University of Economics, Prague
Faculty of Informatics and Statistics
Department of Econometrics

Doctoral Thesis

Analysis of Client and Product Base: Specifics of Retail Transaction Data

Ing. Ondřej Sokol

Supervisor: prof. RNDr. Ing. Michal Černý, Ph.D.
Study Field: Econometrics and Operational Research
Prague 2020

Title: Analysis of Client and Product Base: Specifics of Retail Transaction Data

Author: Ing. Ondřej Sokol

Supervisor: prof. RNDr. Ing. Michal Černý, Ph.D.

Abstract: Every retail transaction must be recorded. This means thousands to millions of transaction data rows per day for retail chains. Automatic processing of this data is crucial for the company's competitiveness. The importance of efficient processing using data driven methods – which minimize the analyst's personal opinion in order to maximize automation – is increasing. The proposed methods are based on the assumption that there is a large amount of data describing the purchase of a particular product, along with additional information such as baskets, links to a particular customer and his properties, and location and time of purchase. Based on this information, methods are proposed for (1) clustering the products into groups of mutual substitutes, (2) segmentation of customers based on their shopping mission, and (3) estimating the number of unique customers of a retail chain. Methods are formulated as optimization problems. Their properties are verified by simulation studies, or compared with alternative approaches. All methods are applied to real data from a Czech retail chain. The results show that the presented methods are better or at least comparable with the common used methods. Their main advantage is a high degree of automation resulting in efficiency. The methods proposed are robust and allow to get additional new information that cannot be obtained by other methods.

Keywords: Retail Business Analytics, Customer Behavior, Cluster Analysis

Název: Analýza klientské a produktové báze: Specifika maloobchodních transakčních dat

Autor: Ing. Ondřej Sokol

Vedoucí: prof. RNDr. Ing. Michal Černý, Ph.D.

Abstrakt: Každá transakce v maloobchodě musí být zaznamenána, což pro obchodní řetězce znamená tisíce až miliony řádků transakčních dat každý den. Automatické zpracování těchto dat je zásadní pro konkurenceschopnost firmy, a tak roste důležitost jejich efektivního zpracování pomocí tzv. data driven metod, při nichž se minimalizuje osobní přístup analytika ve prospěch maximální automatizace. Metody vycházejí z předpokladu, že je k dispozici velké množství dat popisujících nákup konkrétního produktu spolu s doplňujícími informacemi, jako je uspořádání do košů, propojení s konkrétním klientem a jeho vlastnostmi či místo a čas provedení nákupu. Na základě těchto informací jsou navrženy metody pro (1) rozdělení produktů do skupin vzájemných substitutů, (2) segmentaci zákazníků na základě obvyklého důvodu nákupu a (3) odhad počtu unikátních zákazníků maloobchodního řetězce. Metody jsou formulovány jako optimalizační úlohy a jejich vlastnosti jsou ověřeny pomocí simulačních studií, případně jsou porovnány s alternativními přístupy. Všechny metody jsou aplikovány na reálná data z českého maloobchodního řetězce. Výsledky ukazují, že prezentované metody jsou lepší nebo alespoň konkurenceschopné v porovnání se standardně používanými metodami, přičemž jejich hlavní výhodou je vysoká míra automatizace, a tím i efektivita. Metody jsou robustní a navíc umožňují získat dodatečné nové informace, které ostatními metodami získat nelze.

Klíčová slova: maloobchodní analýza dat, nákupní chování zákazníků, shluková analýza

Preface

In my Ph.D. thesis, I study the statistical methods using retail transaction data. The thesis is focused on processing transaction data as a specific data structure with the aim to reveal properties of various objects involved in transaction.

The thesis is a compilation of following articles, to which I significantly contributed, with unifying theme of data driven methods over transaction data:

1. Vladimír Holý, Ondřej Sokol, Michal Černý. Clustering Retail Products Based on Customer Behaviour. *Applied Soft Computing*, 2017. `https://doi.org/10.1016/j.asoc.2017.02.004`.

2. Ondřej Sokol, Vladimír Holý. Customer Segmentation Based on a Shopping Mission in the Retail Business. Submitted to *International Journal of Market Research* in November 2019.

3. Ondřej Sokol, Vladimír Holý. How Many Customers Does a Retail Chain Have? Submitted to *Marketing Science* in June 2019.

The articles had been already submitted to the journals and one was published in February 2020. The others are still in the review process.

The papers are included in this thesis with each paper forming one chapter. In each chapter a data-driven method inspired by a retail industry problem involving transaction data is proposed. While the setup is similar for each chapter, the goal of each method is different. Important part of two chapters is extensive simulation studies. The simulations are conducted in order to investigate the robustness of the methods in dependence of assumption's violation.

The work on the thesis was supported by the Internal Grant Agency of the University of Economics, Prague Project No. F4/63/2016 (Analysis of Financial High-Frequency Data: Estimates in the Presence of Market Microstructure Noise), F4/58/2017 (Modern Methods of Dealing with Uncertainty in Statistical and Optimization Models), F4/19/2019 (Aproximation of unobservable factors in

classification tasks) and by the Czech Science Foundation Project No. 19-02773S (Streaming financial data and related identification and optimization problems).

I wish to thank my supervisor Michal Černý for his guidance and collaboration and my colleague Vladimír Holý for collaboration on the key papers.

I declare that this thesis and the work presented in it are my own except for the shared authorship of the indicated parts. The literature and supporting materials are mentioned in the bibliography.

February 16, 2020, Prague

Ondřej Sokol

Contents

1. Introduction

1.1 Transaction Data

At present, more data than ever before are available every day and their processing is crucial for competitiveness in every industry area. As the data amount is growing, it is important to process the data automatically with only limited expert's input. Henceforth, the data driven methods allowing us to find properties of the data not yet seen or data generating process are a highly discussed and demanded topic.

The focus is placed especially on the transaction data in retail business, sometimes called market basket data. Retail chains have huge amount of market basket data available. Analysts have access to every single basket – a set of items that were bought together along with the revenue. The specific aspect of the Czech retail market is also a big popularity of loyalty programs which allows to link receipts to a single customer.

Transaction data, in the general sense, is the information recorded from a transaction. The result of the transaction is summarized in a receipt. A receipt is, in fact, a set of products related to individual transaction along with the time dimension and the information about a buyer and a seller. Sometimes more data can be added such as product or buyer/seller properties.

The example from retail industry allows us to easily describe transaction data. A transaction occurs when a customer selects a set of products and proceed to cashier to pay. After the payment, the receipt is created listing products bought together in one basket along with information about the price of purchase, date and time of

transaction and link to certain customer in case the customer is a member of loyalty program. Moreover, the properties of each product sold in retail store are usually available in a seller database and can be used to further describe the transaction. Similarly, the information about the customer (if he is a member of loyalty program) can also be added to the transaction. Exploring this data structure allows us to infer new information about customer behavior which is hard to retrieve by standard statistical approaches.

While the retail industry is the well-known for having a huge amount of transaction data, this feature is not limited only for retail industry. For example, a visit of a web page has similar features – user visits a web page where he visits a set of articles and their images, performs a various actions such as clicking on a link. All of it is done in one session, and creates a dataset similar to the structure of a receipt.

1.2 Terminology

The transaction dataset can be described as follows: A *product* is characterized by the brand, physical properties, and purpose. Note that the price of the product and whether the product is in sales promotion can vary over time and therefore it is assigned to the receipt data. Because there are usually many products, it is useful to aggregate them into *product categories*. The categorization can be done in many ways, e. g. by product purpose, price-level or package size. It is assumed that each purchased product belongs to a single level of a product category.

An individual product bought by the customer is referred to as the *purchased product*. A *purchased basket* is a set of purchased products. A specific purchased basket is a subset of all purchased products. A *receipt* is a purchased basket with additional information about the customer ID, prices, sales promotion, date and time.

A *customer history* is a set of purchased baskets. A specific customer history is a subset of all purchased baskets. A *customer* is a customer history with additional

information about the contact, gender, age, number of children and other demographic information. The customer history is by default available only if the customer is a member of loyalty program and identifies himself using loyalty card during a transaction or in case another customer identification is available, for example credit card number.

1.3 Motivation

Proper analysis of retail transaction data allows us to reveal not only product sales trends, seasonal variations or customer peak traffic periods but also identify customer behavior, discover customer shopping patterns and trends. Such findings are crucial for qualified decision making. It can provide various information covering each aspect of the industry needs for improving performance. It helps to achieve better customer retention and satisfaction as the retailer can go through customer's past history and align the marketing strategy according to it.

The customer's past history along with personal information on the customer, which can be gathered via customer's membership in loyalty program, allows even more. We can predict customer potential, improve customer loyalty or even forecast demand for specific products based on the analysis of customers' behavior. Knowledge about the differences in shopping behavior of different groups of customers generally helps to retain the old customers and acquire new ones. In particular, it can be used to predict the effect of sales promotions.

Another well-studied area is market basket analysis – study of natural affinities between two (or more) products. Usually, market basket analysis focuses on the probabilities of joint purchases of two products. Such knowledge can be used for optimal product placement on shelves in a store or for planning of advertising or promotion sales in marketing.

However, it is very difficult to discover relation between products or customer

shopping behavior from the data itself. For this reasons, the proper data analysis is often to a large extent complemented by expert's input which may be flawed for many reasons. In this thesis, I propose new data driven methods for understanding the customer base using their transaction data with very limited expert's input.

1.4 Contribution and Outline

The thesis focuses on analysis of transaction data from retail industry, drugstore retail transaction data in particular. These are used for all of the proposed methods. The thesis tries to answer various retail-inspired questions:

- Which products are perceived by customers as substitutes?
- How to evaluate customer behavior?
- How to estimate the number of unique customers?

In order to find answers to these questions, the methods are formulated as optimization problems and subsequently solved by numerical techniques. On the contrary to the basic data mining methods which rely mostly on testing of simple hypothesis, we focus on much more complex hypothesis the answer to which cannot be found using random search approaches.

Each chapter represents an article which was submitted to a journal. As such, the chapters deal with different problems and while they share common area of interest and similar approach, they are meant to be independent. Each chapter is introduced with short structured problem statement and abstract.

Chapter 2 focuses on the categorization of retail products. It is a common practice to classify products using their quantitative and qualitative characteristics. In this chapter, a purely data-driven approach is used. The applied clustering of products is based exclusively on a customer behavior. We propose a method for clustering retail

products using market basket data. The model is formulated as an optimization problem and is solved by a genetic algorithm. It is demonstrated on simulated data in order to show how our method behaves for different data types. The application using real data from a Czech drugstore chain shows that proposed method leads to similar results when compared with the classification by experts. The number of clusters is a parameter of the algorithm presented. It is demonstrated that when more clusters than the original number of categories is used, the method yields additional information about the structure of the product categorization.

The chapter is based on the paper Holý, Sokol and Černý (2017) which was published in Applied Soft Computing in November 2017. My contribution lied in designing the method, proving of computational complexity and conducting the empirical study. Up to February 2020, the paper was cited 12 times (not including auto citations) according to Google Scholar. Three citations are in journals with impact factor. The paper was awarded the Prize of Dean of the Faculty for publication activities of PhD students 2017 and Prize of Rector of University of Economics, Prague for prestigious publication of PhD students 2018.

Chapter 3 deals with customer segmentation. In retailing, it is crucial to understand customer behavior and determine customer value. A useful tool to achieve this goal is a cluster analysis. Typically, a customer segmentation is based on the recency, frequency and monetary value of shopping or the structure of purchased products. Here, the segmentation is based on a shopping mission – a reason why a customer visits the shop. Shopping missions include the emergency purchase of a specific product category and the general purchase. In the applied part, the new segmentation is applied to a drugstore chain. It is shown that proposed segmentation brings unique information about customers and should be used together with the traditional methods of segmentation such as RFM, see Putra, Cahyawan and Shavitri H. (2012).

Chapter 3 is an extended version of working paper Sokol and Holỳ (2018), in which I have designed the method and verified it using the real retail data. The paper was

submitted in November 2019 to International Journal of Market Research.

In Chapter 4, the method for estimating the number of unique customers is proposed. It is assumed that a portion of customers is monitored and easily counted due to the loyalty program while the rest of customers is not monitored. The behavior of customers in both groups may significantly differ which makes the estimation of the number of unmonitored customers a non-trivial task. Firstly, shopping patterns of several customer segments are identified which allows estimating the distribution of customers without a loyalty card. For this purpose, the least squares and maximum likelihood methods are utilized. In the simulation study, the maximum likelihood estimator was found as the most robust method. The proposed approach is applied in an empirical study of a drugstore chain. The actual number of customers estimated by the proposed method is much higher than the number suggested by the naive estimate assuming the constant customer distribution. The proposed method can also be utilized to determine penetration of the loyalty program in individual customer segments. The original idea was to extend the usage of customer segmentation proposed in Chapter 3; however, this was not possible for reasons mentioned in 4.2.2 as more assumptions involving expert's decisions would be needed.

Chapter 4 is based on the working paper Sokol and Holý (2019) which was submitted to Marketing Science in June 2019. In the paper, I have designed the method, which was firstly presented in Sokol (2018a), and provided empirical analysis.

The thesis concludes with summary in Chapter 5 in which the results of the thesis and potential problems of the methods are discussed together with the opportunities for future research.

2. How to Cluster Products using Customer Behavior

Vladimír Holý, Ondřej Sokol, Michal Černý.
Clustering Retail Products Based on Customer Behaviour.
Applied Soft Computing, 2017.

Problem statement

Input: Transaction data – set of receipts in a form of incidence matrix, with the indication which products appeared on the receipt.

Output: Clustering of products such that products within the same clusters are similar in a sense of substitution.

Method: Clustering method as a nonlinear optimization problem minimizing the joint appearance of products from the same cluster on the same receipt.

Abstract

The categorization of retail products is essential for the business decision-making process. It is a common practice to classify products based on their quantitative and qualitative characteristics. We use a purely data-driven approach. Our clustering of products is based exclusively on the customer behavior. We propose a method for clustering retail products using market basket data. Our model is formulated as an optimization problem which is solved by a genetic algorithm. How our method behaves in different settings is demonstrated on simulated data. The application using real data from a Czech drugstore company shows that our method leads to similar results in comparison with the classification by experts. The number of clusters is a parameter of our algorithm. We demonstrate that if more clusters are allowed than the original number of categories, the method yields additional information about the structure of the product categorization.

2.1 Introduction

Categorization is important in retail business decision-making process. Product classification and customer segmentation belong to the most frequently used methods. The customer segmentation focuses on getting knowledge about the structure of customers and is used for targeted marketing. For example Jonker, Piersma and Van den Poel (2004) dealt with customer segmentation and its usability in marketing. Another approach to determining customer segmentation was used by Lockshin, Spawton and Macintosh (1997). Seret, Verbraken and Baesens (2014) proposed a customer segmentation based on self-organizing maps with a priori prioritization in direct marketing.

The product categorization finds even more applications in marketing, e.g. new product development, optimizing placement of retail products on shelves, analysis of cannibalization between products and more general analysis of the affinity between products. Gruca and Klemz (2003) proposed a genetic algorithm to identify optimal new product position. A placement of retail products on shelves was studied by Borin, Farris and Freeland (1994). Finding the right categories is also crucial for sales promotions planning. Cross-category sales promotion effect was studied in detail by Leeflang, Parreño Selva, Van Dijk and Wittink (2008) and Hruschka, Lukanowicz and Buchta (1999).

Retail chains try to minimize costs everywhere. Among others their aim is to minimize the costs of product storage in stores. Storage management reaches the stage when stores often have no reserves in the drugstore storeroom because they are supplied dynamically two or more times per week. Therefore, it may happen that a store runs out of some products. The task is to answer the following questions:

1. How to fill a free place on shelves until the storage is restored.
2. How to find the best substitutes for the original product.

Sold-out products are usually replaced by other ones from the same category, but it is not clear how to best define the categories from this viewpoint. This is the main business motivation behind this chapter.

Products are almost always categorized according to their purpose, package properties, e.g. package size, brand and price level. However, there are different approaches to product categorization. For example Srivastava, Leone and Shocker (1981) used hierarchical

clustering while Zhang, Jiao and Ma (2007) promoted fuzzy clustering. Another interesting possibilistic approach to clustering both customers and products was published by Ammar, Elouedi and Lingras (2016).

Retail chains have huge amount of market basket data, containing sets of items that a buyer acquires in one purchase, which can be used to efficiently model customer behavior, e.g. Tsai and Chiu (2004). However, these data are rarely taken into account in the product categorization. Data from market baskets are usually used for analysis of cross-category dependence for a priori given categories, e.g. Russell and Petersen (2000), Mild and Reutterer (2003), Decker and Monien (2003) and Manchanda, Ansari and Gupta (1999).

This chapter proposes a new method for choosing categories utilizing market basket data. Our method classifies products into clusters according to their common occurrences in the shopping baskets. Sets of products in individual shopping baskets as they were registered by the receipts are the only data used by the method which assigns each product to just one category. The method determines product categories under given assumptions of product dependency in the same category. It is based on the assumption that a customer buys only one product per category. Experience shows that customers who buy one product from a given category are generally less likely to buy also another product from the same category. The method applies a genetic algorithm to market basket data in order to find the best clusters of products based on their joint occurrence in shopping baskets.

Retail companies usually inspect affinity relationship between single products, e. g. sales in the same basket standardized by total sales. However, clustering of products based solely on market basket data in this area is not so common. It can help mainly in organizing shelf and/or maximizing effect of promotional activities such as newsletter promotions with a significant discount. This kind of promotion should attract customers who do not regularly visit the store. For example, the promotion of two products from the same category is not effective as customer usually buys only one of them. The interesting article focused on marketing strategies of associated products was published in Weng (2016) in which the author deals with the problem of association rules when product is marketed later. Our method for clustering may be helpful mainly in markets with high proportion of sales in promotion, such as Czech drugstore market where over half of sales is in promotion.

k-means is usually taken as base method for clustering. Although k-means was proposed

over 50 years ago, it is still one of the most used method. The overview of k-means and its modification can be found in Jain (2010).

The overview of methods for automatic clustering using nature-based metaheuristic methods such as genetic algorithms or swarm intelligence can be found in José-García and Gómez-Flores (2016). Interesting approach using fuzzy chromosome in a genetic algorithm was published by Yang, Kuo, Chien and Quyen (2015). Another possibilistic approach was presented by Ammar, Elouedi and Lingras (2015). Combination of k-means and ant colony optimization was published in Niknam and Amiri (2010).

The objective of this chapter is to present a new method of retail product clustering based on shopping behavior of customers. Usually products which are commonly bought together are clustered into the same cluster. In our method we use different point of view. Our goal is to find clusters that minimize the number of occurrences in which two products from the same cluster are in the same shopping basket. This is the main and the most important difference of our method.

The resulting categorization can be used not only for choosing the products suitable to replace sold-out ones but also for optimizing placement of retail products on shelves or for maximizing profit of sales promotions. To maximize the profit it is more effective to spread promotion across different categories instead of stacking multiple promotions in the same category. The resulting clustering can also help in persuading customer into buying more expensive alternative, e.g. promotion which includes discount on a more expensive product when a product from same category is bought.

The article is organized as follows: In Section 2.2, we present the general idea of the proposed method and its assumptions. In Section 2.3, we test the method using synthetic data to illustrate its performance. We also show how the violation of the assumptions affects the method's results. In Section 2.4 the application to drugstore market basket is presented and its potential to detect clusters in real datasets which have not been found before is demonstrated. The chapter concludes with a summary in Section 2.5.

2.2 Methods

In this section, we propose a new method for clustering retail products based on customer behavior. We also present our approach to evaluate resulting clustering.

To clarify terminology in this article we use *categories* meaning the original product category that was defined expertly based on the character and the purpose of the products. On the other hand, *clusters* are results of our method. Clusters are determined using market basket data only.

2.2.1 Clustering Using Genetic Algorithm

We formulate clustering of retail products as an optimization problem. The goal is to find clustering that minimizes the number of products within the same cluster in one shopping basket. It is based on the idea that in general customers will not buy more than one product from each cluster (products in clusters are similar – thus they need only one). We define a cost function which penalizes a violation of this behavior. For given clustering the cost function calculates weighted number of violations of the assumption that in each basket there is one product from a cluster at most.

We approach clustering as a series of decisions. For each pair of products there is a decision whether these two products should be in the same cluster or not. It is inspired by the Rand index (formulated later in Subsection 2.2.3). Specifically, we minimize the average ratio of incorrect clustering decisions. Here, incorrect is meant in the sense that they lead to multiple products within the same cluster in one shopping basket.

Let n_B be the number of baskets, n_P the number of products and n_C the maximum number of clusters. We define matrix $\boldsymbol{A}$ with n_B rows, n_P columns and elements $a_{i,j}$ as

$$a_{i,j} = \begin{cases} 1 & \text{if product } j \text{ is present in basket } i, \\ 0 & \text{otherwise.} \end{cases} \tag{2.1}$$

A possible clustering is defined as $\boldsymbol{x} = (x_1, \ldots, x_{n_P})'$, where x_j is an integer and $1 \leq x_j \leq n_C$. Elements of vector $\boldsymbol{x}$ correspond to products and their values represent assignment of a product to a cluster.

For each basket $b = 1, \ldots, n_B$ we calculate the total number of decisions D_b as

$$D_b = \binom{d_b}{2}, \qquad d_b = \sum_{j=1}^{n_P} a_{b,j} \tag{2.2}$$

and the number of decisions that lead to multiple products within the same cluster V_b as

$$V_b(\boldsymbol{x}) = \sum_{c:v_{b,c}(\boldsymbol{x})>1} \binom{v_{b,c}(\boldsymbol{x})}{2}, \qquad v_{b,c}(\boldsymbol{x}) = \sum_{j:x_j=c} a_{b,j}. \tag{2.3}$$

The number of violating decisions V_b is dependent on clustering $\boldsymbol{x}$. Finally, we define the cost function as

$$f_{cost}(\boldsymbol{x}) = \frac{1}{n_B} \sum_{b=1}^{n_B} \frac{V_b(\boldsymbol{x})}{D_b}. \tag{2.4}$$

Hence, the cost function equals the average ratio of decisions in which two products from the same cluster are in the same shopping basket. The range of the cost function is from 0 to 1. If there is no basket containing products from the same cluster, then the cost function is 0. On the other hand, if every basket contains only products from the same cluster, then the cost function is 1.

We have considered several other objective function formulations, most notably:

- the ratio of total number of multiple products within the same cluster,
- the average ratio of multiple products within the same cluster over all baskets,
- the average ratio of multiple products within the same cluster over all products,
- the ratio of total number of baskets with multiple products within the same cluster,

but the best clusterings were given by the objective function (2.4) inspired by Rand index.

The whole optimization problem is formulated as

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & f_{cost}(\boldsymbol{x}) \\ \text{s. t.} \quad & x_i \leq n_C \quad \text{for } i = 1, \ldots, n_P, \\ & x_i \in \mathbb{N} \quad \text{for } i = 1, \ldots, n_P. \end{aligned} \tag{2.5}$$

The cost function (2.4) is minimized over all possible clusterings $\boldsymbol{x}$. The maximum number of clusters n_C is a fixed number. If we did not limit the number of clusters, each product would be assigned to its own cluster. Optimization problem (2.5) is an integer non-linear

programming, which is NP-hard as is shown in following Subsection 2.2.2. To efficiently solve this or at least to get approximate solution we use heuristic *genetic algorithm*. Details of our genetic algorithm parameters are discussed in Subsection 2.3.1. Genetic algorithm for solving an integer non-linear program was already used by Jiang, Wang, Chu and Yu (1997) and more recently by Yang, Kuo, Chien and Quyen (2015).

To ensure that resulting clustering is meaningful we need to make following assumptions:

(A1) The probability that customer buys at least two products from one category is strictly less than 50 %.

(A2) The true number of clusters is known.

(A3) Each customer has the same nonzero probability of buying a specific product and the probability is constant in time.

Assumption (A1) tells us that although our model allows customers to buy more than one product within the same cluster in one shopping basket, this behavior is not considered standard but a model error. Assumption (A2) reflects formulation of our optimization problem in which we specify the maximum number of clusters. In almost all cases the resulting number of clusters is the maximum number of clusters (more clusters are preferred by the objective function). Finally, Assumption (A3) is meant to prevent situations in which some customers buy product A and never product B while other customers buy product B and never product A. This behavior would result in assigning products A and B into the same cluster even if they are completely different. Assumption (A3) ensures that with a large enough dataset, all combinations of products will appear in some baskets with probability approaching 1. Later, in Section 2.3 we discuss in more detail how violation of these assumptions would affect the resulting clustering.

The resulting clustering $\boldsymbol{x}$ allows symmetries in the solution space, for example, solution $\boldsymbol{x} = (3, 3, 2, 1, 3)$ is identical to $\boldsymbol{x} = (2, 2, 1, 3, 2)$. This shortcoming can be remedied by introducing a simple rule where cluster numbers are renumbered to start from the smallest based on the first occasion of each cluster, e. g. the solution above would be renumbered to $\boldsymbol{x} = (1, 1, 2, 3, 1)$.

For finding optimal or at least near-optimal solution we use *GeneticAlg.int* function from R package *gramEvol*. The package was firstly published in Noorian et al. (2016). The method is specifically designed to solving integer optimization problems. *GeneticAlg.int* implements

evolutionary algorithms with chromosomes created from integer values in given range, in our case integers from 1 to the number of clusters n_C. The algorithm first creates an initial population and randomly generated individuals. Cost of each individual is evaluated using the cost function. Iteratively, the evolutionary operators including selection, cross-over and mutation are applied to the population to create a new generation. This iteration is continued until the number of generations exceeds given iterations.

In order to improve the results, a simple form of local search heuristic is added to the algorithm. If used, in each iteration is checked whether a cost function of given number of best individuals, set by a parameter, can be improved by change of one gene. In the simulation study 2.3.5, we discuss the appropriate parameters.

2.2.2 Alternative Model Formulation and Computational Complexity

The model can be straightforwardly transformed from integer program to binary program. Let

$$y_{j,k} = \begin{cases} 1 & \text{if product } j \text{ is assigned to cluster } k, \\ 0 & \text{otherwise,} \end{cases} \tag{2.6}$$

and let $\boldsymbol{y}$ is a matrix with elements $y_{j,k}$.

Define

$$V'_b(\boldsymbol{y}) = \sum_{c=1}^{n_C} \binom{v'_{b,c}(\boldsymbol{y})}{2}, \qquad v'_{b,c}(\boldsymbol{y}) = \sum_{j=1}^{n_P} a_{b,j} y_{j,c}. \tag{2.7}$$

The number of violating decisions V'_b is dependent on clustering $\boldsymbol{y}$. Similarly to 2.5, we define the cost function as

$$f'_{cost}(\boldsymbol{y}) = \frac{1}{n_B} \sum_{b=1}^{n_B} \frac{V'_b(\boldsymbol{y})}{D_b}. \tag{2.8}$$

The whole binary programming model is then

$$\begin{aligned} \min_{\boldsymbol{y}} \quad & f'_{cost}(\boldsymbol{y}) \\ \text{s. t.} \quad & \sum_{k=1}^{n_C} y_{j,k} = 1 \quad \text{for } j = 1, \dots, n_P, \\ & y_{j,k} \in \{0, 1\} \quad \text{for } j = 1, \dots, n_P, \quad k = 1, \dots, n_C. \end{aligned} \tag{2.9}$$

Now, we show that the problem is NP-hard.

Lemma 2.2.1. *Problem 2.5 is NP-hard.*

Proof. In order to proof it, we reduce problem 2.9 to the simplest case. Set $n_C = 2$, i.e. let there are only two categories of products, and let $\sum_{j=1}^{n_P} a_{b,j} = 2$ for all b , i.e. the size of all baskets is set to 2. Therefore $D_b = 1$. Also, as we have only two product categories, we can define

$$z_j = \begin{cases} 1 & \text{if product } j \text{ is assigned to cluster 1,} \\ 0 & \text{if product } j \text{ is assigned to cluster 2} \end{cases} \tag{2.10}$$

and let

$$p_{i,j} \coloneqq \sum_{b=1}^{n_B} a_{b,i} a_{b,j} \quad \text{for } i, j = 1, \ldots, n_P, \quad i \neq j, \tag{2.11}$$

we have computed the number of occurrences of every pair of products in the same basket.

The problem can now be formulated as

$$\begin{aligned} \min_{z} \quad & \sum_{\forall i<j} p_{i,j} \mathbb{I}(z_i = z_j) \\ \text{s. t.} \quad & z_j \in \{0, 1\} \quad \text{for } j = 1, \ldots, n_P, \end{aligned} \tag{2.12}$$

where $\mathbb{I}(\cdot)$ denotes the indicator function. Also, we can rewrite the objective function to maximization form as

$$\begin{aligned} \max_{z} \quad & \sum_{\forall i<j} p_{i,j} \mathbb{I}(z_i \neq z_j) \\ \text{s. t.} \quad & z_j \in \{0, 1\} \quad \text{for } j = 1, \ldots, n_P. \end{aligned} \tag{2.13}$$

The decision problem form of 2.13 can be formulated as

Is there a binary vector $z = \{z_1, z_2, \ldots, z_{n_P}\}$ such that for a given n_P, p and W holds

$$\sum_{\forall i<j} p_{i,j} \mathbb{I}(z_i \neq z_j) \geq W? \tag{2.14}$$

Therefore, an instance is given by $[n_P, p, W]$. Now, we show that Max-Cut problem can be reduced to Problem 2.14.

Max-Cut is an NP-hard decision problem (see proof of NP-hardness in Karp (1972)) defined as follows: having a graph $G = (V, E)$, weighting function $w : E \to \mathbb{Z}$ and a positive integer W, is there a set $S \subset V$ such that

$$\sum_{\{i,j\} \in V, i \in S, j \notin S} w(\{i, j\}) \geq W? \tag{2.15}$$

Hence, an instance of Max-Cut problem 2.15 is given by $[V, E, w, W]$.

In order to proof, that Max-Cut problem is reducible to Problem 2.14, we need to shown that every instance of Problem 2.15 is reducible to Problem 2.14. The reduction function is following:

$$f : [V, E, w, W] \mapsto [n, p, W]. \tag{2.16}$$

V is mapped to a vector $(1, 2, \ldots, n)$ where n is the number of vertices. Values of weighting function $w(i, j)$ is directly translated into $p_{i,j}$. If graph $G = (V, E)$ is not complete, then $p_{i,j}$ is set to zero for missing edges. A value of W remains the same.

Note that Max-Cut problem does not assume non-negativity weighting function w and in our setup $p_{i,j}$ are naturally non-negative as it is number of instances that two products are in the same basket; however, this is not a problem. As it is shown in reduction of Knapsack problem to Max-Cut problem through Partition problem (Karp (1972)), there are two cases when the $w(i, j)$ are negative.

1. If the sum of all items in Knapsack form is lower than the capacity, then $w(i, j)$ may be negative. However, such instance is trivial to solve.
2. If the weight of i-th item $w_i^{(k)}$ in Knapsack problem is negative, then $w(i, j)$ may be negative. However, this case can be easily transformed to instance without negative weights. Such transformation consists in multiplication of negative weight $w_i^{(k)}$ by -1, switching affected binary variable $x_i^{(k)}$ to $1 - x_i^{(k)}$ and increase in the total capacity of knapsack by $-w_i^{(k)}$.

Therefore, the cases, in which $w(i, j)$ are negative, are easy to solve.

As a result, every *hard* instance of Weighted Max-Cut problem given by $[V, E, w, W]$ can

be transformed to Problem 2.14 using reduction function f and therefore Problem 2.5 is NP-hard. □

2.2.3 Evaluation of Clustering

We use three different statistics to evaluate our resulting clustering when true categories are known. The first one is *purity* used for example by Manning, Raghavan and Schütze (2008). It is computed in a very straightforward way. Each estimated cluster is assigned to category that is the most frequent in the cluster. Purity is then the ratio of products with correctly assigned categories. It is calculated as

$$I_{PUR} = \frac{1}{n_P} \sum_{i=1}^{n_C} \max_j |C_i \cap R_j|, \tag{2.17}$$

where n_P is the number of products, n_C is the number of estimated clusters, C_i is the set of products in estimated cluster i and R_j is the set of products in real category j. Bad clusterings have the purity close to 0, the perfect clustering has purity equal to 1. However, if the number of estimated clusters is much larger than the number of real categories the purity always has a high value and in this case the purity statistic is not very meaningful.

For this reason we also use a modification of the purity in which we reverse the role of true categories and estimated clusters. *Reverse purity* is defined as

$$I_{REV} = \frac{1}{n_P} \sum_{j=1}^{n_R} \max_i |C_i \cap R_j|, \tag{2.18}$$

where n_R is the number of real categories.

With more clusters than original categories the statistics becomes limited from above. For example if we have 100 products in 10 equally sized categories and 20 equally sized clusters, than the purity is limited to 0.5 from above as each category contains maximum of 5 products from the same cluster. Therefore, purity as well as reverse purity is not appropriate evaluation statistic in case of different number of categories and clusters.

The last statistic we use is the *Rand index* proposed by Rand (1971). It is based on the idea that clustering is a series of decisions that either put two products into the same cluster or put them to different clusters. Therefore the number of decisions is the number of product pairs. Rand index is defined as a ratio of correct decisions and is calculated as

$$I_{RAND} = \frac{P_{TP} + P_{TN}}{P}, \tag{2.19}$$

where P_{TP} is the number of pairs correctly assigned to the same cluster, P_{TN} is the number of pairs correctly assigned to different clusters and P is the number of all pairs. Accurate clusterings have Rand index close to 1.

2.3 Simulation Study

To reveal properties of our method we perform several simulations. In Subsection 2.3.1 we compare different parameters of our genetic algorithm. In three remaining subsections we simulate a violation of assumptions (A1), (A2) and (A3). In all simulations we consider a dataset consisting of 10 000 shopping baskets, each with 4 categories that have one or two products. We have a total of 10 categories with 10 products each. We simulate the behavior of a customer who selects a set of categories he needs (sets of 4 categories are selected with the same probability, apart from Subsection 2.3.4). Then he selects which products from that category he wants to buy (products are selected with the same probability and there is also a 10% chance to buy two products instead of one). We have chosen this characteristics of simulated data because they are similar to the size and the structure of the real dataset we use later in Section 2.4.

2.3.1 Choosing Genetic Algorithm Parameters

To properly use genetic algorithm we need to choose several parameters. The first one is the *size of population*. With bigger population of individuals more possible clusterings are explored, which can result in finding better solution. We set the population size to 500 individuals due to computational complexity.

Initial population is generated randomly and then the algorithm iteratively selects new populations. The number of iterations is another parameter called the *number of generations*. We always use the fixed number of 1 000 generations. However, our simulations show that most clusterings converge much faster. Each candidate solution is represented by an individual. Properties of candidate solutions are encoded in chromosomes of individuals. At each generation a percentage of individuals with lowest values of cost function (called the elite population) passes to the next generation without any alteration. We consider the *ratio of elite population* from 0 to 0.2. In our case elite population does not have big impact

on the best individuals in the last generation, all values result in perfect or almost perfect clustering. It is meaningful to carry over at least the best individuals from the previous generation so the quality of solution will not decrease. We set the elite ratio to 0.1.

Chromosomes of the rest of the new individuals are generated by crossover and mutation operations. For each new individual (called the child) crossover selects two individuals from the last generation (called the parents). The parents are selected randomly with weights according to the sorting by their cost function. The child is then created by combining chromosomes of both parents. We use one-point crossover which means that a random number of chromosomes c is selected. Then the first c chromosomes of the child are taken from one parent while the remaining chromosomes are taken from the other parent. Finally the mutation operation is performed. Each chromosome of the child has a probability of changing its value to a random one. This probability is the last parameter called the *mutation chance*. We consider this parameter to be from 0 to 0.2. Figure 2.1 shows that positive mutation chances up to 0.04 result in perfect or almost perfect clustering. Simulation with no mutation chance gives far worse results showing the importance of mutation. We set the mutation chance to 0.01 to allow algorithm to concentrate on improving one point while retaining some exploratory ability of mutation.

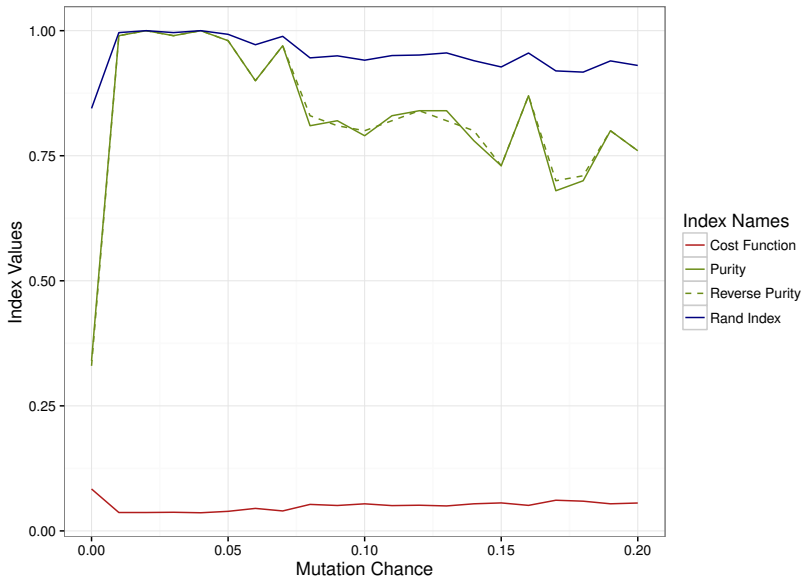

Fig. 2.1: Evaluation statistics for genetic algorithm with different mutation chances.

Next, we analyze combinations of the mentioned parameters. In Figure 2.2 cost functions of the best individuals in each generation are shown for 6 different parameter settings. We combine population sizes $P = (50, 200, 500)$ with mutation chances $M = (0.01, 0.1)$ while the ratio of elite population is set to 0.1. In Table 2.1 several statistics of the resulting clustering are presented. We can see that lower mutation chance leads to faster convergence. There is a risk of lower mutation chance to end up in a local minimum but the results show this is not the case. The population size of 500 individuals with mutation chance of 1% resulted in perfect clustering. In the rest of the chapter, these are the genetic algorithm parameters we use.

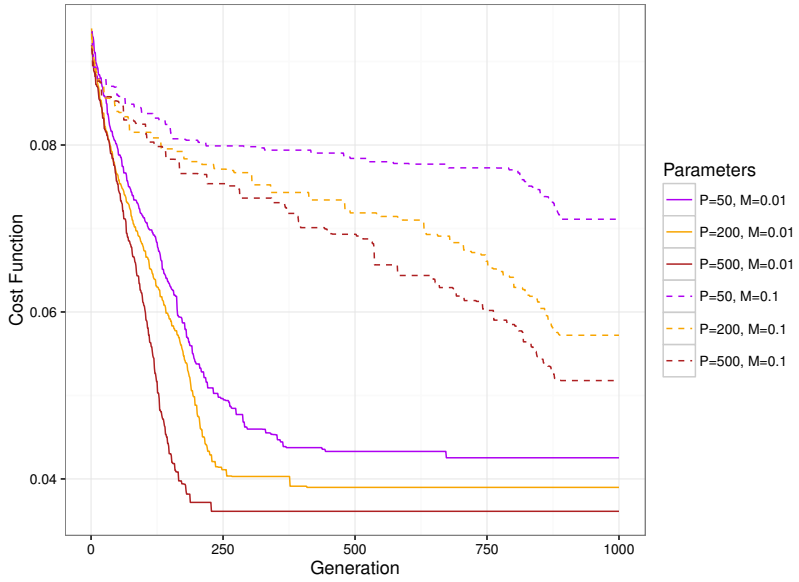

Fig. 2.2: Cost function of best individuals in each generation for different population sizes P and mutation chances M.

2.3.2 Multiple Products Within the Same Category in One Shopping Basket

In this subsection we study the sensitivity of the proposed method to the situation, in which customers buy more than one product from the same category. We simulate data for different probabilities of buying the second product. Results are shown in Figure 2.3.

Tab. 2.1: Statistics of the best individual in the final generation for different population sizes P and mutation chances M.

Population Size	50	200	500	50	200	500
Mutation Chance	0.01	0.01	0.01	0.1	0.1	0.1
Cost Function	0.043	0.039	0.036	0.071	0.057	0.052
Purity	0.940	0.980	1.000	0.550	0.740	0.081
Reverse purity	0.940	0.980	1.000	0.550	0.750	0.081
Rand index	0.978	0.993	1.000	0.886	0.930	0.946

As we can see the method gives almost perfect clustering for the probability of the second product up to 0.18. At probability 0.20 there is a significant decrease in accuracy. This is caused by the loss of relevant information contained in shopping basket data supplied to the cost function. If we could increase the number of observed shopping baskets or the average number of products in a shopping basket we would get more precise results even for the second product probability of 0.20 or higher.

However, if we are sure that some products are substitutes which are commonly bought together, it may be useful to aggregate such products. A good example are oral nutritional supplements, which are usually purchased in large quantities and with many different flavors. This is commonly done in the retail market that variations of same products are taken as one product. However, this contradicts our initial motivation and we recommend to use it only in special cases such as the mentioned supplements.

2.3.3 Unknown Number of Clusters

So far, we have assumed in our simulations that the true number of categories is known. Now, we inspect the behavior of our method when it is used with different number of clusters. Results are shown in Figure 2.4. The question is whether we can identify the correct number of categories. In Figure 2.4 we can see that the purity, the reverse purity and the Rand index have value of 1 for 10 clusters, indicating the perfect clustering. However, in a real application we do not know the true categorization and therefore we cannot calculate the purity statistics or the Rand index. A way to determine how many clusters should be used

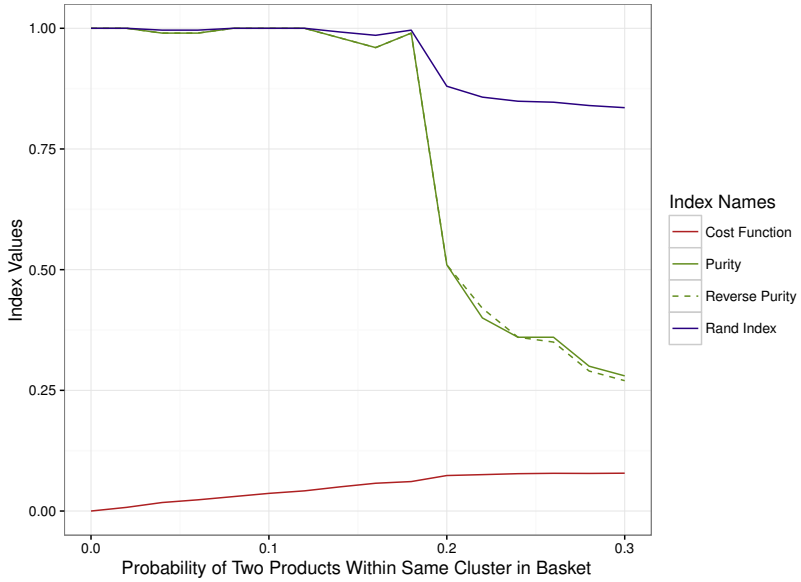

Fig. 2.3: Evaluation statistics for clustering data with different probabilities of second product in the same category in one shopping basket.

is to analyze the shape of the cost function. For a number of clusters smaller than the true number the cost function is significantly decreasing with increasing number of clusters. When the true number of categories has been reached the cost function continues to decrease only by a small amount. As we can see in Figure 2.4 the cost function stops rapidly decreasing with around 10 clusters which is the true number of categories.

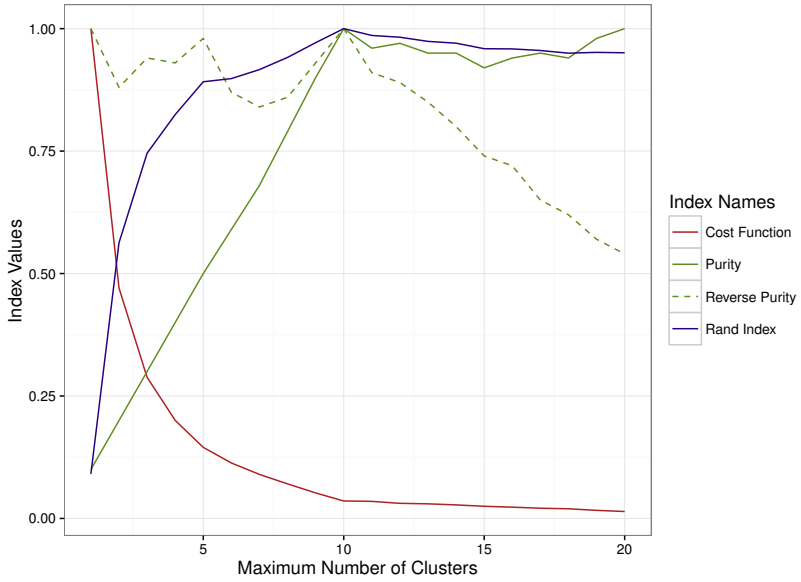

Fig. 2.4: Evaluation statistics for clustering when number of categories is unknown.

2.3.4 Different Types of Customers

We discuss a violation of Assumption (A3). We consider three types of customers. Customer A can buy products from all categories with equal probability. Customer B can buy products only from a half of categories while customer C can buy products only from the other half of categories. We study the behavior of the proposed method for customer structures ranging from all customers being of type A (this was the case of all previous simulations) to half customers being type B and half type C. Results are shown in Figure 2.5. If the customers violating assumption (A3) are in minority the resulting clustering is not affected. However, from the point where customer composition is 50% type A, 25% type B and 25% type C the resulting clustering becomes quite chaotic.

2.3.5 Local Search

Finally, we added a local search feature to the original algorithm and discuss suitable values of its parameters. The local search consists in generating all possible *neighbor* individuals,

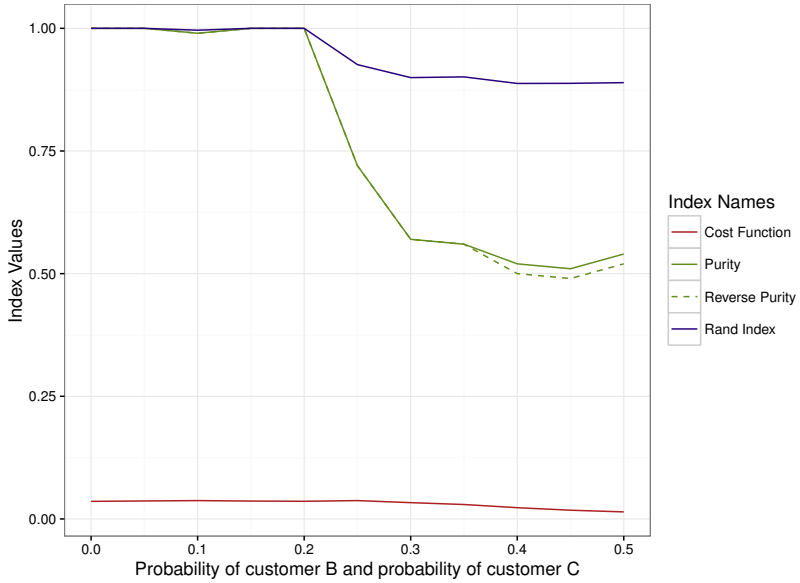

Fig. 2.5: Evaluation statistics for clustering data with different probabilities of occurrence of customers B and C.

created by exchange of one gene, for each of already selected sets of best individuals in each iteration and checking whether the new individuals improve the cost function. If new individual gives better result than the original one, then the original is replaced by the new one. If the individual has already been checked in the previous iterations, then no local search is executed. We study the optimal number of individuals, LS, which are checked by Algorithm 1.

Let GB is an array of best genomes, LS is a given local search parameter, n_C the number of clusters, n_P the number of products and $cost(x)$ is cost function defined in 2.4. The used local search algorithm is following.

Algorithm 1 Local Search algorithm

Require: $GB, n_C, n_P, LS, cost(x)$

for $i \in \{1, \ldots, LS\}$ **do**
  $G := GB[i,]; nCost := cost(G)$
  for $j \in \{1, \ldots, n_P\}$ **do**
    $x = G$
    for $k \in \{1, \ldots, n_C\}$ **do**
      $x[j] = k$
      if $cost(x) < nCost$ **then** $nCost := cost(x); nG := x$
    end for
  end for
  $GB[i,] := nG$
end for
return GB

Because we use the implementation of the genetic algorithm in R, we try to avoid for cycles that are very slow in the R environment. It is much faster to first create a matrix of all adjacent neighbor individuals and then to run the *cost* function in bulk using *apply* function. However, for better clarity, we included the description with for cycles in Algorithm 1.

Figure 2.6 illustrates how the algorithm performs for various numbers of individuals used in the local search method. We show the development of cost function for algorithm without local search and for 1, 2, 3, 4, 5 and 10 numbers of checked individuals. In simulation study, we found that increasing number of checked individuals, quickly decreases the number of iteration needed to find the best solution. On the other hand, the local search method is much more time consuming. In Table 2.3.5 we show the results of simulation study based on 100 generated datasets. The parameters for generating datasets and genetic algorithm are the same as in previous sections. The time values are shown relative to the fastest time needed to converge to the best solution. The cost values are shown relative to the best solution found over the same dataset.

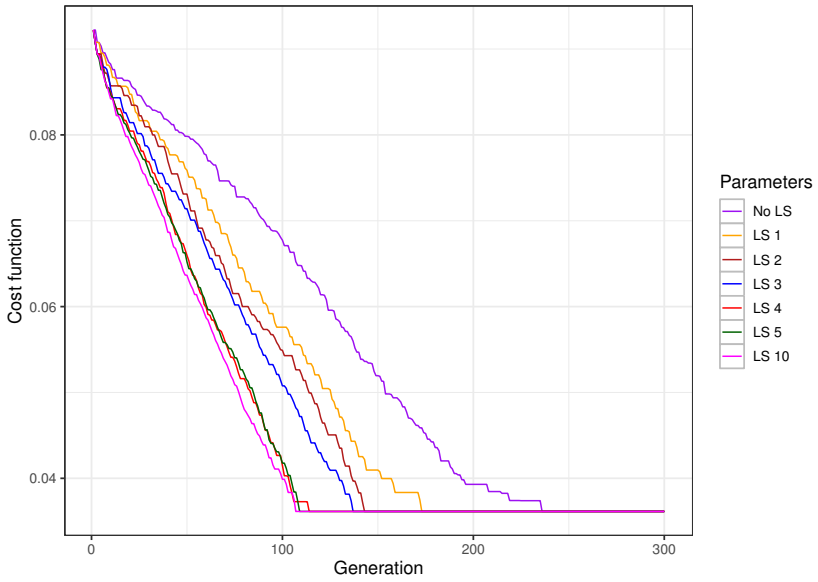

Fig. 2.6: Cost function of best individuals in each generation for different number of chromosomes checked by local search (LS).

Tab. 2.2: Comparison of average results based on local search parameter.

LS	Iterations	Time	Cost
0	322	1.000	1.045
1	278	1.178	1.009
2	205	1.466	1.000
3	164	1.758	1.000
4	142	2.231	1.000
5	130	2.753	1.000

In some cases, we got apparently lower value of the cost function using the algorithm with local search implementation in comparison to the original algorithm. We also found that in all cases the two checked individuals provided same result as larger number of checked individuals. The algorithm without local search is the fastest in terms of time. Algorithms

with parameters of 1 and 2 checked individuals are 17.8 and 46.6 percent slower, and the algorithms with higher parameters are noticeable slower with no benefit. Therefore, we recommend using the local search algorithm with checking two best individuals.

2.4 Data Analysis

In this section, we use our method with a sample of real data. Our dataset consists of individual purchase data of one of drugstore retail chains in the Czech Republic. We take original categories applied to the retail chain that are defined by experts, according to the character and the purpose of the products, as a reference classification.

Customer behavior in the Czech drugstore market is specific. As there is a high density of malls and hypermarkets, customers who visit drugstores usually buy just a few products. The average number of items in a shopping basket in the drugstore is around 3.

2.4.1 Description of Dataset

In this chapter we use a sample of receipts containing at least 4 products from the whole year 2015. The size of our dataset is 10 608 baskets containing 10 best-selling products from 10 most-popular categories defined by the drugstore. Original categories are based mostly on product purpose and price level. Thus, we have 100 different products. The sample is meant to illustrate how our method work and how the clusters are defined – that is the reason why we have chosen exactly 100 products.

Parameters of the genetic algorithm are the same as presented in Section 2.3.1.

To clarify the terminology in this section, original categories are denoted by letters while clusters found by our method are denoted by numbers.

2.4.2 Model with 10 Clusters

We applied the proposed method with the maximum number of 10 clusters and we compared the found clusters with the original categories. Using the real dataset we have found that Assumption (A1) about consumers' behavior was violated in approximately 19% of baskets

on average. The ratios of violations significantly differ for each category as shown in Table 2.3.

The assignment of products to clusters is shown in Table 2.4. It is apparent that the proposed method had a problem with assigning products from categories *C, D* and *E*. Those are the categories with the highest percentage of violations of Assumption (A1).

The method produced 10 clusters which was the maximum allowed. Evaluation statistics of the results of this test are shown in Table 2.5. Purity as well as reverse purity statistics show that some of the products were not assigned as in original categorization. The cost function value of the assignment is 0.0153 which is lower than the value of the expert estimate assignment which is 0.0182. The reason is that the method minimizes the number of products bought together within one cluster. Therefore, if a category suffers from a violation of Assumption (A1), our method puts together products from different original categories into one cluster to minimize the cost function.

Using our method we have found that categories *C, D* and *E* are perceived differently by the customers and by the management. This finding can be further used in designing new product categorization or in defining subcategories.

Tab. 2.3: Violations for each category.

Category	Name	Occurences	Violation ratio
A	Dishwashing liquid	4387	0.022
B	WC liquid cleaners	4212	0.047
C	Handkerchieves and napkins	5769	0.077
D	Soap	2026	0.179
E	Tampons	1837	0.104
F	Toilet paper	6993	0.020
G	Trash bags	4543	0.068
H	Paper towels	4544	0.012
I	Cotton wool and cotton buds	3991	0.026
J	Facial pads	5224	0.012

Tab. 2.4: Assignment of 100 products from original categories *A*-*J* to clusters 1–10.

Original categories									
A	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*J*
7	6	8	3	5	1	4	3	9	10
7	5	2	2	5	1	4	3	9	10
7	6	8	2	5	1	4	3	9	10
7	6	8	5	6	1	4	3	9	10
7	6	2	5	5	1	4	3	9	10
7	6	8	2	5	1	5	3	9	10
7	6	8	2	7	1	4	3	9	10
7	6	8	2	5	1	4	3	9	10
7	6	8	5	5	1	4	5	9	10
7	6	2	2	5	1	4	3	5	10

2.4.3 Model with 8 Clusters

In our next test we assign the same 100 products of 10 categories into 8 clusters. The results are shown in Table 2.6. There is a good correspondence between 8 clusters and 8 categories, *A, B, C, F, G, H, I* and *J*. Products from the *problematic* categories *D* and *E* were assigned quite randomly to clusters 2 to 8. Evaluation statistics of this model are in Table 2.7. Results confirmed that categories *A, B, C, F, G, H, I* and *J* are perceived similarly by the customers and by the managers. As expected, purity statistic is lower than in the test of Section 2.4 with more clusters and the cost function has higher value. Purity

Tab. 2.5: Evaluation statistics for model with 10 clusters.

Number of classes	10
Purity	0.870
Reverse purity	0.870
Rand index	0.955
Cost function value	0.0153

has to be lower as the size of categories is generally larger than the size of clusters if the cost function is minimized.

Tab. 2.6: Assignment of 100 products from original categories *A*-*J* to clusters 1-8.

Original categories									
A	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*J*
8	5	2	4	3	1	7	3	6	4
8	5	2	6	8	1	7	3	6	4
8	5	2	3	5	1	7	3	6	4
8	5	2	2	3	1	7	3	6	4
8	5	2	5	8	1	7	3	6	4
8	5	2	3	5	1	7	3	6	4
8	5	2	6	8	1	7	3	6	4
8	5	2	4	5	1	7	3	6	4
8	5	2	2	7	1	7	4	5	4
8	5	5	6	7	1	7	3	6	4

Tab. 2.7: Evaluation statistics for model with 8 clusters.

Number of classes	8
Purity	0.770
Reverse purity	0.830
Rand index	0.931
Cost function value	0.027

2.4.4 Model with 13 Clusters

In this model we tried to assign 100 products from 10 categories into 13 clusters – a slightly higher number of clusters than the number of given categories. The resulting assignment is shown in Table 2.8.

The method created cluster 4 which contains products of 3 different categories. Again we

can see that category D (soap), which has the highest violation ratio, tends to be split up. On the other hand, category J (facial pads) which has the lowest violation ratio remains the same. Category G (thrash bags) is split up into two exclusive clusters. That makes sense as this category includes both thick and thin thrash bags. It is apparent that categories C, D and E were splitted into more clusters. Therefore customers buying items from these categories are more likely to buy more different products within the same category. This finding could help in planning promotions where customer gets discount on more expensive product when a product from the same category is bought – those promotions are more effective for clusters which are not splitted.

Evaluation statistics are shown in Table 2.9. Reverse purity statistics is lower than in the previous cases. That is an expected result as we estimated more categories than the number of the original ones.

Tab. 2.8: Assignment of 100 products from original categories A-J to clusters 1–13.

Original categories									
A	B	C	D	E	F	G	H	I	J
11	9	13	1	6	10	2	8	12	3
11	1	5	4	6	10	2	8	12	3
11	9	13	6	4	10	7	8	12	3
11	9	5	13	7	10	2	8	12	3
11	9	5	1	6	10	2	4	12	3
11	9	5	6	4	10	7	8	12	3
11	9	13	4	4	1	7	8	12	3
6	9	5	1	4	10	2	8	4	3
11	9	5	1	4	10	7	4	12	3
6	9	5	4	1	10	7	8	2	3

2.4.5 Model with 20 Clusters

We have shown that the proposed method can determine categories which were originally defined expertly based on the nature of the products if the assumption (A1) is not

Tab. 2.9: Evaluation statistics for model with 13 clusters.

Number of classes	13
Purity	0.840
Reverse purity	0.730
Rand index	0.946
Cost function value	0.0083

significantly violated. In this test we assign products to 20 clusters. The resulting assignment is shown in Table 2.10.

From Table 2.10 it follows that categories *A, B, C, F, G, H, I* and *J* were split into two or three clusters which can be used to define subcategories. Conversely, the categories *D* and *E* contain more clusters. None of these clusters are limited only to a single category. Categories *D* and *E* violate the assumption (A1) more than other categories. On the other hand the method created some interesting and reasonable clusters. For example in category *B* all four WC liquid cleaners were clustered with pine aroma. That leads us to a fact that customers usually do not buy more liquid cleaners with the same aroma. Hence, products with the same function and aroma should be placed next to each other instead of sorted by brand as customers are choosing one within the products with the same aroma.

Some other clusters can not be easily described. Finding not so obvious clusters is the advantage of our method.

Evaluation statistics are shown in Table 2.11. Reverse purity statistics is again significantly lower as we assign to more clusters. The cost function value is also significantly lower than in the model described in Table 2.4 as expected.

2.4.6 Computational Complexity

The behavior of our implementation of the genetic algorithm is regular. Regarding the actual time, it took approximately two hours to finish 1000 generations using common PC (i7 CPU with 4 cores). It seems that it is not needed to include such a large number of generations. As can be seen in Figure 2.7, the final assignment is found approximately by the 400th

Original categories									
A	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*J*
19	4	11	18	18	3	5	16	17	13
10	15	8	18	15	1	14	20	2	13
19	7	6	17	12	3	9	20	2	13
19	7	6	11	15	1	14	16	17	13
10	4	8	12	18	3	5	20	2	13
10	7	8	15	12	3	14	16	17	13
10	4	11	12	5	1	9	16	2	13
10	4	8	12	18	1	14	20	2	4
10	7	8	18	5	3	9	15	11	19
10	4	8	15	12	1	5	20	11	13

Tab. 2.10: Assignment of 100 products from original categories *A*-*J* to clusters 1-20.

generation of our genetic algorithm using our dataset based on real data. The rest of the computation was not needed. As we use the genetic algorithm to minimize the cost function and we do not know the optimal solution beforehand; we cannot prove that the solution is indeed optimal. To reduce computational time, it may be useful to stop the algorithm after a given number of generations without improvement, e. g. stop the computation if the 50 iterations do not improve the cost function and call the current solution final. In our cases this would greatly reduce the computational time without affecting the final solution.

Number of classes	20
Purity	0.820
Reverse purity	0.510
Rand index	0.931
Cost function value	0.0036

Tab. 2.11: Evaluation statistics for model with 20 clusters.

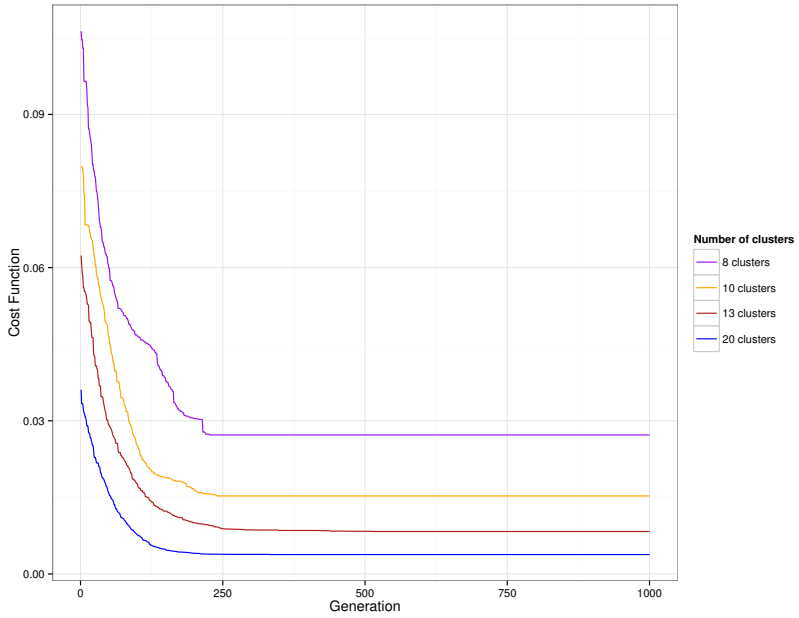

Fig. 2.7: Cost function of best individuals in each generation in models with 8, 10, 13 and 20 clusters.

2.4.7 Summary

The evaluation statistics depend on the number of clusters. Dependency on the number of clusters on real data is similar to the one presented using simulated data in Section 2.3.3 as can be seen in Fig. 2.8.

We have found that the product categorization of retail chain is not perfect. The proposed method was able to find clusters which lead to interesting subcategories. This may be used for example in choosing products which are sold in small stores where space on shelves is limited.

Splitting up categories into more exclusive clusters can help organizing shelves, e. g. not ordering products by the brand but by other characteristic (which may be found by our method) while the products are in the same category. We remind that in this dataset we tried to find what were the reasons that made clusters. That may not be necessary needed in the real business with sufficient amount of data.

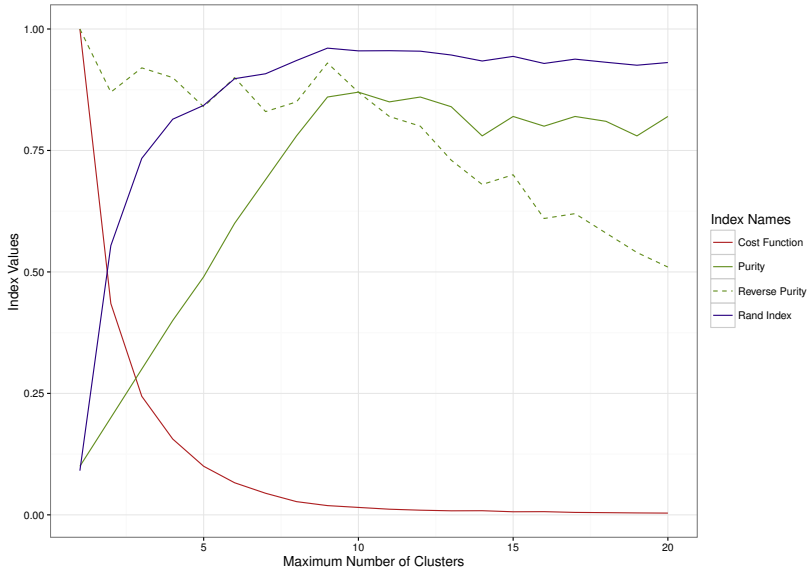

Fig. 2.8: Evaluation statistics for different number of clusters using the proposed method.

In order to maximize utility, the results that are obtained by using our method should be combined with other methods, such as categorizing of products by function, brand or price appeal.

The problem we have encountered is that common drugstore basket contains only a few items while drugstore assortment is usually much larger than that of supermarket. If we have more data the method would give better results. Therefore, we expect our method will work better on supermarket's market basket data with larger amount of different items in the basket and also with thiner range of asortment.

2.4.8 Comparison with other methods

We compared the evaluation statistics of our method with the basic method – in particular k-means and Ward's hierarchical clustering. To use this various methods we had to transform market basket data to *charasteristics* of given products. We count the percentage of occassions in which products are bought together. Therefore, we get square

matrix with dimensions of 100 – number of products in our sample. Hence, during the data transformation there is a loss of information.

We used implementations of these methods in R, particulary *kmeans* and *hclust* functions from package *stats* and *som* function from so called package. For parameters of k-means method we set 1000 starting locations, maximal number of iterations to 1000 and Hartigan-Wong's algorithm. In Ward's hierarchical clustering we set basic Euclidean distance, others parameters remain default. These setups gave the best results.

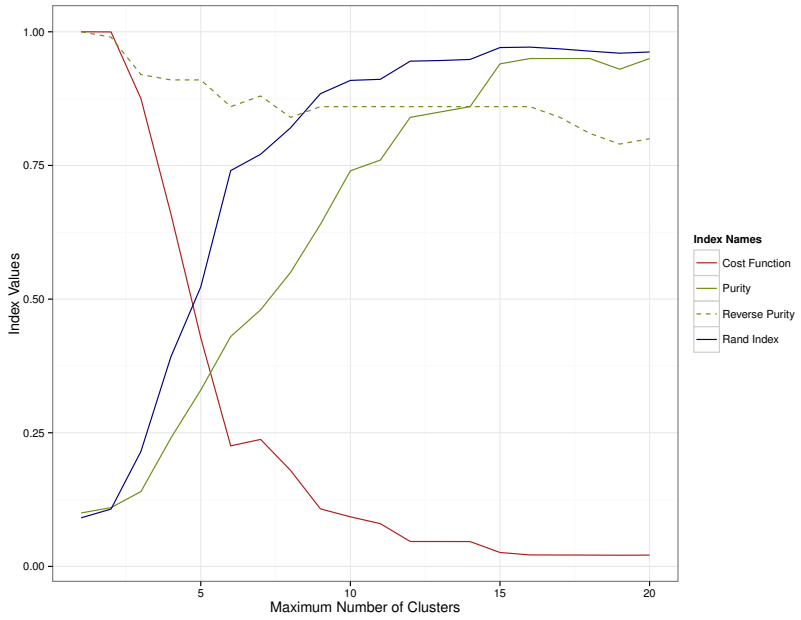

Fig. 2.9: Evaluation statistics for different number of clusters using k-means.

The results were interesting. Ward's hierarchical clustering and k-means method gave almost identical results in every evaluation statistics. According to evaluation statistics such as purity, reverse purity and Rand index, for lower number of cluster our proposed method gave significantly better results; however, with adding more clusters Ward's hierarchical clustering and k-means were more accurate. On the other hand, cost function of Ward's hierarchical clustering and k-means is significantly larger for every number of clusters – as we had to use transformation of data, we could not optimize by our cost function. As a result, the resulting cost function is often more than 10 times larger than in our method.

Figure 2.9 shows resulting statistics for k-means method based on the number of clusters.

It is worth noting that evaluation of statistics purity, reverse purity and Rand index are based on the belief that original categories were set correctly (e. g. they fit our assumptions). Only cost function is purely a data driven statistic and as we show in Section 2.2.1, the objective function we proposed should be more appropriate for our goal as we describe in Section 2.2.1.

Self-organized map did not work well under two-dimensional setup. We suspect that the bottleneck of this approach is the dimension-reduction step which is not appropriate method here. Products may not be easily represented in 2D space when our goal is to cluster products which are not commonly bought together.

2.5 Conclusion

We introduced a new method for the product categorization based solely on the market basket data. The method uses a genetic algorithm for dividing products into a given number of clusters.

We tested the method using synthetic and real data. The method performs well at synthetic data even if the assumptions are violated to some extent. We verified our method using real market basket data from a drugstore retail market. We found that the method accurately identified categories which do not significantly violated the assumptions. When the assumption that customers buy one product from each category at most is violated then the products from that category were spread into several clusters instead of assigned to one cluster. It is worth noting that the original categories were subjectively chosen. Our method identified several *hidden* subcategories using only market basket data that may be widely used in marketing and in general in decision-making processes.

We found out that a common feature of customer's behavior in the Czech drugstore market is that there are not enough receipts with a larger amount of different products, which leads to a violation of the method's assumptions. If we had more data, we suppose that the method would give even more accurate results. Simulations using synthetic data strongly support this hypothesis.

3. How to Segment Customers by Shopping Mission

Ondřej Sokol, Vladimír Holý.
Customer Segmentation Based on a Shopping Mission in the Retail Business.
Submitted to *International Journal of Marketing Research* in November 2019.

Problem statement

Input: Transaction data – set of receipts connected to customer with revenue aggregated by product category.

Output: Clustering of customers based on the usual reason to visit retail store.

Method: Two-phase clustering using k-means. In the first phase, the receipts are clustered by category revenue and total value. In the second phase, customers are clustered using the types of receipts in their purchase history.

Abstract

In the retail business, it is important to understand customer behavior and determine customer value. A useful tool to achieve this goal is the cluster analysis. Typically, a customer segmentation is based on the recency, frequency and monetary value of shopping or the structure of purchased products. We base our segmentation on a shopping mission – a reason why a customer visits the shop. Shopping missions include the focused purchase of a specific product category and the general purchase. We estimate it using market basket data. Applied to a Czech drugstore chain, we show that proposed segmentation brings unique information about customers and should be used together with the traditional methods.

3.1 Introduction

Retail chains have a huge amount of sales data available. An analysis of these data strives to understand the *customer behavior* and determine the *customer value* in order to increase profits. The information about customers can be utilized in various areas such as new product development (Li et al., 2012), product positioning (Gruca and Klemz, 2003), cross-category dependence (Hruschka et al., 1999; Russell and Petersen, 2000; Leeflang et al., 2008), product complements and substitutes determination (Srivastava et al., 1981; Chib et al., 2002), category management (Duchessi et al., 2004), promotions planning (Trappey et al., 2009), online marketing (Chen et al., 2009), targeted advertising (Jonker et al., 2004; Zhang et al., 2007), product recommendation (Liu and Shih, 2005), product association rules (Weng, 2016) and stock optimization (Borin et al., 1994). One of the tools used to achieve these goals is the *cluster analysis*.

There are many applications of the cluster analysis in retail business. Products sold by the shop can be clustered according to their characteristics in order to find substitutes and complements (Srivastava et al., 1981) or target market (Zhang et al., 2007). A product categorization based solely on customer shopping patterns was proposed by Holý et al. (2017). Customers can be segmented according to their demographics and lifestyle or their shopping behavior. A popular approach is to segment customers based on the *recency, frequency and monetary value (RFM)* of their shopping (Kahan, 1998; Miglautsch, 2000; Yang, 2004; Chen et al., 2009; Khajvand and Tarokh, 2011; Putra et al., 2012; Peker et al., 2017; Boon and Ofek, 2016). Another approach is to segment customers based on the *purchased products structure (PPS)* (Russell and Kamakura, 1997; Manchanda et al., 1999; Andrews and Currim, 2002; Tsai and Chiu, 2004). Lingras et al. (2014) and Ammar et al. (2016) simultaneously clustered both products and customers. Overall, the cluster analysis brings useful insight into the customer behavior and helps in the decision-making process, especially when combined with the business knowledge (Seret et al., 2014).

We deal with the *customer segmentation* using data from receipts. The traditional RFM and PPS segmentations answer the questions:

- When was the last time customers visited the shop?
- How often do customers visit the shop?
- How much money do customers spend?

- What product categories do customers buy?

In our analysis, we propose to segment customers based on their *shopping mission (SM)*. The proposed segmentation provides background for the crucial business questions:

- What is the purpose of customer visits?
- Do customers visit the shop because of a specific product category?
- Do customers buy products in other shops?

The proposed approach brings a new insight into the structure of customers. As a result, it can be used in many marketing areas such as the promotion planing and shelf management.

The main idea of the proposed approach is as follows. We utilize the transaction data in a form of receipts which are linked to customers through the loyalty program. We first cluster individual baskets and then use this information to segment customers. This is illustrated in Figure 3.1 along with a comparison to the PPS segmentation. The *k-means* method is utilized for both the basket clustering and the customer segmentation. An analysis of a Czech drugstore chain shows that there are some customers who visit the shop due to an *focused purchase* of a specific product category while others prefer *general purchase*. Two-stage segmentation based on a shopping mission was presented by Reutterer et al. (2006). The main difference from our proposed method is that we aggregate by product categories and also consider the value of the basket instead of using incidence matrix of products. The proposed method is not meant to replace the RFM or PPS segmentation but rather to be used alongside them and to bring a new perspective. The combination of RFM, PPS and SM approaches forms a versatile segmentation based on a broad range of customer characteristics not tied to a single specific purpose. We emphasize the interpretability and usability by marketing departments and other experts involved in the retail decision-making process.

The rest of the chapter is structured as follows. In Section 3.2, we describe the general structure of transaction data in retail business. In Section 3.3, we review the segmentation based on the recency, frequency, and value of a shopping with an application to our dataset. In Section 3.3, we review the segmentation based on the structure of purchased products and again apply it to our dataset. In Section 3.5, we propose a novel segmentation based on the shopping mission of a customer with an application to our dataset. In Section 3.6,

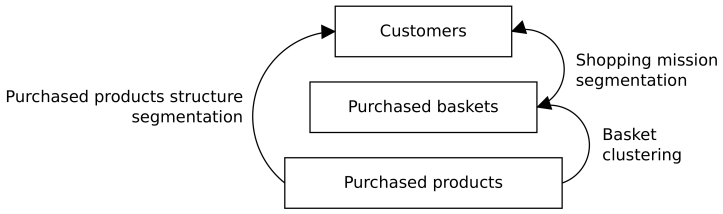

Fig. 3.1: The process of the PPS segmentation and the SM segmentation of customers.

we show how all three segmentations can be combined. Then, we show the implications for practice in Section 3.7. We conclude the chapter in Section 3.8.

3.2 Transaction Data

We perform the analysis of retail business using the *transaction data*. The hierarchical structure of these data is illustrated in Figure 3.1.

A *product* is characterized by the brand, physical properties, and purpose. Note that the price of the product and whether the product is in sales promotion can vary over time and therefore we put it to the receipt data. Because there are many products, it is useful to aggregate them into *product categories*. We denote the product categories as $K = \{K_i : i = 1, \ldots, n_K\}$. An individual product bought by the customer is referred to as the *purchased product*. We denote the purchased products as $P = \{P_i : i = 1, \ldots, n_P\}$. Each purchased product belongs to a single product category.

A *purchased basket* is a set of purchased products. We denote the purchased baskets as $B = \{B_i : i = 1, \ldots, n_B\}$. A specific purchased basket B_i is a subset of all purchased products, i.e. $B_i \subset P$, $i = 1, \ldots, n_B$. A *receipt* is a purchased basket with additional information about the customer ID, prices, sales promotion, date and time.

A *customer history* is a set of purchased baskets. We denote the customer history as $C = \{C_i : i = 1, \ldots, n_C\}$. A specific customer history C_i is a subset of all purchased baskets, i.e. $C_i \subset B$, $i = 1, \ldots, n_C$. A *customer* is a customer history with additional information about

the contact, gender, age, number of children and other demographic information.

In the chapter, we analyze a sample of real data. Our dataset consists of individual purchase data of one of the retail chains in the drugstore market of the Czech Republic. We use a three-month dataset of receipts which include more than 5.6 million baskets bought by more than 1.5 million customers with the loyalty card. Each row in a receipt stands for a single purchased product. The retail chain sells over 10 thousand products which are divided into 55 categories based on their purpose. This categorization was done by an expert.

3.3 Segmentation Based on Recency, Frequency and Monetary Value

One of the most popular way to segment customers is the clustering using data about *recency, frequency and monetary value (RFM)* of their shopping. This method is fast and simple as its original purpose is to provide an easy-to-implement framework for quantifying customer behavior (Kahan, 1998; Miglautsch, 2000).

Yang (2004) described some shortcomings of RFM method (e.g. the inability of RFM to generate the real differences among RFM cells) and introduced a single predictor which is consolidated from the three variables of RFM. A method for the sequential pattern mining using RFM segmentation was presented by Chen et al. (2009). Khajvand and Tarokh (2011) improve RFM segmentation by using the adapted RFM in order to estimate the customer lifetime value. Expansion of RFM segmentation by the fusion with ART2 algorithm to cluster the customers in the retail company was presented by Putra et al. (2012). Another expansion of RFM by Peker et al. (2017) proposes to include customer relation length and periodicity to the customer segmentation.

The first RFM characteristic of a customer is the *recency (Rec)*. Customers are segmented by the time of the last purchase occurrence. With the knowledge of the recency of the last purchase, the retailer can use different marketing techniques to attract customers who had been in the shop in the last week and customers who had not been there for months. In the broader concept, analysis of the occurrences of shopping in time can help for example to identify leaving customers, i.e. those who used to visit the shop frequently but their shopping behavior changed in the recent history. A company can use this knowledge to

retain the customer using various marketing technique. A customer C_i, $i = 1, \ldots, n_C$ is assigned to a recency segment C_j^{Rec}, $j = 1, \ldots, k_{Rec}$ according to the variable

$$IC_i^{Rec} = \min\{\text{Days since a purchase by customer } i\}. \tag{3.1}$$

The segments are either found by some clustering algorithm or defined by an expert.

The second RFM characteristic of a customer is the *frequency (Frq)*. The purchase frequency is defined as the number of visits of a customer during a given time frame. Along with the average value of a basket, it is one of the most tracked key performance indicators. The marketing department can use the information about frequency and distinguish frequent customers from the ones who go to the shop just in the case of emergency. While the goal for the loyal customers is to preserve their shopping behavior, the customers who rarely visit the shop should be recruited. A customer C_i, $i = 1, \ldots, n_C$ is assigned to a frequency segment C_j^{Frq}, $j = 1, \ldots, k_{Frq}$ according to the variable

$$IC_i^{Frq} = \frac{\text{Number of baskets purchased by customer } i}{\text{Time frame}}. \tag{3.2}$$

As in the case of the recency, the segments are either found by some clustering algorithm or defined by an expert.

The third RFM characteristic of a customer is the *monetary value (Mon)*. A customer segmentation by monetary value can be done in various ways. A common approach is to compute either the sums of all sales during a given time frame or the average value of baskets in a given time frame for each customer. The latter is used in a combination with the frequency analysis. Retailers can also focus on margins instead of sales. In our case, we assign a customer C_i, $i = 1, \ldots, n_C$ to a monetary value segment C_j^{Mon}, $j = 1, \ldots, k_{Mon}$ according to the variable

$$IC_i^{Mon} = \frac{\text{Total value of products purchased by customer } i}{\text{Time frame}}. \tag{3.3}$$

Again, segments are either found by some clustering algorithm or defined by an expert.

The combination of the above mentioned approaches forms the RFM segmentation. One approach to derive RFM segmentation is to create the 3-dimensional matrix of all combinations of the $C^{Rec} = \{C_i^{Rec} : i = 1, \ldots, k_{Rec}\}$, $C^{Frq} = \{C_i^{Frq} : i = 1, \ldots, k_{Frq}\}$ and $C^{Mon} = \{C_i^{Mon} : i = 1, \ldots, k_{Mon}\}$ segmentations with size $k_{Rec} \times k_{Frq} \times k_{Mon}$. However, the number of clusters $k_{RFM} = k_{Rec}k_{Frq}k_{Mon}$ can be quite large. To reduce the number of segments, another approach may be used. For a given k_{RFM}, a customer C_i, $i = 1, \ldots, n_C$

can be assigned to a RFM segment C_j^{RFM}, $j = 1, \ldots, k_{RFM}$ according to the vector variable

$$IC_i^{RFM} = \left[IC_i^{Rec}, IC_i^{Frq}, IC_i^{Mon}\right]. \tag{3.4}$$

The segments can be found by some clustering algorithms such as the k-means method. Despite its shortcomings, the RFM segmentation is commonly used across retail business for its simplicity and straightforward interpretation.

3.4 Segmentation Based on Purchased Products Structure

Products in retail shops are often categorized based on their properties such as a purpose, price, pack size and brand. The categorization of products can be done either by an expert or by an algorithm (Srivastava et al., 1981; Zhang et al., 2007; Holý et al., 2017). Subsequently, customers can be segmented using their receipts. We refer to this clustering as *purchased product structure (PPS)* segmentation. The knowledge of their purchases is directly linked to product categories. Such analysis reveals commonly bought categories and therefore helps in targeting of marketing campaigns.

Segmentation of customers based on their category purchases was studied in Russell and Kamakura (1997), where authors segmented customers with respect to brand preference using household purchase data. Another approach of using product categorization on household data to analyze customers behavior was published by Manchanda et al. (1999). A method for identifying customer segments with identical choice behaviors across product categories using logit model was presented by Andrews and Currim (2002). Tsai and Chiu (2004) dealt with clustering customers based on their purchase data linked to product categories and presented a methodology to ensure the quality of the resulting clustering. Lingras et al. (2014) and Ammar et al. (2016) utilized an iterative meta-clustering technique that uses clustering results from one set of objects to dynamically change the representation of another set of objects. The method is applied on product categorization and customer segmentation using supermarket basket data.

We segment customers based on ratios of their purchases in each category. For a customer C_i, $i = 1, \ldots, n_C$, the PPS segmentation is based on information about a product category

j, $j = 1, \ldots, n_K$ given by

$$IC_i^{Cat,j} = \frac{\text{Total value of products in } j \text{ purchased by } i}{\text{Total value of products purchased by } i}. \tag{3.5}$$

A customer C_i, $i = 1, \ldots, n_C$ is then assigned to a PPS segment C_j^{Frq}, $j = 1, \ldots, k_{PPS}$ according to the vector variable

$$IC_i^{PPS} = \left[IC_i^{Cat,1}, \ldots, IC_i^{Cat,n_K} \right]. \tag{3.6}$$

We find PPS segments using the k-means method. The optimal number of clusters k_{PPS} is chosen according to the ratio of between cluster variance and total variance and Davies-Bouldin index alongside with a reasonable interpretation of resulting clusters.

We perform a customer segmentation in a Czech drugstore chain according to 55 product categories. We find that in our case the optimal number of clusters is 12. Figure 3.2 shows that there are 11 specialized segments and 1 general segment. The centers of specialized segments are composed of about 60% spendings in a single category. On the other hand, the general segment is fairly uniformly composed of over 15 popular categories. The distribution of customers assigned to the individual segments is alongside with labels of the dominant product categories shown in Table 3.1. It is worth noting that despite having 55 categories, top 15 categories comprise of over 98% of the total revenue.

We use the PPS segmentation as a base customer clustering according to categories they purchase. The next step is to compare it with a more complex approach featuring an intermediate step and adding the basket information to the segmentation. Figure 3.1 shows the process of both segmentations.

3.5 Segmentation Based on Shopping Mission

We propose an addition to the RFM and PPS segmentations. The above mentioned segmentations lack the information about the reason why customers visit the shop. Some customers visit the shop just to buy one product they need. This is called the *focused purchase*. Other customers purchase large amount of various products. This is called the *general purchase*. Our goal is to design a reasonable segmentation reflecting the reasons why customers come to the shop, i.e. their *shopping mission*. We use the term shopping mission in the sense of what products customers actually buy. We can only guess the true

Tab. 3.1: Distribution of the PPS customer segmentation.

Cluster	Description	Share
P01	General	32.0%
P02	Specialized – Detergents	6.1%
P03	Specialized – Laundry detergents	5.6%
P04	Specialized – Body products	6.1%
P05	Specialized – Face products	6.2%
P06	Specialized – Dental products	5.7%
P07	Specialized – Hair products	7.5%
P08	Specialized – Beauty products	9.6%
P09	Specialized – Products for men	5.3%
P10	Specialized – Products for children	5.0%
P11	Specialized – Parfumes	7.3%
P12	Specialized – Seasonal products	3.7%

motivation behind the visit (i.e. a customer is thirsty and therefore buys a bottle of water). From the marketing point of view, customers who come to the shop just for certain categories probably buy everything else in some other shop, therefore the goal is to transform them into regular customers. On the other hand, customers who already fulfill a majority of their needs in the shop are the most valuable and the goal of the marketing department is to retain them.

In the literature, the shopping mission or shopping motivation is often approached from a qualitative point of view. Hedonic shopping motivation and its effect in utilitarian enviroments was studied by Yim et al. (2014) using a field survey. Studies based on transaction data are present in the literature as well. Schröder (2017) analyzed multi-category purchase decisions on the weekly basis using the item response theory models which allows to reveal characteristics of households for purchase decisions. Underlying latent activities of shoppers are also focus of the study by Hruschka (2014) using topic models. Analysis of baskets using self-organizing maps was presented by Decker and Monien (2003). Reutterer et al. (2006) introduced a two-stage method of clustering customers using the basket clustering. In the first phase, baskets are clustered

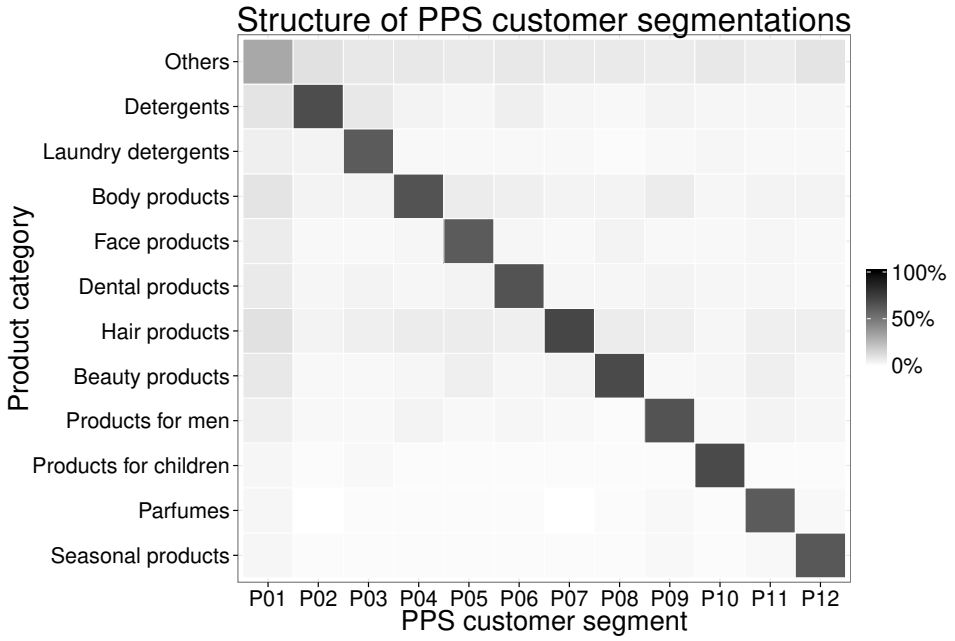

Fig. 3.2: Ratios of product value from the PPS customer segments split into the product categories.

based on the purchased products, this is done using information whether the product appeared in the basket or not. In the second phase the customers are segmented based on their baskets. A method for identifying shopping mission using basket value and variety was proposed in Sarantopoulos et al. (2016).

For the clustering of baskets, we utilize purpose categories as well as the basket value resulting in easily interpretable clusters. In order to get reasonable clustering, the value of basket is standardized using the 95% quantile of a basket value while the baskets with the value over this quantile are set to 1 due to a skewed distribution of the basket value. See Figure 3.3 for the kernel density function of the standardized basket value. In the clustering, each basket is then represented by a vector of non-negative ratios with unit sum and a standardized basket value ranging from 0 to 1. The interpretation is that we give similar weights to both the structure and the value of the basket. For aasket B_i, $i = 1, \ldots, n_B$, the PPS clustering is based on information about product a category j,

$j = 1, \ldots, n_K$ given by

$$IB_i^{Cat,j} = \frac{\text{Total value of products in category } j \text{ in basket } i}{\text{Total sales of products in basket } i} \tag{3.7}$$

and the information about value given by

$$IB_i^{Val} = \min\left\{\frac{\text{Total value of products in basket } i}{q_{95}}, 1\right\}, \tag{3.8}$$

where q_{95} is the 95% quantile of all basket values. A basket B_i, $i = 1, \ldots, n_B$ is then assigned to a SM cluster B_j^{PPS}, $j = 1, \ldots, k_B$ according to the vector variable

$$IB_i^{SM} = \left[IB_i^{Cat,1}, \ldots, IB_i^{Cat,n_K}, IB_i^{Val}\right]. \tag{3.9}$$

We find the SM basket segments using the k-means method and choose the optimal number of clusters k_B according to the ratio of between cluster variance and total variance and Davies-Bouldin index.

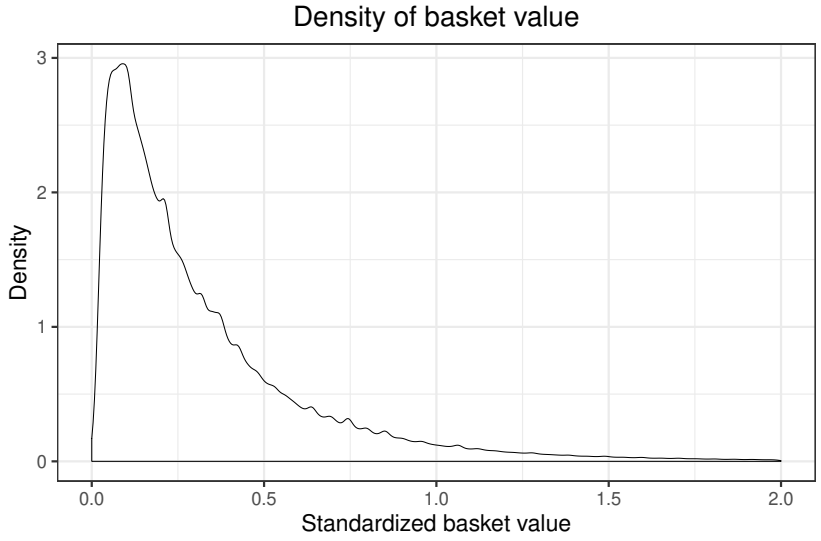

Fig. 3.3: Estimated kernel density function of standardized basket value.

Our basket segmentation has the following geometric interpretation. Let us denote

$$IB_i^{Cat} = \left[IB_i^{Cat,1}, \ldots, IB_i^{Cat,n_K}\right]. \tag{3.10}$$

For a given basket i, IB_i^{Cat} then represents a point in a simplex of dimension n_K while IB_i^{Val} adds a depth to this simplex. The clustering is simply a dissectioning of this space. For

simplicity, we focus on a low dimension of three product categories. The ratio of spending in a product category is then represented by a point in a triangle, whose vertices represent the exclusivity of a category in the basket. The value of basket adds a depth to the triangle and forms a prism. Each basket is a point in this space. The centers of resulting clusters are inside the prism as well. In Figure 3.4, we show two cuts of the prism with 4 cluster centers and defined cluster area for low and high basket value. The center C1 represents a point with a higher value than the others. Therefore its cluster area in cuts by basket value expands with a higher value of the basket. The baskets are assigned to the nearest center using the standard Euclidean distance. Let us consider the following example. First, we denote S1 the basket with hair products worth of 5$, body products worth of 3$ and no face products. Second, we denote S2 the basket with hair products worth of 25$, body products worth of 15$ and no face products. Both baskets have the same ratio of the three product categories and therefore have the same position in the simplex. Their value, however, differs making their depth different as well. As a result, they are assigned to different clusters. Basket S1 is closest to center C4 while basket S2 is closest to center C1.

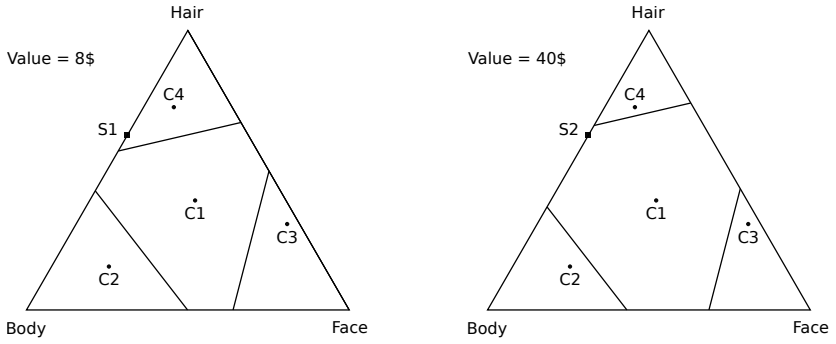

Fig. 3.4: Illustration of the SM segmentation for different basket value levels.

We cluster baskets in a Czech drugstore chain according to 55 product categories. As expected the basket clusters are formed around previously mentioned well-selling categories. We find that 12 is the optimal number of clusters. Two clusters represent small and big universal baskets with no dominant category while the 10 others are focused on a single dominant category. The between cluster variance ratio is 0.8 with 12 clusters while Davies-Bouldin index for 12 clusters has similar value to the other possible choices. The resulting

distribution of baskets to the cluster as well as the interpretation of each cluster is shown in Table 3.2. Each cluster is named after the dominant category similarly to the PPS segmentation. The structure of each basket cluster (archetypes) is shown in Figure 3.5. It is evident that general baskets tend to have a higher value than the others. A larger number of basket clusters leads to the creation of sparsely populated basket clusters with small product categories. Furthermore, these clusters vary depending on the season, which could cause problems in subsequent interpretation.

Tab. 3.2: Distribution of the SM basket clustering.

Cluster	Description	Share
B01	General – Big	13.0%
B02	General – Small	13.9%
B03	Specialized – Detergents	9.8%
B04	Specialized – Laundry detergents	7.9%
B05	Specialized – Body products	6.5%
B06	Specialized – Face products	6.0%
B07	Specialized – Dental products	8.8%
B08	Specialized – Hair products	11.3%
B09	Specialized – Beauty products	9.7%
B10	Specialized – Products for men	2.2%
B11	Specialized – Products for children	6.4%
B12	Specialized – Feminine hygiene	4.5%

In the second step, we determine the customer segments based on the ratio of baskets archetypes they bought. We do not use the absolute number of the baskets because of purely practical reasons. Our goal is to estimate the shopping mission of ordinary customers. However, some people visit the retail chain to supply their own small business. They buy an enormous number of products with a huge frequency. Clustering algorithms, in that case, are likely to create numerous clusters just for a very small number of customers. This is a logical and right way. However, this information about the value and frequency is already included in the RFM segmentation. Therefore we normalize the number of baskets by using the ratio of baskets archetypes bought by a customer. Customers with unusual shopping

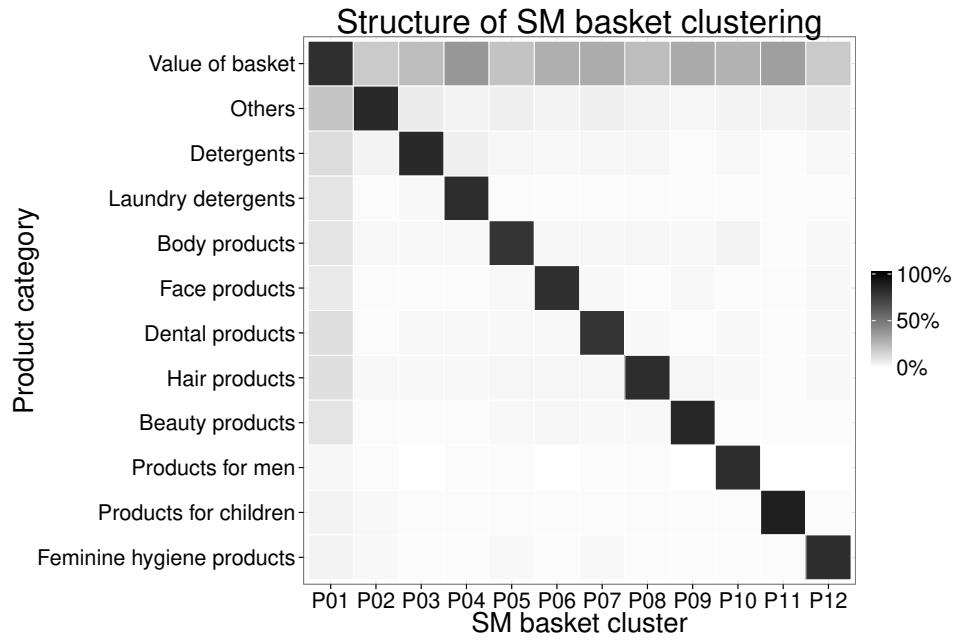

Fig. 3.5: Ratios of product value from the SM basket clusters split into the product categories.

behavior are easily detectable using the RFM and SM segmentations as a whole, so we do not need to exclude them at all. For customer C_i, $i = 1, \ldots, n_C$, the SM segmentation is based on the information about basket cluster j, $j = 1, \ldots, k_B$ given by

$$IC_i^{Bas,j} = \frac{\text{Number of baskets in } j \text{ purchased by } i}{\text{Total number of baskets purchased by } i}. \tag{3.11}$$

Customer C_i, $i = 1, \ldots, n_C$ is then assigned to SM segment C_j^{SM}, $j = 1, \ldots, k_{SM}$ according to the vector variable

$$IC_i^{SM} = \left[IC_i^{Bas,1}, \ldots, IC_i^{Bas,k_B}\right]. \tag{3.12}$$

For the second phase we also use the k-means algorithm and select the optimal number of clusters according to the Davies-Bouldin index and ratio of between cluster variance.

We continue with our empirical analysis and segment customers of a Czech drugstore chain. We find the optimal number of clusters to be 18 using the between cluster variance ratio statistics. For the description of each segment, we use its center ratios of each basket

type in the customer history. This allows us to distinguish three main types of customers. The *general* customers buy variety of categories in their baskets. As a customer with bulk purchases has a significantly different shopping motivation compared to a customer with very small yet varied purchases, the general group is further divided into more segments based on the prevailing purchase size. The *focused* customers focus only on one type of category in each of their purchases and visit the store with a very straightforward motivation. They are looking for specific products and may be buying the rest of the drugstore products elsewhere. The proposed segmentation further divide focused customers into segments based on the category they prefer in majority of their purchases. The *mixed* customers are a combination of the above customer types. Overall, the clustering consists of 5 general segments with different basket values, 12 focused segments formed around a single basket type and 3 mixed segments of both general and focused baskets. The interpretation of the segments along with the percentage of assigned customers is shown in Table 3.3. The structure of the baskets archetypes in clusters is shown in Figure 3.6.

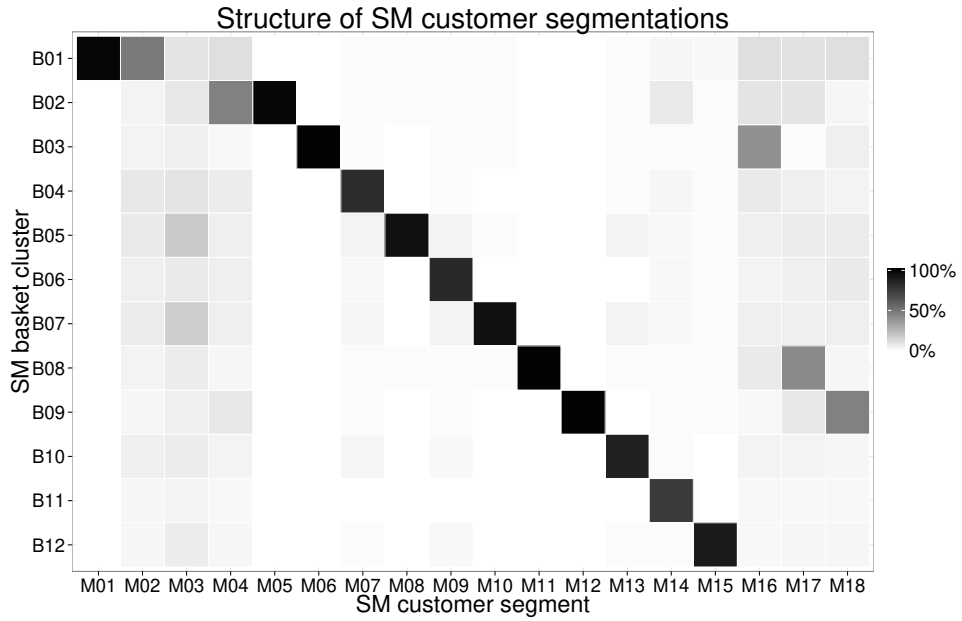

Fig. 3.6: Ratios of baskets from the SM customer segments split into the SM basket clusters.

The division into the three main customer types and the subdivision into the specific

Tab. 3.3: Distribution of the SM customer segmentation.

Cluster	Description	Share
M01	General – Exclusively small	6.0%
M02	General – Mainly small	8.8%
M03	General – Small and big	8.4%
M04	General – Mainly big	11.3%
M05	General – Exclusively big	6.8%
M06	Focused – Detergents	2.9%
M07	Focused – Laundry detergents	4.3%
M08	Focused – Body products	3.3%
M09	Focused – Face products	3.6%
M10	Focused – Dental products	3.1%
M11	Focused – Hair products	4.2%
M12	Focused – Beauty products	4.2%
M13	Focused – Products for men	3.0%
M14	Focused – Products for children	4.3%
M15	Focused – Feminine hygiene	1.4%
M16	Mixed – Detergents	8.2%
M17	Mixed – Hair products	9.6%
M18	Mixed – Beauty products	6.4%

segments is important in choosing suitable marketing strategies and is not contained in the common RFM and PPS segmentation techniques. Using the knowledge of experts in the field and the analysis of customer characteristics, each segment can be further described. For example, the segment of customers focused exclusively on the big baskets is distinguished by high proportion of promotion sales. Not surprisingly, the segment of customers focused on products for children are usually parents in their 20s and 30s. Such type of information is crucial for practical applications including marketing targeting and optimization of promotion sales. The proposed approach therefore offers a novel insight into the shopping behavior of customers.

Overall, our approach is similar to Reutterer et al. (2006) but there are some distinctive differences. Their first phase consists of assigning shopping baskets to the basket prototypes. This is done using clustering techniques applied to incidence matrix of purchased products leading to a high number of basket prototypes. We also need to assign shopping baskets to the clusters but instead of incidence matrix we use the percentage of total spending by product category along with the standardized total value of the basket. This allows us to consider lower number of clusters. The main difference between the methods, however, lies in the second phase. Reutterer et al. (2006) simply count occurrences of each customer's basket prototype during a given period. Customers are then assigned to behaviorally persisting segment if the number of purchased basket of some type exceed a user-defined threshold value. Instead, we use k-means for the ratios of purchased basket types for each customer. The approach of Reutterer et al. (2006) is intended to be used in longer time-frames and utilizes series of historical data. Our method does not utilize the historical data which may be an advantage.

3.6 Comparison of Segmentations

First, we compare the RFM, PPS and SM segmentations. Our goal is to find if the segmentations are similar or if each segmentation brings unique information to the customer analysis. We adopt the *purity* measure for comparison. Let us assume we have n objects clustered by methods I and II with k_I and k_{II} clusters. The purity is then defined as

$$\text{Purity} = \frac{1}{n}\sum_{i=1}^{k_I}\max_j |C_i^I \cap C_j^{II}|, \tag{3.13}$$

where C_i^I is the set of objects in the cluster i of the method I and C_j^{II} is the set of objects in the cluster j of the method II. Similar clusterings have the purity close to 1 while different clusterings have the purity close to 0. Note that the purity is not symmetrical. Table 3.4 reports the purity for segmentations based on recency (Rec), frequency (Frq), monetary value (Mon), purchased product structure (PPS) and shopping mission (SM). We can see that each segmentation is unique as there are no segmentations with a high similarity. However, a medium similarity does exist. For example, the F and PPS segmentations are related to the SM approach due to Frq/SM purity 0.523 and PPS/SM purity 0.428. Reverse relationships have much lower purities because the SM segmentation has significantly more clusters than other segmentations.

Tab. 3.4: Purity between customer segmentations in rows and columns (rounded).

	Rec	Frq	Mon	PPS	SM
Rec	1.0	0.3	0.2	0.2	0.3
Frq	0.5	1.0	0.4	0.3	0.5
Mon	0.3	0.4	1.0	0.3	0.3
PPS	0.4	0.4	0.4	1.0	0.4
SM	0.2	0.2	0.1	0.1	1.0

Next, we investigate the relationship between the SM and Frq segmentations in more detail. We compare 6 Frq segments and SM segments described in Table 3.3. Figure 3.7 shows how customer segments from the SM approach are divided into the Frq segments. The general segments M02–M04 and mixed segments M16–M18 have relatively high frequency while general segments M01 and M05 and the specialized focused segments M05–M15 have quite low frequency. This is an expected result as loyal customers with a general shopping visit the shop more often than customers that mainly shop elsewhere and visit the shop only for emergencies.

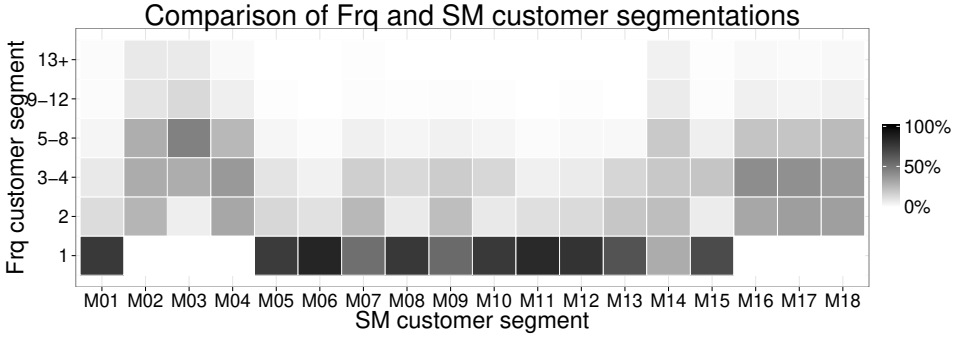

Fig. 3.7: Ratios of customers from the SM segments split into the Frq segments.

Finally, we investigate the relationship between the SM and PPS segmentations in more detail. We compare PPS segments described in Table 3.1 and SM segments described in Table 3.3. Figure 3.8 shows how customer segments from the SM approach are divided into the PPS segments. We can see that specialized segments P02–P10 correspond to

segments M05–M14. General segment P01 is split among M01–M05 clusters according to the value of typical baskets and to clusters M16–M18 with dominant products. Interestingly, feminine hygiene products form their own segment M15. This is because many customers visit the drugstore specifically for these products but also purchase them in general baskets. Specialized segments P11 and P12 are clustered into general segments M01–M05. This is because customers do not visit shop specifically for these products but rather buy them together with other products.

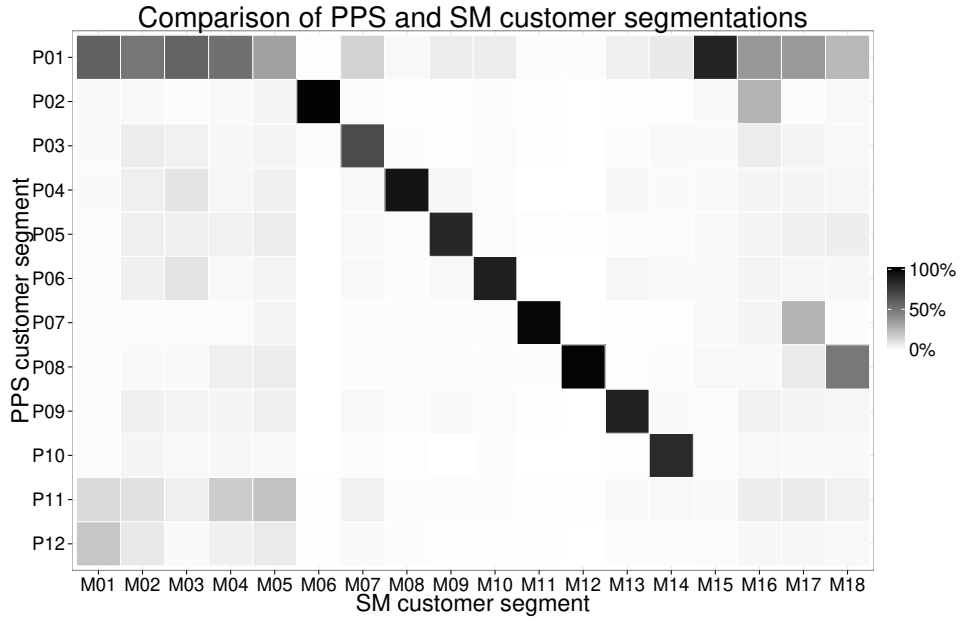

Fig. 3.8: Ratios of customers from the SM segments split into the PPS segments.

All considered segmentations should be used together in the analysis of customers as each segmentation has a unique structure and interpretation and brings different information about customers.

3.7 Implications for Practice

Finally, we discuss the application of the proposed approach to a Czech drugstore. The main use of the proposed segmentation lies in better planning of promotional activities and

targeted communication. Based on the proposed customer segmentation, the communicated message can be modified into a few variants according to the customer's interest. A major advantage of this approach is its simple interpretation as managers can easily select groups of customers based on their shopping mission. The typical situation in the e-mail trageting is the following. Every two weeks dozens of products are discounted and promoted. In e-mail communication, however, only a subset of them is announced. In order to maximize the effect of such promotions, it is necessary to select the products that will attract the customer most. In our experience, only a few variations of communicated message to customer have good results and are the most cost-effective approach.

In order to examine the effect of the proposed segmentation, we conducted the following experiment. We divided customers from each segment randomly into the test and control groups. The standard e-mail was sent to the customers in the control group while targeted e-mails designed according to the customer segmentation were sent to the customers in the test group. We focused mainly on three large groups of customers – customers focused on products for children (segment *M14*), customers focused on detergents (segments *M06, M07* and *M16*) and customers focused on hair and beauty products (segments *M11, M12, M17* and *M18*). After the implementation of the SM segmentation, we observed an increase in the opening of e-mails for all test groups in comparison to control groups. The increase was 11 percent for the childrens group, 3 percent for the detergents group and 2 percent for the hair and beauty group. We also observed an increase in average spending in the promoted products by customers who opened the e-mail. The difference was 6 percent for the childrens group, 2 percent for the detergents group and 5 percent for the hair and beauty group.

Furthermore, we compared the newly introduced SM segmentation in the current period with the PPS segmantion used by the managers in the previous period. In order to avoid bias due to irrelevant changes between the two periods, we standardized the open rate and spending indicators according to the average performance of the total e-mail wave. We observed an increase in the open rate in all three groups. The increase was less than one percent percent for the childrens group, 4 percent for the detergents group and 4 percent for the hair and beauty group. Negligible increase in the test group of customers focused on children products was caused by the high degree of similarity between the PPS and SM segments focused on children. Next, we focused on the average spending on the promoted products. Similar to the open rate, the children products group remained at the same level. In contrast, the spending increased by 3 percent in detergents group and by 6 percent in the

hair and beauty group. Overall, the proposed SM segmentation brought the highest effect for premium groups focused on hair and beauty products. The effect on customers focused on detergents, which is generally less premium category, was smaller. The difference between the customer group focused on children products was minimal compared to the previous period due to the fact that the PPS and SM segments are very similar for customers with children.

3.8 Conclusion

We dealt with a segmentation of customers in retail business according to their shopping behavior. The chapter has two main contributions.

- First, we propose a new segmentation approach based on a shopping mission of a customer. The shopping mission answers the question why the customer visits the shop. Possible shopping missions include the purchase of a specific product category and the general purchase.
- Second, we show how various segmentations can be combined in a real application. Besides the proposed method, we also consider recency, frequency and monetary value approach as well as the approach based on the structure of purchased products. The results show that the proposed segmentation brings useful insight into the analysis of customer behavior.

The proposed segmentation has been introduced in a major Czech drugstore chain and is currently used mainly in the e-mail targeting. The customer reaction indicators in targeted emailing campaigns such as the open rate and click rate have significantly improved in comparison to the previously used PPS segmentation.

4. How Many Customers Does a Retail Chain Have

Ondřej Sokol, Vladimír Holý.
How Many Customers Does a Retail Chain Have? [1]
Submitted to *Marketing Science* in June 2019.

Problem statement

Input: Transaction data – set of receipts assigned to the basket type, some of them linked to loyalty program members, others unlinked.

Output: Estimated number of non-member customers by customer segment.

Method: Nonlinear constrained optimization program using the shopping behavior of loyalty program members by customer segment.

[1] The preliminary results were presented in Sokol (2018a) and Sokol and Holý (2019)

Abstract

The knowledge of the number of customers is the pillar of retail business analytics. In our setting, we assume that a portion of customers is monitored and easily counted due to the loyalty program while the rest is not monitored. The behavior of customers in both groups may significantly differ making the estimation of the number of unmonitored customers a non-trivial task. We identify shopping patterns of several customer segments. This allows us to estimate the distribution of customers without the loyalty card. For this purpose, we utilize the least squares and maximum likelihood methods. In a simulation study, we find that the maximum likelihood estimator is slightly more robust method. In an empirical study of a drugstore chain, we illustrate the applicability of the proposed approach in practice. The actual number of customers estimated by the proposed method is significantly higher than the number suggested by the naive estimate assuming the constant customer segment distribution. The proposed method can also be utilized to determine penetration of the loyalty program in the individual customer segments.

4.1 Introduction

Retail transaction data contain information about the composition of the shopping basket, the price of the purchased goods and possibly the ID of customers. Analysis of such data is an important aspect of retail business analytics as it provides a valuable insight into the customer behavior. For an overview of marketing and retail business analytics and its impact, see Lilien et al. (2013), Germann et al. (2014), Roberts et al. (2014), Bradlow et al. (2017) and France and Ghose (2019).

The most basic question in retail analytics is *how many customers does a retail chain actually have*. In some cases, customers are monitored through the loyalty program. In other cases, customers can visit the store repeatedly without any possibility of identifying them. In our setting, we assume that only a portion of customers is monitored. Clearly, the number of customers with the loyalty card is known while the number of customers without the loyalty card is unknown. With absolute certainty, we can say that there is at least one customer without the loyalty card – a single customer could theoretically purchase all unmonitored shopping baskets. We also know that there are at most as many customers without the loyalty card as unmonitored receipts – each customer could visit the store only once. To be able to give a more specific estimate, we need to adopt further assumptions about customer behavior. One may assume that the number of customers is proportional to the number of sales, i.e. customers with the loyalty card shop as often as customers without the loyalty card. In that case, the number of customers can be estimated simply by the number of sales divided by the average frequency of customer visits. However, this assumption is too strict and unrealistic as customers without the loyalty card are likely to visit the store less frequently. The question is how less frequently. Our work deals with estimation of the number of customers without the loyalty card under less restrictive and more realistic assumptions.

The knowledge of the number and structure of customers without the loyalty card finds its use in marketing departments involved in the retail decision-making process. It can be used to improve mass marketing communications, especially customer targeting and promotional sales planning. The number of customers in certain segment can also be used as an input in prediction of demand for products during the promotion sales as each segment has different shopping behavior. The proper prediction then allows to minimize sold-out situations. All of the above mentioned applications play an important role in increasing sales and profits

for the company.

In the literature, counting of customers is most often approached in the context of *active customers* that are monitored. Schmittlein et al. (1987) proposed the *Pareto/NBD Model* to determine the probability that a customer with a given pattern of transactions is still active. This model was further studied and extended by Fader et al. (2005), Batislam et al. (2007), Jerath et al. (2011), Abe (2009), Glady et al. (2009), Ma and Büschken (2011) and Mzoughia et al. (2018). Schumacher (2006) dealt with *faux-new customers* who seem to be new customers due to the absence of past transaction history but are actually regular customers. In contrast, we count customers that are never monitored.

Transaction data are often used for clustering of customers, products and baskets. *Customer segements* were determined by Tsai and Chiu (2004), Konuş et al. (2008), Putra et al. (2012), Lingras et al. (2014), Ammar et al. (2016), Peker et al. (2017) and Sokol and Holý (2018), *product categories* by Zhang et al. (2007), Lingras et al. (2014), Ammar et al. (2016) and Holý et al. (2017) and *basket types* by Decker and Monien (2003), Reutterer et al. (2006), Sarantopoulos et al. (2016), Griva et al. (2018) and Sokol and Holý (2018). We utilize clustering of customers as well as baskets.

Our procedure consists of the following steps. First, we determine basket types based on the value, price level and diversity of the products in the basket. Second, we determine customer segments based on the bought baskets using history of transactions linked to loyalty cards. For each customer segment, we have the average ratios of bought basket types and the average shopping frequency. Third, we estimate the distribution of customer segments using the observed distribution of basket types not linked to loyalty cards. For this purpose, we utilize the *least squares (LS)* estimator and the *maximum likelihood (ML)* estimator. Finally, we estimate the number of customers from the distribution of customer segments and average shopping frequencies.

In a simulation study, we compare the LS and ML estimators with the naive estimator that assumes the distribution of customer segments is the same for customers with and without the loyalty card. In all scenarios differing in violation of different assumptions or the size of monitored or unmonitored sample, the ML estimator has slightly lower error and also slightly lower standard deviation of the error than the LS estimator. The differences between ML and LS method are, however, minimal. The naive approach gives much worse estimates than the both methods if the distribution of customer segments differ between monitored or

unmonitored sample.

In an empirical study, we study behavior of customers in a drugstore chain. We find that the actual total number of customers without the loyalty card is significantly times higher than the number suggested by the naive estimate assuming the same distribution of customer segments for members and non-members of the loyalty program. As expected, we find that the distribution of customers without the loyalty card includes much more casual customers who visit the store rarely and purchase smaller baskets.

The rest of the chapter is structured as follows. In Section 4.2, we propose a procedure for determining the number and distribution of unmonitored customers. In Section 4.3, we examine the behavior of the proposed method using simulations. In Section 4.4, we illustrate the applicability of the proposed method in practice. We conclude the chapter in Section 4.5.

4.2 Methodology

4.2.1 Stochastic Framework

We introduce our probabilistic framework and notation. Let a denote the number of transactions. Further, let there be n basket types and m customer segments. Each transaction has a single basket type and a customer segment. Let $B_k \in \{1, \ldots, n\}$ be a random variable denoting the basket type of transaction k and $C_k \in \{1, \ldots, m\}$ a random variable denoting its customer segment. The random number of transactions with basket type i is then $X_i = \sum_{k=1}^{a} \mathbb{I}\{B_k = i\}$, $i = 1, \ldots, n$ while the random number of transactions with customer segment j is $Y_j = \sum_{k=1}^{a} \mathbb{I}\{C_k = j\}$, $j = 1, \ldots, m$, where $\mathbb{I}(\cdot)$ denotes the indicator function. The random number of transactions with basket type i and customer segment j is $Z_{i,j} = \sum_{k=1}^{a} \mathbb{I}\{B_k = i \wedge C_k = j\}$, $i = 1, \ldots, n$, $j = 1, \ldots, m$. Clearly, we have

$$\sum_{i=1}^{n} X_i = \sum_{j=1}^{m} Y_j = \sum_{i=1}^{n} \sum_{j=1}^{m} Z_{i,j} = a. \tag{4.1}$$

Using a vector and matrix notation, we have $B = (B_1, \ldots, B_a)'$, $C = (C_1, \ldots, C_a)'$, $X = (X_1, \ldots, X_n)'$, $Y = (Y_1, \ldots, Y_m)'$ and $Z = (Z_{i,j})_{i=1,j=1}^{n,m}$. We denote $b = (b_1, \ldots, b_a)'$ the observed basket types and $c = (c_1, \ldots, c_a)'$ the observed customer segments. We further denote $x = (x_1, \ldots, x_n)'$ the observed numbers of transactions with given basket type,

$y = (y_1, \ldots, y_n)'$ the observed numbers of transactions with given customer segment and $z = (z_{i,j})_{i=1,j=1}^{n,m}$ the observed numbers of transactions with given basket type and customer segment. We assume that random variables $\{B_k : k = 1, \ldots, a\}$ are independent and identically distributed with the probability of basket type i denoted as $p_i = \mathrm{P}[B_k = i]$, $i = 1, \ldots, n$. Similarly, we assume that $\{C_k : k = 1, \ldots, a\}$ are independent and identically distributed with the probability of customer segment j denoted as $q_j = \mathrm{P}[C_k = j]$, $j = 1, \ldots, m$. Finally, the probability of basket type i conditional on customer segment j is denoted as $r_{i,j} = \mathrm{P}[B_k = i|C_k = j]$, $i = 1, \ldots, n$, $j = 1, \ldots, m$. Using a vector and matrix notation, we have $p = (p_1, \ldots, p_n)'$, $q = (q_1, \ldots, q_m)'$ and $r = (r_{i,j})_{i=1,j=1}^{n,m}$. Using the law of total probability, we have

$$p_i = \mathrm{P}\left[B_k = i\right] = \sum_{j=1}^{m} \mathrm{P}\left[B_k = i|C_k = j\right] \mathrm{P}\left[C_k = j\right] = \sum_{j=1}^{m} r_{i,j} q_j, \qquad i = 1, \ldots, n. \tag{4.2}$$

In matrix notation, we simply have $p = \mathrm{R}q$. We denote $r_{i,j}$ as elements of matrix R. This is the key equation used in the proposed approach.

First, let us assume that we observe both basket types b and customer segments c. In this case, we can simply estimate the probability of basket type i as $\hat{p}_i = x_i/a$, $i = 1, \ldots, n$, the probability of customer segment j as $\hat{q}_j = y_j/a$, $j = 1, \ldots, m$ and the probability of basket type i conditional on customer segment j as $\hat{r}_{i,j} = z_{i,j}/y_j$, $i = 1, \ldots, n$, $j = 1, \ldots, m$.

Next, let us assume that we observe only basket types b. In this case, we can estimate p in the same way but cannot estimate q nor R. However, if we additionally assume that the matrix of conditional probabilities R is known, we are able to estimate the customer segment distribution q. In the rest of this section, we propose several estimators of q under this setting.

4.2.2 Matrix of Conditional Probabilities

Our goal is to estimate the distribution of customer segments $\hat{q}$ using the observed vector of basket types b and the given matrix of conditional probabilities R. Note that the matrix R has n rows and m columns. In the estimation, we utilize the equation $\hat{p} = \mathrm{R}\hat{q}$, where $\hat{p} = x/a$ is the estimated distribution of basket types. There are three distinctive cases based on the structure of the matrix R.

(a) If the matrix R has linearly dependent columns, the estimation faces identifiability

issues. Specifically, if we further assume linearly independent rows, the vector of customer probabilities $\hat{q}$ is not identifiable as the equation $\hat{p} = \mathrm{R}\hat{q}$ with variable $\hat{q}$ has infinite number of solutions. Note that a matrix with $n < m$ has always linearly dependent columns. In practice, we recommend avoiding this case by finding appropriate dimension of basket types and customer segments.

(b) If the matrix R has linearly independent columns and is of square form, i.e. $n = m$, the estimation is quite simple. In this case, we can invert the matrix R and obtain the straightforward estimator $\hat{q} = \mathrm{R}^{-1}\hat{p} = \mathrm{R}^{-1}x/a$. However, finding a meaningful basket types and customer segments satisfying the independence and square restriction can be quite difficult. Therefore, this case is theoretically ideal but unlikely to occur in practice.

(c) If the matrix R has linearly independent columns and is tall, i.e. $n > m$, the estimation is more complex. In this case the equation $\hat{p} = \mathrm{R}\hat{q}$ can be inconsistent, i.e. without a solution for $\hat{q}$. However, we can find suitable $\hat{q}$ by minimizing an error between $\hat{p}$ and $\mathrm{R}\hat{q}$. For this purpose, we utilize the least squares and maximum likelihood methods. As this case is the most realistic, we devote to it the rest of this section.

The three cases are illustrated in Figure 4.1. We discuss how to obtain the matrix R in practice in Section 4.2.6.

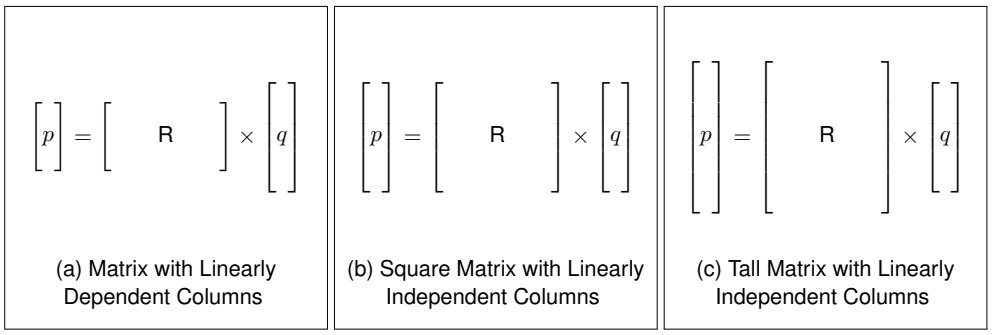

Fig. 4.1: Illustration of the three possible structures of the conditional probability matrix R.

4.2.3 Least Squares Estimation

Let us assume that we observe the vector of basket types b. The total number of transactions is denoted as a while the vector of the numbers of transactions with given basket type is denoted as x. Let us further assume that we know the conditional probability matrix R and it has linearly independent columns with $n > m$. The *least squares (LS)* method minimizes the squared error given by

$$SE(q) = \sum_{i=1}^{n} \left(\hat{p}_i - \sum_{j=1}^{m} r_{i,j} q_j \right)^2 , \tag{4.3}$$

where $\hat{p} = x/a$. The estimates $\hat{q} = (\hat{q}_1, \ldots, \hat{q}_m)'$ are found by solving the quadratic optimization problem

$$\begin{aligned} \hat{q} \in \arg\min_{q} & \sum_{i=1}^{n} \left(\sum_{j=1}^{m} r_{i,j} q_j \right)^2 - 2 \sum_{i=1}^{n} \sum_{j=1}^{m} \hat{p}_i r_{i,j} q_j \\ \text{s.t.} & \sum_{j=1}^{m} q_j = 1, \\ & 0 \leq q_j \leq 1, \qquad j = 1, \ldots, m. \end{aligned} \tag{4.4}$$

4.2.4 Maximum Likelihood Estimation

Let us assume the same setting as in Section 4.2.3. The *maximum likelihood (ML)* method maximizes the likelihood function or, equivalently, the logarithm of the likelihood function. The logarithmic likelihood function is given by

$$\begin{aligned} L(q) &= \ln \mathrm{P}\left[B = b | Q = q\right] \\ &= \sum_{k=1}^{a} \ln \mathrm{P}\left[B_k = b_k | Q = q\right] \\ &= \sum_{i=1}^{n} x_i \ln \mathrm{P}\left[B_k = i | Q = q\right], \end{aligned} \tag{4.5}$$

where the second equality holds as $\{B_k\}$ are independent and the third equality holds as $\{B_k\}$ are identically distributed. Using the law of total probability, we have

$$\mathrm{P}\left[B_k = i | Q = q\right] = \sum_{j=1}^{m} \mathrm{P}\left[B_k = i | C_k = j, Q = q\right] \mathrm{P}\left[C_k = j | Q = q\right]. \tag{4.6}$$

The logarithmic likelihood function is then

$$L(q) = \sum_{i=1}^{n} x_i \ln \left(\sum_{j=1}^{m} r_{i,j} q_j \right). \tag{4.7}$$

The estimates $\hat{q} = (\hat{q}_1, \ldots, \hat{q}_m)'$ are found by solving the nonlinear optimization problem

$$\begin{aligned} \hat{q} \in \arg\max_{q} & \sum_{i=1}^{n} x_i \ln \left(\sum_{j=1}^{m} r_{i,j} q_j \right) \\ \text{s.t.} & \sum_{j=1}^{m} q_j = 1, \\ & 0 \leq q_j \leq 1, \qquad j = 1, \ldots, m. \end{aligned} \tag{4.8}$$

Note that this is equivalent to minimizing the Kullback–Leibler divergence from Rq to $\hat{p}$

$$KL(q) = \sum_{i=1}^{n} \hat{p}_i \ln \left(\frac{\hat{p}_i}{\sum_{j=1}^{m} r_{i,j} q_j} \right), \tag{4.9}$$

where $\hat{p} = x/a$. The *Kullback–Leibler divergence* also known as *relative entropy* measures how one probability distribution differs from a second probability distribution.

4.2.5 Number of Customers

Finally, we estimate the number of customers. Suppose that we have the vector $\hat{q} = (\hat{q}_1, \ldots, \hat{q}_m)'$ containing the estimated probabilities of the customer segments of a transaction. Let us assume that we also know the vector $f = (f_1, \ldots, f_m)'$ containing the average frequencies of visits by a customer in a given customer segment. We discuss how to obtain the vector f in practice in Section 4.2.6. If we observe a transactions, the number of transactions with customer segment j is estimated as $\hat{q}_j a$ for $j = 1, \ldots, m$. The average number of customers in customer segment j is estimated as $\hat{q}_j a / f_j$ for $j = 1, \ldots, m$. The average total number of customers is estimated as

$$\hat{d} = \sum_{j=1}^{m} \frac{\hat{q}_j a}{f_j}. \tag{4.10}$$

The probability of a customer belonging to customer segment j is then estimated as

$$\hat{u}_j = \frac{\hat{q}_j a}{f_j \hat{d}}, \qquad j = 1, \ldots, m. \tag{4.11}$$

Note that $\hat{q} = (\hat{q}_1, \ldots, \hat{q}_m)'$ denotes the estimated distribution of the customer segments of a *transaction* while $\hat{u} = (\hat{u}_1, \ldots, \hat{u}_m)'$ denotes the estimated distribution of the customer segments of a *customer*.

4.2.6 Observed Samples

Let us remind the assumptions of our proposed approach for the estimation of the average total number of customers $\hat{d}$.

(a) The vector of basket types b is observed.

(b) The conditional probability matrix R is known and has linearly independent columns.

(c) The frequency vector f is known.

The question is how can R and f be obtained in practice. We consider that we observe two samples of transactions. In the *monitored sample*, the transactions are linked to loyalty cards and we observe basket types denoted as b^0 and also customer segments denoted as c^0. The *unmonitored sample* contains transactions by customers without the loyalty card and we observe only basket types b. So far, we have focused on the unmonitored sample as the number of customers is known in the monitored sample. However, we can use transactions from the monitored sample to estimate the conditional probability matrix R^0 and the frequency vector f^0. This can be done straightforwardly as we have the complete knowledge of customers in the monitored sample. We then assume that $\mathrm{R} = \mathrm{R}^0$ and $f = f^0$. In the context of the monitored and unmonitored samples, the assumptions (b) and (c) can be reformulated in the following way.

(b*) The conditional probability matrix R is equal to the conditional probability matrix R^0 of the monitored sample and has linearly independent columns.

(c*) The frequency vector f is equal to the frequency vector f^0 of the monitored sample.

In other words, we assume that the behavior of customers in a given customer segment is the same with and without the loyalty card while the distribution of customers segments itself can differ. This setting is quite realistic as we discuss in the empirical study in Section 4.4.

4.3 Simulation Study

4.3.1 Design of the Simulation Study

We investigate the finite-sample properties of the proposed approach using simulations. For simplicity, we focus on the situation with $n = 6$ basket types and $m = 3$ customer segments. We consider both the basket types and the customer segments to be given in advance. We utilize several simulation scenarios with various sample sizes and violated assumptions. Let ν denote the number of simulations of a specific scenario. Furthermore, let d_s denote the true number of customers and $\hat{d}_s$ its estimate in simulation s, $s = 1, \ldots, \nu$.

For the evaluation of the accuracy of the estimates, we use the *mean absolute percentage error (MAPE)* given by

$$MAPE\left(\hat{d}_1, \ldots, \hat{d}_\nu\right) = 100\% \cdot \frac{1}{\nu} \sum_{s=1}^{\nu} \left| \frac{d_s - \hat{d}_s}{d_s} \right|. \tag{4.12}$$

It is the average absolute difference between the true values and the estimated values divided by the true values. The estimates with lower MAPE are preferred. In some simulation scenarios, we have different values of the true number of customers. For example, the true number of unmonitored customers in scenario (ii) in Table 4.1 is 500 000 while it is 500 in scenario (iv). The naive estimates assuming the same distribution for customers in the monitored and unmonitored samples are 300 000 in scenario (ii) and 300 in scenario (iv). The absolute error is 200 000 and 200 respectively while the MAPE is 40 percent in both cases. The MAPE therefore allows us to compare differently scaled scenarios. This is the main motivation for utilizing the MAPE measure in our simulation study.

Next, we present the values of the parameters for the *benchmark scenario* denoted as (ii) in Table 4.1. The customer distribution vector, the conditional probability matrix and the frequency vector for the monitored and unmonitored samples are respectively given by

$$q^0 = \begin{pmatrix} 0.60 \\ 0.20 \\ 0.20 \end{pmatrix}, \; q = \begin{pmatrix} 0.20 \\ 0.20 \\ 0.60 \end{pmatrix}, \; \mathrm{R}^0 = \begin{pmatrix} 0.40 & 0.10 & 0.10 \\ 0.20 & 0.10 & 0.10 \\ 0.10 & 0.40 & 0.10 \\ 0.10 & 0.20 & 0.10 \\ 0.10 & 0.10 & 0.40 \\ 0.10 & 0.10 & 0.20 \end{pmatrix}, \; \mathrm{R} = \mathrm{R}^0, \; f^0 = \begin{pmatrix} 6.00 \\ 3.00 \\ 1.50 \end{pmatrix}, \; f = f^0. \tag{4.13}$$

In the monitored sample, most customers shop with relatively high frequency while in the unmonitored sample, most customers shop with much lower frequency. Each customer segment has its own two dominant basket types. Similar behavior is observed in the empirical study in Section 4.4. Furthermore, the number of observations is $a^0 = 1\,000\,000$ for the monitored sample and $a = 1\,000\,000$ for the unmonitored sample. Such values of the customer distribution vector, the frequency vector and the number of observations result in the number of customers $d^0 = 300\,000$ in the monitored sample and $d = 500\,000$ in the unmonitored sample. Note that the customer distribution vector changes in the monitored and unmonitored samples while the conditional probability matrix and the frequency vector remain the same. This is consistent with our assumptions. If the customer distribution vector does not change, the naive estimator is superior and there is no need for the proposed approach as we demonstrate in scenario (i) in Table 4.1. This is, however, unrealistic in practice.

We base all the other scenarios on the benchmark scenario (ii) with some modifications. Scenarios (iii)–(v) are designed to assess the effects of sample sizes while scenarios (vi)–(ix) are designed to investigate the effects of assumption violations.

Tab. 4.1: Mean absolute percentage errors with standard deviations of the naive, least squares (LS) and maximum likelihood (ML) estimates using 50 000 simulations with 9 scenarios.

Simulation Scenario		Mean			Standard Deviation		
No.	Description	Naive	LS	ML	Naive	LS	ML
(i)	No change in q	0.00	0.19	0.19	0.00	0.14	0.14
(ii)	Benchmark scenario	40.00	0.19	0.18	0.00	0.14	0.14
(iii)	Small $a^0 = 1000$	40.00	5.30	5.20	0.00	4.16	4.05
(iv)	Small $a = 1000$	40.00	2.67	2.62	0.00	2.00	1.97
(v)	Small $a^0 = a = 1000$	40.00	5.90	5.79	0.00	4.57	4.45
(vi)	Change in R	40.00	5.29	5.10	0.00	3.97	3.84
(vii)	Change in f	40.92	14.79	14.79	11.00	10.95	10.95
(viii)	Change in R and f	40.98	15.67	15.60	10.95	11.57	11.52
(ix)	Linear dependence in R	40.00	30.19	30.08	0.00	16.02	15.97

4.3.2 Sample Size

We consider various sizes of the monitored and unmonitored samples. Specifically, the size of the monitored sample a^0 ranges from 100 to 10 000 in the left plot of Figure 4.2 and is set to 1 000 in scenarios (iii) and (v) in Table 4.1. Similarly, the size of the unmonitored sample a ranges from 100 to 10 000 in the right plot of Figure 4.2 and is set to 1 000 in scenarios (iv) and (v) in Table 4.1. In the extreme case of 100 observations, the probability that some customer segment is not present in the monitored data sample is $4 \cdot 10^{-10}$ for the customer distribution vector and the frequency vector given by (4.13). In our simulations, all customer segments are always present. As the customer segmentation is build on the monitored sample in real applications, the situation with missing customer segments would not make any sense.

We focus on the performance of the maximum likelihood (ML) estimator. As we can see in Figure 4.3, the ML estimator performs quite well even for very small monitored and unmonitored data samples. The size of the monitored sample is more important in comparison to the size of the unmonitored sample as the monitored sample is used to estimate the conditional probability matrix and the frequency vector. Overall, the sample size is not really an issue as retail stores have typically large amount of data available.

4.3.3 Violation of Assumptions

First, we examine the change in the conditional probability matrix in the unmonitored sample. Let r_j^0 denote the j-th column of matrix R^0 and r_j denote the j-th column of matrix R. For the purposes of the simulation study, we consider the columns r_j to be random vectors following the Dirichlet distribution with the concentration parameter equal to $\alpha^{\mathrm{R}} r_j^0$. Parameter α^{R} controls to what degree is the assumption $\mathrm{R} = \mathrm{R}^0$ violated. Small values indicate significant deviation from R^0 while for $\alpha^{\mathrm{R}} \to \infty$, random vector R converge to deterministic value R^0. The value of α^{R} ranges from 10 to 1 000 in the left plot of Figure 4.3 and is set to 100 in scenarios (vi) and (viii) in Table 4.1. This parameter is, however, a bit cumbersome to interpret. From the Bayesian point of view, $\alpha^{\mathrm{R}} r_j^0$ is the increment of the concentration parameter of the Dirichlet distribution after α^{R} random variables distributed according to r_j^0 are observed. The average absolute difference between the elements of R^0 and R is 0.08 for $\alpha^{\mathrm{R}} = 10$, 0.03 for $\alpha^{\mathrm{R}} = 100$ and 0.01 for $\alpha^{\mathrm{R}} = 1\,000$.

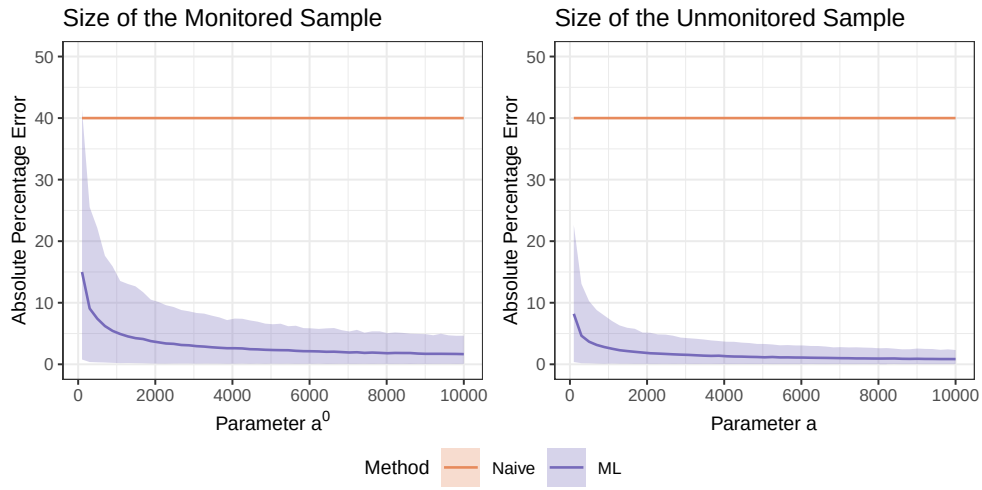

Fig. 4.2: Mean absolute percentage errors with 95% confidence intervals of the naive and maximum likelihood (ML) estimates using 5 000 simulations with various sample sizes.

Second, we examine the change in the frequency vector in the unmonitored sample. In a similar fashion, we consider the frequency vector f to be random and to follow the Dirichlet distribution. In this case, however, the sum of the random vector is set to the sum of f^0 instead of 1. Again, we control to what degree is the assumption $f = f^0$ violated using parameter α^f. The value of α^f ranges from 10 to 1 000 in the right plot of Figure 4.3 and is set to 100 in scenarios (vii) and (viii) in Table 4.1. The average absolute difference between the elements of f^0 and f is 1.10 for $\alpha^f = 10$, 0.36 for $\alpha^f = 100$ and 0.11 for $\alpha^f = 1\,000$. Note that a change in the frequency vector f causes a change in the number of unmonitored customers d, which is also a random variable in this case.

Finally, we examine linear dependence in the conditional probability matrix. In scenario (ix) in Table 4.1, we set the last column in matrices R^0 and R to be the average of the other columns.

We focus on the performance of the maximum likelihood (ML) estimator. The method is more sensitive to changes in the frequency vector f than to changes in the conditional probability matrix R as we can clearly see in Figure 4.3. However, even for value $\alpha^f = 100$ with average absolute difference 0.36 in scenario (vii), the ML significantly outperforms

the naive estimator. As shown in scenario (ix), linear dependence in R can cause serious problems. The error caused by linear dependence is affected by the frequency vector f. For example, if there are two identical columns in R and their corresponding frequencies are the same, there is no error. If the frequencies, however, significantly differ (as they do in our case), the error can be substantial. Luckily, linear dependence in R can be avoided by selecting suitable segmentation using the monitored sample.

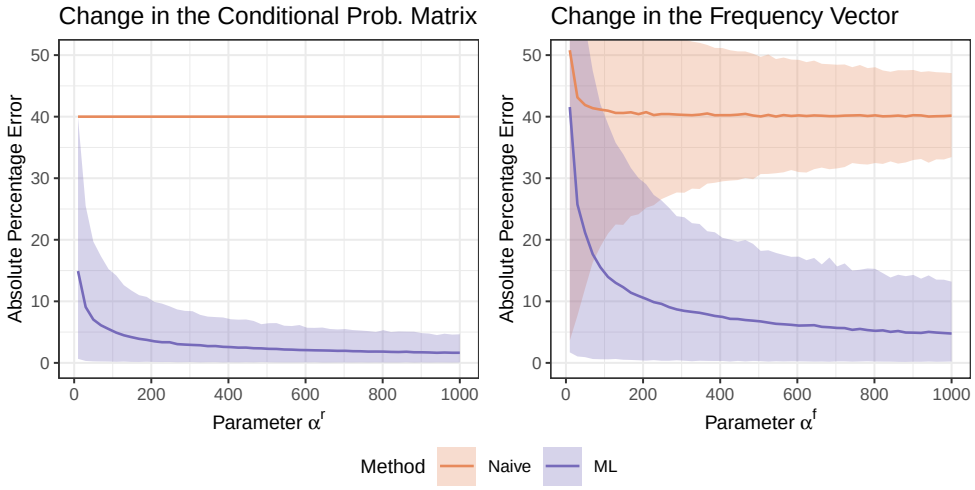

Fig. 4.3: Mean absolute percentage errors with 95% confidence intervals of the naive and maximum likelihood (ML) estimates using 10 000 simulations with violated assumptions.

4.3.4 Comparison of Methods

Finally, we focus on differences between the LS and ML estimators. We can see in Table 4.1 that estimation errors are very similar in both methods. In all scenarios (i) to (ix), the ML estimator has slightly lower error and also slightly lower standard deviation of the error. The differences are, however, minimal. Therefore, we focus on both LS and ML estimators in the rest of the chapter.

4.4 Empirical Study

4.4.1 Motivation

The loyalty program of the studied Czech retail chain is well-known in the Czech Republic as it is frequently advertised not only in stores but in commercials on TV, radio and newspaper as well. The loyalty program brings a lot of benefits to its members, such as an extra discount for members or gifts for buying specific products. Even more, customers can collect bonus points for spending and then exchange them for 10 or 5 percent discount on the whole basket. These loyalty perks resulted into high popularity of the loyalty program among the customers and nowadays the majority of receipts are linked to a specific customer with the loyalty card.

The knowledge of the number of customers who are not members of the loyalty program is crucial for various marketing strategies. Among the management of the company it is believed that the loyalty program may be already saturated since approximately 2015 and the vast majority of customers who are willing to join the loyalty program are already a members. In this study, we focus on quantification of this hypothesis.

Our *monitored sample* consists of the loyalty club members and our *unmonitored sample* are customers without the loyalty card. The goal is to reliably estimate the number of unique customers in the unmonitored sample. We do not have any prior knowledge about the distribution of customers aside of expectation that frequent customers are already members. To find the appropriate segmentation, we propose an automatic approach consisting in two-phase clustering of baskets and customer using k-means. We chose k-means for this purpose because of easy application and at the same time well interpretable clusters.

4.4.2 Transaction Data

We use both proposed methods on a sample of real transaction data. Our dataset consists of individual purchase data of one of the retail chains in the drugstore market of the Czech Republic. The retail chain sells over 10 thousand products which are divided into 55 categories. This categorization is done by an expert's opinion based on the product properties and purpose. Each product is also assigned to one of the three price-levels –

low-end, standard and high-end. We use quarterly data from the table of all transactions during years 2017 and 2018. The dataset includes every single product sold. It also includes identification of the basket and in some cases a link to a specific customer. We remove monitored customers with extremely high frequency (more than 15 visits per month) from the dataset as they are a special type of customers and we believe that no customer without the loyalty card exhibits this behavior. The cleaned dataset consists of over 50 million receipts, with 69 percent linked to a specific customer through their loyalty card.

4.4.3 Basket Types

In order to get meaningful customer segmentation, we use two phase method involving segmentation of both baskets and customers. Particularly, in the first phase, we define basket types and in the second phase, we define customer segments.

Basket types are estimated using k-means on the four dimensions – the *value* of the basket, the share of *premium products*, the share of *products for children* and the *diversity* of the basket. The value of the basket is normalized by the 95% quantile, therefore for each basket we get a number between 0 and 1 with 1 assigned to baskets with the value of at least 95% quantile. The share of premium products is a number between 0 and 1 determined on the value spend on the premium products and the whole value of the basket. Similarly, we define the share of products for small children. Products for children include diapers, wipes and baby food mostly for children up to age 3. The diversity of the basket is determined by the number of products normalized by 90% quantile as we want to distinguish whether the basket contain multiple categories or just one or two.

The dataset is adjusted such that every dimension has the mean of 0 and the variance of 1. The combination of these four dimensions then allows us to distinguish 13 different types of baskets using k-means. The number of basket types is determined using Davies-Bouldin statistic which is minimized in the case of 13 centers. The clusters are relatively easy to interpret using centers of the clusters. The proposed method is robust to the choice of the number of basket types. The differences in the results were minimal with selections from 10 to 15 basket types.

4.4.4 Customer Segments

Similarly to the first phase, we use four dimensions – the *total value* of the customer during the chosen time period, the share of *premium products* and the share of *products for children* in customer purchase history during the given time-frame and the average *diversity* of the customer's baskets. The share of children products is included as the parents of small children are usually frequent customers often focused on price and sales promotions.

After adjusting of the dataset such that every dimension has the mean of 0 and the variance of 1, we use k-means to determine 5 customer segments. Again, we used Davies-Bouldin statistic for estimation of optimal number of clusters as we have found out, that the accuracy of the method is maximized following Davies-Bouldin statistic. With higher number of customer segments, k-means tends to find segments with similar probability of basket type but different frequency. This results in various inaccuracies in estimates. We recommend checking the rows of matrix R to be similar. In the case of high level of similarity, consider the suitability of the dimensions used for segmentation.

The description of the customer segments and their shopping frequency are described in Table 4.2. The largest group among customers with the loyalty card are two types of general customers distinguished by the share of premium products. These are the archetype of general customer who exhibits standard behavior and moderate shopping frequency. Another segment of customers are the suppliers. These are the customers who visit the store rarely but their value is high and typical basket value very high. The customers who buy products for small children are one of the most valuable ones as their frequency is very high and their revenue moderate. However, margins of their most frequently purchased products are below standard. For this reason, we distinguish them as a special segment as the goal is not only to maximize their revenue but also redirect them to the products with higher margins. The most valuable customers belongs to Loyal Customer segment with very high frequency, moderate average diversity and overall very high value. While segments Loyal Customer and Suppliers make just around one quarters of all members of loyalty program, they make over half, 55 percent on average, of revenue.

Note that for every quarter we find the segments using k-means on the quarterly data. However, the centers of segments used to be very similar regardless the chosen period and the description in Table 4.2 is valid for all of the quarters.

Tab. 4.2: Description of the customers segments.

Segment	Value	Premium	Children	Diversity	Frequency
General Budget	Very Low	Very Low	Low	Low	Moderate
General Premium	Low	High	Low	Low	Moderate
Loyal Customer	Very High	Moderate	Low	Moderate	Very High
Supplier	High	Moderate	Low	High	Low
With Children	Moderate	Low	Very High	Low	High

4.4.5 Empirical Validation

In order to validate the method using empirical data, we take the *monitored sample* as defined above – customers with loyalty card, and then divide in two samples. In the first sample are included customers who added e-mail address to their loyalty account and the second sample is the rest of customers. Therefore, the customers in Sample 1 must have provided an additional data, while the customers in Sample 2 did not have to take any further action aside of joining the loyalty program. This partition should emulate the differences between *monitored sample* and *unmonitored sample* used in this chapter. We look closer at the second quarter of 2018 which is chosen because we have available an expert estimation of the number of non-member customers for this period and which is in line with our estimate.

Figure 4.4 shows the comparison of estimated distribution by segments in the second quarter of 2018. It is apparent that both LS and ML methods give much better results than naive approach.

In Table 4.3 is highlighted the performance of the method for different time periods with 13 basket types and 5 customer types. For weekly data, our model estimates are the most accurate with approximately half the error of a naive estimate. The accuracy decreases as the time period increases. Compared to the naive estimation, the estimates using LS and MS methods have significantly better results. The statistics are calculated using data from 2017 and 2018, so we have only 2 observations for the annual data. Note that in most cases, the results are underestimated by all methods.

Decreasing accuracy of estimates seems to be related to inaccuracy due to segmentation.

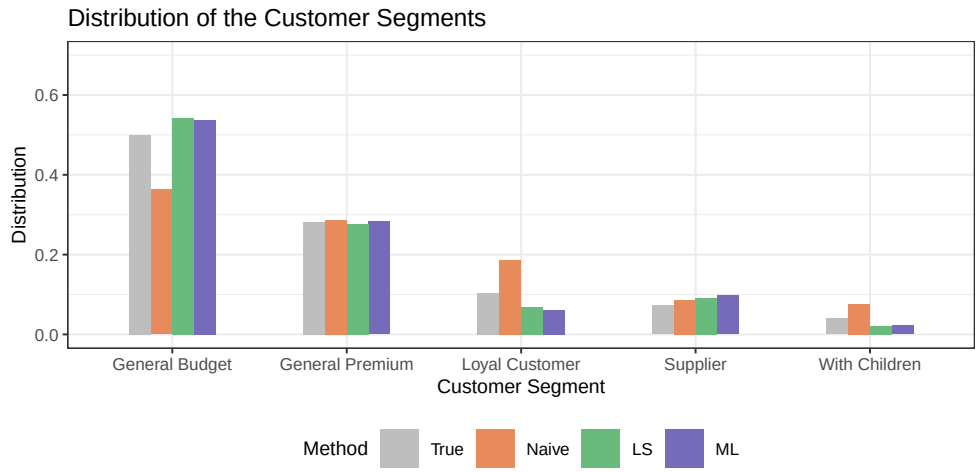

Fig. 4.4: Comparison of True, Naive, ML and LS estimated distributions of customer segments in 2018 Q2.

For example, in the case of weekly data, the difference in frequencies of monitored and unmonitored samples is significantly smaller than in the case of quarterly and especially yearly data. For very long time periods, it is advisable to combine more customer segment estimation approaches than just k-means. k-means in these cases, even with a higher number of segments, defines most segments with relatively large shopping frequencies. As a results, the estimated numbers of customers are underestimated.

Contrary to the simulation study, the LS estimator is more precise than the ML estimator on the limited sample of two years data. In the rest of the section we interpret the results primary using the ML estimator. However, the LS estimator gives only minimally different results when estimating the total number of customers.

4.4.6 Distribution of Non-member Customers

The next part deals with the estimation of non-member customers in the second quarter of 2018 using unmonitored sample. As we have shown before, the results of both methods are close in terms of total number of customers using both simulation study and empirical validation with ML method having better results in extensive simulation study and LS

Tab. 4.3: Mean absolute percentage errors with standard deviations of the naive, least squares (LS) and maximum likelihood (ML) estimates using empirical dataset from years 2017 to 2018.

	Mean			Standard Deviation		
Time period	Naive	LS	ML	Naive	LS	ML
Week	2.42	1.04	1.18	0.41	0.52	0.51
Month	13.33	3.33	3.53	1.56	2.04	3.58
Quarter	24.52	4.38	5.75	2.00	2.09	2.64
Year	34.90	16.79	17.16	2.39	1.88	1.49

method having slightly better results in empirical validation on limited sample. For this reason, we show results of both methods. Primary, we focus on ML results with LS results mentioned in parenthesis.

It is reasonable to assume that the distribution of members and non-members of the loyalty program is substantially different and our case study of the drugstore retail chain supports it. Loyal customers and suppliers are not present in non-member group. This is expected, as the customer who frequently visit the shops get significantly better prices and other advantages and joining the loyalty program is very simple and fast – it can be even done during checkout. The distribution of other customer segments also change. While general customers form just 69.3 percent of the loyalty program members, in non-member sample they form 96.2 percent (98.9 percent using LS). Similarly, customers with children make 5.7 percent of customers with the loyalty card and only 3.8 percent (1.0 percent using LS) of customers without the loyalty card. The comparison between distributions of members and non-members of the loyalty program in each customer segment is shown in Figure 4.5.

The results support the initial hypothesis that the loyalty program is already saturated and customers who are willing to join the loyalty program are already members. The differences in customer segment distributions between members and non-members of the loyalty program are crucial for estimation of unique non-members customers. The number estimated by the naive approach assuming the same distribution of members and non-members is 23.7 percent (25.2 percent using LS) lower than the number estimated by the

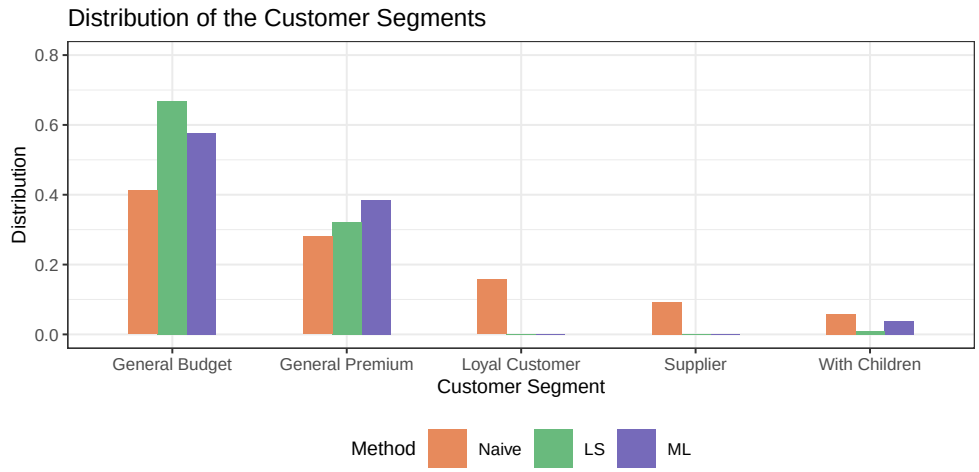

Fig. 4.5: Comparison of Naive, ML and LS estimated distributions of customer segments in 2018 Q2.

proposed method which takes the differences into account. The naive approach estimates 748 196 customers while the proposed method estimates 980 568 (ML) and 1 000 115 (LS) customers without loyalty card. This result is in line with the expert's estimate and also expectations of the management of the company.

Using our estimates, we are able to study the penetration of the loyalty program in each customer segment. The results support the hypothesis that the loyalty program is close to saturation as the most valuable customers are already members of the loyalty program. Premium general customers are members of the loyalty program in 53.6 percent (57.5 percent using LS) and budget general customers in 53.2 percent (49.0 percent using LS). For customers with children, we got a different results based on the method as we have 71.0 percent (89.5 percent using LS) participation rate. Here, a large percentage difference between our methods is evident and we do not know how to exactly explain the difference. We suppose that the difference is generally due to the small absolute number of these customers, which in turn affects the percentage difference. However, the results supports hypothesis that majority of customers with children are already members of loyalty program. The high participation rate is probably caused by the recent focus of the company on attracting customers with children into the loyalty program by additional club

discounts on products for children and other promotional activities.

4.5 Conclusion

We propose a method for estimation of the total number of unique customers using retail transaction data. The method also estimates the number of unique customers in each customer segment. We verify the proposed method using synthetic data in the extensive simulation study. The method performs quite well even if its assumptions are violated to some extent. Therefore, we expect the method to give reliable results in real applications.

In the empirical part, we use retail transaction data from a Czech drugstore's retail chain. We quantify the initial proposition that the loyalty program is very popular among the most valuable customers. This is something which has been done only by expert's opinion before and therefore the estimates were contaminated by high uncertainty. The reliable quantification is the main contribution of the method. Overall, the resulting estimates are crucial not only for planning of promotional sales and other marketing techniques but may also be used in predictions of demand for specific types of products.

5. Summary and Future Research

The thesis focuses on the analysis of retail transaction data. The goal is to develop new data driven methods to deal with various business inspired problems. Specifically, three topics in retail transaction data analysis are investigated in order to answer following questions:

- Which products are perceived by customers as substitutes?
- How to evaluate customers behavior?
- How to estimate the number of unique customers?

The main results of doctoral thesis are following:

1. A method for clustering substitutes is proposed and verified using extensive simulation study and then applied using real drugstore dataset. The clustering is formulated as an optimization problem in which joint occurrences of products from the same cluster in the same basket are minimized. The number of clusters is given by analyst and it is useful to try the method with different numbers of clusters. The results are highly accurate and the method allows to discover previously unknown relationships between products.
2. A new way to segment customers by their shopping behavior is proposed. The new segmentation uses a different approach than the methods commonly used. Firstly, the baskets are clustered using k-means by the category penetration and the total value of the basket. Secondly, the customers are clustered by k-means using the information of the share of each basket type in their shopping history. This is a new perspective to consider customer behavior as is shown in comparison. The proposed segmentation was successfully implemented in practice.
3. A new approach to the estimation the number of unique customers is presented. Again, two phase clustering is used. Firstly, basket types are defined and secondly,

customer segments are set. The estimation is based on the fitting of known customers' segments obtained from the members of the loyalty program to the set of non-members receipts. Using shopping behavior of members of loyalty program while allowing different distribution of customer segment between members and non-members lead to accurate estimates. The results are in line with the expert's estimation; however, the method also allows the quantification of customer segments distribution within non-member customers with only a very limited expert's input.

As shown, the goals of the thesis have been fulfilled; however, some problems remain for future research. For example the method for clustering products presented in Chapter 2 is computationally expensive. In my recent work Sokol (2018b), I dealt with the same problem using different approach to modeling in order to reduce computational demands. Instead of finding clustering which minimize the co-occurrence of the products in the same basket, the approach uses a special way of computing the similarity matrix on which Ward's hierarchical clustering method is applied. The similarity matrix stems from the co-occurrence of products in the same basket as utility data. Products are denoted similar when they have similar co-occurring products and simultaneously are not frequently present in the same basket. The method is reasonably fast even for huge dataset of tens of millions rows and the preliminary results are promising.

Bibliography

Abe, M. 2009. "Counting Your Customers" One by One: A Hierarchical Bayes Extension to the Pareto/NBD Model. *Marketing Science*. Volume 28. Issue 3. Pages 541–553. ISSN 0732-2399. https://doi.org/10.1287/mksc.1090.0502.

Ammar, A., Elouedi, Z., Lingras, P. 2015. Semantically Segmented Clustering Based on Possibilistic and Rough Set Theories: Semantically Segmented Clustering. Volume 30. Issue 6. Pages 676–706. ISSN 08848173. https://doi.org/10.1002/int.21723.

Ammar, A., Elouedi, Z., Lingras, P. 2016. Meta-Clustering of Possibilistically Segmented Retail Datasets. *Fuzzy Sets and Systems*. Volume 286. Pages 173–196. ISSN 0165-0114. https://doi.org/10.1016/j.fss.2015.07.019.

Andrews, R. L., Currim, I. S. 2002. Identifying Segments with Identical Choice Behaviors Across Product Categories: An Intercategory Logit Mixture Model. *International Journal of Research in Marketing*. Volume 19. Issue 1. Pages 65–79. ISSN 0167-8116. https://doi.org/10.1016/S0167-8116(02)00048-4.

Batislam, E. P., Denizel, M., Filiztekin, A. 2007. Empirical Validation and Comparison of Models for Customer Base Analysis. *International Journal of Research in Marketing*. Volume 24. Issue 3. Pages 201–209. ISSN 0167-8116. https://doi.org/10.1016/j.ijresmar.2006.12.005.

Boon, E., Ofek, N. 2016. Deal of the Day: Analysing Purchase Frequency-Based Subscriber Segmentation. *International Journal of Market Research*. Volume 58. Issue 1. Pages 95–118. ISSN 1470-7853. https://doi.org/10.2501/ijmr-2016-006.

Borin, N., Farris, P. W., Freeland, J. R. 1994. A Model for Determining Retail Product Category Assortment and Shelf Space Allocation. *Decision Sciences*. Volume 25. Issue

3. Pages 359–384. ISSN 0011-7315. https://doi.org/10.1111/j.1540-5915.1994.tb01848.x.

Bradlow, E. T., Gangwar, M., Kopalle, P., Voleti, S. 2017. The Role of Big Data and Predictive Analytics in Retailing. *Journal of Retailing*. Volume 93. Issue 1. Pages 79–95. ISSN 0022-4359. https://doi.org/10.1016/j.jretai.2016.12.004.

Chen, Y.-L., Kuo, M.-H., Wu, S.-Y., Tang, K. 2009. Discovering Recency, Frequency, and Monetary (RFM) Sequential Patterns from Customers' Purchasing Data. *Electronic Commerce Research and Applications*. Volume 8. Issue 5. Pages 241–251. ISSN 15674223. https://doi.org/10.1016/j.elerap.2009.03.002.

Chib, S., Seetharaman, P. B., Strijnev, A. 2002. Analysis of Multi-Category Purchase Incidence Decisions Using IRI Market Basket Data. *Advances in Econometrics*. Volume 16. Pages 57–92. ISSN 0731-9053. https://doi.org/10.1016/S0731-9053(02)16004-X.

Decker, R., Monien, K. 2003. Market Basket Analysis with Neural Gas Networks and Self-Organising maps. *Journal of Targeting, Measurement and Analysis for Marketing*. Volume 11. Issue 4. Pages 373–386. ISSN 0967-3237. https://doi.org/10.1057/palgrave.jt.5740092.

Duchessi, P., Schaninger, C. M., Nowak, T. 2004. Creating Cluster-Specific Purchase Profiles from Point-of-Sale Scanner Data and Geodemographic Clusters: Improving Category Management at a Major US Grocery Chain. *Journal of Consumer Behaviour*. Volume 4. Issue 2. Pages 97–117. ISSN 1472-0817. https://doi.org/10.1002/cb.162.

Fader, P. S., Hardie, B. G. S., Lee, K. L. 2005. "Counting Your Customers" the Easy Way: An Alternative to the Pareto/NBD Model. *Marketing Science*. Volume 24. Issue 2. Pages 275–284. ISSN 0732-2399. https://doi.org/10.1287/mksc.1040.0098.

France, S. L., Ghose, S. 2019. Marketing Analytics: Methods, Practice, Implementation, and Links to Other Fields. *Expert Systems with Applications*. Volume 119. Pages 456–475. ISSN 0957-4174. https://doi.org/10.1016/j.eswa.2018.11.002.

Germann, F., Lilien, G. L., Fiedler, L., Kraus, M. 2014. Do Retailers Benefit from Deploying Customer Analytics? *Journal of Retailing*. Volume 90. Issue 4. Pages 587–593. ISSN 0022-4359. https://doi.org/10.1016/j.jretai.2014.08.002.

Glady, N., Baesens, B., Croux, C. 2009. A Modified Pareto/NBD Approach for Predicting Customer Lifetime Value. *Expert Systems with Applications*. Volume 36. Issue 2. Pages 2062–2071. ISSN 0957-4174. https://doi.org/10.1016/j.eswa.2007.12.049.

Griva, A., Bardaki, C., Pramatari, K., Papakiriakopoulos, D. 2018. Retail Business Analytics: Customer Visit Segmentation Using Market Basket Data. *Expert Systems with Applications*. Volume 100. Pages 1–16. ISSN 0957-4174. https://doi.org/10.1016/j.eswa.2018.01.029.

Gruca, T. S., Klemz, B. R. 2003. Optimal New Product Positioning: A Genetic Algorithm Approach. *European Journal of Operational Research*. Volume 146. Issue 3. Pages 621–633. ISSN 0377-2217. https://doi.org/10.1016/s0377-2217(02)00349-1.

Holý, V., Sokol, O., Černý, M. 2017. Clustering Retail Products Based on Customer Behaviour. *Applied Soft Computing*. Volume 60. Pages 752–762. ISSN 1568-4946. https://doi.org/10.1016/j.asoc.2017.02.004.

Hruschka, H. 2014. Linking Multi-Category Purchases to Latent Activities of Shoppers: Analysing Market Baskets by Topic Models. *Marketing: ZFP - Journal of Research and Management*. Volume 36. Issue 4. Pages 267–278. ISSN 0344-1369. https://doi.org/10.15358/0344-1369_2014_4_267.

Hruschka, H., Lukanowicz, M., Buchta, C. 1999. Cross-Category Sales Promotion Effects. *Journal of Retailing and Consumer Services*. Volume 6. Issue 2. Pages 99–105. ISSN 0969-6989. https://doi.org/10.1016/S0969-6989(98)00026-5.

Jain, A. K. 2010. Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*. Volume 31. Issue 8. Pages 651–666. ISSN 0167-8655.

Jerath, K., Fader, P. S., Hardie, B. G. S. 2011. New Perspectives on Customer "Death" Using a Generalization of the Pareto/NBD Model. *Marketing Science*. Volume 30. Issue 5. Pages 866–880. ISSN 0732-2399. https://doi.org/10.1287/mksc.1110.0654.

Jiang, J.-H., Wang, J.-H., Chu, X., Yu, R.-Q. 1997. Clustering Data Using a Modified Integer Genetic Algorithm (IGA). *Analytica Chimica Acta*. Volume 354. Issue 1-3. Pages 263–274. ISSN 0003-2670. https://doi.org/10.1016/S0003-2670(97)00462-5.

Jonker, J.-J., Piersma, N., Van den Poel, D. 2004. Joint Optimization of Customer Segmentation and Marketing Policy to Maximize Long-Term Profitability. *Expert Systems*

with Applications. Volume 27. Issue 2. Pages 159–168. ISSN 0957-4174. https://doi.org/10.1016/j.eswa.2004.01.010.

José-García, A., Gómez-Flores, W. 2016. Automatic Clustering Using Nature-Inspired Metaheuristics: A Survey. *Applied Soft Computing*. Volume 41. Pages 192–213. ISSN 1568-4946. https://doi.org/10.1016/j.asoc.2015.12.001.

Kahan, R. 1998. Using Database Marketing Techniques to Enhance Your One-to-One Marketing Initiatives. *Journal of Consumer Marketing*. Volume 15. Issue 5. Pages 491–493. ISSN 0736-3761. https://doi.org/10.1108/07363769810235965.

Karp, R. M. 1972. Reducibility among combinatorial problems. In *Complexity of computer computations*. Springer. Pages 85–103.

Khajvand, M., Tarokh, M. J. 2011. Estimating Customer Future Value of Different Customer Segments Based on Adapted RFM Model in Retail Banking Context. *Procedia Computer Science*. Volume 3. Pages 1327–1332. ISSN 1877-0509. https://doi.org/10.1016/j.procs.2011.01.011.

Konuş, U., Verhoef, P. C., Neslin, S. A. 2008. Multichannel Shopper Segments and Their Covariates. *Journal of Retailing*. Volume 84. Issue 4. Pages 398–413. ISSN 0022-4359. https://doi.org/10.1016/j.jretai.2008.09.002.

Leeflang, P. S. H., Parreño Selva, J., Van Dijk, A., Wittink, D. R. 2008. Decomposing the Sales Promotion Bump Accounting for Cross-Category Effects. *International Journal of Research in Marketing*. Volume 25. Issue 3. Pages 201–214. ISSN 0167-8116. https://doi.org/10.1016/j.ijresmar.2008.03.003.

Li, Y. L., Tang, J. F., Chin, K. S., Luo, X. G., Han, Y. 2012. Rough Set-Based Approach for Modeling Relationship Measures in Product Planning. *Information Sciences*. Volume 193. Pages 199–217. ISSN 0020-0255. https://doi.org/10.1016/j.ins.2011.12.017.

Lilien, G. L., Roberts, J. H., Shankar, V. 2013. Effective Marketing Science Applications: Insights from ISMS-MSI Practice Prize Finalist Papers and Projects. *Marketing Science*. Volume 32. Issue 2. Pages 229–245. ISSN 0732-2399. https://doi.org/10.1287/mksc.1120.0756.

Lingras, P., Elagamy, A., Ammar, A., Elouedi, Z. 2014. Iterative Meta-Clustering Through Granular Hierarchy of Supermarket Customers and Products. *Information*

Sciences. Volume 257. Pages 14–31. ISSN 0020-0255. https://doi.org/10.1016/j.ins.2013.09.018.

Liu, D. R., Shih, Y. Y. 2005. Integrating AHP and Data Mining for Product Recommendation Based on Customer Lifetime Value. *Information and Management*. Volume 42. Issue 3. Pages 387–400. ISSN 0378-7206. https://doi.org/10.1016/j.im.2004.01.008.

Lockshin, L. S., Spawton, A. L., Macintosh, G. 1997. Using Product, Brand and Purchasing Involvement for Retail Segmentation. *Journal of Retailing and Consumer Services*. Volume 4. Issue 96. Pages 171–183. ISSN 0969-6989. https://doi.org/10.1016/S0969-6989(96)00048-3.

Ma, S., Büschken, J. 2011. Counting Your Customers from an "Always a Share" Perspective. *Marketing Letters*. Volume 22. Issue 3. Pages 243–257. ISSN 0923-0645. https://doi.org/10.1007/s11002-010-9123-0.

Manchanda, P., Ansari, A., Gupta, S. 1999. The "Shopping Basket": A Model for Multicategory Purchase Incidence Decisions. *Marketing Science*. Volume 18. Issue 2. Pages 95–114. ISSN 0732-2399. https://doi.org/10.2307/19321195.

Manning, C. D., Raghavan, P., Schütze, H. 2008. *Introduction to Information Retrieval*. First Edition. New York. Cambridge University Press. ISBN 978-0-521-86571-5. https://doi.org/10.1017/cbo9780511809071.

Miglautsch, J. R. 2000. Thoughts on RFM Scoring. *Journal of Database Marketing & Customer Strategy Management*. Volume 8. Issue 1. Pages 67–72. ISSN 1741-2439. https://doi.org/10.1057/palgrave.jdm.3240019.

Mild, A., Reutterer, T. 2003. An Improved Collaborative Filtering Approach for Predicting Cross-Category Purchases Based on Binary Market Basket Data. *Journal of Retailing and Consumer Services*. Volume 10. Issue 3. Pages 123–133. ISSN 0969-6989. https://doi.org/10.1016/S0969-6989(03)00003-1.

Mostafa, M. M. 2019. Clustering Halal Food Consumers: A Twitter Sentiment Analysis. *International Journal of Market Research*. Volume 61. Issue 3. Pages 320–337. ISSN 1470-7853. https://doi.org/10.1177/1470785318771451.

Mzoughia, M. B., Borle, S., Limam, M. 2018. A MCMC Approach for Modeling Customer Lifetime Behavior Using the COM-Poisson Distribution. *Applied Stochastic Models in Business and Industry*. Volume 34. Issue 2. Pages 113–127. ISSN 1524-1904. https://doi.org/10.1002/asmb.2276.

Neunhoeffer, F., Teubner, T. 2018. Between Enthusiasm and Refusal: A Cluster Analysis on Consumer Types and Attitudes Towards Peer-to-Peer Sharing. *Journal of Consumer Behaviour*. Volume 17. Issue 2. Pages 221–236. ISSN 1472-0817. https://doi.org/10.1002/cb.1706.

Niknam, T., Amiri, B. 2010. An efficient hybrid approach based on PSO, ACO and k-means for cluster analysis. Volume 10. Issue 1. Pages 183–197. ISSN 15684946. https://doi.org/10.1016/j.asoc.2009.07.001.

Noorian, F., Silva, A., Leong, P. 2016. gramEvol: Grammatical Evolution in R. *Journal of Statistical Software, Articles*. Volume 71. Issue 1. Pages 1–26. ISSN 1548-7660. https://doi.org/10.18637/jss.v071.i01. https://www.jstatsoft.org/v071/i01.

Peker, S., Kocyigit, A., Eren, P. E. 2017. LRFMP Model for Customer Segmentation in the Grocery Retail Industry: A Case Study. *Marketing Intelligence & Planning*. Volume 35. Issue 4. Pages 544–559. ISSN 0263-4503. https://doi.org/10.1108/mip-11-2016-0210.

Putra, I. K. G. D., Cahyawan, A. A. K. A., Shavitri H., D. 2012. Combination of Adaptive Resonance Theory 2 and RFM Model for Customer Segmentation in Retail Company. *International Journal of Computer Applications*. Volume 48. Issue 2. Pages 18–23. ISSN 0975-8887. https://doi.org/10.5120/7320-0110.

Rand, W. M. 1971. Objective Criteria for the Evaluation of Clustering Methods. *Journal of the American Statistical Association*. Volume 66. Issue 336. Pages 846–850. ISSN 0162-1459. https://doi.org/10.2307/2284239.

Reutterer, T., Mild, A., Natter, M., Taudes, A. 2006. A Dynamic Segmentation Approach for Targeting and Customizing Direct Marketing Campaigns. *Journal of Interactive Marketing*. Volume 20. Issue 3-4. Pages 43–57. ISSN 1094-9968. https://doi.org/10.1002/dir.20066.

Roberts, J. H., Kayande, U., Stremersch, S. 2014. From Academic Research to Marketing Practice: Exploring the Marketing Science Value Chain. *International Journal*

of Research in Marketing. Volume 31. Issue 2. Pages 127–140. ISSN 0167-8116. https://doi.org/10.1016/j.ijresmar.2013.07.006.

Russell, G. J., Kamakura, W. A. 1997. Modeling Multiple Category Brand Preference with Household Basket Data. *Journal of Retailing*. Volume 73. Issue 4. Pages 439–461. ISSN 0022-4359. https://doi.org/10.1016/s0022-4359(97)90029-4.

Russell, G. J., Petersen, A. 2000. Analysis of Cross Category Dependence in Market Basket Selection. *Journal of Retailing*. Volume 76. Issue 3. Pages 367–392. ISSN 0022-4359. https://doi.org/10.1016/s0022-4359(00)00030-0.

Sarantopoulos, P., Theotokis, A., Pramatari, K., Doukidis, G. 2016. Shopping Missions: An Analytical Method for the Identification of Shopper Need States. *Journal of Business Research*. Volume 69. Issue 3. Pages 1043–1052. ISSN 0148-2963. https://doi.org/10.1016/j.jbusres.2015.08.017.

Schmittlein, D. C., Morrison, D. G., Colombo, R. 1987. Counting Your Customers: Who Are They and What Will They Do Next? *Management Science*. Volume 33. Issue 1. Pages 1–24. ISSN 0025-1909. https://doi.org/10.1287/mnsc.33.1.1.

Schröder, N. 2017. Using Multidimensional Item Response Theory Models to Explain Multi-Category Purchases. *Marketing: ZFP – Journal of Research and Management*. Volume 39. Issue 2. Pages 27–37. ISSN 0344-1369. https://doi.org/10.15358/0344-1369-2017-2-27.

Schumacher, N. 2006. The Butterfly Effect: Estimating "Faux-New" Customers. *Journal of Consumer Marketing*. Volume 23. Issue 1. Pages 43–46. ISSN 0736-3761. https://doi.org/10.1108/07363760610641154.

Seret, A., Verbraken, T., Baesens, B. 2014. A New Knowledge-Based Constrained Clustering Approach: Theory and Application in Direct Marketing. *Applied Soft Computing*. Volume 24. Pages 316–327. ISSN 1568-4946. https://doi.org/10.1016/j.asoc.2014.06.002.

Sokol, O., Holỳ, V. 2018. Customer Segmentation Based on a Shopping Mission in the Retail Business.

Sokol, O., Holý, V. 2019. How Many Customers Does a Retail Chain Have? *arXiv preprint arXiv:1904.10199*.

Sokol, O. 2018a. Estimating the Number of Customers Using Market Basket Data. In *Proceedings of the 7th International Conference on Management*. Prešov. Bookman s.r.o. Pages 763–767. ISBN 978-80-8165-301-8.

Sokol, O. 2018b. Hierarchical Clustering of Products Using Market-basket Data. In *2nd International Conference On Advances In Business Management And Law (ICABML). Preprint*. Dubai.

Sokol, O., Holý, V. 2018. *Customer Segmentation Based on a Shopping Mission in the Retail Business*. Working Paper.

Srivastava, R. K., Leone, R. P., Shocker, A. D. 1981. Market Structure Analysis: Hierarchical Clustering of Products Based on Substitution-in-Use. *Journal of Marketing*. Volume 45. Issue 3. Pages 38–48. ISSN 0022-2429. https://doi.org/10.2307/1251540.

Trappey, C. V., Trappey, A. J. C., Chang, A.-C., Huang, A. Y. L. 2009. The Analysis of Customer Service Choices and Promotion Preferences Using Hierarchical Clustering. *Journal of the Chinese Institute of Industrial Engineers*. Volume 26. Issue 5. Pages 367–376. ISSN 1017-0669. https://doi.org/10.1080/10170660909509151.

Tsai, C. Y., Chiu, C. C. 2004. A Purchase-Based Market Segmentation Methodology. *Expert Systems with Applications*. Volume 27. Issue 2. Pages 265–276. ISSN 0957-4174. https://doi.org/10.1016/j.eswa.2004.02.005.

Weng, C.-H. 2016. Identifying Association Rules of Specific Later-Marketed Products. *Applied Soft Computing*. Volume 38. Pages 518–529. ISSN 1568-4946. https://doi.org/10.1016/j.asoc.2015.09.047.

Yang, A. X. 2004. How to Develop New Approaches to RFM Segmentation. *Journal of Targeting Measurement and Analysis for Marketing*. Volume 13. Issue 1. Pages 50–60. ISSN 0967-3237. https://doi.org/10.1057/palgrave.jt.5740131.

Yang, C.-L., Kuo, R., Chien, C.-H., Quyen, N. T. P. 2015. Non-Dominated Sorting Genetic Algorithm Using Fuzzy Membership Chromosome for Categorical Data Clustering. *Applied Soft Computing*. Volume 30. Pages 113–122. ISSN 1568-4946. https://doi.org/10.1016/j.asoc.2015.01.031.

Yim, M. Y.-C., Yoo, S.-C., Sauer, P. L., Seo, J. H. 2014. Hedonic Shopping Motivation and Co-Shopper Influence on Utilitarian Grocery Shopping in Superstores. *Journal of the*

Academy of Marketing Science. Volume 42. Issue 5. Pages 528–544. ISSN 0092-0703. https://doi.org/10.1007/s11747-013-0357-2.

Zhang, Y., Jiao, J., Ma, Y. 2007. Market Segmentation for Product Family Positioning Based on Fuzzy Clustering. *Journal of Engineering Design*. Volume 18. Issue 3. Pages 227–241. ISSN 0954-4828. https://doi.org/10.1080/09544820600752781.

List of Tables

List of Figures