# A combined method to solve the distribution centre location problem in a green-field approach

## Bachelor Thesis


| | |
|---|---|
| Date: | August 23nd 2023 |
| Author: | Erik-Jan Damhof (s2057875) |
| Company Supervisor: | Tom Bode |
| University Supervisor: | Dr. Patricia Rogetzer (1st) |
| | Dr. Mahak Sharma (2nd) |

# Colophon

University of Twente, Industrial Engineering and Management

PO Box 217, 7500 AE Enschede

Tel. +31(0)53489911

Department of Industrial Engineering and Management

Faculty of Behavioural, Management and Social Sciences

Bachelor thesis: A combined method to solve the distribution centre location problem in a greenfield approach.

Erik-Jan Damhof – s2057875

First supervisor: Dr. Patricia Rogetzer

Second supervisor: Dr. Mahak Sharma

Company supervisor: Tom Bode

Publication date: August 29th 2023

2nd edition

# Preface

Dear reader,

Infront of you lies my bachelor thesis report on “A combined method to solve the distribution centre location problem in a green field approach” conducted for ViVochem within the Industrial Engineering and Management program of the University of Twente.

The past six months I have had the pleasure of working on this assignment. Throughout my Industrial Engineering and Management study I have developed a keen interest in supply chain management. I am therefore grateful for the opportunity that ViVochem has offered me through this assignment. The ability to work in a subject area which has my interest has made the past six months a pleasant experience.

I would like to thank my company supervisor, Tom Bode, for guiding me throughout this process and always being prepared to answer questions. I am thankful for the atmosphere that he, Marco and Han created at the office. The pleasant and fun work environment meant I always enjoyed coming to ViVochem to work on my thesis.

Furthermore, I would like to express my gratitude to my supervisors Patricia and Mahak for their great supervision. Their guidance, support and help was invaluable to this thesis. They were always prepared to help me, even if it was last minute. Our meetings were helpful, and the feedback I received helped me significantly improve my thesis.

Finally, I would like to thank my family and friends for their continuous support throughout my thesis and bachelor.

This thesis marks the end of my journey as a student at the University of Twente. Now, I am looking forward to the challenges and opportunities that lie ahead.

Yours sincerely,

Erik-Jan Damhof

Enschede, August 2023

# Management summary

ViVochem is a chemical wholesaler located in Almelo, the Netherlands. In 2022 they partnered with Ferr-Tech, a chemical startup producing a powerful environmentally-friendly oxidiser named FerSol. Its application is in the agrifood, dairy, steel, industrial, oil and gas sectors and for the regional water authorities. Through the partnership, ViVochem took the responsibility to store and transport the chemical to its customers. As of now FerSoL is primarily sold in the BeNeLux, with the biggest customer located in the Netherlands. Ferr-Tech and ViVochem have the ambition to scale up their operations, and start selling their product throughout the entire world. In order to do so efficiently ViVochem needs to enlarge their supply chain network, which currently mostly serving the BeNeLux and German markets.

The current lack of a supply chain network to other regions of the world means that they have to start from zero, a so called greenfield approach. This means that it is necessary to first determine the number of distribution centres required to serve their customers and where to place them. It is vital that this is determined correctly, as a misplacement of distribution centre or placing one to many or little can result in large and unnecessary expenses. The aim of this research is therefore to determine a method which assists ViVochem in making the right distribution centre location decisions.

While conducting a literature review it became apparent that the majority of scientific papers related to this problem either optimised the location of one distribution centre, or of a pre-determined number of centres. These solutions, however, are not applicable to ViVochem's problem. ViVochem does not yet know the exact number of distribution centres they want to establish. Therefore it was decided to develop a new method, based on pre-existing methods commonly used in supply chain management. The combination of which offer a complete solution to the problem ViVochem is currently facing. The methods are both quantitative and qualitative to allow for a more holistic solution to the problem. The need for a combination of quantitative and qualitative methods to determine distribution centre locations was suggested in scientific literature.

Through the literature review three methods were identified which are commonly used to determine distribution centre locations. These are: $K$-means clustering analysis, the centre of gravity of method and the analytic hierarchy process. The first two are both quantitative methods used to first, determine the number of distribution centre required with the cluster analysis. The analysis clusters potential customers based on geographic proximity to one another. These groups of customers are each assigned a distribution centre. Secondly, the results of the cluster analysis are used in the centre of gravity method to determine the optimal location of the distribution centre. In this second step the demand of each customer within the cluster is used to calculate the optimal location of the centre. The location will shift towards the customers with larger demand, or areas within the cluster which have a high density of customers. The results are the coordinates of the optimal locations for distribution centres. However, these coordinates do not take into account geographical locations or any qualitative criteria. Therefore, the final step is the analytic hierarchy process. A shortlist of candidate locations in proximity to the optimal coordinates are compared in the last step. All possible locations are evaluated based on criteria identified through scientific literature. The literature review identified the following criteria: infrastructure, proximity to market, land availability, government support, labour supply and total costs. Each criteria is assigned a weight based on a pairwise comparison by experts within ViVochem. The candidate locations are consequently scored for each criteria. The final output is a ranking of candidate locations based on their score per criteria, and the weight assigned to each criteria.

The analytic hierarchy process used the pairwise comparison of three experts employed at ViVochem to determine the weights assigned to each criteria, these are shown in Table 1.

| Rank | Criteria | Weight |
|---|---|---|
| 1 | Proximity to market | 27% |
| 2 | Infrastructure | 22% |
| 3 | Total costs | 19% |
| 4 | Labour supply | 18% |
| 5 | Land availability | 7% |
| 6 | Government support | 6% |

*Table 1 Final ranking and weight of each criteria.*

The results show a clear priority for ViVochem to place their distribution centres in close proximity to market. These is inline with recommendations found in scientific literature. The method applied in this research also ensures a close proximity to market through the application of the centre of gravity method, with determines optimal coordinates close to areas with high customer density or large demand.

The method was tested for validity and reliability by applying it to six different data sets, each with different parameters. One of these data sets represented potential customers in France and Spain provided by ViVochem. The clustering analysis determined that the customers could best be divided in two clusters. One cluster was populated with French customers, and the other with Spanish customers. This means that both markets are best served by two independent distribution centres. The centre of gravity calculated the optimal coordinates of the centres based on the distribution of demand within the country, it was observed that the optimal location shifted towards the area within the clusters where customer density was highest. As a result the final locations were placed in close proximity to market. Using this information ViVochem can make a short list of candidate locations, located close to the optimal coordinates. Each location is to be scored per criterion, after which a ranking can be determined with the weights resulting from the analytic hierarchy process.

# Table of Contents

# List of Abbreviations

| | |
|---|---|
| ADR | The Agreement concerning the International Carriage of Dangerous Goods by Road |
| AHP | Analytic Hierarchy process |
| B2B | Business to Business |
| BeNeLux | Belgium, the Netherlands and Luxembourg |
| BV | Besloten Vennootschap |
| COG | Centre of Gravity |
| IBC | Intermediate Bulk Container |
| ICIS | Independent Commodity Intelligence Services |
| IEM | Industrial Engineering and Management |
| MCDA | Multi-Criteria Decision Analysis |
| MRQ | Main research question |
| NL | the Netherlands |
| PRP | Plastic Recycling Plant |
| SKU | Stock keeping unit |
| SQ | Sub question |
| 3PL | Third-party logistics |

# 1. Introduction

In chapter one, the background of the research is covered. Section 1.1 briefly introduces the company, followed by the problem description in section 1.2. Next are sections 1.3, 1.4, 1.5 and 1.6, which discuss this research's aim, scope, deliverables and main research question, respectively. Finally, a brief outline of the thesis is given in section 1.7.

## 1.1 Company description

ViVochem is a wholesaler and distributor of chemicals established in 1961. Since 2011, they have been a subsidiary of the BÜFA group, a German company in the chemical trade industry founded in 1883. BÜFA consists of three divisions: chemicals, cleaning and composites. Each operates in its respective business sectors. ViVochem falls under the chemicals division. In 2022, the Independent Commodity Intelligence Services (ICIS) ranked BÜFA group number 19 out of 100 in the European chemical distribution leaders ranking, totalling $552.7 million in sales (Creswell et al., 2023).

ViVochem is located in Almelo, the Netherlands, close to Germany. Their location is ideal for serving the BeNeLux and German markets, although their focus remains the Netherlands. The customers are active in the agricultural, technical, food, cleaning, and personal care industries (ViVoChem, 2023). ViVochem's services include warehousing, drumming and distribution of chemicals. The warehouse located in Almelo has two docking stations and a capacity for 9,000 pallets. Due to the nature of the chemicals, the warehouse is certified according to the Agreement concerning the International Carriage of Dangerous Goods by Road (ADR), meaning additional safety requirements to handle and store the chemicals are in place.

Additionally to warehousing, they offer drumming services, which is the conversion from tankers to 1,000-litre Intermediate Bulk Containers (IBC), 200-litre drums and 20-litre cans. In order to facilitate the distribution, they have a fleet of nine trucks. Their fleet transports part of the chemicals, and part of the transportation is outsourced. Due to ViVochem's operations, their supply chain is classified as "distributor storage with carrier delivery". Figure 1 shows such a general design. The distributors hold inventory in intermediate warehouses and transport the product from the warehouse to the customers. Doing so alleviates the manufacturers' warehousing and transportation problems (Chopra & Meindl, 2016). ViVochem's primary supply chain strategy is based on its high responsiveness to customer demand due to its warehouse storage. It is essential that their customer orders can be delivered as fast as possible if required.

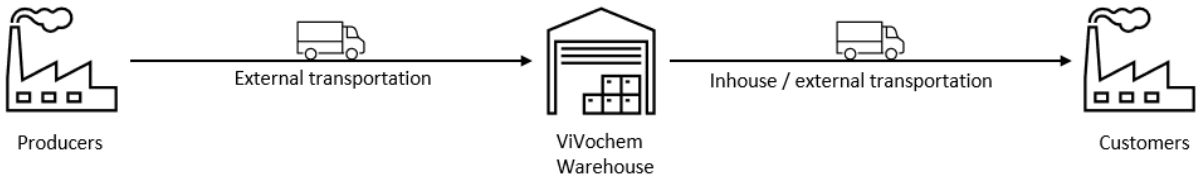


*Figure 1 Distributor storage with carrier delivery*

In 2022, ViVochem partnered with Ferr-Tech, positioning ViVochem as their preferred 3PL partner and sales department (Vivochem and Ferr-Tech Official Partners, 2022). Ferr-Tech is a startup in Meppel, the Netherlands, producing the chemical FerSol. This chemical is an environmentally friendly oxidiser used in wastewater treatment (Ferr-Tech – Revolutionaire Waterzuivering, 2023). Its application is in the agrifood, dairy, steel, industrial, oil and gas sectors and for the regional water authorities. The chemical exists in Potassium Ferrate VI ($K_2FeO_4$) and Sodium Ferrate VI ($Na_2FeO_4$). Both forms serve different uses. Sodium Ferrate VI primary users are plastic recycling plants (PRP), which use FerSol to clean the plastics before recycling. The largest user is a plastic recycling company in the Netherlands.

For the sake of simplicity, this report refers to the Sodium Ferrate VI form each time FerSol is mentioned. The current FerSol supply chain does not differ significantly from ViVochem's overall design. The product is picked up at the production plant in Meppel, stored at Almelo and finally transported to the customer by ViVochem. Ferr-Tech can only store up to six IBCs. Therefore, all produced FerSol is immediately stored at ViVochem.

## 1.2 Problem description

Section 1.2 provides the problem description that ViVochem is currently facing. It starts with the problem context in section 1.2.1, after which the action problem is described in section 1.2.2. Finally, after the problem cluster is discussed in section 1.2.3, the core problem is determined in section 1.2.4.

### 1.2.1 Problem context

In order to arrive at the core problem, it is crucial to understand the context entirely. The commitment ViVochem made by becoming Ferr-Tech's preferred logistics partner means that they are responsible for all FerSol transportation to its customers. Together, they are planning to increase sales rapidly. The current only significant customer is a Dutch plastic recycling company. Most other shipments are test kits delivered to potential customers. This means that the only steady stream of FerSol is to the Netherlands. Since ViVochem primarily serves customers based in Belgium, the Netherlands, Luxembourg (BeNeLux) and Germany, it does not possess a supply corridor to other parts of Europe. As such, there is little known about supply chain opportunities for those parts. The ambition of ViVochem and Ferr-Tech is to sell their product throughout Europe and eventually globally.

The lack of any supply chain network to other regions of the world means that ViVochem has to make use of a greenfield approach. A greenfield approach relies on a completely fresh start, so to speak, to start “on the green field” (*Greenfield- vs. Brownfield-Approach - Definition & Explanation*, 2022). Within supply chain management, a greenfield approach means determining the optimal number of distribution centres and the best location to place them (*Solving Facility Location Problem with Greenfield Analysis,* 2022). According to Chopra and Meindl (2016), supply chain network design consists of four phases: Phase one, define supply chain strategy. Phase two: define regional facility configuration. Phase three, select a set of desirable potential sites. Finally, phase four: location choices. Following these phases, ViVochem is currently situated at phase two. Therefore, developing a new supply chain to other parts of the world starts with determining the number of facilities as well as the location of them.

ViVochem's current problem is that they do not have a tool or methodology to do so. It is essential to do this in a structured and methodological manner since the distribution centre location significantly impacts supply chain network design and performance. A misplaced facility can incur large and unnecessary costs.

### 1.2.2 Action Problem

The problem description described in section 1.2 is used to formulate the action problem. An action problem is a discrepancy between the norm and reality, as perceived by the problem owner (Heerkens & van Winden, 2017). For ViVochem, that is:

*There is no supply chain design to facilitate the increased product flow of FerSol to the rest of Europe.*

By the end of 2025, the turnover is planned to increase by 234% compared to 2023. To properly facilitate this, ViVochem must design a robust supply chain that can scale up in line with the increased demand. Failing to do so could result in unnecessary costs and failure to meet the set-out milestones.

Additionally, it would decrease ViVochem's responsiveness to customer demand and lead to a loss of reputation. The start is to determine the number and location of the distribution centres.

### 1.2.3 Problem Cluster

Figure 2 shows the problem cluster of ViVochem regarding the supply chain design to facilitate the increased flow of FerSol to Europe. The problems are denoted as the action, general, and (potential) core problems. A problem cluster allows more insight into the relationships between different problems to find the core problem (Heerkens & van Winden, 2017).

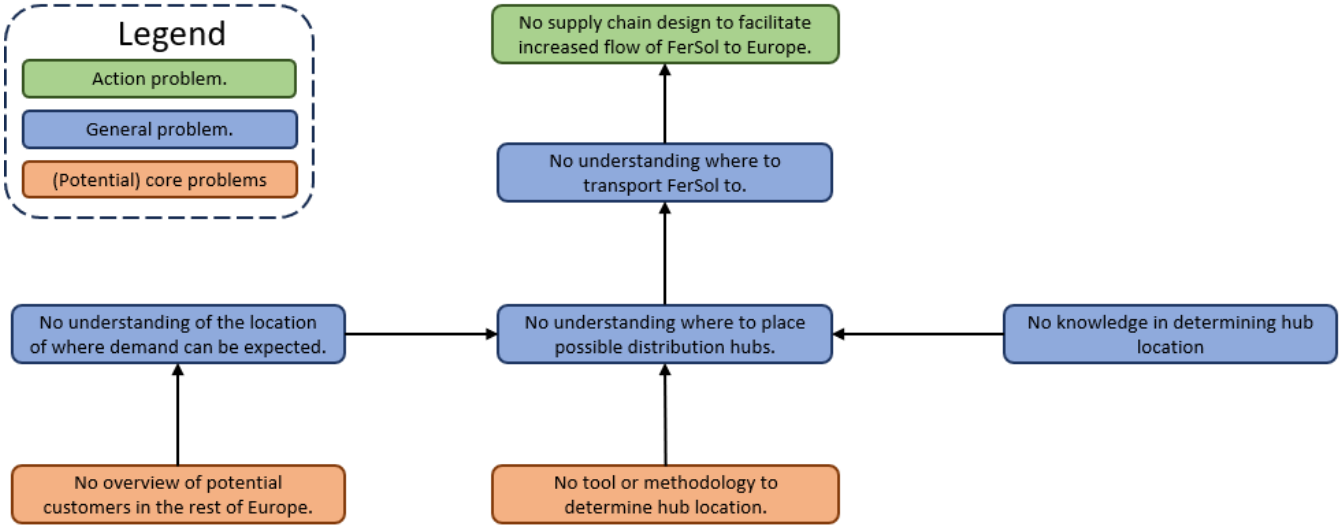


*Figure 2 Problem cluster ViVochem*

The two potential core problems that have been identified are:

- No overview of potential customers in the rest of Europe,
- No tool or methodology to determine the hub location is available.

### 1.2.4 Core problem

To select the main core problem, Heerkens and van Winden (2017) provide four rules of thumb:

- To include a problem in the cluster, one has to be confident that the issue exists and is related to the other problems,
- Follow the problems back to their root causes since problems often are the consequence of other problems,
- When dealing with multiple core problems, choose the one that is the most important,
- Without control over a problem, it cannot be the core problem.

With these four rules of thumb in the back of the mind, the following core problem is chosen:

*There is no tool or method to determine a distribution centre location.*

This choice is made as it is considered the most important problem of the two potential problems. Although the missing overview of potential customers results in a lack of important data, it is also a problem ViVochem can solve relatively easily. They have already started to conduct market research to gather the relevant information. However, the missing tool and methodology to determine a hub location are not easily acquired. It requires specialised and in-depth knowledge of the methods

available. They do not have a structured methodology to make this strategic decision, even though it is crucial to make the right choice. As such, adding such a tool to the toolset of ViVochem and a methodology to choose the optimal distribution hub location can add significant value to their operation.

## 1.3 The aim of this research

This research aims to provide ViVochem with the right tool to determine distribution centre locations to export FerSol to Europe based on academic literature and expert opinion. The tool should assist ViVochem in making this strategic decision as effective as possible.

## 1.4 Scope

So far, the terms supply chain, supply chain network and supply chain design have been used interchangeably. However, according to Chopra and Meindl (2016, p13), a supply chain "consists of all parties involved, directly or indirectly, in fulfilling a customer request". Following this definition of a supply chain, it starts with raw material extraction, ends with the customers, and includes all activities in-between. It is essential to define the role of ViVochem within the supply chain. They are only involved in warehousing products and transporting them to customers. All activities before and after are outside the scope of ViVochem's operations. As such, they cannot control activities outside their operational scope.

The scope of this research falls within the operational activities of ViVochem. This research is not aimed at the transport of raw materials to Ferr-Tech or the transport of FerSol to ViVochem's warehouse. The goal is to give ViVochem a tool to determine a location for a distribution centre of FerSol in other parts of Europe. Optimising the entire supply chain design, optimal route to the distribution centre, planning and improving operational efficiency are considered outside the scope of this bachelor assignment.

## 1.5 Deliverables

This research's primary deliverable is a tool ViVochem can use to determine the location of its new distribution centre(s). The methods used in this tool are based on academic literature and industry knowledge. The tool is delivered in the form of a Microsoft Excel document. Additionally, insights into important location criteria are provided to ViVochem to deepen its understanding and use this knowledge in future endeavours.

## 1.6 Main research question

The problem-solving approach makes use of well-known principles within operations research. These are identified through a literature review and are discussed in Chapter 2. At the end of this research, the following main research question is answered:

*How can ViVochem determine the locations of potential future distribution hubs?*

This question encompasses all necessary aspects of the core problem that ViVochem is facing. By answering this question they will be able to determine the locations of distribution hubs to facilitate their growth.

## 1.7 Outline

The outline of this thesis is as follows: Chapter two discusses the literature review completed in this research. The theoretical framework of this research is discussed, and relevant criteria are identified and defined. Chapter three implements the first step of an integrated approach to determine the distribution centre locations, namely the $K$-means cluster analysis. The implementation as well as results are evaluated on their validity and reliability. Chapter four focusses on the second step of the

approach, the centre of gravity method (COG). Chapter five discusses the final step of the approach, the Multi-Criteria Decision Analysis (MCDA). After which, the overall results, limitations, further research and conclusion are discussed in chapters six, seven, eight and respectively.

# 2. Literature review

This chapter presents the theoretical framework used in this research. It aims to investigate the relevant methods for determining distribution centre locations for ViVochem, which is accomplished through a literature review. The following two research questions are addressed:

1. *What method best determines the optimal distribution centre location?*

2. *What criteria are relevant to evaluate a distribution centre location?*

It is important to define a distribution centre within the context of supply chain management. Langevin and Riopel (2005, pp. 67–99) describe distribution centres as important nodes in a supply network. They preform functions that support the movement of materials. These functions are storing goods, processing products, de-aggregating vehicle loads, creating stock-keeping units (SKU) assortments and assembling shipments. In the case of ViVochem, the primary role of a distribution centre is to store FerSol and, consequently, transport it to the customers. The consensus is that deciding on a location starts with a preselection of a set of potential (candidate) locations (Vieira & Luna, 2016). From this set of candidate locations, the preferred one is chosen based on well-defined criteria. This research aims to follow a similar structure. The first research question aims to find the methods used to determine the coordinates of the optimal location. The second research question explores the criteria for determining the preferred location from a set of candidate locations.

## 2.1 Determining hub locations

The design of a distribution system is a strategic issue that almost every company faces (Klose & Drexl, 2005). The problem of determining distribution centre locations and the number of locations is central to this issue. Due to the familiarity of this challenge within operations research, much research has been conducted addressing this problem. There are multiple methods available, which often fall into one of two categories: multi-criteria models and single-criterion models. As the name suggests, multi-criteria models consider multiple (often conflicting) criteria in determining optimal locations. Whereas single criteria models only consider one criterion, such as costs or distance. Although most publications exclusively use one or the other method, Vieira and Luna (2016) suggest considering aspects of both categories. This results in a two-stage analytical approach.

This research follows the suggestion of Vieira and Luna (2016) and splits the approach into two stages. The first stage consists of the single-criterion models, which conducts a $K$-means cluster analysis and the COG method suggested by Cai et al. (2020). These methods both minimise the distance of a centre's distribution centres to customers to find the optimal coordinate(s). They do not, however, take into account geographical constraints. Consequently, the optimal coordinate(s) could be located in infeasible locations. Therefore, candidate locations are selected based on these optimal coordinate(s). The second stage consists of an MCDA to select the preferred hub location of the set of candidates.

While conducting the literature review, it became evident that the existing body of academic papers on the subject either optimized a single distribution centre location problem or attempted to optimize the locations of a predetermined number of locations. Due to the greenfield approach in this research, it is not yet understood whether or not one distribution centre is sufficient or if there is a need for multiple centres. It was, therefore, decided to formulate a new method based on tried and proven methods already researched. This method integrates the $K$-means clustering, COG and MCDA first to determine the number of distribution centres required based on customer clusters, then determine the optimal location for each cluster, and finally provide a reliable method to compare candidate locations for each cluster using an MCDA.

### 2.1.1 K-Means Cluster Analysis

Cluster analysis is a statistical technique to group similar observations into several clusters (Chain & Arunyanart, 2019). Its application is in data mining, data compression, vector quantisation, pattern classification and pattern recognition (Kanungo et al., 2002). For this research, its primary function is to cluster potential customers based on their proximity to each other. Due to the greenfield approach, it is required to determine the number of distribution centres needed to supply the customers. The customer clusters are assigned to a corresponding distribution centre, following in the footsteps of Cai et al. (2020). The number of clusters identified with the cluster analysis serves as the number of distribution centres required.

In $K$-means clustering, there is a set of $n$ data points in a $d$-dimensional space ($R^d$). The problem is to determine a set of $k$ points in $R^d$ to minimise the squared distance from each data point to its nearest point (Kanungo et al., 2002). A more in-depth explanation is given in chapter three.

For this research, the $n$ data points are the coordinates of potential customers. Since all customers have an $X$ and $Y$ coordinate, the space is two-dimensional ($R^2$). A crucial step in the analysis is determining the $k$ points, which function as the centroids of the $k$ clusters. Kodinariya and Makwana (2013) discuss six approaches to determine the number of clusters, one of which is the Silhouette method. This research uses this method to determine the optimal number of clusters. An in-depth explanation of this is provided in chapter three.

### 2.1.2 Centre of Gravity

The centre of gravity method determines the location of a single distribution centre (Liu & Zhao, 2014). The primary consideration is the distance between data points and the value assigned to each. In the case of this research, the datapoint represents a customer and the value assigned to its corresponding demand. The expected demand functions as a weight to that particular customer. The result is the optimal distribution hub location coordinates based on the weighted coordinates of the customers. The $X$ coordinate is the weighted average of all customer $X$ coordinates. Similarly, the $Y$ coordinate is the average of all customer $Y$ coordinates.

Due to the limitation of only calculating the position of a single distribution centre, it is used in combination with the $K$-means cluster analysis. As suggested by Cai et al. (2020), combining both approaches results in a method which transforms a single-centre location problem into a multiple-centre location problem. Cai et al. (2020) concluded that combining both approaches does not affect the feasibility of the results.

### 2.1.3 Multi-Criteria Decision Analysis

The location of distribution centre is a complex problem, where decisions are affected by context, availability of information and the importance given to the evaluation criteria (Vieira & Luna, 2016). Therefore, decisions must be made based on multiple criteria supported by quantitative and qualitative data. MCDA allows for both types of data to be considered as well as conflicting criteria. Decision makers can then evaluate these to arrive at the preferred location out of the candidate locations. Long and Grasman (2012) confirm the need for qualitative criteria to be considered when determining a distribution hub location. They further emphasise that qualitative and quantitative performance criteria should be considered for a more holistic view.

Multi-criteria models can be solved through multiple approaches. Vieira & Luna (2016) surveyed multiple papers and the methods they employed. The most adopted methods are fuzzy sets and AHP. Overall there is no noticeable difference in output between the two, and their use depends on the researchers choice. In the case of this research AHP is used, due to its relative ease of implementation.

### 2.1.4 Conclusion

The literature review showed a gap between existing scientific literature and the method necessary to provide a complete solution to the greenfield approach in this research. As such there was a need to combine multiple methods into a single one. The $K$- means clustering analysis, COG and AHP methods were therefore chosen to be combined. All three are well regarded and research within the field of supply chain management, as is evident by the large body of scientific research present. The $K$-means clustering analysis provides the number of distribution centres necessary, the COG determines the optimal location of each centre, and the AHP allows for an objective comparison of possible candidate location. Therefore, the first research question has been successfully answered.

## 2.2 Distribution hub criteria

In order to perform an MCDA it is necessary to identify relevant criteria with which to compare candidate locations. The relevant criteria were found using a literature review and are further elaborated in the following sections.

### 2.2.1 Infrastructure

Infrastructure encompasses the availability of roads, railroads, airports and multimodal terminals that provide access to markets (Long & Grasman 2012). Especially the availability of rail service and road infrastructure capacities are important factors affecting the sustainability of any transport systems. Özmen and Aydoğan (2019) further emphasise the existence of intermodal transport infrastructure. The importance of infrastructure is further corroborated by El-Nakib (2010) who identified infrastructure in the form of port, airport and intermodal transport facilities as critical in determining the success of a distribution centre.

### 2.2.2 Proximity to market

Proximity to market represent how close the location is to the demand of the product (Long & Grasman 2012). Market proximity and infrastructure are somewhat related to each other. A large market reach calls for better transportation infrastructure, and better infrastructure increases a region accessibility to its surrounding market. Özmen and Aydoğan (2019) also emphasise the importance of proximity to customers, stating that demand must be met at minimal transportation costs which is achieved by locating as close to customers as possible.

### 2.2.3 Land availability

Long and Grasman (2012) identified land availability as an important consideration to take into account when deciding on distribution centre locations. This is confirmed by Özmen and Aydoğan (2019) as well as El-Nakib (2010). All authors express the important of land availability in the rapidly urbanisation and expansion of cities. Without the available land for warehouses, terminals and other related infrastructure the development opportunities stagnate.

### 2.2.4 Government support

Long and Grasman (2012), Vieira and Luna (2016) and El-Nakib (2012) all identify government and industrial support an important criterion to base a decision on. Support of government is determined to play a big role in accelerating the progression of logistics projects.

### 2.2.5 Labour supply

Without the supply of quality labour to operate machinery and manage the overall freight system the logistics capability would be severely diminished (Long & Grasman 2012). El-Nakib (2012) study of relevant criteria identified the presence of a skilled and qualified labour force as crucial criterion.

### 2.2.6 Total cost

Although not explicitly mentioned by the four papers discussed previously, the total cost is an important aspect to take into consideration. This final criterion was included after experts at ViVochem expressed the need to do so. The total cost refers to the comprehensive assessment of all expenses associated with establishing and operating the distribution centre.

### 2.2.7 Summary

On overview of the criteria mentioned per paper is shown in Table 2. As can be seen in Table 2, most criteria are discussed in multiple papers independently from one another. The fact that multiple scientific papers mention them is a testimony their importance. With the identification of infrastructure proximity to market, land availability, government support and labour supply, the second research question has been answered. In addition, the criteria total costs has been identified through expert opinion within ViVochem.

| | Infra-structure | Proximity to market | Land availability | Government support | Labor supply |
|---|---|---|---|---|---|
| Long & Grasman (2012) | ✓ | ✓ | ✓ | ✓ | ✓ |
| Vieira & Luna (2016) | ✓ | ✓ | | ✓ | |
| El-Nakib (2010) | ✓ | ✓ | ✓ | ✓ | ✓ |
| Özmen & Aydoğan (2019) | ✓ | ✓ | ✓ | | |

*Table 2 Overview criteria per paper.*

## 2.3 Conclusion

The goal of the literature review was to answer two research questions. The first, what method best determine the optimal distribution centre location has been successfully answered. While conducting the literature review it became apparent that there was a need to develop a new method in determining distribution centre locations in a greenfield approach. The need of which arises from the fact that current research focusses on single location problems, or location problems where the number of distribution centres is already known. Both of these scenarios on their own are not applicable for this research. In-order to provide a method that is suitable, the methods $K$- means clustering, COG and AHP are combined.

The second research question, what criteria are relevant to evaluate a distribution centre location, has also been successfully answered. Through a literature review the following criteria were identified: infrastructure, proximity to market, land availability, government support and labour supply. Additionally the total costs was identified by experts within ViVochem as an important criteria as well. These criteria serve as input for the AHP.

The following chapter start with the first step of the method, namely the $K$- means clustering analysis.

# 3. $K$ -means Clustering Analysis

This chapter discusses the implementation of the $K$-means clustering analysis. First, Section 3.1 elaborates on the general algorithm, after which the $K$-mean++ centroid initialization and silhouette method variations are discussed in sections 3.2 and 3.3 respectively. Finally, the implementation of the analysis is examined, and the overall method is evaluated in section 3.4.

## 3.1 General $K$ -means algorithm

The algorithm in its basic form consists of five sequential steps. First, the relevant variables are defined:

$N =$ number of data points in the data set, $N \in \mathbb{N}$
$K =$ number of clusters, $K \in \mathbb{N}$
$X_i =$ data point $i$, $i = 1, 2, \dots, N$
$X_{ix} =$ x coordinate data point $i$
$X_{iy} =$ y coordinate data point $i$
$\mu_j =$ centroid $j$, $j = 1, 2, \dots, K$
$\mu_{jx} =$ x coordinate centroid $j$
$\mu_{jy} =$ y coordinate centroid $j$
$\mu'_j =$ new centroid of cluster
$D_{ij} =$ distance from data point i to centroid $j$
$C(i) =$ cluster to which data point $X_i$ is assigned, $C(i) \in \{1, 2, \dots, K\}$
$N_j =$ number of data points in cluster j, $j = 1, 2, \dots, K$
$P(X_i) =$ probability assigned to $X_i$
$F(X_i) =$ Cumulative probability distribution

These variables are used in the analysis following the next steps:

**Step one:** Initialize centroids.

For $K$ number of clusters, create random centroids denoted as $\mu_1, \mu_2, \dots, \mu_K$

**Step two:** Calculate Euclidean distance from data point $i$ to centroid $j$. Which is calculated by the square root of the sum of the squared distances.

$$D_{ij} = \sqrt{(X_{ix} - \mu_{jx})^2 + (X_{iy} - \mu_{jy})^2}, \quad \text{for all } i, j.$$

**Step three:** Assign all datapoints to a cluster based on the minimum distance to each centroid.

$$C(i) = min(j)\{D_{ij}\}, \quad \text{for all } i, j.$$

**Step four:** Determine the new centroid of each cluster by calculating the mean coordinates of all data points assigned to that cluster.

$$\mu'_{jx} = \frac{\sum X_{ixC(j)}}{N_j}, \quad for\ all\ i, j.$$
$$\mu'_{jy} = \frac{\sum X_{iyC(j)}}{N_j}, \quad for\ all\ i, j. \qquad \mu'_j = (\mu_{jx}, \mu_{jy})$$

**Step five**: Repeat steps two to five until cluster assignment converges and cluster assignment no longer changes.

The $K$ -means cluster analysis is sensitive to the initial placement of the centroids. For that reason, the five steps are repeated for many iterations. Finally, the quality of the cluster assignments for each iteration is evaluated. From this, the best cluster assignments are chosen for each $K$.

In the basic version of the analysis, the centroid initialization occurs randomly. By nature, $K$-means aims to minimise distances between data points and centroids. This is achieved by iteratively updating the centroids and reassigning data points to clusters until convergence is reached. The algorithm's convergence to a solution depends on the initial position of the centroids, if they are poorly initialized, the algorithm may converge to a suboptimal solution (Gul & Rehman, 2023). To solve these problems the cluster initialization in this implementation is achieved through a more sophisticated technique named $K$-means++ initialisation, which is an effective alternative to determine initial centroids (Celebi et al., 2013).

## 3.2 $K$ -Means++ Centroid Initialization

$K$-means++ initialization aims to spread out the initial cluster centroids by randomly selecting a data point and choosing the remaining data points based on a probability proportional to the distance away from a given point's nearest centroid. The effect are centroids spaced as far away from each other as possible, covering as much of the data set as possible. This method follows seven steps, listed below:

**Step one:** Select the first centroid.

Choose the first centroid randomly from the data set.

**Step two:** Calculate Euclidean distance to the nearest existing centroid for each data point.

$$D_{ij} = \sqrt{(X_{ix} - \mu_{jx})^2 + (X_{iy} - \mu_{jy})^2}, \quad \text{for all } i, j.$$

**Step three:** Convert the distances to form probabilities.

In order to convert the distances into probabilities, the normalized probability is calculated. This is achieved by dividing each distance by the sum of all distances.

$$P(X_i) = \frac{D_{ij}}{\sum D_{ij}}, \text{for all } X_i \text{ in the input data.}$$

**Step four:** Determine the cumulative probability distribution $F(X_i)$ for all data points.

**Step five:** Select a new centroid.

A random value is created between zero and one to select the next centroid. After which, the data point $X_i$ is found where $F(X_i)$ is the smallest value such that the $F(X_i)$ is greater than or equal to the random value. Data point $X_i$ is chosen as the next centroid.

**Step six:** Repeat steps two to five until $K$ centroids have been chosen.

**Step seven:** Continue with $K$-mean cluster analysis as usual with initial centroids.

The code used to implement the $K$-means++ initialization is shown in Appendix A.

With the general algorithm and $K$-means++ initialization understood, it is crucial to understand how to optimal number of clusters is determined. This is explained in section 3.3.

## 3.3 Silhouette method

The Silhouette method is an alternative to determine the optimal number of clusters in a given data set. It was first introduced by Rousseeuw (Yuan & Yang, 2019) and combines two factors: cohesion $a(i)$ and separation $b(i)$.

Cohesion measures the similarity of data points to other data points within the same cluster. It is called an intra-cluster metric (Baelung, 2023). In clustering, similarity is the distance between two data points within the same cluster. The cohesion of point $X_i$ in its cluster is the mean distance between $X_i$ and the other data points in the same cluster. Cohesion is defined as:

$$a(i) = mean_{x_j \in C}\big(distance(X_i, X_j)\big), \quad \text{where } i, j = 1,2, \dots, N \text{ and } i \neq j$$

Separation refers to the degree to which the clusters do not overlap. Therefore, it is called an inter-cluster metric (Baelung, 2023). It is the minimum mean distance between $X_i$ and other clusters, defined as:

$$b(i) = \min_{C_2 \neq C_1} \left( mean_{x_j \in C_2} \left( distance\big(X_i, X_j\big) \right) \right), \quad \text{where } i, j = 1,2, \dots, N \text{ and } i \neq j$$

Both cohesion and separation are used to calculate the silhouette width $S(i)$ for data point $X_i$. Silhouette width is defined as:

$$S(i) = \frac{(b(i) - a(i))}{Max(a(i), b(i))}, \qquad -1 \leq S(i) \leq 1$$

$S(i)$ is calculated by subtracting the cohesion from the separation and dividing the result by either the cohesion or separation, which ever one is largest.

The silhouette width indicates how well each data point fits into the cluster they have been assigned. A value close to -1 indicates that the data point is misclassified, and a value of 1 indicates that the data point is well placed within the cluster.

The Silhouette method is utilized at the end of the $K$-means cluster analysis. After the algorithm has run several iterations, the silhouette score is determined per cluster per iteration. Finally, the iteration with the highest score is saved. The code used to implement the Silhouette method is shown in Appendix B.

In order to understand the evaluation of the cluster analysis the understanding some basic terms within the context of cluster analysis are required. The quality of the cluster is linked to the cohesion within the cluster and separation to other clusters. A low cohesion indicates that that the data points within the cluster are spread out and not tightly cluster together, while high cohesion is an indication that the data points are closely packed together. Separation indicates the dissimilarity between data points in one cluster compared to the other neighbouring clusters. Good separation means that different clusters are distinct from one another and there is a clear boundary between them. Poor separation implies that clusters are not well-separated or overlapping.

If the cluster quality is deemed to be high, it means that the datapoints within the clusters are closely packed together and the clusters are well separated. With the Silhouette method this is indicated with a silhouette width of 1. Similarly, low cluster quality implies that the data points within a cluster are spread out, and there is little to no separation between the clusters. This is indicated by a silhouette width of -1. The silhouette width can range anywhere between -1 and 1, a score of zero indicates that data points are on or close to the boundary between two neighbouring clusters.

## 3.4 Implementation

The $K$-means cluster algorithm, as described in sections 3.1 to 3.3, has been implemented in Visual Basic, an overview of the code is in Appendix C. In order to test the reliability and validity of the implementation, the cluster analysis has been applied to six different data sets. These, along with the results, are discussed in this section.

### 3.4.1 Data sets

In order to test the reliability and validity the cluster analysis is applied to six different data sets. Data set one (Appendix D) are potential customers in France and Spain provided by ViVochem, represented by their respective latitude and longitude coordinates in decimal degree form. In order to test the cluster analysis implementation in different scenarios data sets two to six (Appendices E to I ) were generated. Each generated data set contains a different number of clusters, each with different shapes, densities and separation. The sets are constructed so that there are clear visual clusters, which range from three to seven per data set. In order to test the robustness of the code, the data sets consists of varying number of data points ranging from 89 to 200 per set. A graphical visualization of the data sets are found in Appendix J. If the cluster analysis is able to identify all clusters in each data set successfully, it can be concluded that the implementation of the cluster analysis is successful and ready for applications across diverse data sets and a testimony for its robustness.

### 3.4.2 Results

Each data set functioned as input for the cluster analysis. In order to reduce the sensitivity to initial centroid placement, 100 iterations per $K$ value were conducted. Following the Silhouette method, the iterations with the highest scores are saved. The results per data set are discussed in the following sections.

#### *3.4.2.1 Results data set one*

Data set one represents the coordinates of potential customers of ViVochem in France and Spain. In total, it consists of 89 data points. The analysis was conducted for two to five clusters, and Table 3 shows the respective Silhouette scores per cluster.

| **Results test data one** | |
|---|---|
| $K$-values | Silhouette score |
| 2 | 0.229946 |
| 3 | 0.184482 |
| 4 | 0.161284 |
| 5 | 0.201348 |

*Table 3 Silhouette score data set one.*

Table 3 shows that the highest silhouette value is achieved when the data is divided into two clusters, which signifies that the data points were relatively well clustered into two distinct groups. However, as the number of clusters increased, the scores declined. The decrease in scores suggests that the quality of the cluster deteriorated as the number of clusters increased. This means that separating into more clusters led to less distinct separation. Consequently, having two clusters would be the best clustering result based on the silhouette score. The visualization of the two clusters is shown in Figure 3.

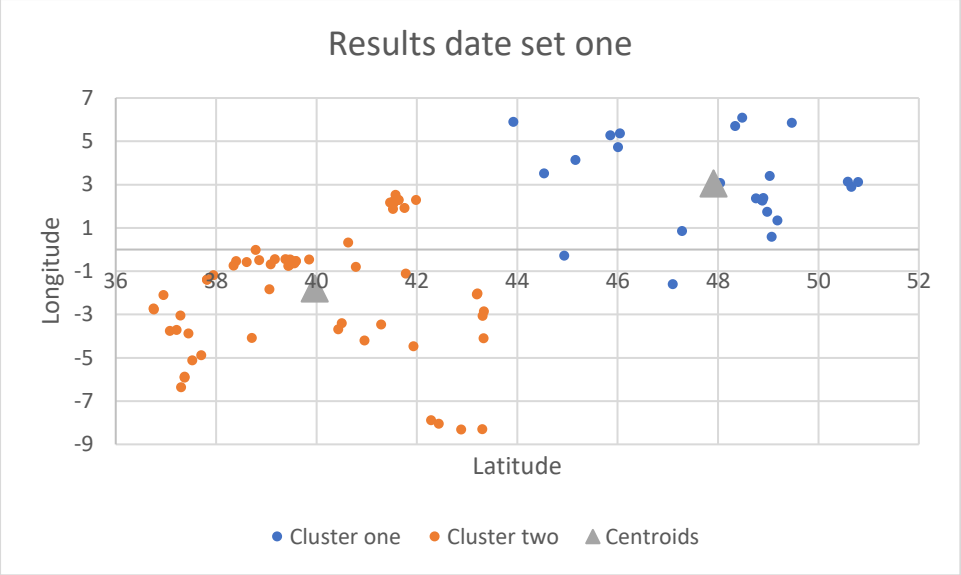


*Figure 3 Cluster results data set one.*

A score close to zero might still indicate effective clustering in particular situations. Separated data in well-defined and well-separated clusters result in higher silhouette scores. When consulting the visualization of data set one (Appendix J), it is concluded that no well-defined and well-separated clusters are present. As such, it can be expected that the resulting scores will not be high.

It is essential to take into account the context of the data. In this case, the data represent potential customers in France and Spain. When further analysing the cluster assignment of each data point, the conclusion is that the clusters represents the French and Spanish markets, respectively. As such, a valid conclusion is that ViVochem needs to consider two distribution hubs, one for each market.

#### *3.4.2.2 Results data set two*

Data set two is a generated test data set and consists of 121 data points. A visualization can be found in Appendix J. The resulting Silhouette scores are shown in Table 4.

| **Results test data two** | |
|---|---|
| $K$-values | Silhouette score |
| 2 | 0.244554 |
| 3 | 0.478959 |
| 4 | 0.358638 |
| 5 | 0.359973 |

*Table 4 Silhouette score data set two.*

Upon examination of the results, the score significantly increased from two to three clusters. After which, it decreases as the number of clusters increases. These scores imply that the clustering effectiveness did not improve beyond the three clusters. Based on the scores, it can be concluded that clustering the test data into three distinct groups produced the most effective partitioning. Each cluster assignment and corresponding centroid are shown in Figure 4.

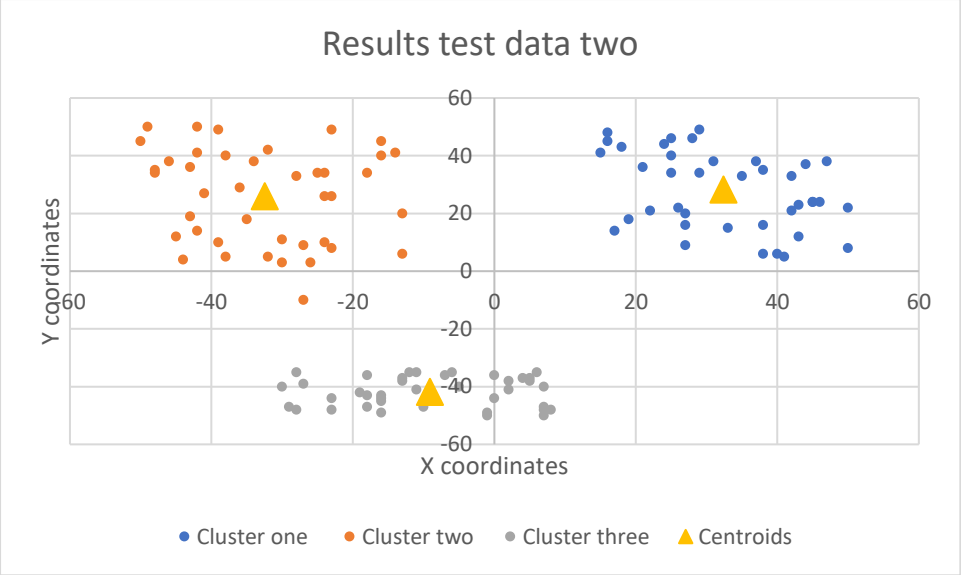


*Figure 4 Cluster result data set two.*

The cluster analysis identified the three clusters in this data set. However, further examining the scores shows a relatively low score of 0.478959 for three clusters. This further indicates that the nature of the data needs to be considered when evaluating the result. In this case, the score could be explained by the varying density within the clusters and the relative spread of intra-cluster data points.

Although such circumstances are essential to consider, the implementation did not fail in successfully identifying the three distinct clusters.

#### *3.4.2.3 Results data set three*

The third data set consisted of 169 data points. When analysing the results shown in Table 5, it is evident that as the number of clusters increased, the Silhouette score generally showed an upward trend until four clusters, where the score reaches its maximum of 0.607783. However, when the number of clusters was increased, the score reflected a significant decrease in cluster effectiveness. Based on the Silhouette score, the most effective clustering solution is achieved with four clusters.

| **Results test data three** | |
|---|---|
| $K$-values | Silhouette score |
| 2 | 0.298634 |
| 3 | 0.550603 |
| 4 | 0.607783 |
| 5 | 0.534843 |

*Table 5 Silhouette score data set three.*

The cluster assignment resulting from the optimal number of clusters is depicted in Figure 5. The Silhouette score achieved with four clusters is the highest so far. The well-defined and well-separated cluster within the data set can explain the high score.

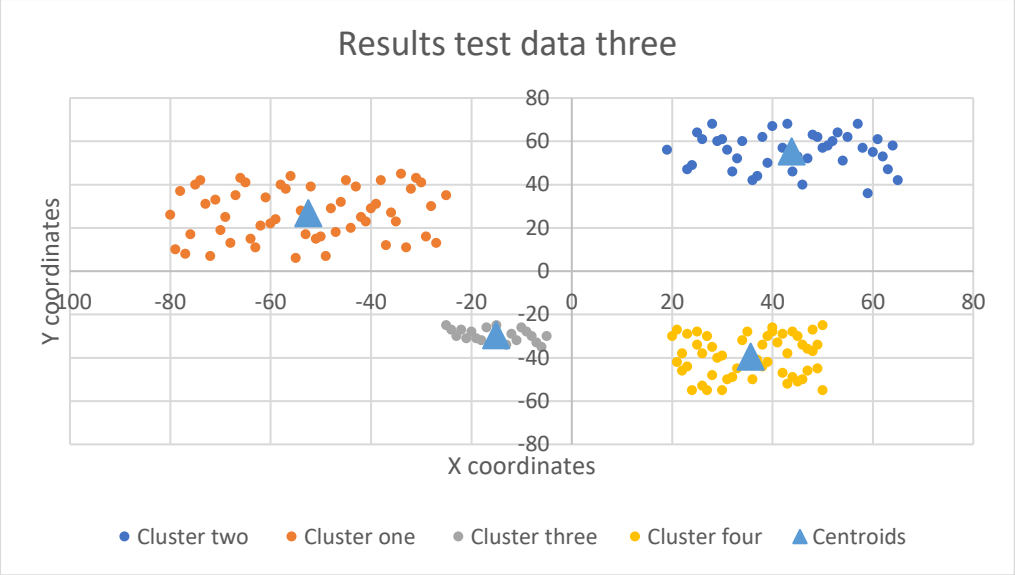


*Figure 5 Cluster result data set three.*

### *3.4.2.4 Results data set four*

The clustering results on data set four for various $K$-values are shown in Table 6. The data set consisted of 190 data points. A general upward trend is observed for increasing cluster numbers. For two up to five clusters, the silhouette score increases, except for three clusters.

| **Results test data four** | |
|---|---|
| $K$-values | Silhouette score |
| 2 | 0.390780958 |
| 3 | 0.373494500 |
| 4 | 0.588832135 |
| 5 | 0.639700659 |
| 6 | 0.574865537 |

*Table 6 Silhouette score data set four.*

All clusters result in good scores, indicating that no matter the $K$-value, clusters have a reasonable level of separation and cohesion. The maximum silhouette score is obtained at five clusters, with a score of 0.639700659. From that point onwards, increasing the number of clusters results in sub-optimal cluster assignments. The final result is depicted in Figure 6.

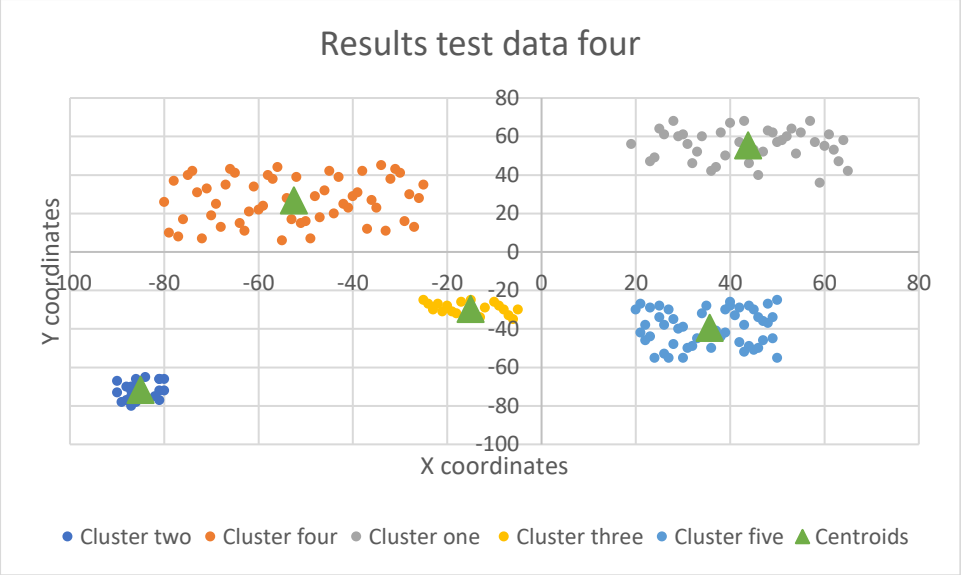


*Figure 6 Cluster result data set four.*

The high silhouette score is attributed to the high level of inherent structure already present in the data set. Which means that there are well defined clusters of data points present withing the data set.

#### *3.4.2.5 Results data set five*

Table 7 shows the results for the second last data set, which consisted of 89 data points. In this experiment, the number of clusters ranged from two to ten to better understand the effect of increased cluster numbers.

| **Results test data five** | |
|---|---|
| $K$-values | Silhouette score |
| 2 | 0.225789 |
| 3 | 0.176330 |
| 4 | 0.371924 |
| 5 | 0.310749 |
| 6 | 0.302215 |
| 7 | 0.247913 |
| 8 | 0.247692 |
| 9 | 0.212191 |
| 10 | 0.172429 |

*Table 7 Silhouette score data set five.*

The silhouette score are relatively low, with the highest being 0.371924 for a $K$ value of four. The optimal cluster assignment is shown in Figure 7.

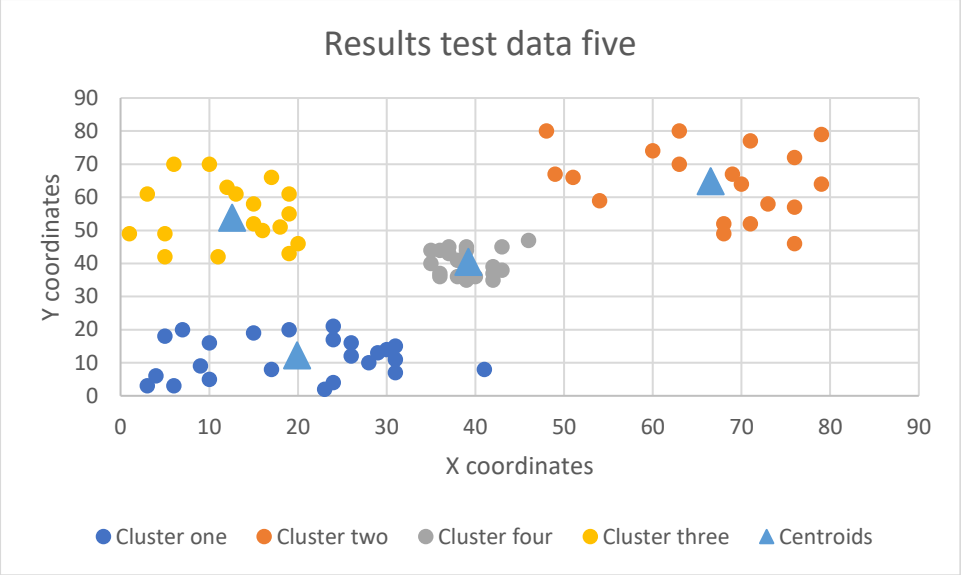


*Figure 7 Cluster result data set five.*

The lower scores can be attributed to the less separated clusters within the data set. It further highlights the impact outliers have on the overall performance of the cluster analysis. However, despite outliers, the results suggest an optimal number of four clusters, which can be confirmed once analysing the data visualization.

#### *3.4.2.6 Results data set six*

The final data set is the most extensive set used in these experiments, consisting of 200 data points. The results of the analysis are depicted in Table 8.

| **Results test data six** | |
|---|---|
| $K$-values | Silhouette score |
| 2 | 0.387132 |
| 3 | 0.192397 |
| 4 | 0.341037 |
| 5 | 0.497962 |
| 6 | 0.481904 |
| 7 | 0.513868 |
| 8 | 0.481649 |
| 9 | 0.466774 |
| 10 | 0.469812 |

*Table 8 Silhouette score data set six.*

The highest silhouette score is realized at $K = 7$ with 0.513868. Notably, starting at five clusters, all scores are within a range of 0.0470094 of the optimum. When consulting the visual representation of the data in Appendix J, it can be seen that it is likely due to the proximity to each other and the fact that some clusters are large enough to be divided further into smaller clusters.

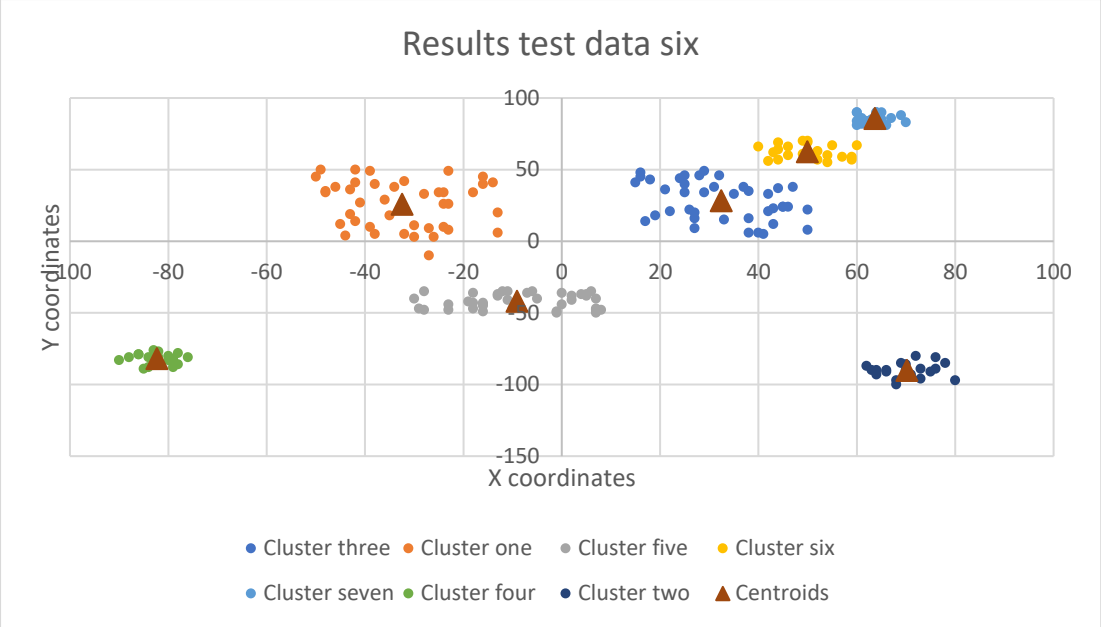


*Figure 8 Cluster result data set six*

Figure 8 shows the cluster assignment for the optimal cluster value of seven. The implementation of the cluster analysis identified the inherent structure of seven clusters.

### 3.4.3 Conclusion

Through analysis of the results of the six different data sets, it can be concluded that the implementation of the $K$-mean cluster analysis in Visual Basic is successful as it could detect all clusters within the different data sets. For data sets two to six, it could identify the inherent structures present in the data. Optimal silhouette scores ranged from 0.37 to 0.64 for data sets two to six. Regardless of the difference between optimal scores, the highest silhouette score was always connected to the number of clusters visible in the data visualization.

The lowest score was obtained for data set number one with 0.23. The low score is explained by the large separation of data points, resulting in low cohesion and separation within and outside the clusters. Regardless of the low silhouette value, the implementation successfully identified the two market segments, Spain and France. From this, it is valid to conclude that ViVochem needs to consider two distribution hubs to supply each market.

The broad range of silhouette scores indicates that critical examination of the input data is required before arriving at a definite solution. A low silhouette score does not necessarily imply bad clustering however needs to be considered within the context of the data. If overall silhouette scores are relatively low, it could indicate that the input data is inherently noisy and unstructured.

The $K$- means cluster analysis is capable of successfully identifying the correct number of cluster within a data set. The clusters identified through the analysis provide in input for the next step, the COG. For each cluster the COG is determined, which is further explained in chapter four.

# 4. Centre of Gravity

This section elaborates on the COG method implemented in Visual Basic. It first discusses the general methodology and mathematical underpinnings, after which the implementation results are examined. The COG follows the $K$-means clustering analysis, and uses the identified clusters as input.

## 4.1 The general method

The COG method is a mathematical method used to determine the optimal location of, for instance, a distribution hub in a supply chain network based on the geographical distribution of demand points. The method aims to minimise the total transportation costs by positioning distribution centres at the so called COG. This COG is calculated using the weighted average of the demand points coordinates, where the weights are represented by the demand values at each location (Liu & Zhao, 2014).

In the case of the tool developed for ViVochem, this method follows after the $K$ - means cluster analysis. The rationale is that after the optimal number of clusters has been identified, the COG allows optimal hub placement to minimise transportation costs within each customer cluster.

In order to accurately convey the mathematical method, the following variables are first defined:

$X_i = \text{x coordinate of data point } i$
$Y_i = \text{y coordinate for data point } i$
$D_i = \text{demand of data point } i$
$X_c = \text{x coordiate centre of gravity}$
$Y_c = \text{y coordinate of centre of gravity}$
$i \in \mathbb{N}$

The COG is calculated as follows:

$$X_c = \frac{\sum_{i=1}^{N}(D_i * X_i)}{\sum_{i=1}^{N} D_i} \quad (1)$$

$$Y_c = \frac{\sum_{i=1}^{N}(D_i * Y_i)}{\sum_{i=1}^{N} D_i} \quad (2)$$

The hub optimal hub is then located at $(X_c, Y_c)$.

Equations (1) and (2) are applied to each cluster. It is essential to understand that only the coordinates and demand of data points assigned to the same cluster are used. For the full implementation of the COG method in Visual Basic, see Appendix K.

It is important to note that the COG method description above assumes that input data is described in Cartesian coordinates. In reality, the data is described in decimal-degree coordinates, where $X$ and $Y$ coordinates are instead the latitude and longitude of a customer location. Although conversion to Cartesian coordinates is possible, it is not necessary. The results remain valid when decimal-degree coordinates are used. Data set one reflects a realistic scenario where the input data is in decimal degree coordinates.

## 4.2 Data sets

The reliability and validity of the COG method was tested using six different data sets. These datasets consisted of coordinates to represent customers, each customer has an assigned value which represents demand. The coordinate data sets remain unchanged compared to the data sets used in the $K$-means clustering analysis and are shown in Appendices G to L . The data set shown in Appendix L are demand estimates of potential customers in France and Spain provided by ViVochem. In order to test the effectiveness of the COG implementation five more demand data sets were generated. These sets followed normal distributions and were created with varying mean and standard deviation parameters. The number of data points ranged from 89 to 200 to match the number of data points used in the test sets for the $K$ -mean cluster analysis. Across different sets, the mean values spanned from 8,000 to 19,000, while standard deviations ranged from 2,000 to 7,500. The generated data sets are shown in Appendices M to Q.

## 4.3 Results

The test data set served as input for the COG method, along with the optimal cluster assignments calculated by the $K$ -means cluster analysis. The coordinates of the COG for each cluster are shown in Appendix R. The results are categorised into two groups. For the first group no notable change is observed between the cluster centroids and the centre of gravity, whereas the second group undergoes a notable change.

The data sets showing a noteworthy change in coordinates are sets one and five, depicted in Figure 1 and Figure 2, respectively. All other results, for the remaining datasets, are visualized in Appendix S.

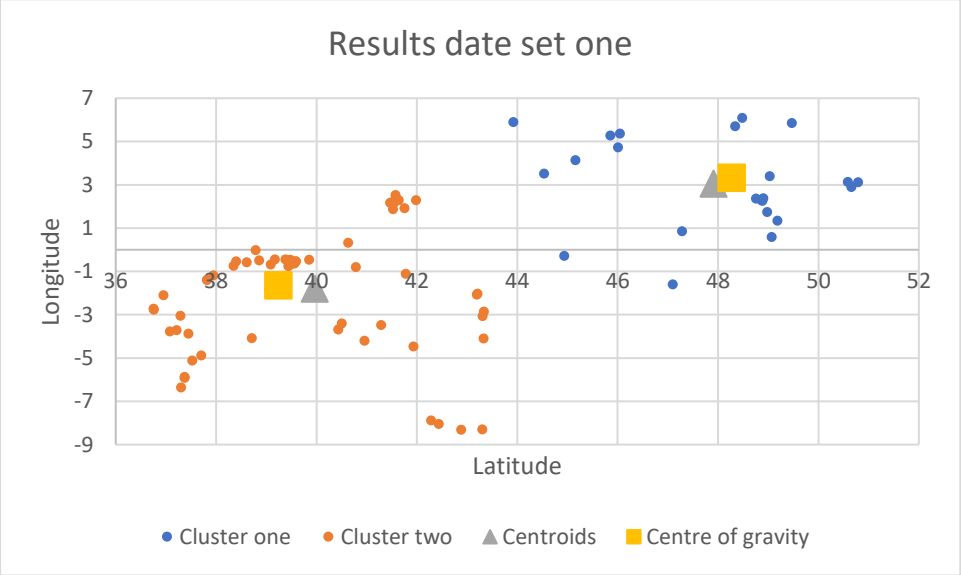


*Figure 9 Centre of gravity results from data set one.*

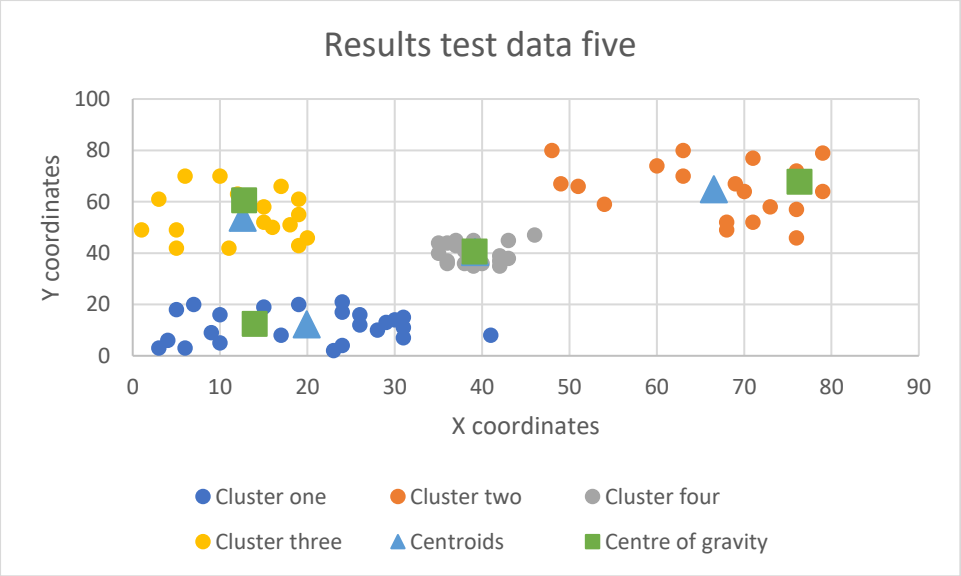


*Figure 10 Centre of gravity results from data set five.*

From the results, it can be concluded that the COG is heavily influenced by the density and symmetry of a cluster, as well as the distribution of demand across the data points. It calculates the average position of data points within a cluster based on their weighted contributions. Therefore, higher demand distribution at a particular side of the cluster or a higher varying density within a cluster can shift the COG. The tendency to shift is fundamental to the method, as higher density or demand results in coordinates placed closer to that region. In Figure 9 the COG moves towards the region within the cluster where customer density is highest. For Figure 10, the shift is the result of a single or a few customers withing the cluster that have considerable larger demand value compared to the other customers.

The results show no notable difference when comparing the centroid position (Appendix T) to the COG of data sets two, three, four, and six (Appendix R). Analysing the visualizations found in Appendix S, it becomes clear that the clusters are symmetrical. In a symmetric cluster, the data points are evenly distributed around the centre, resulting in the average position of the data points being close to the middle of the cluster. At the same time, the clustering algorithm aims to minimize the sum of the distances of data points assigned to a particular cluster. The effect moves the centroids towards the centre of the clusters. The result is that the COG and cluster centroids converge to each other when clusters are symmetrical.

## 4.4 Conclusion

Despite the tendency of the COG and $K$-means cluster centroid to converge for data sets two, three, four and six, the results from data sets one and five show that the COG method is correctly implemented within the context of its objective - calculating a representative point based on the demand distribution within each cluster. It effectively captures the centre of mass, considering the demand data, and yields a single representative point for each cluster. It's crucial to emphasize that the COG method is not a clustering algorithm like $K$ -Means. Instead, it provides a different perspective on identifying representative points for clusters based on demand data. Its simplicity and ability to

capture demand-driven characteristics make it a useful approach in scenarios where demand distribution plays a significant role in cluster analysis.

This chapter implements the second step of the method, the COG method. The COG coordinates represent the optimal location of the distribution centre. The coordinates minimise the total transportation costs based on the demand associated with each customer, however, do not take into account geographical constraints or qualitative assessments of the location. Therefore, once the COG of each cluster is known a list of candidate location has to be composed which can be compared to each other based on the method used in the next chapter.

# 5. Multi-Criteria Decision Analysis

This chapter determines the preferences of ViVochem towards the distribution hub location criteria using an MCDA. First, the AHP methodology is presented, after which the implementation is discussed. Finally, the results are evaluated.

## 5.1 Analytical Hierarchy Process

The AHP is a multi-criteria decision-making method that calculates the criteria weights using a pairwise comparison. The method handles multiple criteria related to decision making with relative ease and successfully manages both quantitative and qualitative data (Sharma & Sehrawat, 2020). It also allows for minor inconsistencies in participant judgement, which is evaluated by the consistency ratio at the end of the process. Winston and Goldberg (2004) break the process down into the following steps:

**Step one**: Define the objective.

**Step two**: Structure the criteria.

**Step three**: Create an $n\ x\ n$ decision matrix based on the number of criteria identified, $n$ is the number of criteria one wishes to compare. With this, a pairwise comparison between criteria is conducted where the entry $a_{ij}$ represents the perceived importance of criteria $i$ over criteria $j$. The importance is measured using a nine-point scale, the interpretation of which is shown in Table 9. For all $i$ , $a_{ii} = 1$ since each criteria is equally important compared to itself.

$$A_{ij} = \begin{pmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \dots & a_{nn} \end{pmatrix}$$

| Value of $a_{ij}$ | Meaning |
|---|---|
| 1 | Criteria $i$ and $j$ are equally important |
| 3 | Criteria $i$ is moderately more important than $j$ |
| 5 | Criteria $i$ is strongly more important than $j$ |
| 7 | Criteria $i$ is very strongly more important than $j$ |
| 9 | Criteria $i$ is extremely more important than $j$ |
| 2,4,6,8 | Intermediate values - for example, 4 is used when criteria $i$ is in-between moderately more important and strongly more important than $j$. |
| $\frac{1}{3}, \frac{1}{5}, \frac{1}{7}, \frac{1}{9}$ | Inverse values, these are applicable if criteria $j$ is perceived as more important than criteria $i$ . The same (inverse) intermediate values apply. |

*Table 9 Interpretation of the nine-point scale used in AHP.*

**Step four:** Normalise the results of the pairwise comparison. Each element is divided by its respective column sum.

$$N_{ij} = \frac{a_{ij}}{\sum_{i=1}^{n} a_{ij}}, \qquad \begin{pmatrix} N_{11} & \cdots & N_{1n} \\ \vdots & \ddots & \vdots \\ N_{n1} & \cdots & N_{nn} \end{pmatrix}, \qquad i, j = 1,2, \dots n$$

**Step five:** Calculate the weight ($W_i$) of the criteria by taking the sum of the matrix normalised rows and dividing it by the number of criteria used.

$$W_i = \frac{\sum_{j=1}^{n} N_{ij}}{n}, \qquad \begin{pmatrix} W_1 \\ \vdots \\ W_n \end{pmatrix}, \qquad i = 1,2, \dots n$$

**Calculate the Consistency ratio.**

The consistency ratio measures how consistent a participant's judgement is throughout the pairwise comparison. By comparing the ratio to a predefined threshold, in the case of this research 10.0%, the consistency of their answer can be judged. If the consistency ratio exceeds this threshold, it indicates a lack of coherence in the decision maker's assessments, necessitating further refinement and adjustment of the pairwise comparisons. Ensuring a low consistency ratio is vital in upholding the integrity of the AHP process and enhancing the credibility of the final decisions. The consistency ratio is computed following the next steps:

**Step one:** Multiply the pairwise comparison matrix with the transpose of the weight vector $(AW^T)$.

**Step two**: Compute $\lambda_{max}$. Which is the average of the sums of the calculated quotients of the i[th] entry in the $AW^T$ and $W^T$.

$$\lambda_{max} = \frac{1}{n}\left(\sum_{i=1}^{n} \frac{\text{ith entry in } AW^T}{\text{ith entry in } W^T}\right)$$

**Step three**: Calculate consistency index $(CI)$. This is achieved by subtracting the number of criteria from the $\lambda_{max}$, and dividing the result by the number of criteria minus one.

$$CI = \frac{\lambda_{max} - n}{n - 1}$$

**Step four**: Compute the consistency ratio $(CR)$. Which is done by calculating the quotient of the consistency ratio and the random index.

$$CR = \frac{CI}{RI}$$

$RI$ is the random index taken from Table 10, where $n$ represents the number of criteria.

| n | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|
| **Random Index** | 0 | 0.58 | 0.90 | 1.12 | 1.24 | 1.32 | 1.41 | 1.45 | 1.51 |

*Table 10 Random Indexes (based on: Winston & Goldberg, 2004).*

Section 5.2 shows the implementation of the steps discussed previously in section 5.1.

## 5.2 Implementation

This section follows the steps laid out previously. After identifying the relevant criteria, three experts working at ViVochem were asked to fill in the pairwise comparison. Their responses are shown in Appendices U to W. An example computation is provided with the pairwise comparison of respondent A:

**Step one:** Define the objective.

The objective is to establish a ranking of distribution hub location criteria.

**Step two**: Structure the criteria.

We identified the criteria through the use of a literature review. The definition of the criteria provided in chapter two are used in the pairwise comparison. The criteria identified were proximity to market, infrastructure, land availability, government support, labour supply and total costs.

**Step three**: Create a $6 \; x \; 6$ decision matrix based on the number of criteria. The pairwise comparison of respondent A is shown in Table 11.

| | Proximity to market | Infrastructure | Land availability | Government support | Labour supply | Total cost |
|---|---|---|---|---|---|---|
| **Proximity to market** | **1** | 1 | 9 | 9 | 1 | 1 |
| **Infrastructure** | 1 | **1** | 9 | 9 | 1 | 1 |
| **Land availability** | $\frac{1}{9}$ | $\frac{1}{9}$ | **1** | 1 | $\frac{1}{9}$ | $\frac{1}{9}$ |
| **Government support** | $\frac{1}{9}$ | $\frac{1}{9}$ | 1 | **1** | $\frac{1}{5}$ | $\frac{1}{5}$ |
| **Labour supply** | 1 | 1 | 9 | 5 | **1** | 1 |
| **Total costs** | 1 | 1 | 9 | 5 | 1 | **1** |
| **Sum** | 4.2222 | 4.2222 | 38 | 30 | 4.3111 | 4.3111 |

*Table 11 Pairwise comparison matrix respondent A.*

**Step four:** Normalise the results of the pairwise comparison. Each element is divided by its respective column sum.

| | Proximity to market | Infrastructure | Land availability | Government support | Labour supply | Total cost |
|---|---|---|---|---|---|---|
| **Proximity to market** | 0,236842 | 0,236842 | 0,236842 | 0,3 | 0,231959 | 0,231959 |
| **Infrastructure** | 0,236842 | 0,236842 | 0,236842 | 0,3 | 0,231959 | 0,231959 |
| **Land availability** | 0,026316 | 0,026316 | 0,026316 | 0,033333 | 0,025773 | 0,025773 |
| **Government support** | 0,026316 | 0,026316 | 0,026316 | 0,033333 | 0,046392 | 0,046392 |
| **Labour supply** | 0,236842 | 0,236842 | 0,236842 | 0,166667 | 0,231959 | 0,231959 |
| **Total costs** | 0,236842 | 0,236842 | 0,236842 | 0,166667 | 0,231959 | 0,231959 |

*Table 12 Normalised results.*

**Step five:** Calculate the weight of the criteria by taking the sum of the normalised rows of the matrix and dividing it by the number of criteria used, in this case six.

| Criteria | Preference vector | Resulting weights |
|---|---|---|
| Proximity to market | 0,245741 | 25% |
| Infrastructure | 0,245741 | 25% |
| Land availability | 0,027305 | 2.7% |
| Government support | 0,033177 | 3.3% |
| Labour supply | 0,223518 | 22% |
| Total costs | 0,223518 | 22% |

*Table 13 Weight per criteria.*

**Calculating the consistency index.**

**Step one:** Calculate $AW^T$.

$$AW^T = \begin{pmatrix} 0.245741 & 0.245741 & 0.245741 & 0.307596 & 0.223518 & 0.223518 \\ 0.245741 & 0.245741 & 0.245741 & 0.307596 & 0.223518 & 0.223518 \\ 0.027305 & 0.027305 & 0.027305 & 0.034177 & 0.024835 & 0.024835 \\ 0.027305 & 0.027305 & 0.027305 & 0.034177 & 0.044704 & 0.044704 \\ 0.245741 & 0.245741 & 0.245741 & 0.170887 & 0.223518 & 0.223518 \\ 0.245741 & 0.245741 & 0.245741 & 0.170887 & 0.22318 & 0.223518 \end{pmatrix}$$

**Step two:** Compute $\lambda_{max}$.

| Criteria | |
|---|---|
| Proximity to market | 6,070852 |
| Infrastructure | 6,070852 |
| Land availability | 6,070852 |
| Government support | 6,012701 |
| Labour supply | 6,062792 |
| Total costs | 6,062792 |
| $\lambda_{max}$ | 6,058473 |

*Table 14 Computation $\lambda_{max}$.*

**Step three**: Calculate consistency index ($CI$).

$$CI = \frac{6.058473 - 6}{5} = 0.011695$$

**Step four**: Compute the consistency ratio ($CR$).

$$CR = \frac{0.011695}{1.24} = 0.09 = 0.9\%$$

The consistency ratio falls within the threshold of 10.0%, therefore it can be concluded that the pairwise comparison performed by respondent A is consistent and coherent.

## 5.3 Results

The methodology described in the previous section was also applied to the pairwise comparisons of respondents B and C. The results are shown in Table 15.

| Criteria | Respondent A | Respondent B | Respondent C |
|---|---|---|---|
| Proximity to market | 25% | 34.9% | 21.4% |
| Infrastructure | 25% | 5.5% | 36.7% |
| Land availability | 2.7% | 16.7% | 2.5% |
| Government support | 3.3% | 9.5% | 6.1% |
| Labour supply | 22% | 12.7% | 18.6% |
| Total costs | 22% | 20.7% | 14.7% |
| **Consistency ratio** | **0.9%** | **9.2%** | **8.0%** |

*Table 15 Weights assigned to criteria and consistency ratio per respondent.*

All consistency ratios fall within the threshold of 10.0%. Therefore, the respondents' judgement in the pairwise comparison is consistent and coherent, and the results can be considered acceptable.

The process aimed to provide a ranking of distribution hub location criteria as perceived by relevant experts within ViVochem. In order to achieve this, the average weight of each criterion across all respondents is computed. The final weights and ranking are shown in Table 16.

| Rank | Criteria | Weight |
|---|---|---|
| 1 | Proximity to market | 27% |
| 2 | Infrastructure | 22% |
| 3 | Total costs | 19% |
| 4 | Labour supply | 18% |
| 5 | Land availability | 7% |
| 6 | Government support | 6% |

*Table 16 Ranking and associated weights of criteria.*

Considering the result, it becomes clear that four of the six criteria carry notably more weight. These are proximity to the market, infrastructure, total costs and labour supply. Together they make up 87% of the weight. Within the top four criteria, proximity to the market scores the highest with 27%, after which the criteria ranked second, infrastructure, scoring 22%. Notably, between the second and fourth ranking criteria, there is only a difference of 4%, indicating that each of these four criteria is of nearly equal importance to ViVochem. The last two, land availability and government support, only score 6% and 7%, respectively. This indicates that ViVochem places nearly no importance on these two criteria.

These results provide ViVochem valuable insights into the importance of distribution centre location criteria. The candidate locations shortlisted after the COG method can be compared using the identified criteria and the weights assigned to them. This is done by scoring each criteria for all locations, and assess them with their associated weights. The result is a ranking of candidate locations, where the location which scores the highest is most desirable for ViVochem.

# 6. Results

In this chapter, the results of the combined approach to determine a distribution hub location are discussed. This studied aimed to combine three methodologies: K-mean clustering, the COG method and AHP. By combining these three techniques, the aim was to create a comprehensive approach that considers both special distribution patterns and multiple decision criteria to guide the selection of the most suitable hub locations.

## 6.1 K-Means clustering

Based on the analysis of the six different data sets, the implementation of K-means cluster analysis in Visual Basic has proven to be successful. For data sets two to six, the algorithm effectively identified underlying structures within the data, as evident from the optimal silhouette scores ranging from 0.37 to 0.64. The highest silhouette score consistently corresponded to the number of clusters visible in the data visualisation.

However, it is important to note that data set number one yielded a lower silhouette score of 0.23. This can be attributed to the absence of clear inherent structure within the provided potential customer data by ViVochem. Nevertheless, the implementation successfully identified two distinct market segments, namely Spain and France. Consequently, it is recommended that ViVochem establishes two distribution hubs to cater to each market.

The varying range of silhouette scores across the datasets emphasises the necessity for a thorough examination of the input data before drawing definitive conclusions. Low scores should be interpreted within the context of the data and silhouette scores for different $K$-values. When overall silhouette scores are relatively low, it may indicate that the input data is noisy and lacks inherent structure.

The reliability of the implementation is satisfactory. The analysis of multiple data sets demonstrates that the $K$-means algorithm successfully identified inherent structures present in most of the data sets (data sets two to six). The consistent relationship between the highest silhouette scores and the visually observed clusters confirms that the algorithm is producing reliable results.

The viability of the implementation is reasonable, but there are some considerations to be made. The optimal silhouette scores for data sets two to six ranged from 0.37 to 0.64, indicating moderate to good clustering quality. However, the lowest silhouette score of 0.23 for data set number one raises concerns about the algorithm's effectiveness in scenarios where the data lacks clear inherent structure. Moreover, the broad range of silhouette scores across different data sets highlights the need for careful examination and critical analysis of the input data before drawing definite conclusions. This indicates that the implementation's viability may be influenced by the quality and nature of the input data.

## 6.2 Centre of gravity

The analysis of the results indicated that the COG is heavily influenced by cluster density, symmetry, and the distribution of demand across data points. It calculates the average position of data points within a cluster, considering their weighted contributions. Consequently, a higher demand distribution on one side of a cluster or varying density within a cluster can cause a shift in the COG. This tendency is inherent in the method, where higher density or demand leads to the coordinate being placed closer to the corresponding region.

The COG method was successful in achieving its objective of calculating a point based on the demand distribution within each cluster. It effectively captures the COG, taking demand data into account and producing a single point for each cluster. Its simplicity and ability to capture demand-driven characteristics make it a valuable approach, particularly in scenarios where demand distribution varies greatly within clusters.

The reliability of the centre of gravity method implementation is satisfactory. The result for each data set were coordinates that shifted towards the concentration of demand. However, the severity of the shift was strongly dependent on the input clusters. Symmetrical clusters showed minimal difference between the cluster centroids and the clusters centre of gravities. Despite the minimal difference observed between the cluster centroids and the clusters' centre of gravities for symmetrical clusters, the implementation can still be deemed reliable.

The validity of the centre of gravity results are acceptable. Despite the lack of difference between centroids and centre of gravities, the centre of gravity coordinates this remain valid. Within symmetrical clusters the most efficient place that minimises distances is in the middle of the cluster. The fact that the centre of gravity coordinates of data sets one and five did significantly shifted towards demand concentration provides proof that the output of the method is valid.

## 6.3 Analytic Hierarchy Process

The objective of this analysis was to determine the ranking of distribution hub location criteria as perceived by experts at ViVochem. To achieve this, AHP was applied to pairwise comparisons of criteria by three respondents.

Upon analysis of the consistency ratios, it was found that the corresponding values of the respondents all fell within the threshold of 10.0%. This result indicated that the respondents' pairwise comparisons were consistent and coherent, making the results acceptable and reliable.

The final ranking from most important to least important, determined by the average of the respondents, was: proximity to the market, infrastructure, total costs, labour supply, land availability and as last, government support. Noteworthy is the difference of importance between the top four, and last two criteria. The top four criteria received a weight between 18% and 27%, whereas the last two 7% and 6% respectively.

The criteria used in the pairwise comparisons were selected based on their relevance to hub location decision-making, as identified by scientific literature. The AHP is a widely recognised and valid technique for measuring relative importance of multiple criteria. By applying this method to multiple experts within ViVochem, along with scientific literature to back up the criteria, the results can be considered valid.

# 7. Limitations

This section discusses the limitations of the combined methods individually. It starts with the $K$-means clustering analysis, COG method, and multi-criteria decision analysis and finally the limitations present in the data sets used.

## 7.1 K-means clustering analysis

The implementation of the $K$-means cluster analysis in Visual Basic was successful, however, there are limitations present. The random seed Excel used to initialise the centroids can influence the resulting cluster. It can lead to potentially different outcomes depending on what seed is used. Moreover, the method becomes computationally expensive for large data sets and increasing K values, demanding significant computational resources and time. This is compounded by the fact that the implementation was done in Excel, which increases computation time further. Increased computation time does not help that the analysis must first compute the results for all $K$- values before deciding the optimal number of clusters. This is relatively inefficient since the optimal number could be realised halfway through the computation. Finally, as became apparent while testing the implementation, outliers in the data significantly impact the final result. The analysis works best on data sets with an inherent structure, which is not always accurate in real life, as evident for data set number one. It was, however, able to cluster the potential customers according to the countries in which they were located. Both clusters represented France and Spain respectively. Subsequent discussion of the result of data set one with company experts confirmed that, indeed, the most efficient approach to serving both markets is to establish distribution centres for each country. This shows that the cluster analysis is a valid method to determine the number of distribution centres. Despite limitations, the result of the $K$-means cluster analysis remains reliable and valid. It is, however, essential to evaluate the results critically based on the data that was provided as input.

## 7.2 Centre of gravity

The centre of gravity method comes with several limitations. The first of which is associated with the distance-based calculation. The method does not consider traffic and transportation networks, instead simplifying distance to straight-line distances between two points. This oversimplification could lead to impractical or inefficient facility locations as it does not consider the complexities of real-life transportation routes and traffic. However, this limitation is balanced with the AHP, allowing for such considerations. Secondly, the method is sensitive to outliers. A single point with extreme demand value can significantly shift the calculated centre, potentially leading to misjudged distribution centre location decisions. Lastly, the methodology only considers transportation costs as the sole factor influencing facility location, disregarding any other crucial cost factors. By neglecting these additional costs, the COG method alone may not provide an accurate solution for optimal facility locations. However, this limitation is also balanced out by the AHP, which considers a more sophisticated cost environment.

## 7.3 Analytic Hierarchy Process

The AHP has several limitations connected to its methodology. One significant limitation is the reliance on the decision-maker's judgement when assessing the pairwise comparisons. Different individuals may have varying opinions, as seen in the three respondents' final results. These differences can potentially lead to inconsistencies in the final results. Moreover, making precise and consistent pairwise comparisons can be challenging, which became apparent throughout this study. While conducting the AHP, multiple experts had to reevaluate their pairwise comparison due to a consistency ratio outside the threshold. Additionally, many factors could influence the respondent's pairwise comparison. There are significant chances that, when asked to participate again in several months, respondents will

provide different answers than they do now. Finally, AHP heavily depends on the expertise and knowledge of the decision-makers, as their understanding of the problem and criteria directly influences the results. It is, therefore, vital to establish a common understanding of the problem and criteria among respondents. In this research, the limitations of the AHP were mitigated by combining the opinion of multiple experts. The result is a ranking of criteria not based on the opinion and possible bias of a single respondent, increasing the reliability and viability of this method.

### 7.4 Limitations in test data sets

When discussing the initial data set of potential customers in Spain and France with company experts, it became clear that it was inadequate and did not accurately reflect reality. As a result, the implementation of the $K$-mean cluster analysis and COG method had to be evaluated using generated test data. The use of this test data came with several limitations to this research. When researchers create test data, they might introduce bias and subjectivity unintentionally, leading to unrealistic patterns that do not represent real-world scenarios. Additionally, a researcher's control over data creation may result in overfitting, where the model performs well on the constructed data but is inadequate for other data. To address these limitations, the research used one data set that best-reflected real-life data which was provided by ViVochem. This made sure that the methodology was not just evaluated on generated data but also realistic data. The generated data served to test the effectiveness of the $K$-means cluster analysis and COG method. If the cluster analysis was able to identify all clusters visible in the test data, it could be concluded that the was successful and the code was working. The same reasoning is used for the COG method, if the implementation is able to successfully determine the COG of the test data, it can be concluded that it was successful. In order to test robustness of the implementations, each data set consisted of a varying number of data points to ensure they hold up in different scenarios.

## 8. Further research

Future research can focus on several key aspects to enhance decision-making and sustainability. Firstly, improvements to cluster analysis techniques can consider the cost trade-off between additional investment costs and transportation costs associated with adding another cluster and opening an additional distribution centre. Incorporating cost factors and transportation efficiency metrics can lead to more optimised and practical cluster configurations, ensuring the efficient allocation of resources and minimising operational costs.

Secondly, conducting sensitivity analysis for AHP can provide valuable insights into the robustness of decision outcomes. Evaluating the impact of varying criteria weights and pairwise comparisons can help decision-makers understand the stability of their choices and identify critical factors that influence the final decisions.

Lastly, given the growing concerns about climate change and its potential disruptions to supply chains, researchers should explore incorporating environmental considerations. This may involve assessing the vulnerability of supply chain networks to extreme weather events, developing climate-resilient logistics strategies, and promoting eco-friendly practices to ensure supply chain sustainability and adaptability in the face of changing environmental conditions. By addressing these research areas, supply chain management can become more efficient, resilient, and environmentally conscious.

# 9. Conclusion

This section concludes the research by discussing the theoretical and practical implications of the methodology proposed in this thesis.

## 9.1 Theoretical implications

From a scientific perspective, this research aims to contribute to existing knowledge regarding the optimal distribution centre location problem. While conducting a literature review, many papers were identified that used either a quantitative approach (linear programming, COG, et cetera) or a qualitative approach (MCDA). Cai et al. (2020) presented the only paper combining both approaches. Vieira and Luna (2016) further emphasised the need for integrated approaches to consider aspects of both categories. The methodology proposed in this thesis uses well-known machine learning, facility location problem and decision-making concepts to provide a well-rounded approach to determining the optimal distribution hub location. It is best suited for a greenfield approach, where a decision-maker can determine the optimal distribution location. The $K$-mean cluster analysis offers insight into customer clusters which serve as an indication of the number of distribution centres needed, after which the COG method minimises the average distance to each customer within a cluster. Finally, to systematically choose a location based on multiple criteria, the AHP provides a solution. The contribution of this research is the proposal of a new method, which combines multiple well-known methods in supply chain management, however, have never been combined before. It offers a starting point for further research and optimisation in the combination of these methods.

## 9.2 Practical implications

The practical implications of this research are primarily for ViVochem decision-makers and top management. By combining the three different approaches in a tool, they can easily, systematically and accurately decide upon the number of distribution centres, and their respective locations. Furthermore, the tool is not restricted to a single geographical location. It can be used on any scale, for any part of the world. This supports ViVochem in further expansion in the future and provides them with a starting point from which they can design their supply chain network in a greenfield approach. Setting up a distribution centre can require large upfront investments, and selecting a wrong location can be costly. The method suggested in this research reduced the risk and enables ViVochem to make well considered decisions based on scientific literature and expert opinion.

A well placed distribution centre will contribute to the cost efficiency of ViVochem by optimising the proximity to its customers. Additionally, they will be able to uphold their responsive supply chain strategy by providing quick deliveries of FerSol, regardless of the location of the customer. Which allows them to maintain their competitive advantage.

# Bibliography

Baelung. (2023, June 13). Silhouette plots. Baeldung. https://www.baeldung.com/cs/silhouette-values-clustering

Cai, C., Luo, Y., Cui, Y., & Chen, F. (2020). Solving Multiple Distribution Center Location Allocation Problem Using K-Means Algorithm and Center of Gravity Method Take Jinjiang District of Chengdu as an example. IOP Conference Series: Earth and Environmental Science, 587(1), 012120. https://doi.org/10.1088/1755-1315/587/1/012120

Celebi, M. E., Kingravi, H. A., & Vela, P. A. (2013). A comparative study of efficient initialisation methods for the k-means clustering algorithm. Expert Systems with Applications, 40(1), 200–210. https://doi.org/10.1016/j.eswa.2012.07.021

Chain, P., & Arunyanart, S. (2019). Using cluster analysis for location decision problem. Proceedings Journal of Physics, 673, 012086–012086. https://doi.org/10.1088/1757-899x/673/1/012086

Chopra, S., & Meindl, P. (2016). Supply chain management: strategy, planning, and operation (6th ed.). Pearson.

Creswell, S., Chang, J., & Becham, W. (2023, May 25). ICIS Top 100 Chemical Distributors. Special Report Top 100 Chemical Distributors, 100 - 101.

El-Nakib, I. (2010, October 8). Location preference of Egyptian firms for logistics hub in Southeast Africa. International Conference on Supply Chain Management and Information Systems: Logistics Systems and Engineering. International Conference on Supply Chain Management and Information Systems: Logistics Systems and Engineering, Alexandria, Egypt. https://l8.nu/qtVl

Langevin, A., & Riopel, D. (2005). Logistics Systems: Design and Optimization (pp. 67–99). Springer Science & Business Media.

Greenfield- vs. Brownfield-Approach - Definition & Explanation. (2022). Dokumentenmanagement Software | EASY SOFTWARE AG. https://easy-software.com/en/glossar/greenfield-vs-brownfield-approach/

Gul, M., & Rehman, A. (2023). Big data: an optimised approach for cluster initialisation. Journal of Big Data. https://doi.org/10.1186/s40537-023-00798-1

Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R., & Wu, A. Y. (2002). An efficient k-means clustering algorithm: analysis and implementation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(7), 881–892. https://doi.org/10.1109/tpami.2002.1017616

Klose, A., & Drexl, A. (2005). Facility location models for distribution system design. European Journal of Operational Research, 162(1), 4–29. https://doi.org/10.1016/j.ejor.2003.10.031

Kodinariya, T., & Makwana, P. (2013). Review on Determining of Cluster in K-means Clustering. International Journal of Advance Research in Computer Science and Management Studies, 1(6), 90–97. https://www.researchgate.net/publication/313554124_Review_on_Determining_of_Cluster_in_K-means_Clustering

Langevin, A., & Riopel, D. (2005). Logistics Systems: Design and Optimization (pp. 67–99). Springer Science & Business Media.

Liu, X., & Zhao, X. (2014). Based on Gravity Method of Logistics Distribution Center Location Strategy Research. Semantic Scholar. https://doi.org/10.2991/lemcs-14.2014.134

Long, S., & Grasman, S. E. (2012). A strategic decision model for evaluating inland freight hub locations. Research in Transportation Business & Management, 5, 92–98. https://doi.org/10.1016/j.rtbm.2012.11.004

Sharma, M., & Sehrawat, R. (2020). A hybrid multi-criteria decision-making method for cloud adoption: Evidence from the healthcare sector. Technology in Society, 61, 101258. https://doi.org/10.1016/j.techsoc.2020.101258

Solving Facility Location Problem with Greenfield Analysis. (2022). Www.anylogistix.com. https://www.anylogistix.com/features/solving-facility-location-problem-with-greenfield-analysis/

ViVoChem. (2023). Your B2B partner in chemical distribution | ViVoChem. Vivochem.com. https://www.vivochem.com/

Vieira, C. L. dos S., & Luna, M. M. M. (2016). MODELS AND METHODS FOR LOGISTICS HUB LOCATION: A REVIEW TOWARDS TRANSPORTATION NETWORKS DESIGN. Pesquisa Operacional, 36(2), 375–397. https://doi.org/10.1590/0101-7438.2016.036.02.0375

Vivochem and Ferr-Tech official partners – Ferr-Tech. (2022, January 31). Ferr-Tech. https://ferr-tech.com/vivochem-and-ferr-tech-official-partners/

Winston, W. L., & Goldberg, J. B. (2004). Operations Research: Application and Algorithms. Brooks/Cole.

Yuan, C., & Yang, H. (2019). Research on K-Value Selection Method of K-Means Clustering Algorithm. J, 2(2), 226–235. https://doi.org/10.3390/j2020016

Özmen, M., & Aydoğan, E. K. (2019). Robust multi-criteria decision making methodology for real life logistics center location problem. Artificial Intelligence Review. https://doi.org/10.1007/s10462-019-09763-y

# 10. Appendences

## Appendix A: Visual basic implementation of the $K$-means++ initialisation

```
Function InitializeCentroidsKMeansPlusPlus(ByVal inputData As Variant, ByRef centroids() As Double, ByVal clusterCount As Integer)
    Dim dataLength As Integer
    Dim randomIndex As Integer
    Dim i As Integer, j As Integer
    Dim minDistance As Double, totalDistance As Double, dist As Double

    dataLength = UBound(inputData, 1)

    ' Select the first centroid randomly from the data points
    Randomize
    randomIndex = Int((dataLength - 1 + 1) * Rnd + 1)
    centroids(1, 1) = inputData(randomIndex, 1)
    centroids(1, 2) = inputData(randomIndex, 2)

    ' Select the remaining centroids using K-means++ strategy
    For i = 2 To clusterCount
        totalDistance = 0

        ' Calculate the distance to the nearest centroid for each data point
        Dim distances() As Double
        ReDim distances(1 To dataLength) As Double

        For j = 1 To dataLength
            minDistance = distance(CDbl(inputData(j, 1)), CDbl(inputData(j, 2)), CDbl(centroids(1, 1)), CDbl(centroids(1, 2)))

            For k = 2 To i - 1
                dist = distance(inputData(j, 1), inputData(j, 2), centroids(k, 1), centroids(k, 2))
                minDistance = WorksheetFunction.Min(minDistance, dist)
            Next k

            distances(j) = minDistance
            totalDistance = totalDistance + minDistance
        Next j

        ' Choose the next centroid with probability proportional to distance squared
        Dim cumulativeProbabilities() As Double
        ReDim cumulativeProbabilities(1 To dataLength) As Double

        cumulativeProbabilities(1) = distances(1) / totalDistance
        For j = 2 To dataLength

            cumulativeProbabilities(j) = cumulativeProbabilities(j - 1) + distances(j) / totalDistance
        Next j

        Dim randomValue As Double
        Randomize
        randomValue = Rnd()
        For j = 1 To dataLength
            If randomValue <= cumulativeProbabilities(j) Then
                centroids(i, 1) = inputData(j, 1)
                centroids(i, 2) = inputData(j, 2)
                Exit For
            End If
        Next j
    Next i
End Function
```

## Appendix B: Visual basic implementation of the silhouette score

```
Function CalculateSilhouetteScore(ByVal inputData As Variant, ByVal clusterAssignments As Variant) As Double
    Dim dataLength As Integer
    dataLength = UBound(inputData, 1)

    Dim silhouetteSum As Double
    Dim dataPointCountInCluster() As Integer
    ReDim dataPointCountInCluster(1 To UBound(inputData, 1))

    ' Calculate the number of data points in each cluster
    For i = 1 To dataLength
        dataPointCountInCluster(clusterAssignments(i, 1)) = dataPointCountInCluster(clusterAssignments(i, 1)) + 1
    Next i

    ' Calculate the silhouette score for each data point
    For i = 1 To dataLength
        Dim clusterIndex As Integer
        clusterIndex = clusterAssignments(i, 1)

        Dim a As Double ' Average distance to points in the same cluster
        Dim b As Double ' Average distance to points in the nearest other cluster
        Dim minB As Double
        minB = 100000

        For j = 1 To dataLength
            If i <> j Then
                If clusterAssignments(j, 1) = clusterIndex Then
                    a = a + distance(CDbl(inputData(i, 1)), CDbl(inputData(i, 2)), CDbl(inputData(j, 1)), CDbl(inputData(j, 2)))
                ElseIf dataPointCountInCluster(clusterAssignments(j, 1)) > 0 Then
                    ' Calculate distance to points in the nearest other cluster and find the minimum
                    Dim currB As Double
                    currB = distance(CDbl(inputData(i, 1)), CDbl(inputData(i, 2)), CDbl(inputData(j, 1)), CDbl(inputData(j, 2)))
                    minB = WorksheetFunction.Min(currB, minB)
                End If
            End If
        Next j

        If dataPointCountInCluster(clusterIndex) > 1 Then
            a = a / (dataPointCountInCluster(clusterIndex) - 1)
        End If

        ' Update b with the minimum distance to points in the nearest other cluster
        b = minB

        silhouetteSum = silhouetteSum + (b - a) / WorksheetFunction.Max(a, b)
    Next i

    ' Calculate the average silhouette score for all data points
    Dim silhouetteScore As Double
    silhouetteScore = silhouetteSum / dataLength

    CalculateSilhouetteScore = silhouetteScore
End Function
```

## Appendix C: K-mean cluster analysis implementation in VBA

```
Sub KMeansClusteringWithSilhouetteScore()
    Dim dataRange As Range
    Dim inputData() As Variant
    Dim kValues As Variant
    Dim i As Integer, j As Integer, iter As Integer
    Dim clusterCount As Integer, minSilhouetteScore As Double
    Dim bestClusterAssignments() As Variant, bestCentroids() As Double
    Dim centroidCount As Integer
    Dim sumDistances() As Double, variance() As Double, minVariance As Double
    Dim ws As Worksheet
    Dim outputSheet As Worksheet
    Dim centroidsSheet As Worksheet
    Dim silhouetteSheet As Worksheet
    Dim outputRow As Integer
    Dim outputCol As Integer

    ' Set the input data range dynamically
    Dim lastRow As Long
    lastRow = ThisWorkbook.Sheets("InputData").Cells(Rows.Count, "A").End(xlUp).Row
    Set dataRange = ThisWorkbook.Sheets("InputData").Range("A2:B" & lastRow)

    ' Read the input data into an array
    inputData = dataRange.Value

    ' Define the k-values
    kValues = Array(2, 3, 4, 5)

    ' Create a new sheet for the results
    Set outputSheet = ThisWorkbook.Sheets.Add(After:=ThisWorkbook.Sheets(ThisWorkbook.Sheets.Count))
    outputSheet.Name = "KMeans Clustering Results"

    ' Write the headers for the results
    outputSheet.Range("A1").Value = "X Coordinate"
    outputSheet.Range("B1").Value = "Y Coordinate"

    ' Create a new sheet for the centroids
    Set centroidsSheet = ThisWorkbook.Sheets.Add(After:=ThisWorkbook.Sheets(ThisWorkbook.Sheets.Count))
    centroidsSheet.Name = "Centroids"
```

```
' Write the headers for the centroids
centroidsSheet.Range("A1").Value = "K Value"
centroidsSheet.Range("B1").Value = "Centroid Number"
centroidsSheet.Range("C1").Value = "Centroid X Coordinate"
centroidsSheet.Range("D1").Value = "Centroid Y Coordinate"

' Create a new sheet for the silhouette scores
Set silhouetteSheet = ThisWorkbook.Sheets.Add(After:=ThisWorkbook.Sheets(ThisWorkbook.Sheets.Count))
silhouetteSheet.Name = "Silhouette Scores"

' Write the headers for the silhouette scores
silhouetteSheet.Range("A1").Value = "K Value"
silhouetteSheet.Range("B1").Value = "Silhouette Score"

' Initialize the output row counter for results
outputRow = 2

' Initialize the output row counter for centroids
Dim centroidOutputRow As Integer
centroidOutputRow = 2

' Loop through each k-value
For i = LBound(kValues) To UBound(kValues)
    clusterCount = kValues(i)
    minSilhouetteScore = -1
    ReDim bestClusterAssignments(1 To UBound(inputData, 1), 1 To 1)
    ReDim bestCentroids(1 To clusterCount, 1 To 2) As Double

    ' Loop 100 times to find the best centroids
    For iter = 1 To 100
        ' Initialize centroids using K-means++ strategy
        centroidCount = clusterCount
        Dim centroids() As Double
        ReDim centroids(1 To centroidCount, 1 To 2) As Double
        InitializeCentroidsKMeansPlusPlus inputData, centroids, clusterCount
```

```
' Assign data points to clusters based on initial centroids
Dim clusterAssignments() As Variant
clusterAssignments = AssignDataPoints(inputData, centroids)

Dim converged As Boolean
Dim iteration As Integer
Dim maxIterations As Integer
converged = False
iteration = 1
maxIterations = 100

' Iterate until convergence or maximum iterations reached
Do While Not converged And iteration <= maxIterations
    ' Update centroids
    Dim newCentroids() As Double
    newCentroids = UpdateCentroids(inputData, clusterAssignments, clusterCount)

    ' Check convergence
    converged = CheckConvergence(centroids, newCentroids)

    ' Copy newCentroids values to centroids
    For j = 1 To clusterCount
        centroids(j, 1) = newCentroids(j, 1)
        centroids(j, 2) = newCentroids(j, 2)
    Next j

    ' Reassign data points to clusters based on updated centroids
    clusterAssignments = AssignDataPoints(inputData, centroids)

    iteration = iteration + 1
Loop

' Calculate sum of distances between centroids and data points
sumDistances = CalculateSumDistances(inputData, clusterAssignments, centroids)

' Calculate the silhouette score for the current clustering
Dim silhouetteScore As Double
silhouetteScore = CalculateSilhouetteScore(inputData, clusterAssignments)
```

```
            ' Select the clustering with the highest silhouette score
            If minSilhouetteScore = -1 Or silhouetteScore > minSilhouetteScore Then
                minSilhouetteScore = silhouetteScore
                bestClusterAssignments = clusterAssignments
                bestCentroids = centroids
            End If
        Next iter

        ' Write the cluster assignments to the output sheet for the best clustering
        outputCol = 3 + i
        outputSheet.Cells(1, outputCol).Value = "Cluster " & clusterCount & " Assignment"
        outputSheet.Cells(2, outputCol).Resize(UBound(bestClusterAssignments, 1), 1).Value = bestClusterAssignments

        ' Store the final centroids in a separate array
        Dim finalCentroids() As Variant
        ReDim finalCentroids(1 To clusterCount, 1 To 2) As Variant
        For j = 1 To clusterCount
            finalCentroids(j, 1) = bestCentroids(j, 1)
            finalCentroids(j, 2) = bestCentroids(j, 2)
        Next j

        ' Write the centroid coordinates to the centroids sheet
        centroidsSheet.Cells(centroidOutputRow, 1).Value = clusterCount

        For j = 1 To clusterCount
            centroidsSheet.Cells(centroidOutputRow, 2).Value = "Centroid " & j ' Centroid Number
            centroidsSheet.Cells(centroidOutputRow, 2 + 1).Value = finalCentroids(j, 1) ' X coordinate
            centroidsSheet.Cells(centroidOutputRow, 2 + 2).Value = finalCentroids(j, 2) ' Y coordinate
            centroidOutputRow = centroidOutputRow + 1 ' Increment the output row counter for centroids
        Next j

        ' Calculate the silhouette score for the best clustering
        silhouetteScore = CalculateSilhouetteScore(inputData, bestClusterAssignments)

        ' Write the silhouette score to the silhouette sheet
        silhouetteSheet.Cells(i + 1, 1).Value = clusterCount
        silhouetteSheet.Cells(i + 1, 2).Value = silhouetteScore

        ' Write X and Y coordinates to the output sheet for the best clustering
        outputSheet.Cells(1, 1).Value = "X Coordinate"
        outputSheet.Cells(1, 2).Value = "Y Coordinate"

         outputRow = 2 ' Reset the output row counter for results

         For j = 1 To UBound(bestClusterAssignments, 1)
             outputSheet.Cells(outputRow, 1).Value = inputData(j, 1) ' X Coordinate
             outputSheet.Cells(outputRow, 2).Value = inputData(j, 2) ' Y Coordinate
             outputRow = outputRow + 1 ' Increment the output row counter for results
         Next j
    Next i

    ' Apply formatting to the output sheet
    outputSheet.Columns.AutoFit
    centroidsSheet.Columns.AutoFit
    silhouetteSheet.Columns.AutoFit
End Sub
```

```
Function CalculateSilhouetteScore(ByVal inputData As Variant, ByVal clusterAssignments As Variant) As Double
    Dim dataLength As Integer
    dataLength = UBound(inputData, 1)

    Dim silhouetteSum As Double
    Dim dataPointCountInCluster() As Integer
    ReDim dataPointCountInCluster(1 To UBound(inputData, 1))

    ' Calculate the number of data points in each cluster
    For i = 1 To dataLength
        dataPointCountInCluster(clusterAssignments(i, 1)) = dataPointCountInCluster(clusterAssignments(i, 1)) + 1
    Next i

    ' Calculate the silhouette score for each data point
    For i = 1 To dataLength
        Dim clusterIndex As Integer
        clusterIndex = clusterAssignments(i, 1)

        Dim a As Double ' Average distance to points in the same cluster
        Dim b As Double ' Average distance to points in the nearest other cluster
        Dim minB As Double
        minB = 100000

        For j = 1 To dataLength
            If i <> j Then
                If clusterAssignments(j, 1) = clusterIndex Then
                    a = a + distance(CDbl(inputData(i, 1)), CDbl(inputData(i, 2)), CDbl(inputData(j, 1)), CDbl(inputData(j, 2)))
                ElseIf dataPointCountInCluster(clusterAssignments(j, 1)) > 0 Then
                    ' Calculate distance to points in the nearest other cluster and find the minimum
                    Dim currB As Double
                    currB = distance(CDbl(inputData(i, 1)), CDbl(inputData(i, 2)), CDbl(inputData(j, 1)), CDbl(inputData(j, 2)))
                    minB = WorksheetFunction.Min(currB, minB)
                End If
            End If
        Next j

        If dataPointCountInCluster(clusterIndex) > 1 Then
            a = a / (dataPointCountInCluster(clusterIndex) - 1)
        End If


        ' Update b with the minimum distance to points in the nearest other cluster
        b = minB

        silhouetteSum = silhouetteSum + (b - a) / WorksheetFunction.Max(a, b)
    Next i

    ' Calculate the average silhouette score for all data points
    Dim silhouetteScore As Double
    silhouetteScore = silhouetteSum / dataLength

    CalculateSilhouetteScore = silhouetteScore
End Function
```

```
Function InitializeCentroidsKMeansPlusPlus(ByVal inputData As Variant, ByRef centroids() As Double, ByVal clusterCount As Integer)
    Dim dataLength As Integer
    Dim randomIndex As Integer
    Dim i As Integer, j As Integer
    Dim minDistance As Double, totalDistance As Double, dist As Double

    dataLength = UBound(inputData, 1)

    ' Select the first centroid randomly from the data points
    Randomize
    randomIndex = Int((dataLength - 1 + 1) * Rnd + 1)
    centroids(1, 1) = inputData(randomIndex, 1)
    centroids(1, 2) = inputData(randomIndex, 2)

    ' Select the remaining centroids using K-means++ strategy
    For i = 2 To clusterCount
        totalDistance = 0

        ' Calculate the distance to the nearest centroid for each data point
        Dim distances() As Double
        ReDim distances(1 To dataLength) As Double

        For j = 1 To dataLength
            minDistance = distance(CDbl(inputData(j, 1)), CDbl(inputData(j, 2)), CDbl(centroids(1, 1)), CDbl(centroids(1, 2)))

            For k = 2 To i - 1
                dist = distance(inputData(j, 1), inputData(j, 2), centroids(k, 1), centroids(k, 2))
                minDistance = WorksheetFunction.Min(minDistance, dist)
            Next k

            distances(j) = minDistance
            totalDistance = totalDistance + minDistance
        Next j

        ' Choose the next centroid with probability proportional to distance squared
        Dim cumulativeProbabilities() As Double
        ReDim cumulativeProbabilities(1 To dataLength) As Double

        cumulativeProbabilities(1) = distances(1) / totalDistance
        For j = 2 To dataLength
            cumulativeProbabilities(j) = cumulativeProbabilities(j - 1) + distances(j) / totalDistance
        Next j


         Dim randomValue As Double
         Randomize
         randomValue = Rnd()
         For j = 1 To dataLength
             If randomValue <= cumulativeProbabilities(j) Then
                 centroids(i, 1) = inputData(j, 1)
                 centroids(i, 2) = inputData(j, 2)
                 Exit For
             End If
         Next j
    Next i
End Function


Function distance(ByVal x1 As Double, ByVal y1 As Double, ByVal x2 As Double, ByVal y2 As Double) As Double
    distance = Sqr((x1 - x2) ^ 2 + (y1 - y2) ^ 2)
End Function
```

```
Function AssignDataPoints(ByVal inputData As Variant, ByVal centroids As Variant) As Variant
    Dim clusterAssignments() As Variant
    Dim i As Integer, j As Integer
    Dim minDistance As Double, minIndex As Integer

    ReDim clusterAssignments(1 To UBound(inputData, 1), 1 To 1)

    For i = 1 To UBound(inputData, 1)
        minDistance = -1
        minIndex = -1

        For j = 1 To UBound(centroids, 1)
            Dim dist As Double
            dist = distance(CDbl(inputData(i, 1)), CDbl(inputData(i, 2)), CDbl(centroids(j, 1)), CDbl(centroids(j, 2)))

            If minDistance = -1 Or dist < minDistance Then
                minDistance = dist
                minIndex = j
            End If
        Next j

        clusterAssignments(i, 1) = minIndex
    Next i

    AssignDataPoints = clusterAssignments
End Function
```

```
Function UpdateCentroids(inputData() As Variant, clusterAssignments() As Variant, ByVal clusterCount As Integer) As Variant
    Dim clusterSums() As Double
    Dim clusterSizes() As Integer
    Dim newCentroids() As Double
    Dim i As Integer, j As Integer

    ReDim clusterSums(1 To clusterCount, 1 To 2) As Double
    ReDim clusterSizes(1 To clusterCount)
    ReDim newCentroids(1 To clusterCount, 1 To 2) As Double

    For i = 1 To UBound(inputData, 1)
        Dim dataX As Double
        Dim dataY As Double
        Dim clusterIndex As Integer

        dataX = CDbl(inputData(i, 1))
        dataY = CDbl(inputData(i, 2))
        clusterIndex = clusterAssignments(i, 1)

        clusterSums(clusterIndex, 1) = clusterSums(clusterIndex, 1) + dataX
        clusterSums(clusterIndex, 2) = clusterSums(clusterIndex, 2) + dataY
        clusterSizes(clusterIndex) = clusterSizes(clusterIndex) + 1
    Next i

    For j = 1 To clusterCount
        If clusterSizes(j) > 0 Then
            newCentroids(j, 1) = clusterSums(j, 1) / clusterSizes(j)
            newCentroids(j, 2) = clusterSums(j, 2) / clusterSizes(j)
        Else
            newCentroids(j, 1) = -9999#
            newCentroids(j, 2) = -9999#
        End If
    Next j

    ' Assign unassigned data points to clusters with zero size
    For i = 1 To UBound(inputData, 1)
        If clusterAssignments(i, 1) = -1 Then
            Dim minDistance As Double
            Dim minIndex As Integer
            minDistance = -1
            minIndex = -1


            For j = 1 To clusterCount
                Dim dist As Double
                dist = distance(CDbl(inputData(i, 1)), CDbl(inputData(i, 2)), CDbl(newCentroids(j, 1)), CDbl(newCentroids(j, 2)))

                If minDistance = -1 Or dist < minDistance Then
                    minDistance = dist
                    minIndex = j
                End If
            Next j

            clusterAssignments(i, 1) = minIndex
        End If
    Next i

    UpdateCentroids = newCentroids
End Function
Function CheckConvergence(ByVal centroids As Variant, ByVal newCentroids As Variant) As Boolean
    Dim i As Integer
    Dim convergenceThreshold As Double

    ' Set the convergence threshold
    convergenceThreshold = 0.00001

    For i = 1 To UBound(centroids, 1)
        ' Check the distance between centroids
        If distance(centroids(i, 1), centroids(i, 2), newCentroids(i, 1), newCentroids(i, 2)) > convergenceThreshold Then
            CheckConvergence = False
            Exit Function
        End If
    Next i

    ' All centroids have converged
    CheckConvergence = True
End Function
```

```
Function CalculateSumDistances(ByVal inputData As Variant, ByVal clusterAssignments As Variant, ByVal centroids As Variant) As Variant
    Dim sumDistances() As Double
    Dim i As Integer, j As Integer

    ReDim sumDistances(1 To UBound(centroids, 1))

    For i = 1 To UBound(inputData, 1)
        Dim pointX As Double
        Dim pointY As Double
        Dim clusterIndex As Integer

        pointX = CDbl(inputData(i, 1))
        pointY = CDbl(inputData(i, 2))
        clusterIndex = clusterAssignments(i, 1)

        Dim centroidX As Double
        Dim centroidY As Double
        centroidX = centroids(clusterIndex, 1)
        centroidY = centroids(clusterIndex, 2)

        sumDistances(clusterIndex) = sumDistances(clusterIndex) + distance(pointX, pointY, centroidX, centroidY)
    Next i

    CalculateSumDistances = sumDistances
End Function
```

## Appendix D: Coordinate data set one

| X coordinate | Y coordinate |
|---|---|
| 45.85531 | 5.26892 |
| 46.00797 | 4.72972 |
| 46.04719 | 5.35678 |
| 45.15881 | 4.13194 |
| 47.28425 | 0.84636 |
| 48.04325 | 3.07164 |
| 48.48522 | 6.08106 |
| 49.47069 | 5.85397 |
| 48.34372 | 5.69444 |
| 50.65631 | 2.89619 |
| 50.78978 | 3.11064 |
| 48.76114 | 2.36297 |
| 49.18261 | 1.33889 |
| 48.86489 | 2.29147 |
| 44.93614 | -0.28292 |
| 44.53775 | 3.51028 |
| 47.09950 | -1.59883 |
| 43.92661 | 5.89333 |
| 49.03517 | 3.39114 |
| 50.58633 | 3.12392 |
| 48.98047 | 1.73856 |
| 48.87353 | 2.30681 |
| 48.88867 | 2.24969 |
| 49.07233 | 0.58311 |
| 48.90225 | 2.37311 |
| 37.70486 | -4.88142 |
| 36.75539 | -2.72300 |
| 37.21653 | -3.72581 |
| 37.30575 | -6.37072 |
| 36.76492 | -2.76664 |
| 37.08275 | -3.77217 |
| 37.37500 | -5.87864 |
| 36.95506 | -2.11106 |
| 37.29228 | -3.05617 |
| 41.78019 | -1.12133 |
| 41.29058 | -3.47622 |
| 41.75369 | 1.91544 |
| 41.63808 | 2.28061 |
| 41.55331 | 2.25264 |
| 41.57883 | 2.51408 |
| 41.59589 | 2.28197 |
| 41.52878 | 1.86825 |
| 41.53906 | 2.13256 |
| 40.63397 | 0.31431 |
| 41.56389 | 2.24592 |
| 40.50753 | -3.40622 |
| 40.43753 | -3.68636 |

| X coordinate | Y coordinate |
|---|---|
| 39.06531 | -1.83781 |
| 38.60933 | -0.58494 |
| 38.85908 | -0.49625 |
| 39.17586 | -0.45708 |
| 39.85631 | -0.46958 |
| 39.58797 | -0.53489 |
| 39.47000 | -0.51564 |
| 39.56108 | -0.63831 |
| 38.34747 | -0.74794 |
| 39.44056 | -0.76069 |
| 39.48369 | -0.58817 |
| 39.46067 | -0.73369 |
| 39.38650 | -0.44831 |
| 38.79144 | -0.02794 |
| 39.59197 | -0.54083 |
| 39.47758 | -0.47269 |
| 40.78772 | -0.80842 |
| 42.43578 | -8.05372 |
| 42.28764 | -7.89550 |
| 43.30814 | -8.29803 |
| 43.33786 | -2.86386 |
| 43.31503 | -3.06353 |
| 43.19717 | -2.07781 |
| 37.82122 | -1.39494 |
| 37.82072 | -1.39725 |
| 37.94783 | -1.18778 |
| 37.37700 | -5.92403 |
| 43.33069 | -4.09989 |
| 39.08978 | -0.68797 |
| 37.52842 | -5.12172 |
| 38.71511 | -4.08886 |
| 37.44864 | -3.88461 |
| 43.20789 | -2.03828 |
| 39.43764 | -0.47731 |
| 41.62694 | 2.28694 |
| 41.46703 | 2.16961 |
| 41.93692 | -4.47839 |
| 38.40475 | -0.54297 |
| 42.88256 | -8.32172 |
| 40.95772 | -4.20408 |
| 41.98364 | 2.28681 |

## Appendix E: Data set two

| X coordinate | Y coordinate |
|---|---|
| 42 | 21 |
| 29 | 34 |
| 26 | 22 |
| 25 | 34 |
| 43 | 12 |
| 29 | 49 |
| 45 | 24 |
| 24 | 44 |
| 18 | 43 |
| 33 | 15 |
| 21 | 36 |
| 16 | 48 |
| 37 | 38 |
| 22 | 21 |
| 27 | 9 |
| 43 | 23 |
| 27 | 20 |
| 46 | 24 |
| 38 | 35 |
| 44 | 37 |
| 42 | 33 |
| 25 | 40 |
| 28 | 46 |
| 19 | 18 |
| 31 | 38 |
| 47 | 38 |
| 40 | 6 |
| 16 | 45 |
| 17 | 14 |
| 38 | 6 |
| 27 | 16 |
| 35 | 33 |
| 45 | 24 |
| 15 | 41 |
| 25 | 46 |
| 38 | 16 |
| 41 | 5 |
| 50 | 8 |
| 50 | 22 |
| 32 | 46 |
| -38 | 40 |
| -18 | 34 |
| -48 | 35 |
| -14 | 41 |
| -48 | 34 |
| -39 | 10 |
| -26 | 3 |

| X coordinate | Y coordinate |
|---|---|
| -42 | 50 |
| -43 | 19 |
| -25 | 34 |
| -23 | 49 |
| -24 | 34 |
| -23 | 26 |
| -24 | 26 |
| -39 | 49 |
| -46 | 38 |
| -28 | 33 |
| -42 | 14 |
| -16 | 45 |
| -50 | 45 |
| -32 | 42 |
| -13 | 6 |
| -24 | 10 |
| -36 | 29 |
| -16 | 40 |
| -27 | 9 |
| -38 | 5 |
| -45 | 12 |
| -44 | 4 |
| -30 | 11 |
| -49 | 50 |
| -32 | 5 |
| -30 | 3 |
| -41 | 27 |
| -23 | 8 |
| -43 | 36 |
| -34 | 38 |
| -13 | 20 |
| -42 | 41 |
| -35 | 18 |
| -27 | -10 |
| 7 | -48 |
| 7 | -50 |
| -18 | -43 |
| -11 | -41 |
| 2 | -41 |
| -18 | -47 |
| -11 | -35 |
| 6 | -35 |
| -13 | -37 |
| -23 | -48 |
| 8 | -48 |
| 7 | -47 |
| -23 | -44 |

| X coordinate | Y coordinate |
|---|---|
| -28 | -35 |
| -19 | -42 |
| 0 | -36 |
| 5 | -37 |
| -18 | -36 |
| -30 | -40 |
| -7 | -36 |
| 7 | -40 |
| -1 | -50 |
| 2 | -38 |
| -13 | -38 |
| -1 | -49 |
| -10 | -47 |
| -16 | -49 |
| 0 | -44 |
| -16 | -43 |
| -5 | -40 |
| 5 | -38 |
| -16 | -44 |
| -12 | -35 |
| -28 | -48 |
| -6 | -35 |
| 4 | -37 |
| -29 | -47 |
| -16 | -45 |
| -27 | -39 |

## Appendix F: Data set three

| X coordinate | Y coordinate |
|---|---|
| 23 | 47 |
| 45 | 53 |
| 34 | 60 |
| 32 | 46 |
| 50 | 57 |
| 25 | 64 |
| 48 | 63 |
| 59 | 36 |
| 62 | 53 |
| 46 | 40 |
| 61 | 61 |
| 53 | 64 |
| 39 | 50 |
| 40 | 67 |
| 24 | 49 |
| 26 | 61 |
| 55 | 62 |
| 33 | 52 |
| 52 | 60 |
| 47 | 52 |
| 19 | 56 |
| 60 | 55 |
| 43 | 68 |
| 44 | 46 |
| 36 | 42 |
| 63 | 47 |
| 29 | 60 |
| 42 | 57 |
| 28 | 68 |
| 38 | 62 |
| 64 | 58 |
| 54 | 51 |
| 49 | 62 |
| 58 | 57 |
| 37 | 44 |
| 51 | 58 |
| 65 | 42 |
| 30 | 61 |
| 57 | 68 |
| 31 | 56 |
| -25 | 35 |
| -46 | 32 |
| -60 | 22 |
| -40 | 29 |
| -27 | 13 |
| -32 | 38 |
| -30 | 41 |

| X coordinate | Y coordinate |
|---|---|
| -50 | 16 |
| -78 | 37 |
| -51 | 15 |
| -42 | 25 |
| -70 | 19 |
| -55 | 6 |
| -47 | 18 |
| -28 | 30 |
| -33 | 11 |
| -66 | 43 |
| -45 | 42 |
| -72 | 7 |
| -71 | 33 |
| -59 | 24 |
| -37 | 12 |
| -31 | 43 |
| -48 | 29 |
| -53 | 17 |
| -65 | 41 |
| -61 | 34 |
| -69 | 25 |
| -39 | 31 |
| -75 | 40 |
| -41 | 23 |
| -54 | 28 |
| -52 | 39 |
| -77 | 8 |
| -62 | 21 |
| -67 | 35 |
| -49 | 7 |
| -74 | 42 |
| -36 | 27 |
| -79 | 10 |
| -56 | 44 |
| -64 | 15 |
| -44 | 20 |
| -34 | 45 |
| -58 | 40 |
| -29 | 16 |
| -57 | 38 |
| -38 | 42 |
| -73 | 31 |
| -43 | 39 |
| -68 | 13 |
| -76 | 17 |
| -63 | 11 |
| -80 | 26 |

| X coordinate | Y coordinate |
|---|---|
| -35 | 23 |
| -26 | 28 |
| -10 | -26 |
| -22 | -27 |
| -8 | -30 |
| -19 | -31 |
| -15 | -25 |
| -13 | -34 |
| -6 | -35 |
| -9 | -28 |
| -25 | -25 |
| -18 | -32 |
| -23 | -30 |
| -16 | -33 |
| -12 | -29 |
| -20 | -28 |
| -21 | -31 |
| -17 | -26 |
| -7 | -33 |
| -24 | -27 |
| -5 | -30 |
| -11 | -32 |
| 21 | -27 |
| 42 | -47 |
| 29 | -40 |
| 37 | -41 |
| 35 | -28 |
| 26 | -53 |
| 24 | -55 |
| 27 | -30 |
| 50 | -25 |
| 38 | -44 |
| 43 | -52 |
| 36 | -43 |
| 32 | -49 |
| 44 | -28 |
| 47 | -36 |
| 33 | -45 |
| 23 | -29 |
| 46 | -50 |
| 25 | -34 |
| 22 | -38 |
| 28 | -48 |
| 48 | -27 |
| 49 | -45 |
| 30 | -39 |
| 45 | -51 |

| X coordinate | Y coordinate |
|---|---|
| 40 | -26 |
| 39 | -42 |
| 34 | -32 |
| 31 | -50 |
| 41 | -33 |
| 20 | -30 |
| 50 | -55 |
| 48 | -37 |
| 46 | -34 |
| 42 | -29 |
| 44 | -49 |
| 25 | -28 |
| 49 | -34 |
| 21 | -42 |
| 47 | -46 |
| 36 | -50 |
| 45 | -30 |
| 39 | -30 |
| 23 | -44 |
| 30 | -55 |
| 26 | -38 |
| 43 | -38 |
| 40 | -28 |
| 22 | -46 |
| 27 | -55 |
| 28 | -35 |
| 38 | -34 |

## Appendix G: Data set four

| X coordinate | Y coordinate | X coordinate | Y coordinate | X coordinate | Y Coordinate | X coordinate | Y coordinate |
|---|---|---|---|---|---|---|---|
| 23 | 47 | -50 | 16 | -35 | 23 | 40 | -26 |
| 45 | 53 | -78 | 37 | -26 | 28 | 39 | -42 |
| 34 | 60 | -51 | 15 | -10 | -26 | 34 | -32 |
| 32 | 46 | -42 | 25 | -22 | -27 | 31 | -50 |
| 50 | 57 | -70 | 19 | -8 | -30 | 41 | -33 |
| 25 | 64 | -55 | 6 | -19 | -31 | 20 | -30 |
| 48 | 63 | -47 | 18 | -15 | -25 | 50 | -55 |
| 59 | 36 | -28 | 30 | -13 | -34 | 48 | -37 |
| 62 | 53 | -33 | 11 | -6 | -35 | 46 | -34 |
| 46 | 40 | -66 | 43 | -9 | -28 | 42 | -29 |
| 61 | 61 | -45 | 42 | -25 | -25 | 44 | -49 |
| 53 | 64 | -72 | 7 | -18 | -32 | 25 | -28 |
| 39 | 50 | -71 | 33 | -23 | -30 | 49 | -34 |
| 40 | 67 | -59 | 24 | -16 | -33 | 21 | -42 |
| 24 | 49 | -37 | 12 | -12 | -29 | 47 | -46 |
| 26 | 61 | -31 | 43 | -20 | -28 | 36 | -50 |
| 55 | 62 | -48 | 29 | -21 | -31 | 45 | -30 |
| 33 | 52 | -53 | 17 | -17 | -26 | 39 | -30 |
| 52 | 60 | -65 | 41 | -7 | -33 | 23 | -44 |
| 47 | 52 | -61 | 34 | -24 | -27 | 30 | -55 |
| 19 | 56 | -69 | 25 | -5 | -30 | 26 | -38 |
| 60 | 55 | -39 | 31 | -11 | -32 | 43 | -38 |
| 43 | 68 | -75 | 40 | 21 | -27 | 40 | -28 |
| 44 | 46 | -41 | 23 | 42 | -47 | 22 | -46 |
| 36 | 42 | -54 | 28 | 29 | -40 | 27 | -55 |
| 63 | 47 | -52 | 39 | 37 | -41 | 28 | -35 |
| 29 | 60 | -77 | 8 | 35 | -28 | 38 | -34 |
| 42 | 57 | -62 | 21 | 26 | -53 | -81 | -77 |
| 28 | 68 | -67 | 35 | 24 | -55 | -81 | -66 |
| 38 | 62 | -49 | 7 | 27 | -30 | -87 | -80 |
| 64 | 58 | -74 | 42 | 50 | -25 | -90 | -67 |
| 54 | 51 | -36 | 27 | 38 | -44 | -87 | -72 |
| 49 | 62 | -79 | 10 | 43 | -52 | -86 | -78 |
| 58 | 57 | -56 | 44 | 36 | -43 | -88 | -77 |
| 37 | 44 | -64 | 15 | 32 | -49 | -86 | -66 |
| 51 | 58 | -44 | 20 | 44 | -28 | -85 | -73 |
| 65 | 42 | -34 | 45 | 47 | -36 | -90 | -73 |
| 30 | 61 | -58 | 40 | 33 | -45 | -82 | -75 |
| 57 | 68 | -29 | 16 | 23 | -29 | -81 | -72 |
| 31 | 56 | -57 | 38 | 46 | -50 | -88 | -70 |
| -25 | 35 | -38 | 42 | 25 | -34 | -84 | -65 |
| -46 | 32 | -73 | 31 | 22 | -38 | -86 | -69 |
| -60 | 22 | -43 | 39 | 28 | -48 | -80 | -72 |
| -40 | 29 | -68 | 13 | 48 | -27 | -89 | -78 |
| -27 | 13 | -76 | 17 | 49 | -45 | -80 | -66 |
| -32 | 38 | -63 | 11 | 30 | -39 | -87 | -70 |
| -30 | 41 | -80 | 26 | 45 | -51 | -81 | -66 |

## Appendix H: Data set five

| X coordinate | Y coordinate |
|---|---|
| 23 | 2 |
| 6 | 3 |
| 24 | 4 |
| 10 | 5 |
| 4 | 6 |
| 31 | 7 |
| 17 | 8 |
| 9 | 9 |
| 28 | 10 |
| 31 | 11 |
| 26 | 12 |
| 29 | 13 |
| 30 | 14 |
| 31 | 15 |
| 26 | 16 |
| 24 | 17 |
| 5 | 18 |
| 15 | 19 |
| 19 | 20 |
| 24 | 21 |
| 79 | 79 |
| 49 | 67 |
| 69 | 67 |
| 73 | 58 |
| 54 | 59 |
| 51 | 66 |
| 71 | 77 |
| 46 | 47 |
| 63 | 70 |
| 70 | 64 |
| 76 | 57 |
| 63 | 80 |
| 76 | 46 |
| 68 | 52 |
| 79 | 64 |
| 68 | 49 |
| 71 | 52 |
| 76 | 72 |
| 60 | 74 |
| 48 | 80 |
| 43 | 45 |
| 42 | 39 |
| 37 | 43 |
| 36 | 36 |
| 39 | 35 |
| 40 | 36 |
| 39 | 38 |

| X coordinate | Y coordinate |
|---|---|
| 35 | 40 |
| 39 | 44 |
| 36 | 37 |
| 39 | 45 |
| 38 | 41 |
| 37 | 45 |
| 43 | 38 |
| 35 | 44 |
| 39 | 39 |
| 42 | 35 |
| 42 | 37 |
| 38 | 36 |
| 36 | 44 |
| 17 | 66 |
| 11 | 42 |
| 20 | 46 |
| 18 | 51 |
| 10 | 70 |
| 19 | 55 |
| 13 | 61 |
| 16 | 50 |
| 5 | 49 |
| 3 | 61 |
| 19 | 61 |
| 1 | 49 |
| 12 | 63 |
| 19 | 43 |
| 6 | 70 |
| 15 | 58 |
| 5 | 42 |
| 15 | 52 |
| 37 | 41 |
| 41 | 8 |
| 44 | 40 |
| 7 | 20 |
| 43 | 45 |
| 10 | 16 |
| 3 | 3 |
| 35 | 41 |
| 15 | 34 |
| 25 | 29 |
| 35 | 40 |
| 15 | 34 |
| 25 | 29 |

## Appendix I: Data set six

| X coordinate | Y coordinate |
|---|---|
| 42 | 21 |
| 29 | 34 |
| 26 | 22 |
| 25 | 34 |
| 43 | 12 |
| 29 | 49 |
| 45 | 24 |
| 24 | 44 |
| 18 | 43 |
| 33 | 15 |
| 21 | 36 |
| 16 | 48 |
| 37 | 38 |
| 22 | 21 |
| 27 | 9 |
| 43 | 23 |
| 27 | 20 |
| 46 | 24 |
| 38 | 35 |
| 44 | 37 |
| 42 | 33 |
| 25 | 40 |
| 28 | 46 |
| 19 | 18 |
| 31 | 38 |
| 47 | 38 |
| 40 | 6 |
| 16 | 45 |
| 17 | 14 |
| 38 | 6 |
| 27 | 16 |
| 35 | 33 |
| 45 | 24 |
| 15 | 41 |
| 25 | 46 |
| 38 | 16 |
| 41 | 5 |
| 50 | 8 |
| 50 | 22 |
| 32 | 46 |
| -38 | 40 |
| -18 | 34 |
| -48 | 35 |
| -14 | 41 |
| -48 | 34 |
| -39 | 10 |
| -26 | 3 |

| X coordinate | Y coordinate |
|---|---|
| -42 | 50 |
| -43 | 19 |
| -25 | 34 |
| -23 | 49 |
| -24 | 34 |
| -23 | 26 |
| -24 | 26 |
| -39 | 49 |
| -46 | 38 |
| -28 | 33 |
| -42 | 14 |
| -16 | 45 |
| -50 | 45 |
| -32 | 42 |
| -13 | 6 |
| -24 | 10 |
| -36 | 29 |
| -16 | 40 |
| -27 | 9 |
| -38 | 5 |
| -45 | 12 |
| -44 | 4 |
| -30 | 11 |
| -49 | 50 |
| -32 | 5 |
| -30 | 3 |
| -41 | 27 |
| -23 | 8 |
| -43 | 36 |
| -34 | 38 |
| -13 | 20 |
| -42 | 41 |
| -35 | 18 |
| -27 | -10 |
| 7 | -48 |
| 7 | -50 |
| -18 | -43 |
| -11 | -41 |
| 2 | -41 |
| -18 | -47 |
| -11 | -35 |
| 6 | -35 |
| -13 | -37 |
| -23 | -48 |
| 8 | -48 |
| 7 | -47 |
| -23 | -44 |

| X coordinate | Y coordinate |
|---|---|
| -28 | -35 |
| -19 | -42 |
| 0 | -36 |
| 5 | -37 |
| -18 | -36 |
| -30 | -40 |
| -7 | -36 |
| 7 | -40 |
| -1 | -50 |
| 2 | -38 |
| -13 | -38 |
| -1 | -49 |
| -10 | -47 |
| -16 | -49 |
| 0 | -44 |
| -16 | -43 |
| -5 | -40 |
| 5 | -38 |
| -16 | -44 |
| -12 | -35 |
| -28 | -48 |
| -6 | -35 |
| 4 | -37 |
| -29 | -47 |
| -16 | -45 |
| -27 | -39 |
| 49 | 70 |
| 46 | 60 |
| 44 | 57 |
| 52 | 63 |
| 46 | 66 |
| 44 | 64 |
| 40 | 66 |
| 43 | 62 |
| 59 | 59 |
| 44 | 69 |
| 57 | 59 |
| 54 | 55 |
| 60 | 67 |
| 42 | 56 |
| 52 | 57 |
| 59 | 57 |
| 49 | 61 |
| 50 | 69 |
| 54 | 60 |
| 55 | 67 |
| 50 | 70 |

| X coordinate | Y coordinate |
|---|---|
| 60 | 90 |
| 67 | 86 |
| 64 | 90 |
| 60 | 84 |
| 64 | 85 |
| 69 | 88 |
| 60 | 90 |
| 65 | 90 |
| 66 | 81 |
| 64 | 90 |
| 61 | 82 |
| 70 | 83 |
| 60 | 81 |
| 62 | 84 |
| 64 | 88 |
| 65 | 81 |
| 65 | 85 |
| 61 | 86 |
| 63 | 85 |
| -90 | -83 |
| -86 | -79 |
| -84 | -81 |
| -78 | -86 |
| -82 | -81 |
| -79 | -82 |
| -80 | -83 |
| -83 | -76 |
| -79 | -88 |
| -80 | -80 |
| -88 | -81 |
| -82 | -77 |
| -78 | -78 |
| -84 | -88 |
| -86 | -79 |
| -76 | -81 |
| -85 | -89 |
| 66 | -90 |
| 78 | -85 |
| 63 | -90 |
| 64 | -90 |
| 68 | -97 |
| 62 | -87 |
| 76 | -89 |
| 64 | -93 |
| 69 | -85 |
| 72 | -80 |
| 70 | -93 |

## Appendix J: visualisations of data sets

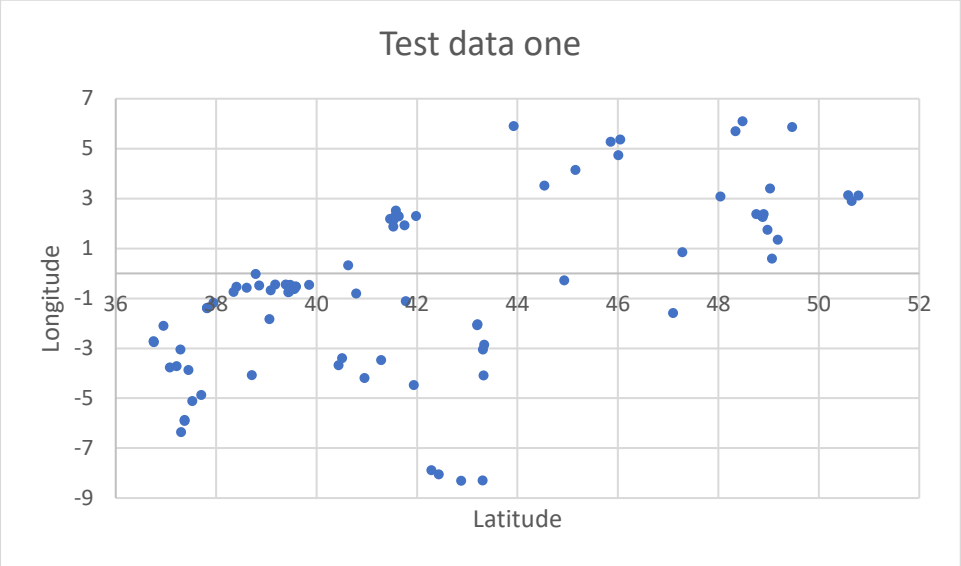


*Figure 11 Visualisation test data set one.*

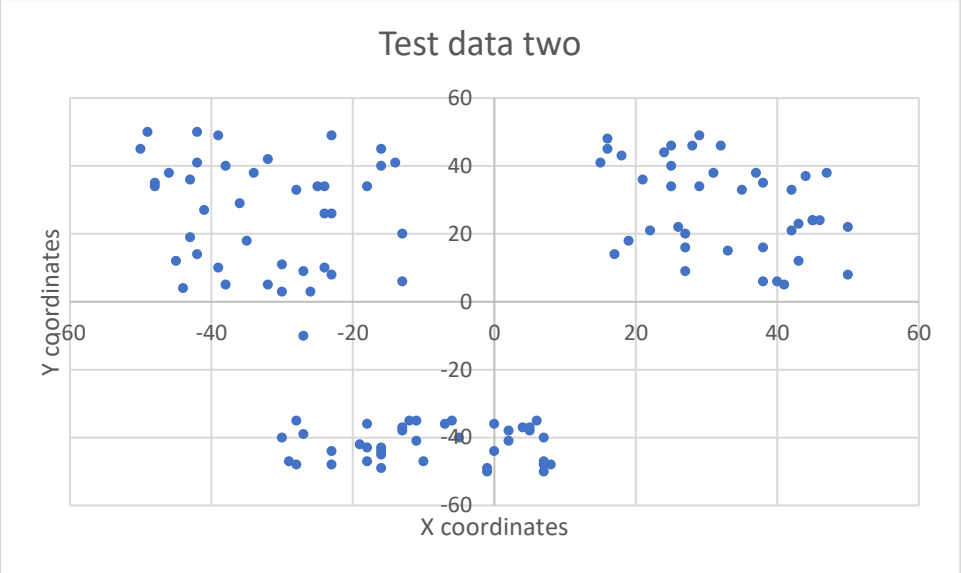


*Figure 12 Visualisation test data set two*

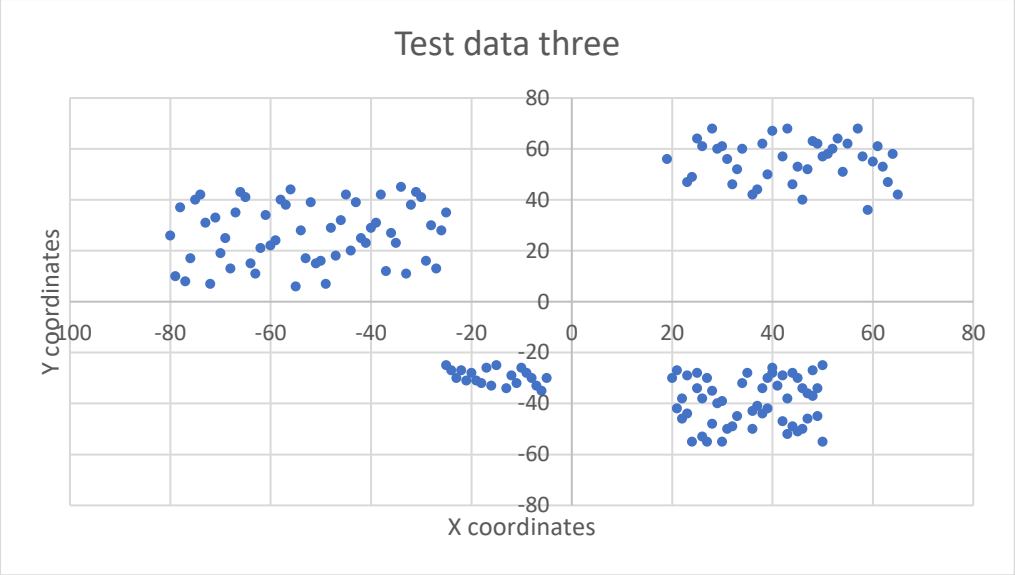


*Figure 13 Visualisation test data set three.*

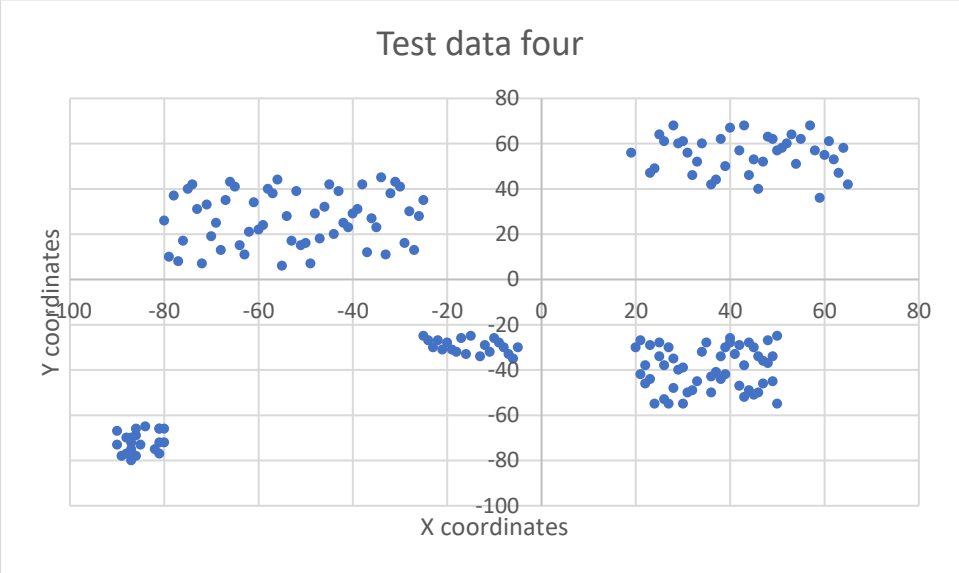


*Figure 14 Visualisation test data set four.*

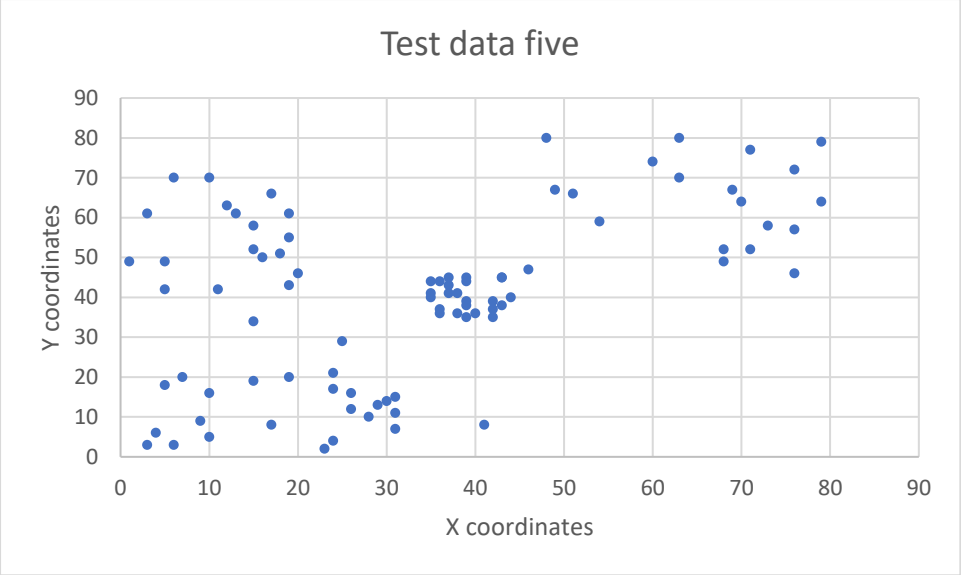


*Figure 15 Visualisation test data set five.*

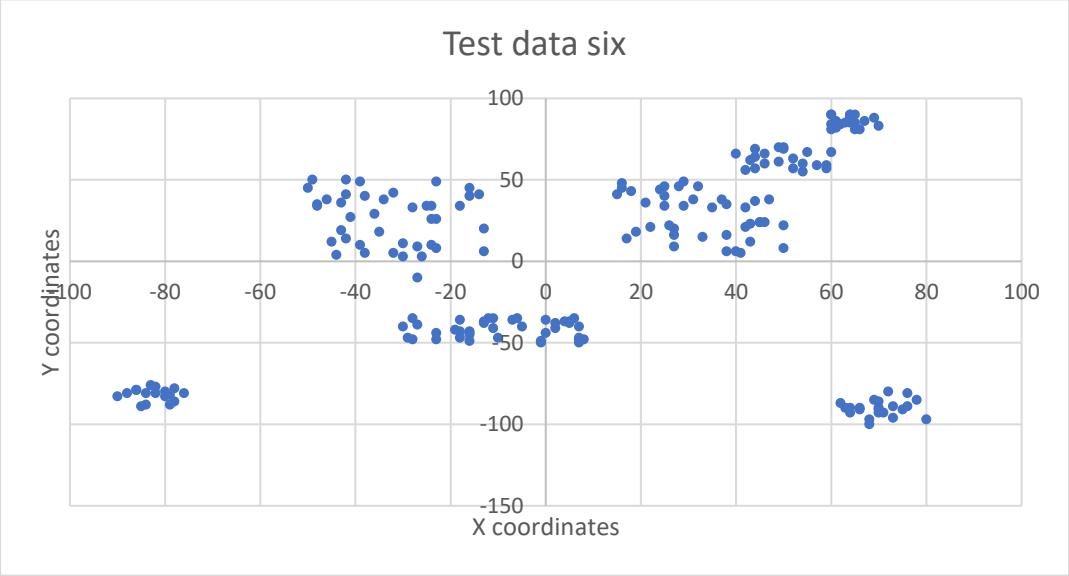


*Figure 16 Visualisation test data set six.*

## Appendix K: Centre of gravity implementation in Visual Basic

```
Option Explicit
Sub CalculateCenterOfGravity()
    Dim ws As Worksheet
    Dim lastRow As Long
    Dim i As Long, j As Long
    Dim sumX As Double, sumY As Double, sumDemand As Double
    Dim centerX As Double, centerY As Double
    Dim cluster As Long
    Dim cogDict As Object

    Set ws = ThisWorkbook.Worksheets("Center of Gravity")

    ' Find the last row with data in column A
    lastRow = ws.Cells(ws.Rows.Count, "A").End(xlUp).Row

    ' Initialize the dictionary to store the center of gravity for each cluster
    Set cogDict = CreateObject("Scripting.Dictionary")

    ' Loop through the data starting from row 2
    For i = 2 To lastRow
        ' Read the cluster assignment from column D
        cluster = ws.Cells(i, "D").Value
        ' Reset the sums for each cluster
        sumX = 0
        sumY = 0
        sumDemand = 0

        ' Loop through the data again to calculate the sums for each cluster
        For j = 2 To lastRow
            If ws.Cells(j, "D").Value = cluster Then
                ' Weighted sum of x and y based on demand
                sumX = sumX + ws.Cells(j, "A").Value * ws.Cells(j, "C").Value
                sumY = sumY + ws.Cells(j, "B").Value * ws.Cells(j, "C").Value

                ' Sum of demand for the current cluster
                sumDemand = sumDemand + ws.Cells(j, "C").Value
            End If
        Next j


        ' Calculate the center of gravity for the current cluster
        If sumDemand > 0 Then
            centerX = sumX / sumDemand
            centerY = sumY / sumDemand
            ' Store the center of gravity coordinates in the dictionary, if not already present
            If Not cogDict.Exists(cluster) Then
                cogDict.Add cluster, Array(centerX, centerY)
            End If
        End If
    Next i

   ' Write the center of gravity coordinates and cluster number to the respective columns
    Dim resultRow As Long
    resultRow = 2 ' Starting row to write the results in the worksheet

    Dim clusterKey As Variant ' New variable for looping through the dictionary keys

    For Each clusterKey In cogDict.Keys
    ' Write the center of gravity coordinates (CenterX and CenterY) in columns K and L
        ws.Cells(resultRow, "K").Value = cogDict(clusterKey)(0) ' CenterX in column K
        ws.Cells(resultRow, "L").Value = cogDict(clusterKey)(1) ' CenterY in column L

        ' Write the cluster number in column M
        ws.Cells(resultRow, "J").Value = "Cluster " & clusterKey

        resultRow = resultRow + 1
    Next clusterKey

End Sub
```

## Appendix L: Demand data set one

| Demand |
|---|
| 18000 |
| 8300 |
| 10000 |
| 12000 |
| |
| 5000 |
| |
| 1000 |
| 52000 |
| |
| 90000 |
| |
| 55000 |
| |
| |
| 50000 |
| 2640 |
| |
| |
| |
| 45000 |
| |
| |
| |
| |
| |
| |
| 12,000 |
| |
| |
| |
| 125,000 |
| 4,500 |
| 40,000 |
| |
| 40,000 |
| 14,400 |
| 50,000 |
| |
| |
| |
| 14,500 |
| |
| |
| 39,000 |
| |
| |

| Demand |
|---|
| 150,000 |
| 40,000 |
| 12,000 |
| 1,000 |
| 70,000 |
| |
| |
| 30,000 |
| 48,000 |
| 10,000 |
| 30,000 |
| |
| 85,000 |
| 4,200 |
| |
| |
| |
| 14,400 |
| |
| |
| |
| |
| |
| 100,000 |
| 7850 |
| |
| 16,000 |
| |
| 7,000 |
| |
| |
| |
| |
| 15,000 |
| |
| 10000 |
| |
| |
| |
| |

## Appendix M: Demand data set two

| Demand | Demand | Demand |
|---|---|---|
| 981 | 1036 | 1024 |
| 981 | 1120 | 1000 |
| 942 | 979 | 961 |
| 1099 | 949 | 943 |
| 979 | 934 | 978 |
| 981 | 1010 | 987 |
| 950 | 1078 | 944 |
| 1008 | 959 | 1062 |
| 1040 | 1003 | 1028 |
| 949 | 997 | 1017 |
| 1011 | 982 | 1022 |
| 947 | 988 | 962 |
| 1040 | 1005 | 963 |
| 931 | 1101 | 997 |
| 1022 | 962 | 1083 |
| 931 | 948 | 1008 |
| 1055 | 954 | 951 |
| 962 | 1081 | 1016 |
| 1005 | 1036 | 1001 |
| 1088 | 948 | 1040 |
| 1027 | 1047 | 1018 |
| 975 | 1115 | 1075 |
| 975 | 986 | 1008 |
| 1163 | 970 | 1041 |
| 1003 | 956 | 953 |
| 923 | 1005 | 1060 |
| 1111 | 959 | |
| 1050 | 1008 | |
| 940 | 1034 | |
| 1077 | 1054 | |
| 1053 | 1002 | |
| 1014 | 1031 | |
| 1042 | 1015 | |
| 995 | 1028 | |
| 965 | 954 | |
| 1048 | 971 | |
| 1108 | 997 | |
| 963 | 970 | |
| 943 | 1004 | |
| 970 | 1010 | |
| 986 | 1001 | |
| 1007 | 996 | |
| 1010 | 1045 | |
| 1096 | 980 | |
| 1045 | 1120 | |
| 927 | 1003 | |
| 977 | 948 | |

## Appendix N: Demand data set three

| Demand |
|---|
| 991 |
| 973 |
| 961 |
| 977 |
| 1071 |
| 1016 |
| 1000 |
| 996 |
| 932 |
| 1093 |
| 995 |
| 963 |
| 1021 |
| 973 |
| 1038 |
| 1031 |
| 1058 |
| 1055 |
| 991 |
| 956 |
| 964 |
| 944 |
| 966 |
| 972 |
| 960 |
| 1022 |
| 986 |
| 1027 |
| 947 |
| 1058 |
| 969 |
| 968 |
| 1005 |
| 986 |
| 1058 |
| 961 |
| 1000 |
| 963 |
| 1017 |
| 944 |
| 954 |
| 1036 |
| 1038 |
| 1051 |
| 990 |
| 972 |
| 982 |

| Demand |
|---|
| 966 |
| 1034 |
| 1031 |
| 966 |
| 1012 |
| 1046 |
| 968 |
| 965 |
| 1011 |
| 973 |
| 1034 |
| 1043 |
| 964 |
| 1018 |
| 1056 |
| 1033 |
| 1085 |
| 1051 |
| 962 |
| 1077 |
| 1000 |
| 1085 |
| 987 |
| 964 |
| 981 |
| 1017 |
| 1050 |
| 1004 |
| 967 |
| 989 |
| 1040 |
| 1016 |
| 972 |
| 1001 |
| 1033 |
| 1053 |
| 1001 |
| 1042 |
| 1037 |
| 1017 |
| 970 |
| 985 |
| 1050 |
| 1005 |
| 993 |
| 1082 |
| 973 |

| Demand |
|---|
| 1024 |
| 1058 |
| 1051 |
| 1014 |
| 1078 |
| 1053 |
| 1026 |
| 1007 |
| 1008 |
| 1082 |
| 995 |
| 1024 |
| 1013 |
| 1004 |
| 1101 |
| 968 |
| 1058 |
| 970 |
| 999 |
| 1012 |
| 995 |
| 994 |
| 1028 |
| 995 |
| 992 |
| 992 |
| 1011 |
| 989 |
| 986 |
| 1002 |
| 1047 |
| 1032 |
| 985 |
| 1020 |
| 1014 |
| 1005 |
| 970 |
| 1046 |
| 1021 |
| 983 |
| 1003 |
| 1031 |
| 1009 |
| 1091 |
| 1031 |
| 1040 |
| 978 |

| Demand |
|---|
| 1056 |
| 1020 |
| 1052 |
| 1054 |
| 998 |
| 1011 |
| 961 |
| 1038 |
| 1022 |
| 994 |
| 972 |
| 977 |
| 1065 |
| 1021 |
| 1017 |
| 1014 |
| 1028 |
| 1022 |
| 998 |
| 1022 |
| 1013 |
| 1062 |

## Appendix O: Demand data set four

| Demand |
|---|
| 1242 |
| 1255 |
| 1043 |
| 1417 |
| 1250 |
| 1205 |
| 1200 |
| 1382 |
| 1163 |
| 1371 |
| 1422 |
| 1536 |
| 1193 |
| 1068 |
| 1302 |
| 1456 |
| 1178 |
| 1251 |
| 1323 |
| 1342 |
| 1086 |
| 1248 |
| 1160 |
| 1221 |
| 1300 |
| 1207 |
| 1131 |
| 1271 |
| 1257 |
| 1278 |
| 1433 |
| 1374 |
| 1343 |
| 1297 |
| 1414 |
| 1264 |
| 1342 |
| 1367 |
| 1102 |
| 1373 |
| 1404 |
| 1161 |
| 1228 |
| 1148 |
| 1299 |
| 1390 |
| 1131 |

| Demand |
|---|
| 1458 |
| 1146 |
| 1288 |
| 1295 |
| 1364 |
| 1238 |
| 1250 |
| 1261 |
| 1448 |
| 1322 |
| 1304 |
| 1180 |
| 1291 |
| 1201 |
| 1412 |
| 1219 |
| 1232 |
| 1257 |
| 1297 |
| 1259 |
| 1369 |
| 1265 |
| 1241 |
| 1192 |
| 1325 |
| 1337 |
| 1403 |
| 1341 |
| 1223 |
| 1316 |
| 1352 |
| 1313 |
| 1440 |
| 1290 |
| 1221 |
| 1379 |
| 1233 |
| 1414 |
| 1143 |
| 1221 |
| 1252 |
| 1265 |
| 1305 |
| 1327 |
| 1435 |
| 1276 |
| 1258 |

| Demand |
|---|
| 1233 |
| 1433 |
| 1173 |
| 1388 |
| 1179 |
| 1368 |
| 1306 |
| 1358 |
| 1228 |
| 1172 |
| 1380 |
| 1198 |
| 1382 |
| 1261 |
| 1336 |
| 1292 |
| 1375 |
| 1386 |
| 1395 |
| 1295 |
| 1355 |
| 1205 |
| 1448 |
| 1162 |
| 1375 |
| 1238 |
| 1197 |
| 1343 |
| 1242 |
| 1163 |
| 1222 |
| 1222 |
| 1197 |
| 1275 |
| 1315 |
| 1375 |
| 1238 |
| 1365 |
| 1294 |
| 1200 |
| 1156 |
| 1366 |
| 1362 |
| 1374 |
| 1278 |
| 1325 |
| 1291 |

| Demand |
|---|
| 1413 |
| 1236 |
| 1285 |
| 1254 |
| 1337 |
| 1307 |
| 1262 |
| 1252 |
| 1249 |
| 1376 |
| 1350 |
| 1359 |
| 1163 |
| 1243 |
| 1206 |
| 1273 |
| 1239 |
| 1377 |
| 1277 |
| 1237 |
| 1305 |
| 1252 |
| 1360 |
| 1211 |
| 1375 |
| 1253 |
| 1223 |
| 1210 |
| 1329 |
| 1391 |
| 1326 |
| 1222 |
| 1255 |
| 1306 |
| 1275 |
| 1201 |
| 1314 |
| 1340 |
| 1212 |
| 1385 |
| 1237 |
| 1371 |
| 1279 |
| 1372 |
| 1185 |
| 1307 |
| 1373 |

## Appendix P: Demand data set five

| Demand |
|---|
| 1451 |
| 2595 |
| 1169 |
| 2118 |
| 6056 |
| 1522 |
| 1280 |
| 6679 |
| 1394 |
| 1195 |
| 1238 |
| 1223 |
| 1153 |
| 1630 |
| 1532 |
| 1184 |
| 13678 |
| 1116 |
| 1186 |
| 1337 |
| 1139 |
| 1322 |
| 1173 |
| 959 |
| 1146 |
| 1428 |
| 1519 |
| 1423 |
| 1497 |
| 995 |
| 965 |
| 1498 |
| 1097 |
| 1366 |
| 100000 |
| 1001 |
| 1138 |
| 100000 |
| 905 |
| 963 |
| 1395 |
| 1126 |
| 1041 |
| 1218 |
| 1322 |
| 1283 |
| 1010 |

| Demand |
|---|
| 1301 |
| 1541 |
| 1392 |
| 1246 |
| 1347 |
| 1456 |
| 1092 |
| 1288 |
| 1126 |
| 1395 |
| 1110 |
| 1555 |
| 1289 |
| 9362 |
| 1338 |
| 1307 |
| 1192 |
| 13678 |
| 1356 |
| 12250 |
| 1291 |
| 1328 |
| 1444 |
| 1335 |
| 1336 |
| 1127 |
| 1325 |
| 1027 |
| 1243 |
| 1077 |
| 1397 |
| 1412 |
| 1227 |
| 1173 |
| 1189 |
| 1370 |
| 1480 |
| 1046 |
| 1380 |
| 1262 |
| 1305 |

## Appendix Q: Demand data set six

| Demand |
|---|
| 7971 |
| 7338 |
| 8167 |
| 7107 |
| 7733 |
| 8091 |
| 8762 |
| 8880 |
| 8377 |
| 7166 |
| 10552 |
| 12102 |
| 13703 |
| 14116 |
| 11131 |
| 13910 |
| 12404 |
| 10836 |
| 15042 |
| 12663 |
| 5430 |
| 5446 |
| 5802 |
| 4240 |
| 5810 |
| 5467 |
| 4512 |
| 5412 |
| 5403 |
| 4532 |
| 18608 |
| 17205 |
| 15456 |
| 11839 |
| 14456 |
| 13519 |
| 16226 |
| 12995 |
| 10996 |
| 14796 |
| 8967 |
| 10012 |
| 8771 |
| 12236 |
| 8805 |
| 8946 |
| 9747 |

| Demand |
|---|
| 9396 |
| 8582 |
| 10111 |
| 22187 |
| 11406 |
| 12133 |
| 19054 |
| 17193 |
| 10082 |
| 16907 |
| 13586 |
| 10025 |
| 19355 |
| 7252 |
| 6178 |
| 5402 |
| 6010 |
| 7638 |
| 6000 |
| 7736 |
| 9340 |
| 5642 |
| 8815 |
| 15122 |
| 14348 |
| 22297 |
| 18625 |
| 16172 |
| 16639 |
| 10769 |
| 17897 |
| 19203 |
| 18175 |
| 9549 |
| 10706 |
| 12334 |
| 9209 |
| 12302 |
| 11081 |
| 11439 |
| 9678 |
| 9949 |
| 11309 |
| 19242 |
| 20016 |
| 21556 |
| 16357 |

| Demand |
|---|
| 16505 |
| 20081 |
| 19817 |
| 17491 |
| 14736 |
| 20137 |
| 8536 |
| 8021 |
| 8440 |
| 6880 |
| 8498 |
| 7637 |
| 8168 |
| 7952 |
| 8100 |
| 6846 |
| 15043 |
| 12418 |
| 9652 |
| 11653 |
| 11294 |
| 14386 |
| 14411 |
| 13307 |
| 10382 |
| 10165 |
| 7691 |
| 6423 |
| 5538 |
| 7177 |
| 6209 |
| 6478 |
| 6351 |
| 5819 |
| 4766 |
| 7176 |
| 10062 |
| 14043 |
| 12926 |
| 13329 |
| 13235 |
| 16328 |
| 13171 |
| 13867 |
| 17240 |
| 14455 |
| 8272 |

| Demand |
|---|
| 8198 |
| 7920 |
| 8733 |
| 9925 |
| 9384 |
| 9486 |
| 9909 |
| 8092 |
| 10931 |
| 14114 |
| 10017 |
| 16695 |
| 18678 |
| 15797 |
| 11539 |
| 15237 |
| 11168 |
| 15654 |
| 13871 |
| 8543 |
| 7109 |
| 7005 |
| 7472 |
| 7240 |
| 8086 |
| 8545 |
| 8772 |
| 5605 |
| 6976 |
| 12216 |
| 18498 |
| 18557 |
| 16177 |
| 20044 |
| 15821 |
| 16368 |
| 18763 |
| 15128 |
| 16339 |
| 11166 |
| 10590 |
| 10664 |
| 10934 |
| 8852 |
| 10793 |
| 10808 |
| 12228 |

## Appendix R: Centre of gravity coordinates per data set per cluster

| **Data set one** | | |
|---|---|---|
| | **X coordinate** | **Y coordinate** |
| **Cluster one** | 48.27384313 | 3.317750741 |
| **Cluster two** | 39.24029162 | -1.644563883 |
| **Data set two** | | |
| | **X coordinate** | **Y coordinate** |
| **Cluster one** | 32.34758 | 28.08666484 |
| **Cluster two** | -32.4706 | 25.99225707 |
| **Cluster three** | -9.08876 | -41.83510897 |
| **Data set three** | | |
| | **X coordinate** | **Y coordinate** |
| **Cluster one** | 43.8760653 | 55.2472 |
| **Cluster two** | -52.56904048 | 26.67944 |
| **Cluster three** | -14.98145193 | -29.6297 |
| **Cluster four** | 35.19905388 | -40.3071 |
| **Data set four** | | |
| | **X coordinate** | **Y coordinate** |
| **Cluster one** | 43.99056 | 55.169 |
| **Cluster two** | -52.5652 | 26.564 |
| **Cluster three** | -15.2144 | -29.582 |
| **Cluster four** | 35.55173 | -39.396 |
| **Cluster five** | -85.1138 | -71.789 |
| **Data set five** | | |
| | **X coordinate** | **Y coordinate** |
| **Cluster one** | 13.99542 | 12.35158 |
| **Cluster two** | 76.35052 | 67.73236 |
| **Cluster three** | 39.15908 | 40.53448 |
| **Cluster four** | 12.75917 | 60.55937 |
| **Data set six** | | |
| | **X coordinate** | **Y coordinate** |
| **Cluster one** | 33.05174653 | 28.08077912 |
| **Cluster two** | -32.58133626 | 27.39415889 |
| **Cluster three** | -9.396655371 | -41.7310915 |
| **Cluster four** | 51.09402724 | 62.11171375 |
| **Cluster five** | 63.63789783 | 85.28104088 |
| **Cluster six** | -82.43171151 | -81.66431302 |
| **Cluster seven** | 70.71244589 | -90.6566083 |

## Appendix S: Visualisation of the centre of gravity and centroids for each data set

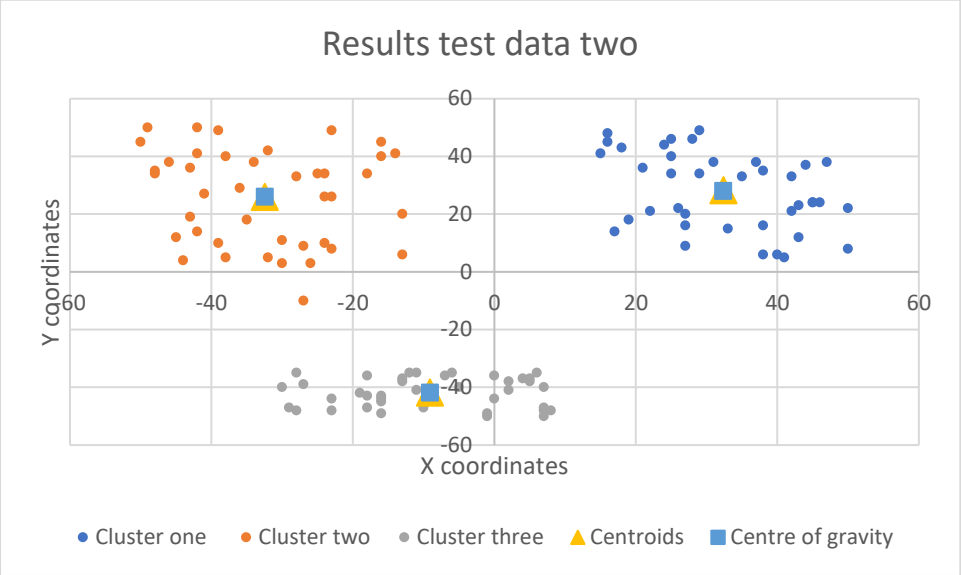


*Figure 17 Visualisation results data set two.*

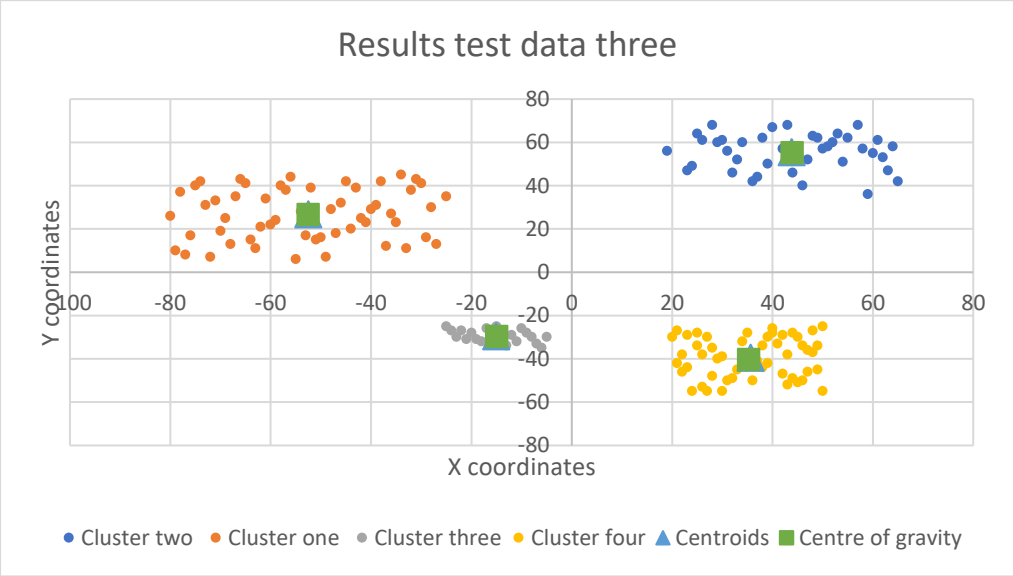


*Figure 18 Visualisation results data set three.*

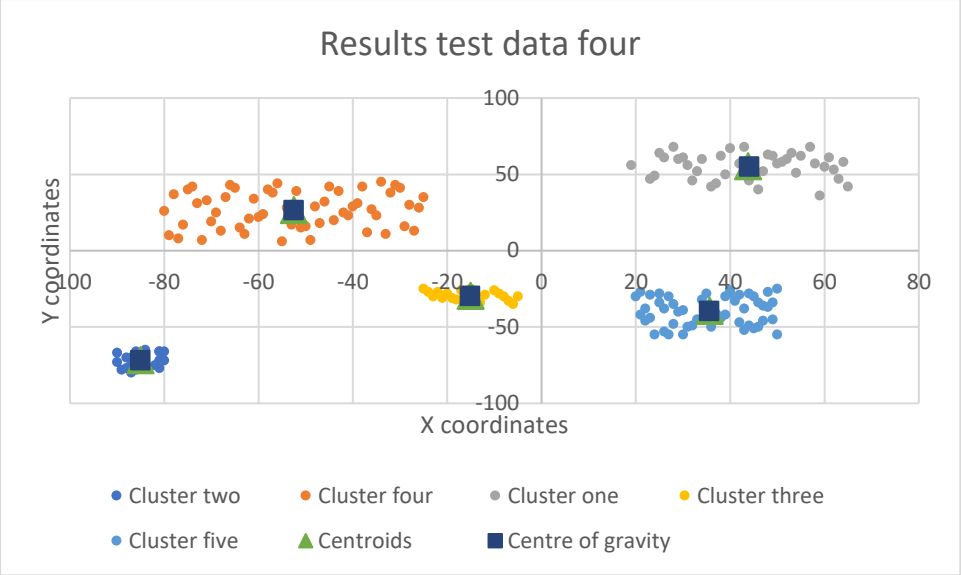


*Figure 19 visualisation results data set four.*

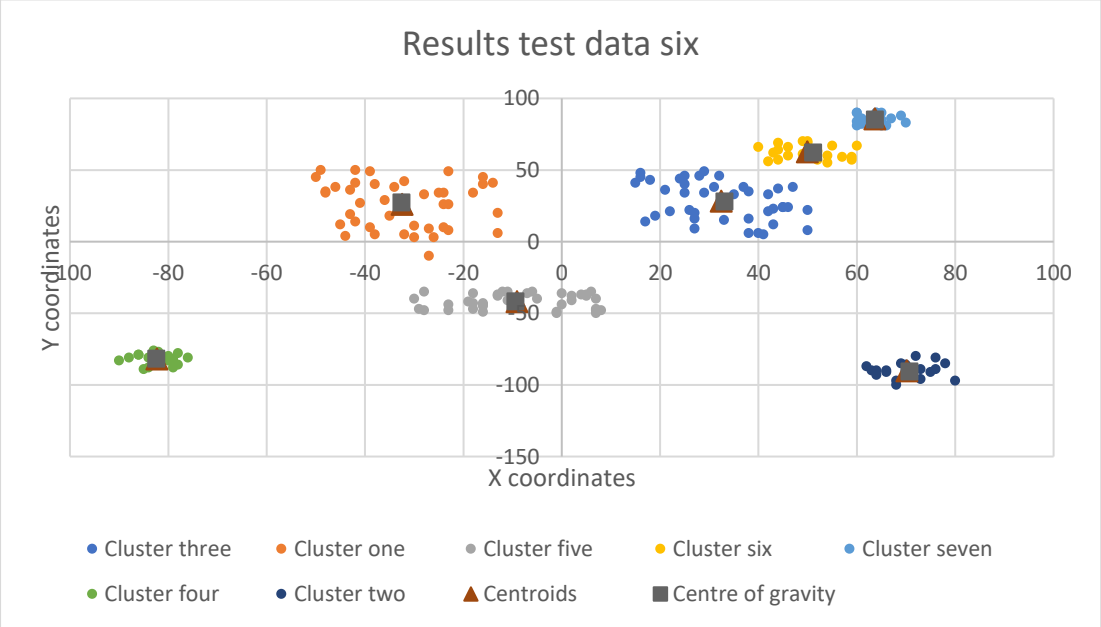


*Figure 20 visualisation results data set six.*

## Appendix T: Centroid coordinates per data set

| **Data set one** | | |
|---|---|---|
| | **X coordinate** | **Y coordinate** |
| **Centroid one** | 47.91159556 | 3.052927776 |
| **Centroid two** | 39.963056 | -1.796699286 |
| **Data set two** | | |
| | **X coordinate** | **Y coordinate** |
| **Centroid one** | 32.4 | 28.25 |
| **Centroid two** | -32.43902439 | 25.92682927 |
| **Centroid three** | -9.102564103 | -41.84615385 |
| **Data set three** | | |
| | **X coordinate** | **Y coordinate** |
| **Centroid one** | 43.8 | 55.375 |
| **Centroid two** | -52.5 | 26.71428571 |
| **Centroid three** | -15.05 | -29.6 |
| **Centroid four** | 35.65384615 | -39.5 |
| **Data set four** | | |
| | **X coordinate** | **Y coordinate** |
| **Centroid one** | 43.8 | 55.375 |
| **Centroid two** | -52.5 | 26.71428571 |
| **Centroid three** | -15.05 | -29.6 |
| **Centroid four** | 35.65384615 | -39.5 |
| **Centroid five** | -85.04761905 | -71.76190476 |
| **Data set five** | | |
| | **X coordinate** | **Y coordinate** |
| **Centroid one** | 19.92 | 12.24 |
| **Centroid two** | 66.52631579 | 64.89473684 |
| **Centroid three** | 39.2 | 40.44 |
| **Centroid four** | 12.57894737 | 53.84210526 |
| **Data set six** | | |
| | **X coordinate** | **Y coordinate** |
| **Centroid one** | 32.4 | 28.25 |
| **Centroid two** | -32.43902439 | 25.92682927 |
| **Centroid three** | -9.102564103 | -41.84615385 |
| **Centroid four** | 49.95238095 | 62.57142857 |
| **Centroid five** | 63.68421053 | 85.73684211 |
| **Centroid six** | -82.35294118 | -81.88235294 |
| **Centroid seven** | 70.18181818 | -90.18181818 |

## Appendix U: Pairwise comparison respondent A

| | Proximity to market | Infrastructure | Land availability | Government support | Labor supply | Total cost |
|---|---|---|---|---|---|---|
| **Proximity to market** | **1** | 1 | 9 | 9 | 1 | 1 |
| **Infrastructure** | 1 | **1** | 9 | 9 | 1 | 1 |
| **Land availability** | $\frac{1}{9}$ | $\frac{1}{9}$ | **1** | 1 | $\frac{1}{9}$ | $\frac{1}{9}$ |
| **Government support** | $\frac{1}{9}$ | $\frac{1}{9}$ | 1 | **1** | $\frac{1}{5}$ | $\frac{1}{5}$ |
| **Labour supply** | 1 | 1 | 9 | 5 | **1** | 1 |
| **Total costs** | 1 | 1 | 9 | 5 | 1 | **1** |

## Appendix V: Pairwise comparison respondent B

| | Proximity to market | Infrastructure | Land availability | Government support | Labor supply | Total cost |
|---|---|---|---|---|---|---|
| **Proximity to market** | **1** | 5 | 4 | 3 | 4 | 1 |
| **Infrastructure** | $\frac{1}{5}$ | **1** | $\frac{1}{4}$ | $\frac{1}{4}$ | 1 | $\frac{1}{4}$ |
| **Land availability** | $\frac{1}{4}$ | 4 | **1** | 3 | 1 | 1 |
| **Government support** | $\frac{1}{3}$ | 4 | $\frac{1}{3}$ | **1** | $\frac{1}{2}$ | $\frac{1}{3}$ |
| **Labour supply** | $\frac{1}{4}$ | 1 | 1 | 2 | **1** | 1 |
| **Total costs** | 1 | 4 | 1 | 3 | 1 | **1** |

## Appendix W: Pairwise comparison respondent C

| | Proximity to market | Infrastructure | Land availability | Government support | Labor supply | Total cost |
|---|---|---|---|---|---|---|
| **Proximity to market** | **1** | $\frac{1}{3}$ | 6 | 7 | 1 | 2 |
| **Infrastructure** | 3 | **1** | 9 | 5 | 3 | 2 |
| **Land availability** | $\frac{1}{6}$ | $\frac{1}{9}$ | **1** | $\frac{1}{6}$ | $\frac{1}{7}$ | $\frac{1}{6}$ |
| **Government support** | $\frac{1}{7}$ | $\frac{1}{5}$ | 6 | **1** | $\frac{1}{6}$ | $\frac{1}{3}$ |
| **Labour supply** | 1 | $\frac{1}{3}$ | 7 | 6 | **1** | 1 |
| **Total costs** | $\frac{1}{2}$ | $\frac{1}{2}$ | 6 | 3 | 1 | **1** |