahead scenario

	M2WC	**M2DC**	**M2WS**	**M2DS**
#1	ARIMA (0.844)	MA (0.778)	ARIMA (0.840)	ES (0.754)
#2	ES (0.941)	ES (0.777)	ES (0.874)	ARIMA (0.761)
#3	LINREG (0.954)	LSTM (0.851)	LINREG (0.943)	LSTM (0.771)

Table 4.13: Best% top 3 per two-step ahead scenario

	M2WC	**M2DC**	**M2WS**	**M2DS**
#1	ARIMA (69%)	ARIMA (86%)	ARIMA (50%)	ARIMA (52.2%)
#2	SVR (9%)	MA (7%)	ES (16%)	LSTM (16.0%)
#3	ES (7%)	ES (3%)	SVR (13%)	ES (13.4%)

3. LSTM particularly impresses in the M2DS scenario, where it is in the top 3 for both best% and average RelRMSE. It can be concluded that LSTM achieves good performance for scenarios with a daily time detail level and its best performance when such a scenario has a store location detail level.

The best performance by far is achieved when for each item the best DF technique is chosen automatically. The automatic best model selection for each individual item results in an improvement that ranges from 17.7% (M2WC) to 32.4% (M2DC) compared to NAIVE. The improvement ranges from 2.1% (M2WC) to 10.1% (M2DC) when comparing to the best individual forecasting technique. The added value of the automatic DF technique selection is greatest for the scenarios with a daily time detail level.

Discussion of results for M3DC to M7DC scenarios

The results tables for the M3DC to M7DC scenarios can be found in tables 4.14 to 4.18. Some interesting trends can be identified from these results. As the forecast is created for more steps ahead, ARIMA is less frequently the best DF technique. For the M2DC scenario, ARIMA was still the best for 86.2% of items, which gradually declines to 40.8% of items in the M7DC scenario. For the three- to six-step ahead scenarios, MA is best for more items, from 9.2% in the M3DC scenario to 25.5% in the M6DC scenario. MA does not perform so well in the M7DC scenario, where the more advanced forecasting techniques perform relatively well. LINREG, ADA and MLP show great performance improvements compared to the fewer steps ahead scenarios, which could for example be because these models can extract more complex patterns in weekly sales differences. The top 3 best DF techniques in terms of lowest average RelRMSE scores remains the same for M3DC to M6DC (#1 MA, #2 ES, #3 LSTM), but has changed for M7DC (#1 LINREG, #2 MLP, #3 ES). The implications of these trends are that depending on how many steps ahead is forecast, it differs which DF techniques perform best, so that the retailer should make a different selection of techniques to implement. Again, in all forecasting scenarios, by far the best performance can be achieved by automatically selecting the best forecasting technique for each individual item, which results in a 12.5% RelRMSE improvement in the M7DC scenario to 35.9% in the M5DC scenario compared to always using NAIVE. Compared to always using the best performing individual DF technique, automatic selection results in a 6.6% RelRMSE improvement in the M7DC scenario to 9.9% in the M3DC scenario.

4.3 Results for External Factors

This section describes the results of the one-step ahead scenarios where forecasts were not only based on historical sales data, but also on external factors including the weather and holidays. Results are compared to the scenarios that only used historical data to gain insight into the added value of using these external factors for perishable food product demand forecasting.

OWCef Scenario results

Table 4.19 shows the results for the OWC scenario with external factors. When comparing the OWCef scenario to the original OWC scenario that only uses historical

Table 4.14: M3DC results (RelRMSE) for 986 items

	avg	std	min	max	%best
NAIVE	1.000	0.000	1.000	1.000	0.1
MA	**0.753**	0.115	0.542	2.098	**9.2**
ES	**0.756**	0.080	0.631	1.091	**3.5**
LINREG	0.867	0.178	0.508	2.969	0.0
ADA	0.881	0.243	0.494	4.934	0.1
ARIMA	0.926	5.753	0.333	145.597	**84.0**
SVR	0.933	0.353	0.556	5.305	0.7
MLP	0.868	0.210	0.467	3.531	0.7
LSTM	**0.835**	0.200	0.500	3.929	1.6
BEST	**0.654**	0.093	0.333	1.000	-

Table 4.15: M4DC results (RelRMSE) for 986 items

	avg	std	min	max	%best
NAIVE	1.000	0.000	1.000	1.000	0.1
MA	**0.731**	0.112	0.479	1.720	**14.0**
ES	**0.743**	0.087	0.594	1.077	**2.9**
LINREG	0.860	0.177	0.508	2.991	0.1
ADA	0.867	0.193	0.493	3.144	0.1
ARIMA	0.920	5.857	0.332	148.268	**80.6**
SVR	0.918	0.334	0.560	4.867	0.7
MLP	0.854	0.180	0.460	2.736	0.4
LSTM	**0.817**	0.174	0.555	3.087	1.0
BEST	**0.643**	0.088	0.332	1.000	-

Table 4.16: M5DC results (RelRMSE) for 986 items

	avg	std	min	max	%best
NAIVE	1.000	0.000	1.000	1.000	0.1
MA	**0.721**	0.104	0.434	1.608	**20.3**
ES	**0.736**	0.077	0.592	1.106	**3.9**
LINREG	0.864	0.154	0.619	2.674	0.0
ADA	0.872	0.180	0.589	3.155	0.3
ARIMA	0.910	5.519	0.368	133.013	**74.2**
SVR	0.925	0.340	0.580	5.253	0.6
MLP	0.861	0.162	0.559	2.825	0.2
LSTM	**0.814**	0.168	0.623	3.515	0.4
BEST	**0.641**	0.075	0.368	1.000	-

Table 4.17: M6DC results (RelRMSE) for 986 items

	avg	std	min	max	%best
NAIVE	1.000	0.000	1.000	1.000	0.1
MA	**0.762**	0.106	0.494	2.306	**25.5**
ES	**0.766**	0.071	0.606	1.138	**7.1**
LINREG	0.881	0.111	0.540	2.124	0.1
ADA	0.896	0.165	0.542	2.847	0.2
ARIMA	0.979	5.802	0.403	141.409	**65.8**
SVR	0.980	0.388	0.576	6.527	0.5
MLP	0.883	0.133	0.531	2.244	0.1
LSTM	**0.847**	0.162	0.567	2.883	0.6
BEST	**0.686**	0.065	0.403	1.000	-

Table 4.18: M7DC results (RelRMSE) for 986 items

	avg	std	min	max	%best
NAIVE	1.000	0.000	1.000	1.000	2.2
MA	1.016	0.150	0.751	4.118	0.0
ES	**0.968**	0.067	0.757	2.027	8.5
LINREG	**0.941**	0.087	0.765	1.775	12.6
ADA	0.973	0.158	0.767	2.501	**16.7**
ARIMA	1.410	9.509	0.442	227.126	**40.8**
SVR	1.210	0.557	0.680	7.661	2.3
MLP	**0.945**	0.103	0.668	2.193	**16.3**
LSTM	1.104	0.238	0.740	3.911	0.5
BEST	**0.875**	0.082	0.442	1.000	-

data, it becomes clear that performance on average has deteriorated. Automatically selecting the best DF technique for each item now results in a 6.4% average RelRMSE improvement compared to NAIVE, which is 0.3% less than in the OWC scenario. LINREG, ADA and LSTM performed worse on average with the external data included and were also less frequently the best DF techniques. MLP performed slightly worse on average, but was now best for a larger percentage of items. SVR performed slightly better on average and also had a slightly higher percentage of items for which it was best. Just like in the original OWC scenario, ARIMA was now still the best DF technique for the largest percentage of items.

ODCef Scenario results

Table 4.20 shows the results for the ODC scenario with external factors. MLP was best for the largest percentage of items (34.9%), followed by LINREG (27.0%) and LSTM (14.3%). MLP and LSTM had the best average performance, resulting respectively in a 19.6% and 13.1% average RelRMSE improvement compared to always using NAIVE. Although LINREG was best for a large percentage of items, it had bad performance consistency resulting in an extremely high average RelRMSE score. Automatically selecting the best DF technique for each item now results in a 30.2% improvement in average RelRMSE compared to always using NAIVE. In the original ODC scenario this was 21.6%, so adding the external factors results in an additional 8.6% improvement in RelRMSE. Adding the external factors enabled the regression and neural networks techniques to outperform ARIMA, which was by far the best technique for the largest percentage of items in the scenario with historical data only.

OWSef Scenario results

Table 4.21 shows the results for the OWS scenario with external factors. Automatically selecting the best DF technique for each item now results in a 14.7% improvement in average RelRMSE compared to always using NAIVE. In the original OWS scenario this was 14.8%, so adding the external factors decreased the average RelRMSE score with 0.1%. The average performance for all DF techniques that used the external factors has now also decreased slightly. For some techniques (such as MLP) the minimum RelRMSE score is lower, indicating that for some individual items it was valuable to add the external factors. ARIMA is still best for the largest amount of items (31.0%), followed by ES (19.3%) and SVR (16.9%).

ODSef Scenario results

Table 4.22 shows the results for the ODS scenario with external factors. LSTM was best for the largest percentage of items (24.3%), followed by SVR (17.7%) and ARIMA (16.5%). ES had the best average performance, resulting in a 21.7% average RelRMSE improvement compared to always using NAIVE, followed by SVR and LSTM with respectively 19% and 18.8%. Automatically selecting the best DF technique for each item now results in a 27.2% average RelRMSE improvement compared to always using NAIVE. In the original ODS scenario this was 24.3%, so adding the external factors resulted in an additional 2.9% improvement in RelRMSE. Adding the external factors enabled LSTM to outperform ARIMA, which was the best for the largest percentage of items in the scenario with historical data only.

Discussion of results across external factor scenarios

Adding external factors has shown to have large benefits for forecasting scenarios with a daily time detail level. For scenarios with a weekly time detail level, adding the external factors even resulted in a minor performance decrease. Therefore, it can be said that adding external factors as input to a demand forecasting technique should only be considered when the forecasts have a daily time detail level. Automatically selecting the best DF technique for each item(-store combination) still resulted in the best performance. The external factors enabled an additional 8.6% average RelRMSE improvement in the ODCef scenario and 2.9% in the ODSef scenario compared to the original scenarios with historical data only when automatically selecting the best DF techniques.

Adding the external factors also changed the top 3 for each scenario. Tables 4.23 and 4.24 respectively show the top 3 DF techniques in terms of the lowest average RelRMSE score and the highest %best for each one-step ahead scenario with external factors. Whereas ARIMA was in first place in terms of %best for all one-step ahead scenarios that used only historical data, it now remains in first place only in the weekly scenarios. ARIMA is third for the ODSef scenario and not even in the top 3 for the ODCef scenario. In the ODCef scenario, MLP, LSTM and MA performed best on average, whereas MLP, LINREG and LSTM performed best for the largest percentage of items. In the ODSef scenario, ES still had the best average performance, but SVR and LSTM followed closely behind. LSTM and SVR were also the techniques with the highest best%.

Table 4.19: OWCef results (RelRMSE) for 909 items

	avg	std	min	max	%best
NAIVE	1.000	0.000	1.000	1.000	**17.9**
MA	1.159	0.519	0.692	8.237	0.4
ES	1.000	0.082	0.721	1.522	9.7
LINREG	1.029	0.182	0.694	2.752	**17.6**
ADA	1.113	0.340	0.698	4.806	7.6
ARIMA	**0.985**	0.105	0.650	1.963	**32.2**
SVR	1.485	0.654	0.716	6.497	5.9
MLP	1.120	0.335	0.692	4.058	7.6
LSTM	1.419	0.468	0.756	5.174	1.0
BEST	**0.939**	0.057	0.650	1.000	-

Table 4.20: ODCef results (RelRMSE) for 986 items

	avg	std	min	max	%best
NAIVE	1.000	0.000	1.000	1.000	1.2
MA	0.871	0.109	0.598	1.509	1.3
ES	0.872	0.095	0.602	1.215	1.5
LINREG	9.42e+8	2.9e+10	0.318	9.3e+11	**27.0**
ADA	0.899	1.087	0.319	30.762	5.3
ARIMA	0.874	2.055	0.443	65.136	11.4
SVR	0.989	0.507	0.431	7.927	3.1
MLP	**0.804**	1.294	0.288	35.132	**34.9**
LSTM	**0.869**	0.327	0.338	4.422	**14.3**
BEST	**0.680**	0.125	0.288	1.000	-

Table 4.21: OWSef results (RelRMSE) for 4278 store-items

	avg	std	min	max	%best
NAIVE	1.000	0.000	1.000	1.000	6.6
MA	1.123	0.421	0.552	7.429	1.5
ES	**0.907**	0.105	0.641	1.528	**19.3**
LINREG	1.010	0.524	0.472	28.623	12.0
ADA	1.097	0.944	0.555	43.484	6.5
ARIMA	**0.909**	0.475	0.552	30.836	**31.0**
SVR	1.130	0.762	0.133	29.943	**16.9**
MLP	1.105	0.670	0.127	34.479	3.2
LSTM	1.151	0.468	0.385	12.772	3.0
BEST	**0.853**	0.086	0.127	1.000	-

Table 4.22: ODSef results (RelRMSE) for 4827 store-items

	avg	std	min	max	%best
NAIVE	1.000	0.000	1.000	1.000	0.4
MA	0.823	0.102	0.380	2.799	1.4
ES	**0.783**	0.060	0.607	1.312	14.1
LINREG	6.3e+10	1.2e+12	0.370	5.4e+13	16.3
ADA	4.1e+10	7.5e+11	0.420	2.4e+13	4.5
ARIMA	0.935	7.430	0.434	501.695	**16.5**
SVR	**0.810**	0.320	0.355	14.476	**17.7**
MLP	0.879	1.989	0.314	126.260	4.9
LSTM	**0.812**	0.443	0.294	25.626	**24.3**
BEST	**0.728**	0.081	0.294	1.000	-

Table 4.23: Average RelRMSE top 3 per external factor scenario

	OWCef	ODCef	OWSef	ODSef
#1	ARIMA (0.985)	MLP (0.804)	ES (0.907)	ES (0.783)
#2	-	LSTM (0.869)	ARIMA (0.909)	SVR (0.810)
#3	-	MA (0.871)	-	LSTM (0.812)

Table 4.24: Best% top 3 per external factor scenario

	OWCef	ODCef	OWSef	ODSef
#1	ARIMA (32%)	MLP (34.9%)	ARIMA (31.0%)	LSTM (24.3%)
#2	NAIVE (17.9%)	LINREG (27.0%)	ES (19.3%)	SVR (17.7%)
#3	LINREG (17.6%)	LSTM (14.3%)	SVR (16.9%)	ARIMA (16.5%)

4.4 DFT Comparison Conclusion & Discussion

This section draws conclusions from the results of the DF technique comparison, discusses the impact of these results for practice and the scientific field, discusses the results' limitations and provides directions for future research.

Conclusion of DFT performance comparison results

When looking at the results, it becomes clear that there is no such thing as the ultimate demand forecasting technique and that performance greatly varies across items and forecasting scenarios. The tables in the previous sections give an overview of which DF technique performs best in which scenario. In general, it can be said that ARIMA performs best for the largest percentage of items for all scenarios that only use historical data. ES performs well in terms of average RelRMSE for many scenarios that only use historical data. ARIMA's average performance tends to be lower for the daily scenarios, which could be due to the fact that this data is less smooth and that there might be days with zero sales. The more complex scenarios, such as the scenarios with a daily time detail level, are also where the more advanced forecasting techniques such as neural networks show their best performance.

By far the best performance overall can be obtained by automatically selecting the best forecasting technique for each individual item(-store combination). The impact of automatic DF technique selection is greatest for forecasting scenarios with a daily time detail level. It results in an average RelRMSE improvement that ranges from 6.4% (OWC scenario) to 32.4% (M2DC scenario) compared to always using NAIVE. The average RelRMSE improvement ranges from 2.1% (M2WC scenario) to 10.1% (M2DC scenario) compared to always using the best individual DF technique.

Additionally using external factors to produce forecasts has shown to be valuable for forecasting scenarios that have a daily time detail level. The best performance was still obtained by automatically selecting the best DF technique for each item(-store combination). The external factors enabled an additional 8.6% average RelRMSE improvement in the ODCef scenario and 2.9% in the ODSef scenario compared to the original scenarios with historical data only. For scenarios with a weekly time detail level, adding external factors slightly decreased performance. Regression and neural networks performance greatly improved compared to scenarios with only historical data and for the daily scenarios some now outperformed ARIMA.

Impact for retailers in practice

Grocery retailers are recommended to ensure their DF decision support system supports a wide range of underlying quantitative forecasting techniques and that it is capable of automatically selecting the best performing one for each item(-store combination). When deciding on the quantitative techniques that should be supported, the grocery retailer can refer to the result tables in this chapter and prioritize the implementation of DF techniques using their performance results.

The automatic selection of the best DF technique for each item results in a substantial forecasting performance improvement. The exact amount of the average RelRMSE improvement depends on the forecasting problem that the particular grocery retailer faces. It ranges from 6.4% in the OWC scenario to 32.4% in the M2DC scenario compared to always using the NAIVE technique. The case of Tesco gives an impression of exactly how large the absolute cost savings can be for a big grocer. Their improved demand forecasting system saved £100 million a year.

For retailers who are not yet ready to modify their DF system itself, the results from this study can be used by forecasters to improve forecasts for a selection of individual items for which current forecasting performance is low. Forecasters can use the results of this study to decide on the most suitable forecasting technique. With some modifications, the evaluation algorithm used in this study can be transformed to a tool that can be used directly by forecasters. They could then upload the sales history of a certain product and receive a suggestion on which DF technique would perform best for that product.

Impact for the scientific field

To the best of our knowledge, this study is the first large-scale comparison of demand forecasting techniques in a food retail context. The results help researchers select a suitable DF technique for their forecasting problem. Before this paper, researchers would have been overwhelmed by the immense range of DF technique variations available. This paper discusses and explains both the most commonly used and most promising new DF forecasting techniques. Finally, the results show researchers which technique performs best in which forecasting scenario. In addition, it is the first study that examines the performance of LSTMs for food demand forecasting.

The results also showed that DF technique performance varies widely across items and scenarios. This has an impact on the research method that should be followed. The literature review showed that currently many researchers base their results on either a single product or a small selection of products. To improve the robustness of their results, we recommend researchers to base their results on as many products that match their research scope (e.g. perishables, dairy) as possible. The robustness of the results in this paper is high, since at least 908 different perishable products were used (with up to 4827 store-item combinations for scenarios with a store location detail level).

Potential limitations and future research directions

The results of the DF technique comparison in this study have some potential limitations. First of all, the performance comparison was conducted using a dataset from a single grocery retailer and performance results might be slightly different for another grocery retailer. In addition, some characteristics of country where this grocer's supermarkets were located (Ecuador) might have influenced the results for

the external factors evaluation. For example, the climate in Ecuador is fairly stable year-round, so the external factors from the weather category might prove to have more predictive power in another country where the climate has more pronounced seasons (such as The Netherlands). Future research could include data from other grocers in other countries to minimize the influence of grocer- and location-specific factors on the results. Another potential limitation is that only lag-1 historical data was used for the linear regression and neural networks techniques in this study, so when more lags are added their performance relative to other DF techniques will likely improve. Future research could add extra lags to the models used in this study, for example lag-7 and lag-31 data could be added to scenarios with a daily time detail level to specifically give the models information about sales in the prior week or month. Future research could also investigate what the optimal size is for the sales history training set. Currently, all data in the training set was indeed used for training. Even though neural networks require a lot of data to perform well, it is not necessarily the case that more historical data is better, because data that is very old might not be representative anymore of current demand patterns. In some cases, using less history to train a neural network could result in a better performance [65].

Due to the anonymized nature of the dataset, there were also some interesting aspects that could not yet be examined. When there is information available on the brand and product names, future research could include external factors such as search trends and social media mentions and sentiment for brands and products in the comparison. When there is information available on how product prices changed over time, this could be used as an extra factor in the demand forecasting models.

Future research could also examine how well neural networks perform when they are trained on the historical sales data (and external factor data) of multiple items. When they are trained on items that are sufficiently similar (e.g. from the same item class such as 'yoghurt'), this might further boost performance, since the neural networks then have more data to learn from.

Future research could also investigate the generalizability of the results to other types of retailers or to other industries. Accurate demand forecasting is also important for for example fashion retailers. Many of their clothing items are seasonal and have a zero or low salvage value at the end of the sales period, so they want to prevent ordering too many items. However, they also don't want to order too few items, since long lead times from factories in Asia make it difficult to order replenishments during the season. The energy industry is an example of another industry where demand forecasting is important, since accurate demand forecasts are required (at a sub-daily time detail level) to maintain the balance in the energy network. It is expected that the results from this study are well generalizable to other retailers or other industries facing forecasting problems with similar characteristics as the ones for grocers.

5
DF Improvement Process

The previous chapters focused primarily on demand forecasting techniques, which form the quantitative core of the demand forecasting process. To give a holistic view, this chapter provides a step-by-step process that can guide DF improvement efforts and gives an overview of other factors that can be considered during the improvement of the wider demand forecasting process. The DF improvement process is depicted in figure 5.1 and contains the following steps:

1. Current forecasting situation assessment
2. Forecasting goals determination
3. Demand forecasting techniques selection
4. DSS implementation and adoption
5. Organizational factor alignment
6. Results evaluation

Step 1 analyses the current forecasting situation (as-is), step 2 defines the organization's forecasting improvement focus (to-be), steps 3 to 5 enable the organization to go from the as-is to the to-be state and step 6 is about evaluating results and making adjustments where necessary. When aiming for continuous improvement, another iteration of the DF improvement process can be started.

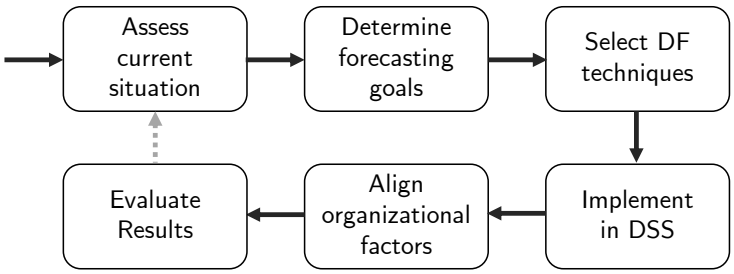

Figure 5.1: Demand forecasting improvement process

5.1 Step 1: Assess Current Situation

The goal of this assessment is to inventory what forecasts are being created in the organization, what decisions they support and what level of forecasting capabilities the organization currently has.

Existing evaluation and audit frameworks already provide useful guidelines on assessing current forecasting capabilities. Mentzer et al. [44] developed a sales forecasting management evaluation framework and benchmark. They identified four dimensions of the forecasting process, including functional integration, approach, systems and performance measurement. For each of these dimensions, four stages are described in detail so companies can benchmark their current forecasting performance. Moon et al. [46] provide a literature review of several forecasting audit frameworks and extended the framework and benchmark from Mentzer et al. [44]. They also identified some strategic themes that many of the companies they interviewed were facing. One important part of the assessment is to evaluate the suitability of the forecasting performance measures that are used. In 2008, Fildes et al. [24] reported that based on observations in practice they had little confidence that forecasting systems and companies used the appropriate forecasting performance measures.

5.2 Step 2: Determine Forecasting Goals

This step encompasses three main activities: focus definition, goal determination and target setting. In the first activity, organizations should define a focus for their demand forecasting process improvement efforts based on the assessment from the previous step. For example, they can choose to focus on improving forecasts that support replenishment decisions and define that they want to forecast on a daily instead of a weekly level. In the second activity, organizations should determine what their goals are. With the earlier example, goals could be to reduce stock-outs or reduce waste. KPIs should be defined that can be used to measure improvements for these goals. In the third activity, specific and measurable targets should be set for each of the defined KPIs. For example, this can include targets for DF technique accuracy improvements, such as a 5% RelRMSE improvement, and/or targets for specific business results, such as 3% less waste.

5.3 Step 3: Select DF Technique(s)

To select the most suitable DF technique(s) for a particular forecasting problem, the following five steps can be used as a guideline.

1. Determine DF problem dimensions
2. Determine DF evaluation criteria
3. Make a pre-selection of DF techniques
4. Get insight into performance of the DF techniques for each criterion
    - Evaluate: Evaluate the DF techniques on large set of products.
    - Benchmark: Use the results from a DF technique comparison study as a benchmark.

5. Select the appropriate DF technique(s)
    - Single: choose the DF technique with highest performance consistency.
    - Multiple: choose the desired number of DF techniques while prioritizing based on the percentage of items for which they were best.

Step 3.1: Determine DF problem dimensions

The first step is about defining the forecasting problem and determining the required forecasting problem dimensions. Determine how forecasts will be used and what decision they will support. Then the most suitable forecasting problem dimensions (in particular the horizon, time detail level and location detail level) can be selected based on this context. For example, one common situation for grocers is that forecasts are used to support the daily replenishment decisions in an individual store. In that case, the most suitable dimensions are a store location level and a daily time detail level. The most suitable horizon can depend on for example the lead time for replenishment orders. If today's orders are delivered tomorrow, the horizon is one step ahead. If orders are delivered three days later, it makes sense to use the three step ahead demand forecasts as a basis for the replenishment decisions.

Step 3.2: Determine DF evaluation criteria

The second step is about defining the evaluation criteria, which may vary for each forecasting situation and between grocers. One important criterion is of course the RelRMSE performance of the DF techniques. Additional criteria that retailers might take into account depending on the situation include the speed, interpretability, required history and implementation complexity of DF techniques. The grocer should then determine how much weight he assigns to each criterion. For example, in a situation where new products are introduced frequently, the grocer may prefer DF techniques that do not require much history. When new forecasts have to be produced very frequently and as quickly as possible, a grocer may prefer a DF technique that has high speed over a DF technique that has slightly better RelRMSE performance but is much slower.

Although the formal evaluation of the performance of the DF techniques for these additional criteria was not the focus of this study, some guidelines can be provided. The lowest amount of historical data is required by NAIVE (which uses no history at all), followed closely by MA and ES, which use a historical window that in our evaluations was and for most grocers will be below 15 observations. LINREG, ARIMA and SVR follow next. Most history is required by the neural networks techniques, MLP and LSTM, which require at least a few months of history.

A similar ranking of DF technique performance can be seen when looking at their speed, where NAIVE, MA and ES are fastest and LSTM the slowest. However, the main speed difference is caused by the fact that several techniques require training, which for for example LSTM takes very long. Training does not have to be repeated each time a forecast is needed since trained models can be stored and reused, so for the actual forecast generation the speed difference between techniques is low. The grocer can determine how often models are retrained, for example each quarter or each year. Further research can investigate the optimal frequency to retrain DF techniques. In terms of interpretability the distribution of techniques is again similar.

With for example LINREG, the outputs of the model can be easily interpreted. For example, it could be seen that the forecast is higher because of the expected temperature for a certain day. The interpretability of neural networks is much lower, since these function more like a black box. The forecaster then has no insight into why certain forecasts were generated. Machine learning researchers are working on explainable machine learning, which could be a valuable approach to improve interpretability of some forecasting techniques in the future.

Step 3.3: Make a pre-selection of DF techniques

In step three, the grocer can make a pre-selection of DF techniques. The literature review and explanations of DF techniques in this thesis greatly simplify this task. The grocer could simply choose the same subset of DF techniques that were considered in this study, or make a further selection based on his capabilities. For example, when a certain forecasting support system that the grocer uses or considers offers only a subset of DF technique as options, that range of DF techniques can be used as the pre-selection so only relevant techniques are evaluated.

Step 3.4: Get insight into DF technique performance

In step four, the grocer can either evaluate the pre-selected set of DF techniques on a large set of his own products, or examine the results from a DF technique comparison study as a benchmark. It is recommended that grocers use a benchmark, because that is by far the fastest way to gain insight. When a grocer wants to conduct his own evaluation, he has to prepare the data for a large number of products and when DF techniques are not implemented yet, they have to be implemented as a pilot. When the pre-selected set of DF techniques match that of this study, the code that was used for this study can be adjusted slightly and used as a tool.

Step 3.5: Select the appropriate DF technique(s)

In step five, the grocer evaluates the performance of all DF techniques on all selection criteria and makes a final selection of techniques that will be used in practice. When the primary selection criterion is RelRMSE performance and the grocer wants to choose a single technique, it is recommended to choose the one with the best performance consistency (lowest average RelRMSE scores). When the grocer wants to select multiple forecasting technique and automatically select the best technique for each individual product, it is recommended to choose the ones that are best for the largest percentage of items. A cost-benefit analysis can also be part of the selection at this stage. The performance increase and the corresponding cost savings resulting from a certain DF technique should generally outweigh the costs that are required for its implementation. The grocer can estimate implementation costs per technique and determine whether those costs can be justified by the performance improvement (and hence waste / stock-out reductions) that each technique generates.

5.4 Step 4: DSS Implementation and Adoption

Because DF techniques are means to improve decision making, their implementation is usually part of a decision support system (DSS), specifically a forecasting

support system (FSS). This section provides literature-based guidelines for the implementation of DF techniques in FSS and suggestions on how to stimulate end user adoption of the new system.

Typically, forecasts are created as part of a two step process, where first a quantitative forecast is created which is then adjusted judgementally by a human forecaster. Research has shown that such judgemental adjustments can be beneficial, but unfortunately in practice often unnecessary and damaging adjustments are made [13]. Two concepts that could be used to improve FSS are restrictiveness and decisional guidance [15], of which the latter seems most promising to maintain system flexibility and prevent user frustration. Decisional guidance can be either informative, for example in the form of performance feedback, or suggestive, for example by proposing the most suitable forecasting technique or parameters. Guidance can be included in 3 modes: predefined (preprogrammed), dynamic (partly responding to users behaviour) or participative (when user can indicate which factors are most important for him/her, like method speed).

Since the results of the DF technique evaluation showed that the performance of DF techniques differs greatly for different products and different forecasting problem dimensions, grocers are advised to include multiple DF techniques in the FSS. Grocers are advised to limit the restrictiveness of the FSS and instead provide human forecasters with suggestive decisional guidance on which DF technique is best for their forecasting dimensions and the specific product under consideration. It would be ideal if this suggestive guidance is participative, such that the forecaster can indicate how important the different selection criteria (such as RelRMSE performance, speed, data requirements) are to him.

When changing a system that forecasters have used for a long time, it should be made sure that they intend to adopt the new system. The intention to use a system is primarily influenced by the perceived usefulness and perceived ease of use of that system [64]. Therefore, it is important to convince forecasters of the perceived usefulness of the system by showing them that the output quality of the new system is high, that there are already demonstrable results and that it makes their jobs easier. One way to achieve this is by doing a pilot with a few stores at first, so actual results can be shown to other forecasters and the results are more tangible to them. Involving forecasters during the selection of new DF techniques and the implementation of a new DSS is also valuable, as it gives them a sense of ownership and makes the system more tangible for them. Perceived ease of use can for example be improved by organizing workshops and training sessions for the forecasters where necessary.

5.5 Step 5: Align Organizational Factors

Sales forecasting performance depends on much more than just the forecasting technique and system that is used. Multiple researchers have investigated the adoption of FSS and the factors influencing forecasting performance [14, 55, 5, 38].

The sales forecasting management (SFM) framework describes how various organizational factors like the sales forecasting climate and capabilities influence performance [14]. The core of the framework is sales forecasting capability, which is determined by information logistics (consisting of IT technology and information processes) and a shared interpretation (obtained through cross-functional communication and ownership). Forecasting support systems (FSS) and new forecasting

techniques would be part of the information logistics part of the framework and in that sense contribute to the company's sales forecasting capability. The sales forecasting climate can contribute to the company's capabilities by providing leadership support, giving credibility to sales forecasts and by aligning rewards for forecasters with the performance of their forecasts. The performance is measured both in terms of forecasting performance, but also in terms of the impact on business performance. Performance results should be used as feedback to improve the sales forecasting climate and capabilities. Fildes et al. [24] mentioned the organizational aspects of forecasting (including information systems issues) as a future research direction, since it has received limited attention in both forecasting and operations research journals.

The forecasting task-technology fit (FTTF) model [55] is a modified version of the original task-technology fit (TTF) model by Goodhue & Thompson [28]. The model shows that FSS characteristics and forecasting procedure quailty and access have a positive relation with FTTF, which in turn has a positive relation with forecasting performance [55].

Asimakopoulos et al. [5] investigated what factors influence FSS adoption and use. A literature review showed that the main concerns for companies were top management support, forecasting accuracy, effective adjustment capabilities and integration with other IT systems [5]. Building on the model of technologies-in-practice they developed an FSS adoption model, which showed that the main factors influencing adoption and use are effective communicative structures and forecasting process support.

The paper by McCarthy [38] gives an overview of criteria to evaluate sales forecasting effectiveness, of which forecasters rank accuracy and credibility as most important. Communication between different departments is also considered to be important, since the survey showed that on average 5 departments contribute information and sometimes also multiple departments jointly have responsibility.

In deciding whether to make adjustments to the demand forecasting process, companies likely also consider return on investment. An interesting question is how much value it has to increase forecasting accuracy and whether the impact of improving the forecasting process will outweigh any increase in complexity and costs. One benchmark for retailers showed that once accuracy reached 75-80% the marginal value of improved accuracy decreased and investments in flexibility could be more effective [41]. Most researchers currently only report accuracy improvements from new forecasting techniques. It could also be evaluated what implications the implementation of advanced quantitative techniques has on the retailers' forecasting process and management practices as well as the impact on business performance.

5.6 Step 6: Evaluate Results

In the final step, the results of the demand forecasting improvement efforts should be evaluated. The results for each of the KPIs that were defined in step 2 should be gathered and compared to their targets. Based on these results, further improvement options might be identified. Even when all targets are met, an organization that is striving for continuous improvement could choose to restart the DF improvement process from the start. In another iteration, for example the focus of the improvement effort can be shifted.

5.7 Conclusion and Discussion

The provided guidelines can be used by retailers to improve their wider demand forecasting process. The DF technique evaluation results in chapter 8 showed how large the impact of using the right forecasting techniques can be and this chapter enables retailers to take action on those results by guiding them not only in the DF technique selection process, but also in improving the wider forecasting process.

Future research could also conduct a survey to gain insight into the relative weights grocers give to the different DF technique selection criteria in different situations. For example in case of daily forecast supporting replenishment decisions, accuracy might be slightly less important and speed more important than for a weekly forecast, since any errors can be corrected the next day already. Future research could also include guidelines on how to asses the return on investment of adjusting the forecasting process (such as implementing new forecasting techniques).

Future research could validate the DF improvement process in practice, for example by applying it during various case studies. Future research could also examine to what extent the existing forecasting evaluation and audit frameworks are still applicable now, whether they can be applied across industries and whether extensions are required to take into account recent advances in forecasting practices.

Part II

Dynamic Pricing of Perishable Food Products

6
Dynamic Pricing Fundamentals

As Warren Buffett once said: "Price is what you pay, value is what you get". This quote illustrates that it is important to always consider customer value when pricing products. Since customer value decreases when perishable (grocery) products deteriorate, it makes sense to adjust prices dynamically over time. This chapter introduces the reader to the fundamentals of dynamic pricing. Section 6.1 describes different pricing strategies and covers how DF and DP are interrelated. Section 6.2 describes the different DP problem dimensions. Section 6.3 covers related work on dynamic pricing for perishable products.

6.1 Pricing Strategies

Sellers can adopt a wide variety of pricing strategies. On a high level, these strategies can be categorized into two main types [20]:

- *posted-price*: price is set by the seller and is a take-it-or-leave-it price
- *price-discovery*: price determined through a bidding process (e.g. auction)

Within the posted-price mechanism, prices can either be static (fixed over time) or dynamic (changing over time). The field of *dynamic pricing* (DP) focuses on finding the optimal product prices over time to maximize profit for the seller.

The height of the price is one of the factors that influence customers' demand for a product: generally the price-response function is downward sloping. Price elasticity e is defined as the fraction of demand Q change over price P change.

$$e = \frac{dQ/Q}{dP/P}$$

For example, when a 2% price increase results in a 1% demand decrease, the price elasticity is -0.5. So if demand for a product changes relatively heavily in response to price changes, that product is said to be relatively elastic. Examples of relatively inelastic products are necessity products like water and bread. The elasticity of a product can differ on the short-term versus on the long-run. For example short-term elasticity for airline travel is low because a travel need exists then, but it is

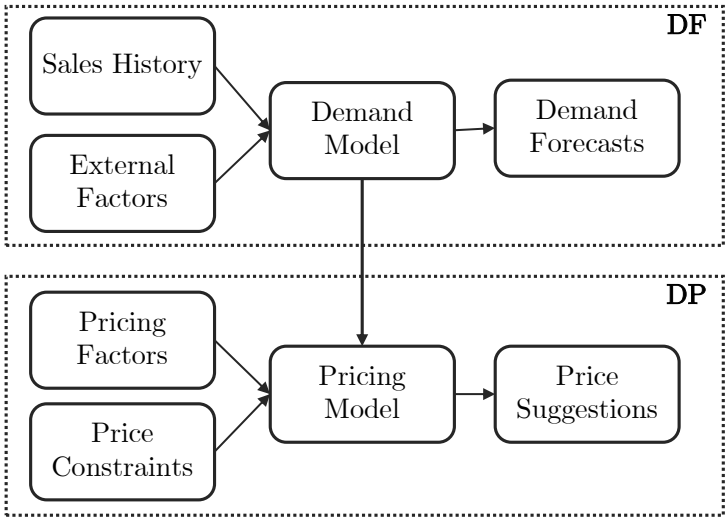

Figure 6.1: Visualization of DF, DP and their link

higher in the long-term. Setting a certain price now does not only affect current demand, but it might also influence future demand through customers' reference price, which is their perception of what they have paid for that product in the past [15]. But customers generally don't have a reference price for all items they buy, so retailers can use that to their advantage. Key value items (KVIs) are the items that mainly drive customer value perception and for these items competitive positioning is especially important and these items should be sharply priced [40].

Another pricing strategy that differs from dynamic pricing is personalized pricing. With dynamic pricing, prices change over time, but all customers will pay an equal price at the same moment in time. However, with personalized pricing, retailers use their knowledge about customers to differentiate pricing on an individual level. This differentiation can for example be based on loyalty card data, but also on channel or browsing history. So dynamic pricing and personalized pricing are two different concepts, but they can also be applied in combination.

Dynamic pricing (DP) is closely related to demand forecasting (DF) and this relation is depicted in figure 6.1. Using the DF techniques discussed in section 2.3 a demand model is derived from the sales history and external factors, which can then be used to produce a forecast for future periods. The demand model is also the relationship between DF and DP, since the pricing model that is used for DP optimization depends strongly on the demand model, which also includes a price elasticity component. In addition, the pricing model takes into account price constraints (see section 6.2) and pricing factors such as remaining inventory levels and supply uncertainty. The pricing model is then used to provide optimized suggestions for product prices over time. Dynamic pricing is a form of prescriptive analytics, since it not only predicts what will happen in terms of demand, but suggests (or even automatically performs) an action which in this case is a price adjustment.

To evaluate a pricing strategy's true performance, it is best to test it in practice.

For example in the case of e-commerce, a/b testing could be employed, to randomly serve part of the customers with the optimized price and the other part with the regular price. For a retailer with brick-and-mortar stores, a few pilot stores could be selected that operate with the optimized prices. However, testing out a pricing strategy in practice does come with quite some risks, so an interesting alternative is to do simulations. One downside of simulations is that the results depend on the validity of the underlying simulation model.

6.2 Dynamic Pricing Problem Dimensions

As mentioned earlier, dynamic pricing is about finding the optimal prices for products to maximize revenue for the seller. Multiple variations of DP problems exist depending on market/product characteristics and the modelling assumptions.

Den Boer [15] identifies two main categories of DP problems for monopolist firms (no competition). In the first category, models have dynamic demand functions, meaning that the demand function changes with changing circumstances. In the second category, demand is static (average demand is same in different periods) and pricing dynamics arise from changes in the marginal value of inventory. A typical example of the second category can be found in the airline industry, where there is a fixed seating capacity on the aircraft and ticket prices are dynamically adjusted based on the marginal value of remaining seats.

Another framework for categorizing different DP problems was proposed by Elmagraby & Keskinocak [20]. In their framework, DP problems vary along 3 dimensions:

- Replenishment vs. no replenishment of inventory (R/NR): whether the inventory is a one-time fixed supply or whether it can be replenished by ordering extra products when needed.
- Dependent vs. independent demand over time (D/I): whether sales now influence demand in the future.
- Myopic vs. strategic customers (M/S): whether customers' buying decisions are influenced by them anticipating future price levels.

For example durable product sales can be considered to be dependent, since a sale now will influence demand in the future - once a customer has bought a car it is much less likely he will buy a car again in the near future. Non-durable products are more independent, since these products are purchased repeatedly. The airline ticket example mentioned earlier can be considered as an NR-D problem, since there is a fixed capacity (NR) in the aircraft and highly dependent demand (D) as customers will not likely buy the same ticket twice.

Several guidelines have been developed that assist in determining the suitability of modelling customers as myopic [20]. These guidelines show that customers can for example be considered myopic in the context of sales of necessity items (like bread), because they need those items at a certain point in time no matter the price they anticipate for it in the future. In the airline ticket example, business travellers might be considered as myopic since they need to travel, but leisure travellers are more likely to be strategic. Each customer has a maximum price they are prepared to pay for a product, which is called the willingness-to-pay (WTP). When a myopic customer has demand for a certain product, he will buy it right away if the product

price is lower than his WTP. However, a strategic customer might not yet buy it in anticipation of a price that is even lower.

A company can impose a wide variety of constraints to the prices that can be set for products. Bitran & Caldentey [7] listed the most common constraints in their literature review.

- Finite set of prices: when only certain prices are allowed, for example only prices ending in .95 or .99.
- Maximum number of price changes: when prices can be changed at most X times over the selling horizon of the product.
- Markups, markdowns, promotions: when a predefined path for price changes is common in the industry, for example markups for airlines, markdowns for fashion retailers and promotions for grocers.
- Joint pricing: when different products can not be priced independently.
- Cost-based pricing: when pricing is based on unit cost of the product and a fixed minimum margin.

For retailers, it is imaginable that constraints could also vary per product or per product category. While optimizing prices for products, all such constraints would have to be taken into account. Some retailers also offer product bundles, where certain items are sold together as one package and the items either are not available separately or are relatively expensive.

6.3 Related Work

This section discusses several related studies that also consider the problem of dynamic pricing for perishable products within the retail industry. For a more exhaustive review of dynamic pricing with inventory considerations in general the reader is referred to Elmaghraby & Keskinocak [20].

Chatwin [10] describes the optimal dynamic pricing for perishable products with stochastic demand and a finite set of allowed prices when the product can not be replenished. They find that for a given inventory level, the optimal price declines as the product's expiration date approaches. For a given remaining shelf life, they found that the optimal price decreases as the remaining inventory level is higher.

Caro and Gallien [9] describe the development and implementation of a dynamic markdown system at Zara, a large fashion retailer. Their problem context has some similarities with ours, in the sense that fashion stores sell perishable products that decrease in value once they 'expire'. One difference between the pricing problem in a fashion and grocery context is that fashion items might still have a salvage value at the end of the sales period (e.g. through outlet sales) while food items become completely worthless. In addition, all individual items of a certain fashion product type have the same 'expiration date', whereas this differs for food products depending on when they were produced. So in a grocery store, items of the same product type with different expiry dates might be on sale at the same time.

Chung and Li [12] considered dynamic pricing for perishable food products and did take into account that products on the shelf have a different remaining shelf life. They investigated the impact of discount frequency for perishable food products

on grocer profit and waste percentage in simulations. Customers made purchases based on their consumption need, meaning that they only considered products for purchase that had at least a minimum number of shelf life days remaining. This minimum number requirement varied between customers and was assumed to be normally distributed with the mean being half the products' maximum shelf life. From the remaining products customers then select the one with the lowest price and longest remaining shelf life. The total discount percentage was fixed at 20%, only the application of this discount varied between pricing strategies. The single-price case applied no discount at all and the two-price case applied the 20% at once. Two multi-period price strategies were considered, namely one that gradually (linearly) applied the discount by changing the price each day and one that changed the price each 2 days. The grocer in their simulation used an order-up-to ordering policy. They found that two-price strategies resulted in better performance than single-price strategies. The multi-price strategies turned out to be more effective in reducing waste and increased profitability when demand could be forecast accurately. Several of the assumptions that were made can be challenged. Customers generally go grocery shopping for that day or for that week, so it seems more likely that their need distribution is at least positively skewed. Because of the way customer behaviour is modelled and because WTP is not considered, the probability that a customer will buy a product with a certain shelf life does not vary with the price.

Lu et al. [36] also evaluated pricing strategies for perishable food products and evaluated how the height of the optimal price changes over time and with the age of the products. They do not consider replenishment, only provide guidelines on how the direction of the optimal price should change and do not provide insight into the impact discounting strategies have on waste percentages. Adenso-Diaz et al. [1] evaluated the performance of a dynamic pricing strategy on perishable food product waste and revenue. They did not take into account replenishments. They found that dynamic pricing strategies resulted in waste reductions, but that the effect on revenue depended strongly on the scenario. Herbon et al. [30] focus on how customer satisfaction and data from RFID tags that measure product quality can be taken into account in a dynamic pricing strategy.

7
Research Method

The research questions for the dynamic pricing part of the thesis are as follows:

DP1. What (dynamic) pricing strategies can be used for the sale of perishable food products in supermarkets?

DP2. What simulation model can be used to simulate perishable food product sales?

DP3. In simulations, which pricing strategy performs best in terms of total revenue, waste and stock-outs?

Question DP1 and DP2 will be answered based on a literature review.This will reveal any related work on dynamic pricing of perishable food products and provide suggestions for feasible pricing strategies. In addition, the literature review will be used to decide on the model components and the necessary assumptions for a simulation. To answer research question DP3, a simulation model will be implemented in Python and run with different pricing strategies to investigate the impact of those strategies on revenue, waste and stock-outs and to determine which performs best.

7.1 Simulation Method

A system is 'a collection of entities that act and interact together towards the accomplishment of some logical end' [34]. A system can be studied by experimenting with the actual system or with a model of the system. In our case, no real-life price experiments can be conducted since that would have a direct financial impact on grocers. Therefore it is suitable to construct a model of the system and because of the high complexity of the system it is studied through simulation. This will help grocers to gain insight into which pricing strategy performs best without having to experiment with prices in-store. The simulation model in this study is dynamic, stochastic and discrete. Hence the simulation represents the system as it evolves over time, the model has probabilistic components and the state of the model can only change at separated points in time [34]. Two main types of simulation time management strategies can be distinguished, being next event time advance (NETA) and fixed increment time advance (FITA) [35]. The simulation in this study follows the FITA strategy, with the fixed increment being 1 day, so each simulation period consists of 1 day.

It is important that the simulation model is valid and credible [35]. For validity, the simulation has to be an accurate representation of the real-life system. For credibility, domain decision-makers have to accept the model as 'correct'. Law [34] defined a 10-step process for conducting a simulation study, which will be followed for this study.

1. Formulate problem and plan the study
2. Collect data, define model and construct assumptions document
3. Check validity of assumptions
4. Construct computer program and verify
5. Make pilot runs
6. Check validity of programmed model
7. Design experiments
8. Make production runs
9. Analyse output data
10. Document, present and use results

The remaining sections in this chapter describe several of these steps in more detail. The first step is covered in section 7.2. Step two and three are covered in section 7.3, which describes the simulation model components, and section 7.4, which describes the simulation model assumptions. The validity of the assumptions was checked (step three) by reviewing literature and by discussing the assumptions multiple times with a simulation subject matter expert. Section 7.5 describes what happens during one day of the simulation. The simulation model was developed in Python (step four). To test the validity of the programmed model, each simulation model element was extensively tested and pilot runs were conducted to test the simulation model as a whole (step five and six). Section 7.7 describes the experiment design and the goals for each experiment (step seven). Then production runs were conducted for each of the experiments and the results from those runs were analyzed (step eight and nine). The results are presented in chapter 8 (step 10).

7.2 Problem Formulation

We consider a monopolist grocer selling a single perishable product with a fixed shelf life. The product can be replenished daily, based on the demand forecast and a safety factor. The grocer wants to evaluate which pricing strategy performs best in terms of total revenue, waste and stock-outs. In terms of the dynamic pricing problem dimensions provided by Elmagraby & Keskinocak [20], this research problem is an R-I-M dynamic pricing problem. Important to note is that in our case the expiry dates of the replenished products differ from the products that were already in the store. This is different from for example fashion, where products are also perishable, but their expiry date is the end of the season. When replenishments are ordered during the fashion season, their expiry date is also the end of the season. Since the grocer sells a single product only, substitution or complementariness effects with other products are not (explicitly) taken into account. Examples of other effects that influence customer behaviour which were also not taken into account include the reference price effect, where customers would adjust there willingness to pay based

on the prices they recently observed, and the presentation effect, where customer demand would be influenced by the amount of remaining inventory.

7.2.1 Pricing Strategies to Evaluate

Four main pricing strategies for perishable products will be evaluated:

PS1. Fixed price, no price change at all as items deteriorate.

PS2. Single fixed price change, with $D\%$ fixed discount on last day before expiration.

PS3. Multiple fixed price changes, with $D\%$ fixed discount spread linearly over the last S days before expiration.

PS4. Single dynamic price change, with a dynamically determined discount on the last day before expiration.

The selected pricing strategies represent a balanced mix of strategies that are already used by grocers in practice and promising new strategies that were identified in the literature review. For the new strategies, an additional inclusion criterion was that they had to be simple enough to be applied in practice.

PS1 and variations of PS2 are currently being applied by Dutch grocers in practice. Jumbo applies PS2 with a 100% discount, meaning that they give products away for free on the last day before expiration. Albert Heijn also applies PS2, but with a 35% discount on the day of expiration. Some supermarkets don't change their price at all and hence apply PS1. PS3 and PS4 were included based on the literature review from section 6.3. PS3 was included because the results from similar simulations by Chung and Li [12] indicated that gradually applying the fixed discount, so spreading it over the last few days before expiry, could improve performance. PS4 was included because dynamically determining the discount based on the inventory levels and expected demand in each period is more flexible than applying a fixed discount percentage in all periods and intuitively this extra flexibility might result in improved performance. PS4 determines the optimal discount for the products that have one day left until expiry by estimating the revenue for that set of products for all discounts in a set of allowed discounts. It relies upon demand forecasts for next period to estimate revenue and chooses the discount that provides the highest expected revenue.

7.2.2 Performance Measures

The performance of these pricing strategies will be evaluated based on several key performance indicators (KPIs). These KPIs include the total revenue, the percentage of inventory wasted and the percentage of customers that wanted to buy a product but faced stock-outs. PS2 to PS4 will be benchmarked against the default pricing strategy, which is PS1. For the comparison, the percentages change in each of the KPIs will be calculated relative to PS1.

$$\text{Stock-outs (\%)} = \frac{\#\text{ Stock-outs}}{\#\text{ Sales} + \#\text{ Stock-outs}} * 100$$

$$\text{Waste (\%)} = \frac{\#\text{ Products Wasted}}{\#\text{ Sales} + \#\text{ Products Wasted}} * 100$$

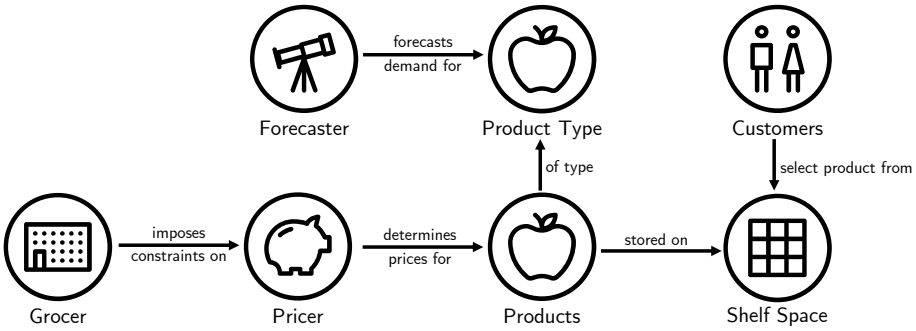

Figure 7.1: Simplified overview of simulation model components

Before systematically evaluating the different strategies, some hypotheses can already be drafted when comparing PS2-4 to PS1. It is expected that PS2 with very high discount percentages result lowest waste by far, since customers like (almost) free products. However, one could argue about how valuable this waste reduction is, since customers might then pick products solely because they were (almost) free without really wanting or needing them, thereby increasing the risk of waste at the customers' homes. A clear downside of PS2 with high discounts is that even though the products are 'sold', the impact on the profit is the same as when they would have been wasted because they generate (almost) zero revenue. That is why PS2 with lower discounts seems more attractive, since then more revenue will be generated for the products that are sold and customers are still attracted with a discount. It is expected that spreading out the discount such as with PS3 will result in higher revenues, as this was suggested in previous research [12]. PS4 is expected to result in the highest total revenues, since it has the ability to give discounts only when necessary (e.g. when excesses are very large),

At first a truly dynamic pricing strategy with multiple price changes over the products' lifetime (with a maximum of 1 price change per day) was discussed as well, however since that is difficult to implement at brick-and-mortar grocery stores it was left out of the research scope for now. With a truly dynamic price strategy, it no longer seems feasible to let an employee change the product price sign or apply new discount stickers on a daily basis.

7.3 Simulation Model Components

The simulation model for perishable food sales consists of several components, which are described at a high level in this section. Some components can be seen as an agent, which is "a computer system that is situated in some environment and that is capable of some autonomous action in this environment in order to meet its delegated objectives" [68]. The agent components are: the customers, grocer, forecaster and pricer. In addition, there are several components that are essential in modelling the system, but which do not act as agents: the product type, the products and shelf space. Figure 7.1 visualizes the relationships between components.

Customers

The customers represent the visitors in the grocery store who might buy a product. Each customer has slightly different characteristics. There are two main types of customers: regular customers and date-checking customers. Regular customers purchase based on product price only, whereas date-checking customers purchase based on price and quality. Date-checking customers pay attention to the remaining shelf life of the product and take that into consideration when choosing a product that satisfies their objective of getting the best value-for-money.

Grocer

The grocer imposes constraints on the pricing agent, such as which discounts are allowed. In all simulations in this study, the dynamic pricing strategies could only use discounts between 0.05 and 1.0 with increments of 0.05. The grocer also has to decide on a safety factor for ordering replenishments of products on the shelves. The objective of the grocer is to minimize waste and maximize revenues.

Pricer

The pricing agent is in control of setting prices for the items in the store. It follows a certain price strategy and optimally adjusts prices accordingly where necessary. With the dynamic pricing strategies, the pricing agent uses demand forecasts from the forecaster to estimate the customers' demand for a certain item. It then determines the optimal price for that item to minimize waste and maximize revenue while taking into account the constraints that were imposed by the grocer.

Forecaster

The forecaster predicts future demand for a certain product type. At the start of the simulation, it automatically determines the best performing forecasting technique. Forecasts are used to support replenishment and pricing decisions.

Product Type

This component represents the product type (SKU). A sales history is available for each product type. In addition, each product type has certain characteristics, such as a maximum shelf life (in days) and a base price.

Product

This component represents the individual products that are available for purchase in the grocery store. Each product has an age, potentially a discount and a current price. In addition, a product can spoil (when it is older than the maximum shelf life) or be sold.

Shelf Space

This component models the shelf space where the products are stored in the grocery store. At the end of each simulation period replenishments are ordered when demand for the next period is expected to exceed the remaining inventory. The order amount

is based on the demand forecast minus the remaining inventory, multiplied with a safety factor that was specified by the grocer.

7.4 Model Assumptions

Several assumptions were made for the development of a simulation model for perishable food product sales. To ensure the simulation represents the real-life situation as best as possible and that validity and credibility are maintained, these assumptions were based on scientific literature or analysis of real-life data whenever possible.

7.4.1 Customer Arrival Rate

The customer arrival rate determines how often customers will visit the store. It is assumed that the arrival rate follows a dynamic Poisson distribution, which is a common assumption in the dynamic pricing literature [20]. The Poisson distribution is dynamic in the sense that it may differ over time during the simulation, depending on the day of the week. Following the central limit theorem, this Poisson distribution can be approximated with a normal distribution.

Assumption 1: Customer arrival can be simulated using a dynamic Poisson distribution which differs per weekday.

To increase the validity of the simulation model, it is important that the number of sales per day follows similar patterns as sales in real life. During the simulation, the distributions that are used for the number of buyers per day are based on the available sales history for each item from the Favorita dataset by doing kernel density estimation (KDE) on each weekday's sales history. To illustrate what these distributions look like, figure 7.2 shows the normal probability distributions for the total number of transactions in one of Favorita's stores obtained through KDE. This clearly shows that the number of transactions differs per weekday. In addition to simulations with this generated sales data, there will also be runs with the actual history of individual products from the Favorita dataset.

There is an infinite customer population with resampling, meaning that the demand distributions do not change based on purchases that are made. This infinite population assumption is suitable in this case, since we consider non-durable products [57]. In addition, it is assumed that demand is independent from the remaining inventory levels, so the 'presentation effect' is not taken into account.

Assumption 2: Demand distributions do not change based on purchases made.

Assumption 3: Demand is independent from inventory levels.

Assumptions also have to be made about how overall demand changes when the price of the product changes. Unfortunately, no price data is contained in the Favorita dataset, so this could not be determined by analysing the data. However, several economists have already conducted research on price elasticity of food demand for several food categories. As a basis for our simulation, the average food price elasticity for perishable food categories will be used that resulted from a literature review on food price elasticity [3] and a recent Canadian study [2].

Assumption 4: Total demand responds to price changes based on the price elasticity of demand, which can be set as a simulation setting.

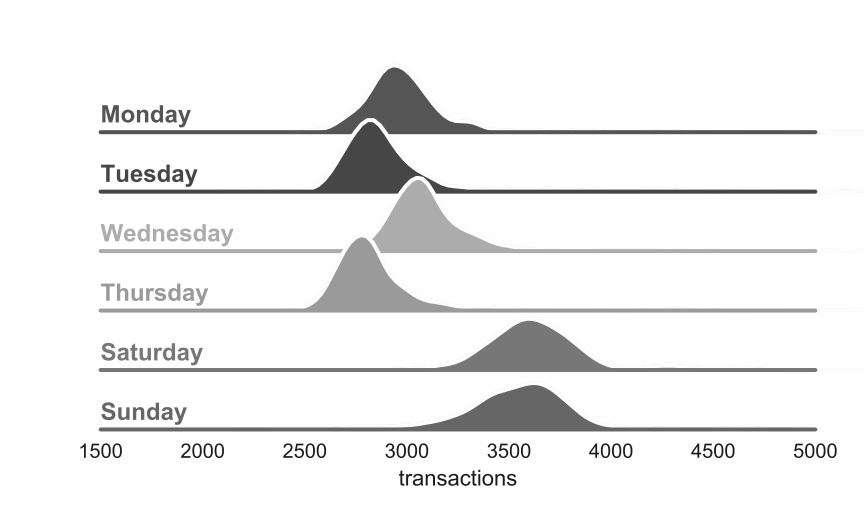

Figure 7.2: Estimated normal probability distributions for total daily sales at one of Favorita's stores

7.4.2 Customer Behaviour Assumptions

Several assumptions have to be made about customers' behaviour and purchase decisions. Firstly, we assume that customers are myopic (as opposed to strategic). Several guidelines have been developed that assist in determining the suitability of modelling customers as myopic [20]. These guidelines show that customers can for example be considered myopic in the context of sales of necessity items, because they need those items at a certain point in time no matter the price they anticipate for it in the future. A large portion of fresh food items are necessity products, such as milk and bread, so the assumption can be justified. So we assume that customers won't postpone their purchase in anticipation of a future lower price.

Assumption 5: Customers are myopic, so they do not strategically postpone purchases in anticipation of a future lower price.

There are two main types of customers: date-checking customers and regular customers. Whether a customer is date-checking is determined from a binomial distribution with a success probability that can be adjusted as a simulation setting.

Assumption 6: The probability of customers being regular customers follows a binomial distribution with a success probability that can be set as a simulation setting.

Each customer has a certain willingness-to-pay (WTP), which is the price they are willing to pay for a product with its entire shelf life remaining. It is assumed that the initial willingness-to-pay for all customers is equal to the product's base price (1.0 by default). Another assumption is that date-checking customers' WTP declines linearly once the product gets closer to its expiry date, because they check expiry dates and are willing to pay less for products that have less remaining shelf life. Customers will only consider products for purchase that have a current price below their current WTP for that product.

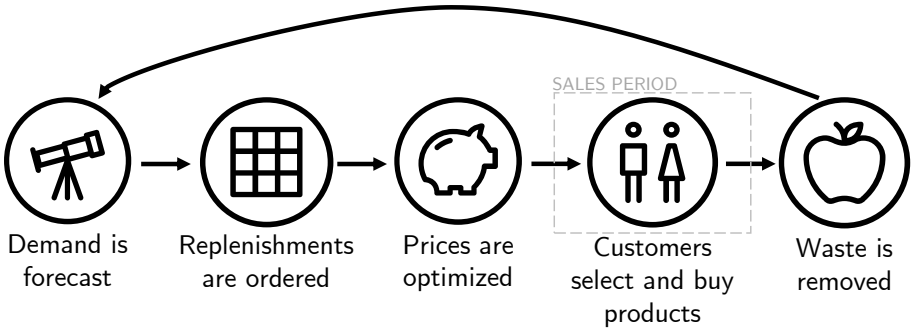

Figure 7.3: Simplified overview of steps per simulation period

Assumption 7: Customers' WTP is initially equal to the products' base price.

Assumption 8: Date-checking customers' WTP declines linearly as a product ages and gets closer to its expiry date.

Assumption 9: Customers only consider a product for purchase when its current price is below their WTP for that product.

Regular customers randomly select a product from the shelf. Regular customers randomly select one of the discounted items with a probability that is at least random, or if it is higher with a probability equal to the discount fraction. So if the discount is 50% the probability that a regular customer purchases one of the discounted products is 50% as well (as long as there are at least two products in the store, otherwise the random probability of picking that product is higher). Date-checking customers are utility maximizers, meaning that they show some semi-strategic behaviour where they will choose the product that gives them the best value-for-money.

Assumption 10: Regular customers randomly pick a product from the shelf, or with a probability equal to the discount percentage if that probability is higher.

Assumption 11: Date-checking customers pick the product that has best value-for-money (highest WTP surplus).

7.5 Walkthrough of a single simulation period

This section walks through one simulation period to help readers understand how all components and assumptions are connected. Figure 7.3 shows a simplified overview of the steps in each simulation period.

At the start of the simulation period, a demand forecast is made for the next sales period and if the current inventory is not expected to be able to meet that demand, a replenishment is ordered based on the difference between the demand forecast and the current inventory level (multiplied by an optional safety factor). Then, the pricer will optimize prices where necessary following the set pricing strategy. Next, the sales period starts and the actual number of customers for that period is determined. That number is either sampled from the dynamic Poisson distributions

(based on the KDEs) or taken directly from the real sales history for that product. When the sales period starts, the customers enter the store. If there is no product available that is below their WTP, customers leave the store immediately without buying anything. When there are affordable products available, it is then determined whether that buyer is a regular buyer or a date-checking buyer. Regular buyers then first decide whether to go for a discounted or for a regular priced product. They can only choose to buy discounted products if they are available. Regular buyers at least randomly select a product from either the discounted or non-discounted set of products. Date-checking buyers select the product that has the highest value-for money out of all 'affordable' products. If no products are available anymore, the stock-outs count goes up by one. Near the end of the simulation period (when all buyers have left the store), the items on the shelf deteriorate (their age goes up by one day) and products that have now past their expiry date go to waste and are removed from the shelf. Afterwards, the next simulation period starts.

7.6 Simulation Settings

Several simulation experiments will be conducted and each simulation experiment requires multiple simulation runs with slightly different settings. The list below gives an overview of the settings that can be easily adjusted before each run. The bold acronyms will also be used in the results tables in chapter 8 and appendix C.

- **#PER**: Number of simulation periods in this run.
- **PS**: Pricing strategy to evaluate during this run.
  - **D**: fixed discount percentage (only used in case of PS2 or PS3).
  - **S**: spread (only used in case of PS3).
- **LOCL**: Location level, which can be country-level ('C') or store-level ('S').
- Product type, which has several options as well:
  - **BP**: Base price
  - **SL**: Maximum shelf life
  - **HT**: Sales history type, data can be real ('R') or generated ('G')
  - **SF**: Order safety factor
- **RP**: Probability of customers being regular customers (as opposed to date-checking).
- **EL**: Price elasticity of total demand.

The default number of maximum simulation periods was set to 155 days (roughly 5 months). Because the simulation is stochastic, each run was repeated 5 times and for 5 different products to make the results more robust and to limit the influence of randomness on results.

Table 7.1: Default Simulation Settings

#REP	#PROD	#PER	LOCL	BP	SL	HT	SF	RP	EL
5	5	155	S	1.0	3	R	1.5	0.4	-0.6

7.7 Experiments and Goals

Several simulation experiments will be conducted, which each have their own specific goals. The first experiment uses default settings and the others each investigate the influence of one simulation setting on the results. This will give insight into the situations for which dynamic pricing strategies are most valuable. It also serves as a sensitivity analysis, which shows how strongly the results of this simulation depend on a certain assumption. If one assumption turns out to strongly influence results, it means that it would be valuable to further validate this assumption in future research. The default settings for all experiments are provided in table 7.1. When a setting is not default, it is mentioned explicitly an an experiment description.

Experiment 1: Default Pricing Strategy Comparison

This experiment uses the default values (see table 7.1) for all simulation settings and then evaluates the performance of all pricing strategies. The default regular probability was based on the results from Tsiros and Heilman [62], since on average 60% of customers indicated that they (almost) always checked expiry dates. The default elasticity was based on results from two food price elasticity studies [3, 2] and was set equal to the average food price elasticity of all perishable food type categories.

Experiment 2: Influence of Product Shelf Life

This experiment varies the maximum shelf life (SL) setting to investigate whether pricing strategy performance differs for products with shorter or longer shelf lives. The maximum shelf lives for which runs will be conducted range from 2 to 10 days. It is expected that the impact of the dynamic pricing strategies is greatest for products with a shorter shelf life, since those are products that spoil quickly and where there is the least time to let any excess inventory resolve naturally or to compensate with lower replenishment orders in future periods.

Experiment 3: Influence of Regular Customer Probability

This experiment varies the regular customer probability (RP) setting, which determines the chance that an individual customer is a regular customer (as opposed to date-checking). The probabilities for which runs will be conducted are: 0.0, 0.25, 0.5, 0.75 and 1.0. So if RP $= 0.75$ and 100 customers enter the store, on average 75 of them will be regular customers and 25 date-checking customers. It is expected that waste will increase when a larger percentage of customers is date-checking. In addition, it is expected that waste will increase more strongly for fixed pricing strategies such as PS1/2 compared to PS3/4. This experiment will also provide insight into the sensitivity of the results with regards to assumption 7 from section 7.4. It is important to vary the percentage of date-checking customers, because

Tsiros and Heilman [62] showed that depending on the perishable product type, between 29% and 93% of customers say they usually or always check the expiration date of a product before they buy it.

Experiment 4: Influence of Safety Factor

This experiment varies the replenishment order safety factor (SF) setting. The safety factors for which runs will be conducted are: 0.8, 1.0, 1.2, 1.5 and 1.7. Varying the safety factor has no direct influence on how prices are set, but it does influence the replenishment order amounts, which are determined by multiplying the demand forecast for the next period with the safety factor. The expectation is that low safety factors on average result in a higher number of stock-outs (since demand is consistently underestimated) and hence in a lower chance of an inventory excess and of products going to waste. So the impact of the dynamic pricing strategies is expected to increase as the safety factor is increased.

Experiment 5: Influence of Price Elasticity of Demand

This experiment varies the price elasticity (EL) setting. The elasticities for which runs will be conducted are: -0.2, -0.4, -0.6, -0.8, -1.0 and -1.2. It is expected that pricing strategies that give discounts have a stronger negative impact on total revenue when the demand in the simulation is more inelastic (so closer to 0). In relatively elastic simulations, part of the discount will be compensated for by more customers being attracted to the product.

Experiment 6: Influence of Sales History Variations

This experiment consists of two sub-experiments. Experiment 6A varies the location level of the product history, which can be at a country-level or a store-level. The country-level sales history time series is expected to be smoother than at store-level and the absolute number of customers will be larger. Experiment 6B varies the type of history data, which can be generated data or real data directly from the Favorita dataset. The generated data is expected to be smoother than the real data, since data is generated from a normal distribution and chances that there are outliers (which may occur on for example holidays) are small. It is expected that there is no significant difference in results for both sales history variations. Experiment 6B will also provide insight into the sensitivity of the results with regards to assumption 1 from section 7.4.

8
Simulation Results

This chapter gives an overview of the simulation results and the performance of different pricing strategies in the different experiments. Performance was measured in terms of the impact on revenue, waste and stock-outs. Tables with detailed simulation results for each experiment can be found in appendix C.

8.1 Results Experiment 1: Default Settings

Table 8.1 provides an overview of results for the first experiment with the default simulation settings. These results show that PS2 achieved a waste reduction of up to 9.2%, but at a great cost to revenue of up to 35.4%. The fixed pricing strategy that achieved the highest waste reduction per percent revenue reduction was PS2 with D = 0.2, which resulted in a 6.9% waste reduction and 3.7% revenue reduction. Following a dynamic pricing strategy (PS4) achieved a waste reduction of 3.8% and reduces revenue by 2.0%. These results show that a dynamic pricing strategy can achieve a similar waste reduction, while having a relatively lower negative impact on revenues compared to the fixed pricing strategies.

When evaluating the strategy that is currently applied by Dutch grocer Albert Heijn (PS2 with D = 0.35) it might be interesting for them to switch to PS2 with D = 0.2, since in simulations this achieved a similar waste reduction (just 0.6% less waste reduction) at a much lower cost to revenue (4.3% less revenue reduction). Dutch grocer Jumbo currently applies PS2 with D = 1.0, but they do not add discount stickers to their products. It is up to the customer to notice that the product is almost expired and then ask to get the product for free. Not all customers will take this effort, so it is expected that the negative revenue impact in practice is much lower for Jumbo than during the simulation (where the discount was advertised).

The results for PS3 were not included in table 8.1, but can be found in the full results tables in appendix C. PS3, which spreads out the fixed discount over multiple days, in this case the last 2 days before expiration, always performed much worse than PS2 and PS4. It achieved the lowest waste reduction results at a high cost to revenue. The relatively higher costs to revenue can be explained by the fact that more products were discounted at an earlier stage, while these discounts may not have been necessary.

All strategies only had a minor influence on stock-outs (up to a 0.9% reduction).

Table 8.1: Experiment 1 results overview

PS	D	Δ%Revenue	Δ%Waste	Δ%Stock-outs
1	0.0	-	-	-
2	0.1	-0.8	-0.6	0.0
2	0.2	-3.7	-6.9	-0.3
2	0.3	-6.4	-7.3	-0.5
2	0.35	-8.0	-7.5	-0.5
2	0.4	-9.9	-8.0	-0.6
2	0.5	-13.3	-8.2	-0.7
2	0.6	-17.2	-8.5	-0.8
2	0.7	-21.3	-8.7	-0.9
2	0.8	-25.5	-8.9	-0.9
2	0.9	-30.3	-9.0	-0.9
2	1.0	-35.4	-9.2	-0.9
4	-	-2.0	-3.8	-0.3

The stock-out percentage is mainly influenced by the ordering policy the grocer has, which will also be discussed in experiment 5. For all other experiments, the stock-out changes are no longer discussed, since the results were very similar to what was found for this experiment.

Overall, all discounting strategies resulted in a waste decrease, but they also all had a negative impact on total revenue in this experiment. It means that purely from an operational revenue perspective, it would not be beneficial for grocers to discount products that approach their expiry dates. However, it will likely still be beneficial from a profit perspective, since costs will be incurred when products go to waste which were not included in this simulation.

8.2 Results Experiment 2: Shelf Life Variations

This section discusses how the impact of the different pricing strategies varied for products with different maximum shelf lives (ranging from 2 days to 11 days).

Products with a shorter shelf life

For extremely fresh products with a shelf life of just 2 days the fixed price strategy with the highest waste reduction per percent revenue reduction was PS2 with D = 0.3. It resulted in a 0.6% waste reduction and decreased total revenue with 16.8%. Virtually no additional waste reduction was achieved with higher fixed discounts and it only had a further negative impact on revenue. A dynamic pricing strategy performed much better, achieving a 8.3% waste reduction while decreasing overall revenue with 7.1%.

Products with a longer shelf life

The impact of applying discounts to older products decreased gradually for products with a longer shelf life. For example, for a product with a maximum shelf life of 7 days, the best performing strategy was PS2 with D = 0.5, which resulted in a 0.1%

revenue decrease and a 1.1% waste decrease. For a product with a maximum shelf life of 11 days, the maximum revenue decrease was 0.5% and the maximum waste decrease 0.2%.

Implications

The results show that applying discounts had the largest impact on waste reduction for products with shorter shelf lives. These results can also be explained intuitively because when the shelf life is longer, there is more time to naturally resolve any excess inventory, for example by compensating for it by ordering less replenishments in future periods. Because of these results, grocers are advised to initially focus their discounting efforts on products with shelf lives between 2 and 7 days, since that is where they can have the largest impact. Dynamic pricing strategies also have more added value compared to fixed pricing strategies in scenarios where products have shorter shelf lives.

8.3 Results Experiment 3: Regular Customer Probability Variations

The results both in terms of the impact of pricing strategies on revenue and waste differ quite strongly depending on the regular customer probability. This section will primarily discuss the two extremes in detail.

Only Date-Checking Customers

When all customers are date-checking (RP = 0.0) applying PS2 with D = 0.2 results in a 5.6% revenue increase and a 12.1% waste decrease. Applying a fixed 35% discount (AH strategy) results in a 12.3% waste decrease, but only increases total revenue with 0.8%. In this scenario, if grocers want to apply a fixed discount strategy they are advised to apply a fixed 20% discount. The dynamic pricing strategy (PS4) performs slightly worse than a fixed strategy, since PS4 achieves a 5.9% waste decrease and a 0.1% revenue increase.

Only Regular Customers

When all customers are regular customers (RP = 1.0), applying PS2 with D = 0.2 results in a 3.8% revenue reduction and a 1.7% waste reduction. In this case a dynamic strategy (PS4) performs better with a 1.5% revenue reduction and a 2.1% waste reduction. When all customers are regular customers, it is never beneficial for retailers from a revenue-perspective to apply discounts to products that approach their expiration dates.

Mix of Customer Types

When 90% of customers in the store were date-checking (RP = 0.1), applying PS2 with D = 0.2 resulted in a 1.7% revenue increase and a 10.5% waste decrease. Applying higher fixed discounts resulted in a slightly higher waste decrease (of up to 12.5%), but at a greater cost to revenue (of up to 31.3%). Therefore, based on the simulation results, if retailers want to choose a fixed discounting strategy, 20% is recommended. PS4 resulted in a 1.3% revenue decrease and a 5.8% waste

decrease. When more than 10% of customers were regular customers (RP $>$ 0.1), applying a discounting strategy to products that approach their expiration dates always had a negative impact on total revenue.

Implications

These results show that discounting products always results in waste decreases, but that the extent of that waste decrease as well as the impact on revenue depends quite strongly on the probability that customers are regular. The results show that it is more beneficial for grocers to apply a discount to products that approach their expiry date when a larger percentage of customers is date-checking. The impact of waste was higher and the negative impact on revenue lower (or even positive) when more customers were date-checking. Grocers are advised to focus their discounting efforts on products where a large portion of customers checks the expiry dates. An example of such a product can be milk, since Tsiros and Heilman [62] showed that for milk, 93% of customers always or usually check the expiry date before purchasing a product. Future research could further investigate for which items customers are very likely to check expiry dates to help grocers focus their discounting efforts.

The more customers are regular, the better the dynamic pricing strategy performs relative to pricing strategy PS2. When at least 40-50% of customers in the simulation were regular customers, the dynamic pricing strategy (PS4) increasingly outperformed PS2 in terms of percentage waste reduction per percentage revenue reduction. This improvement in the relative performance of PS4 can be largely explained by the fact that when more customers are regular, PS2 more frequently discounts unnecessarily. This shows that grocers would benefit from being more flexible in the pricing strategies they apply. In this case, grocers would be advised to apply a fixed 20% discount percentage to items for which less than 40% of customers is regular and applying a dynamic pricing strategy for items with more regular customers.

8.4 Results Experiment 4: Elasticity Variations

The results from experiment 4 show that when price elasticity of demand is higher, applying discounts to products that approach their expiration date is more effective. When the product is more elastic, discounting results in a less negative impact on revenue and larger waste decrease percentages. For example, PS2 with D $=$ 0.2 results in a 6.5% waste decrease and a 5.9% revenue decrease when elasticity is -0.002 compared to a 7.2% waste decrease and a 0.1% revenue decrease when elasticity is -0.012.

Implications

In general, results show that applying discounts to products that approach their expiration date is more effective when price elasticity is higher. This can also be explained intuitively: when the product is more elastic, the impact of discounting on revenue is less negative because increased overall demand makes up for part of the discount and the impact on waste reduction is stronger. It is recommended that retailers initially focus their discounting efforts on products with a relatively high price elasticity of demand.

8.5 Results Experiment 5: Safety Factor Variations

As mentioned earlier, when deciding on how much inventory to order and what safety factor to use, there is a trade-off between waste and stock-outs. A low safety factor underestimates demand and results in a high number of stock-outs and a low waste percentage. A high safety factor overestimates demand and results in a low number of stock-outs and a high waste percentage. Figure 8.1 shows the influence of safety factor variations for the waste and stock-out levels with the baseline pricing strategy, PS1. Similar trends can be seen for all other pricing strategies. Although optimizing the grocers' ordering policy is not the focus of this study, it was shortly examined how different ordering policies could influence pricing strategy performance.

Lower safety factor

When the safety factor was 0.8 the overall waste percentage with PS1 was already very low (0.8%). Therefore the initial expectations were that there was so little room for improvement that discounting strategies would have no real added value. However, it was interesting to see that a fixed pricing strategy can be valuable to grocers even in this case and even result in an overall revenue increase. The best pricing strategy in this case was PS2 with D=0.2, which resulted in a 0.6% revenue increase and a 0.7% waste reduction.

Higher safety factor

When the safety factor is high, there was a relatively large amount of waste and discounting strategies could have a larger impact. The dynamic pricing strategy also outperformed the best fixed price strategy when the safety factor was at least 1.4 (and hence the initial waste percentage was at least 8%).

Implications

These results show that even for products that already have a low waste percentage, grocers should use a discounting strategy because it has a positive impact on revenue and reduces waste. However, the waste reduction percentages will be much higher for products where waste was initially high. When the initial waste is higher, the results also showed that dynamic pricing strategies outperform fixed pricing strategies.

8.6 Results Experiment 6: Sales History Variations

The input data history type and location level did have some influence on the results. However, it does not likely influence the main findings of this study because the default settings used real data at a store-level, which will also be the most common situation in practice when making pricing decisions.

Experiment 6A: History Type Variations

When generated data was used instead of true data, the waste reduction percentages were similar (up to 0.4% more waste reduction), but the negative impact on revenue was much higher (up to 8.4% higher).

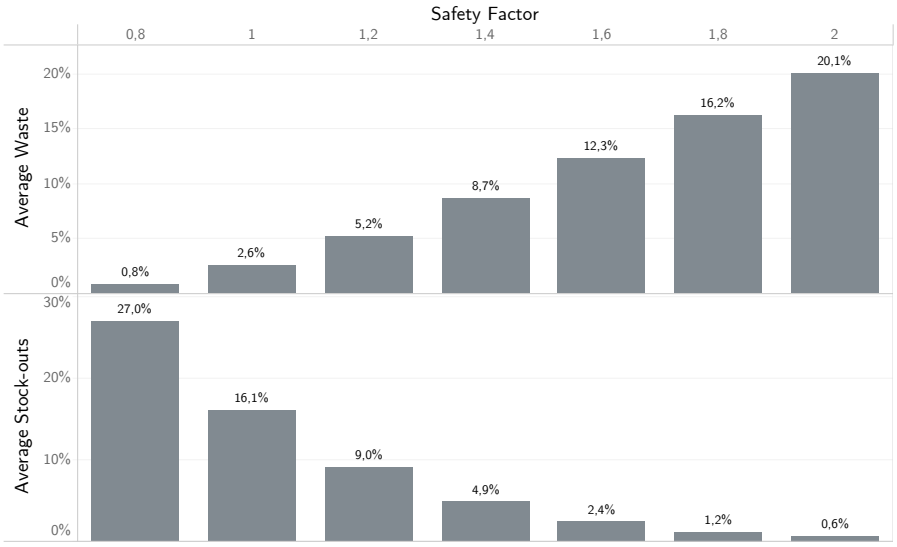

Figure 8.1: Influence of safety factors on waste and stock-outs

Experiment 6B: Location Level Variations

When true data was used with a country location level instead of a store location level, there was less waste reduction (up to 2.2% less) and less revenue decrease (up to 7.8% less).

Implications

It is advised to use real data at the store location level in simulations wherever possible, because this more accurately reflects the situation in practice. When this is not possible due to data constraints, or when data needs to be augmented with generated data, it is important to adjust the final results accordingly. For example, when generated data needs to be used, the impact on revenue will likely be less severe than the simulation results show.

8.7 DP Simulation Conclusion & Discussion

Conclusion of DP simulation results

In general, the results showed that pricing strategies which discounted products that approach their expiration dates always achieved a waste reduction, but sometimes at a great cost to revenue. Discounting was most beneficial in simulations where a product was more elastic and/or where more customers were date-checking, because that resulted in higher waste reductions and less negative (or even positive) changes to revenue at the same time. Discounting strategies were most effective in terms of waste reduction when products had a shorter shelf life, more customers were date-checking, price elasticity of demand was higher and when initial waste percentages for that product were higher.

Overall, the fixed pricing strategy that most frequently delivered the best results in simulations was PS2 with a fixed discount of 20% on the last day before expiry. Spreading out the fixed discount over multiple days (PS3) before expiry always resulted in a more negative impact on revenue and a lower reduction of waste. Interestingly, PS2 frequently outperformed the dynamic pricing strategy (PS4). This could be due to the fact that the dynamic pricing strategy relied upon demand forecasts that were not always accurate. In addition, the forecaster in the simulation only had access to NAIVE, MA, ES, LINREG and SVR forecasts (out of which it automatically chose the best) to limit the time needed to run the simulations. When the simulations will be run with more advanced and (as we saw in the previous part of this thesis) more accurate DF techniques available to the forecaster, the performance of PS4 relative to PS2 will likely improve. A dynamic pricing strategy already outperformed a fixed strategy when initial waste levels for a product were high or when a large percentage of customers were regular customers.

Impact for retailers in practice

Food retailers can use these results as a benchmark to determine what impact a certain pricing strategy would have on their revenue and waste. So whereas the direction of change in revenue or waste could sometimes be predicted intuitively, this study has quantified the actual impact of different pricing strategies in simulations. Food retailers are advised to determine which of the products in their assortment are most elastic and for which products most customers check dates. They should then initially focus their discounting efforts on those products, since in those cases a high waste reduction can be achieved at the relatively lowest cost to revenue (or even slightly increase revenue). When a grocer wants to apply a fixed pricing strategy, they are advised to use a fixed discount of 20% on the last day before expiry, since that was the fixed pricing strategy that most frequently achieved the best results in simulations. Retailers are advised to use a dynamic pricing strategy when initial waste percentages are high or when a large percentage of customers are regular customers.

Impact for the scientific field

As the review of related work showed, there are not many existing studies that examine the impact of dynamic pricing for perishable food products. The simulation in this study is different from the simulations conducted in previous work in several ways. For example, contrary to Chung and Li [12], customers in our simulation decide on purchases based on their quality-adjusted WTP and the percentage of customers that is date-checking will be varied. For the fixed pricing strategies, different discounts are tested and for the dynamic strategies the percentage discount can be different each time. Contrary to what Chung and Li [12] found, spreading a fixed discount over multiple days was not beneficial in this simulation. Spreading discounts might still be beneficial when the discount is spread over more than two days (e.g. linearly over the entire product's shelf life) or when discounts are spread not linearly but according to a different pattern, since that was not tested in this simulation. This study is also shows that dynamic pricing strategies for discounting products are not always better than dynamic pricing strategies, likely because this study uses a simulation model which more closely resembles the situation in practice in that dynamic pricing algorithms rely on imperfect demand forecasts.

Potential limitations and future research directions

One potential limitation is that some of the assumptions used in the simulation model were not validated in practice nor included in the sensitivity analysis. For example, the way regular customers or date-checking customers picked products was not included in the sensitivity analysis and might influence overall results. In addition, while the sensitivity analysis studied the influence of changing one simulation setting on results, it did not investigate how changes in multiple simulation settings at the same time might have influenced results. Future research could further validate each of the assumptions, for example by doing customer surveys, observing customer behaviour in practice, or by gathering a more extensive historic sales dataset. For example, if a historic sales dataset is available that also includes product prices over time, the true price sensitivity of customers per product can be estimated and directly used in simulations.

Another potential limitation is that the simulation was now run for only 5 products and although these products were selected randomly, it might be possible that this relatively small sample influenced the final results. Future research could run the simulation for more products (e.g. all 986 items in the Favorita dataset instead of 5 randomly selected ones) and with more repetitions per product (e.g. 10 instead of 5), to also enable significance testing of results.

The time detail level of the simulations in this study was 1 day and future research could also consider using a more detailed time detail level, for example 1 hour. This could provide retailers with insight into not only on which day, but at which time of the day discounts should be applied, which is especially useful for products with an extremely short shelf life (e.g. 1 or 2 days).

Future research could also expand the scope of these simulations. For example, the grocer in the simulation could be selling more than one product, so substitution effects can be investigated. Another option is to include more types of customers, in particular in terms of how they select products from the shelf. Some examples of additional customer types are customers that always pick the first product on the shelf, customers that are more likely to buy products on the front parts of the shelf and customers that always pick the product with the longest remaining shelf life. Or additional customer behaviour concepts could be included, such as the presentation effect and the reference price effect. In addition, the simulation could include information on the price the grocer pays to buy products and the costs incurred when products go to waste, so the impact of different pricing strategies on overall profit can be examined.

In addition, future research could evaluate even more advanced pricing strategies, where prices are changed more often during the product lifetime. It might then be important to also include menu costs in the model, which are the costs incurred per price change (e.g. labour costs for sticking a discount sticker to a product). Since the best performing pricing strategies in our simulation only changed the prices once, these menu costs will not influence the main results from this study (although relative performance of PS3 might further decrease, as it changed prices twice).

Future research could also model how customer behaviour differs for different product categories. In particular, it might be interesting to distinguish high perceived quality risk products from low perceived quality risk products in future versions of the simulation model. When customers perceive products as having a high quality risk, their willingness-to-pay deteriorates exponentially instead of linearly [62].

Part III

Conclusion and Discussion

9
Conclusion and Discussion

This chapter summarizes the main findings from this study by answering the research questions. In addition, the contributions to retail practice and the scientific field are summarized, the validity of the results is discussed and several categories for future research are identified.

9.1 Results Summary

This section summarizes the main results of this study by answering each of the research sub questions that were defined in chapter 1.

DF1: Which quantitative demand forecasting technique performs best in forecasting perishable product demand?

The results showed that there is no such thing as the ultimate demand forecasting technique. The performance of DF techniques depends strongly on the product for which the forecast is created and the forecasting scenario. The forecasting scenarios that were considered in this study varied in terms of the horizon (one or multi-step ahead forecasts), time detail level (daily or weekly) and location detail level (store or country level). For each of these scenarios, the performance of 9 DF techniques was evaluated on a dataset of 986 perishable items from Favorita Corporacion, a food retailer in Ecuador. Chapter 4 discusses all results in detail and provides an overview of the top 3 best performing DF techniques for each forecasting scenario. More advanced forecasting techniques, such as neural network variations, show their best performance in forecasting scenarios with high complexity, such as scenarios with a daily time detail level.

By far the best performance overall can be achieved by automatically selecting the best forecasting technique for each individual item. The impact of this automatic selection is greatest for forecasting scenarios with a daily time detail level and results forecasting performance improvement of up to 32% compared to always using a naive forecast (using the sales from the previous period as the forecast for next period) and up to 10.1% compared to always using the best individual DF technique. For more details on the exact performance improvement in each forecasting scenarios, the reader is referred to chapter 4.

It was also investigated how much added value it has to include external factors in the demand forecasting process. The external factors that were considered in

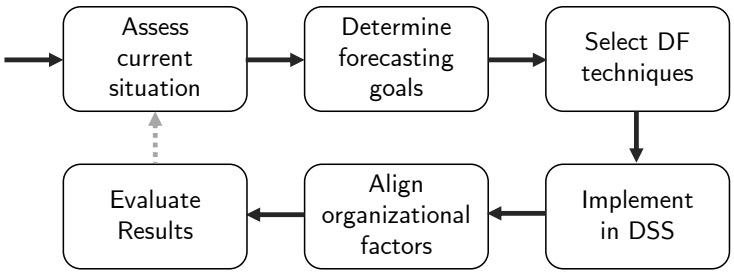

Figure 9.1: Demand forecasting improvement process (copy of figure 5.1)

this study included factors from the weather, economic, events and promotions categories. Using these external factors has shown to be valuable for forecasting scenarios that have a daily time detail level and it enabled an additional forecasting performance improvement of up to 9%.

DF2: What guidelines can be provided to improve the wider forecasting process in practice?

While research sub question DF1 primarily focused on the quantitative core of the demand forecasting process, this sub question also takes into account other factors that may influence demand forecasting performance, such as the implementation of DF techniques as part of a decision support system (DSS) and the alignment of organizational factors. A DF improvement process was developed that can be used by retailers to guide their improvement efforts. Figure 9.1 shows an overview of this process. Each of these steps were described in detail in chapter 5. For the selection of DF techniques (step 3) a separate sub-process with 5 steps was defined, which helps retailers decide which DF techniques are most suitable in their forecasting situation.This sub-process for DF technique selection is also described in detail in chapter 5.

DP1: What (dynamic) pricing strategies can be used for the sale of perishable food products in supermarkets?

Several pricing strategies were identified through a literature review and by examining what strategies are currently applied in practice by grocery retailers. The four pricing strategies that were considered in this study were:

PS1. Fixed price, no price change at all as items deteriorate.

PS2. Single fixed price change, with D% fixed discount on last day before expiration.

PS3. Multiple fixed price changes, with D% fixed discount spread linearly over the last S days before expiration.

PS4. Single dynamic price change, with a dynamically determined discount on the last day before expiration.

The first pricing strategy is used as a baseline and the changes in total revenue, waste percentages and stock-out percentages for the other pricing strategies are all

calculated using this baseline. The second and third pricing strategies always apply a fixed discount percentage that is determined in advance, but strategy 2 applies that discount on the final day before expiry, while strategy 3 gradually applies that discount over the last 2 days before expiry. The fourth strategy estimates the optimal discount percentage dynamically.

DP2: What simulation model can be used to simulate perishable food product sales in supermarkets?

The simulation model components and assumptions were described respectively in sections 7.3 and 7.4. On a high level, the simulation considers a monopolist grocer selling a single product with a fixed maximum shelf life. The inventory can be replenished during the simulation, meaning that products in the store have different ages. Customers are considered to be myopic, meaning that they won't postpone their purchase in anticipation of a lower price in the future. There are two types of customers: one is a regular customer and the other is a date-checking customer who buys the product that gives the best value-for-money. At the end of each simulation day, the products that have reached their maximum shelf life and are not sold are turned into waste. In addition, a demand forecast is created for the next day that is used to decide on inventory replenishments. Finally, before the next simulation day starts, prices are optimized according to the specified pricing strategy and any constraints that the grocer may have imposed (e.g. only a certain set of discounts ia allowed).

DP3: In simulations, which pricing strategy performs best in terms of total revenue, waste and stock-outs?

The main performance differences between pricing strategies strategies showed in terms of their impact on total revenue and waste. The change in stock-out percentages was minimal for all strategies (no more than 1%). In general, all pricing strategies that applied a discount to products that approached their expiry dates resulted in waste reductions (of up to 13%), but sometimes at a great cost to revenue. Different experiments were conducted to determine which simulation settings had the strongest influence on the performance of different pricing strategies.

The fixed pricing strategy that most frequently achieved the best results was applying a 20% discount on the last day before expiry (PS2 with $D = 0.2$). Spreading out the discount over multiple days (PS3) always achieved a lower waste reduction at a higher cost to revenue. In addition, the best fixed pricing strategy (PS2) most frequently outperformed a dynamic pricing strategy (DP).

Discounting was most beneficial in simulations where a product was more elastic and/or where more customers were date-checking, because that resulted in higher waste reductions and less negative (or even positive) changes to revenue at the same time. Discounting strategies were most effective in terms of waste reduction when products had a shorter shelf life, more customers were date-checking, price elasticity of demand was higher and when initial waste percentages for that product were higher. A dynamic pricing strategy increasingly outperformed a fixed pricing strategy when initial waste levels for a product were high or when a large percentage of customers were regular customers. For a more detailed discussion of results for each of the experiments, the reader is referred to chapter 8.

How can enhanced demand forecasting and dynamic pricing contribute to reducing food waste at the retailer level?

In summary, results from this study showed that both improving the demand forecasting process and discounting perishable products that approach their expiry date can achieve high waste reductions. Grocers are advised to initially focus on improving their demand forecasting process, since this will reduce the frequency and severity of inventory excesses. Solving the majority of the waste problem from the demand forecasting side is most ideal for grocers, since discounting strategies frequently have a negative impact on revenue. Grocers can follow the demand forecasting improvement process and the demand forecasting technique selection process to guide their improvement efforts. Once the demand forecasts are reasonably accurate, the grocer can apply discounting strategies to reduce any inventory excesses that still occur. The grocer can use the simulation results as a benchmark to decide which pricing strategy is most optimal in his situation and he can use the guidelines to focus his discounting efforts on the set of products where discounting is likely to have the largest waste reduction impact at the lowest cost to revenue.

9.2 Contributions

This study makes the following contributions to both retail practice and the scientific fields of demand forecasting and dynamic pricing.

First of all, it provides a robust performance comparison of commonly used and most promising new demand forecasting techniques based on large set of perishable products from a real-life dataset. To the best of our knowledge, no such comparison is currently available in a retail context and both retailers and researchers can use this study as a benchmark to select the best performing DF technique for their forecasting problem.

In addition, results from this study showed that automatically selecting the best DF technique for each individual product resulted in by far the best performance overall compared to always using one single forecasting technique for all products. Manually selecting the best DF technique for each product in a food retailer's assortment would consume so much time that it is virtually impossible. As part of this study, an algorithm was developed that automatically selects the best forecasting technique for each individual product and determines the optimal (hyper)parameters for each technique, allowing retailers and researchers to more easily use a wide variety of DF techniques.

This study did not only show how to improve the performance of the quantitative core of the demand forecasting process, but it also provided a process to guide retailers and researchers in their wider demand forecasting improvement efforts. This process also considers other factors that influence demand forecasting performance, such as DSS implementation, adoption and the alignment of organizational factors.

For the dynamic pricing part of the study, a robust performance comparison was conducted for different pricing strategies that mark down perishable products that approach their expiry date, giving food retailers insight into the impact of discounting on their total revenue, waste and stock-outs. In addition, this study developed a simulation model for the sales of perishable products in grocery stores that can be easily reused in future research and a simulation tool where assumptions can be easily varied.

9.3 Validity

This section discusses the validity of results for both the demand forecasting and the dynamic pricing part of this study, both in terms of internal and external validity.

There is little reason to doubt the validity of the DF technique evaluation results. Internal validity was ensured in several ways. First of all, DF techniques were evaluated on a large sample of 986 perishable food products (and 5 stores), which limited the influence that one single product (or store) had on the overall results. Finally, performance was compared across multiple forecasting problem dimensions, a robust performance measure was chosen (RelRMSE) and the code that was used for the evaluations was extensively tested before doing the final evaluations. The external validity (generalizability) of results is high within the context of perishable food retail. The set of demand forecasting techniques that was used represents both commonly used and promising new DF techniques and they are evaluated across multiple forecasting scenarios that are all also used by retailers in practice. When the performance impact of automatically selecting the best forecasting technique for each item is discussed, it is compared against always using the naive forecast and always using the best individual forecast. The external validity of results might be impacted negatively by the fact that the dataset that was used originates from Favorita in Ecuador. However, generalizability to other grocers in general is considered high, also when their product mix differs from that from Favorita, because DF technique performance was shown not to differ across heavily across product categories. However, the results for including the external factors might change for food retailers in other regions of the world, as for example the climate there might vary more heavily throughout the year. Results might even also be generalizable to other retail sectors, or other industries, but that should be confirmed through future research. The validity of the DF improvement process was not formally tested yet (for example in case studies), but it does already provide retailers with useful guidelines on how to approach DF improvement and enables them to take action on the results from this study.

We are also confident in the validity of the DP simulation results. Although the simulation inevitably required a lot of assumptions, the majority was based on previous research. In addition, a sensitivity analysis was conducted through multiple experiments to evaluate how the impact of different pricing strategies changes when some of those simulation settings change. Therefore, we have confidence in the internal validity of the results from the DP simulations, but it can be further improved by validating more of the simulation assumptions in practice. Results are generalizable for retailers facing the same pricing problem dimensions (R-I-M) for their perishable products, so when replenishment is possible, demand is independent and customers are myopic. The external validity of results might be negatively impacted by the fact that some customer behaviour concepts were not explicitly included in the simulation. For example substitution may influence results in practice: discounting one product (e.g. milk of brand 1) might lead to cannibalization of demand for another product (e.g. milk of brand 2). Results might not be directly generalizable to practice for high perceived quality risk products, since for those products customer willingness to pay declines exponentially (instead of linearly) as the expiration date nears. The generalizability of results to practice can be improved by widening the scope of the simulation to include more of these customer behaviour concepts. Results are not generalizable to other retailers or industries when they

have different pricing problem dimensions, for example when replenishment is not possible, or when customers are strategic (e.g. they postpone purchases in anticipation of future lower prices), since in that case different pricing strategies are likely to be more optimal.

9.4 Suggestions for Future Work

This section shortly summarizes the main directions for future research that were provided throughout this thesis. More detailed future research directions for the DF part were provided in sections 4.4 and 5.7, for the DP part in section section 8.7.

First of all, for the demand forecasting part of the research the scope of the evaluation could be expanded by including more (variations of) DF techniques. In addition, some of the existing DF techniques can be further improved, for example by adding extra lags or adding extra external factors. In addition, the generalizability of results could be investigated in future research, by including data from more supermarkets in more geographic regions or even by doing a similar evaluation on data from other types of retailers or industries. Future research could also evaluate the DF improvement process and the DF selection process in case studies. For the dynamic pricing part of the research, the existing assumptions underlying the simulations could be validated in practice in future research, for example through customer surveys or by observing behaviour in stores. In addition, the scope of the simulation could be expanded, for example by including extra customer behaviour concepts such as substitution.

Bibliography

[1] Adenso-díaz, B., Lozano, S., and Palacio, A. Effects of dynamic pricing of perishable products on revenue and waste. *Applied Mathematical Modelling 45* (2017), 148–164.

[2] Agriculture and Agri-Food Canada (AAFC). The Estimation of Food Demand Elasticities in Canada, 2007.

[3] Andreyeva, T., Long, M. W., and Brownell, K. D. The Impact of Food Prices on Consumption: A Systematic Review of Research on the Price Elasticity of Demand for Food. *American Journal of Public Health 100*, 2 (2010), 216–222.

[4] Aras, S., Kocakoç, p. D., and Polat, C. Comparative study on retail sales forecasting between single and combination methods. *Journal of Business Economics and Management 18*, 5 (2017), 803–832.

[5] Asimakopoulos, S., and Dix, A. Forecasting support systems technologies-in-practice : A model of adoption and use for product forecasting. *International Journal of Forecasting 29*, 2 (2013), 322–336.

[6] Basak, D., Pal, S., and Patranabis, D. C. Support Vector Regression. *Neural Information Processing 11*, 10 (2007), 203–224.

[7] Bitran, G., and Caldentey, R. An Overview of Pricing Models for Revenue Management. *Manufacturing & Service Operations Management 5*, 3 (2003), 203–229.

[8] Brockwell, P. J., and Davis, R. A. *Introduction to Time Series and Forecasting*. Springer, 2016.

[9] Caro, F., and Gallien, J. Clearance Pricing Optimization for a Fast-Fashion Retailer. *Operations Research 60*, 6 (2012), 1404–1422.

[10] Chatwin, R. E. Optimal dynamic pricing of perishable products with stochastic demand and a finite set of prices. *European Journal of Operational Research 125* (2000), 149–174.

[11] Chen, F. L., and Ou, T. Y. Sales forecasting system based on Gray extreme learning machine with Taguchi method in retail industry. *Expert Systems With Applications 38* (2011), 1336–1345.

[12] Chung, J., and Li, D. A simulation of the impacts of dynamic price management for perishable foods on retailer performance in the presence of need-driven purchasing consumers. *The Journal of the Operational Research Society 65*, 8 (2014), 1177–1188.

[13] Cross, R. G., Higbie, J. A., and Cross, Z. N. Milestones in the application of analytical pricing and revenue management. *Journal of Revenue and Pricing Management 10*, 1 (2010), 8–18.

[14] Davis, D. F., and Mentzer, J. T. Organizational factors in sales forecasting management. *International Journal of Forecasting 23* (2007), 475–495.

[15] den Boer, A. V. Dynamic pricing and learning: Historical origins, current research, and new directions. *Surveys in Operations Research and Management Science 20*, 1 (2015), 1–18.

[16] Doganis, P., Alexandridis, A., Patrinos, P., and Sarimveis, H. Time series sales forecasting for short shelf-life food products based on artificial neural networks and evolutionary computing. *Journal of Food Engineering 75* (2006), 196–204.

[17] Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A., and Vapnik, V. Support Vector Regression Machines. *Proceedings of the 9th Conference on Advances in Neural Information Processing Systems* (1997).

[18] ECR Europe, and Roland Berger. Optimal Shelf Availability, 2003.

[19] Efendigil, T., Önüt, S., and Kahraman, C. A decision support system for demand forecasting with artificial neural networks and neuro-fuzzy models: A comparative analysis. *Expert Systems with Applications* (2008).

[20] Elmagraby, W., and Keskinocak, P. Dynamic pricing in the presence of inventory considerations: Research overview, current practices and future directions. *Management Science 49*, 10 (2003), 1287–1309.

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

[22] Ferreira, K. J., Lee, B. H. A., Simchi-levi, D., Hong, B., Lee, A., and Simchi-levi, D. Analytics for an Online Retailer: Demand Forecasting and Price Optimization. *Manufacturing & Service Operations Management 18*, 1 (2016), 69–88.

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

[24] Fildes, R., Nikolopoulos, K., Crone, S. F., and Syntetos, A. A. Forecasting and operational research: a review. *The Journal of the Operational Research Society 59*, 9 (2008), 1150–1172.

[25] Food and Agriculture Organization of the United Nations. Global food losses and food waste, 2011.

[26] Freund, Y., and Schapire, R. E. A Short Introduction to Boosting. *Journal of the Japanese Society for Artificial Intelligence 14*, 5 (1999), 771–780.

[27] Geva, T., Oestreicher-singer, G., Efron, N., and Shimshoni, Y. Using Forum and Search Data for Sales Prediction of High-Involvement Projects. *MISQ 41*, 1 (2017), 65–82.

[28] Goodhue, D. L., and Thompson, R. L. Task-Technology Fit and Individual Performance. *MISQ 19*, 2 (1995), 213–236.

[29] Harvey, L. Using data analytics to reduce food waste. `https://blogs.mathworks.com/headlines/2017/03/28/using-data-analytics-to-reduce-food-waste/`, March 2018. Web article, accessed 05-04-2018.

[30] Herbon, A., Levner, E., and Cheng, T. C. E. Perishable inventory management with dynamic pricing using time temperature indicators linked to automatic detecting devices. *International Journal of Production Economics 147* (2014), 605–613.

[31] Hochreiter, S., and Schmidhuber, J. Long Short-Term Memory. *Neural Computation 9*, 8 (1997), 1735–1780.

[32] Hyndman, R. J., and Koehler, A. B. Another look at measures of forecast accuracy. *International Journal of Forecasting 22* (2006), 679–688.

[33] Kayser, V., and Blind, K. Extending the knowledge base of foresight: The contribution of text mining. *Technological Forecasting and Social Change* (2016).

[34] Law, A. M. *Simulation modeling and analysis*, 4th ed. ed. McGraw-Hill, Boston, MA, 2007.

[35] Law, A. M. How to build valid and credible simulation models. *Proceedings of the 2009 Winter Simulation Conference* (2009), 24–33.

[36] Lu, J., Zhang, J., Lu, F., and Tang, W. Optimal pricing on an age-specific inventory system for perishable items. *Operational Research* (2017).

[37] Makridakis, S., and Hibon, M. The M3-Competition: results, conclusions and implications. *International Journal of Forecasting 16* (2000), 451–476.

[38] McCarthy, T. M., Davis, D. F., Golicic, S. L., and Mentzer, J. T. The Evolution of Sales Forecasting Management: A 20-Year Longitudinal Study of Forecasting Practices. *Journal of Forecasting 324* (2006), 303–324.

[39] McKinsey & Company. A fresh take on food retailing, 2013.

[40] McKinsey & Company. Pricing in retail: Setting strategy, 2015.

[41] McKinsey & Company. My supply chain is better than yours - or is it?, 2017.

[42] Mellers, B., Stone, E., Murray, T., Minster, A., Rohrbaugh, N., Bishop, M., Chen, E., Baker, J., Hou, Y., Horowitz, M., Ungar, L., and Tetlock, P. Identifying and Cultivating Superforecasters as a Method of Improving Probabilistic Predictions. *Perspectives on Psychological Science 10*, 3 (2015), 267–281.

[43] Mena, C., Adenso-Diaz, B., and Yurt, O. The causes of food waste in the supplier-retailer interface: Evidences from the UK and Spain. *Resources, Conservation & Recycling 55* (2011), 648–658.

[44] Mentzer, J. T., Bienstock, C. C., and Kahn, K. B. Benchmarking Sales Forecasting Management. *Business Horizons*, May-June (1999), 48–56.

[45] Mitrea, C. A., Lee, C. K. M., and Wu, Z. A Comparison between Neural Networks and Traditional Forecasting Methods: A Case Study. *International Journal of Engineering Busienss Management 1*, 2 (2009), 19–24.

[46] Moon, M. A., Mentzer, J. T., and Smith, C. D. Conducting a sales forecasting audit. *International Journal of Forecasting 19* (2003), 5–25.

[47] Oliver Wyman. A retailer's recipe: fresher food and far less shrink, 2014.

[48] Papargyropoulou, E., Lozano, R., Steinberger, J. K., Wright, N., and bin Ujang, Z. The food waste hierarchy as a framework for the management of food surplus and food waste. *Journal of Cleaner Production 76* (2014), 106–115.

[49] Phaal, R., Farrukh, C. J. P., and Probert, D. R. Technology roadmapping A planning framework for evolution and revolution. *Technological Forecasting and Social Change 71* (2004), 5–26.

[50] Qi, M., and Zhang, G. P. An investigation of model selection criteria for neural network time series forecasting. *European Journal of Operational Research*, 132 (2001), 666–680.

[51] Qu, T., Zhang, J. H., Chan, F. T. S., Srivastava, R. S., Tiwari, M. K., and Park, W.-y. Demand prediction and price optimization for semi-luxury supermarket segment. *Computers & Industrial Engineering 113* (2017), 91–102.

[52] Ramos, P., Santos, N., and Rebelo, R. Performance of state space and ARIMA models for consumer retail sales forecasting. *Robotics and Computer Integrated Manufacturing 34* (2015), 151–163.

[53] Ren, S., Chan, H.-l., and Ram, P. A Comparative Study on Fashion Demand Forecasting Models with Multiple Sources of Uncertainty. *Annals of Operation Research 257* (2017), 335–355.

[54] Sivanandam, N., and Ahrens, D. A hybrid seasonal autoregressive integrated moving average and quantile regression for daily food sales forecasting. *Intern. Journal of Production Economics 170* (2015), 321–335.

[55] Smith, C. D., and Mentzer, J. T. Forecasting task-technology fit : The influence of individuals , systems and procedures on forecast performance. *International Journal of Forecasting 26*, 1 (2010), 144–161.

[56] Sun, Z.-l., Choi, T.-m., Au, K.-f., and Yu, Y. Sales forecasting using extreme learning machine with applications in fashion retailing. *Decision Support Systems 46* (2008), 411–419.

[57] Talluri, K. T., and Van Ryzin, G. J. *The Theory and Practice of Revenue Management*. Springer, 2006.

[58] Tashman, L. J. Out-of-sample tests of forecasting accuracy: an analysis and review. *International Journal of Forecasting 16* (2000), 437–450.

[59] Taylor, J. W. Forecasting daily supermarket sales using exponentially weighted quantile regression. *European Journal of Operational Research 178* (2007), 154–167.

[60] Teschner, F., and Weinhardt, C. A macroeconomic forecasting market. *Journal of Business Economics 85* (2015), 293–317.

[61] Thomassey, S. Sales forecasts in clothing industry: The key success factor of the supply chain management. *International Journal of Production Economics 128* (2010), 470–483.

[62] Tsiros, M., and Heilman, C. M. The Effect of Expiration Dates and Perceived Risk on Purchasing Behavior in Grocery Store Perishable. *Journal of Marketing 69*, April (2005), 114–129.

[63] United Nations. Transforming our world: the 2030 agenda for sustainable development, 2015.

[64] Venkatesh, V., and Davis, F. A Theoretical Extension of the Technology Acceptance Model: Four Longitudinal Field Studies. *Management Science 46*, 2 (2000), 186–204.

[65] Walczak, S. An empirical analysis of data requirements for financial forecasting with neural networks. *Journal of Management Information Systems 17*, 4 (2001), 203–222.

[66] Wijnhoven, F., and Plant, O. Sentiment Analysis and Google Trends Data for Predicting Car Sales. *ICIS 2017 Proceedings* (2017), 1–16.

[67] Wong, W. K., and Guo, Z. X. A hybrid intelligent model for medium-term sales forecasting in fashion retail supply chains using extreme learning machine and harmony search algorithm. *International Journal of Production Economics 128* (2010), 614–624.

[68] Wooldridge, M. *An Introduction to Multi-Agent Systems*, 2nd ed. Wiley, 2009.

[69] Zhang, G., Patuwo, B. E., and Hu, M. Y. Forecasting with artificial neural networks: The state of the art. *International Journal of Forecasting 14* (1998), 35–62.

Appendices

A
DFT Implementation Specifics

A.1 Package Use

Several existing Python packages were used in the implementation of the demand forecasting techniques. Functionality from the `pandas` package was sufficient for the basic DFTs, where `shift()` was used for NAIVE, `rolling()` for MA and `ewm()` for ES. For ARIMA the `SARIMAX` class from the package `statsmodels` was used (without using the seasonal or extra factor components). The package `sklearn` was used for linear regression (`LinearRegressor` class), AdaBoost (`AdaBoostRegressor` class), support vector regression (`SVR` class) and for the multi-layer perceptron (`MLPRegressor` class). The packages `keras` and `tensorflow` were used to build the LSTM networks.

B
DFT Evaluation Significance Tests Results

This chapter contains figures that show the results from the Wilcoxon signed rank tests for each scenario. This statistical test was most suitable in our case, since the data is continuous (RelRMSE scores), two paired samples have to be compared each time (two DFTs evaluated on the same set of items) and because the Wilcoxon test is non-parametric, so it does not assume that data is normally distributed like the paired t-test does.

Each figure contains a matrix with for each set of two DFTs the p-values that resulted from the Wilcoxon signed rank test. To help guide the interpretation of the diagrams, they are coloured based on their p-values. The darkest shade green indicates a p-value < 0.001, medium green a p-value < 0.01 and light green a p-value < 0.05. When a field is red, it means that no significant difference between the two RelRMSE scores distributions could be found.

For the one-step ahead scenarios, figures B.1 to B.4 show the resulting p-values from the Wilcoxon signed-rank tests. For the two-step ahead scenarios, figures B.5 to B.8 show the resulting p-values from the Wilcoxon signed-rank tests. For the two- to seven-step ahead scenarios, figures B.9 to B.13 show the resulting p-values from the Wilcoxon signed-rank tests. For the one-step ahead scenarios with external factors, figures B.14 to B.17 show the resulting p-values from the Wilcoxon signed rank tests.

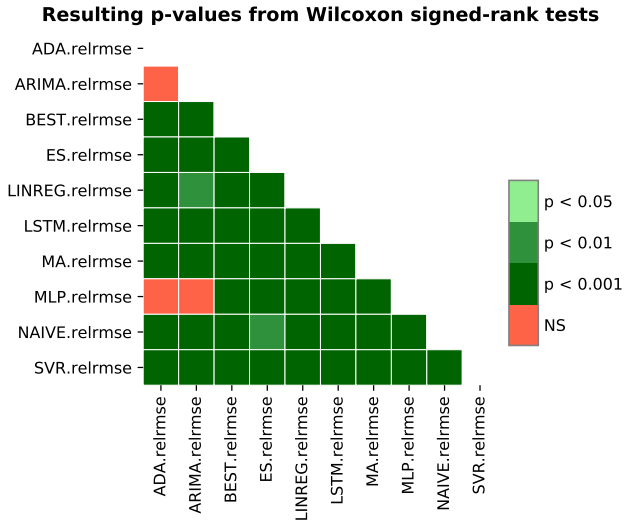

Figure B.1: OWC scenario Wilcoxon results

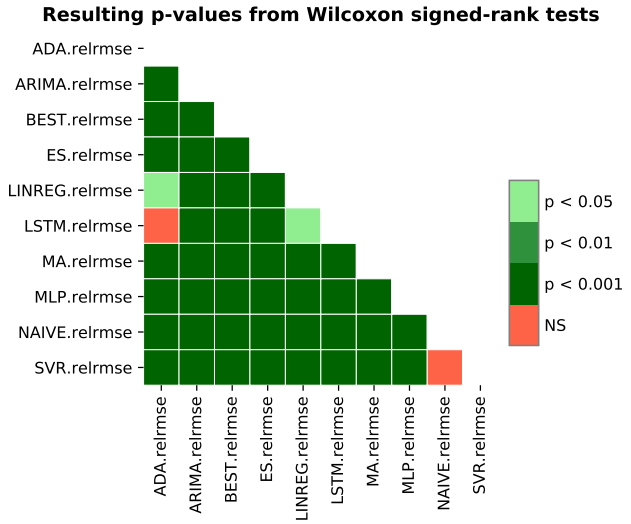

Figure B.2: ODC scenario Wilcoxon results

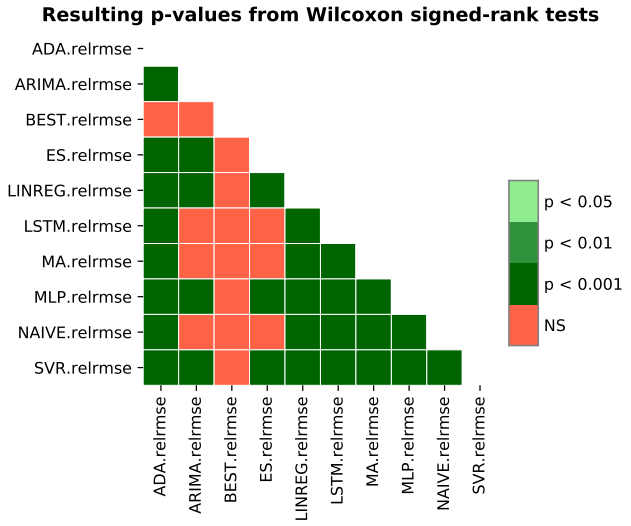

Figure B.3: OWS scenario Wilcoxon results

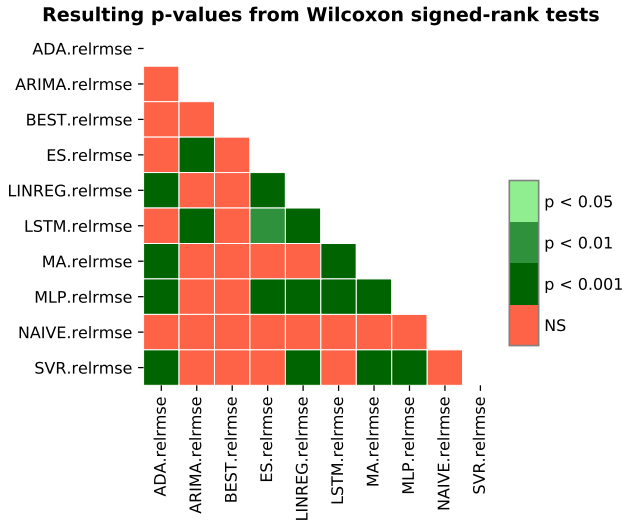

Figure B.4: ODS scenario Wilcoxon results

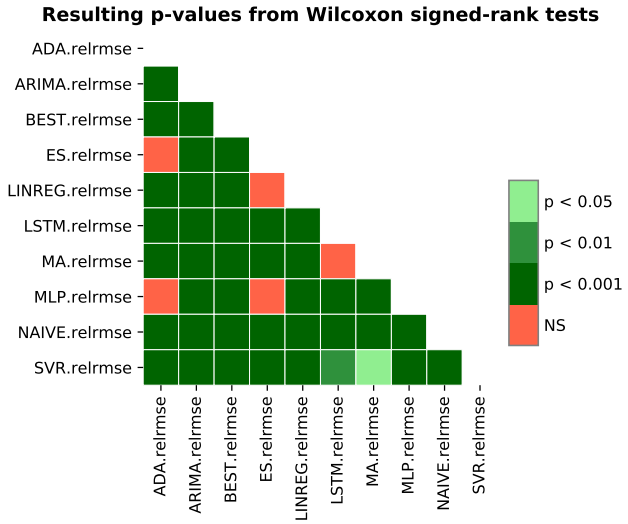

Figure B.5: M2WC scenario Wilcoxon results

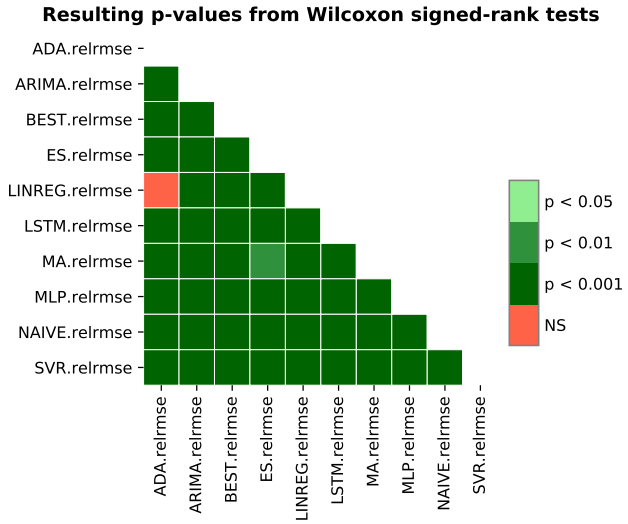

Figure B.6: M2DC scenario Wilcoxon results

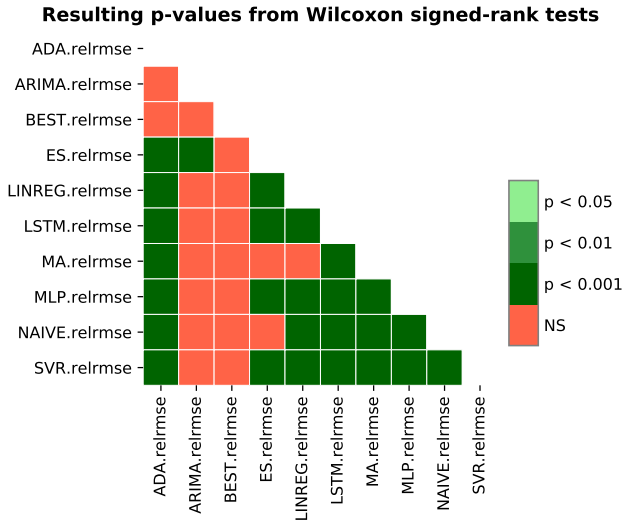

Figure B.7: M2WS scenario Wilcoxon results

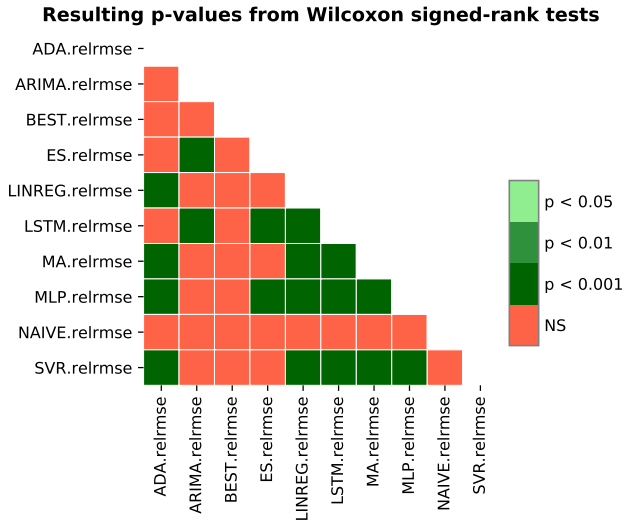

Figure B.8: M2DS scenario Wilcoxon results

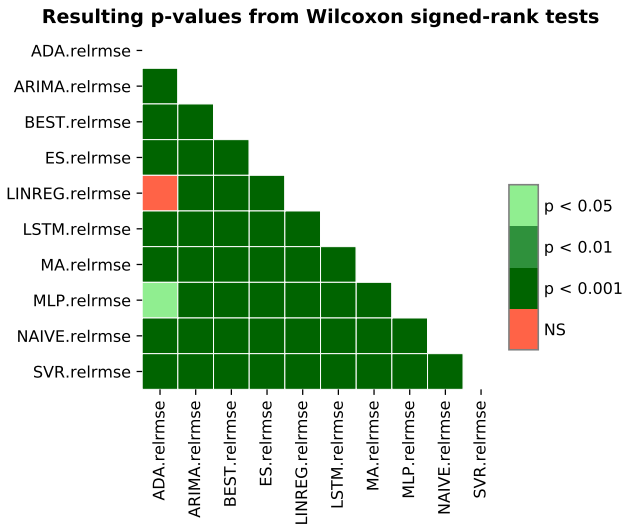

Figure B.9: M3DC scenario Wilcoxon results

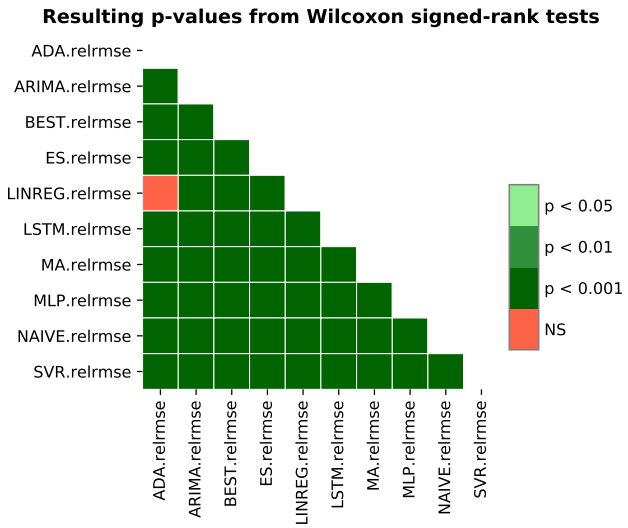

Figure B.10: M4DC scenario Wilcoxon results

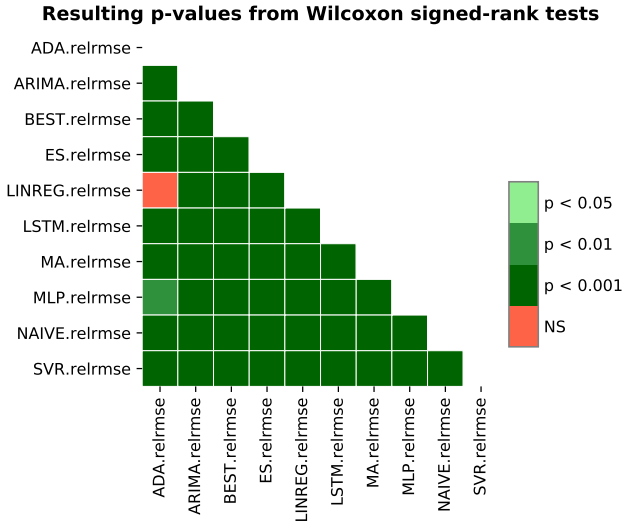

Figure B.11: M5DC scenario Wilcoxon results

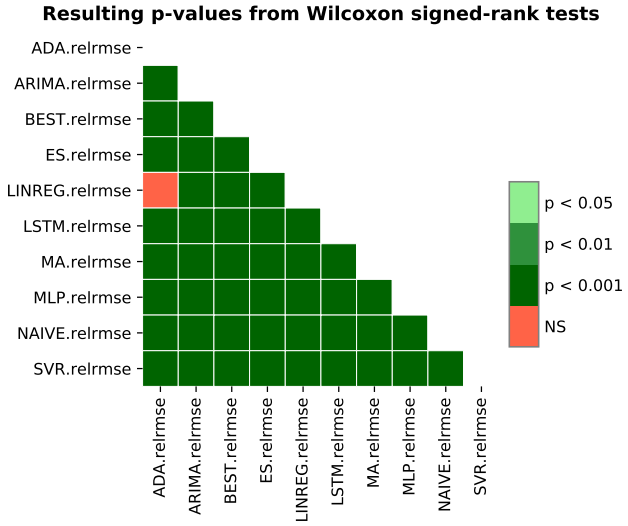

Figure B.12: M6DC scenario Wilcoxon results

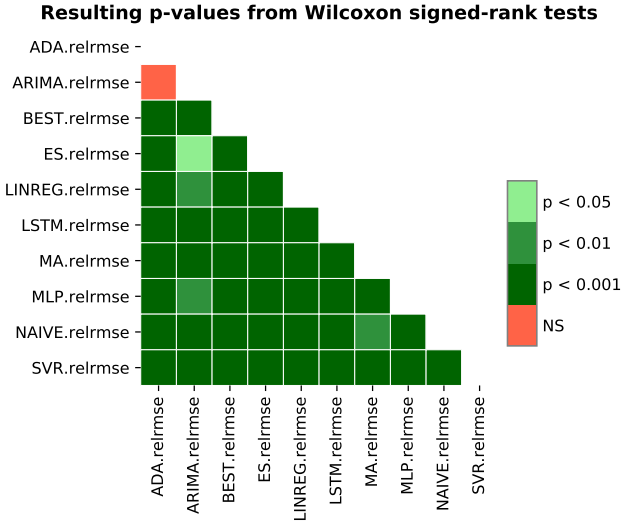

Figure B.13: M7DC scenario Wilcoxon results

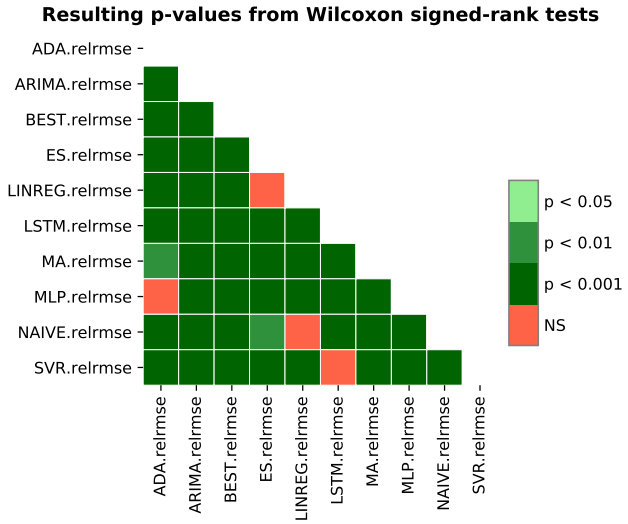

Figure B.14: OWCef scenario Wilcoxon results

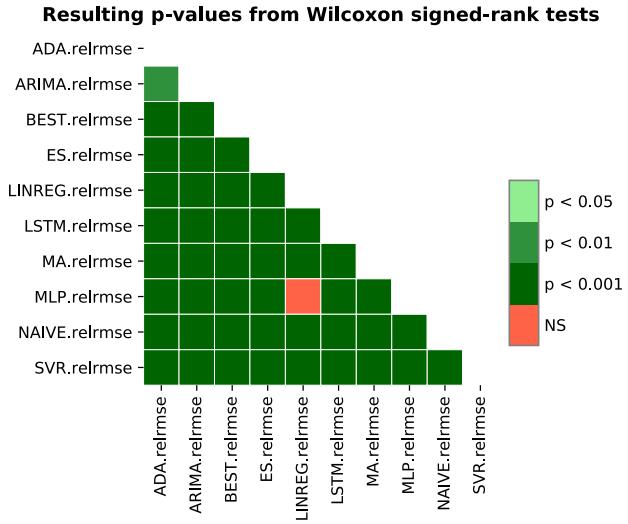

Figure B.15: ODCef scenario Wilcoxon results

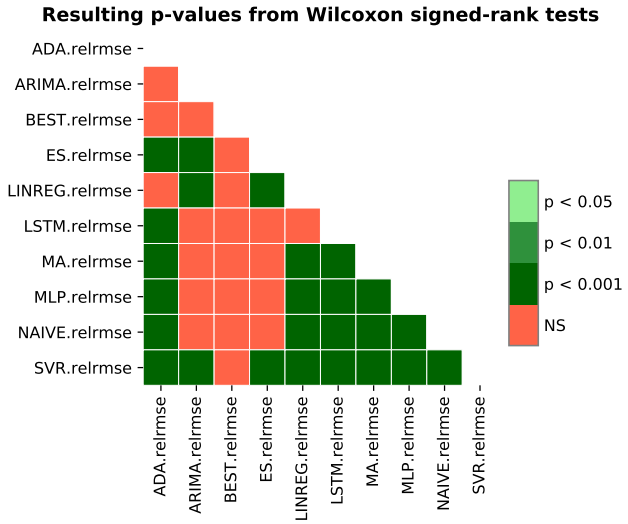

Figure B.16: OWSef scenario Wilcoxon results

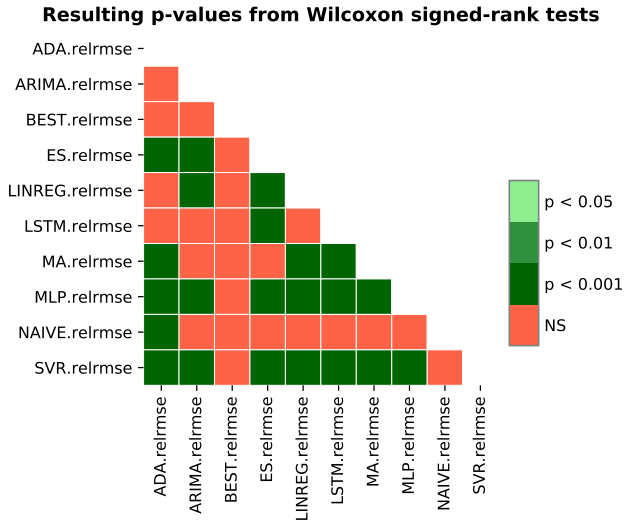

Figure B.17: ODSef scenario Wilcoxon results

C
DP Simulation Results

The tables on the next pages contain the results from the pricing simulations that were conducted. Table C.1 contains the results for PS1-PS4 for experiment 1 (baseline with default simulation settings). Tables C.2 to C.6 show the results from respectively experiment 2 to experiment 6.

To limit the length of this appendix, not all simulation results were included here. Since the pricing strategies only had a minor influence on the stock-out percentages in all experiments and to reduce the amount of space needed for the tables, the stock-out changes were not included in the tables for experiment 2 onwards, but they are available for request. For all experiments, with the exclusion of experiment 3 (which varies the probability of customers being regular) the default regular probability was used. However, since that setting turned out to have a large influence on results, all other experiments were re-run to also include a 0.0 and 1.0 RP. Those results are not reported in this appendix either, but are also available for request.

Table C.1: Experiment 1 results, baseline.

PS	D	S	Δ%Revenue	Δ%Waste	Δ%Stock-Outs
1	0.0	1	-	-	-
2	0.1	1	-0.8	-0.6	0.0
2	0.2	1	-3.7	-6.9	-0.3
2	0.3	1	-6.4	-7.3	-0.5
2	0.35	1	-8.0	-7.5	-0.5
2	0.4	1	-9.9	-8.0	-0.6
2	0.5	1	-13.3	-8.2	-0.7
2	0.6	1	-17.2	-8.5	-0.8
2	0.7	1	-21.3	-8.7	-0.9
2	0.8	1	-25.5	-8.9	-0.9
2	0.9	1	-30.3	-9.0	-0.9
2	1.0	1	-35.4	-9.2	-0.9
3	0.1	2	-2.9	-0.0	-0.0
3	0.2	2	-4.0	-0.2	-0.1
3	0.3	2	-6.9	-0.4	-0.2
3	0.35	2	-8.6	-0.4	-0.2
3	0.4	2	-15.3	-6.2	-0.2
3	0.5	2	-20.7	-6.3	-0.2
3	0.6	2	-26.8	-6.4	-0.3
3	0.7	2	-33.5	-6.5	-0.3
3	0.8	2	-41.2	-7.2	-0.3
3	0.9	2	-49.1	-7.2	-0.3
3	1.0	2	-57.4	-7.2	-0.3
4	-	1	-2.0	-3.8	-0.3

Table C.2: Experiment 2 results, different maximum shelf lives.

SL	PS	D	Δ%R	Δ%W	Δ%SO	SL	PS	D	Δ%R	Δ%W	Δ%SO
2	1	0.0	-	-	-	5	1	0.0	-	-	-
2	2	0.1	-5.2	-0.1	0.1	5	2	0.1	-0.6	-0.5	0.1
2	2	0.2	-11.0	-0.1	0.1	5	2	0.2	-0.9	-1.0	0.1
2	2	0.3	-16.8	-0.6	0.1	5	2	0.3	-1.3	-1.4	0.2
2	2	0.35	-20.8	-0.6	0.1	5	2	0.35	-1.6	-1.6	0.1
2	2	0.4	-25.0	-0.6	0.2	5	2	0.4	-2.4	-3.3	0.1
2	2	0.5	-34.5	-0.7	0.2	5	2	0.5	-3.2	-3.4	0.0
2	2	0.6	-45.1	-0.8	0.3	5	2	0.6	-4.2	-3.5	-0.0
2	2	0.7	-56.9	-0.8	0.3	5	2	0.7	-4.9	-3.5	-0.0
2	2	0.8	-69.9	-0.8	0.4	5	2	0.8	-5.7	-3.5	-0.1
2	2	0.9	-84.2	-0.7	0.5	5	2	0.9	-6.5	-3.5	-0.2
2	2	1.0	-99.6	-0.8	0.5	5	2	1.0	-7.3	-3.5	-0.1
2	4	-	-7.1	-8.3	-0.7	5	4	-	-0.7	-1.2	0.1
3	1	0.0	-	-	-	6	1	0.0	-	-	-
3	2	0.1	-0.6	-0.6	-0.0	6	2	0.1	-0.2	-0.5	0.1
3	2	0.2	-3.5	-6.8	-0.4	6	2	0.2	-0.3	-1.0	0.2
3	2	0.3	-6.3	-7.3	-0.5	6	2	0.3	-0.7	-1.3	0.2
3	2	0.35	-7.8	-7.5	-0.6	6	2	0.35	-1.2	-1.3	0.2
3	2	0.4	-9.8	-7.9	-0.7	6	2	0.4	-0.6	-2.0	0.0
3	2	0.5	-13.1	-8.2	-0.8	6	2	0.5	-0.8	-2.0	-0.1
3	2	0.6	-17.1	-8.5	-0.9	6	2	0.6	-1.5	-2.0	-0.1
3	2	0.7	-21.2	-8.6	-0.9	6	2	0.7	-1.7	-2.0	0.0
3	2	0.8	-25.5	-8.8	-1.0	6	2	0.8	-2.2	-2.0	-0.1
3	2	0.9	-30.3	-9.1	-1.0	6	2	0.9	-2.7	-2.0	-0.0
3	2	1.0	-35.2	-9.1	-1.0	6	2	1.0	-3.3	-2.0	-0.0
3	4	-	-2.3	-3.7	-0.2	6	4	-	-0.2	-1.0	-0.1
4	1	0.0	-	-	-	7	1	0.0	-	-	-
4	2	0.1	-0.4	-0.7	0.0	7	2	0.1	-0.4	-0.5	0.0
4	2	0.2	-1.0	-1.3	0.0	7	2	0.2	-0.5	-0.6	0.1
4	2	0.3	-3.4	-5.6	-0.1	7	2	0.3	-0.7	-0.8	0.0
4	2	0.35	-4.1	-5.7	-0.1	7	2	0.35	-0.6	-0.9	0.0
4	2	0.4	-4.7	-5.8	-0.2	7	2	0.4	-1.0	-1.0	0.0
4	2	0.5	-6.1	-6.0	-0.4	7	2	0.5	-0.1	-1.1	-0.1
4	2	0.6	-7.6	-6.0	-0.5	7	2	0.6	-0.3	-1.1	-0.0
4	2	0.7	-9.3	-6.1	-0.6	7	2	0.7	-0.3	-1.1	-0.1
4	2	0.8	-11.2	-6.1	-0.5	7	2	0.8	-1.0	-1.1	-0.1
4	2	0.9	-12.8	-6.1	-0.7	7	2	0.9	-1.2	-1.1	-0.1
4	2	1.0	-14.5	-6.1	-0.7	7	2	1.0	-1.7	-1.1	-0.2
4	4	-	-1.0	-1.7	0.1	7	4	-	-0.1	-0.5	0.1

Table C.3: Experiment 3 results, different regular probabilities.

RP	PS	D	Δ%R	Δ%W
0.0	1	0.0	-	-
0.0	2	0.1	0.3	0.1
0.0	2	0.2	5.6	-12.1
0.0	2	0.3	2.4	-12.2
0.0	2	0.35	0.8	-12.3
0.0	2	0.4	-1.7	-13.4
0.0	2	0.5	-5.5	-13.5
0.0	2	0.6	-9.7	-13.6
0.0	2	0.7	-14.0	-13.7
0.0	2	0.8	-18.5	-13.8
0.0	2	0.9	-23.3	-13.8
0.0	2	1.0	-28.4	-13.9
0.0	4	-	0.1	-5.9
0.1	1	0.0	-	-
0.1	2	0.1	0.5	-0.1
0.1	2	0.2	2.0	-10.5
0.1	2	0.3	-1.1	-10.8
0.1	2	0.35	-2.7	-10.9
0.1	2	0.4	-4.9	-11.8
0.1	2	0.5	-8.7	-11.9
0.1	2	0.6	-12.6	-12.0
0.1	2	0.7	-16.9	-12.2
0.1	2	0.8	-21.2	-12.3
0.1	2	0.9	-25.9	-12.4
0.1	2	1.0	-31.1	-12.5
0.1	4	-	-1.3	-5.8
0.2	1	0.0	-	-
0.2	2	0.1	-0.2	-0.3
0.2	2	0.2	-0.8	-9.2
0.2	2	0.3	-3.8	-9.6
0.2	2	0.35	-5.4	-9.7
0.2	2	0.4	-7.4	-10.4
0.2	2	0.5	-11.0	-10.5
0.2	2	0.6	-15.0	-10.7
0.2	2	0.7	-19.1	-10.8
0.2	2	0.8	-23.4	-11.0
0.2	2	0.9	-28.1	-11.1
0.2	2	1.0	-33.1	-11.3
0.2	4	-	-2.4	-5.9
0.3	1	0.0	-	-
0.3	2	0.1	-0.6	-0.5
0.3	2	0.2	-2.7	-8.0
0.3	2	0.3	-5.5	-8.4
0.3	2	0.35	-7.0	-8.5
0.3	2	0.4	-9.1	-9.1
0.3	2	0.5	-12.5	-9.3
0.3	2	0.6	-16.4	-9.5
0.3	2	0.7	-20.5	-9.7
0.3	2	0.8	-24.8	-9.9
0.3	2	0.9	-29.4	-10.0
0.3	2	1.0	-34.5	-10.2
0.3	4	-	-2.5	-5.0
0.4	1	0.0	-	-
0.4	2	0.1	-0.6	-0.5
0.4	2	0.2	-3.5	-6.7
0.4	2	0.3	-6.3	-7.2
0.4	2	0.35	-7.9	-7.4
0.4	2	0.4	-9.7	-7.9
0.4	2	0.5	-13.1	-8.1
0.4	2	0.6	-17.0	-8.3
0.4	2	0.7	-21.2	-8.6
0.4	2	0.8	-25.5	-8.8
0.4	2	0.9	-30.2	-9.0
0.4	2	1.0	-35.3	-9.1
0.4	4	-	-2.0	-3.4
0.5	1	0.0	-	-
0.5	2	0.1	-0.8	-0.7
0.5	2	0.2	-4.1	-5.9
0.5	2	0.3	-6.8	-6.4
0.5	2	0.35	-8.3	-6.5
0.5	2	0.4	-10.1	-7.0
0.5	2	0.5	-13.6	-7.3
0.5	2	0.6	-17.4	-7.6
0.5	2	0.7	-21.5	-7.7
0.5	2	0.8	-25.9	-8.0
0.5	2	0.9	-30.7	-8.2
0.5	2	1.0	-35.8	-8.4
0.5	4	-	-1.6	-2.9
0.6	1	0.0	-	-
0.6	2	0.1	-1.1	-0.8
0.6	2	0.2	-4.5	-4.9
0.6	2	0.3	-7.2	-5.3
0.6	2	0.35	-8.7	-5.6
0.6	2	0.4	-10.4	-5.9
0.6	2	0.5	-13.9	-6.4
0.6	2	0.6	-17.6	-6.6
0.6	2	0.7	-21.8	-6.8
0.6	2	0.8	-26.4	-7.2
0.6	2	0.9	-31.2	-7.3
0.6	2	1.0	-36.3	-7.5
0.6	4	-	-2.1	-3.2
0.7	1	0.0	-	-
0.7	2	0.1	-1.3	-0.9
0.7	2	0.2	-4.5	-4.0
0.7	2	0.3	-7.1	-4.6
0.7	2	0.35	-8.6	-5.0
0.7	2	0.4	-10.3	-5.3
0.7	2	0.5	-13.8	-5.7
0.7	2	0.6	-17.5	-5.9
0.7	2	0.7	-21.8	-6.2
0.7	2	0.8	-26.4	-6.5
0.7	2	0.9	-31.3	-6.7
0.7	2	1.0	-36.5	-7.0
0.7	4	-	-2.1	-3.1
0.8	1	0.0	-	-
0.8	2	0.1	-1.5	-0.9
0.8	2	0.2	-4.3	-3.1
0.8	2	0.3	-6.9	-3.8
0.8	2	0.35	-8.3	-4.0
0.8	2	0.4	-10.0	-4.3
0.8	2	0.5	-13.3	-4.8
0.8	2	0.6	-17.3	-5.2
0.8	2	0.7	-21.6	-5.5
0.8	2	0.8	-26.2	-5.8
0.8	2	0.9	-31.3	-6.0
0.8	2	1.0	-36.7	-6.2
0.8	4	-	-2.3	-3.2
0.9	1	0.0	-	-
0.9	2	0.1	-1.7	-1.0
0.9	2	0.2	-4.1	-2.5
0.9	2	0.3	-6.6	-3.1
0.9	2	0.35	-8.0	-3.4
0.9	2	0.4	-9.6	-3.8
0.9	2	0.5	-13.0	-4.2
0.9	2	0.6	-17.1	-4.6
0.9	2	0.7	-21.5	-5.0
0.9	2	0.8	-26.0	-5.2
0.9	2	0.9	-31.3	-5.5
0.9	2	1.0	-36.7	-5.7
0.9	4	-	-2.1	-2.9
1.0	1	0.0	-	-
1.0	2	0.1	-1.9	-0.9
1.0	2	0.2	-3.8	-1.7
1.0	2	0.3	-6.2	-2.4
1.0	2	0.35	-7.6	-2.8
1.0	2	0.4	-9.2	-3.0
1.0	2	0.5	-12.6	-3.6
1.0	2	0.6	-16.7	-4.1
1.0	2	0.7	-21.1	-4.4
1.0	2	0.8	-26.0	-4.7
1.0	2	0.9	-31.2	-4.9
1.0	2	1.0	-36.7	-5.2
1.0	4	-	-1.5	-2.1

Table C.4: Experiment 4 results, different price elasticities of demand.

EL	PS	D	Δ%R	Δ%W
-0.012	1	0.0	-	-
-0.012	2	0.1	0.9	-0.7
-0.012	2	0.2	-0.1	-7.2
-0.012	2	0.3	-1.6	-7.7
-0.012	2	0.35	-2.6	-7.9
-0.012	2	0.4	-3.8	-8.3
-0.012	2	0.5	-6.6	-8.6
-0.012	2	0.6	-9.8	-8.8
-0.012	2	0.7	-13.6	-9.0
-0.012	2	0.8	-18.0	-9.2
-0.012	2	0.9	-22.9	-9.4
-0.012	2	1.0	-28.4	-9.5
-0.012	4	-	-0.4	-3.7
-0.010	1	0.0	-	-
-0.010	2	0.1	0.2	-0.6
-0.010	2	0.2	-1.4	-7.0
-0.010	2	0.3	-3.2	-7.5
-0.010	2	0.35	-4.3	-7.7
-0.010	2	0.4	-5.9	-8.2
-0.010	2	0.5	-8.8	-8.5
-0.010	2	0.6	-12.2	-8.7
-0.010	2	0.7	-16.2	-8.9
-0.010	2	0.8	-20.6	-9.1
-0.010	2	0.9	-25.4	-9.2
-0.010	2	1.0	-30.6	-9.4
-0.010	4	-	-0.6	-3.4
-0.008	1	0.0	-	-
-0.008	2	0.1	-0.2	-0.7
-0.008	2	0.2	-2.5	-6.9
-0.008	2	0.3	-4.9	-7.4
-0.008	2	0.35	-6.2	-7.7
-0.008	2	0.4	-7.8	-8.1
-0.008	2	0.5	-11.0	-8.3
-0.008	2	0.6	-14.6	-8.7
-0.008	2	0.7	-18.6	-8.8
-0.008	2	0.8	-23.0	-9.0
-0.008	2	0.9	-27.8	-9.2
-0.008	2	1.0	-33.0	-9.4
-0.008	4	-	-1.3	-3.6
-0.006	1	0.0	-	-
-0.006	2	0.1	-0.7	-0.7
-0.006	2	0.2	-3.6	-7.0
-0.006	2	0.3	-6.4	-7.4
-0.006	2	0.35	-7.9	-7.6
-0.006	2	0.4	-9.7	-8.0
-0.006	2	0.5	-13.2	-8.2
-0.006	2	0.6	-17.1	-8.5
-0.006	2	0.7	-21.1	-8.7
-0.006	2	0.8	-25.6	-8.9
-0.006	2	0.9	-30.2	-9.1
-0.006	2	1.0	-35.3	-9.3
-0.006	4	-	-2.0	-3.7
-0.004	1	0.0	-	-
-0.004	2	0.1	-1.1	-0.6
-0.004	2	0.2	-4.7	-6.8
-0.004	2	0.3	-8.0	-7.2
-0.004	2	0.35	-9.8	-7.4
-0.004	2	0.4	-11.8	-7.9
-0.004	2	0.5	-15.5	-8.1
-0.004	2	0.6	-19.6	-8.3
-0.004	2	0.7	-23.9	-8.5
-0.004	2	0.8	-28.3	-8.7
-0.004	2	0.9	-33.1	-8.8
-0.004	2	1.0	-38.0	-9.0
-0.004	4	-	-2.5	-3.7
-0.002	1	0.0	-	-
-0.002	2	0.1	-1.6	-0.4
-0.002	2	0.2	-5.9	-6.5
-0.002	2	0.3	-9.7	-6.9
-0.002	2	0.35	-11.7	-7.1
-0.002	2	0.4	-13.9	-7.5
-0.002	2	0.5	-18.1	-7.8
-0.002	2	0.6	-22.5	-7.9
-0.002	2	0.7	-26.9	-8.1
-0.002	2	0.8	-31.4	-8.3
-0.002	2	0.9	-36.3	-8.5
-0.002	2	1.0	-41.0	-8.6
-0.002	4	-	-2.9	-3.2

Table C.5: Experiment 5 results, different safety factors.

SF	PS	D	%W	Δ%R	Δ%W
0.8	1	0.0	0.8	-	-
0.8	2	0.1	0.7	-0.4	-0.1
0.8	2	0.2	0.1	0.6	-0.7
0.8	2	0.3	0.1	0.3	-0.7
0.8	2	0.35	0.1	0.1	-0.8
0.8	2	0.4	0.1	-0.1	-0.7
0.8	2	0.5	0.0	-0.5	-0.8
0.8	2	0.6	0.0	-0.9	-0.8
0.8	2	0.7	0.0	-1.3	-0.8
0.8	2	0.8	0.0	-1.8	-0.8
0.8	2	0.9	0.0	-2.2	-0.8
0.8	2	1.0	0.0	-2.6	-0.8
0.8	4	-	0.5	-0.3	-0.3
1.0	1	0.0	2.6	-	-
1.0	2	0.1	2.3	-0.2	-0.3
1.0	2	0.2	0.3	-0.0	-2.2
1.0	2	0.3	0.3	-0.9	-2.3
1.0	2	0.35	0.2	-1.4	-2.3
1.0	2	0.4	0.2	-1.8	-2.4
1.0	2	0.5	0.2	-2.8	-2.4
1.0	2	0.6	0.2	-3.7	-2.4
1.0	2	0.7	0.2	-4.6	-2.4
1.0	2	0.8	0.2	-5.5	-2.4
1.0	2	0.9	0.1	-6.6	-2.5
1.0	2	1.0	0.1	-7.5	-2.5
1.0	4	-	1.8	-0.3	-0.8
1.2	1	0.0	5.2	-	-
1.2	2	0.1	4.8	-0.5	-0.5
1.2	2	0.2	0.9	-1.2	-4.3
1.2	2	0.3	0.8	-2.8	-4.4
1.2	2	0.35	0.7	-3.6	-4.5
1.2	2	0.4	0.5	-4.6	-4.7
1.2	2	0.5	0.5	-6.3	-4.7
1.2	2	0.6	0.4	-8.3	-4.8
1.2	2	0.7	0.4	-10.3	-4.9
1.2	2	0.8	0.3	-12.4	-4.9
1.2	2	0.9	0.3	-14.7	-4.9
1.2	2	1.0	0.3	-16.8	-5.0
1.2	4	-	3.7	-0.7	-1.5
1.4	1	0.0	8.6	-	-
1.4	2	0.1	8.0	-0.5	-0.6
1.4	2	0.2	2.5	-2.7	-6.0
1.4	2	0.3	2.2	-5.2	-6.4
1.4	2	0.35	2.1	-6.5	-6.5
1.4	2	0.4	1.7	-8.1	-6.9
1.4	2	0.5	1.5	-11.0	-7.1
1.4	2	0.6	1.3	-14.2	-7.3
1.4	2	0.7	1.1	-17.6	-7.4
1.4	2	0.8	0.9	-21.2	-7.6
1.4	2	0.9	0.9	-25.0	-7.7
1.4	2	1.0	0.7	-29.2	-7.9
1.4	4	-	5.8	-1.1	-2.7
1.6	1	0.0	12.4	-	-
1.6	2	0.1	11.8	-0.4	-0.6
1.6	2	0.2	4.8	-4.3	-7.6
1.6	2	0.3	4.3	-7.6	-8.0
1.6	2	0.35	4.1	-9.4	-8.3
1.6	2	0.4	3.5	-11.5	-8.8
1.6	2	0.5	3.2	-15.8	-9.2
1.6	2	0.6	2.9	-20.2	-9.5
1.6	2	0.7	2.7	-25.0	-9.7
1.6	2	0.8	2.4	-30.2	-9.9
1.6	2	0.9	2.3	-35.7	-10.1
1.6	2	1.0	2.0	-41.7	-10.3
1.6	4	-	8.3	-2.2	-4.0
1.8	1	0.0	16.1	-	-
1.8	2	0.1	15.6	-0.6	-0.5
1.8	2	0.2	7.5	-6.3	-8.6
1.8	2	0.3	7.1	-10.3	-9.0
1.8	2	0.35	6.8	-12.6	-9.3
1.8	2	0.4	5.8	-15.5	-10.3
1.8	2	0.5	5.4	-20.7	-10.7
1.8	2	0.6	5.1	-26.4	-11.0
1.8	2	0.7	4.8	-32.5	-11.3
1.8	2	0.8	4.6	-39.0	-11.5
1.8	2	0.9	4.2	-46.2	-11.9
1.8	2	1.0	3.9	-53.8	-12.2
1.8	4	-	10.7	-3.8	-5.4
2.0	1	0.0	20.0	-	-
2.0	2	0.1	19.4	-0.6	-0.6
2.0	2	0.2	10.7	-7.6	-9.3
2.0	2	0.3	10.2	-12.4	-9.8
2.0	2	0.35	10.0	-14.8	-10.0
2.0	2	0.4	8.8	-18.4	-11.1
2.0	2	0.5	8.4	-24.4	-11.6
2.0	2	0.6	7.9	-31.3	-12.1
2.0	2	0.7	7.4	-38.9	-12.5
2.0	2	0.8	7.0	-46.9	-12.9
2.0	2	0.9	6.7	-55.5	-13.3
2.0	2	1.0	6.3	-64.9	-13.7
2.0	4	-	12.0	-6.2	-8.0

Table C.6: Experiment 6 results, different sales history variations.

LOCL	HT	PS	D	Δ% R	Δ% W
C	False	1	0.0	-	-
C	False	2	0.1	-0.2	-0.8
C	False	2	0.2	-3.2	-6.6
C	False	2	0.3	-5.9	-6.8
C	False	2	0.35	-7.3	-6.8
C	False	2	0.4	-8.8	-7.0
C	False	2	0.5	-11.6	-7.0
C	False	2	0.6	-14.6	-7.0
C	False	2	0.7	-17.7	-7.0
C	False	2	0.8	-20.8	-7.0
C	False	2	0.9	-24.1	-7.0
C	False	2	1.0	-27.4	-7.0
C	False	4	-	-1.5	-3.6
C	True	1	0.0	-	-
C	True	2	0.1	-0.0	-0.9
C	True	2	0.2	-0.9	-9.2
C	True	2	0.3	-5.1	-9.9
C	True	2	0.35	-7.4	-10.2
C	True	2	0.4	-10.8	-12.2
C	True	2	0.5	-16.2	-12.4
C	True	2	0.6	-21.8	-12.5
C	True	2	0.7	-27.6	-12.6
C	True	2	0.8	-33.8	-12.6
C	True	2	0.9	-40.1	-12.6
C	True	2	1.0	-46.7	-12.7
C	True	4	-	-1.5	-5.6
S	False	1	0.0	-	-
S	False	2	0.1	-0.5	-0.5
S	False	2	0.2	-3.5	-6.9
S	False	2	0.3	-6.3	-7.3
S	False	2	0.35	-7.8	-7.4
S	False	2	0.4	-9.8	-8.0
S	False	2	0.5	-13.2	-8.1
S	False	2	0.6	-17.0	-8.4
S	False	2	0.7	-21.2	-8.7
S	False	2	0.8	-25.4	-8.8
S	False	2	0.9	-30.1	-9.0
S	False	2	1.0	-35.2	-9.2
S	False	4	-	-1.9	-3.5
S	True	1	0.0	-	-
S	True	2	0.1	-0.9	-0.4
S	True	2	0.2	-3.3	-6.1
S	True	2	0.3	-6.7	-6.7
S	True	2	0.35	-8.6	-6.9
S	True	2	0.4	-10.9	-7.6
S	True	2	0.5	-15.4	-8.0
S	True	2	0.6	-20.3	-8.4
S	True	2	0.7	-25.5	-8.8
S	True	2	0.8	-31.1	-9.1
S	True	2	0.9	-37.3	-9.3
S	True	2	1.0	-43.6	-9.6
S	True	4	-	-2.3	-3.0