Aalto University
School of Science
Master's Programme in Mathematics and Operations Research

Tuuli Aaltonen

Feasibility of using choice models with loyalty card data in quantifying product substitution

Master's Thesis
Espoo, July 12, 2021

Supervisor: Professor Antti Punkka
Advisor: Tuomas Viitanen D.Sc. (Tech.)

Aalto University
School of Science
Master's Programme in Mathematics and Operations Research

ABSTRACT OF MASTER'S THESIS

Author: Tuuli Aaltonen

Title:
Feasibility of using choice models with loyalty card data in quantifying product substitution

Date:	July 12, 2021	**Pages:**	vi + 64
Major:	Systems and Operations Research	**Code:**	SCI3055
Supervisor:	Professor Antti Punkka		
Advisor:	Tuomas Viitanen D.Sc. (Tech.)		

In today's competitive retail market, retailers constantly thrive for more and more optimised operations to maximise their profits. Substitute products play an important role in many retail optimisation problems, such as sales forecasting, campaign planning and assortment optimisation. The accurate detection of substitute product pairs and measuring the magnitude of the substitution effect is therefore crucial for retailers.

This thesis evaluates the feasibility of using choice modelling in measuring the magnitude of product substitution, when there is customer-specific choice data available through loyalty card data. Three choice models based on multinomial logit and probit models are developed to model product choice probability within three product categories of different sizes. The performance of the models is evaluated by comparing their accuracy in forecasting sales decreases of products, which are induced by price discounts of their substitute products. The forecasting accuracy of the models is compared to a benchmark sales correlation method which only utilises daily aggregate sales data in forecasting the sales. For data privacy reasons, the research was conducted on data sets that were generated using simulation.

The results of the thesis indicate that it is possible to gain improvements in quantifying substitution when choice models are used together with customer-specific choice data in comparison to only using daily aggregate sales. However, the models did not perform well in the larger product categories, so further research is needed on the possibility of reducing the number of product alternatives using, for instance, product attribute data. The models also need to be validated with real retail data in order to account for the phenomena in customer behaviour that could not be simulated.

Keywords: retail, substitution, loyalty card data, demand forecasting, choice modelling, multinomial logit, multinomial probit

Language: English

Aalto-yliopisto
Perustieteiden korkeakoulu
Matematiikan ja operaatiotutkimuksen maisteriohjelma

DIPLOMITYÖN TIIVISTELMÄ

Tekijä: Tuuli Aaltonen

Työn nimi:
Valinnan mallien soveltuvuus tuotesubstituution mittaamiseen kanta-asiakaskorttidatan avulla

Päiväys:	12. heinäkuuta 2021	**Sivumäärä:**	vi + 64
Pääaine:	Systeemi- ja operaatiotutkimus	**Koodi:**	SCI3055
Valvoja:	Professori Antti Punkka		
Ohjaaja:	TkT Tuomas Viitanen		

Nykypäivän vähittäistavarakaupassa vallitsee ankara kilpailu, mikä saa vähittäistavarakauppiaat pyrkimään kohti yhä optimoidumpia operaatioita voittojen maksimoimiseksi. Substituuttituotteilla on tärkeä rooli monissa vähittäistavarakaupan optimointiongelmissa, kuten kysynnän ennustamisessa, kampanjoiden suunnittelussa ja valikoiman optimoimisessa. Substituuttituoteparien tunnistaminen sekä tuotteiden välisen substituution voimakkuuden mittaaminen on siis äärimmäisen tärkeää vähittäistavarakauppiaille.

Tässä diplomityössä arvioidaan diskreettien valinnan mallien soveltuvuutta tuotteiden välisen substituution voimakkuuden mittaamiseen, kun käytössä on kanta-asiakaskorttiohjelman avulla kerättyä asiakaskohtaista valintadataa. Kolme valinnan mallia kehitetään mallintamaan tuotteiden valinnan todennäköisyyttä kolmessa eri kokoisessa tuoteryhmässä. Mallien suorituskykyä arvioidaan vertailemalla niiden tarkkuutta ennustaa myyntiä tilanteissa, joissa tuotteen myynti laskee substituuttituotteen hinnanalennuksen takia. Mallien ennustustarkkuuden vertailukohtana pidetään yksinkertaisen, ainoastaan päiväkohtaisia myyntejä hyödyntävän myyntikorrelaatiomallin ennustustarkkuutta. Tietosuojasyistä tutkimuksen data generoitiin simuloimalla.

Tutkimuksen tulokset osoittavat, että valinnan malleilla on mahdollista parantaa substituution voimakkuuden arvioita, kun käytössä on päivittäisten myyntien lisäksi asiakaskohtaista valintadataa. Valinnan mallit eivät kuitenkaan suoriutuneet hyvin suurissa tuotekategorioissa, joten mahdollisuuksia vähentää tuotevaihtoehtojen lukumäärää esimerkiksi tuoteattribuuttien avulla tulisi tutkia lisää. Mallit tulisi myös validoida todellisella kanta-asiakaskorttidatalla, jotta voitaisiin ottaa huomioon asiakkaiden käyttäytymisen ilmiöt, joita ei ollut mahdollista simuloida.

Asiasanat:	vähittäiskauppa, substituutio, kysynnän ennustaminen, valinnan mallintaminen, multinomiaalinen logit-malli, multinomiaalinen probit-malli
Kieli:	Englanti

Acknowledgements

This thesis has been the most challenging project of my life but also perhaps the most rewarding one to finish. During this thesis, I have gained confidence in my writing, and writing is one of the things I will miss most moving forward. Perhaps I need to come up with some new writing projects in the future.

There are some incredible people whose support made this thesis possible. I would like to thank my employer for providing me with a great topic and the means to make this thesis possible. I want to thank my thesis advisor Tuomas for your continuous support and guidance throughout this project and for the invaluable feedback on all drafts of this thesis. Likewise, I would like to thank my supervisor Antti for the great feedback as well as for providing a set of fresh eyes from outside the retail bubble.

Finishing this thesis also concludes my studies at Aalto University, and I would like to give my thanks to everyone who made these six years the best ones of my life. Words cannot express the gratitude I have for the Guild of Physics, which has been the most intelligent, innovative and fun community I could have ever hoped to be a part of. Volunteering in the Guild has provided me invaluable lessons in leadership and teamwork that I could have never gotten otherwise. Thank you to Raati 18 for the incredible board year and all the great times ever since and to VujuTMK for your peer support during this conflicting time of growing up and graduating. Thank you also to my former team lead Markus for your guidance in kicking off my professional journey. Lastly, I would like to thank my family for being there for me all my life, and Lassi, for your love.

Espoo, July 12, 2021

Tuuli Aaltonen

Contents

Chapter 1

Introduction

In the highly competitive world of retail, maximisation of profits and minimisation of losses has become increasingly important. Forecasting the future demand of products as accurately as possible is important in making sure that the right number of products are ordered to each store, as well as in estimating future workload needed for, for example, shelving. Marketing campaigns and promotional discounts need to be planned so that they bring in as much revenue as possible with as small a cost as possible. In addition, the retailer needs to make decisions on which products to include in the product assortment of each store, so that profits are maximised. Substitution effects play a significant role in all these problems.

Substitute products are products that have the same occasion of use and are typically found in the same product category. A decrease in a product's price during, for instance, promotional discounts prompts some customers to switch to purchasing the discounted product instead of its substitute products. Not accounting for the substitution effect can lead to too high sales forecasts for the substitutes, which in turn lead to excessive replenishment of the products, increased inventory costs and the risk of spoilage. Substitution should also be accounted for when planning which products to discount and promote in the first place: to determine how profitable a promotion is, it is important to know if a promoted product simply transfers sales from substitute products or if it also brings in new buyers to the product category. Lastly, substitute products play a key role in assortment optimisation. When making assortment decisions, it is important to make sure that if a customer cannot find their first-choice product from the store due to, for example, a stock-out, the customer is likely to find a substituting product that they are willing to purchase, instead of taking their money to a competitor. So, the

accurate detection of substitute product pairs, and measuring the magnitude of the substitution effect between product pairs is crucial for retailers in order to make optimal business decisions.

When measuring the magnitude of substitution between two products, the relevant question that needs to be answered is: "How much does the demand of substitute B decrease, when the price of product A decreases?" A commonly used measure for this is the cross-price elasticity (Nicholson and Snyder, 2008). There are a few ways to measure the elasticity by using total daily sales per stock-keeping-unit (SKU), which is the most abundant data retailers have and most often use as the data source for sales forecasting and other supply chain -related calculations (Bradlow et al., 2017). Since the sales of substitute products are negatively correlated, simple sales correlation coefficients like the Pearson correlation coefficient can be used to measure the correlation between the demand of two potential substitutes. Finding statistically relevant correlation this way may be difficult between products that lack sufficient sales data or price variation. Another popular approach is to use multiplicative regression models, where SKU sales are explained by both the product's own price and the prices of its substitute. The Scan*Pro model is perhaps the most notable and well-known example of such models (Van Heerde et al., 2002). A significant drawback in this approach is related to the number of model parameters. In a modern grocery store, there is fierce competition among commonly purchased goods, and a single product category can contain dozens of SKUs. Since in Scan*Pro all prices of the SKUs are used as regressors to predict the sales of all SKUs, the number of model parameters can grow too large.

The total sales of an SKU are comprised of individual purchases made by different customers. If sales are recorded on point-of-sales (POS) level, the total daily sales of a product can be disaggregated to receipt level data. Moreover, many retailers have the possibility to identify some of the customers who made the purchases through loyalty programmes (Dowling and Uncles, 1997). This data makes it possible to analyse product sales as customer purchases on an individual level, providing data on how often and what products any individual loyalty card holder purchases. Instead of modelling total daily sales, with this kind of data it is possible to model customer behaviour using choice models like the multinomial logit and probit models. Substitution is at the core of these choice models, and they are designed to answer questions like: Out of a group of potential substitute products, what is the probability of a customer choosing one of the products during each visit? How are the choice probabilities of the products affected when the price of one product decreases? With loyalty card data, it is possible to implement heterogeneity

in the product preferences and price sensitivities between customers in the choice models. The ability to detect heterogeneity increases the information in the models, which could potentially lead to better model predictions, in comparison to models with no heterogeneity (Van Heerde and Neslin, 2017).

This thesis evaluates the feasibility of using choice modelling in measuring the magnitude of substitution effects during price promotions, when there is customer-specific choice data available through, for example, loyalty card information. This is done by comparing the forecasting accuracy of three choice models in a product category of close substitute products during price promotions, so that the main question is: if a product in the category is promoted, how well are the models able to forecast the potential sales decrease of the other products in the category? The performance of the choice models is compared to a simple sales correlation-based method. A particular focus is cast on situations, where the amount of past sales data is somehow limited, so that either the sales volume or data length is reduced. The analysis is performed on three data sets that resemble loyalty card data, generated by simulating customer visits and purchasing choices in product categories of 10, 20 and 30 SKUs.

The rest of the thesis is structured as follows. Chapter 2 introduces the basic concepts of price promotions and substitution effects, theoretical background on random utility choice models as well as previous research done using loyalty card data. Chapter 3 presents the generated data sets and the methods used to create it. Chapter 4 presents the three choice models in comparison as well as the benchmark sales correlation method, and results on the forecasting accuracy of the models are covered in Chapter 5. Finally, conclusions on the results and ideas for future research are discussed in Chapter 6.

Chapter 2

Background

2.1 Retail as business

Retail is the selling of goods directly to consumers to satisfy their demand. Retail outlets hold different assortments of SKUs that consumers purchase. Grocery retailers typically sell food and consumable household items like cleansers and hygiene products in grocery stores, like supermarkets and convenience stores. Like most businesses, retailers aim to maximise their profits. In this effort, they face a myriad of problems that need to be solved through analysis, modelling and optimisation, including ensuring availability and minimising costs through accurate forecasting of sales and optimal replenishment, price optimisation, campaign planning and assortment planning.

To satisfy consumer demand, retailers need to provide the right number of products to the right place at the right time. To determine how many units of each SKU should be ordered from the supplier, sales forecasts are used to predict future demand based on past sales data. Accurate forecasting and optimal replenishment methods are extremely important in making sure that stock levels are optimal. On one hand, too low inventory with respect to demand leads to stock-outs, causing losses in potential sales. Excessive stock, on the other hand, causes higher inventory costs, and in the case of fresh and spoiling goods, the products may spoil before they are sold, causing additional losses.

Sales forecasting methods typically use the past daily or weekly sales of an SKU to forecast future demand (Hyndman and Athanasopoulos, 2018). Various phenomena that affect a product's sales patterns need to be accounted

for in forecast models. Typical sales patterns of an SKU include seasonal changes, long-term upwards or downwards trends and weekday profiles. The demand of some products may increase due to special events, such as national holidays like Christmas or Easter, or local events like a summer festival. Different kinds of campaigns, like advertisements and in-store displays increase the visibility of the promoted products and consequently their sales temporarily. Promotional discounts and other changes in a product's price also significantly affect the sales of the product, as will be discussed later in this section and in Section 2.3. One of the most common methods of forecasting sales is time-series modelling, where the past sales patterns of the product are modelled to estimate future sales, assuming that the sales patterns will be similar in the future (Hyndman and Athanasopoulos, 2018).

Promotional discounts and other changes in a product's price also significantly affect the sales of a product (Nicholson and Snyder, 2008). When the price increases, the demand of almost any product decreases. The sensitivity of demand to sales price varies by product, and can be measured using the price elasticity of demand

$$E = \frac{\Delta Q/Q}{\Delta P/P},$$

where $\Delta Q/Q$ is the percentage change in demand and $\Delta P/P$ is the percentage change in the price of the product. Since an increase in price causes lower demand, E is almost always negative. The higher the absolute value of E is, the more sensitive demand is to the product's price. Price elasticity of a single product is affected by various factors, like the necessity of the product, brand image, prices for the same product in competing stores, and overall competition in the product category. The demand of a product also reacts to the price of its substitutes, which is discussed further in Section 2.2. Price elasticity also affects the sales increase during promotional discounts, which are discussed in Section 2.3.

Product prices have a significant effect on a retailer's profits. Each product in a retailer's assortment brings a profit of *sales quantity* $\times$ *sales margin*, where sales margin is the difference between the sales price and purchase price from the supplier. Retailers attempt to price the goods they sell so that the total profit is maximised, which requires them to deal with the trade-off between sales margin and sales quantity: although a higher sales price brings a higher margin, the higher the price, the lower the sales quantity (Phillips, 2020). Pricing is a vastly complex subject and out of the scope of this thesis, but for example Phillips (2020) provides an in-depth review of price optimisation.

2.2 Substitute products

A substitute product refers to a product which can be used to fulfil the same need as some other product in the market (Nicholson and Snyder, 2008). Substitute products have the same or similar product characteristics and occasion of use. Since substitute products can be purchased to replace each other, they are typically competing with each other in the market. Since substitute products compete for the same customers, a change in the demand of a product also affects the demand of its substitute products. If products A and B are substitutes, then when the price of product A increases, some customers will switch to purchasing product B instead, causing the demand of A to decrease and the demand of B to increase. This means that the demands of substitute products are negatively correlated.

The level of change of a product's demand to another product's price can be measured using the cross-price elasticity of demand. For products A and B, cross-price elasticity is given as

$$E_{A,B} = \frac{\Delta Q_A / Q_A}{\Delta P_B / P_B},$$

where $\Delta Q_A / Q_A$ is the percentage change in sales quantity for product A and $\Delta P_B / P_B$ is the percentage change in price of product B (Besanko et al., 2009). For substitute product pairs, cross-price elasticity of demand is positive. The closer the products are too each other, the higher the elasticity: for instance, two products of packaged minced beef of different brands would have a higher cross-price elasticity than a minced beef product and a minced chicken product.

Cross-price elasticity of demand is not necessarily symmetric for two substitute products (Shocker et al., 2004). This means that the demand of product B might be more elastic to the price change of product A than the other way around. For example, customers purchasing Pepsi might be more prone to switching to Coca-Cola during promotions than the other way around, leading to a higher cross-price elasticity of the demand of Pepsi on the price of Coca-Cola than the elasticity of the demand of Coca-Cola on the price of Pepsi.

2.3 Price promotions

Price promotions refer to the temporary lowering of a product's price to gain higher sales volumes in exchange for the smaller sales margin. In addition

to selling more units, price promotions are used to introduce new products or brands to customers, to build a low-price image and to attract higher store traffic (Dawes, 2012). Price promotions can also be set to counter promotions set by a competitor (Van Heerde and Neslin, 2017). In addition to the temporary sales increase of the promoted product, promotions can also affect the product's own sales before and after the promotion occurs as well as the sales of its substitutes.

Van Heerde and Neslin (2017) divide the effects of price promotions on sales into three categories: immediate effects, or the impact on sales during the promotion period; medium-term effects, or the impact on sales on the weeks surrounding the promotions; and long-term effects, or the impact on sales beyond the medium-term effects. Furthermore, they divide the immediate effects of sales promotions into three categories: category growth, switching, and timing effects.

Category growth refers to the sales increase during the promotion which does not take away from the sales of other products in the product category (Van Heerde and Neslin, 2017). Category growth is composed of purchases, where the consumer extends their budget during the store visit to purchase the promoted item, without leaving any other purchases out. So, category growth is the true increase of purchases caused by the promotion in the product category, that is, the sales increase that is not caused by effects that decrease the sales of other products in the product category, or the sales of the product's own sales before or after the promotion.

Switching refers to consumers switching to purchase the promoted product from some other product that they intended to buy (Gupta, 1988; Van Heerde and Neslin, 2017). This means that sales are transferred to the promoted product from other products within the same category or store. Switching can occur between different brands, in which case it is often called *brand switching*; within the same brand, in which case it is called *cannibalisation*; or across product categories, in which case it is called *category switching*. Switching typically occurs between substitute products. In this thesis, cannibalisation is used as a general term for switching effects.

Finally, the timing effects of promotions can be divided to effects before the promotions, called *deceleration*, and effects after the promotion, called *acceleration* (Van Heerde and Neslin, 2017). In deceleration, consumers learn to expect an upcoming promotion and delay their purchases to the time of the promotion. Conversely, in acceleration consumers advance their purchases to the time of the promotion, often purchasing more than they normally would and thus transferring their purchases of the product from a later time to the

promotion period. This is also called consumer stockpiling.

Some long-term effects of price promotions include purchase-event feedback and consumer learning. Purchase-event feedback refers to the effect that a consumer's past purchases have on their current purchase decisions (Guadagni and Little, 1983; Van Heerde and Neslin, 2017). For example, after purchasing a promoted product, some consumers may permanently switch to purchasing the product due to the information they learned from the promotion purchase. In consumer learning, consumers learn about the frequency of promotions, which then affects their price and promotion sensitivity (Krishna et al., 1991; Van Heerde and Neslin, 2017). For instance, frequent promotions within a product category may train consumers to only buy on deal, increasing stockpiling effects.

Decomposing the effects of price promotions is extremely important for retailers in forecasting the promoted product's and its substitutes' sales before, during and after the promotion (Gupta, 1988). Recognising switching effects of a promotion on other products results in more accurate forecasts, which in turn results in smaller inventory levels and potentially reduces spoilage. On the other hand, the decomposition of sales effects is valuable information that can be used to design more profitable promotions. For example, promotions that induce significant category growth are more profitable than promotions where the sales increase is attributed only to timing effects. Promotions that induce switching are more profitable when the promoted product has a higher sales margin than the other products, in comparison to situations where the sales margin is lower.

2.4 Modelling consumer choice

Discrete choice modelling refers to the study of a decision maker's choice among a set of alternatives (Train, 2009). The decision maker can be any decision-making unit like an individual person, a household, a firm, or a government entity. The decision-making situation can refer to any situation where the decision maker chooses between a set of alternatives, called the choice set, such as deciding whether to commute to work by car, bicycle, or train; choosing which charity to donate to; or choosing which carton of milk to buy at the supermarket. The choice set has three restrictions: it must include all available alternatives, it must be finite, and it must be defined so that the decision-maker can only choose one alternative out of the choice set. Choice models give the probability of choosing each alternative as a function of variables specific to the alternative or decision-maker.

In the context of retail, discrete choice modelling is often used to model and predict customer choice among a category of near-perfect substitute products, like ground beef, soft drinks or laundry detergents (Chandukala et al., 2008). Discrete choice models quantify customer decision processes in order to gain insights into the origins of their preferences (Allenby et al., 2017). Choice models allow retailers to estimate how customer decisions are affected by various marketing mix variables, such as price, promotions and advertisement, or any other factors, such as product attributes, product placement in the store and brand loyalty. The obtained models can be utilised in the estimation of price and cross-price elasticities and customer segmentation among other use cases. Choice models are very useful in modelling brand switching and cannibalisation effects during promotions. If product A is on 20% discount, how much will the choice probability of product A increase and how much will the choice probabilities of other alternatives in the product category decrease? Choice models directly answer this question.

2.4.1 Random utility choice models

Discrete choice models are often formulated following the assumption that consumers aim to maximise their personal utility function (Train, 2009). A utility function is a representation of an individual's preferences that quantifies the value of a consumer good or service beyond its pure monetary value (Nicholson and Snyder, 2008). In discrete choice situations, each choice set alternative brings a certain amount of utility to the consumer according to their own utility function. Consumers are assumed to behave rationally and choose the option that provides maximum utility. When making purchasing decisions, consumers also have the possibility not to purchase anything and hold on to their money for future purchasing decisions. This option is called the *no-buy alternative* (Chandukala et al., 2008). In retail applications, the unit of analysis in choice models is typically brand or brand-size (Fader and Hardie, 1996).

Random utility models attempt to model consumer utility functions while also allowing for unknown factors that affect utility (Train, 2009). There are certain observable variables that are known to affect consumer preferences: the less money consumers spend, the more they have left to purchase other products, so the lower the price of the product, the higher utility consumers gain from the purchase (Chandukala et al., 2008). However, there are always factors that affect the utility function that cannot be observed based on pure shopping behaviour. Buying a product for a visitor or to a certain event, wanting to try a new product for variety, and impulse buying are examples

of situations, which affect a consumer's utility function, but are unobservable to the researcher (Trijp et al., 1996).

Random utility models solve this issue by constructing a two-part utility function that separates the observed and unobserved factors in the utility (Train, 2009). The utility of product j for consumer h at time t is

$$U_{jt}^h = V_{jt}^h + \varepsilon_{jt}^h,$$

where V_{jt}^h is the deterministic component of the utility function and ε_{jt}^h denotes the random variation in the customer's utility over time t, capturing all factors that affect the customer's decision that cannot be observed. For clarity of notation, indices t and h are left out for the rest of this section.

Following the principle of utility maximisation, a consumer chooses product j only if $U_j > U_k$ for all $k \neq j$. The probability of choosing alternative j is therefore

$$\begin{aligned} p_j &= P(U_j > U_k \text{ for any } k \neq j) \\ &= P(V_j + \varepsilon_j > V_k + \varepsilon_k \text{ for any } k \neq j) \\ &= P(\varepsilon_j - \varepsilon_k > V_k - V_j \text{ for any } k \neq j). \end{aligned}$$

The error terms can be removed by integrating over them. If $f(\varepsilon)$ is the joint distribution of the error terms of the alternatives, the probability becomes

$$p_j = \int_{\varepsilon} I(\varepsilon_j - \varepsilon_k > V_k - V_j) f(\varepsilon)\, d\varepsilon, \tag{2.1}$$

where I is an indicator function, which equals 1 if the condition is true and 0 otherwise. This is a multidimensional integral over the joint distribution of the error terms. Therefore, the distribution of the error terms dictates the final formulation of the probability as well as what assumptions are made of the unobserved customer preferences. The choice of error distribution leads to different types of choice models, the most common of which are discussed in the next section.

When defining a random utility choice model, one needs to both define the form of the deterministic component V_j and choose the error term distribution $f(\varepsilon)$. V_j is often constructed as a linear function $V_j = \beta X_j$, where X_j is a vector of variables and β is a vector of model parameters. In addition to price, common variables included in X_j are product-specific intercepts, indicator variables for different campaigns like in-store displays and features, referring to advertisements in for instance newspapers, and customer-specific

variables such as household income (Van Heerde and Neslin, 2017). So, one example form of the utility could be

$$\begin{aligned} V_{hjt} &= \beta_j X_{hjt} \\ &= u_j + \beta_1 \text{price}_{jt} + \beta_2 \text{display}_{jt} + \beta_3 \text{feature}_{jt} + \beta_4 \text{income}_{ht}, \end{aligned}$$

where u_j is the product-specific intercept. Price, feature and display variables depend on both product and time and household income depends on the household and time.

2.4.2 Multinomial logit, nested logit and multinomial probit

In this section, three of the most well-known random utility choice models are presented. The simplest and most widely used random utility choice model is the *multinomial logit* (MNL) model, which was first introduced by McFadden (1973). In MNL, the error terms are assumed to be distributed identically and independently and to follow the Extreme value type I distribution, also known as the Gumbel distribution. When the error terms have this form, the integral in Equation (2.1) has a simple closed-form solution and the probability of choosing product j out of K alternatives is

$$p_j = \frac{\exp(V_j)}{\sum_{k=1}^{K} \exp(V_k)}.$$

The model parameters are easy to estimate using log-likelihood maximisation.

The primary fallback of MNL is that it assumes the independence of irrelevant alternatives property (IIA) (Chandukala et al., 2008). IIA states that in a discrete choice situation, the relative preference between alternatives in the choice set should not be affected by changes in the presence or properties of other alternatives in the choice set. It is easy to prove that MNL assumes the IIA property by inspecting the ratio of the choice probabilities of two alternatives i and j:

$$\frac{p_i}{p_j} = \frac{\exp(V_i)}{\exp(V_j)}$$

The probability ratio does not depend on parameters of alternatives other than i and j. This indicates that if the choice probability of some other alternative k increases, the choice probabilities of alternatives i and j decrease in a proportional manner. It is argued that IIA is violated whenever there are dissimilarities between the alternatives in the choice set. For example, when

modelling the demand of three soda brands, 7up, Pepsi and Coke, it would be expected that if the price of Coke changes, more people would substitute it with Pepsi than 7up, meaning that the choice probability of Pepsi should increase more than the choice probability of 7up.

IIA can be relaxed by allowing for correlation between the error terms using other types of error distributions. In *nested logit* models, the error distribution is assumed to follow a type of the generalised max value (GEV) distribution that features a block correlation structure (Fok, 2017). In this approach, the choice set is divided into "nests", so that the alternatives within each nest are in some way similar to each other, and different across nests. The similarity can be defined based on, for example, a certain product attribute. The correlation structure relaxes the IIA property so that IIA holds within each nest but does not hold across the nests. Nested logit is often used to model sequential decision processes, where it is assumed that consumers first pick a product based on some key attribute and then only consider alternatives that possess that attribute. For example, when modelling the choice of milk, it could be assumed that consumers first choose which type of milk they want (whole, semi-skimmed or skimmed milk) and then choose between the available brands of that type. The sequences can have multiple levels of hierarchy.

Nested logit relaxes the IIA property between nests, but the model still has a few limitations. First, the attributes that define the nests need to be well justified and defined separately for each choice set, so the approach cannot be generalised for different product categories. Secondly, IIA still holds within each nest, which means that dissimilarity of the products within the nest is not considered (Train, 2009).

In the *multinomial probit* (MNP) model, the error term is assumed to follow the multivariate normal distribution which allows for any type of correlation structure between error terms, making any substitution pattern among the choice alternatives possible (Chandukala et al., 2008). MNP relaxes IIA flexibly and with no need for a pre-defined structure. One disadvantage of the model is that when the error terms are normally distributed, the integral (2.1) has no closed-form solution, and the model parameters have to be evaluated numerically, which is computationally heavy.

2.4.3 Consumer heterogeneity

Even though utility functions are individual to each consumer by definition, in practice researchers are often interested in the choice probabilities of a

large number of customers. Modelling consumers as a homogeneous group with a shared utility function is convenient in terms of estimation (Guadagni and Little, 1983), but it comes with some drawbacks. Consumers are individuals with different life-stages, budgets, dietary restrictions and personal preferences, which means that in reality, there is heterogeneity in their brand and pack-size preferences as well as their sensitivity to changes in price and promotions. Incorporating heterogeneity provides more information for the choice models, which should cause the prediction abilities of the models to increase (Van Heerde and Neslin, 2017).

Van Heerde and Neslin (2017) divide heterogeneity into two groups: observed heterogeneity and unobserved heterogeneity. Observed heterogeneity refers to the inclusion of household-specific variables in V_j that have different values across households or choice occasions, but all response coefficients are still common to all households. Different types of loyalty variables are a very commonly featured way of including observed heterogeneity, first discussed by Guadagni and Little (1983). Loyalty variables describe the household's loyalty to a certain brand, pack-size or some other product attribute that is often calculated as a weighed moving average over the household's past purchases, so the more often the household has purchased a product with some feature, the higher their loyalty towards the feature is (Guadagni and Little, 1983). Another way of creating loyalty variables is to calculate the portion of the number of purchases of each brand or size out of the total number of the shopper's purchases, during an initialisation period in the data (Ailawadi et al., 1999).

Unobserved heterogeneity refers to the inclusion of heterogeneity in the parameters of the model, so that some or all response coefficients can vary across households (Van Heerde and Neslin, 2017). The heterogeneity can be modelled in multiple different ways. In mixture models, households are divided into segments so that each consumer belongs to exactly one segment, and some or all the model parameters are estimated so that they are segment-specific (Kamakura and Russell, 1989). In addition to the regular parameters of the model, the researcher also needs to estimate the number of segments and the probability of a consumer belonging to each segment. Mixture models are often mixtures of multinomial logit models, called mixture of logits, but the base model can just as well be a multinomial probit model or a nested logit model. A mixture of logits model relaxes IIA on the aggregate level due to the difference in model parameters in different segments, but the IIA property still applies within each segment (Van Heerde and Neslin, 2017). On top of improving models by accounting for heterogeneity, mixture models are a natural way to segment consumers by their purchase behaviour (Kamakura

and Russell, 1989). Another way to include unobserved heterogeneity is to assume that the customer-level response coefficient follows a common, continuous distribution (Chintagunta et al., 1991). This means that instead of estimating the actual response coefficients themselves, the mean and variance of the response coefficients are estimated instead. These models are often called *random effects models*.

These various ways of including heterogeneity can also be mixed with each other to form a very large selection of models (Van Heerde and Neslin, 2017). For example, in a mixture model, the mixed base model can be assumed to be an MNL, MNP or a nested logit model, or a random effects variation of any of these models. The model designer also needs to decide which model parameters are pooled across segments and which are segment-specific, and whether to also include unobserved heterogeneity in the form of, for example, loyalty variables in the model.

2.4.4 Attribute-based choice models

The unit of analysis in the choice models discussed so far has been brand or brand-size, which are the most typical units of analysis in literature (Fader and Hardie, 1996). In reality, a single product category can contain several variants of the same brand-size, which means that a category with a handful of brand-sizes can contain dozens of SKUs. In brand and brand-size level models the choice alternatives can be defined in multiple different ways. One option is to pool all SKUs within a specific brand-size together. This means that when, for instance, modelling the choice among yogurts, each group would contain yogurt products of a certain brand and size, but the yogurt SKUs in the group could vary by, for example, flavour (Pedrick and Zufryden, 1991). Another option is to separate choice sets within the product category to a more specific level by attribute (Fader and Hardie, 1996). In the yogurt example, this would mean that the unit of analysis is still brand-size, but the choice sets would be separated by flavour to 'natural-flavoured yogurts', 'strawberry-flavoured yogurts', and so on.

Consumers purchase individual SKUs and not brand-sizes, which makes SKU the most appealing unit of analysis (Fader and Hardie, 1996). If variant SKUs are pooled together under the same brand-size, it is not possible to study the cross-SKU substitution effects between, for example, the different yogurt flavours of the same brand-size. Separating the choice sets by attribute does not allow for the study of these effects either, since by separating the choice sets, it is by definition assumed that there is no substitution between alternatives that belong in different choice sets. On the other hand, there

may be so many SKUs in a single product category that changing the unit of analysis to SKU without making any separation in the choice set can make the model infeasible (Van Heerde and Neslin, 2017).

This problem can be addressed by utilising choice models that include product attributes in the model, so that SKUs are described as a combination of their attributes. Attributes are physical characteristics of a product that can be quantified or categorised in some way. Brand and package size are examples of attributes that are common to most product categories. In food products, other common attributes are flavour; dietary values like low-fat or regular fat products; in milk products lactose-free, low lactose and regular products; and in sodas sugar and sugar-free products. The attributes included in the model should be chosen so that they are relevant to the category and observable to the consumer (Fader and Hardie, 1996). Representing SKUs in terms of their attributes not only makes it possible to analyse larger choice sets in comparison to models that do not utilise attributes, but also allows the analysis of substitution effects with respect to product attributes.

2.4.5 Purchase incidence

Choice models describe the probability of a person purchasing a certain product out of a set alternatives, during each choice occasion. However, they do not determine, how often the choice occasions occur or how many units are purchased during each shopping occasion. As discussed in Section 2.3, the sales increase during a promotion can also be attributed to factors other than the increase of choice probability of a promoted product. Total daily sales of a product in a store are affected by how often consumers visit the store, how often they purchase from the product category and how many units consumers buy (Gupta, 1988). Category growth has to be modelled as well as product choice probabilities, to account for increase of overall purchases in a product category during, for instance, a promotion. Category growth is often caused by the increase in the probability of customers shopping in the product category, which is called *purchase incidence*.

One possibility of accounting for purchase incidence is to add a "no-buy" option as one of the choice alternatives in the choice model (Chandukala et al., 2008). The advantage of this approach that it is simple to implement to the existing choice model and that there is no need to build a separate model for purchase incidence. However, one problem with the method is that it is not clear, how the deterministic utility V_0 should be defined for the no-buy. It is intuitive to set the price of the no-buy to 0, but in reality, even when not making a purchase from the product category, customers allocate

their money somewhere else.

Another option is to model the choice of when to purchase from the product category separately from the choice of what to buy (Gupta, 1988). These models are called purchase incidence models. When purchase incidence models are combined with choice models, the choice model gives the conditional probability of which item to purchase from the category, when something is purchased from the product category in the first place according to the purchase incidence model. Purchase incidence models allow for the incorporation of explanatory variables to the probability of purchasing from a category, such as estimators for household inventory – the number of units the household should still have in stock from the last purchase – and the potential utility provided by the product category.

Gupta (1988) models the time between product category purchases, the inter-purchase time, as a function of explanatory variables describing the product category and household inventory. The inter-purchase time follows the Erlang-2 distribution, so that the probability density depends on explanatory variables such as prices and promotional activities in the product category and estimates of household inventory.

Bucklin and Gupta (1992) utilise a category incidence model, paired with a multinomial logit choice model. The category incidence model is a binary nested logit model that includes the category utility as an explanatory variable, derived from the MNL choice model. The probability of purchasing from the product category is

$$P^h_{\text{buy},t} = \frac{\exp(\gamma_0 + \gamma Y^h_t)}{1 + \exp(\gamma_0 + \gamma Z^h_t)},$$

where Z^h_t is a function explanatory variables for category purchase decision, including category value

$$CV^h_t = \log \sum_{k=1}^{K} \exp(V^h_k).$$

CV^h_t is the log-denominator of the multinomial logit choice model, describing the overall utility provided by the alternatives to the household.

2.5 Loyalty card data

Retailers have different data sources to utilise in their analysis of substitute products and customer behaviour. The most traditional data used by nearly

all retailers are the aggregated or total daily sales per SKU obtained from universal product code (UPC) scanners, which is combined with inventory data from the retailer's Enterprise Resource Planning (ERP) system (Bradlow et al., 2017). Retailers utilise total sales in many analyses, such as in analysing forecasting sales patterns by SKU, estimating price elasticities, and assessing the effects of promotions and events on sales. Analysis is also possible on the receipt or point-of-sales (POS) level, which allows for the analysis of, for example, item co-occurrences in market baskets. Total sales is abundant, easy to store and consistent. However, total sales only show the total effect of marketing decisions, but it cannot be used to analyse the behaviour of individual customers.

Another traditional source of purchasing data is household scanner panel data. In household scanner panels, a random sample of households are asked to log all their supermarket purchases using a UPC scanner for a certain time period (National Research Council, 2005). The purchases are manually linked to the store they were purchased at to obtain store-level data. Scanner panel data is an ideal data source for random utility choice models. It provides longitudinal, household-specific choice data while also recording the factors that most likely affect purchasing decisions, like price, promotions, and assortment information. As discussed in Section 2.4.3, accounting for customer heterogeneity is crucial in order to create good choice models: household panel data provides choice data on the most disaggregated, heterogeneous level. A major shortcoming in household panel data is that since the panellists need to be recruited and compensated, it is only possible to include a small portion of all households in the panel (Bradlow et al., 2017). Also, the data is not continuous, but restricted to a fixed time period, so the panel needs to be repeated at regular intervals. Guadagni and Little (1983) and Kamakura and Russell (1989) among many others use household panel data in their studies that model customer choice on an individual level.

A modern form of household panel data is loyalty card data, which can be collected automatically through retail loyalty programmes. Loyalty programmes and the way they are used to collect loyalty card data is described in the next section. Then, some previous research that utilises loyalty card data is presented and lastly, some potential biases that can occur in loyalty card data are discussed.

2.5.1 Loyalty programmes

Loyalty programmes are initiatives, where customers are rewarded with various benefits in exchange for their loyalty to the company (Cortiñas et al.,

2008; Mauri, 2003). In retail loyalty programmes, members are often offered monetary bonuses, when their spending within a certain time period reaches a predefined limit, as well as promotions offered exclusively to programme members (Dowling and Uncles, 1997). Customers are identified at checkout by their personal loyalty card. The goal of these actions is to increase customer loyalty to the firm and thus increase revenue.

Loyalty programmes in retail became increasingly popular in the 1990s and have since matured in many countries so that a very large portion of the population is a member in at least one loyalty programme (Dowling and Uncles, 1997). For instance in Finland, out of the 5.5 million inhabitants, 3.8 million people hold the loyalty card of the country's largest retailer (S-Ryhmä, 2020). The effectiveness of loyalty card programmes has been the target of extensive research and the results are somewhat mixed. While many studies suggest that loyalty programmes have a positive effect on the retailer's total revenue, it has also been argued that the large number of competing loyalty schemes has made the effect redundant (Cortiñas et al., 2008; Dowling and Uncles, 1997). Due to the competition that has diminished the original goal of the loyalty programmes, the data that is supplied by programme members to the retailer has become an increasingly valuable part of the benefits of loyalty programmes (Cortiñas et al., 2008).

Loyalty programmes allow for extensive data collection of loyalty programme members. First of all, the retailer has access to the shopper's demographic data that is given upon registration, which may include information such home address, age and gender (Bradlow et al., 2017). Secondly, since most often programme members need to show their loyalty card upon each purchase to receive the bonuses, the retailer has the possibility to add the customer's personal loyalty card ID to each purchase they make at any of the retailer's stores. This allows the retailers to analyse the purchases of loyalty programme members as a time series of POS data points on an individual level, similar to household panel data. From now on, this data is dubbed as loyalty card data.

2.5.2 Previous applications

Public research that explicitly utilises loyalty card data is sparse. In comparison to traditional data sources like total store sales data and household scanner panel data, few studies have utilised loyalty card data in research. Loyalty card data has been utilised in various applications, such as modelling promotional impact (Felgate and Fearne, 2015), promotion targeting (Pauler and Dick, 2006) and the analysis of customer loyalty (Allaway et al., 2006;

Buckinx et al., 2007). A common factor in most research utilising loyalty card data is the segmentation of customers based on either their shopping behaviour or demographic data (for instance, Allaway et al. (2006); Pauler and Dick (2006)).

Felgate and Fearne (2015) analyse the impact of different promotional activities on fresh beef products using the UK retailer Tesco's club card data. They hypothesise that the sales growth caused by a certain promotion type is moderated by the consumer's life-stage, meaning if they are, for instance, young adults, young families or old families. They fit a simple regression model to the club card data, where sales per store are explained by the presence of different types of promotions, like price reductions and multi-buy promotions, as well as the composition of the customers' life-stages. They find that the model that includes customer life-stage represents the sales better than when it is left out and that the sales increase by promotion type varies by the life-stage.

Pauler and Dick (2006) develop a model for optimising the prices of a food retailer by using their loyalty card data to segment customers by their value to the retailer. They segment customers to eight groups using three measures of customer value: total sales, total profit and sales gap coefficient, which is a measure of the consumer's spending on groceries outside the retailer. Segmenting customers lead to better estimates of price and promotion elasticities, which in turn lead to higher profits after price optimisation. In addition, the model allows the identification of the best and worst customers to the retailer, which can be used in promotion targeting and planning among other applications.

Analysing customer loyalty on an individual or household level is a natural application of loyalty card data. Allaway et al. (2006) segment loyalty programme members based on their shopping habits, using five measures of loyalty-related behaviour, such as total number of purchase occasions, average purchasing interval and total number of dollars spent at the retailer. K-means clustering is used to segment customers to size loyalty groups, which were then profiled and analysed further. Finally, membership in the clusters is predicted by geographical variables, such as distance from the store and distance from rival stores, determined from the registered address of the member. Buckinx et al. (2007) conduct a loyalty survey on customers of a retail chain to assess the true loyalty of customers. The survey answers are then linked to the customer's loyalty card demographic data and purchasing history. They are then able to fairly reliably predict the customers' "true" loyalty based on the survey with predictive variables derived from the loyalty

card data, such as monetary spending, purchasing frequency and distance to the store.

2.5.3 Missing data and bias in loyalty card data

Although loyalty card data is a rich source of data, it could give a biased representation of the total customer base. Not all customers of a retailer join the loyalty programme, which means that loyalty programme members only represent a portion of all customers. Loyalty scheme members may behave very differently in comparison to non-members. Since customers sign-up for the programme voluntarily, there can be selection bias in what type of customers join the programme. For example, if the loyalty card scheme provides numerous member-only promotions, customers who join the programme may be more sensitive to promotions than others (Cortiñas et al., 2008). The bias in sensitivity may also vary by product category. On the other hand, if the programme awards monetary bonuses for customers with high monthly spending, the programme could be more attractive to people who are already frequent and high-spending customers at the store (Smith et al., 2003).

Some studies have been conducted on the differences between loyalty programme members and non-members. Loyalty programme members tend to shop bigger shopping baskets and their average spending per visit is higher (Cortiñas et al., 2008; Smith et al., 2003). Cortiñas et al. (2008) study the difference in price and promotion sensitivity between members and non-members of a loyalty programme, where on each shopping trip, programme members are awarded points that can be exchanged to gifts. They find that price sensitivity is similar between members and non-members, but sensitivity to promotions varies between the groups, depending on the product category. They argue that since the average shopping basket size of purchases made with a loyalty card is larger compared to purchases without loyalty card, the marginal cost of increasing the shopping basket size is higher for purchases made with the card, which causes card holders to be less sensitive to promotions in product categories with large pack-sizes.

In addition to loyalty card data not including purchases from all customers, it also does not include all purchases of loyalty card members. In the diary study conducted by Smith et al. (2003), loyalty card holders use their loyalty cards on only 84% of purchases. Mauri (2003) study the activity of loyalty card use at an Italian grocery retailer and find that a large portion of customers who hold a loyalty card do not use it frequently. Potential explanations for this could be that the shopper does not feel the need to show the card when doing small purchases compared to their regular basket size,

or that the loyalty scheme is generally of little interest to the card holder. Card holders might not necessarily have sufficient knowledge on the loyalty programme to know how to profit from it. The lack of commitment to using the loyalty card means that the POS data points that do not have a loyalty ID also include purchases made by loyalty programme members. Another problem with loyalty card data is the fact that shoppers tend to spread their purchases across multiple stores, which means that the loyalty card database of one retailer only covers a part of the purchases of any household. This means that it is not possible to determine true consumption based on loyalty card data.

Chapter 3

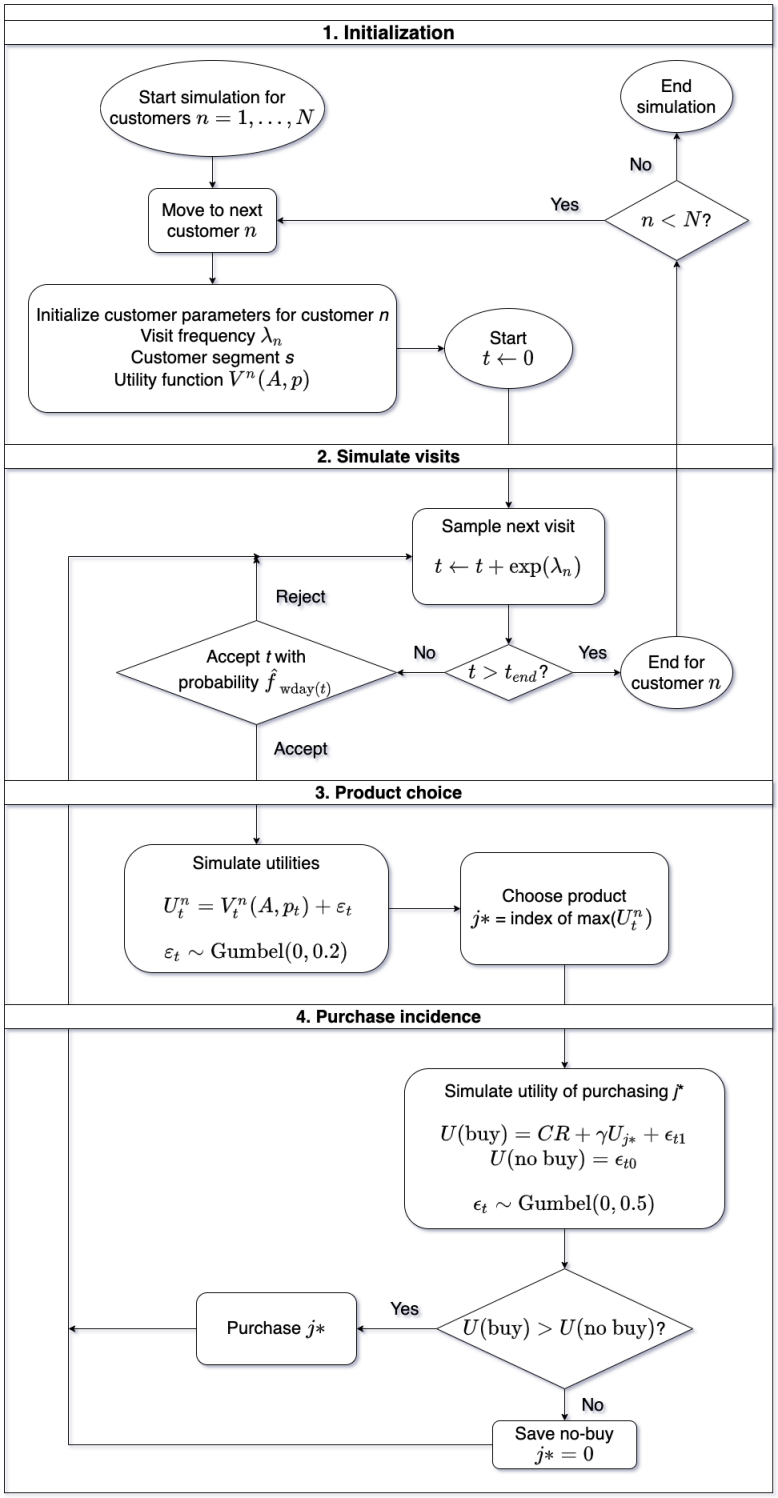
Generation of loyalty card data

Due to data privacy issues, real loyalty card sales data could not be used in this thesis. Instead, the loyalty card data is simulated. The purchases of N customers in product categories of 10, 20 and 30 SKUs in a retail store are simulated to generate three data sets of product choice data over a time period of four years. The shopper of each purchase can be identified like in loyalty card data. In the simulation, different products are frequently promoted, so that large promotional increases and substitution effects are created.

The simulation consists of four steps that are repeated for each customer: initialisation of simulation parameters, simulation of store visits, determining product choice out of the product category, and determining purchase incidence. The simulation is visualised in the diagram in Figure 3.1. Customer visit times are simulated as independent Poisson processes of varying frequencies, pictured in part 2 of the diagram. Product choice is simulated using attribute-based multinomial logit models that are unique to each customer, pictured in part 3 of the diagram. Lastly, purchase incidence is simulated using a binary nested logit approach, so that the utility of making a purchase depends on the utility that the chosen product would provide for the customer. Simulation of purchase incidence is described in part 4 of the diagram.

In the rest of this chapter, methods used in generating the data and the resulting data sets are presented. Section 3.1 elaborates on some desired qualities of the data and Section 3.2 describes the product assortments. The different steps of the simulation are presented in Sections 3.3 and 3.4. Finally, some characteristics of the final data sets are presented in Section 3.5.

Figure 3.1: Diagram of simulation.



3.1 Desired qualities for generated data

Three central goals for the generated data are set:

1. On daily aggregate level, the sales data should resemble real retail sales data to a reasonable extent.
2. Customers should be extremely heterogeneous in their product preferences and sensitivities to product price, so that it is not possible to reverse-engineer the underlying choice model by fitting.
3. There should be distinct substitution effects between products so that the IIA assumption is not obeyed on aggregate level.

To meet the first goal, customer visit times are simulated as a Poisson process. The Poisson process is a common way to simulate customer visits and it generates sales whose variance resembles the variance in real-life sales data. The day of week affects the probability of visiting, so that customers tend to shop more often at the end of the week than at the beginning of the week. To achieve this, the customer visits are simulated using a non-homogeneous Poisson process, where the visit rate depends on the weekday.

The second and third goals are closely related to each other. The two goals are met by simulating product choice using a mixture of attribute-based multinomial logit models with random effects. The mixture structure and random effects ensure that there are considerable differences between customers' utility functions and that each of them is unique. The mixture structure also relaxes IIA on aggregate level through customer heterogeneity, so that the third goal is also met. For simplicity, IIA is allowed to apply to the decisions of each customer, which allows the use of an MNL model as the ground model.

3.2 Product assortments, attributes and prices

Three data sets are simulated for product assortment sizes of 10, 20 and 30 SKUs and different product attributes are defined for each SKU. Setting product attributes makes it easier to design substitution patterns that break the IIA assumption. For example, if some customer segments have a high preference for brand A, customers in that segment are more likely to substitute within the brand rather than switching to competing brands. On aggregate level, when a product of brand A is promoted, the sales of the

brand's other products are then cannibalised more than would be expected according to the IIA assumption. The 10 SKU and 20 SKU assortments are a subset of the 30 SKU assortment, and the products are named using a numerical index in the original 30 SKU assortment. So, for example, product 26 is included in all 3 assortments.

The SKUs have seven different attributes with different numbers of attribute levels. The number of levels depends on the assortment size, since in the smaller assortments, all SKUs may possess the same attribute level for some attribute, in which case the attribute is excluded from the simulation. The three main attributes include brand, pack-size and type. Brand has 3, 6 and 8 different labels in the 10, 20 and 30 SKU assortments, respectively. Pack-size has 3 different size classes in the 10 SKU assortment and 4 classes in the 20 and 30 SKU assortments, and there are 4 different types in the 10 SKU assortment and 5 types in the other assortments. Additionally, 2, 4 and 5 different binary attributes are used in the 10, 20 and 30 SKU assortments, respectively. These kinds of attributes could describe whether the product is, for example, organic or not. All product attribute levels of product j are represented as a vector $A_j \in \{0,1\}^K$, where $A_{jk} = 1$, when product j has attribute level k and $A_{jk} = 0$ otherwise. Table 3.1 presents the A_j vectors and prices of the products in the 10 SKU assortment.

attribute	A_1	A_2	A_7	A_{10}	A_{11}	A_{12}	A_{13}	A_{17}	A_{25}	A_{26}
brand 1	1	1	1	0	0	1	1	0	0	0
brand 2	0	0	0	1	1	0	0	1	0	0
brand 3	0	0	0	0	0	0	0	0	1	1
size 1	0	0	0	1	0	0	0	0	0	0
size 2	1	0	0	0	1	1	0	1	1	1
size 3	0	1	1	0	0	0	1	0	0	0
type 1	0	0	0	0	1	1	1	1	0	0
type 2	1	1	0	0	0	0	0	0	0	0
type 3	0	0	1	1	0	0	0	0	0	0
type 4	0	0	0	0	0	0	0	0	1	1
not organic	1	1	1	1	1	1	1	1	1	0
organic	0	0	0	0	0	0	0	0	0	1
not flavoured	1	1	1	1	0	1	1	1	1	1
flavoured	0	0	0	0	1	0	0	0	0	0
price	5.3	5.1	6.7	7.92	8.87	8.1	7.6	8.7	9.5	9.2

Table 3.1: Attributes and prices of the products in the 10 SKU assortment.

In order to simulate substitution effects, the prices of the products need to vary. This is done by setting price promotions for the products. For simplicity, each of the products has a fixed price that only changes when the product is promoted. During the first three years of simulated data, that is during the initialisation and training periods, each of the products is promoted in numerical order in cycles, so that after the last product has been on promotion, the promotion cycle is starts again from SKU 1. When on promotion, the price of a product is discounted by a percentage that is uniformly drawn from discounts of 5, 10, 15 or 20%. The length of each promotion and the time periods between two consecutive promotions are determined at random, so that the length of each promotion varies from 5 to 14 days and that there is at minimum 4 days and at maximum 12 between any two promotions. During the last year of simulated data, the validation data period, each product is promoted once by a 15% discount, so that the length of the promotion is 7 days in the 20 and 30 SKU assortments and 14 days in the 10 SKU assortment. Only up to 20 products are promoted in each data set, meaning that in the 30 SKU assortment, only 20 SKUs are promoted.

3.3 Simulation of customer visits

Customer visit times are simulated as independent Poisson point processes so that each customer has a unique visit frequency. Consumers tend to shop more often at the end of the week than in the beginning of the week. This means that when customer visits are simulated using a Poisson process, the rate parameter of the exponential distribution needs to vary by weekday, resulting in a non-homogeneous Poisson process. The distribution of sales by the day of week is called a *week profile*. The simulation of customer visits is visualised in part 2 of the diagram in Figure 3.1. The number of customers N is 5000 in all three simulations.

Each customer n visits the store according to a non-homogeneous Poisson process with rate function $\lambda_n(t)$. The rate function has the form

$$\lambda_n(t) = \lambda_{n,\text{wday}(t)},$$

where wday is a function that indicates the weekday of time t, so that $\text{wday}(t) = 1$, if the date is a Monday, $\text{wday}(t) = 2$, if the date is a Tuesday, and so on.

The non-homogeneous Poisson process can be efficiently implemented using *thinning* (Lewis and Shedler, 1979; Pasupathy, 2011). The idea be-

hind thinning is to sample from a homogeneous Poisson process with rate $\lambda_{\max}$ equivalent to the maximum rate of the non-homogeneous Poisson process's rate function $\lambda(t)$. Each sample t is then rejected with probability $1 - \lambda(t)/\lambda_{\max}$. The resulting samples are equivalent to samples drawn from a non-homogeneous Poisson process with rate function $\lambda(t)$.

Applying thinning to account for week profiles is straightforward. The week profile is described by the relative distribution of weekly sales between the weekdays $f \in \mathbb{R}^7_{\geq 0}$, $\sum f = 1$, so that f_1 represents the portion of weekly sales that occur on Monday, and so on. The mean of samples per time unit, in this case day, generated by a Poisson process is equivalent to the rate λ. For weekday w, the average number of visits is therefore

$$\lambda_w = f_w \sum_{i=1}^{7} \lambda_i,$$

which is also the rate of the non-homogeneous Poisson process on weekday w. The acceptance probability of sample t is therefore

$$\hat{f}_{\text{wday}(t)} = \frac{\lambda_{\text{wday}(t)}}{\lambda_{\max}} = \frac{f_{\text{wday}(t)} \sum_{i=1}^{7} \lambda_i}{f_{\max} \sum_{i=1}^{7} \lambda_i} = \frac{f_{\text{wday}(t)}}{f_{\max}}.$$

The week profile and the resulting acceptance probabilities used in simulation are presented in the table below. For simplicity, the same week profile is used for all customers.

	Mon	Tue	Wed	Thu	Fri	Sat	Sun
f_w	0.11	0.13	0.13	0.14	0.15	0.16	0.18
$\hat{f}_w$	0.61	0.72	0.72	0.78	0.83	0.89	1

Since the visiting frequencies should vary between customers, an individual rate λ_n is drawn for each customer from a gamma distribution with shape $k = 1.5$ and rate $r = 0.06$, so that

$$1/\lambda_n \sim \Gamma(k, r).$$

The resulting rates λ_n are capped at 1, which corresponds to visiting the store once per day on average. The maximum rate $\lambda_{n,\max}$ used in simulation is derived from the drawn rates as $\lambda_{n,\max} = 7 f_{\max} \lambda_n$.

There is no customer churn, which means that no customers begin or end visiting the store during the simulation period, in other words all customers are frequent shoppers for the entire simulation time.

3.4 Product choice simulation

Product choice was simulated using a mixture of attribute-based multinomial logit models with random effects, inspired by Fader and Hardie (1996). The deterministic utility of each customer is a function of product price, which varies by time, and product attributes which are time-invariant. Each customer is assigned to a customer segment and the attribute-based part of their utility function is generated by adding a small random variation to a base utility function that is defined by the segment. Price sensitivity is drawn from a log-normal distribution for each customer. Following the definition of MNL models, the final utility of each product at each choice occasion is determined by adding a Gumbel-distributed error term to the utility of each product. Customers always choose the product that provides maximum utility.

The deterministic utility function of customer h is a function of product attributes A_j and price p_{jt}

$$V^h(A_j, p) = A_j\alpha^h + p_{jt}\beta^h,$$

where $\alpha^h \in \mathbb{R}^K$ and $\beta^h \in \mathbb{R}$ are customer-specific response coefficients for product attributes and product price, respectively. To obtain the customer-specific parameters, each customer h is assigned randomly to one of $S = 72$ customer segments with uniform probability. The product attribute preferences are generated as

$$\alpha^h = \alpha^s + \xi^h,$$

where α_s is the preference vector of the customer segment that the customer belongs to and ξ^h is a random term, generated by sampling independently from a zero-mean normal distribution, so that $\xi^h_k \sim \mathcal{N}(0, 0.08)$, for each $k = 1, \ldots, K$.

Price sensitivity β^h is generated using a log-normal distribution, so that

$$\beta^h \sim -\text{Lognormal}(0.5, 0.75)/3.$$

The negative of a log-normal distribution is used to ensure that price sensitivity is always negative.

An error term is added to V to determine the final utility U of each product at each choice occasion, so that

$$U^h_{jt} = V^h_{jt} + \varepsilon_{jt}.$$

The error term ε_{jt} is drawn from a zero-mean Gumbel distribution with probability density

$$f(\varepsilon) = \frac{1}{\beta}\exp(-(\frac{x}{\beta} + \exp(-\frac{x}{\beta}))),$$

where the scale $\beta = 0.2$. During each purchasing occasion, the customer chooses the product that provides maximum utility. The maximum utility is denoted as $U_{\max} = \max_{j=1,\ldots,J}(U_{jt}^h)$.

As the final step, the purchase incidence is determined based on the maximum utility $U_{\max}$. The probability of purchasing is determined using a binary logit approach, so that utilities of buying and not buying are determined as

$$\begin{aligned} U(\text{buy}) &= -4 + 2U_{\max} + \epsilon_1 \\ U(\text{no-buy}) &= \epsilon_0 \end{aligned}$$

The error terms are drawn independently from the Gumbel distribution, so that $\epsilon \sim \text{Gumbel}(0, 0.5)$. If $U(\text{buy}) > U(\text{no-buy})$, the customer makes the purchase and otherwise does not. The no-buy decisions are also recorded, similar to product choices.

3.5 Characteristics of simulated data

The simulations result in 3 data sets of product choice data of 5000 customers, out of product assortments of 3 different sizes. All simulated data sets include four years of data. Up to 20 products in each data set are regularly promoted, so that during the first three years, promotion lengths and discounts vary and during the final year of the data, each product is discounted by 15% once for 14 days in the 10 SKU assortment, and for 7 days in the other assortments. From now on, the data set with 10 SKUs will be called **A10**, the data set with 20 SKUs will be called **A20** and similarly the 30 SKU data set will be called **A30**, to improve readability.

Table 3.2 recaps some key figures for each data set. The number mean visits per day is almost the same in all data sets, since customer visit frequencies were generated with the same methods and parameters. The number of mean purchases per day varies between the data sets, so that most purchases are made in A30 (446.4 purchases per day on average), and the least purchases are made in A10 (around 359.9 purchases per day). Consequently, the number of no-buy decision is the highest in A10.

	mean purchases per day	mean no-buys per day	unique SKUs per customer	highest sellers	lowest sellers
A10	359.9	134.9	3.6	2 (86.0)	25 (6.2)
				26 (76.2)	7 (14.2)
A20	441.9	52.9	5.1	5 (66.4)	25 (0.3)
				4 (41.2)	7 (3.2)
A30	446.4	47.6	5.4	5 (56.4)	25 (0.3)
				28 (37.4)	20 (0.6)

Table 3.2: Key figures of product sales per day and per customer for each assortment. The highest and lowest sellers columns are given in format "Product name (mean daily sales)".

The third column in Table 3.2 shows the mean number of unique SKUs each customer purchases during the simulation period. As expected, the number is highest in A30, where the number of product alternatives is the largest. The values are relatively small, considering that the assortments are large and the simulation time is long. This means that customers are fairly loyal to their favourite products.

Figure 3.2 shows plots for daily sales of the products in A10 during the second training data year in A10, as well as the number of no-buy decisions. The sales spikes that occur during campaign periods are very high. The sales data shows clear cannibalisation effects between the products, and some promotion periods that cannibalise other products are highlighted in the plots. Looking at the plot for no-buys, it can be seen that the number of no-buys decreases significantly during promotions on products 1 and 2. These products are the two cheapest products in the assortment, so it makes sense that these two products increase purchase incidence in the category the most.

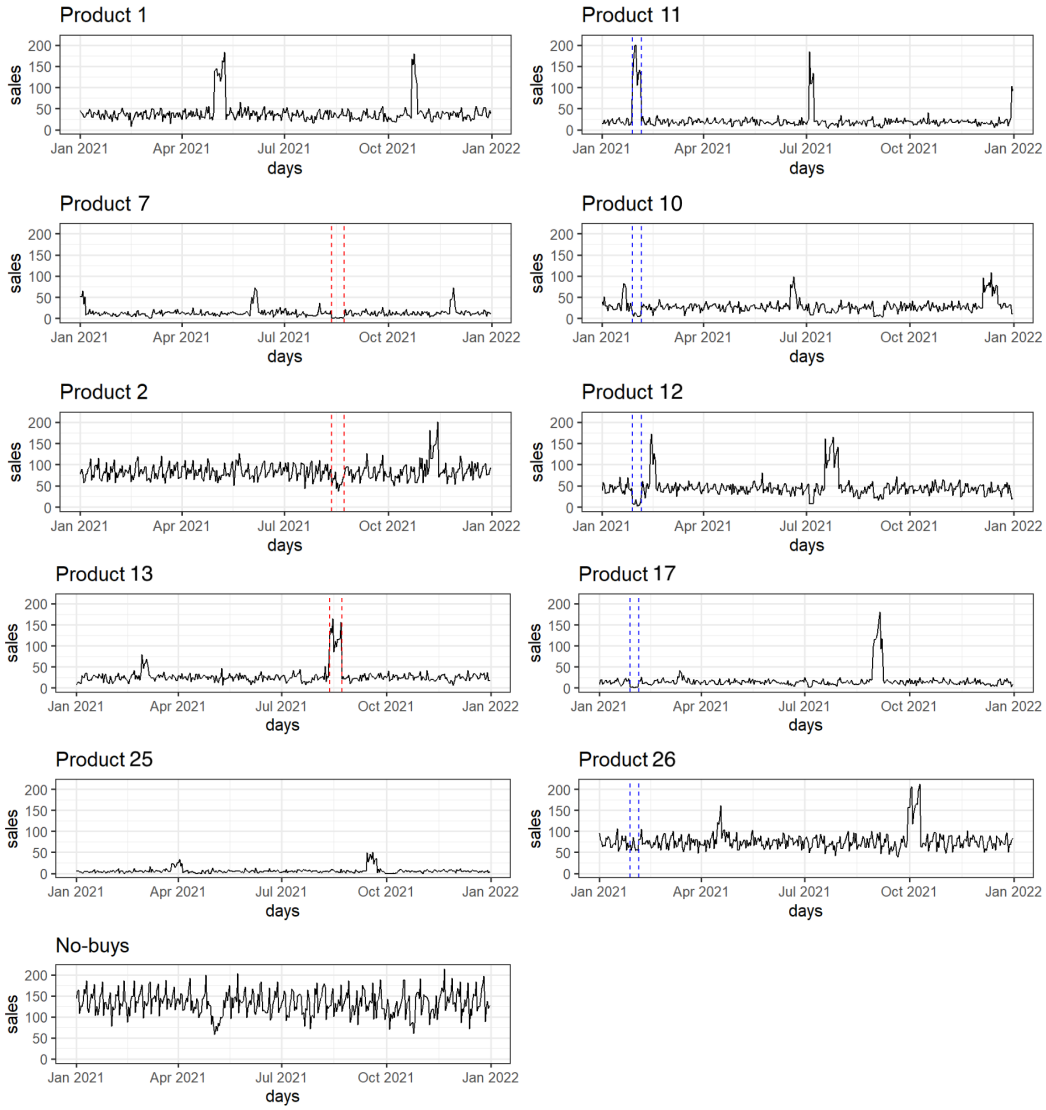


Figure 3.2: Daily product sales during the second year of training data in A10. The red lines define a promotion period for product 13, which visibly cannibalises the sales of products 7 and 2. The blue lines define a promotion period for product 11, which in turn cannibalises products 10, 12, 17 and 26.

Chapter 4

Methods in modelling customer choice

This chapter introduces the three random utility choice models used to model customer choice in the three generated data sets, using product price and loyalty variables, which are calculated using the initialisation period data. The first two models are called the pooled MNL and the k-means mixture of MNLs, and the third is simply called MNP. The models are evaluated against a benchmark sales correlation method that only uses daily aggregates of the generated data.

This section has the following structure. First, in Section 4.1, some basic principles used in constructing the choice models are discussed. Section 4.2 presents the loyalty variables used in the methods and Section 4.3 presents the three choice models that were developed. For each choice model, the deterministic utility function that is used by the model is presented and the methods of calculating model fits and sales forecasts are discussed. Section 4.4 presents the sales correlation model that is used as a benchmark model for the choice models. Finally, Section 4.5 presents the methods used in estimating the choice models.

4.1 Discussion on choice model construction

There are a few things to consider when applying MNL and MNP models to retail loyalty card data. First of all, as discussed in Section 2.4, SKU-level sales data is usually pooled to brand or brand-pack size level so that the number of product alternatives ranges from 5 to 10 products. In SKU-

level models that were discussed in Section 2.4.4, products are presented as a combination of product attributes which makes estimation more efficient. In retail, SKU is the main planning unit for forecasting and replenishment. So, the methods need to be able to handle a much larger number of alternatives than simply the number of brand-sizes in the product category. However, consistent attribute information is usually difficult to obtain on a large scale. Retailers may not have attribute information available for all brands or all relevant attributes, and if it does exist, the attribute labels are not necessarily stored in a standard format across brands. Getting consistent attribute information for all products would therefore require significant standardisation of the data through manual work or natural language processing. Therefore, in this research, SKU choice probabilities are modelled without attribute information.

Another important thing to consider is how customer heterogeneity is included in choice models. As discussed in Section 2.4.3, accounting for heterogeneity in the customer base is key in choice modelling. Observed heterogeneity is very straightforward to implement in choice models from the past purchases of each customer using loyalty card data. Unobserved heterogeneity in, for instance, price sensitivity between customers, on the other hand, is more difficult to implement. Latent mixture models are computationally heavy and overly complex for the purposes of this thesis. For this reason, observed heterogeneity is included in all models and unobserved heterogeneity is not implemented in any way. In the pooled MNL and MNP models, loyalty variables are included as model regressors to capture heterogeneity in the models. The k-means mixture of MNLs approach attempts to also include heterogeneity in price sensitivity by using a simple k-means clustering approach.

Lastly, a way to handle no-buy decisions needs to be decided. In the simulated data, the decision to make or not to make a purchase depends on the utility that the chosen product would bring to the shopper. To avoid having to model purchase incidence separately from the choice models, the recorded no-buy decisions are included in all choice models as a choice alternative similar to the other products. The price of the no-buy is set to 0 and loyalty variables are calculated for the no-buy decisions similar to other products. The number of products is denoted with J, so the total number of choice alternatives with the no-buy included is $J + 1$.

4.2 Loyalty variables

Observed heterogeneity was implemented using SKU-level customer-specific loyalty variables. As discussed in Section 2.4.3, the two options for a loyalty variable are the moving average-type variable from Guadagni and Little (1983), which changes depending on the choice occasion, and the preference variable introduced in Ailawadi et al. (1999), which is constant for each choice occasion. The latter one was chosen as it was the simpler of the two.

The SKU-loyalty variables are calculated using the 1 year initialisation period in the data set. Loyalties are calculated for each customer and each SKU and the no-buy alternative, so that there are $N \times (J+1)$ loyalty variables in total. The loyalty of customer h to product j is

$$L_j^h = \frac{\text{number of times customer } h \text{ chose alternative } j}{\text{total number of visits of customer } h}. \tag{4.1}$$

Consequently for each customer, the sum $\sum_{j=1}^{J+1} L_j^h = 1$. If the customer has no visits during the initialisation period, all loyalty variables are set equal so that they sum to 1, $L_j^h = 1/(J+1)$.

4.3 Choice models in comparison

Next, the three choice models in comparison are presented. As all three models are random utility choice models, they share some similar features. Like discussed in Section 2.4, all random utility choice models have a deterministic utility function, which gives the utility of each product alternative at each choice occasion before random variation. In this thesis, linear utility functions are used, so that the utility of product j at choice occasion i is

$$V_{ji} = X_{ji}\theta,$$

where θ is a vector of model parameters and X_{ji} is the data matrix. As only the relative utilities of the alternatives on each choice occasion matter, utilities are scaled to represent their utility difference to a base alternative. SKU 1 is used as the base alternative so that $V_{1i} = 0$ at all times. The final utility is defined as

$$W_{ji} = V_{ji} + \varepsilon_{ji}, \tag{4.2}$$

where ε_{ji} is an error term whose distribution depends on whether the model is an MNL or an MNP model. The utilities W are not observed from the data

but instead the data only shows which alternative was chosen, that is, which alternative had the highest utility at each choice occasion. The observation Y_i is an indicator of which alternative had maximum utility and is given by the latent variable W_i as

$$Y_{ji}(W_i) = \begin{cases} 1 & \text{if } W_{ji} \geq \max(W_i, 0) \\ 0 & \text{otherwise.} \end{cases}$$

The observed choices Y are used in the model estimation methods to derive optimal model parameters θ^*.

4.3.1 Pooled MNL

The simplest fitted model is called the pooled MNL model. In the model, all customers share the same deterministic utility function and the base model is a multinomial logit model. The deterministic utility function is defined as

$$V_{jhi} = \alpha_j + \beta_1 \text{price}_{ji} + \beta_2 L_j^h, \tag{4.3}$$

where α_j is the intercept term of alternative j, L_j^h is the loyalty of customer h to product j defined in Equation (4.1) and price_{ji} is the price of product j at time i. MNL outputs the choice probabilities of each product given the product prices and loyalty variables, so that the probability of customer h choosing product j at time t is

$$p_{jt}^h = \frac{\exp(\alpha_j + \beta_1 \text{price}_{jt} + \beta_2 L_j^h)}{\sum_{k=1}^{J} \exp(\alpha_k + \beta_1 \text{price}_{kt} + \beta_2 L_j^h)}. \tag{4.4}$$

Optimal model parameters are found using maximum likelihood estimation (MLE), which is described in more detail in Section 4.5.2.

Fitting the estimated MNL model to training data is a straight-forward calculation. When calculating model fits, for each choice occasion the product prices and customer are simply inserted in Equation (4.4) to obtain product choice probabilities for that choice occasion. The probabilities of each product are then summed together by day, which results in model fits of daily sales.

Calculating model forecasts for the validation data is trickier. In the training data, the number of customer visits for each day and the customers' loyalty variables are known, which makes the estimation of choice probabilities simple. However, in forecasting, there is no knowledge of which customers visit the store each day. So, there is only data on future daily prices but no data on

loyalty variables. Simply using the mean of all loyalty variables in forecasting is not sufficient, since the IIA assumption would apply to the forecasts. Instead, the customer populations' loyalty variables are summarised using k-means clustering, which is explained in more detail in Section 4.5.1. The forecasts are then calculated using product prices in the validation data set and each loyalty variable cluster centre. Note that in this step, instead of clustering customers, the loyalty variables of each *choice occasion* are clustered. This means that in the data that is clustered, there are duplicates of the same loyalty variable vectors of each customer, depending on how often the customer visited the store.

Lastly, the weighed average of the clusters' product choice probabilities is calculated, using the cluster sizes as weights. The average product choice probabilities are then multiplied with the average number of customer visits per day. This results in forecasts that are constant unless there is a change in product prices, which causes the magnitudes of the forecasts to shift.

4.3.2 K-means mixture of MNLs

The pooled MNL method captures differences in the population's SKU loyalties which slightly relaxes the IIA assumption. One problem with the pooled model is that the other model parameters, product intercepts α and price sensitivity β_1 are the same for all customers. In the choice model used for simulating the data, there are significant differences between customers in these parameters.

In the k-means mixture of MNLs, customers are first clustered by both their SKU loyalty variables L and the mean price of their purchases during the initialisation period, so that customers are divided into fixed segments S'. Note that unlike in the pooled approach, the clustering in this method is performed on customers instead of choice occasions. The deterministic utility of the model is then

$$V_{ji}^s = \alpha_j^s + \beta^s \text{price}_{ji}, \tag{4.5}$$

so that all model parameters depend on the customer segment $s \in S'$ of the customer, which is determined in the initial clustering. Unlike in the pooled MNL model, the loyalty variables are not used as regressors in the model.

Model fits to the training data are obtained by calculating product choice probabilities for each choice occasion, using the model parameters of the cluster that the customer belongs to. In forecasting, product choice probabilities are first calculated separately for each cluster, using the product prices of

each date in the validation data. The probabilities are then weighed with the cluster's size, which is calculated based on the average visit frequency of the customers in that cluster, and summed together to obtain product choice probabilities for the entire customer base. Finally, the choice probabilities are multiplied with the average number of customer visits per day.

The k-means mixture model has two drawbacks worth mentioning. First, some customer groups may not have any observations for a product alternative. In these cases, the alternative must be omitted from the fitted model, since it was noticed that the MNL models are not able to give a reliable model estimates if one choice alternative has no observations. Another drawback is that the number of model parameters increases by almost the number of clusters in comparison to the pooled MNL.

4.3.3 MNP

The final choice model is the multinomial probit model, which is formulated the same way as the pooled MNL model, except that the base choice model used is a multinomial probit model. The same deterministic utility function is used as in the pooled MNL approach, defined in Equation (4.3). In MNP, the error term added to the deterministic utility is normal distributed, so that the model formulation is

$$W_{ij} = X_{ij}\theta + \varepsilon_{ij}, \quad \varepsilon_i \sim \mathcal{N}(0, \Sigma).$$

In addition to the parameters θ, in the MNP model, also the covariance matrix Σ needs to be estimated. As explained in Section 2.4.2, MNP does not have a closed-form solution like the MNL model, which makes model estimation much more difficult. The model parameters are estimated using a simulation approach called Markov Chain Monte Carlo estimation, which is described in further detail in Section 4.5.3. In fitting and forecasting sales with the MNP model, a similar approach to the pooled MNL approach is used.

4.4 Benchmark sales correlation method

Alternative cannibalisation campaign forecasts are calculated using a simple sales correlation method. In the sales correlation method, a correlation test is performed between the sales increase of a promoted product and the sales decrease of each of the other products. If the correlation is considered significant on a predefined significance level, a linear regression model is fit to

predict the sales decrease of the affected product as a function of the sales increase of the promoted product. Otherwise, it is assumed that the other product is not affected by price changes of the promoted product.

In order to compare sales increase and sales decrease, approximations of baseline sales for each product are needed. Additionally, to forecast the sales decrease, a separate campaign forecast is required for the promoted product. Daily baseline sales and campaign forecasts are estimated using one of the fitted choice models, referred to as the *reference choice model* from now on. The correlation tests are performed on total daily sales level. Both sales from the initialisation and training periods are used in the correlation tests.

Consider a correlation test between the sales increase of a promoted product l and the sales decrease of product m. First, daily sales of the products are fetched for dates when product l was promoted, denoted as group T_l^p with length N_p. Daily sales increase of promoted product l is calculated as

$$y_{lt}^{\text{inc}} = y_{lt} - b_l, \quad t \in T_j^p,$$

where y_{lt} denotes the sales of product l on day t and b_l are the daily baseline sales of product l, calculated using the reference choice model. Sales decrease of product m is similarly calculated as

$$y_{mt}^{\text{dec}} = y_{mt} - b_m, \quad t \in T_j^p.$$

A correlation test is performed between y_{lt}^{inc} and y_{mt}^{dec} using the Pearson correlation coefficient, which measures linear correlation between two variables. The coefficient is calculated for a data sample as

$$r(y_l^{\text{inc}}, y_m^{\text{dec}}) = \frac{\sum_{t \in T_l^p}(y_{tl}^{\text{inc}} - \bar{y}_l^{\text{inc}})(y_{tm}^{\text{dec}} - \bar{y}_m^{\text{dec}})}{\sqrt{\sum_{t \in T_l^p}(y_{tl}^{\text{inc}} - \bar{y}_l^{\text{inc}})^2}\sqrt{\sum_{t \in T_l^p}(y_{tm}^{\text{dec}} - \bar{y}_m^{\text{dec}})^2}}$$

Student's t-test is then performed on the coefficient with $N_p - 2$ degrees of freedom and test statistic

$$t^* = r\sqrt{\frac{N_p - 2}{1 - r^2}}.$$

The null-hypothesis is that there is no linear correlation between the two variables and the alternative hypothesis is that there is linear correlation. The null-hypothesis is rejected when p-value < 0.05.

If correlation is detected, a linear regression model is fit between the sales decrease and sales increase, so that

$$y_m^{\text{dec}} = c + dy_l^{\text{inc}}.$$

Since the sales correlation does not have independent baseline or campaign increase forecasts, only cannibalisation forecasts are calculated using the method, and the baseline and campaign increases forecasts are set equal to the forecasts given by the reference choice model. During promotions, for products other than the promoted product, if correlation was detected between the promoted product l and the other product m, the forecast of the other product is set as

$$\begin{aligned} \hat{y}_m &= b_m + y_m^{\text{dec}} \\ \Rightarrow \hat{y}_m &= b_m + c_{ml} + d_{ml}(\hat{y}_l - \hat{b}_l). \end{aligned}$$

If no correlation was detected, the forecast of the other product is set to its baseline forecast

$$\hat{y}_m = b_m.$$

4.5 Methods used in estimation

In this section, methods used in model estimation are presented. K-means clustering is used in both the pooled MNL and the k-means mixture of MNLs model, and is discussed in Section 4.5.1. Maximum likelihood estimation is used in estimating the MNL models in both the pooled MNL and k-means mixture of MNLs, and is presented in Section 4.5.2. Section 4.5.3 presents the Markov Chain Monte Carlo method which is used to estimate the MNP model.

4.5.1 K-means clustering

The goal of k-means clustering is to partition the data set x into k sets $S' = \{S'_1, \ldots, S'_k\}$ so that the total within-cluster sum of squares (TSS) is minimised. The problem formulation is

$$\arg\min_{S'} \text{TSS} = \arg\min_{S'} \sum_{i=1}^{k} \sum_{x \in S'_k} ||x - \mu_i||^2,$$

where μ is the mean of data points in cluster i.

The clustering is done using the popular algorithm of Hartigan and Wong (1979). The basic idea in the Hartigan–Wong algorithm is to iterate through each data point, so that on each iteration, the data point is assigned to its closest cluster and cluster centres are recalculated. The algorithm is initialised by assigning k random data points as the cluster centres, calculating the Euclidian distance between each data point and cluster centre, and assigning each point to the cluster that is closest.

The main algorithm alternates between two stages where data points are reassigned to cluster centres. In the *optimal transfer stage*, each data point I is iterated through and re-assigned to its current closest cluster centre. After each re-assignment, cluster centres are recalculated. In the *quick-transfer stage*, each point is iterated through again, but this time it is only checked if the point's second-closest cluster that was saved in the optimal transfer stage has since become the point's closest cluster, and appropriate re-assignments are made. The algorithm alternates between the optimal transfer stage and the quick-transfer stage until there are no more reassignments.

The number of clusters in the k-means algorithm is taken as an input, which means that some method of determining the number of clusters in the data needs to be implemented. In this thesis, the clustering is first performed with different number of clusters k and the TSS of each clustering scheme is saved. A penalty term is then added to each TSS value, so that the higher the number of clusters is, the larger the penalty term is. The penalty term used in this thesis is inspired by the Bayesian Information Criterion (BIC). So, for each number of clusters k, the following value is calculated

$$\text{TSS} + \frac{1}{2} k \; m \ln n, \tag{4.6}$$

where n is the number of observations and m is the length of the data point vectors. The clustering scheme that minimises Equation (4.6) is chosen.

4.5.2 Maximum likelihood estimation

The MNL model is estimated using maximum likelihood estimation (MLE). In MLE, the model parameters are estimated by maximising the likelihood of the data sample, so that the sample is most probable under the assumed statistical model. The likelihood of a data sample is the joint probability, or the product of the probability of each data point given the model parameters (Chandukala et al., 2008). The model parameters that are estimated with MLE are called maximum likelihood estimates (MSE).

Let us evaluate the log-likelihood function of a linear MNL model given the observed data sample of T choices out of $J+1$ alternatives. The probability of choosing product j at choice occasion i is

$$p_{ji} = \frac{\exp(\theta X_{ji})}{\sum_{k=1}^{J} \exp(\theta X_{ki})}.$$

where X_k a is vector of the model variables and θ a vector of model parameters. The likelihood of choice occasion i is

$$p_i = \prod_j p_{ji}^{Y_{ji}},$$

where $Y_{ji} = 1$, if product j was chosen and 0 otherwise. The likelihood function is obtained by multiplying the probabilities of each choice in the data sample. For computational convenience, the logarithmic transformation of the likelihood function is maximised instead of the regular likelihood function. The log-likelihood function for the MNL model is

$$\begin{aligned}
\log \mathcal{L}(\theta) &= \log \prod_i \prod_j p_{ji}^{y_{ji}} = \sum_i \sum_j y_{ji} \log p_{ji} \\
&= \sum_i \sum_j y_{ij} \log(\frac{\exp(\theta X_{ij})}{\sum_k \exp(\theta X_{ik})}) \\
\log \mathcal{L}(\theta) &= \sum_i \sum_j y_{ij} \theta X_{ij} - \sum_i \sum_j y_{ij} \log(\sum_k \exp(\theta X_{ik})) \qquad (4.7)
\end{aligned}$$

The Newton-Raphson method is used to find maximum likelihood estimates $\theta^\star$ that maximise Equation (4.7) given the observed choice indicators Y. The basic principle of Newton-Raphson is to iteratively move towards the gradient of the log-likelihood function by a step size that is the negative inverse of the Hessian matrix of the log-likelihood (Süli and Mayers, 2003). The log-likelihood is globally concave, which means that there is a unique global maximum and that the Newton-Raphson method is guaranteed to increase the likelihood function at each iteration. The iterations are repeated until the increase in log-likelihood is less than a predefined convergence parameter.

4.5.3 Markov Chain Monte Carlo

The MNP model is fit using Markov Chain Monte Carlo (MCMC) simulation. MCMC is a type of Monte Carlo integration that uses Markov Chains in sampling data from complex probability distributions (Gilks et al., 1996).

MCMC is widely applied in Bayesian modelling, where it is used to sample from complex posterior distributions. In the estimation of multinomial probit models, MCMC is used to estimate the model parameters.

MCMC combines the principles of Monte Carlo integration and Markov Chains (Gilks et al., 1996). Monte Carlo integration is a method of numerical integration that uses samples of random numbers (Gilks et al., 1996). In Monte Carlo integration, an integral is evaluated by drawing random points from a target distribution and evaluating the integrand at each point. A Markov Chain is a probabilistic process that models a sequence of random variables corresponding to the states of a system, so that the probability of the next state depends only on the state of the previous observation (Ching and Ng, 2006). The Markov Chain can be described using a *transition kernel* $P(X_{t+1}|X_t)$, which gives the probability of the next observation given the current state. Under certain regularity conditions, the Markov Chain has a stationary distribution π, so that when a sequence of samples are drawn using the transition kernel, the distribution of the samples will converge towards the stationary distribution π.

Monte Carlo integration assumes that the target distribution ϕ can be sampled efficiently, which may not always be possible. The idea behind MCMC is to construct a Markov Chain so that its stationary distribution π equals the target distribution ϕ. The samples can then be generated using the Markov Chain instead of the target distribution, since as long as the number of drawn samples is large enough, the resulting sample will resemble a sample drawn directly from the target distribution.

The most commonly used way of constructing the Markov Chain is *Gibbs sampling* (Gilks et al., 1996). Gibbs sampling relies on the result that it is possible to draw from the joint distribution of a collection of random variables by drawing successively from their conditional distributions (McCulloch and Rossi, 1994). In a bivariate example case, provided by McCulloch and Rossi (1994), in order to sample from the joint distribution of $\pi = (\pi_1, \pi_2)$, one would first draw from the conditional distribution $\pi_1'|\pi_2$ and then from the distribution $\pi_2'|\pi_1'$. The sequence of drawn variables forms the Markov Chain that converges to a sample drawn from the target joint distribution of π.

The MCMC algorithm used in the estimation of the MNP model follows Scheme 1 of Algorithm 1 proposed by Imai and Van Dyk (2005). The goal of the algorithm is to estimate posterior values for the deterministic utility function's β parameters and the covariance matrix Σ of the normal distributed error term, conditional on the observed choices Y. The estimation is done using the latent utilities W, so that posterior values are also drawn for the

W parameters. So, a Gibbs sampler is constructed to draw from the target distribution $W, \beta, \Sigma|Y$. A posterior distribution is defined for each parameter, conditional on the current estimates of the other parameters, and the samples for each parameter are drawn in sequence. The samples are drawn in the following steps:

Step 1: For each sample i and alternative $j = 1, \ldots, J-1$ draw W_{ij}^t given $W_{i,-j}^t$, θ^{t-1}, and Y, where

$$W_{i,-j}^t = (W_{i,1}^t, \ldots, W_{i,j-1}^t, W_{i,j+1}^{t-1}, \ldots, W_{i,J-1}^{t-1}).$$

The distribution $W_{ij}^t|W_{i,-j}^t, \theta^{t-1}, Y$ has a univariate normal form that is truncated to positive, if $Y_{ij} = 1$ and truncated to negative if $Y_{ij} = 0$.

Step 2: Draw from β^t given W^t and Σ^{t-1}, using the normal posterior distribution.

Step 3: Draw Σ^t given β^t and $(W_i^t - X_i\beta^t)$ for all samples i, using the Wishart posterior distribution.

The algorithm is continued for a pre-defined number of iterations, and the parameter estimates are calculated as the mean of all draws. Typically, some number of samples are discarded from the beginning of the algorithm, called a *burning period*. The samples can also be thinned, so that some of the samples are removed at a regular *thinning interval*.

Chapter 5

Results

The results are presented in three parts. Section 5.1 presents initial results for model fits in full data sets and an overview of the forecasting accuracies of the different models. Since the full data sets have abundant sales data, it is expected that the sales correlation performs well in comparison to the choice models in the full data sets. However, when there is less data available, detecting statistically significant correlation becomes more difficult, and the choice models should outperform the sales correlation model. Model performance was therefore tested also in data sets that were reduced from the full data sets. The results of fitting the models to the reduced data sets are presented in Section 5.2. Lastly, some findings of the cross elasticity estimates between the products that are calculated using the models are presented in Section 5.3.

For clarity, the four models are abbreviated for figures and tables. The pooled MNL model is simply called 'pooled', the k-means mixture MNLs is called 'mixed', the MNP model is called 'probit' and the sales correlation method is called 'corr'. Consistent names for different kinds of forecasts are also used. A *baseline forecast* is a forecast during a time period when no product is on promotion. A *campaign increase forecast* is the total sales forecast of a product during a time period when the product itself is promoted. A *cannibalisation forecast* is a forecast for a product during a time period when some other product, but not the product itself, is promoted. Following the same naming pattern for actual product sales, *baseline sales* refer to sales during a time when no product is on promotion, *campaign sales* refer to the total sales of a product during a time period when the product is promoted, and *cannibalisation sales* refer to the sales of a product during a time period when another product is promoted.

5.1 Full data sets

The choice models were first analysed by fitting all three models to each data set, and calculating model fits for the training data as well as forecasts for the validation data period. Based on initial testing of fitting the MNP model, the parameter estimates did not change much after around 100 iterations. When fitting the model in A10 and A20, the number of iterations was set to 200 and a burning period of 10 iterations was used. In A30, the number of iterations was 400 and a burning period of 20 iterations and thinning period of 2 was used.

All models fit nicely to the data, and the model parameters of the deterministic utility functions defined in Equations (4.3) and (4.5) received expected values: the coefficient for price was negative in all models and the coefficient for loyalty was positive in the models where it was used as a regressor. As discussed in Section 4.4, the sales correlation method needs separate campaign increase and baseline forecasts, which are provided by a reference choice model. The k-means mixture model was chosen as the reference choice model based on the fit and forecast plots, as well as forecast error calculations, which are presented shortly.

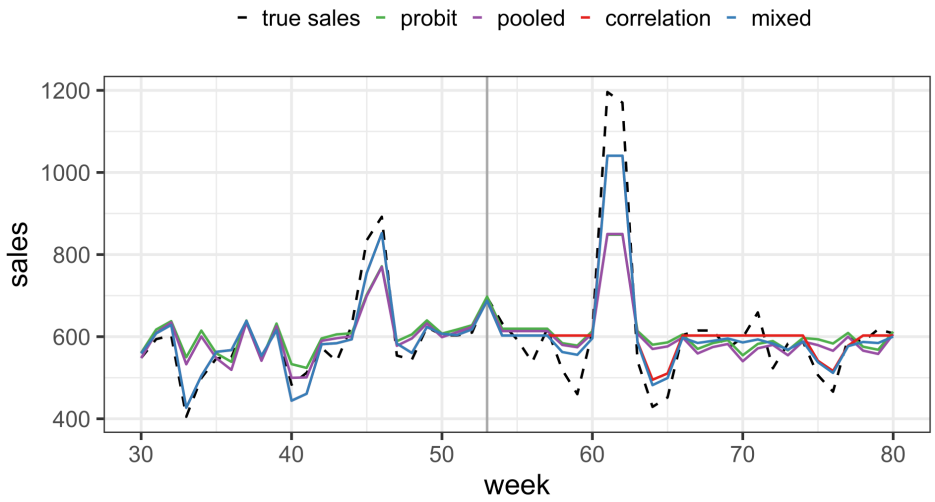
Table 5.1 shows the number of clusters chosen by each model in each assortment, as well as the number of detected substitute product pairs in the sales correlation method. The same clusters were used for both the pooled MNL and MNP models. The number of detected clusters is reasonable in comparison to the underlying choice model used in simulation: in the simulated data, the number of customer segments is 36 in A10 and 72 in A20 and A30, on top of which each customer has a unique price sensitivity parameter and there is additional random variation in the utility function of each customer. The number of detected substitute pairs with the sales correlation method was lower in A30 than in A20, even though there are more product pairs in A30. This is explained by the fact that in A30, there is less sales data per product than in A20, as the number of customers is the same in each data set. What is more, only 20 of the 30 products are set on promotion, which decreases the number of potential substitute product pairs.

	no of clusters in pooled and probit	no of clusters in mixed	no of substitute product pairs detected with sales correlation
A10	53	13	27
A20	56	11	53
A30	43	10	49

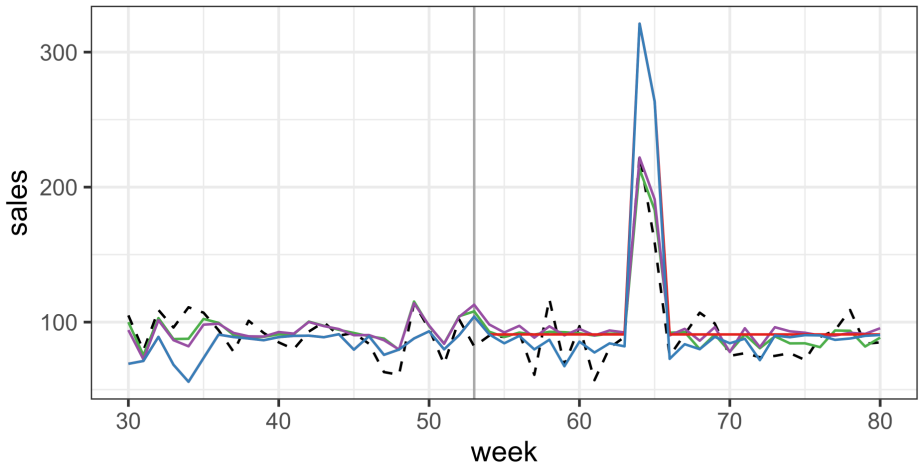
Table 5.1: Number of clusters and detected substitute product pairs by assortment size and fitted model.

When comparing the cannibalisation forecasts of the mixed MNL and the sales correlation method, the cannibalisation forecasts were often similar to each other if the substitution effect was very prominent. However, when the substitution was less evident, there were larger differences between the two models. Looking again at Figure 5.1a, the correlation method and mixed MNL give very similar forecasts during the cannibalisation campaign near week 65. However, between week 66 and 70, there are multiple promotions for other products, which cause the MNL model to forecast a slight decrease in the sales of product 2. Since the sales correlation method does not find significant correlation between the sales increases of the promoted products and the sales decrease of product 2, the forecast is simply set to the baseline sales.

In rare cases, the mixed MNL tended to over-forecast some campaign increases. An example of this is seen in Figure 5.1b which shows model fits and forecasts for product 7 in A20. This could be a sign of the model slightly over-fitting to the data. The pooled MNL and MNP generally gave very similar forecasts to each other in A10 and A20. In A30, however, the MNP model gave generally poor forecasts. This presented as some extreme campaign increases forecasts, as well as some visibly poor baseline forecasts. An example of the latter is shown in Figure 5.2.



(a) Product 2 in data set A10.



(b) Product 7 in data set A20.

Figure 5.1: Model fits and forecasts of weekly sales data during the last year of training data (starting from week 30) and first year of validation data for two examples products in A10 and A20. The plots show that all choice models and the sales correlation method give fairly good models fits and forecasts in these data sets. The grey line marks the separation between training and validation data. Note that the magnitudes of the y-axis change between plots.

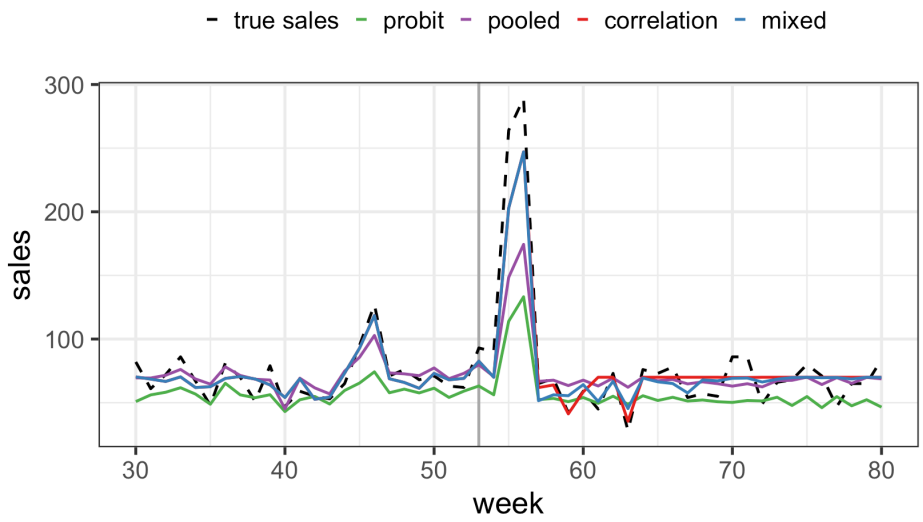


Figure 5.2: Model fits and forecasts of the sales of product 1 in A30 demonstrate that the MNP model fitted poorly to the 30 SKU data set.

Next, the models are compared by their forecasting accuracy in the validation data using forecast error measures. Forecast error is measured using the symmetric mean absolute percentage error (SMAPE) measure proposed by Flores (1986), which measures the relative forecasting error of each forecasting unit. SMAPE is calculated as

$$\text{SMAPE} = \frac{1}{C}\sum_{c=1}^{C} 200\% \frac{|\hat{y}_c - y_c|}{|\hat{y}_c| + |y_c|}, \tag{5.1}$$

where $\hat{y}_c$ is the sales forecast and y_c is the true sales, c is an index of the fitted data point and C is the total number of fitted points c. The errors can be calculated on different levels depending on the purpose. For example, if SMAPE is calculated on product-week level, the indices $c = (t, j)$ are a collection of sales weeks t by product j, and each y_c and $\hat{y}_c$ in (5.1) are the total summed sales and forecast for product j on week t. The advantage of SMAPE is that it is bounded, so that it does not receive extreme error values even when the sales y are close to zero.

In order to get a clear picture of how much the models under-forecast or over-forecast, a forecast bias measure is calculated as well. The bias measure used is

$$\text{bias} = 100\%\left(\frac{\sum_{c=1}^{C} \hat{y}_c}{\sum_{c=1}^{C} y_c} - 1\right).$$

Negative bias values mean that there is under-forecasting and positive bias values mean that there is over-forecasting. Unlike SMAPE, the bias measure is not relative, which means that products with larger sales volumes affect the measure more than low-selling products.

SMAPE and bias values were calculated separately for baseline forecasts, campaign increase forecasts and cannibalisation forecasts in the full data sets. Baseline errors and biases were calculated on product-week level, so that only full sales weeks when no products were on promotion were included in the calculation. Campaign increase errors and biases were calculated on a product-campaign level, so that only campaign sales and forecasts from each product's own campaign periods were included in the calculation. Cannibalisation errors and biases were also calculated on campaign-product level, so that each for each campaign period, the total sales of products other than the promoted product during the promotion period were included in the calculation. Calculating the measures on product-campaign level is possible, since all promotions in the validation period are of equal length within each data set. Promotions are 14 days long in A10 and 7 days long in A20 and A30.

The errors and biases are presented in Table 5.2. Out of the three choice models, the mixed MNL has the smallest baseline, campaign increase and cannibalisation forecast errors in all data sets. The probit model has the largest errors in all data sets. In A10, all methods have large negative biases for campaign increase forecasts, which means that campaign increases are under-forecast. However, in A20 and A30, the mixed MNL model even tends to over-forecast campaign increases, while the probit and pooled MNL models still under-forecast. Comparing the correlation method and the mixed MNL, there are very small differences in the cannibalisation SMAPE errors. The bias, however, is consistently larger in the correlation method forecasts than in the mixed MNL forecasts.

The probit model under-forecasts the sales across the entire time period. A potential explanation for the poor forecasts is that when the number of products increases to 30, the MNP model is too complex to fit well with the available amount of sales data. This is because the number of model parameters increases exponentially by the number of product alternatives, due to the estimation of the error covariance matrix.

	SMAPE (%)				bias (%)			
	pooled	probit	mixed	corr	pooled	probit	mixed	corr
A10								
baseline	6.6	8.8	6.3	6.3*	0.4	0.2	-0.4	-0.4*
increase	19.2	23.5	10.6	10.6*	-12.4	-15.5	-6.5	-6.5*
cann.	15.4	19.2	8.5	10.7	2.3	2.9	2.6	7.8
A20								
baseline	10.7	13.2	10.0	10.0*	0.7	0.4	0.3	0.3*
increase	27.4	37.0	21.0	21.0*	-2.6	-5.1	1.7	1.7*
cann.	19.0	21.8	17.4	17.2	1.5	1.7	0.8	6.8
A30								
baseline	18.6	21.5	18.6	18.6*	0.5	0.3	0.0	0.0*
increase	28.9	45.6	23.7	23.7*	-14.9	-15.9	2.8	2.8*
cann.	21.1	28.6	19.6	20.0	1.8	1.7	-0.6	6.3

Table 5.2: SMAPE forecasting errors and biases in full data sets show that the k-means mixture of MNLs gives the smallest baseline and campaign increase errors out of the three choice models, but that its forecasting accuracy during cannibalisation campaigns is similar to the sales correlation method. Baseline SMAPE errors are calculated on week-product level and campaign increase and cannibalisation errors on campaign-product level. Bias measure values are calculated from total sales and forecasts in the data sets during either baseline, campaign increase or cannibalisation periods.

* Since the correlation method uses the mixed MNL model's baseline and campaign increase forecasts, the corresponding forecast errors and biases are the same between the mixed model and the sales correlation method.

5.2 Reduced data sets

As expected, the sales correlation method performed well in comparison to the choice models in terms of forecast error. Next, the four models were tested on data sets that were reduced from the full data sets in one of two different ways. In the first way, the length of the training data period was limited, which decreased the number of campaign days and data points in the sales correlation test. This simulates a situation, where the products are newly

introduced or less frequently promoted. In the second way, sub-samples of the customers were taken, so that there was less sales data available per product. This simulates a situation where, for example, the product category is less frequently purchased or the visited store is smaller. Only cannibalisation forecasts are compared in this section.

First, results are presented for the experiments where the models were fit to data sets where the length of training data was reduced. In the original data sets, there is 1 year of initialisation data and 2 years of training data, so 3 years of data to train the models in total. In the reduced data sets, the total length of the initialisation and training data was shortened to 2 years and 1 year. The proportion of initialisation and training data was kept the same, so that the first third of the shortened data was used as initialisation data, and the last two thirds were used as training data. The number of actual data used by the sales correlation method depends on the number of campaign days in the sales data. The average number of campaign days in A10 is around 16 per SKU per year, and in A20 and A30 around 9.

In this section, instead of using the relative SMAPE measure, the forecast error is measured using absolute forecast errors. The error is calculated simply as

$$\text{error} = \sum_{c=1}^{C} |\hat{y}_c - y_c|. \tag{5.2}$$

The errors are also examined separately by the sums of over-forecast and under-forecast periods. This is demonstrated by dividing the forecast error in Equation (5.2) to two parts, based on whether the difference between forecast $\hat{y}$ and true sales y during period c is positive (over-forecast) or negative (under-forecast). The exact formulas are

$$\text{over} = \sum_{c=1}^{C} |\max(\hat{y}_c - y_c, 0)|, \quad \text{and}$$
$$\text{under} = \sum_{c=1}^{C} |\min(\hat{y}_c - y_c, 0)|.$$

The obtained errors are shown in Figure 5.3. As expected, the less training data there is, the larger the forecast errors are across all models. The mixed MNL model gives the smallest total forecasting errors and also the smallest over-forecast errors in all data lengths and assortment sizes. However, the sales correlation model gives the smallest under-forecast error in all cases,

which is expected since the model is positively biased. Out of the three choice models, the mixed MNL model gives smallest under-forecast errors. There are also some interesting differences in the errors between the three data sets. In A10, total absolute errors in sales correlation forecasts are almost twice as large as the mixed MNL model's forecast errors when there is only 1 year of data. However in A20 and A30, the differences are smaller.

Considering forecast bias is important in determining if the choice models are superior to the correlation method. Over-forecasting sales may cause excess inventory or even spoilage in case of spoiling product categories, but excess inventory or products that are expiring can be cleared using discounts. Under-forecasting, however, can lead to lost sales and stock-outs, which in turn distorts sales data and induces substitution effects. Under-forecasting is therefore usually avoided at all costs. So, the forecast errors presented in Table 5.3 present an interesting trade-off between the mixed MNL model and the sales correlation method: on one hand, the mixed MNL over-forecasts much less and overall provides more accurate forecasts when the length of the data is reduced. On the the other hand, the sales correlation method has a smaller risk of under-forecasting.

Next, the forecasting accuracy of the models was tested by fitting the models to data sets that had the original 3 years of initialisation and training data, but the number of customers visiting the store was reduced. This was done by taking multiple sub-samples of the customers and fitting the models to choice data of only the sampled customers. Three different sample sizes were determined so that either 50%, 25% or 5% of the original customer base of 5000 customers were sampled. Samples were taken by the customer segments that were used in simulating the data, so that the samples would resemble the original data as much as possible. So, when sampling 50% of customers, half of the customers in each customer segment were randomly sampled, rounded upwards. Since the MNP model performed poorly in the previous tests, it was excluded from this analysis.

Multiple customers samples for each sample size were drawn and the pooled MNL, mixed MNL and sales correlation models were fitted to each sample. The number of customer samples per sample size was 500 for A10 and A20 data sets, and 250 for A30. The cannibalisation forecast error on product-campaign level was calculated using the total error measures defined in Equation (5.2). Figure 5.4 shows the mean of the total error in the customer samples of the size defined in the data portion column, in each of the data sets A10, A20 and A30. The figure shows that while the mixed MNL model has the smallest mean total error in all data sample and assortment

sizes, there is very little difference between the mixed model and the sales correlation method.

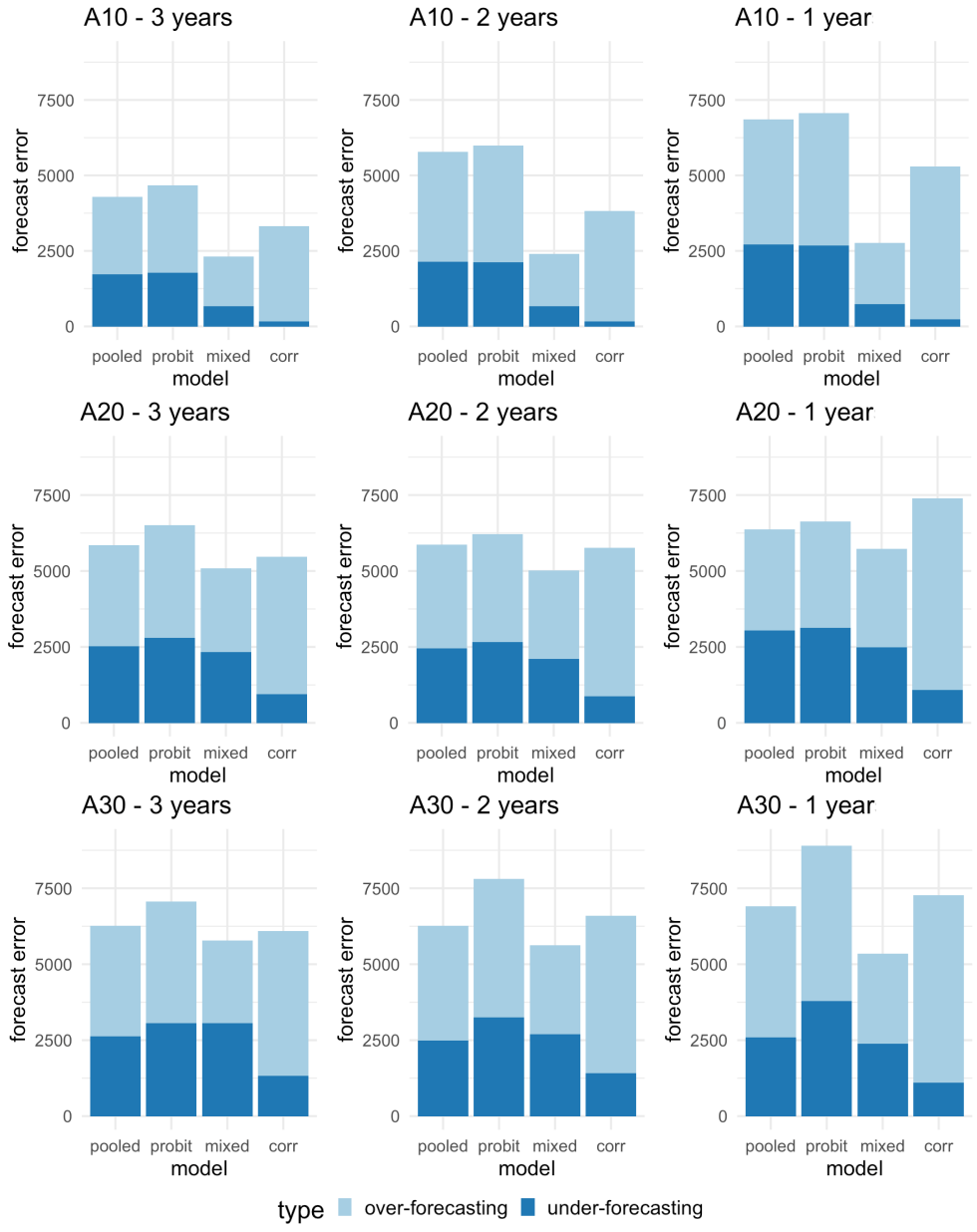


Figure 5.3: Total absolute cannibalisation forecast error in the three data sets with different lengths of training data show that the less data there is, the better the choice models perform in comparison to the sales correlation approach. The choice models still under-forecast sales during cannibalisation more than the sales correlation method in all scenarios.

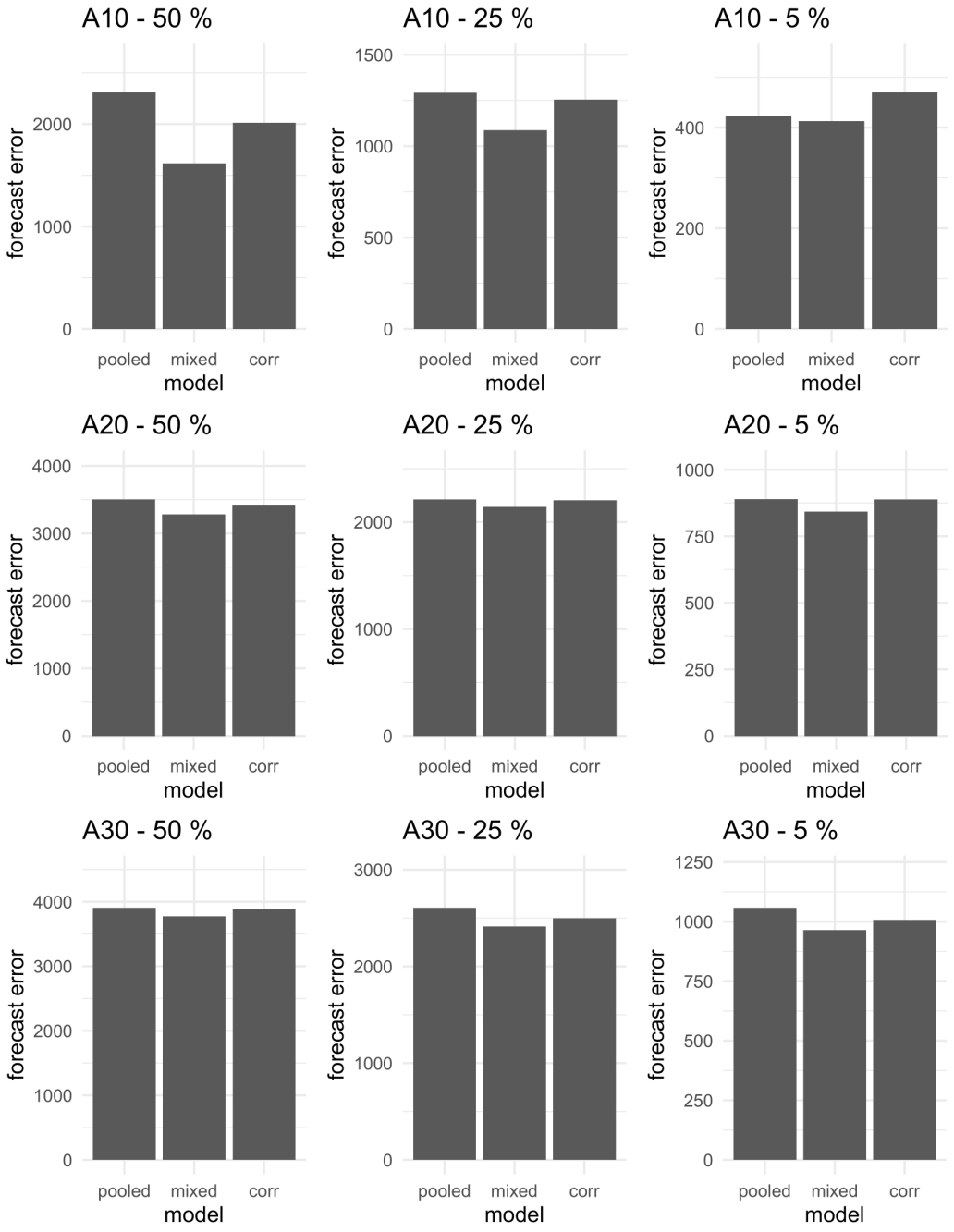


Figure 5.4: Mean of total absolute cannibalisation forecasting error in data sets, where only a portion of the visiting customers were sampled show that when the total sales in the category is reduced, the choice models do not improve the forecasting accuracy much in comparison to the sales correlation method. Note that the y-axes change in magnitude between the plots.

In conclusion, the choice models improved cannibalisation forecasts in comparison to the sales correlation method, when the length of the training data, and thus the number of campaign days used by the sales correlation method, was decreased. However, when the number of customers was decreased, there was almost no improvement in the forecasting accuracy of the choice models to the sales correlation method. So, when the sales volume per product is decreased in the category, the choice models suffer from the lack of data in a similar way as the sales correlation method. A potential explanation for this is that when the number of visiting customers is decreased, the choice models receive less data points to teach the models, as each customer choice is used as an observation. The sales correlation method, however, has the same number of data points available for the correlation test as in the full data sets, and only the daily sales volumes are decreased. When the length of the training data is reduced, the number of data points available to the sales correlation model is also decreased, as there are less campaign days available.

5.3 Cross elasticities

Finally, cross elasticities between the products in the A10 data set were calculated. As discussed in Section 2.2, the magnitude of the substitution between two products can be measured using cross-price elasticity. However, since this research deals with large discounts instead of small incremental changes in price, and since product prices in the simulated data sets have not been discussed much previously in this thesis, cross elasticity is calculated with respect to the sales increase of the discounted product instead of price increase. So, in this section, cross elasticity is defined as a measure of how many percentages the sales of a product change when the *sales* of another product increase by 1%. The exact formula is

$$E_{A,B} = \frac{\Delta Q_A / Q_A}{\Delta Q_B / Q_B}. \tag{5.3}$$

For substitute products, E is negative and a cross elasticity of for example -0.5 would mean that for each 1% increase in sales of product B, the sales of product A would decrease by 0.5%. Cross elasticities are calculated between each promoted product and all other products. The percentage changes are calculated using the campaign increase forecast or the cannibalisation forecast $\hat{y}$ and the baseline forecast b, so that

$$\frac{\Delta Q_A}{Q_A} = \frac{\hat{y}_A - b_A}{b_A}.$$

Table 5.3 shows cross elasticities between a selection of product pairs in A10, using Equation (5.3) and campaign and cannibalisation forecasts calculated using the pooled and mixed MNL models and the sales correlation method. The left-most column shows the chosen promoted products, 2, 13 and 26, and the header shows a selection of other products, 2, 7, 10 and 25. So, the values in the table are the cross elasticities of products 2, 7, 10 and 25 on the demand of products 2, 13 and 26. For the MNL models, cross elasticities are also calculated for no-buys, treating no-buy decisions similarly to product purchases.

B	$E_{2,B}$	$E_{7,B}$	$E_{10,B}$	$E_{25,B}$	$E_{\text{no-buy},B}$
2	-	-0.47	-0.26	-0.40	-0.18
13	-0.04	-0.18	-0.08	-0.10	-0.04
26	-0.13	-0.29	-0.23	-0.41	-0.15

(a) Elasticities calculated with pooled MNL model.

B	$E_{2,B}$	$E_{7,B}$	$E_{10,B}$	$E_{25,B}$	$E_{\text{no-buy},B}$
2	-	-0.36	-0.06	-0.05	-0.41
13	-0.08	-0.29	-0.07	-0.01	-0.04
26	-0.14	-0.12	-0.17	-0.78	-0.22

(b) Elasticities calculated with mixed MNL model.

B	$E_{2,B}$	$E_{7,B}$	$E_{10,B}$	$E_{25,B}$	$E_{\text{no-buy},B}$
2	-	0.0	0.0	0.0	N/A
13	-0.08	-0.35	0.0	0.0	N/A
26	0.0	0.0	0.0	-0.86	N/A

(c) Elasticities calculated with sales correlation model.

Table 5.3: Cross elasticities for a sample of product pairs in A10. Rows show the product that was set on promotion, or the product in the denominator of Equation (5.3). Columns show the cannibalised product, or the nominator product in the equation.

A property of MNL choice models is that when the choice probability of one alternative increases, the probability of all other alternatives decreases. Although the change can be very small for some products, all cross elasticities calculated between any products are always at least slightly negative. When cross elasticities are calculated with the sales correlation method, the value is always zero if no correlation was detected. Therefore, the results of the

sales correlation method are easier to interpret, since there is always a binary outcome depending on whether sales correlation was detected between the products. If a retailer wanted a similar binary result from the choice models, some sort of threshold value for the cross elasticities would need to be decided, so that only product pairs reaching a certain elasticity would be considered substitutes.

Looking at Table 5.3, elasticities calculated using the pooled MNL model have less variation than the ones estimates with the mixed MNL model: for example, looking at the cross elasticities on the demand of product 2, that is, when $B = 2$, the estimates calculated using the pooled model have a similar magnitude to each other, and for example $E_{25,2}$ and $E_{7,2}$ are quite close to each other. In the estimates calculated using the mixed model, however, $E_{7,2}$ is clearly larger than $E_{25,2}$ and $E_{10,2}$. The same happens with when $B = 26$: in estimates made with the sales correlation method and mixed MNL model, product 25 has clearly the largest elasticity on the demand of product 26. In the pooled model, product 25 has the largest elasticity, but the difference to other elasticities on the same row is not very large.

An advantage of using loyalty card data is that no-buy decisions can also be extracted from the data. If the no-buys are treated in a similar way to product choices, cross elasticities can also be calculated between the promoted products and no-buys. Based on these elasticities, conclusions can then be drawn on which product promotions increase the total sales volume in the product category, instead of simply cannibalising other products. For example, looking at estimates on the no-buy elasticity given both by the pooled and mixed MNL models, the largest elasticity of no-buys is on the sales of product 2. Like it was noticed earlier in Figure 3.2, a promotion on product 2 causes the number of no-buys to clearly decrease.

Chapter 6

Conclusion

In this thesis, three methods based on random utility choice models that utilise customer-specific product choice data were used to estimate the magnitude of substitution effects during price promotions. This was done by forecasting sales during discount campaigns for both the promoted product and its potential substitute products using the three models. The forecasting accuracies of the three models for cannibalised products were then compared to benchmark forecasts obtained using a simple sales correlation method that only used aggregate daily sales data. The focus was on situations where there is insufficient sales data for the sales correlation model to detect meaningful correlation, so the product sales data was reduced by either shortening the training data or reducing the total daily sales. The data used in the research was generated using a simulation.

Out of the three choice models, the k-means mixture of MNLs provided the best campaign increase and cannibalisation forecasts. The MNP model provided the worst forecasts out of the choice models, especially in the largest assortment of 30 SKUs. Furthermore, the model was computationally extremely heavy in the larger assortments. The performance of the pooled MNL was between the other two models: it did not reach the forecasting accuracy of the k-means mixture of MNLs, but it also did not provide bad forecasts in any assortment like the MNP model. The relative forecast errors increased along with the size of the product category with all models.

When comparing the forecasting accuracies of the best choice model, the k-means mixture of MNLs, against the sales correlation method in the reduced data sets, the choice model performed slightly better. The k-means mixture of MNLs outperformed the sales correlation method especially when

the length of the training data was shorter. However, when the sales volume was decreased, there was little difference in the forecasting accuracy of the two models. A major difference between the choice models and the sales correlation method was that the choice models tended to under-forecast sales of products that were cannibalised, whereas the sales correlation method tended to over-forecast the sales. The sales correlation method performed surprisingly well, considering that the p-value threshold used in the identification of cannibalisation pairs was not even optimised beyond setting it to the commonly used 0.05. A potential explanation is that since all products had many campaign days with different levels of price discounts, the generated data was very clean compared to real retail data. So, the sales correlation most likely performed better in this study due to the high quality of the data, than it would perform if real data was used.

Based on the results of the thesis, choice models have the most potential to improve measurements of substitution effects for products that have little sales data for past price promotions, for example new products and other products that have been rarely promoted. In particular, the models could be used in smaller product categories and in situations where the risk of under-forecasting is not a great problem, such as for spoiling goods. The k-means mixture of MNLs could potentially also be used in customer segmentation to get insights into the customer base through the model parameters like price sensitivity.

The research raised multiple potential topics for further research. First, as the results were obtained using simulated data, the logical next step would be to test the methods using real loyalty card sales data. It would be useful to test the models in situations where the loyalty card data does not follow the assumptions that the choice models make. For example, there can be errors in the retailer's product category definitions, so that the category could include a complement product to the other products in the category, or alternatively, there could be some key substitutes that have been categorised to some other product group. In the generated data used in this thesis, customers always maximised their utility, but in real life, jealousy effects, for example, can cause customers to behave irrationality when making purchasing choices (Feinberg et al., 2002). Another important research direction would be to include product attribute information in the models to reduce the number of model parameters in the data, which could make the MNP model feasible in larger product categories. Adding attribute information would also provide more information of the products to the model, which could potentially further improve the choice models in comparison to the sales correlation method. A very different direction of future research would

be to explore using loyalty card data with other methods than choice models. For example, the loyalty card data could be utilised in adding heterogeneity to models that only use total daily sales, following, for instance, the research of Pauler and Dick (2006).

Bibliography

K. L. Ailawadi, K. Gedenk, and S. A. Neslin. Heterogeneity and purchase event feedback in choice models: An empirical analysis with implications for model building. *International Journal of Research in Marketing*, 16(3): 177–198, 1999.

A. W. Allaway, R. M. Gooner, D. Berkowitz, and L. Davis. Deriving and exploring behavior segments within a retail loyalty card program. *European Journal of Marketing*, 40(11/12):1317–1339, 2006.

G. M. Allenby, J. Kim, and P. E. Rossi. Economic models of choice. In B. Wierenga and van der Lans R, editors, *Handbook of Marketing Decision Models*, pages 199–222. Springer, 2017.

D. Besanko, D. Dranove, M. Shanley, and S. Schaefer. *Economics of strategy*. John Wiley & Sons, 2009.

E. T. Bradlow, M. Gangwar, P. Kopalle, and S. Voleti. The role of big data and predictive analytics in retailing. *Journal of Retailing*, 93(1):79–95, 2017.

W. Buckinx, G. Verstraeten, and D. Van den Poel. Predicting customer loyalty using the internal transactional database. *Expert systems with applications*, 32(1):125–134, 2007.

R. E. Bucklin and S. Gupta. Brand choice, purchase incidence, and segmentation: An integrated modeling approach. *Journal of Marketing Research*, 29(2):201–215, 1992.

S. R. Chandukala, J. Kim, T. Otter, and G. M. Allenby. *Choice models in marketing: Economic assumptions, challenges and trends*. Now Publishers Inc, 2008.

W.-K. Ching and M. K. Ng. *Markov chains: Models, algorithms and applications*. Springer, 2006.

P. K. Chintagunta, D. C. Jain, and N. J. Vilcassim. Investigating heterogeneity in brand preferences in logit models for panel data. *Journal of Marketing Research*, 28(4):417–428, 1991.

M. Cortiñas, M. Elorz, and J. M. Múgica. The use of loyalty-cards databases: Differences in regular price and discount sensitivity in the brand choice decision between card and non-card holders. *Journal of Retailing and Consumer Services*, 15(1):52–62, 2008.

J. G. Dawes. Brand-pack size cannibalization arising from temporary price promotions. *Journal of Retailing*, 88(3):343–355, 2012.

G. R. Dowling and M. Uncles. Do customer loyalty programs really work? *Sloan management review*, 38(4):71–82, 1997.

P. S. Fader and B. G. Hardie. Modeling consumer choice among skus. *Journal of marketing Research*, 33(4):442–452, 1996.

F. M. Feinberg, A. Krishna, and Z. J. Zhang. Do we care what others get? a behaviorist approach to targeted promotions. *Journal of Marketing research*, 39(3):277–291, 2002.

M. Felgate and A. Fearne. Analyzing the impact of supermarket promotions: a case study using tesco clubcard data in the uk. In D. on M, editor, *The Sustainable Global Marketplace. Developments in Marketing Science: Proceedings of the Academy of Marketing Science*, pages 471–475. Springer, 2015.

B. E. Flores. A pragmatic view of accuracy measurement in forecasting. *Omega*, 14(2):93–98, 1986.

D. Fok. Advanced individual demand models. In P. S. Leeflang, J. E. Wieringa, T. H. Bijmolt, K. H. Pauwels, et al., editors, *Advanced methods for modeling markets*. Springer, 2017.

W. R. Gilks, S. Richardson, and D. J. Spiegelhalter. *Markov Chain Monte Carlo in Practice*. Springer Science+Business Media Dordrecht, 1996.

P. M. Guadagni and J. D. Little. A logit model of brand choice calibrated on scanner data. *Marketing science*, 2(3):203–238, 1983.

S. Gupta. Impact of sales promotions on when, what, and how much to buy. *Journal of Marketing research*, 25(4):342–355, 1988.

J. A. Hartigan and M. A. Wong. A k-means clustering algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1):100–108, 1979.

R. J. Hyndman and G. Athanasopoulos. *Forecasting: principles and practice*. OTexts, 2018.

K. Imai and D. A. Van Dyk. A bayesian analysis of the multinomial probit model using marginal data augmentation. *Journal of econometrics*, 124 (2):311–334, 2005.

W. A. Kamakura and G. J. Russell. A probabilistic choice model for market segmentation and elasticity structure. *Journal of marketing research*, 26 (4):379–390, 1989.

A. Krishna, I. S. Currim, and R. W. Shoemaker. Consumer perceptions of promotional activity. *Journal of Marketing*, 55(2):4–16, 1991.

P. A. Lewis and G. S. Shedler. Simulation of nonhomogeneous poisson processes with degree-two exponential polynomial rate function. *Operations Research*, 27(5):1026–1040, 1979.

C. Mauri. Card loyalty. a new emerging issue in grocery retailing. *Journal of Retailing and Consumer Services*, 10(1):13–25, 2003.

R. McCulloch and P. E. Rossi. An exact likelihood analysis of the multinomial probit model. *Journal of Econometrics*, 64(1-2):207–240, 1994.

D. McFadden. Conditional logit analysis of qualitative choice behavior. In P. Zarembka, editor, *Frontiers in Econometrics*. New York: Academic Press, 1973.

National Research Council. *Improving Data to Analyze Food and Nutrition Policies*. National Academies Press, 2005.

W. Nicholson and C. Snyder. *Microeconomic theory: Basic principles and extensions*. Thomson/South-Western, Mason, Ohio, 2008.

R. Pasupathy. Generating homogeneous poisson processes. In J. J. Cochran, L. A. Cox, P. Keskinocak, J. P. Kharoufeh, and J. C. Smith, editors, *Wiley encyclopedia of operations research and management science*. Wiley Online Library, 2011.

G. Pauler and A. Dick. Maximizing profit of a food retailing chain by targeting and promoting valuable customers using loyalty card and scanner data. *European Journal of Operational Research*, 174(2):1260–1280, 2006.

J. H. Pedrick and F. S. Zufryden. Evaluating the impact of advertising media plans: A model of consumer purchase dynamics using single-source data. *Marketing Science*, 10(2):111–130, 1991.

R. Phillips. *Pricing and revenue optimization*. Stanford university press, 2020.

S-Ryhmä. Monella on jemmarahaa, josta he eivät tiedä -- jopa 350 miljoonaa euroa käyttämätöntä rahaa s-pankissa asiakasomistajien tileillä, 2020. URL `https://s-ryhma.fi/uutinen/monella-on-jemmarahaa-josta-he-eivat-tieda-jopa-35/11WAXq9Hd8uZ26X8PRUu6k`.

A. D. Shocker, B. L. Bayus, and N. Kim. Product complements and substitutes in the real world: The relevance of "other products". *Journal of Marketing*, 68(1):28–40, 2004.

A. Smith, L. Sparks, S. Hart, and N. Tzokas. Retail loyalty schemes: results from a consumer diary study. *Journal of Retailing and Consumer Services*, 10(2):109–119, 2003.

E. Süli and D. F. Mayers. *An Introduction to Numerical Analysis*. Cambridge University Press, 2003.

K. E. Train. *Discrete choice methods with simulation*. Cambridge university press, 2009.

H. C. V. Trijp, W. D. Hoyer, and J. J. Inman. Why switch? product category–level explanations for true variety-seeking behavior. *Journal of marketing research*, 33(3):281–292, 1996.

H. J. Van Heerde and S. A. Neslin. Sales promotion models. In B. Wierenga and R. van der Lans, editors, *Handbook of marketing decision models*, pages 13–77. Springer, 2017.

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