GIMA
Geographical Information Management and Applications

Project A-Locate: Using location-allocation modelling to optimise human resources in retail environments

Thesis report

Name: Yannick Brangers
S6031137
Email: y.z.brangers@student.utwente.nl

Thesis report

Project A-Locate: Using location-allocation modelling to optimise human resources in retail environments

Some data is hidden with **<confidential>** marks for business data purposes in this public version. This data is available upon request (on certain conditions). Reach out to me: yannick@brangers.nl

Student:	Yannick Brangers
Supervisor GIMA:	Dr. C. Maat
Responsible professor:	Prof. dr. ir. P.J.M. van Oosterom
Thesis coordinator:	Ir. E. Verbree
External supervisors:	Rinske ten Hoopen
	Dr. Bart Voorn
Submission date:	1 March 2019

Preface

This thesis report which, from the beginning, is named 'Project A-Locate' is named so because it aims to allocate employees to the most optimal (A) location. During the thesis project, the case study organization has been an 'A location' for me by creating the work environment of a nice, diverse and friendly team and by providing the coffee. Most importantly they supplied me with all the data and resources for my research that enabled me to write this thesis. I am very honoured that they gave me this opportunity.

Apart from all the data, they also provided me with a lot of help, feedback, knowledge and support when needed. I want to give a special thanks to Rinske ten Hoopen and Bart Voorn who helped me guiding through the organization to get the project going with the needed support across different departments. Even the support from a professional cinematographer and animator were hired to produce professional (animation) videos to help explain the project to the organization and generate enough survey responses. I made huge progress for the academic learning goals, but with this thesis project, I made huge progress in personal development as well by working within a big organization. I learned a lot about how to get things done in a professional working environment.

Special thanks also to Maarten Goos and Ronja Röttger from Utrecht University, who helped me to set up the employee survey. It was an honour to collaborate with you and I am sure we were able to use our different skills and mindsets to the utmost. I am happy that Ronja is writing her PhD on the topic of my thesis. I hope my thesis is a good starting point for her research and that my methods will contribute to her research. I am looking forward to the more extensive results in the upcoming years!

Last but definitely not least, I want to thank all academic staff involved from TU Delft. To start with, I want to thanks Dongliang Peng, who hugely helped me to program the integer linear programming optimal allocation model. The sessions we had to create this program were intensive for me but definitely interesting and really informative. I learned a lot! I also want to thank the supervisors from the MSc GIMA program for their reviews, comments, feedback, help and other thoughts. The feedback I got was not always what I wanted to hear, but I am sure that it always helped me to improve my thesis. I want to especially thank my daily supervisor, Kees Maat, Peter van Oosterom as the responsible professor and lastly Edward Verbree as the thesis coordinator, thanks for being part of this thesis project.

Enjoy reading!

Yannick Brangers

Zaandam, 1 March 2019

Abstract

All stores of a brand have their characteristics. Store employees, in turn, have specific characteristics, skills and preferences as well. How can the supply of such unique human resource be matched optimally with the demand for specific types of stores on specific locations?

In this study, we cluster stores based on their sales and socio-demographic characteristics by performing cluster analysis using K-means, which gives insights in the main drivers of store sales differentiation and is a driver for the content of the conducted employee survey. By clustering the different characteristics for each cluster are clear to allow specific allopcation of employee skills. The link of sales data with the socio-demographic characteristics of the service area of a store is researched by comparing clustering on both data types. We found that spatial patterns can be found for clusters of stores. However, the clusters based on sales data do not completely overlap the clusters based on socio-demographic variables. The sales data provides the skill demand of a store for optimization.

After clustering, store employees are surveyed to gather their skills and preferences. The skills and work preferences of the employees are ranked and quantified using this survey, which also captures travel preferences and professional background. This survey supplied the data for the skill supply from an employee to be used for optimization. The survey showed the central tendency bias, where employees tend to give themselves just above average, but not perfect scores. The average rating for the different product groups is within a small range.

Lastly, we explore the options of the optimal allocation of employees, in a way the team skills match the store characteristics as good as possible while limiting travelling time and distance, by developing an allocation model using Integer Linear Programming. In doing so, this research generates unique knowledge using micro-organized data. Additionally, the methods developed in this research can be used in an applied contact to make a fact-based decision within human resources and store operations for the long term.

The Integer Linear Programming allocation model shows that the optimal solution involves relocating **<confidential>** % of the employees to another store. This optimization shows an improvement of skill supply and demand of **<confidential>** % while the average commute distances can be reduced with 14.33%.

Table of Contents

1. Introduction

1.1. Problem statement

Generating sales in store chains in part requires that employees advise customers about the relevant products (Söderlund, 2018), known as the Mincer equation (Heckman, Lochner, & Todd, 2005). Therefore, a reason why stores sell more of certain products could be that employees differ in providing a high-quality service to customers about the different products sold. For example, an employee of a drug store who has children might know more about baby-related products, whereas an employee who watches Nikkietutorials (a series of make-up related films on YouTube) might be particularly good at selling beauty products (Goos, 2018). Another reason certain stores sell different products can be the location of the store with an environment of people living close to the store with certain characteristics. Store location still represents a crucial driver of store performance in modern retail environments (Pan & Zinkhan, 2006).

If the service employees provide in selling certain products is important for store sales, recruiting, selecting, educating and ultimately also scheduling the right employees through thorough HR policies, is of pivotal importance to the organization to achieve their business objectives.

For employees, their skill match with the job has a strong effect on his or her job satisfaction (Allen & van der Velden, 2001). The person-organization fit (Deniz et al., 2015), is therefore of importance for the employer as well as the employee. Another factor for employee well-being is job stress, which is an important reason to change jobs (Deniz et al., 2015).

However, research shows that not only a person-job fit is important for high performing employees. The commute or travel time from home to work, including possible traffic jams, can cause stress and burn-out symptoms (Université de Montréal, 2015). Creating an allocation optimization model that takes into account the travel distance, therefore, could improve the person-job fit and possibly optimizes store performance.

These three factors, job satisfaction, person-organization fit and travel time to work, are the main pillars of this research. They underline the importance of allocating employees with the right skills and preferences, to stores with the same characteristics while limiting travel distances.

Yet, allocating skills to stores is not easy. In general, optimization problems are widely addressed mathematical and spatial problems, for example in creating an optimal planning for the resources to treat cancer (Vieira, Demirtas, van de Kamer, Hans, & van Harten, 2018). As proposed by Zhao et al (2018), optimization can also be used to assign roles to employees (Barnhorst, Betro, & Haq, 2001) or for optimizing public transportation (Bussieck, Winter, & Zimmermann, 1997).

This study aims to combine those outcomes in a single model, to create an optimal skills and spatial allocation of employees and stores. It, therefore, aims to fill the void in both practical and academic knowledge around the optimal allocation of the right people skills to the right stores. In the three main steps, this research tries to answer the following research question, which will be introduced in chapter 1.4.

1.2. Related work

This chapter describes related literature and to what extent it is viable for our research. This chapter is split down into the main subjects of this research. Store classification, geo-demographic data analysis and allocation optimization.

The first main subject of this research is the classification and analysis of (differences) of stores in terms of sales and the environment. According to Mason & Jacobson (2007), clustering enables the data to be classified into groups, "making the data more manageable for analysis purposes, each group having a particular profile" (Mason & Jacobson, 2007), p.1), which helps to understand the differences and driving forces of a stores' success. In their research, they propose a fuzzy clustering method which allows data points to be in more than one cluster. Their conclusion is that the proposed clustering method provides the capability of applying population and distance effects into a geodemographic cluster analysis and their methodology increases spatial interactions and homogenization in the clusters. Interestingly, such distance effects are also addressed in our research for optimizing travel times for store employees to stores.

Other researchers propose different clustering techniques. Such as, Holy, Sokol, & Cerny (2017), who propose a methodology to cluster retail products of customers in a drug store by using the K-means algorithm for a Market Basket Analysis, for a dataset of 10.000 shopping baskets. A Market Basket Analysis is "a field of modelling techniques based upon the theory that if you buy a certain group of items, you are more (or less) likely to buy another group of items" (Tatiana & Mikhail, 2018). Holy, Sokol, & Cerny (2017) use the Market Basket analysis to classify products into clusters according to their occurrence in the same shopping routine from a customer. Such methods are used to optimize decisions for customers in a retail context such as the placement of products in a store (Valle, Ruz, & Morrás, 2018). Our research has a case study for drug stores but is not particularly interested in market baskets. However, since Valle et al. (2018) discuss the product placement question that can be addressed using such techniques, similar to determine the location of products in stores, customer choice and experience can be influenced by the placement of people -staff- in a store. Hence, it could be explored if the methods used to assess product placement could be transferred to people placement as well. Our research will aim to do so.

The K-means clustering method is one of the most commonly used methods according to Holy et al. (2017) and Jain (2010). The K-means methodology assigns each data point to a single cluster, called a hard assignment (Jain, 2010). Jain (2010) elaborates on different clustering methodologies and discusses their main challenges. He concludes his research with the statement for tighter integration of the clustering algorithms used and the application needs. Which, for our research means that each store should not be fuzzy allocated so stores can belong to different clusters, but should be allocated to a single cluster and thus K-means clustering is a viable option.

Pan & Zinkhan (2006) researched predictors that are related to shoppers' retail choice. Their research, in which previous empirical studies are reviewed, showed that a stores' selection of products is the most important driver for a potential customers' choice and that both services

of store employees and store location have a high correlation with store choice as well. They also found that demographic data is a good predictor of shopper frequency. Thus, for our research, since the location is considered as an important driver for a stores success (Pan & Zinkhan, 2006), this research should take geo-demographic data of store service areas into account. Similarly, Carpenter & Moore (2006) aim to provide a general understanding of a population of 454 grocery customers in the United States for a certain store. In their research, all characteristics and choices of consumers are statistically tested for significance. It turned out that for some shopper groups the product selection of a store and the courtesy of store employees are of significant importance (Carpenter & Moore, 2006). Those methods will be implemented in our research for testing differences in socio-demographic variables of service areas of stores. Our research is not particularly interested in customer behaviour. However, the characteristics of potential customers in the service area of a store will influence the local store Carpenter and Moore (2006).

For calculating such service area of stores, Dramowicz (2005), in her research introduces the Huff model for calculating service areas. Her research explains the parameters of the model, in which the distance of customers and the attractiveness of the stores are important. Remarks are made that attractiveness is measured as an attribute as square footage. Without changing the model, other quantitative attributes such as turnover could be used to measure attractiveness, and thus determine the size of the service area. Therefore, this model can be the start of determining service areas in our research.

The geographical characteristics of the service areas can be used for clustering, as introduced by Mason & Jacobson (2007). The analysis of geo- or socio-demographic data (Geo-Demographic Analysis (GDA)) is defined as 'the analysis of spatially referenced geo-demographic and lifestyle data' (See & Openshaw, 2001, p.269). This is in line with our research, namely to understand the socio-demographic characteristics of the service areas of stores.

Understanding the geo-demographic characteristics of service areas allows retail companies to start localization (Rigby & Vishwanath, 2016). Rigby & Vishwanath (2016) study the shift from standardization to the localization of retail chains. However, the important takeout of this work is that localization is expensive. Shifting to localization needs clustering based on geo- and socio-demographic data to create store types that suit stores in comparable locations (Rigby & Vishwanath, 2016).

The second main subject of this research is the optimization and matching of a store with certain skill needs with the team of employees with the most comparable skills. A well-known example of an optimization model is the travelling-salesman problem which by definition "asks for the shortest tour through all vertices of a graph with respect to the weights of the edges" (Oswin et al., 2017, p.521). The travelling- salesmen problem is used to calculate the most optimal routes between X number of stores with Y number of agents. Adding an extra city to the travelling-salesman problem the number of possibilities increases exponential (Senthilkumar, Nallakaruppan, Chandrasegar, & Prasanna, 2014). In their research, Senthilkumar et al. (2014), address the problems of a travelling-salesman algorithm both in

terms of quality and model run time. In their study, they modify the genetic algorithm to improve the performance and quality and apply it to a synthetic data set of nodes. Their conclusion is that a travelling salesmen problem can be successfully improved both in terms of runtime and tour lengths. The research for optimizing distance by Senthilkumar et al. (2014) is comparable to the optimizing distance in our research since both methods do not allow duplicate allocations in the model. Therefore, in our study, we explore the usage of optimizing distance in the context of assigning employees to stores.

Current literature describes assigning people to locations with location-allocation models for different purposes, such as multi-period location-allocation for nursing at home (Khodaparasti, Bruni, Beraldi, Maleki, & Jahedi, 2018). In their study, Khodaparasti et al. (2018) propose a model for locating mobile nursing facilities. Their model involves accessibility and deals with changing demand over time and location. The research aims to forecast the demand for nursing homes to be able to improve the service level on a given time and location. Our research is a one-time allocation of employees. However, the optimization of demand and supply in their research is interesting for optimizing the optimal amount of employees to a store.

Gokbayrak & Kocaman (2017) in their research introduce a new location-allocation problem where the studied facilities, such as power plants, have a fixed opening cost and coverage distance is limited. They apply their distance-limited model to energy and water networks. For our research, the maximum coverage distance of the Gokbayrak & Kocaman (2017) is interesting, since maximum distance will be limited between a store and an employee to allow allocation. In doing so, Gokbayrak & Kocaman (2017), they added a constraint that limits the distance to their model.

Integer Linear Programming (ILP) is widely used for optimization. Kilci, Kara, & Bozkaya (2015) used ILP to locate temporary shelter areas after an earthquake in Turkey. They improve the original model of the Turkish Red Crescent (Red Cross) by adding distances between districts and shelter areas to the model. And take distances to roads and health supply locations into account. Their model is tested using 3000 data points from Istanbul in Turkey and then used for a case study for an earthquake in 2011. The model manages to come up with an exact solution in less than a second. The model of Kilci et al. (2015) thus allocates people to locations, taking into account distances, and decides whether to open a location based on 10 criteria. Our research aims for optimizing distance but does not have to decide to open a location, since each location/store will be opened. Important in our research is the optimization of skills supply of the employees and the skills demand of the stores. The supply and demand of such skills can be seen as a capacity for modelling, which is addressed by Paul & Batta (2008). ILP is thus proposed as an optimization method in this research.

Paul & Batta (2008) in their research aim to optimize the capacity supply and demand for hospital facilities after a natural disaster as effective as possible. They consider them as optimal when supply minus demand is as close to zero as possible. A surplus of supply means empty hospitals, where a surplus of demand means a need to build extra hospitals or re-allocate existing hospital capacity (Paul & Batta, 2008). Next to effectiveness, Paul & Batta

(2008) also address two potential methods regarding computational power that can be of use for this research. In their research, they propose Heuristic solutions and small size tests. The Heuristic methods are used to give a near-perfect solution without extreme use of computational power. The other solution is using a small-size problem to test whether the model is viable. Both methods can be considered for our research to decrease the use of computational power if the model, because of complexity, is too slow to run for all stores and employees. Therefore, in this study, we explore the usage of ILP modelling in the context of assigning employees to stores with the aim of optimizing effective skill allocation.

As proposed earlier, service of store employees is an important driver for choosing a store (Pan & Zinkhan, 2006). According to Bruecker, Bergh, Beliën, & Demeulemeester (2015), a group of employees is often very heterogeneous and a manager should not just employee preferences constraints into account, but skills that workers possess as well. Skills are defined as "the ability of a worker to perform certain tasks well" (Bruecker et al., 2015, p.1). Bruecker et al. (2015) in their research further point out that the main requirement for optimal skill allocation is to classify skills. They show options to measure skills by stating the link between skills and employee background and education. Yet, the work only subtly refers to travelling distance from home to work, since the research is about allocating skills within a single store, instead of all stores of a brand. This research thus lacks a broader consideration of distance, which can and will be covered for our research based on other literature such as Kilci, Kara, & Bozkaya (2015) and Paul & Batta (2008).

To conclude, clustering is thus used as reviewed by Holy et al. (2017), who used clustering for performing a MBA and Valle et al. (2018) who used it for analysing product placement, which for our research is explored to use for people placement. The importance of the service area of a store is addressed by Pan & Zinkhan (2006), who concluded that a stores product selection is the most important driver for a potential customer to choose for a store, after which employee service and a stores' location are most important. Optimization in our research is covered using ILP as proposed by Kilci, Kara, & Bozkaya (2015) who focusses on distance and Paul & Batta (2008) who focusses on effective capacity optimization. Lastly, skill allocation for our research is covered by Bruecker et al. (2015).

All in all, our research uses a combination of scientific work as explored in the chapter above and is unique due to the combination of skills- and allocation optimization including distance effects.

1.3. Case study

This research will examine and test our methodologies using an actual case within a company. This company is a nationwide brand that operates of drugstores, which is one of the brands of a global retailer. Within this company, there is a separate HR Analytics department which uses an academic mindset to make the practice of management more rigorous. Modelling and researching HR is exactly what the department HR Analytics does to take data-driven decisions (Rasmussen, Ulrich, & Likert, 2015). HR must deal with many data sets that have a geographic component because both the stores and all employees have a location.

Some of these problems can be considered location-based problems. For example, employees and the route and distance they have to travel to their work every day. Or challenges in planning employees with the right skills in the store with the right characteristics at the right time.

The size of an organization partly determines the volume of information, or data, available. For HR, this implies that a larger headcount could be associated with a larger volume of data. The volumes of those data can become problematic to be analysed by humans. To get insights into these data, data analytics can be used. For example, clustering algorithms can be used to create groups of similar stores with sales or environment data as variables (Nerurkar, Shirke, Chandane, & Bhirud, 2018). Those models cannot only be used for the case study since they can be changed to be suitable for other organizations.

1.4. Research questions

This research focusses on the optimal allocation of employees with a specific set of skills over all stores of a brand. It takes store type, employee skills and commute distance into account. The main question that this research focusses on is:

"How can employees with specific skills be allocated to stores with specific (geographic-) characteristics in a way that both store team composition and individual travel distances are optimized?"

To answer the main research question, four sub-questions are used, which are:

Sub-questions:

1. *What data sources and methods can be used for store clustering? (step 1)*
2. *How can employees be rated for different skills and preferences? (step 2)*
3. *How can store-specific services areas be calculated and used for verifying sales data as used in step 1? (step 3)*
4. *What location-allocation method can be used? (step 4)*

1.5. Research steps

This research has four steps, which all try to help to create more suitable, more locally relevant, employee teams for each specific store, based on a match between employee skills, store characteristics and the characteristics of the service area of the store. Figure 1.1 shows the four research steps. Step 1, 2 and 4 form the main research, where step 3 focusses on the options to improve the research in steps 1 and 2. Step 3 is thus linked to those steps with a dotted line.

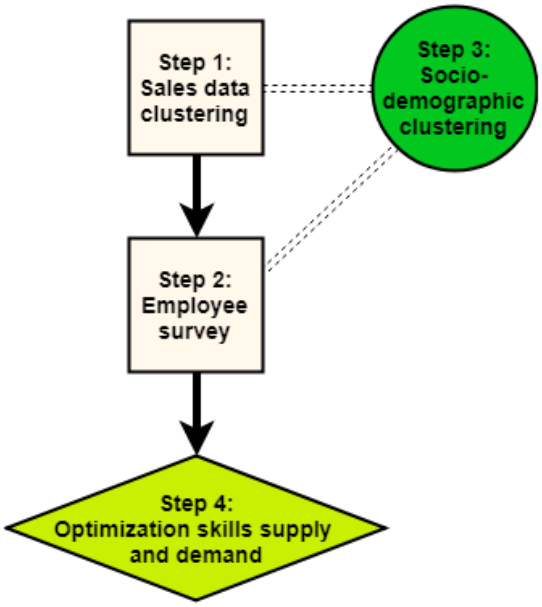


Figure 1.1 Research steps

In Figure 1.2 the situation without the skills and preferences of the employees or the store characteristics known is schematically shown. Store employees are thus linked to stores without knowing the store characteristics and employee preferences, there is no differentiation. Same goes for the service areas of each store, which is the same for every store.

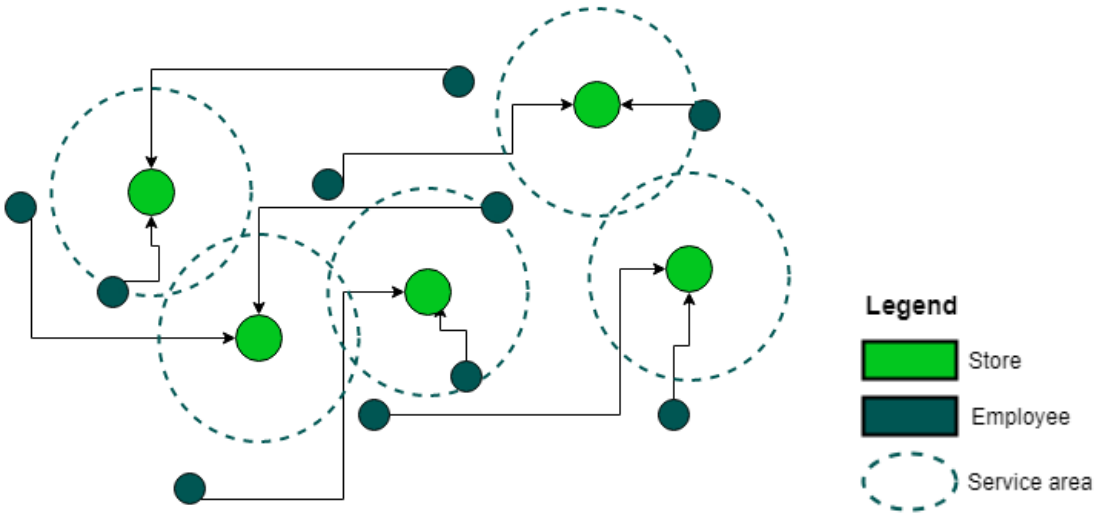


Figure 1.2 Schematic model current situation

The first step is to create clusters of stores with comparable customer profiles based on stores sales data. Classification and clustering options will be discussed and reviewed to get to the best clustering method. More about the considered options and the findings will be discussed in chapter three, methodology and chapter four, results. After this clustering, the model changes to the model in Figure 1.3. With this step, store characteristics are made clear to differentiate stores. Nothing is changed on the employee side and the service areas.

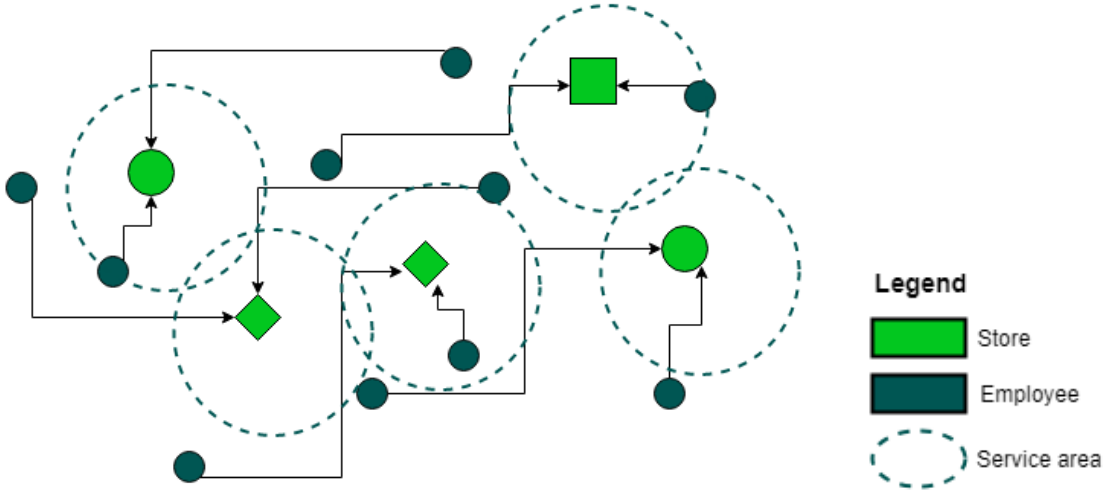


Figure 1.3 Schematic model after completing step 1

The second step aims to research the skills and preferences of employees by doing a survey. This does not involve changes in the allocation of employees to stores but allows differentiation in the preferences of employees. This step is visualized in Figure 1.4.

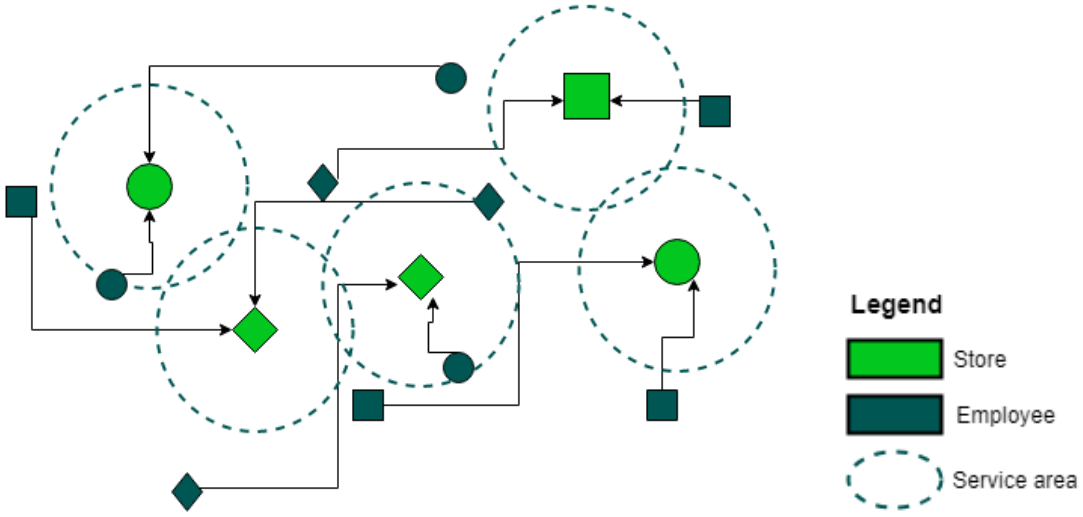


Figure 1.4 Schematic model after completing step 2

The third step is creating store-specific service areas using a model as explained in chapter 3. For those service areas, characteristics will be calculated such as average age and average income. Those calculations are based on socio-economic and demographic data, which will be further discussed in chapter 3, methodology. The result is shown in Figure 1.5. The characteristics from the service area will be used to verify if the cluster that a store is given in step 1 is legit or that the sales data might be biased, for example by current employees that influence the sales for a specific product category.

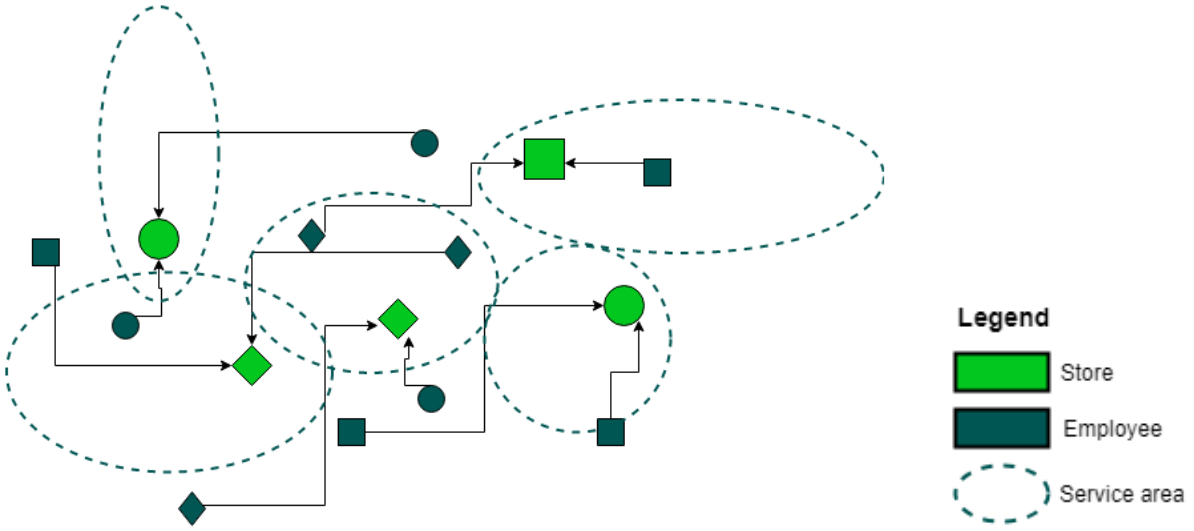


Figure 1.5 Schematic model after completing step 3

The fourth and last step is the location-allocation of store employees. This analysis aims to allocate employees with the right skills to the store with the same specific needs. This analysis does not consider the business problems of re-allocating store employees since it is a mathematical optimization. The main parameter is an optimal match of employee skill and store demand. Travelling time or distance for each individual store employee is used as a variable in this model. After this location-allocation step, the model will look like the one in Figure 1.6. This figure is a schematic model, which is not exact representation of the real world. However, it is seen that most lines are for commute distance are shorter than in Figure 1.5. Some are longer, because of the overall optimization, but will be limited to a maximum commute distance. The surplus that remains after matching and allocating employee preferences with store characteristics shows the room for improving the teams with extra skills, for example by additional training for specific employees. Optimal allocation minimizes this surplus.

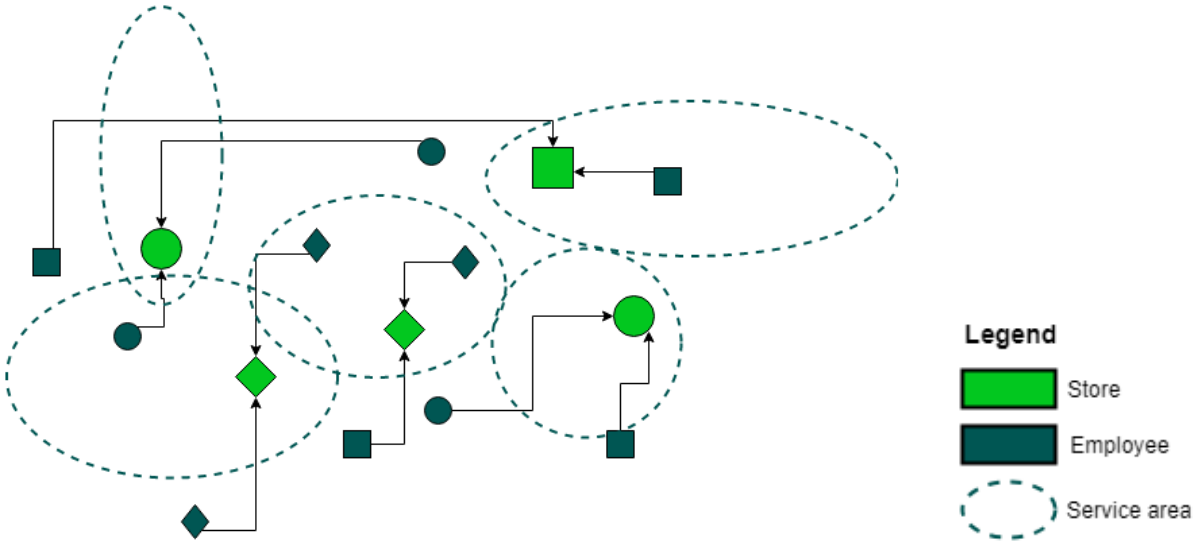


Figure 1.6 Schematic final model after step 4

1.6. Reading guide

After the first chapter which contains an introduction to the topic, a literature study and the research steps, the methodology will follow in chapter 2. Chapter 2 describes the overall workflow of the research and all steps are described in depth, together with the used data, tools, techniques and a schematic workflow for each step.

The first step is clustering stores by using the K-means clustering algorithm based on sales data, to gain insights and understand the data and its main drivers. This first step is correlated with step 3 in which the calculation of a service area for each store is included. For this service area averages of the socio-demographic characteristics are calculated and stores are then clustered with this data to be compared with the sales data clustering from step 1. This comparison should give insights into the suitability of the store in its service area, based on its sales data. Step 2 is a survey amongst store employees to measure their skills, preferences and background. This data will be used to optimize in the last step. Lastly, step 4 uses Integer Linear Programming (ILP) to solve the optimization of allocating the right store employees to the right store.

Chapter 3 contains the detailed results and findings for every step after which a conclusion follows. This report closes with a discussion and recommendations for future research, to give an overview of this research in the bigger, societal, context and to propose possible next steps in the research.

2. Methodology

2.1. Introduction

As described in chapter one, this research will exist of four major steps. Those steps, including what data and methodology are used will be described in this chapter and were schematically shown in Figure 1.1. In the first step store characteristics are made clear using clustering, the second step lists preferences of employees with a survey. Step 3 researches the options to further improve step 1 and 2 by using socio-demographic characteristics. Lastly, step 4 is the optimization of sales data and employee survey data using the creation of an optimization model.

The steps are described more in-depth in Figure 2.1. Stores are clustered based on sales data and using socio-demographic data of the service areas of the stores. The input sales data for clustering is also used for optimization, together with the data that is gathered using the employee survey. The same data is used to perform a reference skills match which allows calculating to what extent the optimization model improves the allocation of skills.

Clustering is done to get insights into the population of stores. This shows the main drivers of differentiation which are used to determine the product categories that are important to extra research in the employee survey.

With their delivered services, knowledge and advice, the current allocation of employees is expected to have an effect on the sales data. To research whether this bias occurs, this research tests to compare the sales data cluster with the characteristics of the service area of a store. Clustering based on the socio-demographic characteristics of the service area of a store is thus done to check if a sales pattern of a store matches the service area characteristics. This check will be done by visualizing the distribution of sales clusters over the service area clusters. Lastly, the ultimate step will be to determine if the differences in skills match can be explained with the service area of a store.

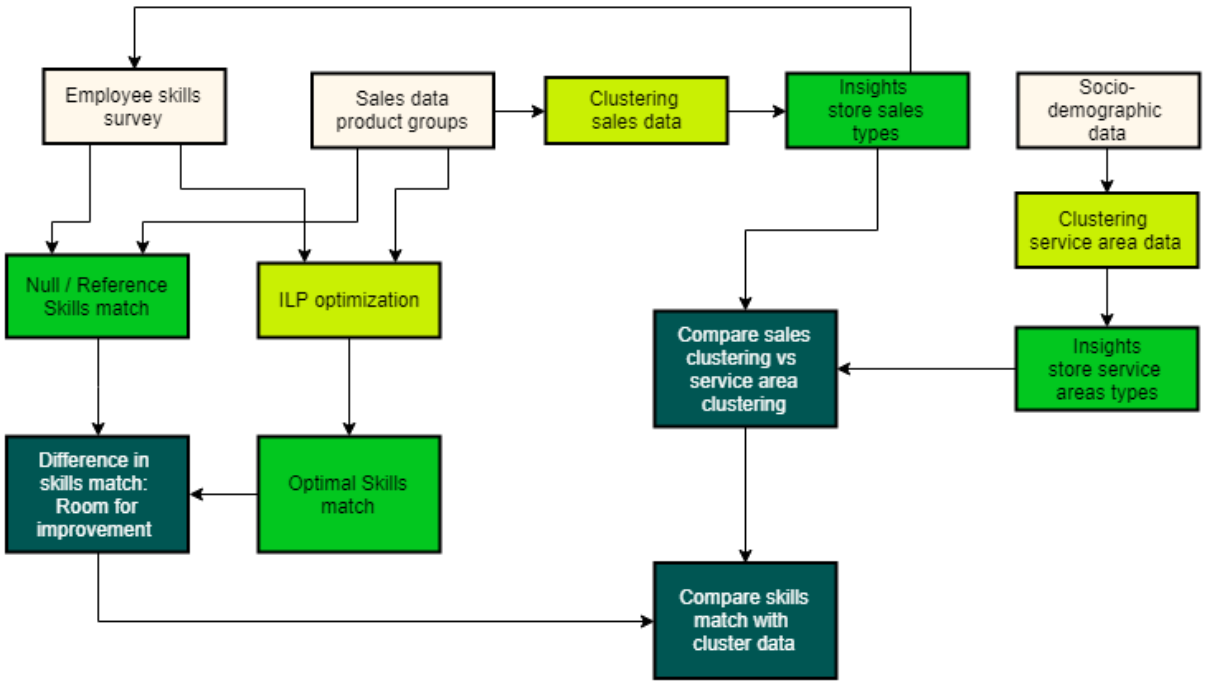


Figure 2.1 Model steps

2.2. Step 1: Clustering with store sales data

The aim of this step is to assign clusters to comparable stores, based on sales data. A cluster is defined as a set of objects which have a higher degree of similarity to each other compared to objects not in the same set (Nerurkar et al., 2018). For example, stores in which the product classes of 'health-related products' are more dominant than in other stores are clustered as 'health stores'. Clustering enables retailers to start localization of their stores to customize them to suit the local customer, without having to develop a customized format for every store (Rigby & Vishwanath, 2016). Clusters are also used to check if the sales pattern of stores from the same cluster have service areas with comparable characteristics, which will be described in-depth in step 3.

Different techniques of clustering stores exist, such as partitional clustering, hierarchical clustering and density-based clustering. With partitional clustering, data is divided into non-overlapping subsets in a way that all data points are assigned to exactly one subset (Wong, 2015). Every store should be assigned exactly one cluster, which makes partitional clustering suitable. K-means is a partitioning method where all data is linked to a centroid of a cluster, which are called seeds. Seeding is done randomly which makes this method vulnerable. Another partitional clustering technique to solve the random seeding problem of k-means is k-means++. With k-means++ seeds are chosen as dispersed as possible, which improves the Rand Index (Wong, 2015). According to Deepashri & Kamath in Figure 2.2, clustering is a descriptive method. The k-means method will be used for all clustering in this research because all stores should be assigned to a single cluster. The random seeding problem of the regular K-means clustering method is taken into account.

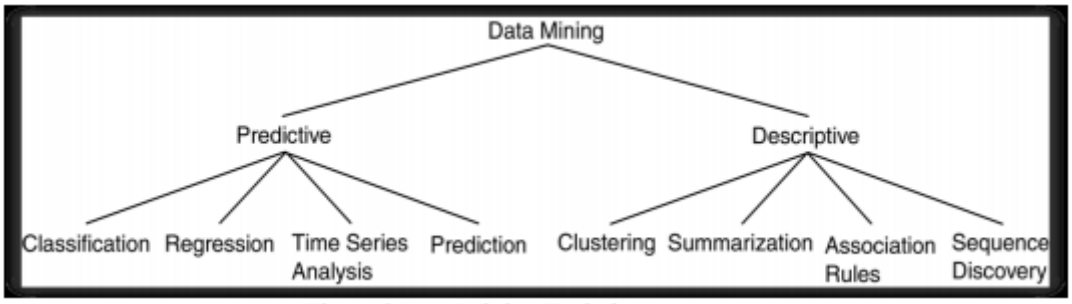


Figure 2.2 Data mining, (Deepashri & Kamath, 2017)

Clustering is done based on store sales data and based on socio-demographic data such as income and age. More about the socio-demographic data can be found in chapter 3.3. Data preparation and clustering will be done using the predictive grouping tools in AlteryX Designer software and using the scripting language R.

Figure 2.3 shows the steps a K-means clustering algorithm takes. The centroid changes after each iteration until convergence has been reached.

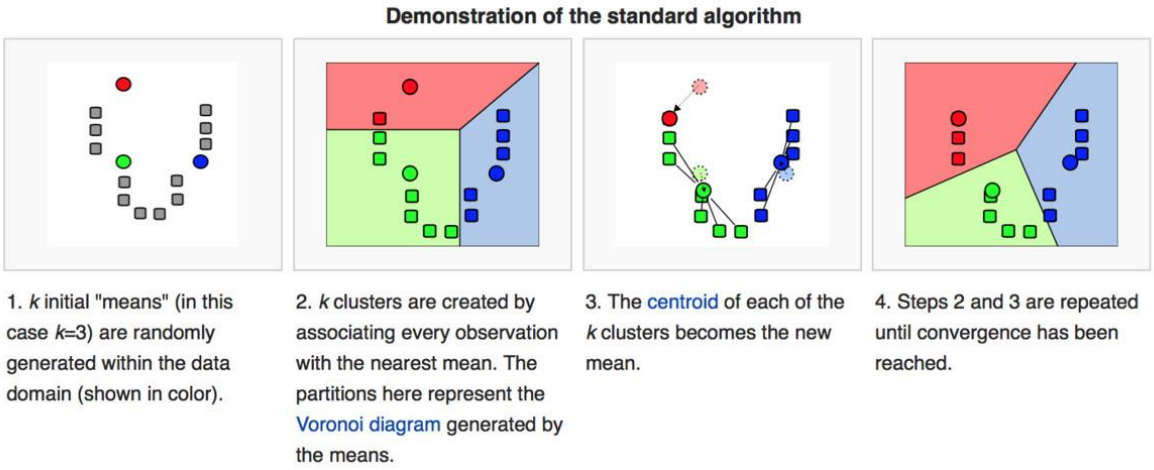


Figure 2.3 K-means clustering method (Pace, 2007)

The optimal number of clusters that will be created can be decided by running the model over with a different number of clusters and then evaluate the results. According to Dolnicar (2003), choosing the number of clusters is not an exact science. In AlteryX the K Centroids diagnostics tool generates an output which contains a boxplot of adjusted Rand indices which can be used to determine the best number of clusters (Zhao et al., 2018). Other, more rigorous, mathematical methods to estimate the optimal number of clusters are the 'Average silhouette' method and the 'Gap statistics method'. The average silhouette method computes the average silhouette of observations for different numbers of clusters. The output with the maximum average silhouette over a range of a possible number of clusters marks the optimal number of clusters (Gentle, Kaufman, & Rousseuw, 1991). The gap statistics method compares the total change in within-cluster dispersion with that expected under an appropriate reference null distribution to estimate the optimal number of clusters (Hastle, Tibshirani, & Walther, 2001). Both methods will be used in this research. The best method is decided by consulting internal experts from within the case study organization.

The workflow that is created for clustering based on sales data is shown in Figure 2.4. Input data sets for this workflow are sales per product group per store and input from internal experts for reclassifying product groups to less and more general groups. Correlation analysis for the attributes is part of the research to check if the expert input for grouping is robust. Apart from those data sets, research for determining the optimal amount of clusters is an input as well.

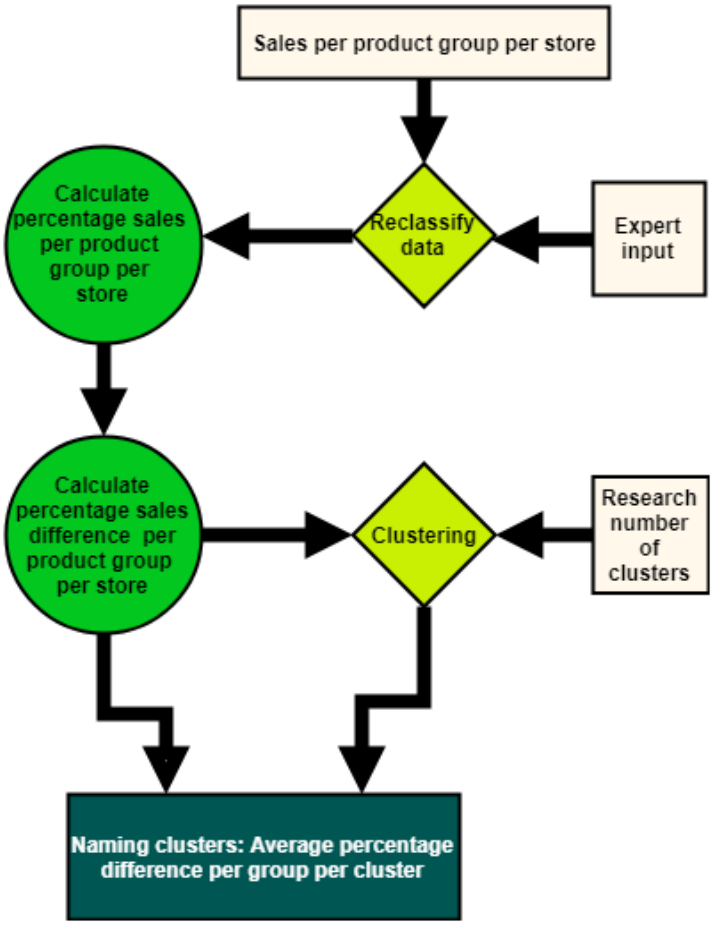


Figure 2.4 Workflow for clustering

A percentage of sales, in volume, for every product group per store is calculated, as well as the percentage difference from the average per product group over all stores. Then the clustering is done using AlteryX and R, which returns an output with a cluster value (ranging from 1 to n clusters) for each store. Stores with the same cluster value are most similar, but all clusters have to be named with a name that is representative for the cluster.

An average of every cluster variable of all stores in a cluster is calculated. The product groups that have the highest percentage are performing the best and thus are used to name the cluster.

To help support the cluster naming decision a correlation matrix is created, which shows the correlation between two variables in a value between 0 and 1. Where 0 means 'no correlation at all', and 1 means 'full correlation', which is the correlation coefficient of the variable with itself.

2.2.1. Normalizing sales data for shelf space

Not every store of a brand has the same percentage of shelf space available for each product group in every store. This research, therefore, tests whether it is possible to normalize for this possible bias by calculating an average sales per shelf unit. The formulas which calculate the sales per shelf unit for each product category for each store are formula 2.1 and 2.2.

Indices:
j: Index for a product group/attribute, in total there are l groups/attributes
i: Index for a store; in total there are m stores

Constants:
U: Sales per shelf unit
T: Sales per product category
P: Shelf space per product category
C: Cluster value store

$$U_{i,j} = \frac{T_{i,j}}{P_{i,j}}$$

Formula 2.1 Shelf space

Not all stores have the same average sales per shelf unit. To calculate comparable cluster values for each product category for each store, the following formula is used.

$$C_{i,j} = \frac{U_{i,j}}{\sum_{j=1}^{l} U_{i,j} \ / \ l}$$

Formula 2.2 Cluster value

With these formulas, it is tested whether it is possible to normalize the cluster values for the amount of shelf space that is available for a product category in a store, and for the average sales for a store. This method may make values comparable and suitable for clustering.
Example data is given in the tables from Appendix 1. If this was real data, store 5678 will be clustered as “health”, where store 1234 will be given the label “Baby” or “Moderate”. Clusters might also be given names by the location they are mostly in. For example when a specific sales pattern occurs for stores in city centres.

2.3. Step 2: Employee survey

2.3.1. Survey content

To allocate employees to the best suitable store, data about the characteristics of the employees are needed. With a survey, data about the geographic location (postal code 4), background, skills and interests of the employees per product category will be gathered. The location of residence of each employee is used to calculate distances and travelling time between the employee and all stores. This travelling distance and time is used as a variable in the optimization model of step 4.
Sampling is considered for this survey. In doing so, a sample from the whole population of stores has to be taken carefully. A too large sample may be a waste of resources, where a too small sample might not be representative of the whole population (Columb & Stevens, 2008). All clusters must be represented equally. Sampling can be done in many ways. The simplest option is to randomly choose a fixed percentage of all stores. This can cause over- and under-representation of certain clusters. Another option is to take a fixed percentage of each cluster, which causes a more even spread of the samples. However, a small cluster might be biased,

because the percentage of all values can be just one sample in the smaller clusters. The option to change a fixed percentage of each cluster to a fixed number of samples of each cluster will fix this problem. However, this raises another problem, which makes small clusters over-represented, compared to bigger clusters. Better options may be sampling through a more scientific approach, such as DENDIS, among others, as proposed by Ros and Guillaume (2016). All in all, there is no single 'all-time-best' solution.
However, for this research, the aim is not to take a sample but conduct the survey for all employees. The main aim is to make sure that every employee from a store participates in the survey, instead of many stores with just one or two respondents. The survey is targeted for all employees of non-franchise stores, which are around 290 stores, with around 3000 employees. More details of the data can be found in chapter 3.1.

2.3.2. Skills and preference measurement

Where possible, the survey in this research aims to measure skills and behaviour indirect instead of direct. The way questions are formulated contributes to the expected response. Questioning people about future behaviour has an impact on this behaviour, which is called the question-behavior effect (Sprott et al., 2006). This effect is an example of a bias occurring when conducting survey research. Those biases will be mitigated where possible.
This research does not aim to change behaviour but to get to know the intrinsic motivation of the employees for all clusters and different product categories. If it is possible to decide on quantitative measurable skills for store employees, a skills assessment form as proposed by Winckel, Reznick, Cohen, & Taylor (1994) can be implemented as part of the survey.

2.3.3. Collaboration Utrecht University

The MSc GIMA programme has no direct links with social-sciences, surveying and setting up scientific correct and valid questionnaires. To be able to gather proper data in this research about the skills and preferences of employees, a collaboration with Utrecht University is started, with professor Maarten Goos and PhD candidate Ronja Röttger (2018). In collaboration with Röttger, a questionnaire to proper measure skills and preferences has been set-up. The created questionnaire is pasted in appendix 4.

2.4. Step 3: Clustering with socio-economic and demographic data

Stores have a specific environment with specific demographic characteristics. Those characteristics have an influence on the store's sales, for different product groups. Data about the store's environment is used in this research to verify whether the given cluster in step 1 is expected for the store.
Socio-economic and demographic data can be used from CBS ('statistics Netherlands'), which is available mostly on an aggregated level such as neighbourhood (Statistics Netherlands, 2018). More detailed data is available as commercial data. An example of such a data set is the one from Whize (2018) customer segmentation. This data set contains, among others, on a postal code 6 digit level, the income level, the household size, education level, stage of life (children, single and family) and social class. The used variables and classes are given in

Appendix 8. The number of inhabitants that will be selected to create a service area is linked to the turnover of the store. A store with more turnover is considered to have a wider range of service than a smaller store with less turnover.

The model in Figure 2.5 selects people that live the closest to the store, on a 6 digit postal code (PC6) level, or a PC4 level if more aggregated data is desired. For this area, an average of the used data is calculated to compare and cluster stores. Calculating the service area is done using the spatial tools set available in AlteryX Designer.

The clusters assigned to stores in step 1 are used to research whether it is possible to detect differences in characteristics of service areas and the clustering of stores on sales data. For example, comparable service areas, but deviant sale patterns. A change in allocated employee skills, as conducted in the fourth step of this research, might change the sale pattern positively, and enhance the store to suit the location better.

Tests will be done if the clustering of stores based on the characteristics of their service areas can be used to categorize them. The outcome of this clustering is compared with the clustering based on sales data. Then, a comparison of the socio-demographic clusters over the sales clusters will be made. This shows whether there is a correlation between the sales pattern and the location of the store.

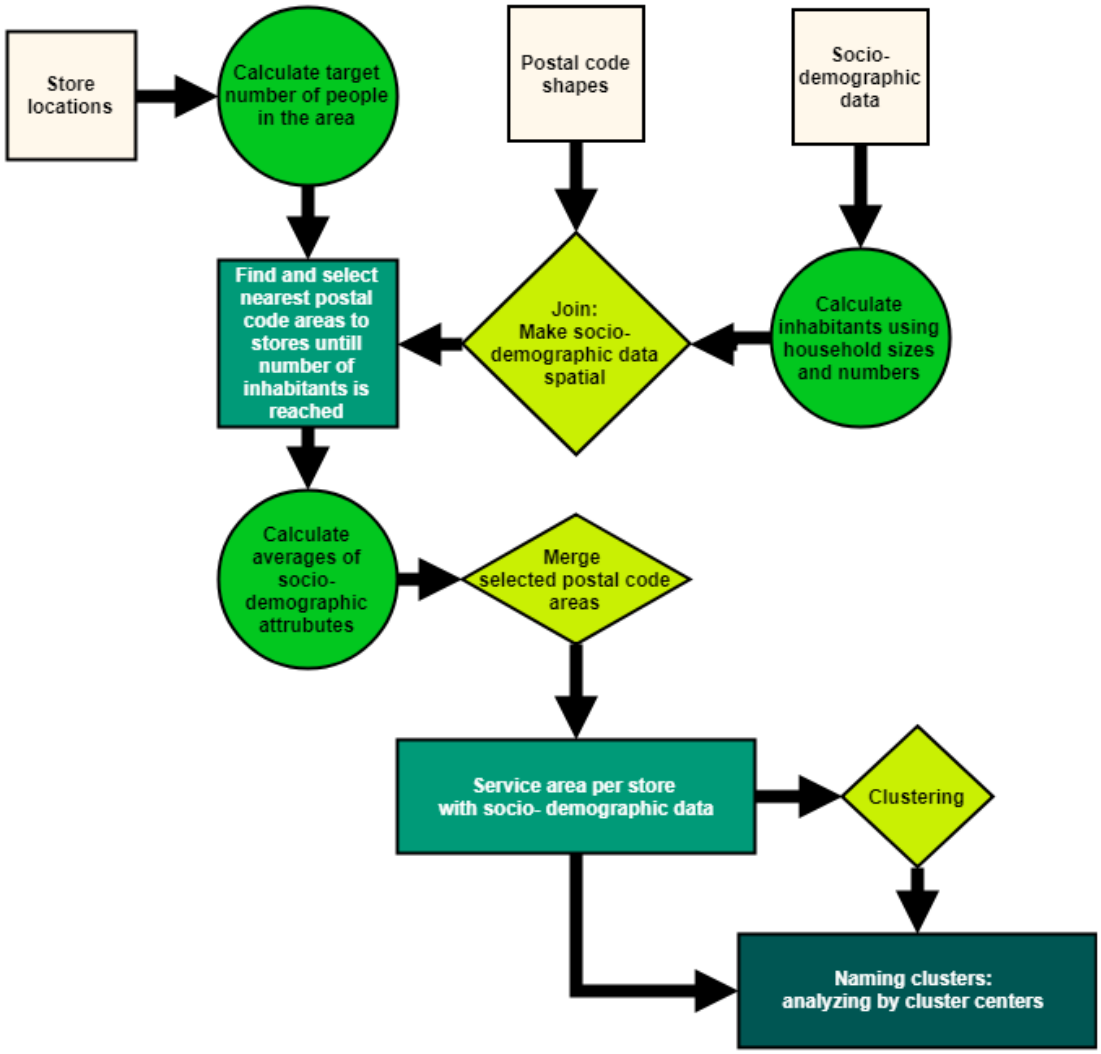


Figure 2.5 Workflow calculating service areas

2.5. Step 4: Employee-store allocation optimization

The fourth step of this research aims to allocate employees with a certain set of skills to the correct stores, with the same specific needs. Previous empirical research shows that optimal allocation will make employees more satisfied about their job, and more productive, because they work in a store where they can use their skills optimally (Yuen, Loh, Zhou, & Wong, 2018). As a positive external effect, the travelling times will be minimized where possible while keeping the optimal team composition.

In the previous steps of this research, we clustered stores and rated employees. Using the sales data and the survey data we assign "points" to each store and employee to determine skill supply and demand. A store with a higher turnover will need a higher summed amount of skills than smaller stores. On the other hand, an employee who works more contract hours than a colleague probably can deliver more skills. To estimate the amount of skill supply and demand, formulas 2.3 and 2.4 are used.

Indices:
j: Index for a product group/attribute, in total there are l groups/attributes
i: Index for a store; in total there are m stores
k: Index for an employee; in total, there are n employees

Constants:
S: Skill supply or demand
C: Cluster value store
R: Rating for an employee for a product group
T: Turnover index for store
H: Contract hours per period of an employee
Q: Coefficient for levelling overall skill supply and demand

$$S_{i,j} = C_{i,j} * T_i$$

Formula 2.3 Store skill demand

$$S_{k,j} = R_{i,j} * \frac{\sum_{j=1}^{l} R_k}{H_k}$$

Formula 2.4 Store skill supply

In the final allocation model, the total points of the store's needs, needs to be equal to the number of assigned points to employees, because the total of stores has the right number of employee contract hours. This is done by indexing all employee values with a constant value, Q, which is determined using formula 2.5.

$$Q = \frac{\sum_{i=1}^{m} \sum_{j=1}^{l} S_{i,j}}{\sum_{k=1}^{n} \sum_{j=1}^{l} S_{k,j}}$$

Formula 2.5 Levelling skill supply and demand coefficient

In doing so, the total sum of skill supply and demand is equal. However, the spread of skill supply over the different clusters can differ from the skill demand.
After these steps, the "scorecards" of both the stores and the employees should be allocated as optimal as possible. This involves calculating distances between stores and employees, which can be done using the spatial tools set in AlteryX Designer. Most importantly, it takes an optimization algorithm. Peng, Wolff, & Haunert (2017) in their research use Integer Linear Programming for solving an area aggregation sequence problem. They state that Integer Linear Programming is relatively easy and thus suitable for solving new problems. The model for optimization in our research can be created by using the Integer Linear Programming method. The result can be as schematically shown in Figure 2.6 for two stores and four employees.

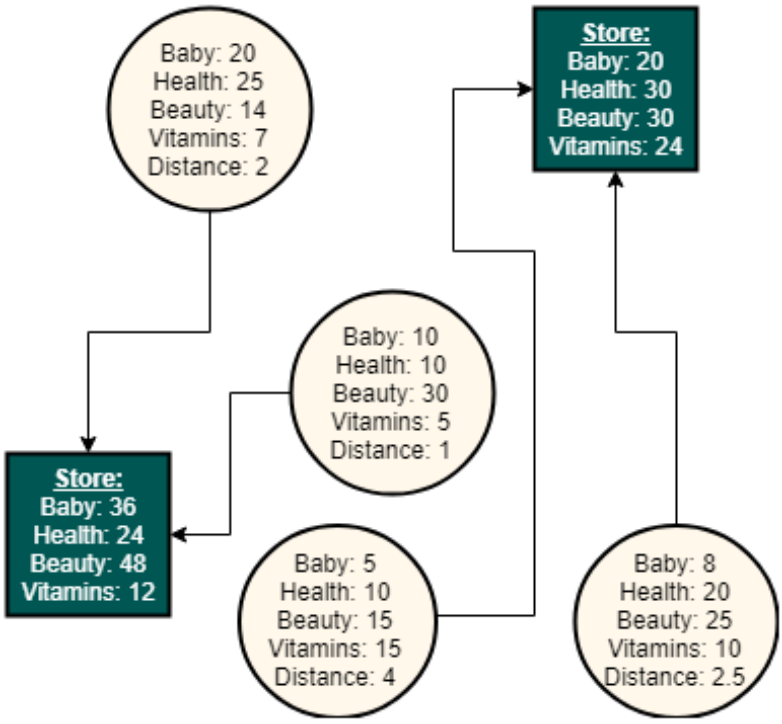


Figure 2.6 Schematic allocation result

The calculation for this example is made in Table 0.1 in appendix 3. The total number of store points and employee points is the same. The value of both stores together for health is negative and for baby positive. This means, overall, there is not enough knowledge about baby products and too much about health products. The hypothetical travel distance from employee to the store is given as an attribute. This value will be used in the model to prevent that employees are allocated on a really remote store. By setting a higher or lower cost constraint for every distance unit that has to be travelled, the importance of distance compared to the skills match can be changed.

2.5.1. (Mixed) Integer Linear Programming

The optimal allocation model will be made using the (mixed) integer linear programming technique. A (Mixed) Integer Linear Programming model consists of an objective, constraints and bounds (Risbeck, Maravelias, Rawlings, & Turney, 2017). The objective for the optimal allocation model for this research will be "minimizing the left-over from the store's values minus all values of the linked employee, including travelling distance". Constraints are that every employee should, and can, be linked to exactly 1 store and that all variables/preferences of an employee should be linked to the same store altogether. Boundaries need to be set for the travelling distance. With this information and techniques, the model can be developed.
Figure 2.7 shows the Integer Linear Programming workflow. The employee matrix contains the values as researched in the survey for each employee. Same goes for the distance matrix, which contains the travelling time or distance from each employee to each store. The store sales matrix contains all clustering values for each store. The objective function determines what has to be optimized, minimized or maximized, which is done by the solver.

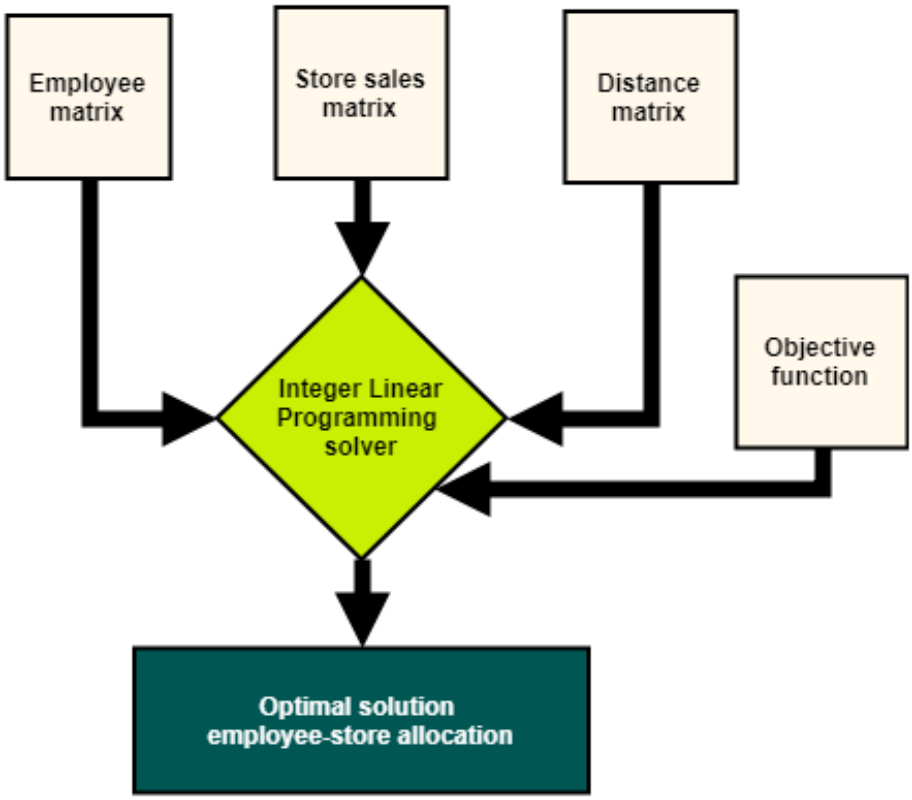


Figure 2.7 Schematic Integer Linear Programming (ILP) model

The conceptual model from Figure 2.7 is converted into a mathematical model, which is described in this chapter. The variables and other inputs from the model are denoted as follows (Peng, 2018)[1]:

Indices:
i: index for a store; in total, there are m stores
j: index for a data attribute (product category); in total, there are l attributes
k: index for an employee; in total, there are n employees

Constants:
$S_{i,j}$: Store i demands $S_{i,j}$ units of attribute j
$E_{k,j}$: Employee k is able to provide $E_{k,j}$ units of attribute j
$D_{k,i}$: The distance between the home of employee k and store i
λ: The weight ('the cost') of distance, where λ = a constant, but can be changed for testing purposes

Variables:

$$x_{k,i} = \begin{cases} 1, \text{if employee } k \text{ is assigned to store } i \\ 0, \text{otherwise} \end{cases}$$

Formula 2.6 Matrix $x_{k,i}$

$x_{k,i}$ is a binary matrix with size $k\ (rows) * i\ (columns)$.

The aim of the optimization model is to minimize the surplus of the supply and demand of employee skills. Therefore the objective function is stated in formula 2.7.

$$\text{minimize} \left(\sum_{i=1}^{m} \sum_{j=1}^{l} \left| S_{i,j} - \sum_{k=1}^{n} \left(E_{k,j} \cdot x_{k,i} \right) \right| + \lambda \sum_{k=1}^{n} \sum_{i=1}^{m} \left(D_{k,i} \cdot x_{k,i} \right) \right)$$

Formula 2.7 Objective function absolute values

The summed surplus of the supply and demand is between absolute signs. Why do we need absolute values? Because the function is minimizing our solution. Without absolute values, it means that a more negative outcome is better than an output close to zero. A strong negative output means that there is an over-allocation of employee skills, which means an unwanted increase in labour costs. The actual optimal output is as close to zero as possible, summed over all stores.
This objective function is not linear in terms of variables $x_{k,i}$ because it uses absolute values. To linearize this function, we introduce two more sets of variables: $y_{i,j}$ and $z_{i,j}$. Then, we can transform the above function to a linear objective function in formula 2.8.

[1] Dongliang Peng, PhD - D.L.Peng@tudelft.nl

$$\text{minimize} \left(\sum_{i=1}^{m} \sum_{j=1}^{l} \left(y_{i,j} + z_{i,j} \right) + \lambda \sum_{k=1}^{n} \sum_{i=1}^{m} \left(D_{k,i} \cdot x_{k,i} \right) \right)$$

Formula 2.8 Linear objective function

This objective function comes with $3ml$ constraints, formula 2.9:

$$\left.\begin{array}{c} S_{i,j} - \sum_{k=1}^{n} \left(E_{k,j} \cdot x_{k,i} \right) + y_{i,j} - z_{i,j} = 0 \\ y_{i,j} \geq 0 \\ z_{i,j} \geq 0 \end{array}\right\} \quad \forall i \in \{1,2,\dots,m\} \text{ and } \forall j \in \{1,2,\dots,l\}.$$

Formula 2.9 Constraints

The first constraint states that the demand of a store for a certain product category, minus the sum of the supply of all allocated employees plus the corresponding $y_{i,j}$ and minus the corresponding $z_{i,j}$ equals 0. The other constraints state that $y_{i,j}$ and $z_{i,j}$ are not negative. Those three constraints go for every combination of i and j, which leads to $3ml$ constraints.
Our mixed-integer linear program needs two more sets of constraints. The first one is that every employee is assigned to exactly one store, which goes for every k:, formula 2.10.

$$\sum_{i=1}^{m} x_{k,i} = 1 \qquad \forall k \in \{1,2,\dots,n\}.$$

Formula 2.10 Constraint employee to exactly one store

We use the second set of constraints to restrict the distances for the employees to commute. In other words, an employee should not be assigned to a store that is too far away. We require that the distance from the home of an employee to the store should be less than the distance $D_{\max}$, that is for every k and i:, formula 2.11.

$$D_{k,i} \cdot x_{k,i} \leq D_{\max} \qquad \forall k \in \{1,2,\dots,n\} \text{ and } \forall i \in \{1,2,\dots,n\}.$$

Formula 2.11 Constraint maximum distance

If the commute distance exceeds $D_{\max}$, the specific employee cannot be allocated to that specific store even if the skill supply and demand of the store and employee are a perfect match.
The mathematical model has to be changed to a program in script to be able to run the model. This is discussed in chapter 3.6.

3. Results

3.1. Data description

This chapter starts with a brief description of the data sources used and their content for this research. Next, the results of the four main steps of this research are described in this chapter. This research mainly uses data sources as provided by the case study organization as described in chapter 1.3. Table 3.1 gives a brief overview of the data that are provided by the case study organization, which includes both company data such as the number of stores and employees as well as commercial data sets such as the socio-demographic Whize (2018) data.

Table 3.1 Case study organization data information, content and size

Data source		Data size / unit
Stores		540 (including franchise)
	For optimisation	287
	Sales data	Sold units per product group per store
	Store information	Location / Turnover / Size
Employees		3070 (without franchise)
	For optimisation	2868
	Employee information	Location / Store number
Survey responses		842 (including 100 franchise)
	Without empty fields	668 (including 81 franchise)
Whize data (socio-demographics)		Postal code 6 (PC6) level of detail (see Appendix 8)

For this research, there was no full access to employee data sets such as the name or age of employees, to ensure anonymity. Table 3.1 shows different values for the number of both stores and employees in the entire data set and the number used for optimisation. Franchise stores are not taken into account for the optimisation. For employees, the number is given without franchise. The difference occurs because of recently closed, or opened, stores, with new store numbers. To mitigate errors, the employees who are not linked to a store number in the store data set are removed. This causes a different number of employees used for optimisation.

3.2. Step 1. Clustering stores - Sales data

As described in chapter 2, the stores are clustered based on sales data. Before clustering, the sales data is reclassified to combine product groups that are similar, based on internal experts. This step decreases the number of product groups from 95 to 32. Those 32 groups are then grouped into 4 main groups and 1 other group. In this research different datasets for clustering are tested to get the best result.

The first option uses the three biggest sub-groups of each main product groups. This results in a dataset with 13 variables, including the 'other' group. The R code that is used for analysis and clustering is given in Appendix 5: R code clustering. The first step of the analysis is to determine the best number of cluster based on the data, Figure 3.1, which is 5 clusters for this data set based on the silhouette method.

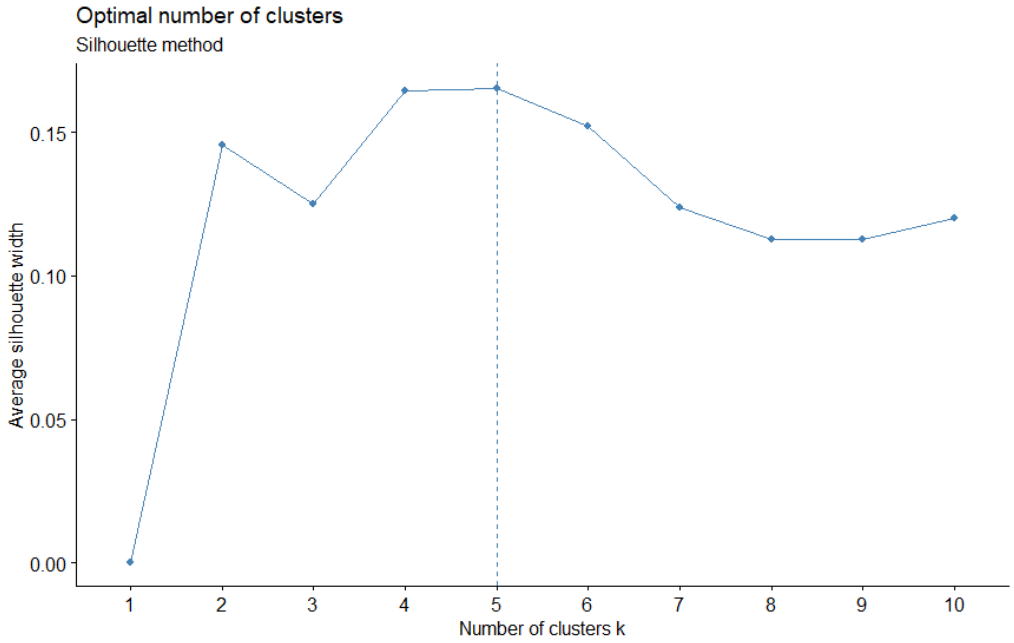


Figure 3.1 Optimal number of clusters 13 groups of data

This data is then clustered using the K-means algorithm, which results in the data in Figure 3.2.

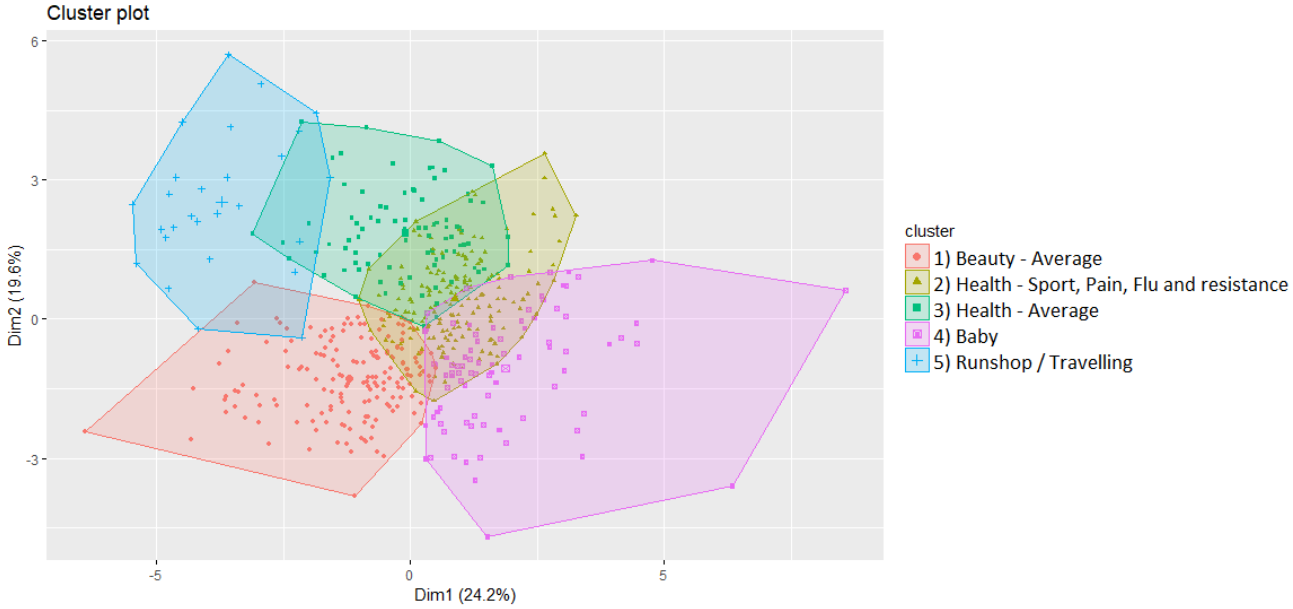


Figure 3.2 Cluster plot result, 13 groups

The output contains 5 clusters which are named by analysing the data as shown in Table 3.2 and observing which categories are, on average, the most above average. Using an average value for each cluster has the problem that an outlier, positive or negative, can be of major impact for the average of the cluster, especially for smaller clusters with fewer stores. Therefore, tests are done if counting which group is most often the highest above average within a cluster works to solve this problem. This did not work out because this method removes the detail of knowing what groups are important for the cluster as well.

Table 3.2 Average cluster values for naming 5 clusters

	Cluster				
	1) Beauty - Average	2) Health - Sport, Pain, Flu and resistance	3) Health - Average	4) Baby	5) Runshop / Travelling
Avg. Baby - Babyverzoring	-9.8	3.5	-12.1	37.5	-65.4
Avg. Baby - Babyvoeding	-18.9	-10.9	5.8	58.2	-60.5
Avg. Baby - Luiers	-10.8	-2.4	-10.0	50.4	-80.5
Avg. Beauty - Deco / Make-up	16.6	-4.4	-18.7	-2.0	-8.9
Avg. Beauty - Gezichtsverzorging	11.1	-4.4	0.7	-8.3	-12.6
Avg. Beauty - Scheren & Ontharen	4.8	-3.5	7.0	-6.2	-6.9
Avg. Care - Bad & Douche	4.8	5.5	-16.2	7.9	-40.8
Avg. Care - Deodorant	5.2	-2.9	-8.1	-2.2	18.9
Avg. Care - Haar	5.7	0.8	-9.9	0.8	-11.7
Avg. Health - Bewegen, Pijn & Griep	-16.8	25.4	-1.9	-11.7	2.2
Avg. Health - Mondverzorging	1.6	-6.6	15.9	-2.5	-10.7
Avg. Health - Weerstand, Gezin & Balans	-9.5	10.6	8.0	-5.4	-9.5
Avg. Other - Onderweg & Kassa	4.3	-23.5	13.4	-17.0	135.5

The number of stores each cluster contains is shown in Figure 3.3. Cluster 5 is small, which can be seen as not favourable for clustering, including calculating averages as explained above. However, this cluster contains stores on public transport facilities, which is a separate cluster.

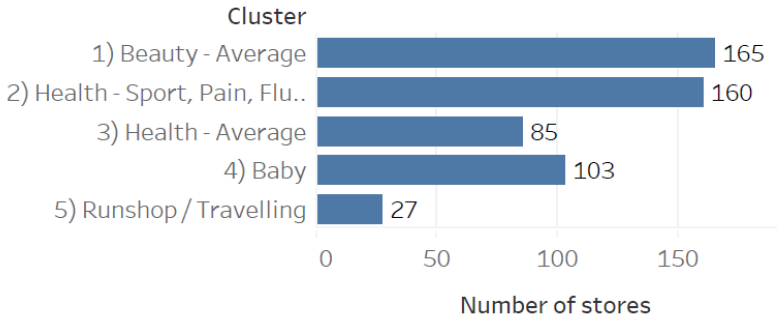


Figure 3.3 Cluster sizes 5 clusters

With the data as used for clustering a correlation matrix as Figure 3.4 can be made. In this matrix, the groups are accented based on the 4 main groups. Except for health, all groups within their main group show above average correlation.

The clustering 13 groups that are used for the clustering together sum to close to 75 per cent of the sales volume. The next clustering tests with all data, 32 groups. The steps and R code is similar to the previous clustering and is shown in Appendix 5: R code clustering.

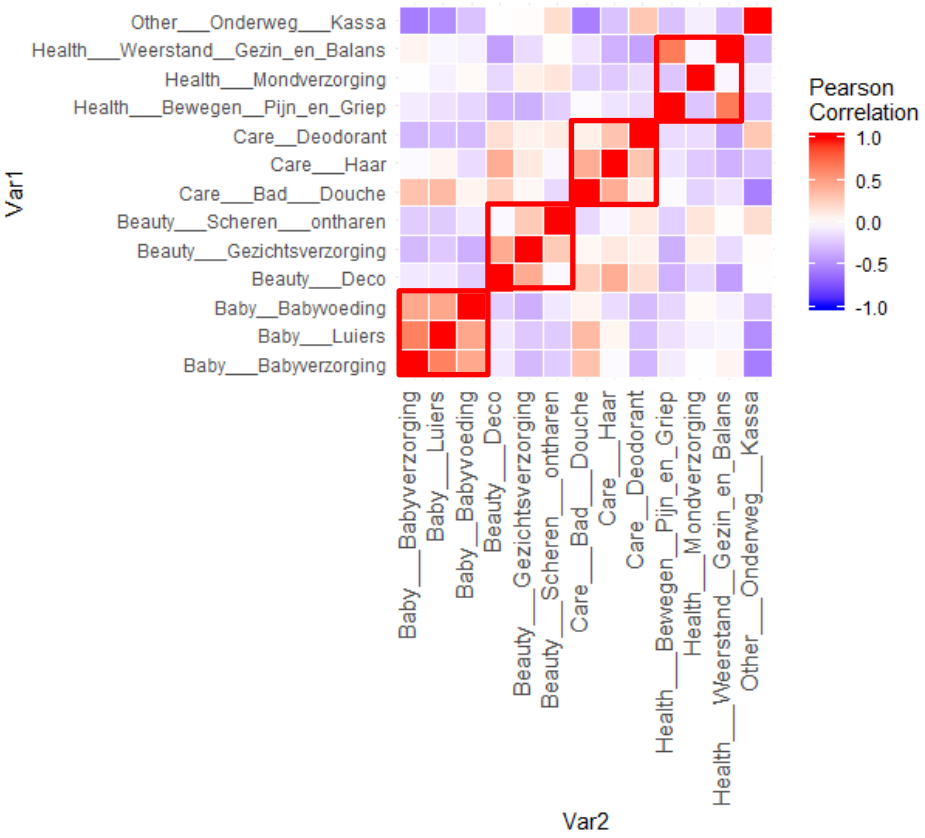


Figure 3.4 Correlation heatmap clustering 13 groups

The first step, estimating the optimal number of clusters, is done in Figure 3.5, by using the gap statistic method. The optimal number of cluster for this data set is 7.

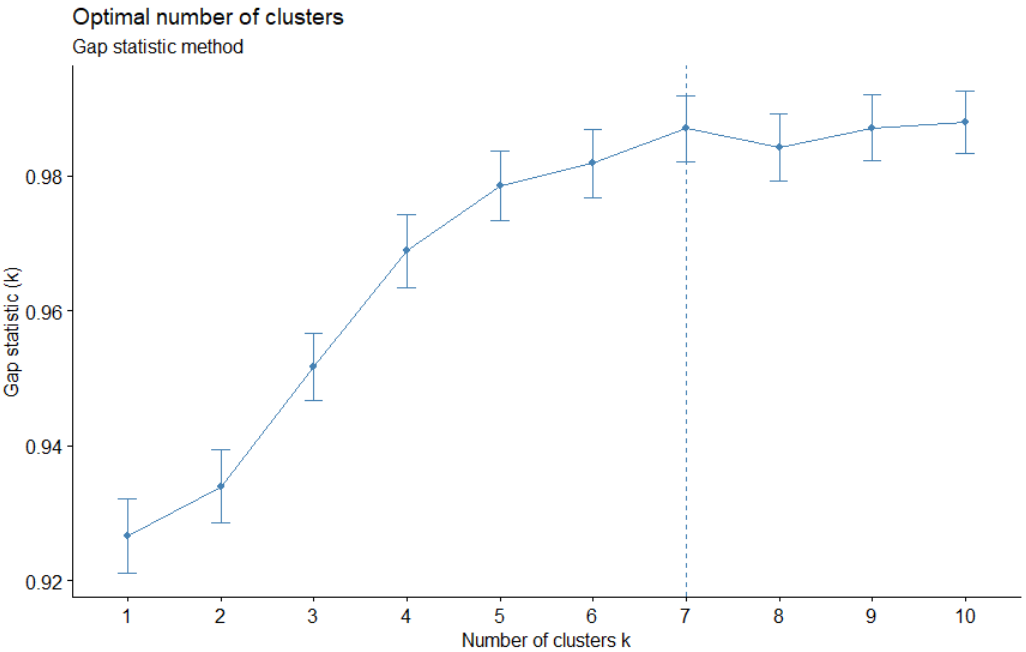


Figure 3.5 Optimal number of clusters 32 groups data

Figure 3.6 Cluster plot result, 32 groups

The clustered data can be plotted as in Figure 3.6. The overlap of clusters is bigger than for the first clustering. However, with a closer look at the legend, the overlapping clusters are mostly from the same main group. The more clusters add more detailed information, what product group is especially important.

The data that is used to name the clusters as in Figure 3.6 is given in Table 3.3. The naming method is the same as for the first run with five clusters. Not all 32 groups are used for naming, but the same 13 groups, which together are about 75 per cent of the sales.

Table 3.3 Average cluster values for naming 7 clusters

	Cluster						
	1) Health - Sport, pain and flu	2) Runshop / Travelling	3) Beauty - Make-up	4) Baby - Average	5) Beauty - Average	6) Baby - Baby food	7) Health - Average
Avg. Baby - Babyverzorging	-5.8	-64.5	5.3	21.9	-33.8	55.0	-13.8
Avg. Baby - Babyvoeding	-17.8	-62.1	-1.4	8.9	-22.0	112.9	5.1
Avg. Baby - Luiers	-8.3	-78.5	7.3	20.7	-33.2	60.3	-9.6
Avg. Beauty - Deco / Make-up	-0.5	-8.8	15.7	-6.1	15.9	-17.9	-19.0
Avg. Beauty - Gezichtsverzorging	1.7	-13.7	6.0	-8.1	18.4	-20.6	1.0
Avg. Beauty - Scheren & Ontharen	-4.1	-7.1	-0.2	-4.0	15.0	-4.0	8.4
Avg. Care - Bad & Douche	4.3	-40.3	9.1	9.2	-9.4	-4.0	-16.3
Avg. Care - Deodorant	-4.4	19.7	3.3	0.1	7.3	-8.1	-7.9
Avg. Care - haar	-1.3	-11.7	6.5	3.2	4.3	-11.8	-8.7
Avg. Health - Bewegen, Pijn & Griep	21.4	3.8	-17.6	13.1	-23.9	-14.5	-0.4
Avg. Health - Mondverzorging	-2.7	-9.8	0.4	-7.4	1.9	17.1	10.3
Avg. Health - Weerstand, Gezin & Balans	9.9	-8.9	-9.9	4.2	-12.0	-5.7	10.1
Avg. Other - Onderweg & Kassa	-20.2	128.7	-4.2	-23.5	34.4	-10.5	11.9

The downside of using the values in Table 3.3 is that it shows an average value for each cluster. Outliers, positively or negatively, will have a huge impact on the final value in the table.
The number of stores each cluster contains is shown in Figure 3.7. There is good spread over the clusters, with cluster 2 and 6 as minor number clusters. For cluster 2 this is favourable, since this cluster contains most stores on public transport stations, as visualized in Figure 3.9.

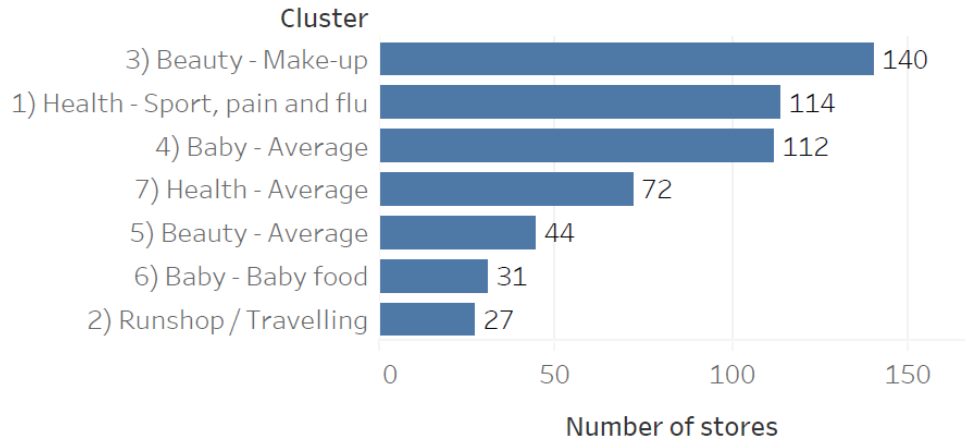


Figure 3.7 Cluster sizes with 7 clusters

The correlation matrix for this data set with 32 variables can be seen in Figure 3.8. The main groups are again marked in this figure. Not all groups in the main groups are correlated, but most are. The most remarkable outlier is lip care (*'lipverzorging'*) which has stronger correlations with the other/runshop group (*'onderweg en kassa'*) or throat care (*'keelverzorging'*).

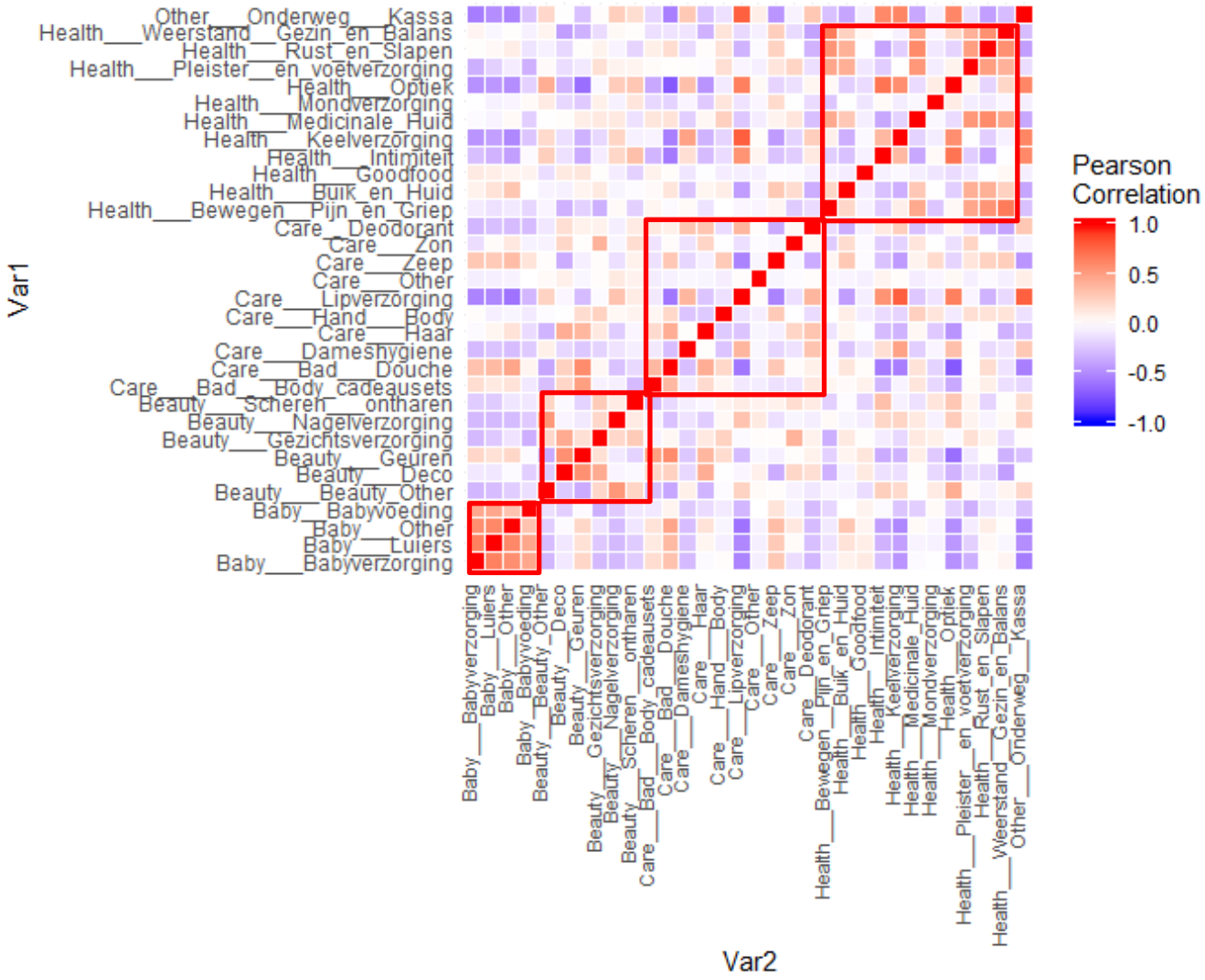


Figure 3.8 Correlation heat map clustering 32 groups

The run with 7 clusters is considered the best one since the output gives 2 clusters for 3 out of 4 main groups and a runshop/ travelling cluster. It is expected that this will enhance the opportunities to use the clusters to optimize skill allocation. However, clustering is not an exact science. Decisions are supported by internal company experts.

The maps in Figure 3.9 show the spatial spread of stores in the cluster grouped by the main groups. It stands out that there is no cluster for the main group 'care', which means that this group has a rather even spread over the stores. All maps are available in appendix 6 in full format. Some clusters are spatially grouped. For example, the 'Health – Average' cluster is mainly in centres of cities, mostly in the Randstad. Especially compared to both of the beauty clusters, which have a more even spread. The runshop / travelling clusters are located on public transport stations. The baby clusters seem to be mostly near city centres, for example close to, but not in, Amsterdam centre.

The location of a store is expected to highly influence the sales pattern. Therefore, the store environment is calculated together with its socio-demographic characteristics in chapter 3.4, which will be compared with the output of the clustering.

A) Beauty

B) Health

<confidential>

<confidential>

C) Baby

D) Runshop

<confidential>

<confidential>

Figure 3.9 Maps of the 7 clusters grouped per main group (high-resolution versions available in appendix 6)

3.2.1. Normalizing for shelf space

As proposed in chapter 2, in this research it is tested whether it is possible to normalize sales data, volumes, for the shelf space a product group is allocated in each store. After extensive testing, it did not give useful results, caused by errors in the source data, in the allocation of shelf space per product group per store. Next to the data error, this method also assumed that a double amount of shelf space for a product group also means double the sales. This turned out not to be the case. Consulting internal company experts, it can be that some stores have a lot of shelf space for a product group which it actually does not need. For this research, the allocated shelf space, therefore, is left out and sales data in volume of sold products is used.

3.3. Step 2. Employee survey data

As described in Figure 2.1 the insights gathered from the clustering based on sales data is used to help determine parts of the content of the employee survey. The survey is created in collaboration with a Professor and a PhD candidate from Utrecht University.[2] The collecting, processing and storing of the data, which contains personal information about employees, will be done by Utrecht University. This means that no personal data collected with the survey ever will be used in databases or models within the organization, outside the University of Utrecht. The case study organization will get the aggregated results to benefit from.

The first part of the survey lets the participants rank the different product groups based on their own skills, preferences and how eager they are to learn in a certain product group. Those ranks are followed by some question about the background, non-work-related interests and characteristics of the participating employee together with his or her living situation.

Next, some question about the position, job, working- days and hours and type of contract will follow. The survey ends with some personal information about the employee.

In the survey, an option is given to fill in an email address, which will be used to send a personal 'score-card' of the employee as individual compared with an average, for example from the store, or a region. This report is used as incentive to gather enough respondents.

The first part of the survey, which rates the respondents on how good they are in selling a product group and how much they enjoy advising a product group, is used for this analysis, together with the store number, postal code and number of contract hours.

For legal reasons, the survey focusses on non-franchise stores and their employees. For the entire population counts that N = 3070. The number of respondents that completely completed the survey is 842, of which 100 are employees of franchise stores. The response rate as a number of participants is: $\frac{742}{3070} * 100 \approx 24.2\%$

When every employee with one or more missing values for the product groups is removed, the actual response rate is: $\frac{587}{3070} * 100 \approx 19.1\%$

From the 3070 employees, there are **<confidential>** employees who work, on contract, a maximum of **<confidential>** hours a week. It is expected that employees who have more contract hours are, on average, more engaged than employees who work fewer hours, and thus a higher response rate. When calculating the response rate in terms of contract hours, the rate is: $< \mathbf{confidential} > * 100 \approx 37.2\%$

Although franchise employees are not the main target of the survey, they were able to fill out the survey. This returned in another 100 respondents, of which 81 without missing values. Those respondents will not be directly used in the optimization. To assign a rating to all employees that did not take part in the survey, the ranking of the nearest employee who did take part in the survey is used. This could cause many employees with the same data, which makes the optimization less meaningful. However, it still acts as a proof of concept, which could be used when a survey will be conducted with a, near, 100% response rate.

[2] Prof. Dr. Maarten Goos – m.goos@uu.nl
PhD candidate Ronja Röttger - r.c.rottger@uu.nl

On average, an employee has 15 contract hours. Some of the 587 responses missed a value for contract hours, for those responses the average of 15 hours is used. When a value of over 40 hours a week is detected, the values are changed to 40.
Every response thus has two ratings for each product group, those values are summed and used as the input value for the optimization as described in chapter 2.5. This means that the selling skills and the rating for how much they enjoy a product group have the same weight for the optimization. For all responses, there was an error in the survey data, which caused that there is no value for the preference of one of the product groups (health_mondverzorging). This error is mitigated by using the value of the selling skills for that product group for the missing value, for every respondent.
Figure 3.10 shows the spread of the rating from employees from the surveys, where the two values are summed. The error for health mouth care becomes clear because every odd value is null. Figure 3.10 shows that employees tend to give themselves above average ratings, without giving the maximum rating. This result is as expected from such a survey, according to the central tendency bias (Sinova, Rosa, & Ángeles, 2014). In chapter 3.6 the amount of contract hours of an individual employee is used to create better suitable individual ratings from the survey data.

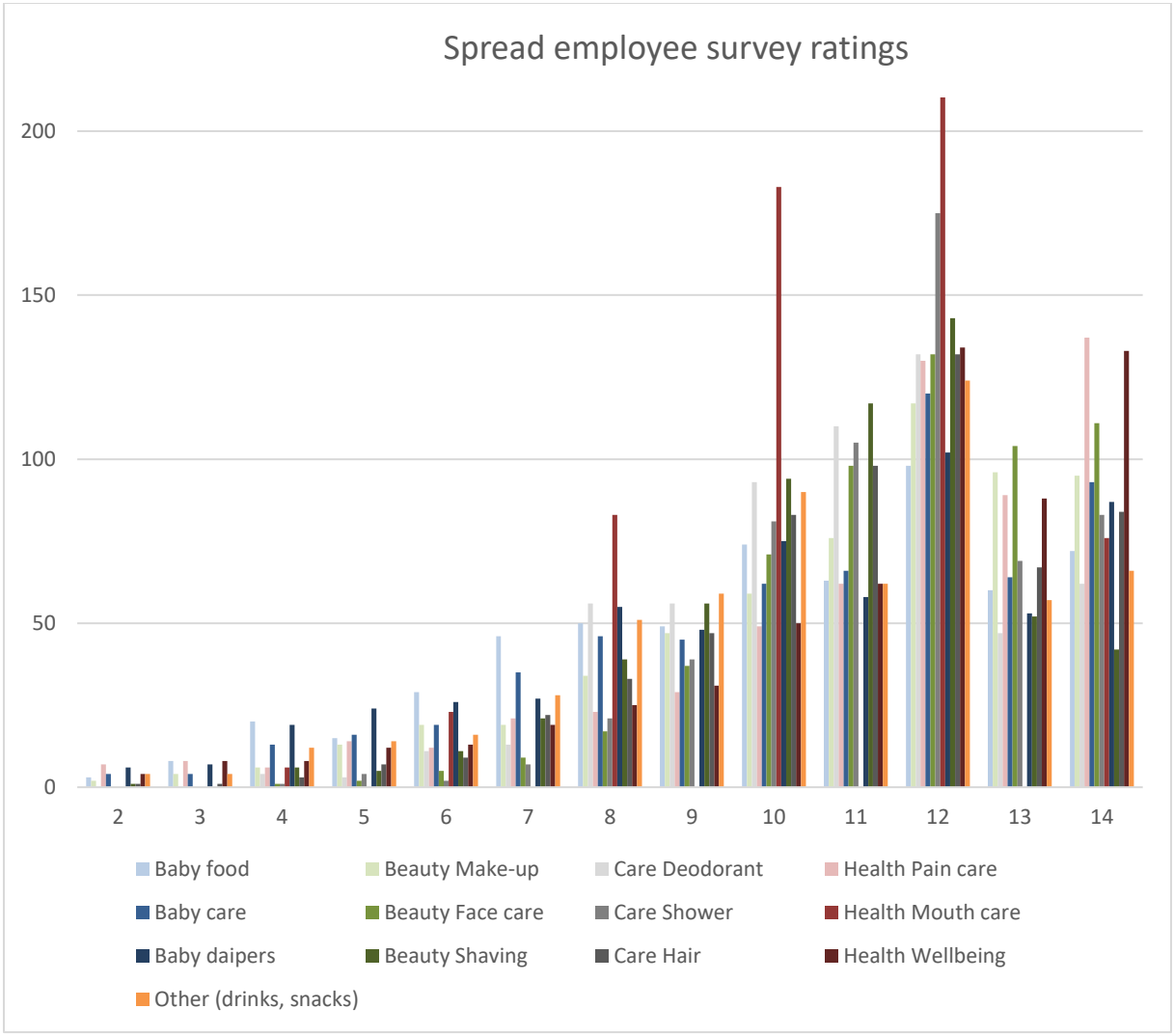


Figure 3.10 Spread employee survey ratings

Figure 3.11 shows the average of the ratings from the survey per product group. The groups are similarly rated since the values range from an average around 10.2 to 11.7. The three product groups for baby and the other group all score well below average. This could mean that there is a lack of knowledge, but for those groups, it is considered that they do not need much advice or specific employee knowledge.

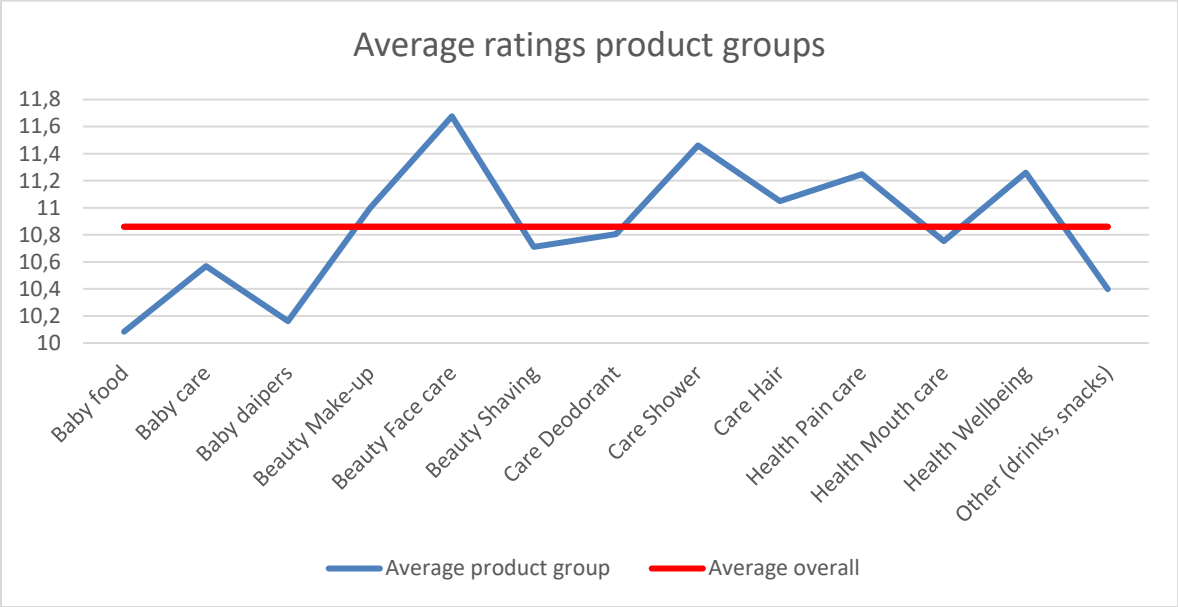


Figure 3.11 Average rating per product group

It was expected that the spread of the data from the survey would be more evenly over the rating groups, and more dispersed over the product groups. This would show that employees have their own specific set of characteristics, which could help when optimizing in step 4.

3.4. Step 3. Clustering stores - Geographical and socio-demographic data

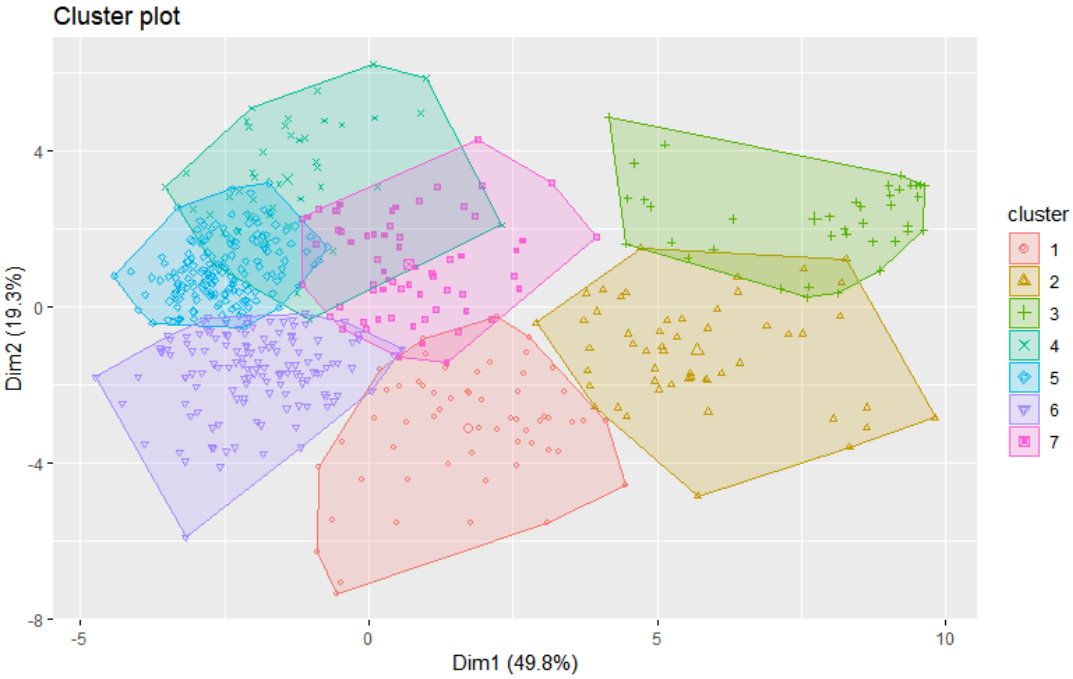
As described in chapter 2.4, for each store a service area is calculated based on the stores' turn over. Taking into account Tobler's' First Law of Geography everything is related, but closer things are more related than distance ones (Goodchild, 2009). For selecting the postal codes for the service area they are ranked based on distance from the store location. The nearest ones are selected, since these are the most likely locations the potential customers live, and thus are important for the characteristics of the store.

The Whize (2018) data set contains socio-economic data for every postal code 6 (PC6). The available attributes are given in Appendix 8. All selected PC6 polygons are summarized and a weighted average value for every attribute is calculated for every store. An example is given in Figure 3.12, which shows the percentage of the inhabitants with a university grade in the service area as created for each store. Just as with the clustering based on sales data, the possibilities of using all socio-economic attributes together for clustering are explored.

The visualization in Figure 3.12 is one of the variables that is used for the clustering in Figure 3.13.
The socio-demographic data can, just as the sales data, be used for clustering, which creates another set of clusters. For equal comparison, the number of clusters is 7. In comparison with the sales data clustering, this clustering looks good because of the relatively little overlap of the clusters, as shown in Figure 3.13. The clusters all contain stores that have a comparable store service area. This clustering uses 4 stages of education level, 5 classes of income, 10 classes for the living situation such as married or living with children and lastly 4 ranks of social classes are variables.
This result can be combined with the first clustering based on sales data, to see where those clusters overlap and most importantly where differences occur. Those differences might show stores that have an unexpected sales pattern for their location. An explanation for this can be due to the employees currently working in the store. More about this in chapter 3.5.

<confidential>

Figure 3.12 Percentage level of education: University grade in store service area



Education	Income	Family situation	Social class
1) Low	Below average	Without children + other	C+D
2) High	Below average	Family with children + other	A+B1
3) Extremely high	More than twice average	Without children young + other	A
4) Moderate	Above average	Married with children	B1+B2
5) Low	1 to 2 times average	Married often with children	B2+C
6) Extremely low	Average	Without children old	B2+C+D
7) Relatively high	1.5 to above 2 average	Without children old	A

Figure 3.13 Store service areas clustering

The number of stores in each of the clusters is shown in Figure 3.14. Compared to Figure 3.7, which shows the spread for the clustering based on sales data, there are some similarities. Two clusters are significantly bigger than average, where two other clusters are smaller and some clusters are about average.

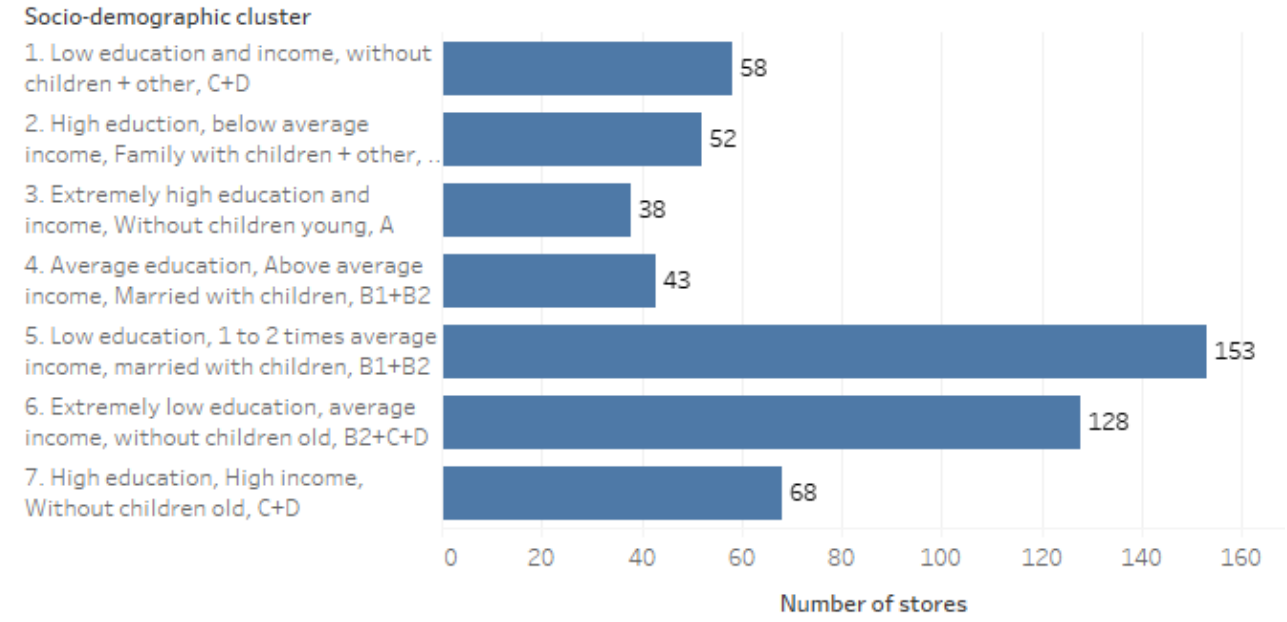


Figure 3.14 Number of stores in each socio-demographic cluster

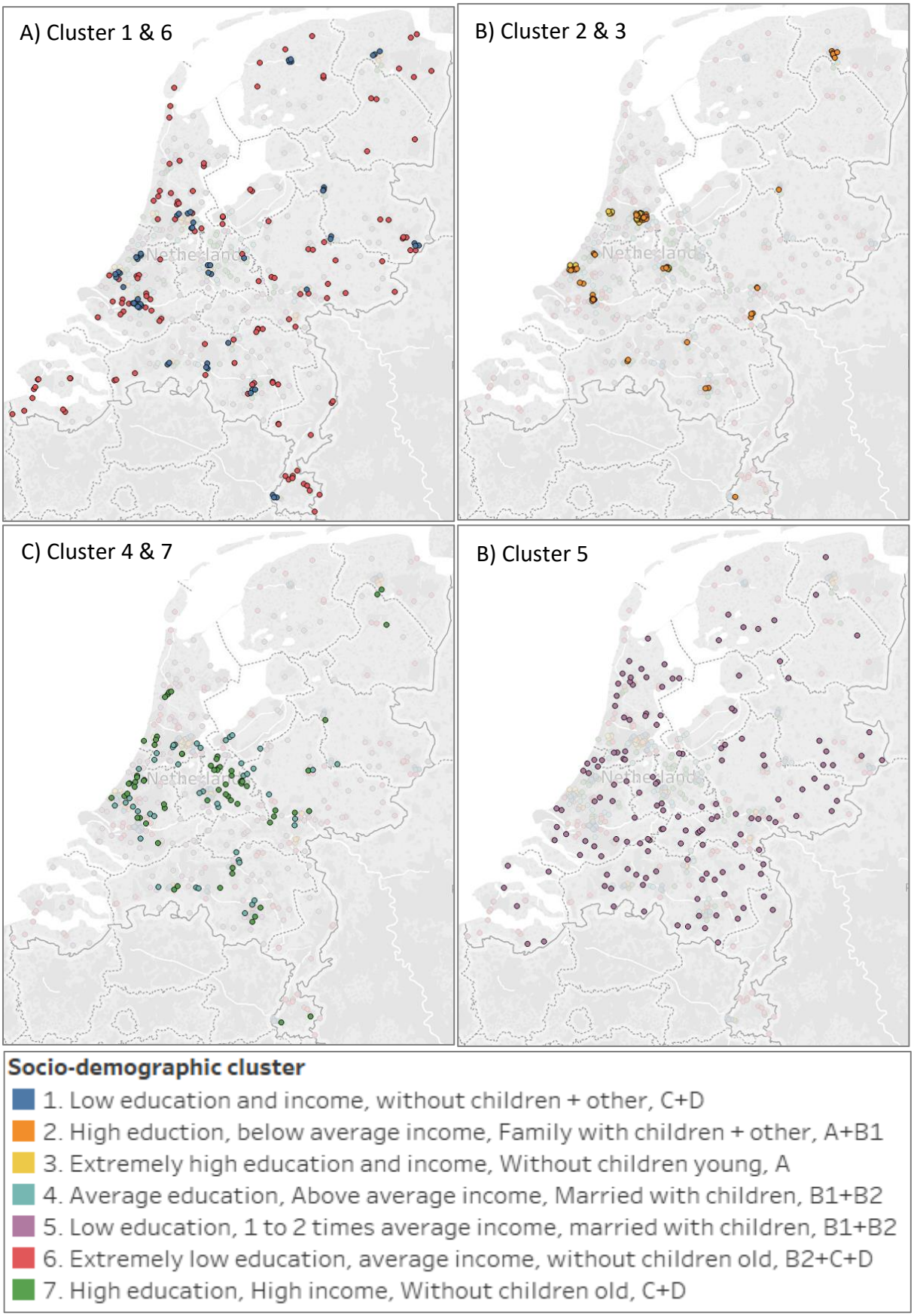


Figure 3.15 Maps of the 7 socio-demographic clusters (high resolution versions available in appendix 9)

3.5. Compare steps 2 & 3: Sales data and socio-demographic clustering

As briefly mentioned in the literature chapter, the stores' team staff is, next to the stores' location and competitors, expected to be one of the drivers for a stores' sales pattern. In this chapter both clustering methods, the first based on sales data and the other based on socio-demographic data, are compared. Firstly, by doing so, we research links between socio-demographic data and sales patterns. Secondly, this methodology can be used to help predict a stores' sales pattern for new locations, even before any sales are generated. This thus helps to predict what team skills are needed when a store opens on that location.

The socio-demographic data gives us insights into the characteristics of the service area of a store or a cluster of stores. This might show interesting trends. An example is given in Figure 3.16, which shows the average age differentiation within one of the health clusters. For this health cluster, above average age is expected, since it focusses on pain care and other most common for elderly complaints. For all classes up and until 45 years old, the stores in this health-related cluster have scores below average. From the categories with people from 45 years and over, the scores for those stores are increasingly more above average. This observation underlines the expectations for this cluster.

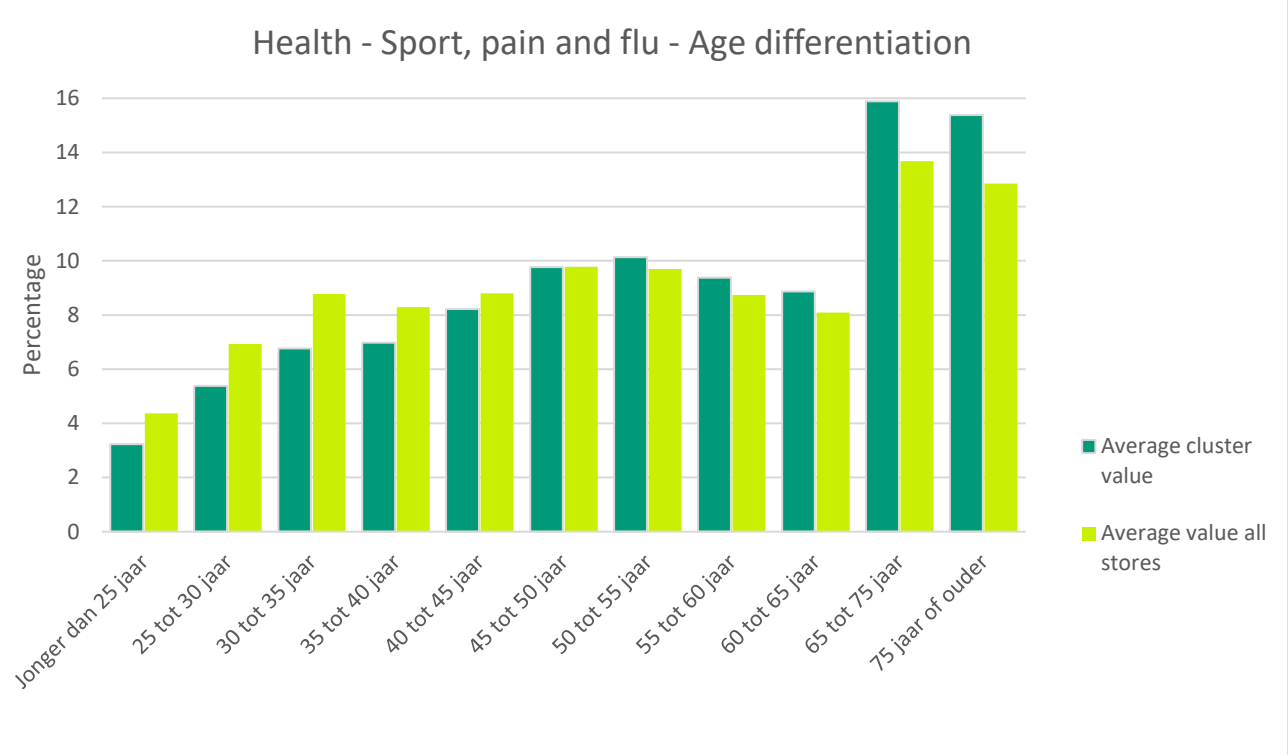


Figure 3.16 Average age differentiation in a health cluster

The runshop cluster, which contains the stores mostly on public transport stations, have on average deviant values for most socio-demographic variables. Since most public transport stations are in city centres, the hypothesis is that the service areas for stores in this cluster have many students and single relatively prosperous and rich people with high average education. The differentiation of education level is shown in Figure 3.17, compared with the education differentiation for the Beauty – Make-Up stores.

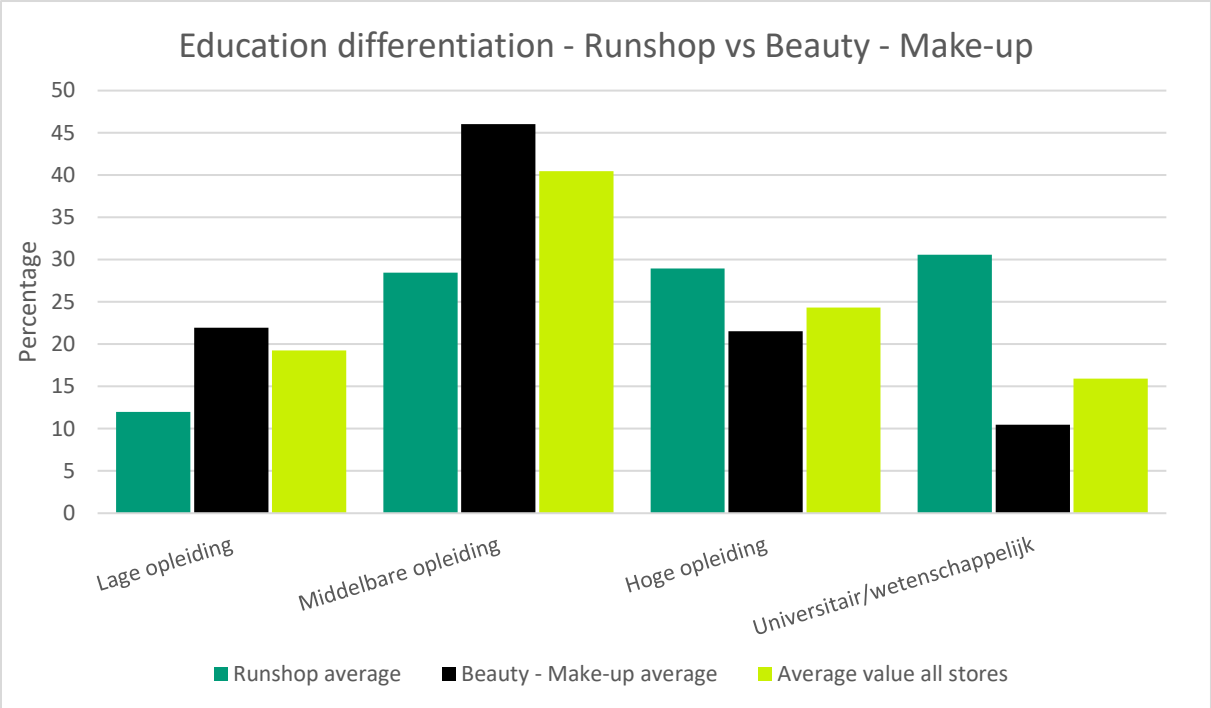


Figure 3.17 Education differentiation - Runshop vs Beauty - Make-up

The percentage of university grade education for runshops is double the average value. Compared to the Beauty – Make-Up clusters the difference in education level is even more significant. The overall trend for education levels is opposite for the runshop cluster, mainly high education, and Beauty – Make-up cluster, mainly lower education.

When looking at the other socio-demographic groups for the Runshop cluster, the percentage of students in the cluster is around three times average and the wealthiest social class, A, is close to twice the size as the average. The hypotheses as stated on the last page, that runshop clusters have a relatively high education level, income, social class and percentage of students is thus accepted.

In Figure 3.18 the data from Figure 3.16 is enhanced with the age data of the runshop cluster, including (linear) trend lines. Although the linear R^2 is about 0.41 for the runshops, against 0.83 for the health stores, the figure shows the opposite trends for average age differentiation in the two selected clusters. The R^2 of the regression can be improved by using polynomial regression, as shown in Figure 3.18 with small dotted lines. The R^2 for runshops is doubled to 0.83 and the new value for the health cluster is 0.89. The trend from the linear regression is still visible.

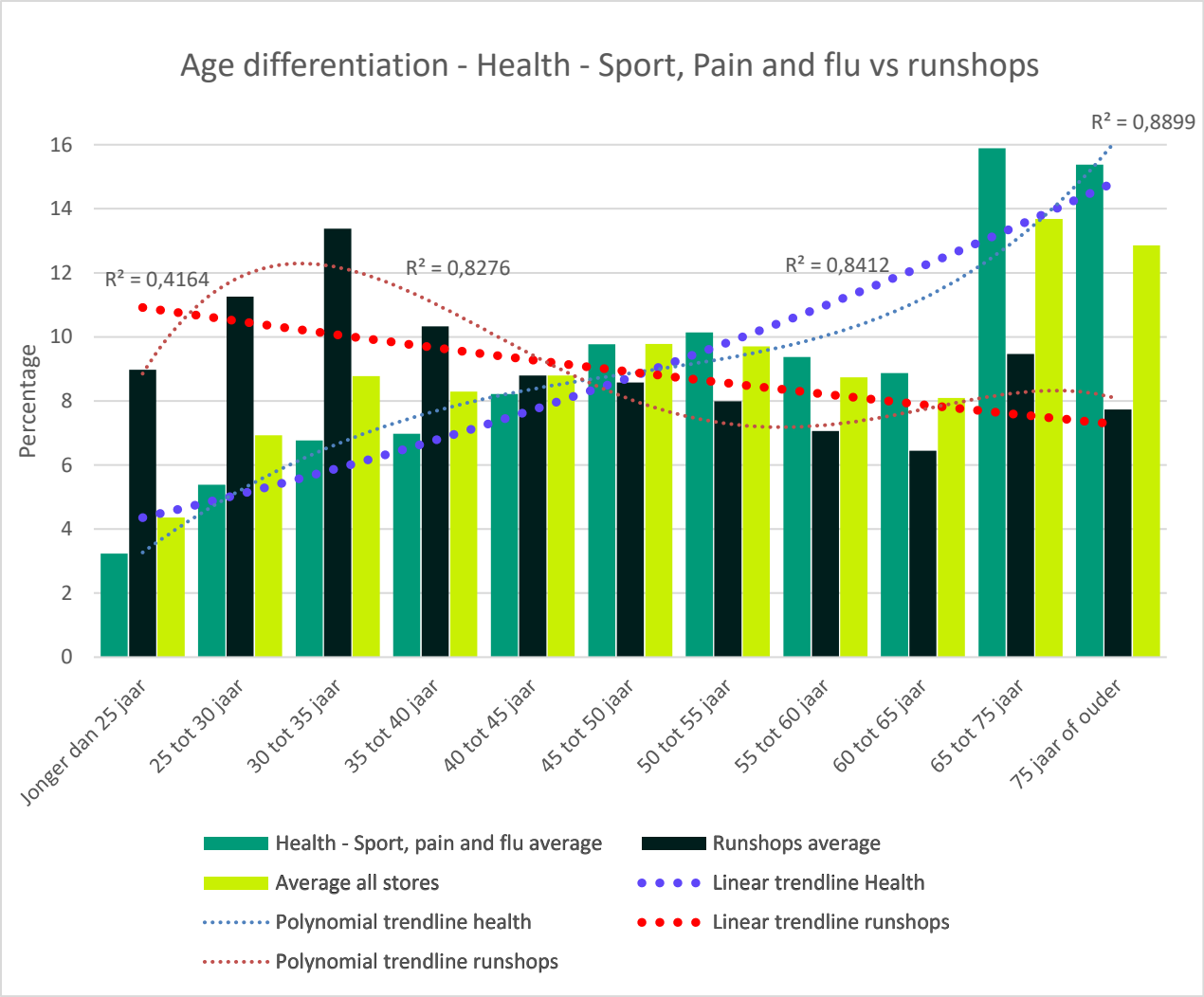


Figure 3.18 Age differentiation and regression for cluster Health - Sport, Pain and flu vs runshops

Analysis as performed above can be extensively done with this data, on both cluster and store level. This analysis uses the clustering based on sales data, together with the socio-demographic data in the service areas. In the next part of this chapter, the clustering on sales data is used together with the clustering on socio-demographic variables to research the possible links or deviations.

Figure 3.19 and Figure 3.20 both show the spread of stores over the two types of clustering, which shows the links between the clustering based on the socio-demographic data and the sales data. Both figures are based on the same data and then pivoted, to help understand the different spread over the clusters.

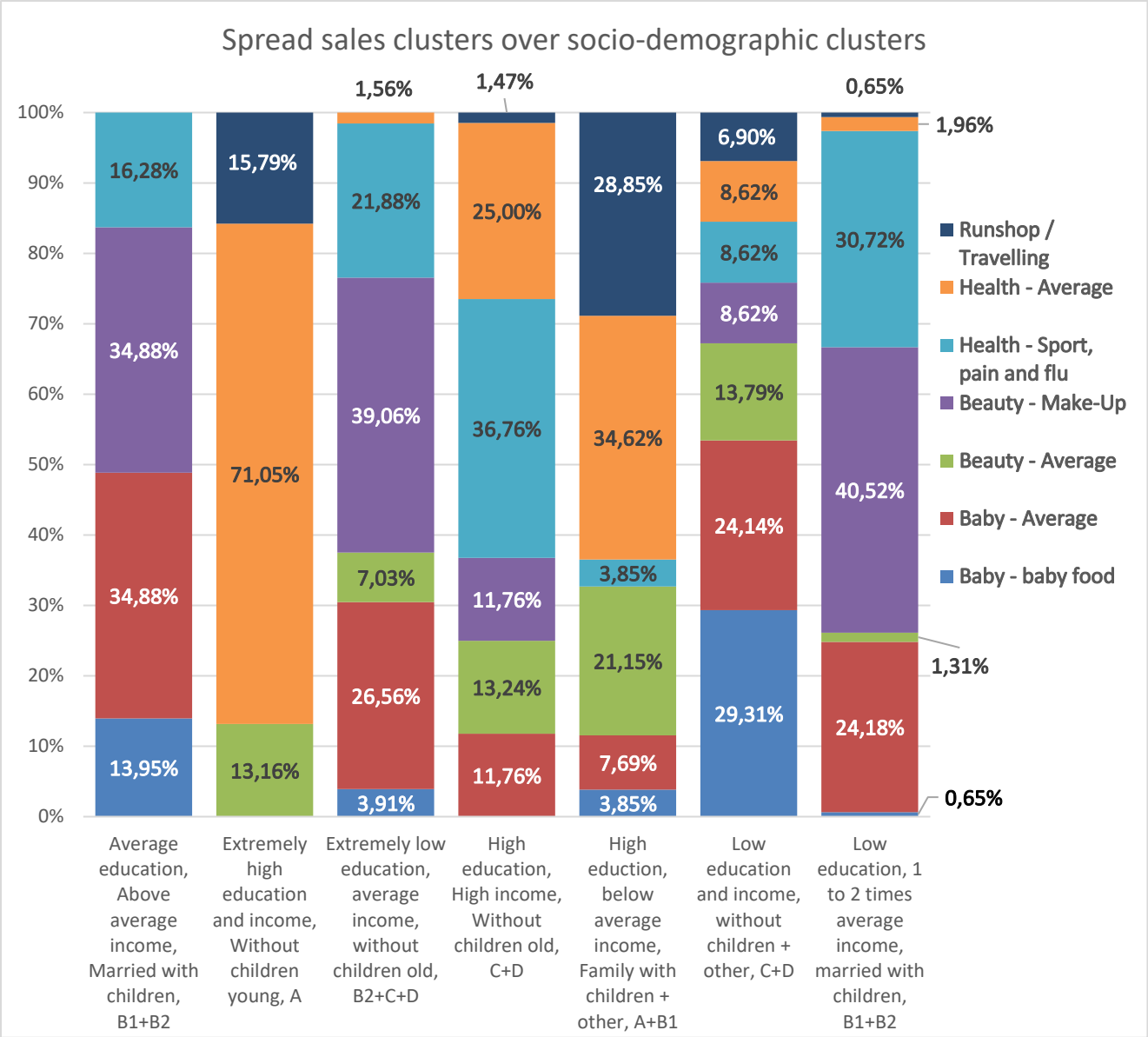


Figure 3.19 Spread sales clusters over socio-demographic clusters

The most significant pattern can be seen in the cluster with extremely high education and income. This cluster only contains stores from three out of seven clusters. For over 71 per cent this cluster contains stores clustered as Health – Average. When looking at the other socio-demographic clusters that are named high education, those two clusters have a significant part of Health –Average stores as well. From this data, we can say that when a service area of a store has a high education, the health cluster on average performs better.

On the other hand, the three socio-demographic clusters with extremely low to low education level and an average or above average income, have the highest percentage of Beauty – Make-Up stores. This is another insight for predicting sales for new locations.

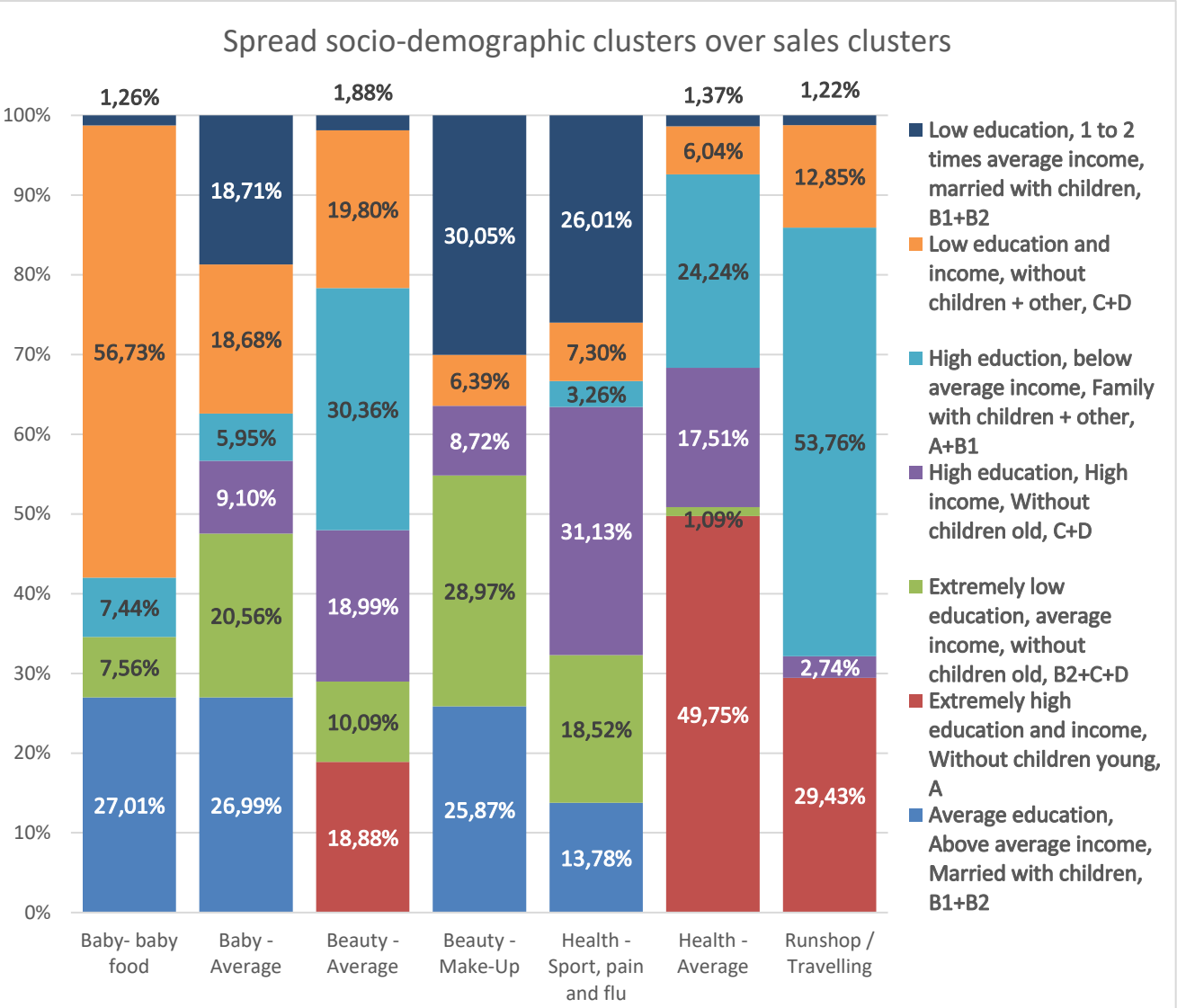


Figure 3.20 Spread socio-demographic clusters over sales clusters

For baby food, it is expected that the major percentage of the stores has a service area that is clustered as families with children. For this cluster, we see an unexpected pattern since more than half (56.73%) has a service area which is not classified as 'family with children', as shown in Figure 3.20.

Research in the underlying sales data explains this significant bias. It turns out that some stores in the Baby – Baby food cluster sell more than **<confidential>**% above average for some baby food products. Since the clustering and cluster naming is based on average values some outliers can have a significant impact on the output. This bias only occurs for certain product groups, which are all closely related to the increase of sales of powdered milk due to safety issues in China such as the melamine scandal around the year 2015 (Wu et al., 2018). Clustering might be improved by removing such extreme outliers.

More insights can be pulled from Figure 3.20. For the Beauty – Make-Up cluster there are three major socio-demographic clusters, which are all classified as (extremely) low to average education, with on average an average income, just as the conclusion from Figure 3.19.

3.6. Step 4. Employee to store allocation optimization

3.6.1. (Mixed) Integer Linear Programming program

In this chapter the development of the integer linear programming optimization model, as proposed in chapter 2.5.1 in script, is explained and discussed, followed by the analysis and interpretation of the results. This optimization aims to minimize the residual between the skill demand of a store and the skill supply, while limiting commute distances.

The development of the Integer Linear Programming program is an important step for creating the optimal allocation of employees to stores. The program is written in C# with Microsoft Visual Studio. In this chapter, the most important steps of the program and code will be discussed. The actual code is pasted in Appendix 7.

The first step is to import the right systems into the program. One of them includes the CPLEX solver that enables the Integer Linear Programming optimization.

Next step is to load the data, as specified in chapter 2.5. The data consists of 3 matrices; the store matrix with a set of skill demands for every store, the employee matrix with a set of skill supply of every employee and the distance matrix with the distance between each employee and every store.

In the Integer Linear Program, some parameters are used, as shown in Appendix 7. The parameter Maximumdistance is the $D_{\max}$ from the model in chapter 2.5.1, which is the maximum commute distance for allocation. The other parameter is dblLambda, the λ in the model in chapter 2.5.1, which is the cost of commute distance. The input unit of distance in the data sets is kilometres, and thus this parameter is in kilometre as well.

The third parameter is dblTimeLimit, which is not specified in the model since it is used to set the runtime of the model in seconds. Since the program is iterative, the result will get more optimal as the model runs for a longer period of time.

Figure 0.4 in Appendix 7 shows the code of the actual CPLEX solver of the Integer Linear Program. This solver uses the inputs and parameters for optimization. To do so, the objective function is generated. This includes the constraints as discussed above. An important constraint is that every employee should be allocated to exactly one store. This prevents unjustified unemployed employees or employees that have to work in multiple stores. The solver will start with a random solution, which will probably be far from optimal, after which it will continue finding other, more optimal, solutions. The model will output the most optimal solution after the maximum runtime has passed. The model produces a .csv file with a matrix filled with binary data. Every row represents an employee, and all columns represent stores. Thus, every value is a zero, except for a single column per row, which is a one. This is the store to which the employee is allocated by the program.

3.6.2. Integer Linear Program testing

The created Integer Linear Programming is tested with a synthetic, randomly generated data set. Runs with different runtimes and distance costs are performed and the results are compared, shown in Figure 3.21 and the blow-up in Figure 3.22.

For the tests, the parameters for distance costs (λ) and the runtime of the model are changed. The longer the runtime, the better the solution should be, and thus the lower the residual of the model should be. This is the first check, which is the case according to Figure 3.21, since every line decreases when then runtime increases. A bigger λ, which means higher travelling costs, cause a less optimal solution. High traveling costs cause the model to select the nearest employees, which increases the chance for a less optimal solution. All runs for testing with a maximum travelling distance ($D_{\max}$) of 45 to prevent that travelling distances become too big. The residual of the most optimal solution decreases significantly in the first minutes of the run. Although the model becomes better and better with an increased run time, only minor improvements are made. This is important when using the model with the real world data, since the first tests can be done with rather short runs to get the first insights.

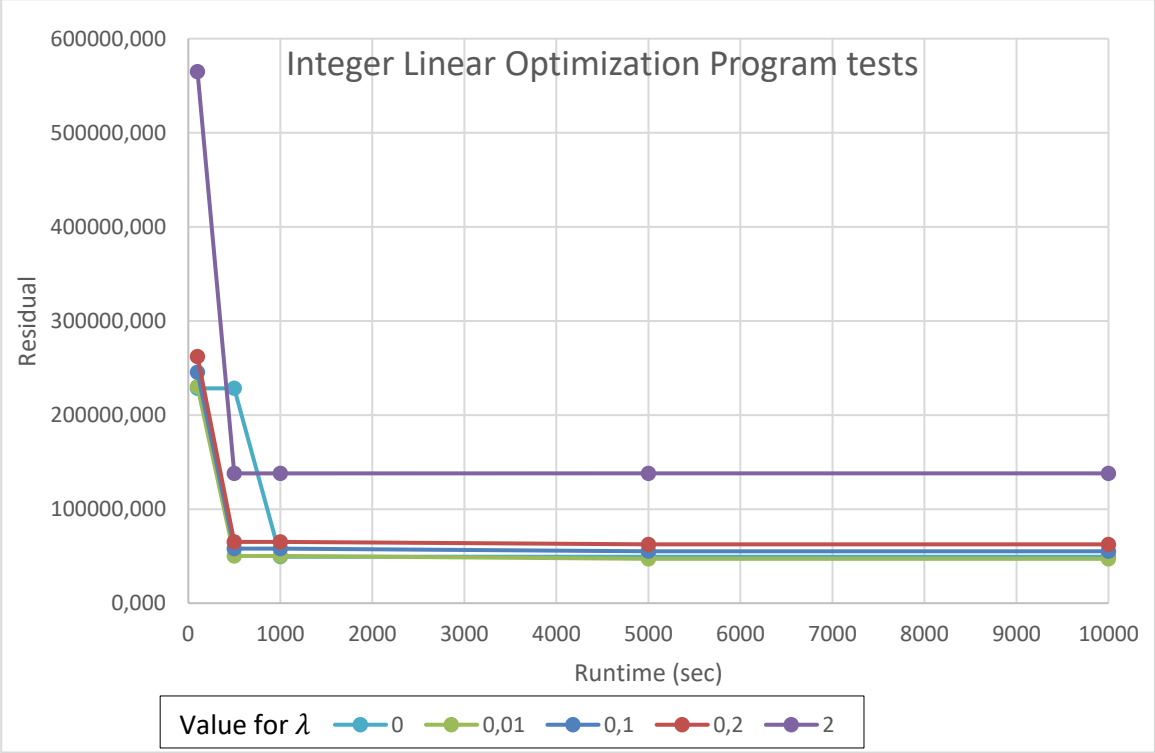


Figure 3.21 Test runs Integer Linear Optimization Program

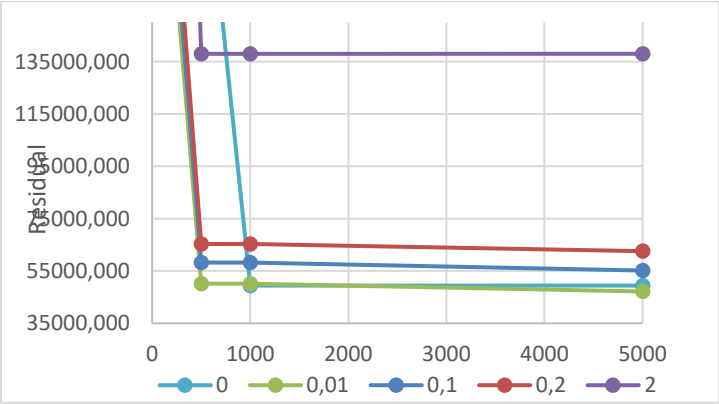


Figure 3.22 Blow-up test runs Integer Linear Program

The second test to check if the model is optimizing correctly is done with small 10 by 10 matrix data sets filled with binary data, where every skill/attribute combination is a 0, except from a single combination of store/attribute and employee/attribute, which is given a 1. The distance matrix is filled with the opposite binary values, where every value is 1 and only one store/employee combination is 0. This means that an employee lives on the exact location of a store. This data set-up should return the optimal solution, with a residue of 0, and should run really fast. The output of this test is shown in Figure 3.23-left. This is the expected output, within a couple of seconds. This shows that the model is running as it should and is thus validated.

```
Variable x:
Employee     1:     1  0  0  0  0  0  0  0  0  0
Employee     2:     0  1  0  0  0  0  0  0  0  0
Employee     3:     0  0  1  0  0  0  0  0  0  0
Employee     4:     0  0  0  1  0  0  0  0  0  0
Employee     5:     0  0  0  0  1  0  0  0  0  0
Employee     6:     0  0  0  0  0  1  0  0  0  0
Employee     7:     0  0  0  0  0  0  1  0  0  0
Employee     8:     0  0  0  0  0  0  0  1  0  0
Employee     9:     0  0  0  0  0  0  0  0  1  0
Employee    10:     0  0  0  0  0  0  0  0  0  1
Solution value = 0
Solution gap = 0
Solution status = Optimal
```

```
Variable x:
Employee     1:     0  1  0  0  0  0  0  0  0  0
Employee     2:     1  0  0  0  0  0  0  0  0  0
Employee     3:     1  0  0  0  0  0  0  0  0  0
Employee     4:     1  0  0  0  0  0  0  0  0  0
Employee     5:     1  0  0  0  0  0  0  0  0  0
Employee     6:     1  0  0  0  0  0  0  0  0  0
Employee     7:     1  0  0  0  0  0  0  0  0  0
Employee     8:     1  0  0  0  0  0  0  0  0  0
Employee     9:     1  0  0  0  0  0  0  0  0  0
Employee    10:     1  0  0  0  0  0  0  0  0  0
Solution value = 21
Solution gap = 0
Solution status = Optimal
```

Figure 3.23 Optimization program small data test: Left) Optimal situation Right) Long commute distance test

When the distance matrix is changed to be completely filled with low values, except from the diagonal which is filled with values 100. This means that every store exactly fits an employee based on skills. However, based on distance, this is the exact opposite. The output as in Figure 3.23-right is again as expected. For employee 1 the distance to store 1 is 100. Although the skill of employee 1 matches the skill demand of store 1 perfectly, the distance is too big. For every other employee, the skill match is perfect with the store with their most remote distance. Apart from this store, the values for every store are equal. The model then chooses the first store for allocation.

3.6.3. Optimization results

This chapter will discuss the results of the skill supply and demand optimization model. The result will consist of the difference between the residual from skill supply and demand prior to the optimization, and the residual after the optimization. Secondary, this will provide insights into the theoretical room for improvement of store employee teams by relocating employees.

The results can be influenced by two sets of inputs and parameters. Firstly, the weights for each product category for the skill needs affect the outputs, such as a higher weight for make-up products. Secondly, the parameters of the integer linear program affect the results, with the constant cost for distance. For the runtime of the program means the longer the better, which is discussed in chapter 3.6.2, but the correlation between output and runtime is more logarithmic than linear.

The inputs for the attributes on the skill demand side for stores are given in Table 3.4. The factors are based on internal experts. For product groups which need more specific skills from an employee, the value is larger with 1.5 and lowered where less specific skills are needed.

Table 3.4 Multiply factors for sales product groups

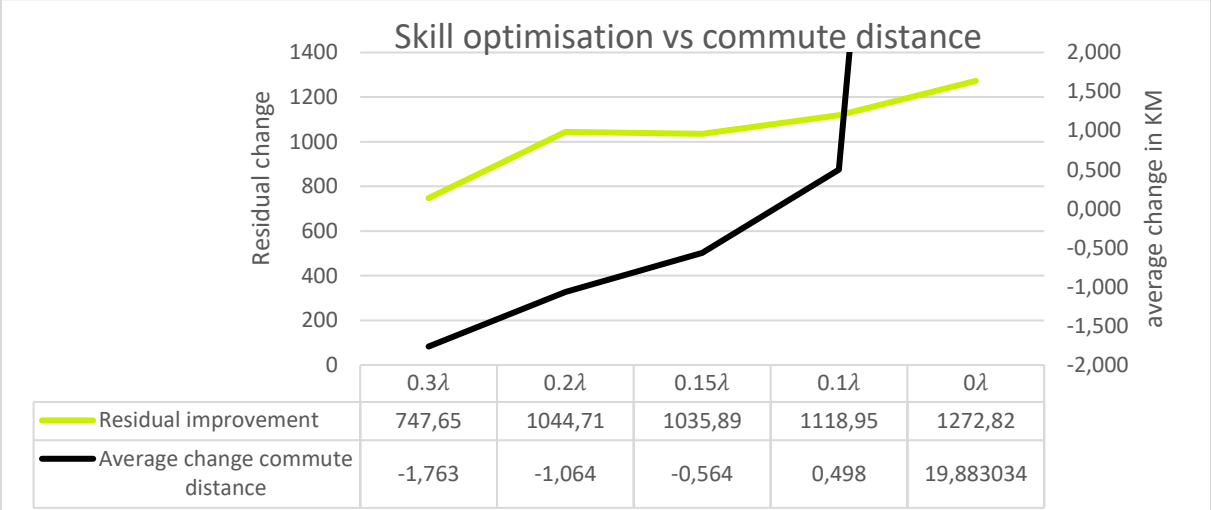
Product group / Optimization attribute	Multiply factor
Baby (care, food, diapers)	1
Beauty (make-up, face care, shaving)	1.5
Care (douche, hair, deodorant)	0.5
Health (mouth care, pain care, wellbeing)	1.5
Other (other, drinks, snacks)	0.5

The model focusses on optimizing skills allocation and travel distance, where improving both is best. However, both sets of parameters are closely related. As seen in Figure 3.21 and Figure 3.22, a higher cost for distance (λ) means a higher residual, and thus a less optimal solution. On the other hand, the travelling distance are reduced with higher costs per travelling unit, which makes the skill gap residual larger.

The third parameter with an impact on the output is the runtime of the program. While testing the model it became clear that from 1000 seconds and up, the improvement of the model is low. The actual optimization model is smaller than the testing model because the franchise stores and employees are removed. The model is run for 3600 seconds, 1 hour, for all test runs, which is seen as a safe margin for the model to reach the heuristic optimal solution, based on Figure 3.21 and Figure 3.22. The maximal commute distance is set equal for all runs at 45 kilometres, which is the maximum distance the case study organization pays for.

Figure 3.24 Current allocation of employees – lines between postal code of employee and current store

Figure 3.24 shows the commute distances in current situation of employee allocation. Each line is between the postal code of an employee and their current allocated store. The commute distance is on average 6.922 kilometres. The residual between skill supply and demand is 41691.4 points **<skill points do not have a unit to keep confidentiality>**. This current situation is the reference point. Other runs are compared with this reference data.
The parameter that is changed for comparison is the distance cost (λ). Runs are done with λ values between 0 and 0.3.



	0.3λ	0.2λ	0.15λ	0.1λ	0λ
Residual improvement	747,65	1044,71	1035,89	1118,95	1272,82
Average change commute distance	-1,763	-1,064	-0,564	0,498	19,883034

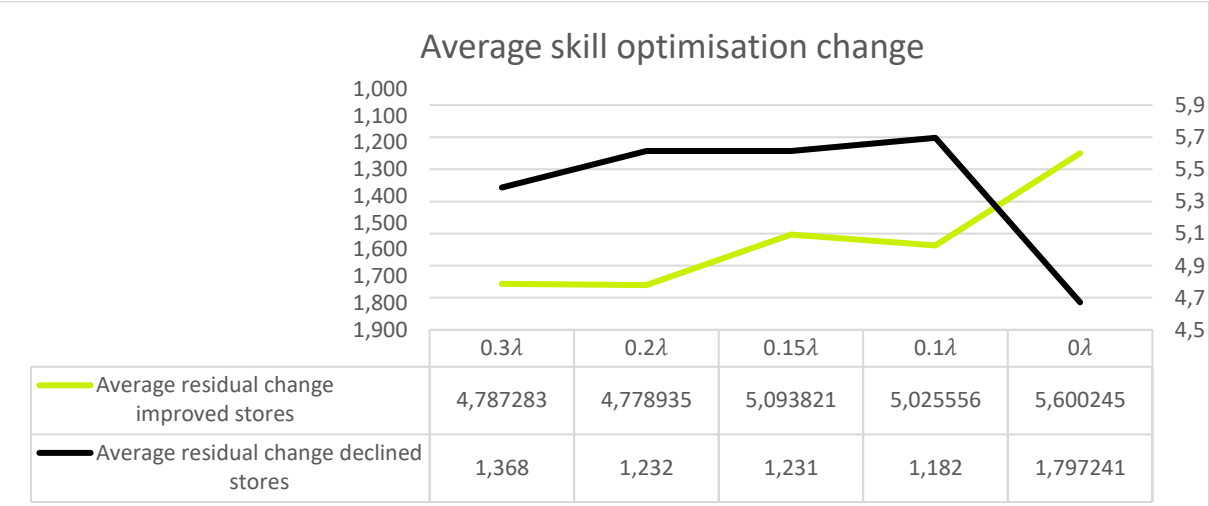
Figure 3.25 Skill optimisation vs commute distance optimisation

Figure 3.25 shows the change in residual and average commute distance, dual axis. All runs, with the chosen λ values, show a residual improvement compared to the current situation. Same goes for the commute distance, which is decreased. Except for the 0.1λ and 0λ runs. For the run with 0λ, the actual distance is not in the chart, since it is significantly larger than the other values. When λ is 0, this means travelling is free. This gives the solver many more options to take into account, which leads to possible longer run times. Figure 3.25 shows the output for runs of 3600 seconds. From this figure as the best option can be considered the λ 0.2 run, where the gap between the two lines is the biggest.

	0.3λ	0.2λ	0.15λ	0.1λ	0λ
#stores improved	154	156	151	148	173
#stores declined	133	131	136	139	114

Figure 3.26 Number of stores improved after optimisation

In Figure 3.26 Number of stores improved after optimisation, the entire population of stores, 287, is split into a part that has an improved residual and a part that became worse. For every run, the number of improved stores is higher than the declined stores. However, the difference is small for most runs. For the run with no travel costs (0λ), the number of stores that is improved is substantially higher. As discussed in Figure 3.25 Skill optimisation vs commute distance optimisation this comes with a cost of significant higher commute distances. Based on Figure 3.26 Number of stores improved after optimisation, the best run is again the 0.2 λ run.



	0.3λ	0.2λ	0.15λ	0.1λ	0λ
Average residual change improved stores	4,787283	4,778935	5,093821	5,025556	5,600245
Average residual change declined stores	1,368	1,232	1,231	1,182	1,797241

Figure 3.27 Skill optimisation change per store

An analysis on the average skill optimisation change for the groups of stores from Figure 3.26 Number of stores improved after optimisation is shown in Figure 3.27. This figure shows the average change in the residual for the different runs. The first conclusion from the data is that the overall effect from the different runs is comparable for both lines and that a smaller λ means a bigger improvement. The second conclusion is that the average skill optimisation change for an improved store is substantially bigger for all runs than the average decline of one of the other stores. This is the desired result from the model.

In Figure 3.28 the change in residual is split in each product group, for every run. Bars show the absolute change of the residual, the lines show the corresponding relative change. A negative value means that the residual has increased, which means that it became worse. This is the case for the 'other' group for every run. Surprisingly, the run with 0 λ has another product group that became a lot worse. This occurs due to the amount of options of combinations of stores that increases when travelling costs are eliminated. With more runtime this error should be mitigated.

Overall, a correlation can be seen between the absolute change and the relative change. The product groups that are most affected by the optimization are the beauty groups, the care – hair group and health – medicine and wellbeing. Since the sum of all values from a run is equal to the values from Figure 3.25, the run with 0.2 λ is considered as the best run.

The 0.2 λ in 3600 seconds decreases the residual with 1044.71 points from 41099.07 to 40704.04. This is an improvement of $\frac{1044.71}{41099.07} * 100 \approx 2.54\%$. With the same run, the average commute distance has decreased with $\frac{1.064}{6.92212} * 100 \approx 15.73\%$.

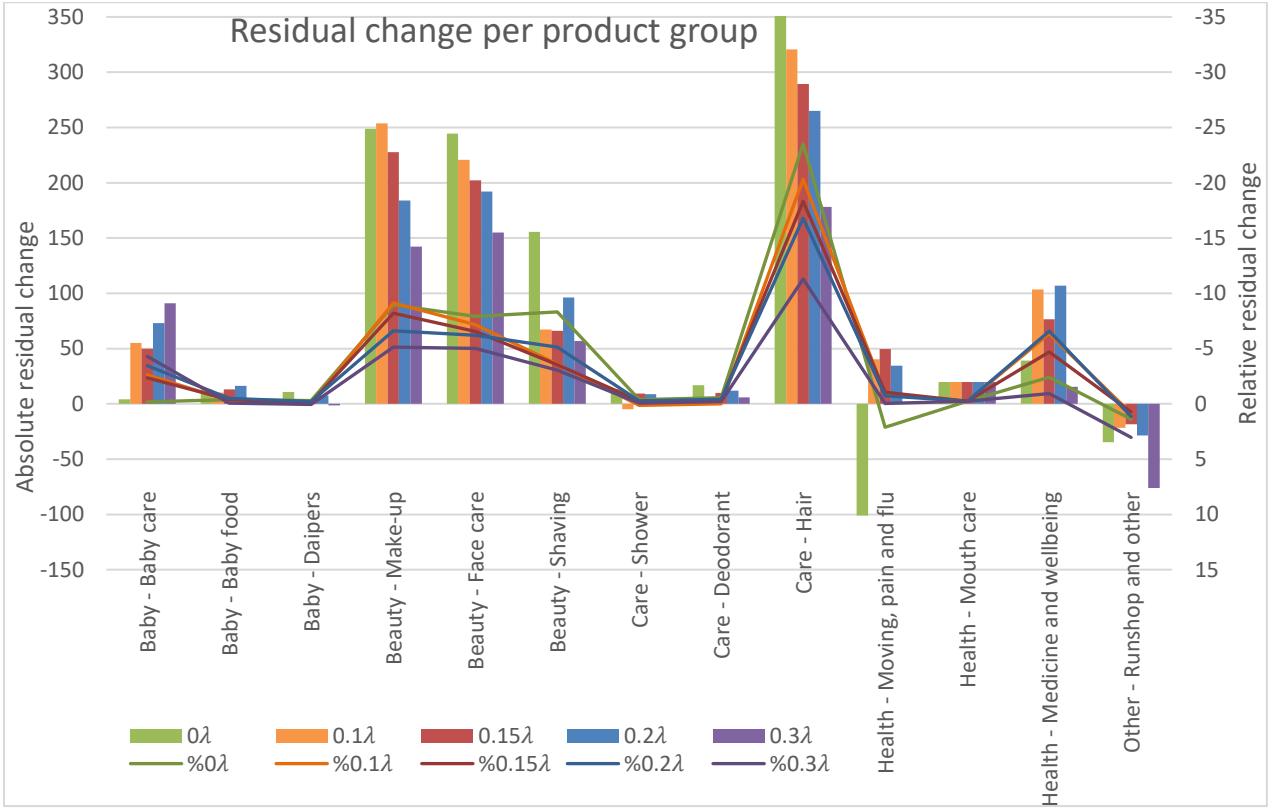


Figure 3.28 Residual change per product group

All in all, the integer linear program method shows that it is possible to optimise both parameters. For all runs with a runtime of 3600 seconds, the solver returns feasible as solution status. As from Figure 3.23, the solver will return optimal when the gap value is 0. With the 0.2λ run, the gap value is 0.0367. To get to the optimal solution the 0.2 λ option is ran significantly longer, for 32.500 seconds. This run from over 9 hours returned a gap value of 0.028, with still feasible as solution status. This run gives an improvement of the residual of $\frac{1213,19}{41099.07} * 100 \approx 2.95\%$. The average commute distance has decreased with $\frac{0.992}{6.92212} * 100 \approx 14.33\%$. In Table 3.5 all statistics are compared for the different run times. All variables are improved, except from the commute distance. When more run time or computer power is available the model can be ran until the output is optimal.

Table 3.5 Compare runs of different run times

	3600 seconds	**32.500 seconds**
Gap	0.0367	0.028
Residual skills	2.54%	2.95%
Distance	15.73%	14.33%
# stores improved	156 (average 4.78)	160 (average 4.78)
# stores declined	131 (average 1.23)	127 (average 1.23)

More than half of the stores will benefit from this outcome. The average benefit is bigger than the average loss for a store. To conclude, from this model can be drawn that optimization using integer linear programming can cut commute distances with about 15%, where skill allocation is still optimized with 2.95%.
However, implementing such a model means a big change for an organization. In this case most of the employees, $\frac{2236}{2868} * 100 \approx 78\%$, will be switched from one store to another.

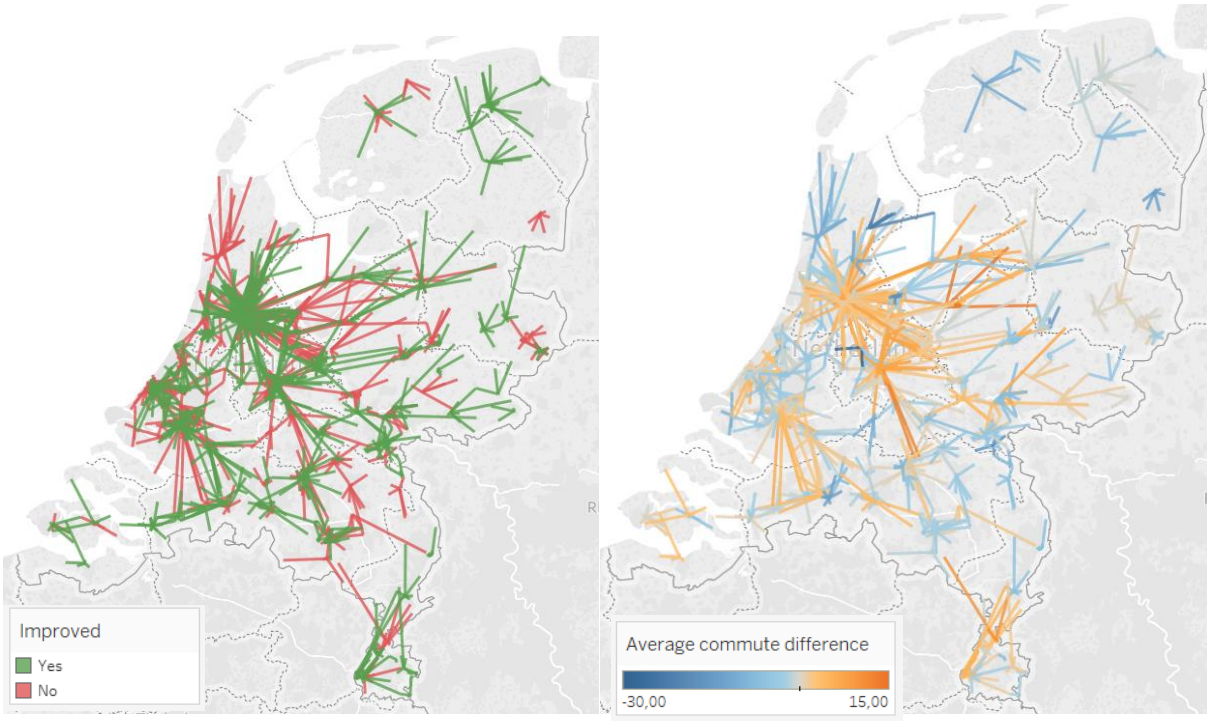


Figure 3.29 Optimized employee allocation (Left: improved stores yes or no, Right: Average change commute distance)

Compared to Figure 3.24, the allocation of employees is shown in Figure 3.29. The left map shows which stores will benefit from this optimized, and which will not. It is striking that most stores that will be improved are in city centres. The map on right shows an opposite spatial pattern for change in commute distance, where the average commute distance in the city centres has become worse and the distances for more remote stores are improved. This visualization suggests that most store employee teams will benefit from either a better average commute distance or a better skill match.

4. Conclusions

The focus of this research is to create better-performing store employee teams by creating a model to optimize the allocation of skill needs of stores and skill supply of employees, including limiting commute distances. To answer this optimization problem insights in stores are gathered by clustering stores using K-means. Employee skills supply is collected by performing a store employee survey. Socio-demographic characteristics of the service area of stores are researched to check whether it is possible to validate the skill demand of a store, which then can be used for new store locations where no sales data is available yet. Lastly, the skills supply and demand, including travelling distances, are optimized using Integer Linear Programming, which aims to improve both the skill gap and commute distances.

For clustering stores based on sales data, and the socio-demographic characteristics, the K-means algorithm is used since this allocates stores to a single cluster. For the clustering in step 1 of this research, sales data is used in quantity sold, per store, per product group. According to the Gap-statistics method, the optimal number of clusters is 7. By analysing the average values for the 13 biggest product groups the clusters are named. This created 2 health clusters, 2 baby clusters, 2 beauty clusters and a runshop/Travelling cluster of which the stores are mainly located on public transport stations. When looking at the spatial spread of the 6 other clusters, it turns out that, from appendix 6, the beauty clusters are mainly in more rural areas, where the health clusters are mainly in densely populated urban areas. The baby clusters are mainly in between, located just outside the main city centres.

In step 2, an employee survey is used to rate the skills and preferences of employees. For the same 13 product groups as used for cluster naming, employees are asked to rate themselves how well they perform in selling such products, as well as advising those product groups. It turned out that, due to the central tendency bias people tend to give above average ratings, which can be concluded from Figure 3.10.

Step 3 of this research focussed on getting insights into socio-demographic characteristics of service areas of stores. This research shows that clustering based on socio-demographic characteristics using K-means is possible and that it can help to research differences in the service areas. When compared to the clustering based on sales data it turns out that it is possible to see trends in the combination of sales data clustering and specific service area characteristics. This can be of benefit for the initial skill allocation for new locations when sales data is not yet available.

Lastly, in step 4, the problem of optimizing skills supply from employees with skills demand from stores is solved using Integer Linear Programming. This research shows that it is possible to minimize the residual of skills supply minus skills demand and taking into account commute distances, and thus travel costs, in the same process. The best values are found when using a λ value in the model of 0.2. This leads to an improvement of more than half of the stores. The average residual of skills is improved with **<confidential>**%, where commute distances would be reduced with **<confidential>**%.

All in all, this research shows that optimization improves the allocation of skills and spatial allocation of employees in order to enable the best possible store teams for a stores success.

5. Discussion & recommendations

The model as proposed in this research is not company specific, which means that, with slight modifications, it can be used for other organizations if the data is available. In this chapter, the impact, positive and negative, as well as the possibilities are discussed for the company, other drug stores, other retail chains and for the society in general.

The benefits for the company in this research are clear, which are better suiting employee teams in the stores while limiting travelling distances. This positively affects employee satisfaction, since commute time can cause stress and burn-out symptoms (Université de Montréal, 2015).

Since this research focusses on optimally allocating employees to stores based on skill supply and demand, there is a possible benefit for every retail chain or organization with multiple locations. For direct competitors of the case study organization, nationwide drug stores, the benefits of using such optimization are expected to be the same. Because this research focusses on skills optimization, it is important for a retail chain to have products which need specific skills to give advice and sell them, to benefit from the methodology from this research, since the service of an employee is an important driver for a stores' success according to Carpenter & Moore (2006). For example, grocery stores are expected to have very little benefit from such methodology, since most groceries do not need much specific skills and knowledge from an employee since it is mainly self-service. For grocery stores, a model that primarily minimizes commute distances for their employees might have more impact, which could still use some principles from this research. On the other hand, store types that sell products that require more specific skills, such as do-it-yourself stores, could use the model from this research. The expected benefits are expected to be comparable to the drug store example from this research. Although such type of stores is mainly located some more remote, for example not in the city centre, which might decrease the importance of commute distance. Those three examples show the possible benefits organizations could have from the principles of this research.

Using such an allocation model as proposed in this research could also have a negative impact. As can be seen from chapter 3.6.3, the output of the optimal allocation could involve that a significant part of the employees, 78% in the example of this research, should change location. People tend to have a bigger preference for things they previously encountered than new things, which is called the mere effect (Zizak & Reber, 2004). This is important for relocating employees since it can cause that switching such significant part of employees at once could be catastrophic for a company or other organization. With this, ethical questions come in mind, since the model, in theory, could decide, without human interaction, where to allocate an employee, which has a big impact on the life of an employee. The European Union's General Data Protection Regulation (GDPR), which is in effect since 2018, restricts "automated individual decision-making that will significantly affect users" (Goodman & Flaxman, 2016). Relocating employees with a model as proposed in this research, without any human interaction, will probably be not allowed to do according to the GDPR.

However, widespread implementation of such optimization techniques can have a positive societal impact as well. When a significant part of the nationwide companies would use such techniques to optimally allocate human resources to their locations, this will mean less travelling, which will possibly decrease the financial and emotional losses caused by traffic jams.

Although the methodology of this research is thoroughly researched, remarks can be made. For clustering, it might be worthy to use turnover per product group instead of volume. This gives impact to more expensive product groups, but might bias the data for high volume product groups with low prices. Possibly even a combination of both will be the best viable option. Clustering might also be improved by removing outliers in the sales data, such as the baby food milk powder example in this research, which might help to create better links between sales data and socio-demographic characteristics. Battaglia (2006) address the problem of outlier detection in time series. They propose a methodology to correct for this bias. Although we do not consider time series, yet, in this research, the proposed methodology might be used to estimate the outlier bias.

Further research is needed to perform regression analyses for the optimization results. This could involve which characteristics of an employee, such as function type, years of experience, motivation, education and age, have the biggest influence on employee skills and thus a stores' success. This could be of use to research the influence of the current employee store teams on the sales data, since the service and advice of an employee is important for a stores' success an proposed by Carpenter & Moore (2006) in chapter 1.2. Those current teams might create a bias of the sales data, and thus the skill demand in the optimization.

Furthermore, this research takes the skill need and skill supply as static data. However, sales data, and thus skills demand, is changing every day as a moving average. Same goes for the skills supply from an employee, which can change over time due to new education, interests or experience. When those data sets can be dynamically used in the optimization model, the skill match might be improved even further. History has many examples where reskilling of labour was needed to prevent job loss. An example is mine workers in South Africa, which have become redundant because of technology. Leeuw & Mtegha (2018) describe the role of modern mining technologies, which as a side effect caused unintended job losses. By reskilling those workers, job losses will be omitted. In the case of optimal employee allocation in this research, those job losses might not be a side effect. However, reskilling or reallocating employees to better suit their job will make employees happier while still improving stores success and limiting societal losses due to traffic jams.

6. References

Allen, J., & van der Velden, R. (2001). Educational mismatches versus skill mismatches: Effects on wages, job satisfaction, and on-the-job search. *Oxford Economic Papers*, *53*(3), 434–452. https://doi.org/10.1093/oep/53.3.434

Barnhorst, B. S., Betro, S. A., & Haq, T. U. (2001). Intelligent System for Dynamic Resource Management.

Battaglia, F. (2006). Bias correction for outlier estimation in time series. *Journal of Statistical Planning and Inference*, *136*(11), 3904–3930. https://doi.org/10.1016/j.jspi.2005.04.002

Bruecker, P. De, Bergh, J. Van Den, Beliën, J., & Demeulemeester, E. (2015). Workforce planning incorporating skills : State of the art. *European Journal of Operational Research*, *243*, 1–16. https://doi.org/10.1016/j.ejor.2014.10.038

Bussieck, M. R., Winter, T., & Zimmermann, U. T. (1997). Discrete Optimization in Public Transportation, *79*(october), 415–444. https://doi.org/10.1007/BF02614327

Carpenter, J. M., & Moore, M. (2006). Consumer demographics , store attributes , and retail format choice in the US grocery market. https://doi.org/10.1108/09590550610667038

Columb, M. O., & Stevens, A. (2008). Power analysis and sample size calculations. *Current Anaesthesia and Critical Care*, *19*(1), 12–14. https://doi.org/10.1016/j.cacc.2007.03.011

Deepashri, K. S., & Kamath, A. (2017). Survey on Techniques of Data Mining and its Applications. *International Journal of Emerging Research in Management &Technology*, *ISSN*(62), 2278–9359. Retrieved from https://pdfs.semanticscholar.org/b738/3df4705133a132f58104b514b80555fe78cb.pdf

Deniz, N., Noyan, A., & Ertosun, Ö. G. (2015). Linking Person-job Fit to Job Stress: The Mediating Effect of Perceived Person-organization Fit. *Procedia - Social and Behavioral Sciences*, *207*, 369–376. https://doi.org/10.1016/j.sbspro.2015.10.107

Dolnicar, S. (2003). A Review of Unquestioned Standards in Using Cluster Analysis for Data-Driven Market Segmentation A Review of Unquestioned Standards in Using Cluster Analysis for Data-. *Australasian Journal of Market Research*, *2002*(December), 2–4.

Dramowicz, E. (2005). Retail Trade Area Analysis Using the Huff Model, 1–12.

Gentle, J. E., Kaufman, L., & Rousseuw, P. J. (1991). Finding Groups in Data: An Introduction to Cluster Analysis. *Biometrics*, *47*(2), 788. https://doi.org/10.2307/2532178

Gokbayrak, K., & Kocaman, A. S. (2017). A distance-limited continuous location-allocation problem for spatial planning of decentralized systems. *Computers and Operations Research*, *88*, 15–29. https://doi.org/10.1016/j.cor.2017.06.013

Goodchild, M. F. (2009). First Law of Geography. *International Encyclopedia of Human Geography*, 179–182. https://doi.org/10.1016/b978-008044910-4.00438-7

Goodman, B., & Flaxman, S. (2016). European Union regulations on algorithmic decision-making and a “right to explanation,” (Whi), 26–30. https://doi.org/10.1609/aimag.v38i3.2741

Goos, M. (2018). The importance of worker competencies in retail services: Study Description. Utrecht.

Hastle, T., Tibshirani, R., & Walther, G. (2001). Estimating the number of clusters in a data set via the gap statistic.pdf. *Journal of the Royal Statistical Society*.

Heckman, J. J., Lochner, L. J., & Todd, P. E. (2005). Earnings functions, rates of return and treatment effects: The Mincer equation. *National Bureau of Economic Research*, (1700). Retrieved from http://www.nber.org/papers/w11544

Holy, V., Sokol, O., & Cerny, M. (2017). Clustering retail products based on customer behaviour. *Applied Soft Computing Journal*, *60*, 752–762. https://doi.org/10.1016/j.asoc.2017.02.004

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, *31*(8), 651–666. https://doi.org/10.1016/j.patrec.2009.09.011

Khodaparasti, S., Bruni, M. E., Beraldi, P., Maleki, H. R., & Jahedi, S. (2018). A multi-period location-allocation model for nursing home network planning under uncertainty. *Operations Research for Health Care*, *18*, 4–15. https://doi.org/10.1016/j.orhc.2018.01.005

Kilci, F., Kara, B. Y., & Bozkaya, B. (2015). Locating temporary shelter areas after an earthquake: A case for Turkey. *European Journal of Operational Research*, *243*(1), 323–332. https://doi.org/10.1016/j.ejor.2014.11.035

Leeuw, P., & Mtegha, H. (2018). The significance of mining backward and forward linkages in reskilling redundant mine workers in South Africa. *Resources Policy*, *56*(June 2017), 31–37. https://doi.org/10.1016/j.resourpol.2018.02.004

Mason, G. A., & Jacobson, R. D. (2007). Fuzzy Geographically Weighted Clustering. *Proc. of the 9th International Conference on Geocomputation*, (1998), 1–7.

Nerurkar, P., Shirke, A., Chandane, M., & Bhirud, S. (2018). Empirical Analysis of Data Clustering Algorithms. *Procedia Computer Science*, *125*, 770–779. https://doi.org/10.1016/j.procs.2017.12.099

Oswin, A., Fischer, A., Fischer, F., Meier, J. F., Pilz, A., Staněk, R., & Pferschy, U. (2017). Minimization and maximization versions of the quadratic travelling salesman problem. *Optimization*, *1934*, 1–22. https://doi.org/10.1080/02331934.2016.1276905

Pace, W. (2007). K-means methodology.

Pan, Y., & Zinkhan, G. M. (2006). Determinants of retail patronage: A meta-analytical perspective. *Journal of Retailing*, *82*(3), 229–243. https://doi.org/10.1016/j.jretai.2005.11.008

Paul, J. Al, & Batta, R. (2008). Models for hospital location and capacity allocation for an area prone to natural disasters Rajan Batta. *International Journal of Operation*, *3*(5), 473–496. https://doi.org/10.1504/IJOR.2008.019170

Peng, D. (2018). Postdoc TU Delft. Retrieved from http://www1.informatik.uni-wuerzburg.de/en/staff/alumni/peng-dongliang/

Peng, D., Wolff, A., & Haunert, J. (2017). Finding Optimal Sequences for Area Aggregation – A ⋆ vs . Integer Linear Programming. *ACM Trans. Spatial Algorithms Syst.*, *0*(0).

Rasmussen, T., Ulrich, D., & Likert, R. (2015). Learning from practice: how HR analytics avoids being a management fad. *Organizational Dynamics*, *44*, 236–242.

https://doi.org/10.1016/j.orgdyn.2015.05.008

Rigby, D. K., & Vishwanath, V. (2016). Localization : The Revolution, 1–20.

Risbeck, M. J., Maravelias, C. T., Rawlings, J. B., & Turney, R. D. (2017). A mixed-integer linear programming model for real-time cost optimization of building heating, ventilation, and air conditioning equipment. *Energy and Buildings*, *142*, 220–235. https://doi.org/10.1016/j.enbuild.2017.02.053

Ros, F., & Guillaume, S. (2016). DENDIS: A new density-based sampling for clustering algorithm. *Expert Systems with Applications*, *56*, 349–359. https://doi.org/10.1016/j.eswa.2016.03.008

Röttger, R. (2018). PhD Utrecht University. Retrieved from https://www.uu.nl/medewerkers/RCRottger

See, L., & Openshaw, S. (2001). Fuzzy geodemographic targeting. *Regional Science in Business*, 269–282.

Senthilkumar, M., Nallakaruppan, M. K., Chandrasegar, T., & Prasanna, S. (2014). A modified and efficient genetic algorithm to address a travelling salesman problem. *International Journal of Applied Engineering Research*, *9*(10), 1279–1288. Retrieved from http://search.ebscohost.com.ezproxy.liv.ac.uk/login.aspx?direct=true&db=edselc&AN=edselc.2-52.0-84897876277&site=eds-live&scope=site

Sinova, B., Rosa, M., & Ángeles, M. (2014). Central tendency for symmetric random fuzzy numbers q. *Information Sciences*, *278*, 599–613. https://doi.org/10.1016/j.ins.2014.03.077

Söderlund, M. (2018). The proactive employee on the floor of the store and the impact on customer satisfaction. *Journal of Retailing and Consumer Services*, *43*(October 2017), 46–53. https://doi.org/10.1016/j.jretconser.2018.02.009

Sprott, D. E., Spangenberg, E. R., Block, L. G., Fitzsimons, G. J., Morwitz, V. G., & Williams, P. (2006). The question – behavior effect : What we know and where we go from here. *Social Influence*, *1*(2), 128–137. https://doi.org/10.1080/15534510600685409

Statistics Netherlands. (2018). CBS Open data StatLine. Retrieved June 18, 2018, from https://opendata.cbs.nl/statline/portal.html?_la=nl&_catalog=CBS

Tatiana, K., & Mikhail, M. (2018). Market basket analysis of heterogeneous data sources for recommendation system improvement. *Procedia Computer Science*, *136*, 246–254. https://doi.org/10.1016/j.procs.2018.08.263

Université de Montréal. (2015). You're driving yourself to burnout, literally. Retrieved June 26, 2018, from https://www.sciencedaily.com/releases/2015/05/150526085136.htm

Valle, M. A., Ruz, G. A., & Morrás, R. (2018). Market basket analysis : Complementing association rules with minimum spanning trees. *Expert Systems With Applications*, *97*, 146–162. https://doi.org/10.1016/j.eswa.2017.12.028

Vieira, B., Demirtas, D., van de Kamer, J. B., Hans, E. W., & van Harten, W. (2018). A mathematical programming model for optimizing the staff allocation in radiotherapy under uncertain demand. *European Journal of Operational Research*, *270*(2), 709–722. https://doi.org/10.1016/j.ejor.2018.03.040

Whize. (2018). Whize - Slim gebruik van data voor klantgesegmenteerde communicatie. Retrieved from https://www.whooz.nl/whize

Winckel, C. P., Reznick, R. K., Cohen, R., & Taylor, B. (1994). Reliability and Construct Validity of a Structured Technical Skills Assessment Form. *American Journal of Surgery*, *167*(April), 423–427. https://doi.org/10.1016/0002-9610(94)90128-7

Wong, K.-C. (2015). A Short Survey on Data Clustering Algorithms, 64–68. https://doi.org/10.1109/ISCMI.2015.10

Wu, X., Lu, Y., Xu, H., Lv, M., Hu, D., He, Z., … Feng, Y. (2018). Trends in Food Science & Technology Challenges to improve the safety of dairy products in China. *Trends in Food Science & Technology*, *76*(October 2016), 6–14. https://doi.org/10.1016/j.tifs.2018.03.019

Yuen, K. F., Loh, H. S., Zhou, Q., & Wong, Y. D. (2018). Determinants of job satisfaction and performance of seafarers. *Transportation Research Part A: Policy and Practice*, *110*(November 2017), 1–12. https://doi.org/10.1016/j.tra.2018.02.006

Zhao, D., Li, J., Tan, Y., Yang, K., Ge, B., & Dou, Y. (2018). Optimization adjustment of human resources based on dynamic heterogeneous network. *Physica A: Statistical Mechanics and Its Applications*, *503*, 45–57. https://doi.org/10.1016/j.physa.2018.02.168

Zizak, D. M., & Reber, A. S. (2004). Implicit preferences: The role(s) of familiarity in the structural mere exposure effect. *Consciousness and Cognition*, *13*(2), 336–362. https://doi.org/10.1016/j.concog.2003.12.003

Appendix 1: Calculating cluster values, example data

Table 0.1 Example data sales data per category, quantitative and relative

Store number	Sales Health	Sales Beauty	Sales Baby	Sales Total
1234	200	500	300	800
5678	300	100	100	500

Table 0.2 Example data available shelf space per category, quantitative and relative

Store number	Shelf space Health	Shelf space Beauty	Shelf space Baby	Shelf space Total
1234	50	100	50	200
5678	30	40	30	100

Table 0.3 Average sales per product category per store

Store number	Cluster Health	Cluster Beauty	Cluster Baby
1234	200/50 = 4	500/100 = 5	300/50 = 6
5678	300/30 = 10	100/40 = 2.5	100/30 ≈ 3.33

Table 0.4 Average sales per shelf unit

Store number	Average sales per shelf unit
1234	800/200 = 4
5678	500/100 = 5

Table 0.5 Cluster values, normalised for clustering

Store number	Cluster Health	Cluster Beauty	Cluster Baby
1234	4/4 = 1	5/4 = 1.25	6/4 = 1.5
5678	10/5 = 2	2.5/5 = 0.5	3.33/5 ≈ 0.666

Appendix 2: Calculating store specific points, example data

Table 0.1 Assigning points to all clusters for a specific store

Cluster	Cluster value	Relative store size (index of sales)	Employee needs for specific store ("points")
Baby	30	120	(30*120)/100 = 36
Health	20	120	(20*120)/100 = 24
Beauty	40	120	(40*120)/100 = 48
Vitamins	10	120	(10*120)/100 = 12

All employees get a rating for each cluster ranging from 1 to n, in the example of Table 0.1 four clusters are used. In this example the amount of points for this specific employee is calculated using the formula in chapter 3.4. This is a sample calculation. In the final allocation the total points of the stores needs must be close to the amount of points as assigned to employees, because all stores have about the right number of employees. However, the spread of preference points over the different clusters can differ.

Table 0.2 Assigning point to employee preferences

Cluster	Employee rating	Employee contract hours	Employee preferences "points"
Baby	1	32	1*(32/10) = 1*3.2=3.2
Health	3	32	3*(32/10) = 3*3.2=9.6
Beauty	4	32	4*(32/10) = 4*3.2=12.8
Vitamins	2	32	2*(32/10) = 2*3.2=6.4
Total	10		32

Appendix 3: Optimization example data

Table 0.1 Calculation data optimization example

Cluster	Store A	Store B	Total
Baby	36-20-10 = 6	20-5-8 = 7	13
Health	24-25-10 = -11	30-10-20 = 0	-11
Beauty	48-14-30 = 4	30-15-25 = -5	-1
Vitamins	12-7-5 = 0	24-15-10 = -1	-1
Total	-1	1	0

Appendix 4: Employee questionnaire

Jouw passies en expertise in de spotlight!

Ben jij 16 jaar of ouder en werk je in een winkel van Etos? Dan willen wij je vragen om mee te doen aan een **onderzoek van de Universiteit Utrecht**. Dit onderzoek gaat over de rol van passies en expertise in de detailhandel. Meedoen is helemaal vrijwillig.
Terwijl je de vragen beantwoordt leer je veel over jezelf. Je wordt bewuster van **jouw passies en expertise**. Op het einde van deze vragenlijst heb je de optie om een overzicht toegestuurd te krijgen van jouw passies en expertises dat je van ons zal ontvangen nadat we alle gegevens hebben verwerkt.
Het is voor ons belangrijk dat **jouw privacy wordt beschermd**. Daarom hebben alleen leden van het onderzoeksteam van de Universiteit Utrecht toegang tot jouw antwoorden. Met Etos worden alleen onderzoeksresultaten anoniem en op groepsniveau gedeeld, zodat jij als individu niet te herkennen bent.
In deze vragenlijst zijn sommige persoonlijke vragen. Als je een vraag liever niet wilt beantwoorden, dan hoeft dat ook niet. Wij willen wel benadrukken dat wij jouw individuele gegevens niet gaan delen. Al jouw gegevens worden vertrouwelijk behandeld.
Meedoen? Het invullen van de vragenlijst duurt ongeveer **10 minuten**. Wil je meedoen, dan hebben wij wel jouw schriftelijke toestemming nodig. Die kun je geven door een vinkje te zetten onderaan. Meer informatie onder “gedragslijn tonen”.

Vragen of klachten? Neem contact op met onderzoeksteam door een email te sturen naar xpertise.use@uu.nl

1. Filiaalnummer

Wat is het filiaalnummer van de winkel waar jij op dit moment werkt?

XXXX (validation wordt gebruikt)

2. Expertise

Waar ligt **JOUW EXPERTISE** en wat vind jij leuk om te doen?

Hoe goed ben je in het verkopen van:

Dubbelklik of klik-en-sleep items van de linkerlijst naar de rechterlijst. Zet in de rechterlijst de items op volgorde van belangrijkheid. Zet het voor u belangrijkste item bovenaan.

Jouw keuzes	Jouw rangschikking
Beauty	
Care	
Health	
Baby	

Hoe graag adviseer je klanten over:

Dubbelklik of klik-en-sleep items van de linkerlijst naar de rechterlijst. Zet in de rechterlijst de items op volgorde van belangrijkheid. Zet het voor u belangrijkste item bovenaan.

Jouw keuzes	Jouw rangschikking
Beauty	
Care	
Health	
Baby	

Hoe graag wil je meer kennis opdoen over:

Dubbelklik of klik-en-sleep items van de linkerlijst naar de rechterlijst. Zet in de rechterlijst de items op volgorde van belangrijkheid. Zet het voor u belangrijkste item bovenaan.

Jouw keuzes	Jouw rangschikking
Beauty	
Care	
Health	
Baby	

Op een schaal van 1 (niet heel goed) tot 7 (uitstekend)

Hoe goed ben je in het verkopen van:

	1	2	3	4	5	6	7
Make-up	O	O	O	O	O	O	O
Geuren	O	O	O	O	O	O	O
Gezichtsverzorging	O	O	O	O	O	O	O
Scheren & ontharen	O	O	O	O	O	O	O
Haar	O	O	O	O	O	O	O
Deodorant	O	O	O	O	O	O	O
Bad & Douche	O	O	O	O	O	O	O
Mondverzorging	O	O	O	O	O	O	O

Bewegen, Pijn & Griep	O	O	O	O	O	O	O
Weerstand, Gezin & Balans	O	O	O	O	O	O	O
Babyverzorging	O	O	O	O	O	O	O
Babyvoeding	O	O	O	O	O	O	O
Luiers	O	O	O	O	O	O	O
Onderweg & Impulsaankoop	O	O	O	O	O	O	O

? Onderweg & Impulsaankoop: bijv. gekoelde dranken en zoetwaren, kassa-meter, panties

Op een schaal van 1 (niet heel graag) tot 7 (heel graag)

Hoe graag adviseer je klanten over:

	1	2	3	4	5	6	7
Make-up	O	O	O	O	O	O	O
Geuren	O	O	O	O	O	O	O
Gezichtsverzorging	O	O	O	O	O	O	O
Scheren & ontharen	O	O	O	O	O	O	O
Haar	O	O	O	O	O	O	O
Deodorant	O	O	O	O	O	O	O
Bad & Douche	O	O	O	O	O	O	O
Mondverzorging	O	O	O	O	O	O	O
Bewegen, Pijn & Griep	O	O	O	O	O	O	O
Weerstand, Gezin & Balans	O	O	O	O	O	O	O
Babyverzorging	O	O	O	O	O	O	O
Babyvoeding	O	O	O	O	O	O	O
Luiers	O	O	O	O	O	O	O
Onderweg & Impulsaankoop	O	O	O	O	O	O	O

? Onderweg & Impulsaankoop: bijv. gekoelde dranken en zoetwaren, kassa-meter, panties

3. Passies

Wat zijn JOUW PASSIES? Denk niet te lang over de vragen na en volg gewoon jouw eerste gedachte.

Volg jij beauty-, fashion- of lifestyle-influencers op social media, bijvoorbeeld op instagram of youtube, zoals Anna Nooshin of Nikkietutorials?

O Ja O Nee

Indien ja:

Welke van de volgende types social influencers volg jij vooral? Kies een van de volgende mogelijkheden.

Ik volg vooral…

O Beauty-influencers (focus op make-up)

O Fashion-influencers (focus op fashion)

O Lifestyle-influencers (focus op fitness, food, design)

O Anders, namelijk: _____

O geen antwoord

? Wij definiëren de social influencer types als volgt: beauty-influencers focussen vooral op make-up, fashion-influencers focussen vooral op fashion en lifestyle-influencers focussen vooral op fitness, food of design.

Indien nee:

Wat is de voornaamste reden om geen beauty-, fashion- of lifestyle-influencers te volgen op sociale media?

O Geen interesse

O Kennis uit andere bronnen

O Te druk

O Weet niet welke er zijn

O Anders:

O Geen antwoord

In het afgelopen halve jaar, heb jij drogisterij vakbladen gelezen?

O Ja O Nee

Indien ja:

Welke van de onderstaande vakbladen lees je vooral? Kies een van de volgende antwoorden:

O Lijfblad

O Kosmetiek

O Careality

O Whocares Zelfzorg

O DrogistenWeekblad

O Anders, namelijk: _____

? Dit kunnen fysieke of online vakbladen zijn.
Hoe vaak lees je vakbladen ongeveer? Kies een van de volgende antwoorden:
O maximaal een keer per jaar

O maximaal twee keer per jaar

O om de drie maanden

O minimaal elke twee maanden

O minimaal elke maand

O Anders, namelijk: _____

O geen antwoord

Indien nee:

Wat was de voornaamste reden om geen vakbladen te lezen? Kies een antwoord.

O Geen interesse

O Kennis uit andere bronnen

O Te druk

O Weet niet welke er zijn

O Anders, namelijk:_____

In de afgelopen 6 maanden, heb jij extra (online-)cursussen of trainingen gevolgd in je vrije tijd?

O Ja O Nee

Indien ja:

Ik heb deze extra (onine-) cursussen of trainingen gevolgd…

Meerdere antwoorden mogelijk

O op drogiweb

O van leveranciers, bijvoorbeeld “L´Oréal” of “vsm”

O Anders, namelijk: _____

Welke van de onderstaande onderwerpen beschrijft het beste waar de cursus(sen) of training(en) over gingen? Kies een van de volgende antwoorden

O Gezondheid/zelfzorg

O Beauty/make-up

O Verzorging

O Baby

O Alle thema´s

O Anders, namelijk: _____

O Geen antwoord

Hoe vaak volg jij cursussen ongeveer? Kies een van de volgende antwoorden

O maximaal een keer per jaar

O maximaal twee keer per jaar

O een cursus om de drie maanden

O minimaal elke twee maanden

O minimaal elke maand

O Anders, namelijk: _____

O Geen antwoord

Indien nee:

Wat was de voornaamste reden om geen (online-)cursussen of trainingen te volgen? Kies een van de volgende antwoorden

O Geen interesse

O Kennis uit andere bronnen

O Te druk

O Weet niet welke er zijn

O Te duur

O Anders, namelijk:_____

O Geen antwoord

De volgende twee vragen gaan over sporten. Of je nou weinig sport of juist veel - dit kan laten zien welke soorten passies jij hebt. Het ene gedrag is niet beter of slechter dan het andere.

In de afgelopen 7 dagen, op hoeveel van deze dagen heb jij tenminste 20 minuten lichamelijke activiteiten verricht waarbij jouw hartslag omhoog is gegaan (bijv. je was aan het zweten en je had moeite met ademen), zoals wielrennen, hardlopen, zwemmen?

Kies een van de volgende antwoorden

O 0

O 1

O 2

O 3

O 4

O 5

O 6

O 7

Als je nu terug denkt aan de afgelopen 4 weken: gemiddeld, op hoeveel dagen per week heb jij tenminste 20 minuten lichamelijke activiteiten verricht waarbij jouw hartslag omhoog is gegaan (bijv. je was aan het zweten en je had moeite met ademen), zoals wielrennen, hardlopen, zwemmen?

Kies een van de volgende antwoorden

O 0

O 1

O 2

O 3

O 4

O 5

O 6

O 7

Stel dat familie of vrienden jou om advies vragen. Welke van de onderstaande mogelijkheden beschrijft het beste om welke soort advies het vooral zou gaan?

Kies een van de volgende antwoorden

O Advies over gezondheid (bijv. wat je moet doen als je ziek bent)

O Advies over make-up (bijv. welke make-up het beste bij jou past)

O Advies over haar- en huidverzorging (bijv. hoe jouw haar mooi en gezond uit gaat zien)

O Advies over alles rond om baby´s (bijv. welke verzorging het beste past)

O geen antwoord

Welke van de onderstaande statements beschrijft het beste wat je op dit moment interessant vind?

Kies een van de volgende antwoorden

O de werking van het menselijke lichaam

O de nieuwste make-up trends

O de nieuwste tips voor haar- en huidverzorging

O de beste voeding en verzorging voor baby´s

O geen antwoord

Op een schaal van 1 (niet heel makkelijk) tot 7 (heel makkelijk)

Hoe makkelijk vind jij het om op klanten af te stappen?

1	2	3	4	5	6	7	geen antwoord
O	O	O	O	O	O	O	O

4. Jouw baan

Welke eigenschappen heeft JOUW BAAN? Wij vragen dit om beter te begrijpen hoe jouw baan eruit ziet en hoe je passies en expertise je helpen in je baan.

Vul deze vragen zo precies mogelijk in. Mocht je de exacte dag bijvoorbeeld niet meer weten, vul dan de eerste dag van de maand in.

Wat is je huidige functie in de Etos winkel waar je nu werkt?

Kies een van de volgende antwoorden

O Vulploegmedewerker

O Aankomend verkoopmedewerker

O Verkoopmedewerker

O Verkoop coördinator

O Winkelmanager

O Anders, namelijk:_____

O geen antwoord

Wanneer ben je begonnen in je huidige functie in de Etos winkel waar je nu werkt?

(dd-mm-yyyy)

Wanneer ben je begonnen bij Etos?

(dd-mm-yyyy)

? Dit kan maar hoeft niet overeenkomen met wanneer je bent begonnen bij de Etos winkel waar je nu werkt.

Wanneer ben je begonnen met werken?

(dd-mm-yyyy)

? Dit kan maar hoeft niet overeenkomen met wanneer je bent begonnen bij Etos.

Welke type contract heb jij?

Kies een van de volgende antwoorden

O Onbepaalde tijd

O Bepaalde tijd

O Stage

O Anders, namelijk: _____

O geen antwoord

Hoeveel contracturen per week heb je?

_____ contracturen per week

In de afgelopen vier weken: hoeveel uren heb je gemiddeld per week gewerkt?

_____ uren per week

In de afgelopen 7 dagen: op welke dagen heb je gewerkt? Kies de dagen die voor jou van toepassing zijn.

Meerdere antwoorden mogelijk

O Maandag

O Dinsdag

O Woensdag

O Donderdag

O Vrijdag

O Zaterdag

O Zondag

O geen antwoord

In de afgelopen vier weken: heb jij altijd op dezelfde dagen van de week gewerkt?

O Ja O Nee

Op welke dagen werk jij meestal?

Meerdere antwoorden mogelijk

O Maandag

O Dinsdag

O Woensdag

O Donderdag

O Vrijdag

O Zaterdag

O Zondag

O altijd flexibel

O geen antwoord

Wat is de reistijd van jouw huis naar de Etos winkel waar je op dit moment werkt?

_____ minuten van deur tot deur

Wat vind je van deze reistijd? Mijn reistijd is…

Kies een van de volgende antwoorden

O niet te lang

O te lang

O geen antwoord

In welke van de onderstaande categorieën valt jouw netto maandloon bij Etos?

Kies een van de volgende antwoorden

O Minder dan 500€

O 500€ tot 999€

O 1000€ tot 1499€

O 1500€ tot 1999€

O 2000€ tot 2499€

O 2500€ en meer

O geen antwoord

Op een schaal van 1 (helemaal niet tevreden) tot 7 (helemaal)

Hoe tevreden ben je met je huidige baan bij Etos?

1	2	3	4	5	6	7	geen antwoord
O	O	O	O	O	O	O	O

Heb je naast jouw huidige functie bij Etos nog een andere baan?

O Ja O Nee

5. Jouw opleiding

Wat is het niveau van jouw hoogst behaalde diploma in voltijds onderwijs?

Kies een van de volgende antwoorden

O Geen diploma

O Praktijkonderwijs

O Voorbereidend beroepsonderwijs (vbo)

O Middelbaar algemeen vormend onderwijs (mavo)

O Voorbereidend middelbaar beroepsonderwijs (vmbo)

O Hoger algemeen voortgezet onderwijs (havo)

O Voorbereidend wetenschappelijk onderwijs (vwo)

O Leerlingwezen (primair, voortgezet en tertiair)

O Middelbaar beroepsonderwijs (mbo)

O Hoger beroepsonderwijs (hbo)

O Wetenschappelijk onderwijs (wo, bachelor)

O Wetenschappelijk onderwijs (wo, master)

O Anders, namelijk: _____

O geen antwoord

In welke studierichting behaalde je dit diploma?

_____

? Je kunt hier de naam van de studierichting aangeven of indien je de exacte naam niet kent, geef een korte omschrijving van de richting van jouw studie

Heb jij een drogisterij diploma behaald, zoals `verkoper in de drogisterij´, `assistent drogist´ of `drogist´?

O Ja O Nee

Indien ja:

Wat is het niveau van jouw hoogst behaalde drogisterij diploma?

Kies een van de volgende antwoorden

O Verkoper in de drogisterij

O Assistent-drogist

O Drogist

O Anders, namelijk: _____

O geen antwoord

Volg je jaarlijkse nascholing via het “Zorg dat je bijblijft” platform?

O Ja O Nee

Heb je je hoogste drogisterij opleiding gehaald voordat je begon bij Etos?

O Ja O Nee

6. Algemene informatie

Ten slotte willen we jou enkele algemene informatie vragen zodat we het onderzoek beter kunnen kaderen. Dit zijn korte vragen over bijvoorbeeld jouw thuissituatie. Ook hier geldt: al jouw gegevens worden vertrouwelijk behandeld zodat jouw privacy wordt beschermd.

Wat is je geslacht?

O Vrouw O Man

Wanneer ben je geboren?

yyyy

Wat zijn de eerste vier cijfers van de postcode waar je woont?

_____

? bijvoorbeeld 1234 (validation wordt gebruikt)

Welke van de volgende antwoorden beschrijft het beste jouw woonsituatie?

Kies een van de volgende antwoorden

Ik woon …

O Samen met mijn partner

O Alleen

O Bij mijn ouders/ bij een van mijn ouders

O Met huisgenoten

O Anders, namelijk:_____

O geen antwoord

Heb je kinderen?

O Ja O Nee

Indien `ja`:

Hoeveel kinderen wonen er bij je?

Kies een van de volgende antwoorden

O 1

O 2

O 3

O 4+

O geen antwoord

In welke van de volgende leeftijdscategorieën valt jouw kind/vallen jouw kinderen? Kies de mogelijkheid/mogelijkheden die voor jou van toepassing zijn:

O 0 tot en met 2 jaar

O 3 tot en met 6 jaar

O 7 tot en met 10 jaar

O 11 tot en met 14 jaar

O 15 tot en met 18 jaar

O Ouder dan 18 jaar

O geen antwoord

Ben je een scholier of student?

O Ja O Nee

Dit is het einde van de bevraging. Dank je voor je antwoorden!

Geef ons je emailadres indien je een overzicht wilt ontvangen van jouw passies en expertises nadat we alle gegevens hebben verwerkt:

_____

Heb je nog opmerkingen over deze vragenlijst?
(Vul uw antwoord hier in:
open answer box)

Verzend uw enquête.

Bedankt voor uw deelname aan deze enquête.

Appendix 5: R code clustering

```r
#load and normalize data
dataYannick=read.csv('Final clusters nonnorm 13 groups.csv')
data.scaled13 <- scale(dataYannick[,-1])

#determine optimal number of clusters
fviz_nbclust(data.scaled13, kmeans, method = "silhouette")+
  labs(subtitle = "Silhouette method")

#set number of cluster
k=5

#clustering including visualization
km.res <- kmeans(data.scaled13, k, nstart=5)
fviz_cluster(km.res, data.scaled13 ,show.clust.cent = T,repel = T, stand = FALSE, geom = "point",  pointsize = 1)

# PCA circle plot
library(ade4)
pca1=dudi.pca(data.scaled13)
s.corcircle(pca1$co, full = TRUE, box = FALSE,clabel = 0.9)

# Heatmap
library(reshape2)
cormat=cor(data.scaled13)
melted_cormat <- melt(cormat)
library(ggplot2)
ggplot(data = melted_cormat, aes(Var2, Var1, fill = value))+
  geom_tile(color = "white")+
  scale_fill_gradient2(low = "blue", high = "red", mid = "white",
                       midpoint = 0, limit = c(-1,1), space = "Lab",
                       name="Pearson\nCorrelation") +
  theme_minimal()+
  theme(axis.text.x = element_text(angle = 90, vjust = 0,
                                   size = 11, hjust = 1))+
  coord_fixed()

```

Figure 0.1 R code clusters 13 product groups

```r
#load and normalize data
dataYannick=read.csv('Final clusters nonnorm 32 groups.csv')
data.scaled32 <- scale(dataYannick[,-1])

#determine optimal number of clusters
set.seed(123)
fviz_nbclust(data.scaled32, kmeans, nstart = 25,  method = "gap_stat", nboot = 50)+
  labs(subtitle = "Gap statistic method")

#set number of cluster
k=7
#clustering including visualization
km.res <- kmeans(data.scaled32, k, nstart=4)
fviz_cluster(km.res, data.scaled32 ,show.clust.cent = T,repel = T, stand = FALSE, geom = "point",  pointsize = 1)

# PCA circle plot
library(ade4)
pca1=dudi.pca(data.scaled32)
s.corcircle(pca1$co, full = TRUE, box = FALSE,clabel = 0.9)

# Heatmap
library(reshape2)
cormat=cor(data.scaled32)
melted_cormat <- melt(cormat)
library(ggplot2)
ggplot(data = melted_cormat, aes(Var2, Var1, fill = value))+
  geom_tile(color = "white")+
  scale_fill_gradient2(low = "blue", high = "red", mid = "white",
                       midpoint = 0, limit = c(-1,1), space = "Lab",
                       name="Pearson\nCorrelation") +
  theme_minimal()+
  theme(axis.text.x = element_text(angle = 90, vjust = 0,
                                   size = 8, hjust = 1))+
  coord_fixed()

```

Figure 0.2 R code clusters 32 product groups

Appendix 6: Clustering maps – sales data

A) Beauty

<confidential>

B) Health

<confidential>

C) Baby

<confidential>

D) Runshop

<confidential>

Appendix 7: Integer Linear Program code

```
using System;
using System.Collections.Generic;
using System.ComponentModel;
using System.Data;
using System.Drawing;
using System.IO;
using System.Linq;
using System.Text;
using System.Threading.Tasks;
using System.Windows.Forms;
using System.Globalization;

using CsvHelper;

using ILOG.Concert;
using ILOG.CPLEX;
```

Figure 0.1 Importing systems for Integer Linear Program

```
            string strPathALocate = System.IO.Path.GetFullPath(@"..\..\..\..\..\");
            string strPathData = strPathALocate + "AlteryX\\Etos_employees\\Output";

            //string strPathALocate = System.IO.Path.GetFullPath(@"..\..\..\..\");
            //string strPathData = strPathALocate + "Data";

            //read store matrix
            var strStoreLtLt = ReadData(strPathData + "\\2000. Optimization data set stores.csv");
            var adblStore = GetValues(strStoreLtLt);

            //read employee matrix
            var strEmployeeLtLt = ReadData(strPathData + "\\2000. Optimization data set employees.csv");
            var adblEmployee = GetValues(strEmployeeLtLt);

            //read distance matrix
            var strDistanceLtLt = ReadData(strPathData + "\\3000. Distance matrix.csv");
            var adblDistance = GetValues(strDistanceLtLt);
```

Figure 0.2 Importing data matrices to Integer Linear Program

```
            //every employee must be assigned to exactly one store <-> the distance between them must be small
            double Maximumdistance = 45;  //km 'as the crow flies'

            //2D variables
            IIntVar[][][] var2;
            double dblLambda = 0.75; //'cost' of travel distance
            PopulateByRow(cplex, out var2, adblStore, adblEmployee, adblDistance, Maximumdistance, dblLambda);

            double dblTimelimit = 10000; //Maximum time in seconds the model has to run to find the most optima
            cplex.SetParam(Cplex.DoubleParam.TiLim, dblTimelimit); //stop if cplex can't find a solution in a
```

Figure 0.3 Integer Linear Programming parameters

```
internal static void PopulateByRow(IMPModeler model, out IIntVar[][][] var2,
   double[,] adblStore, double[,] adblEmployee, double[,] adblDistance,
   double dblDisThreshold, double dblLambda)
{
    var2 = new IIntVar[1][][];


    //i: index for stores
    //j: index for attributes
    //k: index for employees
    var intEmployeeCount = adblEmployee.GetLength(0); //number of all employess
    var intStoreCount = adblStore.GetLength(0);
    var x = new IIntVar[intEmployeeCount][];  //main variable
    for (int k = 0; k < intEmployeeCount; k++)
    {
        x[k] = model.BoolVarArray(intStoreCount);
    }
    var2[0] = x; //in order to display x later

    //y and z are assistant variables
    //because we want to transform absolute expressions to linear expressions
    //e.g., https://www.researchgate.net/post/How_can_I_make_linear_the_Absolute_Value_function_x_in_optimization_problems
    //e.g., https://optimization.mccormick.northwestern.edu/index.php/Optimization_with_absolute_values
    var intAttributeCount = adblEmployee.GetLength(1);
    var y = new INumVar[intStoreCount][];
    var z = new INumVar[intStoreCount][];
    double dblLB = 0;
    double dblUB = 100000; //we will define this upper bound according to our data; the smaller, the better
    for (int i = 0; i < intStoreCount; i++)
    {
        y[i] = model.NumVarArray(intAttributeCount, dblLB, dblUB);
        z[i] = model.NumVarArray(intAttributeCount, dblLB, dblUB);
    }

    //generate the objective that we want to minimize
    //lne: linear number expression
    ILinearNumExpr lneCost = model.LinearNumExpr(); //objective linear expression
    for (int i = 0; i < intStoreCount; i++) //want to minimize differences between demands and supply of skills
    {
        for (int j = 0; j < intAttributeCount; j++)
        {
            lneCost.AddTerm(y[i][j], 1);
            lneCost.AddTerm(z[i][j], 1);
        }
    }

    for (int k = 0; k < intEmployeeCount; k++)  //want to minimize travel costs
    {
        for (int i = 0; i < intStoreCount; i++)
        {
            lneCost.AddTerm(x[k][i], dblLambda * adblDistance[k, i]);
        }
    }

    //our optimization
    model.AddMinimize(lneCost);

    //add constraints: use model.AddGe(), model.AddEq(), or model.AddLe()
    //Each employee gets assigned to exactly one store; in total, k constraints
    //sum_i x_{k,i} = 1, for each k
    for (int k = 0; k < intEmployeeCount; k++)
    {
        ILinearNumExpr lneEmployeeToOneStore = model.LinearNumExpr();
        for (int i = 0; i < intStoreCount; i++)
        {
            lneEmployeeToOneStore.AddTerm(x[k][i], 1);
        }
        model.AddEq(lneEmployeeToOneStore, 1);
    }


    //The distance for each employee to his/her store is less than dblDisThreshold, e.g., 40 km
    // in total, intEmployeeCount*intStoreCount constraints
    //x_{k,i} * d_{k,i} <= dblDisThreshold, for each k, i
    for (int k = 0; k < intEmployeeCount; k++)
    {
        for (int i = 0; i < intStoreCount; i++)
        {
            ILinearNumExpr lneEmployeeDis = model.LinearNumExpr();
            lneEmployeeDis.AddTerm(x[k][i], adblDistance[k, i]);
            model.AddLe(lneEmployeeDis, dblDisThreshold);
        }
    }

    //add the relationships between x and y,z
    for (int i = 0; i < intStoreCount; i++)
    {
        for (int j = 0; j < intAttributeCount; j++)
        {
            ILinearNumExpr lneEmployee = model.LinearNumExpr();
            for (int k = 0; k < intEmployeeCount; k++)
            {
                lneEmployee.AddTerm(x[k][i], adblEmployee[k, j]);
            }
            INumExpr neDiff = model.Sum(adblStore[i, j], model.Negative(lneEmployee));
            INumExpr neAssitant = model.Sum(neDiff, model.Prod(y[i][j], 1), model.Negative(z[i][j])); //diff + y[i][j] - z[i][j]

            model.AddEq(neAssitant, 0);
```

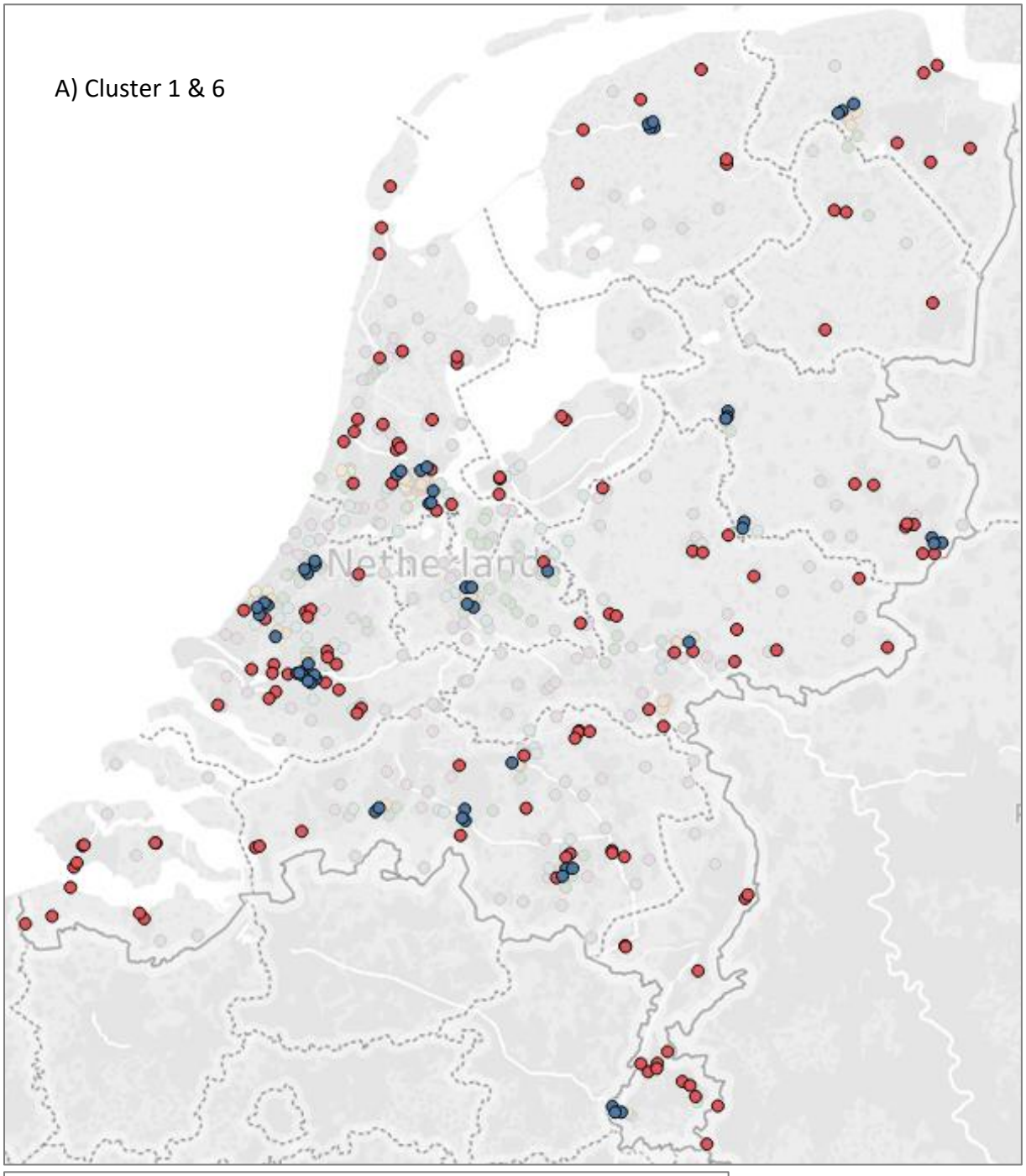
Figure 0.4 CPLEX solver code

Appendix 8: Whize data set variables and classes

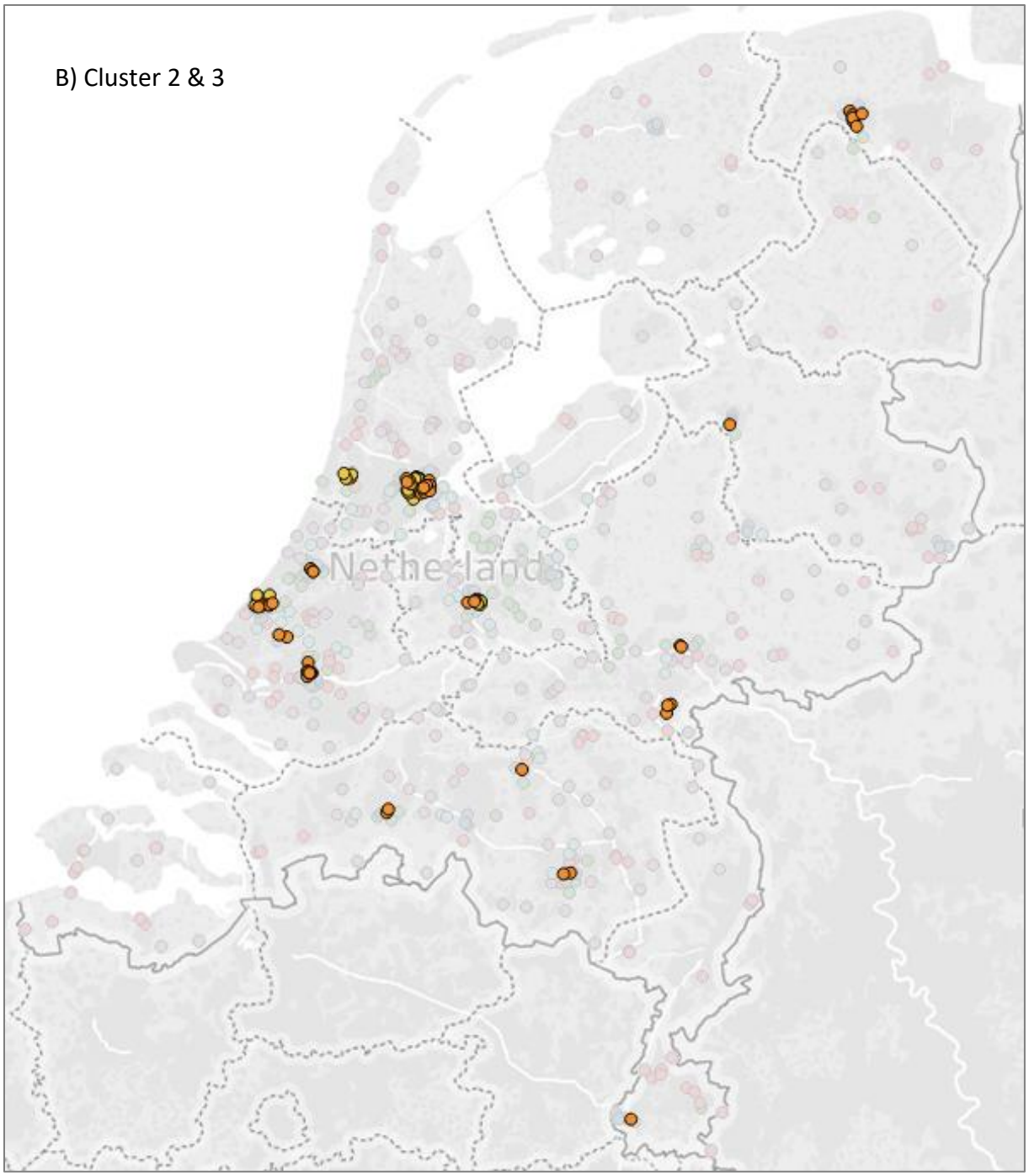
Variable	Class
Income	Low (€0 - €26.000)
	Average (€26.000 - €40.000)
	1.5 times average (€40.000 - €54.000)
	2 times average (€54.000 - €70.000)
	Above 2 times average (over €70.000)
Household size	One person
	Two persons
	Three persons
	Four persons
	Five or more persons
Age	Below 25 years
	Values between 25 and 75 in classes of 5 years (e.g. 25-30 years)
	Above 75 years
Education	Low education
	Average education
	High education
	Academic education
Purchasing power	10 deciles from lowest 10% to highest 10%
Family situation	Family with children (oldest child between 0-5 years)
	Family with children (oldest child between 6-12 years)
	Family with children (oldest child between 13-19 years)
	Family with children (oldest child over 20 years)
	No children (younger than 35 years)
	No children (between 35-49 years)
	No children (between 50-64 years)
	No children (older than 65 years)
Car ownership	One car
	Two or more cars
	No car
Marital status	Married, partner registration, living together
	Other
Social class	Social class A
	Social class B1
	Social class B2
	Social class C
	Social class D
Work situation	Part-time
	Full-time
	Student
	Retired
	Unemployed

Appendix 9: Clustering maps – socio-demographic data

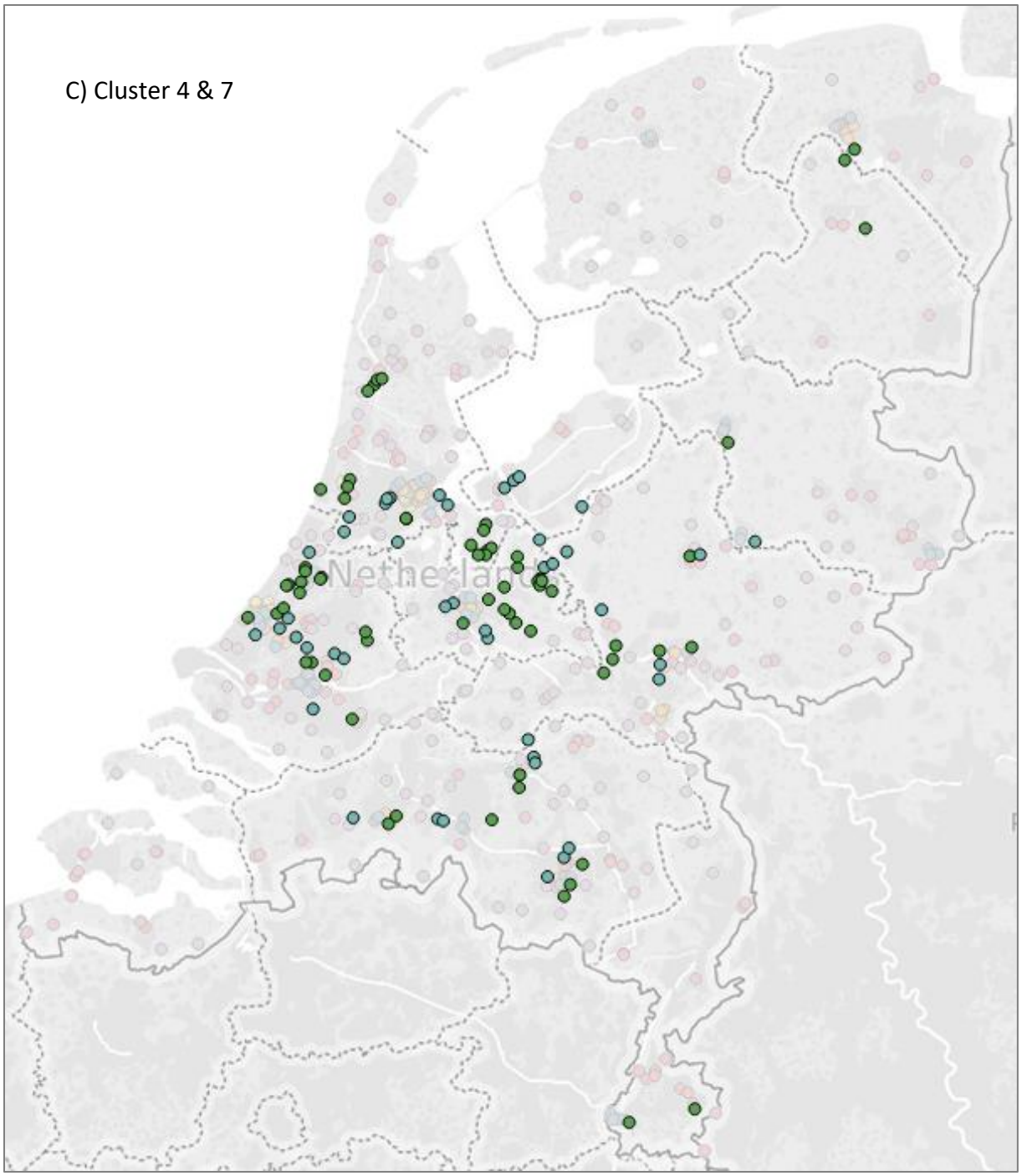
A) Cluster 1 & 6



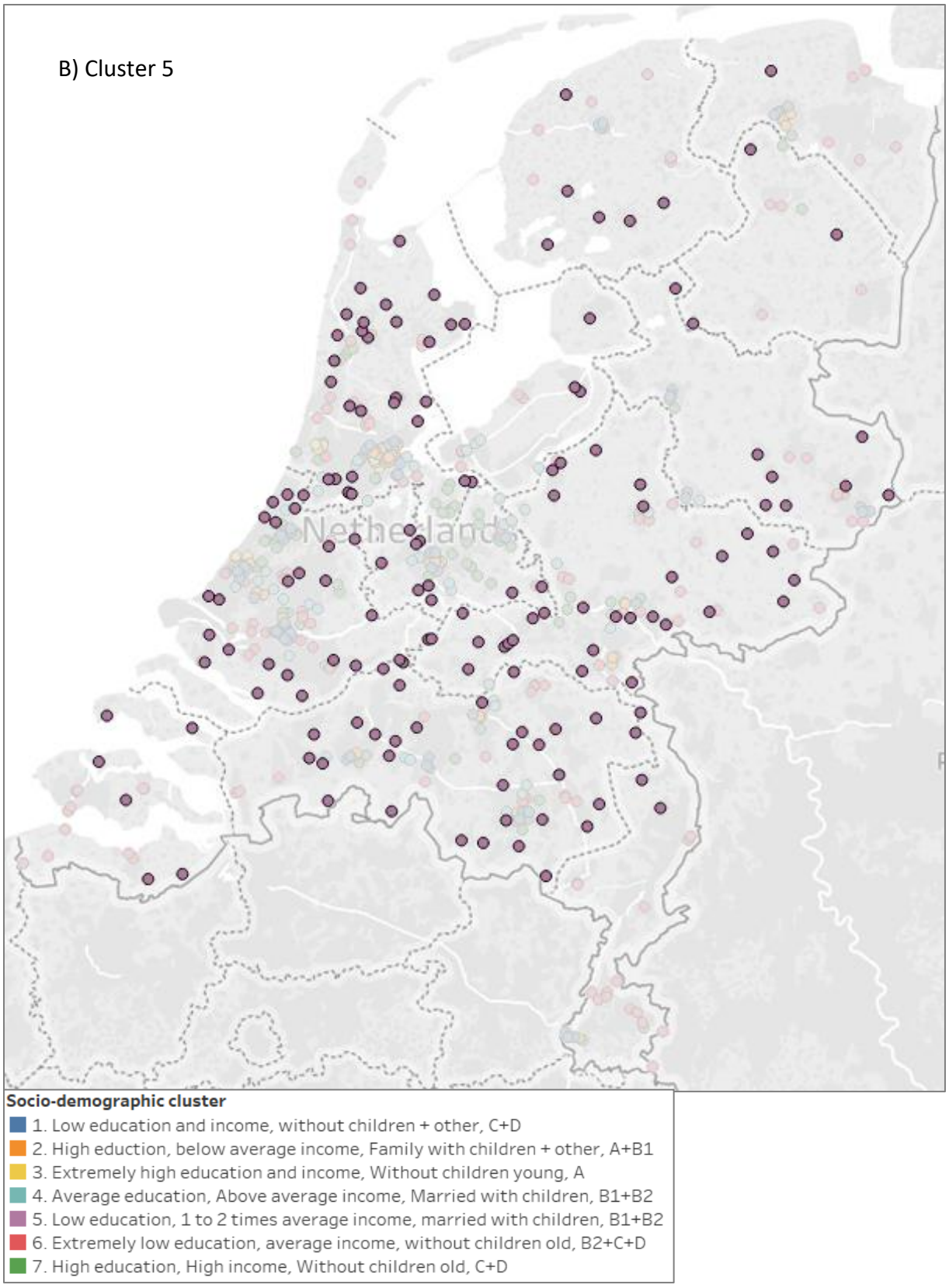
B) Cluster 2 & 3


Netherlands
Socio-demographic cluster
1. Low education and income, without children + other, C+D
2. High eduction, below average income, Family with children + other, A+B1
3. Extremely high education and income, Without children young, A
4. Average education, Above average income, Married with children, B1+B2
5. Low education, 1 to 2 times average income, married with children, B1+B2
6. Extremely low education, average income, without children old, B2+C+D
7. High education, High income, Without children old, C+D

C) Cluster 4 & 7


Socio-demographic cluster
1. Low education and income, without children + other, C+D
2. High eduction, below average income, Family with children + other, A+B1
3. Extremely high education and income, Without children young, A
4. Average education, Above average income, Married with children, B1+B2
5. Low education, 1 to 2 times average income, married with children, B1+B2
6. Extremely low education, average income, without children old, B2+C+D
7. High education, High income, Without children old, C+D

B) Cluster 5


Netherland
Socio-demographic cluster
1. Low education and income, without children + other, C+D
2. High eduction, below average income, Family with children + other, A+B1
3. Extremely high education and income, Without children young, A
4. Average education, Above average income, Married with children, B1+B2
5. Low education, 1 to 2 times average income, married with children, B1+B2
6. Extremely low education, average income, without children old, B2+C+D
7. High education, High income, Without children old, C+D