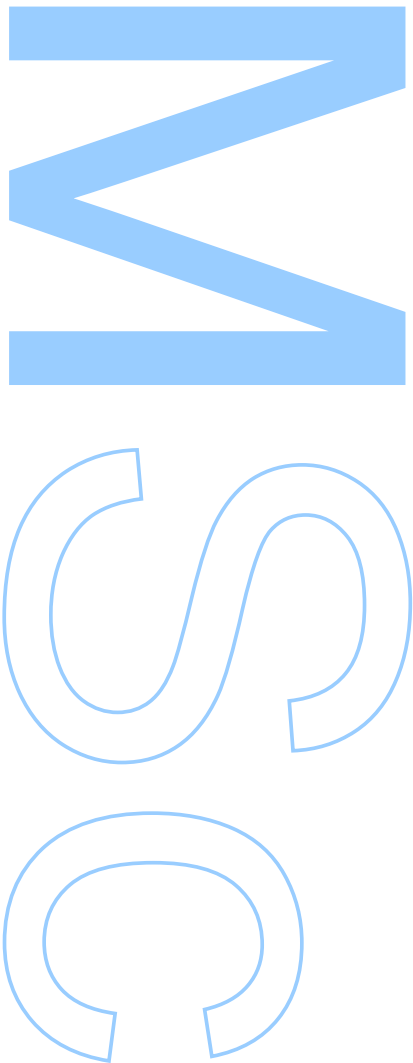
Clustering of longitudinal data: Application to COVID-19 data

André Alexandre da Silva Galvão Garcia
Engenharia Matemática
Departamento de Matemática
2020

Orientador
Joaquim Fernando Pinto da Costa, Professor Auxiliar, FCUP



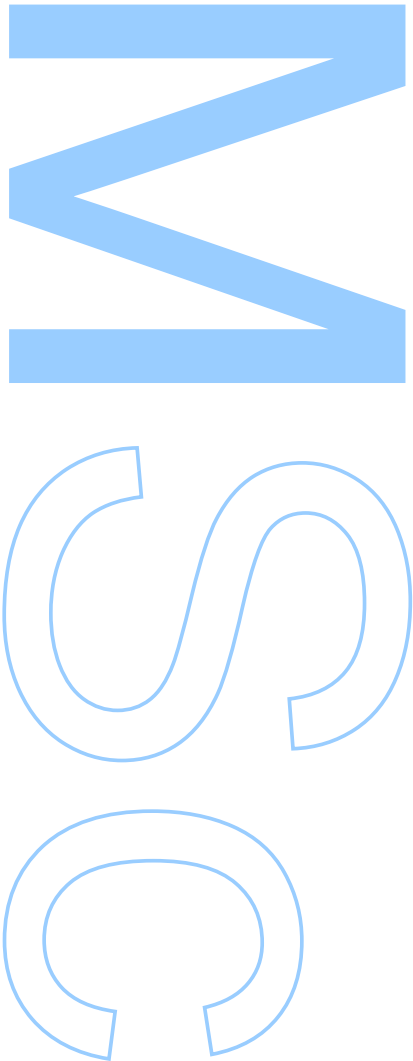
U.PORTO

FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Todas as correções determinadas pelo júri, e só essas, foram efetuadas.

O Presidente do Júri,

Porto, _____/_____/_____



"Essentially, all models are wrong, but some are useful"

- George E.P. Box

Acknowledgements

This thesis marks the end of my education journey and therefore I would like to briefly express my appreciation to the ones that supported me the most throughout the years.

First of all, I would like to express my utmost gratitude and respect to my supervisor, Professor Joaquim Costa, without whom none of this would be possible. His guidance and knowledge made my Master's Degree as enjoyable as possible.

I would also like to thank my group of friends for our innumerable conversations and lunches as well as for their encouragement during this journey.

To my family for all the help and support, without whom I would not have been able to pursue my studies and dreams.

Last but not least, I am deeply grateful to my girlfriend for all of her patience, support and motivation throughout the years and for always making me be a better person.

Resumo

Nas últimas décadas, o aumento de problemas de saúde gerou uma necessidade acrescida de estudos longitudinais em áreas como epidemiologia e genética. Ao contrário de estudos transversais, nos quais os sujeitos são observados num único momento, os estudos longitudinais são caracterizados por observarem indivíduos repetidamente ao longo de um período. Uma possível abordagem para a análise deste tipo de dados, é através do clustering, que permite encontrar a existência de grupos homogêneos que poderão não ser visível inicialmente a olho nu.

O objetivo principal desta dissertação é explorar e solidificar a utilização de métodos de clustering especificamente para dados longitudinais. Dada a incerteza na escolha do número ótimo de clusters, ou na escolha do método a usar, são também revistos vários índices de validação para acudir essas decisões. Estas técnicas são também aplicadas a dados relacionados com o novo coronavirus que, para além de testar as limitações dos métodos usando dados reais, pretendem descobrir tendências ou padrões importantes que possam auxiliar a comunidade científica na batalha do COVID-19.

Em particular, na aplicação a dados artificiais, ver-se-á que os métodos não paramétricos superam os métodos baseado em modelos, sendo o K-médias o método com os melhores resultados, e que o Calinski-Harabsz é o índice a sugerir o número ótimo de clusters mais frequentemente.

Na aplicação do clustering a dados do COVID-19, são encontrados dois grupos nos dados dos *Relatórios de Mobilidade da Comunidade* da Google e dois grupos nos dados das mortes causadas pelo novo coronavirus.

Keywords: Dados Longitudinais; Análise de Clusters; Clustering Longitudinal; K-médias; Hierarquico; Baseado-em-Modelo; Não Paramétricos; R; Coronavirus; COVID-19; Relatórios de mobilidade da comunidade

Abstract

Over the last decades, the increase in health problems generated an increased need to conduct longitudinal studies in fields like epidemiology and genetics. In contrast to cross-sectional studies, in which subjects are observed at a single moment, longitudinal studies are characterized for observing individuals repeatedly at multiple time points. One possible approach to analyse such data is through clustering, which allows the finding of homogenous groups that may not be visible at first sight.

This dissertation has as its primary goal, to explore and solidify the application of clustering methods specifically to longitudinal data. Given the uncertainty in the choices of the optimal number of clusters, or the method to be used, several validity indices are also reviewed to assist those decisions. These techniques are also applied to data related with the new coronavirus that, besides testing the limitations of each method using real data, intends to discover important trends or patterns that can assist the scientific community in the COVID-19 battle.

In particular, with the use of artificial data, it will be seen that non-parametric methods outperform model-based methods, with K-means being the method with the best results, and that Calinski-Harabsz is the index that more often suggests the optimal number of clusters.

In the clustering application to the COVID-19 data, two groups are found in both the *Community Mobility Reports* published by Google and in the deaths caused by the coronavirus dataset.

Keywords: Longitudinal data; Cluster analysis; Longitudinal Clustering; K-means; Hierarchical; Model-Based; Non-parametric; R; Coronavirus; COVID-19; Community Mobility Reports

Contents

List of Figures

List of Tables

Acronyms

AR	Autoregressive Model
BHC	Bayesian Hierarchical Clustering
CI	Confinement Index
CMR	Community Mobility Reports
DTW	Dynamic Time Warping
EM	Expectation-Maximization
HCA	Hierarchical Clustering Analysis
PCA	Principal Component Analysis
PV	Personal Variation
RV	Residual Variation
SSB	Sum-of-squares Between Clusters
SSW	Sum-of-squares Within Clusters

Chapter 1

Introduction

1.1 General considerations

Over the years, the occurrence and development of many diseases has been increasing [57, 86, 91, 115], leading to the necessity of cohort studies in fields like genetics and epidemiology. This type of studies are a particular form of longitudinal studies, in which individuals are observed over a certain time period, gathering, what is known as, longitudinal data. Observations can be seen as repeated measurements over time, making each individual produce his own trajectory in a geometric plane.

Univariate time series data typically occur from the collection of several data points over time from a single source, whereas longitudinal data typically arises from gathering observations over time from several sources. Thus, the collection of longitudinal data is naturally much more costly than the collection of time series or cross-sectional data [22]. However, it has become widely available in both developed and developing countries [58]. Longitudinal studies are mainly encountered in the health science field. Nonetheless, they can also be found in areas like climatology [66], business [24] or bank industry [56].

There are several advantages of longitudinal studies over cross-sectional studies or case-control studies. Firstly, a single longitudinal study can examine various outcome variables. For instance, while examining smokers, it can simultaneously look at deaths from cardiovascular, lung and cerebrovascular diseases. This opposes case-control studies as they measure only one outcome variable [81].

Moreover, to achieve a similar level of statistical power, less subjects are needed in a longitudinal study. Repeated measurements from a single subject provide more information than a single measurement obtained from a singular individual [30].

In addition, longitudinal studies allow the researcher to separate aging effects (changes over time within individuals) from cohort effects (differences between subjects at baseline). Such cohort effects are often mistaken for changes occurring within individuals. With the absence of longitudinal data, one cannot distinguish these two alternatives [30, 81].

Finally, with cross-sectional data, the representation of the population is at a single period in time. Thus the temporal aspects of a specific subject is not necessarily accessible. Longitudinal data can provide information about individual changes, whereas cross-sectional data cannot. Statistical estimates of personal trends can be used to better understand heterogeneity in the population and to determinate the change at an individual level [30].

In many situations, like in clinical trials, there is a natural heterogeneity among individuals in terms of how diseases develop and progress. This heterogeneity is owed to many factors, but mainly to genetics [4, 16, 123]. Hence, it is important to study and analyse the existence of homogeneous groups among the individuals trajectories. This helps to identify relevant trajectories patterns and track subjects who are very likely to be involved in the same or similar processes. This process of finding homogeneity is not a novelty in the medical field and is often called *disease subtyping*, which is, by definition, the task of identifying sub-populations of similar patients that can conduct treatment decisions for a given individual [109].

Over time, different statistical approaches have been conducted in order to analyse and model longitudinal data, such as Generalized Linear Models [73], Generalized Estimating Equations [127], Structural Equation Modelling [74] and Support Vector Machine [31]. Regardless, one efficient approach to detect patterns and structures in the data is through cluster analysis.

Clustering is an unsupervised learning method that seeks to separate objects (that can be individuals, cities, countries, etc.), into homogeneous groups. It relies on an algorithm with the purpose of minimizing the within-cluster variability (for compacter clusters) and maximizing the between-cluster variability (for a good separation). The relevant groups obtained by clustering share common characteristics and play an important role in the composition of the data.

As of today, there are a few different types of clustering, yet there isn't a general consensus in how to categorize them. The main reason for having so many clustering methods might

be because the notion of *cluster* is not precisely defined [35]. However, Han and Kamber [52] organized them into 5 categories: Partitioning, Hierarchical, Density-based, Grid-based and Model-based.

Partitioning methods are the simplest and most fundamental forms of cluster analysis [52]. These methods organize n objects of a certain dataset into K partitions, where each partition is a cluster containing at least one object and K is the number of clusters chosen beforehand ($K \leq n$). The clusters are formed to optimize an objective partitioning criterion, such as a function based on distances, ensuring that the objects within-clusters are similar to one another and dissimilar to objects in other clusters in terms of the dataset attributes [52]. The most known and commonly used partitioning methods are the K-means [79] and K-medoids [68].

Different from partitioning algorithms, a hierarchical clustering method groups data objects into a hierarchy, also called tree, or dendrogram, of clusters. This method has usually two types: divisive and agglomerative. Divisive methods start with all objects in one large cluster and they iteratively split it to smaller clusters. Agglomerative are the opposite, they start with individual objects as clusters and they are iteratively merged to form bigger clusters. Additionally, it also uses a linkage criteria to measure the dissimilarity between groups. A great advantage that this method has in comparison to other clustering algorithms, is that hierarchical clustering does not require the number of cluster to be given beforehand. Still, it can also encounter difficulties, for instance, in choosing where to cut the tree in order to get the best partition. This method is often forgotten in the application of longitudinal data, even though it is commonly used in time series clustering, considering that the interest in time series classification has been increasing in the last decades [10, 27, 59]. This approach can be extended to longitudinal data when considering each time series as an individual trajectory.

Density-based methods, like DBSCAN [34], are based on the idea that clusters are high density regions in the data space separated by regions of low density. This separation is done when the density in the neighbourhood exceeds some threshold. The observations staying in the low density regions are typically considered outliers or noise.

Grid-based methods take a space-driven approach by partitioning the object space into a finite number of *cells*, that forms a grid structure on which the clustering operations are performed [52]. They have the advantage of being fast because they are typically independent of the number of objects in the data. However they depend on the number of cells in each dimension in the quantized space. STING [121] is typically the most used among the grid-base

methods.

Last but not least, model-based methods assume that a mixture of underlying probability distributions generates the data and that it can be described using a standard statistical model [49, 99, 111, 130]. In model-based clustering algorithms, the parameters of each distribution are usually estimated by maximizing the likelihood. Thus, a particular clustering method can be expected to work well when the data fits the model. Model-based clustering has a long history and the notion of defining a cluster as a component in a mixture model was put forth by Tiedeman in 1955 [119]. Since then, the use of mixture models for clustering has grown into an important subfield of classification, especially since 1965, wherein a work published by Wolfe [124] traces the evolution of model-based clustering with the use of Gaussian mixtures.

Because model-based algorithms have a wide range of procedures, it is also common to classify all of the clustering methods as either model-based or non-parametric [19, 45, 53]. Model-based approaches are gaining popularity and have been applied in several contexts of longitudinal data over the years [48, 49, 84, 85, 111, 130].

Unlike model-based approaches, non-parametric clustering methods have no assumptions on how the data was generated and explicitly focus on defining similarity between subjects and clusters. They mainly focus on the dissimilarity measure, the clustering algorithm and the number of clusters [53]. Non-parametric clustering methods may be referred to as the traditional approaches. The K-means algorithm is by far the most used non-parametric method and has already been extended and adapted to longitudinal data [19, 43–45].

Using different algorithms for clustering usually leads to different results. Even when using the same algorithm, the changing of parameters can alter the output. Hence, some type of criteria is necessary to evaluate the results in order to give the users some reliance. The process of evaluating the results of a clustering algorithm is known as cluster validity (or cluster validation). This process consists of a set of techniques to find the set of clusters that best fits natural partitions of the data without any a priori class information [12, 75].

Based on external clustering validity indices, one can test if a method is efficient when the "true" partition is known beforehand. These indices are a measure of agreement between two partitions where one partition is the a priori known clustering structure, and the second is the partitioning with the groups obtained from clustering [32].

Even so, most of the clustering algorithms require the prior knowledge of the number of clusters as an input. However, since there is no reliable method to determine the "optimum"

number of clusters in a dataset [37], various internal clustering validity indices have also been proposed. These indices are used to measure the goodness of a clustering structure without external information [118]. This decision is done by executing the clustering algorithm several times with a different number of clusters in each run in order to choose the one with the best index results. Still, this issue remains mostly unsolved.

1.2 COVID-19 Pandemic

In late 2019, a local outbreak of pneumonia of originally unknown causes was reported in Wuhan, China and was quickly determined to be caused by a novel coronavirus, which has subsequently affected multiple countries worldwide [125, 129]. It has since been identified as a zoonotic coronavirus, resembling to MERS coronavirus and SARS coronavirus and designated as COVID-19 [76]. This type of pneumonia, caused by the SARS-CoV-2 virus, is a highly infectious disease, and the ongoing outbreak has been declared by the World Health Organization (WHO) as a public health emergency of international concern, posing a high risk to countries with vulnerable health systems.

It is known that epidemics tend to grow exponentially. Nevertheless, the growth cannot be exponential forever. Eventually, the virus will run out of new people to infect/kill either because most people have already been infected/killed or because we as a society manage to control it. Since countries did not follow the same restrictions or responded equally in the COVID-19 pandemic, different situations were seen throughout the world. For instance, considering the number of inhabitants, European countries like Italy or Belgium had high number of obits while countries like Turkey or Czechia had relatively low numbers. Thus, it is very pertinent to compare the response given by different countries in the COVID-19 battle.

Several approaches have been done to study the trajectories generated by the coronavirus infections (or deaths) in a attempt to model or predict future outcomes, while a vaccine is still yet to be found [55, 71, 98, 101]. Even so, as of today, the application of longitudinal clustering to these trajectories has not been found among the literature.

Besides that, in the course of the COVID-19 pandemic, multiple governmental authorities dictated quarantines in an effort to control the rapid spread of the virus. In late march, around a third of the world's population were under some form of lockdown. Consequently, the economy and other sectors suffered big repercussions with the reduced mobility.

1.3 Objective

In this work we aim to enhance the use of clustering in longitudinal data. To increase the potential and optimization of the discussed methods, we resort to validity indices that are evaluated in specifically longitudinal data test, which has not been done yet. With the application of longitudinal clustering to two different types of COVID-19 data, we also have the intention to discover relevant patterns or trends in the data that can possibly aid the scientific community in the coronavirus battle.

1.4 Thesis Outline

This thesis is organized in five chapters. In Chapter 2 we discuss the theory behind the clustering algorithms. We review a total of 11 methods, 7 of which are non-parametric while the remaining 4 are model-based. In Chapter 3 we introduce several available internal and external clustering validity indices. Besides, we also exploit the use of artificial longitudinal datasets to evaluate the internal indices as well as the clustering methods. For Chapter 4, we explore two different real datasets related to the COVID-19. One consists on the Community Mobility Reports published by Google while the other provides information about the obits caused by the new coronavirus. To complete, we share our conclusions and future work in Chapter 5.

1.5 Software used

There are several softwares that can be used in statistics and machine learning like `R`, `SAS`, `Python`, `SPSS`, `STATA` and many more. However, we chose to use the software `R` for many reasons. Besides being free, `R` was used a lot throughout the bachelor and master's degrees and it is still one of the most used programming languages in the world. It is relatively simple to use and has more than 10000 packages that help to perform a wide variety of functions, such as data manipulation, statistical modelling, and graphic visualization. Since the code scripts performed are very vast and use a considerably number of lines, we decided to not display them.

Chapter 2

Longitudinal Clustering Methods

In this chapter we introduce several possible approaches to cluster longitudinal data. Among the vastly clustering methods, K-means is presumably the most popular due to its simplicity and low time complexity [64]. In Section 2.1 we review 7 non-parametric methods, wherein five of them use K-means as a foundation. We also discuss two hierarchical clustering procedures, one with the traditional distance and another with an alternative metric. Further ahead, in Section 2.2, we consider four model-based approaches that use a probabilistic setting to assign clusters to observations.

2.1 Non-parametric methods

K-means is a hill-climbing algorithm that can also be seen as a particular case of a Expectation-Maximization (EM) algorithm for an iterative convergence [15, 45]. The idea behind K-means is to simply separate a dataset with n observations into K homogenous groups. For the initialization of this method, the number of clusters K is usually needed beforehand as an input, where each observation is then assigned to a cluster. To reach the optimal partition, the algorithm alternates between two steps: Expectation (E) and Maximization (M). In the E step, the distance between the observations and the centroids of each cluster are calculated. The M step then consists of assigning each observation to its nearest cluster [45]. This two phases are done repeatedly and iteratively until a stabilization in the cluster assignment is reached.

For the formulation of the K-means algorithm, let us consider a dataset with n observations

$X = \{x_1, \ldots, x_n\}$ to be clustered into a set of K groups $C = \{C_k,\ k = 1, \ldots, K\}$. In the longitudinal context, each observation x_i represents a trajectory produced by a certain individual while centroids represent the mean trajectories. As a result, each observation x_i represents a trajectory generated by the observed values of an individual i at t different times $(l = 1, \ldots, t)$, which can be represented as $x_i = \{x_{i_1}, x_{i_2}, \ldots, x_{i_t}\}$, where x_{i_l} represents the measured value of subject i at time l. Thus, each observation x_i $(i = 1, \ldots, n)$ represents a trajectory to be clustered.

K-means tries to find a partition that minimizes the total distances between each observation and its cluster centroid. Let z_k represent the centroid of cluster C_k, which can be calculated as the average of the observations that belong to that cluster. The squared distance between z_k and all of the observations x_i in cluster C_k can be defined as [63]:

$$SD(C_k) = \sum_{x_i \in C_k} ||x_i - z_k||^2 \tag{2.1}$$

The objective of K-means is to minimize the following sum of the squared distances over all K clusters [63]:

$$\sum_{k=1}^{K} \sum_{x_i \in C_k} ||x_i - z_k||^2 \tag{2.2}$$

2.1.1 K-means with Euclidean distance

The most popular metric for continuous features is the Euclidean distance [64]. Using K-means with this metric can also be referred to as the usual, or traditional, approach. This method was adapted for longitudinal data in the `R` package `kml` by Genolini et al. in 2011 [45] and has been applied throughout the years in several fields [28, 96, 113, 117]. This package is simple to use and also works with other distances such as Manhattan, Canberra or Chebyshev.

The Euclidean distance between two trajectories x_1 and x_2 is given by [45]:

$$d(x_1, x_2) = ||x_1 - x_2|| = \sqrt{\sum_{l=1}^{t} (x_{1_l} - x_{2_l})^2} \tag{2.3}$$

2.1.2 K-means with Fréchet distance

The Fréchet distance is a similarity measure for geometric shapes introduced by Maurice Fréchet in 1906 [42]. Unlike Euclidean distance, the Fréchet distance treats each trajectory as a curve and is able to identify clusters based on their shape rather than classical distances [44].

The first algorithm for the computation of this distance was provided by Alt and Godau in 1992 [5], where they also introduced the concept of a free-space diagram. This distance is often represented informally by an example of a person walking his dog with a leash. A person walks in one curve x_1 while his dog walks on another curve x_2, where they both can vary their speed, but neither of them can move backwards. The Fréchet distance between x_1 and x_2 is the minimum length of the leash that is required for both to complete their path without interruption.

More formally, a curve $x_1 \subseteq \mathbb{R}^2$ is a continuous mapping from $[0,1]$ to $\mathbb{R}^2$ and a reparameterization is a continuous non-decreasing surjection $\alpha : [0,1] \to [0,1]$ such that $\alpha(0) = 0$ and $\alpha(1) = 1$. The Fréchet distance between two curves x_1 and x_2 is defined as the infimum over all reparameterizations α and β of $[0,1]$ of the maximum over all $t \in [0,1]$ of the distance between $x_1(\alpha(t))$ and $x_2(\beta(t))$ [1, 6]:

$$Fr(x_1, x_2) = \inf_{\alpha,\beta} \max_{t\in[0,1]} \|x_1(\alpha(t)) - x_2(\beta(t))\| \tag{2.4}$$

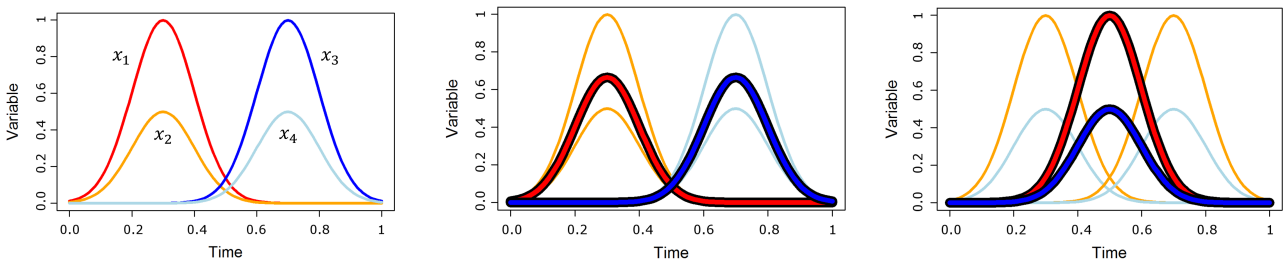
Going back to the example of a person walking his dog, we can think of the parameter t as time, $x_1(\alpha(t))$ as the human and $x_2(\beta(t))$ as the dog, where the length of the leash between them at time t is the distance between $x_1(\alpha(t))$ and $x_2(\beta(t))$.

As mentioned in Section 2.1, K-means requires the calculation of the mean trajectory in the Maximization phase. Informally, the Fréchet mean between two trajectories is the middle of the leash that links the dog to the person. More precisely, the Frétchet mean is defined as [44]:

$$MeanFrechet = \frac{x_1(\alpha(t)) + x_2(\beta(t))}{2} \tag{2.5}$$

As it is shown in Figure 2.1, performing K-means with the traditional Euclidean distance to a dataset with four trajectories $X = \{x_1, x_2, x_3, x_4\}$, the observations x_1 and x_2 (in orange) were assigned to the same cluster while x_3 and x_4 (light blue) were assigned to another cluster. The mean trajectories of the clusters are red and deep blue respectively. Using the same method with the Fréchet distance, the trajectories x_1 and x_3 (in orange) belong to the same cluster while x_2 and x_4 (light blue) belong to another. The mean trajectories naturally also differ from the classical method.

For the application of this method, we resort to the `kmlShape` package provided by Genolini et al. [44].



(a) Trajectories **(b)** K-means - Euclidean **(c)** K-means - Fréchet

Figure 2.1: Clustering according to Euclidean distance versus Fréchet. (Figure adapted from [44])

2.1.3 K-means with Mahalanobis distance with pre-defined correlation matrices

The next method we will explore was proposed recently by Joaquim Costa et al. [19, 38] which has shown good results in comparison to the traditional method (K-means with Euclidean distance).

This method considers that we have a multivariate longitudinal response, taking values in $\mathbb{R}^p$. More specifically, X_{il} represent a vector of length p containing the values of p continuous response variables, each one observed at time $l = 1, \ldots, t$. Furthermore, the model takes correlation into account by considering the following Mahalanobis distance between two trajectories x_1 and $x_2 \in \mathbb{R}^{p \times t}$ [19]:

$$d_M(x_1, x_2) = (x_1 - x_2)^T \Sigma^{-1} (x_1 - x_2) \tag{2.6}$$

where $\Sigma \in \mathbb{R}^{(p\times t)\times(p\times t)}$ is the following diagonal block matrix

$$\Sigma = \begin{bmatrix} \Sigma_1 & 0 & \cdots & 0 \\ 0 & \Sigma_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \Sigma_p \end{bmatrix}$$

assuming that the variance-covariance matrix between different time instances for variable k is Σ_k $(k = 1, \ldots, p)$.

This distance may present advantages over the usual Euclidean distance. First, it accounts for the fact that the variances in each direction may be different. Secondly, it also accounts for the covariances between the responses at different times. Finally, it reduces to the usual euclidean distance for uncorrelated data with unit variance.

For instance, in the case of just one response variable ($p = 1$) and equal variances for the different times, the matrix Σ reduces to the correlation matrix between different times. Thus, as in time series, we can model this matrix by an autoregressive correlation matrix $\Sigma \in \mathbb{R}^{txt}$ of order 1, AR(1) [19]:

$$\Sigma = \begin{bmatrix} 1 & r & r^2 & r^3 & \dots & r^{t-1} \\ r & 1 & r & r^2 & \dots & r^{t-2} \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ r^{t-1} & r^{t-2} & r^{t-3} & r^{t-4} & \dots & 1 \end{bmatrix}$$

An estimate of r can be obtained from the sample correlation coefficient between the vector of observations for times $1, 2, \dots, t-1$ and the correspondent lag 1 vector. For instance, if there are 4 different times, then: $\hat{r} = \text{Cov}\left(c\left(x_{l=1}, x_{l=2}, x_{l=3}\right), c\left(x_{l=2}, x_{l=3}, x_{l=4}\right)\right)$.

Still under the assumptions of a single variable and equal variances across the time instances, another possible Σ is [19]:

$$\Sigma = \begin{bmatrix} 1 & r & r & \dots & r \\ r & 1 & r & \dots & r \\ \vdots & \vdots & \vdots & & \vdots \\ r & r & r & \dots & 1 \end{bmatrix}$$

which reflects a compound symmetry structure corresponding to a uniform correlation between different time points. In fact, any structure of a correlation matrix could be used. However, we chose to use the autoregressive correlation matrix.

The next step in this clustering process consists on the application of the K-Means algorithm with the above Mahalanobis-type distance. For this part, we considered three different types of input variables: raw data and two versions of profile data.

Occasionally, the interest in longitudinal clustering relies on the relative behaviour of the individual trajectories rather than on their absolute values. For instance, if we have the prices of two stocks over t different times, we might be interested in grouping together stocks with a similar behaviour (increasing and decreasing at the same time even if taking very different values) rather than stocks having similar price values but behaving very differently. For these situations we propose to use profiles, that is, to perform one of the following initial operations

before proceeding to the clustering:

$$I)\ (W_{i_1},\ldots,W_{i_t}) \leftarrow \frac{(x_{i_1},\ldots,x_{i_t})}{\sum_{j=1}^{n} x_{i_j}};\quad II)\ (W_{i_1},\ldots,W_{i_t}) \leftarrow \frac{(x_{i_1},\ldots,x_{i_t})}{\|x_i\|}$$

Throughout the thesis, we will refer to these profiling processes as Profile I and II, respectively. While in the former the sum across coordinate values of each individual curve is 1, in the latter all individual trajectories take values on the unit sphere.

We now treat the application of our longitudinal clustering algorithm. This stage may not be an easy task for someone wanting to use the methodology once it involves of the Mahalanobys-type distance above and the correspondent estimator. Fortunately, the application can be simplified. In fact, one of our main points now is to consider the following transformation:

$$U = \Sigma^{-1/2} X \tag{2.7}$$

The existence of $\Sigma^{-1/2}$ follows from the Cholesky decomposition for Σ. Then, for the usual Euclidean distance:

$$\begin{aligned} d^2(U_1,U_2) &= (U_1-U_2)^T (U_1-U_2) \\ &= \left(\Sigma^{-1/2}x_1 - \Sigma^{-1/2}x_2\right)^T \left(\Sigma^{-1/2}x_1 - \Sigma^{-1/2}x_2\right) \\ &= (x_1-x_2)^T \left(\Sigma^{-1/2}\right)^T \Sigma^{-1/2} (x_1-x_2) \\ &= d_M^2(x_1,x_2) \end{aligned}$$

This shows that the Euclidean distance between the transformed data coincides with Mahalanobis distance of the original data. In the above calculations we have used the symmetry of $\Sigma^{-1/2}$, which is a consequence of the symmetry of Σ. In conclusion, by applying the above transformation we can thus make use of a well-known algorithm, `kml`, to cluster the data according to the new methodology.

2.1.4 Hierarchical

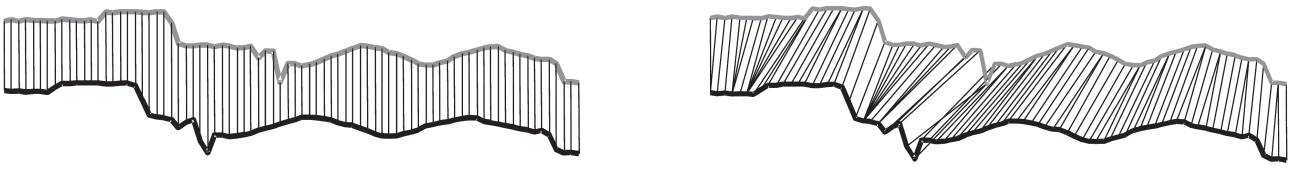
Hierarchical clustering analysis (HCA) is another technique to perform cluster analysis that, unlike most methods, does not require the pre-specification of the number of clusters. Hierarchical algorithms typically cluster the data based on distances, even though the use of distance functions is not mandatory. Some hierarchical algorithms use other clustering methods, like density or graph based, as a tool for the construction of the hierarchy [2].

Hierarchical algorithms can be categorized as either divisive (top-down) or agglomerative (bottom-up). The latter one treats each element as a single cluster and then successively

agglomerates pairs of clusters until all clusters have been merged into one that contains them all. In contrast, divisive methods start with all elements in a single cluster and split iteratively to smaller clusters. One relevant choice required in hierarchical clustering is how to measure the distance between clusters. Popular distances between clusters (referred to as *linkages*) include complete, single, average, centroid, among others. The hierarchical structure obtained from HCA is usually represented graphically trough a dendrogram (from the Greek, *dendron*, for tree). The variety of shapes and rotations of dendrograms provide major benefits in several areas such as biology, wherein it can be used to classify biological species, like the feline family for instance [39].

In the longitudinal data context, HCA is not so frequently used when comparing to partitioning or model-based methods. This may be due to the fact that hierarchical clustering is generally higher in time complexity, which can be a major problem in big data. Nevertheless, a big interest has been rising over the last decades in time series clustering [10, 27, 59].

For this method we will consider two distance measures: Euclidean and Dynamic Time Warping (DTW). The DTW was first introduced by Bellman and Kalaba in 1959 [11] and broadly explored and adjusted later in speech recognition applications [107, 108]. Nowadays, plenty of applications can be found in many fields of data mining, including time series clustering [36, 51, 67, 92]. With the premise that the Euclidean distance is a very fragile measure, the DTW allows an elastic shifting of the time axis, to accommodate sequences that are similar but out of phase. As it is shown in Figure 2.2, while two time series may have an overall similar shape, they might not be aligned in the time axis. The Euclidean distance, assumes that the ith point of one sequence is aligned with the ith point of the other. Contrariwise, the DTW allows a more intuitive distance measure.



(a) Euclidean **(b)** DTW

Figure 2.2: Time alignment of two time-dependent sequences. (Figure adapted from [69])

For the formulation of the DTW, let us consider two time series $x_1 = \{x_{1_1}, \ldots, x_{1_u}\}$ and $x_2 = \{x_{2_1}, \ldots, x_{2_v}\}$ with the length of u and v respectively. To align the two sequences using DTW, we construct a $u \times v$ matrix , denominated as M, that represents the point-to-point

correspondence relationship between x_1 and x_2. Each element M_{ij} indicates the alignment between x_{1_i} and x_{2_j} obtained by the distance $d(x_{1_i}, x_{2_j})$ between x_{1_i} and x_{2_j} (usually the Euclidean distance).

A warping path W is a set of matrix elements that defines a mapping between x_1 and x_2, where its ℓth element is defined as $w_\ell = (i, j)_\ell$. This is illustrated in Figure 2.3

So, we have [17]:

$$W = \{w_1, \ldots, w_L\}, \quad max(u, v) \leq L \leq v + u - 1 \tag{2.8}$$

The warping path is typically subject to several constraints, however we are interested only in the path which minimizes the warping cost [17]:

$$DTW(x_1, x_2) = \min \left\{ \frac{1}{L} \sqrt{\sum_{\ell=1}^{L} w_\ell} \right\} \tag{2.9}$$

where L in the denominator is used to take into account that warping paths may have different lengths. To find this path, dynamic programming is used to evaluate the following recurrence which defines the cumulative distance $D(i, j)$ as the distance $d(i, j)$ found in the current cell and the minimum of the cumulative distances of the adjacent elements [17]:

$$D(\mathrm{i}, \mathrm{j}) = d\left(x_{1_i}, x_{2_j}\right) + \min\{D(i-1, j-1), D(i-1, j), D(i, j-1)\} \tag{2.10}$$

For our purpose, we can consider each temporal sequence (time serie) as an individual trajectory. The DTW discussion in this work is necessarily brief. For a more detailed version, we recommend the reference [70] for the interested reader.

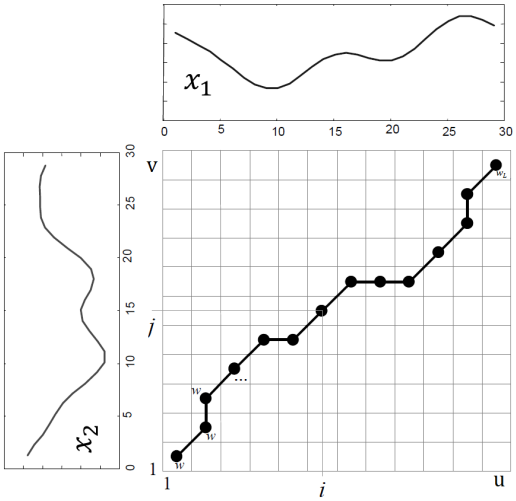


Figure 2.3: A warping path between two time series x_1 and x_2. (Figure adapted from [17])

2.2 Model-based methods

Unlike K-means, which assigns a definite cluster to each observation, model-based methods gives them probabilities of belongingness. These models work by using finite mixture models, wherein it is assumed that the data is generated by a mixture of probability distributions, in which each component represents a different cluster. Thus, a particular clustering method can be expected to work well when the data fits the model.

If we knew which mixture component each observation came from, estimating the mixing proportions and the parameters of each component distribution would be simpler. Since that is not known, given an initial guess of the parameters, we can probabilistically assign a component to each observation and then get a better estimation of the parameters based on these assignments. This has some similarity with K-means, but it occurs in a probabilistic setting, with a proof that the algorithm will reach a (local) maximum of the likelihood.

The most popular family of mixture models is the MCLUST family [65], which is implemented in the `mclust` package in `R` [41]. Members of this family display eigen-decomposed component covariance matrices, where, in a general case, the component covariance structure is of the form $\Sigma_k = \lambda_k H_k A_k H_k'$, where H_k represents the matrix of eigenvectors of Σ_k, A_k the diagonal matrix with entries proportional to the eigenvalues of Σ_k, and λ_k the relevant constant of proportionality [46].

A random vector X is said to emerge from a parametric finite mixture distribution if, for all $x \subset X$, its density can be written as [83]:

$$f(x|\vartheta) = \sum_{k=1}^{K} \pi_k f_k\left(x|\theta_k\right) \tag{2.11}$$

where

$$\pi_k > 0, \quad \sum_{k=1}^{K} \pi_k = 1 \tag{2.12}$$

Here, π_k is the probability that an observation belongs to the kth component (group) and θ_k is parameters of the kth component in the mixture. When x is continuous, f_k is often taken to be Gaussian and, in that case, $\theta_k = (\mu_k, \Sigma_k)$. The term $f_k\left(x|\theta_k\right)$ represents the kth component density and $\vartheta = (\pi_1, \ldots, \pi_K, \theta_1, \ldots, \theta_K)$ represents the vector of parameters. When the number of mixture components K is known, only ϑ has to be estimated. When K is not provided, we have to additionally estimate the number of components in the mixture. In the

clustering context, the component densities $f_1(x|\theta_1), f_2(\mathbf{x}|\theta_2), \ldots, f_K(\mathbf{x}|\theta_K)$ are usually taken to be of the same kind, i.e., $f_k(x|\theta_k) = f(x|\theta_k)$, $\forall g$.

Let $r_i = (r_{i1}, \ldots, r_{iK})$ represent the component membership of observation x_i, so that $r_{ik} = 1$ if observation x_i belongs to component k and $r_{ik} = 0$ otherwise. Suppose we observe n vectors $x_1, \ldots, x_n$ and that all of them are unlabelled. Considering that $f_k(x|\theta_k) = f(x|\theta_k)$, $\forall k$, the likelihood for the mixture model with K components is given by [83]:

$$L(\vartheta|x) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k f(x_i|\theta_k) \tag{2.13}$$

After the parameters have been estimated, the predicted classification results are obtained by the following *a posteriori* probabilities [83]:

$$\hat{r}_{ik} = P(x_i \in k|x_i) = \frac{\hat{\pi}_k f\left(x_i|\hat{\theta}_k\right)}{\sum_{h=1}^{K} \hat{\pi}_h f\left(x_i|\hat{\theta}_h\right)} \tag{2.14}$$

for $i = 1, \ldots, n$.

2.2.1 Gaussian mixture model with Cholesky decomposed covariance structure

One of the most used distributions in mixture models is the Gaussian distribution. The gaussian mixture model has its density of the form [82]:

$$f(x|\vartheta) = \sum_{k=1}^{K} \pi_k \phi(x|\mu_k, \Sigma_k) \tag{2.15}$$

where π_k represents the probability of membership of group k and $\phi(x|\mu_k, \Sigma_k)$ denotes the density of a multivariate Gaussian distribution with mean μ_k and covariance matrix Σ_k.

Prior to the work done by McNicholas, Paul et al. [84] there had been published some literature about mixture models applied to longitudinal data [20, 94], but none of those models had a covariance structure specifically designed for the analysis of longitudinal data. The method we will discuss was developed by McNicholas, Paul et al. [84] where a family of mixture models with a covariance structure specifically designed for the model-based clustering of longitudinal data is introduced. Considering that the outcome variable is recorded in a ordered time, a covariance structure that concretely accounts for the relationship between measurements at different time points is necessary.

The concept of modified Cholesky decomposition was introduced in 1999 by M. Pourahmadi [97], where he exploited the fact that a covariance matrix Σ can be decomposed

as $T\Sigma T' = D$, where T is a unique lower triangular matrix with diagonal elements of 1's and D is a unique diagonal matrix with strictly positive diagonal entries. Another way to formulate the modified Cholesky decomposition is in the form of $\Sigma^{-1} = T'D^{-1}T$, where the values of T and D have interpretations as generalized autoregressive parameters and innovation variances, respectively.

The method we present in this section assumes a Gaussian mixture model, with a modified Cholesky-decomposed covariance structure for each mixture component. Thus, the density of an observation x_i in group k is given by [84]:

$$f\left(x_i|\mu_k, T_k, D_k\right) = \frac{1}{\sqrt{(2\pi)^p \left|D_k\right|}} \exp\left\{-\frac{1}{2}\left(x_i - \mu_k\right)' T_k' D_k^{-1} T_k \left(x_i - \mu_k\right)\right\} \tag{2.16}$$

where T_k is the $p \times p$ lower triangular matrix and D_k is the $p \times p$ diagonal matrix that follow from the modified Cholesky decomposition of Σ_k.

EM algorithm is then implemented by assuming that there are some missing observations, namely the group identifiers, that when combined with the observed data, gives the so-called complete data. The data that is missing is taken to be the group membership labels, which is denote as r, where $r_{ik} = 1$ if the observation x_i is in group k and $r_{ik} = 0$ otherwise. Joining the known data x and the missing data r, gives the complete-data (x, r). The complete-data likelihood for the mixture model is then given by [84]:

$$L_c\left(\pi_k, \mu_k, T_g, D_k\right) = \prod_{i=1}^{n} \prod_{k=1}^{K} \left[\pi_k f\left(x_i|\mu_k, T_k, D_k\right)\right]^{r_{ik}} \tag{2.17}$$

and the expected value of the complete-data log-likelihood for the mixture model is [84]:

$$Q\left(\pi_k, \mu_k, T_k, D_k\right) = \sum_{k=1}^{K} n_k \log \pi_k - \frac{np}{2} \log 2\pi - \sum_{k=1}^{K} \frac{n_k}{2} \log \left|D_k\right| - \sum_{k=1}^{K} \frac{n_k}{2} \operatorname{tr}\left\{T_k S_k T_k' D_k^{-1}\right\} \tag{2.18}$$

where the r_{ik} have been replaced by their expected values:

$$\hat{r}_{ik} = \frac{\hat{\pi}_k f\left(x_i|\hat{\mu}_k, T_k, D_k\right)}{\sum_{h=1}^{K} \hat{\pi}_h f\left(x_i|\hat{\mu}_h, \hat{T}_h, \hat{D}_h\right)}, \tag{2.19}$$

and where $n_k = \sum_{i=1}^{n} \hat{r}_{ik}$ and $S_k = (1/n_k) \sum_{i=1}^{n} \hat{r}_{ik} \left(x_i - \mu_k\right)\left(x_i - \mu_k\right)'$. Now maximizing Q with respect to π_k and μ_k gives $\hat{\mu}_k = \sum_{i=1}^{n} \hat{r}_{ik} x_i / \sum_{i=1}^{n} \hat{r}_{ik}$ and $\hat{\pi}_k = n_k/n$ respectively.

Expectation-Maximization algorithm plays a critical role by alternating between two steps. In the E step, $\hat{r}_{ik}$ is calculated. In the M step, the parameters that maximize the expected log-likelihood, with the fixed value r_{ik} obtained from the E step, are computed. With new estimates obtained in the this last step, the E is repeated to recompute new membership values and this entire process keeps repeating until model parameters converge.

2.2.2 Mixture of multivariate t-distributions models with Cholesky decomposed covariance structure

The model density for a mixture of multivariate t-distributions with K components has the form [85]:

$$f(x|\vartheta) = \sum_{k=1}^{K} \pi_k f\left(x|\mu_k, \Sigma_k, v_k\right) \tag{2.20}$$

where $f\left(x|\mu_k, \Sigma_k, v_k\right)$ is the density of a multivariate t -distribution with mean μ_k, covariance matrix Σ_k, and v_k degrees of freedom.

To take in consideration the longitudinal context of the time course, the covariance matrix is parametrized to account for the correlation structure of the data. Like in the Gaussian mixture model discussed in Section 2.2.1, this method also uses the modified Cholesky decomposition. In model-based clustering applications, the mean is not usually modelled. However, to capture the trend of the time course, this algorithm uses a linear model for the mean.

In this method, we consider a mixture of multivariate t-distributions model, where the component precision matrix is modified Cholesky-decomposed so that $\Sigma_k^{-1} = \mathrm{T}_k'\mathrm{D}_k^{-1}\mathrm{T}_k$, with T_k and D_k as defined in Section 2.2.1 and the mean modelled using a linear model.

Now suppose that each observation x_i in the data is measured at the folloing t time points: $q_1, \ldots, q_t$. To simplify the modelling of the mean of component k, the authors define a intercept a_k and a slope b_k, where [85]:

$$Q = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ q_1 & q_2 & \cdots & q_t \end{pmatrix}^T, \quad \beta_k = \begin{pmatrix} a_k \\ b_k \end{pmatrix}$$

The likelihood for this mixture model can then be written as [85]:

$$L(\vartheta|x) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k f\left(x_i|Q\beta_k, \left(I_k' D_k^{-1} T_k\right)^{-1}, v_k\right) \tag{2.21}$$

where $f\left(x_i|Q\beta_k, \left(r_k^{-1}T_k\right)^{-1}, v_k\right)$ represents the density of a p-dimensional t-distributed random vector with precision matrix $T_k' D_k^{-1} T_k$, mean $Q\beta_k$ and v_k degrees of freedom.

Let r_{ik} denote the component membership of observation x_i where $r_{ik} = 1$ if observation belongs to component k and $r_{\mathrm{ik}} = 0$ otherwise. In this model, a latent variable U_{ik} is introduced so that [85]:

$$x_i|u_{ik} \cdot r_{ik} = 1 \sim \mathcal{N}\left(Q\beta_k \left(T_k D_k^{-1} T_k\right)^{-1} / u_{ik}\right) \tag{2.22}$$

independently, for $i = 1, \ldots, n$, and $U_{ik}|r_{ik} = 1$ follows a gamma distribution with parameters $(v_k/2, v_k/2)$, independently. The observed plus missing data, known as the complete-data,

then consist of the observed x_i, the unobserved r_{ik}, and the latent u_{ik}. The complete-data log-likelihood can then be written in the form $l_c(\mathcal{B}) = l_{1c}(\pi) + l_{2c}(v) + l_{3c}(\zeta)$, where $\pi = (\pi_1, \ldots, \pi_K), v = (v_1, \ldots, v_K), \zeta = (\beta_1, \ldots, \beta_K, T_1, \ldots, T_K, D_1, \ldots, D_K)$, and [85]:

$$l_{1c}(\pi) = \sum_{k=1}^{K} n_k \log \pi_k \tag{2.23}$$

$$l_{2c}(v) = \sum_{k=1}^{K}\sum_{i=1}^{n} r_{ik}\left[-\log\Gamma\left(\frac{v_k}{2}\right) + \frac{v_k}{2}\log\left(\frac{v_k}{2}\right) + \frac{v_k}{2}\left(\log u_{ik} - u_{ik}\right) - \log u_{ik}\right] \tag{2.24}$$

$$l_3(\zeta) = \frac{np}{2}\log(2\pi) - \sum_{k=1}^{K}\frac{n_k}{2}\log|D_k| - \sum_{k=1}^{K}\frac{n_k}{2}\operatorname{tr}\left(\mathrm{Tg}_g \operatorname{Tr}' D_k^{-1}\right) \tag{2.25}$$

where $n_k = \sum_{i=1}^{n} r_{ik}$ and $S_k = (1/n_k)\sum_{i=1}^{n} r_{ik}u_{ik}\left(x_i - Q\beta_k\right)\left(x_i - Q\beta_k\right)'$, for $k = 1, \ldots, K$.

At each Expectation step, the missing data, r_{ik} and u_{ik} are replaced by their conditional expected values [85]

$$\hat{r}_{ik} = \frac{\hat{\pi}_k f\left(X_i|Q\hat{\beta}_k\left(\hat{t}_k'\hat{D}_k^{-1}\hat{T}_k\right)^{-1}, \hat{v}_k\right)}{\sum_{h=1}^{K}\hat{\pi}_h f\left(x_i|Q\hat{\beta}_k, \hat{T}_k'\hat{D}_k^{-1}\hat{T}_k\right)^{-1}, \hat{v}_k\right)} \tag{2.26}$$

and

$$\hat{u}_{ik} = \frac{\hat{v}_k + p}{\hat{v}_k + \delta\left(x_i Q\hat{\beta}_k|\hat{T}_k'\hat{D}_k^{-1}\hat{T}_k\right)^{-1}\right)} \tag{2.27}$$

In the Maximization step, parameters are updated. Because this model is significantly vast and complex, more detailed information about the parameters estimation can be read in [85]. For the implementation of the previous and current method, we mainly used the `longclust` package.

2.2.3 Gaussian mixed-effects model with smoothing spline estimation

The following method employs mixed-effects models with a non-parametric smoothing spline fitting. This method was developed by Golumbeanu, Monica et al. in 2018 [47] and used in clustering of genes expression of HIV-1 replications [48]. Again, let $X = \{x_1, \ldots, x_n\}$ be a set of n observations, where each observation $x_i = \{x_{i_1}, \ldots, x_{i_t}\}$ is an individual trajectory measured at t different times ($l = 1, \ldots, t$). The task is then to cluster the n observations into K groups based on their trajectories behaviour.

In this procedure, each individual trajectory x_i, at time l, is modelled as a linear combination of a fixed effect $\xi_k(l)$ associated to the cluster k, a random effect β_i, and an error term ϵ_{il} [47]:

$$x_{i_l} = \xi_k(l) + \beta_i + \epsilon_{il} \tag{2.28}$$

where $\beta_i \sim \mathcal{N}(0, \theta_k)$ and $\epsilon_{il} \sim \mathcal{N}(0, \theta)$. The fixed effect ξ_k corresponds to the general trajectories trend or baseline associated to cluster k, the random effect β_i captures any systematic shift from the general trend, and ϵ_{il} corresponds to the measurement error. Consequently, x_i follows a multivariate normal distribution $\mathcal{N}(\xi_k, \Sigma_k)$ with the covariance Σ_k defined as [47]:

$$\Sigma_k = \theta_k I_t + \theta J_t = \begin{pmatrix} \theta_k + \theta & \theta & \dots & \theta \\ \theta & \theta_k + \theta & \dots & \theta \\ \dots & \dots & \dots & \dots \\ \theta & \theta & \dots & \theta_k + \theta \end{pmatrix}$$

where I_t represents the unity matrix of dimension t and J_t a squared matrix of ones with dimension t. This clustering issue is moulded into a mixture model, where each cluster can be described with a Gaussian distribution with the parameters $\mathcal{N}(\xi_k, \Sigma_k)$ [47]:

$$x_i \sim \sum_{k=1}^{K} \pi_k \mathcal{N}(\xi_k, \Sigma_k) \tag{2.29}$$

with π_k representing the mixing coefficients of the mixture model.

Rather than following a parametric approach for the choice of the baseline ξ_k, the model uses a less restrictive and non-parametric way by using smoothing splines. In 2014, Gu, Chong [50] shown that when fitting smoothing splines to a set of Gaussian random variables, the usual residual sum of squares (RSS) minimization problem has an identical maximum-likelihood formulation. When we try to fit a cubic spline ξ_k to a set of observations, we try to find the ξ_k that minimizes the following score [47]:

$$\sum_{l=1}^{t} (x_{i_l} - \xi_k(l))^2 + \lambda_k \int (\xi_k''(t))^2 \, dt \tag{2.30}$$

where $\sum_{l=1}^{t} (x_{i_l} - \xi_k(l))^2$ quantifies the deviation of the observed values from the curve ξ_k, and $\lambda_k \int (\xi_k''(t))^2 \, dt$ penalizes the roughness of the curve. If the variables x_{i_l} follow a normal distribution, the first term of the score becomes proportional to their minus log likelihood, leading to the following penalized likelihood score [50]:

$$-L(x_i) + \lambda_k \int (\xi_k''(t))^2 \, dt \tag{2.31}$$

where $L(x_i)$ stands for the log likelihood of the data. Hence, the problem can be formulated as a special case of maximum-likelihood estimation and can be solved using Expectation-Maximization [78].

The log-likelihood function in the context of the above-defined mixture of Gaussian mixed effects models is [78]:

$$L(X) = \log P(X|\vartheta) = \log \prod_{i=1}^{n} P(x_i|\vartheta) \tag{2.32}$$
$$= \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k \mathcal{N}(x_i|\xi_k, \Sigma_k)$$

where ϑ represents the complete set of model parameters. For the application of this method, we used the `TMixClust` package, which itself resorts to the `gss` package. The latter one is used to perform a non-parametric smoothing spline fitting for the Gaussian random variables. It works by finding the cubic smoothing spline that minimizes the penalized likelihood score described previously and by estimating the parameters of the associated multivariate Gaussian distribution, more specifically, the mean vector ξ_k and the covariance matrix Σ_k.

In essence, this method starts by initializing clusters and then consists on mainly two stages: Expectation and Maximization. In the Expectation phase, the posterior probabilities are computed and trajectories are assigned to clusters based on their posterior probabilities. In the Maximization phase, the penalised likelihood score is maximized and the parameters are updated. When the convergence is achieved, the maximum likelihood solution is returned.

2.2.4 Bayesian Hierarchical

We will now consider an algorithm that performs hierarchical clustering in a Bayesian setting. Similar to the traditional agglomerative hierarchical clustering, the Bayesian hierarchical clustering (BHC) initializes each observation in its own cluster and iteratively merges pairs of clusters, except it uses a statistical hypothesis test to choose which clusters to merge. This method was firstly introduced by Heller and Ghahramani [54] and then applied in the clustering of gene expression data [18, 110].

Using the notations from [54, 110], let $X = \{x_1, \ldots, x_n\}$ denote the dataset and $\mathcal{D}_i \subset x$ the set of observations at the leaves of the subtree T_i. The algorithm is initialized with n trees $\{T_i : i = 1, \ldots, n\}$, each containing a single observation $\mathcal{D}_i = \{x_i\}$. At each step of the algorithm there is an examination to consider all possible merging of two trees. As is shown in Figure 2.4, if T_i and T_j are merged into some new tree T_k then the associated set of data is $\mathcal{D}_k = \mathcal{D}_i \cup \mathcal{D}_j$.

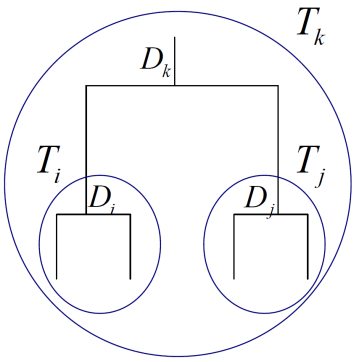


Figure 2.4: Considering a portion of a tree, T_i and T_j are merged into T_k, and consequently the associated data sets D_i and D_j are merged into D_k. (Figure adapted from [54])

To consider each merge, two hypotheses are addressed. The first hypothesis $\mathcal{H}_1^k$ is that all of the data in $\mathcal{D}_k$ was generated independently and identically from the same probabilistic model $p(\mathbf{x}|\theta)$ with unknown parameters θ. We also assume that this probabilistic model is a multivariate Gaussian, with parameters $\theta = (\mu, \Sigma)$. The alternative hypothesis $\mathcal{H}_2^k$ is that the data in $\mathcal{D}_k$ has two or more clusters in it.

To evaluate the probability of the data under hypothesis $\mathcal{H}_1^k$, we need to specify some prior over the parameters of the model, $p(\theta|\beta)$ with hyperparameters β. The probability of the data $\mathcal{D}_k$ under $\mathcal{H}_1^k$ is [110]:

$$p\left(\mathcal{D}_k|\mathcal{H}_1^k\right) = \int p\left(\mathcal{D}_k|\theta\right) p(\theta|\beta) d\theta = \int \left[\prod_{x_i \in \mathcal{D}_k} p\left(x_i|\theta\right)\right] p(\theta|\beta) d\theta \tag{2.33}$$

This computes the probability that all of the data in $\mathcal{D}_k$ were generated from the same parameter values, assuming a model of the form $p(\mathrm{x}|\theta)$.

The probability of the data under the alternative hypothesis $\mathcal{H}_2^k$ is simply a product over the subtrees $p\left(\mathcal{D}_k|\mathcal{H}_2^k\right) = p\left(\mathcal{D}_i|T_i\right) p\left(\mathcal{D}_j|T_j\right)$ where the probability of a data set under a tree (e.g. $p\left(\mathcal{D}_i|T_i\right)$) is defined below.

Combining the probability of the data under hypotheses $\mathcal{H}_1^k$ and $\mathcal{H}_2^k$, weighted by the prior that all points in $\mathcal{D}_k$ belong to one cluster $\pi_k = p\left(\mathcal{H}_1^k\right)$, we obtain the marginal probability of the data in tree T_k [110]:

$$p\left(\mathcal{D}_k|T_k\right) = \pi_k p\left(\mathcal{D}_k|\mathcal{H}_1^k\right) + \left(1 - \pi_k\right) p\left(\mathcal{D}_i|T_i\right) p\left(\mathcal{D}_j|T_j\right) \tag{2.34}$$

The first part of the equation considers the hypothesis that there is a single cluster in D_k while the second term efficiently sums over all other clusterings of the data in D_k which are

consistent with the tree structure.

The posterior probability of the merged hypothesis $r_k = p\left(\mathcal{H}_1^k|\mathcal{D}_k\right)$ is then obtained using Baye's rule [110]:

$$r_k = \frac{\pi_k p\left(\mathcal{D}_k|\mathcal{H}_1^k\right)}{p\left(\mathcal{D}_k|T_k\right)} = \frac{\pi_k p\left(\mathcal{D}_k|\mathcal{H}_1^k\right)}{\pi_k p\left(\mathcal{D}_k|\mathcal{H}_1^k\right) + (1-\pi_k)\, p\left(\mathcal{D}_i|T_i\right) p\left(\mathcal{D}_j|T_j\right)} \tag{2.35}$$

The posterior probability is used to decide which two trees to merge. If $r_k > 0.5$, it means that it is more likely that the observations contained in the trees were generated from the same underlying function, and therefore are merged. If $r_k < 0.5$ then the branches constitute separate clusters.

The implementation of this method was possible using the `BHC` package, which can be downloaded from *Bioconductor's* [1] website.

A brief description of the methods we address in this thesis are displayed in Table 2.1. There are a total of 11 models, wherein 4 are model-based and 7 are non-parametric. The M's notation for the models will be used throughout the next chapter in order to save some space.

Table 2.1: Outline of the 11 methods discussed in this work

	Method	Type
M1	K-means with Euclidean distance	Non-Parametric
M2	K-means with Fréchet distance	Non-Parametric
M3	K-means with Mahalanobis distance: Raw data	Non-Parametric
M4	K-means with Mahalanobis distance: Profile I	Non-Parametric
M5	K-means with Mahalanobis distance: Profile II	Non-Parametric
M6	Gaussian mixture with Cholesky decomposition	Model-Based
M7	Mixture of multivariate t-distributions models with Cholesky decomposition	Model-Based
M8	Gaussian mixed-effects model with smoothing spline estimation	Model-Based
M9	Hierarchical with Euclidean distance	Non-parametric
M10	Hierarchical with DTW distance	Non-parametric
M11	Bayesian Hiearchical	Model-Based

[1] Bioconductor is a open source and open development software project that provides tools for the analysis and comprehension of high-throughput genomic data

Chapter 3

Clustering Validity

"The validation of clustering structures is the most difficult and frustrating part of cluster analysis. Without a strong effort in this direction, cluster analysis will remain a black art accessible only to those true believers who have experience and great courage."

- Jain and Dubes

In this chapter, we examine the prominence of the cluster validity. We start discussing internal indices and present them along with their historical date context. Afterwards, we elaborate several tests on artificial data to evaluate them. Later on, we also review some of the external indices and provide various tests with artificial data.

3.1 Internal Indices

Internal validity indices aim to discover if the structure obtained by clustering is intrinsically appropriate for the data. Thus, it can be used to determine the number of clusters necessary in the clustering. In order to do so, clustering methods are executed with different values of K, followed by a comparison of the obtained values. The instance that yields the best results, suggests the optimal number of clusters to be used. The indices we will discuss in this work are presented in Table 3.1 along with their abbreviations, goals and references in a chronological order. The "*goal*" represents if the corresponding index aims to maximize or minimize its value, e.g, the higher the PBM index the better, whereas in C index, the lowest value is desired. For

the calculation of the indices, we mainly used the `clusterCrit` package [25], that contains nearly all of the internal and external indices we will consider.

Table 3.1: Internal indices with their respective abbreviation, goal, date and reference

Index	Abbreviation	Goal	Date	Ref
Ball-Hall	BH	$\uparrow max$	1965	[9]
Calinski-Harabasz	CH	$\uparrow max$	1974	[13]
Gamma	Γ	$\uparrow max$	1975	[8]
C	C	$\downarrow min$	1976	[61]
Davies-Bouldin	DB	$\downarrow min$	1979	[23]
Sillhouette	S	$\uparrow max$	1987	[106]
Banfield-Raftery	BR	$\downarrow min$	1993	[65]
Ray-Turi	RT	$\downarrow min$	1999	[103]
Wemmert-Gancarski	WG	$\uparrow max$	2000	[122]
PBM	PBM	$\uparrow max$	2004	[93]

In cluster analysis, the within group variance and between group variance can be calculated by the sum-of-squares within clusters (SSW) and sum-of-squares between clusters (SSB) respectively. Some of the internal indices are based on these two sums. The SSW measures the dispersion within clusters. More specifically, it is the total of the squared distances between each element and the centroid of its cluster [128]:

$$SSW = \sum_{k=1}^{K} \sum_{x_i \in C_k} ||x_i - z_k||^2 \tag{3.1}$$

On the other hand, SSB measures the dispersion of the clusters by calculating the weighted sum of the squared distance between each centroid of the cluster to the centroid of all elements in the data [128]:

$$SSB = \sum_{k=1}^{K} n_k ||z_k - \bar{x}||^2 \tag{3.2}$$

where K represents the number of clusters; x_i an observation; $\bar{x}$ the centroid of all the observations in the data, i.e, the mean value of the whole dataset; z_k the centroid of cluster C_k which is equal to the average of all observations that belong to that cluster and n_k the number of observations in cluster C_k. The previous notation will be used throughout the indices formulation.

1. **Ball-Hall (BH)** [9]

   The Ball-Hall index is given by the dispersion within clusters divided by the total number of clusters K:

$$BH = \frac{SSW}{K} \tag{3.3}$$

2. **Calinski-Harabasz (CH)** [13]

   This index aims to maximize the dispersion between clusters and minimize dispersion within clusters. The CH index can also be seen as the Pseudo F statistic [90].

$$CH = \frac{n-K}{K-1}\frac{SSB}{SSW} \tag{3.4}$$

   where K is the number of clusters and n the number of observations.

3. **Gamma (Γ)** [8]

   The Gamma index is given by:

$$\Gamma = \frac{s^+ - s^-}{s^+ + s^-} \tag{3.5}$$

   where s^+ represents the number of times that the distance between a pair of objects from the same cluster is strictly smaller than the distance between a pair of objects not belonging to the same cluster. Likewise, s^- is the number of times that the distance between a pair of objects belonging to the same cluster are strictly greater than the distance between a pair of objects not belonging to the same cluster. Better partitions are expected to have higher values of s^+ and lower values of s^-, therefore having higher values of Γ. Since this index varies between -1 and 1, the desired value to obtain from Γ is 1.

4. **C** [61]

$$C = \frac{d_w - min(d_w)}{max(d_w) - d_w} \tag{3.6}$$

   Where d_w is the sum of the distances of all pairs of observations in the same cluster. If we consider that n_p is number of pairs of observations in the same cluster, then $min(dw)$ is the sum of the n_p smallest distances if all pairs of objects are considered. Similarly,

$max(d_w)$ is the sum of the n_p largest distances out of all pairs. The C index is limited to the interval $[0, 1]$ and should be minimized.

5. **Davies-Bouldin (DB)** [23]

$$DB = \frac{1}{K} \sum_{k=1, k \neq k'}^{K} max\left(\frac{d_k + d_{k'}}{||z_k - z_{k'}||} \right) \tag{3.7}$$

where d_k (resp. d'_k) is the average of the distances between all elements of cluster C_k (resp. $C_{k'}$) to the respective centroid z_k (resp. $z_{k'}$) and the denominator $||z_k - z_{k'}||$ is the distance between the centroids of the clusters C_k and $C_{k'}$.

6. **Silhouette (S)** [106]

The Silhouette index is given by:

$$S = \frac{1}{K} \sum_{k=1}^{K} s(C_k) \tag{3.8}$$

where $s(C_k)$ is the Silhouette width for the kth cluster C_k, and can be expressed as follows:

$$s(C_k) = \frac{1}{n_k} \sum_{x_i \in C_k} s(x_i)$$

where n_k is the number of observations in C_k and $s(x_i)$ is the Silhouette width for the observation x_i that can be written as:

$$s(x_i) = \frac{b(x_i) - a(x_i)}{max(a(x_i), b(x_i))}$$

where $a(x_i)$ is the average dissimilarity of the observation x_i to all other objects of the same cluster; $b(x_i)$ is the smallest of the mean distances of x_i to the observations belonging to the other clusters.

7. **Banfield-Raftery (BR)** [65]

This index is the weighted sum of the logarithms of the traces of the variance-covariance matrix of each cluster:

$$BR = \sum_{k=1}^{K} n_k log\left(\frac{tr(WG_k)}{n_k} \right) \tag{3.9}$$

where WG_k represents within-group scatter matrix for each cluster C_k and n_k, the number of observations in cluster C_k.

8. **Ray-Turi (RT)** [103]

   RT is the quotient between the distance of intra-clusters and the distance of inter-clusters:

$$RT = \frac{intra}{inter} \tag{3.10}$$

   where intra-cluster is the the squared distance between each observation and the centroid of its cluster (SSW), divided by the total number of observations:

$$intra = \frac{SSW}{n}$$

   For the inter-cluster we consider the minimum of the squared distances between all the of the clusters centroids:

$$inter = min\ ||z_k - z_{k'}||^2$$

   where $k = 1, \ldots, K-1$ and $k' = k+1, \ldots, K$

9. **Wemmert-Gancarski (WG)** [122]

$$WG = \frac{1}{n}\sum_{k=1}^{K} max\{0, 1 - \frac{1}{n_k}\sum_{x_i} R(x_i)\} \tag{3.11}$$

   where, for a observation x_i in cluster C_k, $R(x_i)$ is the quotient between the distance of x_i to the centroid of the cluster C_k (denominated as z_k) and between the smallest distance of x_i to the centroids of all the other clusters: $R(x_i) = \frac{||x_i - z_k||}{\underset{k \neq k'}{min} ||x_i - z_{k'}||}$

10. **PBM** [93]

    The PBM index (acronym formed of the initals of the names of its authors, Pakhira, Bandyopadhyay and Maulik) is given by:

$$PBM = \left(\frac{1}{K}\frac{E_T}{E_W}D_B\right)^2 \tag{3.12}$$

    where K is the number of clusters, E_W the sum of the distances of the points of each cluster to their centroid:

$$E_W = \sum_{k=1}^{K} \sum_{x_i \in C_k} ||x_i - z_k||$$

E_T the sum of the distances of all the observations to the centroid $\bar{x}$ of the entire dataset:

$$E_T = \sum_{x_i=1}^{n} ||x_i - \bar{x}||$$

and D_B the largest distance between two cluster centroids:

$$DB = \max_{k<k'} ||z_k - z_{k'}||$$

3.1.1 Implementations in artificial data

In order to determine which internal index works best, we performed all of the 11 clustering methods to 12 artificial datasets, where the true number of clusters K is known. Since K is known, one can compare its value to the one suggested by a given index and establish which index suggests the optimum number of clusters more often.

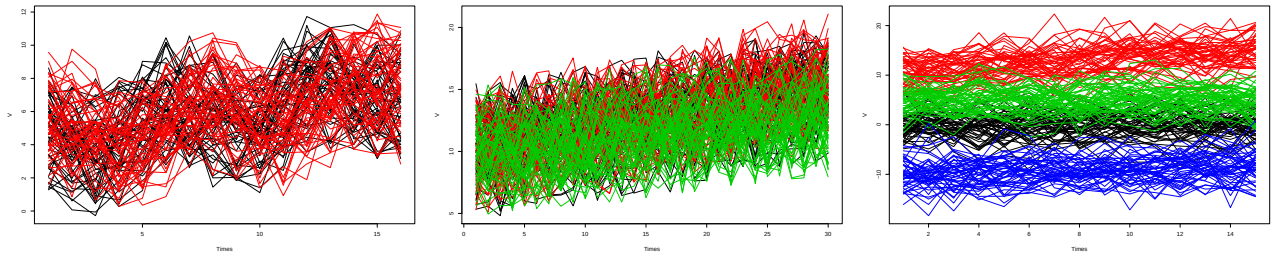
In 1985, Milligan and Cooper performed a study comparing 30 internal indices for four different hierarchical clustering methods while varying the number of clusters from 2 to 5 [87]. They concluded that even though Gamma and C index yielded good results, the Calinski-Harabasz outperformed all of the indices. Motivated by this study, in 2005, Shim et al. [112] compared 16 different indices for non-hierarchical methods and also deduced that the Calinski-Harabsz was the best among them. This latter study also indicated that Ray-Tury and Davies-Bouldin indices also yield interesting outcomes. Even more recently, in 2013, Arbelaitz et al. [7] tested 30 indices and revealed Silhouette, Davies-Bouldin and Calinski-Harabsz to be the best ones among its artificial datasets. Notwithstanding, both of these findings were done on non longitudinal artificial data. Thus, we present a new comparison of internal indices applied specifically to longitudinal data.

For the generation of the longitudinal artificial datasets, we used the function `gald` of the `kml` package. Using this function, one can choose the number of clusters, which trajectories are produced and how much noise is added. We chose to vary the number of clusters from 2 to 6. For the noise, two different types can be added: Personal Variation (PV) and Residual Variation (RV). The first one defines the personal variation between an individual and the mean trajectories of its cluster, while the second describes the noise of each trajectory within its own cluster. As it is shown in Table 3.2, we used a considerable amount of different trajectory

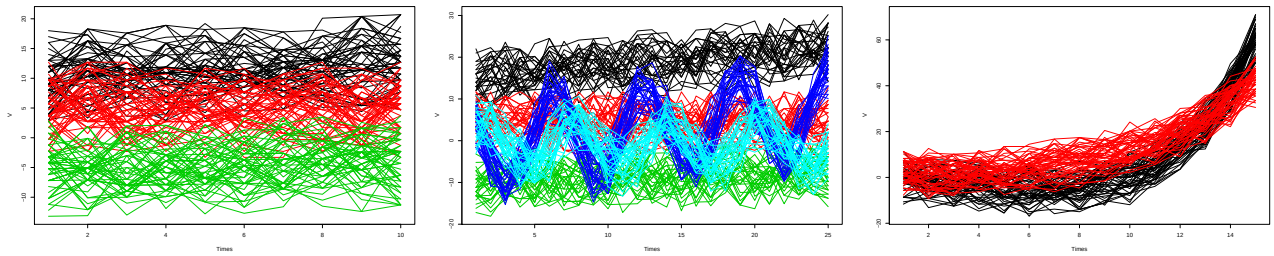
functions and different types of noise. The plots produced by the artificial datasets can be seen in Figure 3.1.

Table 3.2: Description of the artificial datasets along with their true number of clusters K

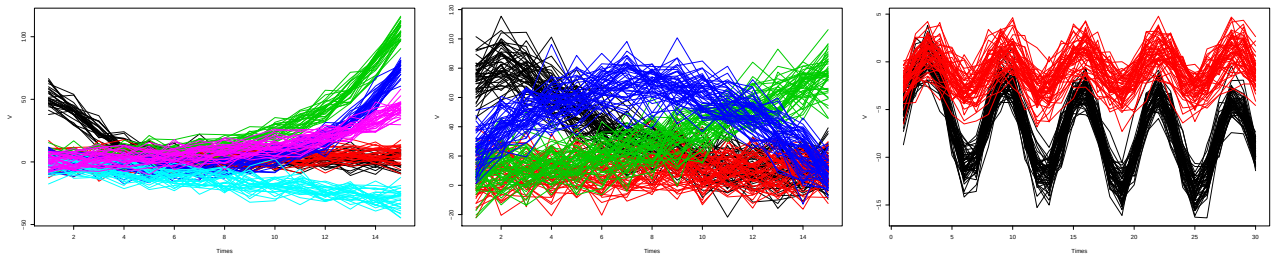
	Description	**K**
Dataset 1	$n = 100$; $f_1(t) = 0.2t - 3 + cos(t)$; $f_2(t) = 0.2t - 3 + sin(t)$ $t \in [1:16]$; $PV = \mathcal{U}(1,6)$; $RV = \mathcal{U}(1,6)$	2
Dataset 2	$n = 150$; $f_1(t) = 0.15t$; $f_2(t) = 0.2t$; $f_3(t) = 0.1t$ $t \in [1:30]$; $PV = \mathcal{U}(2,8)$; $RV = \mathcal{U}(2,8)$	3
Dataset 3	$n = 200$; $f_1(t) = 0.02t$; $f_2(t) = 0.3t + 10$; $f_3(t) = 5$; $f_4(t) = 0.1t - 10$ $t \in [1:15]$; $PV = \mathcal{N}(0,2)$; $RV = \mathcal{N}(0,2)$	4
Dataset 4	$n = 150$; $f_1(t) = 0.3t + 10$; $f_2(t) = 5$; $f_3(t) = 0.1t - 5$ $t \in [1:10]$; $PV = Random(-3,3)$; $RV = Random(-5.3,5.3)$	3
Dataset 5	$n = 150$; $f_1(t) = 0.3t + 15$; $f_2(t) = 5$; $f_3(t) = 0.1t - 10$, $f_4(t) = 0.3t + 10cos(t)$; $f_5(t) = 5sin(t)$; $t \in [1:25]$; $PV = Random(-3,3)$; $RV = Random(-5.3,5.3)$	5
Dataset 6	$n = 100$; $f_1(t) = -2t + e^{0.3t}$; $f_2(t) = e^{0.25t}$ $t \in [1:15]$; $PV = Random(-5,5)$; $RV = Random(-7,7.3)$	2
Dataset 7	$n = 180$; $f_1(t) = 140t/e^t$; $f_2(t) = e^{0.1t}$; $f_3(t) = e^{0.31t}$; $f_4(t) = -2t + e^{0.31t}$; $f_5(t) = -2t + e^{-0.3t}$; $f_6(t) = e^{0.25t}$; $t \in [1:15]$; $PV = \mathcal{U}(-5,5)$; $RV = \mathcal{N}(1,5)$	6
Dataset 8	$n = 200$; $f_1(t) = 100t/e^{0.5t}$; $f_2(t) = e^{-254t}$; $f_3(t) = 0.3t^2$; $f_4(t) = 15t - t^2$ $t \in [1:15]$; $PV = \mathcal{U}(-10,10)$; $RV = \mathcal{N}(10,10)$	4
Dataset 9	$n = 100$; $f_1(t) = -2log(t) - 5cos(t)$; $f_2(t) = -2cos(t) + e^{0.01t}$ $t \in [1:30]$; $PV = RV = \mathcal{N}(-1,1)$;	2
Dataset 10	$n = 100$; $f_1(t) = -2log(t) - 5cos(t)$; $f_2(t) = -2cos(t) + e^{0.01t}$; $f_3(t) = 5t - 0.3t^2$; $f_4(t) = 5t^{1/2}cos(t)$; $f_5(t) = 2t^{1/2}sin(t)$; $t \in [1:15]$; $PV = RV = \mathcal{N}(-1,1)$	5
Dataset 11	$n = 90$; $f_1(t) = \sqrt{\lvert 2log(t)\rvert + 5log(t)}$; $f_3(t) = 0.055t^2 - t$; $f_3(t) = t/2$ $t \in [1:30]$; $PV = RV = \mathcal{U}(0,5)$	3
Dataset 12	$n = 180$; $f_1(t) = 54t/e^{0.3t}$; $f_2(t) = e^{-2t}$; $f_3(t) = 0.3t^2$; $f_4(t) = 12t - t^2$; $f_5(t) = log(t)e^{0.16t}$; $f_6(t) = -e^{0.2t}$; $t \in [1:15]$; $PV = \mathcal{U}(-5,5)$; $RV = \mathcal{N}(5,5)$	6



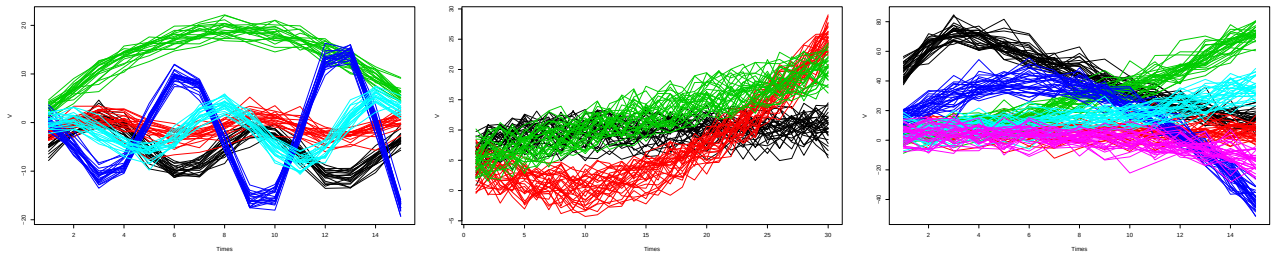
(a) Dataset 1 **(b)** Dataset 2 **(c)** Dataset 3



(d) Dataset 4 **(e)** Dataset 5 **(f)** Dataset 6



(g) Dataset 7 **(h)** Dataset 8 **(i)** Dataset 9



(j) Dataset 10 **(k)** Dataset 11 **(l)** Dataset 12

Figure 3.1: Trajectories of the artificial datasets

Table 3.3: Number of clusters suggested by the internal indices for each of the clustering methods troughout the datasets

Dataset 1

K=2	BH	CH	Γ	C	DB	S	BR	RT	WG	PBM
M1	2	2	2	2	2	2	6	2	2	2
M2	2	2	2	2	2	2	6	2	2	2
M3	3	2	2	2	2	2	6	2	2	2
M4	2	3	4	6	6	6	6	3	3	5
M5	2	3	4	6	6	2	6	6	2	2
M6	2	2	5	6	2	2	6	2	4	2
M7	6	2	3	3	2	2	4	2	2	2
M8	2	2	2	2	2	2	2	2	2	2
M9	2	2	2	3	2	2	6	2	2	2
M10	2	2	2	2	2	2	6	2	2	2
M11	2	2	2	3	2	2	4	2	2	2

Dataset 2

K=3	BH	CH	Γ	C	DB	S	BR	RT	WG	PBM
M1	2	2	3	3	2	2	6	2	2	2
M2	2	2	2	3	2	2	6	2	2	2
M3	2	2	6	6	2	2	6	2	2	2
M4	2	2	6	6	6	2	6	6	6	2
M5	2	2	6	6	6	3	6	6	6	2
M6	2	3	5	5	2	2	5	2	2	3
M7	2	6	6	6	6	2	6	6	4	4
M8	2	2	3	5	2	2	6	2	2	2
M9	2	2	4	4	2	2	6	2	2	4
M10	2	2	4	4	2	2	6	2	2	2
M11	2	2	4	4	2	2	6	2	2	2

Dataset 3

K=4	BH	CH	Γ	C	DB	S	BR	RT	WG	M1
M1	2	3	3	6	3	3	6	3	3	3
M2	2	3	3	5	3	3	6	3	3	3
M3	2	2	6	6	6	4	6	6	6	2
M4	3	6	2	2	2	-	-	2	2	2
M5	4	2	6	6	5	2	6	6	5	5
M6	2	3	3	4	3	3	6	3	3	3
M7	5	3	3	3	3	3	6	3	3	3
M8	2	4	4	6	3	3	6	3	2	3
M9	2	4	6	6	2	2	6	3	2	3
M10	2	3	3	6	2	2	6	3	3	3
M11	2	3	3	6	2	2	6	3	3	3

Dataset 4

K=3	BH	CH	Γ	C	DB	S	BR	RT	WG	M1
M1	2	2	3	5	2	2	6	2	2	3
M2	2	2	3	6	2	2	6	2	2	3
M3	2	2	6	6	5	2	6	4	6	2
M4	4	4	3	3	2	-	-	2	2	3
M5	2	2	6	6	6	4	6	6	6	2
M6	2	2	3	3	2	2	6	2	2	3
M7	3	2	2	6	2	2	6	2	2	2
M8	2	2	3	5	2	2	6	2	2	3
M9	2	2	2	5	2	2	6	2	2	3
M10	2	2	2	6	2	2	6	2	2	3
M11	2	2	2	6	2	2	6	2	2	2

Dataset 5

K=5	BH	CH	Γ	C	DB	S	BR	RT	WG	M1
M1	2	5	5	5	2	2	6	2	2	3
M2	2	3	6	6	3	3	6	3	3	3
M3	2	2	2	2	2	3	6	2	2	2
M4	2	6	3	3	3	-	-	2	2	3
M5	4	2	4	4	2	2	5	2	2	2
M6	2	4	6	6	4	4	6	4	4	4
M7	6	5	5	5	5	5	2	5	5	5
M8	2	5	5	5	2	2	6	2	2	3
M9	2	5	5	5	2	2	6	2	2	3
M10	2	5	5	5	2	2	6	2	2	3
M11	2	3	3	3	2	2	6	2	2	3

Dataset 6

K=2	BH	CH	Γ	C	DB	S	BR	RT	WG	M1
M1	2	2	4	4	2	2	6	2	2	2
M2	2	2	2	2	2	2	4	2	2	2
M3	2	2	6	6	6	2	6	6	6	2
M4	2	5	3	3	4	-	-	2	2	3
M5	2	2	6	6	2	2	6	2	2	3
M6	2	4	4	4	4	4	5	5	4	4
M7	2	4	4	4	5	4	5	4	4	4
M8	2	2	5	5	2	2	6	2	2	2
M9	2	2	2	6	2	2	6	2	2	2
M10	2	2	2	2	2	2	6	2	2	2
M11	2	2	6	6	2	2	6	2	2	3

Dataset 7

K=6	BH	CH	Γ	C	DB	S	BR	RT	WG	M1
M1	2	6	6	6	5	5	6	5	6	4
M2	2	6	6	6	6	6	6	2	6	6
M3	2	6	2	2	2	-	-	2	2	4
M4	3	2	2	2	2	-	-	2	2	2
M5	4	3	5	5	2	2	6	2	2	2
M6	3	6	6	6	6	6	6	6	6	6
M7	2	6	6	6	5	5	6	5	6	6
M8	2	6	6	6	6	6	6	2	6	6
M9	2	6	6	6	5	5	6	5	6	4
M10	2	6	6	6	5	5	6	5	4	5
M11	2	5	5	5	2	4	6	4	4	3

Dataset 8

K=4	BH	CH	Γ	C	DB	S	BR	RT	WG	PBM
M1	2	4	4	4	4	4	6	4	4	4
M2	2	4	4	4	4	4	5	4	4	4
M3	2	2	6	6	3	3	6	3	3	2
M4	4	3	2;3	2;3	2	-	-	2	2	2
M5	5	2	6	6	2	3	6	5	5	6
M6	2	4	4	4	4	4	6	4	4	4
M7	5	4	4	4	4	4	6	4	4	4
M8	2	4	4	4	4	4	6	4	4	4
M9	2	4	4	4	4	4	6	4	4	4
M10	2	4	4	4	4	4	6	4	4	4
M11	2	4	4	4	4	3	6	4	4	4

Dataset 9

K=2	BH	CH	Γ	C	DB	S	BR	RT	WG	M1
M1	2	2	2	2	2	2	6	2	2	2
M2	2	2	2	2	2	2	6	2	2	2
M3	2	2	2	2	2	2	6	2	2	2
M4	4	6	2	2	2	-	-	2	2	4
M5	2	2	6	6	2	2	4	2	2	2
M6	3	2	2	2	2	2	6	2	2	3
M7	2	3	3	3	6	3	6	5	3	3
M8	2	2	2	2	2	2	6	2	2	2
M9	2	2	2	2	2	2	6	2	2	2
M10	2	2	2	2	2	2	6	2	2	2
M11	2	2	2	2	2	2	6	2	2	2

Dataset 10

K=5	BH	CH	Γ	C	DB	S	BR	RT	WG	M1
M1	2	5	3	3	3	3	6	3	3	5
M2	2	6	3	3	3	3	6	3	3	4
M3	2	4	2	2	2	2	6	2	2	4
M4	3	2	2	2	2	-	-	2	2	2
M5	2	4	6	6	4	4	6	3	6	4
M6	2	5	3	3	3	3	6	3	3	5
M7	2	5	3	3	3	3	6	3	3	5
M8	2	6	2	2	2	2	6	2	2	6
M9	2	5	3	3	3	3	6	3	3	4
M10	2	6	3	3	3	3	6	3	3	4
M11	2	2	2	2	2	2	6	2	2	2

Dataset 11

K=3	BH	CH	Γ	C	DB	S	BR	RT	WG	M1
M1	2	3	3	3	3	3	6	3	3	3
M2	2	3	3	3	3	3	6	3	3	3
M3	2	2	6	6	6	2	6	4	6	2
M4	2	2	6	6	6	-	-	6	6	3
M5	4	2	2	4	3	2	6	2	2	2
M6	4	5	5	5	2	2	5	2	5	5
M7	4	2	3	2	3	2	6	2	2	2
M8	2	3	3	3	3	3	6	3	3	3
M9	2	3	3	3	3	3	6	3	3	3
M10	2	3	3	3	3	3	6	3	3	3
M11	2	2	5	5	2	2	6	2	2	3

Dataset 12

K=6	BH	CH	Γ	C	DB	S	BR	RT	WG	PBM
M1	2	5	5	5	4	4	6	4	5	5
M2	2	5	5	5	5	5	6	5	5	5
M3	2	2	4	4	4	4	6	2	4	2
M4	2	6	3	5	6	-	-	3	3	2
M5	2	2	6	6	3	2	6	3	2	2
M6	2	5	5	5	4	4	6	4	5	5
M7	2	5	5	5	4	4	6	4	5	5
M8	2	5	5	5	4	4	6	4	5	5
M9	2	5	6	6	3	3	6	3	5	5
M10	2	5	6	6	3	3	6	3	5	5
M11	2	4	5	5	4	4	6	2	2	4

As Table 3.3 and Table 3.4 reveal, after running all of the clustering methods with the 12 artificial datasets, the Calinski-Harabsz was the best index with a total of 59 correct guesses, followed by the Gamma index with 54. Moreover, Silhouette and Banfeld-Raftery indices did not converge in every situation in oppose to the others, therefore indicating less reliability.

Table 3.4: Correct numbers of K suggested by the indices

	BH	CH	Γ	C	DB	S	BR	RT	WG	PBM
Dataset 1	9	9	7	5	9	10	1	9	9	10
Dataset 2	0	1	2	2	0	1	0	0	0	1
Dataset 3	1	2	1	1	0	1	0	0	0	0
Dataset 4	1	0	5	2	0	0	0	0	0	1
Dataset 5	0	4	4	4	1	1	1	1	1	1
Dataset 6	11	8	3	2	7	8	0	8	8	6
Dataset 7	0	8	7	7	3	3	8	1	6	4
Dataset 8	0	8	8	8	8	7	0	8	8	8
Dataset 9	9	9	9	9	10	9	0	10	10	8
Dataset 10	0	4	0	0	0	0	0	0	0	3
Dataset 11	0	5	6	5	7	5	0	5	5	7
Dataset 12	0	1	2	2	1	0	11	0	0	0
Total	**31**	**59**	**54**	**47**	**46**	**45**	**21**	**42**	**47**	**49**

Hence, our study seems to corroborate the conclusions achieved by Milligan and Cooper [87] and Shiem et al [112] that the CH is the best index. Furthermore, looking individually at each clustering method, the CH was almost the best, or among the best, indicating the correct number of clusters. For this reason, from now on, we will consider CH as the main indicator of the number of clusters in our work.

3.2 External Indices

After running any type of clustering method, some kind of evaluation is needed to perceive if the method is efficient. Just knowing that a clustering method is more suitable to identify the right number of clusters is not an indicator of better outcomes. A method can correctly guess the number of clusters but still be less accurate than a method that failed the real number of clusters. Identifying the right number of clusters is not the only aim. External validity indices can be used to compare the true partition with the partition obtained from the clustering. However, in order to do so, a knowledge of the true partition is needed, albeit in the real world that very rarely happens. Hence, a common way to unravel this problem is with the use of simulated data.

It is relevant to note that some clustering algorithms may find the right clusters but label them in a different way. For instance, consider a dataset with six observations $X = \{x_1, x_2, x_3, x_4, x_5, x_6\}$ and three clusters C_1, C_2 and C_3 in its composition: $C_1 = \{x_1, x_2\}$, $C_2 = \{x_3, x_4\}$ and $C_3 = \{x_5, x_6\}$. After running a clustering algorithm with the right number of clusters, it might still indicate them as $C_1 = \{x_3, x_4\}$, $C_2 = \{x_5, x_6\}$ and $C_3 = \{x_1, x_2\}$. In this situation, the algorithm correctly guessed the clusters with a proper grouping but labelled C_1 as cluster C_2, C_2 as cluster C_3 and C_3 as cluster C_1. This is known as *label switching* and was first mentioned by Redner and Walker in 1984 [104].

To overcome this, all of the external indices are based on the occurrence of pairs of observations in each partition. Pair counting was first applied scientifically by Thurstone in 1927 through his *Law of Comparative Judgment* [77]. To apply a pair counting approach, firstly all the members of one clustering partition are incrementally paired [95]. These pairs are then compared to all the similarly paired members of the other clustering partition. Given two partitions P and Q, the contingency table exposes the number of pairs classified as belonging or not belonging to them. The use of pairs to evaluate clustering methods can be found abundantly in the literature [7, 14, 29, 60].

Let us consider two different partitions $P = \{p_1, p_2, p_3, \ldots, p_r\}$ and $Q = \{q_1, q_2, q_3, \ldots, q_s\}$ into r and s clusters respectively. Also, let n_{ij} denote the number of objects that are both in clusters p_i and q_j. Then a pair of objects is called concordant if they belong to the same cluster in both partitions or if they do not belong to the same cluster in both partitions:

Table 3.5: Contigency table for the comparinson of two partitions

	q_1	q_2	$\dots$	q_s
p_1	n_{11}	n_{12}	$\dots$	n_{1s}
p_2	n_{21}	n_{22}	$\dots$	n_{2s}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
p_r	n_{r1}	n_{r2}	$\dots$	n_{rs}

A simplified version of this table consists in considering the following two-by-two concordance table [25, 33, 95]:

Table 3.6: Concordance table for each pair of observations

		Q	
		Pairs in Q	Pairs not in Q
P	Pairs in P	a	b
	Pairs not in P	c	d

As is demonstrated in the previous table, there are 4 possibilities (a, b, c, d) to take into account:

a: The pair of observations belong to the same cluster according to P and Q

b: The pair of observations belong to the same cluster according to P but not Q

c: The pair of observations belong to the same cluster according to Q but not P

d: The pair of observations do not belong to the same cluster according to either P or Q

Example 3.2.1. Let us consider a simple example adapted from [114] where X is a dataset with four observations $X = \{x_1, x_2, x_3, x_4\}$. If $P = \{\{x_1, x_2, x_4\}, \{x_3\}\}$ and $Q = \{\{x_1, x_4, x_3\}, \{x_2\}\}$, then:

	x_1x_2	x_1x_3	x_1x_4	x_2x_3	x_2x_4	x_3x_4	*total*
a	0	0	1	0	0	0	1
b	1	0	0	0	1	0	2
c	0	1	0	0	0	1	2
d	0	0	0	1	0	0	1

In this example we have $a = 1$, $b = 2$, $c = 2$ and $d = 1$ and can write the concordance table as:

		Q	
		Pairs in Q	Pairs not in Q
P	Pairs in P	1	2
	Pairs not in P	2	1

In the information retrieval context, one of the most used metrics to evaluate results are the precision, recall and F1-score:

$$Precision = \frac{a}{a+c} \tag{3.13}$$

$$Recall = \frac{a}{a+b} \tag{3.14}$$

$$F1 = 2\frac{Precision \cdot Recall}{Precision + Recall} \tag{3.15}$$

In the clustering context, there are a lot more approaches to evaluate the results via indices. With a considerable amount of external indices available, we chose to explore 10 of them. The external indices we will discuss are presented in Table 3.7 along with their abbreviations and references in a chronological order. All of the indices use the notation established in the previous concordance table.

Table 3.7: External indices with their respective abbreviation, date and reference

Index	Abbreviation	Date	Ref
Jaccard	JA	1908	[62]
Kulczynski	KU	1927	[72]
Sørensen–Dice	SD	1945	[26]
Russel-Rao	RR	1948	[102]
Rogers-Tanimoto	RT	1960	[105]
Sokal-Sneath 1	SS1	1963	[116]
Sokal-Sneath 2	SS2	1963	[116]
Rand	RI	1971	[100]
Folkes-Mallow	FM	1983	[40]
Adjusted Rand	ARI	1985	[60]

Unlike internal indices, each one of the external indices aims to maximize its value. The value 1 represents the perfect agreement between the partitions, except Russel-Rao index, where its

maximum value depends on the number of clusters but cannot surpass 0.5. Some indices like the ARI can sporadically get negative values, meaning that the the value is under the expected.

1. **Jaccard (JA)** [62]

$$\text{Jaccard} = \frac{a}{a+b+c} \tag{3.16}$$

2. **Kulczynski (KU)** [72]

$$\text{Kulczynski} = \frac{1}{2}\left(\frac{a}{a+b} + \frac{a}{a+c}\right) \tag{3.17}$$

The Kulczynski index can also be seen as the arithmetic mean of the precision and recall coefficients:

$$\text{Kulczynski} = \frac{1}{2}(Precision + Recall) \tag{3.18}$$

3. **Sørensen–Dice (SD)** [26]

$$\text{SD} = \frac{2a}{2a+b+c} \tag{3.19}$$

This index can also be expressed as the harmonic mean of the precision and recall coefficients. It is also identical to the F1-score defined previously.

$$\text{SD} = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}} = 2\left(\frac{Precision \cdot Recall}{Precision + Recall}\right) \tag{3.20}$$

4. **Russel-Rao (RR)** [102]

$$\text{RR} = \frac{a}{a+b+c+d} \tag{3.21}$$

5. **Rogers-Tanimoto (RT)** [105]

$$\text{RT} = \frac{a+d}{a+d+2(b+c)} \tag{3.22}$$

6. **Sokal-Sneath 1 (SS1)** [116]

The Sokal-Sneath index has multiple variations. The first one is given by:

$$\text{SS1} = \frac{a}{a+\frac{1}{2}(b+c)} \tag{3.23}$$

7. **Sokal-Sneath 2 (SS2)** [116]

   The second version of the SS index is:

$$\text{SS2} = \frac{a+d}{a+d+\frac{1}{2}(b+c)} \tag{3.24}$$

   We consider two variants of the SS index, but there are more that can bee seen in [3].

8. **Rand (RI)** [100]

   The Rand index is defined as the fraction of agreement, i.e. the number of pairs of observations that are either in the same cluster in both partitions or in different clusters according to both partitions, divided by total of pairs:

$$\text{Rand} = \frac{a+d}{a+b+c+d} \tag{3.25}$$

9. **Folkes-Mallows (FM)** [40]

$$\text{FM} = \frac{a}{\sqrt{(a+b)(a+c)}} \tag{3.26}$$

   Identically, it also equals the geometric mean between the Precision and Recall coefficients:

$$\text{FM} = \sqrt{Precision \cdot Recall} \tag{3.27}$$

10. **Adjusted Rand (ARI)** [60]

    The ARI, also known as Corrected Rand Index, is a rectified version of the RI. One problem that RI yields is that its expected value of two random partitions does not take a constant value, like zero for example. Whereas in RI, the range of values is only between 0 and 1, ARI adjusts the score so that the expected value is 0 (for two random partitions) and the index has a range of values between -1 and 1 [126]. On top of that, Milligan and Cooper shown in 1986 [88], that the ARI outperforms the Jaccard, RI, and FM. They even recommended ARI when comparing partitions with different number of clusters.

$$ARI = \frac{a - \frac{(a+c)(a+b)}{a+b+c+d}}{\frac{2a+b+c}{2} - \frac{(a+b)(a+c)}{a+b+c+d}} \tag{3.28}$$

3.2.1 Implementations in artificial data

For the evaluation of clustering algorithms, we used the artificial datasets introduced in Section 3.1.1 (displayed in Table 3.2). For the tests, all of the clustering algorithms were executed with their real number of clusters. Given that all of the indices aim to maximize their value, we will seek for the methods that obtain the highest values overall.

Table 3.8: Evaluation of clustering methods with the use of external indices

Dataset 1

	JA	KU	SD	RR	RT	SS1	SS2	RI	FM	ARI
M1	0.345693	0.513804	0.513776	0.256162	0.346939	0.208965	0.68	0.515151	0.51379	0.030347
M2	0.341231	0.508846	0.508833	0.253131	0.343466	0.205713	0.676648	0.511313	0.508839	0.022626
M3	0.368056	0.538085	0.538071	0.267677	0.370242	0.225319	0.701639	0.540404	0.538078	0.080808
M4	0.510908	0.677599	0.676293	0.350101	0.497957	0.3431	0.79869	0.664849	0.676954	0.330244
M5	0.503293	0.669977	0.669588	0.339596	0.497957	0.336267	0.79869	0.664849	0.669782	0.32996
M6	0.33624	0.50363	0.503263	0.249293	0.340374	0.202096	0.673633	0.507879	0.503263	0.015673
M7	0.330931	0.497262	0.49718	0.249293	0.329573	0.198201	0.662885	0.495758	0.497221	-0.00832
M8	0.363243	0.533345	0.53291	0.271515	0.355051	0.221929	0.687699	0.52404	0.533128	0.048543
M9	0.349731	0.518525	0.518224	0.262828	0.343466	0.211924	0.676648	0.511313	0.518374	0.023009
M10	0.341231	0.508846	0.508833	0.253131	0.343466	0.205713	0.676648	0.511313	0.508839	0.022626
M11	0.34818	0.516605	0.516519	0.25899	0.346939	0.210786	0.68	0.515151	0.516562	0.030457

Dataset 2

	JA	KU	SD	RR	RT	SS1	SS2	RI	FM	ARI
M1	0.21417	0.352937	0.352784	0.118479	0.394	0.119928	0.722273	0.56528	0.352861	0.025727
M2	0.219226	0.359996	0.359616	0.122237	0.393392	0.123107	0.721762	0.564653	0.359806	0.03038
M3	0.211393	0.34916	0.349008	0.117226	0.391483	0.118188	0.720151	0.562685	0.349084	0.019978
M4	0.207055	0.343075	0.343075	0.112931	0.396177	0.115483	0.724097	0.567517	0.343075	0.020725
M5	0.208613	0.345217	0.345211	0.114005	0.396177	0.116453	0.724097	0.567517	0.345214	0.022352
M6	0.195508	0.327086	0.327071	0.108277	0.383558	0.108345	0.713373	0.554452	0.327078	-0.00593
M7	0.202871	0.337383	0.337311	0.112573	0.386649	0.112886	0.716034	0.557673	0.337347	0.005472
M8	0.216842	0.356427	0.356401	0.11821	0.401605	0.121605	0.728597	0.573065	0.356414	0.037027
M9	0.24392	0.402432	0.392179	0.153468	0.355285	0.1389	0.687918	0.524295	0.397273	0.017504
M10	0.217722	0.357591	0.357589	0.117852	0.405042	0.122159	0.731411	0.576555	0.35759	0.041788
M11	0.22517	0.368597	0.367573	0.127606	0.389753	0.126868	0.718684	0.560895	0.368085	0.032699

Dataset 3

	JA	KU	SD	RR	RT	SS1	SS2	RI	FM	ARI
M1	0.733133	0.846022	0.846021	0.208593	0.858858	0.578698	0.960537	0.92407	0.846021	0.795632
M2	0.655597	0.79198	0.791976	0.195427	0.81379	0.487649	0.945891	0.897337	0.791978	0.723829
M3	0.149097	0.258144	0.257988	0.065126	0.454944	0.07997	0.769516	0.625377	0.258066	0.007627
M4	0.244524	0.60848	0.39296	0.238995	0.150489	0.139292	0.414722	0.261608	0.488987	0.000308
M5	0.209818	0.39569	0.346859	0.131658	0.337051	0.117205	0.670364	0.504171	0.370471	0.021173
M6	0.679756	0.809477	0.809351	0.201809	0.826358	0.514872	0.95009	0.904925	0.809414	0.746036
M7	0.679756	0.809477	0.809351	0.201809	0.826358	0.514872	0.95009	0.904925	0.809414	0.746036
M8	0.524446	0.702056	0.688048	0.197286	0.696505	0.355423	0.901766	0.821106	0.695016	0.566123
M9	0.628803	0.772125	0.772104	0.191106	0.797245	0.45858	0.940221	0.887186	0.772115	0.697146
M10	0.557046	0.735302	0.715517	0.210754	0.712933	0.386046	0.908542	0.832412	0.725342	0.601216
M11	0.699539	0.823426	0.823211	0.20603	0.837404	0.537917	0.953705	0.911508	0.823318	0.764214

Dataset 4

	JA	KU	SD	RR	RT	SS1	SS2	RI	FM	ARI
M1	0.749228	0.856643	0.282327	0.827324	0.599013	0.950409	0.950409	0.905503	0.856641	0.786164
M2	0.702034	0.824938	0.824936	0.271767	0.793164	0.540873	0.938797	0.884653	0.824937	0.738928
M3	0.356219	0.525401	0.525312	0.175034	0.519375	0.216707	0.812118	0.683669	0.525357	0.288172
M4	0.316474	0.608417	0.480791	0.291723	0.226943	0.187983	0.540075	0.369933	0.540852	0.002521
M5	0.248401	0.400602	0.39795	0.142461	0.397574	0.141814	0.725261	0.568949	0.399274	0.065696
M6	0.740907	0.85118	0.851174	0.280716	0.821219	0.588445	0.948384	0.901834	0.851177	0.777939
M7	0.231138	0.375543	0.375487	0.125011	0.41259	0.130671	0.737502	0.584161	0.375515	0.063866
M8	0.749228	0.856643	0.856639	0.282327	0.827324	0.599013	0.950409	0.905503	0.856641	0.786164
M9	0.60715	0.756666	0.755561	0.258345	0.713563	0.435905	0.908798	0.832841	0.756114	0.628837
M10	0.7248	0.841013	0.840445	0.283758	0.805477	0.568381	0.943063	0.89226	0.840729	0.759199
M11	0.733318	0.846143	0.846143	0.2783	0.816187	0.578928	0.946699	0.898792	0.846143	0.770738

Dataset 5

	JA	KU	SD	RR	RT	SS1	SS2	RI	FM	ARI
M1	0.973243	0.98644	0.98644	0.192036	0.989462	0.94788	0.997354	0.99472	0.98644	0.983162
M2	0.769138	0.869689	0.869506	0.171723	0.901966	0.624878	0.973546	0.948456	0.869598	0.8374
M3	0.658667	0.794213	0.794212	0.15472	0.851545	0.491054	0.958236	0.919821	0.794213	0.744423
M4	0.193081	0.572256	0.323668	0.184788	0.128503	0.106856	0.370991	0.227741	0.430373	0.001119
M5	0.361164	0.539916	0.530669	0.118837	0.652618	0.220378	0.882556	0.789799	0.532728	0.398215
M6	1	1	1	0.194631	1	1	1	1	1	1
M7	0.155359	0.269623	0.268937	0.055123	0.538832	0.084222	0.823746	0.700313	0.26928	0.081065
M8	0.635447	0.791522	0.777093	0.174855	0.817664	0.465682	0.947195	0.899687	0.784274	0.713888
M9	0.78483	0.880099	0.879445	0.175928	0.907973	0.645861	0.975288	0.951767	0.879768	0.849326
M10	0.79798	0.888089	0.887641	0.176734	0.914347	0.663866	0.977117	0.955257	0.887865	0.859725
M11	0.543332	0.704109	0.704103	0.13745	0.792877	0.372997	0.938696	0.884474	0.704106	0.632329

Dataset 6

	JA	KU	SD	RR	RT	SS1	SS2	RI	FM	ARI
M1	1	1	1	0.5	1	1	1	1	1	1
M2	1	1	1	0.5	1	1	1	1	1	1
M3	1	1	1	0.5	1	1	1	1	1	1
M4	0.49	0.737475	0.657718	0.485051	0.329038	0.324503	0.662343	0.495515	0.696456	0
M5	0.625497	0.76961	0.769607	0.381616	0.628022	0.455071	0.871023	0.771515	0.769608	0.543001
M6	0.330458	0.496785	0.496759	0.247677	0.331719	0.197934	0.665049	0.498182	0.496772	-0.00359
M7	0.330458	0.496785	0.496759	0.247677	0.331719	0.197934	0.665049	0.498182	0.496772	-0.00359
M8	1	1	1	0.5	1	1	1	1	1	1
M9	1	1	1	0.5	1	1	1	1	1	1
M10	0.617778	0.764868	0.763762	0.393131	0.60871	0.446945	0.861546	0.756768	0.764302	0.513875
M11	0.493896	0.666606	0.661218	0.359596	0.46147	0.327929	0.774146	0.631515	0.663907	0.264408

Dataset 7

	JA	KU	SD	RR	RT	SS1	SS2	RI	FM	ARI
M1	1	1	1	0.162011	1	1	1	1	1	1
M2	0.628588	0.778128	0.771942	0.137306	0.849917	0.458351	0.95772	0.91887	0.775029	0.723077
M3	0.468689	0.638243	0.638241	0.1036	0.789011	0.30607	0.937615	0.882557	0.638242	0.568141
M4	0.260046	0.567048	0.412757	0.139789	0.430855	0.149456	0.751743	0.602235	0.487899	0.220624
M5	0.346272	0.526208	0.514416	0.098014	0.687706	0.209389	0.898047	0.81496	0.520279	0.403266
M6	0.681225	0.822118	0.810391	0.1491	0.86956	0.516559	0.963854	0.93023	0.816234	0.768356
M7	0.71396	0.847957	0.833112	0.155556	0.882669	0.555162	0.967837	0.937679	0.840502	0.795618
M8	0.679487	0.81991	0.80916	0.148045	0.869452	0.514563	0.963821	0.930168	0.814518	0.767109
M9	1	1	1	0.162011	1	1	1	1	1	1
M10	0.955447	0.977217	0.977216	0.158411	0.985335	0.914695	0.996293	0.992613	0.977216	0.972808
M11	0.624153	0.768664	0.768588	0.12576	0.859204	0.45365	0.960645	0.924271	0.768626	0.723322

Dataset 8

	JA	KU	SD	RR	RT	SS1	SS2	RI	FM	ARI
M1	1	1	1	0.246231	1	1	1	1	1	1
M2	0.98	0.989899	0.989899	0.243769	0.9901	0.960784	0.997506	0.995025	0.989899	0.986599
M3	1	1	1	0.246231	1	1	1	1	1	1
M4	0.383713	0.661727	0.554614	0.228493	0.463074	0.237404	0.775272	0.633015	0.605808	0.319712
M5	0.59274	0.763145	0.744302	0.217437	0.740043	0.421201	0.919271	0.850603	0.753665	0.642419
M6	1	1	1	0.246231	1	1	1	1	1	1
M7	1	1	1	0.246231	1	1	1	1	1	1
M8	1	1	1	0.246231	1	1	1	1	1	1
M9	1	1	1	0.246231	1	1	1	1	1	1
M10	1	1	1	0.246231	1	1	1	1	1	1
M11	0.691371	0.817841	0.817527	0.205327	0.832075	0.528317	0.95197	0.908342	0.817684	0.756359

Dataset 9

	JA	KU	SD	RR	RT	SS1	SS2	RI	FM	ARI
M1	1	1	1	0.5	1	1	1	1	1	1
M2	1	1	1	0.5	1	1	1	1	1	1
M3	1	1	1	0.5	1	1	1	1	1	1
M4	0.49	0.737475	0.657718	0.485051	0.329038	0.324503	0.662343	0.495152	0.696456	0
M5	0.459277	0.653261	0.629458	0.385051	0.376147	0.298092	0.706897	0.546667	0.641249	0.097499
M6	1	1	1	0.5	1	1	1	1	1	1
M7	1	1	1	0.5	1	1	1	1	1	1
M8	1	1	1	0.5	1	1	1	1	1	1
M9	1	1	1	0.5	1	1	1	1	1	1
M10	1	1	1	0.5	1	1	1	1	1	1
M11	1	1	1	0.5	1	1	1	1	1	1

Dataset 10

	JA	KU	SD	RR	RT	SS1	SS2	RI	FM	ARI
M1	1	1	1	0.191919	1	1	1	1	1	1
M2	0.648148	0.803664	0.786517	0.176768	0.824885	0.479452	0.949602	0.90404	0.795044	0.72632
M3	1	1	1	0.191919	1	1	1	1	1	1
M4	0.254841	0.562459	0.406177	0.164849	0.34959	0.146027	0.682459	0.51798	0.47797	0.159973
M5	0.65748	0.800946	0.793349	0.168687	0.834401	0.489736	0.954041	0.912121	0.797138	0.738215
M6	1	1	1	0.191919	1	1	1	1	1	1
M7	1	1	1	0.191919	1	1	1	1	1	1
M8	0.63037	0.788023	0.773285	0.171919	0.816847	0.460249	0.946921	0.899192	0.780619	0.710022
M9	1	1	1	0.191919	1	1	1	1	1	1
M10	0.500645	0.690398	0.66724	0.156768	0.72956	0.333907	0.915187	0.843636	0.67872	0.569483
M11	0.370253	0.548636	0.540416	0.118182	0.665265	0.227185	0.888265	0.79899	0.54451	0.414236

Dataset 11

	JA	KU	SD	RR	RT	SS1	SS2	RI	FM	ARI
M1	1	1	1	0.325843	1	1	1	1	1	1
M2	0.482176	0.676402	0.650632	0.263421	0.558972	0.317675	0.835248	0.717104	0.663392	0.427788
M3	0.37487	0.545374	0.54531	0.179526	0.539201	0.230671	0.823961	0.700624	0.545345	0.322199
M4	0.487959	0.68652	0.657534	0.269663	0.561404	0.324324	0.836601	0.719101	0.671871	0.435995
M5	0.49161	0.688511	0.659167	0.270662	0.562622	0.325917	0.837277	0.7201	0.673679	0.438393
M6	0.318819	0.448354	0.383492	0.159051	0.492732	0.18964	0.795308	0.660175	0.483513	0.230316
M7	0.658667	0.794213	0.794212	0.15472	0.851545	0.491054	0.958236	0.919821	0.794213	0.744423
M8	1	1	1	0.325843	1	1	1	1	1	1
M9	1	1	1	0.325843	1	1	1	1	1	1
M10	1	1	1	0.325843	1	1	1	1	1	1
M11	0.639528	0.783033	0.780137	0.270662	0.735269	0.470078	0.917421	0.847441	0.781583	0.663993

Dataset 12

	JA	KU	SD	RR	RT	SS1	SS2	RI	FM	ARI
M1	0.916758	0.956573	0.956572	0.155183	0.97221	0.84631	0.992905	0.985909	0.956573	0.948162
M2	0.727635	0.858988	0.842348	0.158535	0.887964	0.571876	0.969422	0.940658	0.850628	0.806667
M3	0.465033	0.635072	0.634843	0.104842	0.784745	0.30296	0.935826	0.879392	0.634957	0.562641
M4	0.160831	0.554521	0.277096	0.153383	0.11092	0.087557	0.332902	0.19969	0.391989	0.000679
M5	0.339453	0.542018	0.506853	0.11018	0.646903	0.204422	0.879928	0.785599	0.524141	0.381031
M6	0.682051	0.822009	0.810976	0.148603	0.870428	0.51751	0.96412	0.930726	0.816474	0.769263
M7	0.977652	0.9887	0.9887	0.160211	0.992702	0.95628	0.998166	0.996338	0.9887	0.986514
M8	0.680627	0.820843	0.809968	0.148293	0.869886	0.515871	0.963954	0.930416	0.815387	0.768067
M9	0.854035	0.921407	0.921272	0.151086	0.949655	0.745254	0.98692	0.974178	0.921339	0.905831
M10	0.769935	0.871156	0.870015	0.146245	0.91626	0.62593	0.977662	0.9563	0.870585	0.843791
M11	0.548031	0.70944	0.708036	0.12005	0.819825	0.37744	0.947918	0.900993	0.708737	0.648567

As it is shown throughout Table 3.8, K-means with Euclidean distance (M1) was the best method overall, followed by Hierarchical with Euclidean distance (M9). K-means with Mahalanobis distance with raw data (M3) and model-based methods like Gaussian mixture model (M6) and mixture of multivariate t-distributions (M7) also reached fine outcomes but presented more inconsistency. The Bayesian Hierarchical (M11) was the worst method by far, as it could not stand out in any of the datasets.

The results also seem to indicate that, for the previous datasets, the non-parametric methods achieved better results than the model-based ones.

Chapter 4

Application to COVID-19 data

This chapter focus on the application of clustering to two different types of data associated with the new coronavirus. One is based on the Community Mobility Reports published by Google, while the other addresses the number of deaths over time. For the first data, before the application of clustering, we initially define a confinement index that takes into account several variables of mobility changes. Regarding the data of the deceased, we employ an additional variable and therefore resort to multivariate clustering.

In both cases, we use the Calinski-Harabasz internal index to indicate the number of clusters, as it was shown in Chapter 3 to outperform the other indices.

4.1 Complexity of COVID-19 analysis

In the analysis of COVID-19 statistics, there is a lot of misconception. Some people argue that the number of infections should be divided by the total number of inhabitants of each country while some argue that it should be divided by the total number of tests. Moreover, the trajectories can also be seen in a logarithmic scale instead of a linear one, thereby possibly providing different conclusions [80]. This uncertainty, along with the incorrect use of statistics concepts, has led to some deceitful results shared throughout the media [21]. To make things worse, the available data might not correspond to the real numbers out there since not everyone takes the test and some non-asymptomatic individuals might have the virus without even knowing it.

Regarding the comparison between different countries, many studies report the number

of cases per million inhabitants. Nevertheless, not everyone agrees with that procedure. For instance, let us consider two countries free of COVID-19 infections, one with 10 million inhabitants and another with 200 million. Imagine 5 infected individuals enter each country and that in a week each one of them infects 5 individuals, leading to 25 new cases. By the end of the week, one country would have 3 cases per million of inhabitants while the other would have 0.15 per million of inhabitants. In the beginning of the epidemic, the values in cases per million inhabitants do not report about the dynamics of the infection and its effect on health care [89].

A country capable of testing more population will likely detect more cases than a country with less capability. However, even when considering the number of tests, there is an added complexity. Not all of the countries are doing the same type of tests, which makes the true positives to vary depending on the type [89].

For these reasons, in order to avoid any erroneous result regarding the infections and number of tests, we decided to not consider them in our analysis.

4.2 Community Mobility Reports

One intriguing aspect to study about the coronavirus pandemic is the mobility during the lockdown period. In the course of the COVID-19 pandemic, multiple governmental authorities enacted quarantines in an effort to control the rapid spread of the virus. In late march, around 2.6 billion people worldwide (around a third of the world's population) were under some form of lockdown. Undoubtedly, people had to adapt to new daily routines. In addition to the physical and psychological impacts of the quarantine [120], the economy and other activities also suffered a big impact.

The Community Mobility Reports (CMR) released by Google aims to provide insights into what has changed, in each country, in response to policies aimed at combating COVID-19. The data measures the following six mobility changes compared to their baseline value (in percentage):

1. **Retail and Recreation** (m_1): Mobility trends for places like restaurants, cafes, shopping centers, theme parks, museums, libraries, and movie theaters.

2. **Grocery and Pharmacy** (m_2): Mobility trends for places like grocery markets, food warehouses, farmers markets, specialty food shops, drug stores, and pharmacies.

3. **Parks** (m_3): Mobility trends for places like national parks, public beaches, marinas, dog parks, plazas, and public gardens.

4. **Transit stations** (m_4): Mobility trends for places like public transport hubs such as subway, bus, and train stations.

5. **Workplaces** (m_5): Mobility trends for places of work.

6. **Residential** (m_6): Mobility trends for places of residence

To take in consideration these 6 variables, we defined a *Confinement Index* (CI) which summarizes the six variables above in a single variable and explains about 85% of the information contained in those variables. The CI was built performing Principal Component Analysis (PCA) to a matrix with the average values of each variable for each country. We then used the coefficients of the first eigenvector $V_1 = (-0.35, -0.25, -0.85, -0.24, -0.19, 0.1)$ obtained from PCA and multiplied them by the mobility variables. Hence, our index is defined as the following:

$$CI = -0.35m_1 - 0.25m_2 - 0.85m_3 - 0.24m_4 - 0.19m_5 + 0.1m_6 \tag{4.1}$$

A country with a high CI value indicates that its nation followed a more restrict quarantine with less mobility.

We will now consider the measurements of the mentioned CI on a daily basis for European countries from the 15th of February till the 9th of May. The last day we chose for the analysis was 9th of May and not further because we want to study the mobility during the lockdown period. In mid May some of the countries were no longer on a mandatory quarantine and some governments even encourage population to go out and buy local products. Since not all of the European countries had available data, we only could use 37 of them in the analysis. A map displaying these countries can be seen in Figure 5.1, in the Appendix section.

Firstly, without the application of clustering, just by looking at the CI trajectories in Figure 4.1, one can observe that at the end of the period between 15h of February and 9th of May, Turkey is the country with highest CI while Denmark is the country with the lowest. That same figure also highlights other relevant countries with different colours.

It should be noted that for some countries with high values of CI near the end of that period, the adherence to the quarantine might have been to late in order to stop spreading the virus. Thus, a high value of CI might not indicate less infections/obits.

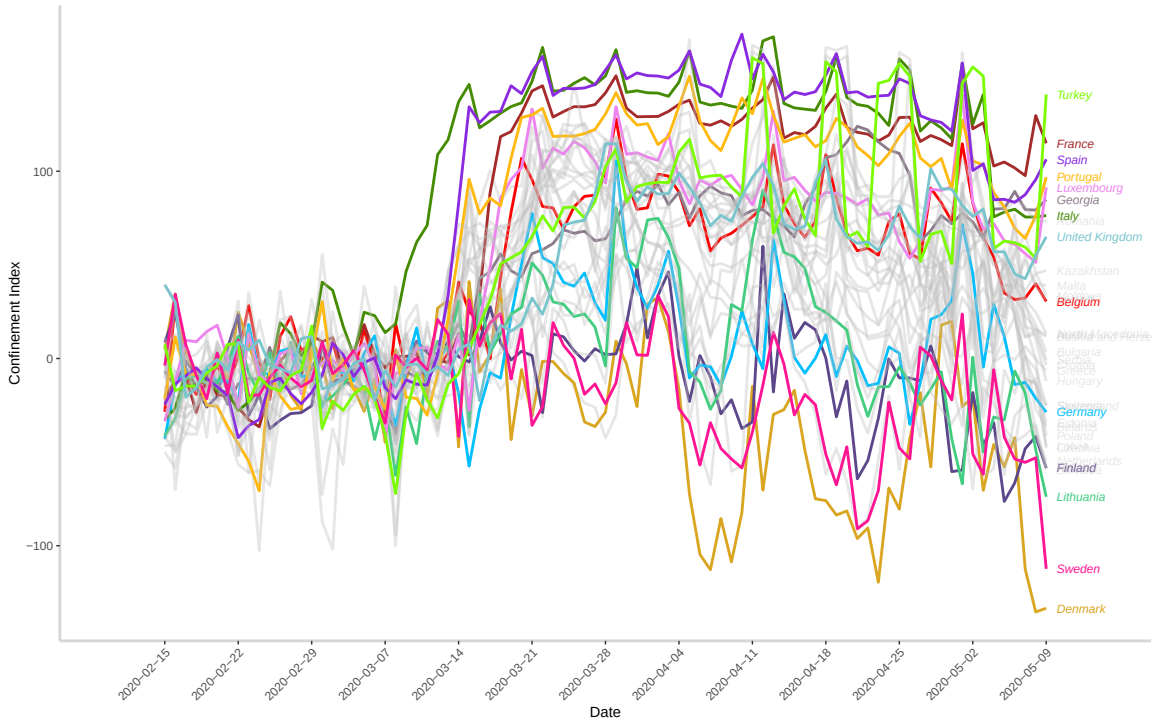


Figure 4.1: CI trajectories with some countries highlighted (prior to any clustering)

Regarding clustering, after running the methods, one can observe that, using the Calinski-Harabasz internal index, most of them suggest two clusters as it is shown in the following Figures.

The interpretation of the plots are very simple. The trajectories with the same colour indicate that they belong to the same group/cluster. The thick trajectories represent the mean trajectories (centroids) of each cluster. The percentages displayed in the top of the plots represent the percentage of countries in each group.

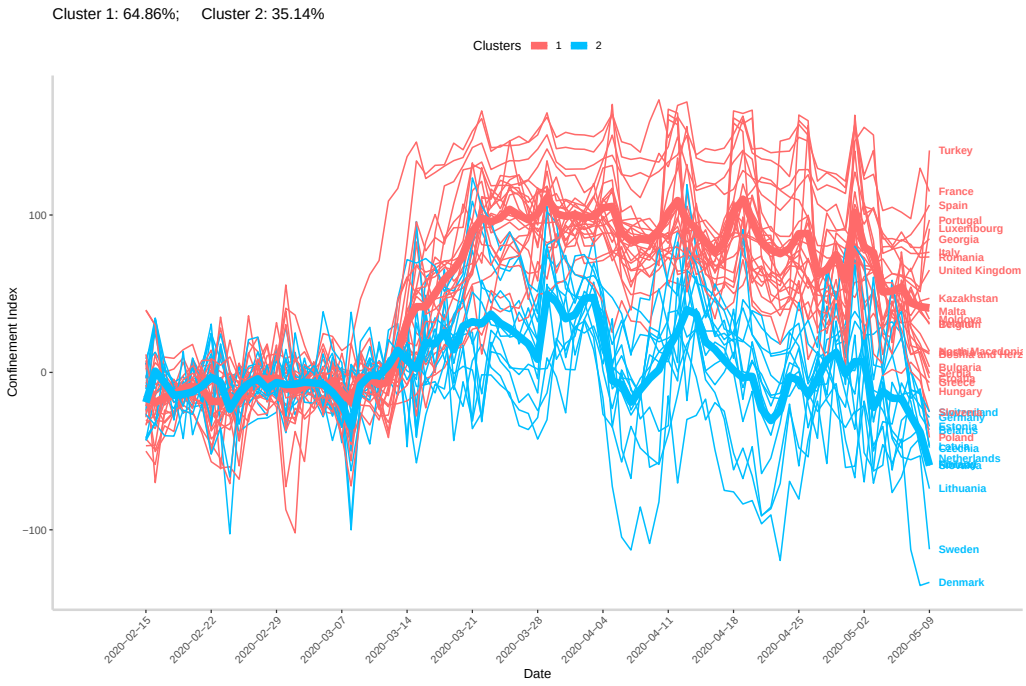


Figure 4.2: CMR Clustering: K-means Euclidean

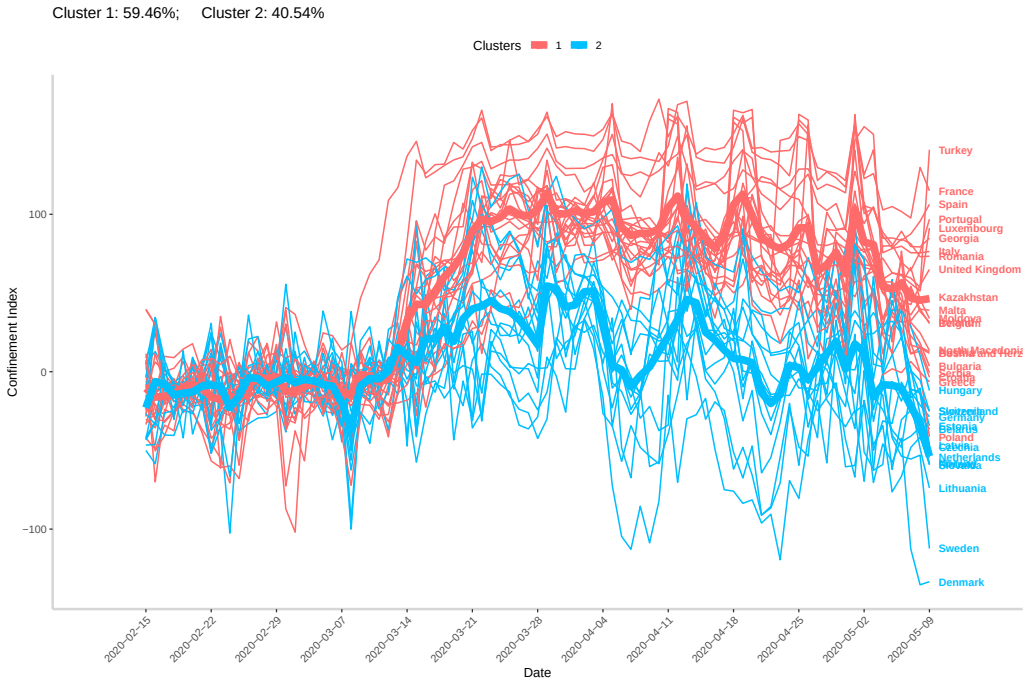


Figure 4.3: CMR Clustering: K-means Fréchet

Despite having different outcomes, the clusters found by K-means with Euclidean and Fréchet distance are quite similar as well as their mean trajectories. The only methods to suggest more than two clusters were the K-means with the Mahalanobis distance (M3, M4, M5); namely, 5, 6 and 4 clusters.

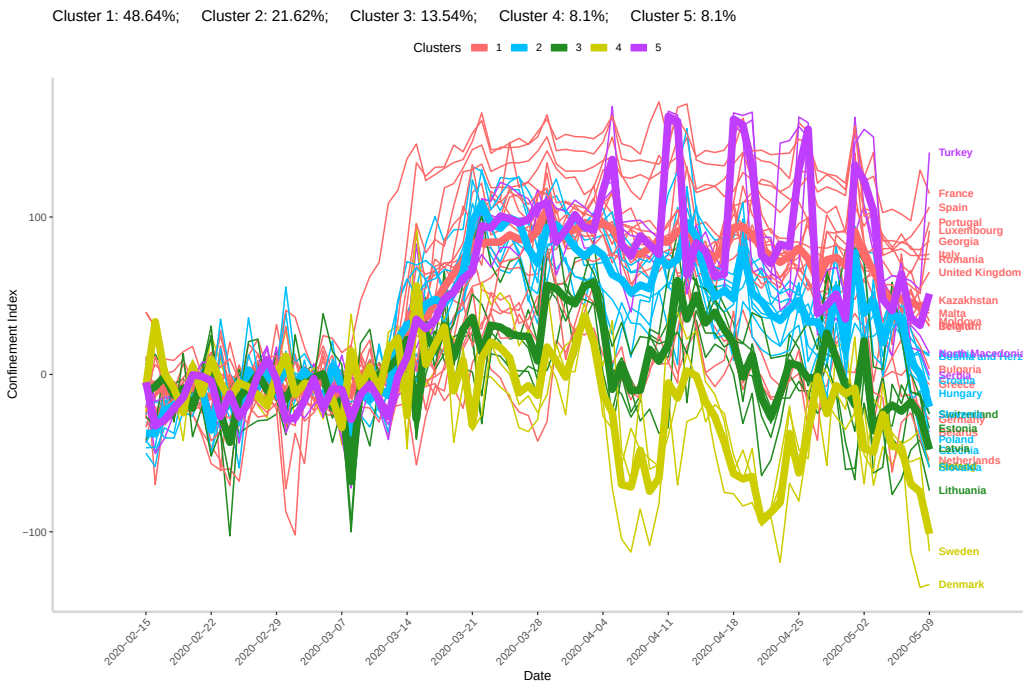


Figure 4.4: CMR Clustering: K-means Mahalanobis - Raw

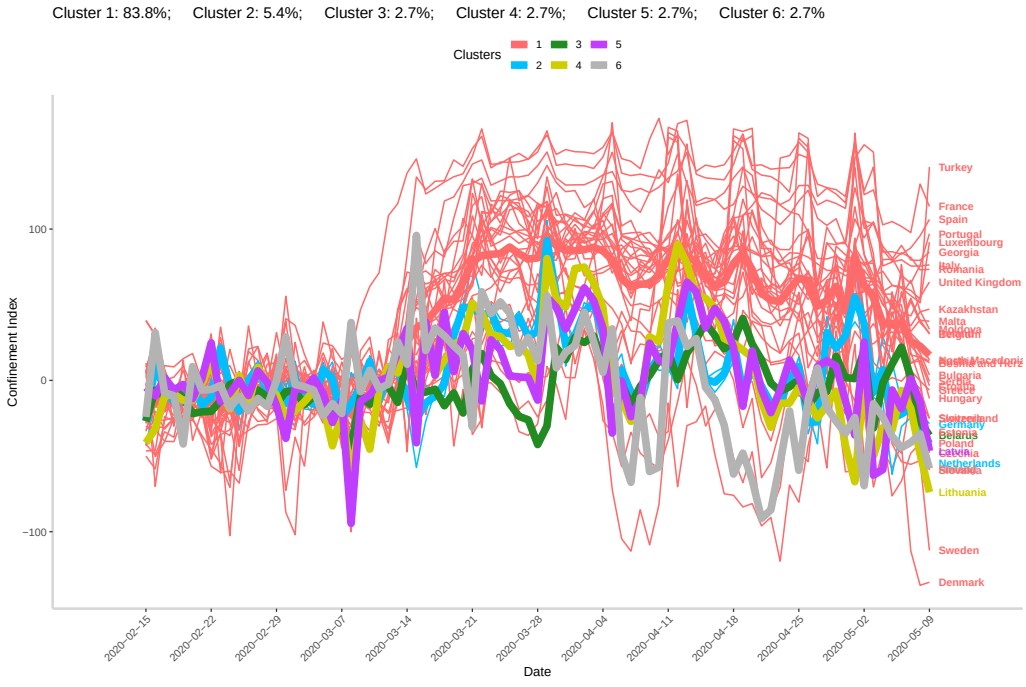


Figure 4.5: CMR Clustering: K-means Mahalanobis - Profile I

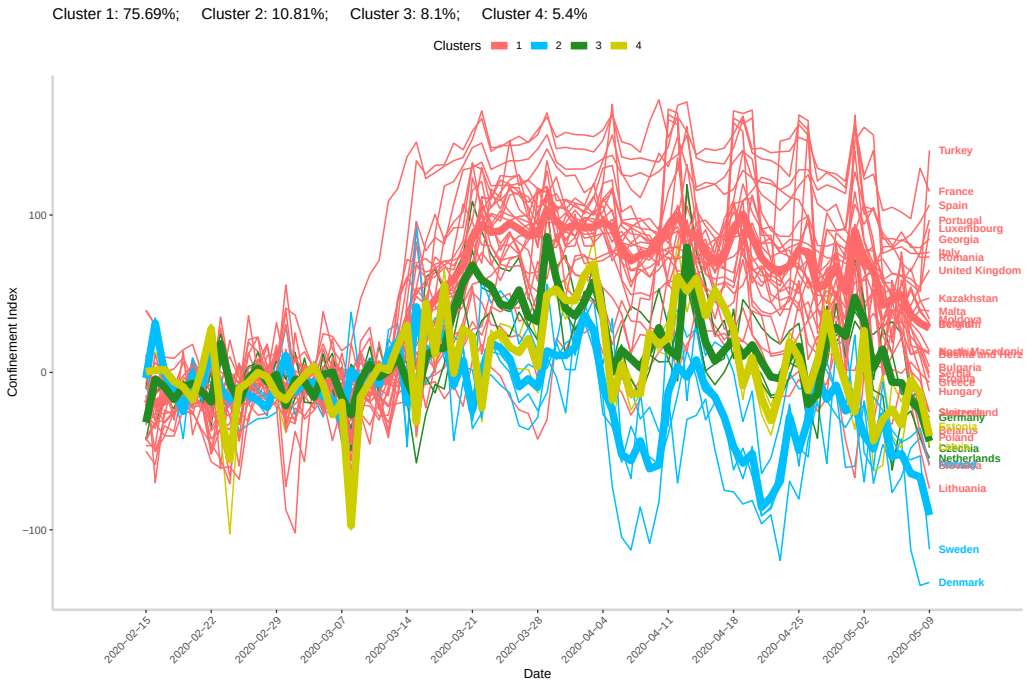


Figure 4.6: CMR Clustering: K-means Mahalanobis - Profile II

The Gaussian mixed-effects model with smoothing spline estimation got exactly the same clusters as the K-means with Euclidean distance.

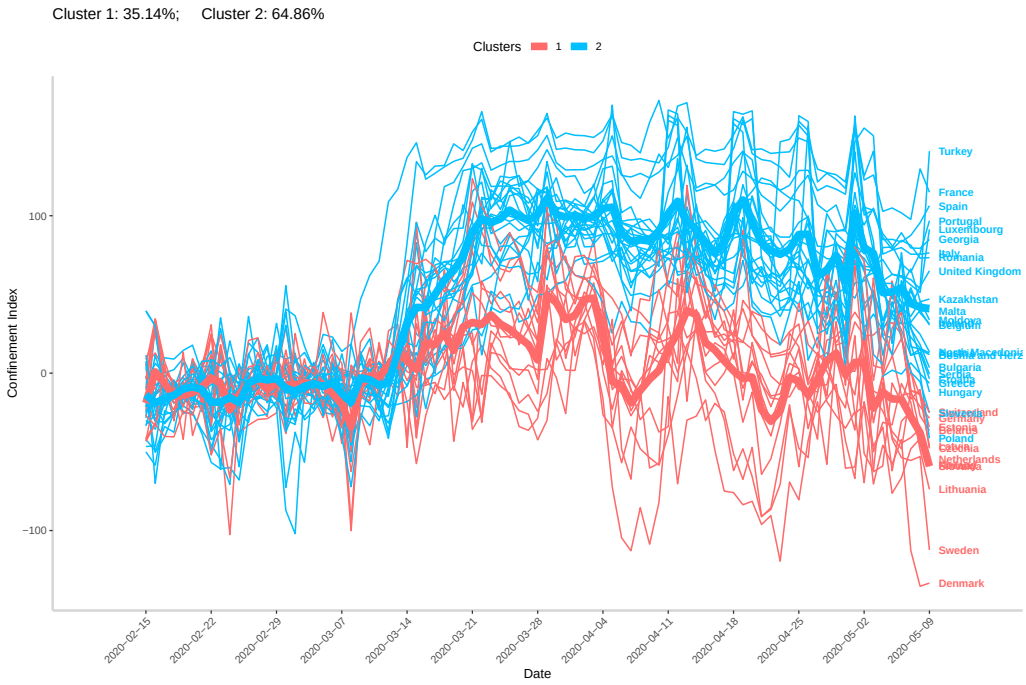


Figure 4.7: CMR Clustering: Mixed-effects model

The hierarchical methods also found two clusters.

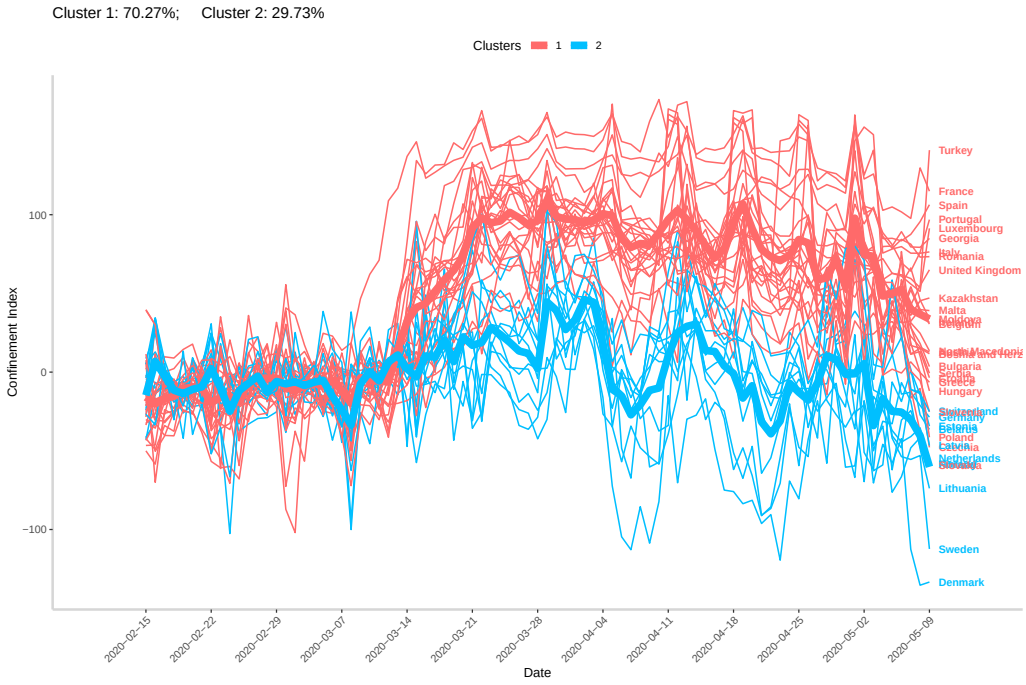


Figure 4.8: CMR Clustering: Hierarchical Euclidean

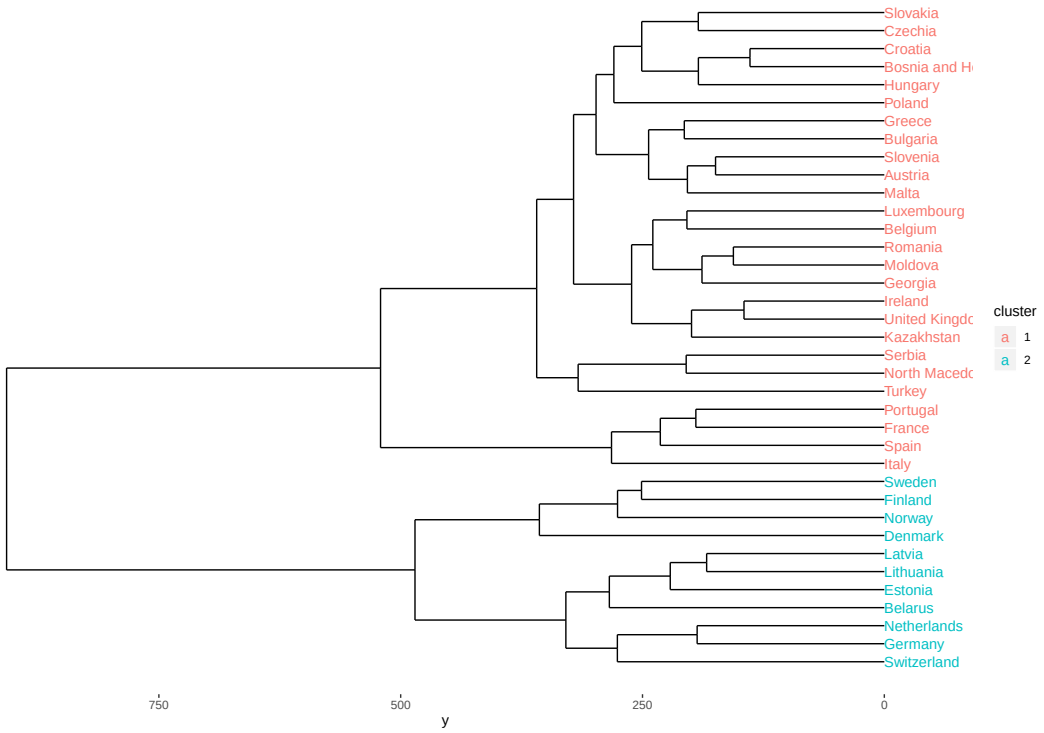


Figure 4.9: Hierarchical clustering dendrogram (Euclidean)

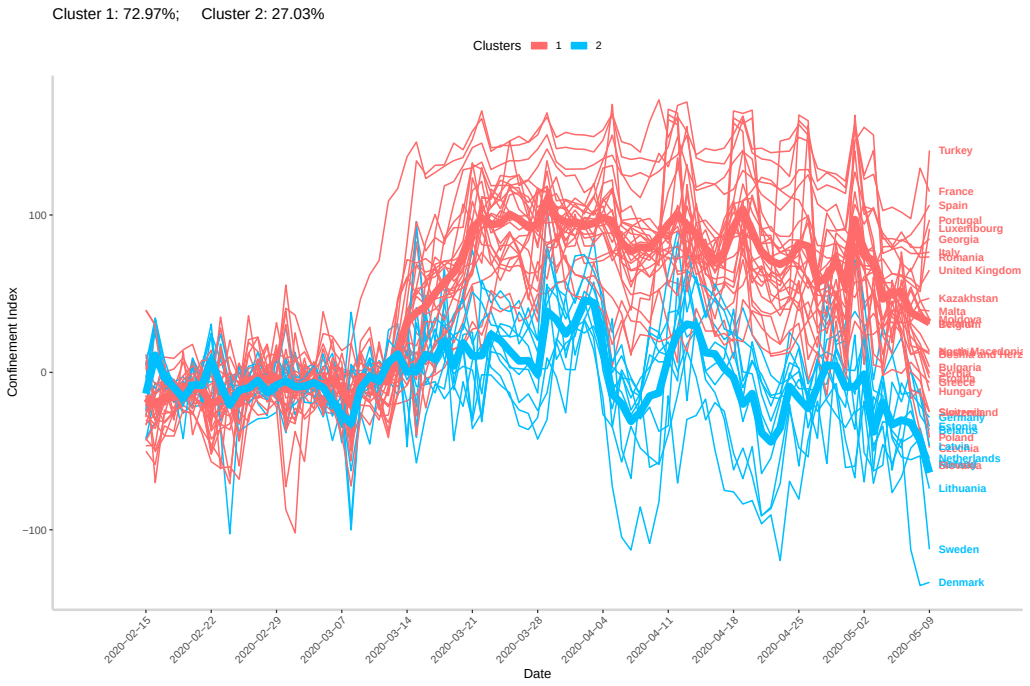


Figure 4.10: CMR Clustering: Hierarchical DTW

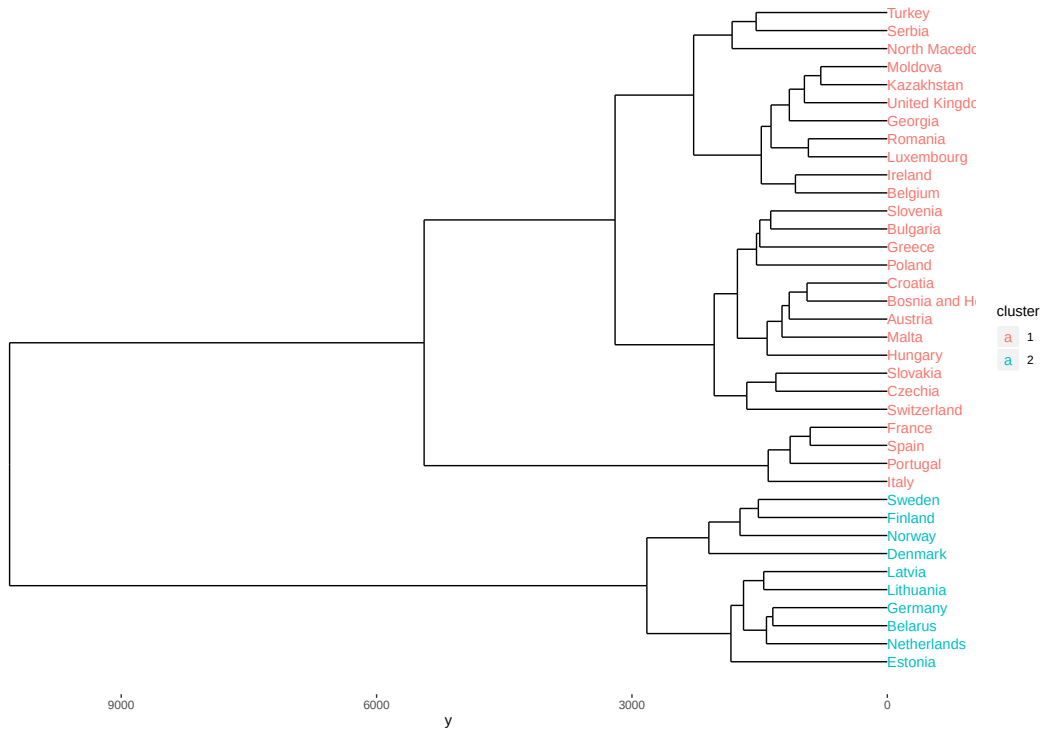


Figure 4.11: Hierarchical clustering dendrogram (DTW)

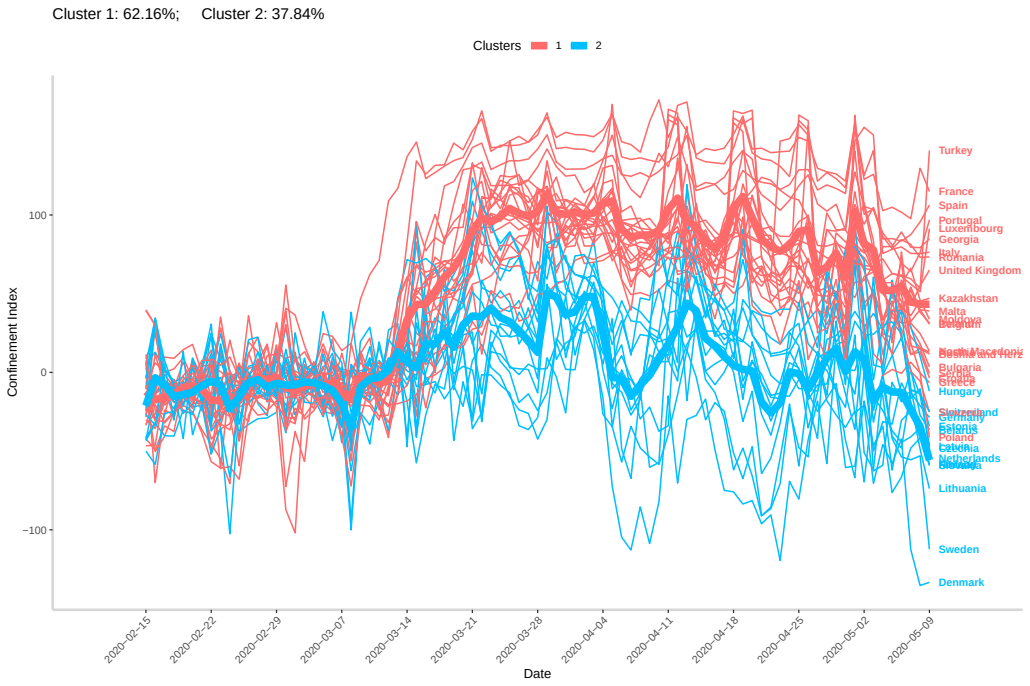


Figure 4.12: CMR Clustering: Bayesian Hierarchical

The clusters obtained from the mixture models with the Cholesky decomposition (M6 and M7) did not converge and therefore are not displayed.

Taking into account that the majority of the methods suggest two clusters, and that the methods with the best performance in Section 3.2 (K-means and Hierarchical clustering with Euclidean distance) got quite similar results, one might deduce the existence of two tendencies of mobility in the period of 15th February to 9th of May. Most of the countries in the analysis (around 65%) were relatively restrictive during the quarantine while a the others (around 35%) did not follow such measures.

Furthermore, by looking meticulously at each group obtained from both of those methods, one can observe that Northern countries like Denmark, Sweden, Norway, Finland and some of the Central Europe nations like Germany and Netherlands belong to the smaller cluster and therefore seem to had less restrictive measures during the quarantine in contrast to countries like Portugal, Spain, Italy, UK, Turkey or France.

In addition, the period in which countries like Denmark and Sweden obtained the lowest confinement index, much lower than the baseline values, was in the week of 18th to 25th of April, intriguingly already in full pandemic (see Figure 4.1). This behaviour might have been explained by the weather during that week. As it is shown in Figure 5.3 and Figure 5.4 in the Appendix section, during that period, cities like Stockholm and Copenhagen had the maximum

temperatures of the whole month.

A part of the previous analysis was published in the Público newspaper and can also be read in their website[1].

4.3 Deaths caused by the coronavirus

The next data we will analyse reports the number of deaths caused by the new coronavirus between the months of February and June. Since most of the COVID-19 statistics are public, this data can be downloaded in several websites.

In such analysis, rather than only focusing on absolute numbers, the change rate should also be considered. Thus, instead of only considering a variable with the total number of casualties over time, we added a variable that, for each day, takes into account the number of deaths in the past week. In other words, for the first variable, each point represents the cumulative number of new deaths whereas the second variable represents the total number of new deaths regarding the past 7 days. This added variable, when comparing to the total number of obits, gives a total different perspective of the matter.

For instance, let us consider the number of deaths in Portugal. If we observe the graphic with the total number of deaths over time (Figure 4.13) and we were around the 30th day since the first death (blue line), one might conclude that the numbers of deaths are still growing at a considerably rate.

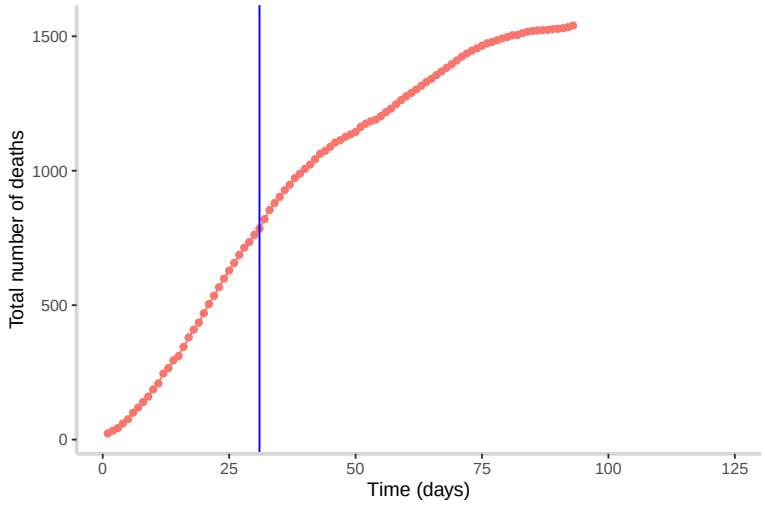


Figure 4.13: Total number of reported deaths in Portugal

However, in the same conditions, but observing the number of deaths regarding their past

[1] https://www.publico.pt/2020/06/06/opiniao/opiniao/confinamento-europa-1919585/amp

week (Figure 4.14), one would notice that the growth rate of deaths has already started do slow down.

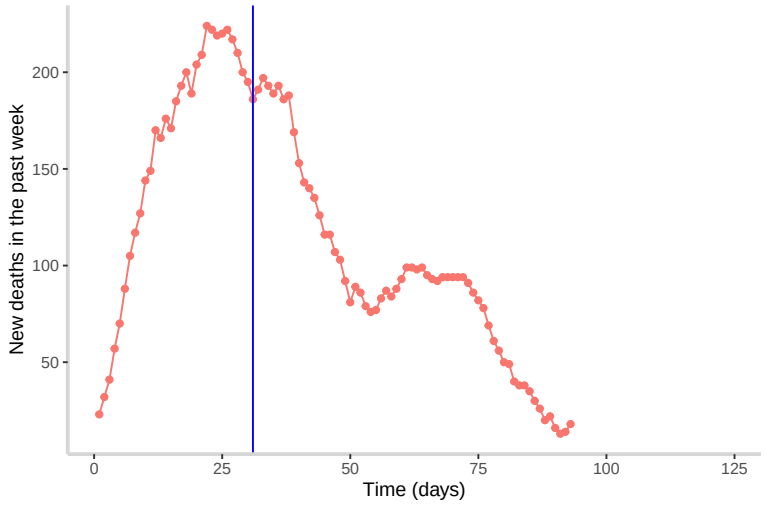


Figure 4.14: Number of reported deaths in relation to the 7 last days in Portugal

The Figure 4.15 illustrates the plot of the deaths regarding the past weeks against the cumulative number of deaths (in Portugal). As it is shown, in the beginning, the weekly values are much more spaced and tend to condense as the total number of deaths increases.

When the total number of deaths were 1300 and until they reached a value of 1450, the reported weekly deaths were very similar. After that value, the growth rate of deaths rapidly started to decrease. The blue line simultaneously indicates the weekly numbers and total numbers of deaths that happened around the 30th day since the first death.

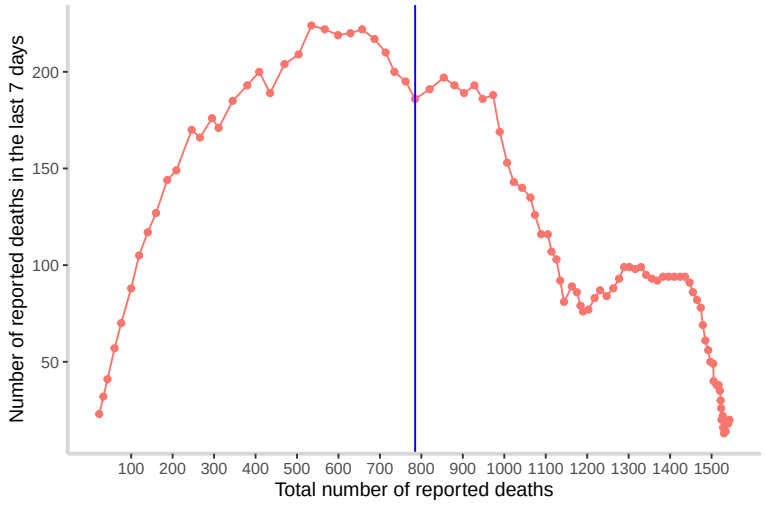


Figure 4.15: Number of deaths in the past week against the total number

The majority of the coronavirus analyses show the infections/obits plotted against time. However, the virus does not care if it is April or May. It only cares about how many cases there are today and how many will be tomorrow, i.e, the total number of infections/obits and the growth rate of infection/obits. Thus, this added variable when plotted against the total number of deaths, gives a more clear view of the casualties growth.

However, to achieve such goals, along with the clustering, we had to resort to multivariate trajectories. This application was possible using the `kml3d` package [43], which clusters several variable-trajectories jointly using K-means with Euclidean distance. The package also provides tools to visualize an interactive 3D dynamic graphs in an R session that can be attached to PDF pages. However, that feature does not work well with all PDF readers. Thus, we will display the 3D-plot obtained from clustering in a non-interactive way, i.e, separately.

Because we are comparing countries that have big differences in population size, we decided to divide the values by their respective number of inhabitants. Although incorporating the population sizes might be problematic, that will not be an issue because we are dealing with deaths instead of infections and we are also aiming to the visualize the number of deaths in the past week against the total amount of deaths.A barplot with the number of fatalities, per million of inhabitants, can be seen in Figure 5.2 in the Appendix section.

Moreover, it is known that the virus did not cause the first death at the same day across the countries. Still, we have made all of the trajectories start at the same moment, i.e, at the first death of the respective country. In such manner, it is more appropriate to compare the trajectories development. Hence, trajectories length vary and depend on the first day of death.

It is worth noting that the data is consistently suffering amendments. For instance, Spain announced in May 25 a different way of collecting data, by counting a death based on when it happened, instead of when authorities were notified about it. As a result, the country's death toll slightly dropped.

Running the multivariate clustering, we discovered 2 clusters in the data. The resulting plot can be seen in the following figures. Again, the colours of the trajectories represent a group while the thick ones indicate the mean trajectories of each group.

A small group (17.8%) was discovered by the multivariate longitudinal clustering with a higher number of fatalities containing the following countries: Belgium, France, Ireland, Italy, Netherlands, Spain, Sweden, United Kingdom (blue trajectories). The bigger group (82.2%) consists on the rest of the countries and has relatively lower numbers of deaths (red trajectories).

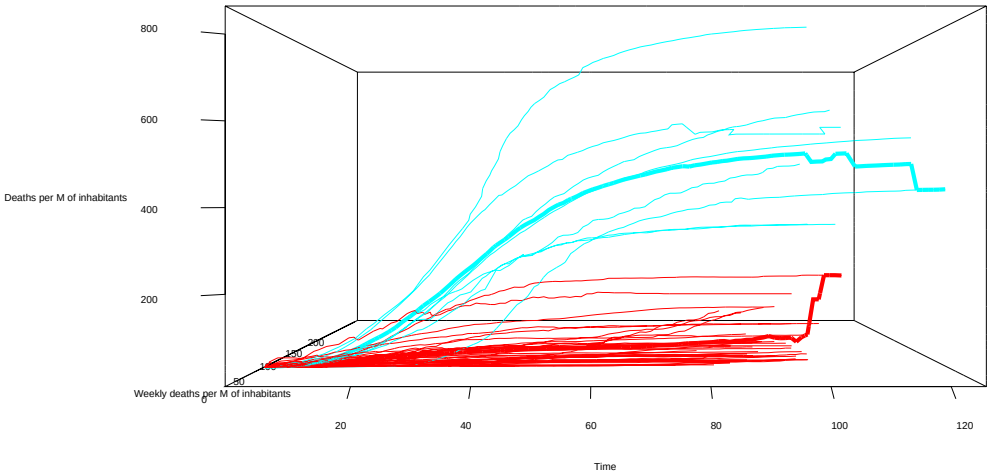


Figure 4.16: Total number of deaths over time (per M of inhabitants)

The number of deaths over time are illustrated in Figure 4.16. The mean trajectory of the blue cluster starts to rapidly grow and reaches a relatively constant state over time. In the red cluster, the opposite is seen.

Furthermore, by looking at Figure 4.17, as well as Figure 4.18, it is notable that all of the trajectories in the smaller cluster (containing the higher number of deaths) have already started to significantly slow the growth of deaths. Observing the mean trajectory of that cluster in Figure 4.17, the growth of deaths apparently started to decrease around the 40th day since the first deaths. Observing that same mean trajectory in Figure 4.18, shows that the decreasing in the growth of deaths occurred around the value 370 deaths per million of inhabitants.

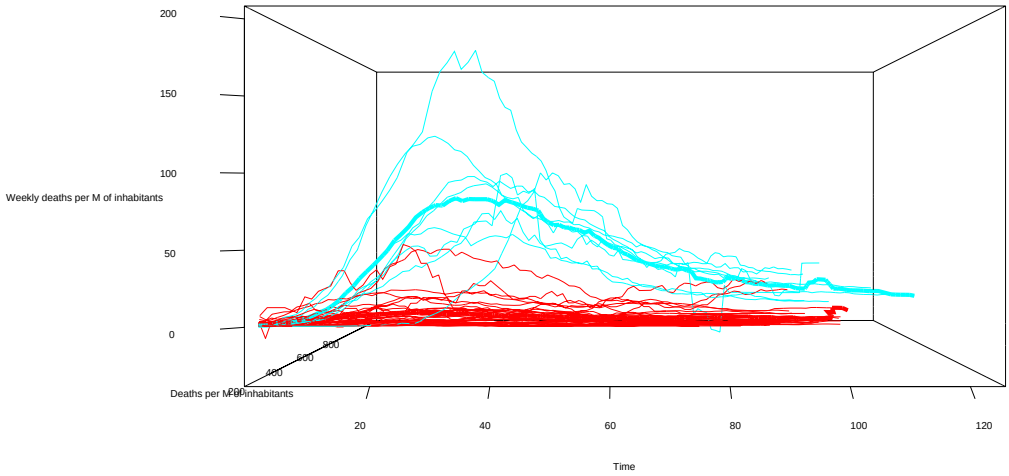


Figure 4.17: Number of deaths regarding the past week over time (per M of inhabitants)

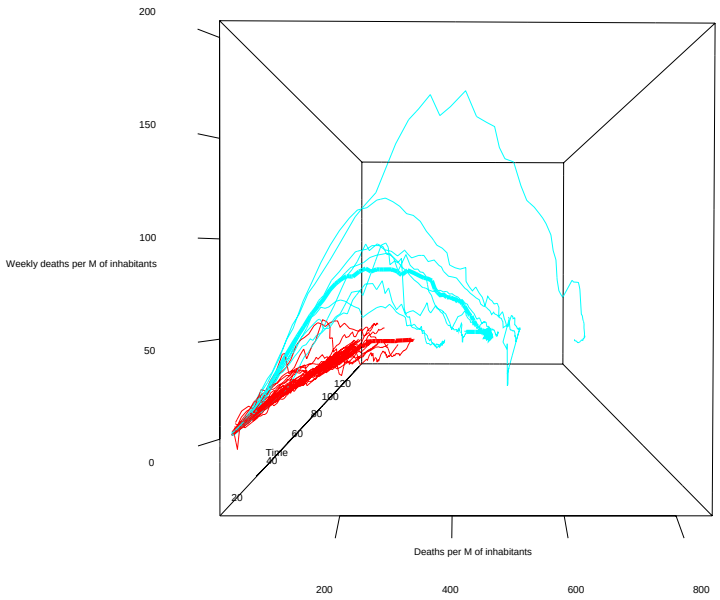


Figure 4.18: Number of deaths regarding the past week against total number of deaths (per M of inhabitants)

Chapter 5

Conclusion and Future Work

In this work, we presented a brief review of some clustering methodologies used in longitudinal data analysis. This study was carried out in order to unravel the best option between non-parametrics and model-based methods, and to identify which particular method is more reliable. Besides that, with the ongoing debate on the selection of the number of clusters, we also intended to identify the index that performs better in this action, particularly when longitudinal data is used.

Resorting to various clustering validity indices and to artificial longitudinal data, we found that non-parametric methods, in addition to being theoretically less complex, had better results and less software limitations. K-means along with Hierarchical clustering, both using Euclidean distance, were the methods to yield the best results, which might reinforce why they are so popular. Moreover, mixture models with the Cholesky decomposition (M6 and M7) revealed some software limitations, as they did not converge in the CMR application. Regarding the number of clusters choice, from the artificial longitudinal dataset tests, the index that suggested the correct number more frequently was the Calinski-Harabsz. Clustering solutions are not unique and strongly depend on the investigator choices. However, with the use of clustering validity indices, those choices might be improved.

From the analysis performed in the Community Mobility Reports dataset, we discovered that, between 15th of February and 9th of May, countries like Turkey, France, Spain and Portugal had higher levels of confinement when comparing to countries like Sweden or Denmark. The study also suggests that possibly there existed two tendencies (groups) in that period.

One major tendency concerning countries with high confinement and another smaller tendency containing countries with less restrictive measures. Looking scrupulously to each group, it is evident that Northern European countries and some of the Central Europe nations like Germany and Netherlands, belonging to the smaller cluster, had more mobility during the lockdown.

In the second dataset, with the application of multivariate clustering to the fatalities caused by the COVID-19 between February and June, we discovered two groups among the data. More explicitly, considering the population sizes, we found a small group with higher number of deaths containing 17.8% of the countries. This group includes Belgium, France, Ireland, Italy, Netherlands, Spain, Sweden and UK. However, looking at its mean trajectory, it was also uncovered that the growth of deaths in that same group started to significantly drop at around the 40th day after the first death. Moreover, the multivariate clustering enabled us to see the plot of the deaths in the past week against the total number of deaths, in which it is evident that the beginning of that drop was also around the 370 deaths per million of inhabitants. The second cluster obtained consists on 82.2% of the remaining countries and exhibits trajectories with a relatively small and steady growth of deaths over time. Even so, this analysis is made with the assumption of trusting the available data. As it was said before, there is a lot of complexity and misconception regarding the COVID-19 data.

The present work can be extended in several directions. Firstly, there are other clustering methods that were not discussed in this thesis that can work with longitudinal data. Besides, those methods can also be employed in the clustering evaluation to test their performance. Secondly, alternative statistical procedures can be done separately, or combined with clustering, to provide additional knowledge in the analysis of longitudinal data. Lastly, regarding the applications of the COVID-19 data, the number of infections and tests were left out and they can add major contributions to the analysis. With the constant amendments in the coronavirus numbers, the data is continuously being updated and future analysis will be needed.

Bibliography

[1] Agarwal, P. K., Avraham, R. B., Kaplan, H., and Sharir, M. (2014). Computing the Discrete Fréchet distance in subquadratic time. *SIAM Journal on Computing*, 43(2):429–449.

[2] Aggarwal, C. C. (2015). *Data Mining*, volume 14. Springer International Publishing, Cham.

[3] Albatineh, A. N. and Niewiadomska-Bugaj, M. (2011). Correcting Jaccard and other similarity indices for chance agreement in cluster analysis. *Advances in Data Analysis and Classification, Springer*, 5(3):179–200.

[4] Alizadeh, A. A., Aranda, V., Bardelli, A., Blanpain, C., Bock, C., Borowski, C., Caldas, C., Califano, A., Doherty, M., Elsner, M., Esteller, M., Pe, D., Polyak, K., Roberts, C. W. M., Siu, L., Snyder, A., and Stower, H. (2015). Toward understanding and exploiting tumor heterogeneity. *Nature Medicine*, 21(8):1–8.

[5] Alt, H. and Godau, M. (1992). Measuring the resemblance of polygonal curves. *Eighth Annual Symposium On Computational Geometry*, pages 102–109.

[6] Alt, H. and Godau, M. (1995). Computing the Fréchet distance between two polygonal curves. *International Journal of Computational Geometry & Applications*, 05(01n02):75–91.

[7] Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Pérez, J. M., and Perona, I. (2013). An extensive comparative study of cluster validity indices. *Pattern Recognition*, 46(1):243–256.

[8] Baker, F. B. and Hubert, L. J. (1975). Measuring the power of hierarchical cluster analysis. *Journal of the American Statistical Association*, 70(349):31–38.

[9] Ball, G. H. and Hall, D. J. (1965). Isodata: A novel method of data analysis and pattern classi cation. *Menlo Park: Stanford Research Institute.*

[10] Batista, G. E. A. P. A. and Keogh, E. J. (2014). A Complexity-Invariant Distance Measure for Time Series. (March).

[11] Bellman, R. and Kalaba, R. (1958). On adaptive control processes. *IRE Transactions on Automatic Control*, 4(2):1–9.

[12] Bolshakova, N. and Azuaje, F. (2003). Cluster validation techniques for genome expression data. *Signal Processing*, 83(4):825–833.

[13] Caliński, T. and Harabasz, J. (2007). A dendrite method for cluster analysis. *Communications in Statistics - Theory and Methods*, 3272(1974).

[14] Celebi, M. E. and Aydin, K. (2016). *Unsupervised learning algorithms*. Springer International Publishing.

[15] Celeux, G. and Govaert, G. (1992). A classification EM algorithm for clustering and two stochastic versions. *Computational Statistics & Data Analysis*, 14(3):315–332.

[16] Cho, J. H. and Feldman, M. (2015). Heterogeneity of autoimmune diseases : pathophysiologic insights from genetics and implications for new therapies. (June).

[17] Chu, S., Keogh, E., Hart, D., and Pazzani, M. (2002). Iterative Deepening Dynamic Time Warping for Time Series. pages 195–212.

[18] Cooke, E. J., Savage, R. S., Kirk, P. D., Darkins, R., and Wild, D. L. (2011). Bayesian hierarchical clustering for microarray time series data with replicates and outlier measurements. *BMC Bioinformatics*, 12.

[19] Costa, J., Mascarello, M., Ferreira, F., and Gaio, R. (2018). Some Developments in the Clustering of Longitudinal Trajectories. *(Submitted)*, pages 1–17.

[20] Cruz Mesia, R. D., Quintana, F. A., and Marshall, G. (2007). Model Based Clustering for Longitudinal Data. pages 1–25.

[21] da Estatística, S. P. Nota sobre a Utilização Incorreta de Conceitos Estatísticos. page 19.

[22] Das, P. (2019). Analysis of Cross Section, Time Series and Panel Data with Stata 15.1. In *Econometrics in Theory and Practice*, pages XXVII, 565. Springer.

[23] Davies, D. L. and Bouldin, D. W. (1979). A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2):224–227.

[24] Delmar, F. and Wiklund, J. (2013). The effect of small business managers' growth motivation on firm growth: A longitudinal study. *New Perspectives on Firm Growth*, pages 82–111.

[25] Desgraupes, B. (2013). Clustering Indices: CRAN Package. *University Paris Ouest*, (April):1–10.

[26] Dice, L. R. (1945). Measures of the Amount of Ecologic Association Between Species Author (s): Lee R . Dice Published by : Ecological Society of America Stable URL : http://www.jstor.org/stable/1932409. *Ecology*, 26(3):297–302.

[27] Ding, H., Trajcevski, G., Scheuermann, P., Wang, X., and Keogh, E. (2008). Querying and Mining of Time Series Data : Experimental Comparison of Representations and Distance Measures. 1:1542–1552.

[28] Divoux, A., Tordjman, J., Lacasa, D., Veyrie, N., Hugol, D., Aissat, A., Basdevant, A., Guerre-Millo, M., Poitou, C., Zucker, J. D., Bedossa, P., and Clément, K. (2010). Fibrosis in human adipose tissue: Composition, distribution, and link with lipid metabolism and fat mass loss. *Diabetes*, 59(11):2817–2825.

[29] Dolnicar, S., Grün, B., and Leisch, F. (2018). *Market Segmentation Analysis: Understanding It, Doing It, and Making It Useful*. Springer International Publishing.

[30] Donald, H. and Robert D., G. (2006). *Longitudinal Data Analysis*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, USA.

[31] Du, W., Cheung, H., and Johnson, C. A. (2015). A longitudinal support vector regression for prediction of ALS score. pages 1586–1590.

[32] Dudoit, S. and Fridlyand, J. (2002). A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome biology*, 3(7):1–21.

[33] Edda, K., Wolfram, L., Christoph, W., Axel, K., Hans, L., and Ralf, H. *Systems Biology: A textbook*. Wiley-Blackwell, 2nd edition.

[34] Ester, M., Kriegel, H.-p., Xu, X., and Miinchen, D. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise.

[35] Estivill-castro, V. Why so many clustering algorithms — A Position Paper. 8.

[36] Euachongprasit, W. and Ratanamahatana, C. A. (2008). Efficient Multimedia Time Series Data Retrieval Under Uniform Scaling and Normalisation. *European Conference on Information Retrieval*, pages 506–513.

[37] Everitt, B. S., Landau, S., Leese, M., and Stahl, D. (2011). *Cluster Analysis*. Wiley Series in Probability and Statistics, 5th edition.

[38] Ferreira, F. (2017). Novos Desenvolvimentos em Análise de Dados (Tese).

[39] Figueiró, H. V., Li, G., Trindade, F. J., Assis, J., Pais, F., Fernandes, G., Santos, S. H., Hughes, G. M., Komissarov, A., Antunes, A., Trinca, C. S., Rodrigues, M. R., Linderoth, T., Bi, K., Silveira, L., Azevedo, F. C., Kantek, D., Ramalho, E., Brassaloti, R. A., Villela, P. M., Nunes, A. L., Teixeira, R. H., Morato, R. G., Loska, D., Saragüeta, P., Gabaldón, T., Teeling, E. C., O'Brien, S. J., Nielsen, R., Coutinho, L. L., Oliveira, G., Murphy, W. J., and Eizirik, E. (2017). Genome-wide signatures of complex introgression and adaptive evolution in the big cats. *Science Advances*, 3(7):1–14.

[40] Fowlkes, E. B. and Mallows, C. L. (1983). A Method for Comparing Two Hierarchical Clusterings. *Journal of the American Statistical Association*, 78(383):553–569.

[41] Fraley, C. and Raftery, A. E. (2006). MCLUST Version 3 for R : Normal Mixture Modeling and. *Technical Report No. 504*, Department(504):1–57.

[42] Fréchet, M. M. (1906). Sur quelques points du calcul fonctionnel. *Rendiconti del Circolo Matematico di Palermo*, 22(1):1–72.

[43] Genolini, C., Alacoque, X., Sentenac, M., and Arnaud, C. (2015). kml and kml3d : R Packages to Cluster Longitudinal Data. *Journal of Statistical Software*, 65(4).

[44] Genolini, C., Ecochard, R., Benghezal, M., Driss, T., Andrieu, S., and Subtil, F. (2016). kmlShape: An Efficient Method to Cluster Longitudinal Data (Time-Series) According to Their Shapes. *PLOS ONE*, 11(6):e0150738.

[45] Genolini, C. and Falissard, B. (2011). Kml: A package to cluster longitudinal data. *Computer Methods and Programs in Biomedicine*, 104(3):e112–e121.

[46] Giudici, P., Ingrassia, S., and Vichi, M. (2013). *Statistical Models for Data Analysis*, volume 3 of *Studies in Classification, Data Analysis, and Knowledge Organization*. Springer International Publishing, Heidelberg.

[47] Golumbeanu, M. and Beerenwinkel, N. (2018). Clustering time series gene expression data with TMixClust. pages 1–16.

[48] Golumbeanu, M., Desfarges, S., Hernandez, C., Quadroni, M., Rato, S., Mohammadi, P., Telenti, A., Beerenwinkel, N., and Ciuffi, A. (2019). Proteo-Transcriptomic Dynamics of Cellular Response to HIV-1 Infection. *Scientific Reports*, 9(1):1–12.

[49] Gong, H., Xun, X., and Zhou, Y. (2019). Profile clustering in clinical trials with longitudinal and functional data methods. *Journal of Biopharmaceutical Statistics*, 29(3):541–557.

[50] Gu, C. (2014). Smoothing spline ANOVA models: R package gss. *Journal of Statistical Software*, 58(5):1–25.

[51] Gu, J. and Jin, X. (2006). A Simple Approximation for Dynamic Time. *International Conference on Intelligent Data Engineering and Automated Learning*, pages 841–842.

[52] Han, J., Kamber, M., and Pei, J. (2012). *Data mining: Concepts and Techniques*. Morgan Kaufmann Publishers, 3rd edition.

[53] Heggeseth, B. C. (2014). *Longitudinal cluster analysis with applications to growth trajectories*. PhD thesis, University of California, Berkeley.

[54] Heller, K. A. and Ghahramani, Z. (2005). Bayesian Hierarchical Clustering.

[55] Hellewell, J., Abbott, S., Gimma, A., Bosse, N. I., Jarvis, C. I., Russell, T. W., Munday, J. D., Kucharski, A. J., Edmunds, W. J., Sun, F., Flasche, S., Quilty, B. J., Davies, N., Liu, Y., Clifford, S., Klepac, P., Jit, M., Diamond, C., Gibbs, H., van Zandvoort, K., Funk, S., and Eggo, R. M. (2020). Feasibility of controlling COVID-19 outbreaks by isolation of cases and contacts. *The Lancet Global Health*, 8(4):e488–e496.

[56] Helliar, C., Cobb, I., and Innes, J. (2002). A longitudinal case study of profitability reporting in a bank. *British Accounting Review*, 34(1):27–53.

[57] Hossain, P., Kawar, B., and El Nahas, M. (2007). Obesity and diabetes in the developing world - A growing challenge. *New England Journal of Medicine*, 356(3):213–215.

[58] Hsiao, C. (2007). Panel data analysis-advantages and challenges. *Test*, 16(1):1–22.

[59] Hu, B., Chen, Y., and Keogh, E. (2013). Time Series Classification under More Realistic Assumptions. *SDM 2013*, (1).

[60] Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 218:193–218.

[61] Hubert, L. and Schultz, J. (1976). Quadratic Assignment As a General Data Analysis Strategy. *British Journal of Mathematical and Statistical Psychology*, 29(2):190–241.

[62] Jaccard, P. (1908). Nouvelles researches sur la distribution florale. *Bulletin de la Société vaudoise des sciences naturelles*, 11(2):37–50.

[63] Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31(8):651–666.

[64] Jain, A. K., Murty, M. N., and Flynn, P. J. (1999). Data clustering: A review. *ACM Computing Surveys*, 31(3):264–323.

[65] Jeffrey D . Banfield and Adrian E . Raftery (1993). Model-Based Gaussian and Non-Gaussian Clustering. *International Biometric Society*.

[66] Johansson, M. A., Cummings, D. A. T., and Glass, G. E. (2009). Multiyear climate variability and dengue - El Niño southern oscillation, weather, and dengue incidence in Puerto Rico, Mexico, and Thailand: A longitudinal data analysis. *PLoS Medicine*, 6(11).

[67] Kahveci, T., Singh, A., and Gürel, A. (2002). Similarity searching for multi-attribute sequences. *Proceedings of the International Conference on Scientific and Statistical Database Management, SSDBM*, 2002-Janua:175–184.

[68] Kaufman, L. and Rousseeuw, P. (1987). Clustering by means of medoids. *Statistical Data Analysis Based on the L1-Norm and Related Methods*.

[69] Keogh, E. and Ratanamahatana, C. A. (2005). Exact indexing of dynamic time warping. (May 2004):358–386.

[70] Kruskal, J. and Liberman, M. (1983). The symmetric time-warping problem: from continuous to discrete. In *Time warps, string edits, and macromolecules: the theory and practice of sequence comparison*.

[71] Kucharski, A. J., Russell, T. W., Diamond, C., Liu, Y., Edmunds, J., Funk, S., Eggo, R. M., Sun, F., Jit, M., Munday, J. D., Davies, N., Gimma, A., van Zandvoort, K., Gibbs, H., Hellewell, J., Jarvis, C. I., Clifford, S., Quilty, B. J., Bosse, N. I., Abbott, S., Klepac, P., and Flasche, S. (2020). Early dynamics of transmission and control of COVID-19: a mathematical modelling study. *The Lancet Infectious Diseases*, 20(5):553–558.

[72] Kulczynski, S. (1927). Die Pflanzenassoziationen der Pieninen. *Bulletin International de l'Academie Polonaise des Sciences et des Lettres, Classe des Sciences Mathematiques et Naturelles, B (Sciences Naturelles)*.

[73] Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22.

[74] Little, T. D., Deboeck, P., and Wu, W. (2015). Longitudinal Data Analysis. In *Emerging Trends in the Social and Behavioral Sciences*, number May 2018, pages 1–17. Wiley.

[75] Liu, L. and Özsu, M. T. (2009). *Encyclopedia of Database Systems*, volume 53. Springer US, New York.

[76] Liu, Y., Gayle, A. A., Wilder-Smith, A., and Rocklöv, J. (2020). The reproductive number of COVID-19 is higher compared to SARS coronavirus. *Journal of Travel Medicine*, (Figure 1):1–4.

[77] Louis Leon Thurstone (1927). A law of comparative judgement. *Psychol Rev*.

[78] Ma, P., Castillo-Davis, C. I., Zhong, W., and Liu, J. S. (2006). A data-driven clustering method for time course gene expression data. *Nucleic Acids Research*, 34(4):1261–1269.

[79] MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *In 5-th Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297.

[80] Maier, B. F. and Brockmann, D. (2020). Effective containment explains subexponential growth in recent confirmed COVID-19 cases in China. *Science*, page eabb4557.

[81] Mann, J. (2003). Observational research methods. Research design II. *Emergency Medicine Journal*, (October 2008):54–61.

[82] Mcnicholas, P. D. (2011). On Model-Based Clustering , Classification , and Discriminant Analysis Model-Based Approaches. 10(2):181–199.

[83] McNicholas, P. D. (2016). Model-Based Clustering. *Journal of Classification*, 33(3):331–373.

[84] Mcnicholas, P. D. and Murphy, T. B. (2010). Model-based clustering of longitudinal data. 38(1):153–168.

[85] McNicholas, P. D. and Subedi, S. (2012). Clustering gene expression time course data using mixtures of multivariate t-distributions. *Journal of Statistical Planning and Inference*, 142(5):1114–1127.

[86] Mendis, S., Lindholm, L. H., Mancia, G., Whitworth, J., Alderman, M., Lim, S., and Heagerty, T. (2007). World Health Organization (WHO) and International Society of Hypertension (ISH) risk prediction charts: Assessment of cardiovascular risk for prevention and control of cardiovascular disease in low and middle-income countries. *Journal of Hypertension*, 25(8):1578–1582.

[87] Milligan, G. W. and Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50(2):159–179.

[88] Milligan, G. W. and Cooper, M. C. (1986). A Study of the Comparability of External Criteria for Hierarchical Cluster Analysis. *Multivariate Behavioral Research*, 21(4):441–458.

[89] Milton Severo, A. C., Santos, Ana Isabel Ribeiro, A., Rocha, Carla Lopes, D., Correia, E. R., Gonçalo Gonçalves, J., Araújo, Makram Talih, M., Tavares, Nuno Lunet, P., Meireles, Raquel Lucas, Rui Camacho, Sílvia Fraga, S., Correia, Susana Silva, T., and Barros, L. e. H. (2020). Factos para compreender a epidemia da covid-19. O que têm de específico as doenças infecciosas? *Público*, pages 10–11.

[90] Nenad Tomašev and Miloš Radovanović (2016). Clustering Evaluation in High-Dimensional Data. In *Unsupervised Learning Algorithms, Springer*.

[91] Ngoma, T. A. (2006). World Health Organization cancer priorities in developing countries. *Annals of Oncology*, 17(SUPPL. 8):9–14.

[92] Niennattrakul, V. and Ratanamahatana, C. A. (2007). On clustering multimedia time series data using k-means and dynamic time warping. *Proceedings - 2007 International Conference on Multimedia and Ubiquitous Engineering, MUE 2007*, pages 733–738.

[93] Pakhira, M. K., Bandyopadhyay, S., and Maulik, U. (2004). Validity index for crisp and fuzzy clusters. 37:487–501.

[94] Pauler, D. K., Laird, N. M., and Carlo, M. (2000). Application to Assessment of Noncompliance. (June):464–472.

[95] Pfitzner, D., Leibbrandt, R., and Powers, D. (2009). Characterization and evaluation of similarity measures for pairs of clusterings. *Knowledge and Information Systems, Springer*, pages 361–394.

[96] Pingault, J.-B., Tremblay, R. E., Vitaro, F., Carbonneau, R., Genolini, C., Falissard, B., and Côté, S. M. (2011). Childhood Trajectories of Inattention and Hyperactivity and Prediction of Educational Attainment in Early Adulthood: A 16-Year Longitudinal Population-Based Study. *American Journal of Psychiatry*, 168(11):1164–1170.

[97] Pourahmadi, M. (1999). Joint Mean-Covariance Models with Applications to Longitudinal Data : Unconstrained Parameterisation. *Biometrika - Oxford Journals*, 86(3):677–690.

[98] Prem, K., Liu, Y., Russell, T. W., Kucharski, A. J., Eggo, R. M., Davies, N., Flasche, S., Clifford, S., Pearson, C. A., Munday, J. D., Abbott, S., Gibbs, H., Rosello, A., Quilty, B. J., Jombart, T., Sun, F., Diamond, C., Gimma, A., van Zandvoort, K., Funk, S., Jarvis, C. I., Edmunds, W. J., Bosse, N. I., Hellewell, J., Jit, M., and Klepac, P. (2020). The effect of control strategies to reduce social mixing on outcomes of the COVID-19 epidemic in Wuhan, China: a modelling study. *The Lancet Public Health*, 2667(20):1–10.

[99] Proust-Lima, C., Philipps, V., and Liquet, B. (2017). Estimation of extended mixed models using latent classes and latent processes: The R package lcmm. *Journal of Statistical Software*, 78(2).

[100] Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850.

[101] Rao, A. S. and Vazquez, J. A. (2020). Identification of COVID-19 Can be Quicker through Artificial Intelligence framework using a Mobile Phone-Based Survey in the Populations when

Cities/Towns Are under Quarantine. *Infection Control and Hospital Epidemiology*, pages 1–5.

[102] Rao, C. R. (1948). The Utilization of Multiple Measurements in Problems of Biological Classification. *Journal of the Royal Statistical Society: Series B (Methodological)*, 10(2):159–193.

[103] Ray, S. and Turi, R. (1999). Determination of number of clusters in k-means clustering and application in colour image segmentation. *Proceedings of the 4th international conference on advances in pattern recognition and digital techniques*, pages 137–143.

[104] Redner, R. and Walker, H. (1984). Mixture Densities, Maximum Likelihood and the Em Algorithm. *Society for Industrial and Applied Mathematics*.

[105] Rogers, D. J. and Tanimoto, T. T. (1960). A computer program for classifying plants. *Science*, 132(3434):1115–1118.

[106] Rousseeuw, P. J. (1987). Silhouettes : a graphical aid to the interpretation and validation of cluster analysis. 20:53–65.

[107] Sakoe, H. and Chiba, S. (1978a). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1):43–49.

[108] Sakoe, H. and Chiba, S. (1978b). Performance Tradeoffs in Dynamic Time Warping Algorithms. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1):43–49.

[109] Saria, S. and Goldenberg, A. (2015). Subtyping: What It is and Its Role in Precision Medicine. *IEEE Intelligent Systems*, 30(4):70–75.

[110] Savage, R. S., Heller, K., Xu, Y., Ghahramani, Z., Truman, W. M., Grant, M., Denby, K. J., and Wild, D. L. (2009). R / BHC : fast Bayesian hierarchical clustering for microarray data. 9:1–9.

[111] Schramm, C., Vial, C., Bachoud-Lévi, A. C., and Katsahian, S. (2018). Clustering of longitudinal data by using an extended baseline: A new method for treatment efficacy clustering in longitudinal data. *Statistical Methods in Medical Research*, 27(1):97–113.

[112] Shim, Y., Chung, J., and Choi, I. C. (2005). A comparison study of cluster validity indices using a nonhierarchical clustering algorithm. *Proceedings - International Conference on Computational Intelligence for Modelling, Control and Automation, CIMCA 2005 and International Conference on Intelligent Agents, Web Technologies and Internet*, 1:199–203.

[113] Silva, A. L., Dawson, S. N., Arends, M. J., Guttula, K., Hall, N., Cameron, E. A., Huang, T. H., Brenton, J. D., Tavaré, S., Bienz, M., and Ibrahim, A. E. (2014). Boosting Wnt activity during colorectal cancer progression through selective hypermethylation of Wnt signaling antagonists. *BMC Cancer*, 14(1):1–10.

[114] Silva, L., Peres, S., and Boscarioli, C. (2016). *Introdução a Mineração de Dados com aplicações em R*. Elsevier.

[115] Smith, K. F., Goldberg, M., Rosenthal, S., Carlson, L., Chen, J., Chen, C., Ramachandran, S., and Smith, K. F. (2014). Global rise in human infectious disease outbreaks. pages 1–6.

[116] Sokal, R. and Sneath, P. (1963). Principles of numerical taxonomy. *W.H. Freeman & Company*.

[117] Sousa, P., Oliveira, A., Gomes, M., Gaio, A. R., and Duarte, R. (2016). Longitudinal clustering of tuberculosis incidence and predictors for the time profiles: The impact of HIV. *International Journal of Tuberculosis and Lung Disease*, 20(8):1027–1032.

[118] Thalamuthu, A., Mukhopadhyay, I., Zheng, X., and Tseng, G. C. (2006). Evaluation and comparison of gene clustering methods in microarray analysis. *Bioinformatics*, 22(19):2405–2412.

[119] Tiedeman, D. (1955). On the study of Types. *Symposium on Pattern Analysis*.

[120] Torales, J., O'Higgins, M., Castaldelli-Maia, J. M., and Ventriglio, A. (2020). The outbreak of COVID-19 coronavirus and its impact on global mental health. *International Journal of Social Psychiatry*, pages 3–6.

[121] Wang, W. and Muntz, R. (1997). STING : A Statistical Information Grid Approach to Spatial Data Mining.

[122] Wemmert, C., Gancarski, P., and Korczak, J. J. (2000). A collaborative approach to combine multiple learning methods. 9(1):59–78.

[123] Williams, D. R., Kontos, E. Z., Viswanath, K., Haas, S., Lathan, C. S., Macconaill, L. E., and Ayanian, J. Z. (2012). Integrating Multiple Social Statuses in Health Disparities Research : The Case of Lung Cancer. pages 1255–1277.

[124] Wolfe, J. (1965). A Computer Program for the Maximum Likelihood Analysis of Types. *Technical Bulletin 65-15, U.S. Naval Personnel Research Activity.*

[125] Xu, Z., Shi, L., Wang, Y., Zhang, J., Huang, L., Zhang, C., Liu, S., Zhao, P., Liu, H., Zhu, L., Tai, Y., Bai, C., Gao, T., Song, J., Xia, P., Dong, J., Zhao, J., and Wang, F. S. (2020). Pathological findings of COVID-19 associated with acute respiratory distress syndrome. *The Lancet Respiratory Medicine*, 2600(20):19–21.

[126] Yeung, K. Y. and Ruzzo, W. L. (1979). Details of the Adjusted Rand index and Clustering algorithms Supplement to the paper "An empirical study on Principal Component Analysis for clustering gene expression data. *Bioinformaticsioin*, 9:175–186.

[127] Zeger, L. and Liang, S. (1986). Longitudinal Data Analysis for Discrete and Continuous. *Biometrics*, 42(1):121–130.

[128] Zhao, Q., Xu, M., and Fränti, P. (2009). Sum-of-squares based cluster validity index and significance analysis. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 5495 LNCS:313–322.

[129] Zhou, F., Yu, T., Du, R., Fan, G., Liu, Y., Liu, Z., Xiang, J., Wang, Y., Song, B., Gu, X., Guan, L., Wei, Y., Li, H., Wu, X., Xu, J., Tu, S., Zhang, Y., Chen, H., and Cao, B. (2020). Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. *The Lancet*, 395(10229):1054–1062.

[130] Zhu, X. and Qu, A. (2018). Cluster analysis of longitudinal profiles with subgroups. *Electronic Journal of Statistics*, 12(1):171–193.

Appendix

Figure 5.1: European countries present in the CMR clustering analysis

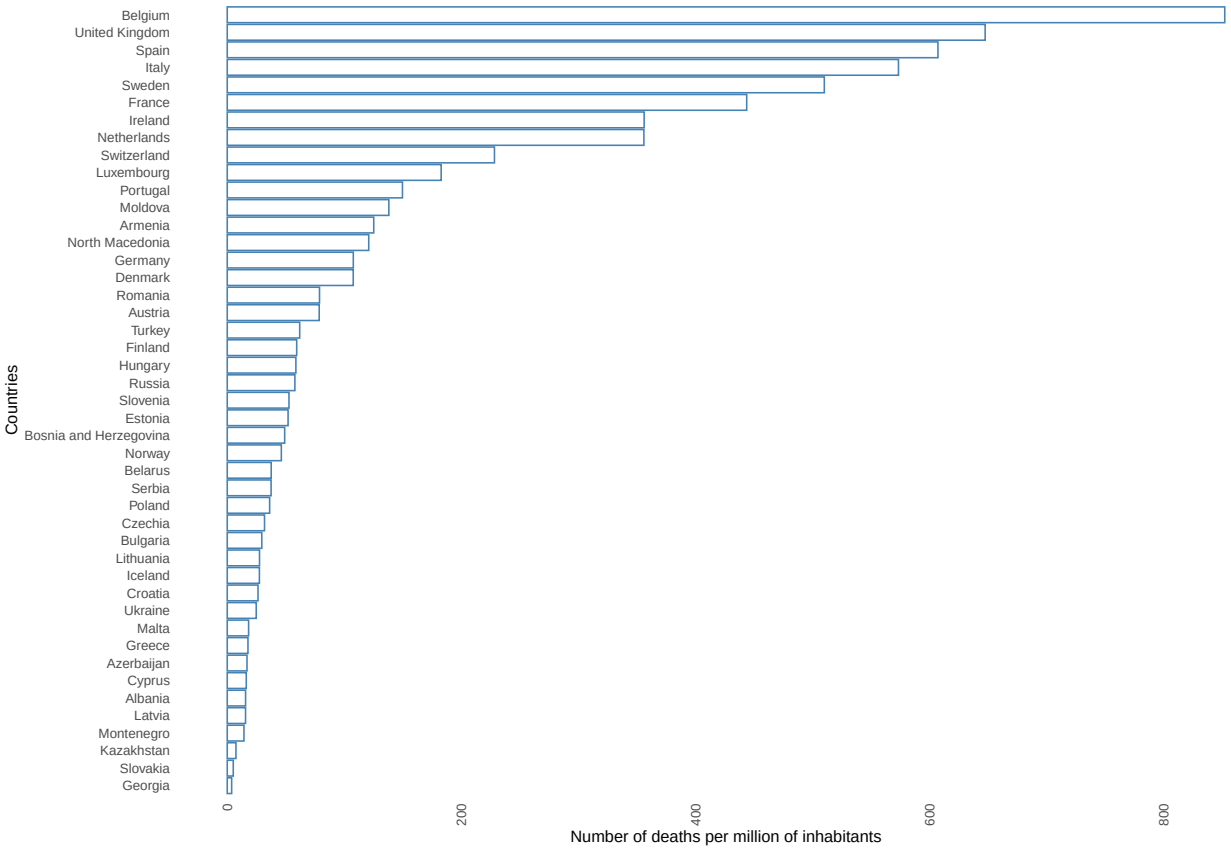


Figure 5.2: Number of deaths per million of inhabitants (Last updated: 24th of June)

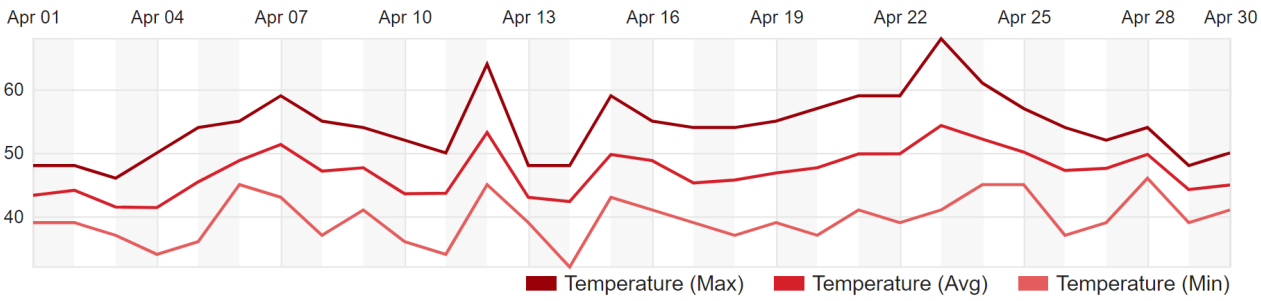


Figure 5.3: Temperature registered in Copenhagen (Denmark) during the month of April

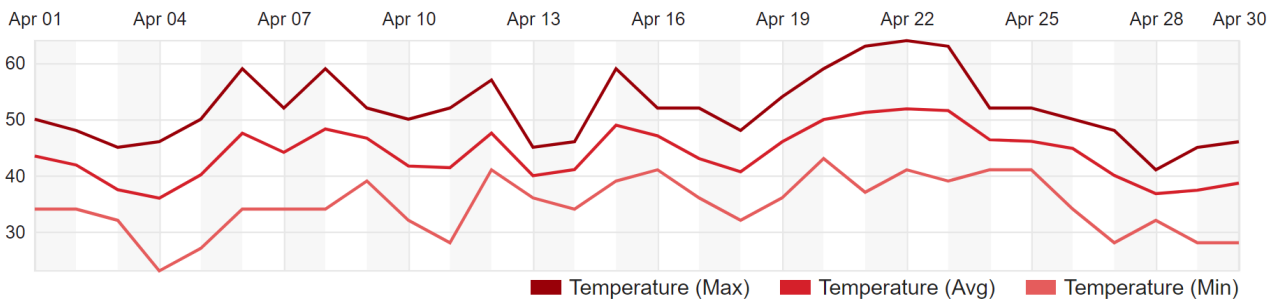


Figure 5.4: Temperature registered in Stockholm (Sweden) during the month of April