The Performance of Different Variations of the K-Means Clustering Method

Bachelor Thesis Econometrics and Operations Research

Major specialisation: Business Analytics and Quantitative Marketing

Marjolein de With

Student ID number: 483866

Supervisor: C. Cavicchia

Second assessor: P. H. B. F. Franses

Abstract

This thesis focuses on comparing K-Means clustering with four different variations: K-Means++, K-Harmonic Means, Fuzzy K-Means, and, developed specifically for this comparison study, K-Harmonic Means++. This leads to the research question 'In what ways can the standard K-Means clustering method be improved?'. To answer this question, I concentrate on the internal clustering validity and the initialisation dependence of the different methods. I use eight different data sets to evaluate the different methods, with a particular focus on a data set on energy dependence in the EU by Eurostat (2015). I find that K-Means++ generally outperforms the other methods, closely followed by both Harmonic Means methods.

Erasmus University Rotterdam

Erasmus School of Economics

July 4, 2020

Contents

1 Introduction

The widely known K-Means clustering method has been around since 1957 (Steinhaus, 1957) and was given its current name by MacQueen (1967). This method is used for clustering analysis, which originated in 1932 (Driver and Kroeber, 1932), and groups objects such that objects in one cluster are more similar to each other than objects in different clusters. Since its first introduction, K-Means, which tries to minimise the within-cluster variance (Steinhaus, 1957), has received both praise and criticism. To battle its shortcomings, many new clustering methods have been developed. Some of these have an approach very different from K-Means, such as Density-based spatial clustering of applications with noise (DBSCAN), developed by Martin, Kriegel, Sander, and Xu (1996), which is a non-parametric algorithm that groups closely packed points together. Others tried to address the shortcomings of K-Means, by slightly adjusting the original method. I focus on these methods in this thesis. I evaluate the performance of different variations of the K-Means clustering method, namely standard K-Means, K-Means++, K-Harmonic Means, Fuzzy K-Means and a combination of K-Means++ and K-Harmonic Means, that I call 'K-Harmonic Means++'. I appoint hierarchical clustering using Ward's procedure as a benchmark. To evaluate the performances, I look at both the internal clustering validity of the methods and at the extent to which they depend on the initialisation of their centres.

The advantage of adjusting K-Means in favour of developing a new clustering method altogether is that the advantages of K-Means, which I discuss later in this thesis, can be preserved, while some of its shortcomings could be eliminated. This thesis focuses on what variation of K-Means achieves this goal best.

The different K-Means variations are applied to eight different data sets, but in particular, I discuss a data set regarding the energy dependence of EU countries. I am extending the research done by Bluszcz (2016), who applied hierarchical clustering using Ward's procedure and standard K-Means on the same data set. The author uses hierarchical clustering to find the number of clusters and then forms these clusters using K-Means. The data set is small, with only 28 observations, namely, the current 27 EU member states plus the United Kingdom, and three variables.

Horuckova (2016) discusses the importance of energy dependence in the EU. After the first oil shock, in 1973, energy security became an important item on the political agenda. Chevalier (2005) defines energy security as:

- a reliable supply of energy;

- reliable transportation of energy;
- a reliable distribution and delivery of supply to the final customer;
- a 'reasonable price' over a continuous period.

Because of the uncertainty since the beginning of this century around the energy dependence of the EU, the Commission of the European Communities published a Green Paper (Commission of the European Communities, 2006). This document discusses the energy mix of the EU and considers what steps member states should take to assure their future energy security. Bluszcz (2016) attempted to cluster the EU member states, based on their dependence on oil, coal and natural gas, such that long-term common energy policies can be formulated.

To make a fitting clustering of the EU member states based on their energy dependence, I formulate the following research question: 'In what ways can the standard K-Means clustering method be improved?'. I divide this question into two sub-questions: 'What are appropriate ways to improve the internal clustering validity of the standard K-Means clustering method?', and 'What are appropriate ways to increase the probability that the standard K-Means clustering method reaches a global optimum?'. These questions will be answered with the use of different evaluation techniques.

I find that for most data sets, K-Means++ gives the best results. K-Harmonic Means and K-Harmonic Means++ do not differ greatly from each other in terms of performance. Fuzzy K-Means gives better results than standard K-Means, but not by much. As the implementation of K-Means++ is easy, it would be useful to have standard K-Means exchanged for it in all sorts of applications.

The rest of this thesis is structured as follows. Section 2 discusses existing literature on the standard K-Means method and the comparison of different clustering algorithms. Then, Section 3 discusses the data set I mainly focus on. I explain the hierarchical benchmark method, standard K-Means and its variations in Section 4. Afterwards, in Section 5, I discuss the evaluation of these methods and make comparisons. Finally, Section 6 gives some concluding remarks and suggestions for future research.

2 Literature Review

Clustering analysis has been around for almost a century. It has its origins in anthropology (Driver and Kroeber, 1932) and was applied to psychology a few years later (Zubin, 1938; Tryon, 1939).

Clustering analysis became widely known when it was used in personality psychology by Cattell (1943). Since then, cluster analysis has become one of the main tools in marketing research (Punj and Stewart, 1983).

My research focuses on comparing several variations of the standard K-Means clustering method. The classic method has some advantages and disadvantages. MacQueen (1967), who came up with one of the first K-Means variants, states that K-Means creates reasonably efficient clusters regarding minimising the variance within the clusters, is easily programmed and computationally easy. Therefore, this classic method is very suitable to process large amounts of data. Xu and Wunsch (2005) evaluate some of its disadvantages. The authors mention the method's sensitivity to its initialisation, meaning that the number of clusters has to be set beforehand. Another problem regarding the initialisation is that the convergence of K-Means algorithms to a global optimum is not guaranteed, but depends on the initial centres that have been chosen. A third disadvantage of K-Means is that it is sensitive to outliers and noise in the data, as objects far away from a cluster centre are still forced in the cluster, thereby distorting the shape of the cluster. K-Means detects spherical clusters, thereby constraining them. Various researchers have made suggestions related to these limitations. Ball and Hall (1967), Pham, Dimov, and Nguyen (2005) and Redmond and Heneghan (2007) researched different ways to decide on the number of clusters. Others focused on the initialisation of the centres (Likas, Vlassis, and Verbeek, 2003; Bradley and Fayyad, 1998; Peña, Lozano, and Larrañaga, 1999) or on the sensitivity to outliers and noise (Estivill-Castro and Yang, 2004). Likas et al. (2003) worked with random restarts, which means running an algorithm several times, with different starting points, and taking the best outcome.

For as long as clustering analysis has existed, there have been comparative studies on different methods. Dougherty, Barrera, and Brun (2002) compared the basis K-Means method with fuzzy K-Means, self-organising maps, hierarchical Euclidean-distance based clustering and correlation-based clustering in the field of genetics. Maulik and Bandyopadhyay (2002) compared K-Means with single linkage clustering and a Simulated Annealing based technique, using among others the Davies-Bouldin Index and Dunn's Index. Another example is the work by De Souto, Costa, Araujo, Ludermir, and Schliep (2008), who studied the comparative performance of K-Means, single, complete and average linkage, multivariate Gaussians, spectral clustering and a nearest neighbour-based method on clustering cancer gene expression data. Finally, Costa, Carvalho, and Souto (2004) studied the comparative performance of K-Means, agglomerative hierarchical clustering, Cluster Identification via Connectivity Kernels (CLICK), dynamical clustering and self-organising

maps, also in the field of genetics, using evaluation indices like the corrected Rand Index and the Hubbert Index.

In my research, I compare the standard K-Means clustering method with some improved K-Means methods. By analysing the internal clustering validity and the initialisation dependence of these variations, I can evaluate the quality and relevance of the adjustments that have been proposed. Hamerly and Elkan (2002) have conducted similar research, by comparing K-Means to K-Harmonic Means, Fuzzy K-Means and two hybrid versions of K-Means and K-Harmonic Means. Their research, however, was mainly focused on what aspects of these methods actually improve standard K-Means. The authors found that soft membership, meaning that objects belong to multiple clusters to a certain degree, is essential for finding good clusterings. In case of methods with a hard membership, where each object belongs to one cluster only, varying weights are beneficial. The authors do not explicitly focus on the internal clustering validity and the initialisation dependence of the different methods. I extend their research by also discussing two other methods, namely K-Means++ and K-Harmonic Means++. Third, Hamerly and Elkan (2002) only compare clustering performance using two artificial data sets that both contain true underlying clusters, thus mainly focusing on the external clustering validity, whereas I use 'real world' data sets, that do not necessarily have a natural clustering.

3 Data

The data used both by Bluszcz (2016) and in this paper can be found in Chapter 2 of the Statistical book by Eurostat (2015). This book provides the total energy dependence of the formerly 28 EU member states, namely the current 27 member states plus the United Kingdom, for the period 2004-13. Furthermore, it provides the countries' energy dependence on solid fuels and its derivatives, petroleum products and natural gas respectively. The energy dependence of a country indicates to what extent it relies on import. The Directorate-General for Economic and Financial Affairs (2013) gives the following formula to calculate this dependence for energy product i:

$$\mathrm{ED}_i = \frac{m_i - x_i}{\mathrm{GIC}_i + \mathrm{BUNK}_i}, \tag{1}$$

where m_i is the import of product i, x_i the export, GIC_i the gross inland consumption and $BUNK_i$ represents the consumption of ships and aircraft on international routes. Table A1 contains the energy dependence of the 27 EU member states plus the United Kingdom in 2013. Table 1 contains descriptive statistics on the dependence on the three energy products. As standardisation of data

is recommended for clustering (Milligan and Cooper, 1988), I present the standardised results in Table A2.

Table 1: Descriptive statistics of the three variables used for clustering

	Mean	Median	St.Dev	Min.	Max.
Fuels	62.07	81.3	41.40	-11.6	111.6
Petroleum	87.18	95.95	25.64	-13.7	106.2
Gas	70.63	95.75	48.49	-86.8	115.6

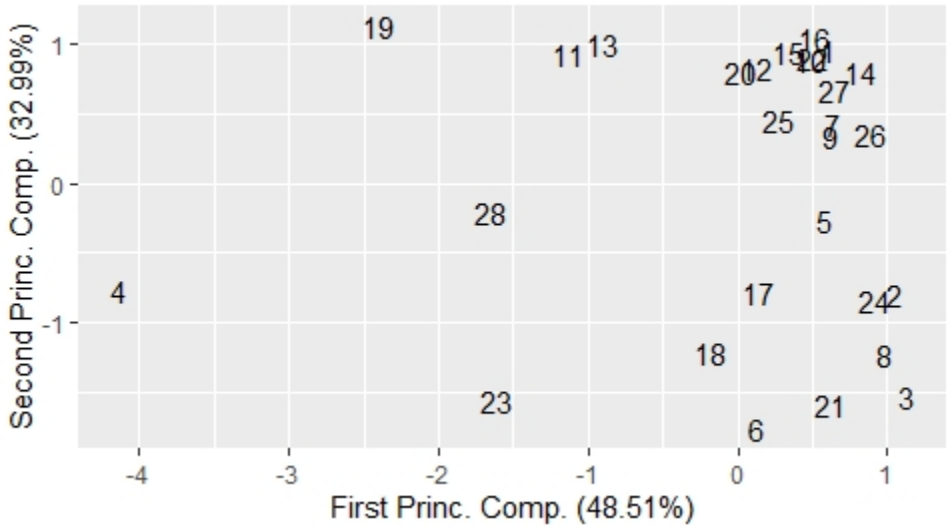

Figure 1: 27 EU member states plus the United Kingdom in two-dimensional space

The first two principal components of the data are shown in Figure 1. In this figure, the numbers correspond with the numbers given to the countries in Table A1. Together, those two principal components explain 81.5% of the total variance.

The other seven data sets I evaluate are discussed in Section 5.2.

4 Methodology

I compare the clustering performance of four different K-Means clustering variations. I evaluate those variations using eight different data sets and discuss the data set on the energy dependence of the 27 EU member states plus the United Kingdom in particular. In Section 4.1, I discuss hierarchical clustering. Then, after deciding the optimal number of clusters using these hierarchical

methods, I will apply several non-hierarchical clustering methods to the data in Section 4.2. Finally, I explain how those methods are evaluated in Section 4.3.

4.1 Hierarchical Methods

Malhotra, Nunan, and Birks (2017) describe hierarchical clustering in Chapter 25. Hierarchical clustering can be *agglomerative* or *divisive*. For divisive clustering, all objects are grouped in a single cluster at first. Then, clusters are divided until each object is in a separate cluster. For agglomerative clustering, all objects are placed in separate clusters at first. Then, objects are iteratively grouped into bigger clusters, until there is one single cluster left. Agglomerative methods include *linkage methods*, *variance methods* and *centroid methods*. Linkage methods cluster objects based on their distance. Variance methods approach clustering differently, namely by trying to minimise the within-cluster variances. A commonly used variance method is *Ward's procedure* (Ward, 1963), also used by Bluszcz (2016). Centroid methods focus on merging clusters with the smallest difference between their centroids. The general algorithm for agglomerative hierarchical clustering, according to Chapter 12.3 of the book by Johnson and Wichern (2007), is described in Algorithm 1. As Bluszcz (2016) only discusses Ward's procedure, I also limit this section to that method.

Algorithm 1: Agglomerative hierarchical clustering algorithm

1. Place all n objects in separate clusters and create an $n \times n$ symmetric distance matrix $\mathbf{D} = d_{ik}$.
2. Search $\mathbf{D}$ for the nearest pair of clusters. Let the distance between these clusters U and V be d_{UV}.
3. Merge U and V. Relabel this newly formed cluster *(UV)*. Update distance matrix $\mathbf{D}$ by replacing the rows and columns corresponding to clusters U and V by a row and column for the distances between the new cluster *(UV)* and the remaining clusters.
4. Repeat steps 2 and 3 $n - 1$ times.

For Ward's procedure, the squared Euclidean distance of each object to the mean of their respective cluster is calculated (Ward, 1963). These distances are summed for all the objects, and then the clusters that would cause the smallest increase in the Error Sum of Squares (ESS) are combined. If there are currently k clusters formed, the ESS is the sum over the ESS_j:

$$\text{ESS} = \text{ESS}_1 + \text{ESS}_2 + ... + \text{ESS}_k. \tag{2}$$

The ESS for cluster j is computed as follows:

$$\text{ESS}_j = \sum_{k \in j} (\mathbf{x}_k - \overline{\mathbf{x}_j})'(\mathbf{x}_k - \overline{\mathbf{x}_j}), \tag{3}$$

where $\mathbf{x}_k$ is the position of the kth object and $\overline{\mathbf{x}_j}$ the mean of all the objects in cluster j.

To find out the optimal number of clusters for our data, I apply the *average silhouette method* to Ward's procedure. This method is introduced in Chapter 2 of the book by Kaufman and Rousseeuw (1990). The method uses the *silhouette width* to evaluate the relationship between the tightness of the objects within the clusters and their separation from objects from different clusters. Its exact specification is

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \tag{4}$$

where $a(i)$ is the average distance between object i and all the other objects in its cluster, and $b(i)$ the average distance between object i and the objects in the other clusters. If a cluster only contains one object, the value of $s(i)$ is set to zero. From Equation (4) follows that $-1 \leq s(i) \leq 1$. If $s(i)$ is close to -1, i should be assigned to a different cluster. If $s(i)$ is close to 1, the cluster i belongs to is well-specified. To find the optimal number of clusters, I calculate the silhouette width of all the individual objects, for different numbers of clusters. The largest average silhouette width indicates the optimal number of clusters.

4.2 Non-Hierarchical Methods

Bluszcz (2016) not only discusses hierarchical but also non-hierarchical clustering methods. The author focuses on the K-Means method. A characteristic of non-hierarchical methods like the K-Means method is that the number of clusters has to be specified in advance. Those methods either start from an initial partition of objects into groups or from an initial random set of cluster centres. In this section, I first discuss the K-Means method. Then, I discuss some adjustments to this method, namely, the K-Means++ method, the K-Harmonic Means method and the Fuzzy K-Means method respectively. Finally, I explain a combination of the K-Means++ method and the K-Harmonic Means method, which I name K-Harmonic Means++.

The original idea for the K-Means method comes from Steinhaus (1957), but the term K-Means itself was first used by MacQueen (1967). I use the algorithm by Hartigan and Wong (1979). K-Means is a method that assigns each item to the cluster with the nearest centre, resulting in the

following objective function:

$$\mathrm{KM}(X, C) = \sum_{j=1}^{k} \sum_{i:c(i)=j} ||x_i - c_j||^2, \tag{5}$$

where $x_i \in X$ is the ith object, $c_j \in C$ the jth cluster centre and $||\cdot||$ the Euclidean distance. The algorithm for K-Means is shown in Algorithm 2.

Algorithm 2: K-Means clustering algorithm

1. Randomly specify k initial centres.
2. Assign each object to the cluster with the nearest centre.
3. For each object, assign it to its nearest centre. If this changes anything, recalculate the centres for the old and new clusters. This is done by finding the arithmetic means of the members of the clusters.
4. Repeat step 3 a pre-specified amount of times or until no more reassignments take place. Assign all x_i to the c_j with the shortest Euclidean distance.

Arthur and Vassilvitskii (2007) first introduced the K-Means++ clustering method. It augments the standard K-Means method with a simple seeding technique. Experiments conducted by Arthur and Vassilvitskii (2007) have shown that this seeding technique often improves the accuracy of K-Means greatly. The intuition behind K-Means++ is that the k initial clusters are optimally spread out, as centres 2,...,k are chosen with a probability directly proportional to the squared distance to the closest existing centre, thereby favouring objects that do not have a nearby centre. After choosing these initial centres, K-Means++ follows the same procedure as K-Means, with the same objective function as shown in Equation (5). The exact algorithm for K-Means++ is shown in Algorithm 3.

Algorithm 3: K-Means++ clustering algorithm

1. Choose centre c_1 uniformly at random from X.
2. For each object, compute $D(x)$, which is the Euclidean distance to the nearest centre that already exists.
3. Choose new centre c_i, choosing $x_i \in X$ with probability $\frac{D(x)^2}{\sum_{x_i \in X} D(x)^2}$.
4. Repeat step 2 until there are k centres.
5. Proceed with steps 2-4 from Algorithm 2.

Zhang, Hsu, and Dayal (1999) recognise a flaw in the K-Means method. Namely, this method is

very sensitive to the initialisation of the centres. Therefore, K-Means does not always lead to a global optimum. To combat this, the authors developed the K-Harmonic Means method, which tries to avoid this by using the harmonic mean of the distances from each object to the centres in its objective function. This means that large weights are assigned to data points that are not close to any centres and small weights to data points that are. Additionally, as K-Harmonic Means is a soft clustering method: every object has a *membership* for every centre, indicating to what degree it belongs to this centre. Eventually, objects are assigned to the cluster for which this membership degree is highest. Mathematically, the objective function for K-Harmonic Means can be described as follows:

$$\text{KHM}(X, C) = \sum_{i=1}^{n} \frac{k}{\sum_{j=1}^{k} \frac{1}{||x_i - c_j||^p}}. \tag{6}$$

Jiang, Yi, Li, Yang, and Hu (2010) suggest choosing $p \geq 2$. I find the optimal value of p by taking 51 points in the interval [2, 7]. As K-Harmonic Means depends on a seed for its initialisation, I calculate which p gives, after a certain amount of iterations, the highest average silhouette width. Jiang et al. (2010) also describe an algorithm for K-Harmonic Means, shown in Algorithm 4.

Algorithm 4: K-Harmonic Means clustering algorithm

1 Randomly specify k initial centres.

2 Calculate the objective function value $\text{KHM}(X, C)$.

3 For each object, calculate its membership to centre c_j:

$$m(c_j, x_i) = \frac{||x_i - c_j||^{-p-2}}{\sum_{j=1}^{k} ||x_i - c_j||^{-p-2}}. \tag{7}$$

4 For each object, calculate its weight:

$$w(x_i) = \frac{\sum_{j=1}^{k} ||x_i - c_j||^{-p-2}}{(\sum_{j=1}^{k} ||x_i - c_j||^{-p})^2}. \tag{8}$$

5 For each cluster, calculate its new centre location:

$$c_j = \frac{\sum_{i=1}^{n} m(c_j, x_i) \times w(x_i) \times x_i}{\sum_{i=1}^{n} m(c_j, x_i) \times w(x_i)}. \tag{9}$$

6 Repeat steps 2-5 a pre-specified amount of times or until $\text{KHM}(X, C)$ converges. Assign all x_i to the c_j with the shortest Euclidean distance, or, equivalently, for which the membership $m(c_j, x_i)$ is highest.

Fuzzy K-Means, or Fuzzy C-Means, is a soft clustering method, just like K-Harmonic Means. Fuzzy K-Means was developed by Dunn (1973) and improved by Bezdek (1981). Fuzzy K-Means works very similar to standard K-Means, but has a different objective function, namely:

$$\text{FKM}(X, C) = \sum_{j=1}^{k} \sum_{i=1}^{n} u_{ij}^{m} ||x_i - c_j||^2, \tag{10}$$

where

$$u_{ij} = \frac{1}{\sum_{l=1}^{k} \left(\frac{||x_i - c_j||}{||x_i - c_l||} \right)^{\frac{2}{m-1}}} \tag{11}$$

for $u_{ij} \in [0, 1]$ is the membership degree and $m \in \mathbb{R}, m \geq 1$. Here, m is the fuzzifier and determines the level of fuzziness. The higher m, the fuzzier the cluster, meaning that a lot of objects could potentially belong to multiple clusters. I find the optimal value of m by taking 196 points in the interval $[1.05, 3]$. As Fuzzy K-Means depends on a seed for its initialisation, I calculate which m gives, after a certain amount of iterations, the highest average silhouette width. The algorithm for Fuzzy K-Means is given in Algorithm 5.

Algorithm 5: Fuzzy K-Means clustering algorithm

1. Randomly specify k initial centres.
2. For each object, assign its membership degrees for the different clusters randomly.
3. For each cluster, calculate its centre as follows:

$$c_j = \frac{\sum_i u_{ij}^m \times x_i}{\sum_i u_{ij}^m}, \tag{12}$$

   where u_{ij} denotes the degree of membership for object x_i to cluster j.
4. For each object, calculate its new membership degrees for the different clusters using Equation (11).
5. Repeat steps 3-4 a pre-specified amount of times or until $\text{FKM}(X, C)$ converges. Assign all x_i to c_j with the largest u_{ij}.

The final method I discuss is a combination of K-Means++ and K-Harmonic Means. Hamerly and Elkan (2002) have done something similar, by combining standard K-Means and K-Harmonic Means. K-Harmonic Means++ uses the same cluster initialisation as K-Means++ but then tries to optimise the objective function for K-Harmonic Means as shown in Equation (6). I use the same p's as used for K-Harmonic Means. The steps for this method are shown in Algorithm 6.

Algorithm 6: K-Harmonic Means++ clustering algorithm

1 Choose centre c_1 uniformly at random from X.

2 For each object, compute $D(x)$, which is the Euclidean distance to the nearest centre that already exists.

3 Choose new centre c_i, choosing $x_i \in X$ with probability $\frac{D(x)^2}{\sum_{x_i \in X} D(x)^2}$.

4 Repeat step 2 until there are k centres.

5 Proceed with steps 2-6 from Algorithm 4.

4.3 Method Evaluation

Finally, I want to evaluate the clustering performance of the methods that have been used. To do this, I look at both the internal clustering validity for those methods and their dependence on their random initialisation. The internal clustering validity will be discussed in Section 4.3.1 and the dependence on random initialisation in Section 4.3.2. As stated in Section 4.2, a major flaw of the K-Means method is that it does not always lead to a global optimum, as it highly depends on the random initialisation of the centres. Therefore, I run the algorithms for the different methods multiple times for different initial seeds before assessing their performance, to obtain reliable results.

4.3.1 Internal Clustering Validity

I evaluate the internal clustering validity of the different clustering methods in two ways. First, I calculate the average silhouette width, as given in Equation (4), for all the clustering methods. Second, I consider the Davies-Bouldin (DB) Index, as proposed by Davies and Bouldin (1979). Just like the average silhouette width, the DB Index is an internal clustering validity measure: the results are evaluated based on the internal information and do not rely on external data. The DB Index can be formulated as follows:

$$\mathrm{DB}(X, C) = \frac{1}{k} \sum_{j=1}^{k} \max_{1 \leq j' \leq k, j \neq j'} \left(\frac{s_j + s_{j'}}{d_{jj'}} \right), \tag{13}$$

where $s_j = \frac{1}{||c_j||} \sum_{i:c(i)=j} d(x_i, c_j)$, $||c_j||$ is the number of objects in cluster j, $d_{jj'} = d(c_j, c'_j)$ and $d(\mathbf{x}, \mathbf{y})$ is the Euclidean distance between x and y. Low DB Index values are preferred over high values.

I run the algorithms for the different methods 100 times for different initial seeds and take the average silhouette width and DB Index value to evaluate their internal clustering validity.

4.3.2 Initialisation Dependence

I evaluate to what extent different clustering methods depend on their centre initialisation. For every time I run the algorithm for that method, I compute its objective function, or, in case of Fuzzy K-Means, the objective function for standard K-Means. I group the five different clustering methods in two groups:

Group 1: K-Means, K-Means++ and Fuzzy K-Means, and

Group 2: K-Harmonic Means and K-Harmonic Means++.

These groupings are made because the methods they contain use different minimisation criteria, and thus cannot be compared. For the objective functions of K-Harmonic Means and K-Harmonic Means++, as shown in Equation (6), I calculate the centres as the arithmetic means of the members of the clusters and do not take weights and memberships into account. I assume that, after running the algorithm for every method a certain amount of times, the optimal clustering for that method must have been reached at least once, for both groups. This has been supported by Fränti and Sieranoja (2019). Thus, I evaluate the initialisation dependence of the different methods as follows:

$$\text{ID} = \frac{\text{Amount of times that global minimum is reached}}{\text{Amount of iterations}}, \tag{14}$$

where I assume that the lowest local minimum found for each group is equal to the global minimum. As the benchmark, hierarchical clustering using Ward's procedure, does not have its own objective function, it is excluded from this measure.

5 Results

In this section, I present and discuss the results for the different clustering methods. The results are obtained using R by the R Core Team (2020). Programming code can be provided upon request. A short description of the code files can be found in Appendix B. First, I discuss the application of the data on European energy dependence by Eurostat (2015) on the hierarchical benchmark method and the different non-hierarchical methods in Section 5.1. As the data set I consider is very limited, I apply the different methods as well to some test data sets to be able to draw stronger conclusions. Their clustering results are briefly discussed in Section 5.2.

5.1 Energy Dependence Data Set

First, I specifically look at the clustering of the data on the energy dependence of the 27 EU member states plus the United Kingdom, as discussed in Section 3. I evaluate the clustering performance of hierarchical clustering using Ward's procedure for the data set on energy dependence. This method serves both as a benchmark for the different K-Means variations and is used to determine the optimal number of clusters. A dendrogram for this clustering can be found in the appendix, Figure A1. In this figure, the numbers correspond with the numbers given to the countries in Table A1.

Before I determine the optimal number of clusters for this data set, I examine an elbow plot, as shown in Figure 2, which describes the explained variation as a function of the number of clusters, for clusters constructed by hierarchical clustering. The elbow plot was introduced by Thorndike (1953). The vertical axis shows the Total Within Sum of Squares, which is the sum of the squared distances between the objects and their cluster centre, or in other words, the variation, when the data is divided into a certain amount of clusters. The intuition behind the elbow plot is that the 'elbow' in the figure denotes for how many clusters the returns of an extra cluster are no longer worth the costs, due to overfitting of the data. At first, a big part of the variation in the data can be explained by adding more clusters, but at some point, the new information an additional cluster gives, decreases sharply. The 'elbow' seems to appear at three or four clusters, but I check this using the average silhouette method, as described in Section 4.1. As can be seen in Figure 3, the average silhouette width is indeed highest for four clusters.

Bluszcz (2016) uses seven clusters in total, so I also evaluate the different K-Means variations for seven clusters, to be able to judge whether I can find a better clustering solution then Bluszcz (2016) in that case.

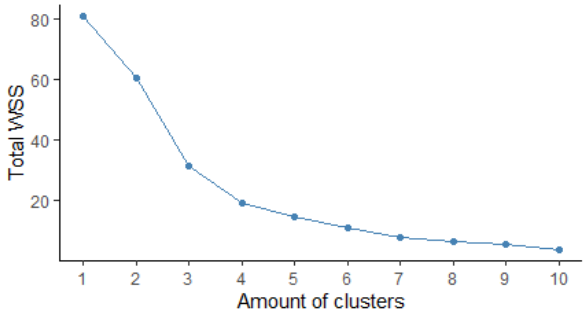

Figure 2: Elbow plot for EU data on energy dependence

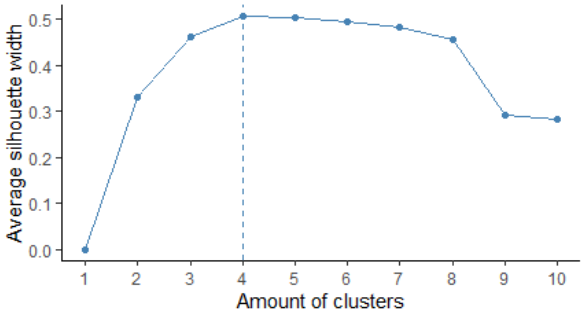

Figure 3: Silhouette plot for EU data on energy dependence

The first two principal components of the data set are depicted again in Figures 4 and 5. The clusters according to Ward's procedure are shown.

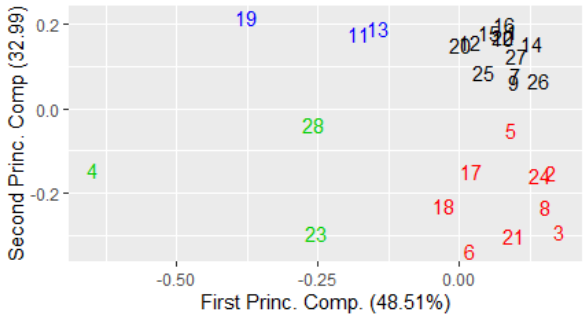

Figure 4: Four clusters

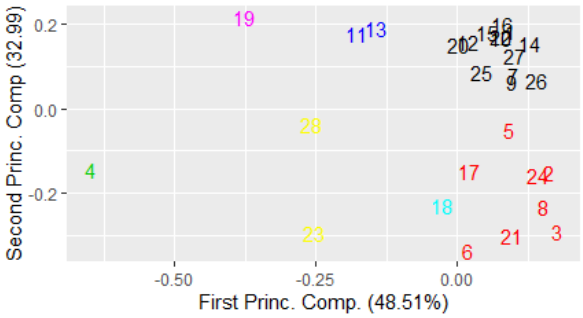

Figure 5: Seven clusters

The values for the different evaluation measures discussed in Sections 4.3.1 and 4.3.2 are shown in Tables 2 and 3 for four and seven clusters respectively.

Table 2: Evaluation results four clusters

	Ward	K-Means	K-Means++	K-Harmonic Means	Fuzzy K-Means	K-Harmonic Means++
Silhouette Width	0.5064	0.4727	**0.5064**	**0.5064**	**0.5064**	**0.5064**
DB Index	0.7519	0.7908	**0.7519**	**0.7519**	**0.7519**	**0.7519**
Initialisation Dependence		0.70	**1.00**	**1.00**	**1.00**	**1.00**

Best results (for K-Means variations) are bold

In the case of four clusters, which was suggested by the average silhouette method, standard K-Means is outperformed by all its variations, even by hierarchical clustering using Ward's procedure. Remarkable is that hierarchical clustering and the four variations all come to the same average values for the evaluation measures. This might be because the data set only consists of 28 objects. Also, a visual inspection of Figure 4 suggests that the distinction between the four clusters is quite clear. For 100 different initialisations, all K-Means variations reach the same objective function in 100% of the cases, while standard K-Means only realises this optimum in 70% of the cases. Clearly, for this data set, standard K-Means depends more on its initialisation than the other methods.

Conducting a t-test on both the average silhouette widths and the DB Indices of K-Means and K-Harmonic Means also indicates that the different K-Means variations significantly perform better than standard K-Means, at a 1% confidence level. An assumption for comparing the means of two samples using a t-test is that they should follow a normal distribution, but, due to the central limit

theorem, under weak assumptions, this is not needed for large samples (Lumley, Diehr, Emerson, and Chen, 2002).

Figure 6 shows the optimal clustering of the 28 countries. This clustering is also listed in Table A3, which includes for each country the Euclidean distance to the centre of its cluster. Figure 6 suggests that most of the countries in Cluster 1 (dark purple) are situated in Western Europe. The countries in this cluster are generally characterised by high levels of energy dependence, especially for petroleum and gas. Most of the countries in Cluster 2 (light purple) are in Eastern Europe. They are generally characterised by a low, or even negative, oil dependence. There does not seem to be any relation between the countries in Clusters 3 and 4 (light green and dark green respectively).

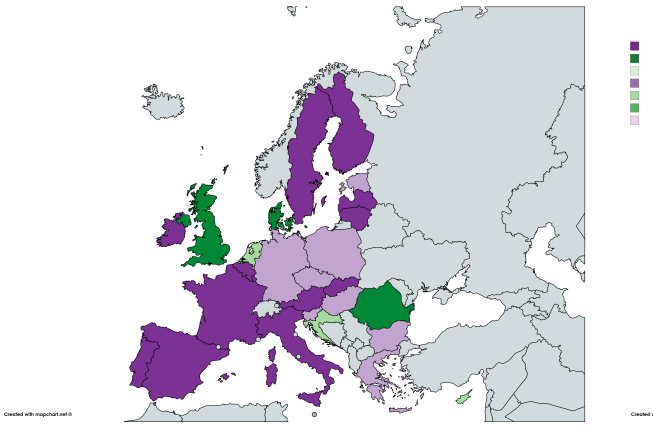

Figure 6: Map of the EU plus the United Kingdom divided into four clusters - Created using MapChart (2020)

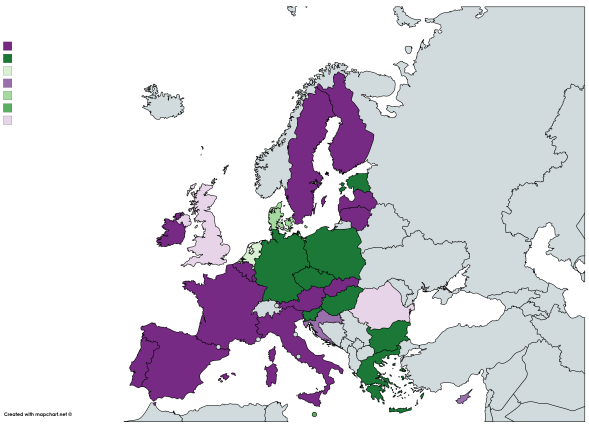

Figure 7: Map of the EU plus the United Kingdom divided into seven clusters - Created using MapChart (2020)

Table 3: Evaluation results seven clusters

	Ward	K-Means	K-Means++	K-Harmonic Means	Fuzzy K-Means	K-Harmonic Means++
Silhouette Width	0.4839	0.3253	**0.4807**	0.4775	0.3738	0.4753
DB Index	0.4847	0.7251	0.4939	**0.4925**	0.6597	0.4986
Initialisation Dependence		0.04	**0.84**	0.32	0.05	0.30

Best results (for K-Means variations) are bold

When clustering the data set into seven clusters, imitating Bluszcz (2016), standard K-Means still seems to be the worst available option. This time, however, there are differences in the performances of the other options. The average silhouette width is highest for K-Means++, followed by K-Harmonic Means and K-Harmonic Means++ respectively. K-Harmonic Means scores best for

the DB index, followed by K-Means++ and K-Harmonic Means++. The performance of Fuzzy K-Means, however better than that of standard K-Means, is also clearly worse than the other three non-hierarchical options. In the case of seven clusters, none of the non-hierarchical methods can come to the same objective function for every iteration with different initialisations. K-Means++ reaches the optimal objective function value for Group 1 in 84% of the cases, followed by K-Harmonic Means and K-Harmonic Means++, that reach the optimal objective function value for Group 2 in only 32% and 30% of the cases respectively. Similar to clustering into four data sets, hierarchical clustering using Ward's procedure performs at least as well as the non-hierarchical K-Means variations.

To confirm the significance of these results, I conduct t-tests on both the average silhouette widths and the DB Indices of any combination of K-Means variations leads to the results in Tables 4 and 5 respectively. These tables show the p-values resulting from a t-test with the null hypothesis that the method in the header column outperforms the method in the header row. I select a 5% confidence level. For example, in Table 4, the null hypothesis that standard K-Means outperforms any of its variations is rejected. Remarkable for the average silhouette width in Table 4, is that the null hypothesis that K-Harmonic Means++ outperforms K-Harmonic Means is rejected, indicating that, for this data set, augmenting K-Harmonic Means is not beneficial. The null hypothesis that K-Means++, which has the highest average silhouette width according to Table 2, outperforms any of the other four methods is not rejected.

Table 4: P-Values based on average silhouette width

	K-Means	K-Means++	K-Harmonic Means	Fuzzy K-Means	K-Harmonic Means++
K-Means		0.0000	0.0000	0.0000	0.0000
K-Means++	1.0000		0.9904	1.0000	0.9999
K-Harmonic Means	1.0000	0.0096		1.0000	0.9790
Fuzzy K-Means	1.0000	0.0000	0.0000		0.0000
K-Harmonic Means++	1.0000	0.0001	0.0210	1.0000	

Table 5: P-Values based on DB Index

	K-Means	K-Means++	K-Harmonic Means	Fuzzy K-Means	K-Harmonic Means++
K-Means		0.0000	0.0000	0.0001	0.0000
K-Means++	1.0000		0.3401	1.0000	0.8912
K-Harmonic Means	1.0000	0.6599		1.0000	0.9920
Fuzzy K-Means	0.9999	0.0000	0.0000		0.0000
K-Harmonic Means++	1.0000	0.1088	0.0080	1.0000	

When looking at Table 5 for the DB Index, it is noteworthy that the null hypothesis that K-Harmonic Means or K-Harmonic Means++ outperforms K-Means++ or the other way around is not rejected. It is again rejected that K-Harmonic Means++ outperforms K-Harmonic Means.

Bluszcz (2016) finds a clustering solution by running standard K-Means just once. This leads to an average silhouette width of 0.4683 and a DB Index of 0.5253. For certain seeds, I find the same clustering solution as Bluszcz (2016). Clearly, this solution is better than the average performance of K-Means and Fuzzy K-Means, as shown in Table 3. However, as Table 3 only depicts the average performance, individual K-Means or Fuzzy K-Means solutions might still be superior. Moreover, even the average performances of K-Means++, K-Harmonic Means and K-Harmonic Means++ outperform the solution by Bluszcz (2016).

Figure 7 shows the optimal clustering of the 28 countries, in case there are seven different clusters. Details on the clustering can be found in Table A4, including the Euclidean distance for each country to the centre of its cluster. As was also true for the case of four clusters, there seems to be a divide between Western and Eastern Europe: Cluster 3 seems to be mainly comprised of Western European countries, Cluster 5 of Eastern European countries.

To summarise, for this data set on the energy dependence of the European Union, it is best to divide the EU member states into four groups. However, in that case, K-Means++, K-Harmonic Means, Fuzzy K-Means and K-Harmonic Means++ all perform equally well. Therefore, I apply all K-Means variations to seven other data sets with different properties, to be able to decide which method is superior. These data sets and their evaluation are discussed in the next section.

5.2 Test Sets

To further evaluate the performance of the different methods I compare in this paper, I apply them to seven built-in data sets in R. I use the *mtcars* data set from the package by the R Core Team (2020), the *iris*, the *Ionosphere*, the *Glass* and the *BostonHousing* data sets from the package by Newman et al. (1998) and the *birth.death.rates.1966* and *acidosis.patients* data sets from the package by Novomestky (2013). These different sets will from now on be referred to as Test Sets 1-7. For each data set, I determine the number of clusters by determining what number gives the highest average silhouette width for hierarchical clustering, as described in Section 4.1. If this results in three or more methods getting equal, optimal results, I change the number of clusters to the second-best number. Then, I apply the different methods to the sets and evaluate them using the performance techniques discussed in Section 4.3.

Table A5 shows some characteristics of the used sets. As can be seen, the sets vary greatly in the number of objects and variables. The first two principal components of each data set are depicted in Figures A2 - A7. The figures also contain the clusters according to hierarchical clustering.

The evaluation results for the test sets are displayed in Tables A6 - A12. I will shortly discuss their internal clustering validity and initialisation dependence.

There are seven different test sets, so, in total, there are 14 different internal clustering indices. For the non-hierarchical methods, K-Harmonic Means++ performs best in seven out of 14 cases, followed by K-Means++, which performs best in six out of 14 cases, and K-Harmonic Means, for four cases. Fuzzy K-Means performs best in three cases, and standard K-Means never performs best for these data sets. In Test Set 4, K-Harmonic Means and K-Harmonic Means++ have the same results, just like the three remaining non-hierarchical methods between themselves, suggesting that this data set is not sensitive to its initialisation. In Test Set 6, K-Harmonic Means and K-Harmonic Means++ again perform equally well, just like K-Means++. The benchmark method, hierarchical clustering using Ward's procedure, outperforms any of the non-hierarchical methods in terms of both average silhouette width and DB Index only for Test Set 1. This might be related to the low number of objects in this data set, which is only 32. This was also true for the hierarchical clustering of the EU data set into four clusters. A last remarkable feature of the evaluation results is that for Test Sets 3 and 6, the indices for the preferred methods are clearly superior to the indices for the worst methods. A possible explanation for this is that upon visual inspection of Figures A4 and A8, the clusters for these data sets seem to have more outliers than those for other data sets.

For the different non-hierarchical methods, K-Means++ has the highest value for the Initialisation Dependence for six out of seven cases, followed by K-Harmonic Means++, having the highest value for three out of seven cases, and K-Harmonic Means, which performs best in two cases. Standard K-Means and Fuzzy K-Means are both only preferred for one out of seven cases. For Test Set 4, all non-hierarchical methods come to a global optimum for all of the 100 iterations, reinforcing the idea that this data set is not sensitive to its initialisation. For Test Sets 6 and 7, K-Means++ also comes to a global minimum for 100% of the iterations. Remarkable is that for Test Set 1, the maximum initialisation dependence is only 20%, namely for K-Harmonic Means++. This might be connected to the fact that this data set only contains 32 objects, but is still divided into nine clusters.

6 Conclusion

This thesis aimed to answer the following question: 'In what ways can the standard K-Means clustering method be improved?', with the sub-questions 'What are appropriate ways to improve the internal clustering validity of the standard K-Means clustering method?' and 'What are appropriate ways to increase the probability that the standard K-Means clustering method reaches a global optimum?' I discussed five different methods, namely, standard K-Means, K-Means++, K-Harmonic Means, Fuzzy K-Means and K-Harmonic Means++. I appointed hierarchical clustering using Ward's procedure as a benchmark. Using eight different data sets, specifically focusing on a data set on energy dependence in the EU, I formulated this answer.

First, I evaluated the internal clustering validity of the different methods, according to their average silhouette width and DB Index. I found that K-Means++ and K-Harmonic Means++, in general, give the best results. K-Means++ augments standard K-Means with an intelligent cluster initialisation. K-Harmonic Means++ is a soft clustering method (meaning that objects belong to multiple clusters to a certain degree), that adopts the same cluster initialisation and, in addition, uses the harmonic mean of the distances from each object to their centre. They are followed by K-Harmonic Means, which is similar to K-Harmonic Means++ but has a random initialisation. Fuzzy K-Means performs better than standard K-Means, however, not by much. Except for the sets with a low number of objects (less than 35), hierarchical clustering using Ward's procedure is generally outperformed by any of the other methods.

Second, I discussed the initialisation dependence of each method, meaning, to what extent the outcome of the clustering depends on the initial clusters. K-Means++ substantially outperforms all other methods for this criterion, followed by K-Harmonic Means++ and K-Harmonic Means. Fuzzy K-Means, although outperforming standard K-Means, does not seem to be an adequate alternative.

To answer the research question: augmenting standard K-Means by first finding a proper cluster initialisation, such as K-Means++ does, consistently improves results. Substituting the standard mean for the harmonic mean, as done by K-Harmonic Means and K-Harmonic Means++, also seems an effective way to augment the standard method, but these methods contain a free parameter which must be estimated, which can lead to substantially higher computation time. This can be solved by fixing this parameter, but this would impair the results. In the future, possibly more intelligent ways of initialising K-Means could be developed, and K-Means with random initialisation might eventually be rendered unnecessary.

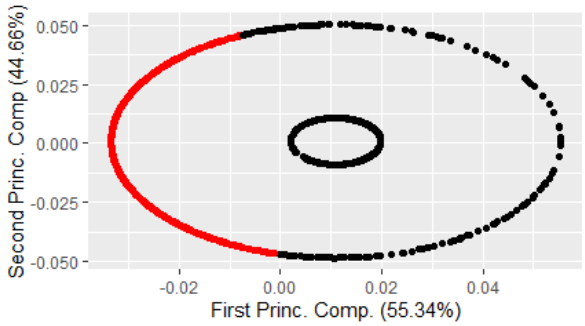

Figure 8: Failure for K-Means: circular data - based on code by Morbieu (2018)

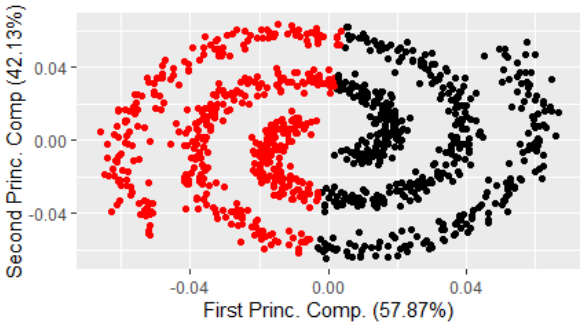

Figure 9: Failure for K-Means: spiral data - based on code by Morbieu (2018)

Of course, for certain data types, other variations than K-Means++ might be more suitable. Moreover, often any K-Means variation, despite greatly improving standard K-Means, might not be the best option for a data set. For example, K-Means variations use the (harmonic) mean as cluster centre, which can lead to complications, as can be seen in Figures 8 and 9, where clustering is done using the K-Means++ method. Clearly, these clusters are wrongly specified. Such issues could be solved by Mean-Shift Clustering (Fukunaga and Hostetler, 1975), DBSCAN (Martin et al., 1996) or Expectation-Maximisation Clustering using Gaussian Mixture Models (Dempster, Laird, and Rubin, 1977). In the future, it could be interesting to research which data types require which clustering methods.

References

D. Arthur and S. Vassilvitskii. K-Means++: The Advantages of Careful Seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1027–1035, Philadelphia, PA, USA, 2007. Society for Industrial and Applied Mathematics.

G. H. Ball and D. J. Hall. A Clustering Technique for Summarizing Multivariate Data. *Behavioral Science*, 12:153–155, 1967.

J. C. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Kluwer Academic Publishers, 1981.

A. Bluszcz. European Economies in Terms of Energy Dependence. *Quality & Quantity*, 51:1531 – 1548, 2016.

P. S. Bradley and U. M. Fayyad. Refining Initial Points for K-Means Clustering. In *Proceedings of the Fifteenth International Conference on Machine Learning*, ICML '98, page 91–99, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc.

R. B. Cattell. The Description of Personality: Basic Traits Resolved Into Clusters. *Journal of Abnormal and Social Psychology*, 38(4):476–506, 1943.

J. M. Chevalier. Security of Energy Supply for the European Union. *European Review of Energy Markets*, 1 (3):1–20, 2005.

Commission of the European Communities. *A European Strategy for Sustainable, Competitive and Secure Energy*. European Commission, 2006.

I. G. Costa, F. A. T. D. Carvalho, and M. C. P. D. Souto. Comparative Analysis of Clustering Methods for Gene Expression Time Course Data. *Genetics and Molecular Biology*, 27(4):623–631, 2004.

D. L. Davies and D. W. Bouldin. A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1:224 – 227, 1979.

M. C. De Souto, I. G. Costa, D. S. D. Araujo, T. B. Ludermir, and A. Schliep. Clustering Cancer Gene Expression Data: A Comparative Study. *BMC Bioinformatics*, 9, 2008.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38, 1977.

Directorate-General for Economic and Financial Affairs. European Economy Member States' Energy Dependence: An Indicator-Based Assessment, 2013.

E. R. Dougherty, J. Barrera, and M. Brun. Inference From Clustering with Application to Gene-Expression Microarrays. *Journal of Computational Biology*, 9(1):105–126, 2002.

H. E. Driver and A. L. Kroeber. *Quantitative Expression of Cultural Relationships*. University of California Press: California, USA, 1932.

J. C. Dunn. A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. *Journal of Cybernetics*, 3(3):32–57, 1973.

V. Estivill-Castro and J. Yang. Fast and Robust General Purpose Clustering Algorithms. *Data Mining and Knowledge Discovery*, 8:127–150, 2004.

Eurostat. *Energy, transport and environment indicators*. European Commission, 2015.

P. Fränti and S. Sieranoja. How Much Can K-Means Be Improved by Using Better Initialization and Repeats? *Pattern Recognition*, 93:95–112, 2019.

K. Fukunaga and L. D. Hostetler. The Estimation of the Gradient of a Density Function, with Applications in Pattern Recognition. *IEEE Transactions on Information Theory*, 21(1):32–40, 1975.

G. Hamerly and C. Elkan. Alternatives to the K-Means Algorithm That Find Better Clusterings. In *Proceedings of the Eleventh International Conference on Information and Knowledge Management*, pages 600–607, New York, NY, USA, 2002. ACM.

J. A. Hartigan and M. A. Wong. A K-Means Clustering Algorithm. *JSTOR: Applied Statistics*, 28(1): 100–108, 1979.

M. Horuckova. Growing Energy Dependence: A Security Risk for the EU. *Proceedings of the Third International Conference on European Integration 2016*, pages 359–367, 2016.

H. Jiang, S. Yi, J. Li, F. Yang, and X. Hu. Ant Clustering Algorithm with K-Harmonic Means Clustering. *Expert Systems with Applications*, 37(12):8679–8684, 2010.

R. Johnson and D. Wichern. *Applied Multivariate Statistical Analysis*. Pearson Education, Inc.: London, UK, 2007.

L. Kaufman and P. J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley-Interscience: Hoboken, NJ, USA, 1990.

A. Likas, N. Vlassis, and J. J. Verbeek. The Global K-Means Clustering Algorithm. *Pattern Recognition*, 36 (2):451–461, 2003.

T. Lumley, P. Diehr, S. Emerson, and L. Chen. The Importance of the Normality Assumption in Large Public Health Data Sets. *Annual Review of Public Health*, 23(1):151–169, 2002.

J. B. MacQueen. Some Methods for Classification and Analysis of Multivariate Observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, Berkeley, CA, USA, 1967. University of California Press.

N. K. Malhotra, D. Nunan, and D. F. Birks. *Marketing Research: An Applied Approach*. Pearson Education, Inc.: London, UK, 2017.

MapChart, 2020. URL `https://mapchart.net`.

E. Martin, H. P. Kriegel, J. Sander, and X. Xu. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, pages 226–231, Portland, OR, USA, 1996.

U. Maulik and S. Bandyopadhyay. Performance Evaluation of Some Clustering Algorithms and Validity Indices. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(12):1650–1654, 2002.

G. W. Milligan and M. C. Cooper. A Study of Standardization of Variables in Cluster Analysis. *Journal of Classification*, 5:181–204, 1988.

S. Morbieu. Generate Datasets to Understand Some Clustering Algorithms Behavior, 2018.

D. Newman, S. Hettich, C. Blake, and C. Merz. UCI Repository of Machine Learning Databases, 1998. URL `http://www.ics.uci.edu/~mlearn/MLRepository.html`.

F. Novomestky. *cluster.datasets: Cluster Analysis Data Sets*, 2013. URL `https://CRAN.R-project.org/package=cluster.datasets`. R package version 1.0-1.

J. M. Peña, J. A. Lozano, and P. Larrañaga. An Empirical Comparison of Four Initialization Methods for the K-Means Algorithm. *Pattern Recognition Letters*, 20(10):1027–1040, 1999.

D. T. Pham, S. S. Dimov, and C. D. Nguyen. Selection of K in K-Means Clustering. *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, 219(1), 2005.

G. Punj and D. W. Stewart. Cluster Analysis in Marketing Research: Review and Suggestions for Application. *Journal of Marketing Research*, 20:134–148, 01 1983.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020. URL `https://www.R-project.org`.

S. J. Redmond and C. Heneghan. A Method for Initialising the K-Means Clustering Algorithm Using K-D Trees. *Pattern Recognition Letters*, 28(8):965–973, June 2007.

H. Steinhaus. Sur la Division des Corps Matériels en Parties. *Bulletin L'Académie Polonaise des Science*, 4 (12):801–804, 1957.

R. L. Thorndike. Who Belongs in the Family? *Psychometrika*, 18(4):267–276, 1953.

R. C. Tryon. *Cluster Analysis: Correlation Profile and Orthometric Analysis for the Isolation of Unities in Mind and Personality*. Edwards Brothers, 1939.

J. H. Ward. Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, 58:236–244, 1963.

R. Xu and D. C. Wunsch. Survey of Clustering Algorithms. *IEEE Transactions on Neural Networks*, 16(3): 645–678, May 2005.

B. Zhang, M. Hsu, and U. Dayal. K-Harmonic Means - A Data Clustering Algorithm. Technical report, HP Labaratories Palo Alto, 1999.

J. Zubin. A Technique for Measuring Like-Mindedness. *The Journal of Abnormal and Social Psychology*, 33 (4):508–516, 1938.

Appendix A

Table A1: Data by Eurostat (2015)

Country	Fuels	Petroleum	Gas
Belgium	95.1	102	100.5
Bulgaria	16.4	103.7	93.2
Czech Republic	-11.6	96.3	100.2
Denmark	90.7	-13.7	-23.1
Germany	44.5	96.1	87.2
Estonia	-0.1	59.9	100
Ireland	72.4	100.2	95.9
Greece	3.2	94.2	100
Spain	70.3	97.4	98.6
France	93.4	98.9	97.4
Croatia	110.1	77.1	31.8
Italy	96.2	90.7	88.1
Cyprus	100	101	0
Latvia	88.8	100.4	115.6
Lithuania	99.7	93.2	100
Luxembourg	100	100.3	99.6
Hungary	29.5	83.9	72.1
Malta	0	104.6	0
Netherlands	111.6	94.7	-86.8
Austria	93.8	92.9	75.5
Poland	-10.4	91.3	74.2
Portugal	95.4	97.2	101.5
Romania	18.9	47	11.9
Slovenia	19.4	95.8	99.6
Slovakia	80.6	88.5	95.6
Finland	65.7	106.2	99.9
Sweden	82.4	101.5	99.1
United Kingdom	82	39.8	50.1

Table A2: Data from Table A1 standardised

Country	Fuels	Petroleum	Gas
Belgium	0.7977	0.5780	0.6159
Bulgaria	-1.1031	0.6443	0.4654
Czech Republic	-1.7793	0.3556	0.6097
Denmark	0.6914	-3.9348	-1.9329
Germany	-0.4244	0.3478	0.3417
Estonia	-1.5016	-1.0641	0.6056
Ireland	0.2495	0.5077	0.5211
Greece	-1.4219	0.2737	0.6056
Spain	0.1987	0.3985	0.5768
France	0.7567	0.4570	0.5520
Croatia	1.1600	-0.3932	-0.8008
Italy	0.8243	0.1372	0.3602
Cyprus	0.9161	0.5389	-1.4566
Latvia	0.6456	0.5155	0.9273
Lithuania	0.9088	0.2347	0.6056
Luxembourg	0.9161	0.5116	0.5974
Hungary	-0.7867	-0.1280	0.0303
Malta	-1.4992	0.6794	-1.4566
Netherlands	1.1962	0.2932	-3.2466
Austria	0.7663	0.2230	0.1004
Poland	-1.7503	0.1606	0.0736
Portugal	0.8050	0.3907	0.6366
Romania	-1.0427	-1.5673	-1.2112
Slovenia	-1.0306	0.3361	0.5974
Slovakia	0.4475	0.0514	0.5149
Finland	0.0876	0.7418	0.6036
Sweden	0.4910	0.5585	0.5871
United Kingdom	0.4813	-1.8481	-0.4234

Table A3: Optimal clustering for four clusters

Country	**Distance from centroid**
Cluster 1	
Belgium	0.262618
Ireland	0.372863
Spain	0.409301
France	0.157183
Italy	0.397420
Latvia	0.390590
Lithuania	0.351688
Luxembourg	0.328572
Austria	0.514880
Portugal	0.215024
Slovakia	0.392808
Finland	0.619531
Sweden	0.192951
Cluster 2	
Bulgaria	0.553540
Czech Republic	0.683695
Germany	0.858394
Estonia	1.327595
Greece	0.441481
Hungary	0.587396
Malta	1.755438
Poland	0.513369
Slovenia	0.476329
Cluster 3	
Croatia	1.168228
Cyprus	0.572383
The Netherlands	1.423456
Cluster 4	
Denmark	1.782601
Romania	1.399741
United Kingdom	1.067968

Most typical countries for each cluster (with minimum distance to its centroid) are italicised

Table A4: Optimal clustering for seven clusters

Country	Distance from centroid
Cluster 1	
Denmark	0
Cluster 2	
Croatia	0.582780
Cyprus	0.582780
Cluster 3	
Belgium	0.262618
Ireland	0.372863
Spain	0.409301
France	0.157183
Italy	0.397420
Latvia	0.390590
Lithuania	0.351688
Luxembourg	0.328572
Austria	0.514880
Portugal	0.215024
Slovakia	0.392808
Finland	0.619531
Sweden	0.192951
Cluster 4	
Netherlands	0
Cluster 5	
Bulgaria	0.544556
Czech Republic	0.634499
Germany	0.836631
Estonia	1.226630
Greece	0.315776
Hungary	0.632639
Poland	0.629000
Slovenia	0.345089
Cluster 6	
Malta	0
Cluster 7	
Romania	0.869195
United Kingdom	0.869195

Most typical countries for each cluster (with minimum distance to its centroid) are italicised

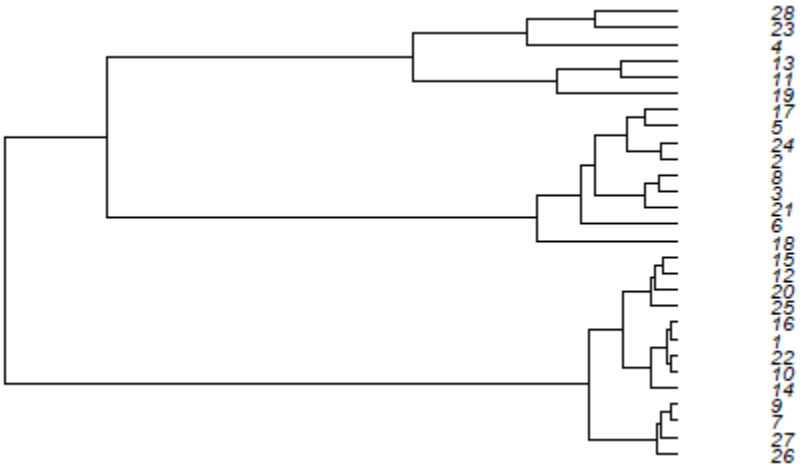

Figure A1: Dendrogram hierarchical clustering

Table A5: Statistics for Test Sets 1-7

	Objects	**Variables**	**Clusters**
mtcars	32	11	9
iris	150	4	4
Glass	214	9	2
BostonHousing	506	13	2
Ionosphere	351	32	4
birth.death.rates.1966	70	2	4
acidosis.patients	40	6	5

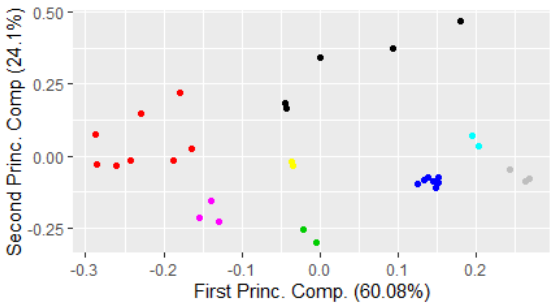

Figure A2: Clustering *mtcars*

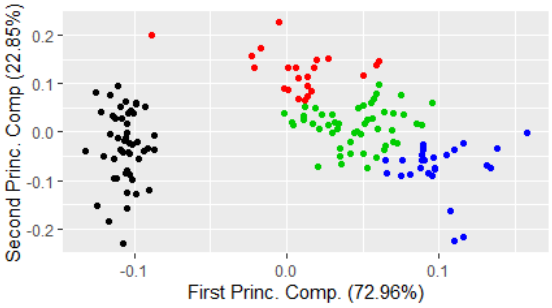

Figure A3: Clustering *iris*

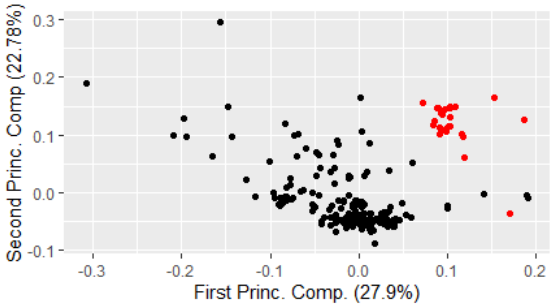

Figure A4: Clustering *Glass*

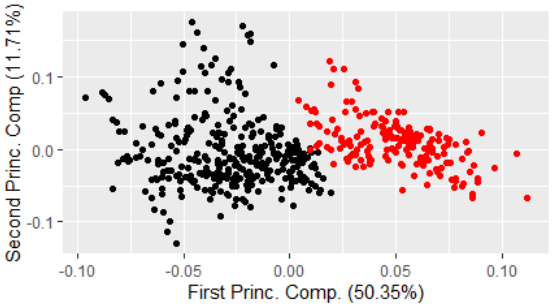

Figure A5: Clustering *BostonHousing*

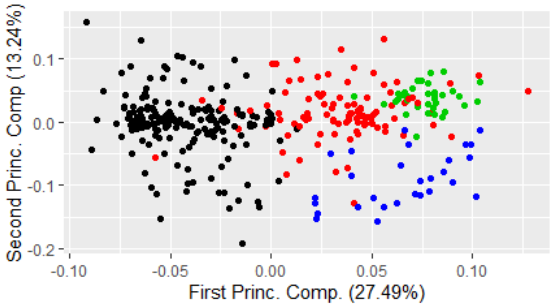

Figure A6: Clustering *Ionosphere*

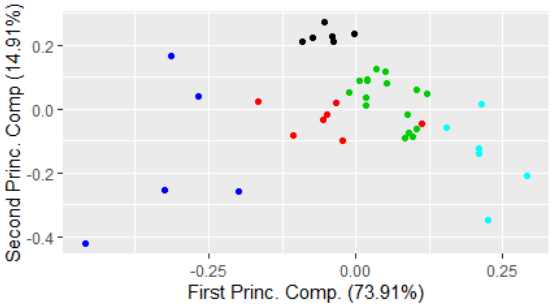

Figure A7: Clustering *birth.death.rates.1966*

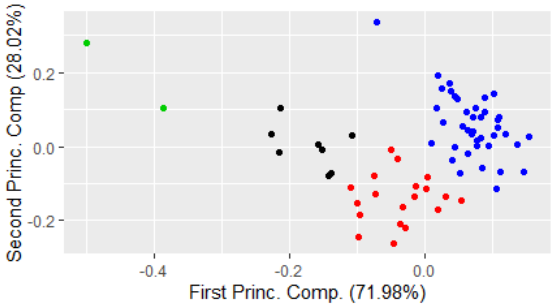

Figure A8: Clustering *acidosis.patients*

Table A6: Evaluation results *mtcars*

Set 1	Ward	K-Means	K-Means++	K-Harmonic Means	Fuzzy K-Means	K-Harmonic Means++
Silhouette Width	0.4644	0.3894	**0.4315**	0.4117	0.4153	0.4065
DB Index	0.6649	0.7884	0.7645	0.8009	**0.7311**	0.7909
Initialisation Dependence		0.03	0.16	0.14	0.16	**0.20**

Best results (for K-Means variations) are bold

Table A7: Evaluation results *iris*

Set 2	Ward	K-Means	K-Means++	K-Harmonic Means	Fuzzy K-Means	K-Harmonic Means++
Silhouette Width	0.4219	0.4033	0.3866	0.4027	0.4016	**0.4046**
DB Index	0.8982	0.8943	**0.8778**	0.8855	0.9122	0.8893
Initialisation Dependence		0.11	**0.62**	0.51	0.36	0.44

Best results (for K-Means variations) are bold

Table A8: Evaluation results *Glass*

Set 3	Ward	K-Means	K-Means++	K-Harmonic Means	Fuzzy K-Means	K-Harmonic Means++
Silhouette Width	0.4018	0.3786	0.4410	0.5975	0.4444	**0.5981**
DB Index	1.0571	1.5480	1.4031	1.2915	1.3951	**1.2858**
Initialisation Dependence		0.31	**0.92**	0.01	0.00	0.00

Best results (for K-Means variations) are bold

Table A9: Evaluation results *BostonHousing*

Set 4	Ward	K-Means	K-Means++	K-Harmonic Means	Fuzzy K-Means	K-Harmonic Means++
Silhouette Width	0.3647	0.3693	0.3693	**0.3701**	0.3693	**0.3701**
DB Index	1.1665	1.1510	1.1510	**1.1474**	1.1510	**1.1474**
Initialisation Dependence		**1.00**	**1.00**	**1.00**	**1.00**	**1.00**

Best results (for K-Means variations) are bold

Table A10: Evaluation results *Ionosphere*

Set 5	Ward	K-Means	K-Means++	K-Harmonic Means	Fuzzy K-Means	K-Harmonic Means++
Silhouette Width	0.2898	0.2567	**0.2935**	0.2743	0.2902	0.2735
DB Index	1.7224	1.7924	**1.6934**	2.0276	1.7266	2.0192
Initialisation Dependence		0.50	**1.00**	0.01	0.00	0.00

Best results (for K-Means variations) are bold

Table A11: Evaluation results *birth.death.rates.1966*

Set 6	Ward	K-Means	K-Means++	K-Harmonic Means	Fuzzy K-Means	K-Harmonic Means++
Silhouette Width	0.5023	0.3912	**0.5096**	**0.5096**	0.4336	**0.5096**
DB Index	0.6133	0.7780	**0.6824**	**0.6824**	0.7505	**0.6824**
Initialisation Dependence		0.13	**1.00**	**1.00**	0.49	**1.00**

Best results (for K-Means variations) are bold

Table A12: Evaluation results *acidosis.patients*

Set 7	Ward	K-Means	K-Means++	K-Harmonic Means	Fuzzy K-Means	K-Harmonic Means++
Silhouette Width	0.3220	0.2977	0.2820	0.2971	**0.3218**	0.2855
DB Index	1.0038	0.9165	0.8580	0.8869	**0.8556**	0.8909
Initialisation Dependence		0.08	**0.88**	0.01	0.00	0.00

Best results (for K-Means variations) are bold

Appendix B

Here, I give a short description of the different R code files used for this thesis.

- *MAIN*: installs packages; imports and prepares data on energy dependence
- *Functions*: gives functions I created for this thesis:
- *EnergyDependence*: EU data set - finds optimal amount of clusters; finds optimal values for p and m; conducts cluster analysis for both four and seven clusters; analyses results; compares results to Bluszcz (2016)
- *mtcars*: *mtcars* data set - finds optimal amount of clusters; finds optimal values for p and m; conducts cluster analysis; analyses results
- *iris*: *iris* data set - finds optimal amount of clusters; finds optimal values for p and m; conducts cluster analysis; analyses results
- *Glass*: *Glass* data set - finds optimal amount of clusters; finds optimal values for p and m; conducts cluster analysis; analyses results
- *BostonHousing*: *BostonHousing* data set - finds optimal amount of clusters; finds optimal values for p and m; conducts cluster analysis; analyses results
- *Ionosphere*: *Ionosphere* data set - finds optimal amount of clusters; finds optimal values for p and m; conducts cluster analysis; analyses results
- *birthdeathrates1966*: *birth.death.rates.1966* data set - finds optimal amount of clusters; finds optimal values for p and m; conducts cluster analysis; analyses results
- *acidosispatients*: *acidosis.patients* data set - finds optimal amount of clusters; finds optimal values for p and m; conducts cluster analysis; analyses results
- *Conclusion*: generates circular and spiral data