Determining the mobile device offering at a large SA retailer

Monique Dagnin

Thesis presented in fulfilment of the requirements for the degree of
MCom (Quantitative Management)
in the Faculty of Economic and Management Sciences at Stellenbosch University

Supervisor: Dr Isabelle Nieuwoudt

March 2023

Declaration

By submitting this thesis electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Date: March 1, 2023

Abstract

The retail industry is one of the biggest industries in the world and an important factor in the success of retailers, is carrying the correct products for their customers. The Retailer in this study, like many other retailers, provides a range of financial services and products to their customers to add value and improve the customers' experience in their stores. Therefore, the aim of this study is to assist The Retailer in determining the best range of mobile devices to keep in stock in their stores.

The Retailer has over 1 280 stores and was seeking device ranges for store groups, rather than a unique range for each individual store. Therefore, stores with similar characteristics are grouped based on several external factors that were identified. Hierarchical clustering is used to group similar stores within each supermarket type based on the number of landlines, rate of population change, population age and population income. Six clusters, two per supermarket type, are found with this method.

For each group of stores, the range of mobile devices to keep in stock is determined using three performance measures, namely rate of sale, total units sold and average units in stock. These measures are calculated to evaluate the performance of mobile devices and rank the devices according to their performance. Two iterative approaches are followed to determine whether a device should be ranged in any of the six clusters. For mobile devices that have not been ranged in a particular store, but should be ranged according to their performance, the required stock level in these stores is determined by estimating the rate of sale per store using a regression tree for each mobile device. To build the regression trees, population age, rate of population change, population income, number of landlines, store size, province in which a store is located, adapted mobile device category rate of sale, average sales amount per store and total number of mobile devices sold in a store are used as independent variables. The methodology is illustrated using a selection of The Retailer's devices currently in stock. Eleven of these 30 sampled mobile devices are ranged using this methodology, suggesting that this methodology succeeds in reducing the variety of mobile devices ranged by The Retailer by removing under performing devices.

Opsomming

Die kleinhandelbedryf is een van die grootste nywerhede ter wêreld en om die regte produkte aan te hou, is 'n belangrike faktor in die sukses van kleinhandelaars. Die Kleinhandelaar in hierdie studie verskaf, soos baie ander kleinhandelaars, finansiële produkte en dienste aan hul kliënte om waarde toe te voeg en die kliënte se ervaring in hul winkels te verbeter. Dus is die doelwit van hierdie studie om Die Kleinhandelaar te help om die beste reeks selfone te bepaal om in hul winkels in voorraad te hou.

Die Kleinhandelaar het meer as 1 280 winkels en was op soek na 'n reeks selfone vir winkelgroepe, eerder as 'n reeks per individuele winkel. Dus word winkels met soortgelyke eienskappe word dan op grond van verskeie eksterne faktore gegroepeer. Hiërargiese groepering word gebruik om soortgelyke winkels binne elke supermarktipe op grond van die aantal landlyne, bevolkingsveranderingskoers, bevolkingsouderdom en bevolkingsinkomste te groepeer. Ses groepe, twee per supermarktipe, word met hierdie metode gevind.

Vir elke groep winkels, word die reeks selfone wat in voorraad gehou moet word, met behulp van 'n drie prestasiemaatstawwe, naamlik die verkoopskoers, die totale eenhede verkoop en die gemiddelde eenhede in voorraad, bepaal. Die prestasiemaatstawwe word bereken om die prestasie van selfone te evalueer en daarvolgens te rangskik. Twee benaderings word gevolg om te bepaal of 'n selfoon in enige van die ses winkelgroepe aangehou moet word. Vir selfone wat nie in 'n spesifieke winkel aangehou word nie, maar volgend hul prestasie aangehou behoort te word, word die vereiste voorraadvlak in hierdie winkel bepaal deur die verkoopkoers, deur middel van 'n regressieboom, te benader. Om die regressiebome te bou, word die bevolkingsouderdom, bevolkingsveranderingskoers, bevolkingsinkomste, aantal landlyne, winkelgrootte, provinsie waarin 'n winkel geleë is, aangepaste verkoopkoers vir die selfoonkategorie, gemiddelde verkoopsprys per winkel en die totale aantal toestelle wat in 'n winkel aangehou word, as onafhanklike veranderlikes gebruik. Die metodologie word deur 'n steekproef van Die Kleinhandelaar se huidige selfone geïllustreer. Elf van die 30 selfone in die steekproef word steeds volgens hierdie metodologie aangehou, wat daarop dui dat hierdie metodologie daarin slaag om die verskeidenheid selfone wat deur Die Kleinhandelaar aangehou word, te verminder deur onderpresterende toestelle te verwyder.

Contents

List of Figures

List of Tables

Reserved Symbols

x_i LSM binary variable (attribute present in the household)

ω_i contribution of attribute i to the classification of the LSM groups

δ probability of buying second choice if preferred product is not available

$m_{2007,i}$ population size for municipality i in 2007

$m_{2016,i}$ population size for municipality i in 2016

$\bar{g}_i$ average distance of data point i to all other data points in the same cluster as point i

$d(i, C)$ distance between data point i and a cluster C, where i is not assigned to cluster C

b_i closest distance between data point i and any cluster not containing i

s_i silhouette coefficient for data point i

u_{ij} number of units of device i sold in store j

t_{ij} number of days device i is in stock in store j

P total number of mobile devices

N total number of stores

r_{ij} ROS of device i in store j

T_j total number of days that any device was in stock in store j

R_j ROS of the mobile device category in store j

I_{ij}^R ROS index of device i in store j

p_j number of devices sold in store j

$\bar{U}_j$ average number of units of all mobile devices sold in store j

I_{ij}^U total-units-sold index of device i in store j

o_{ijk} number of units of device i in stock in store j on day k

$\bar{o}_{ij}$ average units of device i in stock in store j per day

$\bar{O}_j$ average number of units of any mobile device in stock in store j per day

I_{ij}^{O} average units in stock index of device i in store j

I_{ij}^{C} combined index of device i in store j

$\bar{I}_{i}^{C}$ average combined index of device i

n_i number of stores in which mobile device i was sold

a_{ijk} total sales amount of device i in store j on day k

$\bar{A}_j$ average sales amount of the mobile device category in store j

τ_{ij} number of days device i is sold in store j

w^R weighting assigned to the ROS index, I_{ij}^{R}

w^U weighting assigned to the total units sold index, I_{ij}^{U}

w^O weighting assigned to the average units in stock index, I_{ij}^{O}

CHAPTER 1

Introduction

Contents

The retail industry is one of the biggest industries in the world and is playing a vital role in every country's economy [6]. According to Fildes, Ma and Kolassa [10], this industry is still growing worldwide, both in-store and online, increasing the competition among retailers. This is no different in South Africa. In the last quarter of 2019, the retail industry contributed to 36% of the total turnover in South Africa [41]. Statistics South Africa (StatsSA) reported an income growth of 108% (between 2009 and 2018) for the retail industry. During the same period, the industry also increased the number of employees by 28% [42].

One of the factors that contributes to the success of a certain retailer is whether the retailer, or a specific store in a retail group, carries the correct products for the target audience, thus which type of clientele is in close proximity to the particular store. The type of clientele of a particular store is determined by, for example, their wealth and living standards. The data of the South African Audience Research Foundation (SAARF) is a handy way to determine the type of clientele of a specific store.

The SAARF was formed in 1973 and its objective is to measure and understand the audience of various media outlets in South Africa by conducting surveys. Their Living Standards Measure (LSM) is the most popular tool whereby the South African population is segmented into 10 household groups, namely LSM 1 (least affluent) to LSM 10 (most affluent), based on their living conditions. In total, 29 factors are considered to measure wealth and living standards, ranging from access to services like water and electricity, to the ownership of various appliances and assets [38]. Table 1.1 provides an outline of the different LSM groups and their access to services and ownership, compiled using survey data collected by SAARF in June 2011. Note that Table 1.1 gives an indication of services and assets most likely accessed by households within each group, and does not necessarily mean each household within a specific LSM group will have the same access and ownership [38]. Appendix A contains the full list of factors considered, as well as the calculation used to classify a household into one of the LSM groups.

In this research project, amongst others, the type of clientele of each store for a specific retailer is used to seek the correct list of products to carry in each store.

LSM	Level of education	Type of dwelling	Services (e.g. electricity)	Ownership of assets (e.g. TV)	Avg household income per month
1	Primary school completed	Traditional hut	Minimal access	Radio	R1 363
2	Some high school	Informal settlement	Communal access to water	Radio and stove	R1 929
3	Some high school	Informal settlement	Electricity and water	Radio and stove	R2 258
4	Some high school	Informal settlement	Electricity and water	TV and electric hotplate	R3 138
5	Some high school	Small urban or rural	Electricity, water and communal toilet	TV, radio, stove and fridge	R4 165
6	Matric and higher	Large urban	Electricity, water and flush toilet	A number of durables and cellphone	R6 322
7	Matric and higher	Urban	Full access	All above and motor vehicle	R10 292
8	Matric and higher	Urban	Full access and bank accounts	Full ownership, including PC	R14 046
9	Matric and higher	Urban	Full access and bank accounts	Full ownership	R19 658
10	Matric and higher	Urban	Full access and bank accounts	Full ownership	R29 614

TABLE 1.1: *Examples of access to services and ownership of a typical household per LSM group.*

1.1 The retailer in this study

One of Africa's largest retailers (hereafter referred to as The Retailer) operates through three different store formats. Each one of these types of supermarkets has different LSM target markets, as well as different competitors, within the market. As indicated in Table 1.2, the first of the three types of supermarkets (type A) is aimed at providing a small range of necessities to lower LSM groups. With larger stores, type B supermarkets can provide a wider range of products to the middle LSM target market. In the last type of supermarket, the focus is to provide a wide variety of products, including speciality lifestyle products, to the higher LSM groups. Also indicated in Table 1.2, is the number of stores in each type of supermarket. It is clear that the middle LSM groups, serviced by the type B supermarkets, are the largest target market of The Retailer.

Supermarket type	Number of stores	Average store size (m^2)	LSM groups served
Type A	394	582	LSM 1 - 4
Type B	604	2 572	LSM 4 - 7
Type C	289	3 868	LSM 8 - 10

TABLE 1.2: *The Retailer's different supermarket types.*

Along with providing groceries and basic household items, The Retailer also provides a range of financial products and services in its stores. According to a survey done by Eighty20 [7],

retailers across South Africa provide financial services, including credit facilities, insurance and airtime, as value added services. The aim of these services is to increase the number of customers in store by providing additional products and services to create a 'one stop shop' and improve customers' shopping experiences.

Mobile devices are one example of these value added services provided by The Retailer. It is this value added service of selling mobile devices and in particular, determining which mobile devices as well as how much to keep in stock, that is the topic of this research project.

1.2 Problem description

Since retailers sell mobile devices on behalf of telecommunications service providers (TSPs), there are many factors influencing the ranges, prices and sales of mobile devices, which are out of The Retailer's control. For example, different areas in the country have different network coverage (2G, 3G, et cetera) and this will definitely have an impact on the type of mobile devices used in a specific area. According to the company, Signal Booster [40], older devices are not always compatible with newer generation networks. Similarly, some features on newer devices require the speed of newer generation networks to function properly and may not work on older generation networks.

In 2008, Kalba [21] suggested that there is a positive relationship between the number of mobile devices and the number of fixed lines in an area. This was partly attributed to the fact that fixed lines increase awareness of the benefits of telecommunications and therefore increase the demand of mobile devices. However, the number of fixed lines in an area is also out of The Retailer's control.

Demographic information is also often used to understand the demand for mobile devices. For example in his study, Cox [5] considered the rate of population birth, death and migration in order to understand changes in the population size over time. These variables were used in forecasting demand for different telecommunication products. Another commonly used factor in mobile device demand analyses is population income, because consumers with a higher disposable income are able to spend more on mobile devices [21]. Age is also a factor often used when studying mobile device choices. Zhou *et al.* [51] found that older adults value visible attributes, like phone and display size, more than younger adults. Further, younger adults in this study were more interested in functions and connectivity to the internet and other devices than older adults.

All of the factors discussed above, that are out of the Retailer's control, are referred to as *external factors* influencing the demand of mobile devices. Apart from these external factors, there are also *phone-specific factors*, in other words attributes of mobile devices that influence consumers' demand. Many studies have already been done on these factors. In their study, Madashi and Raghupataiah [27] assess the difference in importance of mobile device price, quality, style, functions and brand between rural and urban markets in India. They found that there was no significant difference in consumers' view of price or style, however, there was a significant difference in their views on quality, functions and brand. Similarly, Karjaluoto *et al.* [23] found that, among students, brand, price and features of mobile devices are the most important influencing factors in mobile phone choices. In their study, Işıklar and Büyüközkan [18] identified physical characteristics, technical features, basic requirements (for example price), functionality, brand choice and customer excitement as decision criteria. These criteria were used to identify the best mobile phone alternative based on the specific user's preferences.

In order for The Retailer's mobile device offering to be successful in adding value to the customers, The Retailer has to consider all of these factors, as well as the knowledge it has on past sales, to understand what its customers are looking for. By doing so, it is able to provide the best range of mobile devices that will meet customer demand.

1.3 Project objectives and layout

The aim of this research is to assist The Retailer in determining the best range of mobile devices to keep in stock in all the stores distributed over the three different supermarket types. To achieve this, the following objectives were identified:

I Determine which external factors influence the demand for mobile devices,

II Collect and manipulate the relevant data needed for all aspects of the project,

III Group stores with similar characteristics, based on the external factors,

IV For each group, determine which mobile devices The Retailer should keep in stock using a number of performance metrics, and

V Calculate how much stock to keep in each store for each of the mobile device ranges.

Objective I was already addressed in §1.2. Chapter 2 provides a literature review of various methods needed to address Objective III to V, while Chapter 3 contains a brief overview of the different steps needed to obtain a mobile device range in this project. Data collection and manipulation, thus addressing Objective II, is also described in Chapter 3. This is followed by a discussion of the process to group similar stores as required in Objective III in Chapter 4. Objectives IV and V are addressed in Chapter 5, which provides a detailed explanation of the process of obtaining a mobile device range with stock levels. This thesis is concluded with the results in Chapter 6 and some final remarks in Chapter 7.

CHAPTER 2

Literature review

Contents

In this chapter, a literature review provides the reader with background of similar studies, as well as an overview of the different types of methods used in this project. In §2.1, relevant studies are discussed. Cluster analysis, which can be used to group similar stores, is discussed in §2.2. This is followed by a description of various regression methods in §2.3.

2.1 Assortment planning

Assortment planning entails the selection of a set of products with unique attributes such as brand, size and colour (also called *stock keeping units* (SKUs)), to be sold in-store so that revenue or profit of the particular retailer is maximised [11].

According to Corsten *et al.* [4], there are two approaches frequently followed in assortment planning. With the first approach, one common assortment of products is determined for all stores. The second approach is to determine individual assortments for each of the stores in a network. Although these two approaches are often used, both have their own drawbacks. In a retail chain with many stores, having only one assortment can results in suboptimal revenue, while having unique assortments for all stores can be difficult to manage. Corsten *et al.* [4] combined these two approaches to determine one basic assortment for all stores, which is then supplemented with region-specific assortments. In their study, Fisher and Vaidyanathan [11] addressed the problem above by placing a limit on the number of assortments allowed, and assessed the impact of a change in this number on revenue.

When doing assortment planning, customers' choices and their behaviour when their preferred choice is not available (referred to as *substitution*) are important considerations [24]. Kök

et al. [25] divide substitution into three types. The first is called *out-of-stock (OOS) substitution* and occurs when a customer's preferred product is ranged in a store, but is out of stock when the customer wants to buy it. In this case, the customer can choose a different product, return on a different day, or go to a competing store to buy their preferred product. The second type of substitution, *assortment-based substitution*, occurs when a customer's preferred product is not included in a store's assortment. Since the product will never be available at the particular store, the customer's only choices are to buy a substitute product or go to a competing store. The last type is called *dynamic substitution* and is a combination of OOS and assortment-based substitution. It describes a situation where a customer arrives at a store and chooses a product based on what is available at that moment, not considering products that are out of stock or only available at other stores. Thus, the customer's only options are to choose the best available product, or to leave the store without purchasing anything [25]. OOS and assortment-based substitution generally apply to food and other products that are purchased often, while dynamic substitution is usually more appropriate for one-time purchases [24]. Finally, Gaur and Honhon [14] also defined *static 'substitution'*, which occurs when a customer only has one desired product with no alternatives. In this case, there are no substitutions and a customer will either buy their preferred product or leave the store with no purchase.

Customers' demand and substitution behaviour are often represented by *choice models*, where the customers' preferences are used to determine products' importance relative to each other [4]. The three choice models most often used are the multinomial logit (MNL) model, exogenous demand models and the locational choice (LC) model [25].

With the MNL model, a customer is assumed to have a utility assigned to each product in a set, while the model also includes a no-purchase option with an associated utility. This utility consists of two parts, the deterministic (expected) component which is calculated from the data, as well as a random component representing the difference between the expected and actual utility (for each individual customer) of a specific product. The deterministic utility value of a product is the same for all customers, but due to the random component, the total utility can vary from one customer to the next [46].

A problem of the MNL model is the so-called *Independence of Irrelevant Alternatives* (IIA) *assumption*, which states that the preference of alternative A over alternative B in the choice set {A, B} should not change to a preference for B over A when including alternative C to the choice set, now {A, B, C}. This assumption does not hold when products within a subcategory have a higher similarity with each other than with products from other subcategories. Kök *et al.* [25] explain this with the 'red bus/blue bus paradox'. Suppose a person has the same probability of driving his car or taking the bus to work, thus $P(car) = P(bus) = \frac{1}{2}$. Then, suppose that there are two buses and the only difference between the two is their colour. One is red and one is blue. Assuming that the colour of the bus makes no difference to this person's decision, it is expected that

$$P(\text{car}) = \frac{1}{2} \quad \text{and} \quad \text{P(red bus)} = \text{P(blue bus)} = \frac{1}{4}.$$

However, with the MNL model

$$P(\text{car}) = \text{P(red bus)} = \text{P(blue bus)} = \frac{1}{3} \text{ [25].}$$

The Nested Logit model is one way to address this problem. In this case, products are divided into subsets and a two-stage process is followed so that customers first choose a subset, and then a product within the relevant subset. The MNL model is used in both stages of this decision making process. In this two-stage process, the IIA assumption does not apply to products

from different subsets, so that the expected probabilities from the 'red bus/blue bus paradox' is achieved. However, applying the Nested Logit successfully requires knowledge of the hierarchy of attributes from the customers' perspectives. For example, in applying this model to shirts, subsets may be based on brand, sizes or colours, depending or which attribute is most important to customers [25].

With MNL, substitution can occur between any two products [25] and the substitution rate is based on the utility value. Therefore, OOS and assortment-based substitution have the same substitution rate [24].

Gaudagni and Little [13] applied the MNL model to the coffee purchases of 100 households over a 32 week period to predict the demand for various brand-size combinations. Van Ryzin and Mahajan [46] studied the trade-off between inventory costs and the benefit of keeping a wider variety of products using the MNL model, while Chen *et al.* [3] used a Nested Logit model under dynamic substitution.

With *exogenous demand models*, the demand for each product and the substitution is specified, and consumer choice can be explained by two assumptions. Firstly, every customer chooses their favourite product. Secondly, if the preferred product is not available, the customer chooses their second choice with a probability of δ or they choose not to purchase with a probability of $1-\delta$. When the second choice is also unavailable, the customer considers their third choice and decides whether or not to purchase. For each attempt, δ could remain the same or be specified differently [25].

LC models are based on Hotelling's study of competition between firms [14]. Hotelling [16] found that if two competing stores have homogeneous products and equal prices, the only differentiating factor is customers' travel cost to each of the two stores. Since customers will always travel to the nearest store, the two competitors will both choose to be located in the middle of the market where the average travel cost to all customers is minimised. This phenomenon can still be seen today, for example in the fast food or motor trade industries, where competitors are found in close proximity.

Lancaster [26] applied Hotelling's location model to consumer choice. In this model, products are viewed as a set of characteristics and a customer's preferred product is identified by a point in this characteristics space. A customer's utility for a specific product is calculated relative to their preferred point. They then choose the product with the highest utility (in other words, the product closest to their ideal point). With LC models, substitution can only occur between products that are close to each other in the characteristics space [25].

Gaur and Honhon [14] used an LC model to determine a single period assortment for a retailer, based on static as well as dynamic substitution. Fisher and Vaidyanathan [11] also used an LC model, which they also applied to real examples, including snack cakes and tires.

2.2 Cluster analysis

With cluster analysis, a dataset is grouped into smaller subsets of elements so that each element is similar to other elements in the subset, and different from elements in other subsets. Cluster analysis has been used in many different areas, including image recognition, anomaly detection and biology. Han *et al.* [15] classified clustering techniques into four categories, namely partitioning methods, hierarchical methods, density-based methods and grid-based methods.

2.2.1 Partitioning methods

Partitioning methods group elements of a dataset into a predetermined number of clusters which are mutually exclusive. Clustering with partitioning methods is usually done based on distance. Because these methods follow an iterative approach, finding the global optimum is often too computationally expensive. Therefore, heuristic approaches that find a local optimum, for example k-means, are more popular. With k-means each cluster is represented by a centroid (the mean of the elements in the cluster) and elements are assigned to clusters so that the sum of squared errors (the Euclidean distance between an element and its cluster centroid) is minimised. [15]. Although k-means is simple to understand, a disadvantage of the method is that the decision maker has to specify the value of k before clustering can be done [19].

Kargari and Sepehri [22] used k-means clustering to cluster stores in a distribution network. By combining this clustering with a priority system and distribution policies within the distribution centre, their model led to a 32% reduction in transportation and distribution costs.

2.2.2 Hierarchical methods

If clusters may have sub-clusters, a *hierarchical clustering* (in other words a set of nested clusters) is obtained. Hierarchical methods can be split into divisive or agglomerative methods. *Divisive methods* start with all elements in a single cluster at the top of the hierarchy and iteratively splits clusters into smaller clusters until each element is in its own cluster or a termination condition is met [15].

Agglomerative methods follow an inverse approach, with each element of the dataset initially in its own cluster. Then, clusters are merged together until the top of the hierarchy is reached (where there is only one cluster containing all elements) or until a termination condition is met [15].

In agglomerative methods, elements are combined based on a dissimilarity function and often the Euclidean distance is used. Once clusters are formed, the dissimilarity between clusters can be calculated using various methods, called *linkage* [19].

In their study, Pagnuco *et al.* [32] used hierarchical clustering to find groups of co-expressed genes to be used when conducting experiments.

2.2.3 Density-based methods

Density-based methods group elements together based on their density, rather than their distance to another element. In other words, elements belong to the same cluster as long as their density (or the number of neighbouring elements within a certain radius) is greater than a given threshold. Generally, density-based clusters are mutually exclusive [15].

One of the most popular density-based methods is called Density-Based Spatial Clustering of Applications with Noise (DBSCAN) which takes two user inputs, the radius of a neighbourhood and the density threshold indicated as a number of elements. An element of the dataset is considered a core object if it has at least the number of neighbours needed to pass the density threshold within its neighbourhood. Data elements are considered *density reachable* if they fall within the specified radius of a core object. Clusters are then formed by grouping these core objects and their density reachable neighbours [15]. Any elements that are not grouped in clusters, are considered noise or outliers [9].

An advantage of DBSCAN is that the use of density, rather than distance, makes it possible to identify clusters of arbitrary shape. This is also an effective method to use when working with large datasets [9].

Pavlis *et al.* [33] applied a modified DBSCAN approach to cluster retail stores in Great Britain. By using a graph representation of the different locations, as well as a list of k-nearest neighbours, DBSCAN can be applied iteratively. This allows for the parameters to be changed for different subgroups of stores, improving the clustering results.

2.2.4 Grid-based methods

Grid-based methods use a grid with a finite number of cells to divide the dataset. All clustering is applied to the cells in the grid, rather than to the data itself [15].

In the *statistical information grid* (STING) method, the decision space is divided into rectangular cells that correspond to different levels of a hierarchical structure. Cells at a higher level of the hierarchy can be divided to introduce lower levels of the hierarchy [15].

Another example of a grid-based clustering method is *clustering in quest* (CLIQUE). In the CLIQUE method, each dimension in a dataset is divided into equal intervals, creating a grid. Dense cells can be identified using a density threshold. Clusters are then formed using these dense cells [15].

In their study, Jannu and Jana [20] combined grid-based clustering with a routing algorithm in the design of a wireless sensor network to improve energy efficiency and to avoid overloading nodes close to the sink nodes, which reduces the number of node failures in the network.

2.2.5 Cluster validation

Once a dataset has been clustered, it is important to assess the validity of those clusters. Theodoridis and Koutroubas [45] define three broad categories of cluster validation criteria, namely external, internal and relative criteria. External criteria are pre-existing structures in the data and can be used to validate clustering results by measuring how well the clusters fit this pre-existing structure. This can be done by comparing the clusters with the partitions in the pre-existing structure, or by measuring the difference between the pre-existing partitions and the proximity matrix of the clusters.

Internal validation criteria are directly related to the data used in the clustering, for example the proximity matrix, and can be used to validate a hierarchy of clusters or a single set of clusters. When using external or internal validation criteria, various statistical tests are performed. In many cases, Monte Carlo techniques are needed to estimate the distribution of these statistics and this can become computationally difficult, rendering these tests inefficient [45].

Relative criteria refer to the use of different parameter values with the same clustering method to evaluate clustering results. These criteria do not require statistical tests, nor the same computational capacity as with external and internal criteria. The goal with relative criteria is to select the best clustering based on a predefined criterion, usually a distance or similarity measure. Some commonly used methods include the silhouette index, gap statistic and, in the case of hierarchical clustering, the dendrogram [45].

2.3 Regression analysis

With regression analyses, the goal is to find relationship between variables in order to estimate or predict the value of some numeric field. Various regression models exist and the best model for a specific dataset depends on the nature of the relationships between the different variables in the dataset [19].

One of the most straightforward regression models is called *simple linear regression* and consists of only one predictor (independent) variable and the dependent variable which is being predicted. The relationship between the dependent and independent variable is assumed to be linear and can be explained by a combination of the regression line,

$$y = mx + c,$$

and an error term ϵ, which accounts for the deviation of real data points from this regression line. This concept can be extended to include more than one independent variable, resulting in what is called *multiple linear regression* [17].

When working with real world data, however, the relationships between variables are seldom linear and multicollinearity[1] often exists between independent variables. For these reasons, other regression models must also be considered.

Ridge and *lasso regression* are two methods based on linear regression, with the goal being to shrink the coefficient estimates in the regression equation. Shrinking the coefficients is useful when dealing with multicollinearity in a dataset, as the smaller coefficients reduce the impact of multicollinearity [19]. With ridge regression, coefficients are reduced towards zero, but can never be exactly zero. On the other hand, lasso regression can perform variable selection by setting coefficients equal to zero, effectively removing them from the regression equation [50].

When the relationship between the dependent and independent variables is not linear, polynomial regression can be used. In this method, new independent variables are added to the regression function by raising the original independent variables to a power. This allows for a non-linear curve, rather than the simple regression line, while still solving a linear regression function,

$$y = m_1 x_1 + m_2 x_1^2 + c + \epsilon \text{ [19, 35].}$$

Other statistical methods, like *decision trees*, can also be used to predict a numeric variable. With decision trees, the dataset is divided into different sections and a prediction is made for each section, usually by calculating the mean or mode of the values in the section. The decision tree consists of nodes and branches, with each node representing a splitting condition. The tree is constructed from the top and follows a greedy approach in which the best split is made at each particular node, without considering future splits. The 'best split' at each node refers to the variable and splitting condition that result in the lowest residual sum of squares (RSS). The tree continues to grow in this way until a stopping criterion is met. The nodes found at the bottom of the tree are called terminal nodes and this is where the predictions are made [19].

Decision trees are easy to interpret, but can overfit to training data if the tree is too complex. A number of approaches have been developed to overcome this overfitting and improve the accuracy of predictions, including bagging, random forests and boosting. In short, *bagging* is the process of taking multiple random training datasets, creating a predictive model for each training set and averaging the results. This method increases the accuracy of predictions in datasets with a high variance. With *random forests*, a similar process is followed. However, at

[1] Multicollinearity exist in a dataset when two or more independent variables are highly correlated [49].

each node in the tree, a random sample of the dependent variables is considered for the split, rather than considering all variables at every node. This is done to reduce the impact of strong variables by ensuring that the different trees are not highly correlated. Lastly, with *boosting*, multiple decision trees are also built. But, unlike with bagging, the trees are built one at a time, with each tree using information from the previous tree to improve the predictions [19].

CHAPTER 3

Overview and data collection

Contents

In the first part of this chapter, the overall methodology approach that will be used in this research project, is given in §3.1. In the second part of the chapter, in §3.2, the various data sources used in this project, are discussed.

3.1 Project overview

The project described in this thesis is divided into several steps, as depicted in Figure 3.1. During Step 1 in Figure 3.1, all the factors that will influence the demand of mobile devices are determined. This step coincides with Objective I in §1.3 and these factors were obtained from literature, as discussed in §1.2.

After the external factors that influence mobile device demand were identified, data for these factors, as well as for other aspects of the project, need to be collected and analysed for each of The Retailer's stores. This is done in Step 2 in Figure 3.1. The raw data are obtained from The Retailer as well as from StatsSA using the Eighty20 Data Portal [8]. The collection and manipulation of this data will be described in detail in §3.2, satisfying Objective II in §1.3.

The Retailer has a total of 1 287 stores and thus determining a unique range for each store will be impractical. On the other hand, having one common assortment of mobile devices for all 1 287 stores, is also not viable, as stated in §2.1. Therefore, in this project, using the factors from Step 1 and the data collected in Step 2, similar stores, thus stores with more or less the same characteristics, are grouped together in Step 3 in Figure 3.1. In this way, a range of mobile devices per store group, rather than a unique range for each of the 1 287 stores, can be determined to address Objective III in §1.3. Stores are grouped using agglomerative hierarchical clustering introduced in §2.2.2, and this grouping process will be discussed in Chapter 4.

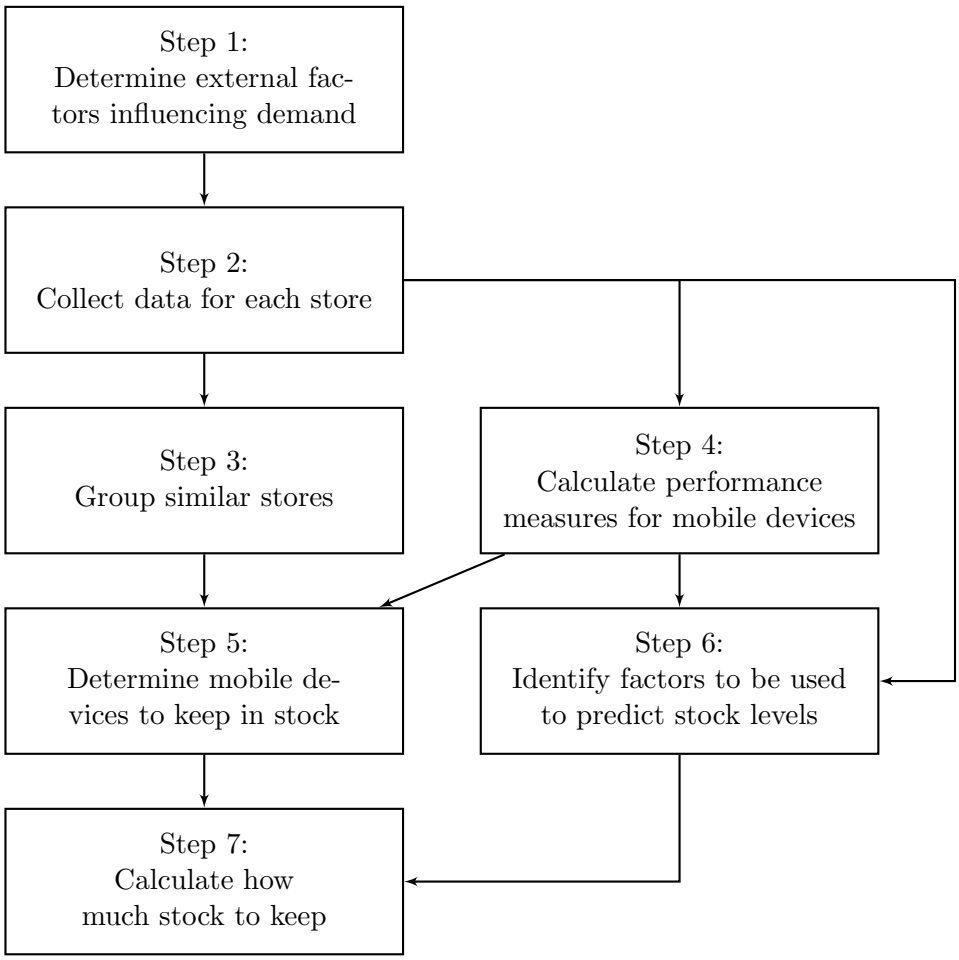


FIGURE 3.1: *Project flowchart*

In Step 4, The Retailer's data are used to calculate performance measures for the different mobile devices. These measures are used to evaluate the performance of the different devices in order to determine the range of mobile devices to keep in stock in each of the store groups identified in Step 3. These ranges are determined in Step 5 in Figure 3.1 and meet Objective IV in §1.3. The methods used to determine the performance measures and the subsequent mobile device ranges are discussed in §5.1.

For the next step, only devices identified in Step 5 that now need to be carried at stores that did not previously carry the device, are considered. To determine the stock levels for these devices, the different factors that could influence the amount of stock to keep, are first identified as indicated in Step 6 in Figure 3.1. This is done by using the performance measures calculated in Step 4, as well as The Retailer's data collected in Step 2. Finally, the amount of stock to keep for each mobile device identified to be ranged, is calculated in Step 7 using regression trees. These Steps address Objective V in §1.3 and are discussed in more detail in §5.2.

3.2 Data collection and manipulation

In this section, the various data sources utilised in this research project will be discussed. First, the data provided by The Retailer is described in §3.2.1. This is followed by discussions regarding

stores' network coverage, as well as population statistics in §3.2.2 and §3.2.3, respectively.

3.2.1 The Retailer's data

Four datasets are received from The Retailer. The first dataset contains daily sales data for mobile devices, while the second dataset contains stock data. The last two datasets provide more information on stores and products. The Retailer's sales and stock data will be used in Steps 4 to 7 in Figure 3.1 in order to calculate performance measures, determine the mobile device ranges for the different store groups and calculate how much stock to keep for each mobile device.

Sales data

The sales dataset received from The Retailer consists of daily mobile device sales per store for one year between June 2020 and June 2021. The data is aggregated to provide the total sales of each mobile device in each store per day. As shown in Table 3.1, this dataset contains the number of transactions, number of units and total sales amount of a particular mobile device sold in each store on a specific day. For example, in store A_001 on day 26, one unit of product 00141 was sold for R150. Similarly, on day 1, two units of product 00258 were sold in store B_016 for a total of R1 500.

Store	Product	Day	Number of transactions	Number of units	Total sales amount
A_001	00141	26	1	1	R150
B_016	00258	1	2	2	R1 500
B_432	00257	1	1	2	R280
A_001	00132	36	1	1	R300

TABLE 3.1: *An extract of mobile device sales data received from The Retailer.*

Figure 3.2 contains the total number of mobile devices sold per day, aggregating all mobile devices across all stores. This graph shows a clear seasonal trend with peaks roughly every 30 days, which could indicate that there is an increase in sales during month-end. On average, 628 mobile device units are sold per day over the one year time period. The highest number of mobile device units sold was 2 164 units on day 36, which coincides with the third wave of Covid-19 infections in South Africa and, as a result, lockdown restrictions escalated to level 4. The lowest number of units sold was 161 units on day 171.

Stock data

The next dataset provides the opening stock levels of mobile devices in each store on a particular day. Table 3.2 contains an example of the stock data received from The Retailer. This dataset consists of the store and product codes, the day as well as the stock quantity at the beginning of the day. For example, in store A_046 on day 206, there were 10 units of mobile device 00284 in stock. Similarly, store B_061 had 16 units of mobile device 00277 in stock on days 184 and 185, indicating that no units were sold on day 184. On day 217, stores B_371 and B_478 had only one and eight units of device 00277 in stock, respectively. Both of these stores are smaller than store B_061 and are also located in municipalities with a smaller population, which could

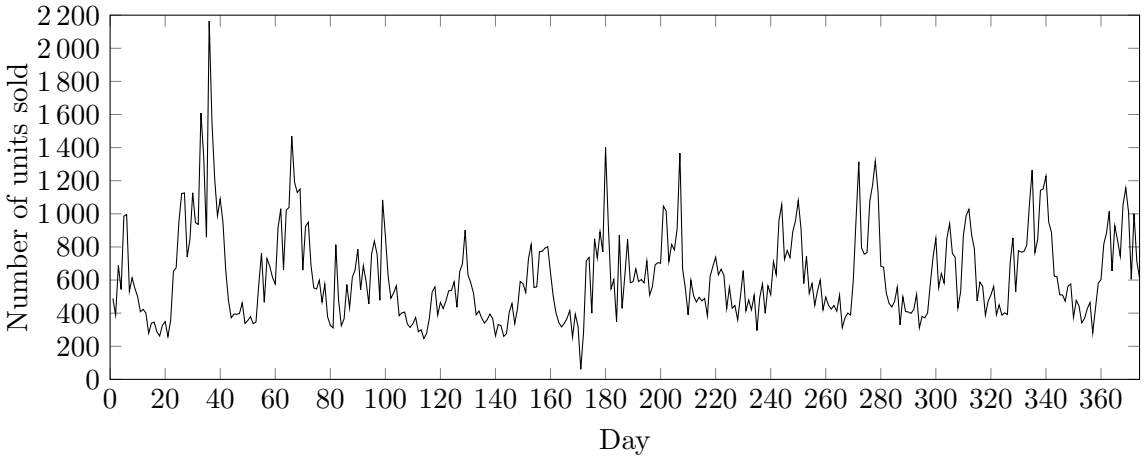


FIGURE 3.2: *The aggregated sales quantity of mobile devices across The Retailer's stores.*

explain their low stock quantity in comparison to store B_061. In total, approximately 9 months of stock data are received between October 2020 and June 2021.

Store	Product	Day	Quantity
A_046	00284	206	10
B_061	00277	184	16
B_061	00277	185	16
B_371	00277	217	1
B_478	00277	217	8

TABLE 3.2: *An extract of mobile device stock data received from The Retailer.*

The stock data, aggregated to the total number of mobile device units available per day over all stores, can be seen in the black line Figure 3.3. The data do not seem to have any trend or seasonality, but stay relatively constant between 120 000 and 150 000 units available until day 323, where there is a large decrease in the number of units to approximately 15 000 units per day. Further, the number of mobile devices for which stock data are recorded, indicated by blue circles in Figure 3.3, is stable between 130 and 140 mobile devices per day, which decreases to approximately 100 mobile devices per day after day 323. This decrease also coincides with a decrease in the number of units in stock for each of the different devices for which stock was recorded after day 323. During 25 days no stock data were recorded, as indicated by the zero values on days 228, 247 and 279, for example. The missing data, as well as the decline in mobile devices with recorded stock data make it reasonable to assume that the decrease in the stock quantity is also due to data recorded incorrectly. Thus, the 25 days with no stock data and the last 35 days with low stock levels are removed from the dataset, resulting in 180 days of data to be used in the analysis.

Stores data

The third dataset contains the supermarket type, store size, province, city and suburb of each store, as illustrated in Table 3.3. The Retailer's stores can be found in all nine provinces in South Africa, with a higher density of stores in the metropolitan areas.

After comparing The Retailer's sales, stock and stores datasets, it is clear that not all of The

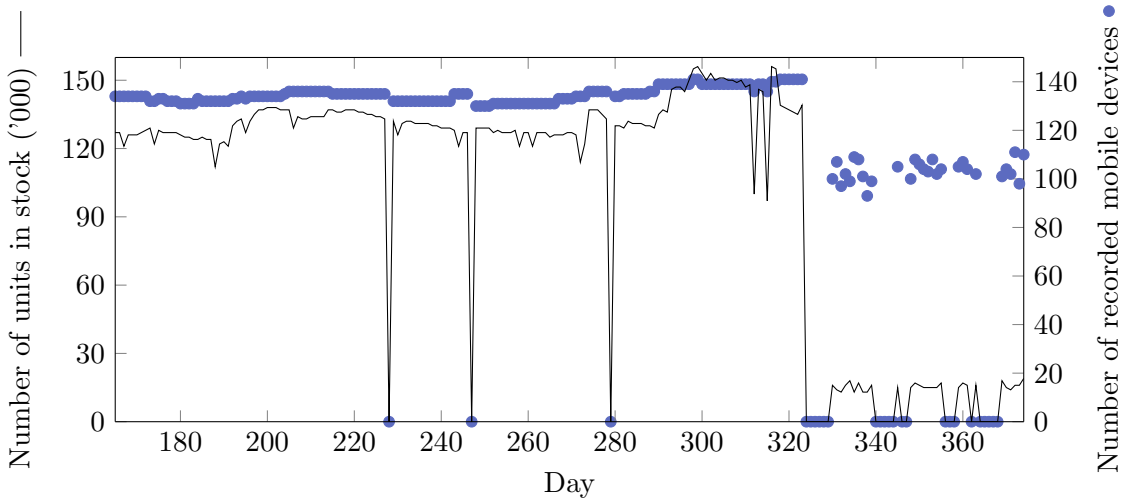


FIGURE 3.3: *The aggregated stock quantity of mobile devices across The Retailer's stores, as well as the number of mobile devices for which stock data are recorded per day.*

Store	Supermarket type	Store size (m^2)	Province	City	Suburb
A_001	Type A	226	Western Cape	Langebaan	Langebaan
B_001	Type B	2 127	Gauteng	Florida	Florida Lake
C_001	Type C	1 9917	Free State	Welkom	Welkom

TABLE 3.3: *An extract of store data received from The Retailer.*

Retailer's stores sell mobile devices. However, as seen in Figure 3.4, most stores and specifically, 350 type A, 469 type B and 238 type C stores, sold mobile devices in the observed time period. Only these 1 057 stores will thus be considered in the remainder of this project.

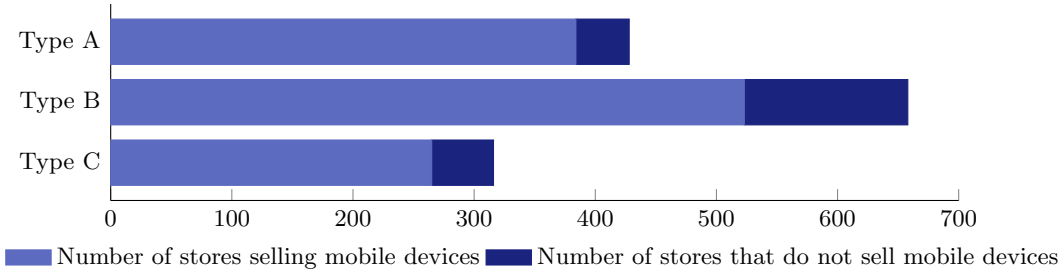


FIGURE 3.4: *The number of stores in each supermarket type that sell and do not sell mobile devices.*

Product data

The product data is the final dataset obtained from The Retailer. In this dataset, a description, brand and, in some cases, colour of each mobile device sold by The Retailer, are given. Table 3.4 contains an extract of the mobile device product data received. As indicated in this table, different variations, for example different colours of the Hauwei P8 Lite 2017, are recorded as individual products. This dataset consists of 335 unique products, however, only 183 unique products were sold in the year for which sales data were received. The Retailer keeps record of all products sold at any given time and the remaining 152 unsold mobile devices are, therefore,

Product	Description	Brand	Colour
00178	P8 LITE 2017	HUAWEI	MIDN BLK
00180	P8 LITE 2017	HUAWEI	PLAT GLD
00225	P8 LITE 2017	HUAWEI	BLUE

TABLE 3.4: *An extract of mobile device product data received from The Retailer.*

products that were sold in the past but are no longer ranged by The Retailer. Figure 3.5 illustrates the distribution of the total number of stores in which the relevant 183 mobile devices are sold. In total, eight mobile devices were sold in only one store each. The value of the first quartile is six, while the median number of stores is 20. The third quartile is found at 48 stores and the maximum is calculated as 83 stores. Thus, 20 mobile devices are identified as outliers based on the number of stores in which they are sold. This indicates that most mobile devices are only ranged in a small number of stores, increasing the complexity of The Retailer's mobile device range.

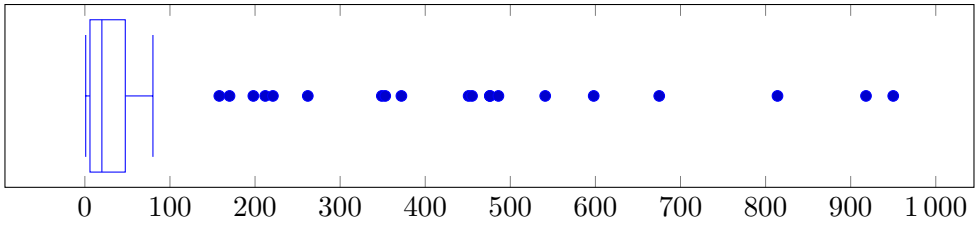


FIGURE 3.5: *The number of stores in which each of the different mobile devices are sold.*

3.2.2 Stores' network coverage

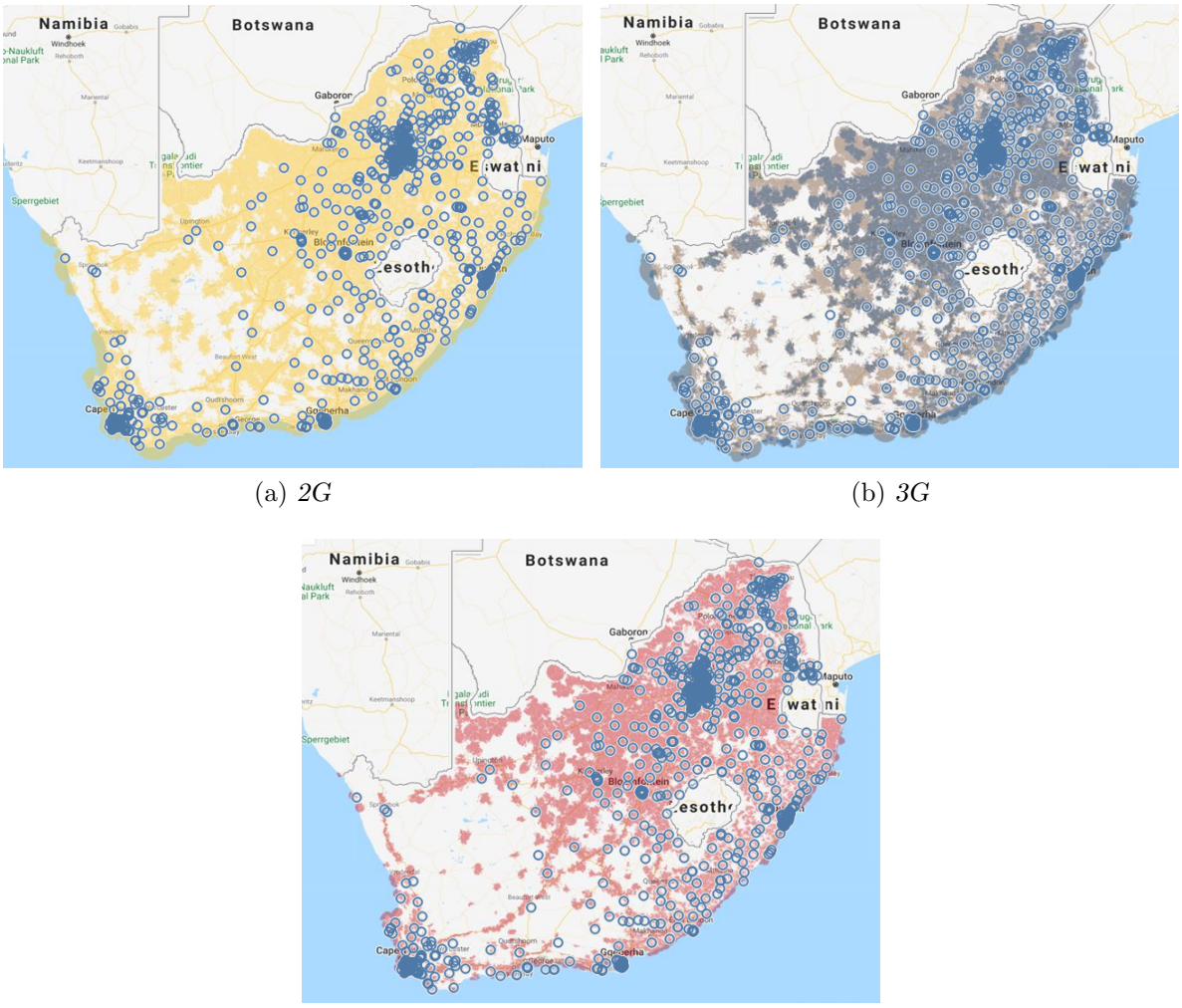
Many TSPs provide maps indicating their network coverage across South Africa. By comparing the suburbs provided by The Retailer to the network coverage of TSPs in South Africa, one can determine the network coverage at each store's location.

The first step to determine each store's network coverage, is to collect network coverage maps from the three largest TSPs in South Africa, Vodacom, MTN and Cell C. After these coverage maps are collected, the suburbs wherein The Retailer has one or more stores, are plotted onto the maps in Tableau [43]. If a suburb falls within the coverage area, it is assumed that all stores in that suburb are covered by the specific network. Similarly, if a suburb falls outside the coverage area, it is assumed that the stores in that suburb are not covered by the specific network. Finally, if a suburb is on the border of the coverage area, that suburb is manually searched on the relevant TSPs website, to confirm whether or not the particular suburb is covered.

For the purpose of this study, 5G coverage is not considered since this is currently only available in very small areas in South Africa. Furthermore, 5G capable devices are currently also not sold by The Retailer.

After following the steps outlined above for each of the networks, the coverage maps for MTN's 2G, 3G and 4G networks can be seen in Figure 3.6 as examples. The suburbs provided by The Retailer can also bee seen on these maps in Figure 3.6, indicated as circles. Similar coverage maps are also found for Vodacom and Cell C, and these maps can be seen in Appendix B. By studying the maps of all three networks, it can be seen that all of The Retailer's stores are

covered by 2G, 3G and 4G networks. Therefore, a store's network coverage no longer needs to be considered as a factor when grouping The Retailer's stores.



(a) *2G*

(b) *3G*

(c) *4G*

FIGURE 3.6: *A comparison of The Retailer's store locations and the coverage maps for MTN's 2G, 3G and 4G networks [30], where the coloured areas in each map indicate the specific type of coverage and the blue circles indicate store locations.*

3.2.3 Population statistics

The population related factors that form part of the external decision criteria identified in the literature, as described in §1.2 are the number of landlines, rate of population change, population age and income. Data for each of these decision criteria are collected from various StatsSA surveys, via the Eighty20 Data Portal [8].

In this research project, the 2011 Census, as well as the 2007 and 2016 Community surveys are used to collect data for all local municipalities in South Africa, since the municipal data associated with a specific store's location, are more relevant than provincial data, for example. The census and community surveys are only conducted every 10 years and are the only surveys with a large enough sample size to produce results for municipalities, rather than just national or provincial data. To ensure that the data from these surveys are still relevant in 2021, the

General Household survey is used. This survey is conducted yearly and its provincial results are used to analyse the change in the above-mentioned criteria over time.

Number of landlines

The landline data are collected from the 2016 Community survey. An extract from this data can be seen in the first five columns of Table 3.5. As seen in columns 2 to 4 in Table 3.5, the survey provided three possible answers for respondents to indicate whether or not their household owns a landline, namely 'Yes', 'No' and 'Unspecified'. The results are expressed as the total number of households that gave each of the three answers in the specified municipality in column 1. For example, in the City of Cape Town, 333 551 households indicated that they own a landline, while 878 286 households do not own a landline. Furthermore, the total number of households in the City of Cape Town is 1 264 949, as displayed in column 5 (Grand total) in Table 3.5.

For the purpose of this study, the 'number of landlines' criterion for a municipality is indicated by the number of households that responded 'Yes', as a percentage of the total number of households in the specific municipality. This is done to ensure that the various municipalities are comparable, even though they are of different sizes. For example, the City of Cape Town has over 1 million households, while the Cederberg municipality has only 15 279 households. Therefore, in the City of Cape Town, the percentage of households owning a landline is 26,4%, or

$$\frac{333\,551}{1\,264\,949} = 0,2638,$$

and this calculated value for the 'percentage owning a landline' variable can be seen in the last column of Table 3.5.

	Own a landline:				
Local municipality	Yes	No	Unspecified	Grand total	Percentage owning a landline
CPT: City of Cape Town	333 551	878 386	53 011	1 264 949	26,4%
WC011: Matzikama	3 235	16 898	687	20 821	15,5%
WC012: Cederberg	2 400	12 573	306	15 279	15,7%

TABLE 3.5: *An extract from the 2016 Community survey landline data.*

Rate of population change

The rate of population change is calculated by using the 2007 and 2016 Community surveys. Table 3.6 contains an example of the population size per municipality in 2007 and 2016. Let $m_{2007,i}$ be the population size for municipality i in 2007 and $m_{2016,i}$ the population size for municipality i in 2016. Then, the rate of population change for municipality i is calculated as

$$\frac{m_{2016,i} - m_{2007,i}}{m_{2007,i}} \div 9$$

to find the average percentage of change per year over the 9-year period from 2007 to 2016, as indicated in Table 3.6.

Local municipality i	$m_{2007,i}$	$m_{2016,i}$	Average % rate of change
CPT: City of Cape Town	3 497 097	4 005 016	1,6%
WC011: Matzikama	46 362	71 045	5,9%
WC012: Cederberg	31 942	52 949	7,3%

TABLE 3.6: *An extract of the 2007 and 2016 Community survey population data and the subsequent calculated percentage rate of population change per year.*

Population age

In the 2016 Community survey, the total number of residents in a municipality in each age group, can be found as well. Table 3.7 contains a small extract of the age data found in the 2016 Community survey. In the City of Cape Town, there are, for example, 74 017 babies under the age of one, while Matzikama has 1 495 babies under the age of one. Across the country the ages of residents range between zero and 116 years, with a total of 213 residents at age 116 years.

	Age (in years)					
Local municipality	0	1	2	3	4	5
CPT: City of Cape Town	74 017	72 290	78 439	70 603	73 070	71 506
WC011: Matzikama	1 495	1 653	1 311	1 350	1 602	1 310
WC012: Cederberg	1 190	1 007	1 108	826	884	861

TABLE 3.7: *An extract of the first six age groups from the 2016 Community survey age data.*

The population age in each municipality is then represented by the average age of all residents in that municipality. The average age over all age groups from 0 to 116 years for the municipalities in Table 3.7 is given in Table 3.8. As indicated in Table 3.8, the average age of residents in the City of Cape Town is 30,62 years, while the average age for Matzikama is 29,59 years.

Local municipality	Average age in years
CPT: City of Cape Town	30,62
WC011: Matzikama	29,59
WC012: Cederberg	30,38

TABLE 3.8: *The average age of residents in the three municipalities in Table 3.7.*

Population income

Annual population income data are collected from the 2011 Census and, as indicated in Table 3.9, are grouped into different income brackets. In the raw population income data, there are also 'Unspecified' and 'N/A' options. For the purpose of this study, the 'Unspecified' and 'N/A' columns are excluded and contribute to a total of 4,2% (or 2,2 million) of the South African population. The remaining 12 income brackets are used to represent the population income in the various municipalities.

As seen in Table 3.9, the difference between the lower and upper limits of each income bracket is not constant, since the upper limit in each range is exactly twice the lower limit -1. In other

Local municipality	Annual household income brackets				
	No income	R1 - R4 800	R4 801 - R9 600	R9 601 - R19 200	R19 201 - R38 400
CPT: City of Cape Town	354 267	86 283	156 589	326 989	575 424
WC011: Matzikama	3 397	1 248	2 370	7 692	15 304
WC012: Cederberg	2 669	313	1 298	5 972	12 036

Local municipality	Annual household income brackets (continued)				
	R38 401 - R76 800	R76 801 - R153 600	R153 601 - R307 200	R307 201 - R614 400	R614 401- R1 228 800
CPT: City of Cape Town	602 817	515 034	430 996	319 404	130 142
WC011: Matzikama	15 165	8 973	6 301	2 709	1 031
WC012: Cederberg	12 937	7 642	2 537	1 540	476

Local municipality	Annual household income brackets (continued)			
	R1 228 801 - R2 457 600	R2 457 601 or more	Unspecified	N/A
CPT: City of Cape Town	34 285	16 598	495	188 042
WC011: Matzikama	235	108	0	2 478
WC012: Cederberg	154	48	0	2 426

TABLE 3.9: *An extract from the 2011 Census data pertaining to household income.*

words, the higher the income bracket is, the wider the income range for that bracket becomes. Furthermore, the last income bracket has no upper limit. This makes it difficult to summarise the population income into just one variable, as the case was for the average population age calculated above. However, it is also not feasible to use 12 income variables when grouping stores, as these variables are correlated and will highly skew the outcome of the grouping. Thus, the above 12 income brackets are summarised into four new income brackets representing the population found in a specific income bracket as a percentage of the total population. Table 3.10 contains the new income brackets and their percentages for the three municipalities given in Table 3.9.

Local municipality	Annual household income brackets			
	R0 - R38 400	R38 401 - R153 600	R153 601 - R614 400	R614 401 or more
CPT: City of Cape Town	42,2%	31,5%	21,1%	5,2%
WC011: Matzikama	46,5%	37,4%	14,0%	2,2%
WC012: Cederberg	46,8%	43,2%	8,5%	1,4%

TABLE 3.10: *The population distribution over different income brackets for the three municipalities for which the raw data were given in Table 3.9.*

CHAPTER 4

Grouping similar stores

Contents

In this chapter, the process of grouping similar stores is discussed to address Objective III in §1.3. The chapter commences in §4.1 where the data, as explained in §3.2, is further examined with a special focus on the scaling of variables. Then, in §4.2, the results of the different clustering methods are compared and the best method is selected to be used in the rest of this project. Finally, the clusters found in §4.2 are evaluated in §4.3.

4.1 Data preparation

Remember from §1.2 that, based on literature, the number of landlines [21], the rate of population change [5], the population age [51] and household income [21] are identified as potential external decision criteria to consider when similar stores are grouped. The Retailer's supermarket types, as well as the network coverage in an area [40] were also considered as potential criteria when grouping stores. However, in §3.2.2 it was found that all of The Retailer's stores have 2G, 3G and 4G coverage. Since there is no difference between the stores, the network coverage in an area is not used to group similar stores together.

Furthermore, The Retailer takes the customers and LSM classifications around a store into consideration when the particular store is classified into one of The Retailer's three supermarket types. Thus, using supermarket types as a criterion might put an inflated weight on the population income. For this reason, the clustering method is applied to each of the three supermarket types individually, which is also in line with the current practice of The Retailer to carry different mobile device ranges for each of the supermarket types. Therefore, the number of landlines, rate of population change, population age and population income remain as the only criteria to group similar stores.

Since clustering methods, in general, are based on a distance measure, it is important to ensure that all variables are on a similar scale. In §3.2.3, the number of landlines and population income variables were transformed into a percentage, thus a decimal number between zero and one. However, the population age variable ranges from 20,94 to 34,41 and the rate of change variable can contain negative values when the population size decreased. Both of these variables are consequently normalised to values between zero and one using `MinMaxScaler` in the `sklearn.preprocessing` Python package [34]. `MinMaxScaler` transforms each variable y so that

$$y_{scaled} = \frac{y - \min(y)}{\max(y) - \min(y)},$$

where y is the original value of the variable and $\min(y)$ and $\max(y)$ are the minimum and maximum values, respectively, that y can be. This transformation is done across all values, regardless of which store or supermarket type corresponds to each value.

Table 4.1 contains the original and scaled values of the average age and rate of population change variables for five stores of supermarket type A. For store A_001, the scaled value for the average age calculated with the above formula is 0,66. Similarly, the scaled value for the average rate of population change for this store is 0,21.

Store	Average age	Scaled average age	Average % rate of change	Scaled average % rate of change
A_001	29,81	0,66	4,5%	0,21
A_002	28,91	0,59	0,8%	0,12
A_003	23,24	0,17	1,7%	0,14
A_004	29,11	0,61	6,0%	0,25
A_005	30,37	0,70	3,0%	0,17

TABLE 4.1: *The original and scaled values of the age and rate of population change variables for five stores.*

Only stores currently selling mobile devices are used in the clustering since The Retailer will continue to sell mobile devices in these stores only. Therefore, as discussed in §3.2.1, 350 type A stores, 469 type B stores and 238 type C stores are subsequently used in the clustering.

4.2 Choosing a clustering method

An agglomerative hierarchical clustering approach is used to group stores, since there is no predefined number of groups that are required. The working of this hierarchical clustering method is discussed in Appendix C and implemented in this project by using the `scipy.cluster.hierarchy` package in Python [44].

To determine the best linkage method, as well as the number of clusters needed to group The Retailer's stores, the relative validation criteria, as discussed in §2.2.5, are used. The average *silhouette coefficient* (SC) for each cluster will be calculated first. This is done using the Euclidean distance and five different linkage methods, namely single, complete, average, centroid and Ward's linkage, as discussed in Appendix C, §C.1. Secondly, dendrograms with the Euclidean distance and the same linkage methods, will also be analysed before choosing the final clusters that will be used in the remainder of this project.

4.2.1 Average Silhouette coefficient

As described in Appendix C, §C.3, the SC of a store compares the distance between each store and the other stores in its cluster with the distance between the original store and all other clusters. This coefficient can take values between -1 and 1, where a higher coefficient means a better fitted clustering. The average SC over all stores for a specific clustering, is calculated with the `sklearn.metrics.silhouette_score` package in Python [34] and the results for The Retailer's three store types are depicted in Figures 4.1, 4.2 and 4.3, respectively.

Forming two clusters with complete linkage, as indicated in green in Figure 4.1, seems to be the best method to cluster the stores in supermarket type A, where the average SC has a value of 0,50. This is followed closely by two clusters formed with Ward's linkage (orange), with a value of 0,49, as well as two clusters formed with average linkage (red), with a value of 0,48. Across all linkage methods, the average SC decreases as the number of clusters increases.

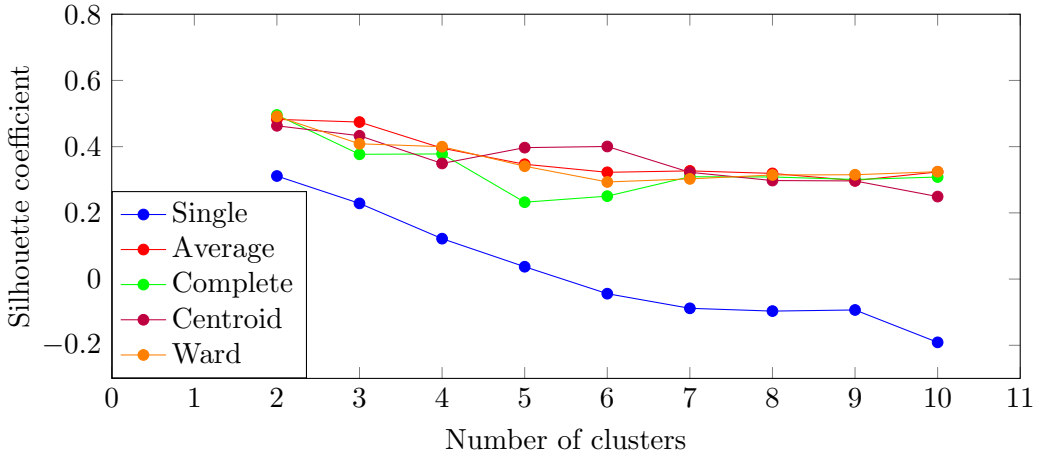


FIGURE 4.1: *The average silhouette coefficient values for different linkage methods and number of clusters formed for the stores in supermarket type A of The Retailer.*

Figure 4.2 contains the average SC values for the stores in supermarket type B. Single, average and centroid linkage have the same coefficient value of 0,67 when two clusters are formed. Then, with three or more clusters, the average SC for all linkage methods start to decline. However, the average SC for complete and Ward's linkage increase towards the right of the graph, where the highest number of clusters are shown.

Finally, in Figure 4.3, the average SC values for the stores in supermarket type C can be seen. In the case of two clusters, all linkage methods except for the Ward's method, have a coefficient value of 0,76. As seen with supermarket types A and B as well, the average SC value decreases when more than two clusters are formed, but with seven or more clusters, the average SC values for the average and Ward's linkage methods increase again.

For all three supermarket types, single linkage performed significantly worse than the other linkage methods in this test. Across the three supermarket types, forming two clusters give the best performance, regardless of the linkage method used.

4.2.2 Dendrograms

Various dendrograms are also drawn using Euclidean distance and different linkage methods, as discussed in Appendix C, §C.2. The default parameter for the distance threshold used in the

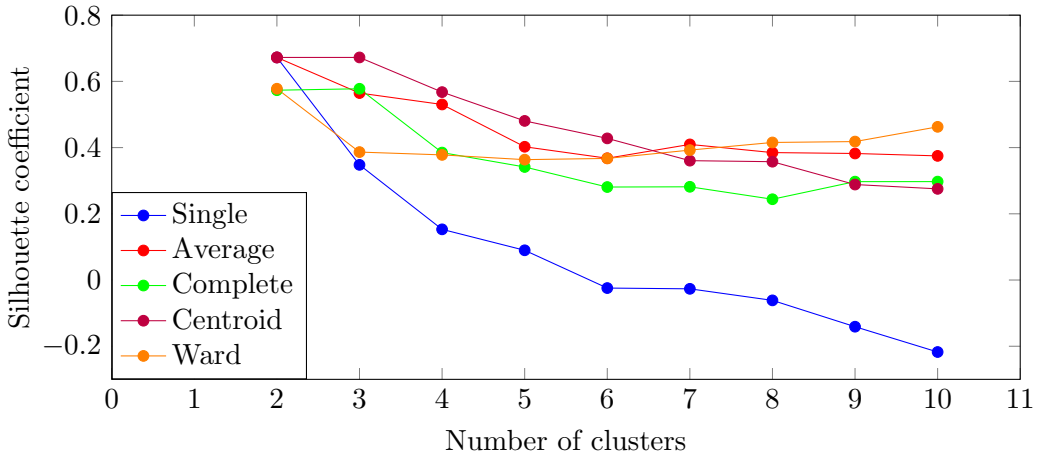


FIGURE 4.2: *The average silhouette coefficient values for different linkage methods and number of clusters formed for the stores in supermarket type B of The Retailer.*

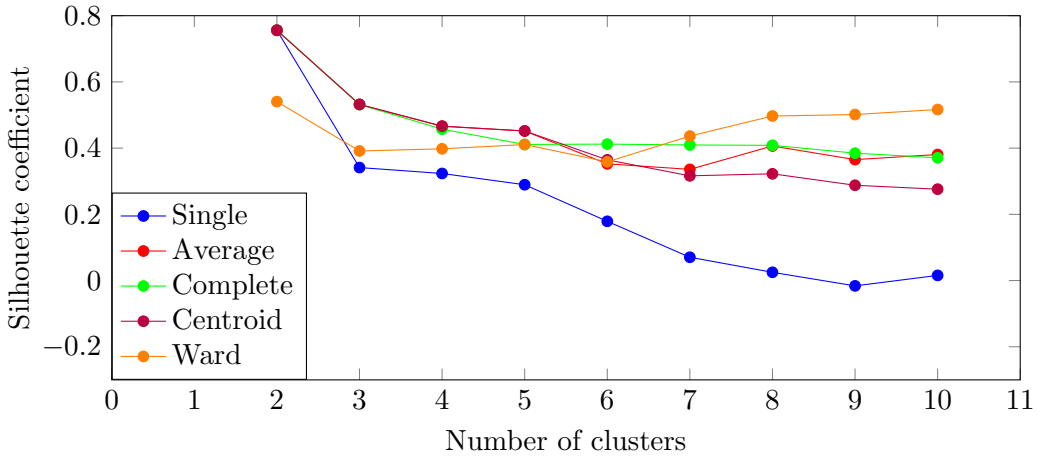


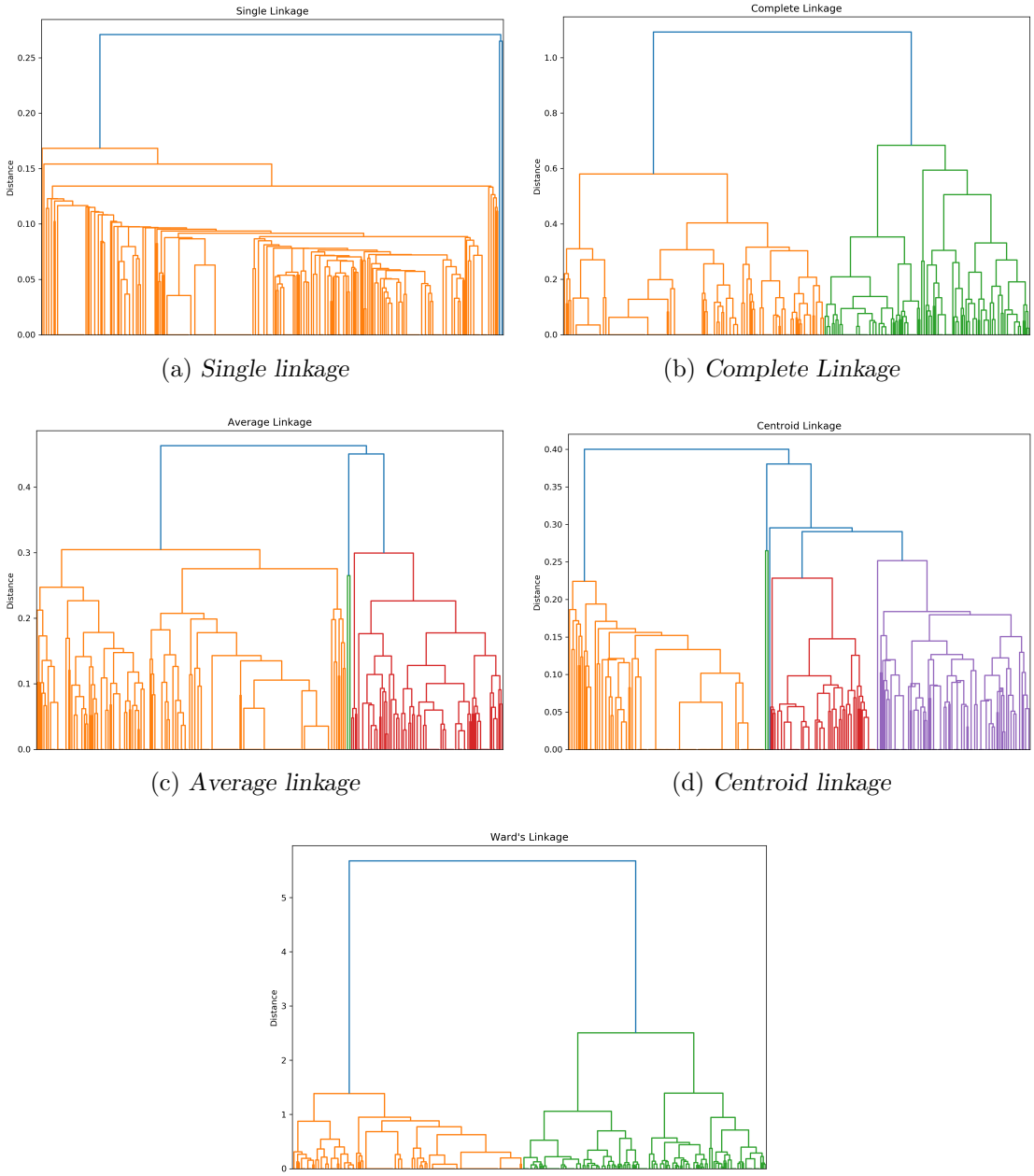
FIGURE 4.3: *The average silhouette coefficient values for different linkage methods and number of clusters formed for the stores in supermarket type C of The Retailer.*

`scipy.cluster.hierarchy.dendrogram` package in Python [44] is used in this project and is calculated as 70% of the maximum distance on the y-axis [44].

Figures 4.4, 4.5 and 4.6 show the dendrograms for each of the three supermarket types using single, complete, average, centroid and Ward's linkage. In these figures, each cluster is highlighted in a different colour and the suggested cut-off is indicated by the blue branches, while the number of stores in each cluster is given in Tables 4.2, 4.3 and 4.4, respectively.

The dendrograms for supermarket type A stores are shown in Figure 4.4. In Figure 4.4(a), single linkage is used and three clusters are formed. As seen in the graph, the default cut-off point is selecting a point of the dendrogram with the largest distance before the next cluster is formed, indicating that stores within these clusters are more similar to each other, and less similar to stores in other clusters. However, as seen in Table 4.2, the cluster sizes found with single linkage are extremely uneven. One cluster, indicated in orange in Figure 4.4(a), contains 346 of the 350 stores, while the other two clusters contain two stores each.

With complete linkage, in Figure 4.4(b), two clusters are formed and again, this clustering is found at the largest distance on the dendrogram. These clusters are more even than those found



(a) *Single linkage* (b) *Complete Linkage*

(c) *Average linkage* (d) *Centroid linkage*

(e) *Ward's linkage*

FIGURE 4.4: *Dendrograms for supermarket type A stores using various linkage methods.*

	Number of stores per cluster				
Linkage method	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Single	346	2	2		
Complete	194	156			
Average	230	4	116		
Centroid	149	4	78	118	1
Ward	169	181			

TABLE 4.2: *The number of supermarket type A stores per cluster using different linkage methods.*

with single linkage and, as shown in the second row in Table 4.2, contain 194 and 156 stores, respectively. Figure 4.4(c) and the third row in Table 4.2 show the three clusters found when average linkage is used. Again, uneven clusters are formed, containing 230, 4 and 116 stores, respectively, as indicated in orange, green and red, respectively, in Figure 4.4(c).

The fourth linkage method, namely centroid linkage, is illustrated in Figure 4.4(d). With this method, five uneven clusters are formed, with one cluster containing only one store and the largest cluster containing 149 stores. However, on this dendrogram it can be seen that the default cut-off value is not necessarily the best cut-off point in this case, as the distance between three and four clusters is larger than the distance between five and six clusters.

Finally, the dendrogram for the Ward's linkage method is shown in Figure 4.4(e). The large distances between the top of the dendrogram and the top of each of the two clusters seen in this graph indicate that this is a good clustering, as the two clusters are not very similar, while the stores within each of the clusters are very similar (there is a small distance between clusters merging within the two big clusters). As seen in the last row of Table 4.2, this method produces the most balanced clusters of all five methods, with only two clusters containing 169 and 181 stores, respectively.

Figure 4.5 provides the dendrograms for supermarket type B stores. With single linkage, as shown in Figure 4.5(a), supermarket type B stores are divided into two clusters. Although this is a good cut-off based on the distance measure, the resulting clusters are skewed again, with two stores in one cluster and all remaining stores in a second cluster, as indicated in Table 4.3.

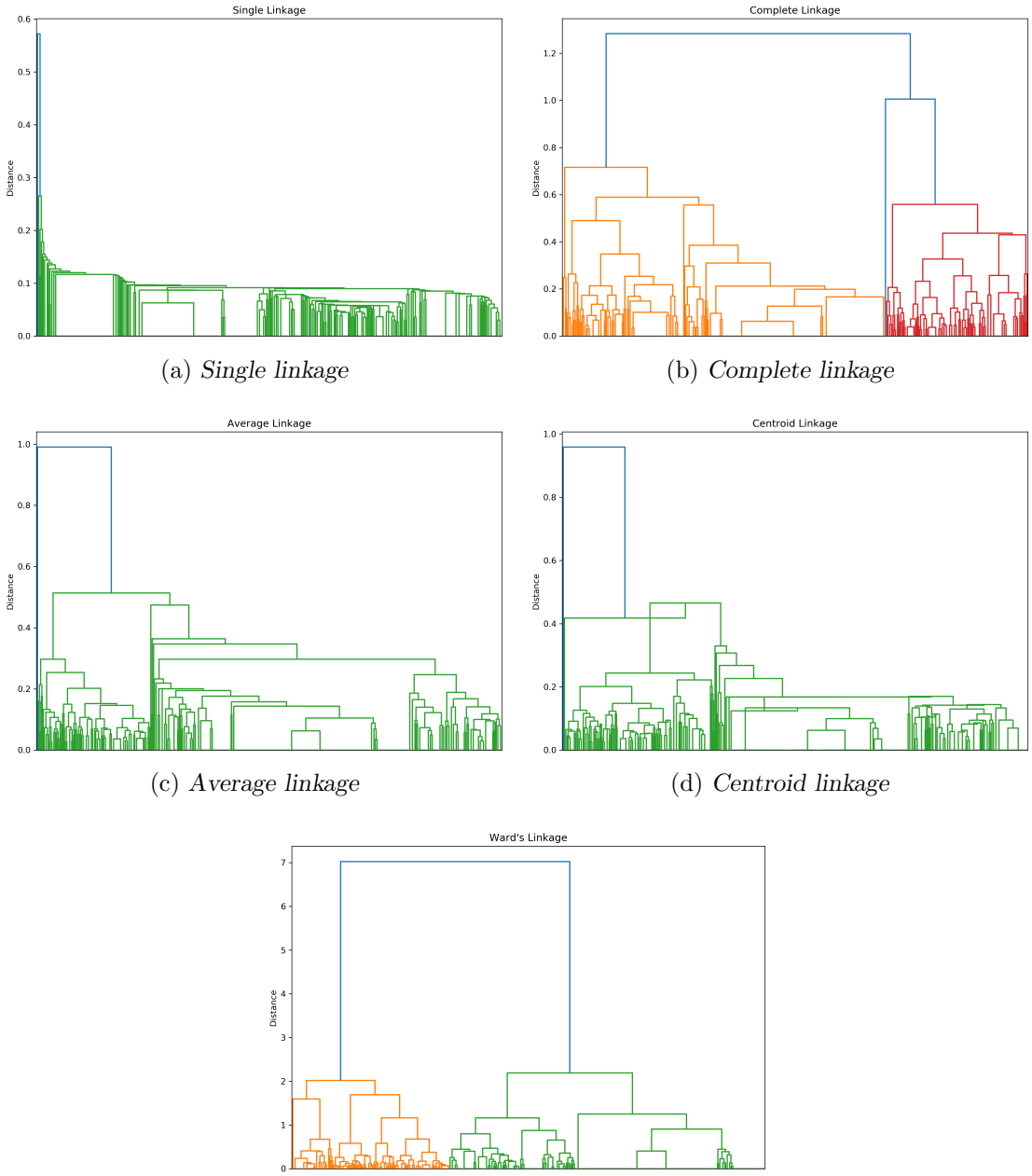
	Number of stores per cluster		
Linkage method	Cluster 1	Cluster 2	Cluster 3
Single	2	467	
Complete	326	2	141
Average	2	467	
Centroid	2	467	
Ward	165	304	

TABLE 4.3: *The number of supermarket type B stores per cluster using different linkage methods.*

Figure 4.5(b) shows that, with complete linkage, three clusters are formed. Based on the distance measure, three clusters are in fact the best grouping, as this is the point of the dendrogram where clusters have the largest distance before the next merging. However, these clusters contain 326, 2 and 141 stores, respectively, and this is also a skewed distribution between the clusters. Average and centroid linkage, seen in Figures 4.5(c) and (d), each produce clusters identical to those found with single linkage, with 2 and 467 stores in each of the two clusters.

With Ward's linkage, two clusters are also formed. Contrary to the other linkage methods, as shown in Table 4.3, these clusters contain 165 and 304 stores, respectively. This is a better clustering for determining product ranges, as it is not viable to maintain a unique product range for only two stores.

For supermarket type C stores, all five linkage methods produce two clusters as this is where all five dendrograms are cut at the largest distance to the next merging of clusters, as seen in Figure 4.6. As Table 4.4 indicates, the single, complete, average and centroid linkage methods produce similar clusters, even though the branches of the dendrogram do not follow the same pattern. The two clusters found with these methods are completely skewed, with only one store in one cluster and the remaining 237 stores in the other cluster.



(a) *Single linkage* (b) *Complete linkage*

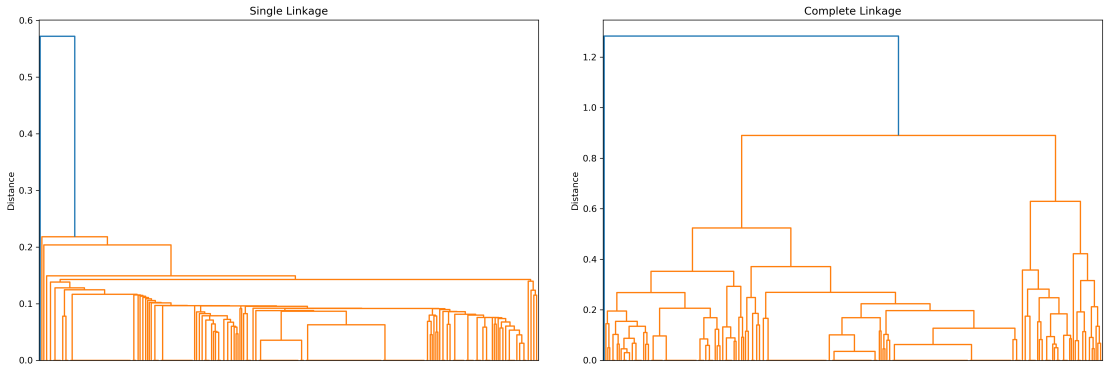
(c) *Average linkage* (d) *Centroid linkage*

(e) *Ward's linkage*

FIGURE 4.5: *Dendrograms for supermarket type B stores using various linkage methods.*

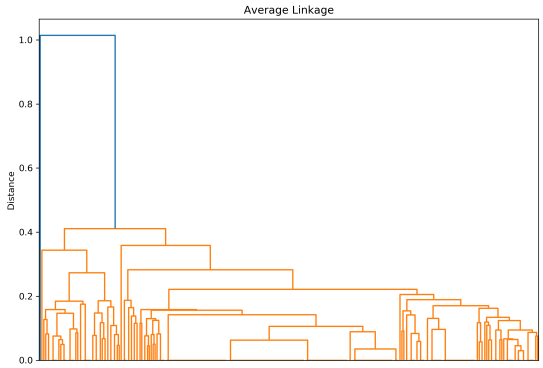
Linkage method	Number of stores per cluster Cluster 1	Cluster 2
Single	237	1
Complete	237	1
Average	237	1
Centroid	237	1
Ward	43	195

TABLE 4.4: *The number of supermarket type C stores per cluster using different linkage methods.*

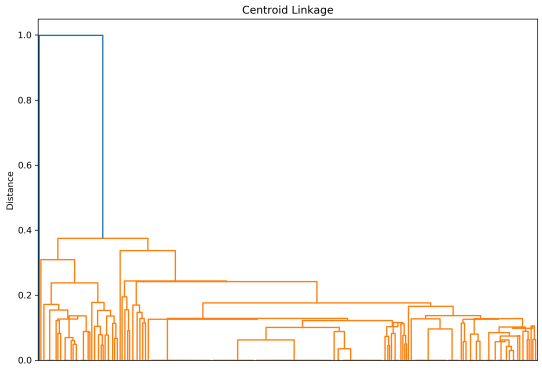


(a) *Single linkage*

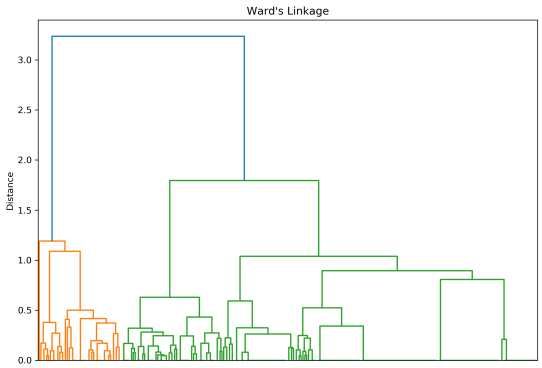
(b) *Complete linkage*



(c) *Average linkage*



(d) *Centroid linkage*



(e) *Ward's linkage*

FIGURE 4.6: *Dendrograms for supermarket type C stores using various linkage methods.*

It can be seen from the dendrogram in Figure 4.6(e) and the last row in Table 4.4 that the clusters formed with Ward's linkage, with 43 and 195 stores, respectively, are a slightly more balanced grouping for the supermarket type C stores.

After analysing the average silhouette coefficients and dendrograms, it can be seen that, for all three supermarket types, two clusters are the most appropriate number to use. The dendrograms indicate that Ward's linkage method produces the most well balanced clusters for each of the three supermarket types. Therefore, the clusterings obtained by Ward's linkage with two clusters for each of the three supermarket types will be used, resulting in six clusters in total.

4.3 Interpreting the clusters

It is clear from the average silhouette coefficients and dendrograms that stores within the final six clusters are similar to other stores within the same cluster, and different to stores in other clusters. To understand where these differences lie, the distributions of the variables within each cluster are considered.

Figure 4.7 contains the distribution of the landlines percentage variable. This graph shows that in each supermarket type, one of the two clusters has a higher percentage of landlines than the other. In supermarket type A, cluster one has a median of 0,16, while cluster two has a median of only 0,03. Similarly in supermarket type B, cluster one has a median of 0,02, while cluster two also has a higher median of 0,16. Finally in supermarket type C, cluster one has a median of 0,05, while cluster two has a median of 0,16.

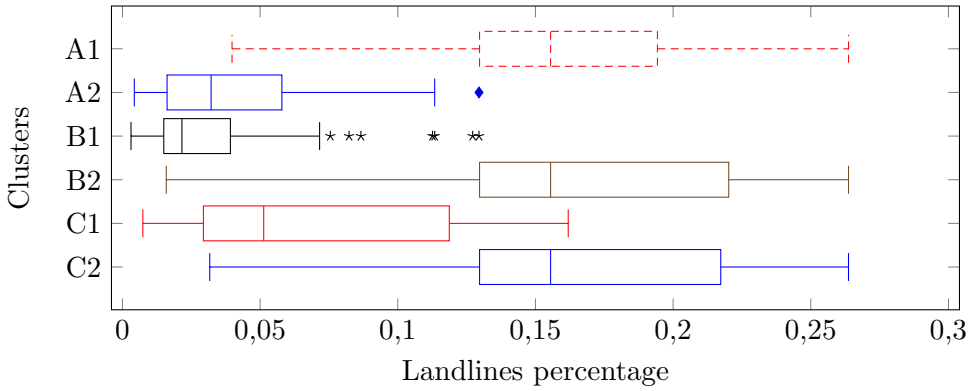


FIGURE 4.7: *The distribution of the landlines percentage variable for each cluster.*

The distribution of the scaled rate of population change variable for the six clusters is shown in Figure 4.8. Although the differences between the cluster values are not as big as the differences seen with the landlines percentage variable, one of the two clusters in each supermarket type still has a slightly higher value than the other. For example, for supermarket type A, the median value for this variable in cluster one is 0,17, while the median value for cluster two is 0,12. Further, it is clear that supermarket type B contains two outliers, one in each cluster.

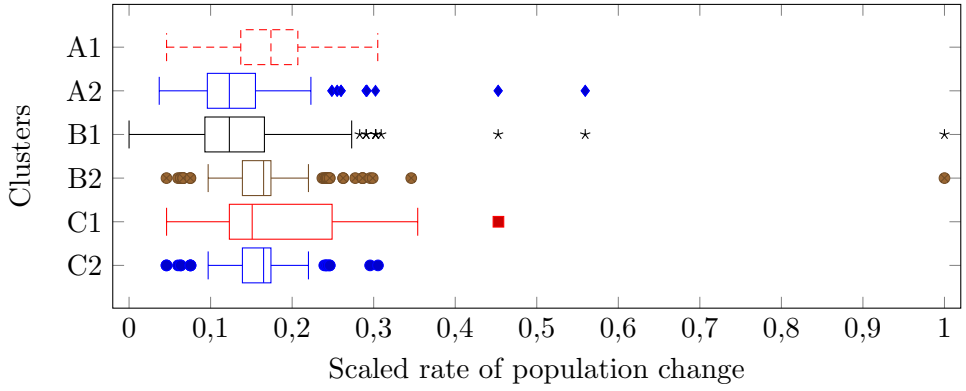


FIGURE 4.8: *The distribution of the scaled rate of population change variable for each cluster.*

In Figure 4.9, the distribution of the scaled population age variable can be seen. Contrary to the scaled rate of population change variable above, the values for the population age variable

are more dispersed and one cluster in each supermarket type has a much higher median value than the other. Clusters A1, B2 and C2 all have a median value of 0,7. Clusters A2, B1 and C1 all have lower median values of 0,4, 0,3 and 0,4, respectively.

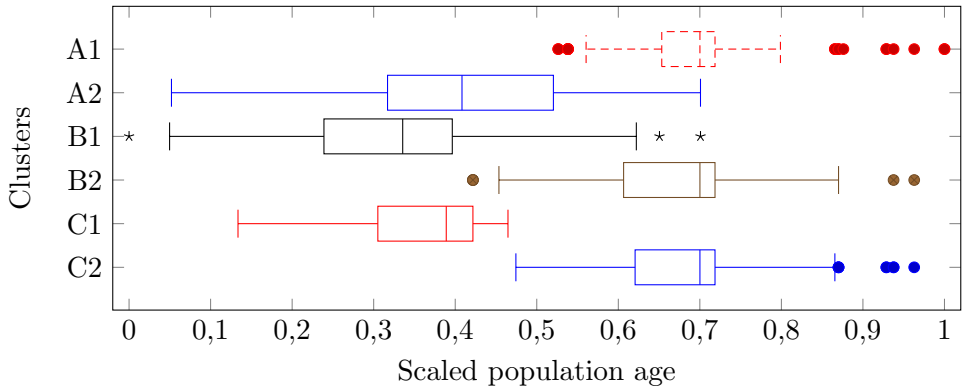
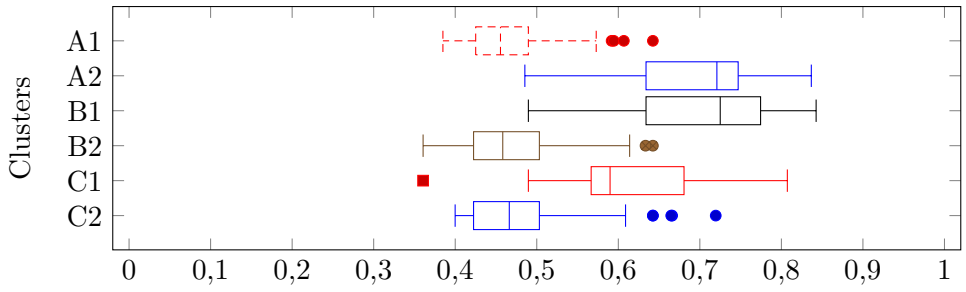
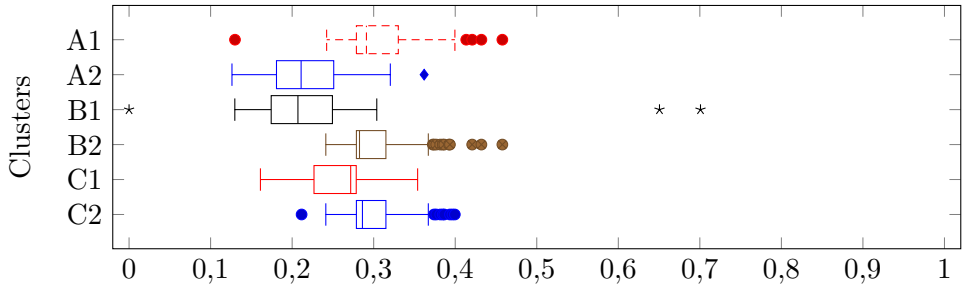


FIGURE 4.9: *The distribution of the scaled population age variable for each of the six clusters.*

Finally, the variables representing population income are shown in Figure 4.10. Across all clusters, the largest proportion of the population is in the lowest income bracket (Figure 4.10(a)) and this proportion gets smaller as the income increases. Although clusters A1, B2 and C2 have a significantly lower proportion than the other three clusters in the lowest income bracket, these clusters again have a higher proportion of the population in the other three income brackets.



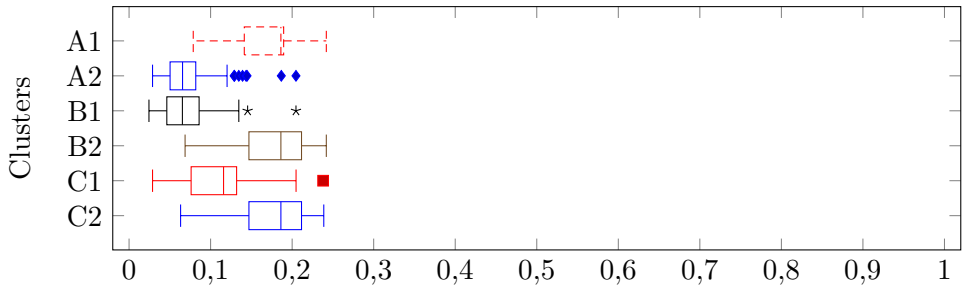
(a) *R0 to R38 400*



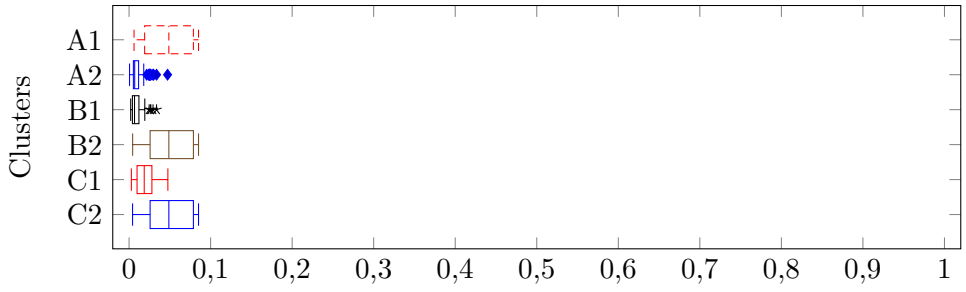
(b) *R38 400 to R153 600*

FIGURE 4.10: *The distribution of the population income variables for each of the six clusters within the different ranges of population income.*

To summarise, as the above graphs display, clusters A1, B2 and C2 have a higher percentage of households with landlines, a higher rate of population change, older population and higher income than clusters A2, B1 and C1. However, before continuing with these six clusters, the stability of this clustering method is tested by dividing the data into two randomly selected



(c) *R153 601 to R614 400*



(d) *R614 401 or more*

FIGURE 4.10: *(continued) The distribution of the population income variables for each of the six clusters within the different ranges of population income.*

samples and re-applying the clustering method. Similar distribution graphs are drawn for each of the clusters found in these two sample datasets and are given in Appendix 4.4. Studying these graphs, it is clear that there is no significant difference between the original cluster values and the values of clusters found with the two sample datasets. Thus, a total of six clusters are identified and a mobile device range for each of these six clusters will, therefore, be determined.

4.4 Cluster validation

In order to verify the store clusters obtained in §4.2, the dataset is randomly split into two sample sets. The clustering process is then applied to both sample datasets.

Landlines percentage

When comparing the distributions of the landlines percentage variable in Figures 4.11 and 4.12 for the two sample datasets, with Figure 4.7, it can be seen that there is no significant difference between the original distribution and that of the sample datasets. In sample dataset 1 (Figure 4.11), cluster A2 has a slightly higher median and upper quartile value than the original dataset, while the lower quartile value for cluster B2 is approximately 0,02 lower than the original dataset. In the full dataset, the values for clusters B2, C1 and C2 are more dispersed than that of sample 2 in Figure 4.12, but clusters A1, B2 and C2 still have higher median values than clusters A2, B1 and C1.

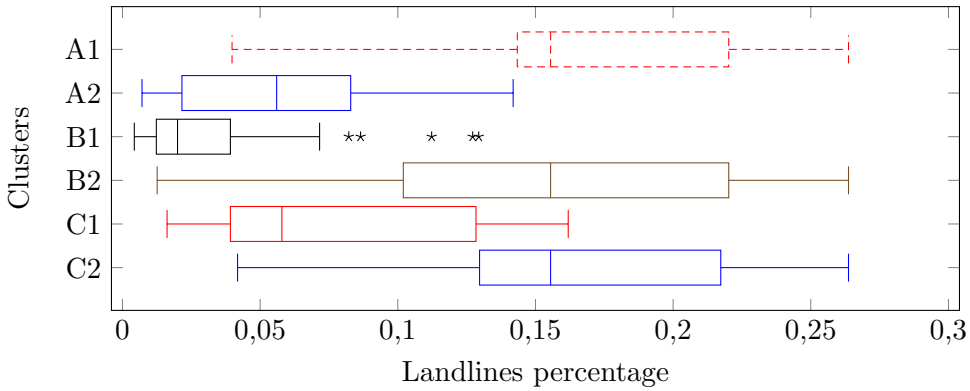


FIGURE 4.11: *The distribution of the landlines percentage variable per cluster for sample dataset 1.*

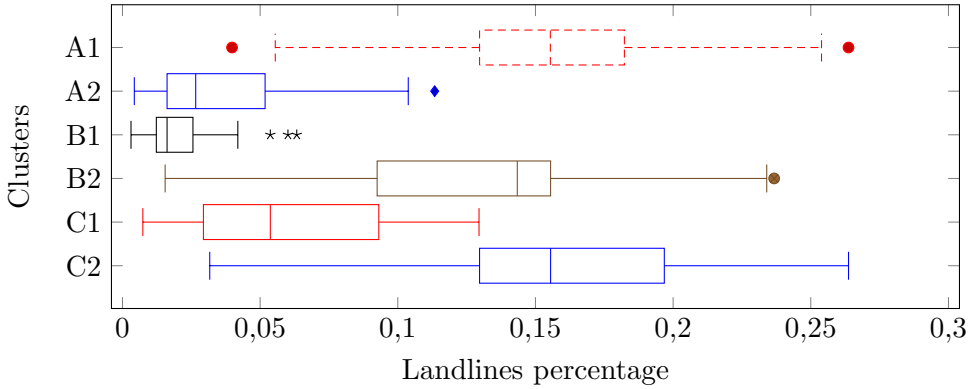


FIGURE 4.12: *The distribution of the landlines percentage variable per cluster for sample dataset 2.*

Scaled rate of population change variable

The distributions of the scaled rate of population change variable for the two sample datasets can be seen in Figures 4.13 and 4.14. When comparing these distributions to Figure 4.8, only a few small differences are noted. In sample dataset 1, clusters B1 and C1 have a smaller distribution than the full dataset seen in Figure 4.8. In sample dataset 2, clusters A1 and B1 have smaller distributions than seen in the original dataset, while cluster C1 has a larger distribution.

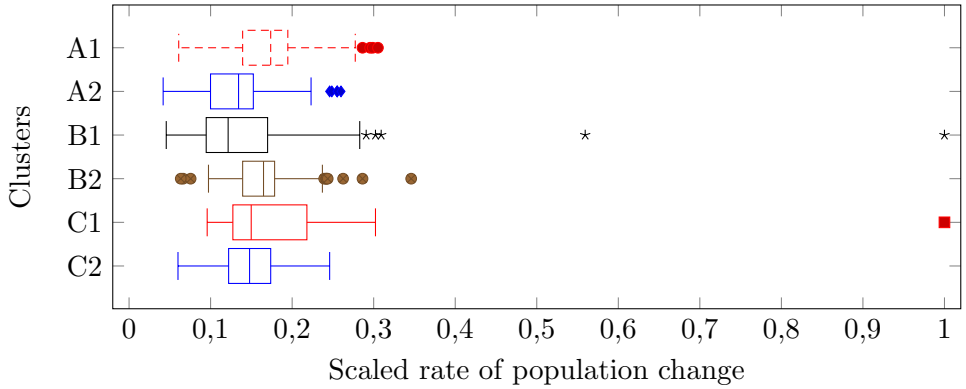


FIGURE 4.13: *The distribution of the scaled rate of population change variable per cluster for sample dataset 1.*

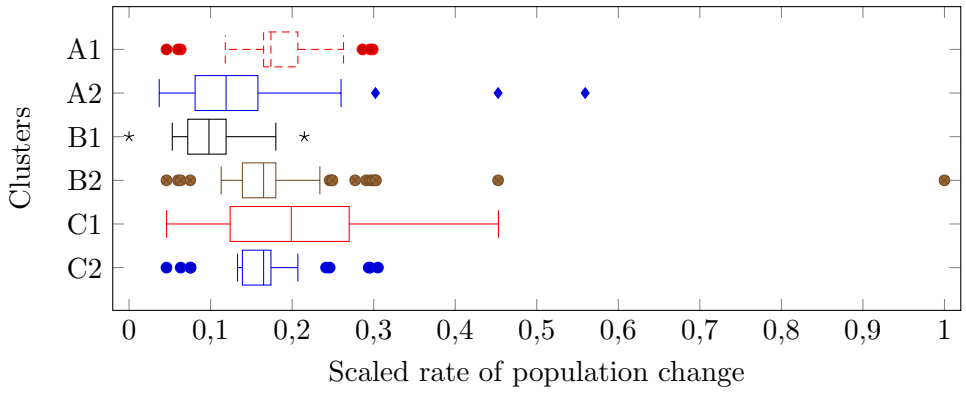


FIGURE 4.14: *The distribution of the scaled rate of population change variable per cluster for sample dataset 2.*

Scaled population age variable

The scaled population age variable distributions of the sample datasets in Figures 4.15 and 4.16 are compared to the distribution in Figure 4.9. It can be seen that, with sample dataset 1, clusters A1, B1 and C1 have a smaller distribution than the original dataset. The values for cluster A2 are more dispersed than seen in Figure 4.9. With sample dataset 2, the values for all clusters except cluster C1 have a wider distribution than that of the original dataset.

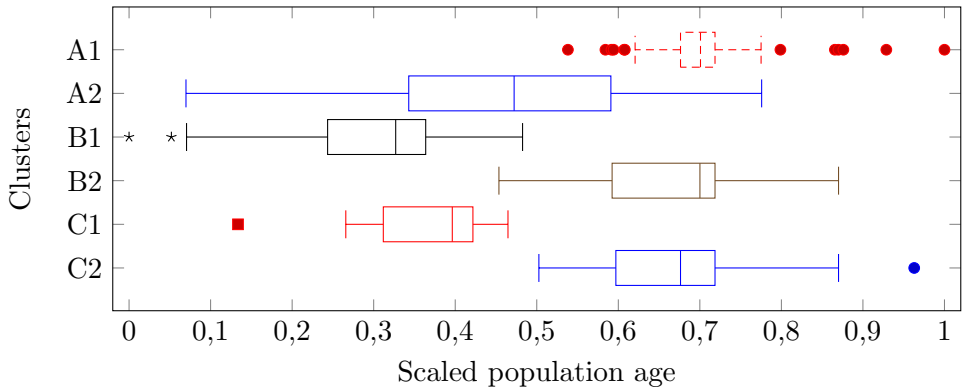


FIGURE 4.15: *The distribution of the scaled population age variable per cluster for sample dataset 1.*

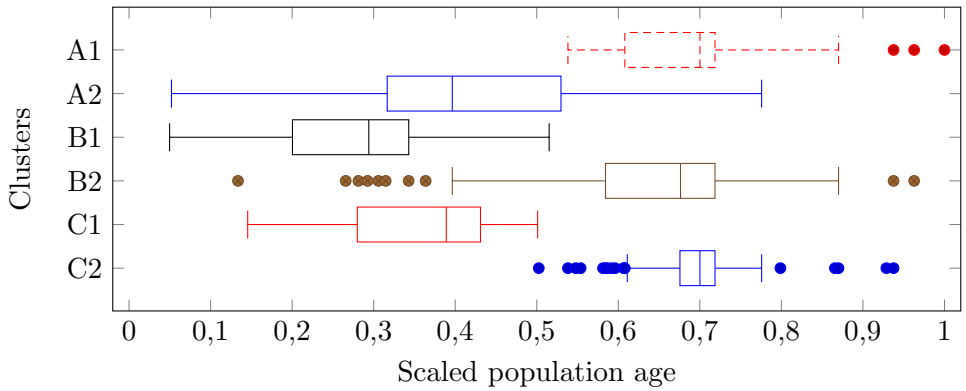
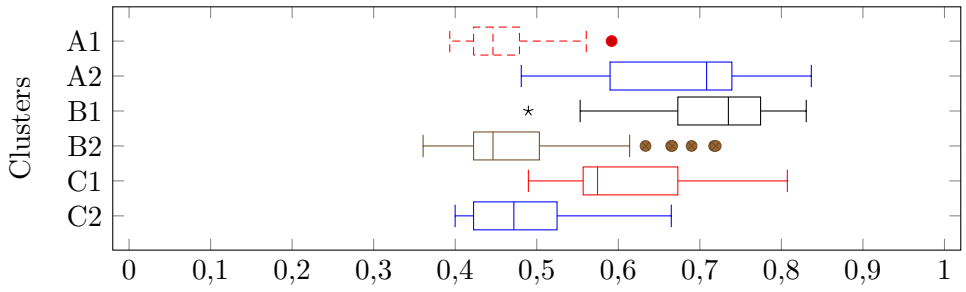


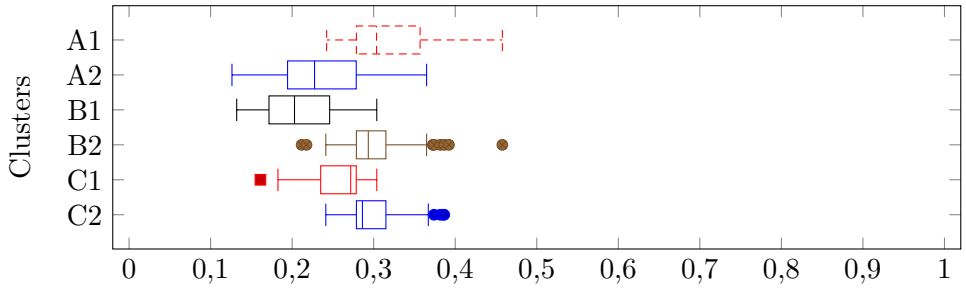
FIGURE 4.16: *The distribution of the scaled population age variable per cluster for sample dataset 2.*

Population income variables

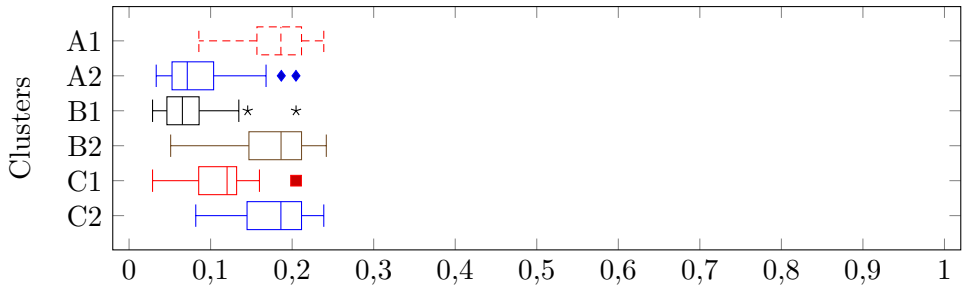
The population income variables' distributions for the two sample sets can be seen in Figures 4.17 and 4.18 and are compared to Figure 4.10. The distributions for all four income variables in the two sample datasets are very similar to the original distributions, with only a few small differences in the whisker and quartile values.



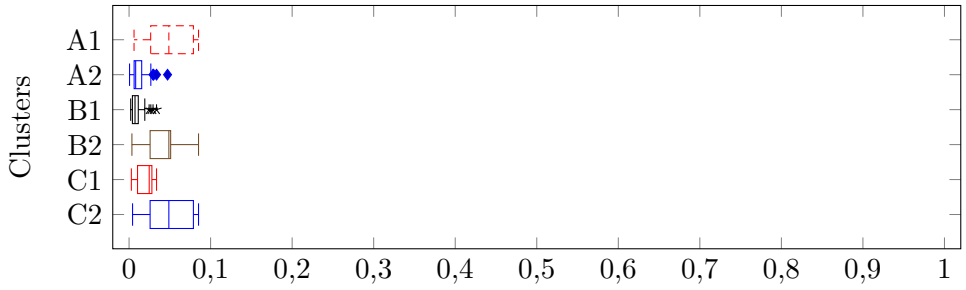
(a) *R0 to R38 400*



(b) *R38 400 to R153 600*

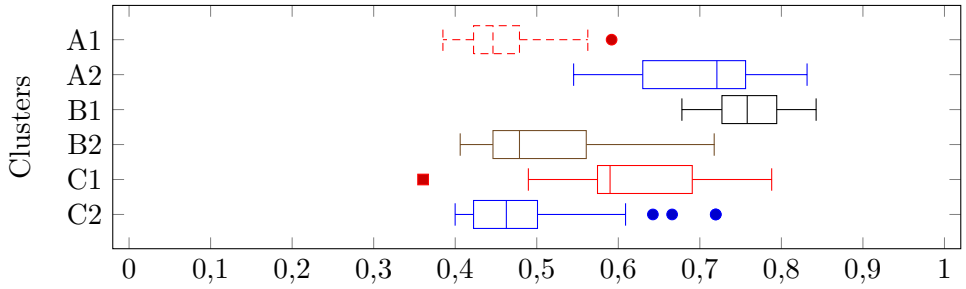


(c) *R153 601 to R614 400*

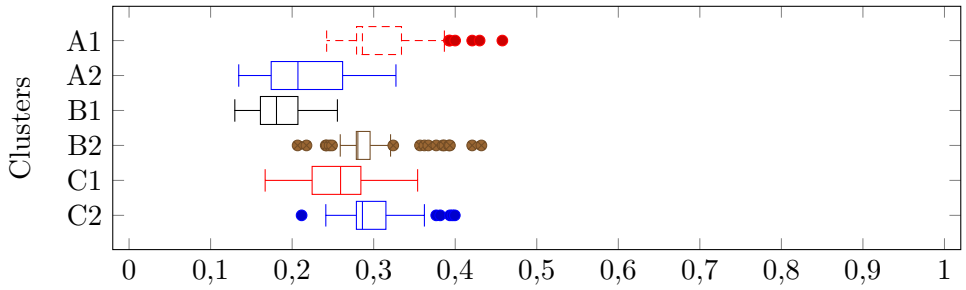


(d) *R614 401 or more*

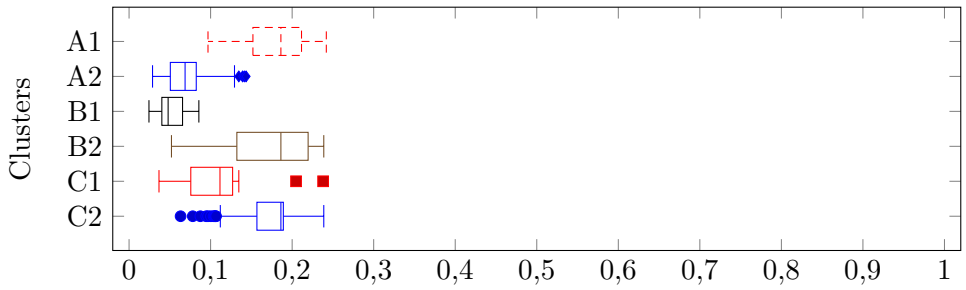
FIGURE 4.17: *The distribution of the population income variables per cluster for sample dataset 1.*



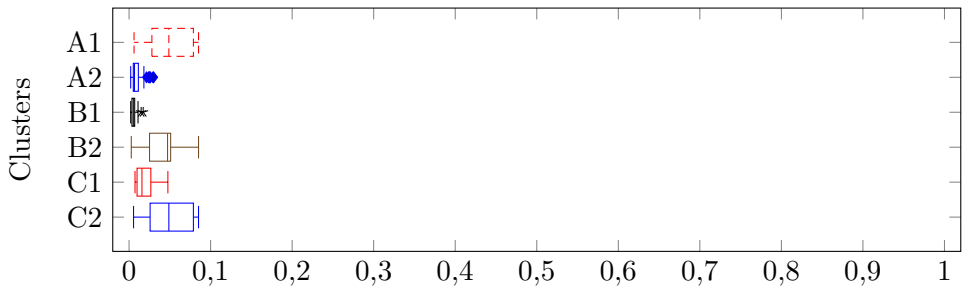
(a) *R0 to R38 400*



(b) *R38 400 to R153 600*



(c) *R153 601 to R614 400*



(d) *R614 401 or more*

FIGURE 4.18: *The distribution of the population income variables per cluster for sample dataset 2.*

CHAPTER 5

Mobile device offering

Contents

The topic of this chapter is the process followed to obtain the mobile device offering of The Retailer. This process is two part, starting with the methodology used to determine mobile device ranges as outlined in §5.1, followed by the method constructed to calculate how much stock to keep, as discussed in §5.2.

5.1 Determining mobile devices to keep in stock

The process to determine which mobile devices to keep in stock consists of multiple steps. First, three measures, calculated to evaluate the performance of each device in the different stores, are discussed in §5.1.1. Next, based on their overall performance in these three measures, the devices sold in at least 40 stores are ranked. Two iterative approaches are then followed to determine whether a device should be ranged in any of the six clusters, starting with the highest and lowest ranked devices, respectively. Devices that sold in less than 40 stores are analysed with the same approach as the one used for the lowest ranked devices. These approaches are discussed in §5.1.2 and §5.1.3, respectively.

For the purpose of this study, only the 10 top and bottom performing devices will be considered after the ranking to demonstrate the methodology and results. However, the same process can be repeated for every device sold by The Retailer.

5.1.1 Performance measures

Three measures are used to evaluate the performance of each device in a particular store. For each of the measures, the performance of each device is expressed relative to the performance of the category[1] in a store, in other words relative to the collective performance of all mobile devices in the particular store.

Measure 1: Rate of sale

In The Retailer's stores, all devices are not ranged for the same period of time and new devices are added within the observed period. Further, there are also days within the observed period where devices were out of stock. For these reasons, the *rate of sale* (ROS) is calculated using the total number of days that a device was in stock, rather than the total number of days in the observed period. Let u_{ij} be the number of units of device i sold in store j in the observed period and t_{ij} the number of days device i is in stock in store j. Further, let P denote the total number of devices and N the total number of stores. Then, the ROS of device i in store j can be written as

$$r_{ij} = \frac{u_{ij}}{t_{ij}} \qquad \begin{cases} i = 1, \ldots, P, \\ j = 1, \ldots, N, \end{cases} \tag{5.1}$$

while the ROS of the mobile device category (thus all devices) in store j is expressed as

$$R_j = \frac{\sum_{i=1}^{p} u_{ij}}{T_j}, \tag{5.2}$$

where T_j is the total number of days that any device was in stock in store j.

After the ROS for each store-device combination (r_{ij} in (5.1)), as well as for the whole mobile device category in a store (R_j above), are calculated, a ROS index for each store-device combination is then calculated as

$$I_{ij}^{R} = \frac{r_{ij}}{R_j} \times 100. \tag{5.3}$$

This is to compare the performance of device i to the overall mobile device category performance in store j. An $I_{ij}^{R} = 100$ indicates that device i has the same ROS as the category in the particular store j. If $I_{ij}^{R} > 100$, the devices has a faster ROS compared to the category in that store, while an $I_{ij}^{R} < 100$ indicates a slower ROS in comparison to the category as a whole.

Various different scenarios can lead to the same ROS or ROS index value for different devices. For example, as shown in Table 5.1, the device with code 00141 sold 30 units over 8 days in store B_077, while device 00321 sold 45 units over 12 days in store B_149. Although device 00321 sold 50% more units than device 00141, both have a ROS of 3,75.

Store	Product	Units sold	Days in stock	r_{ij}
B_077	00141	30	8	3,75
B_149	00321	45	12	3,75

TABLE 5.1: *Example of how the ROS of different devices can lead to the same value.*

[1] All mobile devices stocked at one of The Retailer's stores are in one category for the particular store.

Similarly, as indicated in Table 5.2, various r_{ij} and R_j combinations can lead to the same ROS index value. Although the ROS values differ in each case, the ratio between r_{ij} and R_j remains the same and therefore results in the same ROS index value.

Store	Product	r_{ij}	R_j	I_{ij}^R
B_157	00321	0,414	0,279	148
B_182	00321	0,710	0,481	148
C_004	00321	0,559	0,378	148

TABLE 5.2: *Examples of the different ways in which the calculation of the ROS index can result in the same value.*

Measure 2: Total units sold

As explained above, many different scenarios can take the same ROS value. To provide more context to the ROS measure, the total units sold is also used as a measure. Again, an index is calculated to compare the sales of a particular device to the overall sales of the category.

In this measure, the number of units of device i sold in store j, u_{ij}, is compared to the average number of mobile device units, $\bar{U}_j$, sold in store j, thus

$$\bar{U}_j = \frac{\sum_{i=1}^{P} u_{ij}}{p_j}, \tag{5.4}$$

where p_j is the number of devices sold in store j. Then, the total-units-sold index is calculated as

$$I_{ij}^U = \frac{u_{ij}}{\bar{U}_j} \times 100. \tag{5.5}$$

If $I_{ij}^U = 100$, the total units sold for device i in store j is the same as the average number of units sold over all devices in store j. A value for $I_{ij}^U > 100$ means an above average number of units of device i is sold in store j. Similarly, if $I_{ij}^U < 100$, the total number of units of device i sold in store j is below average for store j.

Measure 3: Average units in stock

Lastly, the average units in stock per day is also used as performance measure. As with *Measure 2*, the average number of units of a specific device in stock is compared to the average number of units of all mobile devices in stock in the relevant store. Let o_{ijk} be the number of units of device i in stock in store j on day k, where $k = 1, \ldots, t_{ij}$. Then, the average units of device i in stock in store j per day can be calculated as

$$\bar{o}_{ij} = \frac{\sum_{k=1}^{t_{ij}} o_{ijk}}{t_{ij}}. \tag{5.6}$$

The average number of units of any mobile device in stock in store j per day is calculated as

$$\bar{O}_j = \frac{\sum_{i=1}^{P} \sum_{k=1}^{t_{ij}} o_{ijk}}{T_j}. \tag{5.7}$$

The index of the average mobile device units in stock is then calculated as

$$I_{ij}^{O} = \frac{\bar{o}_{ij}}{\bar{O}_j} \times 100. \tag{5.8}$$

Similar to *Measures 1* and *2*, $I_{ij}^{O} = 100$ means that the average stock level of device i in store j is equal to the average stock level of all devices in store j. If $I_{ij}^{O} > 100$, device i generally has more units in stock than the average device in store j and if $I_{ij}^{O} < 100$, device i generally has fewer units in stock than the average device in store j.

All three measures for device i in store j, are then combined into one final index, I_{ij}^{C}, that will be used to rank all the mobile devices. The combined index, I_{ij}^{C}, is determined by simply taking the weighted average of the three indices described above, thus

$$I_{ij}^{C} = w^{R} I_{ij}^{R} + w^{U} I_{ij}^{U} + w^{O} I_{ij}^{O}, \tag{5.9}$$

where the weights w^{R}, w^{U} and w^{O} are initially all set equal to $\frac{1}{3}$. A sensitivity analysis is done on these measures by changing the weighting in the final index. The results of this sensitivity analysis will be discussed in §6.3.1.

For each mobile device i, its average combined index, $\bar{I}_i^{C}$, is used to rank the mobile devices from highest to lowest ranked devices. The value of $\bar{I}_i^{C}$ is calculated as

$$\bar{I}_i^{C} = \frac{\sum_{j=1}^{N} I_{ij}^{C}}{n_i}, \tag{5.10}$$

where n_i is the number of stores in which mobile device i was sold. The highest and lowest ranked devices are then identified by simply splitting the ranked devices in half.

5.1.2 Ranging of highest ranked devices

For the highest ranked devices based on the average combined index, $\bar{I}_i^{C}$, for each device i, their performance in each individual store is ranked to find the best and worst performing stores for each device. These stores are then used to determine whether or not a particular cluster should be ranging the specific mobile device. This process will be explained with the aid of mobile device 00141 as example.

Device 00141 is currently sold in 913 stores across all six clusters, and ranked 5^{th} using the methodology describe in §5.1.1. Let mobile device 00141 be denoted as device k. To determine whether device k should remain ranged in all six clusters, the 913 stores are ranked from best to worst performing using the three measures calculated in §5.1.1, thus using I_{kj}^{C} for each store j. The top and bottom 20% of the 913 stores that range device k, or in this case 183 stores, are selected and the remaining 60%, or 547, stores are removed. This is done to only consider the most extreme stores for each device, while remove the middle stores with a more average performance. The subset of top and bottom stores is used to determine the ranging status of the device in each cluster and this ranging status then applies to all stores within the relevant cluster, including those with an average performance.

The top-and-bottom subset of stores are then summarised into their relevant clusters, to arrive at Table 5.3. In cluster A1, 32 stores performed in the top 20% for device k, while 33 stores performed in the bottom 20% of all stores for this device. Therefore, a total of 65 stores in cluster A1 are being considered in this analysis of device 00141. Cluster B2, which is the largest of the six clusters, also contains the most stores in the subset of stores ranging device 00141. If more

Cluster	Number of stores in the top 20%	Number of stores in the bottom 20%	Total number of stores in subset	Ranging status
A1	32	33	65	Maybe
A2	34	20	54	Yes
B1	45	25	70	Yes
B2	55	73	128	Maybe
C1	4	7	11	No
C2	13	25	38	No
Total	183	183	366	

TABLE 5.3: *The best and worst performing stores for device 00141, summarised into their relevant clusters.*

than 60% of the subset stores in a specific cluster are in the top 20%, the cluster is assigned a ranging status of 'Yes' and device 00141 is ranged in all stores in the particular cluster, including those previously removed from the top-and-bottom subset of stores. For example, 34 of the 54 stores in cluster A2 are in the top 20% of all stores for device k. Thus, 63% of all stores in cluster A2 are in the top 20% stores and are assigned a ranging status of 'Yes'. If less than 40% of the subset stores in a cluster are in the top 20%, the cluster receives a ranging status of 'No' and the device is not ranged in any store in the specific cluster. Finally, if between 40% and 60% of the subset stores are in the top 20%, the cluster is assigned a ranging status of 'Maybe' and further investigation is required to determine the performance in the cluster.

The rules applied to the 'Maybe' clusters can be changed by The Retailer to increase or decrease the total number of mobile devices ranged. However, for the purpose of this study, the following rule is applied. If device k is currently ranged in at least half of the stores in the cluster and device k's performance, I^C_{kj}, is above the average performance of all devices in at least half of the ranged stores, the device is ranged in all stores in the cluster, effectively changing the ranging status of the cluster from 'Maybe' to 'Yes'. If this condition is not satisfied, the process of ranging the lowest ranked devices, discussed in the next section, applies to the cluster.

The ratio of 60%:40% to determine the ranging status of a cluster is chosen to ensure that there are more well performing stores in a cluster for a device to be ranged. Column 5 in Table 5.3 indicates the ranging status of each cluster for device 00141. This device should be ranged in all stores in clusters A2 and B1. Its performance in clusters A1 and B2 should be investigated, and device 00141 should not be ranged in any type C store.

5.1.3 Ranging of lowest ranked devices

When a device is currently ranged in less than 40 stores in total or the device is one of the lowest ranked devices, it's performance in each store is considered individually.

Since they are not sold in many stores, or perform worse than other devices, these devices are not ranged in all stores in a cluster as in the case of the highest performing devices discussed in §5.1.2. These devices will only be ranged in a store if it is already ranged and if their performance is above the average of all devices in that store, as determined by the performance measures described in §5.1.1. Again, the process to determine whether a device should be ranged in a store is discussed with the aid of an example.

Consider device 00263, which is currently ranged in 17 stores and, based on its average combined index, $\bar{I}^C_\ell = 27$, is ranked 76^{th} overall. Let device 00263 be denoted as device ℓ. As shown in

Table 5.4, device ℓ is only outperforming the rest of the devices in one store, with a combined index of 122 in store B_481. In the other 16 stores, device ℓ has a combined index of 52 or less, meaning that it is performing well below average and is not contributing significantly to the sales of these stores. As a result, this device will only be ranged in store B_481.

Store	Cluster	$I^R_{\ell j}$	$I^U_{\ell j}$	$I^O_{\ell j}$	$I^C_{\ell j}$
B_481	B2	25	129	213	122
C_079	C2	3	113	41	52
C_129	C2	10	68	61	46
C_091	C1	3	90	29	41
C_147	C1	8	46	47	34
B_480	B2	2	51	46	33
C_220	C1	7	34	26	22
C_227	C1	6	32	15	18
C_213	C1	4	28	19	17
C_218	C1	4	26	14	15
C_089	C1	1	13	24	13
C_033	C2	3	16	14	11
B_479	B2	1	17	15	11
C_219	C1	3	13	13	10
C_053	C1	2	15	11	9
B_482	B2	1	5	8	5
B_346	B2	1	5	5	4

TABLE 5.4: *The performance of device 00263 (denoted as device ℓ) per store.*

5.2 Calculating how much stock to keep

In §5.1.2, the process of ranging the highest ranked devices was discussed, where a device was assigned a ranging status in each of the six clusters. When a device receives a ranging status of 'Yes' in a particular cluster, it is ranged in all stores in that cluster, whether or not the device is currently ranged in all the relevant stores. Thus, in stores where a particular device is not currently sold, the necessary stock levels for the device must be estimated.

To determine how much stock to keep, the ROS is used. In this project, regression analysis is used to estimate the ROS of a device in stores where the device was not previously sold, by training the regression model on stores in which the device is currently sold. In §5.2.1, the variables used in the regression analysis are discussed, followed by a discussion of the regression analysis used in this project in §5.2.2.

5.2.1 Data preparation

In this regression analysis, the ROS r_{ij} of device i in store j as described in §5.1.1, is used as the dependent variable, while the clustering variables, as listed in §4.1, and store attributes described in §3.2, along with some newly defined sales related variables, are considered as independent variables. These variables are listed in Table 5.5.

The first store attribute used, is store size and this is considered as there is generally a correlation between the size of The Retailer's stores and its sales. Stores located in areas with a higher population have a larger store size to provide for the high number of customers. The province in which a store is located, is also considered as a store attribute since The Retailer is already

Independent variables
Scaled average age
Scaled average % rate of change
Population income: R0 - R38 400
Population income: R38 401 - R153 600
Population income: R153 601 - R614 400
Percentage owning a landline
Store size (m^2)
Province
Adapted mobile device category ROS
Average sales amount per store
Total number of mobile devices sold

TABLE 5.5: *The independent variables used to predict the ROS, r_{ij}, of mobile devices.*

using this variable when determining product ranges. Both of these variables can be seen in Table 3.3.

Three new variables are defined and are related to the sales of the mobile device category in a particular store. The first of these variables is similar to the mobile device category ROS, R_j for store j, described in §5.1.1, but the sales of the device for which ROS is being estimated, is excluded from the calculation. Table 5.6 contains an example of the values of this variable for five random devices currently sold in store B_001. Column 3 of Table 5.6 shows the ROS for five devices in store B_001, while column 4 contains the mobile device category ROS in store B_001 as explained in §5.1.1. Column 5 contains the new adapted mobile device category ROS in store B_001, excluding the sales for the device in column 1.

Product	Store	r_{ij}	R_j	Average ROS of other devices
00132	B_001	0,05	0,30	0,25
00141	B_001	0,01	0,30	0,28
00144	B_001	0,04	0,30	0,29
00258	B_001	0,06	0,30	0,24
00259	B_001	0,03	0,30	0,27

TABLE 5.6: *An example of the adapted mobile device category ROS, in which the sales of the device, listed in column 1, for which ROS is being estimated, are excluded.*

The adapted ROS will be the same as R_j when estimating the ROS in a store where the particular device is not currently sold, as r_{ij} is 0 in this case.

The next variable is the average sales amount per store, again excluding the sales of the device for which ROS is being estimated. In Table 5.7, different values of the average sales amount of devices currently sold in store B_001, can be seen for illustrative purposes. Let a_{ijk} be the total sales amount of device i in store j on day k (as seen in column 6 in Table 3.1) and let τ_{ij} be the total number of days on which device i was sold in store j. The average sales amount of the mobile device category in store j, shown in column 3 of Table 5.7, can be calculated as

$$\bar{A}_j = \frac{\sum_{i=1}^{p_j} \sum_{k=1}^{\tau_{ij}} a_{ijk}}{\sum_{i=1}^{p_j} u_{ij}}. \tag{5.11}$$

Finally, the average sales amount per store, excluding the device for which the ROS is being estimated, can be seen in column 4 of Table 5.7.

Product	Store	Average sales amount for Mobile device category	Adapted mobile device category
00132	B_001	R483,64	R524,32
00141	B_001	R483,64	R499,52
00144	B_001	R483,64	R489,07
00258	B_001	R483,64	R424,44
00259	B_001	R483,64	R489,00

TABLE 5.7: *An example of the calculations of the average sales amount for the mobile device category, including and excluding, respectively, the device in column 1.*

When estimating the ROS for a particular device in a store in which it is not currently sold, one can see that $\bar{A}_j$ is the same as the adapted average sales amount per store, excluding the device for which ROS is being estimated.

The last variable is the total number of devices sold in a particular store j, not counting the device for which ROS is being estimated. In other words, $p_j - 1$ is used when training the regression, and p_j is used when applying the regression model to stores in which the relevant device is not currently sold.

The regression analysis is done via a decision tree and since decision trees do not use a distance metric, it is not necessary to scale continuous variables. However, since the province is a categorical variable, binary dummy variables are introduced and used as independent variables instead.

5.2.2 Regression analysis

A decision tree is built for each device for which ROS must be estimated. The method of using decision trees for regression analysis is discussed in Appendix D. This model is trained and tested using stores that already sell the device in order to estimate the ROS for the device for stores that do not have a sales history for the particular device. Continuing with device 00141 as before, the regression analysis will be explained in this section, using device 00141 as example.

As found in §5.1.2, device 00141 has a ranging status of 'Yes' in Table 5.3 for clusters A2 and B1. Therefore, the ROS of device 00141 will be estimated for all stores that do not currently sell this device in these two clusters. In clusters A2 and B1, 198 and 174 stores can be found, respectively. Of these stores, 170 stores in cluster A2 and 160 stores in cluster B1 are currently selling device 00141. These 330 stores, along with 583 stores in other clusters, will be used to build the decision tree in order to estimate the ROS of device 00141 for the remaining stores in clusters A2 and B1.

The 913 stores in total are split into training and test sets using the `sklearn.model_selection.train_test_split` package in Python, so that the test set consists of 20% of the stores (183 stores in total). Then, the `sklearn.tree.DecisionTreeRegressor` package is used to build various decision trees with different values for the `max_depth`, `min_samples_split` and `min_sample_leaf` parameters, as described in §D.2. The decision tree with the largest coefficient of determination (R^2) for the test set is used as the final decision tree.

In the final decision tree chosen for device 00141, parameter `max_depth` has a value of 4 and parameter `min_sample_leaf` has a value of 12. Any value of 24 or less for the `min_samples_split` parameter is irrelevant, as the `min_sample_leaf` of 12 implies that there must be at least 24 data points before each split can be made. Therefore, no value is specified and the default value of 2 is used for `min_samples_split` instead. The resulting R^2 values for this tree are 0,64 for the training set and 0,65 for the test set. The final tree can be seen in Figure 5.1 and shows that there are 13 terminal nodes used to make the ROS prediction. Furthermore, it can be seen that only five variables are used in the tree, with 'ROS of other devices' used six times and 'Sales area' and 'Average age' used twice. Following the branches on the left of the tree in Figure 5.1

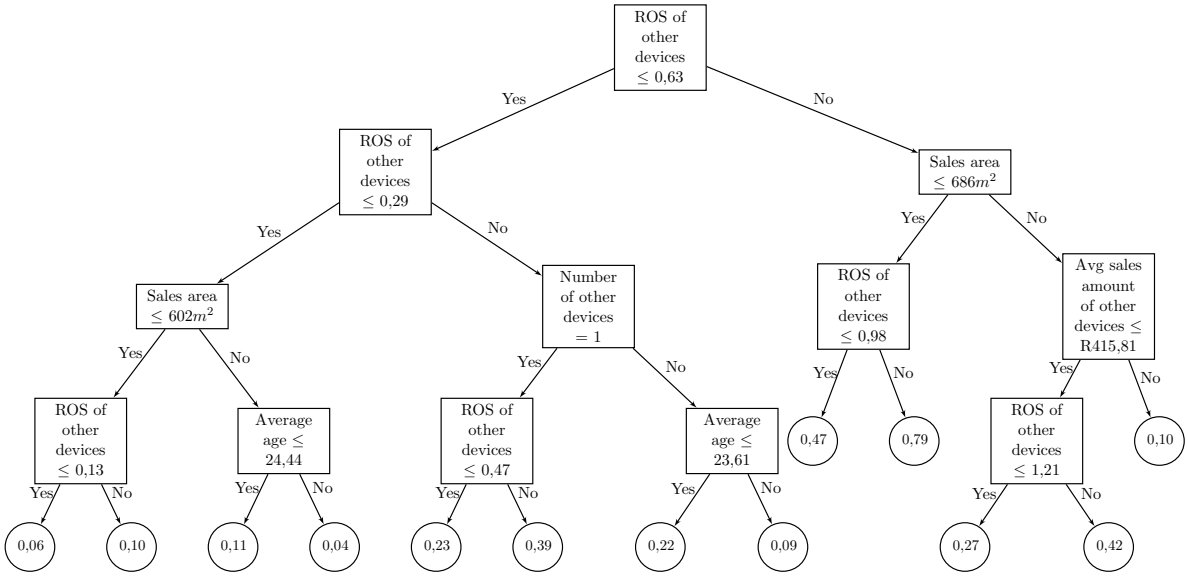


FIGURE 5.1: *The decision tree for device 00141.*

shows that a store with an adapted ROS less or equal to 0,13 and a sales area less or equal to $602m^2$ has a predicted ROS of 0,06.

This decision tree can now be used to predict the ROS for the remaining 42 stores that do not currently sell device 00141. Consider store A_037 in cluster A2. This store, say k, does not currently sell device 00141 and the ROS will, therefore, be predicted using the decision tree in Figure 5.1. This store has a sales area of $415m^2$, R_k of 0,43 and is currently selling 6 different devices. Furthermore, this store has an average age of 26,18 years. Following the relevant branches in the decision tree, it can be seen that the predicted ROS for device 00141 in store A_037 is 0,09. As The Retailer orders new stock weekly, a ROS of 0,09 units per day is equal to 0,63 units per week. Rounding up, this results in ordering one unit of device 00141 per week.

CHAPTER 6

Results

Contents

In this chapter, the final results of the device ranging and regression analysis, described in Chapter 5, will be discussed. Emphasis will be placed on the top and bottom 10 devices to illustrate the results of the described methodology, although it can be easily applied to all devices. In §6.1, the results of the mobile device ranging will be discussed, followed by the results of the regression trees in §6.2. The chapter concludes, in §6.3, with some scenario testing and the effect of these scenarios on the mobile device range obtained in §6.1.

6.1 Mobile devices to keep in stock

As described in §5.1.2, devices are ranged in all stores belonging to a cluster with a ranging status of 'Yes'. The results of these devices and clusters are presented in §6.1.1. If a cluster receives a ranging status of 'Maybe', it was suggested that the performance of stores in the cluster are further investigated. The results of the top 10 devices and associated clusters with a 'Maybe' ranging status are discussed in §6.1.2.

6.1.1 Top 10 devices and their corresponding clusters with a 'Yes' status

Table 6.1 contains the mobile devices from the overall top 10 performing devices, that have at least one cluster with a ranging status of 'Yes'. It can be seen that only six of the top 10

devices have at least one cluster with a ranging status of 'Yes', where these clusters are listed in column 3 of Table 6.1. Devices 00132, 00321 and 00259 are ranged in three clusters each, devices 00141 and 00297 are ranged in two clusters each, while device 00298 is ranged in only one cluster.

Further, all stores in cluster A2 will range devices 00141 and 00297. Stores in cluster B1 will all range devices 00132 and 00141, while all stores in cluster B2 will range devices 00321, 00297 and 00259. Finally, all stores in store types C1 and C2 will range devices 00132, 00321 and 00259, with stores in cluster C2 ranging device 00298 as well.

Rank	Product	Cluster	Number of stores currently selling device	Number of stores for which ROS must be predicted	Total number of stores in the cluster
4	00132	B1	159	6	165
4	00132	C1	36	7	43
4	00132	C2	171	24	195
5	00141	A2	170	11	181
5	00141	B1	160	5	165
5	00321	B2	70	234	304
7	00321	C1	5	38	43
7	00321	C2	18	177	195
8	00297	A2	175	6	181
8	00297	B2	252	52	304
9	00298	C2	39	156	195
10	00259	B2	209	95	304
10	00259	C1	23	20	43
10	00259	C2	146	49	195

TABLE 6.1: *Mobile devices from the top 10 performing devices and the corresponding clusters containing these devices that received a ranging status of 'Yes'.*

A breakdown of the total number of stores in each cluster (column 6) is also provided in Table 6.1, giving the number of stores already selling a particular device in column 4, as well as the number of stores for which the ROS must be predicted in column 5. It can be seen that four of the six devices are already sold in more than half of the stores in the specific cluster, with a small number of stores for which to predict a ROS. For example, device 00132 is currently sold in 83% of all the stores in the three clusters (B1, C1 and C2), leaving only 17% of stores in these clusters with no ROS.

With devices 00321 and 00298, only 15% and 18% of the stores in the corresponding clusters are currently selling these devices. However, these devices performed so well in these stores that the clusters received a ranging status of 'Yes' for these devices.

6.1.2 Top 10 devices and their clusters with a ranging status of 'Maybe'

As described in §5.1.2, a cluster's 'Maybe' status can be changed to 'Yes' if the performance condition is met. In Table 6.2, a few of the top performing devices and their corresponding clusters with a ranging status of 'Maybe' can be seen. The number of stores in the cluster in which each device is currently sold, along with the number of these stores performing above-average, are displayed in columns 3 and 4, respectively. Further, this table also contains the total number of stores in the cluster, as well as the average combined index for all stores with an above-average performance.

Product	Cluster	Number of stores currently selling this device	Number of stores with above-average performance	Total number of stores in the cluster	Average combined index of the stores with an above-average performance
00222	B2	8	4	304	268
00222	C2	41	21	195	152
00287	B1	39	16	165	164
00287	B2	172	99	304	145
00131	B2	37	21	304	160
00132	A2	176	67	181	142
00141	A1	153	68	169	143
00141	B2	265	92	304	157
00291	B1	92	48	165	147
00291	B2	65	26	304	146
00321	A1	66	31	169	145
00321	B1	86	40	165	133
00297	A1	150	57	169	138
00297	B1	145	57	165	147
00259	B1	116	38	165	150

TABLE 6.2: *Mobile devices from the top 10 performing devices and the corresponding clusters containing these devices, that received a ranging status of 'Maybe'.*

Based on the performance condition described in §5.1.2, it is clear that device 00291 meets this criterion in cluster B1, currently sold in 52% of the stores in the cluster with an above-average performance. Thus, the ranging status for device 00291 in cluster B1 is changed from 'Maybe' to 'Yes' and the results of the ROS estimation for the remaining 73 stores in cluster B1 will be given with the devices and clusters discussed in §6.1.1.

The ranging status of the other devices and clusters in Table 6.2 remain unchanged. Therefore, the methodology described in §5.1.3 applies to these devices and they are only ranged in the stores references in column 4 of Table 6.2. Column 6 in this table contains the average combined index of each device in these stores, indicating that all of the devices are performing well above the average of all devices in these stores.

6.1.3 Bottom 10 devices

Table 6.3 contains an extract from the results of the bottom 10 mobile devices currently ranged by The Retailer. Only clusters with at least one store with an above-average performance are shown in this table, resulting in only four of the bottom 10 devices included in the table. The total number of stores currently selling a particular device in each of the clusters, the total number of stores in which the device has an above-average performance, as well as the total number of stores in the cluster. Column 6 in Table 6.3 contains the average combined index of the stores in column 4, as these are the only stores that will include these devices in their range.

As discussed in §5.1.3, devices currently sold in less than 40 stores are ranged with the same methodology and the bottom 10 of these devices are shown in Table 6.4. The number of stores in which these devices are currently sold, ranges from 1 to 5 in a single cluster. As seen in Table 6.4, these mobile devices have an average combined index ranging from only 3 to 22 in the different clusters. This indicates that these devices are performing well below-average in all

Product	Cluster	Number of stores currently selling the device	Number of stores with above-average performance	Total number of stores in the cluster	Average combined index of the above-average performance stores
00248	B2	46	1	304	106
00248	C2	13	2	195	135
00264	B1	137	1	165	114
00264	B2	189	2	304	115
00264	C1	21	1	43	113
00264	C2	128	4	195	142
00305	A1	99	2	169	123
00305	A2	119	2	181	115
00289	B1	81	6	165	130
00289	B2	71	3	304	122
00289	C2	7	1	195	164

TABLE 6.3: *Four of the bottom 10 performing mobile devices with an average combined index greater than 100, and the clusters in which these devices are sold.*

of these stores and, thus, are not included in the new range for any stores.

Product	Cluster	Total number of stores currently selling device	in the cluster	Average combined index
00267	B1	1	165	14
00267	B2	2	304	3
00161	C2	1	195	7
00098	C2	1	195	8
00221	C2	1	195	9
00273	B2	1	304	10
00208	C1	1	43	12
00208	C2	2	195	10
00210	C2	1	195	11
00241	B1	1	165	22
00241	B2	3	304	8
00251	C2	3	195	12
00274	B2	1	304	8
00274	C1	1	43	12
00274	C2	5	195	13

TABLE 6.4: *The bottom 10 devices that sold in less than 40 stores and the clusters in which these devices are sold.*

After considering the ranging status and the performance of the devices above, the final range per store cluster can be found. Table 6.5 contains the six clusters and the mobile devices ranged in each one. Column three in this table indicates whether the device is ranged in all stores in the cluster or only in selected stores. Finally, column four provides the number of stores ranging the particular mobile device in each cluster, if the device is not ranged in all stores in the cluster.

Considering only the devices listed in Table 6.5, one device is ranged in two stores in cluster A1, while seven devices are ranged in the stores in cluster C2. Four of these mobile devices are ranged in all the stores in the cluster, one device is ranged in four stores in cluster C2, while

Cluster	Product	Range in all stores?	Number of stores ranging device
A1	00305	No	2
A2	00141	Yes	-
A2	00297	Yes	-
A2	00305	No	2
B1	00132	Yes	-
B1	00141	Yes	-
B1	00291	Yes	-
B1	00264	No	1
B1	00289	No	6
B2	00321	Yes	-
B2	00297	Yes	-
B2	00259	Yes	-
B2	00248	No	1
B2	00264	No	2
B2	00289	No	3
C1	00132	Yes	-
C1	00321	Yes	-
C1	00259	Yes	-
C1	00264	No	1
C2	00132	Yes	-
C2	00321	Yes	-
C2	00298	Yes	-
C2	00259	Yes	-
C2	00248	No	1
C2	00264	No	4
C2	00289	No	1

TABLE 6.5: *The final devices, from the selection for illustrative purposes, ranged in each cluster.*

two devices are ranged in only one store each.

In total, seven mobile devices are ranged in all the stores of at least one cluster. Thus, the ROS should be predicted for these devices for the stores not currently selling the device.

6.2 How much stock to keep

The ranging of the top 10 devices, discussed in §6.1.1 and §6.1.2, shows that the ROS should be predicted for seven devices, namely 00132, 00141, 00321, 00297, 00298, 00259 and 00291 (see Table 6.5).

Table 6.6 contains the total number of stores in which each of the devices is currently ranged, the number of stores used to train and test the regression tree, as well as the number of stores for which ROS must be predicted. As seen in this table, device 00132 is currently sold in the most stores and, therefore, has the largest training dataset of all seven devices. Device 00298 is sold in the fewest number of stores and has only 42 stores in the training dataset. From Table 6.6, it can also be seen that the number of stores for which ROS must be predicted is higher than the number of stores used to train the regression trees for devices 00321 and 00298.

The decision trees built for the seven mobile devices listed above, and the predicted ROS for each store for these mobile devices will be discussed in §6.2.1 and §6.2.2, respectively. Emphasis will be placed on devices 00132 and 00298.

Mobile device product	Number of stores currently selling device	Number of stores used for training	Number of stores used for testing	Number of stores for which ROS must be predicted
00132	943	754	189	37
00141	913	731	182	16
00321	351	281	70	449
00297	809	648	161	58
00298	52	42	10	156
00259	673	539	134	164
00291	157	126	31	73

TABLE 6.6: *The devices for which regression trees are built, with 80% of the stores currently selling the device used for training, and 20% used for testing.*

6.2.1 Regression trees constructed

The regression trees built for device 00132 and 00298 are given in Figures 6.1 and 6.2, respectively, as examples, while the regression trees for the remaining mobile devices are provided in §E.1. The decision tree for device 00141 was also provided in Figure 5.1.

The regression tree for device 00132 in Figure 6.1 consists of 13 decision nodes and 14 terminal nodes. Further, five different variables are used to construct the tree.

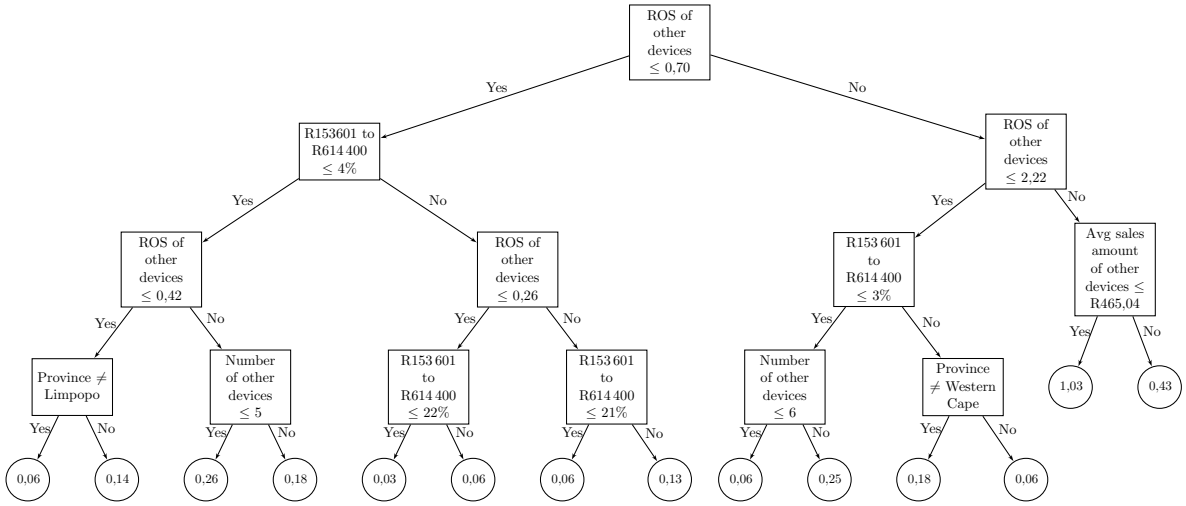


FIGURE 6.1: *The regression tree for device 00132.*

Figure 6.2 contains the regression tree for device 00298. This tree is made up of seven decision nodes and eight terminal nodes, with a total of five different variables.

A summary of the regression tree parameters and results of all seven mobile devices can be seen in Table 6.7. These trees are built using the information provided in Table 6.6. Five of the seven devices have a tree depth of 4, while devices 00297 and 00259 have a depth of 5. Furthermore, the default value for the `min_sample_split` parameter is used for all devices, as a change in this parameter does not lead to an improvement in the R^2 values. The `min_sample_leaf` parameter is set to values between three and 30, with six of the devices having a value below 14.

Although most of the trees are built with similar parameters, the variance in the number of

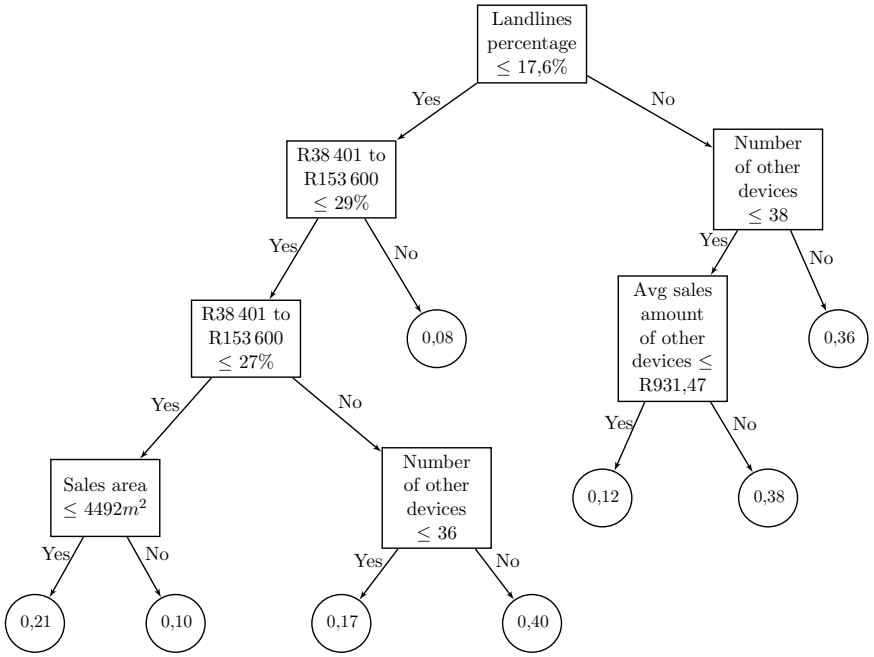


FIGURE 6.2: *The regression tree for device 00298.*

Product	`max_depth`	`min_sample_split`	`min_sample_leaf`	Terminal nodes	Train R^2	Test R^2
00132	4	2	6	14	0,66	0,65
00141	4	2	12	13	0,64	0,65
00321	4	2	11	9	0,54	0,56
00297	5	2	30	12	0,63	0,62
00298	4	2	5	8	0,42	0,45
00259	5	2	3	11	0,57	0,59
00291	4	2	13	7	0,52	0,52

TABLE 6.7: *The parameter values used for the various regression trees.*

stores provided in Table 6.6 leads to the seven devices having regression trees of different sizes. Consider devices 00132 and 00298 with a `min_sample_leaf` value of six and five, respectively. As seen in Table 6.6, device 00132 is currently sold in the most stores, while device 00298 is currently sold in the least stores. This difference of 891 stores leads to device 00132 having the largest of the seven trees with 14 terminal nodes. On the other hand, the tree for device 00298 has only eight terminal nodes.

The R^2 values of the training and test datasets of each device can also be seen in Table 6.7. Only three of the seven devices have a test R^2 higher than 0,6, three devices have a test R^2 above 0,5 and device 00298 has a test R^2 of only 0,45. However, the R^2 values for the training and test datasets are similar, indicating that the regression trees are not overfitting the training data. Further, comparing the size of the training dataset and the performance of the resulting regression tree in Figure 6.3 shows that there is a positive correlation.

Another element of these regression trees that should be discussed, is the variables used in the decision nodes. All of the variables are used at least once in the seven regression trees built and Figure 6.4 contains these variables, along with the number of times each variable is used in a decision node or in a tree. 'ROS of other devices' is the most commonly used variable,

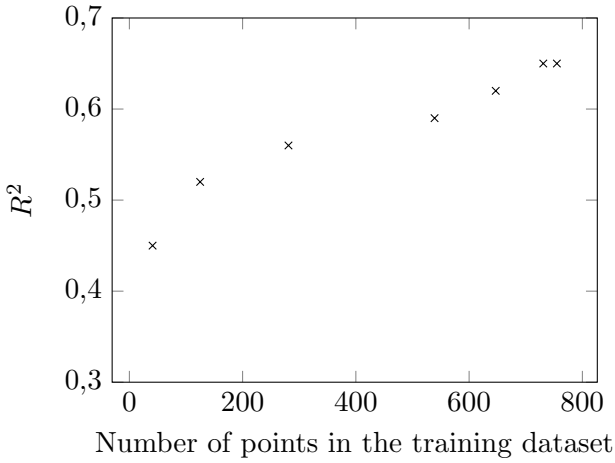


FIGURE 6.3: *The performance of the different regression trees compared to the size of the training dataset.*

used 24 times in six of the seven regression trees. This variable is also used in the first decision node of four regression trees, making it the most important variable when predicting the ROS of a device. This is followed by 'Sales area', used 11 times in six regression trees. The 'Rate of change' variable is only used once in the regression tree for device 00297 and is, therefore, the least important variable when predicting the ROS of a device.

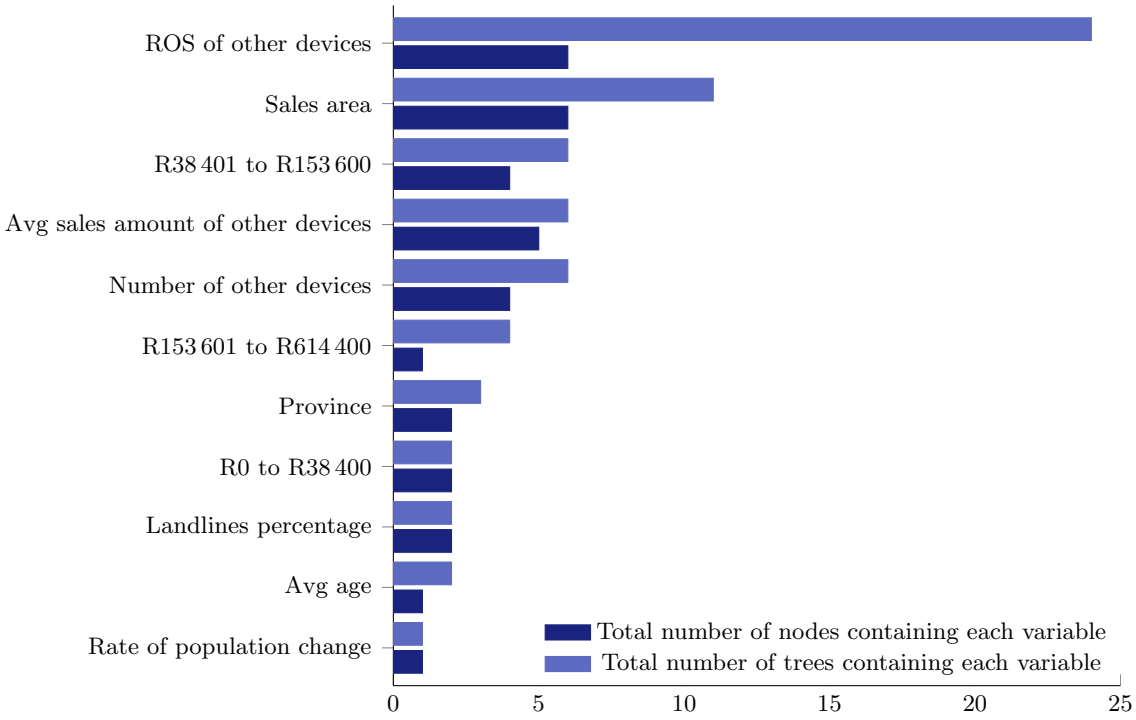


FIGURE 6.4: *The regression variables and the total number of nodes and trees in which they are used.*

6.2.2 Predicted ROS

After applying these regression trees to the stores not currently selling the particular mobile devices, the predicted ROS values, provided in §E.2, are found. An extract of Table E.1 for device 00132 is given in Table 6.8 for illustrative purposes.

Store	Sales area (m^2)	Province	R153601 to R614400	ROS of other devices	Avg sales amount of other devices	Number of other devices	Predicted ROS, r_{ij}
B_143	2 100	KwaZulu-Natal	10%	0,44	R451,54	14	0,06
B_202	1 197	Free State	6%	0,26	R447,63	11	0,03
B_207	992	Limpopo	7%	0,01	R425,00	2	0,03
B_297	1 12	Limpopo	5%	0,06	R296,67	6	0,03
B_340	1 251	Mpumalanga	2%	0,16	R476,09	9	0,06
B_402	1 895	Mpumalanga	6%	0,55	R273,66	10	0,06
C_003	1 970	Western Cape	21%	0,07	R573,00	6	0,03
C_007	2 115	Eastern Cape	14%	0,03	R336,00	3	0,03
C_027	1 863	Gauteng	18%	0,15	R325,45	6	0,03
C_035	2 040	Gauteng	23%	0,13	R700,00	1	0,06

TABLE 6.8: *An extract of the predicted ROS for device 00132 provided in Table E.1.*

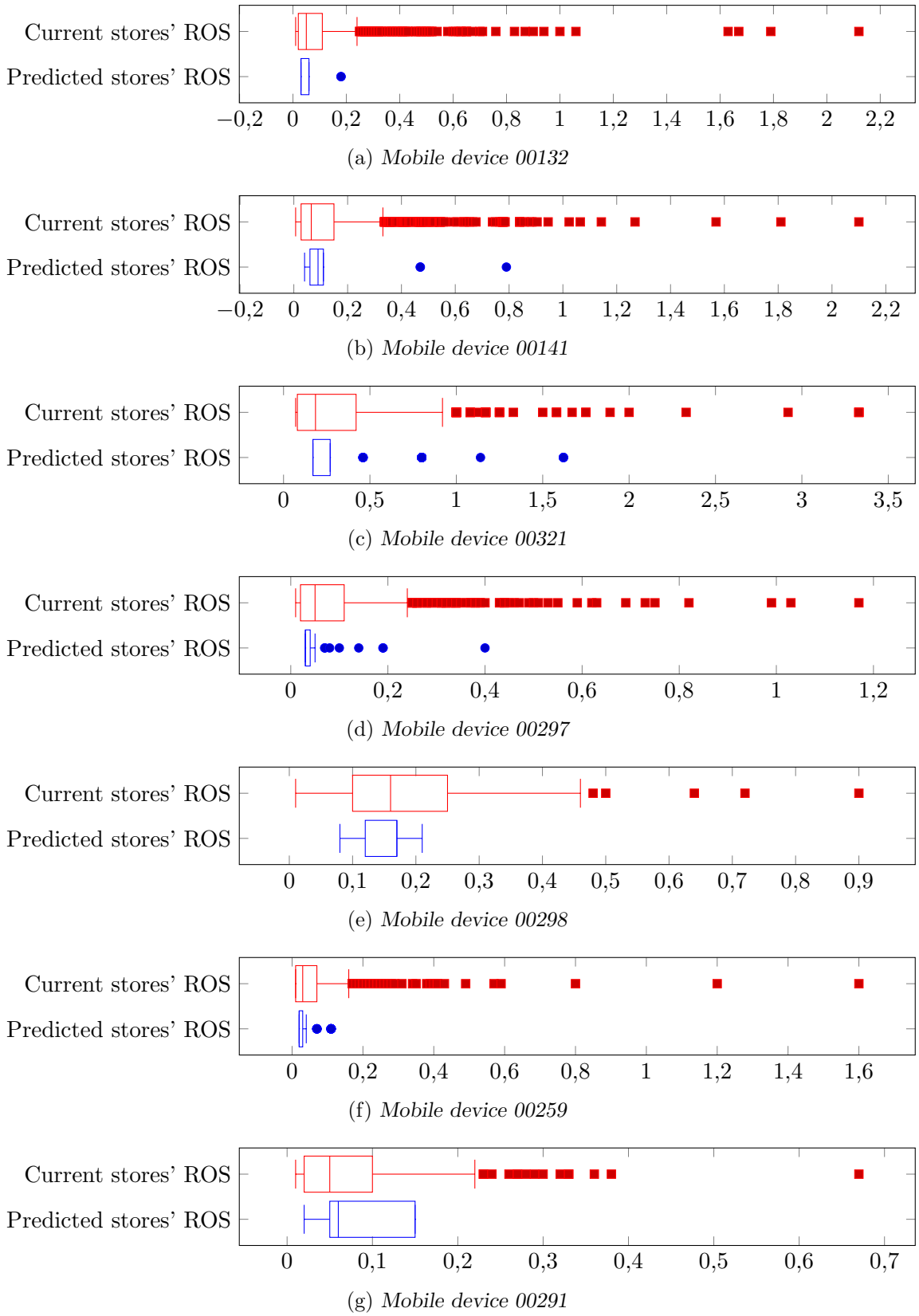
The predicted ROS values range between 0,02 and 1,62 units per day across the seven mobile devices. For each of the seven mobile devices, these predicted ROS values are compared to the ROS values of stores currently selling the device and the distributions of these values are illustrated in Figure 6.5. It can be seen that, for six of the mobile devices, most of the predicted ROS values fall within the interquartile range (IQR) of the known ROS values of stores currently selling the mobile device. Further, some outliers are visible with most of the mobile devices, but these outlier values are not as high as that of the stores currently selling these mobile devices. Store B_143 has a predicted ROS, r_{ij}, of 0,06 units of device 00132 per day, as seen in the first row of Table 6.8. This store should, therefore, order 0,42, rounded up to 1 unit, of device 00132 per week. In general, based on the predicted ROS, The Retailer's stores should order between 1 and 12 units of these devices per store per week.

6.3 Scenario testing

In determining the list of mobile devices to be ranged by The Retailer, equal weights were used for the three performance measures, as described in §5.1.1. Three other weight scenarios were tested and the results are presented in §6.3.1. Other scenarios for the ratio for stores used in calculating the ranging status of mobile devices were also tested and these results are given in §6.3.2.

6.3.1 Performance measures

In order to test the device performance ranking to each of the three performance measures, different weightings are applied in the calculation of the combined index. Initially, all three measures are given equal weights of 0,33. Table 6.9 provides the top 10 mobile devices with their measure values when all three measures have an equal weighting. The weights are then



(a) *Mobile device 00132*

(b) *Mobile device 00141*

(c) *Mobile device 00321*

(d) *Mobile device 00297*

(e) *Mobile device 00298*

(f) *Mobile device 00259*

(g) *Mobile device 00291*

FIGURE 6.5: *The distribution of the ROS of the stores currently selling each mobile device and the predicted ROS for stores not yet selling the particular mobile device.*

adapted so that each of the three measures, in turn, are given a weight of $0,6$ while the other two measures receive a weight of 0,2. The resulting three weight combinations are referred to as Weighting 1 to 3 and are discussed separately.

Rank	Product	I_{ij}^R	I_{ij}^U	I_{ij}^O	I_{ij}^C
1	00222	5	114	233	117
2	00287	18	129	170	106
3	00131	32	203	82	106
4	00132	18	163	131	104
5	00141	25	185	93	101
6	00291	13	120	168	100
7	00321	56	43	202	100
8	00297	17	142	127	95
9	00298	12	188	79	93
10	00259	11	102	130	81

TABLE 6.9: *The top 10 performing devices based on equal weighting.*

Weighting 1

With the first weighting, the ROS index, I_{ij}^R, is given a weight of 0,6, while I_{ij}^U and I_{ij}^O each have a weight of $0,2$. Table 6.10 provides a list of the top 10 devices based on this weighting. Column 6 indicates the change in each device's ranking from the original ranking according to the equal weighting.

Rank	Product	I_{ij}^R	I_{ij}^U	I_{ij}^O	Original I_{ij}^C	New I_{ij}^C	Change in ranking
1	00144	110	85	40	78	91	+12
2	00321	56	43	202	100	83	+5
3	00131	32	203	82	106	76	0
4	00222	5	114	233	117	72	-3
5	00287	18	129	170	106	71	-3
6	00141	25	185	93	101	71	-1
7	00132	18	163	131	104	70	-3
8	00291	13	120	168	100	65	-2
9	00297	17	142	127	95	64	-1
10	00298	12	188	79	93	61	-1

TABLE 6.10: *The top 10 performing devices when the ROS index, I_{ij}^R, has the highest weighting.*

Across all devices, the average absolute change in ranking from the ranking based on equal weighting, is 2,33 positions. In total, 7 devices' ranking did not change and the biggest change in ranking is device 00146 moving up 16 positions from rank 30 to 14.

Weighting 2

Next, the total units sold index, I_{ij}^U is given a weight of 0,6, while I_{ij}^R and I_{ij}^O each have a weight of 0,2. In Table 6.11, the top 10 performing devices according to this weighting, can be seen.

Again, the change in ranking from the equal weighting to Weighting 2 for each device can be seen in column 6.

Rank	Product	I_{ij}^R	I_{ij}^U	I_{ij}^O	Original I_{ij}^C	New I_{ij}^C	Change in ranking
1	00131	32	203	82	106	145	+2
2	00141	3	131	315	101	142	+3
3	00298	40	164	144	93	135	+6
4	00132	25	185	93	104	135	0
5	00222	12	188	79	117	131	-4
6	00287	18	163	131	106	128	-4
7	00297	5	114	233	95	116	+1
8	00291	18	129	170	100	115	-2
9	00311	17	142	127	80	114	+2
10	00259	13	120	168	81	108	0

TABLE 6.11: *The top 10 performing devices when the total units sold index, I_{ij}^U, has the highest weighting.*

Overall, there is an average absolute change of 2,29 positions from the equal weighting to Weighting 2. The biggest change in ranking is 10 positions from device 00321, which moved from 7 to 17. Again 7 devices, including two of the top 10 ranked devices, did not change positions with this weighting.

Weighting 3

Lastly, the average units in stock index, I_{ij}^O, is given a weight of 0,6, while I_{ij}^R and I_{ij}^U each have a weight of 0,2. Table 6.12 provides the top 10 devices based on this fourth weighting. Column 6 once again indicates each device's change in ranking from the equal weighting.

Rank	Product	I_{ij}^R	I_{ij}^U	I_{ij}^O	Original I_{ij}^C	New I_{ij}^C	Change in ranking
1	00222	5	114	233	117	163	0
2	00321	56	43	202	100	141	+5
3	00287	18	129	170	106	131	-1
4	00291	13	120	168	101	128	+2
5	00132	18	163	131	104	115	-1
6	00297	17	142	127	95	108	+2
7	00259	11	102	130	81	101	+3
8	00141	25	185	93	101	98	-3
9	00131	32	203	82	105	96	-6
10	00317	19	101	119	80	96	+2

TABLE 6.12: *The top 10 performing devices when the average units in stock index, I_{ij}^O, has the highest weighting.*

With this weighting, there is an average absolute change in ranking of 2,95 positions. Device 00144 has the biggest change in ranking and moved 10 positions from 13 to 23. In total, two devices remain in the same position as with the equal weighting, including the top ranked device.

6.3.2 Ranging status

As described in §5.1.2, a subset of the best and worst performing stores is used to determine the ranging status of a device in a particular cluster. Table 6.13 contains the number of clusters with a ranging status of 'Yes', 'Maybe' or 'No' for each of the top 10 devices when using the 60%:40% ratio for stores, as described in §5.1.2. In total, 14 device-cluster combinations have a ranging status of 'Yes', 15 have a ranging status of 'Maybe' and 31 have a ranging status of 'No'.

		Ranging status		
Rank	Product	Yes	Maybe	No
1	00222	0	2	4
2	00287	0	2	4
3	00131	0	1	5
4	00132	3	1	2
5	00141	2	2	2
6	00291	0	2	4
7	00321	3	2	1
8	00297	2	2	2
9	00298	1	0	5
10	00259	3	1	2
Total		14	15	31

TABLE 6.13: *The number of clusters with a ranging status of 'Yes', 'Maybe' or 'No' for each of the top 10 devices using the 60%:40% ratio.*

To test the effect of this ratio on the results, three different ratios, 55%:45%, 65%:35% and 70%:30% are tested. These ratio's will be referred to as Ratio 1, 2 and 3, respectively. The ratio can be changed so that more or fewer stores are required in the top 20% for a device to be ranged in the cluster, but more than half of the subset must be top performing stores for the ratio to indicate a well performing cluster. On the other hand, if the ratio becomes too skewed towards the top performing stores, the stores in the bottom 20% will no longer have an impact on the result and it would not be necessary to consider the bottom 20% of stores anymore.

Ratio 1

With the first ratio, at least 55% of the subset must be in the top 20% of stores. Table 6.14 contains the number of cluster that received each of the three ranging statuses with this ratio.

With this ratio, 18 device-cluster combinations have a ranging status of 'Yes', 8 have a ranging status of 'Maybe' and 34 have a ranging status of 'No'. When comparing these results to the 60%:40% ratio, four device-cluster combinations changed from a 'Maybe' ranging status to 'Yes' and three device-cluster combinations changed from 'Maybe' status to 'No'. The ranging status of devices 00222, 00291 and 00298 did not change at all, indicating that even with a weaker requirement, these devices should still not be ranged in the particular cluster.

Ratio 2

The second ratio tested is 65%:35%, meaning that at least 65% of the subset stores must be in the top 20% and a summary of the results of this ratio can be seen in Table 6.15. This ratio results in 9 device-cluster combinations with a ranging status of 'Yes', indicating that with more

		Ranging status		
Rank	Product	Yes	Maybe	No
1	00222	0	2	4
2	00287	0	1	5
3	00131	1	0	5
4	00132	3	0	3
5	00141	2	1	3
6	00291	0	2	4
7	00321	4	1	1
8	00297	3	1	2
9	00298	1	0	5
10	00259	4	0	2
Total		18	8	34

TABLE 6.14: *The number of clusters with a ranging status of 'Yes', 'Maybe' or 'No' for each of the top 10 devices using the 55%:45% ratio.*

		Ranging status		
Rank	Product	Yes	Maybe	No
1	00222	0	2	4
2	00287	0	2	4
3	00131	0	1	5
4	00132	2	2	2
5	00141	0	5	1
6	00291	0	2	4
7	00321	2	3	1
8	00297	1	3	2
9	00298	1	0	5
10	00259	3	1	2
Total		9	21	30

TABLE 6.15: *The number of clusters with a ranging status of 'Yes', 'Maybe' or 'No' for each of the top 10 devices using the 65%:35% ratio.*

strict requirements, fewer devices will immediately be ranged in a cluster. This ratio also results in more device-cluster combinations with a ranging status of 'Maybe', namely 21, and finally 30 device-cluster combinations with a ranging status of 'No'. When comparing the results of this ratio to the initial 60%:40% ratio, it can be seen that only four devices have a change in ranging status. Overall, five device-cluster combinations changed from a ranging status 'Yes' to 'Maybe', while one device-cluster combination changed from a ranging status 'No' to 'Maybe'. All other clusters are still assigned the same ranging status as with the initial 60%:40% ratio.

Ratio 3

The third ratio tested, is the most extreme change of the three and requires at least 70% of the subset stores to be in the top 20%. With this ratio, only 6 device-cluster combinations have a ranging status of 'Yes', compared to the original 14. A total of 37 device-cluster combinations now have a ranging status of 'Maybe' and 17 have a ranging status of 'No'. With this ratio, the ranging status of only two devices (00287 and 00291) remains unchanged from the initial 60%:40% ratio. In total, eight device-cluster combinations changed from ranging status 'Yes' to 'Maybe' and 14 device-cluster combinations changed from ranging status 'No' to 'Maybe'. With most device-cluster combinations having a 'Maybe' status, this ratio is not ideal as it results in uncertainty and further investigation is required to determine the final ranging status of these devices.

		Ranging status		
Rank	Product	Yes	Maybe	No
1	00222	0	3	3
2	00287	0	2	4
3	00131	0	3	3
4	00132	1	5	0
5	00141	0	6	0
6	00291	0	2	4
7	00321	2	4	0
8	00297	0	6	0
9	00298	0	3	3
10	00259	3	3	0
Total		6	37	17

TABLE 6.16: *The number of clusters with a ranging status of 'Yes', 'Maybe' or 'No' for each of the top 10 devices using the 70%:30% ratio.*

CHAPTER 7

Conclusion

Contents

In this chapter, a summary of this study and results are provided in §7.1. This is followed by a discussion of limitations and future work in §7.2.

7.1 Project summary

In this study, the aim is to identify the best range of mobile devices for The Retailer to keep in stock. As The Retailer's stores have very different characteristics, multiple mobile device ranges are found to best fit The Retailer's needs and this is done through various steps.

Firstly, The Retailer's stores are grouped into clusters so that the stores within each cluster have similar characteristics. The results, provided in §4.3, show that a total of six clusters, two per store type, are formed. Analysing the distributions of each of the store characteristics in the different clusters, the data suggest that the stores are well separated into the different clusters, as there is a clear distinction between the higher and lower valued data points within each store type. Consider, for example, the distribution of the scaled age variable in Figure 7.1, also shown in §4.3. For store types A and C there is a 0,3 difference between the two cluster medians, while store type B has a 0,4 difference between the two cluster medians. The landlines percentage and income distributions show a similar differentiation between the two clusters within each store type. However, the median values for the rate of change variable are much closer, due to two outliers in store type B.

Once the clusters were obtained, three performance measures, namely ROS, total units sold and average units in stock, were used to determine the best range of mobile devices for each of the six clusters. The mobile devices are ranked based on their performance in these three measures and some scenario testing is performed on the weighting of these measures in the final ranking of the devices. In turn, each measure is given a weighting of 0,6 while the other two measures are given equal weights. This is to observe the change in ranking when more emphasis is placed on a particular measure. With each of the three weightings, the average absolute change in ranking is between two and three positions. Further, with the first two weightings (I_{ij}^R and I_{ij}^U each having a weight of 0,6) seven devices stayed in the same position, while only two devices stayed

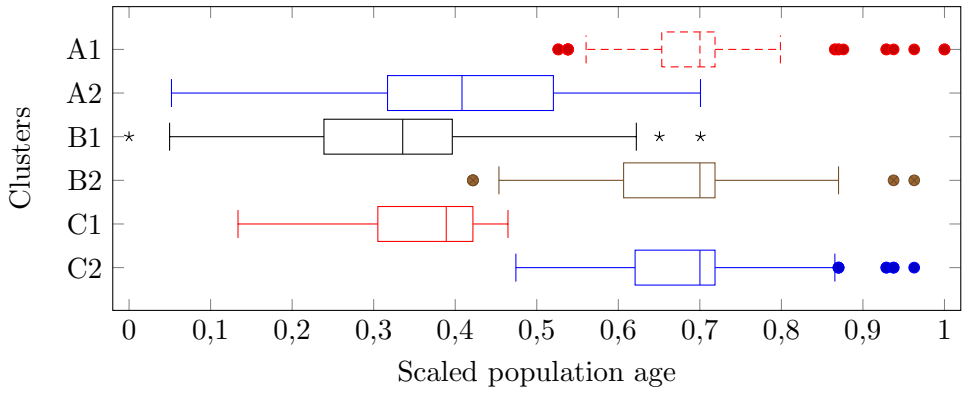


FIGURE 7.1: *The distribution of the scaled age variable per cluster.*

in the same position with the third weighting (I_{ij}^{O} having a weight of 0,6). The biggest change in ranking is device 00146 moving up 16 positions, device 00321 moving down 10 positions and device 00144 moving down 10 positions when more emphasis is placed on measures I_{ij}^{R}, I_{ij}^{U} and I_{ij}^{O}, respectively. Although the largest change in position occurs when more emphasis is placed on measure I_{ij}^{R}, it also has the lowest average absolute change in ranking. Further, the highest average absolute change in ranking is seen when more emphasis is placed on measure I_{ij}^{O} and only two devices with no change in ranking. Based on these results, the ranking of devices is the most sensitive to a change in I_{ij}^{O} and the least sensitive to a change in I_{ij}^{U}.

After the ranking, two approaches are followed to determine whether or not a device should be ranged, starting with the highest and lowest ranked devices, respectively. For the highest ranked devices, a ranging status is assigned to each cluster so that clusters with a higher percentage of top performing stores for the specific device, receive a ranging status of 'Yes' and clusters with a lower percentage of top performing stores for the specific device, receive a ranging status of 'No'. Finally, clusters with an average performance receive a ranging status of 'Maybe'. Once again, some scenario testing is done to analyse the change in ranging status when the ratio between the top and bottom performing stores changes. Increasing the initial ratio with 5% leads to an increase of six device-cluster combinations with a ranging status of 'Maybe'. Similarly, decreasing this ratio with 5% leads to a decrease of seven device-cluster combinations with a ranging status of 'Maybe'. An increase of 10% to 70%:30% leads to the biggest change in ranging status, with a total of 22 device-cluster combinations changing to a ranging status of 'Maybe'.

With the lowest ranked devices, devices are kept in a store's range when this device performs above the average for the specific store. If the device performs below average, it is removed from the store's range.

A sample of the highest and lowest ranked mobile devices is taken to exhibit the findings from the final list of devices that should be ranged in certain stores of The Retailer, according to this study. The results, as provided in Chapter 6, show that seven of the 10 highest ranked mobile devices are ranged in at least one full cluster, while the other three devices are not ranged. Four devices from the 10 lowest ranked devices are ranged in between three and ten stores each across the different clusters. Thus, 11 of the 30 sampled mobile devices are ranged using this methodology, suggesting that this methodology succeeds in reducing the variety of mobile devices ranged by The Retailer by removing under performing devices. Reducing the number of different devices ranged, can significantly reduce stock-keeping and administration costs. Further, reducing the stock in a store can also have other benefits, for example a decrease

in lost or damaged stock as well as an improvement in the cash flow of the business.

Lastly, the ROS of each device is used to determine how much stock to keep in those stores in which a device is not currently ranged, but is included in the new proposed range. This provides The Retailer with an estimate of the number of units to keep in stock when first adding these devices to a store's range. To achieve this, the store characteristics, used to group stores, are used again. Other variables regarding the performance of other devices in each store are also used to build a regression tree for each device. Although the R^2 values for the trees are not very high, this method can still provide a more accurate estimate of the amount of stock needed when compared to The Retailer's 'best guess'. Further, Figure 6.3 suggests that there is a strong correlation between the amount of training data and the performance of the regression tree. Therefore, the assumption can be made that the accuracy of this method will improve once the additional stores range the relevant devices.

7.2 Some limitations and future work

Although the results indicate that this methodology can assist The Retailer in determining the best mobile devices to range in its stores, this study is not without limitations. Firstly, only six months of stock data are available, making it difficult to accurately identify seasonal patterns. Therefore, as the ROS calculation is dependent on this data, it is not possible to predict the ROS over time. As a result, one ROS value is used without considering seasonality. Secondly, some mobile devices are currently ranged in a very small number of stores and, although their performance in these stores can be calculated, their performance cannot be compared to that of other devices with much confidence. Furthermore, their performance cannot be extrapolated to new stores as there is not enough data to identify the relationship between the ROS and independent variables. Thirdly, The Retailer stores very little information beyond a mobile device's name. Although data for device characteristics are widely available for large manufacturers, data for smaller manufacturers are not as accessible. Different models or production years of mobile devices can often lead to these devices having the same name, but different characteristics. As the naming conventions in The Retailer's data are not consistent throughout the dataset, it is not always possible to match a particular product to the correct device characteristics. This lack of mobile device characteristics could potentially cause the proposed range to include nearly identical devices, reducing the efficacy of the range.

Based on the results and limitations of this study, suggestions for future improvements can be made. With a longer time period of data available, a time series forecast can be used to estimate the ROS. A moving ROS can then be calculated, rather than the overall ROS used in this study, and this will allow for any seasonal patterns to be considered. This will lead to a more accurate ROS estimate, effectively improving the results of the regression trees used in this study. Another recommendation for future study is to consider mobile device characteristics in the device ranging and ROS estimation. By clustering mobile devices with similar characteristics and using these clusters when determining the mobile device range, two limitations of this study can be addressed. Firstly, determining the mobile device ranges based on these characteristics can ensure that highly substitutable products are not placed in the same range, resulting in a smaller, unique range. Furthermore, combining devices into clusters can increase the amount of data available when predicting the ROS. As identified in §6.2.1, increasing the amount of training data can improve the accuracy of the predictions of the regression trees. However, not including the mobile device characteristics in the methodology can also be considered a benefit of the proposed methodology, as it can be easily adapted for other products sold by The Retailer.

Bibliography

[1] Aldenderfer M and Blashfield R, 2011, *Cluster Analysis*, Sage Publications.

[2] Cell C, 2021, *Coverage map*, [Online], [Cited on February 14th, 2021], Available from `https://www.cellc.co.za/cellc/coverage-map`.

[3] Chen X, Shi C, Wang Y and Zhou Y, 2021, *Dynamic assortment planning under nested logit models*, Production and Operations Management, **30(1)**, pp. 85 – 102.

[4] Corsten H, Hopf M, Kasper B and Thielen C, 2018, *Assortment planning for multiple chain stores*, OR Spectrum, **40**, pp. 875 – 912.

[5] Cox LA, 2001, *Forecasting demand for telecommunications products from cross-sectional data*, Telecommunications Systems, **16(3-4)**, pp. 437 – 454.

[6] Dekimpe MG, 2020, *Retailing and retailing research in the age of big data analytics*, International Journal of Research in Marketing, **37**, pp. 3–14.

[7] Eighty20, 2014, *Retailers' motivation for offering financial services*, [Online], [Cited on October 10th, 2020], Available from `https://cenfri.org/publications/understanding-retailers-motivation-for-providing-financial-products-and-services-in-south-africa/`.

[8] Eighty20, 2021, *Eighty20 Data Portal*, [Online], [Cited on June 16th, 2021], Available from `https://dataportal.eighty20.co.za/`.

[9] Ester M, Kriegel H, Sander J and Xu X, 1996, *A density-based algorithm for discovering clusters in large spatial databases with noise*, **19**, pp. 226–231.

[10] Fildes R, Ma S and Kolassa S, 2019, *Retail forecasting: Research and practice*, International Journal of Forecasting.

[11] Fisher ML and Vaidyanathan R, 2014, *A demand estimation procedure for retail assortment optimization with results from implementation*, Management Science, **60(10)**, pp. 2401 – 2415.

[12] Forina M, Armanino C and Raggio V, 2002, *Clustering with dendrograms on interpretation variables*, Analytica Chimica Acta, **454(1)**, pp. 13–19, Available from `https://www.sciencedirect.com/science/article/pii/S0003267001015173`.

[13] Gaudagni PM and Little JDC, 1983, *A logit model of brand choice calibrated on scanner data*, Marketing Science, **2(3)**, pp. 203–238.

[14] Gaur V and Honhon D, 2006, *Assortment planning and inventory decisions under a locational choice model*, Management Science, **52(10)**, pp. 1528–1543.

[15] Han J, Kamber M and Pei J, 2012, *Data Mining: Concepts and Techniques*, 3rd Edition, Morgan Kaufmann Publishers.

[16] Hotelling H, 1929, *Stability in competition*, The Economic Journal, **39(153)**, pp. 41–57.

[17] Hyndman RJ and Athanasopoulos J, 2018, *Forecasting: Principles and Practice*, 2nd Edition, OTexts, Available from `https://otexts.com/fpp2/arima.html`.

[18] Işiklar G and Büyüközkan G, 2007, *Using a multi-criteria decision making approach to evaluate mobile phone alternatives*, Computer Standards & Interfaces, **29**, pp. 265–274.

[19] James G, Witten D, Hasie T and Tibshirani R, 2013, *An Introduction to Statistical Learning: with Applications in R*, Springer Science and Business Media.

[20] Jannu S and Jana PK, 2016, *A grid based clustering and routing algorithm for solving hot spot problem in wireless sensor networks*, Wireless Netw, **22**, p. 1901–1916.

[21] Kalba K, 2008, *The adoption of mobile phones in emerging markets: Global diffusion and the rural challenge*, International Journal of Communication, **2**, pp. 631 – 661.

[22] Kargari M and Sepehri MM, 2012, *Stores clustering using a data mining approach for distributing automotive spare-parts to reduce transportation costs*, Expert Systems with Applications, **39**, p. 4740–4748.

[23] Karjaluoto K, Karvonen J, Kesti M, Koivumäki T, Manninen M, Pakola J, Ristola A and Salo J, 2005, *Factors affecting consumer choice of mobile phones: Two studies from Finland*, Journal of Euromarketing, **13(3)**.

[24] Kök AG and Fisher ML, 2007, *Demand estimation and assortment optimisation under substitution: Methodology and application*, Operations Research, **55(6)**, pp. 1001–1021.

[25] Kök AG, Fisher ML and Vaidyanathan R, 2009, *Assortment planning: Review of literature and industry practice*, International Series in Operations Research and Management Science, **122**, pp. 99–153.

[26] Lancaster KJ, 1966, *A new approach to consumer theory*, Journal of Political Economy, **74(2)**, pp. 132–157.

[27] Madashi K and Raghupataiah C, 2014, *Buying behaviour towards mobile phones: A comparative analysis of rural and urban consumers*, Journal of Commerce and Management Thought, **1**, pp. 119–135.

[28] Mantovani RG, Horváth T, Cerri R, Junior SB, Vanschoren J and de Leon Ferreira de Carvalho ACP, 2018, *An empirical study on hyperparameter tuning of decision trees*, ArXiv, **abs/1812.02207**.

[29] Mithrakumar M, 2019, *How to tune a Decision Tree?*, [Online], [Cited on December 15th, 2021], Available from `https://towardsdatascience.com/how-to-tune-a-decision-tree-f03721801680`.

[30] MTN (PTY) LTD, 2021, *Coverage map*, [Online], [Cited on February 11th, 2021], Available from `https://www.mtn.co.za/Pages/Coverage_Map.aspx`.

[31] Nidheesh N, Nazeer KAA and Ameer PM, 2019, *A hierarchical clustering algorithm based on silhouette index for cancer subtype discovery from genomic data*, Neural computing and applications, **32(15)**, pp. 11459–11476.

[32] Pagnuco IA, Pastore JI, Abras G, Brun M and Ballarin VL, 2017, *Analysis of genetic association using hierarchical clustering and cluster validation indices*, Genomics, **109(5)**, pp. 438–445, Available from `https://www.sciencedirect.com/science/article/pii/S0888754317300575`.

[33] Pavlis M, Dolega L and Singleton A, 2018, *A modified dbscan clustering method to estimate retail center extent*, Geographical analysis, **50(2)**, pp. 141–161.

[34] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M and Duchesnay E, 2011, *Scikit-learn: Machine learning in Python*, Journal of Machine Learning Research, **12**, pp. 2825–2830.

[35] Ragsdale CT, 2011, *Managerial Decision Modeling*, 6th Edition, South Western, Cengage Learning.

[36] Ros F and Guillaume S, 2019, *A hierarchical clustering algorithm and an improvement of the single linkage criterion to deal with noise*, Expert Systems with Applications, **128**, pp. 96–108.

[37] Rousseeuw PJ, 1987, *Silhouettes: A graphical aid to the interpretation and validation of cluster analysis*, Journal of Computational and Applied Mathematics, **20**, pp. 53–65, Available from `https://www.sciencedirect.com/science/article/pii/0377042787901257`.

[38] SAARF, 2017, *Living Standards Measure*, [Online], [Cited on October 10th, 2020], Available from `http://www.saarf.co.za/lsm/lsms.asp`.

[39] Schorfheide F and Wolpin KI, 2012, *On the use of holdout samples for model selection*, The American economic review, **102(3)**, pp. 477–481.

[40] SignalBoostercom, 2018, *Does smartphone make and model impact cellular signal?*, [Online], [Cited on November 8th, 2020], Available from `https://www.signalbooster.com/blogs/news/does-smartphone-make-and-model-impact-cellular-signal`.

[41] Statistics SA, 2020, *Quarterly financial statistics, March 2020*, [Online], [Cited on October 5th, 2020], Available from `https://www.statssa.gov.za/?page_id=1854&PPN=P0044`.

[42] Statistics SA, 2018, *Retail trade industry, 2018*, [Online], [Cited on October 5th, 2020], Available from `https://www.statssa.gov.za/?page_id=1854&PPN=Report-62-01-02`.

[43] Tableau Software, LLC, 2021, *Tableau*, [Online], [Cited on February 19, 2022], Available from `https://www.tableau.com`.

[44] The Scipy Community, 2021, *scipy.cluster.hierarchy.dendrogram*, [Online], [Cited on August 26th, 2021], Available from `https://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.hierarchy.dendrogram.html`.

[45] Theodoridis S and Koutroubas K, 2009, *Pattern Recognition*, 4th Edition, Academic Press.

[46] Van Ryzin G and Mahajan S, 1999, *On the relationship between inventory costs and variety benefits in retail assortments*, Management science, **45(11)**, pp. 1496–1509.

[47] Vodacom, 2014, *Coverage map*, [Online], [Cited on February 20th, 2021], Available from `https://www.vodacom.co.za/vodacom/coverage-map?PageSpeed=noscript`.

[48] Ward JH, 1963, *Hierarchical grouping to optimize an objective function*, Journal of the American Statistical Association, **58(301)**, pp. 236–244, Available from `http://www.jstor.org/stable/2282967`.

[49] Wilcox RR, 2019, *Multicolinearity and ridge regression: results on type i errors, power and heteroscedasticity*, Journal of applied statistics, **46(5)**, pp. 946–957.

[50] Wu Y, 2021, *Can't ridge regression perform variable selection?*, Technometrics, **63(2)**, pp. 263–271.

[51] Zhou J, Rau PLP and Salvendy G, 2014, *Age-related difference in the use of mobile phones*, Universal access in the information society, **13(4)**, pp. 401–413.

APPENDIX A

LSM grouping

The Living Standards Measure (LSM) of the South African Audience Research Foundation (SAARF) is used to describe the clientele of each of The Retailer's stores as discussed in Chapter 1. The determination of each of the LSM groups are given in this appendix.

There are 29 attributes, as indicated in Table A.1, used to determine the appropriate LSM group for any given household. Each of these attributes can be represented by a binary variable x_i, where

$$x_i = \begin{cases} 1 & \text{if the attribute is present in the household,} \\ 0 & \text{if not.} \end{cases}$$

A weight, ω_i, representing the contribution of attribute i to the classification of the LSM groups, is assigned to each attribute i. A particular household is then assigned to a LSM group according to the sum of the weights of the attributes applicable to the specific household and a constant of -0.81052, *i.e.*

$$\sum_{i=1}^{29} \omega_i x_i - 0.81052. \tag{A.1}$$

If the value of the weight-sum (A.1) is within the weight-sum range of a particular LSM group (as shown in Table A.2), the household is assigned to this LSM group [38].

	Attribute	Weight
1	Hot running water from a geyser	0,185224
2	Computer - Desktop / Laptop	0,311118
3	Electric Stove	0,163220
4	No domestic workers or household helpers	-0,301330
5	0 or 1 radio set in household	-0,245000
6	Flush toilet in/outside house	0,113306
7	Motor vehicle in household	0,167310
8	Washing machine	0,149009
9	Refrigerator of combined fridge/freezer	0,134133
10	Vacuum cleaner/floor polisher	0,164736
11	Pay TV subscription	0,127360
12	Dishwashing machine	0,212562
13	3 or more cellphones in household	0,184676
14	2 cellphones in household	0,124007
15	Home security service	0,151623
16	Deep freezer - free standing	0,116673
17	Microwave oven	0,126409
18	Rural rest	-0,129360
19	House/cluster house/town house	0,113907
20	DVD player / Blu Ray Player	0,096070
21	Tumble dryer	0,166056
22	Home theatre system	0,096072
23	Home telephone	0,104531
24	Swimming Pool	0,166031
25	Tap water in house	0,123015
26	Built-in kitchen sink	0,132822
27	TV set	0,120814
28	Air conditioner (excl fans)	0,178044
29	Metropolitan dweller (250 000+)	0,079321

TABLE A.1: *LSM attributes and their weights*

LSM	Minimum Weight-sum	Maximum Weight-sum
1		-1,390140
2	-1,390139	-1,242000
3	-1,242001	-1,011800
4	-1,011801	-0,691000
5	-0,691001	-0,278000
6	-0,278001	0,382000
7	0,381999	0,801000
8	0,800999	1,169000
9	1,168999	1,745000
10	1,744999	

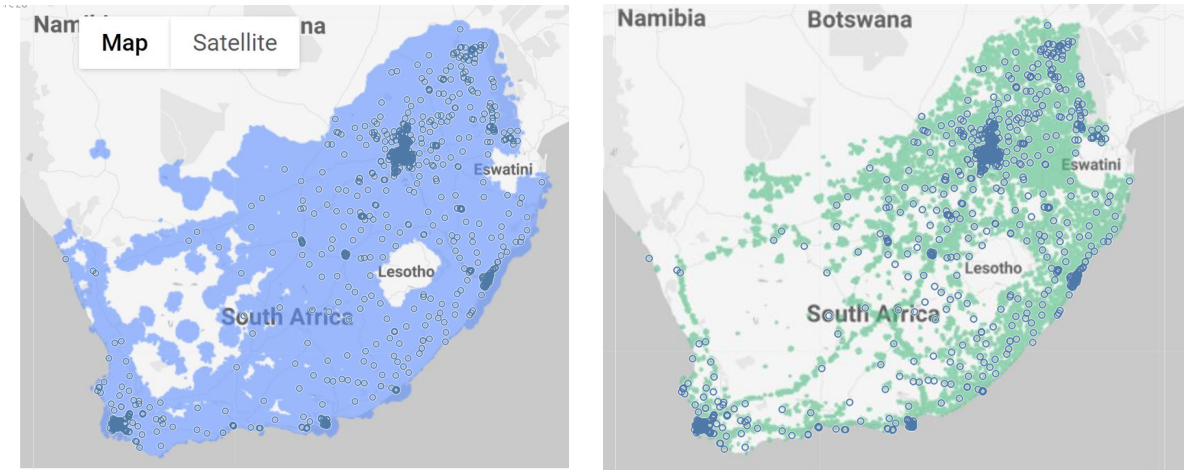
TABLE A.2: *LSM groups and their weight-sum ranges*

APPENDIX B

Stores' network coverage maps

The process to determine the networks coverage of The Retailer's stores is described in §3.2.2. The network coverage maps for MTN were given in §3.2.2 and in this appendix, the network coverage maps for Cell C and Vodacom are provided.

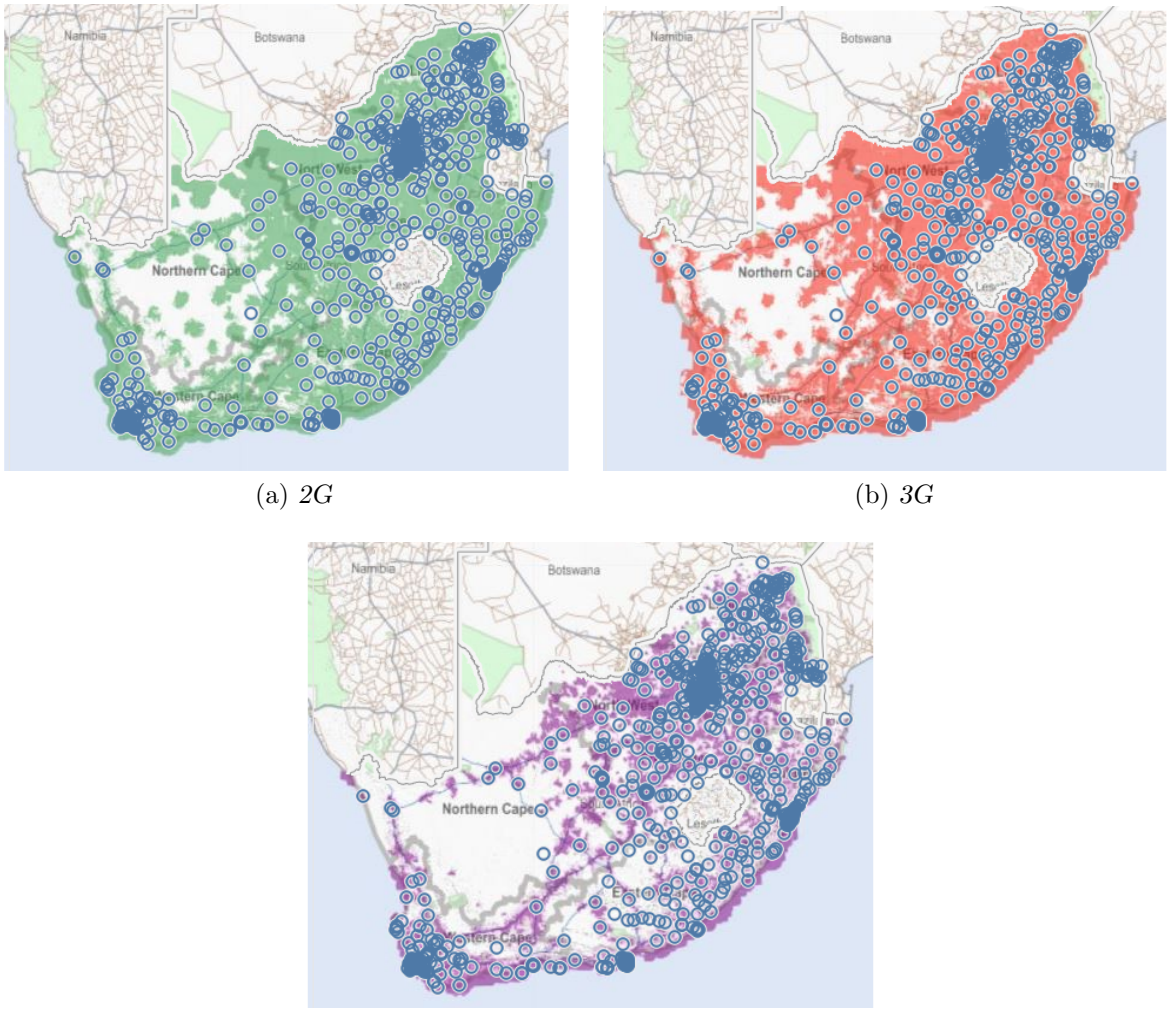
The 2G and 4G coverage maps for Cell C can be seen in Figure B.1 (a) and (b), respectively, and it is clear that all of The Retailer's stores are covered by both networks. The coverage map for Cell C's 3G network was not available at the time of data collection, however, as all stores are covered by the 4G network, it is safe to assume that all stores are also covered by the 3G network.



(a) *2G* (b) *4G*

FIGURE B.1: *A comparison of The Retailer's store locations and the coverage maps for Cell C's 2G and 4G networks [2], where the coloured areas in each map indicate the specific type of coverage and the blue circles indicate store locations.*

In Figure B.2 (a), (b) and (c), respectively, the coverage maps for Vodacom's 2G, 3G and 4G networks are illustrated. Once again, all of The Retailer's stores are covered by the three Vodacom networks, as seen with MTN and Cell C as well.



(a) *2G* (b) *3G*

(c) *4G*

FIGURE B.2: *A comparison of The Retailer's store locations and the coverage maps for Vodacom's 2G, 3G and 4G networks [47], where the coloured areas in each map indicate the specific type of coverage and the blue circles indicate store locations.*

APPENDIX C

Agglomerative hierarchical clustering

Agglomerative hierarchical clustering starts by considering each data point to be in its own cluster. Then, these clusters are merged based on a dissimilarity function until only one cluster remains [15]. The dissimilarity function consists of two elements, called the *distance measure*, or *dissimilarity*, and *linkage*. For the distance measure, Euclidean or Manhattan distances are often used, which is then combined with a linkage method to form the specific dissimilarity function [19].

Different linkage methods are described in C.1 and the so-called dendrogram formed by applying agglomerative clustering is discussed in C.2. The methods described in this appendix are used in Chapter 4 to find the best grouping of The Retailer's stores. Finally, in §C.3, the silhouette coefficient, a measure used to determine the number of cluster in a dataset, is discussed.

C.1 Linkage

Various linkage methods are available in agglomerative hierarchical clustering, however, some are more commonly used than others. This section will describe five of the most commonly used linkage methods in more detail.

Single linkage clustering is one of the most common agglomerative hierarchical clustering methods. In this case, to calculate the dissimilarity for any two clusters, a dissimilarity matrix is built for all elements in the two clusters. The minimum dissimilarity found in the matrix is then recorded as the dissimilarity between the two clusters in question. At each level of the hierarchy, clusters with the smallest dissimilarity are combined [19]. An advantage of single linkage clustering is that it does not require complex mathematical formulations and is easy to understand. On the other hand, the single linkage clustering method only considers each data point once, and a bad split in the data early in the process cannot be changed or improved later [1].

Contrary to single linkage clustering, *complete linkage clustering* uses the maximum dissimilarity between elements in two clusters and records this as the dissimilarity between the two clusters. The two clusters with the smallest dissimilarity are still combined, as with single linkage clus-

tering [19]. This method of using the maximum dissimilarity ensures that all elements within the two clusters are relatively close to each other. However, a drawback of this method is that outliers have a large effect on the clusters formed [1].

Another common linkage method, called *average linkage clustering*, was developed to find a solution between the extremes of single and complete linkage clustering [1]. The dissimilarity between two clusters is recorded as the average of all values in the dissimilarity matrix, that is the average distance between all pairs of data points. Using this method reduces the effect of outliers in the data [19].

With *centroid linkage clustering*, the dissimilarity between two clusters is recorded as the distance between the centroids (or mean) of the two clusters. As opposed to other linkage methods, only one distance value is calculated when using centroid linkage clustering and, therefore, saves computation time. However, it is possible that the distance to the centroid of a new combined cluster can be smaller than the distance to the two original cluster centroids, and this makes the dendrogram difficult to interpret [19].

Finally, with *Ward's linkage clustering*, the goal is to minimise the error sum of squares (ESS), or the total variance from cluster centres [48]. Thus, at each step, clusters are combined in such a way that leads to the smallest increase in the total squared distance between cluster elements and their relevant new combined cluster centres [45]. This leads to more compact clusters compared to some other linkage methods, as the cluster elements are all located close to the cluster centre and, thus, to each other. The resulting clusters found with this method are often also more balanced than when using other linkage methods [36].

C.2 Dendrogram

Hierarchical clustering results are represented in a tree, called a *dendrogram*. Individual data points, or *leaves*, are found at the bottom of the dendrogram. These are joined together by *branches* as clusters are formed, until only one cluster remains at the top of the tree. The y-axis indicates the distance measure, or dissimilarity, between clusters when they are joined. Thus, data points in clusters formed at the bottom of the tree are very similar, with a small dissimilarity, while data points in clusters formed at the top of the tree have larger dissimilarities and are, therefore, more different [19].

A dendrogram is a visual way of identifying clusters for a specific dataset. First, a horizontal line is drawn across the tree graph. The number of clusters is the number of branches intersecting with the horizontal line. The number of clusters used, can be increased or decreased by moving this line along the y-axis. Moving the line down the graph will increase the number of clusters found in the data, while moving the line upwards will decrease the number of clusters [19]. When using predefined code, such as the `scipy.cluster.hierarchy.dendrogram` package in Python, a threshold for this horizontal line can be set, or a default value is used [44]. With this package, the threshold is calculated as a percentage of the maximum distance found on the y-axis, with the default value set to 70%. This package then identifies the required number of clusters and colours individual clusters based on this threshold.

Since agglomerative hierarchical clustering is based on distance measures, the dendrogram can also be used to visually validate clusters found with other methods. In clustering, data points within clusters should be very similar to each other, while being very different from data points in other clusters. The dendrogram is an easy way to validate this similarity where the best set of clusters is the one found by selecting the clusters obtained by the horizontal line where the

distance between a clustering and the next group of clusters on the y-axis is the largest [12].

C.3 Silhouette coefficient

The *silhouette coefficient* is often used with a range of clustering methods and is a measurement of how compact each cluster is, as well as how far apart the different clusters are from one another. This measurement of the goodness of a specific clustering is one of the most popular ways of determining the number of clusters in a dataset [31]. To calculate the silhouette coefficient of a specific clustering, the distances between all data points are needed, as well as the cluster to which each data point is assigned. Assuming data point i is assigned to cluster A, the average distance of data point i to all other data points in cluster A is denoted as $\bar{g}_i$. Furthermore, the distance $d(i, C)$ between data point i and a cluster C, where $C \neq A$, is calculated as the average distance between data point i and all data points in cluster C. Once the distance from data point i to each cluster is calculated, the closest distance b_i between data point i and a cluster is

$$b_i = \min_{\text{all clusters } C \neq A} \{d(i, C)\}.$$

The silhouette coefficient s_i for data point i can be calculated as

$$s_i = \frac{b_i - \bar{g}_i}{\max\{\bar{g}_i, b_i\}}.$$

Thus, s_i can take any value between -1 and 1. It can also be seen that numbers closer to 1 are better, indicating that the distances within cluster A are smaller than the distances between clusters A and B, the cluster for which $d(i, B) = b_i$ [37, 45].

In python, the silhouette coefficient can be calculated using the `sklearn.metrics.silhouette_score` package and returns the average silhouette coefficient of all data points. By default, this package uses the Euclidean distance, although other distance measures can be specified. Further, this package enables the user to calculate the average silhouette coefficient on a random sample to reduce the computation time when working with large datasets. To sample data, the user specifies the required sample size as a parameter and the package automatically selects a random sample of that size [34].

To choose the best number of clusters, the average silhouette coefficient is calculated for different numbers of clusters and the best number of clusters is the one with the highest average silhouette coefficient.

APPENDIX D

Decision tree regression

In this project, the decision tree approach is used to predict the ROS for mobile devices that are not currently stocked at the specific store. These ROS predictions are done in §5.2. In this appendix, the process of creating and validating a decision tree is discussed.

D.1 Decision tree

A decision tree consists of a series of questions or conditions, creating splits in the dataset and can be represented in a tree graph. Each node in the graph represents a question or condition that splits the data. As depicted in Figure D.1, each node has only two possible outcomes, leading to either another condition or a terminal node. The different outcomes found at terminal nodes are mutually exclusive. At each terminal node, a prediction is made by calculating the mean of all data points at the particular node.

The tree continues to grow by adding more nodes, until some termination criteria is met. A greedy algorithm is used to find the next best condition at each step of the tree, selecting the variable and value that will result in the smallest increase in the total error value [19].

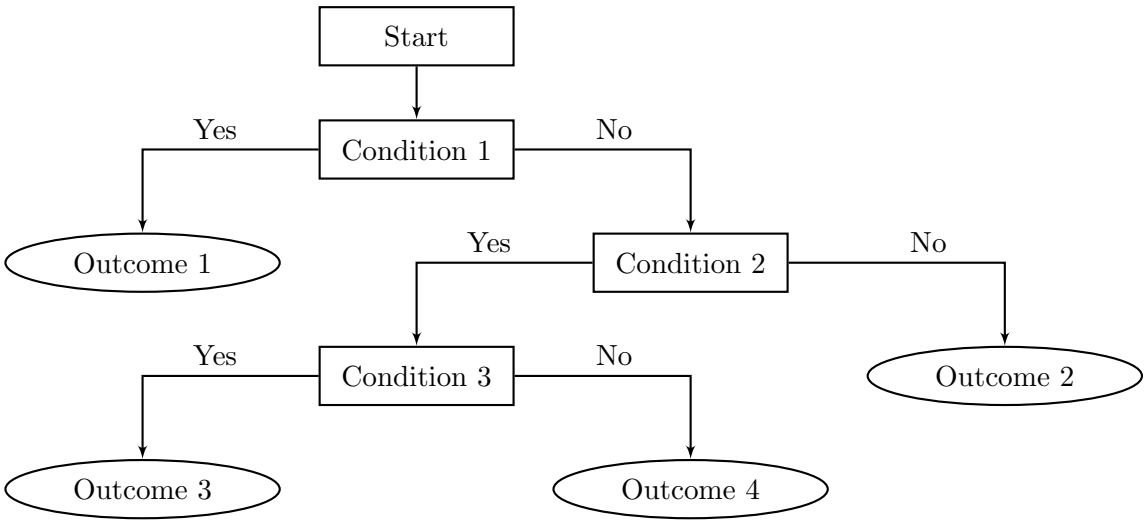


FIGURE D.1: *An example of a general decision tree.*

The decision tree regression can be implemented using the `sklearn.tree.DecisionTreeRegressor` package in Python and only requires the input dataset. All parameters have default values, which can be adjusted to improve the model performance [34].

D.2 Choosing model parameters

When creating a decision tree, the termination criteria must be specified and the model parameters are used to do this. A number of parameters are available, and different combinations of these parameters can lead to different trees.

Firstly, the maximum depth of the tree can be specified and this limits the number of splits allowed in the decision tree. The deeper a decision tree grows, the more complex the model becomes and the more likely it is to overfit to training data [29]. In the `sklearn.tree.DecisionTreeRegressor` package, this parameter is called `max_depth` and by default no value is specified, meaning that the model relies on other termination criteria [34].

The user can also specify the minimum number of data points needed to allow a new split in the tree [29]. Thus, this refers to the number of data points at a node before the split is made, or before each question in Figure D.1 is answered. According to Mantovani *et al.* [28], the ideal value for this parameter ranges between 2 and 40. Generally, a higher value can avoid overfitting to training data. This parameter is called `min_samples_split` and has a default value of 2 when using the `sklearn.tree.DecisionTreeRegressor` package [34].

Thirdly, the minimum number of data points leaving a node can be specified. Contrary to the second parameter, this refers to the number of points in each of the two branches leaving a node after a split is made. In Figure D.1, these branches are called 'Yes' and 'No'. Mantovani *et al.* [28] suggest that the ideal value for this parameter is between 1 and 20. In the `sklearn.tree.DecisionTreeRegressor` package, this parameter is called `min_sample_leaf` and has a default value of 1 [34].

Other, less popular parameters include the maximum number of variables allowed in the tree, as well as the maximum number of nodes [29, 34].

D.3 Tree validation

To test the accuracy of the model and to ensure that it is not overfitting to the data, the dataset is split into two subsets, namely a training and a test (also called holdout) set. Thus, the model is trained on the majority of the data, but a sample is kept aside to use for testing afterwards. This allows the user to compare the accuracy of the predictions made on the training and test sets to identify overfitting and have a better understanding of how accurate the model would be when making predictions on new data. If the model is much more accurate on the training data than on the test data, it can be concluded that the model is overfitting to the training data and will not perform as well with new predictions. On the other hand, if the accuracy of the training and test data are similar, there is more confidence that the model will perform well on new data [35, 39].

In Python, the `sklearn.model_selection.train_test_split` package is used to split a dataset into training and test data. To implement this, the user must select the proportion of the dataset to be used for testing [34].

APPENDIX E

Regression tree results

In this appendix, the remaining regression trees for devices ranged in all stores for at least one cluster are given in §E.1. These regression trees are built and used to estimate the ROS for seven devices according to the methodology in §5.2.2. The results of these trees are described and compared in §6.2. Section E.2 contains the predicted ROS per store for all seven devices that will be ranged in all stores for at least one cluster, and were obtained from the regression trees.

E.1 Regression trees

In Figure E.1, the regression tree for device 00321 is shown. This tree has eight decision nodes with five different variables, resulting in nine terminal nodes.

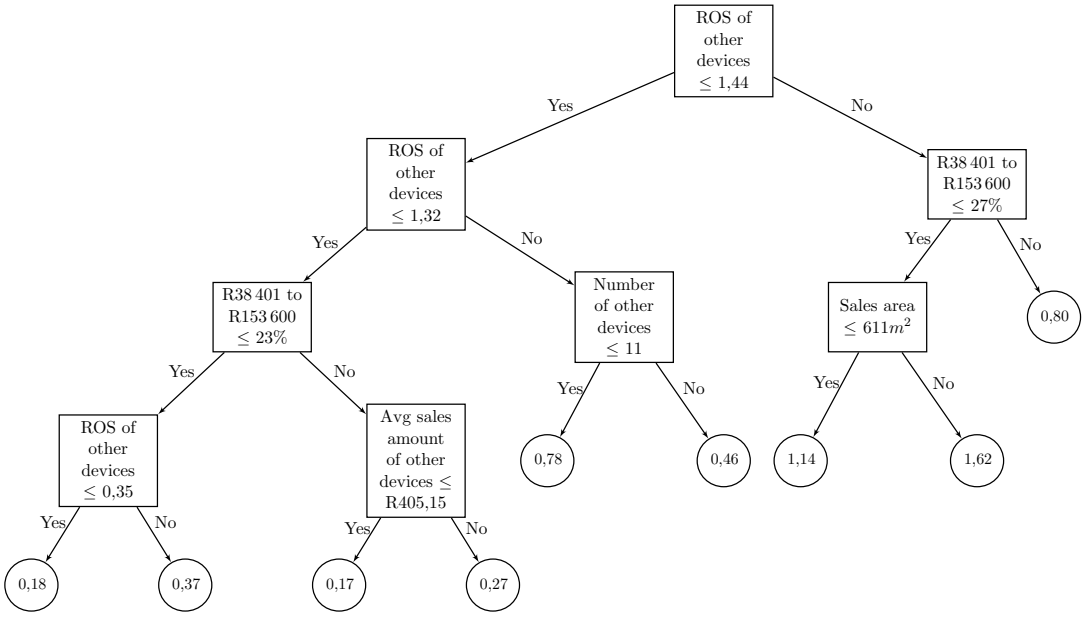


FIGURE E.1: *The regression tree for device 00321.*

The regression tree for device 00297 can be seen in Figure E.2. Four different variables are used to construct this tree, resulting in 11 decision nodes and 12 terminal nodes.

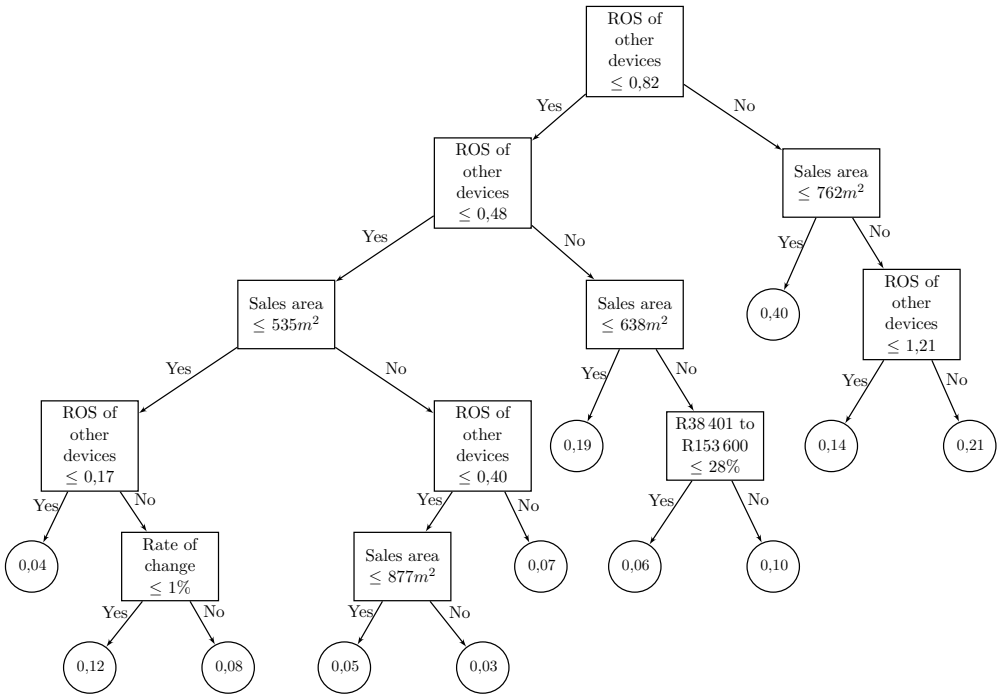


FIGURE E.2: *The regression tree for device 00297.*

In Figure E.3, the regression tree for device 00259 is displayed. This tree consists of 10 decision nodes and 11 terminal nodes. Furthermore, five different variables are used to construct the tree.

Finally, the regression tree for device 00291 is the smallest of the seven trees and can be seen in Figure E.4. This tree consists of six decision nodes and seven terminal nodes. Further, five variables are used in the tree, with the sales area being used twice.

E.2 Predicted ROS

In this section, the predicted ROS per store is provided for each of the seven devices. Table E.1 contains the predicted ROS of device 00132 for stores in clusters B1, C1 and C2.

Store	Sales area	Province	R153601 to R614400	ROS of other devices	Avg sales amount of other devices	Number of other devices	Predicted ROS, r_{ij}
B_143	2 100m^2	KwaZulu-Natal	10%	0,44	R451,54	14	0,06
B_202	1 197m^2	Free State	6%	0,26	R447,63	11	0,03
B_207	992m^2	Limpopo	7%	0,01	R425,00	2	0,03
B_297	1 127m^2	Limpopo	5%	0,06	R296,67	6	0,03
B_340	1 251m^2	Mpumalanga	2%	0,16	R476,09	9	0,06
B_402	1 895m^2	Mpumalanga	6%	0,55	R273,66	10	0,06
C_003	1 970m^2	Western Cape	21%	0,07	R573,00	6	0,03
C_007	2 115m^2	Eastern Cape	14%	0,03	R336,00	3	0,03
C_027	1 863m^2	Gauteng	18%	0,15	R325,45	6	0,03
C_035	2 040m^2	Gauteng	23%	0,13	R700,00	1	0,06
C_042	2 866m^2	Northern Cape	16%	0,34	R427,65	8	0,06
C_045	5 700m^2	Gauteng	18%	1,17	R894,74	34	0,18
C_061	1 670m^2	Western Cape	11%	0,48	R494,23	13	0,06

C_069	1 755m^2	Limpopo	11%	0,01	R350,00	1	0,03
C_081	1 772m^2	Gauteng	18%	0,01	R180,00	1	0,03
C_095	1 066m^2	Gauteng	18%	0,03	R370,00	3	0,03
C_102	1 450m^2	Western Cape	12%	0,27	R456,00	12	0,06
C_110	1 549m^2	Mpumalanga	2%	0,01	R350,00	1	0,06
C_115	1 350m^2	Gauteng	22%	0,21	R353,55	10	0,06
C_129	4 109m^2	North West	15%	1,51	R1037,58	38	0,18
C_135	1 920m^2	Eastern Cape	9%	0,05	R445,71	6	0,03
C_139	1 611m^2	Gauteng	18%	0,06	R458,89	6	0,03
C_141	1 593m^2	North West	7%	0,21	R483,23	8	0,03
C_144	2 772m^2	North West	11%	0,17	R383,60	8	0,03
C_150	1 810m^2	North West	11%	0,08	R457,50	6	0,03
C_165	2 000m^2	North West	9%	0,07	R458,00	4	0,03
C_172	1 183m^2	Western Cape	18%	0,14	R412,86	11	0,03
C_186	2 015m^2	Western Cape	21%	0,32	R431,70	14	0,06
C_189	1 540m^2	Limpopo	7%	0,02	R666,67	2	0,03
C_192	2 461m^2	Gauteng	22%	0,03	R466,67	2	0,06
C_216	2 676m^2	Gauteng	18%	0,06	R493,33	4	0,03
C_231	2 059m^2	KwaZulu-Natal	3%	0,01	R800,00	1	0,06
C_234	1 800m^2	Western Cape	10%	0,07	R450,00	4	0,03
C_239	1 395m^2	Gauteng	18%	0,05	R548,57	3	0,03
C_247	1 805m^2	Gauteng	18%	0,02	R450,00	3	0,03
C_258	5 548m^2	Western Cape	7%	1,75	R733,63	52	0,06
C_261	3 240m^2	KwaZulu-Natal	8%	0,20	R434,83	8	0,03

TABLE E.1: *The predicted ROS for device 00132.*

In Table E.2, the predicted ROS of mobile device 00141 is provided for stores in clusters A2 and B1.

Store	Sales area	ROS of other devices	Avg sales amount of other devices	Number of other devices	Avg age	Predicted ROS, r_{ij}
A_037	415m^2	0,43	R277,14	6	26,18	0,09
A_038	395m^2	0,95	R335,47	4	28,84	0,47
A_095	297m^2	0,03	R252,50	3	24,28	0,06
A_108	250m^2	0,11	R302,94	5	25,96	0,06
A_121	746m^2	0,03	R308,00	3	23,89	0,11
A_229	107m^2	0,05	R390,00	2	28,32	0,06
A_250	221m^2	0,29	R340,63	5	26,28	0,1
A_325	431m^2	0,07	R347,14	2	24,9	0,06
A_356	256m^2	0,75	R330,00	2	24,75	0,47
A_363	399m^2	0,44	R292,73	4	25,57	0,09
A_384	402m^2	0,59	R274,38	4	24,16	0,09
B_143	2 100m^2	0,44	R451,54	14	25,2	0,09
B_297	1 127m^2	0,06	R296,67	6	25,21	0,04
B_324	630m^2	1,16	R397,97	11	25,74	0,79
B_343	750m^2	0,10	R471,54	4	23,74	0,11
B_381	2 099m^2	0,24	R318,61	9	24,73	0,04

TABLE E.2: *The predicted ROS for device 00141.*

The predicted ROS of device 00321 for stores in clusters B2, C1 and C2 are given in Table E.3.

Store	Sales area	R38401 to R153600	ROS other	Avg sales amount of other devices	Number of other devices	Predicted ROS, r_{ij}
B_001	2 127m^2	27%	0,30	R483,64	14	0,27
B_002	1 384m^2	31%	0,66	R408,14	15	0,27
B_003	912m^2	31%	0,13	R433,68	10	0,27
B_005	2 661m^2	28%	0,11	R223,75	5	0,17
B_007	1 533m^2	28%	0,07	R460,00	6	0,27
B_008	968m^2	28%	0,16	R302,50	5	0,17
B_012	1 878m^2	26%	0,58	R336,51	14	0,17
B_013	2 307m^2	35%	0,29	R296,98	12	0,17
B_014	1 835m^2	26%	0,35	R384,23	14	0,17
B_016	1 456m^2	27%	0,12	R281,11	6	0,17
B_018	1 021m^2	31%	0,22	R373,13	15	0,17
B_019	2 755m^2	27%	0,14	R295,00	11	0,17
B_021	1 610m^2	31%	0,20	R380,34	10	0,17
B_022	883m^2	24%	0,15	R419,55	12	0,27
B_024	1 432m^2	28%	0,22	R233,03	10	0,17
B_025	814m^2	26%	0,32	R382,50	16	0,17
B_026	2 137m^2	35%	0,55	R379,75	15	0,17
B_027	1 324m^2	27%	0,42	R325,32	13	0,17
B_030	1 518m^2	27%	0,19	R354,29	9	0,17
B_033	1 830m^2	27%	0,33	R274,08	18	0,17
B_036	2 218m^2	28%	0,09	R376,43	7	0,17
B_039	1 874m^2	29%	0,31	R333,48	15	0,17
B_040	1 083m^2	38%	0,26	R418,21	13	0,27
B_042	2 038m^2	31%	0,09	R346,15	6	0,17
B_046	2 460m^2	31%	0,87	R410,16	14	0,27
B_056	1 680m^2	35%	0,18	R402,69	9	0,17
B_058	1 594m^2	30%	0,70	R298,93	13	0,17
B_060	2 442m^2	28%	0,27	R321,00	10	0,17
B_061	2 390m^2	27%	0,55	R336,34	15	0,17
B_063	1 847m^2	31%	0,11	R446,88	9	0,27
B_064	4 571m^2	31%	0,95	R421,14	14	0,27
B_067	1 703m^2	34%	0,43	R348,44	10	0,17
B_070	1 431m^2	29%	1,37	R321,67	16	0,46
B_072	1 922m^2	34%	0,44	R456,92	11	0,27
B_074	2 300m^2	24%	0,10	R292,67	9	0,17
B_075	2 879m^2	24%	0,06	R342,22	5	0,17
B_076	3 241m^2	29%	0,43	R464,92	10	0,27
B_078	1 809m^2	31%	0,83	R416,02	11	0,27
B_079	902m^2	29%	0,18	R357,41	9	0,17
B_080	1 249m^2	27%	0,49	R331,11	10	0,17
B_082	2 603m^2	36%	0,41	R412,13	13	0,27
B_084	1 692m^2	27%	0,16	R380,87	14	0,17
B_088	2 425m^2	29%	0,25	R457,84	9	0,27
B_091	1 500m^2	26%	0,10	R388,67	11	0,17
B_092	2 100m^2	27%	0,36	R311,32	12	0,17
B_097	1 980m^2	31%	0,97	R400,21	14	0,17
B_098	1 560m^2	36%	0,15	R211,82	8	0,17
B_101	870m^2	33%	0,14	R360,48	8	0,17
B_102	2 181m^2	35%	0,07	R284,00	7	0,17
B_105	1 547m^2	28%	0,14	R365,50	11	0,17
B_109	1 423m^2	37%	0,20	R481,38	8	0,27
B_112	2 134m^2	35%	0,03	R325,00	3	0,17
B_113	1 913m^2	27%	0,35	R252,50	15	0,17
B_115	580m^2	27%	0,01	R315,00	2	0,17
B_116	750m^2	28%	0,08	R222,50	5	0,17
B_117	1 613m^2	28%	0,11	R284,71	5	0,17
B_122	1 447m^2	31%	0,11	R293,75	5	0,17

B_129	1 954m^2	27%	0,16	R382,50	10	0,17
B_132	1 600m^2	28%	0,16	R375,22	8	0,17
B_133	729m^2	28%	0,23	R335,29	12	0,17
B_136	840m^2	35%	0,58	R422,09	16	0,27
B_145	1 731m^2	27%	0,14	R362,86	10	0,17
B_146	1 203m^2	27%	0,22	R340,94	13	0,17
B_148	1 880m^2	29%	0,39	R324,04	9	0,17
B_151	1 709m^2	29%	0,32	R346,88	13	0,17
B_153	2 100m^2	26%	0,26	R255,00	12	0,17
B_156	1 121m^2	38%	0,28	R297,14	9	0,17
B_162	967m^2	27%	0,38	R230,36	13	0,17
B_163	967m^2	27%	0,03	R330,00	2	0,17
B_164	1 667m^2	28%	0,39	R275,79	10	0,17
B_168	1 775m^2	27%	0,09	R228,46	6	0,17
B_172	2 000m^2	27%	0,28	R282,62	11	0,17
B_173	1 572m^2	31%	0,54	R329,88	14	0,17
B_174	1 950m^2	28%	0,25	R325,14	10	0,17
B_175	2 275m^2	29%	0,84	R290,97	12	0,17
B_181	1 198m^2	31%	0,21	R423,23	11	0,27
B_182	841m^2	31%	0,38	R365,18	12	0,17
B_185	1 584m^2	27%	0,18	R285,56	13	0,17
B_186	1 250m^2	36%	0,07	R340,91	7	0,17
B_189	814m^2	27%	0,32	R261,06	8	0,17
B_191	1 164m^2	27%	0,14	R370,00	13	0,17
B_195	555m^2	27%	0,18	R350,00	12	0,17
B_196	380m^2	29%	0,80	R323,22	15	0,17
B_198	1 505m^2	27%	0,67	R342,83	16	0,17
B_200	1 155m^2	27%	0,09	R327,14	6	0,17
B_204	1 029m^2	31%	0,28	R405,48	11	0,27
B_205	1 421m^2	31%	0,15	R326,82	10	0,17
B_211	756m^2	29%	0,26	R358,42	14	0,17
B_213	1 785m^2	31%	0,44	R384,31	13	0,17
B_215	1 122m^2	28%	0,29	R262,56	11	0,17
B_218	1 514m^2	27%	0,35	R350,96	17	0,17
B_219	221m^2	38%	0,18	R376,15	8	0,17
B_220	1 344m^2	31%	0,34	R346,86	13	0,17
B_221	1 276m^2	28%	0,20	R219,00	10	0,17
B_224	605m^2	28%	0,03	R278,00	4	0,17
B_226	1 488m^2	26%	0,11	R390,63	7	0,17
B_228	693m^2	45%	1,22	R434,94	16	0,27
B_229	1 276m^2	27%	0,07	R284,55	6	0,17
B_233	1 275m^2	27%	0,03	R527,50	3	0,27
B_235	857m^2	39%	0,57	R488,69	17	0,27
B_237	578m^2	31%	0,83	R414,55	13	0,27
B_238	1 051m^2	38%	0,24	R415,14	14	0,27
B_240	1 317m^2	29%	0,60	R303,60	11	0,17
B_242	1 258m^2	31%	0,19	R285,71	10	0,17
B_243	1 357m^2	31%	0,77	R419,82	12	0,27
B_244	1 199m^2	31%	0,83	R422,85	12	0,27
B_245	1 529m^2	31%	0,23	R410,00	11	0,27
B_246	1 430m^2	31%	0,41	R378,83	9	0,17
B_247	1 518m^2	31%	0,25	R243,78	11	0,17
B_248	1 216m^2	38%	0,51	R365,53	15	0,17
B_250	1 157m^2	31%	0,45	R439,55	11	0,27
B_251	1 175m^2	31%	0,70	R477,96	10	0,27
B_252	1 850m^2	31%	0,69	R433,43	14	0,27
B_255	1 487m^2	28%	0,06	R320,00	3	0,17
B_258	961m^2	28%	0,16	R232,17	7	0,17
B_259	953m^2	38%	0,22	R390,61	8	0,17
B_262	1 532m^2	27%	0,61	R333,33	14	0,17
B_264	1 251m^2	34%	1,06	R325,48	15	0,17

B_265	$980m^2$	28%	0,48	R323,94	17	0,17
B_266	$953m^2$	27%	0,24	R365,00	9	0,17
B_267	$2\,419m^2$	27%	0,27	R412,25	15	0,27
B_268	$1\,851m^2$	28%	0,26	R305,13	19	0,17
B_270	$1\,105m^2$	28%	0,22	R257,88	5	0,17
B_272	$1\,560m^2$	27%	0,57	R448,00	13	0,27
B_274	$1\,331m^2$	31%	0,04	R430,00	1	0,27
B_275	$1\,427m^2$	31%	0,16	R307,08	7	0,17
B_276	$1\,835m^2$	24%	0,27	R358,75	11	0,17
B_277	$1\,719m^2$	28%	0,14	R347,50	10	0,17
B_280	$1\,671m^2$	31%	0,75	R362,43	13	0,17
B_281	$1\,600m^2$	31%	0,62	R418,70	13	0,27
B_282	$2\,100m^2$	31%	0,66	R375,77	16	0,17
B_283	$1\,540m^2$	26%	0,07	R375,00	8	0,17
B_284	$1\,612m^2$	27%	0,39	R240,70	12	0,17
B_286	$1\,626m^2$	28%	0,30	R303,56	8	0,17
B_289	$1\,384m^2$	29%	1,48	R288,39	12	0,8
B_292	$1\,506m^2$	28%	0,42	R350,16	14	0,17
B_295	$333m^2$	32%	0,28	R406,90	10	0,27
B_305	$1\,567m^2$	27%	0,30	R323,41	8	0,17
B_308	$1\,594m^2$	28%	0,44	R262,00	16	0,17
B_309	$1\,550m^2$	28%	0,14	R318,00	11	0,17
B_315	$961m^2$	31%	0,46	R540,29	12	0,27
B_318	$1\,950m^2$	28%	0,18	R396,30	11	0,17
B_319	$1\,487m^2$	28%	0,20	R233,33	7	0,17
B_323	$1\,583m^2$	25%	0,28	R523,90	10	0,27
B_327	$1\,675m^2$	28%	0,78	R331,13	11	0,17
B_331	$1\,493m^2$	27%	0,18	R157,04	5	0,17
B_332	$1\,300m^2$	31%	1,01	R407,99	12	0,27
B_335	$1\,914m^2$	31%	0,11	R368,00	5	0,17
B_338	$768m^2$	27%	0,19	R375,36	9	0,17
B_339	$1\,396m^2$	24%	0,09	R350,00	4	0,17
B_344	$747m^2$	27%	0,38	R283,75	11	0,17
B_345	$1\,658m^2$	29%	0,51	R354,08	15	0,17
B_349	$1\,594m^2$	26%	0,14	R377,50	8	0,17
B_353	$1\,559m^2$	27%	1,28	R297,57	13	0,17
B_355	$1\,200m^2$	26%	0,34	R283,33	11	0,17
B_356	$1\,242m^2$	39%	0,39	R358,45	14	0,17
B_359	$824m^2$	36%	0,24	R358,06	11	0,17
B_360	$1\,549m^2$	31%	1,93	R308,00	16	0,8
B_363	$2\,667m^2$	31%	0,45	R339,70	14	0,17
B_364	$2\,989m^2$	31%	0,43	R406,51	11	0,27
B_365	$3\,126m^2$	31%	0,64	R403,72	13	0,17
B_366	$994m^2$	28%	0,08	R283,33	7	0,17
B_367	$626m^2$	28%	0,16	R390,00	8	0,17
B_368	$1\,400m^2$	27%	0,56	R378,07	14	0,17
B_369	$1\,746m^2$	31%	0,28	R381,67	11	0,17
B_370	$1\,158m^2$	31%	0,34	R367,45	12	0,17
B_371	$765m^2$	35%	0,30	R436,22	10	0,27
B_372	$1\,687m^2$	31%	0,20	R346,33	12	0,17
B_375	$1\,866m^2$	27%	0,20	R289,67	11	0,17
B_380	$1\,507m^2$	27%	0,05	R435,00	5	0,27
B_383	$2\,990m^2$	29%	0,74	R351,09	13	0,17
B_385	$1\,835m^2$	32%	0,18	R241,48	7	0,17
B_388	$1\,170m^2$	29%	0,45	R316,52	17	0,17
B_390	$632m^2$	27%	0,08	R463,64	5	0,27
B_393	$2\,807m^2$	27%	0,16	R484,78	9	0,27
B_394	$1\,658m^2$	26%	0,52	R209,35	16	0,17
B_396	$1\,568m^2$	27%	0,27	R365,50	7	0,17
B_397	$1\,761m^2$	28%	0,07	R321,82	7	0,17
B_401	$1\,824m^2$	27%	0,37	R265,56	11	0,17

B_403	2 862m^2	26%	0,63	R460,00	14	0,27
B_405	2 426m^2	27%	0,15	R298,18	4	0,17
B_406	4 033m^2	27%	0,27	R351,25	9	0,17
B_414	1 569m^2	32%	0,15	R307,27	8	0,17
B_415	3 781m^2	27%	1,45	R316,45	13	0,8
B_416	3 668m^2	27%	1,58	R282,69	16	0,8
B_417	2 227m^2	35%	0,19	R313,57	8	0,17
B_424	2 350m^2	28%	0,06	R342,22	4	0,17
B_425	2 456m^2	30%	0,28	R242,20	10	0,17
B_426	2 200m^2	28%	0,07	R175,00	3	0,17
B_429	1 862m^2	27%	0,09	R466,92	6	0,27
B_431	1 481m^2	31%	0,13	R425,79	8	0,27
B_432	1 115m^2	28%	0,13	R277,37	9	0,17
B_433	2 473m^2	28%	0,12	R285,56	5	0,17
B_434	3 316m^2	28%	0,24	R338,61	15	0,17
B_437	2 640m^2	27%	0,30	R341,14	14	0,17
B_438	1 973m^2	27%	0,21	R314,52	9	0,17
B_442	1 723m^2	29%	0,07	R442,73	5	0,27
B_443	759m^2	29%	0,39	R402,07	10	0,17
B_445	2 100m^2	28%	0,25	R307,03	11	0,17
B_449	1 435m^2	27%	0,45	R286,42	13	0,17
B_450	1 350m^2	28%	0,10	R288,00	8	0,17
B_452	1 248m^2	28%	1,03	R394,01	16	0,17
B_453	1 800m^2	28%	0,20	R463,00	8	0,27
B_457	1 371m^2	25%	0,18	R481,11	8	0,27
B_460	1 600m^2	28%	0,10	R311,33	9	0,17
B_461	1 550m^2	28%	0,23	R260,88	8	0,17
B_463	785m^2	31%	0,43	R351,88	14	0,17
B_464	1 907m^2	35%	0,17	R276,80	7	0,17
B_470	1 200m^2	27%	0,09	R284,29	6	0,17
B_471	839m^2	27%	0,25	R325,95	11	0,17
B_472	1 000m^2	27%	0,09	R274,29	7	0,17
B_473	1 328m^2	27%	0,64	R264,21	11	0,17
B_474	1 463m^2	28%	0,01	R275,00	2	0,17
B_478	1 723m^2	29%	0,53	R306,20	11	0,17
B_479	77m^2	31%	3,84	R559,96	49	0,8
B_480	60m^2	26%	2,24	R563,02	42	1,14
B_484	1 519m^2	27%	0,28	R256,75	9	0,17
B_485	725m^2	27%	0,14	R375,24	9	0,17
B_486	1 550m^2	28%	0,02	R220,00	2	0,17
B_487	1 106m^2	27%	0,16	R380,00	8	0,17
B_488	1 880m^2	28%	0,41	R309,84	13	0,17
B_491	1 437m^2	27%	0,17	R322,86	9	0,17
B_499	1 358m^2	31%	0,49	R362,47	13	0,17
B_501	1 435m^2	27%	2,08	R310,65	15	0,8
B_502	1 773m^2	24%	0,22	R380,31	8	0,17
B_503	2 460m^2	27%	0,46	R346,18	9	0,17
B_507	778m^2	28%	0,38	R477,86	10	0,27
B_510	850m^2	37%	0,32	R410,00	15	0,27
B_511	746m^2	36%	0,57	R363,88	16	0,17
B_513	1 420m^2	28%	0,05	R218,75	4	0,17
B_517	1 650m^2	28%	0,32	R217,45	13	0,17
B_519	1 572m^2	27%	0,24	R265,83	8	0,17
B_521	1 829m^2	28%	0,32	R432,67	7	0,27
B_522	722m^2	28%	0,06	R466,67	6	0,27
C_001	1 991m^2	27%	0,21	R322,90	6	0,17
C_003	1 970m^2	31%	0,07	R573,00	6	0,27
C_004	1 482m^2	29%	0,22	R380,63	13	0,17
C_006	2 674m^2	29%	0,60	R864,94	15	0,27
C_007	2 115m^2	26%	0,03	R336,00	3	0,17
C_008	1 176m^2	31%	0,14	R572,00	9	0,27

C_009	992m^2	24%	0,18	R508,15	6	0,27
C_010	2 620m^2	31%	0,27	R451,00	12	0,27
C_011	2 284m^2	31%	0,39	R815,96	22	0,27
C_012	1 651m^2	27%	0,09	R473,57	8	0,27
C_013	1 325m^2	27%	0,16	R270,43	7	0,17
C_015	1 797m^2	27%	0,05	R427,14	4	0,27
C_016	4 283m^2	16%	3,35	R1057,28	27	1,62
C_018	2 057m^2	26%	0,17	R434,40	10	0,27
C_019	2 050m^2	31%	0,05	R353,75	6	0,17
C_021	1 585m^2	29%	0,39	R399,14	14	0,17
C_022	1 319m^2	29%	0,28	R458,33	10	0,27
C_023	4 787m^2	28%	2,38	R874,52	43	0,8
C_024	1 102m^2	27%	0,30	R384,32	14	0,17
C_027	1 863m^2	27%	0,15	R325,45	6	0,17
C_028	2 256m^2	29%	0,24	R328,86	14	0,17
C_029	1 539m^2	35%	0,21	R413,23	10	0,27
C_030	4 919m^2	28%	2,68	R949,02	40	0,8
C_031	2 754m^2	24%	0,32	R327,23	10	0,17
C_032	1 435m^2	31%	1,03	R836,01	25	0,27
C_033	6 817m^2	28%	1,39	R1191,94	33	0,46
C_035	2 040m^2	24%	0,13	R700,00	1	0,27
C_036	2 829m^2	24%	0,11	R523,75	7	0,27
C_037	1 722m^2	27%	0,26	R371,79	9	0,17
C_038	3 117m^2	27%	0,09	R342,14	8	0,17
C_039	2 823m^2	31%	0,10	R373,33	9	0,17
C_040	3 095m^2	27%	0,04	R443,33	4	0,27
C_041	5 745m^2	31%	1,68	R1173,59	40	0,8
C_042	2 866m^2	31%	0,34	R427,65	8	0,27
C_045	5 700m^2	28%	1,17	R894,74	34	0,27
C_046	1 520m^2	16%	0,31	R323,00	8	0,18
C_047	1 900m^2	28%	0,11	R367,50	7	0,17
C_049	2 188m^2	27%	0,26	R472,05	8	0,27
C_050	1 800m^2	27%	0,21	R387,42	9	0,17
C_051	2 340m^2	25%	0,22	R403,94	11	0,17
C_052	1 456m^2	26%	0,20	R422,76	10	0,27
C_053	6 175m^2	31%	1,47	R1015,55	33	0,8
C_054	1 984m^2	27%	0,39	R398,77	13	0,17
C_055	2 082m^2	28%	1,70	R766,11	24	0,8
C_056	2 300m^2	18%	0,28	R364,88	9	0,18
C_057	1 539m^2	36%	0,22	R399,39	12	0,17
C_058	3 011m^2	29%	0,15	R491,36	8	0,27
C_059	2 124m^2	27%	0,08	R441,67	8	0,27
C_060	2 631m^2	27%	0,36	R373,33	14	0,17
C_061	1 670m^2	38%	0,48	R494,23	13	0,27
C_062	1 756m^2	27%	0,17	R271,20	8	0,17
C_063	1 700m^2	27%	0,07	R308,18	6	0,17
C_065	2 970m^2	27%	0,15	R659,55	12	0,27
C_067	2 017m^2	24%	0,03	R560,00	3	0,27
C_068	1 517m^2	29%	0,41	R374,26	14	0,17
C_069	1 755m^2	30%	0,01	R350,00	1	0,17
C_070	1 583m^2	28%	0,30	R464,77	13	0,27
C_071	1 974m^2	34%	0,26	R393,16	11	0,17
C_072	1 578m^2	24%	0,36	R455,37	11	0,27
C_074	6 028m^2	27%	3,05	R976,61	29	0,8
C_075	1 478m^2	29%	0,24	R370,00	11	0,17
C_076	1 698m^2	31%	0,50	R450,27	10	0,27
C_077	1 653m^2	24%	0,29	R376,74	14	0,17
C_078	1 870m^2	27%	0,16	R391,25	10	0,17
C_079	5 200m^2	28%	1,44	R1105,21	40	0,46
C_080	2 350m^2	27%	0,36	R342,22	6	0,17
C_081	1 772m^2	28%	0,01	R180,00	1	0,17

C_082	1 863m^2	28%	0,09	R275,38	6	0,17
C_083	2 094m^2	28%	0,28	R410,30	10	0,27
C_084	1 697m^2	27%	0,44	R372,77	10	0,17
C_085	1 912m^2	27%	0,20	R327,59	6	0,17
C_086	5 923m^2	27%	1,33	R1004,16	32	0,46
C_087	1 944m^2	31%	0,05	R324,29	5	0,17
C_088	1 750m^2	37%	0,28	R495,71	13	0,27
C_089	8 855m^2	31%	2,16	R952,19	42	0,8
C_090	1 588m^2	27%	0,16	R370,00	10	0,17
C_091	5 561m^2	29%	0,97	R960,35	32	0,27
C_092	3 400m^2	27%	0,05	R325,71	4	0,17
C_093	1 301m^2	28%	0,17	R344,40	9	0,17
C_094	839m^2	32%	0,12	R412,78	7	0,27
C_095	1 066m^2	27%	0,03	R370,00	3	0,17
C_096	1 143m^2	27%	0,43	R357,46	11	0,17
C_097	1 232m^2	27%	0,39	R367,93	11	0,17
C_098	2 044m^2	25%	0,05	R474,29	4	0,27
C_099	2 217m^2	29%	0,23	R580,88	11	0,27
C_100	2 195m^2	29%	0,17	R396,40	8	0,17
C_101	2 866m^2	31%	0,32	R676,25	17	0,27
C_103	2 148m^2	31%	0,16	R378,26	10	0,17
C_104	1 204m^2	28%	0,38	R307,68	8	0,17
C_105	1 109m^2	26%	0,14	R324,76	8	0,17
C_106	1 183m^2	38%	0,24	R410,86	13	0,27
C_107	1 473m^2	27%	0,07	R280,00	5	0,17
C_108	1 615m^2	28%	0,09	R365,71	6	0,17
C_109	2 298m^2	27%	0,90	R145,11	6	0,17
C_110	1 549m^2	16%	0,01	R350,00	1	0,18
C_111	2 490m^2	38%	0,20	R370,00	11	0,17
C_112	1 420m^2	31%	0,13	R462,63	11	0,27
C_113	1 241m^2	28%	0,55	R467,90	10	0,27
C_114	2 210m^2	31%	0,24	R364,57	9	0,17
C_115	1 350m^2	27%	0,21	R353,55	10	0,17
C_116	2 513m^2	29%	2,42	R943,04	29	0,8
C_117	1 226m^2	27%	0,46	R338,53	11	0,17
C_119	1 301m^2	27%	0,04	R238,33	3	0,17
C_120	1 928m^2	27%	0,23	R461,18	11	0,27
C_122	1 578m^2	28%	0,05	R264,29	3	0,17
C_123	1 267m^2	29%	0,17	R360,40	7	0,17
C_125	1 588m^2	16%	0,28	R346,00	7	0,18
C_127	2 911m^2	28%	0,07	R429,00	7	0,27
C_129	4 109m^2	35%	1,51	R1037,58	38	0,8
C_130	1 452m^2	23%	0,27	R309,50	11	0,17
C_131	2 574m^2	27%	0,20	R409,31	9	0,27
C_132	2 353m^2	26%	0,07	R497,27	6	0,27
C_133	1 122m^2	30%	0,20	R493,67	13	0,27
C_134	3 064m^2	31%	0,70	R644,37	26	0,27
C_135	1 920m^2	32%	0,05	R445,71	6	0,27
C_137	1 717m^2	32%	0,01	R205,00	2	0,17
C_138	1 589m^2	29%	0,13	R438,95	11	0,27
C_139	1 611m^2	28%	0,06	R458,89	6	0,27
C_140	2 121m^2	31%	0,48	R556,06	13	0,27
C_141	1 593m^2	24%	0,21	R483,23	8	0,27
C_142	2 055m^2	35%	0,11	R479,41	9	0,27
C_143	2 125m^2	27%	0,37	R347,82	9	0,17
C_144	2 772m^2	26%	0,17	R383,60	8	0,17
C_145	1 785m^2	28%	0,29	R226,28	8	0,17
C_147	5 478m^2	29%	1,45	R1210,37	33	0,8
C_148	915m^2	38%	0,16	R403,48	9	0,17
C_150	1 810m^2	28%	0,08	R457,50	6	0,27
C_151	1 606m^2	28%	0,09	R354,29	8	0,17

C_152	2 000m^2	28%	0,16	R373,04	10	0,17
C_153	1 600m^2	27%	0,19	R422,86	10	0,27
C_154	2 029m^2	28%	0,34	R326,08	8	0,17
C_156	1 885m^2	27%	0,21	R417,10	11	0,27
C_157	2 229m^2	35%	0,38	R417,68	11	0,27
C_158	1 872m^2	39%	0,44	R642,31	15	0,27
C_159	2 900m^2	24%	0,14	R424,00	5	0,27
C_161	1 200m^2	24%	0,09	R322,14	5	0,17
C_162	2 200m^2	27%	0,19	R363,57	8	0,17
C_164	2 145m^2	27%	0,25	R424,59	9	0,27
C_165	2 000m^2	32%	0,07	R458,00	4	0,27
C_166	1 600m^2	27%	0,09	R349,29	6	0,17
C_167	1 703m^2	28%	0,40	R321,86	9	0,17
C_168	1 938m^2	27%	0,26	R418,95	11	0,27
C_169	2 204m^2	31%	0,26	R400,77	13	0,17
C_170	3 000m^2	27%	0,28	R320,98	13	0,17
C_172	1 183m^2	38%	0,14	R412,86	11	0,27
C_174	1 300m^2	27%	0,08	R382,50	7	0,17
C_175	5 473m^2	24%	1,97	R1119,55	31	1,62
C_178	1 895m^2	28%	0,23	R547,06	8	0,27
C_179	2 200m^2	35%	0,14	R492,38	11	0,27
C_180	1 718m^2	30%	0,23	R555,59	10	0,27
C_182	1 192m^2	36%	0,17	R278,40	8	0,17
C_183	1 103m^2	39%	0,69	R491,08	16	0,27
C_184	2 346m^2	31%	0,40	R451,69	11	0,27
C_185	2 423m^2	24%	0,12	R476,67	6	0,27
C_186	2 015m^2	31%	0,32	R431,70	14	0,27
C_187	5 178m^2	31%	1,18	R1082,82	29	0,27
C_189	1 540m^2	20%	0,02	R666,67	2	0,18
C_190	2 020m^2	27%	0,04	R518,33	5	0,27
C_191	1 900m^2	24%	0,01	R220,00	2	0,17
C_192	2 461m^2	27%	0,03	R466,67	2	0,27
C_193	1 794m^2	30%	0,14	R316,67	7	0,17
C_194	1 440m^2	37%	0,36	R459,62	14	0,27
C_195	1 700m^2	31%	0,31	R396,52	11	0,17
C_196	5 763m^2	31%	1,24	R857,83	35	0,27
C_197	1 944m^2	35%	0,26	R411,32	12	0,27
C_198	5 528m^2	31%	1,06	R926,24	28	0,27
C_199	1 619m^2	31%	0,28	R370,98	11	0,17
C_200	2 571m^2	29%	0,19	R352,14	11	0,17
C_201	1 986m^2	29%	0,15	R437,27	8	0,27
C_204	1 018m^2	29%	0,50	R325,95	7	0,17
C_205	1 944m^2	25%	1,69	R1302,40	36	1,62
C_206	1 396m^2	27%	0,44	R432,00	12	0,27
C_207	1 864m^2	27%	0,16	R416,09	6	0,27
C_208	4 651m^2	29%	1,93	R964,14	41	0,8
C_210	1 655m^2	34%	0,18	R391,54	10	0,17
C_211	1 426m^2	27%	0,09	R242,86	5	0,17
C_212	1 615m^2	27%	0,18	R318,46	8	0,17
C_213	6 494m^2	28%	1,13	R1021,20	46	0,27
C_214	3 067m^2	27%	0,18	R305,38	8	0,17
C_215	2 500m^2	25%	0,20	R826,55	12	0,27
C_216	2 676m^2	27%	0,06	R493,33	4	0,27
C_217	5 317m^2	27%	2,21	R995,47	46	0,8
C_218	5 690m^2	27%	1,26	R1030,16	48	0,27
C_219	5 807m^2	27%	1,76	R944,67	35	0,8
C_220	6 757m^2	28%	1,18	R1031,03	30	0,27
C_222	7 546m^2	27%	1,20	R1025,34	37	0,27
C_223	7 438m^2	27%	1,25	R1115,78	29	0,27
C_224	5 813m^2	28%	1,69	R1153,20	36	0,8
C_225	4 950m^2	28%	2,06	R1122,03	36	0,8

C_226	7 502m^2	28%	1,39	R1091,55	38	0,46
C_227	5 915m^2	27%	1,79	R1136,42	43	0,8
C_228	5 202m^2	26%	0,40	R1126,44	27	0,27
C_229	1 774m^2	28%	0,26	R366,05	11	0,17
C_231	2 059m^2	21%	0,01	R800,00	1	0,18
C_232	2 106m^2	31%	0,53	R367,82	13	0,17
C_233	1 872m^2	31%	0,27	R491,28	12	0,27
C_234	1 800m^2	26%	0,07	R450,00	4	0,27
C_235	1 629m^2	39%	0,41	R492,79	9	0,27
C_237	2 675m^2	31%	0,10	R362,00	10	0,17
C_239	1 395m^2	27%	0,05	R548,57	3	0,27
C_242	1 100m^2	28%	0,05	R462,86	4	0,27
C_243	1 920m^2	31%	0,51	R406,40	15	0,27
C_244	1 600m^2	27%	0,09	R403,57	8	0,17
C_245	2 139m^2	27%	0,15	R439,55	7	0,27
C_246	1 795m^2	27%	0,30	R339,11	9	0,17
C_247	1 805m^2	28%	0,02	R450,00	3	0,27
C_249	1 700m^2	18%	0,14	R275,71	4	0,18
C_251	1 874m^2	27%	0,04	R530,00	2	0,27
C_252	1 500m^2	32%	0,44	R260,00	9	0,17
C_255	4 826m^2	27%	1,28	R1162,32	37	0,27
C_257	7 235m^2	27%	2,23	R810,36	39	0,8
C_258	5 548m^2	23%	1,75	R733,63	52	1,62
C_259	4 700m^2	27%	1,56	R1006,59	34	0,8
C_261	3 240m^2	22%	0,20	R434,83	8	0,18
C_263	1 600m^2	27%	0,01	R315,00	2	0,17

TABLE E.3: *The predicted ROS for device 00321.*

Table E.4 contains the predicted ROS of mobile device 00297 for stores in clusters A2 and B2.

Store	Sales area	R38401 to R153600	ROS other	Rate of change	Predicted ROS, r_{ij}
A_038	395m^2	28%	0,95	0%	0,4
A_058	501m^2	20%	0,06	0%	0,04
A_095	297m^2	14%	0,03	1%	0,04
A_229	107m^2	28%	0,05	0%	0,04
A_325	431m^2	26%	0,07	0%	0,04
A_356	256m^2	13%	0,75	0%	0,19
B_003	912m^2	31%	0,13	1%	0,03
B_008	968m^2	28%	0,16	2%	0,03
B_039	1 874m^2	29%	0,31	0%	0,03
B_053	1 257m^2	31%	0,55	1%	0,1
B_088	2 425m^2	29%	0,25	0%	0,03
B_091	1 500m^2	26%	0,10	2%	0,03
B_101	870m^2	33%	0,14	5%	0,05
B_102	2 181m^2	35%	0,07	4%	0,03
B_112	2 134m^2	35%	0,03	4%	0,03
B_115	580m^2	27%	0,01	3%	0,05
B_116	750m^2	28%	0,08	1%	0,05
B_117	1 613m^2	28%	0,11	0%	0,03
B_132	1 600m^2	28%	0,16	2%	0,03
B_163	967m^2	27%	0,03	3%	0,03
B_175	2 275m^2	29%	0,84	0%	0,14
B_185	1 584m^2	27%	0,18	3%	0,03
B_196	380m^2	29%	0,80	0%	0,19
B_221	1 276m^2	28%	0,20	2%	0,03
B_233	1 275m^2	27%	0,03	3%	0,03
B_258	961m^2	28%	0,16	2%	0,03
B_260	1 607m^2	27%	0,39	4%	0,03

B_266	$953m^2$	27%	0,24	0%	0,03
B_270	$1\,105m^2$	28%	0,22	2%	0,03
B_274	$1\,331m^2$	31%	0,04	1%	0,03
B_295	$333m^2$	32%	0,28	6%	0,08
B_308	$1\,594m^2$	28%	0,44	2%	0,07
B_316	$1\,764m^2$	28%	0,08	2%	0,03
B_331	$1\,493m^2$	27%	0,18	3%	0,03
B_339	$1\,396m^2$	24%	0,09	3%	0,03
B_347	$1\,119m^2$	24%	0,34	1%	0,03
B_355	$1\,200m^2$	26%	0,34	2%	0,03
B_366	$994m^2$	28%	0,08	2%	0,03
B_375	$1\,866m^2$	27%	0,20	0%	0,03
B_390	$632m^2$	27%	0,08	3%	0,05
B_396	$1\,568m^2$	27%	0,27	4%	0,03
B_398	$681m^2$	24%	0,43	1%	0,07
B_401	$1\,824m^2$	27%	0,37	4%	0,03
B_417	$2\,227m^2$	35%	0,19	4%	0,03
B_424	$2\,350m^2$	28%	0,06	2%	0,03
B_425	$2\,456m^2$	30%	0,28	3%	0,03
B_426	$2\,200m^2$	28%	0,07	2%	0,03
B_429	$1\,862m^2$	27%	0,09	3%	0,03
B_433	$2\,473m^2$	28%	0,12	2%	0,03
B_442	$1\,723m^2$	29%	0,07	0%	0,03
B_457	$1\,371m^2$	25%	0,18	0%	0,03
B_460	$1\,600m^2$	28%	0,10	2%	0,03
B_464	$1\,907m^2$	35%	0,17	4%	0,03
B_470	$1\,200m^2$	27%	0,09	3%	0,03
B_484	$1\,519m^2$	27%	0,28	3%	0,03
B_485	$725m^2$	27%	0,14	2%	0,05
B_486	$1\,550m^2$	28%	0,02	2%	0,03
B_513	$1\,420m^2$	28%	0,05	2%	0,03

TABLE E.4: *The predicted ROS for device 00297.*

In Table E.5, the predicted ROS of device 00298 are provided for stores in cluster C2.

Store	Sales area	R38401 to R153600	Landlines percentage	Avg sales amount of other devices	Number of other devices	Predicted ROS, r_{ij}
C_001	$1\,991m^2$	27%	6,%	R322,90	6	0,17
C_003	$1\,970m^2$	31%	26%	R573,00	6	0,12
C_004	$1\,482m^2$	29%	22%	R380,63	13	0,12
C_007	$2\,115m^2$	26%	16%	R336,00	3	0,21
C_008	$1\,176m^2$	31%	26%	R572,00	9	0,12
C_009	$992m^2$	24%	9%	R508,15	6	0,21
C_010	$2\,620m^2$	31%	26%	R451,00	12	0,12
C_011	$2\,284m^2$	31%	26%	R815,96	22	0,12
C_014	$1\,451m^2$	26%	10%	R435,80	15	0,21
C_015	$1\,797m^2$	27%	14%	R427,14	4	0,17
C_017	$1\,565m^2$	27%	14%	R304,39	9	0,17
C_018	$2\,057m^2$	26%	16%	R434,40	10	0,21
C_019	$2\,050m^2$	31%	26%	R353,75	6	0,12
C_020	$1\,995m^2$	27%	14%	R268,29	11	0,17
C_021	$1\,585m^2$	29%	22%	R399,14	14	0,12
C_022	$1\,319m^2$	29%	22%	R458,33	10	0,12
C_024	$1\,102m^2$	27%	15%	R384,32	14	0,17
C_026	$2\,140m^2$	36%	17%	R424,42	13	0,08
C_027	$1\,863m^2$	27%	15%	R325,45	6	0,17
C_028	$2\,256m^2$	29%	22%	R328,86	14	0,12

C_029	1 539m^2	35%	19%	R413,23	10	0,12
C_034	2 650m^2	27%	6%	R638,90	24	0,17
C_035	2 040m^2	24%	18%	R700,00	1	0,12
C_036	2 829m^2	24%	9%	R523,75	7	0,21
C_037	1 722m^2	27%	15%	R371,79	9	0,17
C_038	3 117m^2	27%	15%	R342,14	8	0,17
C_039	2 823m^2	31%	26%	R373,33	9	0,12
C_042	2 866m^2	31%	14%	R427,65	8	0,08
C_044	2 042m^2	24%	9%	R367,65	17	0,21
C_049	2 188m^2	27%	14%	R472,05	8	0,17
C_050	1 800m^2	27%	15%	R387,42	9	0,17
C_052	1 456m^2	26%	16%	R422,76	10	0,21
C_054	1 984m^2	27%	6%	R398,77	13	0,17
C_057	1 539m^2	36%	12%	R399,39	12	0,08
C_058	3 011m^2	29%	22%	R491,36	8	0,12
C_061	1 670m^2	38%	16%	R494,23	13	0,08
C_062	1 756m^2	27%	14%	R271,20	8	0,17
C_063	1 700m^2	27%	14%	R308,18	6	0,17
C_065	2 970m^2	27%	14%	R659,55	12	0,17
C_068	1 517m^2	29%	22%	R374,26	14	0,12
C_070	1 583m^2	28%	8%	R464,77	13	0,17
C_071	1 974m^2	34%	10%	R393,16	11	0,08
C_072	1 578m^2	24%	4%	R455,37	11	0,21
C_075	1 478m^2	29%	22%	R370,00	11	0,12
C_076	1 698m^2	31%	26%	R450,27	10	0,12
C_077	1 653m^2	24%	9%	R376,74	14	0,21
C_078	1 870m^2	27%	15%	R391,25	10	0,17
C_080	2 350m^2	27%	15%	R342,22	6	0,17
C_081	1 772m^2	28%	12%	R180,00	1	0,17
C_083	2 094m^2	28%	12%	R410,30	10	0,17
C_084	1 697m^2	27%	15%	R372,77	10	0,17
C_085	1 912m^2	27%	15%	R327,59	6	0,17
C_087	1 944m^2	31%	26%	R324,29	5	0,12
C_088	1 750m^2	37%	15%	R495,71	13	0,08
C_090	1 588m^2	27%	15%	R370,00	10	0,17
C_092	3 400m^2	27%	15%	R325,71	4	0,17
C_093	1 301m^2	28%	12%	R344,40	9	0,17
C_094	839m^2	32%	12%	R412,78	7	0,08
C_095	1 066m^2	27%	15%	R370,00	3	0,17
C_096	1 143m^2	27%	15%	R357,46	11	0,17
C_097	1 232m^2	27%	15%	R367,93	11	0,17
C_100	2 195m^2	29%	22%	R396,40	8	0,12
C_101	2 866m^2	31%	26%	R676,25	17	0,12
C_103	2 148m^2	31%	26%	R378,26	10	0,12
C_104	1 204m^2	28%	8%	R307,68	8	0,17
C_105	1 109m^2	26%	16%	R324,76	8	0,21
C_106	1 183m^2	38%	19%	R410,86	13	0,12
C_107	1 473m^2	27%	15%	R280,00	5	0,17
C_108	1 615m^2	28%	9%	R365,71	6	0,17
C_111	2 490m^2	38%	15%	R370,00	11	0,08
C_112	1 420m^2	31%	26%	R462,63	11	0,12
C_113	1 241m^2	28%	8%	R467,90	10	0,17
C_114	2 210m^2	31%	14%	R364,57	9	0,08
C_115	1 350m^2	27%	14%	R353,55	10	0,17
C_117	1 226m^2	27%	15%	R338,53	11	0,17
C_119	1 301m^2	27%	15%	R238,33	3	0,17
C_120	1 928m^2	27%	14%	R461,18	11	0,17
C_121	1 639m^2	31%	26%	R465,71	8	0,12
C_122	1 578m^2	28%	12%	R264,29	3	0,17
C_123	1 267m^2	29%	21%	R360,40	7	0,12
C_126	1 677m^2	30%	5%	R361,30	13	0,08

C_127	2 911m^2	28%	10%	R429,00	7	0,17
C_128	1 761m^2	29%	4%	R409,68	9	0,08
C_132	2 353m^2	26%	16%	R497,27	6	0,21
C_133	1 122m^2	30%	11%	R493,67	13	0,08
C_134	3 064m^2	31%	26%	R644,37	26	0,12
C_135	1 920m^2	32%	12%	R445,71	6	0,08
C_137	1 717m^2	32%	4%	R205,00	2	0,08
C_138	1 589m^2	29%	22%	R438,95	11	0,12
C_139	1 611m^2	28%	12%	R458,89	6	0,17
C_140	2 121m^2	31%	26%	R556,06	13	0,12
C_141	1 593m^2	24%	5%	R483,23	8	0,21
C_144	2 772m^2	26%	3%	R383,60	8	0,21
C_145	1 785m^2	28%	12%	R226,28	8	0,17
C_148	915m^2	38%	19%	R403,48	9	0,12
C_150	1 810m^2	28%	9%	R457,50	6	0,17
C_151	1 606m^2	28%	12%	R354,29	8	0,17
C_152	2 000m^2	28%	9%	R373,04	10	0,17
C_153	1 600m^2	27%	15%	R422,86	10	0,17
C_154	2 029m^2	28%	12%	R326,08	8	0,17
C_156	1 885m^2	27%	7%	R417,10	11	0,17
C_157	2 229m^2	35%	18%	R417,68	11	0,12
C_158	1 872m^2	39%	20%	R642,31	15	0,12
C_161	1 200m^2	24%	18%	R322,14	5	0,12
C_162	2 200m^2	27%	15%	R363,57	8	0,17
C_163	1 696m^2	27%	14%	R355,97	17	0,17
C_164	2 145m^2	27%	15%	R424,59	9	0,17
C_165	2 000m^2	32%	4%	R458,00	4	0,08
C_166	1 600m^2	27%	15%	R349,29	6	0,17
C_167	1 703m^2	28%	12%	R321,86	9	0,17
C_168	1 938m^2	27%	14%	R418,95	11	0,17
C_169	2 204m^2	31%	24%	R400,77	13	0,12
C_170	3 000m^2	27%	6%	R320,98	13	0,17
C_172	1 183m^2	38%	19%	R412,86	11	0,12
C_173	1 215m^2	27%	14%	R405,79	13	0,17
C_174	1 300m^2	27%	15%	R382,50	7	0,17
C_178	1 895m^2	28%	12%	R547,06	8	0,17
C_179	2 200m^2	35%	18%	R492,38	11	0,12
C_180	1 718m^2	30%	9%	R555,59	10	0,08
C_182	1 192m^2	36%	7%	R278,40	8	0,08
C_183	1 103m^2	39%	19%	R491,08	16	0,12
C_184	2 346m^2	31%	23%	R451,69	11	0,12
C_185	2 423m^2	24%	9%	R476,67	6	0,21
C_186	2 015m^2	31%	26%	R431,70	14	0,12
C_190	2 020m^2	27%	15%	R518,33	5	0,17
C_192	2 461m^2	27%	14%	R466,67	2	0,17
C_193	1 794m^2	30%	9%	R316,67	7	0,08
C_194	1 440m^2	37%	19%	R459,62	14	0,12
C_195	1 700m^2	31%	23%	R396,52	11	0,12
C_197	1 944m^2	35%	19%	R411,32	12	0,12
C_199	1 619m^2	31%	26%	R370,98	11	0,12
C_202	2 282m^2	29%	22%	R388,45	15	0,12
C_204	1 018m^2	29%	22%	R325,95	7	0,12
C_206	1 396m^2	27%	15%	R432,00	12	0,17
C_210	1 655m^2	34%	10%	R391,54	10	0,08
C_211	1 426m^2	27%	14%	R242,86	5	0,17
C_212	1 615m^2	27%	14%	R318,46	8	0,17
C_214	3 067m^2	27%	14%	R305,38	8	0,17
C_216	2 676m^2	27%	15%	R493,33	4	0,17
C_229	1 774m^2	28%	12%	R366,05	11	0,17
C_232	2 106m^2	31%	24%	R367,82	13	0,12
C_233	1 872m^2	31%	26%	R491,28	12	0,12

C_234	1 800m^2	26%	16%	R450,00	4	0,21
C_235	1 629m^2	39%	14%	R492,79	9	0,08
C_236	1 935m^2	21%	7%	R440,97	14	0,21
C_237	2 675m^2	31%	26%	R362,00	10	0,12
C_238	1 918m^2	31%	26%	R458,13	9	0,12
C_239	1 395m^2	27%	15%	R548,57	3	0,17
C_242	1 100m^2	28%	10%	R462,86	4	0,17
C_243	1 920m^2	31%	26%	R406,40	15	0,12
C_244	1 600m^2	27%	15%	R403,57	8	0,17
C_245	2 139m^2	27%	15%	R439,55	7	0,17
C_247	1 805m^2	28%	12%	R450,00	3	0,17
C_251	1 874m^2	27%	15%	R530,00	2	0,17
C_252	1 500m^2	32%	4%	R260,00	9	0,08
C_264	1 912m^2	31%	26%	R354,29	12	0,12

TABLE E.5: *The predicted ROS for device 00298.*

The predicted ROS of device 00259 for stores in clusters B2, C1 and C2 are given in Table E.6.

Store	Sales area	R0 to R38400	Landlines percentage	ROS other	Avg sales amount of other devices	Predicted ROS, r_{ij}
B_005	2 661m^2	48%	9%	0,11	R223,75	0,02
B_006	1 014m^2	44%	15%	1,01	R340,40	0,11
B_013	2 307m^2	46%	3%	0,29	R296,98	0,04
B_021	1 610m^2	49%	14%	0,20	R380,34	0,03
B_030	1 518m^2	41%	14%	0,19	R354,29	0,03
B_036	2 218m^2	55%	9%	0,09	R376,43	0,02
B_043	1 079m^2	41%	14%	1,35	R317,00	0,11
B_061	2 390m^2	41%	14%	0,55	R336,34	0,11
B_067	1 703m^2	44%	5%	0,43	R348,44	0,04
B_072	1 922m^2	50%	10%	0,44	R456,92	0,07
B_074	2 300m^2	58%	9%	0,10	R292,67	0,02
B_075	2 879m^2	46%	18%	0,06	R342,22	0,02
B_082	2 603m^2	48%	8%	0,41	R412,13	0,07
B_084	1 692m^2	44%	15%	0,16	R380,87	0,02
B_091	1 500m^2	56%	16%	0,10	R388,67	0,02
B_098	1 560m^2	48%	8%	0,15	R211,82	0,02
B_099	650m^2	55%	9%	0,10	R357,33	0,02
B_101	870m^2	38%	6%	0,14	R360,48	0,02
B_102	2 181m^2	46%	3%	0,07	R284,00	0,02
B_109	1 423m^2	46%	6%	0,20	R481,38	0,03
B_112	2 134m^2	46%	3%	0,03	R325,00	0,02
B_113	1 913m^2	44%	15%	0,35	R252,50	0,04
B_115	580m^2	44%	15%	0,01	R315,00	0,02
B_116	750m^2	55%	9%	0,08	R222,50	0,02
B_117	1 613m^2	54%	8%	0,11	R284,71	0,02
B_122	1 447m^2	42%	26%	0,11	R293,75	0,02
B_129	1 954m^2	44%	15%	0,16	R382,50	0,03
B_136	840m^2	47%	18%	0,58	R422,09	0,11
B_146	1 203m^2	44%	15%	0,22	R340,94	0,03
B_163	967m^2	44%	15%	0,03	R330,00	0,02
B_168	1 775m^2	44%	15%	0,09	R228,46	0,02
B_177	1 917m^2	41%	14%	0,80	R339,15	0,11
B_185	1 584m^2	44%	15%	0,18	R285,56	0,03
B_186	1 250m^2	48%	8%	0,07	R340,91	0,02
B_189	814m^2	44%	15%	0,32	R261,06	0,04
B_200	1 155m^2	41%	14%	0,09	R327,14	0,02
B_205	1 421m^2	42%	26%	0,15	R326,82	0,02

B_206	1 309m^2	42%	26%	0,34	R432,80	0,07
B_215	1 122m^2	47%	12%	0,29	R262,56	0,04
B_218	1 514m^2	44%	15%	0,35	R350,96	0,04
B_224	605m^2	47%	12%	0,03	R278,00	0,02
B_229	1 276m^2	44%	15%	0,07	R284,55	0,02
B_233	1 275m^2	44%	15%	0,03	R527,50	0,02
B_241	987m^2	54%	8%	0,18	R311,85	0,03
B_258	961m^2	47%	12%	0,16	R232,17	0,02
B_260	1 607m^2	41%	14%	0,39	R479,66	0,07
B_266	953m^2	60%	6%	0,24	R365,00	0,03
B_270	1 105m^2	47%	12%	0,22	R257,88	0,03
B_273	1 125m^2	44%	15%	0,14	R332,38	0,02
B_293	1 551m^2	44%	15%	0,32	R385,96	0,04
B_299	1 912m^2	41%	14%	0,47	R367,39	0,04
B_305	1 567m^2	41%	14%	0,30	R323,41	0,04
B_308	1 594m^2	47%	12%	0,44	R262,00	0,04
B_316	1 764m^2	47%	12%	0,08	R388,33	0,02
B_319	1 487m^2	55%	9%	0,20	R233,33	0,03
B_328	1 600m^2	44%	15%	0,22	R195,15	0,03
B_331	1 493m^2	44%	15%	0,18	R157,04	0,03
B_335	1 914m^2	42%	26%	0,11	R368,00	0,02
B_339	1 396m^2	46%	18%	0,09	R350,00	0,02
B_344	747m^2	60%	6%	0,38	R283,75	0,04
B_351	1 337m^2	58%	9%	0,12	R270,56	0,02
B_354	1 999m^2	58%	11%	0,29	R348,14	0,04
B_367	626m^2	47%	12%	0,16	R390,00	0,02
B_372	1 687m^2	42%	26%	0,20	R346,33	0,03
B_380	1 507m^2	44%	15%	0,05	R435,00	0,02
B_385	1 835m^2	55%	4%	0,18	R241,48	0,03
B_386	1 558m^2	50%	22%	1,31	R305,05	0,11
B_390	632m^2	44%	15%	0,08	R463,64	0,02
B_396	1 568m^2	41%	14%	0,27	R365,50	0,03
B_405	2 426m^2	41%	14%	0,15	R298,18	0,02
B_406	4 033m^2	41%	14%	0,27	R351,25	0,03
B_414	1 569m^2	55%	4%	0,15	R307,27	0,02
B_415	3 781m^2	41%	14%	1,45	R316,45	0,11
B_417	2 227m^2	46%	3%	0,19	R313,57	0,03
B_418	3 325m^2	41%	14%	0,64	R354,15	0,11
B_425	2 456m^2	54%	5%	0,28	R242,20	0,03
B_426	2 200m^2	48%	9%	0,07	R175,00	0,02
B_432	1 115m^2	47%	12%	0,13	R277,37	0,02
B_433	2 473m^2	47%	12%	0,12	R285,56	0,02
B_434	3 316m^2	55%	9%	0,24	R338,61	0,03
B_437	2 640m^2	44%	15%	0,30	R341,14	0,04
B_438	1 973m^2	44%	15%	0,21	R314,52	0,03
B_450	1 350m^2	47%	12%	0,10	R288,00	0,02
B_460	1 600m^2	47%	12%	0,10	R311,33	0,02
B_461	1 550m^2	47%	12%	0,23	R260,88	0,03
B_472	1 000m^2	41%	14%	0,09	R274,29	0,02
B_473	1 328m^2	41%	14%	0,64	R264,21	0,11
B_474	1 463m^2	48%	9%	0,01	R275,00	0,02
B_484	1 519m^2	44%	15%	0,28	R256,75	0,03
B_486	1 550m^2	47%	12%	0,02	R220,00	0,02
B_487	1 106m^2	44%	15%	0,16	R380,00	0,03
B_500	817m^2	41%	14%	0,64	R257,16	0,11
B_502	1 773m^2	58%	9%	0,22	R380,31	0,03
B_513	1 420m^2	47%	12%	0,05	R218,75	0,02
B_519	1 572m^2	41%	14%	0,24	R265,83	0,03
C_001	1 991m^2	60%	6%	0,21	R322,90	0,03
C_015	1 797m^2	49%	14%	0,05	R427,14	0,02
C_029	1 539m^2	43%	19%	0,21	R413,23	0,03

C_035	2 040m^2	46%	18%	0,13	R700,00	0,02
C_036	2 829m^2	58%	9%	0,11	R523,75	0,02
C_042	2 866m^2	49%	14%	0,34	R427,65	0,07
C_047	1 900m^2	60%	8%	0,11	R367,50	0,02
C_049	2 188m^2	41%	14%	0,26	R472,05	0,03
C_056	2 300m^2	73%	2%	0,28	R364,88	0,03
C_058	3 011m^2	50%	22%	0,15	R491,36	0,02
C_062	1 756m^2	41%	14%	0,17	R271,20	0,03
C_067	2 017m^2	66%	2%	0,03	R560,00	0,02
C_069	1 755m^2	55%	4%	0,01	R350,00	0,02
C_076	1 698m^2	42%	26%	0,50	R450,27	0,07
C_077	1 653m^2	58%	9%	0,29	R376,74	0,04
C_078	1 870m^2	44%	15%	0,16	R391,25	0,03
C_081	1 772m^2	47%	12%	0,01	R180,00	0,02
C_082	1 863m^2	60%	6%	0,09	R275,38	0,02
C_085	1 912m^2	44%	15%	0,20	R327,59	0,03
C_087	1 944m^2	42%	26%	0,05	R324,29	0,02
C_091	5 561m^2	50%	22%	0,97	R960,35	0,11
C_092	3 400m^2	44%	15%	0,05	R325,71	0,02
C_098	2 044m^2	62%	7%	0,05	R474,29	0,02
C_099	2 217m^2	56%	12%	0,23	R580,88	0,03
C_102	1 450m^2	57%	12%	0,27	R456,00	0,03
C_107	1 473m^2	44%	15%	0,07	R280,00	0,02
C_108	1 615m^2	55%	9%	0,09	R365,71	0,02
C_109	2 298m^2	57%	3%	0,90	R145,11	0,07
C_110	1 549m^2	80%	1%	0,01	R350,00	0,02
C_114	2 210m^2	49%	14%	0,24	R364,57	0,03
C_115	1 350m^2	41%	14%	0,21	R353,55	0,03
C_119	1 301m^2	44%	15%	0,04	R238,33	0,02
C_121	1 639m^2	42%	26%	0,09	R465,71	0,02
C_122	1 578m^2	47%	12%	0,05	R264,29	0,02
C_127	2 911m^2	56%	10%	0,07	R429,00	0,02
C_128	1 761m^2	54%	4%	0,21	R409,68	0,03
C_132	2 353m^2	56%	16%	0,07	R497,27	0,02
C_135	1 920m^2	56%	12%	0,05	R445,71	0,02
C_137	1 717m^2	55%	4,%	0,01	R205,00	0,02
C_142	2 055m^2	36%	9%	0,11	R479,41	0,02
C_145	1 785m^2	47%	12%	0,29	R226,28	0,04
C_148	915m^2	39%	19%	0,16	R403,48	0,02
C_151	1 606m^2	47%	12%	0,09	R354,29	0,02
C_153	1 600m^2	44%	15%	0,19	R422,86	0,03
C_156	1 885m^2	60%	7%	0,21	R417,10	0,03
C_159	2 900m^2	58%	3%	0,14	R424,00	0,02
C_174	1 300m^2	44%	15%	0,08	R382,50	0,02
C_178	1 895m^2	47%	12%	0,23	R547,06	0,03
C_185	2 423m^2	58%	9%	0,12	R476,67	0,02
C_189	1 540m^2	70%	1%	0,02	R666,67	0,02
C_190	2 020m^2	44%	15%	0,04	R518,33	0,02
C_191	1 900m^2	58%	3%	0,01	R220,00	0,02
C_192	2 461m^2	41%	14%	0,03	R466,67	0,02
C_200	2 571m^2	56%	12%	0,19	R352,14	0,03
C_201	1 986m^2	58%	11%	0,15	R437,27	0,02
C_207	1 864m^2	57%	3%	0,16	R416,09	0,02
C_214	3 067m^2	41%	14%	0,18	R305,38	0,03
C_216	2 676m^2	44%	15%	0,06	R493,33	0,02
C_223	7 438m^2	44%	15%	1,25	R1115,78	0,11
C_224	5 813m^2	47%	12%	1,69	R1153,20	0,11
C_231	2 059m^2	73%	3%	0,01	R800,00	0,02
C_233	1 872m^2	42%	26%	0,27	R491,28	0,03
C_234	1 800m^2	59%	16%	0,07	R450,00	0,02
C_247	1 805m^2	47%	12%	0,02	R450,00	0,02

C_249	1 700m^2	73%	2%	0,14	R275,71	0,02
C_251	1 874m^2	44%	15%	0,04	R530,00	0,02
C_252	1 500m^2	55%	4%	0,44	R260,00	0,04
C_257	7 235m^2	41%	14%	2,23	R810,36	0,11
C_263	1 600m^2	53%	16%	0,01	R315,00	0,02

TABLE E.6: *The predicted ROS for device 00259.*

In Table E.7, the predicted ROS of device 00291 are provided for stores in cluster B1.

Store	Sales area	Province	R0 to R38400	R38401 to R153600	ROS other	Predicted ROS, r_{ij}
B_015	1 229m^2	Mpumalanga	80%	16%	0,29	0,15
B_020	1 046m^2	Limpopo	74%	17%	0,55	0,15
B_023	978m^2	North West	71%	21%	0,48	0,05
B_044	1 915m^2	KwaZulu-Natal	62%	25%	0,72	0,13
B_047	1 820m^2	North West	75%	18%	0,27	0,02
B_050	2 120m^2	KwaZulu-Natal	56%	29%	0,57	0,13
B_051	1 516m^2	Northern Cape	66%	24%	0,46	0,02
B_054	1 215m^2	Mpumalanga	57%	27%	0,48	0,05
B_066	1 080m^2	Eastern Cape	60%	28%	1,11	0,06
B_073	1 868m^2	Limpopo	61%	25%	0,27	0,02
B_090	1 913m^2	North West	77%	18%	0,53	0,13
B_096	813m^2	Eastern Cape	78%	16%	0,26	0,15
B_103	1 910m^2	Limpopo	71%	21%	0,84	0,13
B_106	740m^2	Eastern Cape	74%	19%	0,42	0,05
B_111	1 771m^2	Mpumalanga	67%	24%	0,36	0,02
B_119	1 519m^2	Limpopo	55%	30%	0,39	0,02
B_123	769m^2	KwaZulu-Natal	73%	21%	0,64	0,1
B_137	1 398m^2	Eastern Cape	77%	16%	0,03	0,15
B_139	1 812m^2	KwaZulu-Natal	76%	17%	0,86	0,15
B_142	1 252m^2	KwaZulu-Natal	71%	22%	0,57	0,1
B_143	2 100m^2	KwaZulu-Natal	62%	25%	0,44	0,02
B_144	1 800m^2	KwaZulu-Natal	67%	22%	0,28	0,02
B_149	1 936m^2	Limpopo	73%	18%	1,57	0,13
B_152	800m^2	Eastern Cape	83%	13%	1,79	0,15
B_157	1 725m^2	Limpopo	61%	25%	0,18	0,02
B_169	946m^2	KwaZulu-Natal	68%	24%	0,10	0,05
B_178	1 720m^2	Mpumalanga	80%	16%	1,34	0,15
B_184	1 682m^2	Free State	76%	18%	0,75	0,13
B_190	1 352m^2	Limpopo	74%	17%	0,30	0,15
B_197	894m^2	Eastern Cape	79%	15%	0,24	0,15
B_202	1 197m^2	Free State	71%	21%	0,26	0,05
B_207	992m^2	Limpopo	74%	17%	0,01	0,15
B_210	895m^2	Mpumalanga	76%	18%	5,11	0,06
B_212	929m^2	Eastern Cape	79%	14%	1,04	0,15
B_222	373m^2	Eastern Cape	82%	14%	1,28	0,15
B_230	1 027m^2	North West	71%	20%	0,67	0,06
B_231	1 079m^2	Eastern Cape	77%	16%	0,53	0,15
B_261	1 344m^2	Limpopo	73%	18%	1,24	0,06
B_285	1 305m^2	Free State	76%	18%	0,17	0,05
B_287	1 030m^2	Mpumalanga	67%	23%	1,50	0,06
B_297	1 127m^2	Limpopo	79%	14%	0,06	0,15
B_301	1 605m^2	Limpopo	74%	17%	2,06	0,15
B_307	1 500m^2	Mpumalanga	57%	27%	1,78	0,06
B_311	1 661m^2	Mpumalanga	80%	16%	0,07	0,15
B_321	2 099m^2	KwaZulu-Natal	48%	27%	0,89	0,13
B_324	630m^2	Northern Cape	78%	17%	1,16	0,15
B_325	1 120m^2	Northern Cape	66%	24%	0,44	0,05
B_334	641m^2	Free State	70%	25%	0,72	0,06

B_336	1 600m^2	Eastern Cape	63%	26%	0,33	0,02
B_337	870m^2	KwaZulu-Natal	75%	20%	0,43	0,05
B_340	1 251m^2	Mpumalanga	80%	16%	0,16	0,15
B_342	742m^2	KwaZulu-Natal	73%	19%	0,45	0,05
B_343	750m^2	Eastern Cape	80%	15%	0,10	0,15
B_379	2 618m^2	KwaZulu-Natal	68%	23%	0,58	0,13
B_381	2 099m^2	KwaZulu-Natal	48%	27%	0,24	0,02
B_387	1 308m^2	KwaZulu-Natal	77%	17%	0,55	0,15
B_389	2 607m^2	KwaZulu-Natal	56%	29%	0,34	0,02
B_399	1 156m^2	Free State	68%	22%	0,24	0,05
B_400	1 263m^2	Free State	76%	18%	0,85	0,06
B_420	1 495m^2	Limpopo	55%	30%	1,55	0,06
B_421	1 874m^2	Limpopo	73%	18%	0,29	0,02
B_422	1 897m^2	Limpopo	74%	17%	2,22	0,15
B_448	1 950m^2	Mpumalanga	57%	27%	0,20	0,02
B_454	735m^2	Limpopo	61%	25%	1,22	0,06
B_467	1 250m^2	Eastern Cape	75%	20%	0,75	0,06
B_468	1 583m^2	Eastern Cape	80%	15%	0,13	0,15
B_477	800m^2	Eastern Cape	80%	16%	0,39	0,15
B_483	40m^2	North West	71%	21%	2,90	0,06
B_494	1 795m^2	Limpopo	60%	28%	0,25	0,02
B_496	1 854m^2	Mpumalanga	57%	27%	0,15	0,02
B_497	787m^2	Limpopo	74%	17%	1,48	0,15
B_504	1 632m^2	Mpumalanga	76%	18%	0,37	0,02
B_508	737m^2	KwaZulu-Natal	66%	24%	3,34	0,1

TABLE E.7: *The predicted ROS for device 00291.*

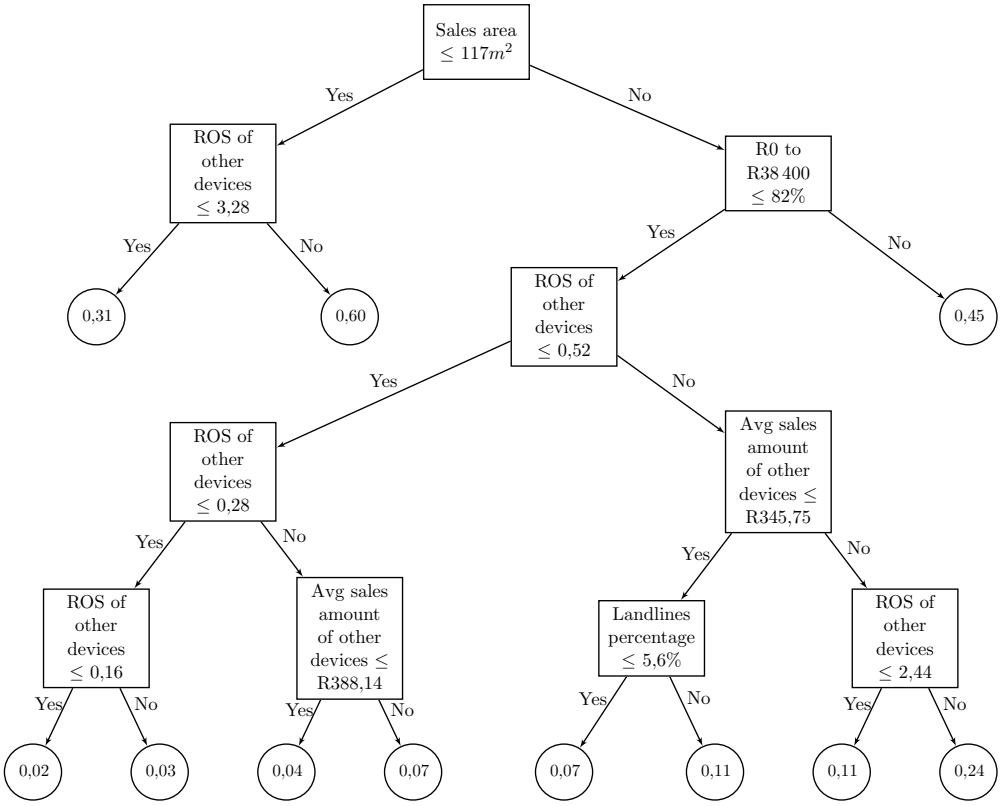


FIGURE E.3: *The regression tree for device 00259.*

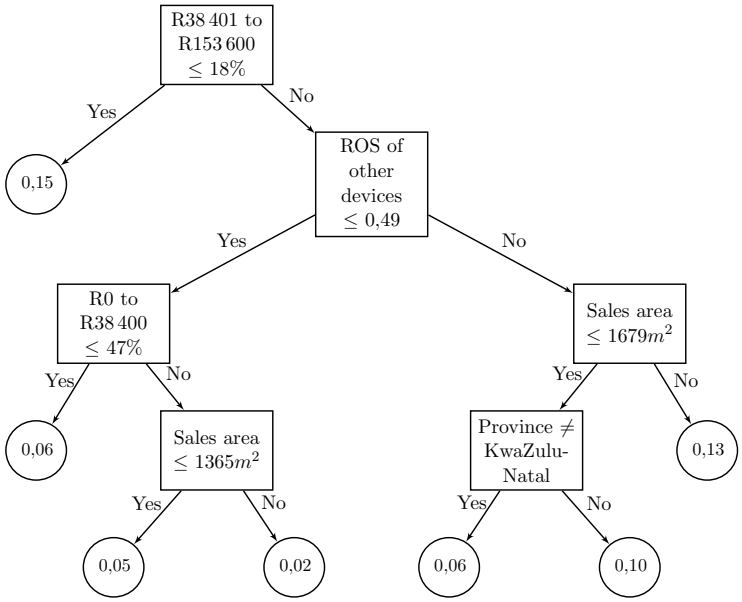


FIGURE E.4: *The regression tree for device 00291.*