

Aalto University
School of Science
Master's Programme in Computer,
Communication and Information Sciences

Ville Vainio

Estimating Assortment-based Substitution in Retail

Master's Thesis
Espoo, July 29, 2019

Supervisor: Senior University Lecturer Vesa Hirvisalo, Aalto University

Advisor: Master of Social Science Erkka Saarinen

Author:	Ville Vainio		
Title:	Estimating Assortment-based Substitution in Retail		
Date:	July 29, 2019	Pages:	viii + 61
Major:	Computer Science	Code:	SCI3042
Supervisor:	Senior University Lecturer Vesa Hirvisalo		
Advisor:	Master of Social Science Erkkä Saarinen		
<p>Deciding which products to include in the assortment of a store is one of the most basic decisions a retailer has to make, but finding out which assortment produces the highest profits or margins is a difficult problem. Failing to offer the optimal products causes loss of profits or even loss of customers for the retailer.</p> <p>When customers are looking for a specific product but find that it is unavailable, they may choose to buy another similar product instead. This is called substitution, and the amount of substitution affects the demand of products. Therefore, the optimal assortment also depends on the amount of substitution, and hence, many assortment planning solutions incorporate the effect of substitution in them. However, only one study was found where substitution has been estimated directly.</p> <p>The goal of this thesis is to find how assortment-based substitution in retail can be directly estimated. In this thesis the methodology for estimating assortment-based substitution is presented and then applied to a retail receipt data set.</p> <p>The results indicate that the accuracy of the substitution estimation method depends heavily on the accuracy of the demand forecasts that are used as inputs for the substitution estimation. The forecasting models used in this thesis are not able to predict demand accurately for slow-moving products, which limits the accuracy of the substitution estimation method. Additionally, the results of this thesis show that the footfall of a store, i.e. how many customers visit a store on a given day, can be estimated very accurately.</p>			
Keywords:	substitution, grocery retail, assortment planning, demand forecasting, data science		
Language:	English		

Tekijä:	Ville Vainio		
Työn nimi:	Valikoimasta johtuvan substituution mallintaminen vähittäiskaupassa		
Päiväys:	29. heinäkuuta 2019	Sivumäärä:	viii + 61
Pääaine:	Tietotekniikka	Koodi:	SCI3042
Valvoja:	Vanhempi yliopistonlehtori Vesa Hirvisalo		
Ohjaaja:	Filosofian maisteri Erkka Saarinen		
<p>Yksi vähittäiskauppiaan olennaisimmista päätöksistä on valita kaupan tuotevalikoimaan sisällytettävät tuotteet, mutta parhaiten tuottavan tuotevalikoiman löytäminen on vaikeaa. Jos asiakkaat eivät löydä valikoimasta haluamiaan tuotteita, johtaa tämä huonompaan myyntiin tai jopa asiakkaiden menetykseen.</p> <p>Kun asiakas etsii jotakin tiettyä tuotetta, mutta huomaa ettei sitä ole saatavilla, saattaa hän ostaa toisen tuotteen tämän sijasta. Tätä ilmiötä kutsutaan substituutioksi, ja se vaikuttaa tuotteiden kysyntään ja siten myös siihen, mikä on optimaalisin tuotevalikoima. Tämän vuoksi monet nykyiset tuotevalikoiman suunnitteluun käytettävät menetelmät sisällyttävätkin substituution vaikutuksen yhdeksi menetelmän osaksi. Itse substituution määrää on kuitenkin mallinnettu vain yhdessä löydetyssä aikaisemmassa tutkimuksessa.</p> <p>Tämän diplomityön tavoite on tutkia, kuinka valikoimasta johtuvan substituution määrää pystytään mallintamaan vähittäiskauppiaan kuittidatalla. Diplomityössä esitetään metodologiaa valikoimasta johtuvan substituution määrän mallintamiseen, ja tämän jälkeen sitä sovelletaan vähittäiskauppiaan kuittidataan.</p> <p>Diplomityön tulokset osoittavat, että substituution mallintamisen tarkkuus riippuu käytettyjen kysynnän ennustamismallien tarkkuudesta. Tässä diplomityössä käytetyt kysynnän ennustamismallit eivät pysty ennustamaan kysyntää tarkasti tuotteiden vähäisestä myynnistä johtuen, mikä rajoittaa substituution mallintamisen tarkkuutta. Tämän lisäksi diplomityön tulokset osoittavat, että kauppojen päivittäistä kävijämäärää on mahdollista ennustaa erittäin tarkasti.</p>			
Asiasanat:	substituutio, vähittäiskauppa, tuotevalikoiman suunnittelu, kysynnän ennustaminen, koneoppiminen		
Kieli:	Englanti		

Acknowledgements

This thesis has been a challenging project for me, and it has taken more of my time and energy than I initially thought. Nevertheless, I have learned a lot while making it, and now that the thesis is finished, I am proud of what I have achieved. However, I could not have done the thesis without the help of other people along the way, and even if I had, the journey would not have been as fun without you. Therefore, I would like to thank all the people who have helped me with the thesis and provided moral support during the process.

First, thanks to my instructor Erkka Saarinen for teaching me the basics of the subject, helping me get started with the thesis and for the fruitful conversations. I would also like to thank my supervisor Vesa Hirvisalo for his advice and feedback throughout the thesis. Also, big thanks to all my colleagues for helping me with my questions along the way, moral support and for the good conversations with you, and thanks to RELEX for giving me the opportunity to write my thesis here.

Last but not least, thanks to my parents for their advice and support, and my friends for giving counterbalance to my work with the leisure time with you.

Espoo, July 29, 2019

Ville Vainio

Abbreviations and Acronyms

SKU	Stock Keeping Unit; A distinct type of product for sale in a retail store
RMSE	Root Mean Square-Error; A measure for absolute amount of error in prediction
MAE	Mean Absolute Error; A measure for absolute amount of error in prediction
MAPE	Mean Absolute Percentage Error; A measure for relative amount of error in prediction
SMAPE	Symmetric Mean Absolute Percentage Error; A measure for relative amount of error in prediction
R^2	R-Squared; Proportion of the variance in the dependent variable that is explained by the explanatory variables
AR^2	Adjusted R-Squared; R-Squared adjusted by the number of explanatory variables in the model
ASDE	After Substitution Demand Estimation
ODE	Original Demand Estimation
HDI	Human Discomfort Index; An index variable used in modeling the effect of weather on customer behavior

Symbols

K	Footfall
π	Purchase incidence
p	Product choice
q	Purchase quantity
B^l	Dummy variable (0 or 1) for weekday l
E^l	Dummy variable (0 or 1) for holiday l
H	Dummy variable (0 or 1) for whether a day is a holiday or not
T	Temperature
A	Dummy variable (0 or 1) for whether a product is in promotion or not
\bar{A}	Average promotion level in a subcategory
R	Unit value of a product
\bar{R}	Average unit value of a subcategory
κ_i	Regression coefficient for footfall
γ_i	Regression coefficient for purchase incidence
β_i	Regression coefficient for product choice
ζ_i	Regression coefficient for purchase quantity
\bar{K}	Average footfall of a store
\bar{N}	Average items per basket for a store
δ	Substitution rate

Contents

Abbreviations and Acronyms	v
Symbols	vi
1 Introduction	1
1.1 Objective and Research Questions	2
1.2 Scope of the Thesis	2
1.3 Thesis Contribution	3
1.4 Structure of the Thesis	3
2 Background	4
2.1 Demand Forecasting	4
2.1.1 Time Series Methods	5
2.1.2 Causal and Machine Learning Methods	5
2.2 Assortment Planning	6
2.3 Substitution	7
3 Methods	9
3.1 Choice-Based Approach for Demand Forecasting	10
3.2 Substitution Rate Estimation	12
4 Implementation	14
4.1 Data Description	14
4.2 Footfall	16
4.3 Choice-Based Model for Estimating Demand	17
4.3.1 Purchase Incidence	17
4.3.2 Product Choice	19
4.3.3 Purchase Quantity	20
4.4 Simple Linear Model for Estimating Demand	21
4.5 ODE and ASDE Models for Capturing Substitution Effect	21
4.5.1 ODE Purchase Incidence	22

4.5.2	ODE Product Choice	23
4.5.3	ODE Purchase Quantity	23
4.6	Calculating Substitution Rate	24
5	Results	26
5.1	Footfall	27
5.2	Choice-Based Model	29
5.3	Demand Forecasting Models' Performance	31
5.4	ODE Model Performances	33
5.5	Substitution Rate	35
5.6	Discussion of Results	36
6	Conclusions	38
6.1	Key Findings	38
6.2	Validity for Real-Life Applications	41
6.3	Potential Future Research	41
A	Complete Data Analysis Results	47

Chapter 1

Introduction

In retail, assortment is defined as the set of products carried in a store at a given point in time [2]. Assortment planning is the process of deciding which products to include in the assortment and setting their inventory levels so as to maximize profits for the retailer [26]. Determining which products to include in the assortment of a store is one of the most basic decisions a retailer has to make. However, finding out which assortment produces the highest profits or margins is a difficult problem because the number of possible products is huge and because retailers are limited by the available shelf space to display products on and their budget for purchasing them. Failing to offer the correct products causes loss of profits or even loss of customers for the retailer. [2, 14, 28] Because of the complexity of the problem, many retailers today use automated processes for assortment planning to decrease costs and increase profitability. However, no solution has emerged dominant in assortment planning [2]. In addition, there exists more and more literature on assortment planning, but the solutions in literature are often theoretical and make a lot of assumptions, which do not always hold in practice. Hence, the advanced assortment planning solutions in the literature have not reached practical applications, but instead the solutions used by most retailers are quite simple [2, 28].

When customers are buying a specific item but find that it is unavailable, they may buy another similar product instead. This is called substitution, and it is important to take it into account in assortment planning since it affects the optimal assortment [26]. Multiple studies have discovered that customers often substitute a product with another if their desired product is unavailable [8, 16, 19]. For example, in a study performed by Andersen Consulting in 1996, it was found that customers substitute a product with

another 60% of the time on average when their desired product is unavailable [11].

1.1 Objective and Research Questions

The objective of this thesis is to find a way to estimate assortment-based substitution rate¹ in a real-world retail receipt data set. Substitution is often incorporated in assortment planning solutions since substitution has an effect on the optimal assortment, and therefore current methods for assortment planning are also briefly reviewed. The only found method in literature for directly estimating assortment-based substitution is also reviewed. Since this method uses demand forecasts to estimate substitution rate, also current methods for demand forecasting in literature are described. In this thesis, it is also studied how accurately a choice-based forecasting model can estimate demand in a real-world retail data set compared to a simple linear regression model. The research questions of this thesis are therefore the following:

- Research question 1: How can assortment-based substitution rate be estimated directly from a real-world retail receipt data set?
- Research question 2: How accurate is a choice-based demand forecasting model compared to simple linear regression model on a real-world retail receipt data set?

1.2 Scope of the Thesis

This thesis focuses on assortment-based substitution estimation in retail and more specifically in grocery retail. However, the results should be applicable to other areas in retail, such as fashion or DIY, as well since the method in this thesis is a general way to estimate assortment-based demand and makes no assumptions about the type of products on sale. This thesis focuses specifically on assortment-based substitution instead of stock-out-based substitution, though Campo et al. state that there are similarities in customer reactions in both of these cases [8].

¹Substitution rate is defined as the proportion of demand from products not in assortment that is split to other products in the same subcategory because of substitution.

1.3 Thesis Contribution

Even though many assortment planning solutions incorporate the effect of substitution in them, only one study was found where substitution rate has been estimated directly. This thesis contributes to existing research by validating the performance of the substitution rate estimation method in that study by applying the method to a real-world setting. The performances of the individual components of the method are also studied in this thesis. Additionally, this thesis presents how a simple linear demand forecasting model can be used as a component of the substitution estimation method and compares the performance of that model to the choice-based model of the previous study.

1.4 Structure of the Thesis

This thesis is split into 5 chapters. In Chapter 2, the general background about demand forecasting and commonly used methods for demand forecasting in retail are presented. After that, background about assortment planning and common methods used for assortment planning are presented, and finally, the concept of substitution in retail is presented. Chapter 3 describes the methodology used in this thesis for estimating assortment-based substitution rate in detail. The chapter first describes in detail a choice-based model for forecasting demand since the demand forecasts are used in estimating the substitution rate. After that, the method for estimating assortment-based substitution rate is presented in detail. Chapter 4 describes how this method is implemented to estimate assortment-based substitution rate from a real-world retail data set. First, the implementations of the choice-based and simple linear demand forecasting models are described, then the implementation of the substitution estimation method is described. Chapter 5 presents how the demand forecasting models and their components perform with different explanatory variables, and then presents the results of the substitution estimation method. After that, the results are discussed and compared to existing research. Finally, Chapter 6 summarizes the key findings of this thesis, discusses the validity of the results for real-life applications and presents suggestions for future research.

Chapter 2

Background

In this chapter, the concept of demand forecasting in retail is introduced and commonly used methods to forecast demand in retail are presented in Section 2.1. After this, the concept of assortment planning is introduced and methods to optimize assortment are presented in Section 2.2. Finally, substitution in retail is described in Section 2.3.

2.1 Demand Forecasting

Being able to forecast demand of products accurately is important for retail companies since these forecasts affect decisions on many functional areas such as marketing, sales, production/purchasing and finance and accounting. Sales forecasts are also the main input when planning distribution and replenishment for retail companies. [15] Accurate demand forecasts reduce the number of stock-outs and lower the safety stocks¹, which contributes to higher profits for the retailer [3].

Forecasting models have been developed and improved largely during the past several decades [15]. Despite this, according to a 2006 study by McCarthy et al., demand forecasting accuracy did not improve or even declined between 1986 and 2006. Additionally, a large portion of retailers used only simple forecasting models or no models at all for forecasting demand. [29] Pe-

¹Safety stock in retail means a level of extra stock that is kept in case of uncertainties in supply or demand. For example, if the demand of a product is higher than expected and there is no safety stock, the product might sell out, making customers unable to buy said product.

terson also discovered in his 1993 research that majority of retailers preferred managerial judgement over mathematical forecasting models [32].

Demand forecasting methods in retail are typically divided into qualitative and quantitative methods. Qualitative methods are based on personal experience and knowledge of managers, i.e. managerial judgement. Quantitative methods on the other hand are mathematical models that predict demand mechanically and do not require any human input or judgement. Quantitative methods in retail are typically split into time-series methods and causal and machine learning methods [4, 25, 27]. The key distinction between these two are that time-series methods use only past demand data to predict future demand, whereas causal and machine learning methods use additional explanatory variables such as price, promotions and weather to predict future demand [4]. Time series methods and causal and machine learning methods are described in Subsections 2.1.1 and 2.1.2 respectively.

2.1.1 Time Series Methods

The simplest time series forecasting method, called Naïve forecasting method, is to predict the future values to be the same as the latest observed value. However, this method is rarely useful in practice for forecasting demand because it cannot predict any fluctuations in demand and also because it is not able to smooth any noise in data and instead includes the noise in the future predictions. [10] A slightly more advanced time series forecasting method is called the simple moving average method, which predicts the future values to be the mean of N past values. Variations of this method such as weighted moving averages and exponential smoothing add weights to past values.[27] The moving average methods are well-known and commonly used in demand forecasting [10]. More advanced time series methods include for example Box-Jenkins ARIMA model [3], Holt-Winters method [10] and Fourier analysis [17].

2.1.2 Causal and Machine Learning Methods

As discussed earlier, causal and machine learning methods use additional explanatory variables such as price, promotions and weather to predict future demand. One simple causal model is linear regression. In linear regression, it is assumed that there exist linear relationships between the explanatory variables and response variable, which in this case is demand. [35] The ben-

enefit of linear regression is that it is simple and easy to compute, which makes it useful for large retailers who have to calculate forecasts for a large number of products quickly [15]. An example of a more advanced causal method is a type of dynamic regression model developed by Divakar et al. which includes additional explanatory variables for own and competitor prices, promotions and seasonality [13]. Weber et al. use a model which combines an ARIMA model and neural networks to forecast demand in a grocery retail environment [1]. More complex machine learning methods have higher generality, which provides potentially more accurate forecasts, but come with the cost of increased danger of overfitting [15].

2.2 Assortment Planning

Assortment in retail is defined as the set of products carried in a store at a given point in time [2]. Determining the assortment of a store is one of the most basic strategic decisions a retailer has to make. However, retailers are constrained by the money they have to buy different products and shelf space to put their products on. [28] According to Quelch and Kenny, the number of SKU's² grew 16% per year between 1985 and 1992 while retail shelf space expanded only by 1.5% per year during the same period [33]. This indicates that assortment planning is more important now than before since only a smaller portion of all possible products can be fit into shelves. Furthermore, Iyengar and Leppar discovered in their empirical research that excessive choices can lead to customers opting not to choose any product at all [24]. Boatwright et al. also found that reducing the assortment sizes in most of the categories increased average sales across all observed categories by 11% [6]. Therefore, retailers must aim to limit their assortments to a reasonable number of products.

Another problem related to assortment planning is shelf space planning. Shelf space planning is usually done after the assortment of a category has been chosen, and it includes deciding how many facings to allocate for each product in a category and where to place the products on the shelf. [23] This problem is not discussed in this thesis though.

When planning the assortment of a category, it is very useful to know how the demand for all products in the category will be distributed. However, this

²In retail, Stock Keeping Unit (SKU) is defined as a distinct type of product for sale. For example a cereal box of some brand, type and size is a SKU and there may be multiple units of said SKU on sale in a store.

problem is very difficult in practice because the demand of a product depends on other products since customers may substitute a product they were going to buy with another product from the same category if the original product is not available [14]. Hence, no solution introduced in academia is dominant in practice, and in addition, most assortment optimization methods used in practice are mathematically very simple because in order to be able to use them in practice they must be fast and robust [2, 28]. However, typical models used in estimating the demand distribution within a category are called customer choice models. Farias et al. define Customer choice models as models that output a probability distribution of the likelihoods that an arriving customer purchases a given product in the set of available products [14]. A simple customer choice model is the multinomial logit (MNL) model, which is commonly used in literature to find the optimal assortment [30, 31, 34]. Other customer choice models used in assortment optimization include the locational choice model [18] and the probit model [5].

2.3 Substitution

When customers are looking to buy a specific product, but find that it is unavailable, they may decide to instead buy a product that is similar to what they were looking to buy in the first place. This is called substitution, and it includes switching to a different product when the assortment does not include the product the customer was originally looking to buy (assortment-based substitution) and switching to a different product when the product has been sold out (stockout-based substitution). [20, 26] A study performed by Campo et al., indicates that there are similarities in customer behavior in stockout-based and assortment-based substitution [8].

Since substitution affects the demand of products, and therefore also the optimal assortment, it is important to take the effect of substitution in account when deciding which products to include in the assortment of a store [26]. For example, if there is a low substitution rate in a subcategory, it is more important to have a larger assortment, especially if the products are fast-moving products, because not having those product in the assortment results in lost sales. On the other hand, if there is a large substitution rate in a subcategory, it is not as important to have a large assortment since customers will substitute their favorite product with a different one if the favorite one is not available. Substitution also affects replenishment because if a category has a low substitution rate, stock-outs cause larger drops in total

category sales than with a large substitution rate since customers will not substitute the product that has ran out with another product. Therefore, it is more important to replenish products in categories with a low substitution rate than in ones with a large substitution rate.

According to an interview study performed in the U.S. by Andersen Consulting in 1996, when facing out-of-stocks, customers decided not to buy anything 34% of the time on average. The proportion also varies heavily on the category: in one of the observed categories customers opted not to buy anything 20% of time when facing out-of-stocks, whereas in one category they opted not to buy anything 47% of the time. [11] According to a summary of previous research by Gruen et al., customers facing out-of-stocks bought a substitute product 22% to 71% of time depending on the study. Gruen et al. state that the large difference among these studies is due to differing data collection methods and categories examined. [19] Because the substitution rates vary by category, retailers cannot assume the substitution rate to be some constant and plan their assortments using that assumption, but instead they should acknowledge that the substitution rate varies by category.

Substitution is usually not modeled independently, but instead it is incorporated into assortment optimization methods [20]. Since substitution is a significant factor in assortment planning, incorporating that into the assortment planning model is important. Accurately estimating substitution is extremely difficult [22], but different mathematical models have been developed that are used in modeling assortment-based substitution. The multinomial logit (MNL) model is typically used to model assortment-based substitution [20]. Ryzin et al. use a MNL model to optimize an assortment under assortment-based substitution with a case of identical prices [34]. Hopp and Xu use the MNL model for a joint assortment optimization and pricing problem under assortment-based substitution, and they show that the optimal assortment also depends on whether the retailer is risk-neutral or risk-averse [21]. In addition to the MNL model, other models are also used in modeling assortment-based substitution. Anupindi et al. use a probit model to optimize a retail assortment under assortment-based substitution, and they show that including customer disutility from having to substitute their favorite product with another product is informative in assortment optimization [5]. Gaur and Honhon use a locational choice model to find the optimal assortment under assortment-based substitution, and they show that the optimal assortment has no assortment-based substitution regardless of customer preferences, and that the optimal assortment can be such that some customers choose not to buy anything from the assortment, and that the optimal assortment does not necessarily include the most popular product [18].

Chapter 3

Methods

Kök & Fisher state that their research is the only one where substitution rate for assortment-based substitution has been estimated directly [26]. This chapter presents their method for estimating substitution rate in detail. The method used in this thesis to estimate substitution rate is based on this method by Kök & Fisher.

Kök & Fisher estimate substitution rate of a category by comparing demand estimates from multiple stores. They estimate the actual observed demand of stores by applying their demand forecasting model to sales data from a particular store. They call this After-Substitution Demand Estimation (ASDE) because the observed demand in the data from these stores might contain substitution if the store does not have a full assortment available, and therefore the ASDE models will also estimate demand so that substitution is included. [26]

On the other hand, stores that have a full assortment do not have assortment-based substitution since all products are on the shelf and therefore available for customers (Kök & Fisher ignore stock-out-based substitution in their research). They apply their demand forecasting model to data from all full-assortment stores so that the regression coefficients in this model are same for all stores. They also use additional explanatory variables to account for differences between stores. They call this Original Demand Estimation (ODE) since the data used for this model is from full-assortment stores only, and therefore does not contain assortment-based substitution, so the ODE models only estimate the original demand of products without additional demand from assortment-based substitution. [26]

The demand estimates from the ASDE and ODE models are then com-

pared to find out the substitution rate of a category since the ASDE model estimates demand including assortment-based substitution, and the ODE model estimates demand without assortment-based substitution [26]. The choice-based demand estimation model and the substitution rate estimation method used by K ok & Fisher are introduced in sections 3.1 and 3.2, respectively.

3.1 Choice-Based Approach for Demand Forecasting

K ok & Fisher estimate demand for products using a causal mathematical model, which they call a "choice-based model" because it models consumer purchase behavior. The model comprises four different components, which individually model one specific part of customer purchase behavior: How many people visit a specific store on a specific day (footfall), what is the probability of a customer buying a product from a specific subcategory (purchase incidence), what is the probability of a customer buying a specific product given that they buy something from the subcategory (product choice) and what is the expected quantity that the customer buys said product (purchase quantity) [26]. Hierarchical models that model customer behavior this way are also common in marketing literature, e.g. Bucklin & Gupta 1992 [7] and Chintagunta 1993 [9]. However, K ok & Fisher also state that the use of this choice-based model is not necessary for estimating the assortment-based substitution, so it could be replaced with another demand estimation method. The mathematical notation for the choice-based model is shown in Equation (3.1) where D_{jht} is the demand for product j on day t in store h and S_h is the assortment of the store for a specific subcategory. K_{ht} is the number of customers who visit the store on that day (footfall) and $(PQ)_{jht}$ is the average demand for product j per customer. π in the model stands for purchase incidence, p_{jht} is the product choice and q_{jht} the purchase quantity.

$$D_{jht} = K_{ht}(PQ)_{jht} = K_{ht}\pi p_{jht}q_{jht}, j \in S_h \quad (3.1)$$

K ok & Fisher calculate footfall from the number of unique receipts at a store for a given day. This approach assumes that each customer who enters the store buys at least one product and also that each customer buys all their products at the same time, i.e. so that all their bought products are on the same receipt. The correlation coefficients for the footfall of a store h are then

estimated using a log-linear regression, which is shown in Equation (3.2). T_t is the temperature for day t , B_t^l is a binary variable for weekday l that is 1 if day t is said weekday and 0 otherwise. Similarly, E_t^l is a binary variable for a holiday l . HDI is the Human Discomfort Index, which is constructed from hours of sunshine and humidity, but the exact method is not discussed in the research by K ok & Fisher.

$$\ln(K_{ht}) = \kappa_1 + \kappa_2 T_t + \kappa_3 HDI_t + \sum_{l=1}^6 \kappa_{3+l} B_t^l + \sum_{l=1}^{14} \kappa_{9+l} E_t^l \quad (3.2)$$

K ok & Fisher model purchase incidence as a binary choice and use logistic regression to estimate the correlation coefficients for purchase incidence. Because logistic regression gives estimates between 0 and 1, it makes sense to use that for estimating purchase incidence instead of e.g. linear regression. They use temperature, weekday, average promotion level in the subcategory, HDI and information on whether a day is a holiday or not as their explanatory variables for subcategory purchase incidence. The mathematical notation for this model is:

$$\ln\left(\frac{\pi_{ht}}{1 - \pi_{ht}}\right) = \gamma_1 + \gamma_2 T_t + \gamma_3 HDI_t + \sum_{k=1}^6 \gamma_{3+k} B_t^k + \gamma_{10} \bar{A}_{ht} + \sum_{l=1}^{14} \kappa_{10+l} E_t^l \quad (3.3)$$

In their research, K ok & Fisher model product choice with the MNL model, which aims to model the utility of a product based on the characteristics of the product, marketing variables and environmental variables. They use absolute price and promotion of the product compared to the average price and average promotion level of the subcategory as their explanatory variables for the model. [26] They use linear regression with log-centered transformation to model the utility of a product. The log-centered transformation is also used in marketing literature [12]. The mathematical notation of the product choice model is shown in Equation (3.4) where $\bar{p}_{ht} = (\prod_{j \in S} p_{jht})^{1/|S|}$ and $I_{jk} = \{1, \text{ if } j = k; 0 \text{ otherwise}\}$. R_{jht} is the price of product j on day t in store h , A_{jht} is a binary variable which is 1 if the product is on promotion and 0 otherwise. \bar{R} and \bar{A} are the average of the price and the average of the binary promotion variable of all the products of the subcategory the product belongs to.

$$\ln\left(\frac{p_{jht}}{\bar{p}_{ht}}\right) = \sum_{k \in N} \beta_k I_{jk} + \beta_{J+1} (R_{jht} - \bar{R}_{ht}) + \beta_{J+2} (A_{jht} - \bar{A}_{ht}) \quad (3.4)$$

The last part of the choice-based model in Kök & Fisher's research is a linear model for purchase quantity. In their research, Kök & Fisher use holidays, weather and promotion as the explanatory variables in their model. The mathematical notation for the purchase quantity model is shown in Equation (3.5)

$$q_{jht} = \sum_{k \in N} \zeta_k I_{jk} + \zeta_{J+1} A_{jht} + \zeta_{J+2} HDI_t + \sum_{l=1}^{14} \zeta_{J+2+l} E_t^l \quad (3.5)$$

The combined choice-based model is then utilized in modeling demand both with and without assortment-based substitution. This is described in more detail in Section 3.2.

3.2 Substitution Rate Estimation

As discussed earlier, the substitution rate estimation method uses the ASDE model to estimate the demand of products in a subcategory so that the estimates include demand from assortment-based substitution, and the ODE model to estimate the demand of products in a subcategory so that the estimates do not include demand from assortment-based substitution. The demand estimates produced by the ODE and ASDE models for unseen data are then used to estimate the substitution rate.

Kök & Fisher present two models for substitution rate: random substitution and proportional substitution. In both cases, only a single substitution rate is estimated for each subcategory, but the difference in the two models is that in random substitution, it is assumed that the demand from substituting a product is split equally for all other products, whereas in proportional substitution, it is assumed that the demand from substituting a product is split for other products in proportion of the original demand rates of the products. Kök & Fisher also show that the results from the proportional and random substitution models are the same or very close to each other for most observed subcategories. [26] In this thesis, only the random substitution model is used for simplicity.

The substitution rate δ is estimated by minimizing the sum of the squared errors across all stores h and time periods t . The mathematical notation for estimating the substitution rate is shown in Equation (3.6). The estimated substitution rate is capped between 0 and 1 since it is defined as a proportion

of demand, and therefore negative values or values above one would not make sense.

$$\delta^* = \arg \min_{0 \leq \delta \leq 1} \sum_h \sum_t (\hat{y}_{ht}(\delta) - y_{ht})^2 \quad (3.6)$$

y_{ht} is the total after-substitution demand of the products in the subcategory produced by the ASDE model in a store h on a time period t , i.e. $y_{ht} = \sum_{j \in S_h} (PQ)_{jht}$ where S_h is the assortment at a store h . $\hat{y}_{ht}(\delta)$ is described in Equation (3.7) (the time period t is omitted from the equation). The superscript o denotes that the parameter estimates are from the ODE models.

$$\hat{y}_h(\delta) = \sum_{j \in S_h} \left((PQ)_{jh}^o + \sum_{k \in N \setminus S_h} \frac{\delta}{|N|} (PQ)_{kh}^o \right) \quad (3.7)$$

The substitution rate estimation method described in this chapter is used in this thesis to estimate assortment-based substitution for a real-world retail data set but with some simplifications, which are explained in Chapter 4.

Chapter 4

Implementation

In this chapter, the retail data set and its properties are first presented. After that, the implementation of a choice-based and a simple linear regression demand forecasting model are presented, and finally, the implementation of the assortment-based substitution rate estimation method is presented. The accuracies of all the presented models with different explanatory variables are later presented in Chapter 5. All the models are fitted using the Ordinary Least Squares (OLS) method¹.

4.1 Data Description

The raw data is from a grocery retailer, and it comprises transactions from 10 different stores for 8 weeks. Product-locations that have only individual sales during this period have been removed from this data as well as special products like coupons and free items. In addition, category- and product-specific models are only constructed for three subcategories. These subcategories were chosen since the products in these subcategories had the most average sales per product-location, but also because the assortments for these subcategories had at least some variance between the stores. Having variance between the assortments in different stores is necessary for calculating the substitution for products. This is because some models are trained with the data from all stores and some only from the stores with a full assortment, and the difference in the output of these models is used in estimating the

¹The Ordinary Least Squares (OLS) method yields the regression coefficients of the models so as to minimize the the summed square differences of the estimated and actual values of the explained variable.

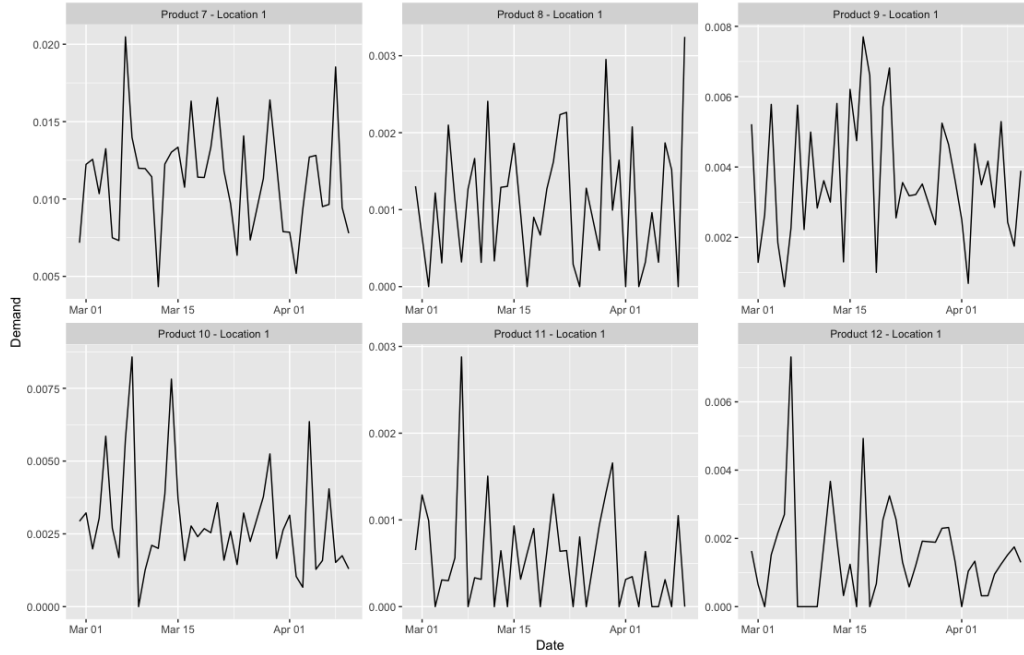


Figure 4.1: Product-location relative demand as a function of time for selected products. Since the daily demand is low, the variance in demand is very high.

substitution. The raw data is split into 6 weeks of data for training the models and 2 weeks for validating the models' performance.

The raw data has the information of a single receipt split into multiple lines where each line comprises one product with information about the value of the bought product, how many units of said products were bought, the sub-group in which the product belongs to, information on whether the product was in promotion at the time of purchase, the location of the transaction and the date of the transaction. In addition, weather data for the store locations was used for capturing the effect of weather.

The raw data was aggregated in order to be able to use it for models that describe location, group-location or product-location level attributes.

Figure 4.1 shows relative demand, i.e. the average amount purchased per customer for a specific date, for selected product-locations. As can be seen from the figure, the demand varies heavily for each day, which makes predicting the exact demand difficult. The variation can also be observed in Table 4.1, which shows that the standard deviation of demand is large relative to the mean of the demand.

Table 4.1: Mean and standard deviation of product-location relative demand for selected products.

Product-location	Demand mean	Demand standard deviation	N
Product 7 - Location 1	0.01129	0.00347	41
Product 8 - Location 1	0.00111	0.00086	41
Product 9 - Location 1	0.00370	0.00179	41
Product 10 - Location 1	0.00291	0.00185	41
Product 11 - Location 1	0.00056	0.00060	41
Product 12 - Location 1	0.00150	0.00146	41

4.2 Footfall

Footfall data is constructed from the raw transaction-level data by aggregating it based on location and date and calculating the footfall based on the number of unique receipts for that location and date.

In the data used in this thesis, footfall depends heavily on location and weekday. For example on Sundays, the average footfall is significantly lower than on Saturdays. Therefore, making the footfall estimation models location- and weekday-specific, or including location and weekday as binary explanatory variables likely increases prediction accuracy when estimating footfall. The mean and standard deviation of footfall per location and per weekday can be seen in the appendix in Table A.1 and Table A.3 respectively.

Temperature on the other hand does not have a statistically significant correlation with footfall for any location, which indicates that adding it as an explanatory variable likely does not increase prediction accuracy. The correlation coefficients and p values per location can be seen in the appendix in Table A.2 and a footfall-temperature scatter plot in Figure A.1.

With these variables, the mathematical notation for the model for estimating footfall for a store h on a day t can be seen in Equation (4.1) where B^l is an index for weekday l and κ are the coefficients to be estimated. Since the model only has binary variables as explanatory variables, the model will estimate a constant footfall for a store and weekday, which is equal to the average footfall of store h on a specific weekday.

$$K_{ht} = \kappa_1 + \sum_{i=1}^6 \kappa_{l+1} B_t^l \quad (4.1)$$

4.3 Choice-Based Model for Estimating Demand

This section describes the implementation of the choice-based model. The mathematical notation of the complete choice-based model is shown in Equation (3.1). In this chapter, the purchase incidence, product choice and purchase quantity models are described, i.e. the $(PQ)_j$ part of that model.

4.3.1 Purchase Incidence

As discussed in Section 3.1, purchase incidence within a subcategory means the probability of a customer buying a product from a subcategory. In the data used in this thesis, purchase incidence depends on the group and weekday. Therefore, including weekday and group as explanatory variables, or making the models group- and weekday-specific is likely to increase the prediction accuracy. Figure A.2 in the appendix shows a box plot for purchase incidence by group and weekday.

The p values in Table 4.2 indicate that there is no statistically significant correlation between temperature and purchase incidence for Group 1, but for Group 2 and Group 3 the correlation is statistically significant. Therefore, adding temperature might increase prediction accuracy for Group 2 and Group 3, but it might make Group 1 overfit.

The data also contains information about the promotion level in the subcategory, i.e. the proportion of products in the subcategory that are on promotion. Intuitively, it would make sense that higher promotion level would cause higher purchase incidence, but the correlation coefficients and p values in Table 4.3 indicate that there is no statistically significant correlation between promotion level and purchase incidence for Group 1, and that there is even a slight negative correlation for Group 2 and Group 3.

With the results of this analysis, the implementation of the model for estimating purchase incidence is shown in Equation (4.2). Since the model only includes binary variables, it is not necessary to use logistic regression

Table 4.2: Pearson correlation coefficient, N and p value for temperature and purchase incidence by group. The p values indicate that the correlation between temperature and purchase incidence is not statistically significant for Group 1, but it is statistically significant for Groups 2 and 3.

Group	Correlation	N	p value
Group 1	0.00	550	0.97
Group 2	0.26	550	$2.89 * 10^{-10}$
Group 3	0.10	550	0.02

Table 4.3: Pearson correlation coefficient, N and p value for promotion level and purchase incidence by group. The p values indicate that the correlation between promotion level and purchase incidence is not statistically significant for Group 1, but it is statistically significant for Groups 2 and 3.

Group	Correlation	N	p value
Group 1	-0.02	550	0.67
Group 2	-0.12	550	0.01
Group 3	-0.16	550	$2.00 * 10^{-4}$

since linear regression will yield the same results. In fact, the estimates of this model will be simply the average purchase incidence for a location on a specific weekday, which is similar to the footfall estimation model in the sense that they both simply estimate averages values of the past data. The model is fitted individually for each subcategory. The results of the model in Section 5.2 also include the accuracy of the model when temperature and promotion level are included as explanatory variables for comparison.

$$\pi_{ht} = \gamma_1 + \sum_{i=1}^6 \gamma_{l+1} B_t^l \quad (4.2)$$

4.3.2 Product Choice

As discussed in Section 3.1, product choice means the probability of a customer buying a product given that they buy something from the subcategory the product belongs to. In the data used in this thesis, the value difference, i.e. the difference between the average price of the subcategory and the price of the product, and product choice have a statistically significant positive correlation for most products in groups 1 and 2 but not for Group 3. This indicates that customers are more likely to buy a product if its value compared to similar products decreases. However, there are some products with statistically significant negative correlation, which indicates that for those products lowering the price of the product (or increasing the price of other similar products) causes customers to buy the product less likely. One explanation for this counter-intuitive behavior can be that there is not enough variation in price, and therefore not enough data points for those product-locations during the period in the data, which causes the correlation to be influenced by randomness in the data. It might also be because of some other variable that affects both value difference and product choice. The correlation coefficients and p values for value difference for all products can be seen in the appendix in Table A.4.

Promotion difference for a product is defined as the difference between the binary promotion variable, which indicates whether the product is on promotion, and the promotion level of the subcategory. The data analysis done indicates that for most products, there is no statistically significant correlation between promotion difference and product choice. The p values and correlation coefficients for promotion difference for all products can be seen in Table A.5 in the appendix.

In the data used in this thesis, product choice varies heavily by product and slightly by weekday, which indicates that the product choice models should be product- and weekday-specific, or alternatively product and weekday should be included as explanatory variables. A box plot of product choice by product and weekday can be seen in the appendix in Figure A.4.

The mathematical notation for the product choice -model is shown in Equation (4.3) where β_{jh} is the correlation coefficient for product j and store h . Since the model has no explanatory variables, the model will predict a constant product choice for the product, which is equal to the average product choice of said product in the data. The results of the model in Section 5.2 also include the accuracy of the model with explanatory variables for comparison.

$$p_{jh} = \beta_{jh} \quad (4.3)$$

4.3.3 Purchase Quantity

As discussed in Section 3.1, purchase quantity means the average amount a customer buys a specific product given that they buy that product in the first place. In the data used in this thesis, average quantity bought per purchase increases if the product is in promotion for almost all products. Therefore, including promotion as an explanatory variable is likely to increase accuracy of the model. A table including the average bought quantity per purchase by product and promotion can be seen in the appendix in Table A.6.

During the period the data is from, there are only a few holidays, and it is uncertain whether their effect on the average bought quantities per product are similar. Therefore, including holidays as an explanatory variable probably does not increase accuracy of the model. Therefore, only promotion will be used as an explanatory variable in the purchase quantity model in this thesis. Equation (4.4) shows the mathematical notation for the implementation of the model. The results of the model in Section 5.2 also include the accuracy of the model with other explanatory variables for comparison.

$$q_j = \zeta_1 + \zeta_2 A \quad (4.4)$$

4.4 Simple Linear Model for Estimating Demand

As discussed in Section 3.1, it is not necessary to use the choice-based model to estimate substitution, but other demand estimation models can be used as well, such as a simple linear regression model. The benefit of modeling demand using linear regression is that the model will be much simpler and faster to compute compared to the choice-based model. When doing the demand estimation this way, the daily aggregate data can be analyzed directly to find correlations between demand and other variables in the data.

The data analysis done indicates that adding promotion as an explanatory variable might increase prediction accuracy since the demand of a product is different depending on whether the product is on promotion. The correlation coefficients and p values by product for demand and promotion can be seen in the appendix in Table A.7. Other explanatory variables that might increase the prediction accuracy are the price of the product and weekday. The accuracy of the simple linear regression model when predicting demand with these explanatory variables is presented in Section 5.3.

4.5 ODE and ASDE Models for Capturing Substitution Effect

As discussed in Chapter 3, substitution rate is estimated by training different demand estimation models using either data from all stores that have a full assortment in a subcategory or data from just a single store. In this thesis, the choice-based demand estimation model for estimating substitution rate is used. The models that use data from a single store (ASDE models) use simply the models described in previous sections to estimate demand. On the other hand, the models that use data from multiple full-assortment stores (ODE models) will have additional explanatory variables to explain variation in consumer purchase behavior between stores. In this thesis, average items in the basket and average footfall are used as additional location-specific explanatory variables. In the following subsections, the implementations for the three components of the choice-based ODE model are presented.

Table 4.4: Pearson correlation coefficient, N and p value for average footfall and purchase incidence by group. The low p values indicate that the positive correlation between average footfall and purchase incidence is statistically significant.

Group	Correlation	N	p value
Group 1	0.19	550	$7.41 * 10^{-6}$
Group 2	0.31	550	$1.23 * 10^{-13}$
Group 3	0.30	550	$3.39 * 10^{-13}$

Table 4.5: Pearson correlation coefficient, N and p value for average items in the basket and purchase incidence by group. The p values indicate that the negative correlation between promotion level and purchase incidence is not statistically significant for Group 1, but it is statistically significant for Groups 2 and 3.

Group	Correlation	N	p value
Group 1	-0.05	550	0.29
Group 2	-0.44	550	$3.09 * 10^{-27}$
Group 3	-0.09	550	0.03

4.5.1 ODE Purchase Incidence

As described earlier, additional explanatory variables for the ODE models are added to explain the variation in consumer purchase behavior between different stores. The correlation coefficients and p values in Table 4.4 indicate that average footfall has a statistically significant correlation with purchase incidence for all groups, and the correlation coefficients and p values in Table 4.5 indicate that purchase incidence has a statistically significant negative correlation with average items in the basket for Group 2 but no clear correlation for the other groups. The correlation between average items in the basket and purchase incidence for Group 2 can also be seen in the appendix in Figure A.5. Equation (4.5) shows the mathematical notation for the implementation of the ODE purchase incidence model with these additional explanatory variables.

$$\pi_t = \gamma_1 + \sum_{i=1}^6 \gamma_{i+1} B_t^i + \gamma_8 \bar{K} + \gamma_9 \bar{N} \quad (4.5)$$

4.5.2 ODE Product Choice

The ODE product choice model uses the same additional location-specific explanatory variables as the ODE purchase incidence model. The analysis done indicates that the correlation between the average footfall of a store and product choice varies heavily by product. For some products, there exists a statistically significant correlation between the average footfall of a store and product choice, but for some products, there is no correlation. This indicates that adding average footfall as an explanatory variable might increase prediction accuracy for some products, but could cause overfitting for some products. More detailed results of the analysis for average footfall and product choice by product can be seen in the appendix in Table A.8 and Figure A.6.

Average items in the basket on the other hand has a more clear correlation with product choice. The data analysis done indicates that in the data used in this thesis, average items in the basket has a statistically significant correlation with product choice for most of the products in Group 1 and Group 2 but not for the products in Group 3. More detailed results can be seen in the appendix in Table A.9 and Figure A.7.

Equation (4.6) shows the mathematical notation for the implementation of the ODE product choice model with these additional variables.

$$p_j = \beta_1 + \beta_2 \bar{K} + \beta_3 \bar{N} \quad (4.6)$$

4.5.3 ODE Purchase Quantity

The last component of the choice-based model is the purchase quantity. Without any additional explanatory variables the mathematical notation of the model is as seen in Equation (4.4).

In the data used in this thesis, there is a varying level of correlation between the average footfall of a store and purchase quantity for products. For most products with statistically significant correlation, the correlation coefficient is

negative, which indicates that in smaller stores customer buy a larger number of the same product at once than in larger stores. The detailed results for the correlation between the average footfall and purchase quantity can be seen in the appendix in Table A.10 and Figure A.8. On the other hand, for most products there is no statistically significant correlation between average items in the basket and purchase quantity. The detailed results for these variables can be seen in the appendix in Table A.11 and Figure A.9.

With these additional explanatory variables added, the mathematical notation for the implementation of the ODE purchase quantity model is as shown in Equation (4.7).

$$q_j = \zeta_1 + \zeta_2 A + \zeta_3 \bar{K} + \zeta_4 \bar{N} \quad (4.7)$$

4.6 Calculating Substitution Rate

As discussed in Chapter 3, the ODE models model demand when there is no substitution, whereas ASDE models model demand so that substitution is included in the estimated demand. When using both ODE and ASDE models to model demand for products in a location with less than a full assortment in a subcategory, the difference in the estimated demands will be the amount of substitution in a subcategory. The difference in ODE and ASDE model estimates is illustrated in Figure 4.2, which shows the demand estimates of the best performing ODE and ASDE models for all product-location-dates for one group. The figure shows that on average, the demand estimates produced by the ASDE model are higher than those produced by ODE model, which means that there is some substitution in the subcategory. The line in the figure is a reference line where the ODE and ASDE demand estimates are equal to each other.

In the substitution rate estimation method described in Section 3.2, substitution rate was defined as the proportion of demand from products not in the assortment that is split for other products in the subcategory, which might include other products not in the assortment. In this thesis, the substitution rate estimation is slightly simplified by defining the substitution rate as the proportion of demand from products not in the assortment that is split for products in assortment. This means that substitution is only accounted in the cases where a customer substitutes a product that they were going to buy with another product from the subcategory that is in assortment, but not in

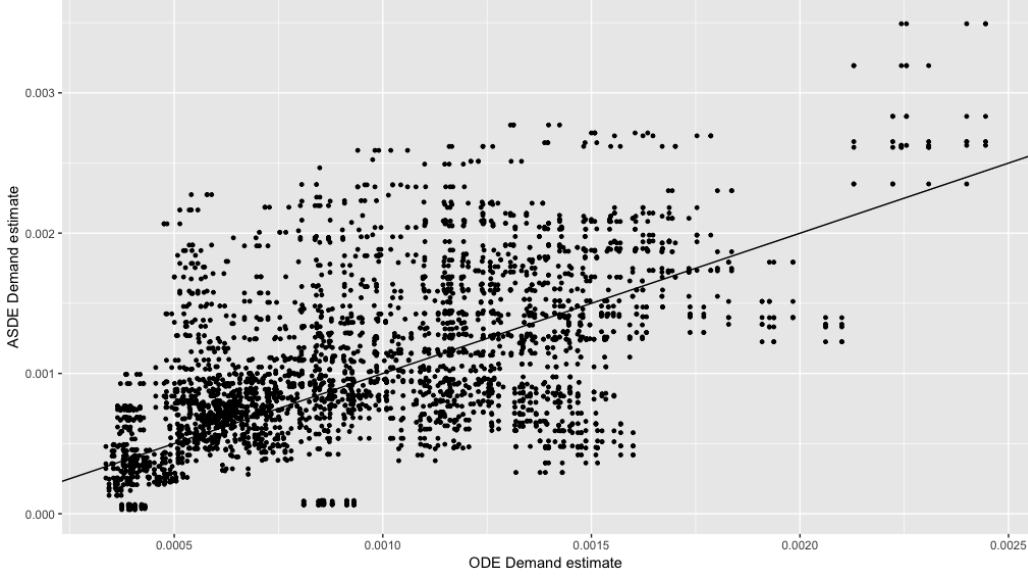


Figure 4.2: ASDE and ODE estimates for relative demand for a group. The reference line shows where the ODE and ASDE demand estimates are equal to each other.

the cases where the substitute product is also not in the assortment.

With this simplification, the absolute amount of substitution in a subgroup from the products not in the assortment will be the sum of the demand estimates for a subgroup produced by the ASDE models subtracted by the sum of the demand estimates produced by the ODE models. The substitution rate estimation described in Section 3.2 minimizes the squared difference in the demand estimates produced by the ODE models summed with the substitution coming from products not in the assortment and the demand estimates produced by the ASDE models for each location and date. As discussed earlier, in this thesis, the substitution rate estimation is slightly simplified. The implementation for the mathematical model for estimating substitution rate is shown in Equation (4.8). x_{ht} is the subgroup total demand for a location h and a date t produced by the ODE model, y_{ht} is the subgroup total demand produced by the ASDE model and z_{ht} is the subgroup total demand produced by the ODE model for products not in the assortment respectively. As with the estimation method described in Section 3.2, substitution rate in the implementation in this thesis also is capped between 0 and 1.

$$\delta^* = \arg \min_{0 \leq \delta \leq 1} \sum_h \sum_t (x_{ht} + \delta z_{ht} - y_{ht})^2 \quad (4.8)$$

Chapter 5

Results

In this chapter, the metrics used to evaluate the models are presented, and then the accuracies of all the models from Chapter 4 with different explanatory variables are presented. After that, the results of the substitution rate estimation performed with the best performing components of the choice-based demand estimation model are presented. Finally, In Section 5.6, the results are compared to results from existing research, and the relation to existing research is discussed.

The accuracy of the models are evaluated using the following metrics: Root Mean Square-Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Symmetric Mean Absolute Percentage Error (SMAPE), Pearson Correlation Coefficient, R-Squared (R^2) and Adjusted R-Squared (AR^2). The formulas for these metrics are presented in Equations (5.1 - 5.7) where d_i is the actual value, f_i is the forecasted value for point i , \bar{d} and \bar{f} are the averages for the actual and forecasted values, and k is the number of explanatory variables in the model used to produce the forecasts.

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (d_i - f_i)^2} \quad (5.1)$$

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |d_i - f_i| \quad (5.2)$$

$$\text{MAPE} = \frac{100\%}{n} \sum_{i=1}^n \left| \frac{d_i - f_i}{d_i} \right| \quad (5.3)$$

$$\text{SMAPE} = \frac{100\%}{n} \sum_{i=1}^n \frac{|d_i - f_i|}{(|d_i| + |f_i|)/2} \quad (5.4)$$

$$\text{Cor} = \frac{\sum_{i=1}^n (d_i - \bar{d})(f_i - \bar{f})}{\sqrt{\sum_{i=1}^n (d_i - \bar{d})^2 (f_i - \bar{f})^2}} \quad (5.5)$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (d_i - f_i)^2}{\sum_{i=1}^n (d_i - \bar{d})^2} \quad (5.6)$$

$$\text{AR}^2 = 1 - (1 - R^2) \left(\frac{n-1}{n - (k+1)} \right) \quad (5.7)$$

5.1 Footfall

Figure 5.1 and Table 5.1 show that even the model that predicts a constant footfall for a weekday and location performs very well when predicting footfall for the validation data set. The table also shows that adding temperature as an explanatory variable decreases the prediction accuracy for the validation dataset because of overfitting.

Table 5.1: Footfall model performance

Explanatory variables	RMSE	MAE	MAPE	SMAPE	Cor	R^2
None	364.769	275.964	0.162	0.147	0.749	0.554
B	109.387	79.261	0.042	0.041	0.984	0.960
$B + T$	118.813	87.150	0.046	0.045	0.979	0.952

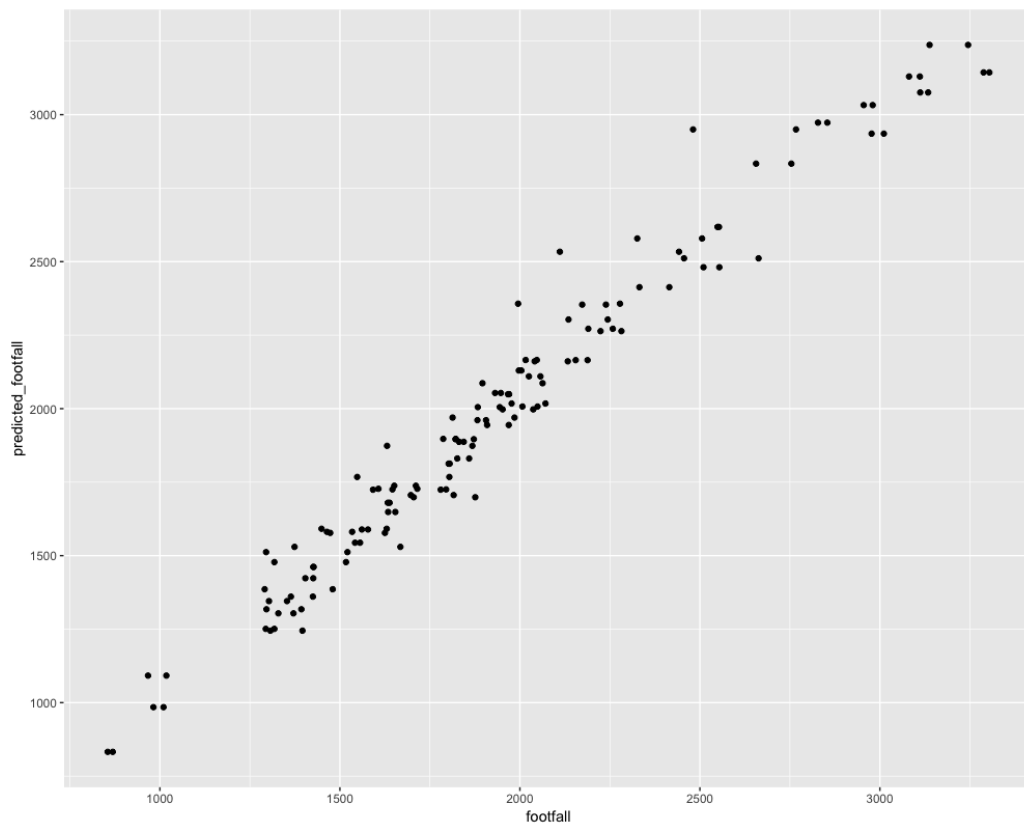


Figure 5.1: Observed and predicted footfall for validation data set

Table 5.2: Purchase incidence model performance

Explanatory variables	RMSE	MAE	MAPE	SMAPE	Cor	R^2	AR^2
None	0.00254	0.00184	0.252	0.234	0.828	0.676	0.676
\bar{A}	0.00254	0.00185	0.255	0.237	0.827	0.674	0.674
T	0.00256	0.00189	0.266	0.238	0.823	0.666	0.666
B	0.00244	0.00178	0.241	0.234	0.845	0.700	0.700

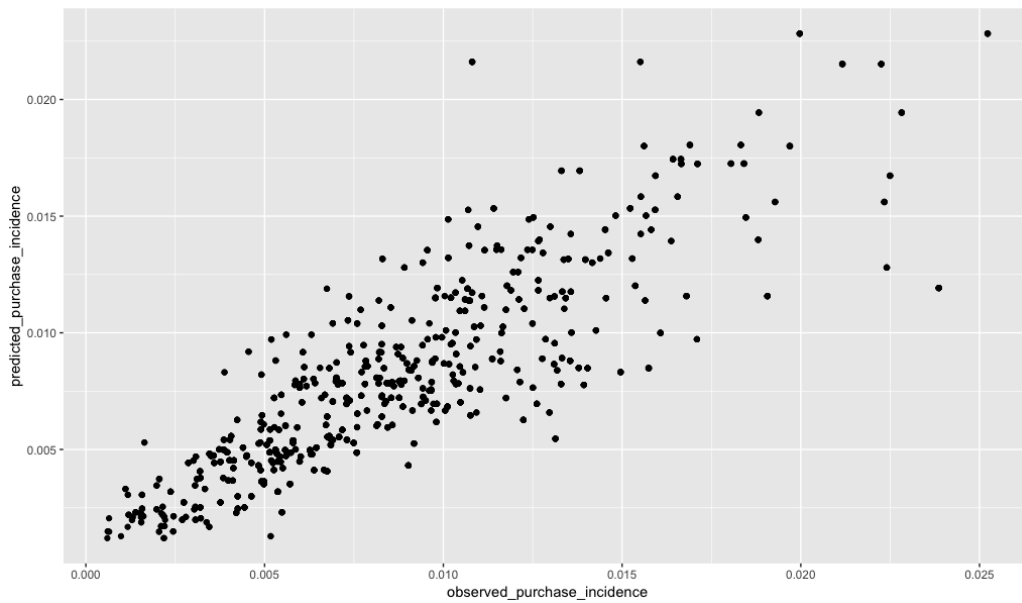


Figure 5.2: Observed and predicted purchase incidence for validation data

5.2 Choice-Based Model

As discussed in Chapter 4, the choice-based model comprises several components, which each model one part of a purchase decision of a customer.

Table 5.2 shows that the only explanatory variable that increases the performance of the purchase incidence model is weekday. However, even a simple model like this provides decent results when predicting purchase incidence, which is illustrated in Figure 5.2 where predicted purchase incidence is shown as a function of observed purchase incidence.

Table 5.3 shows that all explanatory variables provide only marginal improve-

Table 5.3: Product choice model performance

Explanatory variables	RMSE	MAE	MAPE	SMAPE	Cor	R^2	AR^2
None	0.115	0.077	0.424	0.430	0.812	0.637	0.636
$\bar{R} - R$	0.115	0.076	0.425	0.431	0.814	0.641	0.640
$\bar{A} - A$	0.116	0.078	0.438	0.446	0.810	0.635	0.634
B	0.121	0.082	0.463	0.476	0.792	0.599	0.598

Table 5.4: Average quantity model performance for products in promotion

Explanatory variables	RMSE	MAE	MAPE	SMAPE	Cor	R^2	AR^2
None	0.989	0.674	0.370	0.327	0.571	0.315	0.315
A	0.974	0.649	0.348	0.318	0.585	0.339	0.339

ments or even reduce the accuracy of the model for predicting product choice when tested on the validation data set. The models are product-location specific, so in this case their prediction is the average product choice for that product-location during the training period.

Table 5.5 shows that adding promotion as an explanatory variable increases prediction accuracy for the quantity model only marginally. Table 5.4 shows that for those products that have any promotions in the first place, the improvement is slightly larger than for all products but still relatively small. Table 5.5 also indicates that adding holiday or temperature as explanatory variables decreases forecasting accuracy.

Table 5.5: Average quantity model performance for all products

Explanatory variables	RMSE	MAE	MAPE	SMAPE	Cor	R^2	AR^2
None	0.807	0.525	0.313	0.279	0.591	0.336	0.336
A	0.793	0.505	0.296	0.271	0.600	0.359	0.359
H	0.822	0.532	0.316	0.282	0.575	0.311	0.311
T	0.862	0.560	0.335	0.296	0.533	0.243	0.243

5.3 Demand Forecasting Models' Performance

Table 5.6 describes the performance of the whole choice-based model using the best performing purchase incidence, product choice and purchase quantity models. MAPE is not used to measure error here since the data contains zeroes, which would make MAPE be infinite regardless of the predictions of the model. Figure 5.3 shows the predicted demand produced by the choice-based model compared to the actual demand. It can be seen that the model predicts demand accurately for fast-moving products but inaccurately for slow-moving products. The line in the figure is a reference line where the observed demand and the predicted demand are equal.

Table 5.6: Choice-based model performance when predicting demand

RMSE	MAE	SMAPE	Cor	R^2
0.00191	0.00116	0.899	0.899	0.808

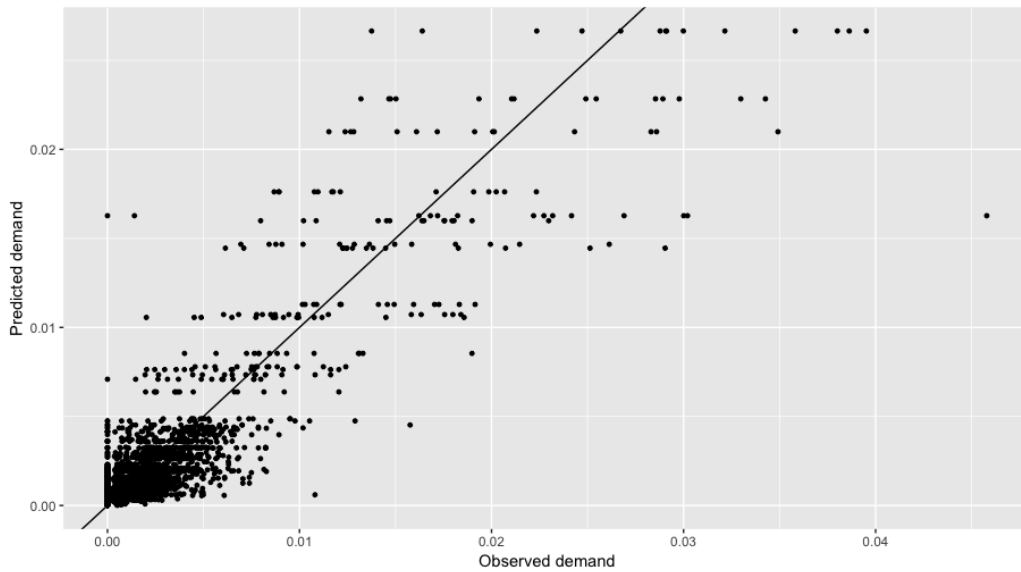


Figure 5.3: Simple linear model predicted relative demand compared to actual relative demand

Table 5.7 describes the performance of the simple linear model with different explanatory variables. As seen from the table, additional variables provide only marginal improvements compared to just taking the average demand of a product-location. As with the choice-based model, the data contains zeroes, so MAPE cannot be used as an error measure. Figure 5.4 shows the predicted demand compared to the observed demand for the simple linear model with no additional explanatory variables. Since there are no explanatory variables, the model predicts a constant demand for a product-location. This can be seen in the figure since the data points form horizontal "lines" for each product-location. As with the figure for choice-based model, the reference line in the figure shows where the observed demand and predicted demand are equal.

Table 5.7: Simple linear model performance when predicting demand

Explanatory variables	RMSE	MAE	SMAPE	Cor	R^2	AR^2
None	0.00205	0.00121	0.892	0.882	0.777	0.777
A	0.00205	0.00120	0.899	0.883	0.779	0.779
B	0.00194	0.00120	0.925	0.896	0.802	0.802
$B + R$	0.00195	0.00121	0.940	0.895	0.799	0.799

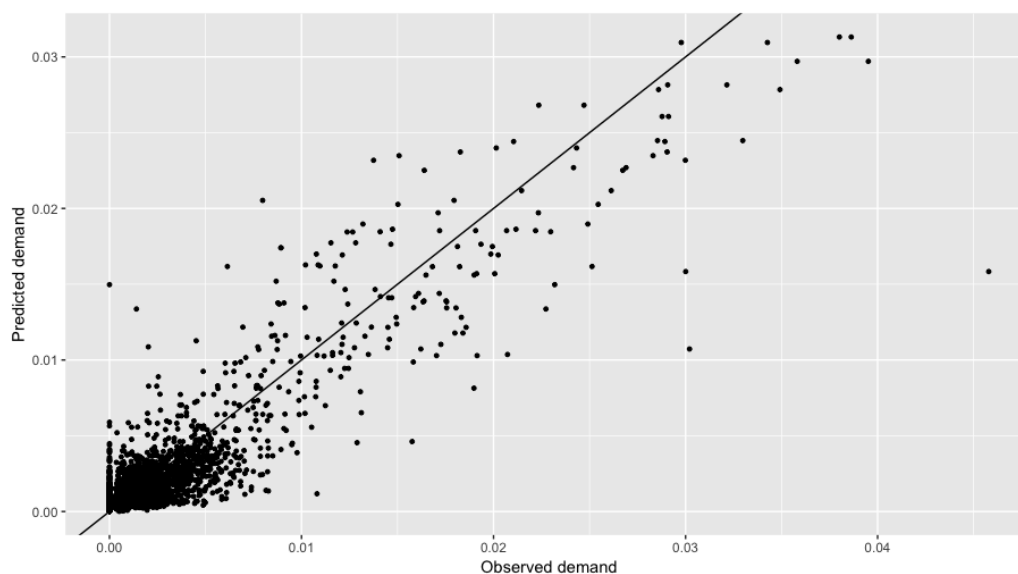


Figure 5.4: Choice-based model predicted relative demand compared to actual relative demand

5.4 ODE Model Performances

The ODE model is constructed from the best performing choice-based model components. This section presents the performances of all the components and finally the performance of the complete ODE model with different location-specific explanatory variables.

Table 5.8 shows the effect of additional explanatory variables to the prediction accuracy when predicting purchase incidence for all locations with the models trained on full-assortment stores. Additional explanatory variables only reduce the performance of the model, and therefore the best model is the one with weekday as the only explanatory variable.

Table 5.9 shows the performance of the ODE product choice model when used to predict product choice for all locations. Adding average items in the basket as an additional explanatory variable slightly increases model performance compared to having no explanatory variables.

Table 5.8: ODE purchase incidence model performance

Explanatory variables	RMSE	MAE	MAPE	SMAPE	Cor	R^2	AR^2
B	0.00353	0.00254	0.481	0.343	0.662	0.430	0.430
$B + \bar{K}$	0.00373	0.00273	0.596	0.367	0.604	0.359	0.359
$B + \bar{N}$	0.00355	0.00256	0.498	0.345	0.657	0.424	0.424
$B + \bar{K} + \bar{N}$	0.00412	0.00291	0.627	0.378	0.515	0.214	0.214

Table 5.9: ODE product choice model performance

Explanatory variables	RMSE	MAE	SMAPE	Cor	R^2	AR^2
None	0.136	0.0937	0.914	0.705	0.488	0.488
\bar{K}	0.141	0.0941	0.929	0.676	0.450	0.450
\bar{N}	0.136	0.0920	0.912	0.706	0.492	0.492
$\bar{K} + \bar{N}$	0.143	0.0943	0.940	0.669	0.437	0.437

Table 5.10 shows the performance of the ODE purchase quantity model when used to predict purchase quantity for all locations. As with the product choice model, adding average items in the basket as an explanatory variable slightly increases model performance for all metrics except MAPE which is slightly better with no additional explanatory variables.

Table 5.11 shows the performance of the complete choice-based ODE model with additional explanatory variables. \bar{N}^* indicates a model where the 'average items in the basket' variable is only added for product choice and purchase quantity models. The best performing model uses 'average items in the basket' as an explanatory variable in those models and no other additional location-specific explanatory variables.

Table 5.10: ODE purchase quantity model performance

Explanatory variables	RMSE	MAE	MAPE	SMAPE	Cor	R^2	AR^2
B	0.865	0.565	0.320	0.300	0.494	0.242	0.242
$B + \bar{K}$	0.864	0.572	0.335	0.303	0.490	0.239	0.239
$B + \bar{N}$	0.861	0.565	0.322	0.300	0.499	0.248	0.248
$B + \bar{K} + \bar{N}$	0.878	0.577	0.340	0.304	0.472	0.215	0.214

Table 5.11: Choice-based ODE model performance when predicting demand

Explanatory variables	RMSE	MAE	SMAPE	Cor	R^2
$A + B$	0.00261	0.00147	1.001	0.803	0.640
$A + B + \bar{K}$	0.00264	0.00150	1.019	0.797	0.630
$A + B + \bar{N}$	0.00255	0.00144	1.000	0.811	0.655
$A + B + \bar{K} + \bar{N}$	0.00284	0.00158	1.039	0.763	0.569
$A + B + \bar{N}^*$	0.00255	0.00144	0.998	0.812	0.656

* \bar{N} only added for product choice and purchase quantity models

5.5 Substitution Rate

Substitution rate is estimated using the best performing choice-based demand estimation model as the ASDE model and the best performing choice-based ODE model. Table 5.12 shows the estimated substitution rates of the subgroups for daily-level data. The substitution rate estimation is a constrained optimization where the values are capped between 0 and 1, which in this case makes value 1 common. Figure 5.5 shows estimation error as a function of substitution rate for Group 3.

Table 5.12: Estimated substitution rates for groups

Group	Estimated substitution rate
Group 1	1.00
Group 2	1.00
Group 3	0.12

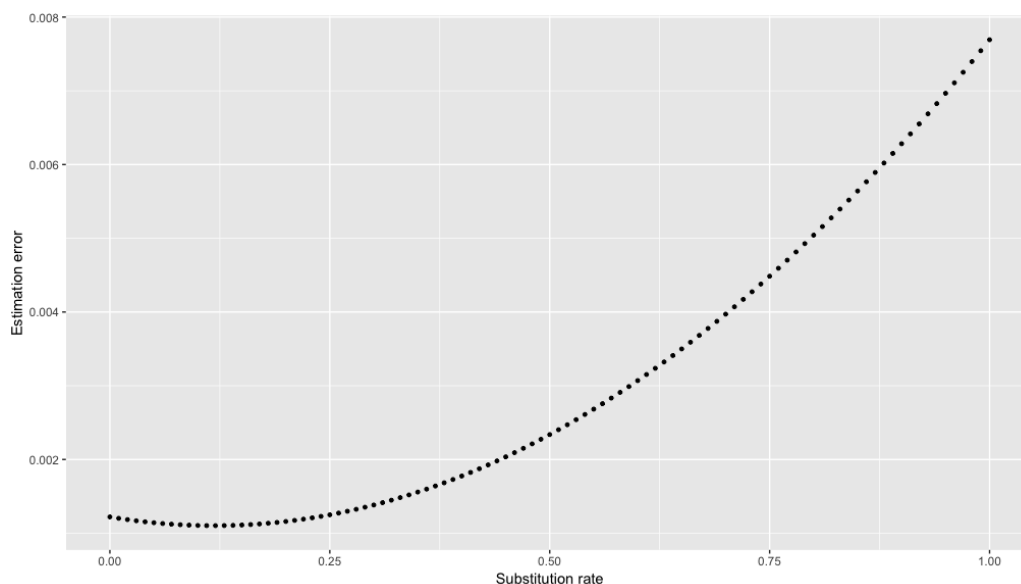


Figure 5.5: Sum of squared errors by substitution rate in substitution rate estimation for Group 3

The results of the substitution rate estimation indicate that there is significant assortment-based substitution in all the three observed subcategories. However, the accuracy of the estimated substitution rate is affected by the accuracy of the ASDE and ODE models used for estimating demand. In this thesis, the demand estimation models could not estimate the demand accurately for the used data set, and therefore the ASDE and ODE models were also somewhat inaccurate with a SMAPE of approximately 0.9 for the ASDE model and over 1.0 for the ODE model. Therefore, the results of the substitution rate estimation might be inaccurate.

5.6 Discussion of Results

As mentioned in Section 2.3, according to an interview study performed by Andersen Consulting, customer substitution behavior varies by category [11]. Therefore, it makes sense for retailers to estimate differing substitution rates for categories instead of assuming a constant substitution rate across subcategories. In Section 2.3, it was also mentioned that substitution is a key factor when deciding which products to include in the assortment of a store [26]. Therefore, it can be concluded that the substitution rate estimation method in this thesis can help retailers in planning their assortments. However, as discussed in Subsection 5.5, the accuracy of the substitution rate estimation method is limited by the accuracy of the demand estimates used in the model.

The accuracy of the substitution rate estimation performed in the research by Kk & Fisher can also be questioned. Kk & Fisher use their substitution rate estimation method to estimate substitution rate for 66 subcategories. In their research, they estimate a substitution rate of 0 for 34 subcategories, substitution rate of 1 for 22 subcategories and a substitution rate between 0 and 1 for only 10 subcategories when assuming random substitution. The substitution rate estimates when assuming proportional substitution are similarly distributed. This indicates that in their research, constraining the optimization affects the end results heavily, and the unconstrained estimated substitution rates would be below 0 or above 1 for the subcategories where the capped substitution rates are 0 or 1, respectively. [26] The results in this thesis are similar to the results from the research by Kk & Fisher in the sense that in both, the estimated substitution rates are 0 or 1 for the majority of the observed subcategories. This might indicate that this method is not suitable for practical applications since the estimated substitution rates

are not necessarily accurate. On the other hand, if a retailer simply needs to know which subcategories have a high substitution rate and which have a low one and not the exact substitution rate, this method could be adequate. Furthermore, even though the solutions for assortment planning that are used today by retailers are very simple, retailers still invest heavily on these systems, which indicates that despite their simplicity, they provide value for the retailers [2]. This indicates that even a simple solution that classifies subcategories to low or high substitution could provide value for retailers in practice since it might help in assortment planning.

Another similarity between the results of this thesis and the research by K ok & Fisher is how the assortments of subcategories differ between stores. In K ok & Fisher’s research, among the 114 subcategories they examined, 48 had a full assortment in all 37 stores, which meant that those subcategories had to be left out from the analysis [26]. The assortments of subcategories in the data set used in this thesis were similar in the sense that it was difficult to find subcategories with differing assortments. This could indicate that assortments that can be used with this methodology are uncommon in retail chains.

Chapter 6

Conclusions

In this chapter, the key findings, validity for real-life applications and ideas for potential future research are discussed.

6.1 Key Findings

Finding 1: Substitution rate estimation relies on accurate demand forecasting. Furthermore, the stores need to be homogeneous in their customer behavior, or there need to be explanatory variables that explain the variation between stores

The estimated substitution rate is calculated using the difference of the demand estimates produced by ODE and ASDE models. In the data used in this thesis, the ASDE model for forecasting demand had a SMAPE of 0.899, which means that the estimates of the model are not very accurate. The ODE model had a SMAPE of 0.998, so the estimates from that model are even less accurate. This means that the substitution rate estimation is not very accurate either because it is built on top of the estimates from the ODE and ASDE models. This is probably one of the biggest reasons that the substitution rate estimates for two of the groups converges towards a very large number, making the result of the estimation counter-intuitive. Furthermore, the stores have to be homogeneous, or there need to be explanatory variables that explain the variation between stores. The results of this thesis show that the prediction accuracy of the ODE models does not improve almost at all

from the additional explanatory variables, which indicates that the variables used in this thesis are not able to explain the variation in customer behavior between different stores.

Finding 2: The choice-based model and the simple linear regression model are equally accurate for estimating demand for slow-moving products

In this thesis, demand was estimated using both a simple linear model and a more complex choice-based model, and the results show that they are approximately equally accurate for estimating demand for the data used. The products in the observed subcategories have low daily demand, so they are so-called slow moving products. Slow-moving products have a very high daily variance, which makes predicting future demand difficult. Another finding was that the simple linear model had the best results with no explanatory variables. With no explanatory variables, the said model predicts future demand to be the average of the observed demand of the product-location that is being forecasted. Also, the choice-based model is marginally more accurate or even less accurate than forecasting the average demand of a product-location depending on which measure of error is used. The linear model with no explanatory variables, i.e. the model that predicts future demand to be the average of the observed demand, has a SMAPE of 0.892, whereas the best performing choice-based model has a SMAPE of 0.899. This result indicates that there are no good explanatory variables to explain the variation in daily demand data for the product-locations in the data used. The data-analysis done also indicates that the variables do not have a clear correlation with the demand. Also, as discussed in Section 4.1, the daily demand varies a lot. This means that predicting the future demand to be the average of the observed demand does not yield very accurate predictions either.

Finding 3: Predicting footfall is very accurate even with weekday as the only explanatory variable

The results of this thesis show that predicting future footfall of a store is very accurate with MAPE of only 0.042 for a model that uses only weekday as an explanatory variable, i.e. a model that predicts the future footfall to be a constant for the same weekday and location. Also, adding

temperature as an explanatory variable decreased prediction accuracy because of overfitting.

Finding 4: The variance in daily demand is high

It was shown in Section 4.1 that the daily variance of demand is high. This can be seen in Figure 4.1 and Table 4.1, where the standard deviation of most product-locations is high compared to the mean. This is because the products are slow-movers, i.e. products that do not have high daily sales, and therefore the daily sales vary because some day there might be no sales at all for a product and the next day there might be several units sold, whereas for products that have high sales per day, the relative change in a high and low daily demand is low, and therefore the variance is low as well.

Finding 5: For noisy and intermittent demand data, simple models perform better for forecasting demand than more complex models

Simpler models, i.e. models with less explanatory variables outperformed more complex models, i.e. models with more explanatory variables, in most cases. For the product choice model, the simple linear model, the ODE purchase incidence model and the ODE purchase quantity model, all explanatory variables decreased the prediction accuracy compared to predicting the average value of the estimated attribute. Also, in the cases where additional explanatory variables increased the prediction accuracy, the increase was tiny compared to not having any explanatory variables.

One explanation for this behavior is that since the sales per day were very low for most products, the data was very noisy and the variation in sales per day was high. Therefore, few variables can pick up the actual causal relationships between the explanatory variable and product demand, and instead they cause overfitting, which reduces the prediction accuracy when testing the model on previously unseen data.

6.2 Validity for Real-Life Applications

The results show that footfall of a location does not vary much for the same weekday. Because of this, predicting future footfall based on the weekday is very accurate and can be useful for example in planning how many workers to have in the store each day. Since footfall does not vary much, it also means that the potential increase in accuracy from more advanced methods is probably not worth the costs because there is not much room left for improvement.

Predicting demand on the other hand was inaccurate mostly because of the products having very little to none sales per day, which made the daily variance in the sales large. To use data science methods to predict future sales, there has to be enough sales for the products. Therefore, using the demand forecasting models like the ones presented in this thesis should be considered only if a retail chain has enough sales per day for the product-locations whose demand they are trying to estimate.

To estimate substitution rate for a subcategory using the methodology described, the retail chain must have more than one store with a full assortment and at least one store with less than full assortment for that subcategory. Therefore, if the retail chain has the same assortment in every store for a subgroup or no store with a full assortment, the substitution rate for that subgroup cannot be estimated using this methodology. Furthermore, the substitution rate estimation works only if the stores of a retail chain are homogeneous in their customer behavior, or if there are explanatory variables that explain the variation of customer behavior between different stores well. There has to be more than one store with a full assortment for a subcategory because otherwise it is impossible to use explanatory variables to explain the variance between stores. With these limitations in mind, estimating substitution rate using these methodologies is not possible for many retailers.

6.3 Potential Future Research

The substitution rate estimation method used requires the data to have a high enough sales per product-location, homogeneous stores or good explanatory variables for stores and at least slightly different assortments in different stores. In the data used, the sales were relatively low, which made the models for estimating demand inaccurate. One potential future research could

therefore be to try out substitution rate estimation with products with higher daily sales to get more accurate estimates for the demand, and therefore also more realistic estimates for substitution rate.

Another potential future research idea would be to use aggregated sales data instead of daily sales. This approach might make the substitution estimation method suitable for slow-moving products as well since the variation in the aggregated sales data is lower than in the daily sales data, and it could also make the substitution rate estimates more accurate overall. Furthermore, regularization could be added to the demand estimation models in order to make the demand estimates more accurate.

In this thesis, assortment-based substitution was estimated but stock-out-based substitution was not. As briefly mentioned in Section 2.3, there are similarities in customer behavior in both cases but also some differences. Therefore, one potential idea for future research would be to estimate stock-out-based substitution using some methodology. Estimating stock-out-based substitution would most likely be simpler than estimating assortment-based substitution since only data from one store where there are some stock-outs would be needed, whereas estimating assortment-based substitution works only in a more specific setting.

One potential future step could also be to find better ways to explain variation between different stores so that models trained with data from all stores would still be accurate. If better ways to explain variation between different stores were found, the results could be utilized in making the substitution rate estimation more accurate. Another benefit would be that models that estimate demand could be trained from the combined data from all stores, which could improve the forecast accuracy of these models.

Another suggestion for future research is to use a different methodology to estimate substitution rate, such as a questionnaire about customer substitution behavior. The results of this method could also be compared to a data science method to see if they produce similar results and to validate the results of the methods.

Bibliography

- [1] ABURTO, L., AND WEBER, R. Improved supply chain management based on hybrid demand forecasts. *Applied Soft Computing* 7, 1 (2007), 136–144.
- [2] AGRAWAL, N., AND SMITH, S. A. *Retail supply chain management*. Springer, 2009.
- [3] ALI, Ö. G., SAYIN, S., VAN WOENSEL, T., AND FRANSOO, J. Sku demand forecasting in the presence of promotions. *Expert Systems with Applications* 36, 10 (2009), 12340–12348.
- [4] ANGERER, A. *The impact of automatic store replenishment on retail: technologies and concepts for the out-of-stocks problem*. Springer Science & Business Media, 2007.
- [5] ANUPINDI, R., GUPTA, S., AND VENKATARAMANAN, M. A. Managing variety on the retail shelf: using household scanner panel data to rationalize assortments. In *Retail Supply Chain Management*. Springer, 2008, pp. 155–182.
- [6] BOATWRIGHT, P., AND NUNES, J. C. Reducing assortment: An attribute-based approach. *Journal of marketing* 65, 3 (2001), 50–63.
- [7] BUCKLIN, R. E., AND GUPTA, S. Brand choice, purchase incidence, and segmentation: An integrated modeling approach. *Journal of Marketing Research* 29, 2 (1992), 201–215.
- [8] CAMPO, K., GIJSBRECHTS, E., AND NISOL, P. Dynamics in consumer response to product unavailability: Do stock-out reactions signal response to permanent assortment reductions? Working Papers 2000024, University of Antwerp, Faculty of Business and Economics, Dec. 2000.

- [9] CHINTAGUNTA, P. Investigating purchase incidence, brand choice and purchase quantity decisions of households. *Marketing Science* 12 (05 1993), 184–208.
- [10] CHRISTOU, I. T. *Quantitative Methods in Supply Chain Management*. Springer London,, 2012.
- [11] CONSULTING, A. Where to look for incremental sales gains: The retail problem of out-of-stock merchandise. *The Coca-Cola Retailing Research Council, Atlanta, GA* (1996).
- [12] COOPER, L. G., AND NAKANISHI, M. *Market-Share Analysis: Evaluating Competitive Marketing Effectiveness. International Series in Quantitative Marketing*. Springer, 1988.
- [13] DIVAKAR, S., RATCHFORD, B. T., AND SHANKAR, V. Practice prize article—chan4cast: A multichannel, multiregion sales forecasting model and decision support system for consumer packaged goods. *Marketing Science* 24, 3 (2005), 334–350.
- [14] FARIAS, V. F., JAGABATHULA, S., AND SHAH, D. A nonparametric approach to modeling choice with limited data. *Management science* 59, 2 (2013), 305–322.
- [15] FILDES, R. A., MA, S., AND KOLASSA, S. Retail forecasting: research and practice. https://mpra.ub.uni-muenchen.de/89356/1/Mpra_paper_89356.pdf/, 2018. Accessed: 2019-07-15.
- [16] FITZSIMONS, G. J. Consumer response to stockouts. *Journal of consumer research* 27, 2 (2000), 249–266.
- [17] FUMI, A., PEPE, A., SCARABOTTI, L., AND SCHIRALDI, M. M. Fourier analysis for demand forecasting in a fashion company. *International Journal of Engineering Business Management* 5, Godište 2013 (2013), 5–30.
- [18] GAUR, V., AND HONHON, D. Assortment planning and inventory decisions under a locational choice model. *Management Science* 52, 10 (2006), 1528–1543.
- [19] GRUEN, T. W., CORSTEN, D. S., AND BHARADWAJ, S. Retail out of stocks: A worldwide examination of extent. *Causes, and Consumer Responses, Grocery Manufacturers of America, Washington DC* (2002).

- [20] HONHON, D., GAUR, V., AND SESHADRI, S. Assortment planning and inventory decisions under stockout-based substitution. *Operations research* 58, 5 (2010), 1364–1379.
- [21] HOPP, W. J., AND XU, X. Product line selection and pricing with modularity in design. *Manufacturing & Service Operations Management* 7, 3 (2005), 172–187.
- [22] HOPP, W. J., AND XU, X. A static approximation for dynamic demand substitution with applications in a competitive market. *Operations Research* 56, 3 (2008), 630–645.
- [23] HÜBNER, A. H., AND KUHN, H. Retail category management: State-of-the-art review of quantitative research and software applications in assortment and shelf space management. *Omega* 40, 2 (2012), 199–209.
- [24] IYENGAR, S. S., AND LEPPER, M. R. When choice is demotivating: Can one desire too much of a good thing? *Journal of personality and social psychology* 79, 6 (2000), 995.
- [25] KALAOGLU, Ö. İ., AKYUZ, E. Ş., ECEMİŞ, S., ERYURUK, S. H., SÜMEN, H., AND KALAOGLU, F. Retail demand forecasting in clothing industry. *Journal of Textile & Apparel/Tekstil ve Konfeksiyon* 25, 2 (2015).
- [26] KÖK, A. G., AND FISHER, M. L. Demand estimation and assortment optimization under substitution: Methodology and application. *Operations Research* 55, 6 (2007), 1001–1021.
- [27] KRAJEWSKI, L., MALHOTRA, M., AND RITZMAN, L. *Operations Management: Processes and Supply Chains, Global Edition*. Pearson Education Limited, 2015.
- [28] MANTRALA, M. K., LEVY, M., KAHN, B. E., FOX, E. J., GAIDAREV, P., DANKWORTH, B., AND SHAH, D. Why is assortment planning so difficult for retailers? a framework and research agenda. *Journal of Retailing* 85, 1 (2009), 71–83.
- [29] 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* 25, 5 (2006), 303–324.
- [30] MILLER, C. M., SMITH, S. A., MCINTYRE, S. H., AND ACHABAL, D. D. Optimizing and evaluating retail assortments for infrequently purchased products. *Journal of Retailing* 86, 2 (2010), 159–171.

- [31] P. CACHON, G., TERWIESCH, C., AND XU, Y. Retail assortment planning in the presence of consumer search. *Manufacturing & Service Operations Management* 7 (10 2005), 330–346.
- [32] PETERSON, R. T. Forecasting practices in retail industry. *The Journal of Business Forecasting* 12, 1 (1993), 11.
- [33] QUELCH, J. A., AND KENNY, D. Extend profits, not product lines. *Make Sure All Your Products Are Profitable* 14 (1994).
- [34] RYZIN, G. V., AND MAHAJAN, S. On the relationship between inventory costs and variety benefits in retail assortments. *Management Science* 45, 11 (1999), 1496–1509.
- [35] STEVENSON, W. *Operations Management. Operations and Decision Sciences*. McGraw-Hill/Irwin, 2011.

Appendix A

Complete Data Analysis Results

Table A.1: Footfall mean and standard deviation by location. The mean of footfall varies for each location, which indicates that using weekday as an explanatory variable most likely increases the prediction accuracy.

Location	Footfall mean	Footfall standard deviation	N
1	2908.98	545.75	41
2	1597.76	385.39	41
3	2247.44	400.56	41
4	2032.46	338.80	41
5	2271.78	427.86	41
6	1954.83	388.59	41
7	1620.17	279.53	41
8	1351.56	261.56	41
9	1827.24	286.42	41
10	2183.78	495.04	41

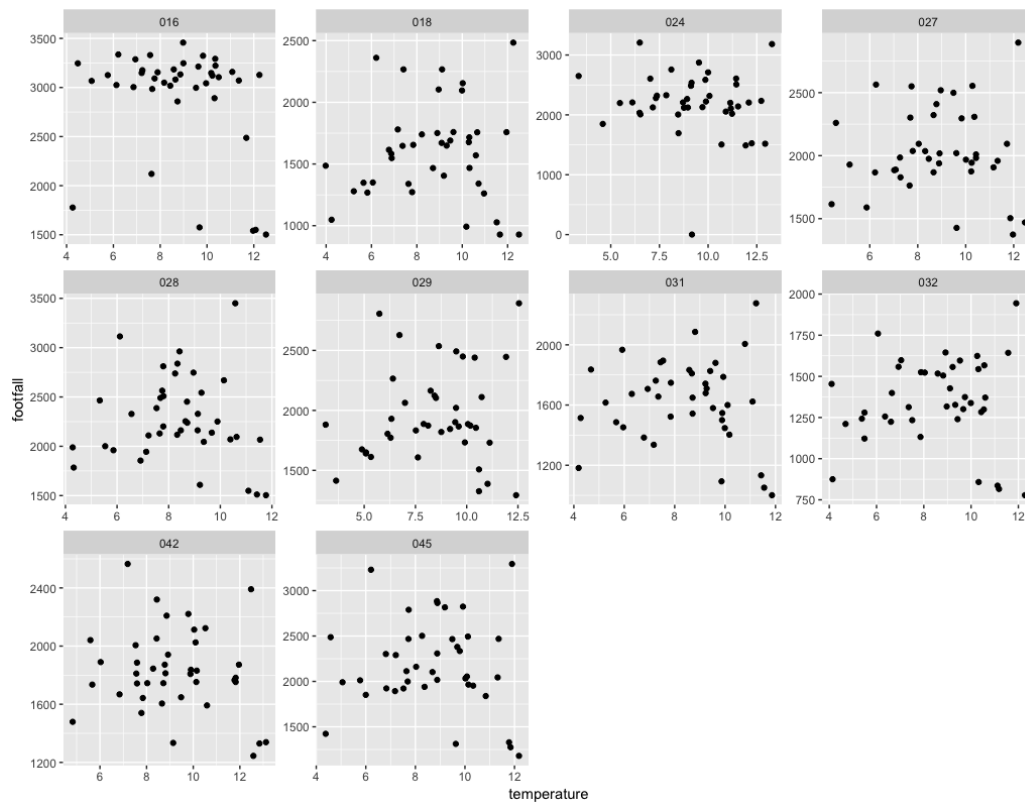


Figure A.1: Footfall-temperature scatter plot. There is no immediately visible correlation between footfall and temperature.

Table A.2: Pearson correlation coefficient, N and p value for temperature and footfall by location. The high p value indicates that the correlation between temperature and footfall is not statistically significant.

Location	Correlation(footfall, temperature)	N	p value
1	-0.08	55	0.56
2	-0.04	55	0.80
3	-0.17	55	0.21
4	-0.09	55	0.50
5	-0.21	55	0.12
6	0.00	55	0.99
7	-0.06	55	0.67
8	-0.00	55	0.98
9	-0.16	55	0.26
10	-0.09	55	0.49

Table A.3: Footfall by weekday. The mean of footfall varies for each day, which indicates that using weekday as an explanatory variable most likely increases the prediction accuracy.

Weekday	Footfall mean	Footfall standard deviance	N
Monday	1851.72	483.74	60
Tuesday	1967.55	467.47	60
Wednesday	1894.13	495.25	60
Thursday	2117.30	528.48	60
Friday	2221.20	471.13	60
Saturday	2513.70	485.15	60
Sunday	1318.00	270.05	50

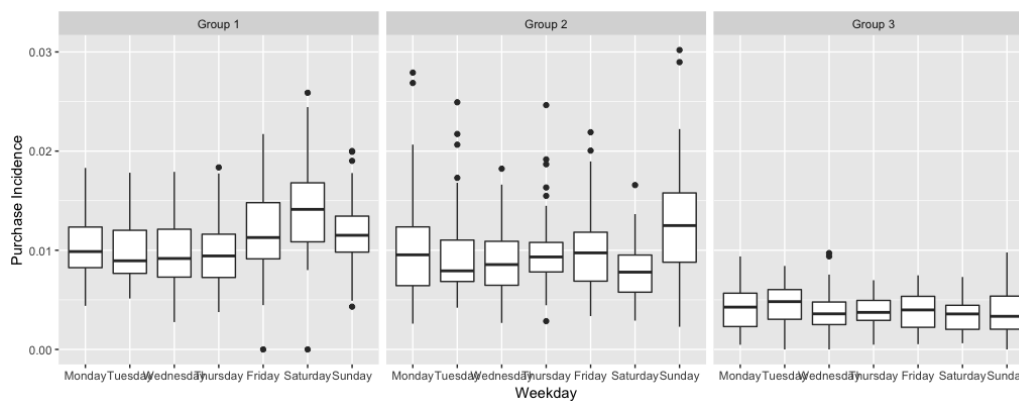


Figure A.2: Purchase incidence by weekday and group. The white boxes indicate the 1st and 3rd quartiles (the 25th and 75th percentiles) and the lines indicate the mean. The plot indicates that purchase incidence varies for both weekday and group.

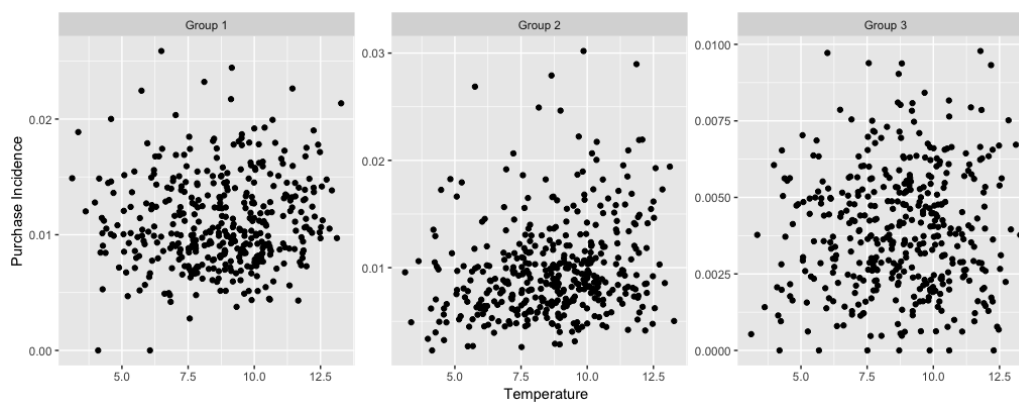


Figure A.3: Purchase incidence by temperature. There is no clearly visible correlation between temperature and purchase incidence.

Table A.4: Pearson correlation coefficient, N and p value for value difference and product choice by product. The majority of products in groups 1 and 2 have statistically significant correlation, but for most of the products in Group 3 there is no statistically significant correlation.

Group	Product	Correlation	N	p value
Group 1	1	0.22	548	$2.80 * 10^{-07}$
	2	0.01	548	0.79
	3	0.22	548	$1.18 * 10^{-07}$
	4	0.17	548	$4.07 * 10^{-05}$
	5	-0.18	440	$1.21 * 10^{-04}$
	6	-0.23	440	$1.10 * 10^{-06}$
Group 2	7	-0.02	550	0.64
	8	0.16	495	$3.31 * 10^{-04}$
	9	0.41	550	$6.64 * 10^{-24}$
	10	0.11	495	0.02
	11	0.34	550	$1.80 * 10^{-16}$
	12	0.18	550	$2.12 * 10^{-05}$
	13	-0.35	220	$7.24 * 10^{-08}$
	14	0.15	110	0.11
	15	0.00	220	0.98
	16	-0.21	220	$1.60 * 10^{-03}$
Group 3	17	-0.07	220	0.32
	18	0.01	440	0.87
	19	0.05	440	0.35
	20	-0.02	440	0.71
	21	0.01	275	0.90
	22	0.06	385	0.21
	23	-0.02	488	0.65
	24	-0.03	275	0.64
	25	0.05	543	0.25
	26	0.17	543	$8.16 * 10^{-05}$
	27	-0.01	330	0.85
	28	0.28	330	$2.76 * 10^{-07}$

Table A.5: Pearson correlation coefficient, amount of distinct values and p values for promotion difference and product choice. The p values indicate that most products have no statistically significant correlation between promotion difference and product choice.

Group	Product	Correlation	Distinct N	p value
Group 1	1	-0.03	8	0.43
	2	0.12	9	0.01
	3	0.02	10	0.60
	4	-0.05	9	0.22
	5	0.02	5	0.68
	6	0.10	7	0.05
Group 2	7	-0.03	5	0.44
	8	0.19	5	$1.69 * 10^{-05}$
	9	0.31	5	$4.11 * 10^{-14}$
	10	0.19	5	$1.84 * 10^{-05}$
	11	0.26	5	$3.06 * 10^{-10}$
	12	0.10	5	0.02
	13	0.01	5	0.94
	14	-0.01	5	0.91
	15	0.01	5	0.88
	16	-0.06	5	0.35
	17	0.14	5	0.04
Group 3	18	-0.14	17	$4.44 * 10^{-03}$
	19	-0.14	20	$2.54 * 10^{-03}$
	20	-0.04	20	0.40
	21	0.10	9	0.10
	22	0.07	15	0.17
	23	0.05	16	0.25
	24	0.02	9	0.80
	25	-0.10	19	0.02
	26	0.05	22	0.24
	27	0.04	14	0.46
	28	0.04	11	0.43

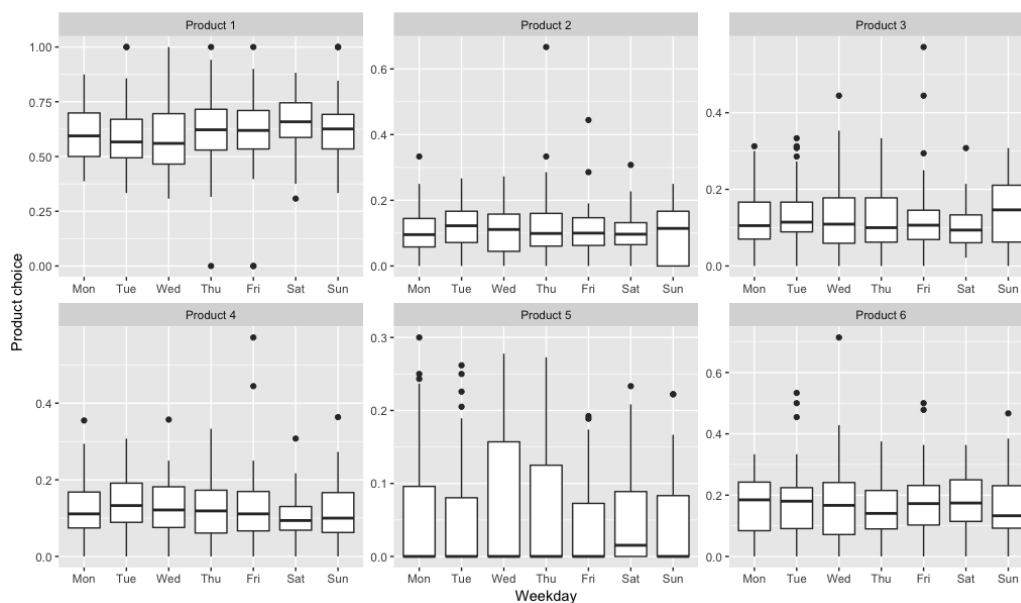


Figure A.4: Product choice by weekday and product for selected products. The white boxes indicate the 1st and 3rd quartiles (the 25th and 75th percentiles) and the lines indicate the mean. Product choice varies significantly by product and slightly by weekday.

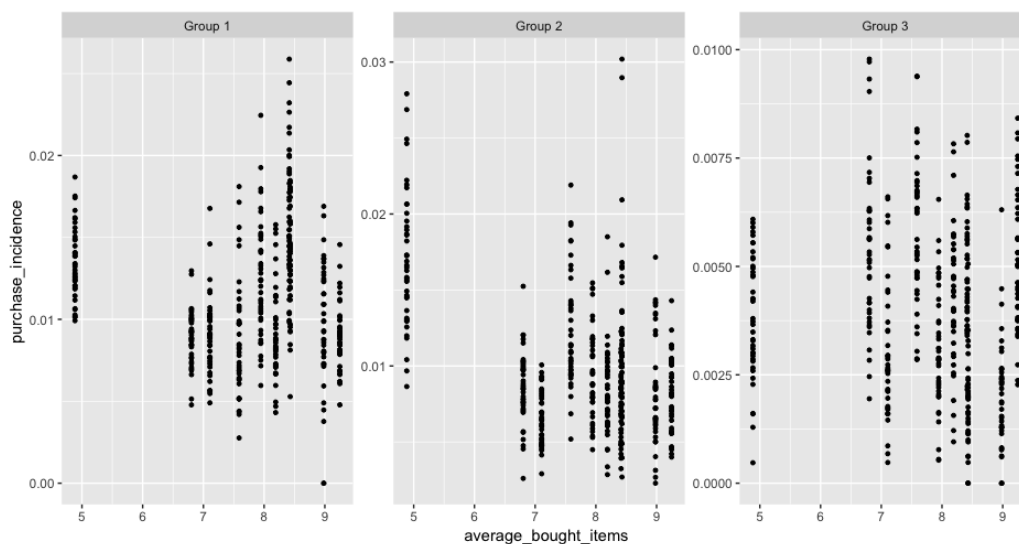


Figure A.5: Purchase incidence by average bought items per group. For Group 2, the average amount of bought items, i.e. average items in basket, seems to negatively correlate with the purchase incidence.

Table A.6: Average purchase quantity by product and promotion. For most products, average purchase quantity increases if the product is on promotion. Some products are not in promotion during the period the data is from so they are not shown in the table.

Product	In promotion	Average purchase quantity
Product 15	0	1.12
Product 15	1	1.56
Product 17	0	1.00
Product 17	1	1.83
Product 18	0	1.34
Product 18	1	1.39
Product 19	0	1.43
Product 19	1	1.41
Product 20	0	1.34
Product 20	1	1.38
Product 2	0	2.67
Product 2	1	2.94
Product 3	0	2.36
Product 3	1	2.74
Product 4	0	2.47
Product 4	1	2.97
Product 25	0	1.61
Product 25	1	1.67
Product 26	0	1.45
Product 26	1	1.75

Table A.7: Pearson correlation coefficient, N, p value for relative demand and promotion. Only some products are both in promotion and out of promotion during the period the data is from, which is why most products have NA correlation coefficient and p value.

Group	Product	Correlation	N	p value
Group 1	1	NA	548	NA
	2	0.14	548	$1.13 * 10^{-03}$
	3	0.08	548	0.07
	4	0.01	548	0.78
	5	NA	440	NA
	6	NA	440	NA
Group 2	7	NA	550	NA
	8	NA	495	NA
	9	NA	550	NA
	10	NA	495	NA
	11	NA	550	NA
	12	NA	550	NA
	13	NA	220	NA
	14	NA	110	NA
	15	0.07	220	0.27
	16	NA	220	NA
	17	0.08	220	0.22
Group 3	18	-0.03	440	0.53
	19	-0.07	440	0.14
	20	0.10	440	0.03
	21	NA	275	NA
	22	NA	385	NA
	23	NA	488	NA
	24	NA	275	NA
	25	0.00	543	0.99
	26	0.21	543	$1.26 * 10^{-06}$
	27	NA	330	NA
	28	NA	330	NA

Table A.8: Pearson correlation coefficient, N and p value for average footfall and product choice by product. The p values and correlation coefficients vary heavily by product.

Group	Product	Correlation	N	p value
Group 1	1	-0.33	548	$1.44 * 10^{-15}$
	2	0.09	548	0.03
	3	-0.03	548	0.49
	4	-0.02	548	0.58
	5	0.58	440	$1.88 * 10^{-40}$
	6	-0.23	440	$6.31 * 10^{-07}$
Group 2	7	0.29	550	$7.50 * 10^{-12}$
	8	-0.34	495	$5.63 * 10^{-15}$
	9	0.06	550	0.17
	10	0.08	495	0.06
	11	-0.34	550	$4.88 * 10^{-16}$
	12	-0.15	550	$4.84 * 10^{-04}$
	13	0.09	220	0.18
	14	0.13	110	0.17
	15	0.14	220	0.03
	16	-0.14	220	0.03
	17	-0.04	220	0.51
Group 3	18	-0.31	440	$4.92 * 10^{-11}$
	19	-0.19	440	$7.66 * 10^{-05}$
	20	-0.23	440	$9.06 * 10^{-07}$
	21	-0.05	275	0.38
	22	0.02	385	0.63
	23	-0.03	488	0.51
	24	-0.13	275	0.04
	25	-0.25	543	$2.30 * 10^{-09}$
	26	-0.38	543	$4.07 * 10^{-20}$
	27	-0.03	330	0.63
	28	0.11	330	0.05

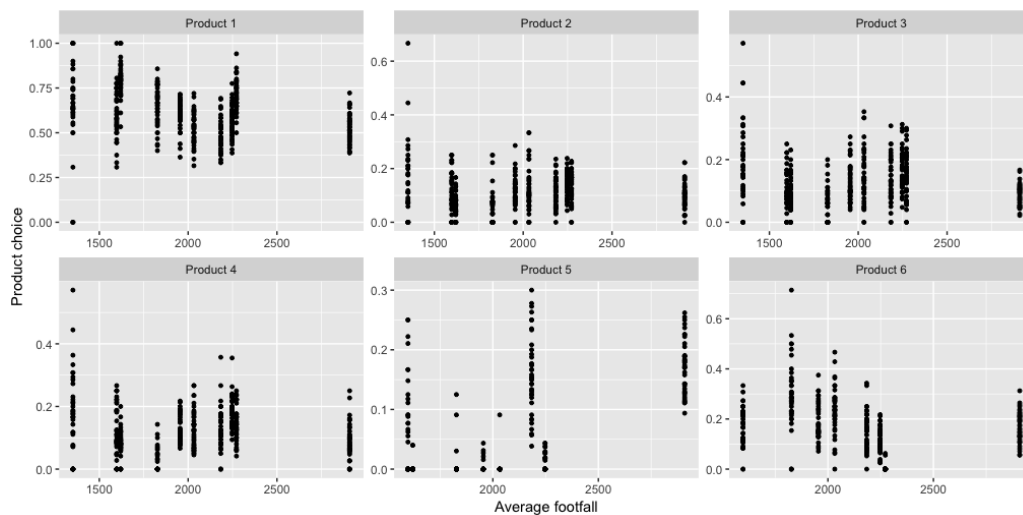


Figure A.6: Product choice by average footfall for selected products.

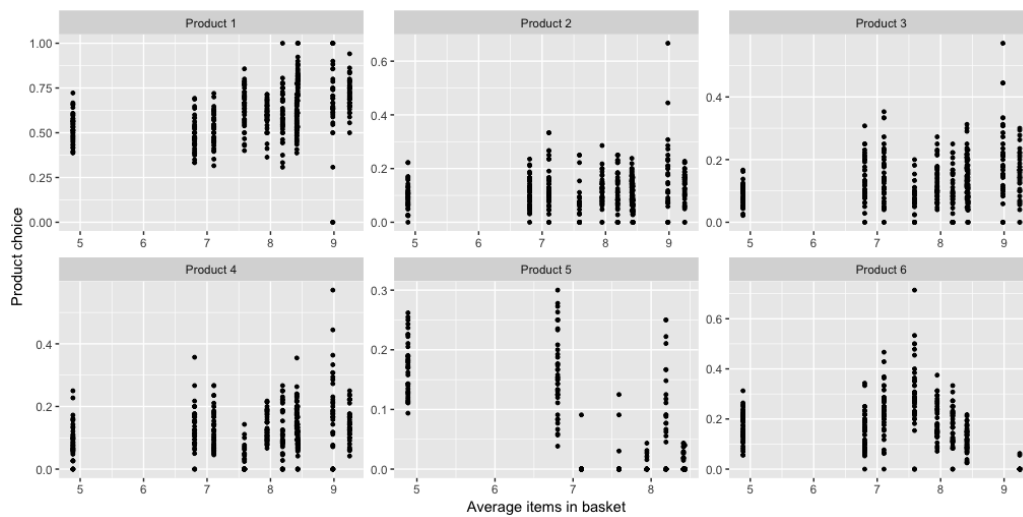


Figure A.7: Product choice by average items in basket for selected products.

Table A.9: Pearson correlation coefficient, N and p value for average items in basket and product choice by product. The low p values in Group 1 and Group 2 indicate that there is statistically significant correlation between average items in basket and product choice in those groups, and the high p values in Group 3 indicate that there is no statistically significant correlation in that group.

Group	Product	Correlation	N	p value
Group 1	1	0.42	548	$3.15 * 10^{-24}$
	2	0.05	548	0.20
	3	0.19	548	$7.06 * 10^{-06}$
	4	0.21	548	$8.46 * 10^{-07}$
	5	-0.68	440	$4.09 * 10^{-61}$
	6	-0.31	440	$1.42 * 10^{-11}$
Group 2	7	-0.23	550	$5.29 * 10^{-08}$
	8	0.30	495	$5.30 * 10^{-12}$
	9	0.05	550	0.25
	10	-0.28	495	$2.31 * 10^{-10}$
	11	0.26	550	$3.37 * 10^{-10}$
	12	0.17	550	$7.56 * 10^{-05}$
	13	-0.16	220	0.01
	14	0.13	110	0.17
	15	-0.24	220	$2.44 * 10^{-04}$
	16	-0.23	220	$4.78 * 10^{-04}$
	17	-0.01	220	0.92
Group 3	18	0.12	440	0.01
	19	0.04	440	0.42
	20	0.01	440	0.82
	21	0.08	275	0.19
	22	0.03	385	0.59
	23	0.02	488	0.73
	24	0.20	275	$6.42 * 10^{-04}$
	25	0.11	543	0.01
	26	0.20	543	$1.91 * 10^{-06}$
	27	-0.06	330	0.26
	28	0.16	330	$2.96 * 10^{-03}$

Table A.10: Pearson correlation coefficient, N and p value for average footfall and purchase quantity by product. The low p values in Group 1 indicate that there is statistically significant correlation between average footfall and purchase quantity in that group. For Group 2 and Group 3 the p values vary, which indicates that in those groups some products have statistically significant correlation and some do not.

Group	Product	Correlation	N	p value
Group 1	1	-0.09	544	0.03
	2	-0.11	468	0.02
	3	-0.12	479	0.01
	4	-0.11	475	0.02
	5	0.19	167	0.01
	6	-0.02	378	0.65
Group 2	7	-0.11	544	0.01
	8	-0.05	385	0.38
	9	-0.04	413	0.44
	10	-0.06	294	0.31
	11	0.01	330	0.87
	12	0.02	398	0.66
	13	-0.20	132	0.02
	14	0.09	55	0.52
	15	-0.31	122	$4.79 * 10^{-04}$
	16	-0.24	109	0.01
	17	-0.24	102	0.02
Group 3	18	-0.10	300	0.08
	19	0.04	275	0.48
	20	0.10	297	0.09
	21	-0.06	225	0.36
	22	-0.12	300	0.04
	23	-0.18	318	$1.06 * 10^{-03}$
	24	-0.24	234	$2.69 * 10^{-04}$
	25	-0.19	416	$1.39 * 10^{-04}$
	26	-0.13	412	0.01
	27	-0.29	132	$6.17 * 10^{-04}$
	28	-0.06	218	0.36

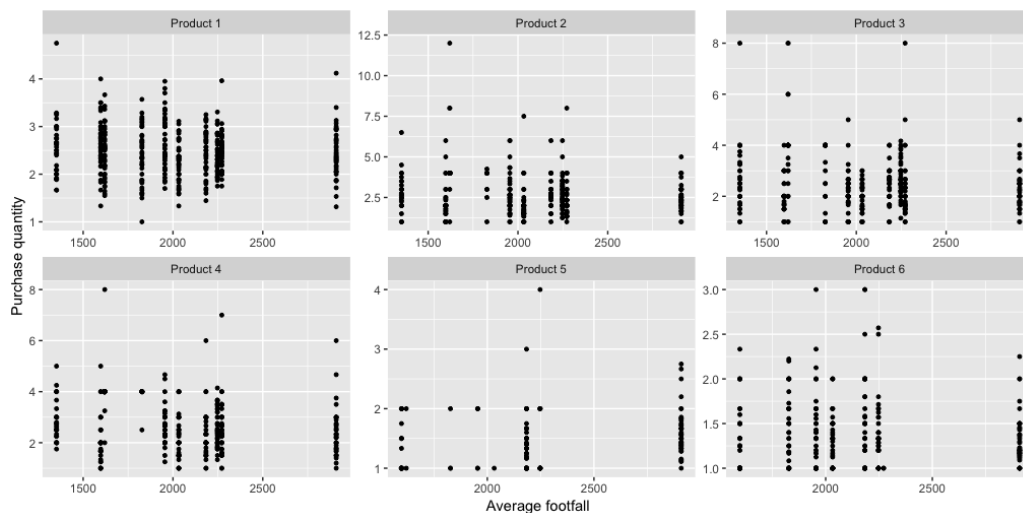


Figure A.8: Purchase quantity by average footfall for selected products.

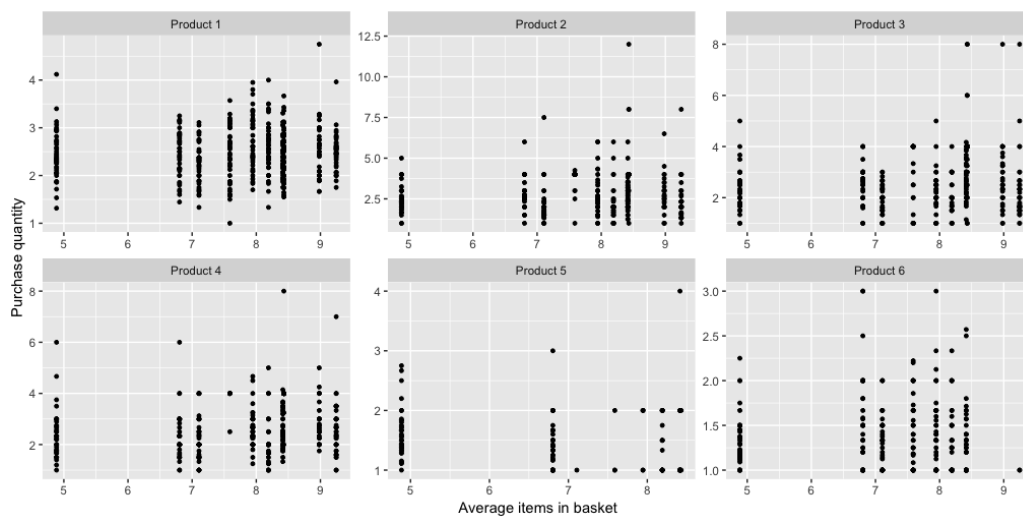


Figure A.9: Purchase quantity by average items in basket for selected products.

Table A.11: Pearson correlation coefficient, N and p value for average items in basket and purchase quantity by product. The high p values indicate that there is no statistically significant correlation for most products.

Group	Product	Correlation	N	p value
Group 1	1	0.08	544	0.05
	2	0.03	468	0.48
	3	0.08	479	0.08
	4	0.06	475	0.21
	5	-0.17	167	0.03
	6	-0.01	378	0.88
Group 2	7	0.09	544	0.04
	8	-0.00	385	0.96
	9	-0.01	413	0.90
	10	0.00	294	0.94
	11	-0.01	330	0.86
	12	-0.06	398	0.25
	13	-0.10	132	0.24
	14	0.09	55	0.52
	15	-0.01	122	0.93
	16	0.01	109	0.92
	17	0.04	102	0.71
Group 3	18	0.03	300	0.59
	19	0.17	275	$4.61 * 10^{-03}$
	20	0.03	297	0.60
	21	0.01	225	0.86
	22	0.03	300	0.59
	23	0.06	318	0.26
	24	0.29	234	$5.94 * 10^{-06}$
	25	0.15	416	$1.82 * 10^{-03}$
	26	0.07	412	0.16
	27	0.15	132	0.09
	28	0.02	218	0.80