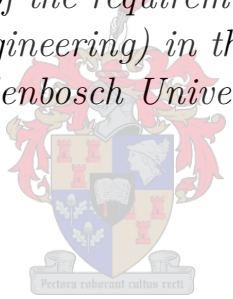
Implementation of clustering techniques for segmentation of Mozambican cassava suppliers

by
Ntebaleng Sharon Matshabaphala

Thesis presented in fulfilment of the requirements for the degree of Master of Engineering (Industrial Engineering) in the Faculty of Engineering at Stellenbosch University



Supervisor: Prof Jacomine Grobler

March 2021

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.

March 2021

Abstract

Although an organisation generally accumulates many suppliers in the course of doing business, some of these suppliers are of little or no importance to the organisation beyond fulfilling a simple order transaction, while other suppliers play a strategic role in the success of an organisation.

The decision to invest in supplier relationships is a major step for an organisation, especially because the value gained from interacting in a supply network rests on the principle of prioritising the right suppliers. The segmentation of suppliers plays a significant role in supplier relationship management. Not only does it offer an effective method of assessing suppliers, but it also provides a resource-efficient decision methodology that specifies appropriate relationships and governance structures for each segment.

In this thesis, three techniques are applied for clustering cassava suppliers in Mozambique. Over 3 000 smallholder farmers supply cassava to a for-profit social enterprise called Dadtco Philafrica. Dadtco Philafrica needs an effective supplier segmentation method to gain insight into how it should direct its resources to where they will have the greatest impact.

The k-means algorithm, agglomerative hierarchical clustering (AHC), and self-organising maps (SOM) with Ward clustering were applied to a real-world case study. Extensive algorithm parameter tuning was conducted in order to ascertain good parameter values for each clustering technique. Performance of the algorithms was evaluated and compared using intra-cluster and inter-cluster distances, and the best performing algorithm, in the context of the case study, was selected. The SOM with Ward clustering outperformed the k-means and AHC, and its results were used to conduct a detailed cluster analysis. The insights gained from the cluster analysis were used to provide recommendations and to suggest suitable intervention strategies to manage each segment of suppliers.

The encouraging results of these algorithms showed that clustering techniques can be utilised effectively in segmenting suppliers. The proposed method offers users the basis of a supplier segmentation system that is more efficient. A user can simply rerun the algorithm using the latest data, to check for suppliers who have moved to a different cluster and to determine cluster allocation of new suppliers. This method relies primarily on historical data to segment suppliers; therefore, it provides an organisation with data-based insight regarding its supply base.

Opsomming

Alhoewel 'n organisasie oor die algemeen heelwat verskaffers deur die loop van sake versamel, is sommige van hierdie verskaffers van min of geen belang vir die organisasie buiten om 'n eenvoudige besteltransaksie uit te voer, terwyl ander verskaffers 'n strategiese rol speel in die sukses van 'n organisasie.

Die besluit om in verskaffersverhoudinge te belê, is 'n belangrike stap vir 'n organisasie, veral omdat die waarde wat uit die interaksie in 'n verskaffingsnetwerk verkry word, berus op die beginsel van prioritisering van die regte verskaffers. Die segmentering van verskaffers speel 'n belangrike rol in die bestuur van verskafferverhoudinge. Segmentering bied nie net 'n effektiewe metode om verskaffers te evalueer nie, maar ook 'n hulpbroneffektiewe besluitnemingsmetodologie wat toepaslike verhoudings en bestuurstrukture vir elke segment spesifiseer.

In hierdie tesis word drie tegnieke toegepas vir die groepering van kassava-verskaffers in Mosambiek. Meer as 3 000 kleinboere lewer kassava aan 'n winsgewende maatskaplike onderneming met die naam Dadtco Philafrica. Dadtco Philafrica benodig 'n effektiewe verskaffersegmenteringsmetode om insig te bekom oor hoe sy hulpbronne aangewend moet word om die grootste impak te maak.

Die k-gemiddelde groepering algoritme, agglomeratiewe hiërargiese groepering (AHC) en selforganiserende afbeelding (SOM) met 'Ward' groepering is toegepas op 'n werklike gevallestudie. Omvattende instelling van algoritme-parameters is uitgevoer om goeie parameter waardes vir elke groeperingstegniek te bepaal. Die uitvoering van die algoritmes is geëvalueer en vergelyk ten opsigte van intra-groep en inter-groep afstande, en die beste presterende algoritme, in die konteks van die gevallestudie, is gekies. Die groepering van die SOM met 'Ward' groepering het beter gevaar as die k-gemiddelde groepering algoritme en AHC, en die resultate daarvan is gebruik om 'n gedetailleerde groepontleding uit te voer. Die insigte wat uit die groepontleding verkry is, is gebruik om aanbevelings te gee en geskikte intervensiestrategieë voor te stel om elke segment van verskaffers te bestuur.

Die bemoedigende resultate van hierdie algoritmes het getoon dat groeperingstegnieke effektief in verskaffersegmentering gebruik kan word. Die voorgestelde metode bied gebruikers die basis van 'n verskaffersegmenteringsstelsel wat meer doeltreffend is. 'n Gebruiker kan eenvoudig die groepontleding oordoen deur die nuutste data te gebruik om verskaffers wat na 'n ander groep beweeg het te identifiseer, en om die groepering van nuwe verskaffers te bepaal. Hierdie metode maak hoofsaaklik staat op historiese data vir verskaffersegmentering; daarom bied dit 'n organisasie data-gebaseerde insig rakende die verskaffer basis.

Acknowledgement

I would like to acknowledge many individuals for helping me during my studies at Stellenbosch University. Particularly, I would like to express my sincere gratitude to my supervisor, Prof Jacomine Grobler, for providing me with the opportunity, knowledge and support throughout this study. I am deeply grateful for her continuous support, patience and guidance.

I would also like to thank my Dadtco Philafrica team for providing me with the dataset used for this study, and for generously offering their time to answer my questions.

A special thanks to my family - my parents and my siblings - for their encouragement and support. I would also like to thank my nephews and nieces for their love.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In today's turbulent and competitive global market, the average number of suppliers that a company needs to manage has increased drastically. A significant increase in the number of suppliers has increased supply chain network complexity. As a result, organisations have been exploring different techniques to integrate with upstream and downstream supply chain partners to increase competitiveness and adapt to rapid changes in market trends. Over the past several years, there has been an emphasis on strategic sourcing that establishes a long-term and mutually beneficial relationship with fewer, but better performing suppliers. Although determination of suitable suppliers in the supply chain has become a key strategic consideration, the nature of the approaches has generally been unstructured. Collaboration between a buying company and its key suppliers can provide protection against supply bottlenecks and inventory shortages, both of which can affect business success [38, 49].

Organisations have been improving their supply chain operations through understanding the importance of effective supplier relationship management (SRM). SRM is one of the core business activities that is applied to gain a competitive advantage needed to operate in global markets in terms of expertise, knowledge, and ability to share risks [15]. The approach towards managing the relationship between suppliers and a buying company has been changing and moving towards a more collaborative approach. Organisations have not only realised that doing business jointly with their strategic suppliers has the potential to enhance their organisational ability to reduce supply risk, but they have also acknowledged that a collaborative approach helps them respond quickly to demand changes [49].

One solution to improve SRM is to divide all suppliers into smaller sets, where the members of each set have a greater degree of similar characteristics. Because of a steep increase in the number of suppliers, it has become exceedingly difficult for a buying company to develop a fully-tailored procurement strategy for each supplier. Supplier segmentation is defined as a process that involves dividing suppliers with different characteristics, needs, and requirements into distinct groups in order to realise value from the exchange between goods and finances [54]. Supplier segmentation is the initial and integral step for an effective SRM initiative, and when applied effectively, it can reduce the purchasing cost and improve corporate competitiveness [7]. Clustering is a standard tool that is commonly used for supplier segmentation. Clustering is defined as dividing data points into groups of similar objects where each group consists of objects that are similar amongst themselves and dissimilar compared to objects of other groups [32, 63].

Clustering has been used in many contexts by researchers in many disciplines. In marketing, decision-makers have moved away from the 'one-size-fits-all' content marketing strategy and are now using data-based techniques to address consumer heterogeneity by grouping consumers into

segments based on similarities. With the changing needs and expectations of customers, segmentation of customers and development of personalised marketing initiatives have become imperative for companies. By efficiently segmenting customers into various groups based on their buying behaviour and spending patterns, organisations can best allocate their marketing budget and yield significant savings [68, 74]. It is important to note that, despite the enormous potential, using clustering techniques has not received much attention in supplier segmentation, where grouping of suppliers based on similarities can enhance the effectiveness of supplier relationship management. An opportunity therefore exists for research into the use of clustering for supplier segmentation.

This thesis focuses on segmenting cassava farmers who supply cassava to cassava processing plants owned by Dadtco Philafrica in Mozambique. Cassava is an important crop for food security as it provides a reliable and inexpensive source of carbohydrates in many developing countries. Cassava is one of the most drought- and disease-resistant crops and it is capable of growing on land that has little or no agricultural value [26, 46]. Cassava is a major staple food in Mozambique, providing a basic diet for over 80% of the country's population. In 2016, Mozambique was ranked eleventh in the world with regard to cassava production [61].

As the only key industrial cassava processor in Mozambique, Dadtco Philafrica plays an enormous role in industrialising the use of cassava. The sustainability and success of the organisation's business model has a great impact on the livelihood of the thousand smallholder farmers who supply to the organisation. Furthermore, the organisation's success will not only be groundbreaking for the Mozambican agriculture sector, but it will have a significant impact on the whole African continent. Cassava is one of the largest produced crops in Africa; therefore, industrialisation of its use will unleash a significant agricultural potential and alleviate poverty in many African countries.

The first objective of the introductory chapter is to provide a rationale for studying segmentation of suppliers. The objectives and contributions of this thesis are further highlighted in section 1.1 and section 1.2 before a brief outline of the rest of this thesis is provided in section 1.3. Please note that throughout this thesis, the word 'organisation' is used to refer to a focal company purchasing from suppliers.

1.1 Research objectives

The main objective of this thesis is to serve as a proof of concept for the use of clustering algorithms to enhance the effectiveness of supplier relationship management through supplier segmentation. This objective is divided into sub-objectives which, once completed, will culminate in the completion of the main objective. These sub-objectives are provided in the following list, along with the chapter within which that sub-objective is addressed.

1. Motivate the necessity for effective supplier relationship management - Chapter 2.
2. Motivate why the selected case study is suitable for the investigation of clustering techniques for supplier segmentation - Chapter 2.
3. Provide a detailed description of how the CRISP-DM (CRoss Industry Standard Process for Data Mining) reference model is applied in a clustering project - Chapter 3.
4. Provide a detailed review of the selected clustering techniques: the k-means algorithm, agglomerative hierarchical clustering and self-organising maps - Chapter 3.

5. Discuss how supplier segmentation can enhance the effectiveness of supplier relationship management - Chapter 4.

6. Provide a detailed review of different methods used in the segmentation of suppliers - Chapter 4.

7. Provide a description of the business problem that the organisation (in the case study) is trying to solve and how the insight gained from clustering techniques will address the stated problem - Chapter 5.

8. Conduct data exploration and provide results obtained from exploring the dataset - Chapter 5.

9. Conduct data preparation considering data cleaning, data normalisation, and feature selection - Chapter 5.

10. Conduct extensive algorithm parameter tuning to ascertain good parameter values for each clustering technique - Chapter 5.

11. Evaluate and compare the performance of clustering techniques, then identify the best performing technique in the context of the case study - Chapter 5.

12. Conduct detailed cluster analysis of the results obtained by the best performing clustering technique - Chapter 6.

13. Discuss the characteristics of the clusters and develop suitable intervention strategies to manage each cluster - Chapter 6.

14. Provide suggestions for future research - Chapter 7.

1.2 Expected contributions

Although supplier segmentation is not new, traditional approaches use methods which primarily rely on human judgement to rate a supplier's perceived importance to the buying company. These human judgement-based approaches are not only subjective but are time-consuming and require the organisation's decision-makers to have worked closely with each supplier for an extended period of time in order to rate suppliers fairly. The reliance on human judgement to measure suppliers' potential is not only exposed to subjectivity, but it is also inefficient. While it might be possible to apply the traditional supplier segmentation methods in an organisation that has a smaller number of suppliers, it would be impractical to apply these traditional methods in an organisation that has thousands of suppliers.

With the explosion and availability of data, relying on qualitative methods for supplier segmentation shows a significant gap in today's complex supply chain environments. Availability of data has grown exponentially over the years, and more organisations are moving to data-driven decision-making as it eliminates inaccuracies caused by possible biases.

The supplier segmentation method proposed in this study follows a static clustering approach where the segment that a supplier belongs to does not change unless the algorithms are rerun with an updated dataset. To the best of the author's knowledge, this study is the first supplier segmentation method to address the following:

1. The proposed method is the first application of clustering to segment cassava suppliers. The benefit of the proposed method is its ability to use multiple-criteria decision analysis on a large dataset.

2. Unlike the existing supplier segmentation methods, the proposed method is the first supplier segmentation method that relies primarily on historical data as input to assess suppliers. The benefit in using historical data, such as historical purchases, is that it contains all the details of transactions between a supplier and a buyer; thus, the results obtained are more reliable. Furthermore, the reliance on historical data instead of human judgement means that this method can be effective even in an organisation where the decision-makers do not have in-depth knowledge of the supplier base.

3. Existing supplier segmentation methods require the end user's involvement in aggregating suppliers' rating and forming clusters among suppliers with similar scores. This kind of involvement not only requires time from the user, but it also requires the user to have some literacy abilities. In the proposed method, all the steps - from rating suppliers, to aggregating scores and assigning suppliers into segments - are done by the clustering techniques.

4. The proposed method offers users the basis of a supplier segmentation system that is more efficient and which can be automated. A user can rerun the algorithm after a certain period, using the latest data, to check for suppliers who have moved into a different cluster and to determine cluster allocation of new suppliers.

1.3 Thesis outline

The remaining chapters of this thesis are organised as follows: Chapter 2 provides background information on the selected case study that will serve as proof of concept to the use of clustering algorithms in supplier segmentation. Chapter 3 discusses the role of clustering and provides a detailed description of how the CRISP-DM reference model is applied in a clustering project. The chapter also includes a detailed review of the clustering techniques, which are implemented in the case study. The three techniques discussed are the k-means algorithm, agglomerative hierarchical clustering, and self-organising maps. Chapter 4 discusses supplier relationship management's role in ensuring that organisations adapt and remain competitive in the dynamic business landscape. Furthermore, the chapter reviews different methods applied in the segmentation of suppliers. The clustering algorithms are applied to a real-world case study in Chapter 5, and the cluster results are evaluated using the inter-cluster distances and intra-cluster distances. The best performing algorithm is selected, and a detailed cluster analysis is conducted. The insight gained from the analysis is used to make recommendations for each cluster. Chapter 7 concludes the thesis with a summary of significant findings from the study, and suggestions for future research opportunities.

Chapter 2

Case study introduction

The overarching goal of this thesis is to serve as a proof of concept for the use of clustering algorithms in supplier relationship management by providing supply chain managers with a method that enables them to segment suppliers efficiently. Dadtco Philafrica, which owns two cassava processing plants in Mozambique, was selected to serve as a proof of concept case study. The purpose of this chapter is to provide background information on cassava processing in terms of its significance in Mozambique and the reasoning behind its selection as a proof of concept case study for the purposes of this thesis.

Section 2.1 explains the significant role that cassava plays in Mozambique's economic and social growth. In section 2.2, the size of land that farmers use to produce cassava is discussed. The processing of cassava and various products that can be made from cassava is discussed in section 2.3. Section 2.4 describes barriers to the development of a viable larger-scale cassava processing industry in Mozambique. Lastly, the motivation for using Dadtco Philafrica to serve as a proof of concept for the use of clustering algorithms in supplier relationship management is discussed in section 2.5.

2.1 The importance of cassava in Mozambique

Cassava is an important crop contributing to Mozambique's overall gross domestic product (GDP). In 2016, agriculture accounted for roughly 18% of GDP, and cassava production's direct share of agricultural output by value was more than one-quarter of the 18%. For this reason, cassava production plays a significant role in the country's social and economic growth, particularly in vulnerable rural populations [14, 61].

There are many different varieties of cassava grown in Mozambique. Generally, varieties are classified according to various traits such as taste, crop duration to maturity, average yield, and disease resistance. In order to provide food under a wide variety of circumstances and periods, some cassava farmers grow a diversity of cassava varieties at the same time in different plots. This approach typically provides a mix of yields and resilience to diseases and drought, as well as the different periods for harvesting [14, 27].

In Mozambique, cassava is largely used in households for direct human consumption. The form in which cassava is consumed varies in the different regions. In the northern regions, cassava flour is mainly boiled with water to make a stiff porridge which can be served with vegetables, fish, or meat. The flour is made by first peeling the roots of the plant. After this, the roots are chipped or fermented, then sun-dried before milling. In the central and southern regions, cassava is mainly consumed fresh. Traditionally, cassava is boiled and served with a green salad

and tea [19, 61]. Examples of cassava meals are shown in Figure 2.1.

Figure 2.1: Cassava dishes

2.2 Cassava Production

Similar to other food crops in Mozambique, cassava is grown largely by subsistence and small-scale family farmers. Below are the three categories that distinguish cassava producers in Mozambique [14]:

- Category 1: There are around 2.5 million smallholder farms, with plot sizes of 1.5 hectares (ha) on average, but these farmers use mainly around 0.4 to 0.6 ha for cassava production.
- Category 2: There are approximately 8 000 medium-size farms with plot sizes between 10 and 20 ha, but most of these farmers use only about 1.2 ha for cassava production.
- Category 3: There are about 115 large commercial cassava farmers, with plot sizes of more than 10 ha.

Cassava is a multiple-year crop and roots can be stored for up to 30 months underground (unharvested). The harvesting season for cassava is considered to be flexible as roots can be harvested between 8 months and three years. Generally, the cassava roots have a bulky shape and contain about 70% water. As a result, the crops require considerable post-harvest effort. Furthermore, a cassava crop has a very short shelf-life; once harvested, the crop needs to be processed within three days. Its rapid post-harvest deterioration is one of the key factors that has limited its market development [19, 61].

2.3 Cassava Processing

Most cassava processing that occurs in Mozambique is non-mechanised. The traditional methods which involve soaking, drying and chipping or grating of cassava, are highly labour-intensive and low profitability work [76].

The industrial use of cassava in Mozambique is less than 0.5% of national cassava production. Although the present industrial use of cassava is very low, the potential is enormous, especially for regional exports. In the present development stage of Mozambican cassava value chains, the existing domestic markets for industrial cassava products are [14, 25, 61]:

- **Animal feed**: An important use of the cassava crop is cassava chips and leaves for animal feed. The leaves are high in protein, and the roots are high in carbohydrates. Therefore, the leaves and the roots are potential substitutes for soybeans and maize.
- **High-quality cassava flour**: High-quality cassava flour (HQCF), also known as Tapioca starch, can be partly used as a wheat flour substitute in bread, pastries and biscuits. HQCF can also be used as a thickener and stabiliser in soups and meat products. It is worth noting that the macroeconomic impact of producing cassava flour locally would make bread more affordable, as imported wheat is usually more expensive.
- **Ethanol**: Ethanol is largely used in the spirit distilling industry. Extra neutral alcohol (ENA), which is its portable form, is blended with water and other flavours to produce many alcoholic beverages. In 2010, there was an initiative in Mozambique to produce ethanol for cooking stoves, but the factory ceased operations after three years.
- **Cassava starch paste**: Cassava roots can also be used to produce alcoholic beverages. In 2011, an initiative led by Mozambique SABMiller began to produce Impala beer, shown in Figure 2.2, using cassava starch paste as an ingredient. Since then, the production of Impala beer has been expanded to other SABMiller plants in Mozambique.

Figure 2.2: Beer made from cassava

2.4 Industrialising the use of cassava

Despite having considerable production capacity, the Mozambican supply chain for cassava is fragile. In Mozambique, cassava production is still a task for rural families, with virtually no organised plantations. The inability to ensure a stable and permanent supply of cassava is considered to be one of the main obstacles to building a strong cassava industry. A cassava processing business can only be successful if the business is supplied with enough good quality raw material on a regular basis [76].

The three major barriers to the development of a viable larger-scale cassava processing industry are supply risk, productivity and transportation. This thesis describes a methodology that can be used to objectively and effectively measure each supplier's performance against these barriers. Suppliers with similar performances are then grouped together to improve supply chain planning and resource allocation processes. The stated three major barriers are described below:

1. **Supply risk**: Supply of cassava to processors is generally highly irregular. One of the causes of these irregularities is that smallholder producers have poor access to quality farming inputs, pesticides, fertilisers, and mechanised ploughs. Non-mechanised farming is labour intensive and a lack of quality farming input usually results in low yields. Some farmers become discouraged which may result in infrequent harvests. Some farmers even switch to a different crop, hoping for better returns [14, 25].

2. **Productivity**: Low productivity emerges as the main constraint preventing farmers from producing at a profit and the processors from sourcing the fresh roots at affordable prices. Mozambican cassava yields varies between 5 and 9 tons per ha, and with the current industry prices, most farmers do not make any profit from producing cassava. From the farmers' point of view, prices offered by the industry are too low. To be commercially viable at the present market prices, cassava producers need to achieve a yield of at least 15 tons per ha [3, 14].

3. **Transportation**: It is important to note that smallholder farmers, who account for almost all cassava production, are poorly organised and spread widely across rural areas which are difficult to reach. As a result of low yields and geographically scattered farms, processors are forced to source from a larger number of farmers, which increases the time and cost of securing an adequate supply of cassava. The inefficient transport system from fields to processing sites has a huge impact on the processor's operational cost [3].

2.5 Motivation for selection of case study

An organisation that has tapped into this enormous potential market of making industrial use of cassava is Dadtco Philafrica. The organisation produces two end products namely HQCF and cassava starch paste. Dadtco Philafrica's approach covers the whole cassava value chain, from providing farming supplies to smallholder farmers, cassava processing, and marketing of the final product in local and international markets [14].

Dadtco Philafrica, which sources raw material from thousands of subsistence and small scale farmers who have no strong farming or business background, seeks to improve its supplier relationship management. Through effective supplier relationship management, the organisation aims to stabilise its operations and prove the viability of its business model. In order to achieve this goal, the organisation has realised the importance of empowering and developing its raw material suppliers. Dadtco Philafrica has observed that without an adequate supply of good quality raw material, the organisation cannot have a sustainable business.

Dadtco Philafrica purchases fresh cassava at the farm gate and uses its own trucks to transport cassava to the processing sites. This type of operation has been beneficial for the small farmers concerned since they can rely on selling their fresh root production for a known and typically higher price, compared to what they would receive by transforming it into chips and selling in the informal market. Being able to sell all their cassava at once allows farmers to replant in an efficient pattern. Traditionally, cassava farmers harvest and sell small quantities throughout the year. The farmers cannot harvest large quantities as once it is harvested, the crop deteriorates within a few days [14].

With over 3 000 smallholder farmers in its database, the organisation needs to find an effective and objective method to determine how to spend its limited resources on its suppliers. First, the organisation understands that the farmers are diverse, and their levels of competency differ

significantly. As a result, the type and intensity of the relationship the organisation can have with each farmer should vary. Furthermore, the farmers will require a different level of development based on the magnitude of their potential. While the organisation aims to empower as many farmers as possible, Dadtco Philafrica understands that it would be a waste of valuable resources to use a 'one-size-fits-all' approach in developing and investing in these farmers.

A key challenge faced by Dadtco Philafrica management is how they can objectively and effectively measure the 'potential' of each supplier. In the existing literature, the way in which 'potential' of suppliers has been measured is largely subjective. Each supplier is given a rating based on the decision-maker's judgement. Although the decision-makers are expected to be knowledgeable employees who have worked with the suppliers for a long period, this method of rating suppliers can still be exposed to bias. Moreover, this approach is suitable for companies with a small number of suppliers; the approach would be impractical to apply it to Dadtco Philafrica, which has over 3 000 suppliers to evaluate.

By providing the means to measure 'potential' of suppliers in an objective and effective manner, clustering algorithms can provide supply chain managers with the insights that are required to implement supplier segmentation. This will improve the organisation's supplier relationship management. Further advantages include improved supply chain planning and more efficient resource allocation. However, implementing clustering algorithms for supplier segmentation requires the availability of data that tracks suppliers' transactions over time, along with clearly defined criteria that are important to the business. The problem addressed by this thesis is whether it is possible for Dadtco Philafrica to implement clustering techniques to evaluate its suppliers and develop intervention strategies suitable for each cluster, with the aim of improving the effectiveness of its SRM approach.

2.6 Chapter summary

This chapter provides background information to the selected case study that will serve as proof of concept to the use of clustering algorithms in supplier relationship management. Firstly, the vital role that cassava production plays in Mozambique is highlighted. Furthermore, the enormous potential of industrialisation of cassava is discussed, followed by the key challenges faced by cassava processors. The chapter also discusses the impact that supplier segmentation would have on the organisation's SRM approach. Finally, the chapter concludes with a motivation as to why Dadtco Philafrica's operations are suitable to serve as a proof of concept for using clustering algorithms to segment suppliers.

Chapter 3

Literature Review: Clustering

In this chapter, the role of clustering and a detailed approach in applying the six phases of CRoss Industry Standard Process for Data Mining (CRISP-DM) [10] is discussed. Section 3.1 provides a detailed description of how the CRISP-DM reference model is applied to a clustering project. Section 3.2 describes the approach used in gaining insight into the business and defining the goal of the project. Section 3.3 describes different data exploration methods. Activities involved in constructing a final dataset are described in section 3.4. Section 3.5 includes a detailed review of the clustering techniques which are implemented in the case study. The three techniques discussed are the k-means algorithm, agglomerative hierarchical clustering (AHC) and self-organising maps (SOM). Lastly, section 3.6 offers a clear description of how the performance of the clustering techniques is evaluated.

3.1 Introduction to CRISP-DM

The CRISP-DM is a highly recommended reference model for data science that provides an overview of the life cycle of a data science project. It defines a process model which provides a framework for carrying out data science projects which are independent of both the industry sector and the technology used. Moreover, it explicitly views the data science process from both an application-focused and a technical perspective [39, 44]. For the purpose of this thesis, data science is defined as a set of fundamental principles that support and guide the principled extraction of information and knowledge from data [52].

The life cycle of a data science project consists of a cycle that comprises six stages which are shown in Figure 3.1. The figure illustrates the flow between each of these phases and emphasises that data is at the heart of the process. The CRISP-DM process is not strictly linear, and some processes are more closely linked than others. For instance, business understanding and data understanding are closely linked, and projects typically spend some time iterating between the two phases. The same applies for the data preparation and modelling phases [34, 44].

3.2 Business understanding

The primary goal in this step is to understand the business problem that the organisation wants to solve and determine the kind of insight that data science can provide to help the organisation address the stated problem. First, the goals that the business wants to achieve need to be

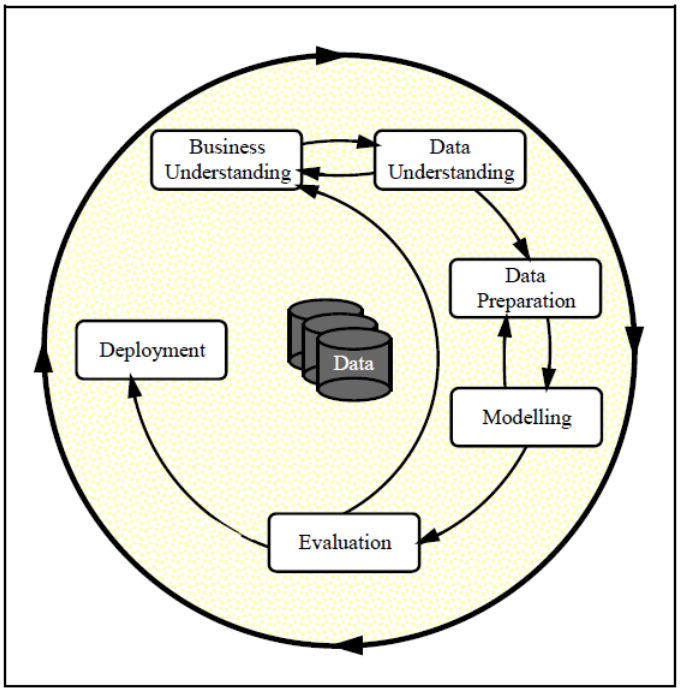


Figure 3.1: CRISP-DM process model [39]

described. Understanding the problem will ensure that the project is focused on goals that are clearly defined, thus increasing the likelihood of the project's success [44, 72].

Second, this phase demonstrates ways a data science project could address the identified organisational problem. Once approaches that address the business problem have been defined, the next task is to determine the volume and availability of the data. The amount of data that is available is important because very small datasets can affect the performance of data science methods [34, 44].

Once the data has been made available, data structures that will be used to build and evaluate data science models need to be designed. This step typically overlaps with the business understanding and data preparation phases [34]. Data in organisations is rarely saved in neat tables ready to be used for data science; therefore, datasets need to be constructed from raw data that may be obtained from diverse sources. Data in an organisation can be obtained from sources such as operational databases, data warehouses, and external feeds. Data obtained from different sources needs to be merged into one dataset to create a data structure suitable for data modelling [34].

3.3 Data understanding

The data understanding phase starts with an initial data collection and proceeds with activities aimed at becoming familiar with the data. It is worth noting that there is a close link between business understanding and data understanding. The formulation of the data science problem requires an adequate understanding of the available data. The first goal in data understanding is to know the data and fully understand its characteristics. The characteristics of each feature in the dataset needs to be studied so that the data types and the data distribution for each feature are understood. The second goal is to determine if data suffers from any data quality issues that could adversely affect the results obtained during data modelling [5, 44, 72].

Section 3.3.1 describes different data types. Section 3.3.2 explains different methods for exploring individual features and section 3.3.3 describes techniques that can be applied when examining relationships between pairs of features.

3.3.1 Data types of features

A key step in understanding the characteristics of the features is to determine their data types. Some of the commonly used data types are stated below. On a high level, data types can be reduced to two types: continuous (encompassing the numeric and interval types) and categorical (encompassing the nominal, ordinal and binary types) [32].

- Numeric: Numeric data is values that are measurable and which allow arithmetic operations (e.g., price, height).
- Interval: The interval data type stores values that represent a span of time. The values generally allow ordering and subtraction, but do not allow other arithmetic operations (e.g., date, time).
- Ordinal categorical: Ordinal categorical data is discrete values that allow ordering but do not permit arithmetic operations (e.g., size measured as small, medium or large).
- Nominal categorical: This data type contains discrete values that cannot be ordered and allow no arithmetic operations (e.g., country name, product type).
- Binary: Binary data is discrete values that can be in only one of two categories (e.g., on or off, yes or no, 1 or 0).
- Textual: The textual data type is free-form, usually short text data (e.g., name, address).

3.3.2 Descriptive statistics for individual features

The descriptive statistics and data visualisation techniques described in this step focus on the characteristics of individual features. The characteristics of each feature in the dataset are described using standard statistical measures of central tendency, standard measures of variation, and standard data visualisation plots [34].

The descriptive statistics described can be visualised using standard data visualisation plots such as bar plots, histograms, and box plots. Figure 3.2 shows a selection of histogram shapes that exhibit characteristics commonly seen when analysing features that have continuous data types. The histograms are indicative of standard, well-known probability distributions [34].

A uniform distribution, as shown in Figure 3.2 (a), indicates a distribution that has constant probability. Sometimes a uniform distribution is indicative of a descriptive feature that contains unique identification codes. Figure 3.2 (b) demonstrates a bell-shaped curve known as a normal distribution. Features following a normal distribution are characterised by a strong tendency toward a central value and symmetrical variation to either side of this central tendency. Having features that exhibit a normal distribution is advantageous as many of the modelling techniques work particularly well with normally distributed data [2, 34].

Figures 3.2 (c) and (d) show unimodal histograms that exhibit skewness. Skewness is described as a tendency toward very high (right-skewed) or very low (left-skewed) values. Figure 3.2

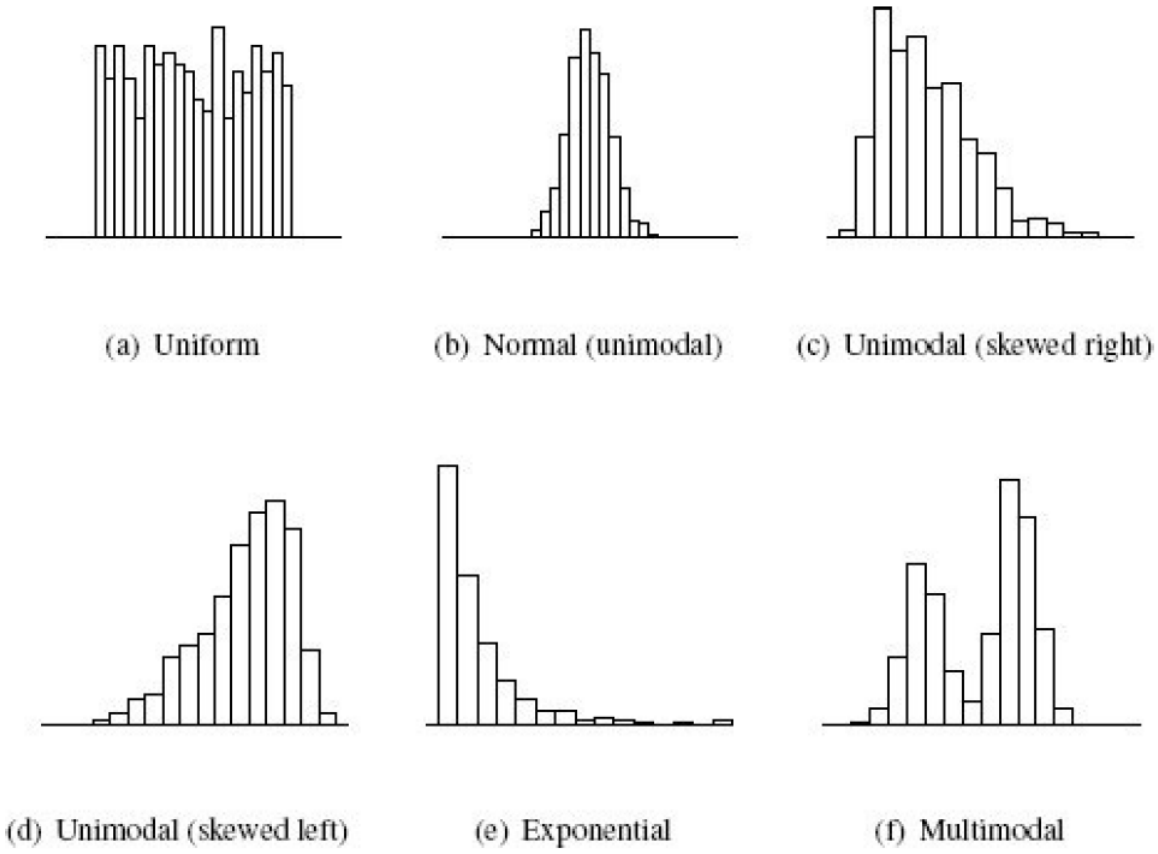


Figure 3.2: Common shapes of histograms [34]

(e) shows a feature following an exponential distribution where the likelihood of low values occurring is very high but diminishes rapidly for higher values. An exponential distribution has a long tail and commonly has high values. Moreover, features that follow an exponential distribution are more likely to contain outliers, which are data points that differ significantly from other observations. Lastly, a feature characterised by a multimodal distribution has two or more very commonly occurring ranges of values that are clearly separated. Figure 3.2 (f) shows a bimodal distribution with two clear peaks. Multimodal distributions tend to occur when a feature contains a measurement made across several distinct groups [2, 34].

A measure of central tendency is a single value that describes a set of data by identifying the central position within a set of data. The measures give a rough estimate of the clustering of the data around the midpoint and an indication of the central value. Measures of central tendency include the mean, median and mode. The arithmetic mean is the most commonly used measure of central tendency, and it measures the average value. It is computed by adding all the values of the feature and dividing by the number of observations. The mean uses every value of the feature and hence it is a good representation of the data. However, the mean is susceptible to the influence of outliers. Therefore, it is not an appropriate measure of central tendency for skewed distributions [16].

The median is the central value obtained when data points are arranged in an ascending or descending order. Unlike the mean, the median is less affected by outliers and skewed data. The mode is the most frequently occurring value in the dataset. The mode is generally used for categorical data used to determine the most common category. Furthermore, the mode would be the most appropriate measure to be used when a feature follows a bimodal or trimodal

distribution [43].

Standard measures of variation are used to describe the amount of variability or spread in a feature. The most common measures of variability are the range, the interquartile range (IQR), and standard deviation. The standard deviation, which is the most widely used measure of variability, is the square root of the average squared difference of the values from the mean. A smaller standard deviation indicates that the values in a feature are grouped closer together, and a larger value indicates the values are more spread out [23].

The interquartile range (IQR) is a measure of variability, based on dividing a dataset into quartiles. Quartiles divide a feature that is ordered in ascending order into four equal parts. The values that divide each part are called the 1^{st}, 2^{nd}, 3^{rd}, and 4^{th} quartiles. The interquartile range is the difference between the third and first quartiles, in which the first and third quartile are the middle value in the first and second half of the feature. The interquartile range is a robust measure of variability as it is less influenced by outliers and can be used on skewed data. Variability can also be defined in terms of how close the scores in the distribution are to the middle of the distribution [23].

3.3.3 Descriptive statistics for pairs of features

This section introduces techniques that can be applied when examining relationships between pairs of features. Correlation, which is a common measure when examining relationships between pairs of features, is a bivariate analysis that calculates strength of association between two continuous features. Correlation values fall into the range between -1 and $+1$ where values close to negative -1 indicate a very strong negative correlation, values close to $+1$ indicate a very strong positive correlation, and values around 0 indicate no correlation. Features that have no correlation are considered to be independent of each other [34].

Correlation is a good measure of the relationship between two continuous features, but it is not by any means perfect. It is important to note that correlation does not necessarily imply causation. Just because the values of the two features are correlated does not mean that an actual causal relationship exists between them. Therefore, before causation is concluded based on a strong correlation, in-depth studies involving domain experts are required [34].

Generally, correlation is a relatively good measure of the relationship between two continuous features. Visualisation methods can also be used to inspect relationships between different features. The scatter plot matrix (SPLOM) is a well-known technique of visual analysis for continuous variables. SPLOM shows scatter plots for a collection of features arranged into a matrix. The relationship between two categorical features can be visualised by using a collection of bar plots. The relationship between a categorical feature and a continuous feature can be visualised by using a collection of box plots [34, 65].

The correlation between two features, $\boldsymbol{a}$ and $\boldsymbol{b}$, in a dataset with n instances can be calculated as follows [34]:

$$corr(\boldsymbol{a}, \boldsymbol{b}) = \frac{cov(\boldsymbol{a}, \boldsymbol{b})}{sd(\boldsymbol{a}) * sd(\boldsymbol{b})} \tag{3.1}$$

where $cov(\boldsymbol{a}, \boldsymbol{b}) = \frac{1}{n-1} \sum_{i=1}^{n} ((a_i - \bar{a}) * (b_i - \bar{b}))$. a_i and b_i are the i^{th} instances of features $\boldsymbol{a}$ and $\boldsymbol{b}$; $\bar{a}$ and $\bar{b}$ are the sample means of features $\boldsymbol{a}$ and $\boldsymbol{b}$, $sd(\boldsymbol{a})$ and $sd(\boldsymbol{b})$ are the standard deviation of $\boldsymbol{a}$ and $\boldsymbol{b}$, respectively.

3.4 Data preparation

The data preparation phase covers all the activities involved in constructing the final dataset that will be used in the data modelling phase. Data preparation tasks are likely to be performed multiple times and not in any prescribed order. Tasks include feature selection, data cleaning and data transformation [44, 72].

After the data understanding phase, the next goal is to identify data quality issues in the dataset. The most common data quality issues are missing values, irregular cardinality, and outliers. Garcia et al. [22] defines a missing value as an item of data that has not been stored or gathered due to factors such as cost restrictions, a flawed sampling process, or limitations in the data acquisition process. Outliers are values that lie far away from the central tendency of a feature. Moreover, irregular cardinality can arise when the cardinality of a feature does not match what is expected [34].

The key methods for handling the missing values is the deletion or imputation approaches. The deletion method, which is also called complete-case analysis, is applied by simply dropping any instance that has missing values. It is important to note that this method can result in a significant amount of data loss and can also introduce a bias into the dataset if the distribution of missing values in the dataset is not entirely random. A general rule is that if the proportion of missing values for a feature is very high, anything above 60%, it is best to simply remove the feature from the dataset as observations are too few [1, 34].

Imputation replaces missing feature values with a plausible estimated value based on the feature values that are present. Imputation techniques generally produce good results and avoid the data loss associated with a deletion approach. The most common approach to imputation is to replace missing values with a measure of the central tendency of that feature [1].

Generally, the mean (or median if there are outliers) is used for continuous features and the mode is used for categorical features. Imputation, however, should not be used for features that have vast numbers of missing values as imputing a large number of missing values will likely change the central tendency of a feature significantly. Imputation should be used with caution on features missing more than 30% of their values and should be avoided on features missing over 50% of their values. Another option to deal with missing values is to 'do nothing', especially if the modelling technique used can work effectively with missing values [1, 28, 34].

Algorithms that use distance measures as their primary measures are generally sensitive and do not perform well in the presence of outliers; therefore, outliers need to be removed before the data modelling phase. There are two kinds of outliers that can be found in a feature: invalid outliers and valid outliers. Invalid outliers are values that have been included in a sample through error and are often referred to as 'noise' in the data. Valid outliers are correct values that are simply significantly different from the rest of the values of a feature. The easiest way to handle outliers is to use a clamp transformation. This method clamps all values above an upper threshold and below a lower threshold, thus removing the outliers [22, 34].

The formula for applying a clamp transformation is defined in equation 3.2 [34]:

$$a_i = \begin{cases} a_l & \text{if } a_i < a_l \\ a_u & \text{if } a_i > a_u \\ a_i & \text{otherwise} \end{cases} \tag{3.2}$$

where a_i is the i^{th} instance of feature $\boldsymbol{a}$, and a_l and a_u are the lower and upper thresholds.

Domain knowledge can be used in setting the upper and lower thresholds manually or the thresholds can be calculated from data. One common way to calculate clamp thresholds is to set the lower threshold to the 1^{st} quartile value minus 1.5 times the interquartile range and the upper threshold to the 3^{rd} quartile plus 1.5 times the interquartile range. The method works effectively and takes into account the fact that the variation in a dataset can be different on either side of a central tendency.

It is important to note that most algorithms that use distance to measure similarity require data input for all features to be converted to numerical values. The one-hot encoding method can be used to transform values to numerical values. In the one-hot encoding method, a nominal input parameter that has z different values is coded as z different binary input parameters, where the input parameter that corresponds to a nominal value has the value of 1, and the rest of the parameters have the value of 0 [22].

Another technique that is generally applied in the data preparation phase is data standardisation. The most frequently used data standardisation method is normalisation, where values of attributes are scaled to fall within a specified range. Having continuous features in a dataset that cover very different ranges can cause difficulties for algorithms that use distance as a key measure [28, 36].

Normalisation techniques are used to change a continuous feature to fall within a specified range while maintaining the relative differences between the values for the feature. It aims to standardise all features to the same level; thus preventing attributes with small numeric ranges from being dominated by those with large numeric ranges [30]. The MinMaxScaler is one of the most commonly used scaling algorithms and it shrinks the range of a feature such that the values are transformed between 0 and 1 (or -1 to 1 if negative values exist). MinMaxScaler follows the following formula for each feature [28, 36]:

$$MinMaxScaler = \frac{a_i - min(\boldsymbol{a})}{max(\boldsymbol{a}) - min(\boldsymbol{a})} \tag{3.3}$$

where a_i is the data point of a feature and $\boldsymbol{a}$ is the vector value for all data points of a feature.

Although adding more descriptive features to a dataset provides more information about each instance, it is worth noting that more features can result in more training time. Furthermore, adding features can result in a more complex model which can be difficult to interpret. High dimensional data may be sparse, making it difficult for a clustering algorithm to find any structure in the data. A feature selection method is generally used to help reduce the number of descriptive features in a dataset to include only features that are the most useful [28, 36].

Feature selection is defined as the process of identifying the most effective subset of the original features to use without compromising overall algorithm performance. The smallest subset of descriptive features can be obtained by eliminating redundant and irrelevant features from

the dataset. A descriptive feature is considered redundant if it has a strong correlation with another descriptive feature. A descriptive feature can be considered to be irrelevant if it does not provide information that is useful in meeting the objective of the algorithm [36].

3.5 Data modelling

Many algorithms are based on a distinct set of assumptions that may be appropriate in one domain and not be effective in another domain. There is no algorithm that can be universally used to solve all problems just as there is no technique that can always outperform all other techniques under all circumstances. Each technique has its merits with data of some specific nature but fails on other types of data [73]. In this phase, different algorithms that can be used for clustering are selected and discussed. Techniques that are discussed in this section are k-means, AHC, and SOM.

Clustering involves the grouping of similar objects into a set known as a cluster. The main goal of clustering is maximising both the homogeneity within each cluster and the heterogeneity among different clusters. In other words, objects that belong to the same cluster should be more similar to each other than objects that belong to different clusters. Clustering, which is used in several research communities to describe methods for grouping of unlabelled data, has been applied in a wide variety of fields including pattern recognition, data mining, image segmentation, medical sciences and marketing. This diversity reflects the critical position of clustering in scientific research [32, 74].

One of the common applications of clustering is market segmentation. Market segmentation is used to identify characteristics of sub-populations which can be targeted for specific purposes, such as marketing aimed at a certain age group or based on purchase histories. An example of clustering is grouping magazine subscribers based on a number of factors such as age, gender, and income. The resulting groups can then be characterised in an attempt to find a business approach which will distinguish those subscribers that will likely renew their subscriptions from those that will not [32, 63].

A cosmetic packaging company which faced difficulties in reaching out to the right set of audience through its marketing initiatives, applied market segmentation. When implementing market segmentation initiatives, the company's main objective was to analyse the regulations and investment options in the cosmetic packaging space to secure potential customers. Moreover, the company wanted to gain better transparency into the market space and devise effective strategies to improve sales performance. With the help of the product segmentation solution, the cosmetic packaging company was able to showcase new products and penetrate into niche market segments. Furthermore, the product segmentation helped in devising a marketing plan to distribute products effectively across the potential markets [35].

Clustering is the process of identifying natural groupings or clusters within multidimensional data based on some similarity measure. Hence, similarity measures are fundamental components of most clustering algorithms. A common way to measure the similarity between two instances, $\boldsymbol{x}$ and $\boldsymbol{y}$, in a dataset, is to measure the distance between the instances in a feature space. A lower distance between the two objects indicates a higher similarity and vice versa. There are many distance metrics and the results obtained by applying each can differ significantly. Common distance measures that are used to measure the similarity between instances are the Euclidean, Manhattan and cosine distance measures [34].

The most widely used distance measure is the Euclidean distance where the distance between two instances $\boldsymbol{x}$ and $\boldsymbol{y}$ in an m-dimensional feature space is defined as [34]:

$$Euclidean(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\sum_{j=1}^{m}(x_j - y_j)^2} \tag{3.4}$$

where x_j is the value of the j^{th} feature of instance $\boldsymbol{x}$ and y_j is the value of the j^{th} feature of instance $\boldsymbol{y}$.

The Manhattan distance between two instances $\boldsymbol{x}$ and $\boldsymbol{y}$ in an m-dimensional feature space is defined as:

$$Manhattan(\boldsymbol{x}, \boldsymbol{y}) = \sum_{j=1}^{m}\left|x_j - y_j\right| \tag{3.5}$$

The cosine similarity between the two instances is the cosine of the inner angle between the two vectors that extend from the origin of feature space to each instance. The cosine similarity is suitable for clustering data of high dimensionality. Its value is between 0 and 1, where 1 indicates maximum similarity and 0 indicates maximum dissimilarity. In an m-dimensional feature space, the cosine similarity between two instances, $\boldsymbol{x}$ and $\boldsymbol{y}$, is defined as:

$$Cosine(\boldsymbol{x}, \boldsymbol{y}) = \frac{\sum_{j=1}^{m}(x_j * y_j)}{\sqrt{\sum_{j=1}^{m} x_j^2} * \sqrt{\sum_{j=1}^{m} y_j^2}} \tag{3.6}$$

Sections 3.5.1, 3.5.2 and 3.5.3 discuss the application of k-means, AHC and SOM. Each section provides background about the technique and explains the implementation process of each technique.

3.5.1 *K*-means algorithm

The k-means algorithm is one of the most well-studied clustering algorithms. The algorithm assigns each instance to a cluster whose centre it is nearest to. The algorithm partitions a set of n instances into K clusters so that the resulting intra-cluster similarity is high while the inter-cluster similarity is low [42, 71].

The k-means algorithm is computationally attractive because of its linear time and space complexity which makes it suitable for very large datasets. This algorithm has a time complexity of O(nKr) and a space complexity of O(K), were n represents the number of instances, K is the number of clusters, and r is the number of iterations taken by the algorithm to converge [32].

In k-means clustering, data is assigned into K clusters by optimising some criterion function. The most frequently used criterion function in k-means clustering is the sum of the squared error (SSE). The SSE is the average distance between instances and their closest centroid. The k-means algorithm starts with an initial clustering and keeps reassigning the instances to clusters based on the similarity between the instance and the cluster centres until the convergence criterion is met [32]. Generally, a convergence criterion is met once the centroids of newly formed clusters are no longer changing. If after multiple iterations, the algorithm continues obtaining the same centroids for all the clusters, it can be concluded that the algorithm is not finding any new solutions. Therefore, the optimisation process can be stopped [11].

The steps for computing the k-means algorithm begin by defining the objective function that the algorithm needs to optimise. The key objective of the k-means algorithm is to minimise the sum of distances between points and their respective cluster centroid. The selected objective function, SSE, is computed as [31, 48]:

$$SSE = \frac{\sum_{k=1}^{K} \sum_{p=1}^{|\boldsymbol{C}_k|} D(\boldsymbol{x}_p, \boldsymbol{c}_k)}{\sum_{k=1}^{K} |\boldsymbol{C}_k|} \tag{3.7}$$

where K is the number of clusters, D is a measure of similarity, $\boldsymbol{x}_p$ is an instance, $\boldsymbol{c}_k$ is the k^{th} cluster's centroid, $|\boldsymbol{C}_k|$ is the number of instances in cluster $\boldsymbol{C}_k$.

After computing the objective function, the following steps in the algorithm are applied [31, 32, 63]:

1. Specify the number of clusters (K).
2. Select initial centroids randomly, based on the number of clusters specified.
3. Assign each instance to the cluster with the closest centroid. The centroids are updated incrementally after each assignment of an instance to a cluster. The closest centroid to an instance is the one with the smallest value with regard to the distance measure applied.
4. When all objects have been assigned, recalculate the positions of the K centroids. Centroids are recalculated as the average vector over all the data points that belong to that centroid.
5. Repeat steps 2 to 4 with the updated means until a defined convergence criterion is met.

When applying the k-means algorithm, the user needs to first specify the number of clusters (K). Although there is no universal method for identifying the optimal number of clusters; there are various techniques which can be used to determine a value for K.

The elbow method is the most well-known method for determining the optimal number of clusters in k-means clustering. In order to determine best K, the k-means algorithm is executed using different K values and a final SSE is calculated. As K value is increased iteratively, the SSE is expected to drop drastically at some value of K, and after that, the value flattens as K increases further. The ideal K is achieved at this point because after this point the new clusters are expected to be very close to some of the existing ones when the number of clusters is increased, thus providing minimal improvement [42].

For instance, in Figure 3.3, the SSE function decreases rapidly and flattens after K reaches 4. It is important to note that using an elbow method to determine the value of K is not advisable because it is quite subjective and does not always produce reliable results, especially if the data does not contain clearly defined clusters [42].

The silhouette coefficient (SC) is another method that can be used to determine the optimal value of K. The SC uses the pairwise difference of between- and within-cluster distances to assess clustering performance. The SC is bounded between -1 and $+1$. Values close to -1 indicate sparse clustering while values close to $+1$ indicate clusters that are dense and well separated. Therefore, a K resulting in the highest value of the SC is considered as the optimal value of K. SC is defined as [40]:

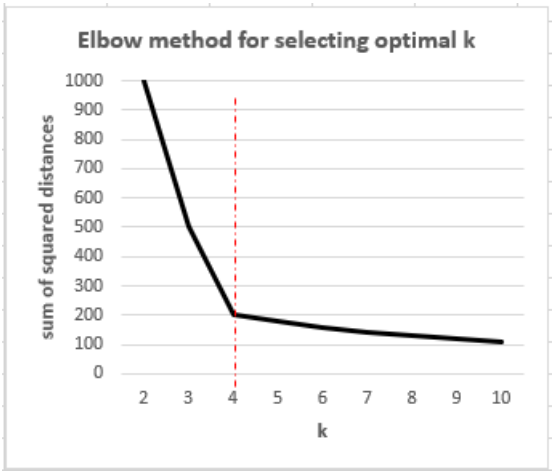


Figure 3.3: Elbow method for selecting value of K

$$sc = \frac{1}{n}\sum_{i=1}^{n}\frac{h_i - d_i}{max\{d_i, h_i\}} \tag{3.8}$$

where n is the total number of instances, d_i is the average distance between point i and all other points in its own cluster, and h_i is the minimum of the average dissimilarities between i and points in other clusters.

In the k-means algorithm, the initialisation of centroids is crucial because it has a direct impact on the final results. The selection of a good initial partition increases the likelihood of the algorithm to find a global minimum value. Random initialisation is commonly used in the k-means initialisation step. In order to increase the likelihood of finding an optimal solution, the random initialisation is repeated multiple times and the best results are selected [6, 66, 74].

Another option is choosing the initial centres more systematically by applying the k-means++ algorithm initialisation method. The objective of the k-means++ initialisation method is to spread out the K initial cluster centres where the first cluster centre is chosen uniformly at random from the instances that are being clustered. After that, each subsequent cluster centre is chosen from the remaining instances with a probability proportional to its squared distance from the instance's closest existing cluster centre. Once the initial centres have been selected, the algorithm proceeds using the standard k-means approach [4, 75].

Despite being a popular method for performing clustering across different disciplines, users have noted some significant drawbacks of k-means. Firstly, k-means is sensitive to noise and outliers. Even if an instance is quite far away from the cluster centroid, the instance is usually forced into a cluster, thus resulting in distortion of the cluster shapes. It is advisable to remove or impute missing values and remove outliers from the dataset before the k-means algorithm is applied. Another drawback is that k-means requires the user to specify the number of clusters (K) in advance. Furthermore, k-means is highly sensitive to the initialisation phase and may converge to a local minimum if the positions of initial centroids are not properly chosen [17, 31]. For instance, Figure 3.4 shows seven two-dimensional patterns. The chosen number of clusters (K) is 3, and centroids (indicated by circles marked with 'z') are randomly placed at different locations. The patterns are assigned to the centroids in closest proximity. In Figure 3.4, the partition from k-means results 1 (on the left) obtained clusters groups of (A, C), (B, F, G), (D, E). However, k-means result 2 (on the right) obtained clusters groups of (A, B, C), (F, G), (D, E). The different results demonstrate that the initial location of the centroids has a significant impact on the final clusters.

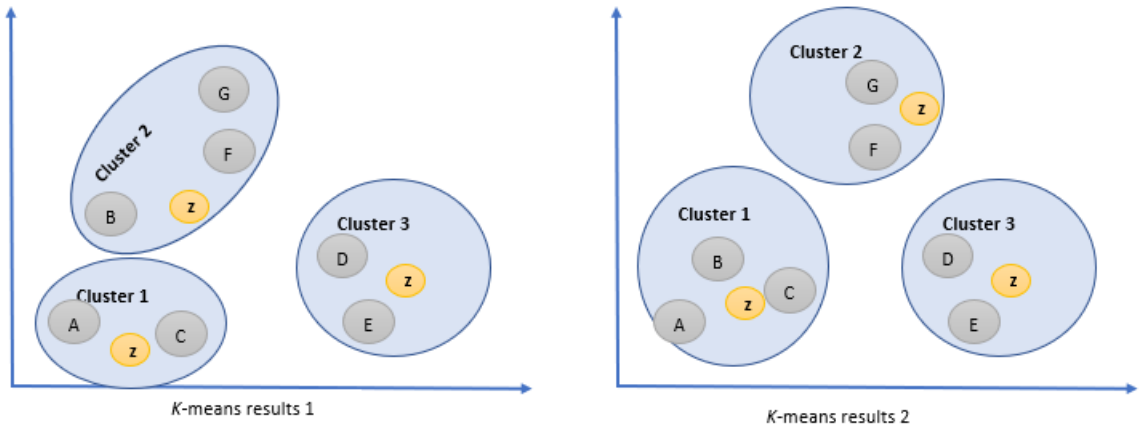


Figure 3.4: K-means clustering example

It is important to also note that the k-means algorithm optimises cluster centres; thus it always assigns an instance to the nearest centroid which can lead to incorrectly defined borders of clusters. Thus the algorithm struggles to handle data with clusters that are of different sizes or different densities. As illustrated in Figure 3.5, the resulting clusters were a mixture of two 'real' clusters [33, 41].

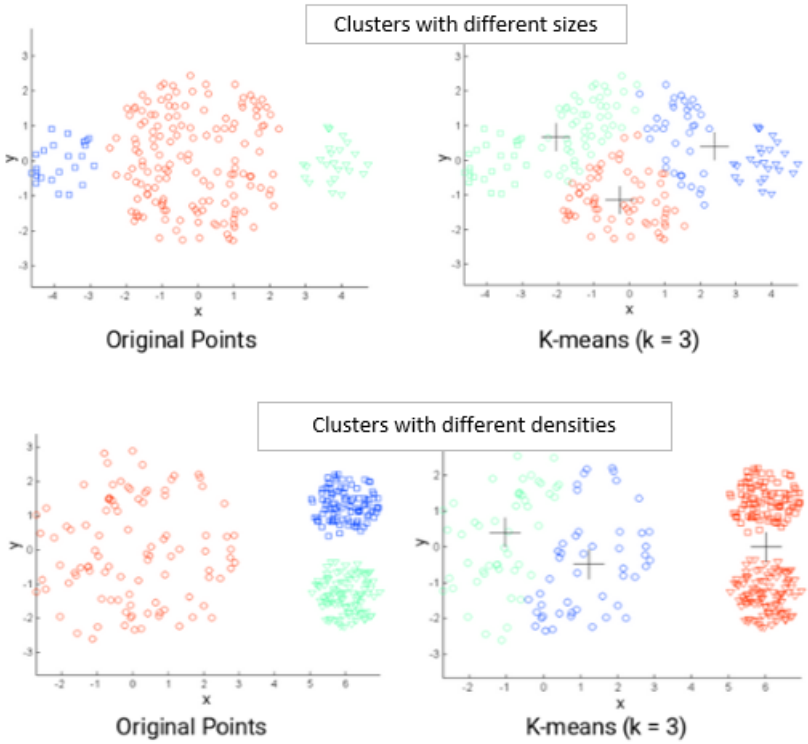


Figure 3.5: K-means clustering when clusters are of different sizes or densities [9]

3.5.2 Agglomerative hierarchical clustering

An agglomerative hierarchical clustering (AHC) algorithm follows a bottom-up approach, which first assigns each instance to its own cluster before merging the instances that have the closest similarity to each other into larger clusters. A simple way of showing the organisational structure of the dataset is by using a tree diagram called a dendrogram. The root node of the dendrogram (the x-axis) represents the entire dataset and each leaf node carries one instance. As the similarity increases, the leaf observations begin to merge into nodes, which carry instances that are similar to each other. The intermediate nodes thus describe the extent to which the objects are proximal to each other; and the height of the dendrogram (y-axis) expresses the distance between each pair of objects or clusters [33, 74].

The operation of the AHC algorithm is illustrated using the two-dimensional dataset in Figure 3.6. The figure depicts seven objects labelled A, B, C, D, E, F and G in three clusters. Based on a similarity matrix applied, objects A, B, C are grouped to form the first cluster. The algorithm first clusters objects B and C, and then A is added. Objects D, E and F, G are grouped to form cluster 2 and 3 respectively [32]. A dendrogram corresponding to the seven objects is also shown in Figure 3.6.

The SC can be used to determine the number of clusters where the highest SC value indicates an optimal number of clusters. On the dendrogram, the optimal number of clusters is demonstrated by making a horizontal cut across the branches of the dendrogram. The number of clusters is the number of vertical lines which lie under the horizontal line on the dendrogram. For instance, the dendrogram in Figure 3.6 shows that the selected number of clusters is three.

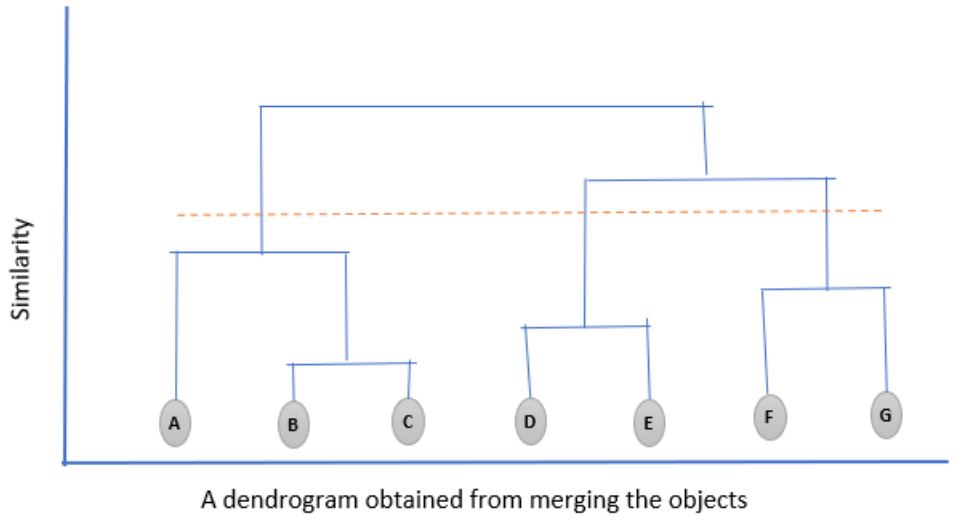


Figure 3.6: Agglomerative hierarchical clustering demonstration

Common methods used to measure similarities between clusters in AHC include single linkage and complete linkage methods illustrated in Figure 3.7. Single linkage algorithms merge the clusters whose distance between their closest patterns is the smallest. Complete linkage algorithms, on the other hand, merge the clusters whose distance between their most distant patterns is the largest [74].

According to Jain et al. [32], clusters obtained by the complete linkage method tend to produce clusters that are tightly bound or more compact than those obtained by the single linkage

method. The single linkage method tends to produce clusters that are elongated. From a practical viewpoint, the complete linkage method is considered more versatile as it tends to produce more useful clusters in many applications.

Equations 3.9 and 3.10 show the computation of the single linkage and complete linkage methods [74]:

$$d_{SL}(\boldsymbol{A}, \boldsymbol{B}) = min\, d_{jj'} \quad j \epsilon \boldsymbol{A}\ , j' \epsilon \boldsymbol{B} \tag{3.9}$$

$$d_{CL}(\boldsymbol{A}, \boldsymbol{B}) = max\, d_{jj'} \quad j \epsilon \boldsymbol{A}\ , j' \epsilon \boldsymbol{B} \tag{3.10}$$

where $d_{SL}(\boldsymbol{A}, \boldsymbol{B})$ is the single linkage distance and $d_{CL}(\boldsymbol{A}, \boldsymbol{B})$ is the complete linkage distance between cluster $\mathbf{A}$ and $\mathbf{B}$. $d_{jj'}$ is the distance between instances j and j'.

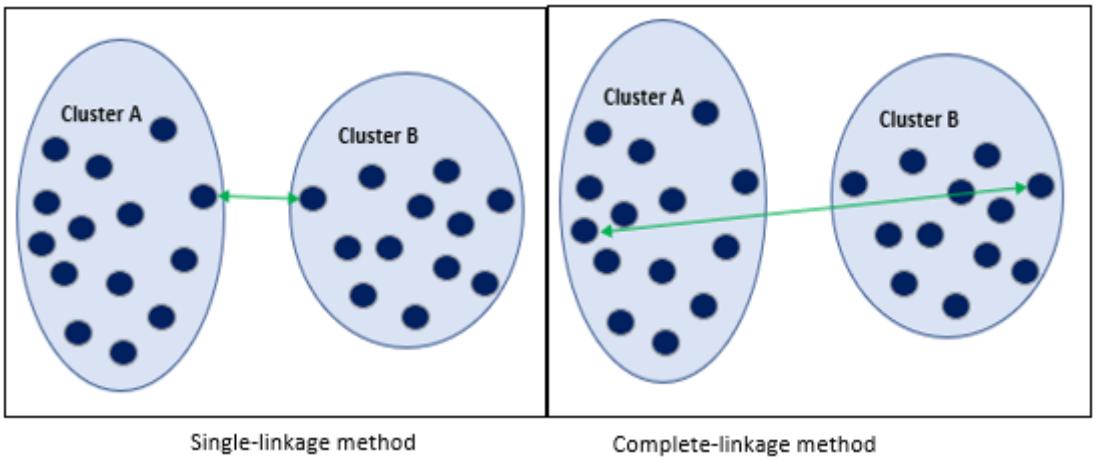


Figure 3.7: Linkage methods

The following steps are followed when applying the AHC algorithm:

1. Start with K clusters, where each cluster consists of one data point.
2. Find the most similar pair of clusters using similarity measures and combine the pair of clusters to form a new cluster.
3. Update the proximity matrix by computing the distances between the new cluster and the other clusters.
4. Repeat step 2 and step 3 until a defined convergence criteria is met. Generally, the algorithm is stopped when all clusters are merged.

The AHC algorithms are considered to be more robust and versatile when compared with the k-means algorithm. Unlike the k-means algorithm, AHC is less impacted by missing values in a dataset. Another advantage is that the number of clusters does not need to be specified in advance and they are independent of the initialisation phase. However, a common criticism is that AHC is computationally expensive, with a computational cost that increases quadratically with the number of instances; thus, AHC is not suitable for very large datasets. Another downside is that AHC is static; once an object is assigned to a cluster, such an instance cannot be reassigned [33, 63, 74].

3.5.3 Self-organising map

The SOM is a multidimensional scaling method to project an I-dimensional input space to a discrete output space, effectively performing compression of the input space onto a set of codebook vectors [20]. The algorithm is commonly used in exploratory data analysis for feature extraction, data visualisation, and cluster analysis. Its key objective is to represent high-dimensional instances with codebook vectors that can be visualised in an output space that is usually a two-dimensional grid, as shown in Figure 3.8 [63].

SOM training is based on a competitive learning strategy. In the process, SOM effectively clusters the instances through a competitive learning process, in which different neurons (elements) in the network compete when an input instance is presented. The winning neuron is found by computing the distance measure, such as the Euclidean distance, from each codebook vector to the instance, and selecting the neuron closest to the instance. The weights of the neurons in the neighbourhood of the winning neuron are then adjusted to be closer to the value of the input instance [6, 20].

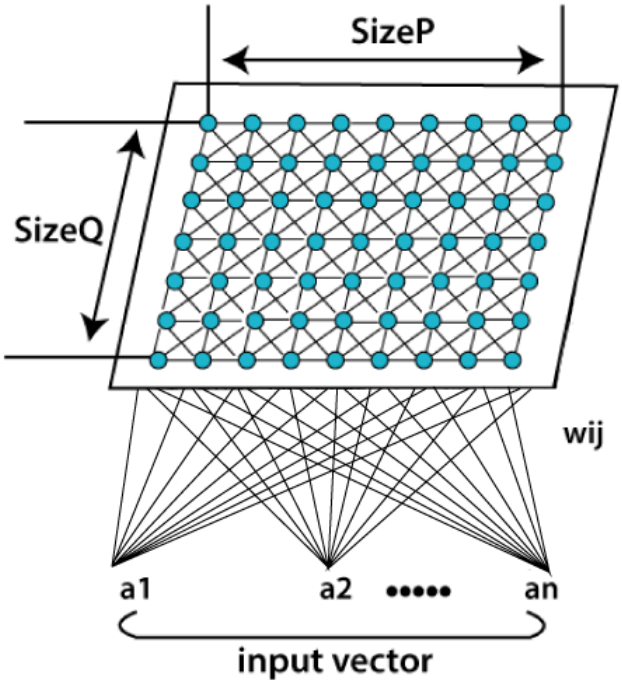


Figure 3.8: Typical SOM structure [6]

The main advantage of SOM is the easy visualisation and interpretation of clusters formed by the maps. SOM is also more robust and unlike other algorithms such as k-means, SOM does not suffer with problems presented by missing values and outliers in a dataset; thus, missing values do not need to be replaced when applying SOM [48].

One of the shortfalls of the SOM method is that the technique is very sensitive to the initialisation phase and may generate suboptimal clusters if the initial weights are not chosen properly. Moreover, its performance is affected by user-dependent parameters such as the size of the map and the neighbourhood function. The neighbourhood function is usually a function of the distance between the coordinates of the neurons as represented on the map. Furthermore, the algorithm naturally depends on the order in which data is presented, but this problem can be

addressed by randomising the choice of data points during each iteration [32, 48].

One of the parameters that the user needs to specify when using SOM is the size of the map. The map size, usually a two-dimensional grid, is expressed by the number of neurons that define the output space of the SOM. Too many neurons may cause overfitting of the instance, as each training pattern is assigned to a different neuron [48]. The computational load increases quadratically with the number of map units. Therefore, too many neurons also cause a substantial increase in computational complexity [70]. However, too few neurons will result in clusters with a high variance among the cluster members. Generally, the map size should be at least equal to the number of independent variables in the training set [20, 24]. According to Vesanto [70], the default number of neurons should be specified in advance using the formula $5 \times \sqrt{n}$, where n is the number of training instances.

In the initialisation phase, the codebook vectors can be initialised by assigning random values to each weight. When random initialisation is applied, the map that will emerge can be far from optimal. Generally, a good strategy is to repeat the random initialisation multiple times and select the best map according to the defined optimisation criterion [24, 70].

Another key parameter in SOM is the learning rate (η). The learning rate determines the extent to which the weights are adjusted during each iteration. If the learning rate is too small, the weight adjustments are correspondingly small. Thus, more learning iterations are required to reach a minimum. A small learning rate also has the disadvantage of being more easily trapped in a poor local minimum [20]. On the other hand, large learning rates will result in large weight updates. Convergence may initially be fast, but the algorithm may eventually oscillate without reaching the minimum. One approach is to select a small value (e.g. 0.1) and increase it if convergence is too slow, or decrease it if the error does not decrease fast enough.

The last key parameter is the neighbourhood function which is a function of the distance between the coordinates of the neurons. The initial spread of neighbouring neurons (σ) is the width of the kernel [8, 20]. A kernel is a parameterised representation of a surface in the space that allows the SOM to operate in the original feature space without computing the coordinates of the data in a higher-dimensional space [70]. There are many neighbourhood functions that can be used to determine the rate of change of the neighbourhood around the winner neuron. The neighbourhood function can decay with distance or can be constrained to be constant around the winner unit. The most popular choice for a neighbourhood function is to use a Gaussian kernel as computed in equation 3.11 [8, 20]:

$$h_{mn,kj}(t) = \eta(t)\, e^{\frac{-\left\|c_{mn}-c_{kj}\right\|_2^2}{2\sigma^2(t)}} \tag{3.11}$$

where coordinates $c_{mn}, c_{kj} \epsilon \mathbb{R}^2$ and mn are the coordinates of the winning neuron.

It is worth noting that the neighbourhood index is independent of the location of the winning neuron and it decreases monotonically to zero as the distance tends to infinity [8, 20]. The functions $\eta(t)$ and $\sigma(t)$ are monotonically decreasing functions and typically decrease over time. The training process ends when the defined stopping criterion is reached. For instance, a stopping criterion may be met when a predetermined number of training cycles (epochs) is reached [24].

The convergence characteristics of SOM can be described by the ability of the network to converge to specified error levels (usually considering the generalisation error). The quantisation

error (QE) is one of the most common measures used as an indication of map accuracy. QE is computed from the average distance of the instance ($\boldsymbol{x}$) to the weight vector of the winning neuron ($\boldsymbol{w}_{mn}(t)$). A SOM with lower average error is considered to be more accurate than a SOM with higher average error [24]. The formula for calculating QE is defined as [45]:

$$QE = \frac{\sum_{i=1}^{n} \|\boldsymbol{x}_i - \boldsymbol{w}_{mn}(t)\|}{n} \tag{3.12}$$

where n is the number of instances used to train the map.

The SOM algorithm is summarised in the following steps [8, 48]:

1. Randomly initialise the codebook vectors ($\boldsymbol{w}_k(0)$).
2. Initialise the learning rate ($\eta(0)$) and the neighbourhood function ($h_{mn,kj}(0)$).
3. Find the winning neuron for each input instance $\boldsymbol{x}_i$. The winning neuron is the unit whose codebook vector has the highest similarity with the input pattern. The similarity is usually defined using a distance measure, typically the Euclidean distance.
4. Use competitive learning to train the codebook vectors such that all neurons within the neighbourhood of the winning neuron move towards $\boldsymbol{x}_i$:

$$\boldsymbol{w}_k(t+1) = \begin{cases} \boldsymbol{w}_k(t) + \eta(t)[\boldsymbol{x}_i - \boldsymbol{w}_k(t)] & k \,\epsilon\, h_{mn,kj}(t) \\ \boldsymbol{w}_k(t) & \text{otherwise} \end{cases} \tag{3.13}$$

   where $\boldsymbol{w}_k(t)$ is the k^{th} codebook vector at time t.
5. Linearly decrease $\eta(t)$ and reduce $h_{mn,kj}(t)$.
6. Repeat steps 3 to 5 until the specified convergence criteria are satisfied.

When the SOM technique is used for clustering, larger cluster groupings are formed by grouping together similar neighbouring neurons. The objective of the SOM training process is to cluster together similar instances while preserving the topology of the input space. The results obtained after training is the set of trained weights with no explicit cluster boundaries. Thus, an additional step is required to find these cluster boundaries [8].

One way to determine and visualise these cluster boundaries is to calculate the unified distance matrix (u-matrix), which contains a geometrical approximation of the codebook vector distribution in the map [20]. The u-matrix generally uses a colour coding strategy to represent distances between neighbouring units in the SOM output space. Generally, units that are near to their neighbours are represented in darker shades; and units distant from their neighbours are represented in lighter shades [8]. Large values within the u-matrix indicate the position of cluster boundaries. Ward clustering of the codebook vectors is generally used to determine the boundaries. Ward clustering follows a bottom-up approach where each neuron initially forms its own cluster. At consecutive iterations, two clusters that are closest to each other are merged [20].

It is worth noting that in order to preserve the topological structure, two clusters can only be merged if they are adjacent. Furthermore, only clusters that have a non-zero number of instances associated with them are merged. Convergence is reached when a set criterion (such

as when the optimal or specified number of clusters has been constructed) is met. The end result of Ward clustering is a set of clusters with a small variance over its members, and a large variance between separate clusters [20].

The Ward distance measure is used to decide which clusters should be merged. The distance measure is defined as [20]:

$$d_{rs} = \frac{n_r * n_s}{n_r + n_s} \|\boldsymbol{w}_r - \boldsymbol{w}_s\|_2^2 \tag{3.14}$$

where r and s are cluster indices, n_r and n_s are the number of instances within the clusters, and $\boldsymbol{w}_r$ and $\boldsymbol{w}_s$ are the centroid vectors of these clusters (i.e. the average of all the codebook vectors within the cluster).

The two clusters are merged if their distance d_{rs} is the smallest. For the newly formed cluster (q),

$$\boldsymbol{w}_q = \frac{1}{n_r + n_s}(n_r * \boldsymbol{w}_r + n_s * \boldsymbol{w}_s) \tag{3.15}$$

and

$$n_q = n_r + n_s \tag{3.16}$$

3.6 Evaluation

At this stage in the project, different algorithms that appear to have high-quality results from a data analysis perspective have been applied. Before proceeding to the final stage of deployment, it is important to evaluate the results more thoroughly and determine if they accurately achieve the business objectives. Effective evaluation criteria are important to provide users with a degree of confidence for the clustering results derived from the implemented technique [62].

These assessments should be objective and have no preference for any technique. The goal of clustering is to ensure that objects within the same cluster are similar and objects in different clusters are distinct. Internal validation assessment, which measures the performance of clustering, is often based on the two criteria: compactness and separation [44, 72].

With regard to compactness, the indices measure how the objects in a cluster are similar to each other. Generally, methods that evaluate cluster compactness measure intra-cluster distances. Intra-cluster distance, shown in Figure 3.9 and computed in equation 3.17, calculates the average distances of all the points within a cluster from the centroid of that cluster. A lower intra-cluster distance indicates better compactness [44, 48, 72].

$$Intra\,cluster\,distance = \frac{\sum_{k=1}^{K}\sum_{p=1}^{|\boldsymbol{C}_k|} D(\boldsymbol{x}_p, \boldsymbol{c}_k)}{\sum_{k=1}^{K}|\boldsymbol{C}_k|} \tag{3.17}$$

where K is the number of clusters, D is a measure of similarity, $\boldsymbol{x}_p$ is an instance, $\boldsymbol{c}_k$ is the k^{th} cluster's centroid and $|\boldsymbol{C}_k|$ is the number of instances in cluster $\boldsymbol{C}_k$.

On the other hand, the separation criterion measures how well-separated or distinct a cluster is from other clusters. Inter-cluster distance,own in Figure 3.10 and computed in equation 3.18,

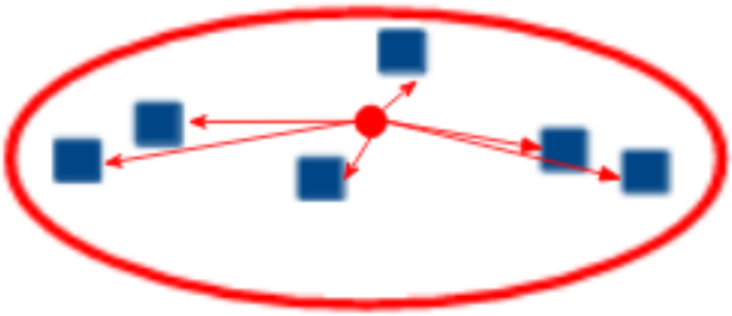


Figure 3.9: Intra-cluster distance measure

is the average distance between each of the cluster centres. Thus, higher inter-cluster distances indicate better separation [44, 63, 74].

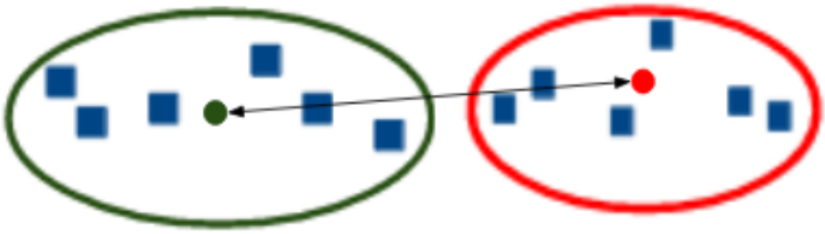


Figure 3.10: Inter-cluster distance measure

$$Inter\ cluster\ distance = \frac{\sum_{k_1 \neq k_2}^{K} D(\boldsymbol{c}_{k_1}, \boldsymbol{c}_{k_2})}{K} \tag{3.18}$$

where K is the number of clusters, D is the Euclidean distance which is used as a similarity measure, $\boldsymbol{c}_{k_1}$ and $\boldsymbol{c}_{k_2}$ are cluster centres of different clusters.

Under the assumption that better compactness within a cluster and separation between cluster centres leads to better clustering, the two criteria can be used as performance indices to determine overall clustering accuracy.

3.7 Deployment

Data science projects aim to serve a purpose within an organisation, and the last phase of CRISP-DM covers all the work that must be done to successfully integrate the results into the processes within an organisation. The ultimate goal of clustering is to provide users with meaningful insights from the original data, so that they can effectively solve the problems encountered. Generally, the knowledge gained needs to be organised and presented in a way that the organisation can use it. Depending on the requirements defined in the business understanding phase, the deployment phase can be as simple as generating a final report with recommendations or as complex as implementing a repeatable data mining process [5, 44, 72].

3.8 Chapter summary

This chapter presents a detailed approach for applying the CRISP-DM in a project. The tasks involved in completing each CRISP-DM phase are explained in detail. Since each algorithm has its strengths and weaknesses, more than one algorithm is selected in the data modelling phase. The evaluation methods to be used in assessing the performance of the three selected algorithms are defined. Finally, the technique that outperforms other techniques based on the defined performance measures is selected for deployment.

Chapter 4

Literature Review: Supplier Relationship Management

In this chapter, the importance of SRM and a detailed approach of its application is discussed. Section 4.1 discusses the changing roles of suppliers and the way organisations are engaging with their suppliers. It also discusses supplier relationship management's role in ensuring that organisations adapt and remain competitive in the ever-changing business landscape. Section 4.2 describes methods that can be used to determine an organisation's requirements from its supply base to ensure the effectiveness of its engagements with suppliers. Lastly, section 4.3 discusses different methods applied in the segmentation of suppliers.

4.1 Background

The world is fast-changing and organisations are forced to change and adapt in order to remain competitive. The role of the supply base and the way organisations are engaging with suppliers is also changing. In order to determine how purchasing and the supply base can add value to the competitiveness of an organisation, the landscape that the organisation operates in needs to be understood. Figure 4.1 shows the four key aspects that highly influence the competitiveness of an organisation [47].

In the past, supply chains were a relatively linear collection of individual entities, each operating independently. Organisations needed to concern themselves only with their immediate suppliers and customers with whom a contractual relationship existed. The landscape that organisations operate in has been changing exponentially, and most organisations have realised that this traditional approach is no longer adequate to manage risk or gain competitive advantage [38].

The number of suppliers that an organisation has to deal with has grown rapidly over the years and organisations are increasingly relying on their suppliers to reduce operational costs, improve quality and develop new products faster than their competitors. Organisations are using SRM to find new ways to involve key suppliers who can help them gain a competitive edge [51]. The approach is used to develop two-way, mutually beneficial relationships with strategic supply partners to deliver greater levels of innovation and competitive advantage that could not easily be achieved by operating independently [47].

SRM consists of three interrelated components that need to be integrated to establish an effective SRM approach. The three focus areas of SRM are the organisation's key requirements

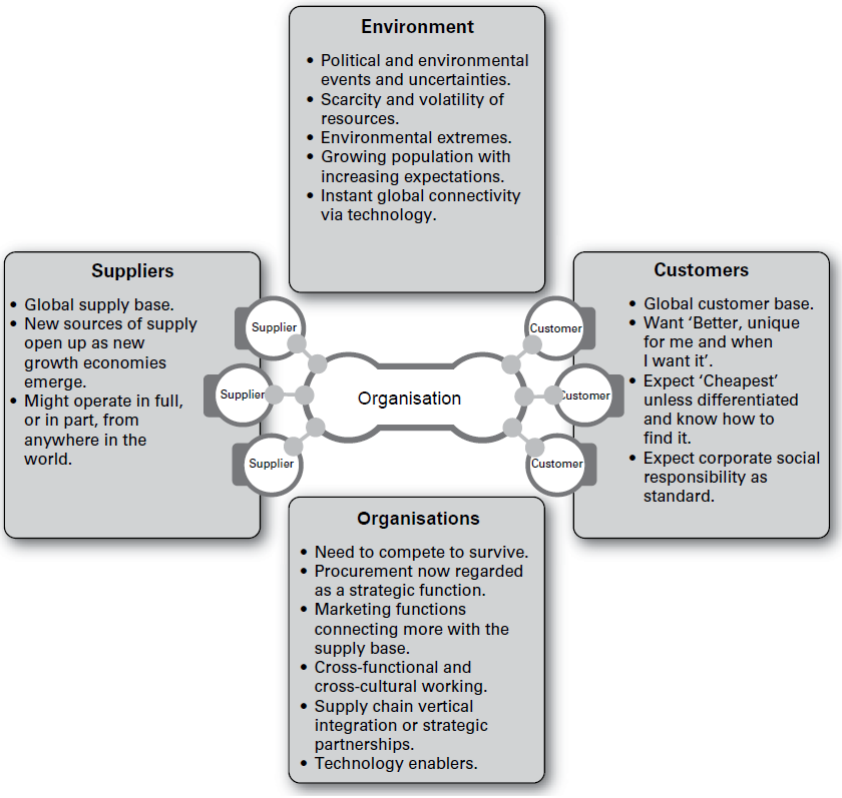


Figure 4.1: The landscapes that impact organisations [47]

from its supply base, the level of importance of each supplier in meeting the organisation's requirements, and possible interventions an organisation can implement to ensure its corporate strategy is achieved. Figure 4.2 shows approaches that are generally used to address the three focus areas of SRM. O'Brien [47] defines corporate strategy as the direction and scope of an organisation over the long term. The scope needs to match an organisation's resources to its changing environment, and in particular, match its markets and customers to meet stakeholders' expectations [15].

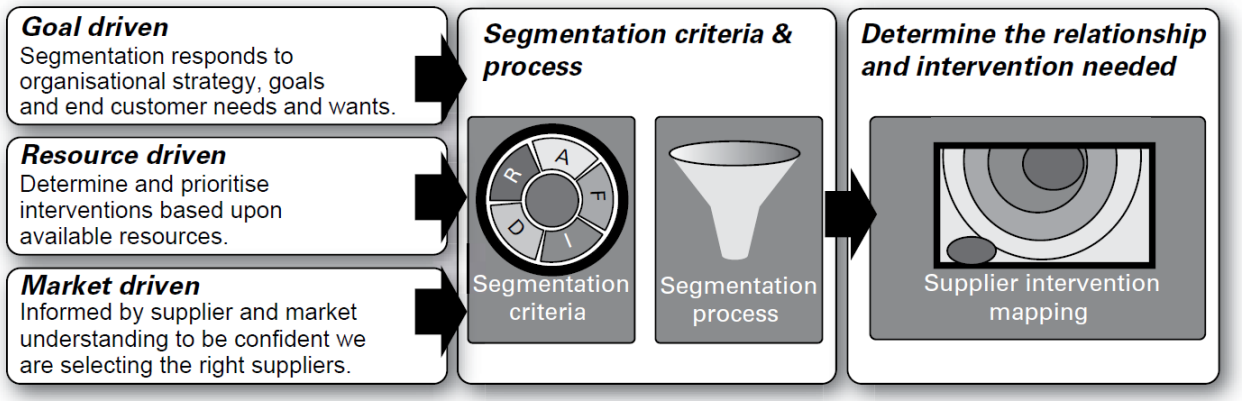


Figure 4.2: The components of segmentations [47]

4.2 Defining requirements for suppliers

For an SRM initiative to be effective, it needs to be goal-driven and to be able to respond to organisational strategy and the needs of customers. Generally, the supply base harbours huge potential in assisting an organisation to achieve its strategic goals. A VIPER model, shown in Figure 4.3, can be used to determine an organisation's requirements from its supply base. VIPER is an acronym representing the key factors that are used to define the requirements of an organisation. The key elements are value, innovation, performance improvements, effectiveness of operations, and risk [47, 50].

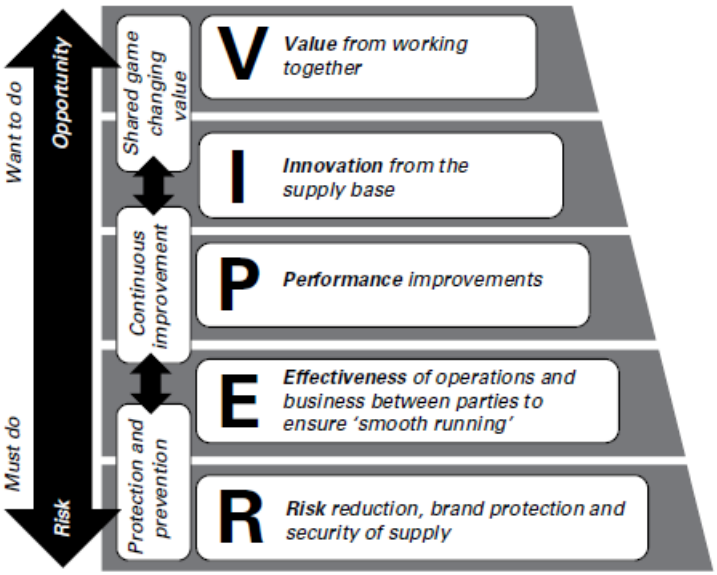


Figure 4.3: The VIPER model [47]

Risk forms the foundation of the VIPER model. There are many instances where a failure in the supply chain can present a significant risk to an organisation. The severity of the damage that risk can present varies significantly. For instance, suppliers delivering incorrect products or not delivering products on time may cause some inconvenience to the organisation, but if an organisation has to stop its production line because one component is not available, the cost of lost time can be immense. The most critical reason for having risk as a foundation for the VIPER model is so that an organisation can work with suppliers in taking steps to prevent crisis or at least be prepared for it. Managing supply risk is one of the greatest sources of value an organisation can secure from its supply base [7].

The second element on the VIPER model is effectiveness of operations, which ensures that an organisation's operations run smoothly and effectively. Effective operation is generally achieved when an organisation necessitates ongoing communication with certain suppliers. The third element, performance improvements, requires an organisation to monitor the performance of its supply base according to the service level agreement. Supplier performance could encompass many areas such as quality, on-time delivery and cost of products [47, 69].

Furthermore, operating in a dynamic environment means businesses need to evolve continuously, and innovation is important in delivering game-changing value. The last factor on the VIPER model is value, which represents additional benefits beyond the traditional list of standard benefits that are possible through working together with certain suppliers. Generally, the

selected suppliers need to possess the capability to help an organisation unlock a new level of potential that can make it stand out from its competitors [47, 69].

4.3 Supplier segmentation

Although an organisation can accumulate many suppliers in the course of doing business, many of these suppliers will be of little or no importance to the organisation beyond fulfilling a simple order transaction. However, there are other suppliers who will play a significant role in the success of an organisation [15].

The decision to invest in supplier relationships is a major step for an organisation, especially when the value gained from interacting in a supply network rests on the principle of prioritising the right suppliers to work with. If SRM is appropriately applied, an organisation can have confidence that it is directing its precious resources where they will have the greatest impact. An organisation must allocate its resources on a selective basis to suppliers from whom it expects to generate the highest return. The allocation of resources entails a careful segmentation of a supply base with the objective of building a portfolio of supplier relationships with varying characteristics that support the firm in different ways [59].

O'Brien [47] defines supplier segmentation as a process whereby suppliers are divided into distinct groups according to their perceived importance to an organisation. Rezaei and Ortt [58] further describe supplier segmentation as the identification of suppliers' capabilities and willingness to cooperate with an organisation. The level of capabilities and willingness to cooperate enables an organisation to engage in strategic partnerships with suppliers regarding a set of evolving business activities and functions in supply chain management. Not only does the supplier segmentation approach provide a means of assessing the supply base, it is also a resource-efficient decision methodology that specifies appropriate relationships and governance structures for each cluster. Managing suppliers in clusters eliminates the need to create a fully tailored procurement strategy for each supplier [7].

The primary factors that drive the SRM initiatives are the corporate strategy and available resources. The corporate strategy informs the development of criteria that are used in assessing suppliers. Section 4.3.1 describes different methods for defining criteria used to measure suppliers. Section 4.3.2 explains different methods for evaluating suppliers and grouping suppliers into distinct groups based on their deemed level of importance to the organisation. Section 4.3.3 describes interventions required for the organisation to manage each group of suppliers in order to achieve its strategic goals.

4.3.1 Segmentation criteria

One of the most common segmentation approaches is the portfolio method, which was first introduced by Kraljic [37]. The method's main objective is to identify the strategic weight of various products to help an organisation in its purchasing strategies. The model is regarded as the most influential purchasing portfolio model and many researchers have endorsed it. The portfolio method uses two predefined segmentation criteria, profit impact and supply risk, to segment suppliers based on the different products supplied to an organisation. The main aim of this method is to minimise the risk of supply while making the most of an organisation's buying power [13, 56].

Although the portfolio method has been praised for its ability to develop purchasing strategies to balance risks with opportunities, the method has also been criticised for limiting organisations since the focus is on only two variables when segmenting suppliers [7, 29]. The lack of an overarching framework that enables organisations to include all the important criteria posed a serious gap as it became more impractical for companies to consider only two segmentation criteria while neglecting other important criteria. Furthermore, the portfolio method only focuses on products and not on suppliers, meaning it is not possible to evaluate suppliers that offer multiple products or services [64, 69].

To address the shortfalls of the portfolio method, Rezaei and Ortt [55] developed a supplier segmentation framework named the supplier potential matrix (SPM). The SPM consists of two dimensions referred to as supplier capabilities and supplier willingness. Capabilities mostly focus on a supplier's skills and willingness focuses on a supplier's motivation to collaborate with an organisation.

Rezaei and Ortt [56] define supplier capabilities as 'complex bundles of skills and accumulated knowledge, exercised through organisational processes that enable firms to coordinate activities and make use of their assets in different business functions that are important for a buyer'. Supplier willingness is defined as 'confidence, commitment and motivation to engage in a long-term relationship with a buyer'. The SPM method enables an organisation to consider all relevant criteria within a given situation. A list of possible criteria for the two dimensions is shown in Table 4.1.

Table 4.1: Examples of different segmentation criteria [56]

Dimensions	Criteria
Capabilities	Industry knowledge
	Design capability
	Price/cost
	Geographic location
	On-time delivery
	Reputation in industry
	After-sales support
	Disclosure of environmental records
	Reliability of products
Willingness	Continuous improvement initiatives
	Supply chain relationship integration
	Waste elimination initiatives
	Communication openness
	Openness to information sharing
	Openness to site audits
	Equipment upgrade initiatives
	Participation in new products' development
	Bidding procedural compliance

While a list of capabilities and willingness dimensions, as demonstrated in Table 4.1, aims to provide an organisation with an extensive list of criteria to choose from, O'Brien [47] believes a comprehensive list of criteria is not necessary as the criteria can be summarised into a few main key areas. Generic segmentation criteria that consist of five key focus areas are:

- **Risk**: Supplier risk is the degree to which an organisation's success can be damaged by a supplier. Risk can take many forms and depends upon the nature of an organisation.

The five key risk areas worth assessing are risk of delay, brand reputation, competitive advantage, cost and quality. Risk of delay refers to the supplier failing to deliver the required products on time. Brand reputation risks are events that can be disastrous to the organisation's brand or reputation. Risks of competitive advantage include a possibility of intellectual property theft by a supplier. Lastly, quality risk refers to the possibility of suppliers delivering products that do not meet the organisation's specifications.

- **Alignment**: Alignment is primarily concerned with the degree to which a supplier could help the organisation or potentially hurt it. The criterion includes alignment of principles, goals, culture and beliefs. For instance, if an organisation brands itself as an environmentally responsible company, but the operations of its strategic suppliers are found to be directly harmful to the environment, the misalignment between an organisation and its supplier can damage the organisation's reputation.

- **Future importance**: Future importance is the degree to which a supplier's importance is likely to increase in the future based on the direction the organisation is intending to take.

- **Current importance**: The criterion assesses all the factors that make a supplier important to an organisation at the current moment. These factors include cost, contractual commitments, operating location, the degree to which the supplier knows the organisation's business and any established relationships that drive preference.

- **Difficulty**: Difficulty is the only criterion that relates specifically to the goods or services being sourced. The criterion assesses all the factors that might restrict freedom of choice when sourcing a particular category of products or services. Difficulty is rated high if a small number of suppliers can supply the required product or if products being sourced are complex. Consequently, it is necessary to work closely with suppliers if there is some form of scarcity in the products or services they supply.

4.3.2 Supplier evaluation

Once a set of criteria has been selected, each supplier is rated against each criterion. Generally, carefully selected participants from different functional areas in the organisation are required when conducting supplier segmentation workshops. The participants, referred to as decision-makers from here onward, are expected to work closely with suppliers and understand the organisation. Figure 4.4 shows an example of a scorecard that can be used to rate each customer based on the selected criteria. Generally, a simple one to five rating for each criterion is used to score suppliers [47, 56].

Once the criteria for segmenting suppliers are defined, the decision-makers need to give each supplier a score against the defined criteria. The key challenge is that the scoring process produces separate scores for each criterion. For example, if five criteria were selected to score the supplier, after the scoring process each supplier will have five independent scores. Summing these scores to produce a grand score could produce skewed results as the organisation could risk including a supplier who has a high total score but yet scores poorly on other critical criteria such as risk [47].

Aggregating ratings of suppliers is complex and require multicriteria decision-making methods that can deal with the process of making decisions in the presence of multiple criteria. Generally, mathematical models are used to aggregate the suppliers' scores according to all the

Supplier Name: **Date:**

	1	2	3	4	5
Risk	No discernable risk	Chance of minor risks only	High likelihood of low severity risks occurring with this supplier or in the supply chain	Some likelihood of a high severity risk occurring with this supplier or in the supply chain	Significant likelihood of high severity risk occurring with this supplier or in the supply chain
	1	2	3	4	5
Alignment	No alignment, conflicting future directions and/or incompatible culture, ethics, policies and ways of working	No apparent misalignment but some aspects of culture, ethics, policies and ways of working are at odds with ours	No apparent misalignment and culture, ethics, policies and ways of working appear compatible	Good degree of alignment of future direction. Compatible culture and ethics, policies and ways of working	Complete alignment of strategic goals, similar culture, ethics and beliefs. Common policies and ways of working
	1	2	3	4	5
Future importance	No future plans or opportunity with this supplier	Possibility that, with effort, we could unlock future contribution from this supplier	Predicted significant demand from this supplier. Likely high spend in the future	Supplier has potential innovation or know-how that could make a dramatic contribution to our business	Predicted significant demand from this supplier and/or great opportunity for them to make a dramatic contribution to our business
	1	2	3	4	5
Current importance	Low spend, no real importance	Med-high spend and some contractual commitments	Med-high spend, supplier is important due to location, know-how or established relationships	High spend and contractual obligations	Very high spend. Supplier has contractual obligations and/or is important due to location, know-how or relationships
	1	2	3	4	5
Difficulty	No difficulty, leverage or acquisition categories and generic, non-complex products. Easy to switch suppliers	Some difficulty of our own creation, eg by our own contracting arrangements. We can switch easily but with effort	Difficult market due to proprietary nature of what we are buying. There is good scope to change this and make the market easier	Difficult market and complexities in what we buy. Difficult but not impossible to switch suppliers	Inability to switch suppliers, little or nothing we can do to change this. Market difficult and/or an essential part of our brand definition

Figure 4.4: Example of a supplier scorecard [47]

criteria met. The aggregated score of each supplier is then used in the segmentation process, where suppliers are grouped together based on their aggregated score [47].

The literature presents a wide range of mathematical models that have been used to aggregate scores for each supplier. A fuzzy rule-based system is one of the commonly used methods to determine suppliers' aggregated scores. Fuzzy rules are linguistic if-then constructions that have the general form 'if A then B' where A and B are collections of propositions containing linguistic variables. According to Rezaei and Ortt [58], the fuzzy rule-based approach is a very flexible approach with the ability to handle the inherent interdependencies and contingencies of segmentation criteria. The most significant hurdle of this approach is that for it to be able to give accurate results, a large number of rules need to be created. For instance, to evaluate six criteria for each dimension, 64 rules need to be created. As a result, this approach would not only be tedious, but it would also be impractical when more variables are to be considered.

Another method used to determine the aggregate score for suppliers is the best-worst method (BWM). In this method, decision-makers need to select the best and the worst criterion among the selected criteria. Pairwise comparison is then conducted between the best criterion and other criteria until the resulting weights are determined. Although BMW have proposed several consistency measurements and is considered to produce more reliable results; there are some deficiencies, including: the lack of a mechanism to provide immediate feedback to the decision-maker regarding the consistency of the pairwise comparisons being provided and the inability to consider the ordinal consistency [59].

Chunguang Bai [13] applied a combination of rough set theory (RST) and VIKOR (Vlse Kriterijumska Optimizacija Kompromisno Resenje - which means multicriteria optimisation and

compromised solution in Serbian) to determine the aggregate scores of suppliers. RST was used to calculate the weight of each criterion for suppliers' capabilities and suppliers' willingness, and VIKOR was used to determine the final aggregated score for each supplier. RST is highly praised for its effectiveness in managing large sets of suppliers' performance criteria. However, the disadvantage of RST is that the method is relatively subjective, and the computational cost can be very high [67]. VIKOR introduces an aggregating function that represents the distance from the ideal solution by considering the comparative importance of all criteria while balancing between total and individual satisfaction [21].

After a selected mathematical model is applied to a supplier's ratings, there would be one aggregated score for each dimension. Therefore, each supplier will have two final scores, one for the capabilities dimension and one for the willingness dimension. The SPM results are typically presented on an x-y axis and subdivided into two levels (high and low), resulting in a 2 x 2 characterisation matrix. The decision-makers need to specify rules that determine the boundaries of the matrix.

SPM does not limit organisations to only two levels; they can use as many levels as they require. In the example illustrated in Figure 4.5, the two levels (low and high) were applied where values smaller than 20 were categorised as low and values greater than 20 as high. The figure on the right shows the resulting segments of suppliers. In the SPM method, one quadrant represents one cluster of suppliers.

Suppliers' aggregated score and score levels

Suppliers	Capabilities score	Level	Willingness score	Level
S1	16	Low	28	High
S2	17	Low	10	Low
S3	26	High	19	Low
S4	15	Low	27	High
S5	18	Low	27	High
S6	16	Low	18	Low
S7	21	High	25	High
S8	25	High	22	High
S9	26	High	13	Low
S10	27	High	26	High

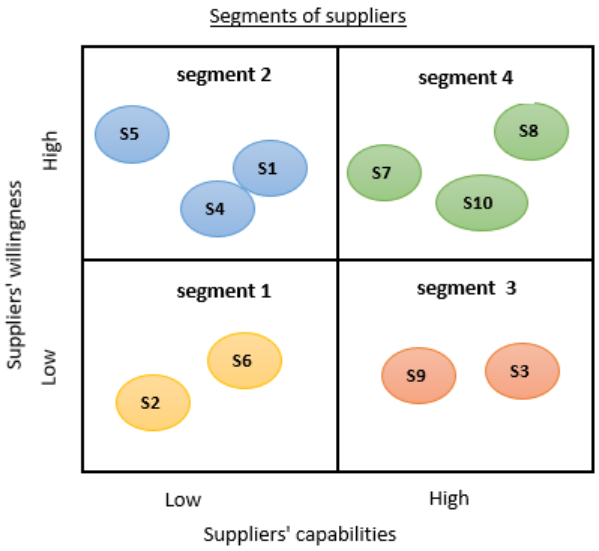


Figure 4.5: Segments of suppliers using SPM

Suppliers in the first quadrant are the worst-performing suppliers with low capabilities and low willingness to cooperate with an organisation. Organisations are advised to either replace these suppliers or maintain an arm's length relationship with them. An arm's length relationship is defined as a simple buyer and seller transactional arrangement mainly guided by a contractual fulfilment. Generally, there is little or no interaction beyond communicating requirements (e.g. via an order) and fulfilment [58].

Although suppliers in the second quadrant are less capable of meeting the organisation's requirements, they have a high level of willingness to cooperate. A general suggestion is for an organisation to assist these suppliers in improving their capabilities by investing in their development. Suppliers in the third quadrant have high capabilities but a low level of willingness to

cooperate with the organisation. Because these suppliers are worth keeping on board, organisations are advised to determine ways to improve their willingness by establishing a partnership based on mutual trust, openness and shared risk that may result in exceptional business performance [56].

Not only are suppliers in the fourth quadrant the most capable, but they are also highly willing to cooperate with the organisation. For this reason, organisations are advised to maintain strong relationships with these suppliers by combining resources and competencies, which in turn will develop a lasting strategic advantage. Organisations are also encouraged to create synergy by circulating and sharing mutually beneficial information and knowledge with these suppliers [56, 57].

Although the segmentation method used by O'Brien [47] follows the generic SPM approach, some steps differ greatly. Not only does he disagree with the way Rezaei and Ortt [56] specifies segmentation criteria, but he also does not support the use of mathematical models to calculate suppliers' aggregate scores. He argues that such approaches are usually flawed or at best suboptimal as suppliers are only regarded as important if they meet multiple criteria. Suppliers that score high but present potentially show-stopping risks could easily be accepted. He adds that a quick, simple method without a complex segmentation system or set of criteria can be highly effective.

According to O'Brien [47], a segmentation process needs to be informed by human judgement. Typically where complex variables and information need to be assimilated, visual tools tend to be most effective at providing a basis for effective human judgement. In this segmentation method, visual representations of the individual supplier's evaluation against the criteria allow rapid multi-supplier evaluation.

Suppliers' scores are marked on the criteria scoreboard and the results obtained illustrate a unique shape for each of them. Figure 4.6 shows two suppliers with two very different shapes created during segmentation. The decision-makers use these unique formed shapes to rate and segment each supplier accordingly. Not only does the visual presentation assist in defining the segments that each supplier belongs to, but they also stimulate discussions that could enable the decision-makers to define the type of relationship that the organisation should have with each segment [47].

Similar to Rezaei and Ortt [56], the results obtained from the segmentation process are presented in four quadrants. However, one quadrant does not necessarily represent one segment of suppliers. Figure 4.7 shows that different types of clusters can fall under the same quadrant. The first quadrant consists of two clusters, namely transactional suppliers and preferred suppliers. Transactional suppliers have a simple transactional arrangement with an organisation where interaction generally does not go beyond communicating the requirements and the fulfilment of requirements. Preferred suppliers have a formally or informally recognised status and are given preference over other suppliers.

Critical suppliers, in the second quadrant, fulfil requirements that an organisation cannot do without, and where organisations cannot easily switch suppliers or source elsewhere. A subcontractor relationship, in quadrant 3, is defined as suppliers that are engaged to complete a specific task as part of a bigger project or to deliver the entire project. The fourth quadrant consists of the types of relationships that focus on partnerships. A merger or group company are

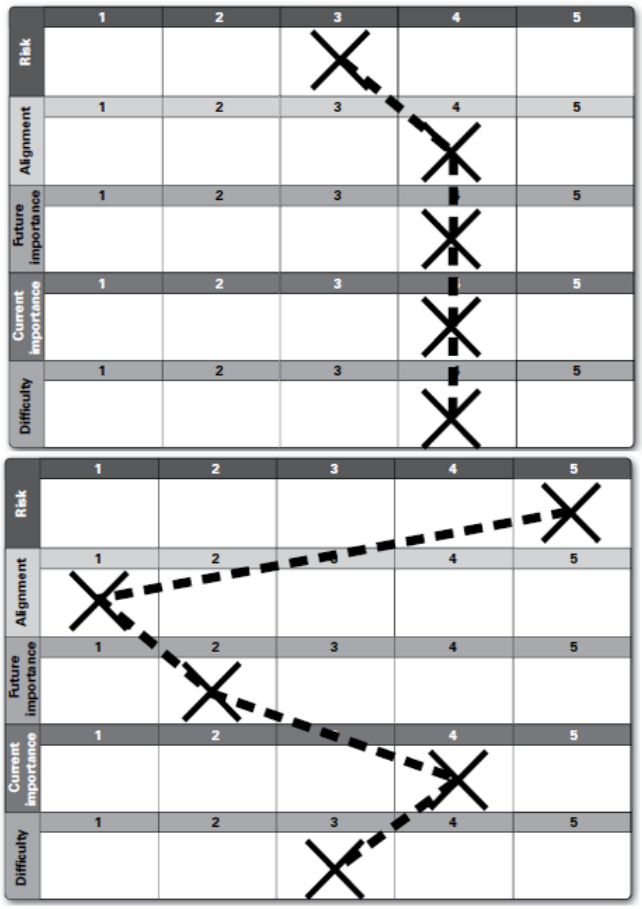


Figure 4.6: Example of segmentation score output [47]

suppliers who are owned by the organisation or owned within the group. Partner relationships such as an alliance partner, technology or creative partner, have an agreement to work together with an organisation. For example, an organisation can work with a partner to develop a new product [47].

Outsourced providers belong between the third and fourth quadrant, and strategic suppliers belong between the second and fourth quadrant. Outsourced providers are suppliers who have taken on the responsibility to fulfil a core activity and function of a company. Examples include an outsourced call centre or information technology (IT) support. Strategic suppliers are of strategic importance and have the potential to help enable an organisation achieve its goals and aspirations [47].

The organisation's corporate strategy generally guides criteria used in supplier segmentation; thus, criteria are unique to an organisation's goals. The quick segmentation method provides a generic segmentation model based on five key criteria. The SPM provides a list of criteria grouped under two categories: capabilities and willingness. Both methods effectively enable organisations to select criteria that are important to their corporate strategy.

Although the SPM method, which resulted from the extensive work of Rezaei and Ortt [56] has been a substantial step forward, SPM still has some limitations worth addressing. The aforementioned SPM methods primarily relied on the input made by the organisation's decision-makers as the only means to evaluate suppliers. Although decision-makers made use of the defined criteria as a guide, and they are expected to be knowledgeable, this method is in-

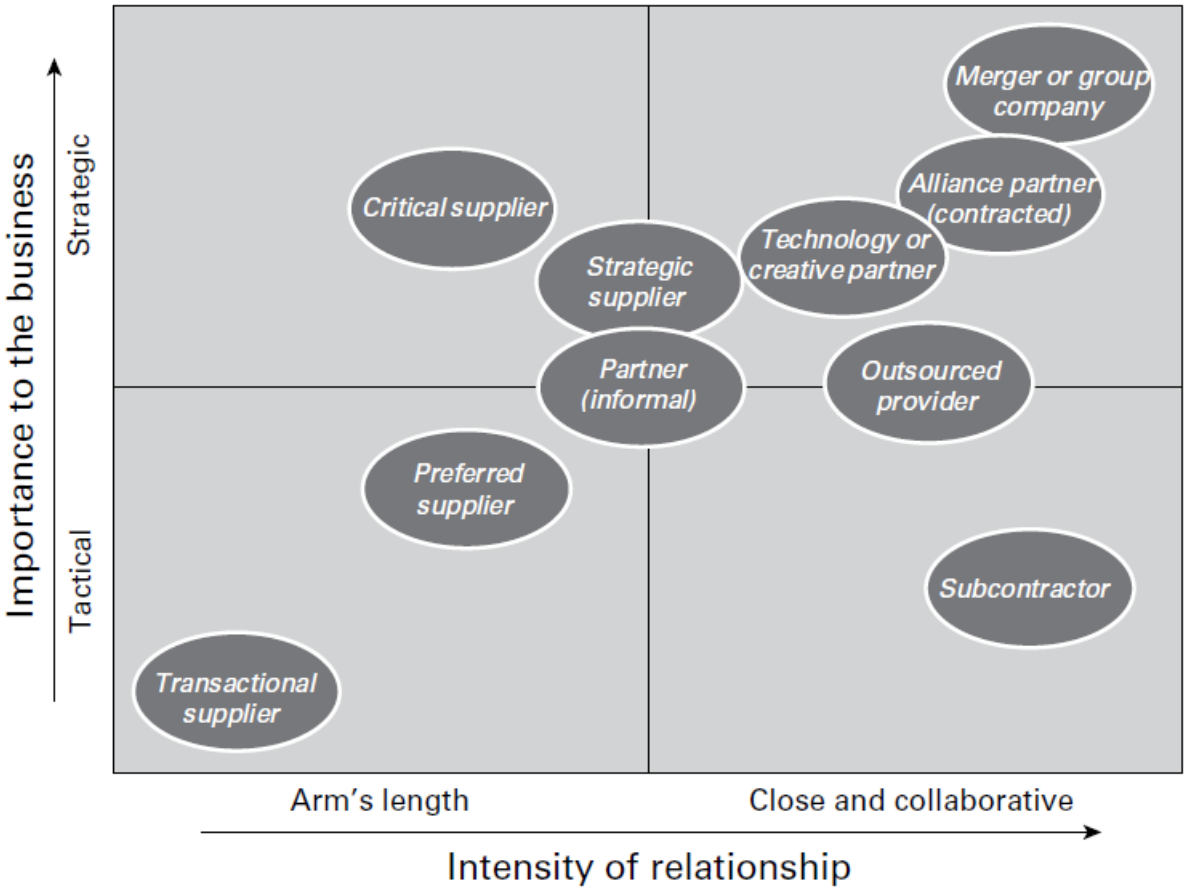


Figure 4.7: Types of supplier relationships [47]

evitably exposed to subjective bias. With the rise of information technologies and the wide adoption of organisational information systems, it is not sufficient to rely only on the decision-makers' judgements to segment suppliers. In order to reduce the possibility of systematic bias, these judgements should be supplemented by insights obtained from data such as past transactions with suppliers.

The size of an organisation's supply base is a significant factor in the SPM segmentation process. For instance, if an organisation has thousands of suppliers, rating every supplier against the selected criteria with any degree of depth would be infeasible. Therefore, the SPM method is practical only for small or medium-sized businesses with a relatively small number of suppliers.

Furthermore, most of the mathematical models used to calculate aggregate scores applied a pairwise comparison matrix to determine the weights of criteria. As a result, a new pairwise comparison matrix would be needed each time an organisation modifies the segmentation criteria or makes an adjustment to the suppliers. The environments that most organisations operate in have become exceedingly competitive and dynamic; their suppliers and segmentation criteria are likely to not only change frequently but unexpectedly too. It can, therefore, be concluded that SPM methods lack the flexibility required in today's fast-paced and competitive business environment.

The reliance on the ratings from an organisation's decision-makers means the effectiveness of a segmentation process is entirely dependent on the opinions of the people chosen in segmenting suppliers. As a result, this method can produce reliable results if an organisation has sufficient individuals who collectively understand its supply base. Organisations, particularly ones who

have a large supply base, may experience challenges in this regard.

Moreover, trends show that the average number of years that an employee stays with one organisation has decreased noticeably. According to Chudzikowski [12], employees aged between 25 and 45 stay with one organisation for about 5 years. Therefore, there is no guarantee that an organisation will have enough employees who have been with the organisation long enough to have sufficient in-depth knowledge to enable them to rate all suppliers fairly. The challenge of not having enough employees with in-depth knowledge poses an additional risk to the segmentation method proposed by O'Brien [47] as it relies solely on human judgement to give an aggregate score to suppliers.

4.3.3 Supplier intervention strategies

A central decision within SRM is to determine specific interventions and interactions an organisation should have with their supply base in order to achieve its strategic goals. There is no best practice type of relationship which applies to all categories of suppliers. Therefore, interventions need to be adapted to the type of relationship an organisation wishes to establish and maintain with each cluster of suppliers.

After the supply base has been divided into different segments, the type of interactions that a buyer need to have with each group is defined. O'Brien [47] divides interactions into five different categories where the different types of recommended interactions depend on the risk involved in the supplier relationship, the potential gain from a supplier relationship and the degree of business impact. Implementation of intervention strategies is a bilateral effort by both an organisation and its suppliers and the initiatives need to be matched against available resources. A supply base intervention map, as demonstrated in Figure 4.8, provides a visual tool for the five interventions. Transactional suppliers do not form part of the map, which means there is no intervention required for this segment [47].

- **Suppliers who need supply chain management (SCM)**: SCM is defined as the management of upstream and downstream relationships with suppliers and customers in order to deliver superior customer value at less cost to the supply chain as a whole. Generally, an organisation needs to assist the selected suppliers by understanding their network and identify where interventions are necessary or beneficial. The SCM intervention's key objective is to help suppliers gain basic capabilities to manage their logistics, demand, information and risk [38, 47].

- **Suppliers who need supplier performance management (SPM)**: SPM is defined as the process of targeted evaluation, measuring and monitoring of supplier performance and practices for the purposes of achieving desired business outcomes and goals. A key factor to effective SPM is the concept of the right amount of measurement used in a way that helps achieve the required outcomes. The degree of measurement needed in SPM interventions is applied differently for each supplier. For instance, an organisation may need to check goods from certain suppliers before acceptance; other suppliers may only need some form of measurement when there is a deviation from the plan. Yet for some suppliers an ongoing regime of measurement and review of various measures may be needed to ensure that an organisation achieves its goals [47].

- **Suppliers who need improvement and development (SI & D)**: SI & D aims to improve suppliers' capabilities by implementing various initiatives with the selected suppliers. An organisation may either abandon or develop suppliers who do not show

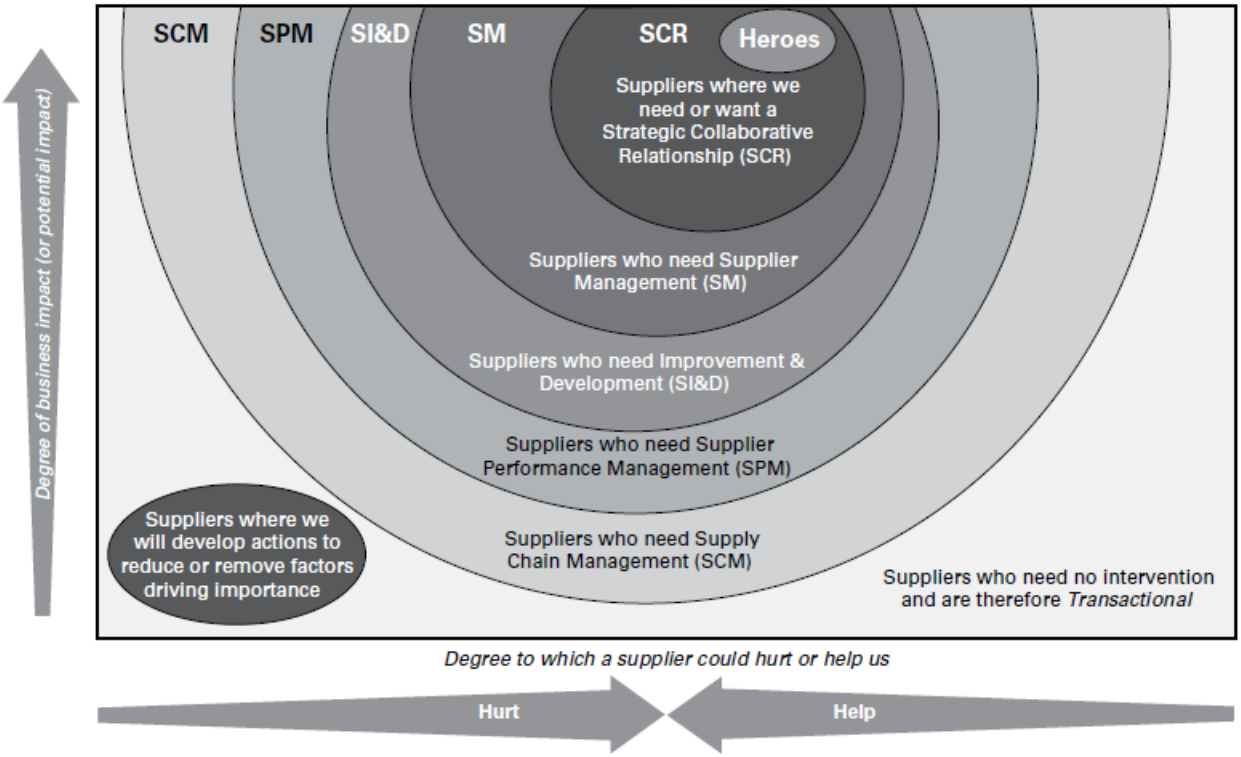


Figure 4.8: Supply base intervention map [47]

improvement after specific interventions have been implemented. The decision to develop or abandon suppliers depends on the supplier's level of importance to the organisation. Generally, supplier development offers an opportunity for both parties to secure greater benefits [47, 51].

- **Suppliers who need supplier management (SM)**: SM refers to a systematic approach used to manage the relationships with important and strategic suppliers to ensure contractual obligations are met and to maximise performance. This intervention can help an organisation manage results, review suppliers and manage relationship interfaces [47].
- **Suppliers who need a strategic collaborative relationship (SCR)**: SCR is the ultimate level of relationship intensity and appropriate only for the critical few suppliers who are of strategic importance and who hold the potential to benefit an organisation dramatically. Effective strategic relationships are mostly about decisions and actions being made in support of a long-term joint relationship. The relationship is born out of commitment, cooperation, and collaboration. If these factors are present, then there will be a natural commitment by parties to build and maintain a high performing relationship [47, 51].

4.4 Chapter summary

This chapter defines a method that organisation can use to define key requirements from their suppliers. The chapter then discusses different approaches used in dividing suppliers into distinct clusters based on their perceived importance to the organisation. Lastly, it proposes various intervention strategies that may be applied in managing different clusters.

Chapter 5

Application of CRISP-DM to the Mozambican cassava supplier segmentation case study

In this chapter, the CRISP-DM processes discussed in Chapter 3 are applied to the historical Dadtco cassava purchasing data. Section 5.1 provides background on Dadtco Philafrica's operations and describes the organisation's key requirements from its supply base. Section 5.2 conducts different data exploration methods to gain more understanding of the purchasing of cassava. Section 5.3 conducts data preparation and constructs a final dataset that is in the modelling phase. In section 5.4, the k-means algorithms, AHC and SOM are implemented and the performance of each technique is evaluated.

5.1 Business understanding

Dadtco Philafrica Cassava Processing is a for-profit social enterprise that manufactures cassava-based products. Their mission is to bridge the gap between smallholder farmers and food companies throughout Africa. Cassava, also called *Manihot esculenta* or Tapioca is a root crop similar to other starch crops such as potatoes. The crop, shown in Figure 5.1, can be found in tropical climates, and some of its advantages are that it is easy to grow as it is drought resistant and adapts well to climate change [18, 60].

Figure 5.1: Cassava root

Dadtco Philafrica's vision is to be a leader in cassava processing throughout Africa, working together with thousands of smallholder farmers. The organisation considers farmers to be its

most important partners as it depends on them for key raw material supply. The two cassava-based products produced by Dadtco Philafrica are cassava starch paste (CSP) and high-quality cassava flour (HQCF). CSP is a semi-wet cassava paste allowing further processing into a variety of downstream-related products such as cassava-based syrups, sorbitol and beer. Impala beer, processed in Mozambique since 2011, is made with 70% CSP. HQCF is a white flour used extensively in the food and beverage industry in the manufacture of culinary cubes, powdered drink products, snacks and soups [60].

The commercialisation of cassava across Africa has been a challenge, especially to smallholder farmers. One of the key challenges is the rapid perishability of the root once it is harvested. The majority of smallholder farmers do not have resources to transport cassava to a processing site, which means the organisation has to collect the crop from them. The transportation of roots from a fragmented farm base can easily lead to high transportation costs and challenging logistics. Figure 5.2 shows one of Dadtco Philafrica's cassava processing plants in Mozambique.

Figure 5.2: Dadtco Philafrica Cassava processing plant [60]

The organisation requires an efficient approach to segment farmers into logical categories based on their similarities, to define the type of relationship it should have with each group in order to achieve its strategic goals. Furthermore, the organisation aims to use the results to define different intervention and development strategies for each cluster. With over 3 000 farmers in its purchase history database, it would not be practical to implement the segmentation methods that currently exist in literature. Not only is the number of suppliers too large, but the organisation does not have employees who have in-depth knowledge about all suppliers to be able to rate them fairly. Lastly, the organisation's supply base is continuously expanding; not only will they benefit from using a more flexible segmentation system, but a method that does not primarily depend on the judgement of employees will be a great asset.

A session to discuss Dadtco Philafrica's key requirements from its supply base was conducted with the company's management team. The management team consisted of the country director, supply chain managers, quality assurance manager and production manager. The following key requirements were defined using the VIPER model:

- **Supply risk**: The company sources raw materials from smallholder farmers with different capacity levels and constraints. Over the years, supply of cassava to Dadtco Philafrica has been very irregular. Poor access to quality farming inputs has been stated as a key cause of these irregularities. The inadequate farming inputs have a negative impact on the

yield, which in turn, negatively affects a farmer's income potential. Some farmers become discouraged which results in long pauses between harvests. Some farmers even switch to a different crop hoping for better returns. The organisation uses a *no. of purchases* feature to measure supply risk. The *no. of purchases* feature measures the number of times a farmer has supplied cassava in the period between February 2018 and April 2020. A higher value indicates that the farmer has replanted cassava after each harvest; thus, the company can rely on them to continue producing.

- **Effectiveness of operations**: It is important to note that smallholder farmers, who account for almost all cassava production, are poorly organised and spread widely across the rural areas. The collection of roots from a fragmented farm base causes logistic challenges to the organisation and results in higher transportation costs. The organisation measures effectiveness of operations using the *amount paid for using own transport* feature.

- **Performance improvements**: In order to grow, the organisation requires a higher quantity of good quality cassava. The performance of farmers is measured by their yields (quantity and quality). The quality of cassava is generally measured by its starch content, where higher starch content indicates good quality. Low productivity is considered as the main constraint preventing farmers from producing at a profit. Cassava yields for most farmers in Mozambique are about 7 tons per hectare (ha), and plot sizes are 0.5 ha. The values indicate that on average, a farmer produces about 3.5 tons of cassava per plot. In order to be commercially viable at the present market prices, cassava producers are expected to achieve at least 15 tons per ha. The organisation measures performance using the *quantity of cassava purchased* and *average starch content* features.

- **Innovation and value**: The organisation's operations are still in the infancy stage and the main focus is to first build a solid foundation for the supply base. For this reason, the organisation will not engage in any innovative and value-enhancing initiatives with its supply base until the basics are in place.

5.2 Data understanding

This section conducts an exploratory data analysis to gain more insight into the features. Section 5.2.1 uses descriptive statistics to analyse data of individual features. Section 5.2.2 uses descriptive statistics to analyse the relationships between features. The historical purchasing data about cassava was received as a comma-separated values (csv) file from an information system called Cropin Smart Farm. The file contained purchasing details for transactions dated between February 2018 and April 2020. Table 5.1 summarises features from the dataset.

5.2.1 Analysis of individual features

This section focuses on descriptive statistics and data visualisation techniques applied to individual features. The key purpose is to describe each feature.

Continuous features

The *latitude of plot* and *longitude of plot* features are continuous variables that provide the geographic coordinates of each farmer's plot. Figure 5.3 shows that the *latitude* and *longitude* of farmers' plots are highly concentrated in two areas; Ribáuè district and Inharrime district. It makes sense to have more plots in these areas as the factories prioritise sourcing cassava from

Table 5.1: Summary of features of the dataset

Feature name	**Description**	**Data type**
Farmer code	A unique identification code given to every farmer	Numeric
Location of factory	The location area where a factory is situated	Categorical
Location of plot	The location area (district) where a farmer's plot is situated	Categorical
Latitude of plot	The latitude coordinates of a plot's location	Numeric
Longitude of plot	The longitude coordinates of a plot's location	Numeric
Field worker	The name of the field worker assigned to a farmer	Categorical
Modified variety?	This field checks if the cassava delivered was a genetically modified variety	Binary
Starch content (%)	Average starch content of cassava delivered	Numeric
Cassava quantity (Kg)	Quantity of cassava delivered	Numeric
Cassava cost (MZN)	Amount paid to a farmer for cassava delivered	Numeric
Transport cost (MZN)	Amount paid for transport to a farmer who organised own transport	Numeric

locations closer to them in order to minimise logistics costs.

The *starch content* feature records the amount of starch detected in the cassava delivered. The key output from cassava processing is starch; therefore, cassava roots with high starch content are ideal. Starch content has a direct impact on the performance of the processing plant. For example, if 2 000 kg of roots that contain an average of 25% of starch is processed, a maximum output of 50 kg can be produced. However, if the cassava roots contain 15% of starch, only 30 kg of output can be produced. The report in Table 5.2 shows a mean value of 18.5%. The value indicates that the organisation can extract about 185 kg of starch from 1 000 kg of cassava input.

Figure 5.3 shows the distribution of the *cassava quantity* feature. The histogram is a right-skewed distribution which indicates that the majority of the farmers delivered less than 4 000 kg of cassava per transaction. According to Costa and Delgado [14], the average yield for most smallholder cassava farmers in Mozambique is about 7 000 kg per hectare. The results indicate that most farmers' plots are less than one hectare in size. Furthermore, Table 5.2 shows a standard deviation of almost 2 000 kg, which indicates a high variance of the plot sizes of the farmers. A closer inspection of the feature revealed a few values less than 1 kg. It is worth noting that a cassava quantity of 1 kg is equivalent to less than 5 roots. As a result, all values less than 1 kg were removed from the dataset as they likely resulted from data entry errors.

The *cassava cost* feature is calculated by multiplying *cassava quantity* feature by the price of cassava. The price of cassava is a fixed value and all cassava, despite their varieties or starch content, have the same price; thus, the total cost depends solely on the quantity of cassava delivered. As expected, the data distribution in Figure 5.3 is similar to a distribution for quantity purchased. The cost is in Mozambican currency which is abbreviated with the symbol MZN.

The *transport cost* feature records the amount that was paid to farmers for organising their own transport. The majority of the farmers, 89%, relied on the organisation's trucks to transport their cassava to the factory. As a result, these farmers were not paid for transport. The histogram in Figure 5.3 shows the data distribution of the 11% farmers who did not use the organisation's transport to deliver roots to the factory. The distribution shows that the major-

ity of the farmers who organised their own transport delivered relatively smaller loads, mainly below 2 000 kg.

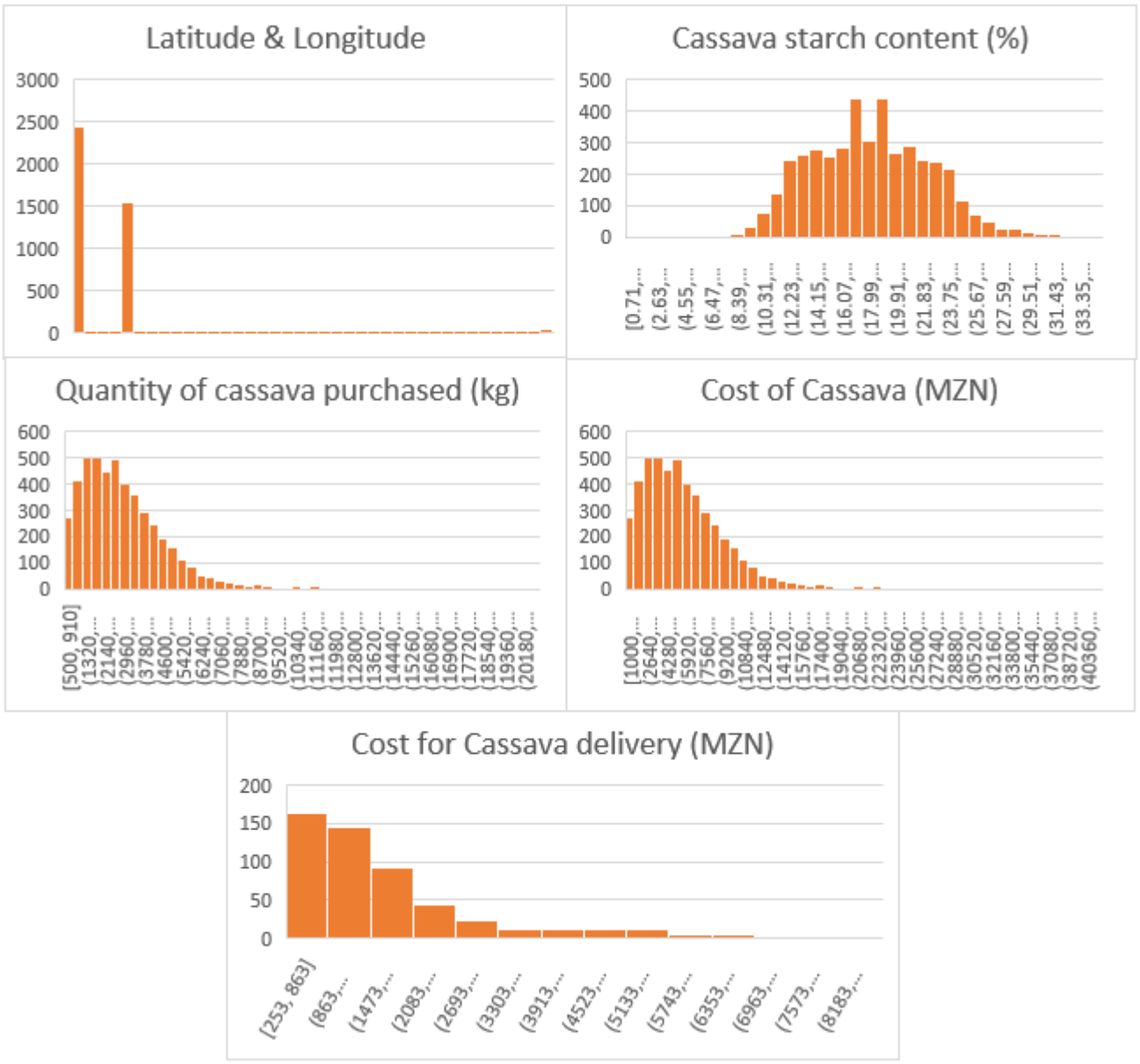


Figure 5.3: Histograms of continuous features

Table 5.2: Data quality report for continuous features

Measure	Latitude	Longitude	Starch content	Cassava quantity	Cassava cost	Transport cost
Count	4 010	4 010	4 370	4 756	4 756	4 756
Missing values	16%	16%	8%	0%	0%	0%
Minimum	-24.47	5.66	0.71	0.3	0.59	0
1st Quartile	-24.44	35.07	15.1	1 731	3 462	0
Mean	-20.22	35.80	18.5	3 127	6 254	189
Median	-24.44	35.07	18.4	2 748	5 495	0
3rd Quartile	-15.05	38.34	21.9	4 014	8 027	0
Maximum	-51.98	39.26	35	20 942	41 884	8 323
Standard deviation	9.46	3.82	4.57	1 962	3 927	710

Categorical feature

The *location of factory* feature records the location of the processing sites. Dadtco Philafrica has two processing plants in Mozambique, one located in Ribáuè district and the other one in Inharrime district. The distance between the two sites is over 1 500 km, as shown in Figure 5.5. The factory in Ribáuè district produces cassava starch paste (CSP) for the local breweries. The factory in Inharrime district has a larger capacity and it produces both cassava starch paste (CSP) for breweries and high-quality cassava flour (HQCF) for meat processing companies. This feature measures the number of transactions concluded at each factory during the analysis period. The data shows that most of the purchasing of cassava occurred in Inharrime district.

The *location of plot* records the locations of cassava farms. The processing plants sourced cassava from 10 districts. Due to the rapid perishability of the crop, the factories prioritise location areas that are closer to each factory. Figure 5.4 shows that the organisation sourced the majority of cassava, 67%, from the districts in which the two factories are located, Inharrime district and Ribáuè district.

The *modified variety?* answers if the types of cassava delivered were modified varieties. The organisation is continuously working with various institutions to find cassava varieties with high starch content and resistance to diseases and climate change. One way to achieve this is through breeding of various varieties. However, some researchers have argued that local varieties generally perform better than most modified varieties. In Mozambique, the research on variety modification is still in its infancy, and few modified cassava stems have been distributed to farmers for trials. Figure 5.4 shows that only 6% of the cassava delivered were modified varieties.

The *field worker* feature records employees that are responsible for the management of sales and relationships with farmers. Figure 5.4 shows that 20% and 17% of all deliveries were from farmers assigned to Moises and Joaquim, respectively. The results show that these two field workers with the highest number of deliveries are based in Inharimme district where the larger factory is located.

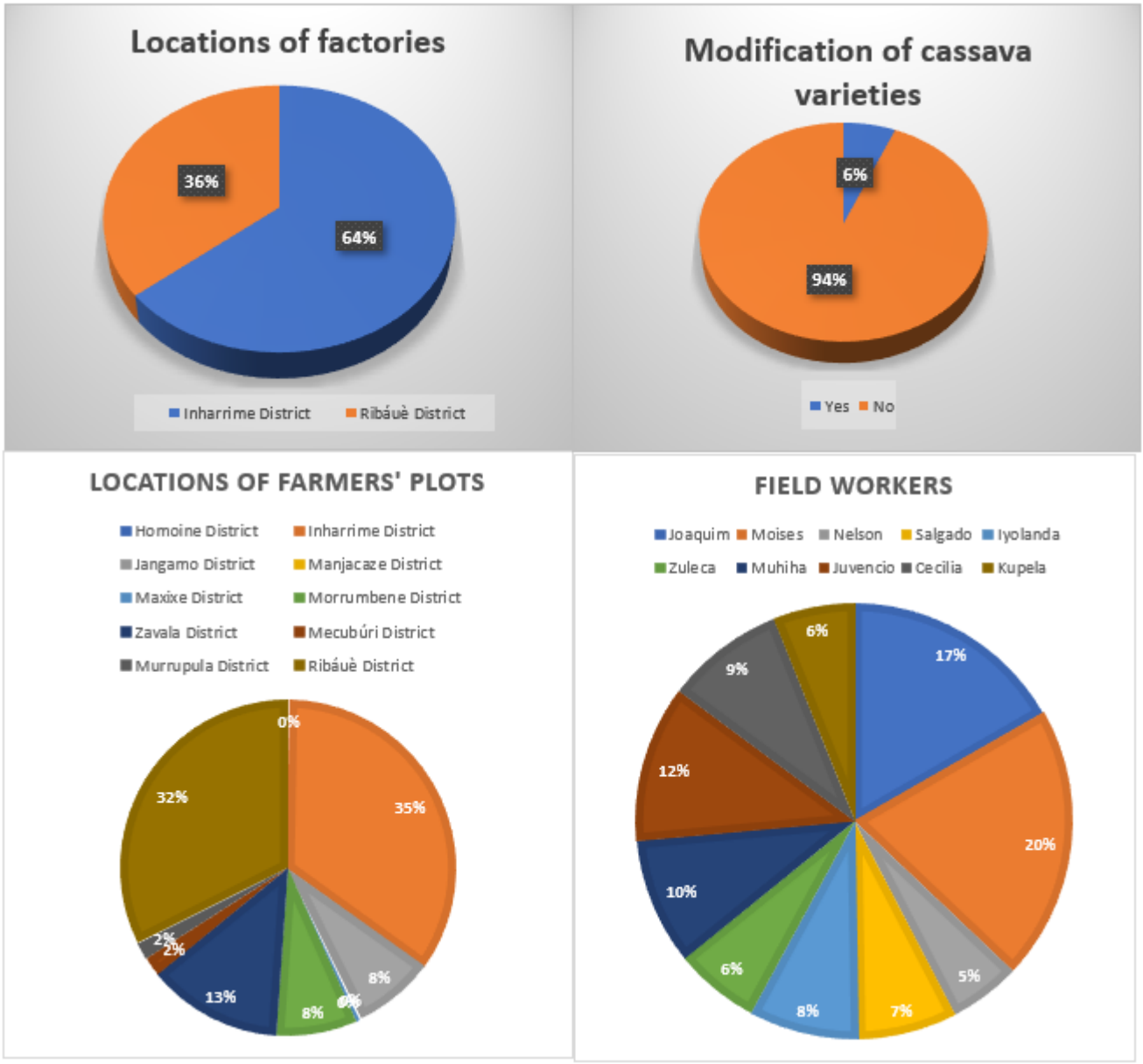


Figure 5.4: Pie charts of categorical features

Table 5.3: Data quality report for categorical features

Measure	Location of factory	Location of plot	Field worker	Modified variety?
Count	4 756	4 756	4 756	4 756
Missing values	0%	0%	0%	0%
Mode	Inharrime	Inharrime	Moises	No
Mode frequency	3 051	1 650	968	4 456
Mode %	64%	35%	20%	94%
2^{nd} mode	Ribáuè	Ribáuè	Joaquim	Yes
2^{nd} mode frequency	1 705	1 540	795	300
2^{nd} mode %	36%	32%	17%	6%

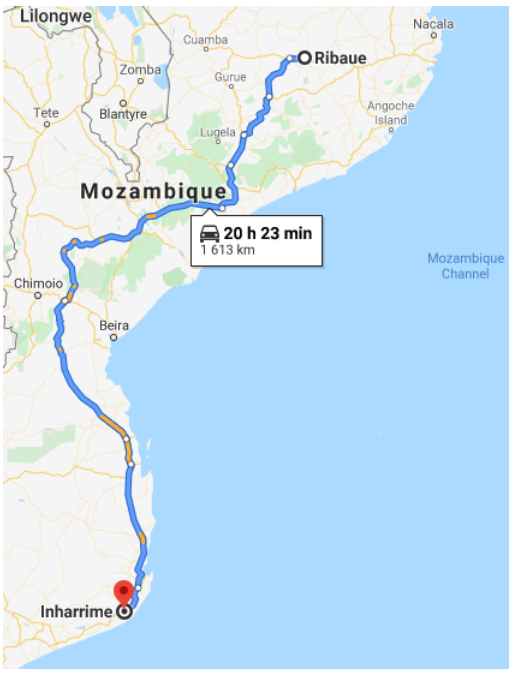


Figure 5.5: Distance between the two processing sites

5.2.2 Analysis of the relationship between features

In this section, the relationships between pairs of features are examined. A scatterplot matrix (splom) is used to examine the relationships between the continuous features and stacked bar graphs are used to examine relationships between the categorical features.

Continuous features

The first row in Figure 5.6 examines the relationship between the *starch content* feature and the other three features. The correlation between the *starch content* and *cassava quantity*, *cassava cost* and *transport cost* is -0.026, $-0,023$ and -0.031 respectively. All these values indicate an extremely weak negative correlation between *starch content* and the other continuous variables.

The correlation between *cassava quantity* and *transport cost* is the same as for the *cassava cost* and *transport cost*. It is worth noting that only 6% of the total cassava was delivered using farmers' own transport; thus farmers were compensated for both cassava and transport. For the 94% of the cassava delivered, farmers were compensated only for cassava and not for transport; hence the relationship between these variables indicates a significantly small positive correlation. The correlation value between *cassava quantity* and *cassava cost* is 1, which indicates a perfect positive correlation between these two features. The strong correlation is expected as the cost of cassava is calculated by multiplying the quantities by a constant value.

Categorical features

Figure 5.7 shows cassava deliveries per locations of plots and field workers. The results shows that the majority of farmers located in the same area are assigned to the same field worker. Figure 5.7 shows that farmers located in Inharrime district are assigned to three field workers; Joaquim, Moises and Salgado. The figure also shows that all four field workers primarily focused on sourcing cassava from Ribáuè district. The majority of farmers are located in Inharrime and Ribáuè districts as the organisation aims to source from location areas that are geographically

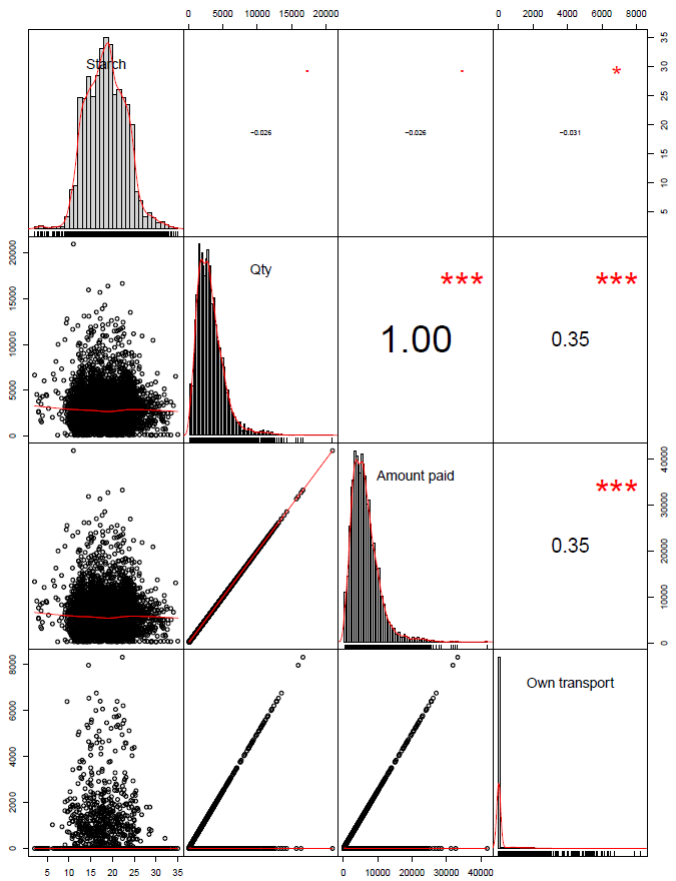


Figure 5.6: Relationships between continuous variables

close to the factories.

Figure 5.8 shows the number of modified cassava varieties per field worker and location of plots, respectively. The results shows that a significant amount of cassava delivered was not of modified varieties. Figure 5.8 reaffirms that most deliveries were from farmers assigned to Joaquim and Moises. Moreover,the figure shows that most deliveries were from farmers located in Inharrime and Ribáuè districts.

5.3 Data preparation

In the data preparation phase, the final dataset to be used in the modelling phase is constructed. Section 5.3.1 conducts data cleaning to address the data quality issues identified in the data understanding phase. Section 5.3.2 conducts feature selection to remove redundant and irrelevant features from the dataset. Section 5.3.3 applies a normalisation method to all features to ensure that there are no features that dominate others due to a significant difference in range. In section 5.3.4, the data is integrated to construct a final dataset that is used in the modelling phase.

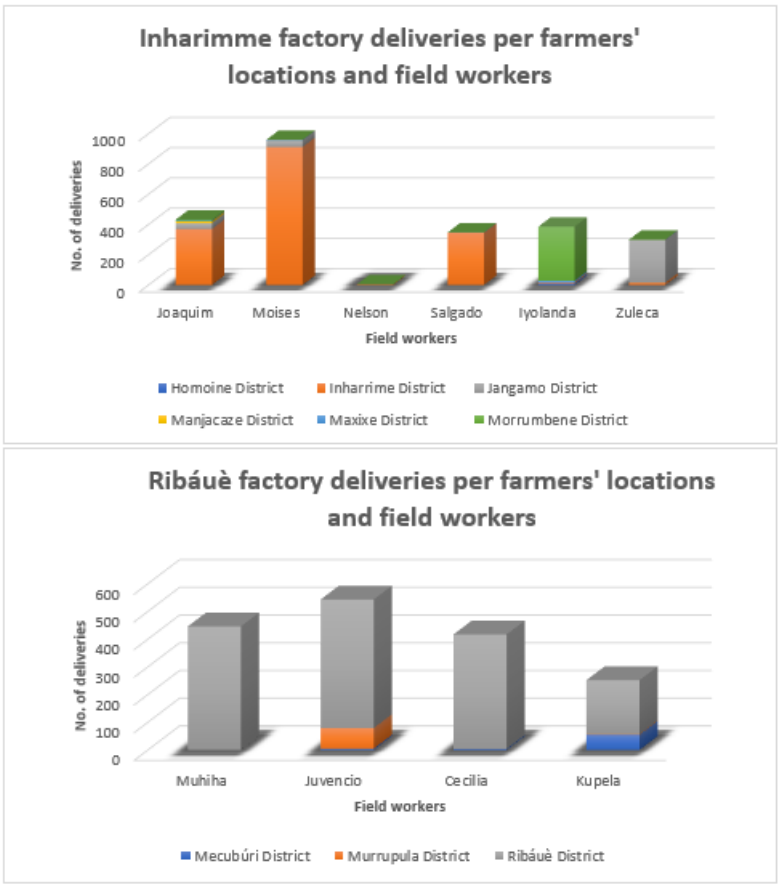


Figure 5.7: Cassava deliveries per field workers and locations of plots

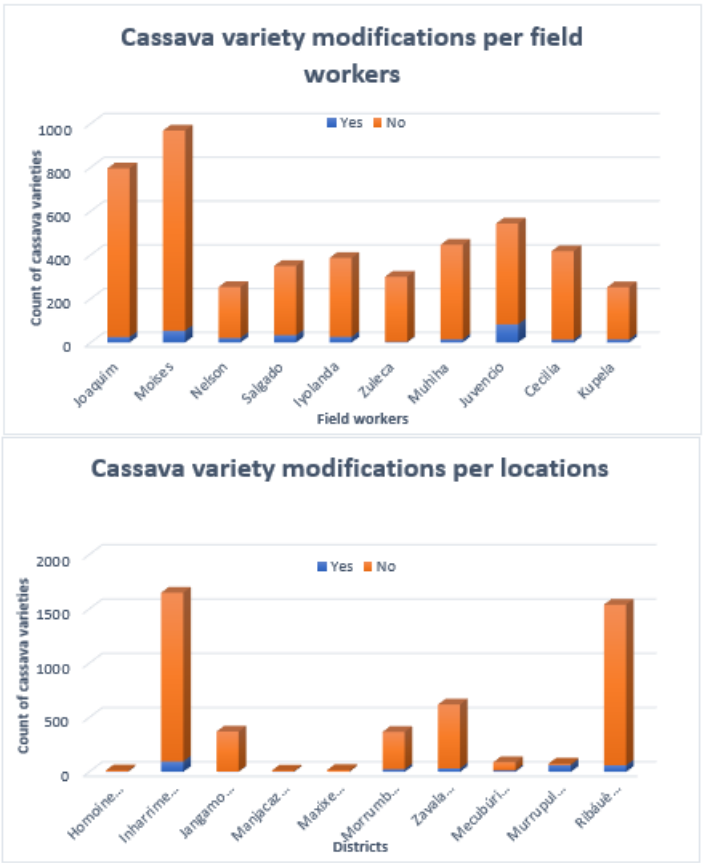


Figure 5.8: Modification of varieties per field workers and locations of plots

5.3.1 Data cleaning

A key objective of data cleaning is to address the data quality issues identified in the data understanding phase. The most common data quality issues identified are missing values and outliers. The *latitude of plot*, *longitude of plot* and *starch content* features are the only features that have missing values. The percentage of the missing values for both *latitude of plot* and *longitude of plot* features is 16%. An imputation approach was thus used to replace the missing values.

For farmers who were supplying to the organisation for the second time, the recorded geographic coordinates from the farmer's previous purchase record were used to substitute the missing coordinates. For farmers who had supplied cassava to the organisation once, average values of all *latitude of plot* and *longitude of plot* of the farmers' respective location areas (districts) were used to replace missing values. For instance, if a farmer's plot is situated in Ribáuè district, the mean value for all geographical coordinates in Ribáuè was used to replace the missing coordinate.

The percentage of the missing values for the *starch content* feature is 8%. The imputation approach was applied in the same way as for the *latitude of plot* and *longitude of plot* features. For farmers who had sold cassava to the organisation more than once, data from the previous purchase records was used to replace missing values of the *starch content* feature.

Figure 5.9 shows box plots of all continuous features of the dataset. The box plot for the *latitude of plot* and *longitude of plot* features shows that farmers' plots, which make up about 2% of the total plots, had positive latitude coordinates, while the rest of the plots had negative latitude coordinates. Having positive latitude coordinates means the location is to the north of the equator while locations on the south of the equator have negative latitude coordinates [53]. The plots were not removed as they are not anomalies and their locations are located in a district that the factory sources from.

The *starch content* feature has a standard deviation of 4.58. A high standard deviation indicates a high variation from the average *starch content* which was confirmed by a huge gap between minimum and maximum values. The supply chain manager confirmed that some crops contained low starch content but stated that values such as 0.71% were unusual and could have resulted from a data entry error or faulty starch detector equipment. He advised that values below 5% could be considered as invalid and should be removed.

The box plot for the *cassava cost* feature shows several data points that could be outliers. The largest value is the maximum value of almost 21 000 kg. The supply chain manager stated that a delivery of 21 000 kg from one farmer is possible but very uncommon. He stated that there are a few farmers who are more established, have multiple plots and they are, thus, able to acquire a yield of up to 15 000 kg per hectare. As a result, a quantity of 21 000 kg is attainable. Based on this explanation, it was concluded that the outliers are a result of valid data. These data points were not removed as they provided information about the different capacities of farmers.

The maximum cassava delivered by a farmer using own transport is 16 645 kg. The supply chain manager confirmed that the outliers are valid as some farmers delivered more than one load in a day and these loads were treated as one transaction in the database. He stated that high quantities delivered was likely to be the result of multiple deliveries.

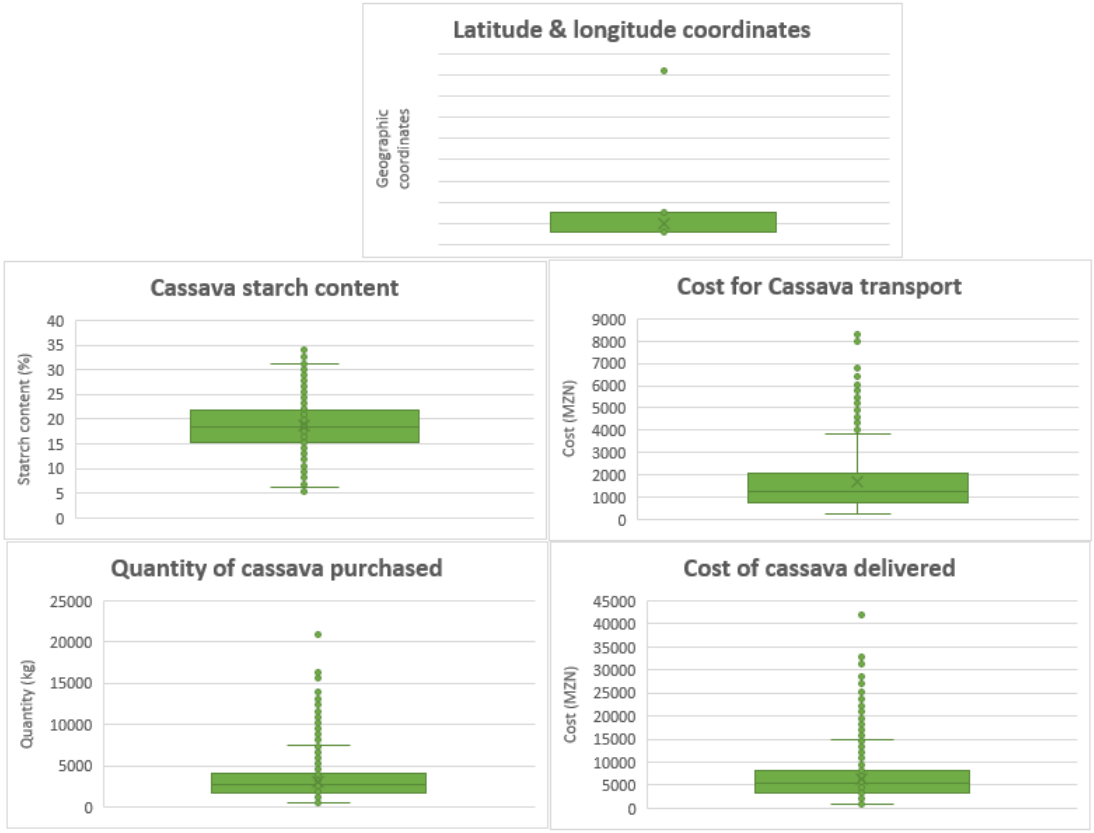


Figure 5.9: Box plots after first data preprocessing

In order to retain as much information as possible, only outliers that resulted from invalid data were removed. The other outliers were discussed with the management of the organisation. Unless the managers stated that outlier values could have resulted from invalid data input, the outliers were retained. It is important to note that most algorithms that use distance to measure similarities are generally sensitive to and do not perform well in the presence of outliers. In order to evaluate the impact of also removing outliers that resulted from valid data input, the dataset was processed further. A clamp transformation method was applied on the *starch content* and *cassava quantity* features. The upper and lower thresholds were determined using the clamp transformation method explained in section 3.4 and the resulting distributions are shown in Figure 5.10.

5.3.2 Feature selection

The results from the splom were used in the feature selection process to identify redundant features. A feature is considered redundant if it has a strong correlation with another feature. The *cassava quantity* and *cassava cost* features have a perfect positive correlation. As a result, the *cassava quantity* feature was removed from the dataset. It is important to note that the *latitude of plot* and *longitude of plot* features provided specific locations of the farmers' plots and the *location of plots* variable provided the names of districts where the plots are located. These two variables provided the same information, which is the location of farmers' plots. As a result, the*latitude of plot* and *longitude of plot* features were removed from the dataset.

Other features that were considered are the *field worker* and *location of factory* features. The two features have a direct link with the *location of plot* feature. A field worker is generally

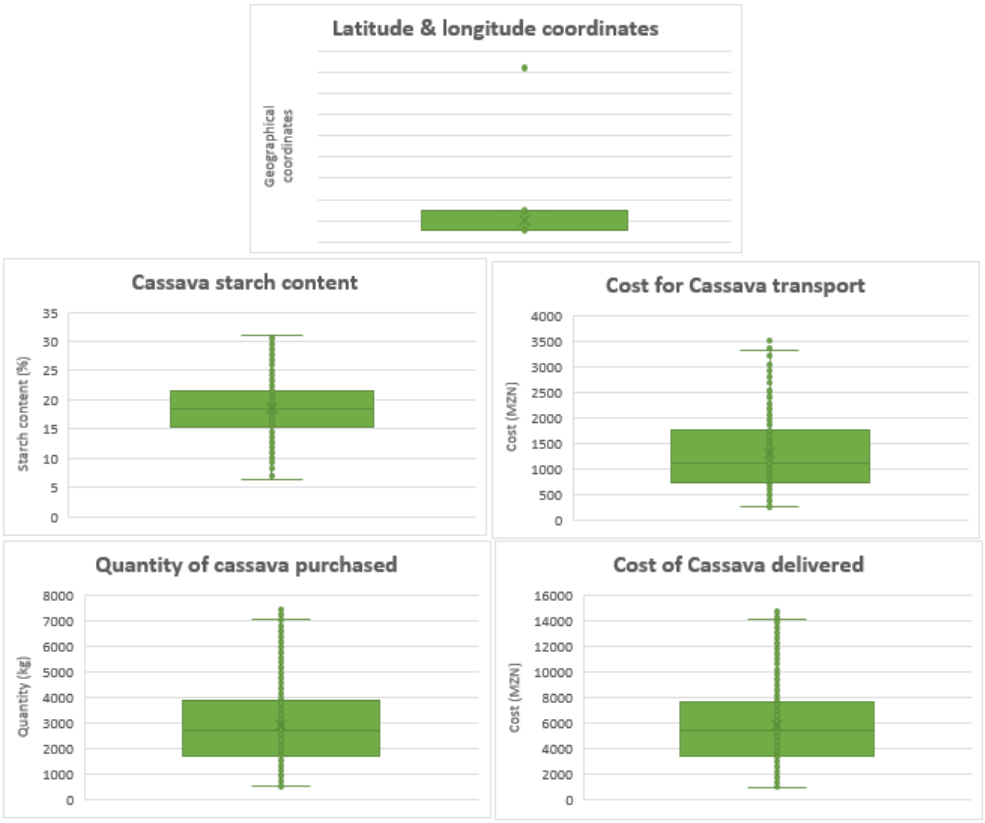


Figure 5.10: Box plots after second data preprocessing

assigned according to the location of farmers' plots. For instance, a field worker would be assigned to all farmers who are based in the same area. Furthermore, the factory that a farmer can supply cassava to is solely dependent on a farmer's location. The two factories are located over 1 500 km apart; therefore, it is not practical for a farmer based in Inharimme district to supply cassava to Ribáuè factory. Consequently, the two variables, *field worker* and *location of factory* features were removed from the dataset.

5.3.3 Data standardisation and transformation

All clustering algorithms used in this study are distance-metric based; thus, it is important for the dataset to be standardised. The MinMaxScaler normalisation method was applied to all features to ensure that there are no features that dominate others due to a significant difference in range. As a result, all variables were transformed to values that ranged between 0 and 1. The original dataset consisted of three categorical features and one binary feature. Since the algorithms in this study use distance to measure similarity; the categorical and binary features were transformed into numerical values using the one-hot encoding method.

5.3.4 Construction of final datasets

The original dataset consists of data recorded per transaction, meaning each instance represents a purchasing transaction and not a farmer's purchasing profile. As a result, farmers who have sold cassava to the organisation more than once appeared in multiple instances. The clustering algorithms evaluate a dataset by considering both the features and instances and group the results accordingly. If the clustering algorithms are applied to the dataset as is, the results would be clusters of purchasing transactions and not of farmers.

In order to have one record per farmer, features were aggregated accordingly. The *transport cost* and *cassava cost* features were summed to one value per farmer. For the *starch content* feature, an average value for all farmers' transactions was used per farmer. There was no data integration required on categorical variables such as the *location of factory*, *field worker* and *location of plot* features, as the values per farmer were the same in all transactions. For the *modified variety?* binary feature, the most recent value was used.

Furthermore, a new feature called *no. of purchases* was added to count the number of transactions for each farmer. As a result, the final dataset that was analysed in the data modelling phase consisted of input variables generated for each farmer. Consequently, dataset1 and dataset2 were implemented in the modelling phase. Dataset1 (DS1) consisted of transactions where outliers that resulted from valid data input were retained. Dataset2 (DS2) consisted of data that was processed further and the outliers were eliminated using the clamp transformation method.

It is important to note that the outlier removal was applied before the transactions were aggregated to form records per farmer. The aggregated datasets, DS1 and DS2, still consisted of outliers. No further data processing was conducted as the outliers showed a true representation of Dadtco Philafrica's supply base. The farmers have different levels of skills and capacities, and as a result, their supplying pattern varied significantly. The final datasets consist of features as shown in Table 5.4. DS1 consists of 3507 instances and 5 features, and DS2 consists of 3387 instance and 5 features.

Table 5.4: Summary of features of the final dataset

Feature name	**Description**	**Data type**
Location of plot	The location area (district) where a farmer's plot is situated	Categorical
Modified variety?	This field checks if the cassava delivered was a genetically modified variety	Binary
Starch content (%)	Average starch content of cassava delivered	Numeric
Cassava cost (MZN)	Amount paid to a farmer for cassava delivered	Numeric
Transport cost (MZN)	Amount paid for transport to a farmer who organised own transport	Numeric
No. of purchases	Total number of transactions with a farmer during the analysis period	Numeric

5.4 Modelling and evaluation

Different techniques with clustering capabilities are implemented and evaluated in this section. Sections 5.4.1, 5.4.2 and 5.4.3 describe steps that are followed in implementing the k-means, AHC and SOM to the case study, respectively. In each section, the techniques are applied to the dataset and the results are evaluated. The final results obtained from each technique are summarised in section 5.4.4.

The main goal of clustering is to maximise the homogeneity within each cluster and the heterogeneity among different clusters. Therefore, the performance of each algorithm was evaluated

using the intra-cluster distance and inter-cluster distance of the clusters. The Euclidean distance was used to measure similarity between objects. Dadtco Philafrica requested that each cluster of suppliers should be allocated to one or more field workers who will be involved in defining and implementing a strategy for that cluster. In order to prevent a situation where one field worker is allocated to two or more clusters, the maximum allowable number of clusters for each algorithm was set to ten clusters.

5.4.1 *K*-means implementation

Figure 5.11 demonstrates the steps that are followed in implementing the k-means algorithm. First, the optimal number of clusters (K) are obtained using the silhouette coefficient (SC). Once the value of K is chosen, the centroids are initialised using the random and k-means++ initialisation methods. The Euclidean distance is used to calculate the distance between data points and centroids, and finally, clusters are formed amongst data points with high similarity.

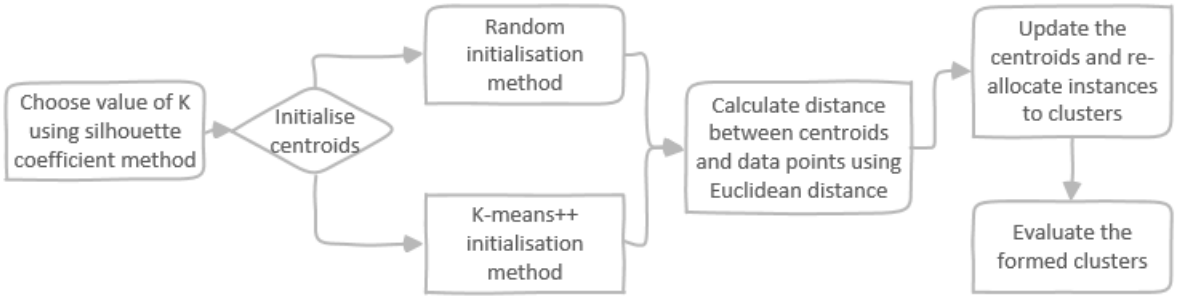


Figure 5.11: Approach used for implementing k-means algorithm

In order to deal with the randomness caused by the centroid initialisation and increase the likelihood of finding an optimal solution, the algorithm was executed for 30 independent simulation runs. The SC method, applied in each run, was used to determine the best value of K and the algorithm was executed for each set of experimental conditions. The experimental conditions tested are different initialisation strategies and different datasets. At the end of the experimental process, each set of experimental conditions consisted of 30 results.

The results indicate that the algorithm obtained a different 'best K value' in each run. Figure 5.12 shows the number of runs where each K value obtained the highest SC value. For the random initialisation method, both DS1 and DS2 obtained most best SC value at K= 7. DS1 consists of valid outliers and in DS2, outliers were removed using the clamp transformation method. From the 30 runs, DS1 obtained the best SC value at seven clusters in 13 runs, and DS2 obtained the best K value at seven clusters from 15 runs. The k-means++ initialisation method, which aims to reduce the impact of the algorithm's randomness obtained the same K value in all 30 runs. DS1 obtained the best SC value at 10 clusters, and DS2's best SC value was at seven clusters.

Inter-cluster results and hypothesis tests

Inter-cluster distance is one of the measures used to evaluate the formed clusters. The measure, which is calculated using equation 3.18, indicates how well-separated or distinct a cluster is from other clusters. A higher inter-cluster distance indicates a better separation between different clusters. Figure 5.13 shows the results obtained from the 30 runs for each set of experimental

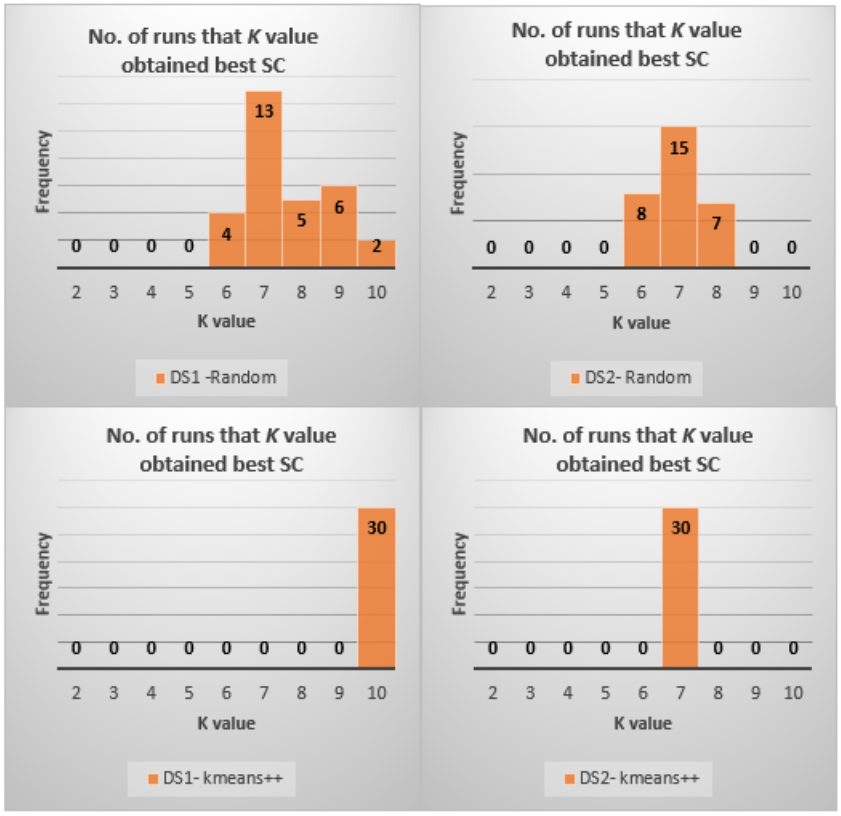


Figure 5.12: No. of runs that K value obtained best SC

conditions. The line graph and the bar graph show the mean and standard deviation for the inter-cluster distance. The first set of experimental conditions, DS1-random, shows the results obtained when DS1 was trained using the random initialisation method. In DS2-random, DS2 was trained using the random initialisation method. The k-means++ initialisation method was respectively applied to DS1 and DS2, and the results are indicated as experimental condition, DS1-k-means++ and DS2-k-means++.

All experimental conditions obtained a very low standard deviation which indicates that the results from each run had small variances between them, and their values were close to the mean. Furthermore, the k-means++ initialisation method obtained a standard deviation of zero, which indicates that the results obtained were constant over all runs. The mean inter-cluster distance obtained by all experimental conditions were very similar with a difference of 0.076 between the highest and lowest mean. The results indicate that all experimental conditions, when compared with each other, achieved relatively similar cluster results.

In order to select the best results for the k-means algorithm and the SOM, the sets of 30 performance metric values of the four-experimental conditions were compared using Mann-Whitney U tests at 95% significance. If the first set of experimental conditions statistically significantly outperformed the second set of experimental conditions, a win was granted for the first set of experimental conditions. A draw was recorded if no statistical difference could be observed. If the second set of experimental conditions outperformed the first set of experimental conditions, a loss was recorded against the first set of experimental conditions. A sum of the wins, draws and losses granted was then recorded for all the experimental conditions.
For instance, if the results in which experimental conditions set A compared with other experimental conditions are recorded as 0-1-2, as shown in Table 5.5, it means that experimental condition set A was granted zero wins, one draw and two losses. The total is a difference

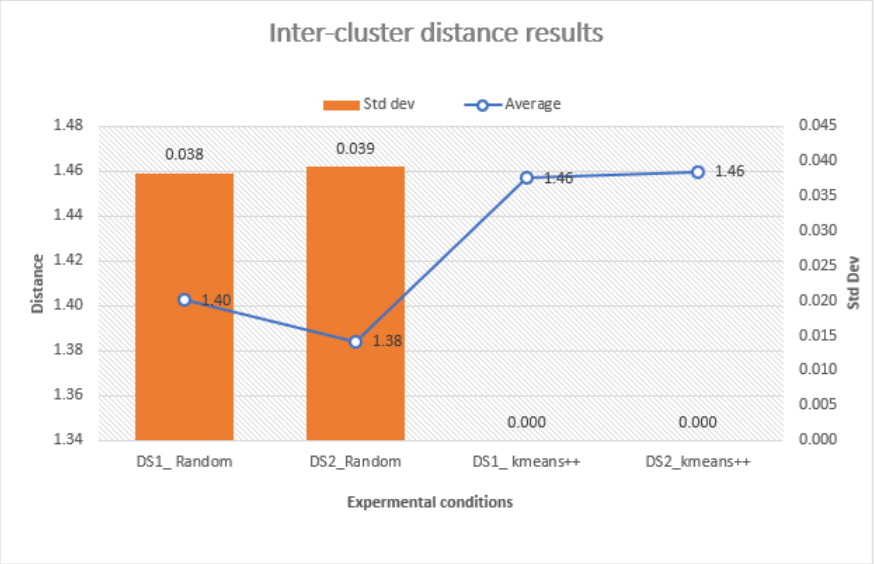


Figure 5.13: K-means algorithms inter-cluster distances

between the number of wins and losses; and a higher value indicates good performance.

Table 5.5 indicates the performance of the k-means algorithm. Experimental conditions set A and B obtained 2 losses, 1 draw and no wins. Experimental conditions set C, which performed better than A and B, obtained 2 wins, 1 loss and zero draws. Experimental conditions set D outperformed the other experimental conditions for inter-cluster distance results. In experimental condition set D, the algorithm was implemented using the DS2 and the k-means++ initialisation method at $K = 7$. The results are aligned with literature stating that k-means++ initialisation method produces better clusters than random initialisation. For DS2, purchase transactions with outliers were removed using the clamp transformation method. K-means is known to be susceptible to outliers; thus, DS2 outperforming DS1 indicates that the removal of outliers gave DS2 some advantage in achieving clusters with a better separation index.

Table 5.5: Hypothesis tests for k-means algorithm inter-cluster distance results

Experimental conditions	Win	Draw	Lose	Total
Experimental conditions set A	0	1	2	−2
Experimental conditions set B	0	1	2	−2
Experimental conditions set C	2	0	1	1
Experimental conditions set D	2	1	0	2

Intra-cluster results and hypothesis tests

The second measure that was used to evaluate clusters was the intra-cluster distance. The intra-cluster distance, which is calculated using equation 3.17, measures how similar the objects in a cluster are to each other. A lower intra-cluster distance indicates better compactness between objects that belong to the same cluster. Figure 5.14 shows the results obtained from the 30 runs. The random initialisation methods obtained very high standard deviations for both DS1 and DS2. A high standard deviation indicates that the results obtained in each run had high variance between the runs and their values were far from the mean. Furthermore, the

k-means++ initialisation method obtained a standard deviation of zero, which indicates that the intra-cluster distances obtained in all runs were the same.

Unlike the results from the inter-cluster distance analysis, there was a significant difference in the mean intra-cluster distances obtained from each set of experimental conditions. The results achieved by the k-means++ initialisation methods were significantly lower than the means of the random initialisation method.

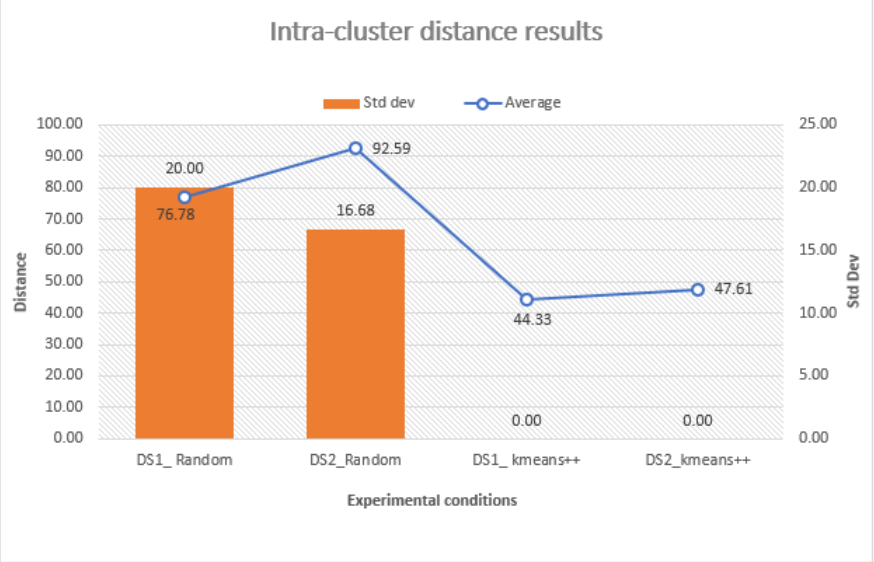


Figure 5.14: K-means algorithms intra-cluster distances

Table 5.6 indicates intra-cluster distance performance for k-means algorithms. For the k-means algorithm, experimental conditions set B, which obtained 3 losses, is the worst performing set. Experimental conditions set A obtained 2 losses, 1 win and zero draws. Experimental conditions set D obtained 2 wins, 1 loss and zero draws. Experimental conditions set C outperformed all other experimental conditions with regard to intra-cluster distance results. In experimental conditions set C, the algorithm was implemented using DS1 and the k-means++ initialisation method at K= 10. Although the removal of outliers seems to have improved the separation of clusters, this removal did not seem to have had much impact on intra-cluster distance as DS1, which has more outliers, performed better than DS2.

Table 5.6: Hypothesis tests for k-means algorithm intra-cluster distance results

Experimental conditions	Win	Draw	Lose	Ttotal
Experimental conditions set A	1	0	2	−1
Experimental conditions set B	0	0	3	−3
Experimental conditions set C	3	0	0	3
Experimental conditions set D	2	0	1	1

5.4.2 Agglomerative hierarchical clustering implementation

Figure 5.15 demonstrates the steps that were followed in implementing the agglomerative hierarchical clustering (AHC) algorithm. The algorithm was run for all possible values of K, which

are 2 to 10 clusters as specified by the organisation's requirements. The two datasets, DS1 and DS2, were trained using the complete linkage and single linkage methods to measure similarities between objects. In single linkage clustering, the distance between two clusters is defined as the shortest distance between two points. Complete linkage clustering measures the longest distance between two points in a cluster. The results obtained from each set of experimental conditions were also evaluated using the inter-cluster and intra-cluster distance.

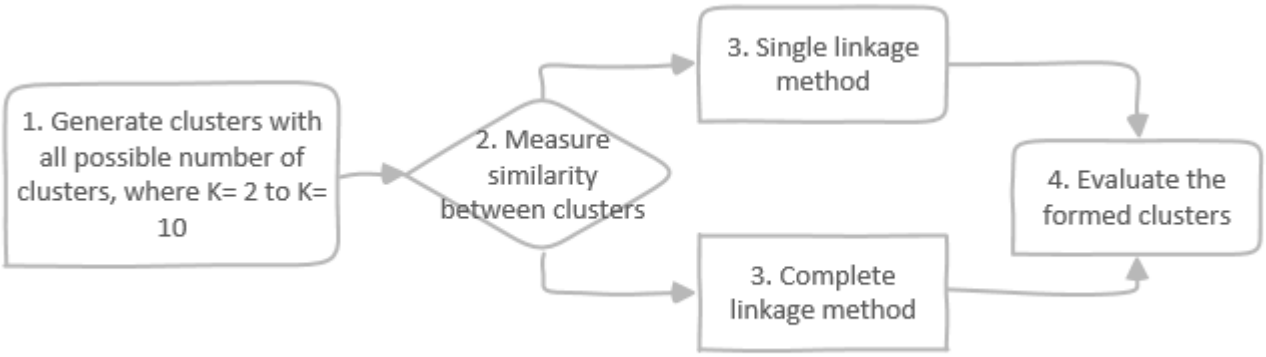


Figure 5.15: Approach used for implementing AHC algorithm

Inter-cluster results

After the clusters were generated for all values for K, the best inter-cluster distance was selected for each set of experimental conditions as shown in Figure 5.16. DS1 obtained the best results at K= 10 for both the single and complete linkage methods. DS2 obtained the best results at K= 10 for the single linkage method, and in the complete-linkage method, the best results were obtained at K= 2.

Figure 5.16: AHC algorithms inter-cluster distances

The AHC algorithm uses a deterministic approach and the results do not change with each run; hence its results were not compared using a Mann-Whitney U test at 95% significance. The

highest inter-cluster distance was achieved by experimental conditions set B with inter-cluster distance of 2.02 at $K = 2$. In experimental conditions set B, the algorithm was implemented using DS2 and the complete linkage method. The complete linkage method tends to produce clusters that are tightly bound or more compact than those obtained by the single linkage method. Furthermore, the complete linkage method also performed better in DS2, which had fewer outliers.

Intra-cluster results

After the clusters were generated for all values of K, the best intra-cluster distances were selected for each set of experimental conditions. Figure 5.17 shows the lowest intra-cluster distance obtained from each set of experimental conditions. DS1 obtained the best results at $K = 9$ for both the single and complete linkage methods. Furthermore, DS2 obtained best results at $K = 10$ for the single linkage method, and the complete linkage method obtained best results at $K = 4$.

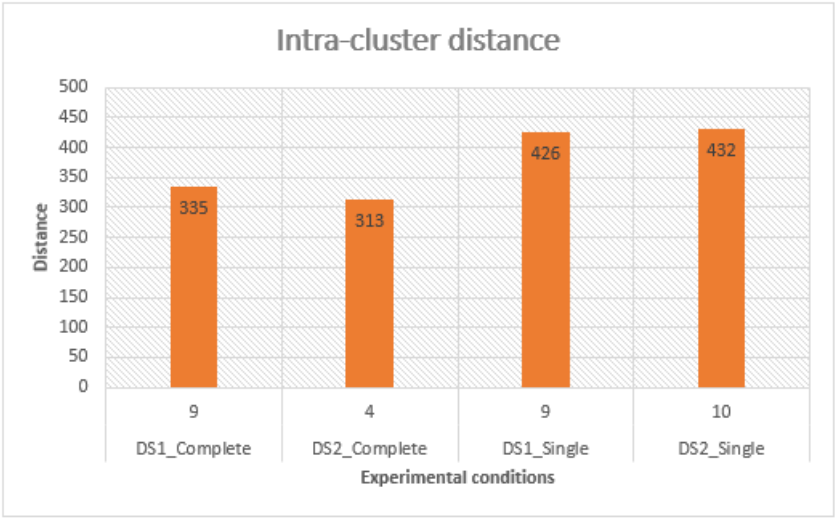


Figure 5.17: AHC algorithms intra-cluster distances

The AHC algorithm achieved the lowest intra-cluster distance at $K = 4$ by experimental conditions set B with intra-cluster distance= 313. In experimental conditions set B, the algorithm was implemented using DS2 and the complete linkage method. The experimental conditions DS2 and complete linkage obtained best results with differing K values for intra- and inter-distance measures.

5.4.3 Self-organising map

Figure 5.18 demonstrates the steps that were followed in implementing a SOM. First, the size of the output map was specified. A square output map where length (x) and width (y) are equal was used. The dimensions of the output map were calculated using the formula defined by Vesanto [70]: $5 \times \sqrt{n}$. This method obtained an output map with 17 by 17 dimensions.

Quantisation error (QE) is one of the common measures used to measure the quality of SOM results and a low QE is an indication of more accurate results. Generally, the QE declines when the output map increases, which means it is not advisable to use QE to evaluate output maps of different sizes. In implementing the SOM, the size of the output map was calculated at the

beginning and kept constant throughout the implementation process.

The random initialisation method was applied for initialising the weight vectors. The initialisation phase was followed by the training process, where the input data was trained using two methods, the random and batch training methods. The random training method's weights are updated after each feature presentation and the batch training method's weight values are updated only after all patterns have been presented.

In the implementation process, the SOM's objective was to determine the values of $\eta(t)$ and $\sigma(t)$ that minimised QE. To obtain clusters, the Ward clustering method was applied to the SOM results. The optimal number of clusters (K) was selected using the silhouette coefficient (SC). In order to address the randomness caused by the weight vectors initialisation, the algorithm was run 30 times for each set of experimental conditions.

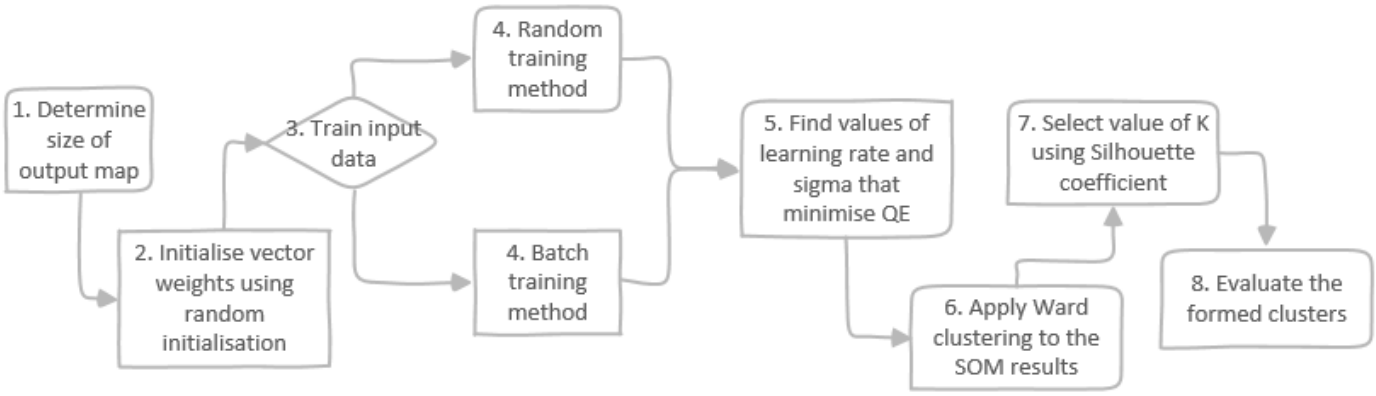


Figure 5.18: Approach used for implementing SOM

The SC method, applied in each run, was used to determine the best value of K. The results show that the algorithm obtained a different 'best K value' for each run. Figure 5.19 shows the number of runs that each K value obtained the highest SC value. For the random training method, DS1 obtained most best SC value at $K = 10$. From the 30 runs, DS1 obtained the best SC value at 10 clusters in seven runs. These results indicate that the 30 runs obtained high SC values from a wide range of K values. DS2 obtained the best SC value at 10 clusters in 16 runs, which shows that in most runs, the high SC value was obtained at 10 clusters. The batch training method, for both DS1 and DS2, also obtained most best SC at $K = 10$. From the 30 runs, DS1 obtained the best SC value at 10 clusters in 22 runs, and DS2 obtained the best SC value at 10 clusters in 16 runs. The results shows that all experimental conditions obtained best results at $K = 10$.

Inter-cluster results and hypothesis tests

Figure 5.20 shows the results obtained from the 30 runs for each set of experimental conditions. The line graph and the bar graph show the mean and standard deviation for the inter-cluster distance. All experimental conditions obtained very low standard deviations which indicates that the results obtained in each run had small variances between the runs, and the results

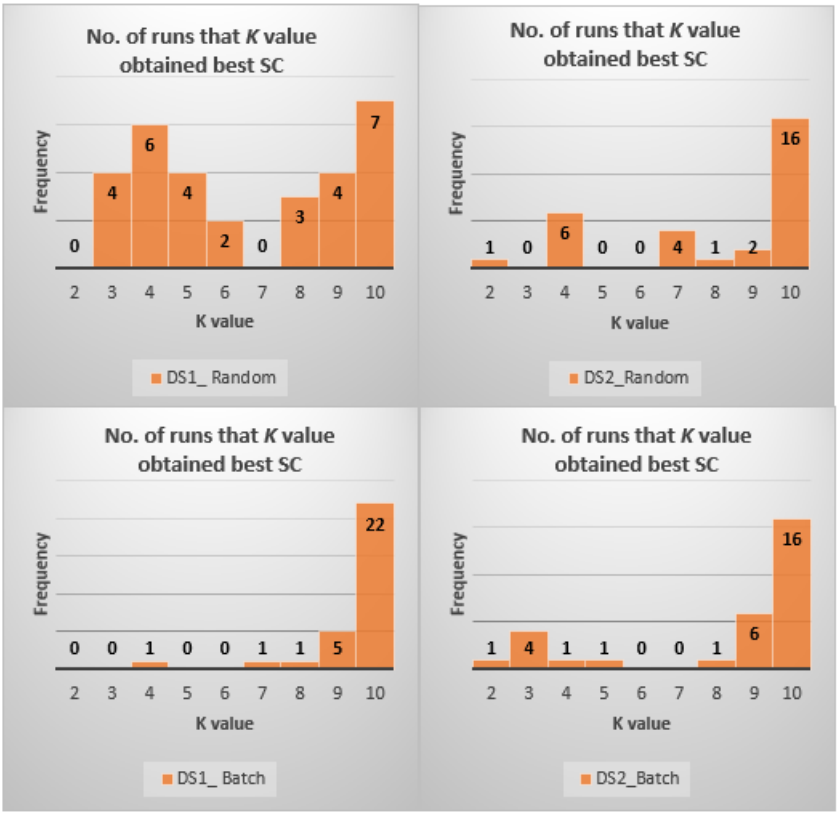


Figure 5.19: No. of runs that K value obtained best SC

from each run were close to the mean. The mean values obtained by all experimental conditions were almost the same.

Table 5.7 indicates the performance of a SOM. Experimental conditions set C is the worst performing and it obtained 2 losses, 1 draw and zero wins. Experimental conditions set B obtained 2 draws, 1 loss and zero win. Experimental conditions set B obtained 1 win, 1 draw and 1 loss. Experimental conditions set A outperformed other experimental conditions for inter-cluster distance results. For experimental conditions set A, SOM was implemented using DS1 and the random training method. Experimental conditions set A obtained the best results at K= 10, where $\eta(t)$= 1.8 and $\sigma(t)$= 3.6.

Table 5.7: Hypothesis tests for SOM algorithm inter-cluster distance results

Experimental conditions	Win	Draw	Lose	Total
Experimental conditions set A	3	0	0	3
Experimental conditions set B	0	2	1	−1
Experimental conditions set C	0	1	2	−2
Experimental conditions set D	1	1	1	0

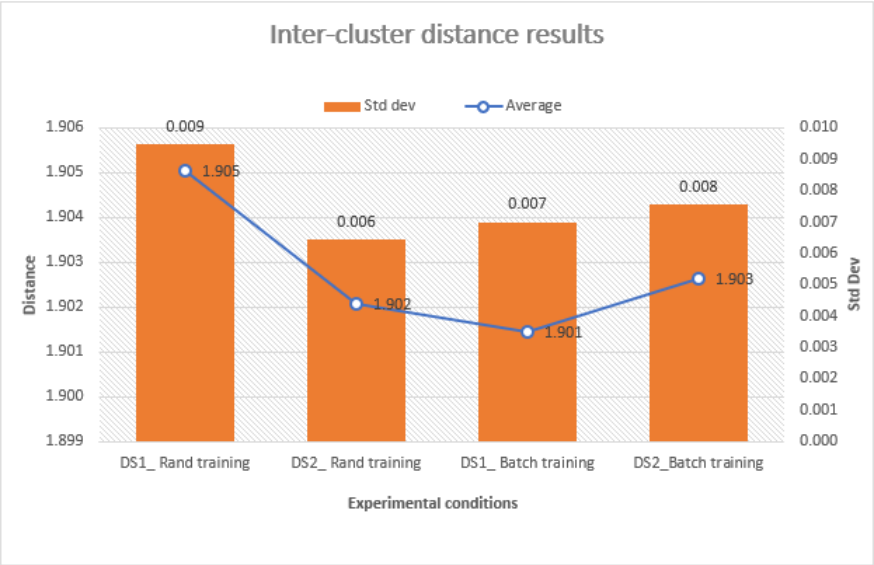


Figure 5.20: SOM algorithms inter-cluster distances

Intra-cluster results and hypothesis tests

Figure 5.21 shows the intra-cluster results obtained from the 30 runs for each set of experimental conditions. All of the experimental conditions obtained low standard deviations and similar mean values with a difference of less than one between the highest and lowest value. The results indicate that the SOM obtained similar results in all the runs for all experimental conditions.

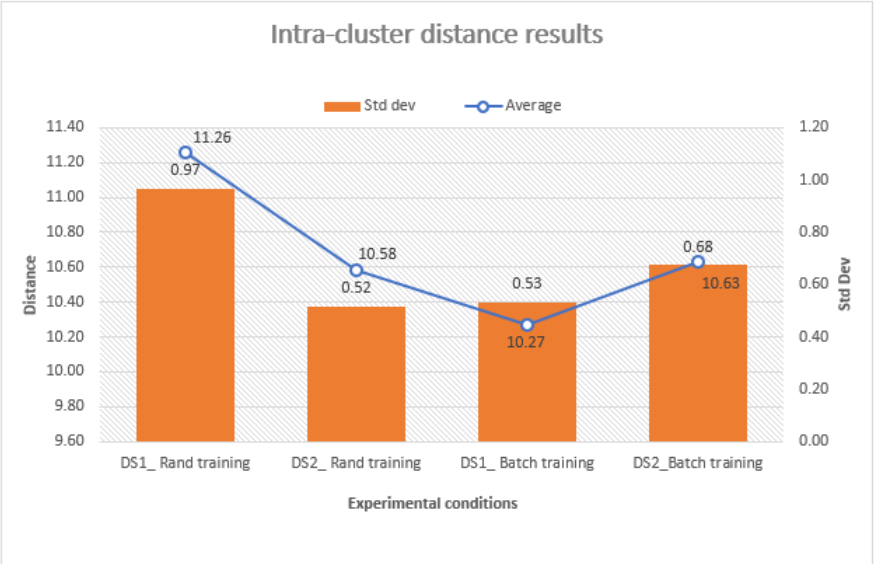


Figure 5.21: SOM algorithms intra-cluster distances

Table 5.8 indicates intra-cluster distance performance for SOM. Experimental conditions set A,B,C and D obtained same results, 3 draws, zero win and zero loss. In this SOM implementation, the hypothesis test results indicate that there was no significant difference in the performance of each set of experimental conditions. The results indicate that all experimental conditions obtained clusters which were equally compact. Consequently, experimental condi-

tions set A, which obtained the best results in the inter-cluster distance analysis, was selected to match the winning set of experimental conditions of the inter-cluster distance analysis.

Table 5.8: Hypothesis tests for SOM algorithm intra-cluster distance results

Experimental conditions	Win	Draw	Lose	Total
Experimental conditions set A	0	3	0	0
Experimental conditions set B	0	3	0	0
Experimental conditions set C	0	3	0	0
Experimental conditions set D	0	3	0	0

5.4.4 Final results summary

To summarise, the following experimental conditions obtained the best inter-cluster distance results in their respective algorithms: Experimental conditions set D in k-means; experimental conditions set B in AHC and experimental conditions set A in SOM. The experimental conditions for each technique are shown in Table 5.9. When the results were compared, the SOM results statistically significantly outperformed both the other algorithms, achieving 2 wins, 0 draws and 0 losses. Therefore, the SOM obtained best clustering results when it was evaluated using the inter-cluster distance method.

Table 5.9: Best parameter combination for each algorithm

Algorithm	Experimental conditions	Conditions	K value
K-means	Experimental conditions set D	DS2, k-means++ init	7
AHC	Experimental conditions set B	DS2, complete linkage	2
SOM	Experimental conditions set A	DS1, random training	10

The following experimental conditions obtained the best intra-cluster distance results in their respective algorithms: Experimental conditions set C in k-means algorithm; experimental conditions set B in AHC and experimental conditions set A in SOM. The experimental conditions for each combination are stated in Table 5.10. When the results were compared, the SOM results statistically significantly outperformed both the other algorithms. Therefore, SOM obtained the best clustering results when evaluated using the intra-cluster distance method.

Table 5.10: Best parameter combination for each algorithm

Algorithm	Experimental conditions	Conditions	K value
K-means	Experimental conditions set C	DS1, k-means++ init	10
AHC	Experimental conditions set B	DS2, complete linkage	4
SOM	Experimental conditions set A	DS1, random training	10

5.5 Chapter summary

This chapter implements various clustering techniques using a real-world dataset. Various experiments are conducted for each technique in order to evaluate the impact that different settings have on cluster results. The CRISP-DM process is used to define the requirements of the business, analyse and understand the available data, apply selected clustering techniques and finally, evaluate the results. The cluster results are evaluated using the inter-cluster distances and intra-cluster distance measures.

Chapter 6

Cluster analysis and recommendations

In this chapter, the resulting models are analysed. Section 6.1 discusses the results obtained from the SOM and what the results mean for the organisation's requirements which were defined in section 5.1. Section 6.2 analyses the clusters obtained from the SOM and uses insights gained to make recommendations to the organisation.

6.1 Cluster analysis of SOM results

The SOM was selected as the best performing technique, because it obtained the best results with respect to both evaluation criteria; the inter-cluster and intra-cluster distances. The SOM was implemented in two phases. The first phase was to train the SOM and to obtain the codebook vectors, and the second phase was to cluster the SOM results using Ward clustering. Figure 6.1 shows the U-matrix after the SOM was trained (first phase). The dark colours depict closely spaced node codebook vectors and lighter colours indicate more widely separated node codebook vectors. Thus, groups of dark colours can be considered as clusters, and the light parts as the boundaries between the clusters. Figure 6.2 shows a dendrogram where the SOM results were further clustered using a Ward clustering algorithm (second phase). In addition to visualising the complete SOM map as illustrated in Figure 6.1, the relative components values in the codebook is illustrated in Figure 6.3 where each component plane is constructed for each feature to visualise distribution of the corresponding weight using a colour scale representation.

Table 6.1 shows the number of farmers belonging to each cluster and the location of the factory that the farmers supplied cassava to. It is worth noting that the *location of factory* feature was not actively used in the clustering process. The variable was removed in the feature selection step as it was perfectly correlated with the *location of plots* feature.

Despite being removed from the clustering process, the *location of factory* feature had a significant influence on the formation of clusters. Almost all clusters formed were highly dependent on the *location of factory* feature as all farmers in the same cluster supplied to the same factory. Cluster 2 was the only cluster with an overlap where 33% of farmers supplied to a different factory.

Unlike the *location of factory* feature which was removed from the dataset, the *location of plots* feature was actively used in the clustering process and the cluster results are summarised in Table 6.2 and Table 6.3. The site in Ribáuè sourced cassava from two districts; Ribáuè district and Mecubúri district. Farmers from clusters 1, 4 and 5 were all from the same district, and cluster 3 had an overlap where 63% of farmers were from Ribáuè district and 37% were from Mecubúri district. The Inharimme site, a larger processing site, sourced from seven districts.

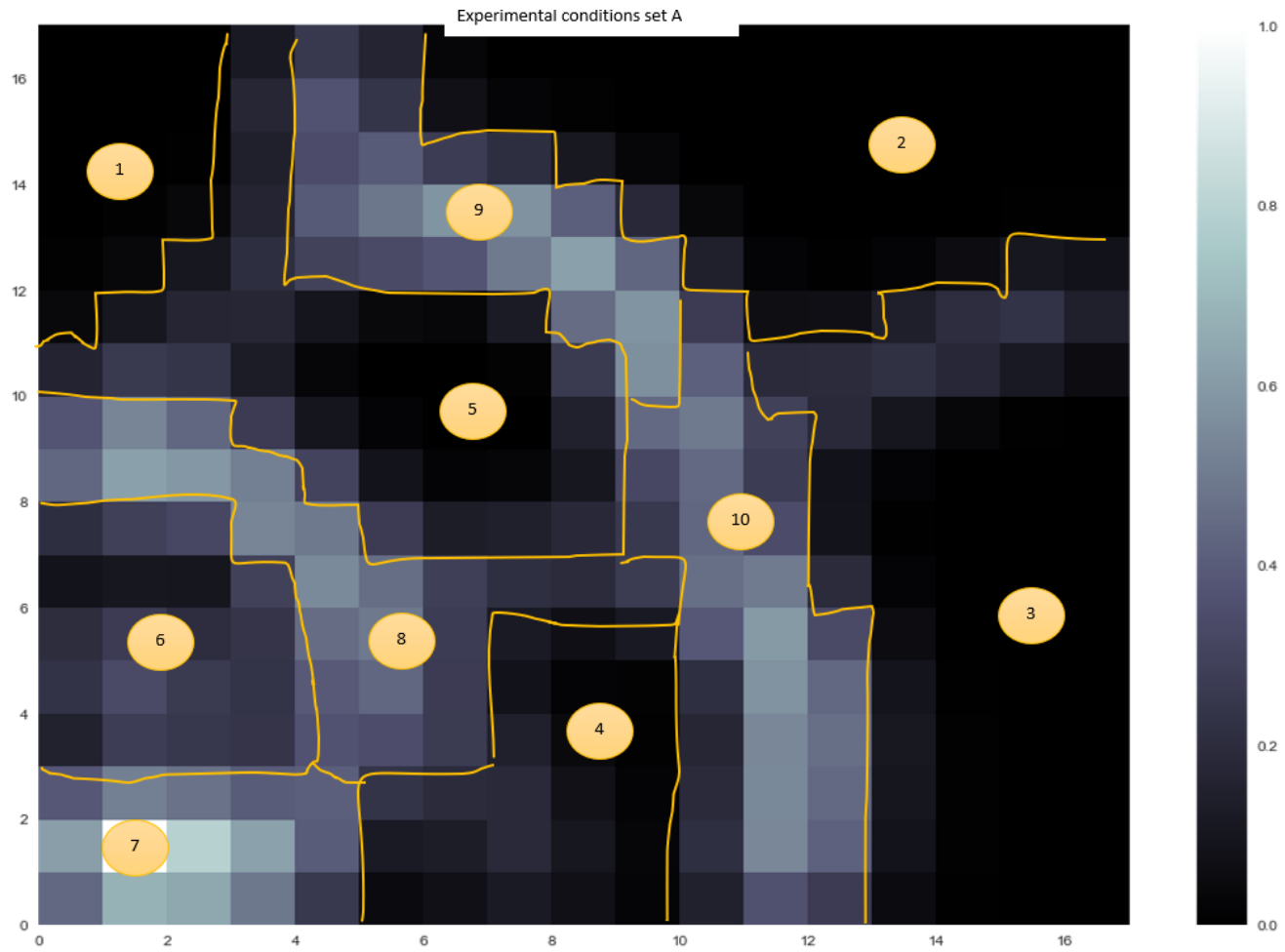


Figure 6.1: Umatrix after SOM training

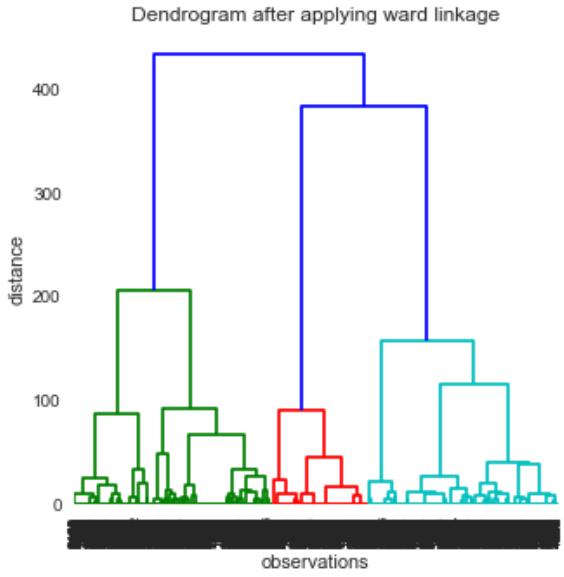


Figure 6.2: Dendrogram after applying Ward clustering to SOM results

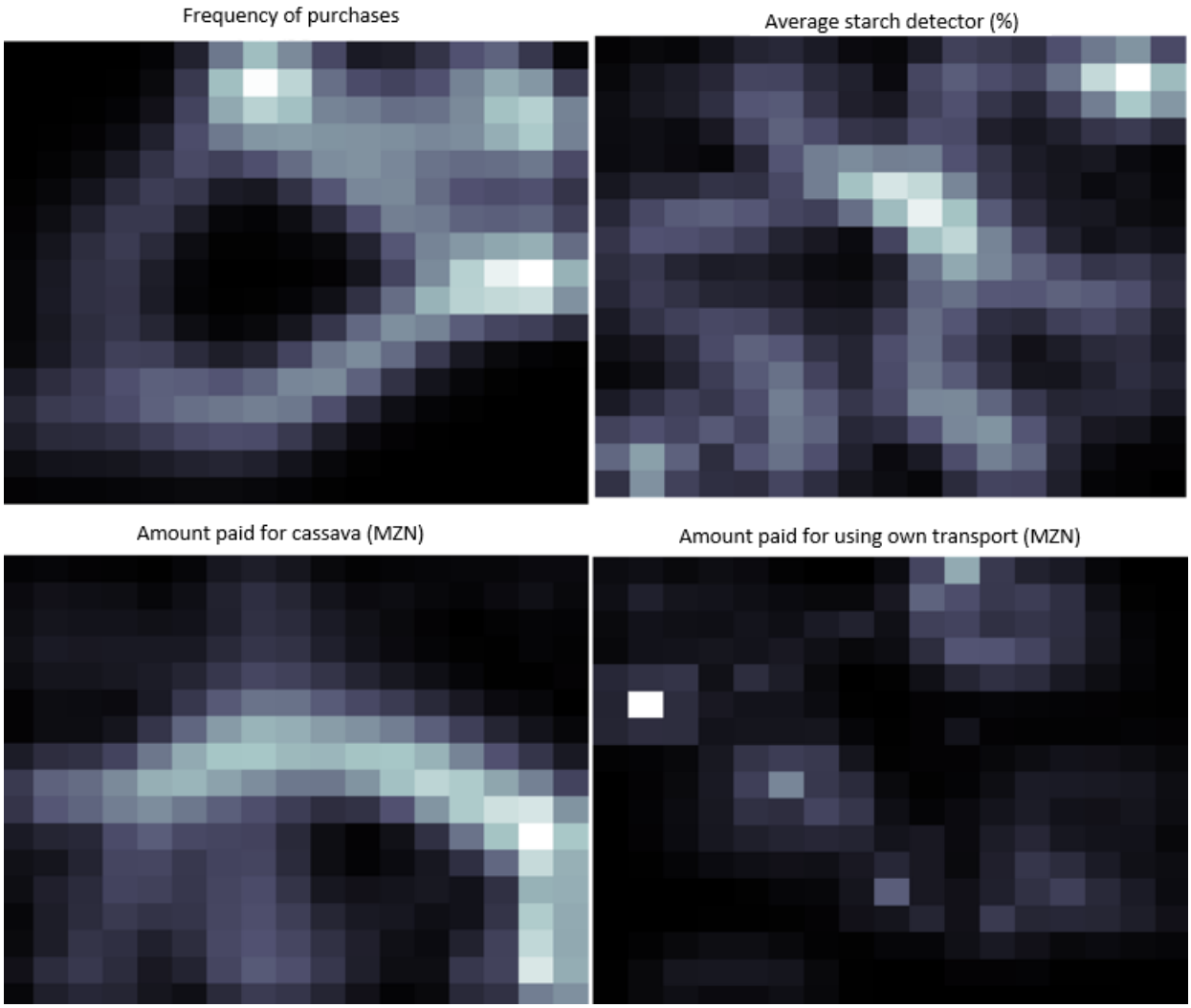


Figure 6.3: component planes for features

Table 6.1: Farmers per location of factory

Cluster	No. of farmers	farmers supply to Ribáuè factory	farmers supply to Inharrime factory
1	408	100%	0%
2	168	67%	33%
3	193	100%	0%
4	337	100%	0%
5	333	100%	0%
6	204	0%	100%
7	491	0%	100%
8	276	0%	100%
9	491	0%	100%
10	605	0%	100%

Table 6.3 shows that farmers from clusters 7, 9 and 10 consist of farmers from same districts.

The results indicate that the location of plots had a substantial impact on the formation of clusters. As defined in the business understanding step, the organisation's clustering criteria were defined as; the risk of supply, quality of raw materials (cassava crop), and effectiveness of the organisation's operations. However, the algorithm clustered farmers based solely on their locations; consequently overlooking other features which also form part of the organisation's clustering criteria. In order to address this issue, the *location of plot* feature was removed from the dataset, and the adjusted dataset was trained using a SOM. The *modified variety?* feature was also removed from the dataset as the quality of raw materials is primarily influenced by the *starch content* feature and not the crops' modification status.

Table 6.2: Nampula farmers per location of plots

Cluster	No. of farmers	Farmers from Ribáuè	Farmers from Mecubúri
1	408	100%	0%
3	193	63%	37%
4	337	100%	0%
5	333	100%	0%

Table 6.3: Inhambane farmers per location of plots

Cluster	No. of farmers	Inharr farmers	Zavala farmers	Janga farmers	Morrumb farmers	Maxixe farmers	Homoi farmers	Manjac farmers
6	204	0%	0%	0%	86%	6%	6%	4%
7	491	0%	100%	0%	0%	0%	0%	0%
8	276	0%	0%	99%	0%	1%	0%	0%
9	491	100%	0%	0%	0%	0%	0%	0%
10	605	100%	0%	0%	0%	0%	0%	0%

After removing the *location of plot* and *modified variety?* features from the dataset, the question arose of whether SOM would still be the best algorithm. As a result, all experiments were rerun. Table 6.4 shows the mean and standard deviations (std dev) of the inter- and intra-cluster distance measures for each algorithm. For each algorithm, the set of experimental conditions that obtained the best results was selected and the SOM again outperformed the k-means algorithm and AHC.

The experimental settings of the updated dataset were compared using Mann-Whitney U tests at 95% significance and the results are shown in Table 6.5. There is no difference between experimental conditions set A, C and D for inter-cluster distance, but experimental conditions set B performs significantly worse than the other three options. Moreover, there is no difference between experimental conditions set A and B for intra-cluster distance, but experimental conditions set C is significantly worse than the other options.

The inter- and intra-cluster distance results indicate that for each measure, more than one experimental condition obtained clusters which were equally separated and equally compact.

Table 6.4: Results of the updated dataset for each algorithm

Algorithm and conditions	Inter-cluster distance		Intra-cluster distance	
	Mean	Std dev	Mean	Std dev
K-means (DS1 $k$means++)	0.357	0.000	167.5	0.000
AHC (DS2 single linkage)	1.70	N/A	73.48	N/A
SOM (DS1 random)	10.60	0.789	1.89	0.005

However, experimental conditions set A is the only setting that obtained the best results in both inter- and intra-cluster distance. For experimental conditions set A, SOM was implemented using DS1 and the random training method. Experimental conditions set A obtained the best results at $K = 10$.

Table 6.5: Hypothesis tests of the updated dataset for SOM algorithm

Experimental conditions	Inter-cluster distance				Intra-cluster distance			
	Win	Draw	Loss	Total	Win	Draw	Loss	Total
Experimental conditions set A	1	2	0	1	1	2	0	1
Experimental conditions set B	0	0	3	−3	1	2	0	1
Experimental conditions set C	1	2	0	1	0	1	2	−2
Experimental conditions set D	1	2	0	1	0	3	0	0

Table 6.6 compares the distribution of farmers' locations in each cluster. The results in Table 6.6 (a) shows the clusters before the *location of plot* feature was removed. In Table 6.6 (b), the algorithm was trained without locations feature, and the results indicate that the clustering process was not influenced by the location of farmers as each cluster consists of farmers from both locations.

Table 6.6: Cluster results per location of factory

(a) Results before removal of feature

Cluster	Ribáuè	Inharrime
1	100%	0%
2	67%	33%
3	100%	0%
4	100%	0%
5	100%	0%
6	0%	100%
7	0%	100%
8	0%	100%
9	0%	100%
10	0%	100%

(b) Results after removal of feature

Cluster	Ribáuè	Inharrime
1	41%	59%
2	42%	58%
3	39%	61%
4	37%	63%
5	38%	62%
6	45%	55%
7	39%	61%
8	38%	62%
9	40%	60%
10	39%	61%

Sections 6.1.1, 6.1.2 and 6.1.3 evaluate cluster analysis results based on the three defined criteria: supply risk, effectiveness of operations and performance improvements.

6.1.1 Criterion 1: Supply risk

The organisation described the risk of supply as its most substantial risk. The organisation sources cassava from smallholder farmers, and does not have a guarantee that a farmer will supply cassava in the long term. Most farmers do not have long-term plan to farm cassava; they either switch to another crop or discontinue farming if a different opportunity comes up. Table 6.7 shows the total number of purchases per cluster and the average number of purchases per farmer. The results show that most of the farmers have supplied cassava to the site only once.

The results were evaluated using the performance indicator outlined in Table 6.8. A score of 1 indicates low risk and 5 represents high risk. Based on these indicators, farmers from cluster 10 were considered to be low risk as they had supplied cassava more than twice in the analysis period. The analysis period was 26 months, thus for farmers to have supplied cassava more than twice not only indicates commitment but good farming skills which enabled those farmers to cultivate cassava in a shorter cycle. Another possibility is that some farmers in cluster 10 have more than one plot and were thus able to supply more frequently.

Table 6.7: No. of deliveries made by farmers

Cluster	No. of farmers	Total no. of purchases	No. of purchases per farmer
1	299	300	1.00
2	366	374	1.02
3	318	590	1.86
4	559	566	1.01
5	194	194	1.00
6	223	223	1.00
7	378	731	1.93
8	606	665	1.10
9	287	369	1.29
10	276	740	2.68

Table 6.8: Risk factor indicators

Risk Factor	No.of purchases
1	3 or higher
2	2.5 to 3
3	2 to 2.5
4	1.5 to 2
5	1 to 1.5

6.1.2 Criterion 2: Effectiveness of operations

The organisation had identified the transportation of roots as one of the key factors that directly impact the effectiveness of its operations. The majority of the cassava suppliers have no means of transporting cassava to the site and rely on the organisation's trucks. Most of the plots are

geographically scattered, which pose a great challenge to the organisation's logistics and cost of operations. Only 13% of farmers were able to organise their own transport to deliver cassava to the site. This percentage is very low, even though farmers are compensated for organising their own transport.

The organisation plans to address this challenge by encouraging farmers to form groups where each group can align to harvest together and hire a larger truck to deliver cassava to the site. The *transport cost* feature was used to measure the impact that farmers have on the effectiveness of the organisation's operations. Clusters with a large number of farmers who delivered cassava using their own transport indicate farmers' high potential in managing their own logistics to deliver cassava to site. Table 6.9 shows the number of farmers from each cluster who organised their own transport to deliver cassava to the site. The results were evaluated using the performance indicator outlined in Table 6.10. A score of 1 indicates low effectiveness and 5 represents high effectiveness. Based on these indicators, farmers from cluster 7 achieved high effectiveness scores as the cluster had a significant number of farmers who organised their own transport to deliver cassava.

Table 6.9: Farmers who organised own transport

Cluster	% of farmers
1	0%
2	3%
3	3%
4	3%
5	0%
6	1%
7	83%
8	4%
9	0%
10	2%

Table 6.10: Effectiveness factor indicators

Effectiveness factor	% of farmers
1	< 20
2	20 to 40
3	40 to 60
4	60 to 80
5	> 80

6.1.3 Criterion 3: Performance improvements

With regard to the performance improvements criterion, the organisation stated that for it to be able to sustain and grow the business, the amount of cassava required will need to increase

exponentially. Furthermore, in order to operate profitably, the cassava delivered needs to be of good quality. These two performance factors were measured using the *cassava quantity* and *starch content* features. Table 6.11 shows the average amount of cassava delivered per farmer. A higher value indicates farmers who delivered high volumes of cassava. Moreover, a higher starch content indicates that the cassava delivered was of higher quality. Farmers who delivered larger volumes of cassava or cassava with high starch content show a greater potential of playing a role in the organisation's growth and profitability.

The results were evaluated using the performance indicator outlined in Table 6.12. A score of 1 indicates low performance and 5 denotes high performance. Based on these indicators, farmers from cluster 10 achieved good performance in terms of volumes delivered. Each farmer delivered an average of 9.6 tons of cassava. On the other hand, cluster 2 delivered cassava with high starch content but in very low volumes. These results show that certain farmers showed strength in one factor but weakness in another factor.

Table 6.11: Performance of farmers

Cluster	Cassava quantity per farmer	Starch content
1	2.5	21.9
2	2.5	26
3	7.2	22.4
4	2.3	17.4
5	1.7	19.3
6	3.1	20
7	7.1	17.8
8	3.1	12.8
9	5.9	16.2
10	10.6	19.9

Table 6.12: Performance factor indicators

Performance factor	Cassava quantity per farmer	Starch content
1	< 2	< 13
2	2 to 5	13 to 18
3	5 to 8	18 to 23
4	8 to 10	23 to 28
5	10 or higher	28 or higher

6.2 Deployment and recommendations

Figure 6.4 shows a summary of scores that each cluster obtained for each factor. It is worth noting that the performance score consisted of two features: *cassava quantity* and *starch content*. In order to obtain one performance score, a weighted average score was calculated with *cassava quantity* feature given 70% weight and *starch content* feature given 30%. The organisation explained that quantities delivered by a farmer were more important because farmers with higher yield would not only increase production, but would also reduce the organisation's

logistic burden as the organisation would thus obtain a full truck-load from fewer farmers. Furthermore, the organisation explained that if farmers with higher yield are provided with good variety cassava stems, they will be able to improve both quantity and quality of cassava delivered.

The results in Figure 6.4 indicate that most clusters are strong in one area but weak in another area. For instance, farmers in cluster 7 are considered to be high risk but also very effective as over 80% of cluster 7 farmers organised their own transport.

Cluster	Risk Factor	Effectiveness Factor	Performance Factor
1	5.0	1.0	2.3
2	5.0	1.0	2.6
3	4.0	1.0	3.1
4	5.0	1.0	2.0
5	5.0	1.0	1.6
6	5.0	1.0	2.3
7	4.0	5.0	2.7
8	5.0	1.0	1.7
9	5.0	1.0	2.7
10	2.0	1.0	4.4

where:	
Excellent =	
Good =	
Average =	
Poor =	
Worse =	

Figure 6.4: Overall results for each cluster

The position of each cluster is discussed in the list below:

1. **Cluster 1**: Farmers in the first cluster are considered to be high risk, with worst effectiveness and poor performance scores. The results show that the majority of these farmers have supplied cassava to the organisation only once, and they have relied on the company for the delivery of cassava. Moreover, they supplied less than two tons of cassava per farmer, and the cassava delivered was of poor quality with starch content of less than 13%.

   These farmers have not shown strong capability in any criterion; thus the organisation should focus on assisting them with basic farming principles. First, the organisation needs to address the high-risk level by understanding why the farmers have not re-supplied cassava to the organisation. One way to address this obstacle is to explain the organisation's vision to the farmers and reassure them that the processing plants will exist for the long term, thus there will be a stable, reliable buyer for the cassava produced. The field workers should also establish demonstration plots which they can use to teach these farmers basic farming practices for growing cassava efficiently. At this stage, the organisation should not form any close relationship with these farmers or provide them with any resources.

The organisation should observe these farmers closely and choose to either abandon or improve them based on their progress after multiple training workshops.

2. **Cluster 2**: The performance of farmers in cluster 2 is similar to farmers in cluster 1, thus, the same intervention approach explained in cluster 1 should be applied for farmers in cluster 2.

3. **Cluster 3**: Farmers in cluster 3 had a slightly lower risk level, with average performance and worst effectiveness. During a period of 26 months, this cluster had a considerable number of farmers who managed to deliver cassava to the organisation more than once. The average volumes and starch content of cassava delivered by each farmer was 7.2 tons and 22%, respectively.

   Farmers in this cluster show high potential and the organisation should consider investing in their development. The development skills should focus on teaching these farmers best farming practices and encourage them to collaborate with other like-minded farmers. Furthermore, the organisation should consider developing customised farming guidelines for these farmers taking into account various conditions such as soil and other climatic factors.

4. **Cluster 4**: The performance of farmers in cluster 4 is similar to cluster 1 farmers, thus, the same intervention approach explained in cluster 1 should be applied for farmers in cluster 4.

5. **Cluster 5**: Farmers in cluster 5 had obtained the worst scores in all criteria. These farmers should be treated as transactional suppliers, where the organisation has a simple buyer and seller transactional arrangement with them, guided mainly by order fulfilment. Furthermore, the farmers should be provided with information about the organisation's growth strategy and their requirements from suppliers. The organisation should define an attainable goal that can be used as an indicator of the farmer's willingness to commit to the organisation. If they meet the defined goal, they should be managed using the strategy defined for cluster 1 farmers.

6. **Cluster 6**: The performance of farmers in cluster 6 is similar to the cluster 1 farmers, thus the same intervention approach explained in cluster 1 should be applied for farmers in cluster 6.

7. **Cluster 7**: Farmers in this cluster had a slightly lower risk and excellent effectiveness. Over 80% of farmers in this cluster delivered cassava using transport organised by themselves. The ability to organise own transport has a significant impact on the effectiveness of the organisation's operations as it reduces the burden on the organisation's logistics. The organisation should form a close relationship with these farmers and organise a workshop that encourages them to collaborate and align with each other. By collaborating and aligning, the farmers can harvest in the same period where they can hire a larger truck to collect cassava. Furthermore, the company should invest in the farmers' development, similar to the interventions suggested for cluster 3 farmers.

8. **Cluster 8**: The performance of farmers in cluster 8 is similar to that of the farmers in cluster 5, thus, the same intervention approach explained in cluster 5 should be applied to the farmers in cluster 8.

9. **Cluster 9**: The performance of farmers in cluster 9 is similar to that of the farmers in cluster 1, thus the same intervention approach explained in cluster 1 should be applied for farmers in cluster 9.

10. **Cluster 10**: Farmers in this cluster had a low risk score and excellent performance. The majority of farmers in this cluster had delivered cassava more than twice, and each farmer had delivered an average of 10.6 tons of cassava. There is a high likelihood that most farmers in this cluster had more than one plot or that their plots are larger, which enabled them to deliver high quantities in 26 months.

Cassava will continue to contribute to food security, but its potential as a raw material for cassava processing can also make the crop an important contributor to alleviating poverty among smallholder farmers. This organisation can also eventually provide significant input to the development of cassava processing in Mozambique. The various challenges faced by farmers from different clusters in this study are an indication that any initiative to process cassava will need a clear strategy considering not only the market demand but all of the conditions needed to build an efficient and feasible industry.

A central decision within SRM is to determine the specific interventions and interactions an organisation should have with their supply base in order to achieve its strategic goals. There is no best practice type of relationship which applies to all categories of suppliers. Therefore, interventions need to be adapted to the type of relationship an organisation wishes to establish and maintain with each cluster of suppliers. The improvement methods discussed have been summarised into four intervention strategies; namely, inform and observe, educate, develop and invest. The four key strategies are explained below:

- **Inform and observe**: In this strategy, the organisation informs the farmers about its growth strategy and its requirements. Then the organisation monitors farmers' progress to determine if they have the willingness and potential to grow. If the farmers show good progress, they move to the *educate* strategy. If they show no progress, the farmer remains at arm's length where the interaction with the organisation is only transactional. This approach applies to clusters 5 and 8.

- **Educate**: The key objective of this strategy is to improve yields substantially. Despite evidence that smallholder farmers can increase productivity, achieving high yield will be a long process that will require strong technical support to farmers through a solid and steady extension services network. The organisation should carry out a massive campaign to mentor smallholder farmers, and the focus should be primarily on improving their capacity to produce high-quality fresh roots to sell in a more demanding market. The organisation should consider distributing improved high-yielding cassava varieties and pesticides to these farmers. Lastly, the organisation should establish demonstration plots to provide farmers with hands-on training. Clusters 1, 2, 4, 6 and 9 should be managed using this strategy.

- **Develop**: The organisation should encourage the farmers to organise themselves into associations, generally with 20 members on average. Furthermore, the organisation should consider introducing mutually beneficial service level agreements with farmers. In addition to ensuring that these farmers are equipped with adequate farming inputs, the organisation should consider enhancing agricultural practices that will lead to a sustainable form of production, for instance, by improving soil fertility and water management. This strategy applies to clusters 3 and 7.

- **Invest**: This strategy aims to form a partnership between farmers and the organisation. The farmers have shown high capability and willingness to grow, and the organisation should empower and invest in them so they can progress into becoming commercial farmers. Farmers should be encouraged to organise into associations to facilitate access to service provider support and reinforce their capacity to practice commercial farming. The organisation should assist farmers with financing, either from specialised financial institutions or from specific development programmes, to support these farmers through tailored financial schemes. This strategy applies to cluster 10.

In the business understanding phase, the organisation requested that the number of clusters should not exceed 10 to avoid a situation where one field worker is responsible for developing more than one intervention strategy. As a result, the number of clusters from each algorithm was limited to 10. The best results obtained using the SOM is where $K = 10$.

Further cluster analysis showed that certain clusters are more similar and can be managed using the same intervention strategy. As a result, the total number of strategies to manage the ten clusters was only four. Field workers need to be allocated to different strategies where they will be responsible for developing the suggested intervention strategies. It is important to note that the field workers are assigned to farmers based on the location of plots. Therefore, once the strategies are developed and approved, each fieldworker will be responsible for implementing all strategies as one location area is likely to have farmers belonging to different clusters.

Figure 6.5 and Figure 6.6 show the distribution of farmers and the selected strategies to manage and develop each of them. Each location area in Figure 6.5 includes farmers in each of the four strategies, which means field workers allocated to these location areas will have to implement all strategies accordingly. The results show that each location consists of farmers with different levels of skills and experiences.

In all locations, the largest proportion of farmers would benefit from the *educate* intervention strategy, and the lowest proportion of farmers qualify for the *invest* intervention strategy. The results indicate that most of the farmers who have supplied cassava to the organisation do not have all the required capabilities, but they have shown some level of willingness to commit and support the organisation's strategic goals.

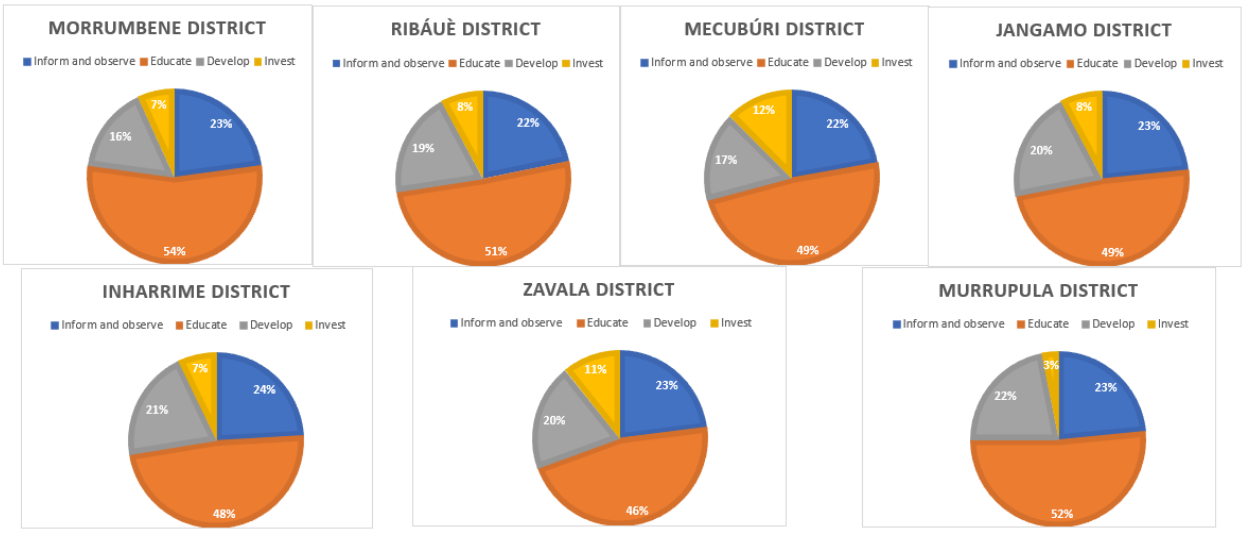


Figure 6.5: Location areas where all four strategies apply

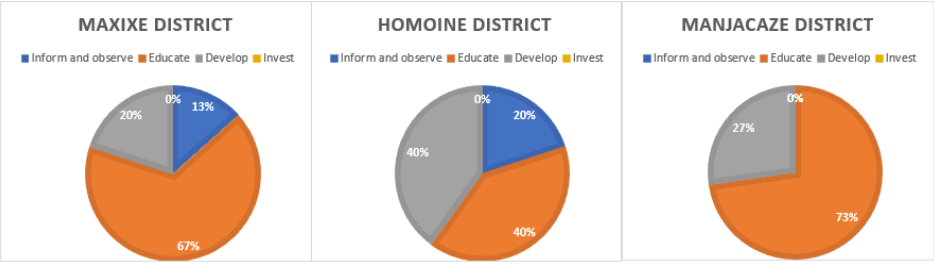


Figure 6.6: Location areas where certain strategies apply

6.3 Chapter summary

In this chapter, the results obtained from the SOM are discussed with regard to the three defined criteria: supply risk, effectiveness of operations and performance improvements. The insight gained from the cluster analysis is used to suggest possible intervention strategies to manage each cluster.

Chapter 7

Conclusion

This thesis considered different clustering techniques and how they can be implemented in segmenting Mozambican cassava suppliers. Section 7.1 summarises the main findings of this study and section 7.2 discusses the identified opportunities for future research.

7.1 Summary

The key purpose of this investigation was to study different clustering algorithms and how they can be applied in supplier relationship management, particularly in supplier segmentation.

In order to achieve this objective, literature discussing different supplier segmentation approaches was studied. Furthermore, various clustering techniques were studied in detail and implemented in a case study of Mozambican cassava suppliers. The techniques considered in this study were the k-means algorithm, agglomerative hierarchical clustering (AHC) and self-organising map (SOM) with Ward clustering. The dataset used was purchasing information from farmers who have supplied cassava to two cassava processing plants in Mozambique.

The CRISP-DM method was used to implement the clustering project. First, the background and objectives of the case study organisation were studied. The dataset was explored by first analysing each feature individually and then analysing the relationships between different features. Data preparation was conducted using methods such as data cleaning, data normalisation and feature selection. The k-means algorithm, AHC, and SOM were implemented and the performance of each technique was evaluated using the inter-cluster distance and intra-cluster distance measures. To ascertain the best possible performance, each technique was evaluated using multiple experimental conditions. The experimental conditions applied considered different initialisation methods and two datasets. In one dataset, outliers were removed using the clamp transformation method and in the other dataset, valid outliers were retained.

An investigation into the clustering techniques' respective experimental conditions was insightful, and the experiments showed how changing one feature can affect the final clustering results. This investigation also involved a comparison between the three techniques. A high inter-cluster distance and low intra-cluster distance indicated that the clusters obtained were of good quality. The SOM with Ward clustering outperformed the k-means algorithm and AHC and was identified as the most suitable algorithm for clustering the cassava suppliers.

The clusters obtained from the SOM were analysed, and the insights gained were used to make recommendations on how clustering can be beneficial for the development of strategies to develop and manage suppliers. The organisation aims to source cassava from many small-

holder farmers. This type of approach, which enables the organisation to divide farmers into distinct groups, is beneficial to the organisation as it enables it to utilise its resources more effectively, while ensuring that the farmers receive appropriate support. It is important to note that, although the SOM obtained the best results at 10 clusters, the cluster analysis showed similarities amongst certain clusters. As a result, only four strategies could be recommended for managing and supporting all ten clusters. The four strategies that were recommended are: inform and observe, educate, develop, and invest strategies. In the inform and observe strategy, the organisation monitors farmers' potential and commitment to growth. The educate and develop strategies aim to upskill farmers and provide them with resources that can enable their sustainable growth. The invest strategy aims to transform the cassava farmers, who mainly consist of subsistence and small-scale family farmers, into commercial farmers.

The method proposed is the first application of clustering to segment cassava suppliers. Unlike the available supplier segmentation methods in literature, the proposed method is the first supplier segmentation method that primarily relies on historical data as input to assess suppliers. Users of the proposed method are provided with the basis of a supplier segmentation system that is more efficient and can be automated. Overall, the SOM was able to identify distinct clusters and provide good insight into farmers' strengths and areas that needed improvement. Furthermore, the algorithms were easy and quick to implement as they did not require users to have in-depth knowledge of farmers or many years of experience with the organisation. However, a number of opportunities for improvement do exist and are listed in section 7.2.

7.2 Future research opportunities

The opportunities for future research identified include the following:

1. The clustering methods applied in this thesis were used to gain insight into farmer capabilities based on the historical purchasing data. There is an opportunity to extend the scope of this project to include yield (quantity and quality) prediction of cassava planted. The opportunities include developing systems that can detect crop patterns and predict the future of the crop, thus highlighting the associated risk and opportunity for stakeholders. This type of study will play an important role in transforming smallholder farmers into commercial farmers as the insight gained from the systems can be used to attract investors to invest in farmers' growth. Moreover, the system can be beneficial when the organisation wants to break new ground, such as building a cassava processing plant in a new area, where historical purchasing information is not available.

2. Another possibility for future studies is to look into remote monitoring systems. The plots of these smallholder farmers are geographically scattered, which makes it difficult for fieldworkers to manage and support farmers effectively. If the field workers can monitor farmers' crops remotely and in real-time, they can plan their farm visits more strategically and efficiently.

7.3 Final words

This thesis has provided a proof of concept for the use of clustering algorithms to enhance the effectiveness of a cassava processing organisation's supplier relationship management efforts through supplier segmentation. Not only can the methods proposed in this thesis improve the organisation's operations, but they can also be used to empower and enhance the livelihood of cassava smallholder farmers in Mozambique.

Bibliography

[1] Khaled A Al-Utaibi and ElSayed M El-Alfy. Intrusion detection taxonomy and data preprocessing mechanisms. *Journal of Intelligent & Fuzzy Systems*, 34(3):1369–1383, 2018.

[2] Mihael Ankerst, Gabi Kastenmüller, Hans-Peter Kriegel, and Thomas Seidl. 3d shape histograms for similarity search and classification in spatial databases. In *International symposium on spatial databases*, pages 207–226. Springer, 1999.

[3] Channing Arndt and Finn Tarp. Agricultural technology, risk, and gender: A cge analysis of mozambique. *World Development*, 28(7):1307–1326, 2000.

[4] David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. techreport 13, Stanford InfoLab, June 2006.

[5] Ana Isabel Rojão Lourenço Azevedo and Manuel Filipe Santos. Kdd, semma and crisp-dm: a parallel overview. Technical report, Informática Comunicações em eventos científicos (ISCAP), 2008.

[6] Fernando Bação, Victor Lobo, and Marco Painho. Self-organizing maps as substitutes for k-means clustering. In *Computer Science*, volume 3516, pages 476–483. International Conference on Computational Science, 2005.

[7] Adam J Brown. *Development of a supplier segmentation method for increased resilience and robustness: A study using agent based modeling and simulation.* PhD thesis, University of Kentucky, 2017.

[8] John A Bullinaria. Self organizing maps: algorithms and applications. Technical Report 17, The University of Birmingham, 2010.

[9] Chinki Chandhok, Soni Chaturvedi, and AA Khurshid. An approach to image segmentation using k-means clustering algorithm. *International Journal of Information Technology (IJIT)*, 1(1):11–17, 2012.

[10] Pete Chapman. The crisp-dm user guide. Technical report, NCR Systems Engineering Copenhagen, 1999.

[11] Ning Chen, Bernardete Ribeiro, Armando Vieira, and An Chen. Clustering and visualization of bankruptcy trajectory using self-organizing map. *Expert Systems with Applications*, 40(1):385–393, 2013.

[12] Katharina Chudzikowski. Career transitions and career success in the 'new'career era. *Journal of Vocational Behavior*, 81(2):298–306, 2012.

[13] Joseph Sarkis Chunguang Bai, Jafar Rezaei. Multicriteria green supplier segmentation. *IEEE Transactions on Engineering Management*, 64(4):515–528, 2017. ISSN doi: 10.1109/TEM.2017.2723639.

[14] Carlos Costa and Christopher Delgado. The cassava value chain in mozambique. Technical report, World Bank, 2019.

[15] Marc Day, Gregory M Magnan, and Morten Munkgaard Moeller. Evaluating the bases of supplier segmentation: A review and taxonomy. *Industrial Marketing Management*, 39(4): 625–639, 2010.

[16] Siddharth Deshpande, Nithya J Gogtay, and Urmila M Thatte. Measures of central tendency and dispersion. *Journal of the Association of Physicians of India*, 64(1):64–66, 2016.

[17] Ayon Dey. Machine learning algorithms: a review. *International Journal of Computer Science and Information Technologies*, 7(3):1174–1179, 2016.

[18] Bert Dijkink and Jan Broeze. Is industrialization in developing countries possible with minimal climate impact? processing cassava in mozambique. *European Journal of Marketing*, 2017.

[19] Cynthia Donovan, Steven Haggblade, Venancio Alexandre Salegua, Constantino Cuambe, João Mudema, and Alda Tomo. Cassava commercialization in mozambique. Technical Report 120, Michigan University, 2011.

[20] Andries P Engelbrecht. *Computational intelligence: an introduction*. John Wiley & Sons, 2007.

[21] AR Fallahpour and AR Moghassem. Evaluating applicability of vikor method of multicriteria decision making for parameters selection problem in rotor spinning. *Fibers and polymers*, 13(6):802–808, 2012.

[22] Salvador Garcia, Sergio Ramirez-Gallego, Julian Luengo, José Manuel Benitez, and Francisco Herrera. Big data preprocessing: methods and prospects. *Big Data Analytics*, 1(1): 9–19, 2016.

[23] Joan Garfield, Robert delMas, and Beth Chance. Using students' informal notions of variability to develop an understanding of formal measures of variability. *Thinking with data*, pages 117–147, 2007.

[24] Jorge Manuel Lourenço Gorricha. Exploratory data analysis using self-organising maps defined in up to three dimensions. Technical report, Teses de Doutoramento, 2015.

[25] Steven Haggblade, Agnes Andersson Djurfeldt, Drinah Banda Nyirenda, Johanna Bergman Lodin, Leon Brimer, Martin Chiona, Maureen Chitundu, Linley Chiwona-Karltun, Constantino Cuambe, and Michael Dolislager. Cassava commercialization in southeastern africa. *Journal of Agribusiness in Developing and Emerging Economies*, 2(1):4–40, 2012.

[26] SK Hahn, NM Mahungu, JA Otoo, MAM Msabaha, NB Lutaladio, and MT Dahniya. *Tropical root crops: Root crops and the African food crisis*. IDRC, 1987.

[27] Sudhanshu Handa and Gilead Mlay. Food consumption patterns, seasonality and market access in mozambique. *Development Southern Africa*, 23(4):541–560, 2006.

[28] MinWei Huang, WeiChao Lin, ChihWen Chen, ShihWen Ke, ChihFong Tsai, and William Eberle. Data preprocessing issues for incomplete medical datasets. *Expert Systems*, 33(5): 432–438, 2016.

[29] Manoj Hudnurkar, Urvashi Rathod, and Suresh Kumar Jakhar. Multi-criteria decision framework for supplier classification in collaborative supply chains: Buyer's perspective. *International Journal of Productivity and Performance Management*, 65(5):622–640, 2016.

[30] Ali Idri, H Benhar, JL Fernández-Alemán, and Ilham Kadi. A systematic map of medical data preprocessing in knowledge discovery. *Computer methods and programs in biomedicine*, 162(1):69–85, 2018.

[31] Anil K Jain. Data clustering: 50 years beyond k-means. *Pattern recognition letters*, 31(8): 651–666, 2010.

[32] Anil K Jain, M Narasimha Murty, and Patrick J Flynn. Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3):264–323, 1999.

[33] Aastha Joshi and Rajneet Kaur. A review: Comparative study of various clustering techniques in data mining. *International Journal of Advanced Research in Computer Science and Software Engineering*, 3(3):1–12, 2013.

[34] John D Kelleher, Brian Mac Namee, and Aoife D'arcy. *Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies*. MIT press, 2015.

[35] Taehyun Kim and Hoon-Young Lee. External validity of market segmentation methods: a study of buyers of prestige cosmetic brands. *European Journal of Marketing*, 2011.

[36] SB Kotsiantis, Dimitris Kanellopoulos, and PE Pintelas. Data preprocessing for supervised leaning. *International Journal of Computer Science*, 1(2):111–117, 2006.

[37] Peter Kraljic. Purchasing must become supply management. *Harvard business review*, 61 (5):109–117, 1983.

[38] Douglas M Lambert. *Supply chain management: processes, partnerships, performance*. Supply Chain Management Inst, 2008.

[39] Deanne Larson and Victor Chang. A review and future direction of agile, business intelligence, analytics and data science. *International Journal of Information Management*, 36 (5):700–710, 2016.

[40] Yanchi Liu, Zhongmou Li, Hui Xiong, Xuedong Gao, and Junjie Wu. Understanding of internal clustering validation measures. In *IEEE International Conference on Data Mining*, volume 1, pages 911–916, 2010.

[41] Yonggang Liu and Robert H Weisberg. *Self Organizing Maps: Applications and Novel Algorithm Design*, chapter 14: A review of self-organizing map applications in meteorology and oceanography, pages 253–264. BoD: Books on Demand, 2011.

[42] Panos Louridas and Christof Ebert. Embedded analytics and statistics for big data. *IEEE Software*, 30(6):33–39, 2013.

[43] S Manikandan. Measures of central tendency: The mean. *Journal of Pharmacology and Pharmacotherapeutics*, 2(2):140–154, 2011.

[44] Óscar Marbán, Gonzalo Mariscal, and Javier Segovia. Data mining and knowledge discovery in real life applications. Technical report, Universidad Europea de Madrid, 2009.

[45] W Natita, W Wiboonsak, and S Dusadee. Appropriate learning rate and neighborhood function of self-organizing map (som) for specific humidity pattern classification over southern thailand. *International Journal of Modeling and Optimization*, 6(1):61–89, 2016.

[46] Hunter H Nielson. The role of cassava in smallholder maize marketing in zambia and mozambique. Master's thesis, Michigan State University, 2009.

[47] Jonathan O'Brien. *Supplier relationship management: Unlocking the hidden value in your supply base*. Kogan Page Publishers, 2018.

[48] Mahamed GH Omran, Andries P Engelbrecht, and Ayed Salman. An overview of clustering methods. *Intelligent Data Analysis*, 6(11):583–605, 2007.

[49] Sevinc Ilhan Omurca. An intelligent supplier evaluation, selection and development system. *Applied Soft Computing*, 13(1):690–697, 2013.

[50] Victor Bernhardsson Oscar Lindgren. Optimal supplier relationship managementa multiple case study of swedish mne within the engineered products industry. Master's thesis, Lund University, 2018.

[51] Jongkyung Park, Kitae Shin, Tai-Woo Chang, and Jinwoo Park. An integrative framework for supplier relationship management. *Industrial Management & Data Systems*, 110(4): 95–515, 2010.

[52] Foster Provost and Tom Fawcett. Data science and its relationship to big data and data-driven decision making. *Big data*, 1(1):51–59, 2013.

[53] Jenny CA Read, Graeme P Phillipson, and Andrew Glennerster. Latitude and longitude vertical disparities. *Journal of Vision*, 9(13):11–11, 2009.

[54] Jafar Rezaei and Hamidreza Fallah Lajimi. Segmenting supplies and suppliers: bringing together the purchasing portfolio matrix and the supplier potential matrix. *International Journal of Logistics Research and Applications*, 22(4):419–436, 2019.

[55] Jafar Rezaei and J Roland Ortt. Two multi-criteria approaches to supplier segmentation. In *IFIP Advances in Information and Communication Technology*, volume 384, pages 317–325, 2011.

[56] Jafar Rezaei and Roland Ortt. A multi-variable approach to supplier segmentation. *International Journal of Production Research*, 50(16):4593–4611, 2012.

[57] Jafar Rezaei and Roland Ortt. Multi-criteria supplier segmentation using a fuzzy preference relations based ahp. *European Journal of Operational Research*, 225(1):75–84, 2013.

[58] Jafar Rezaei and Roland Ortt. Supplier segmentation using fuzzy logic. *Industrial Marketing Management*, 42(4):507–517, 2013.

[59] Jafar Rezaei, Jing Wang, and Lori Tavasszy. Linking supplier development to supplier segmentation using best worst method. *Expert Systems with Applications*, 42(23):9152–9164, 2015.

[60] SalesDept. Dadtco philafrica : Cassava processing, January 2018. URL `https://dadtco-philafrica.com/about-us`.

[61] EM Salvador, Vanessa Steenkamp, and Cheryl Myra Ethelwyn McCrindle. Production, consumption and nutritional value of cassava (manihot esculenta, crantz) in mozambique: An overview. *Journal of Agriculture Biotechnology and Sustainable Development*, 6(3): 29–38, 2014.

[62] Robert G Sargent. Verification, validation and accreditation of simulation models. In *Winter Simulation Conference Proceedings*, volume 1, pages 50–59. IEEE, 2000.

[63] Amit Saxena, Mukesh Prasad, Akshansh Gupta, Neha Bharill, Om Prakash Patel, Aruna Tiwari, Meng Joo Er, Weiping Ding, and ChinTeng Lin. A review of clustering techniques and developments. *Neurocomputing*, 267(1):664–681, 2017.

[64] Marina Segura and Concepción Maroto. A multiple criteria supplier segmentation using outranking and value function methods. *Expert Systems with Applications*, 69(1):87–100, 2017.

[65] Lin Shao, Nelson Silva, Eva Eggeling, and Tobias Schreck. Visual exploration of large scatter plot matrices by pattern recommendation based on eye tracking. In *Proceedings of the 2017 ACM Workshop on Exploratory Search and Interactive Data Analytics*, pages 9–16, 2017.

[66] Ali Seyed Shirkhorshidi, Saeed Aghabozorgi, Teh Ying Wah, and Tutut Herawan. Big data clustering: a review. In *International conference on computational science and its applications*, volume 8583, pages 707–720, 2014.

[67] Jianguo Tang, Kun She, Fan Min, and William Zhu. A matroidal approach to rough set theory. *Theoretical Computer Science*, 471(1):1–11, 2013.

[68] Michael Tuma and Reinhold Decker. Finite mixture models in market segmentation: A review and suggestions for best practices. *Electronic Journal of Business Research Methods*, 11(1):1–13, 2013.

[69] Simon Vänttinen. Supplier segmentation from the perspective of internal knowledgesharing: A case study in the retailing business. Master's thesis, Hanken School of Economics, 2018.

[70] Juha Vesanto. Neural network tool for data mining: Som toolbox. In *Proceedings of symposium on tool environments and development methods for intelligent systems (TOOLMET2000)*, pages 184–196. Citeseer, 2000.

[71] Shouyi Wang, Wanpracha Chaovalitwongse, and Robert Babuska. Machine learning algorithms in bipedal robot control. *IEEE Transactions on Systems*, 42(5):728–743, 2012.

[72] Rüdiger Wirth and Jochen Hipp. Crisp-dm: Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, volume 1, pages 29–39, London, UK, 2000. Springer-Verlag.

[73] David H Wolpert and William G Macready. No free lunch theorems for optimization. *IEEE transactions on evolutionary computation*, 1(1):67–82, 1997.

[74] Rui Xu and Donald C Wunsch. Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, 16(3):645–678, 2005.

[75] Yujie Xu, Wenyu Qu, Zhiyang Li, Geyong Min, Keqiu Li, and Zhaobin Liu. Efficient k-means++ approximation with mapreduce. *IEEE Transactions on Parallel and Distributed Systems*, 25(12):3135–3144, 2014.

[76] Ednah Zvinavashe, H Wolter Elbersen, Maja Slingerland, Sicco Kolijn, and Johan PM Sanders. Cassava for food and energy: exploring potential benefits of processing of cassava into cassava flour and bioenergy at farmstead and community levels in rural mozambique. *Biofuels, Bioproducts and Biorefining*, 5(2):151–164, 2011.