FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Comparison of hierarchical clustering analyses of neurodegenerative diseases

Mariana Silva Santos

Mestrado Integrado em Bioengenharia

Supervisors: Alexandra Oliveira and Luís Paulo Reis

February 1, 2021

Comparison of hierarchical clustering analyses of neurodegenerative diseases

Mariana Silva Santos

Mestrado Integrado em Bioengenharia

February 1, 2021

Abstract

Parkinson's disease is a neurodegenerative disease, that doesn't have a cure yet. The symptoms of this disease consist of tremors, bradykinesia, and rigidity of movements.

There are other neurodegenerative diseases, such as Huntington's disease and Amyotrophic Lateral Sclerosis, whose symptoms also consist of gait problems. In Amyotrophic Lateral Sclerosis the motor nerves' from the cortex to the spinal cord's anterior horn, starts dying limiting movements. In Huntington's disease, there is a gradual degeneration of parts of the basal ganglia responsible for motor coordination.

Since the most common symptoms of these neurodegenerative diseases are associated with the motor domain, a deeper analysis of gait may help in a diagnose.

The gait data collection can be done using pressure sensors placed on the feet of the subject under study. The data obtained from these sensors can be called time series.

A time series is a collection of observations made sequentially over time. In order to find similar times series patterns, some clustering techniques can be applied. These methods are called clustering and make time series analysis and study much more straightforward.

This dissertation aims to study in-depth clustering methods in order to applied to time series and group individuals with similar gait characteristics. Firstly, it is studied Parkinson's disease, Huntington's disease, and Amyotrophic Lateral Sclerosis. Then it is studied clustering, in particular, Hierarchical Clustering. Two datasets were analysed: one with data of the gait of patients with Parkinson's disease and healthy patients; and the other with data of the gait of the three diseases understudy and control patients.

Finally, it was created a hierarchical clustering algorithm to group elements with similar gait characteristics. To see which method achieves the higher purity, four distance measures were combined with seven aggregation methods.

In conclusion, the distance measure combined with the aggregation method that achieved the highest purity percentage in the first dataset was the Manhattan distance with Weighted method. However, in the second dataset, the distance measure combined with the aggregation method that achieved the highest purity percentage was the Fourier distance with Weighted method. So, in both datasets, the aggregation method that achieved the best results was the Weighted. However, the best distance measure was different in both datasets.

Resumo

A Doença de Parkinson é uma doença neurodegenerativa, ainda sem cura. Os sintomas desta doença consistem em tremores, lentidão e rigidez dos movimentos.

Existem outras doenças neurodegenerativas, tais como a doença de Huntington e a Esclerose Lateral Amiotrófica, cujos sintomas também consistem em problemas na marcha. Na Esclerose Lateral Amiotrófica os músculos estriados esqueléticos ficam mais fracos, limitanto o movimento. Na doença de Huntington ocorre uma degeneração gradual de partes dos gânglios basais, sendo que estes são responsáveis pela coordenação motora.

Dado que os sintomas mais comuns destas doenças estão associados à parte motora, uma análise aprofundada da marcha de um indivíduo poderá ser ajudar no diagnóstico.

O estudo da marcha poderá ser feito utilizando sensores de pressão colocados nos pés do indivíduo em estudo. Os dados obtidos a partir destes sensores, enquanto o indivíduo é sujeito realiza uma pequena caminhada, podem ser chamados de séries temporais.

Uma série temporal é uma coleção de observações feitas sequencialmente ao longo do tempo. De modo a analisar este tipo de dados, existem métodos que agrupam os objetos semelhantes das séries. Estes métodos são chamados métodos de *clustering* e facilitam muito a análise e estudo das séries temporais.

Os objetivos desta dissertação consistem em estudar aprofundadamente métodos de *clustering* de séries temporiais, com o fim de agrupar elementos com características de marcha semelhantes.En primeiro lugar, estudar a doença de Parkinson, a doença de Huntington e a Esclerore Lateral Amiotrófica. Em seguida, estudar os métdos de *clustering*, em particular, o *clustering* hierarquico. Foram analisadas duas bases de dados: uma com dados da marcha de pacientes com Parkinson e pacientes Saudáveis; e outra com dados da marcha das três doenças em estudo e de pacientes controlo.

Por fim, foi criado um algorimo de *clustering* hierarquico com o objetivo de agrupar elementos com caracteristicas na marcha semelhantes. De modo a perceber qual o método que alcança uma percentagem de pureza mais elevada, quatro medidas de distância foram combinadas com sete métodos de agregação.

Em conclusão, a medida de distância combinada com o método de agregação que alcançou a maior percentagem de pureza no primeiro *dataset* foi a distância de Manhattan com o método Weighted. No entanto, no segundo *dataset*, a medida de distância combinada com o método de agregação que alcançou a percentagem de pureza mais elevada foi a distância de Fourier com o método Weighted. Assim, em ambos os conjuntos de dados, o método de agregação que alcançou os melhores resultados foi o método Weighted. Contudo, a melhor medida da distância foi diferente em ambos os conjuntos de dados.

Acknowledgements

I would like to thank my supervisors, Prof. Alexandra Oliveira and Prof. Luís Paulo Reis, for all their guidance in developing this dissertation. Thank you for everything you teach me during this semester.

I want to thank my friends for helping me in hard times and make my academic life better. Thank you for your friendship.

Furthermore, I want to thank my family, in special my brother, for their support and caring.

A big thank you for my mother, who is my role model and my light in the dark.

And a special thank for my boyfriend, João, for always being there and for his love, patience and help.

"If you think you are too small to make a difference,
try sleeping with a mosquito."

Dalai Lama

Contents

List of Figures

List of Tables

Abbreviations

DTW	Dynamic time warping
ALS	Amyotrophic Lateral Sclerosis

Chapter 1

Introduction

A time series is a set of observations made over time. In present times, there is much information and, consequently, there are large volumes of time-correlated data, to help analyse it, there is a need to look for patterns of similar behaviour. Especially in the area of health, there is also this need to find patterns.

Neurodegenerative diseases are incurable and debilitating diseases that result in gradual degeneration or death of humans brain's neurons. Indeed, they can cause problems of movement or mental function. The neurodegenerative diseases that are studied in this dissertation are: Parkinson, Huntington and ALS diseases.

Due to the motor problems associated with these conditions, people who suffer from neurodegenerative diseases may have different gait patterns. So, we can compare groups of people over time to distinguish and identify normal or abnormal temporal patterns, in order to help in diagnosis.

In this dissertation, Hierarchical Clustering algorithms are applied, in order to compare different gait patterns.

1.1 Background and Motivation

Parkinson's disease is a neurodegenerative disease that affects 1 % of the world's population over 65 years of age [10]. This disease has no known cure. The symptoms consist of tremor, rigidity, bradykinesia (slow movement) and postural instability (balance problems). Thus, Parkinson's is diagnosed through scales that classify the disease's severity, which essentially evaluate patients' motor capacities. In this way, an in-depth analysis of Parkinson's patients' gait is also an asset for diagnosing the disease.

The Amyotrophic Lateral Sclerosis is a neurodegenerative disease that has an annual incidence of about 1.9 per 100,000 [11]. This disease causes the death of motor nerves from the cortex to the anterior horn of the spinal cord, limiting movement [12]. The main symptoms of Amyotrophic Lateral Sclerosis are dysphagia, dysarthria, breathing difficulties, pain and psychological disorders.

Huntington's disease is a rare neurodegenerative disease that has an incidence of 0.38 per 100,000 per year [13]. This disease affects the patient's movement, causing involuntary and undesired movements. People who suffer from this disease have a gradual degeneration of parts of the basal ganglia responsible for motor coordination.

Humans' gait is a periodic movement that can be studied by inserting pressure sensors at the base of an individual that records the variation of pressure in his feet while walking.

To analyze a large amount of data such as time series, clustering methods can be used. These methods group similar objects, in order to make their analysis make it easier to analyse. In this dissertation's case, these methods group the data from pressure sensors from different individuals, which are similar by some similarity criteria.

It is expected that the walking patterns of individuals with distinct conditions will differ. Hence, the search for similar gait patterns could be relevant for diagnostic purposes.

1.2 Goals

The main goal of this dissertation is to compare different gait patterns between individuals with neurodegenerative diseases. In order to complete this goal, some tasks have to be done:

1. Define the concept of Time Series as well as define and analyse the state of the art time series clustering methods;
2. Analyse neurodegenerative diseases: Parkinson, Huntington and Amyotrophic Lateral Sclerosis;
3. Define human gait and related context variables;
4. Preprocessing the data;
5. Analyse and implement distances measures between Time Series;
6. Analyse and implement aggregation measures for hierarchical clustering methods;
7. Choose results evaluation metrics;
8. Compare the solutions obtained the several combinations of distance metrics and aggregation measurements.

1.3 Dissertation Structure

This dissertation has six chapters.

In chapter 2, the Parkinson's disease, the symptoms of the patients and the forms of diagnosis are described. It is also explained two other neurodegenerative diseases: Huntington's disease and Amyotrophic Lateral Sclerosis. Furthermore, it explains the importance of walking to diagnose neurodegenerative diseases. The gait is also a relevant part in this chapter, and a gait of a normal

and healthy individual is explained. Another variable that is relevant for gait patterns is also explored: the age.

In chapter 3, it is presented the state of the art and the methodology. Time series and one of the methods to analyze it is presented. This method, named clustering, is a Machine Learning technique that consists of aggregate elements with similar features. Clustering can be applied using different algorithms, such as Hierarchical Clustering. This algorithm was the chosen one to apply in this dissertation. A bibliographic review is done based on the concepts explored in the chapter 2 and Time Series. And, in the last section, the methodology is presented.

In the chapter 4 the first dataset, *Gait in Parkinson's Disease*, is analysed. Demographic data are first analysed using graphs. Then the entire data processing is explained before applying the hierarchical clustering. Finally, the results were interpreted, and an evaluation of the distance methods and aggregation measures was done.

In chapter 5 the second dataset, named *Gait in Neurodegenerative Disease*, is analysed. The procedure is the same as the first dataset.

Finally, in chapter 6 some conclusions drawn from this study are presented. Limitations of the study are also described.

Chapter 2

Neurodegenerative diseases and Gait

2.1 Parkinson's Disease

Parkinson's disease is a neurodegenerative disease that affects the central nervous system. It is characterized by nerve cells' inability to produce sufficient neurotransmitters in the brain, which are called dopamine. Dopamine controls muscle activity and, for this cause, the typical symptoms of the disease are essentially tremors, slowness and stiffness of movement [14].

Parkinson has no cure, but some drugs can reduce the symptoms. The treatment choice for Parkinson's disease depends on the patient's specific symptoms and age. If the motor symptoms are very severe, potent treatment such as a dopamine agonist or levodopa are administrated. Levodopa gives the most significant symptomatic advantage, but can cause dopaminergic motor complications, like dyskinesia. If the motor symptoms are not severe, a monoamine oxidase type B inhibitor is administrated [15]. The drug that appers to have more effect in reduce gait problems is the levodopa. [15].

The most used scale to classified the severity of the Parkinson's disease is the Hoehn and Yahr scale. The progression of the disease can be classified into five states, according to this scale. State 1 is characterized by tremors and other movement-related symptoms that occur only on one side of the body, not affecting the patient's daily routine. State 2 involves tremor and stiffness of movements, which affect both sides of the patient's body. In this state, the patient can still live alone; however, the daily tasks become more difficult. In state 3, the patient begins to lose balance and shows a slowness in performing movements. Essential activities like dressing and eating become complicated for the patient. State 4 is characterized by a very significant increase in symptoms and may require a walker. The patient begins to need help to perform their daily tasks and becomes incapable of living alone. Finally, in state 5, the patient becomes incapable of walking or standing. He may also suffer from hallucinations or delirium [16].

Beyond this scale for classifying the severity of the disease, there are others that evaluate the patient's motor capacity. Now, the *Unified Parkinson's Disease Rating scale* (UPDRS) is the scale that best measures the disability of Parkinson's patients. The minimum number on this scale is zero (totally healthy patients), and the maximum number on the scale is 199 (for totally

disabled patients). Studies using this scale conclude that Parkinson's disease does not have a linear detectorization and it is faster at an early stage of the disease [17, 18].

Since the previous scale only assessed Parkinson's disease's motor signs, there was a need to modify it to assess the impact of this disease's non-motor aspects on patients' daily life experiences (UPDRSm). The limits of this scale are similar to those of its mother scale [18].

Further, the *Timed Up & Go* (TUG) scale is present. It consists of a physical ability test that evaluates the patient's ability to get up, walk three meters, turn around, walk backwards, and sit down again. The longer it takes the patient to do this test, the less mobile they are and consequently, the more probable they are to fall [19].

Approximately 60 000 Americans are diagnosed with this disease every year, and there is no effective diagnosis for this disease [20]. Diagnosis is now essentially made either through the previous scales or through neuronal imaging analysis. Nevertheless, in the early stages of the disease, this is not an efficient method [20]. Hence, there is a need to create new ways of diagnosing the condition.

In this way, given that one of the disease's primary symptoms is movement instability, an analysis of the patient's gait could be an essential asset in the Parkinson's diagnosis.

2.2 Huntington's disease

Huntington's disease is a neurodegenerative disease that passes from generation to generation. It is a disease that affects the patient's movement, causing involuntary and undesired movements. It also affects their cognitive component and leads to mental disorders [21].

Approximately 30,000 people in the United States have this disease, and 200,000 people are at chance of developing it [22].

The disease has no cure, but some drugs can reduce the symptoms [21].

All patients develop hypokinesia, akinesia, and stiffness of movements. These motor symptoms are treated with dopamine receptor blocking agents [23].

Huntington's disease is caused by the expansion of a gene that codifies a protein named Huntington's. This expansion leads to the death of neurons in some areas of the brain. There is a detection test for this gene, and it can be done at any stage of development, even before the patient presents symptoms [21].

2.3 Amyotrophic Lateral Sclerose

Amyotrophic Lateral Sclerosis is a neurodegenerative disease, as is Parkinson's and Huntington's Disease. The cause of this disease is the death of motor nerves from the cortex to the anterior horn of the spinal cord[12].

The main symptoms of amyotrophic lateral sclerosis are dysphagia, dysarthria, breathing difficulties, pain and psychological disorders [12].

Approximately 6 000 Americans are diagnosed with this disease every year. Patients with ALS (amyotrophic lateral sclerosis) have a prognosis for survival of 2 to 5 years. The symptoms that allow the diagnosis appears 9 to 12 months after the development of the disease, and by then 50% to 70% of the motor neurons are already nonfunctional [24].

The drug used until 2017 for ALS treatment was only Riluzole and it inhibits the release of glutamate. This drug was approved for the FDA in 1995. In 2017 a new intravenous drug was approved for the FDA: Edaravone. This new drug can be taken with the previous. In Europe,this drug is still not used so patients with take Rizulose in order to reduce the symptoms [25].

2.4 Gait

The gait analysis is the study of human locomotion. As a general rule, sensors, such as accelerometers or force sensors, are used to perform this analysis, connected to some part of the human body [1].

Usually, the gait of an individual is a periodic movement. Therefore, the human gait pattern can be divided into eight different phases grouped into two periods: the stance period (a) and the swing period (b). The stance phase happens when the foot is in contact with the ground, and the swing phase happens from the moment the foot rises from the ground, as suggested in Figure 2.1. During the stance phase in one of the feet, the body's whole weight is supported at this extremity [2].

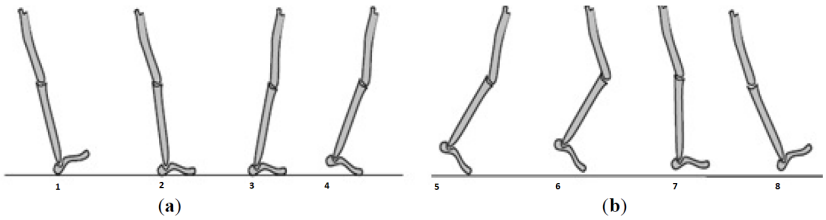


Figure 2.1: Phases of the Human Gait: (a) stance period and (b) swing period (Adapted from [1])

Each of these periods is divided into four phases [2, 1]:

1. Initial Contact: is the moment when the heel touches the ground;
2. Loading Response: the foot of the individual adheres to the floor, taking the form of a flat foot. This phase only ends when the other foot is raised to start the swing phase;
3. Med stance: the moment when the bodyweight is all in one foot;
4. Terminal stance: at this moment the bodyweight is supported on the forefoot. This phase ends when the other foot finishes the swinging phase and reaches the ground;
5. Pre-swing: the objective of this phase is to position the limb for the swing. It ends with the tip of the hallux leaving the floor;

6. Initial swing: is approximately 1/3 of the swing period. It starts when the foot rises from the ground and ends when the swinging foot is in the opposite position of the foot in the stance phase;

7. Mid swing: happens when the foot is under the body;

8. Terminal swing: starts when the tibia is in a vertical position and ends when the foot touches the ground.

The normal distribution of the contact periods with the ground is 60 % for the stance period and 40 % for the swing period. One member's stance period is equal to the other member's swing period since, during a walk, when one foot lands, the other foot rises. The walking cycle time varies according to the walking speed [2].

If an individual walks at a speed of 80 meters per minute, the stance and swing periods represent 62% and 38 % of the walking cycle, respectively. Thus, the duration of the walking cycle is inverse to the walking speed. That is the support and balance period decrease as the walking speed increases [2].

As the speed decreases, the greater the time differences between the support period and the swing period. If the person is walking, there is a time range when both feet contact the ground. This happens while one of the limbs moves to the swing phase and the other to the stance phase. When this period is very short, it means that the individual has gone from walking to running. [2]

The concept of Stride is used to describe a single-member gait cycle. Stride time is the interval between two contacts on the ground by the same member. One Step is the time until the foot in contact with the ground switches. In this way, there are two steps in each Stride.

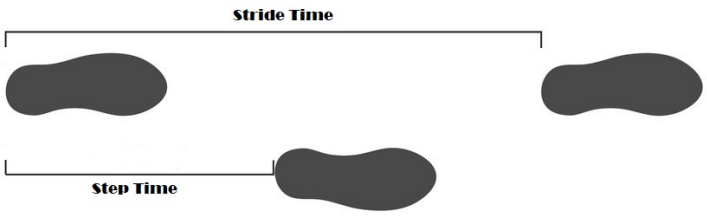


Figure 2.2: Stride Time and Step Time (Adapted from [2])

One of the ways to evaluate an individual's gait is through force sensors. In this way, if force sensors are placed on the base of a subject's foot, sequential data are obtained over time. These data are called time series and will be studied in the next Chapter.

The Ground Reaction Forces (*GRF*) caused by the gait of a normal subject moving at approximately 82m/min have two local maximums. At this speed, typically, the maximums' values are approximately 110% and 80% of body weight. In this sense, the first peak occurs at the beginning of the Middle Stance phase represented by (3) of the Figure 2.1. In the middle of the two peaks, is created a valley by the rise of the centre of gravity, as the body moves forward over the stationary

foot. Moreover, the second peak occurs at the terminal swing phase represented by (8) of the Figure 2.1(Figure 2.3) [2].

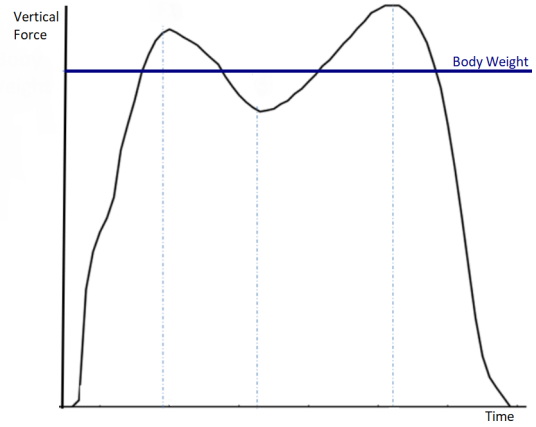


Figure 2.3: Ground Reaction Force to the ground (*GRF*) of a normal individual while walking. (Adapted from [3])

For individuals suffering from neurodegenerative diseases such as Parkinson's, are expected gait changes result in reduced gait speed, shortened stride time and reduced swing time[26]. The swing time reduces because people with gait problems are expected to spend less time with feet on the air. So, it is expected that they present notorious differences comparing to normal individuals[26].

2.5 Gait and Age

Elderly is followed by neuromuscular alterations such as modified muscle force, stability, joint movement, and cardiovascular health [27]. These alterations can contribute to a deterioration in mobility and execution of actions of the quotidian. Decreased mobility is a notable concern in elderly individuals and is a cause for morbidity, disability, and mortality [27]. In ageing, degeneration of one or more sensory systems happen and may compromise stability during gait [28].

Modifications in the nervous system due to age include enhanced reaction times, enhanced brain loss rate, and decreased level of neurotransmitter creation like dopamine [28]. As referred before, dopamine is the principal responsible for controlling muscle activity.

Individuals who have 65 years have a lower number of dopamine-producing cells, and when compared to a 20-year-old, they have less 30–50% cells. [29].

A review article from Katherine A. Boye et al.[27] analysed 29 studies of healthy older adults (71.6 ± 3.3 years) compared to young adults (27.2 ± 2.1 years). The conclusion was that there are standardised effects of age for ankle kinematics, for propulsive ankle moments and powers, and ground reaction forces. In 21 of 23 studies, the stride length is bigger for younger individuals.

Furthermore, the article [30] also concludes that the stride length decreases with age. In this way, the results show that the stride lengths averaged for young individuals (between 20 to 40

years) is 151 to 170 cm. However, old individuals' stride lengths (between 60 to 80 years) averaged is 135 to 153 cm.

The swing time is a variable related to the stride length, and therefore, it can be said that the age affects also the swing time.

In conclusion, individuals older than 65 have relevant differences in their gait pattern compared to younger individuals.

Chapter 3

State of the art and Methodology

3.1 Concept of Time Series

Time series are present in many daily life situations, and for this reason, they have been studied by experts with various profiles. One of the most significant areas of application of time series is medicine. These can be used to detect brain activity or the diagnosis of pathologies, for example [5].

A univariate time series is a collection of observations made sequentially over time. Therefore, it can be said to be a type of time data that usually presents large dimensions. It is important to consider the dependence between neighbouring observations.

Thus, a time series, defined in a continuous time interval, can be defined as ([31]):

$$x(t), t^* \leq t \leq t^{**} \tag{3.1}$$

The standard graphic of a continuous-time series is a line that varies in time. An example of this type of graph is the graph with data from sensors located at the foot of an individual, recording the total pressure variation during a short walking task. Thus, Figure 3.1 presents two time series, one illustrating the variation in the ground reaction force of a healthy individual's right foot and the other showing the variation in the ground reaction force of a healthy individual's left foot.

In order to classify a large amount of data, there are techniques commonly used in artificial intelligence called clustering techniques.

3.2 Clustering

In the review article of Aghabozorgi in [5], it is analysed the technique of clustering. This technique aims to group objects that present similar characteristics, which makes their analysis much more manageable. It is used to analyse the time series since they are too large to be analysed manually by humans.

Given a set of time series D, defined as $D = \{F_1, F_2, ..., F_n\}$, the process of splitting into clustering groups transforms D into multiple clusters C_i, defined as $C = \{C_1, C_2, ..., C_k\}$, $k \leq n$. These

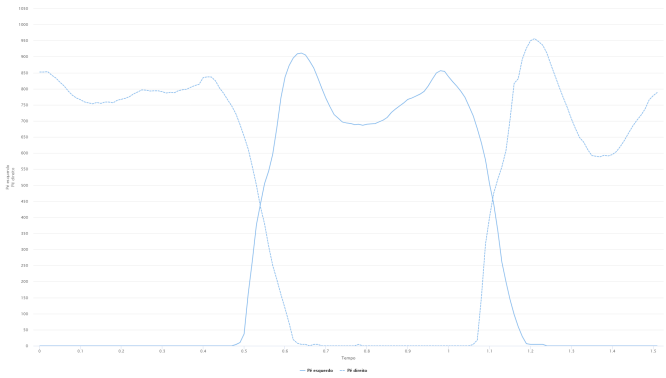


Figure 3.1: Time Series that illustrate the gait of a healthy individual. (Graph generated from data extracted from the project *Physionet* [4])

clusters are homogeneous groups and aggregated with a certain similarity criterion. Therefore, C_i is called a cluster, where:

$$D = \cup_{i=1}^{k} C_i \tag{3.2}$$

$$C_i \cap C_j = \emptyset \tag{3.3}$$

para $i \neq j$.

Time series clustering can be classified into three categories: Whole time-series, Subsequence time-series clustering and time point clustering as seen in the scheme of Fig 3.2. ([5]):

1. *Whole time-series clustering*, process where a set of similar individual time series are joined together.

2. *Subsequence time-series clustering*, where small segments of a time series are grouped, extracted through the sliding window algorithm.

3. *Time point clustering*, where time points are grouped based on the combination of time proximity and the similarity of the value corresponding to the points. This process is applied in a single time series.

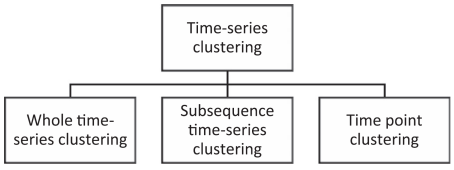


Figure 3.2: Taxonomy of time series clustering. [5]

3.3 Whole time-series clustering

In this dissertation, the focus will be on clustering using the entire time series - Whole time-series clustering. For time series clustering, there are essentially three ways, (Fig. 3.3):

1. *Model-based*: this approach considers that the time series data were created by a parametric model and attempts to produce the original model from the data. The model that we obtain determines clusters. [32]

2. Feature-based: in this approach, raw time series are transformed into a feature vector, smaller than the original time series. A conventional clustering algorithm is then applied to this feature vector, such as k-means, and the outcome is clustered. [32]

3. Shape-based: in this approach, two time series forms are harmonized as better as possible, contracting and expanding the time axis scale. It normal works with raw data.[5]

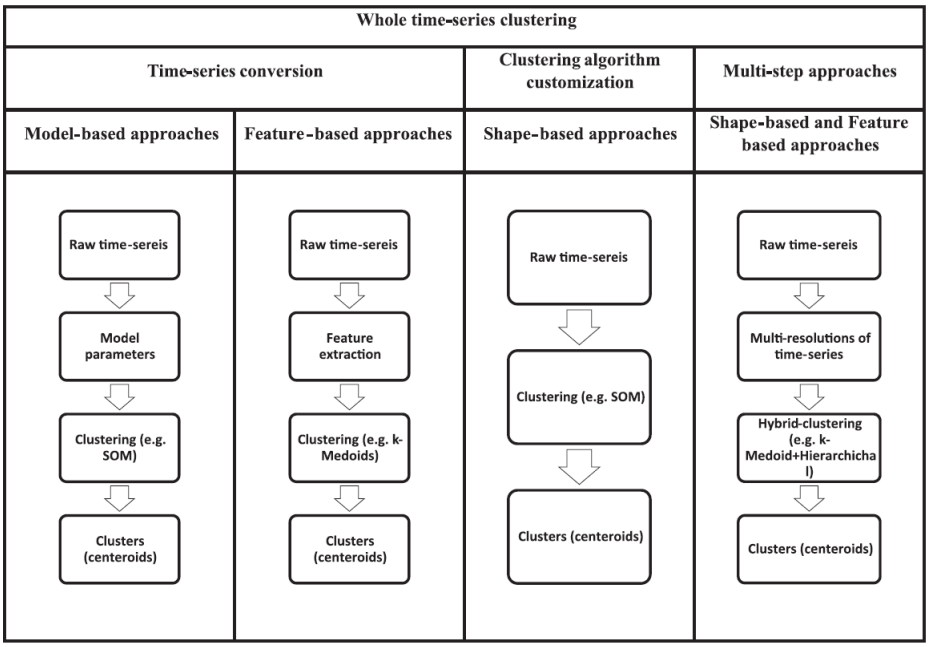


Figure 3.3: Approaches for the Whole time-series clustering method (Adapted from [5])

3.3.1 Components

The Whole Time-Series clustering method consists essentially of four components: representation of the time series, the definition of similarity or distance, determination of the prototype/representative and definition of clusters. The components are represented in Fig. 3.4.

To cluster a time series using this method may require all these components or just a few [5].

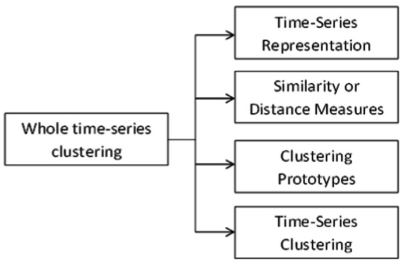


Figure 3.4: Whole time-series clustering method components (Adapted from [5])

3.3.1.1 Time series representation or size reduction

The first component is the time series representation or the reduction of its size. This component projects the series in another smaller dimension space. This reduction is significant since it reduces memory necessity since a time series is too big to be in the main memory. Indeed, a reduction in the series's size also makes it easier to calculate the distance, reducing clustering time.

Given a time series:

$$F_i = \{f_1, ...f_t, ..., f_T\} \tag{3.4}$$

the series representation is to transform it into a reduced vector, such as [5]:

$$F_i = \{f_1, ..., f_x\} \tag{3.5}$$

where $x < T$.

3.3.1.2 Similarity or dissimilarity between time series

The second component is the similarity or dissimilarity between time series. To explain the similarity between two time series, we have to introduce the concept of distance, because they are very connected. The distance is the size of the path that connects two objects, so it is a measure of similarity. The bigger the distance between two time series, the less similar they are [33].

There are various distance measurements designed to specify the similarity between time series and there are distance measures designed for objects that can be applied to time series with adaptations [5]. For example, to measure the distance between time series it can be used the Dynamic Time Warping (DTW), Euclidean, Manhattan and Fourier distance [33].

There are three primary objectives in the clustering of time series, all of which demand distinct approaches.

1. Similarity in time: the first method is to group series that vary in the same way at the same time moments. One way to measure this type of distance is to use correlation-based distances, Euclidean distance or Manhattan distance. Figure 3.5 shows two clusters in which two time series change similarly at each instant of time. [6]

2. Similarity in shape: the next possible method is to group the series using their common forms, independently of time. Generally, the two approaches to achieving this goal are either

transforming the data using techniques such as dynamic time warping (DTW) or developing specific algorithms for the matching of subsequence patterns. Figure 3.6 presents three time-series clusters with the same characteristics in their shape. [6]

3. Similarity in change: the last method is to group time series that vary in the same way over time. This method generally uses the Markov Occult Model (HMM) or an autoregressive moving average process (ARMA), which fits the series and is grouped based on the parameters' similarity in the models. The figure 3.7 illustrates two clusters of time series, the left one being produced by the autoregressive order 2 model (AR(2)) and the right one by the autoregressive order 2 ARMA and the order 2 ARMA in the moving average component (ARMA(2,2)). Although the series belonging to the same cluster have different shapes, they vary in the same way in each instant of time, in other words, each point depends on the previous instant of time similarly. [6]

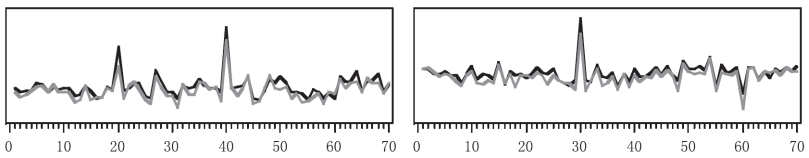


Figure 3.5: Similarity in time (Adapted from [6])

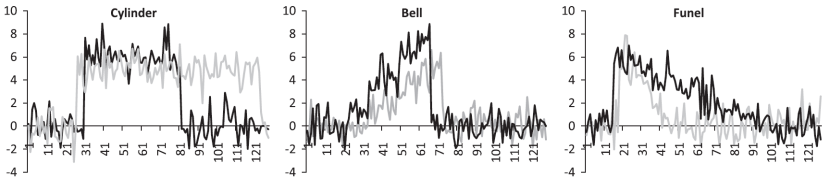


Figure 3.6: Similarity on shape (Adapted from [6])

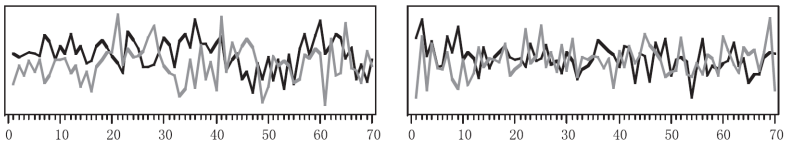


Figure 3.7: Similarity in change (Adapted from [6])

In this dissertation, we will identify five different distance measures: Euclidean, Manhattan, Dynamic Time Warping and Fourier Discrete Transform.

The most used and simplest time domain distance measure is derived from the Minkowski distance and it is named Euclidean Distance. The Minkowski is defined as [33]:

$$D_{Minkowski}(X(t),Y(t)) = [\sum_{i=1}^{N}(|x_t - y_t|)^r]^{\frac{1}{r}} \tag{3.6}$$

The X(t) and Y(t) are two time series and N is their length. This distance is defined in vector space R^N.

The Minkowski distance is equal to the Euclidean distance when r = 2 in and to the Manhattan distance when r = 1 [33].

$$D_{Euclidean}(X(t),Y(t)) = [\sum_{i=1}^{N}(|x_t - y_t|)^2] \tag{3.7}$$

$$D_{Manhattan}(X(t),Y(t)) = [\sum_{i=1}^{N}(|x_t - y_t|)] \tag{3.8}$$

The most commonly used measure of similarity is the Euclidean distance because it is easy to implement, fast and efficient in time and space ([34]). However, this measure is susceptible to distortion on the time axis and has low precision [35].

The most popular measurement of similarity for time series analysis is Dynamic Time Warping (DTW). This distance scales the time series on the time axis to harmonise them and achieve the greatest similarity possible. In this way, it is possible to find patterns between measurements of events with different periods [33]. The disadvantage of DTW when compared to the Euclidean distance is the speed of processing and the advantage is that can deal with temporal drifts. Here we have the formula of this distance [33]:

$$D^2_{DTW}(X,Y) = D^2_{i=1:end}(X_i,Y_i) + min \begin{cases} D^2_{DTW}(X_i, Rest(Y); \\ D^2_{DTW}(Rest(X), Y_i); \\ D^2_{DTW}(Rest(X), Rest(Y)) \end{cases} \tag{3.9}$$

The cell (i,j) corresponds to the adjustment of element x_i with y_i and D corresponds to the distance function.

Finally, the Discrete Fourier transform (DFT) is a classic data reduction method supported by the Discrete Wavelet Transform (DWT) and groups the series using their common forms (similarity in shape). DFT is used to map long time series in the frequency domain to allow the representation of time series by a set of fundamental functions [33].

$$D_{DFT}(a) = \frac{1}{\sqrt{p}} \sum_{i=0}^{p-1} b * e^{\frac{-j2\pi F_i}{p}} \tag{3.10}$$

3.3.1.3 Prototype or Representative of time series clusters

The choice of the representative used for clustering is very relevant for the process's success, particularly in algorithms such as K-Means or Fuzzy C-Means. If the time series are already clustered, the prototype cluster R_j minimizes the distance between all-time series in the cluster

and its prototype. Thus, a time series named R_j that minimizes the distance $E(C_i, R_j)$ is referred to as a Steiner sequence and has the following definition ([36]):

$$E(C_i, R_j) = \frac{1}{n} \sum_{x=1}^{n} dist(F_x, R_j), \tag{3.11}$$

being:

$$C_i = \{F_1, F_2, ...F_n\} \tag{3.12}$$

There are different methods to calculate the representative R_j of a cluster. The most frequently approaches are [5]:

- Use the medoid (centroid, which is part of the database) as a representative;
- Use the average sequence of the cluster as a representative;
- Search the cluster for a representative.

3.3.1.4 Time series clustering algorithms

There are several algorithms for clustering time series, the most commonly used are those shown in Figure 3.8.

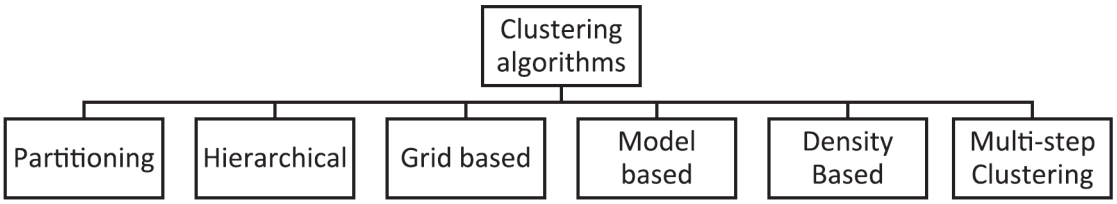


Figure 3.8: Most common algorithms for time series clustering [5].

Partitioning clustering is a method that makes several clusters from k objects, with each group having at least one object. The best-known algorithm of this type of clustering is the *k-means*, where each cluster has its prototype, that is the average value of its objects. Its primary goal is to minimize the objects' total distance to the centre of the cluster [5]. Thus, the *k-means* is applied as follows [37]:

1. Define the value of k (arbitrary number of clusters);
2. Initializing the centers of the k *clusters*;
3. Add each object N to its nearest cluster center;
4. Reestimate the k cluster centers, assuming that point 3 is right;
5. If none of the N objects changes clusters, the algorithm is finished. If not, go back to point 3.

This algorithm has some disadvantages, such as, it is challenging to predict the number of clusters k, and it is hard to compare the quality of the clusters. [38]

Indeed, the *k-means* builds clusters "hard", but there is an algorithm named *Fuzzy C-Means* that builds clusters "softs". This algorithm enables each object in the time series to be part of two or more clusters.

On the other hand, the *Model-based clustering* assumes that the data to be analysed has been generated by a specific model, and tries to return the original model. In this way, it assumes a model for each cluster and finds the time series that have the best fit for the model in hand [5].

The *density-based clustering* approach locates areas of higher object density separated from one another by areas of low density. The best-known algorithm using the density-based model is DBSCAN, and it requires two essential parameters: the *eps* and the *minPoints*. The first parameter says how close the objects should be, so that they can be considered a cluster, in meaning the objects should be located at a distance equal to or less than the *eps*. The second parameter indicates the minimum number of objects for a region to be considered dense [39]. The *density-based clustering* algorithms have some disadvantages, such as high sensitivity to input parameters' characteristics and low cluster descriptors [40].

The *Grid-based clustering* model consists of clustering the space around objects, not the objects themselves. In other words, the space around the objects is covered with a grid to build a spatial summary of the data. Clustering is applied to the cells inside the grid, clustering dense neighbouring cells [41]. This method has a relevant disadvantage: it results in inaccurate cluster despite the technique's fast processing time. [42]

The *Multi-step clustering* consists of some hybrid clustering methods created by a few authors. Zhang's article in [6] presents a new method for *shape-based clustering*. Initially, a network is built based on the similarity between time series, using the triangle distance to measure similarity. The triangle distance is the cosine of the triangle between two vectors so that the distance can take values from -1 to 1. Then, a hierarchical clustering method is applied at some selected points of the network. DTW is also applied to measure the distances between each point in time. Finally, this method reduces the time series' size and improves the efficiency of the *shape-based clustering* method.

Lastly, we have the Hierarchical Clustering Algorithm, which is the focus of this dissertation.

3.3.2 Hierarchical Clustering Algorithm

Hierarchical clustering divides time series into clusters using either the agglomerative method or the divisive method. The agglomerative method starts considering each individual time series as a cluster which are merged gradually with each other, creating an increasingly larger group. On the other hand, the divisive method considers that all time series are a unique cluster which is divided into smaller groups at each step. The limitation of hierarchical clustering is that it is impossible to change the clusters after they have been divided or merged.

In 2009, Hierarchical Clustering Algorithm was the most widely used clustering method in practice[43]. This algorithm's advantages are that it is not mandatory to introduce the number

of clusters as an initial parameter ; presents flexibility regarding the detail level; is adequate for problems connecting point linkages; can be used for outlier detection and has a low computational cost [43][**?**].

In order to implement the Hierarchical Clustering Algorithm, we have to choose aggregation measures. There are seven aggregation measures:

1. *Single linkage*: This method aims to find the elements of different clusters that are closest to one another. The distance between clusters as the minimum distance between these two elements, and it is used to determine which clusters to merge next. This method is easy to implement. [44][45]

$$d(u,v) = min(dist(u[i],v[i]) \tag{3.13}$$

   (All the objects i in cluster u and j in cluster v). [46]

2. *Complete linkage*: It defines the distance between two clusters as the maximum distance between two elements from the cluster. [44][45]

$$d(u,v) = max(dist(u[i],v[i]) \tag{3.14}$$

   (Happens for all the objects i in cluster u and j in cluster v). [46]

3. *Average linkage*: The distance between the two clusters is the average between all existing pairs of points in the two clusters. [44][45]

$$d(c_i,c_j) = \frac{1}{|u| * |v|} \sum_{i,j} d(u[i],v[i] \tag{3.15}$$

   (Happens for all the objects i and j in cluster u and v, where |u| and |v| are the cardinalities of each cluster). [46]

4. *Centroid linkage*: The centroid linkage method consists of taking the distance between the centroids of the data points in clusters.

$$dist(s,t) = ||c_s - c_t||_2 \tag{3.16}$$

   (Happens when C_s and C_t are the centroids of the clusters s and t. The cluster s and t are aggregated into one cluster, and the centroid is computed over the original objects in the previous clusters). [46]

5. *Median linkage*: Attributes the centroid the same way as the previous method. The cluster s and t are aggregated into one cluster and the centroid of the new cluster is the average of the centroids of the cluster s and t [45] [46].

6. *Ward linkage*: Method that uses the Ward variance minimization algorithm, descrived as:

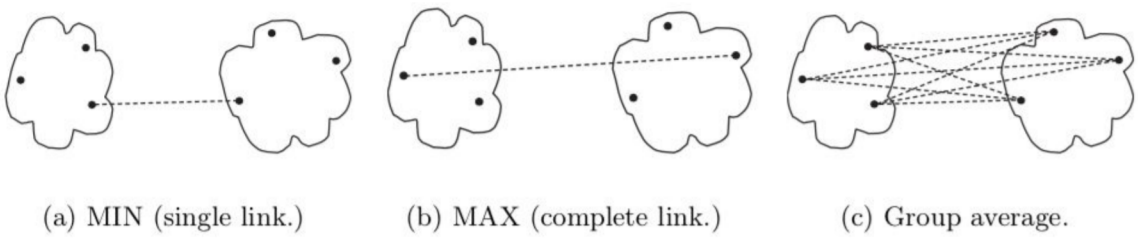
$$d(u,v) = \sqrt{\frac{|v|+|s|}{T}d(v,s)^2 + \frac{|v|+|t|}{T}d(v,t)^2 - \frac{|v|}{T}d(s,t)^2} \tag{3.17}$$

(Cluster s, t and v are the components of the cluster u. Cluster v is a cluster in the same space. $T = |v|+|s|+|t|$ is the cardinality of the cluster u argument) [46].

7. *Weighted*:

$$d(u,v) = \frac{dist(s,v) + dist(t,v)}{2} \tag{3.18}$$

(Cluster s, t and v are the components of the cluster u. Cluster v is a cluster in the same space). [46]



(a) MIN (single link.) (b) MAX (complete link.) (c) Group average.

Figure 3.9: Aggregation Measures: single linkage, complete linkage and average linkage (Adapted from [7])

3.3.2.1 Clustering evaluation measures

In Keogh's article in [47], it is discussed some methods for the evaluation of clustering algorithms results, and some topics are proposed, such as:

- Algorithms must be tested on a wide diversity of data.
- The experiments should be without deviations due to implementation.
- New methods for estimating similarity should be compared with simple methods such as Euclidean distance.
- Where possible, the data and codes used in the experiments should be available free of charge to allow others to duplicate the discoveries.

The accuracy measures for classifying the validity of clusters are classified into two types:

1. External Criteria: used to measure the similarity between the *ground truth* and the clustering procedure done. As the *ground truth* is an ideal clustering, defined by experts [48].

2. Internal Criteria: Compares solutions based on the best fit between each cluster and the data. In this way, it evaluates the clustering using only the characteristics and information present within the database [5].

The most frequently used evaluation methods are the external criteria, so this will be our focus. There are different methods within the external criteria, including purity, cluster similarity measurement (CSM), Jaccard Score, F-measure, and entropy.

The **purity** is a simple and transparent measure to assess the quality of clustering. Considering G as the ground truth and C as the clusters constructed from the algorithm to be evaluated defined as:

$$\mathrm{G} = \{\mathrm{G}_1, \mathrm{G}_2, ...\mathrm{G}_\mathrm{M}\}$$

$$\mathrm{C} = \{\mathrm{C}_1, \mathrm{C}_2, ...\mathrm{C}_\mathrm{M}\}$$

To evaluate the purity of C, each cluster is classified according to its most frequent class. Then it is counted the number of correct objects placed and dividing the number of objects in each cluster. Purity is more easily acquired when we have a larger number of clusters. Purity has a value close to 0 when the clustering is worse and has a value of 1 when it is perfect. [5].

The Cluster Similarity Measure (CSM) is a simple measure defined as [32]:

$$Sim(G,C) = \frac{1}{k}\sum_{i=1}^{k} \max_{1\leq j\leq k} Sim(G_i, C_j) \tag{3.19}$$

$$Sim(G_i, C_j) = \frac{2|G_i \cap C_j|}{|G_i| + |C_j|} \tag{3.20}$$

Another method of clustering evaluation is the Jaccard Score, defined as [49]:

$$Jaccard = \frac{a}{a+b+c} \tag{3.21}$$

Considering $G = \{G_1, G_2, ...G_n\}$ as the dataset clusters, and $A = \{A_1, A_2, ...A_n\}$ as being the clustering algorithms under study. It is a the number of pairs that belong to a G cluster and is grouped in the A cluster. Being b the number of pairs that belong to a G cluster, however, they are not grouped in A. Moreover, c is the number of pairs that belong to a A cluster and is not grouped in the G cluster.

The F-measure is also a widely used evaluation method in the time series domain. This method indicates how similar a cluster is to the ground truth. It can assume the values of 0 or 1 and is defined as the harmonic mean between accuracy and recall. [50]:

$$F = \frac{2 \times P \times R}{P + R} \tag{3.22}$$

where P represents accuracy and R the recall.

Finally, there is entropy. This measure shows how dispersed the classes are in the clusters. The less dispersed, the higher the quality of the clustering performed. [5]

3.4 Methods for the diagnosis of Parkinson's disease using time series

In Joshi's study in [20], the wavelet transform is applied to time series data. These data are concerning the gait of individuals with Parkinson's disease and healthy individuals. The aim is to distinguish these two groups. This transform is used to decompose time-domain functions and facilitate data analysis for patients with and without the disease. It also helps to make it easier to compare the data derived from the patient's left and right leg. The clustering method used was the *Whole Time-Series*. Obtain an accuracy rate of 90.32%.

On the other hand, Wu in [51] analyses the walk variability of patients with Parkinson's disease using the non-parametric Parzen-window method. This method estimates the probability density function (PDF) of the past intervals and their subphases: the balance interval and the posture interval. The clustering method used was Subsequence time-series clustering. It also uses machine learning to determine the properties of gait time series to differentiate Parkinson's patients from healthy patients. The results were evaluated with the Leave One Out cross-validation method, and it was obtained an accuracy rate of 90.32 %.

Pham in [16] uses the fuzzy recurrence plot (FRPs) method to transform one-dimensional time-series signals into two-dimensional objects. In this way, they extract the time-series' characteristic texture from the patients' gait, classifying individuals into diseased and healthy. The clustering method used was the Whole Time-Series. The results were evaluated with the Leave One Out cross-validation method, and an accuracy rate of 77.42% was obtained.

Khorasani's article in [52] Hidden Markov Model (HMM) is used to distinguish Parkinson's patients from healthy ones through their gait data. The clustering method used was the Whole Time-Series. The results were evaluated using the Leave One Out cross-validation method, and an accuracy rate of 90.3% was obtained.

The Zeng study [9] involves using a deterministic learning approach to differentiate Parkinson's patients from a healthy one. A GRF (ground reaction force) sensor system is used and placed on the patients' feet. Thus, the gait characteristics are studied by analysing the vertical GRFs during the steps the patient under study takes. They applied the neural networks radial basis function (RBF) method to classify the patient's condition. The aim is to analyse whether this method will distinguish Parkinson's patients from healthy ones. For the validation of the results obtained, the cross-validation method Leave One Out was used, and an accuracy rate of 93.37% was obtained. The data used in this study comes from the first database used in this dissertation, which contains the gait data of 93 Parkinson's patients.

The article by Linares in [53], unlike all the others, does not use time series derived from physiological signals. Instead, it uses time series based on demographic counts to analyse daily deaths and hospital admissions due to Parkinson's disease in Madrid. These data are then related to heatwave data in the city concerned. Average levels of pressure, humidity and chemical air pollution in the study period are also considered. It is possible to establish a relationship between mortality and hospital admission of Parkinson's patients with the city's heatwaves.

The table 3.1 presents the name of the authors, the method used, the clustering method applied and the respective accuracy.

Table 3.1: Methods for the diagnosis of Parkinson's disease

Authors	Method	Clustering Method	Accuracy Rate
Joshi et al. (2017)	Wavelet Transform	Whole Time-Series	90,32 %
Wu e Krishnan (2010)	Non-parametric Parzen-window	Subsequence time-series clustering	90,32 %
Pham (2018)	Fuzzy Recurrence Plot (FRPs)	Whole Time-Series	77,42%
Khorasani e Daliri (2014)	Hidden Markov Model (HMM)	Whole Time-Series	90,3%
Zeng et al. (2016)	Radial Basis Function (RBF) neural networks	Whole Time-Series	93,37%
Linares et al. (2016)	Poisson Regression	-	-

3.5 Methodology

The datasets chosen to analyse were obtained through the website *Physionet* [4]. The goal is to compare the gait pattern from individuals with neurodegenerative diseases and healthy individuals.

The first dataset named *Gait in Parkinson's Disease Dataset* [54] has a group of individuals with Parkinson and a group of healthy individuals as Control. The demographic data available is the gender, age, weight and height. It also includes measures of disease severity and the average speed of each individual. Hence, it incorporates the vertical ground reaction force records of individuals, with and without Parkinson, as they walked for about 2 minutes on level ground.

The second dataset named *Gait in Neurodegenerative Disease Dataset* has four groups to study: individuals with Parkinson's disease, individuals with Huntington's disease, individuals with Amyotrophic Lateral Sclerosis and healthy individuals. This dataset includes a file with information such as age, gender, height, weight, walking speed, and a measure of disease severity or duration. Includes also a file for each subject with stride time, swing time, swing time percentage, stance time, stance time percentage and double support interval.

As seen before, gait patterns may differ between individuals with neurodegenerative diseases, elderly individuals and healthy young individuals. The gait patterns vary depending on the individual's age, physical condition, taking medication for disease control, if he has some prosthesis, etc. On both datasets, the goal is to distinguish different gait patterns.

The datasets used for this dissertation do not report all the factors that the gait depends on. So, the age and the disease were the only parameters that we could consider to distinguish gait patterns.

To achieve the goal of create homogeneous groups of gait patterns, it was used a clustering method named hierarchical clustering. This method was chosen because it has a low computational cost, it is fast and does not need the number of clusters as input. Figure 3.10 explains briefly the clustering methodology.

The time series used to apply the hierarchical clustering were the swing percentages of each individual's gait.

This dissertation will use the entire time series to do the clustering, so the focus is on Whole time-series clustering.

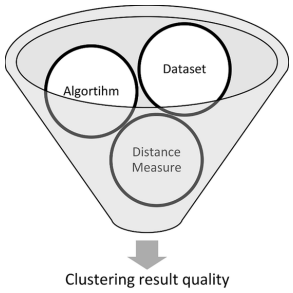


Figure 3.10: Methodology of the clustering (Adapted from [8])

The method used is a multi-step approach, using the Shape-based and Feature-based approaches.

The Whole time series clustering has four components, but some algorithms just need a few. The Hierarchical Clustering algorithm needs the similarity or distance measure. The Hierarchical Clustering strategy chosen was agglomerative clustering ("bottom-up"). And, the number of clusters used was ten, and it was empirically chosen. The Figure 3.11 shows the process of Hierarchical clustering followed in this dissertation.

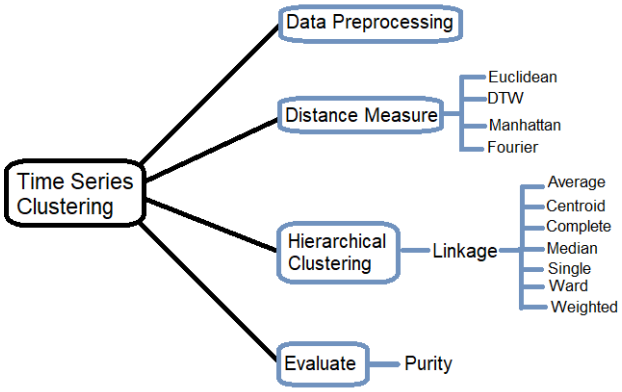


Figure 3.11: Hierarchical Clustering

Four measures of distance were tested, combined with seven measures of aggregation.

The distance measures chosen were Euclidean, Manhattan, DTW and Fourier. The Euclidean was used because it is the most common distance, is fast and efficient, and it is good to compare to the other ones. The Manhattan distance is used because it is similar to the Euclidean, so it can be relevant to compare. Both search for similarity in time.

The DTW distance was chosen because it is the most suitable for measuring similarity for time series analysis. Furthermore, it groups the series by similarity in shape.

The Fourier distance was chosen it is used to map long time series in the frequency domain In this way, it can get different patterns from the other three. This distance groups the series by similarity in shape.

The aggregation measures chosen were: Average, Centroid, Complete, Median, Single, Ward and Weighted.

The Hierarchical clustering was implemented on Python, using the library *Scipy*. The aggregation measures were also done in Python. However, the distances measures chosen were not implemented in the Python library. In this way, it was integrated R in Python. The R's package used is *TSdist*, and it is used for distance measures for Time Series Data.

VBA (on excel) was the software chosen for preprocessing the data before apply the Hierarchical Clustering.

The table 3.2 shows the aggregation measures used (linkage) combined with the distance measure and each distance similarity type.

Table 3.2: List of aggregation and distance measures used in this dissertation

Linkage	Distance	Similarity
Average	Euclidean	In time
Centroid	Euclidean	In time
Complete	Euclidean	In time
Median	Euclidean	In time
Single	Euclidean	In time
Ward	Euclidean	In time
Weighted	Euclidean	In time
Average	Manhattan	In time
Centroid	Manhattan	In time
Complete	Manhattan	In time
Median	Manhattan	In time
Single	Manhattan	In time
Ward	Manhattan	In time
Weighted	Manhattan	In time
Average	DTW	In shape
Centroid	DTW	In shape
Complete	DTW	In shape
Median	DTW	In shape
Single	DTW	In shape
Ward	DTW	In shape
Weighted	DTW	In shape
Average	Fourier	In shape
Centroid	Fourier	In shape
Complete	Fourier	In shape
Median	Fourier	In shape
Single	Fourier	In shape
Ward	Fourier	In shape
Weighted	Fourier	In shape

Due to having more clusters than groups (different gait patterns), the evaluation measure chosen was Purity as it does not imply the number of groups to be equal to the number of clusters. Without recurring to this method, the number of clusters was limited to the number of groups.

Chapter 4

Gait in Parkinson's Disease

In this chapter, the Dataset *Gait in Parkinson's Disease* is analysed. First, it is done a description of the data, and it is explained how it is obtained. The second section refers to the dataset's demographic data, such as age, gender, and body weight. The third section analyses the scales that measure Parkinson's disease's severity and how they are distributed in the dataset's elements. The next section analyses the gait patterns of individuals with Parkinson diseases in different stages and healthy individuals.

4.1 Gait in Parkinson's Disease Dataset

Gait in Parkinson's Disease Dataset [54] is the first dataset in study. It contains data from 93 patients (58 male and 35 female with Parkinson) and 73 healthy patients (40 male and 33 female), used as control. To collect the data, the patient was asked to walk for 2 minutes with 8 force sensors on each foot in order to measure the vertical ground reaction forces (GRF). The output of each of these 16 sensors has been digitized and recorded at 100 samples per second. The location of the sensors is represented in Figure 4.1.

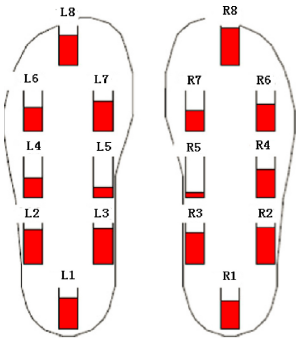


Figure 4.1: Location of sensors on the patient's foot (Adapted from [9])

While the patient walked, the sensors on each foot remained still. The outputs of these sensors were recorded at 100Hz frequency. Thus, each patient's gait data is represented in 19 columns,

with the first column being time, columns 2-17 displaying the data for each sensor for each foot, and the last two columns presenting the total force of the left foot and the right foot.

In order to make the individuals characteristics clear, we change subjects ID to this format:

1. If has Parkinson, the format is pk - age - stage of disease on HY scale; for example, pk-70-2 is an individual with Parkinson disease, with 70 years and stage 2 on HY scale.

2. If is Control, the format is co - age; for example, co-60 is a healthy individual with 60 years.

4.2 Demographic data

The sample is composed mainly of males (60 percent).

By looking at the women's, they are equally distributed across the control groups and the Parkinson's group. Whereas, in men, there are sicker than healthy individuals. (Fig. 4.2).

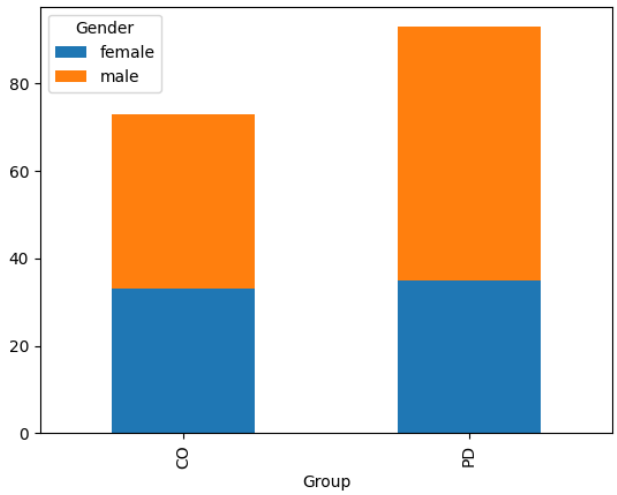


Figure 4.2: Gender distribution (On the left Control; On the right Parkinson)

The maximum and minimum ages are approximately the same in both groups; however, in the case of the Parkinson group, there is a higher concentration of higher age values. The median in the Parkinson's group is 68, and in the healthy group, it is 64. (Fig.4.3).

Comparing the body weight between the two groups, the maximum in the healthy group is 101, slightly lower than in the Parkinson's group, which is 105. Also, the median value is approximately the same. In the case of the control group, there is a more significant dispersion of 50% of the central values (Fig.4.4).

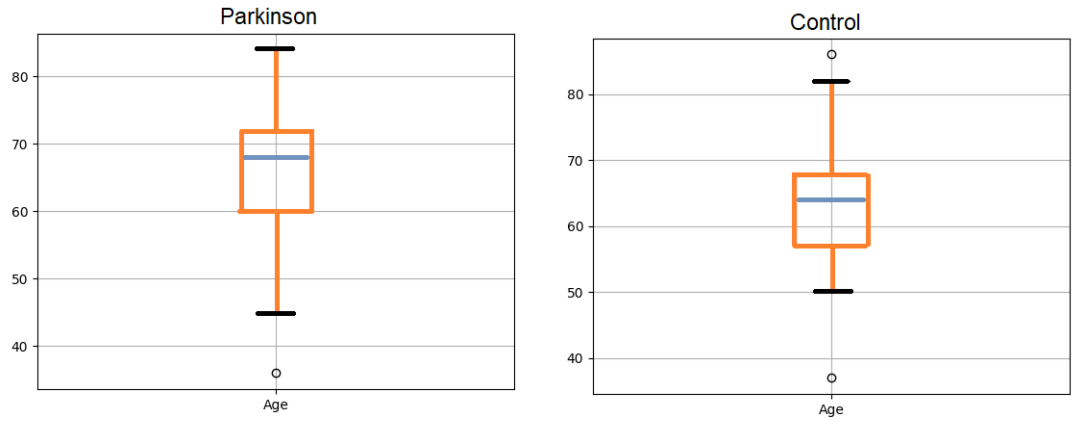


Figure 4.3: Age distribution

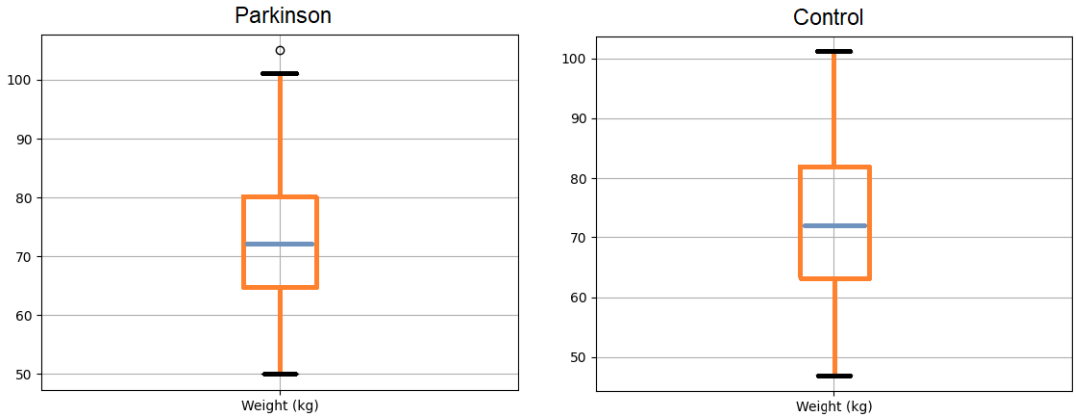


Figure 4.4: Weight distribution

4.3 Scales that measure the severity of Parkinson's disease

In terms of the Hoehn Yahr scale, most Parkinson's patients were found to be at an early stage of the disease. The most advanced stage is stage 3, with a small percentage of cases being the stage 2 the most common (Fig. 4.5).

Looking at the graph of the *Unified Parkinson's Disease Rating Scale* (on the left on Fig. 4.6 there is a concentration on the lower values of the scale, indicating that individuals are in lower states of the disease. The maximum value on the graph is 70, and the median is 30, with the scale varying from 0-199. Concerning the UPDRSm scale on the right on Fig. 4.6, the maximum value is 44. This value is much lower than the previous scale. Indeed, it is also verified that there is a concentration on the lower values of the scale.

The *Timed Up and Go test* shows a concentration in the lower times, meaning that people, in general, take a short time to perform the proposed test. The median is 11.12 minutes (Fig 4.7).

In conclusion, all scales suggest that the large majority of Parkinson's patients who have participated in this study were in the early stages of the disease. This makes sense given that people in

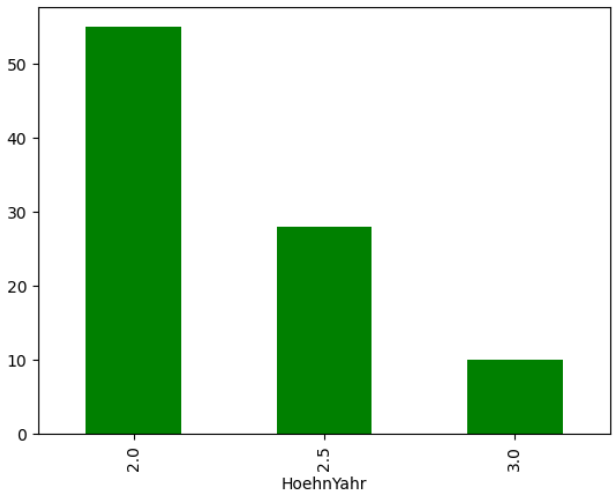


Figure 4.5: Hoehn Yahr Scale

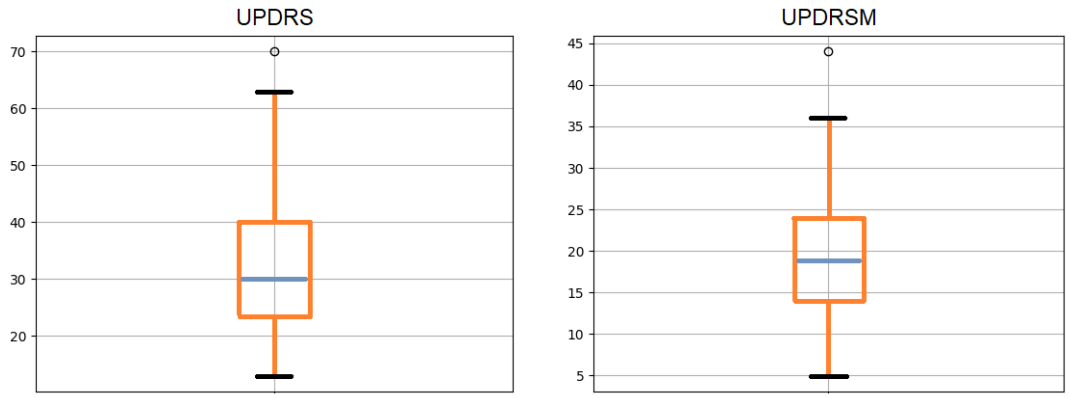


Figure 4.6: UPDRS and UPDRSm

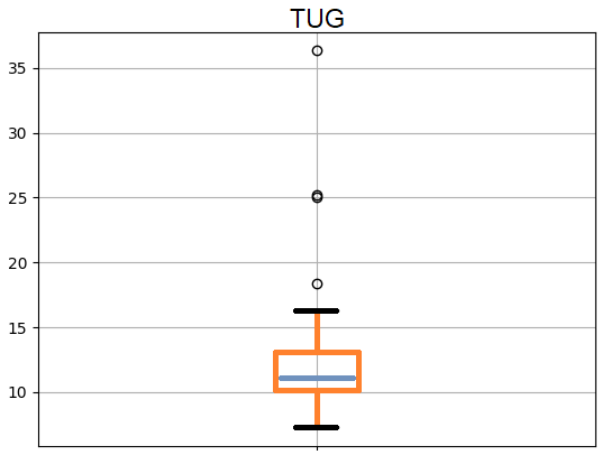


Figure 4.7: Timed Up and Go test

more advanced states may no longer be predisposed or able to perform the kind of tasks required. However, it can make the task of distinguishing between Parkinson and controls a difficult one.

It was also measured the time it takes for a sick individual to take a short walk, as we can see on the left Fig. 4.8 (Speed-01). The median is 1.075, the highest value is 1.423 m/s, and the lowest value is 0.36 m/s.

On the right side of the Fig. 4.8 (Speed-10) it illustrates the average speed at which a patient takes a short walk during subtraction calculations. The minimum time is 0.228m/s, and the maximum time is 1.253m/s. As expected, the speed is lower in this situation.

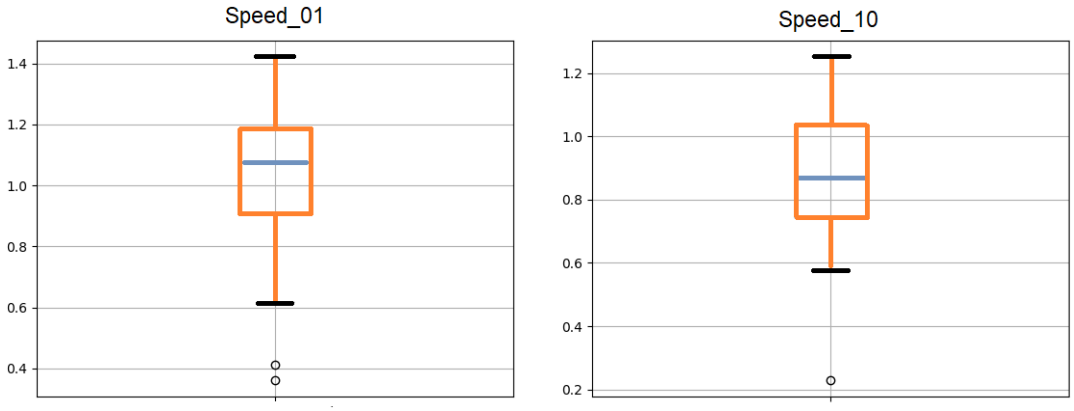


Figure 4.8: Speed-01 and Speed-10

4.4 Gait Data

In order to observe the changes in gait patterns, we made graphics of individuals with Parkinson disease gait patterns. The individuals were divided according to the level of severity of the disease, following the Hoehn Yahr scale. Note that this graphics were made after the preprocessing of the data, explained on the next section.

Thus, the first figure 4.9 presents the control group and the other three figures present individuals with Parkinson with a different stage of the disease.

First, by looking at figure 4.9, we can perceive the characteristic peaks of a typical individual's gait, as explained in chapter 2. The last individual of this figure, represented in red, is 86 years old individual. Contrary to what was supposed, this individual has very defined peaks in at least a part of the sequence.

Figure 4.10 represents individuals in stage 2 of the HY scale. The first individual (blue) has 72 years and has no defined peaks, as expected. The second and third individuals do not have defined peaks either and have 68 and 70 years. Unexpected, the last individual has defined peaks and 82 years.

Figure 4.11 represents individuals in stage 2.5 of HY scale. There is no individual with defined peaks as expected.

Lastly, figure 4.12 represents individuals in stage 3 of the HY scale. The first three individuals have gait patterns as expected. However, the last one has a very irregular gait pattern and some defined peaks. Several reasons justify this: he is taking his medicine against Parkinson's disease, or the disease attacked different muscle that do not interfere with movement.

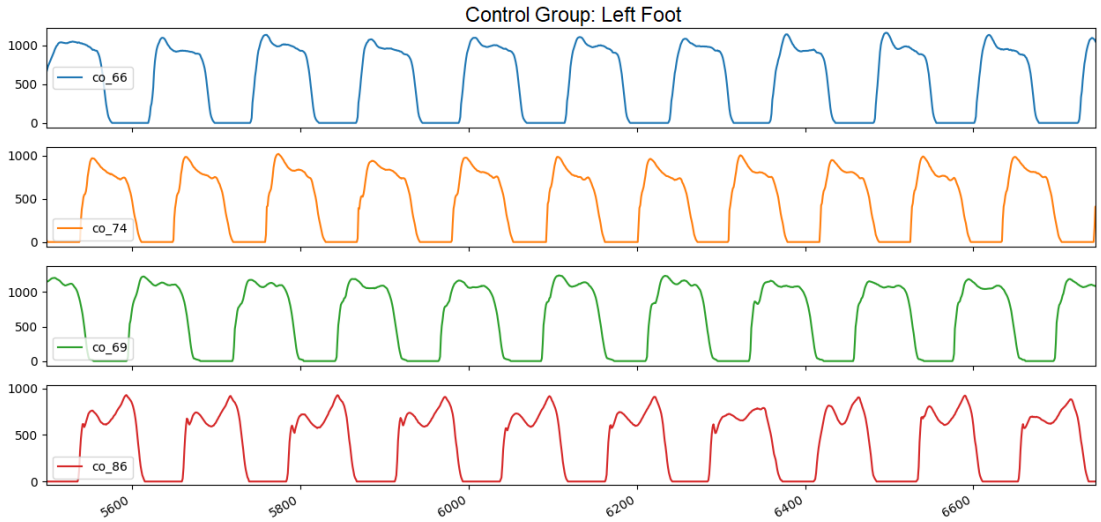


Figure 4.9: Control Group

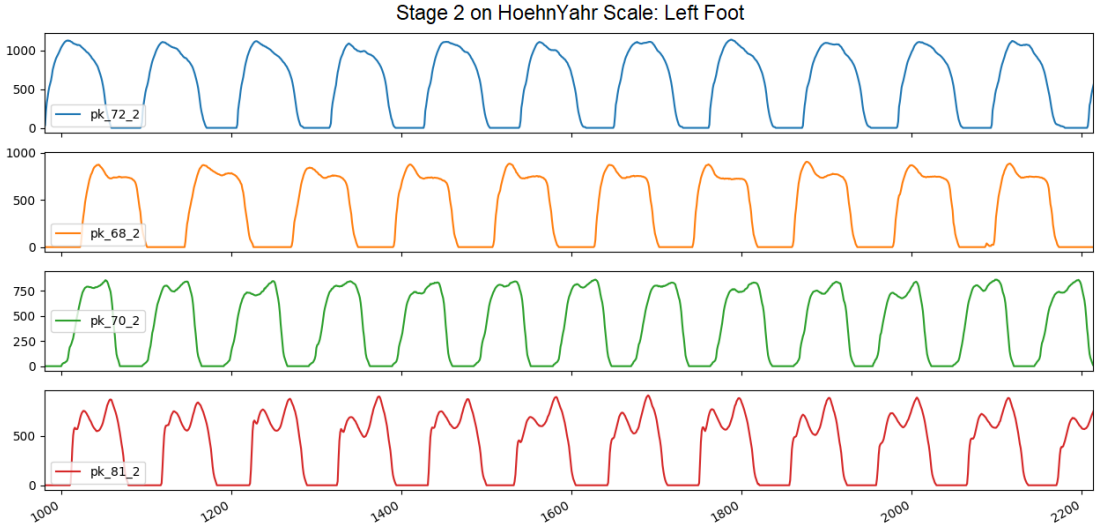


Figure 4.10: Parkinson Group: stage 2 Hoehn Yahr scale

As explained in Chapter 2, the walking patterns of healthy individuals have two peaks. Figure 4.9 present data obtained from individuals from the control group and, except the individual co-69 (third line of the graphic), it is possible to observe the two typical peaks. The individual co-69 has not very pronounced peaks and may be due to several reasons, such as age or some associated physical problem.

On the other hand, the gait patterns of Parkinson's patients have some differences. Figures 4.10, 4.11 and 4.12 suggest that people with Parkinson's do not have such pronounced peaks.

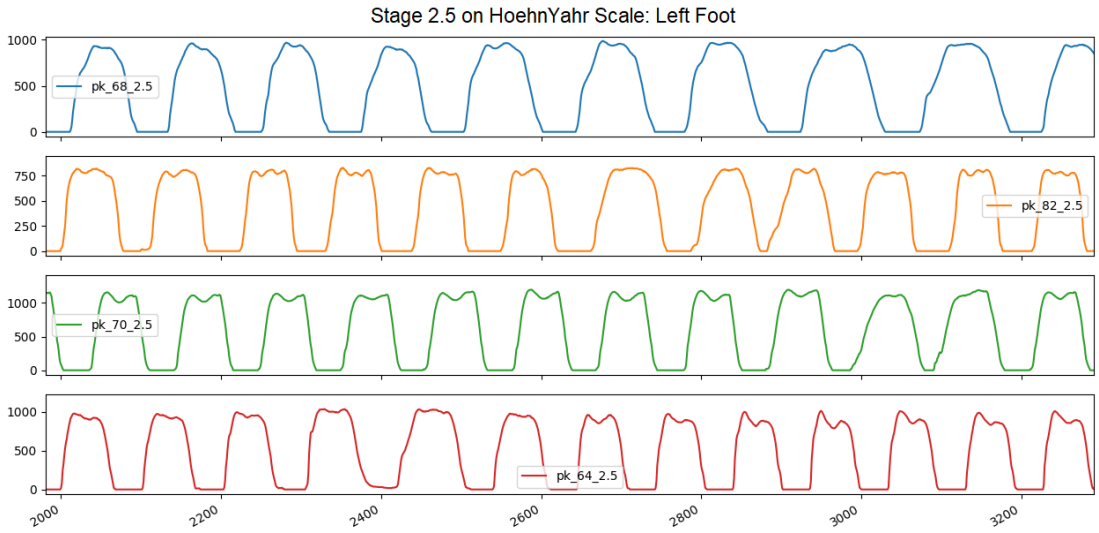


Figure 4.11: Parkinson Group: stage 2.5 Hoehn Yahr scale

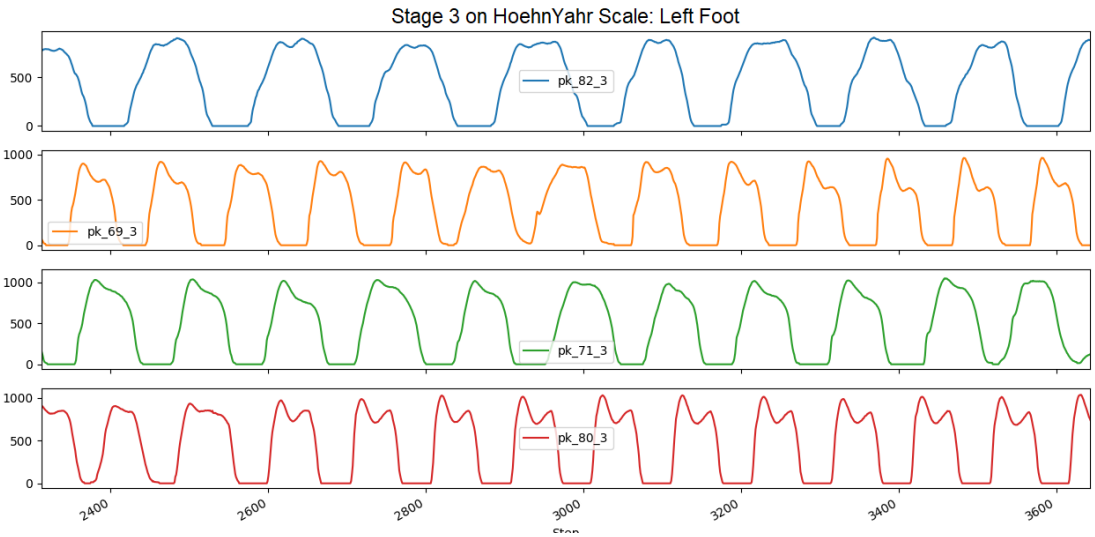


Figure 4.12: Parkinson Group: stage 3 Hoehn Yahr scale

By looking at the graph that exposes an early stage of the disease (stage 2) figure 4.10 we can see that the first individual (pk-72-2) is 72 years old, and has rounded peak, as the second (pk-68-2) and the third individuals (pk-70-2). However, the last individual (pk-81-2) is old and has accentuated peaks. This can be explained for several reasons, such as the part of the walk that was taken for exemplification might not reflect the whole gait. Alternatively, the disease did affect things other than the gait pattern.

Looking at the graph on figure 4.11 we can see that neither individual has peaks well defined. The pattern is irregular as it is shown on the last individual with no defined peaks initially, and then it shows little peaks.

Lastly, in Figure 4.12, the first individual has a very characteristic Parkinson's gait because it has no peaks, and the stride time is high. The last individual is very irregular, and in this part of

the walk, it shows accentuate peaks. However, during the walk, it varies as the first three steps, it has no peaks.

Another variable that is different in walking patterns from ill individuals and healthy ones is the swing percentage. That is the percentage each individual has his foot on the air, while the other foot is in the stride phase. In the graphics it is the time percentage that the sensors value is zero. By looking at the graphics it is noticeable that sick people spend less time with the foot on the air, so they have smaller swing percentage.

In conclusion, it seems that Parkinson individuals act like they are dragging their feet and spend less time with the foot in stride phase. As the stage of the disease increases, the peaks become more rounded. It is also notorious that people's gait is varying through the time they walk. This different can be problems on the sensors, or it can be problems on people's gait.

4.5 Data Preprocessing

Data preprocessing is a data mining technique to convert raw data into cleaner information to be easier analysed.

The data under study were the total pressure of the left foot, the sum of the eight sensors. The first step was to observe the data.

The next step was to determine the best way to calculate the swing percentage using the data that came from 8 force sensors. As explained in Chapter 2, the swing percentage is the percentage of time that the foot is in the air. This percentage is one of the most characteristic between people with Parkinson and healthy people. In order to calculate this, we have to calculate the swing time and stride time first. The swing time is the amount of time that the force sensors are equal to zero. The stride time is the time of a stride.

The algorithm used to calculate these variables was done on Python. The swing percentage was calculated, dividing the swing time of each stride to the stride time. The challenge here was to determine the start and end of each stride. The stride's beginning was calculated by running the force sensor data of each subject and, when the next value is greater than the previous one, being the previous one zero, we save this time. This time is the start of the stride. The stride's end occurs when we find a value equal to zero, and the following is different from zero. So, subtracting the time when the stride begins from the time when it ends, the result is the stride time. The swing time was calculated by the time when the zeros end subtracting the time when the zeros started. Figure 4.13 shows an example of two strides of an individual.

After calculating the swing time, some graphics were done and the data was analysed (Figure 4.14). After observing the result, we noticed a few people with abnormal and unreal data and swing time close to zero (Figure 4.14). We observed the sensors' data and realized that sometimes they failed, and there was a sensor that shot without reason in the middle of a swing time. These values were all derived from only one sensor causing the result to be always smaller than eleven. Due to this, all the pressures under eleven have been replaced by zero.

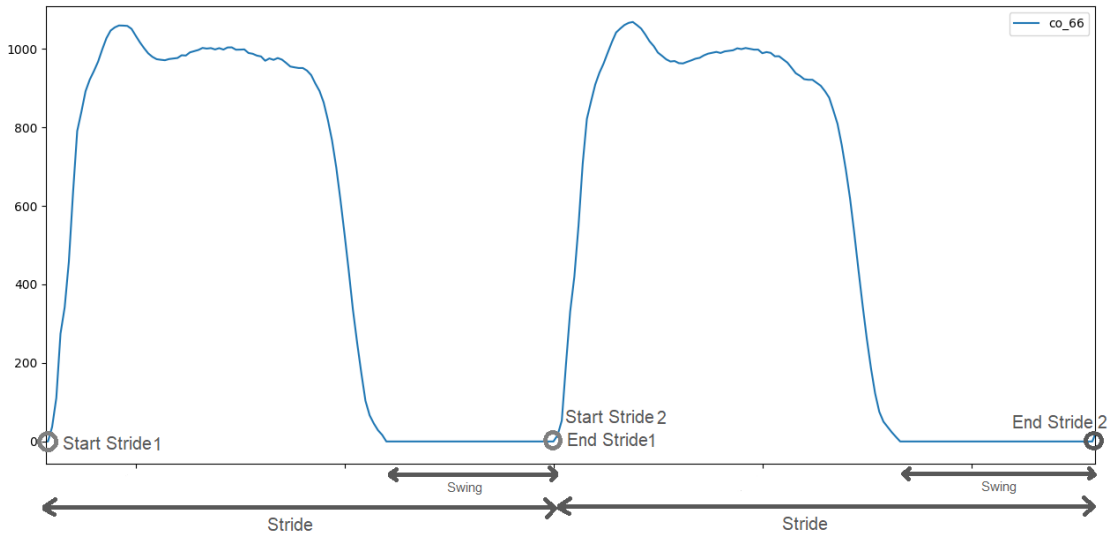


Figure 4.13: Stride and swing time

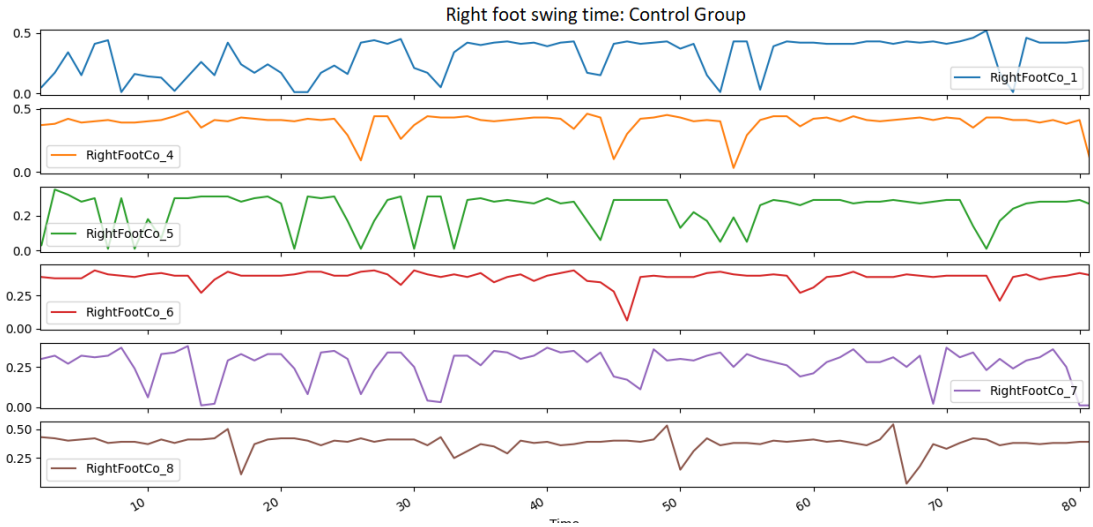


Figure 4.14: Swing Time of the Control Group without preprocessing

After this preprocessing on the force sensors data, the algorithm explained before to determine the swing percentage is applied. At this time, the swing percentage has reliable values. So, the preprocessing ends, and we use this data to initiate machine learning.

Here, we present some graphics that illustrate the swing percentage obtained for each individual.

The figure 4.15 presents results from the swing percentage of the control group. Here we can see that the maximum swing percentage is almost forty and the subjects show a constant swing percentage rounding this value. There are some valleys at values often, when the swing time is shorter. The most irregular subject is the older one, with 86 years, as expected (co-86, colour red).

Figure 4.16 shows individuals with Parkinson in stage 2 of the disease. This figure has some relevant differences from the first one. The swing percentage is much variable. Note that the

younger patient (pk-68-2, colour yellow) is the one with the gait more similar to the control individuals.

Figure 4.17 represents the swing percentage of individuals with Parkinson in stage 2.5 of the disease. The variability is even more significant than the figure 4.16. In this case, the most irregular pattern is not from the older element but from the first element that has 68 years.

The figure 4.18 represents subjects in stage 3 of the disease. The pattern in the swing percentage is very irregular, comparing to the other figures. The most irregular is the first subject and older with 83 years.

It was possible to conclude that the swing percentage oscillates considerably with patients' ageing and the stage of disease increases.

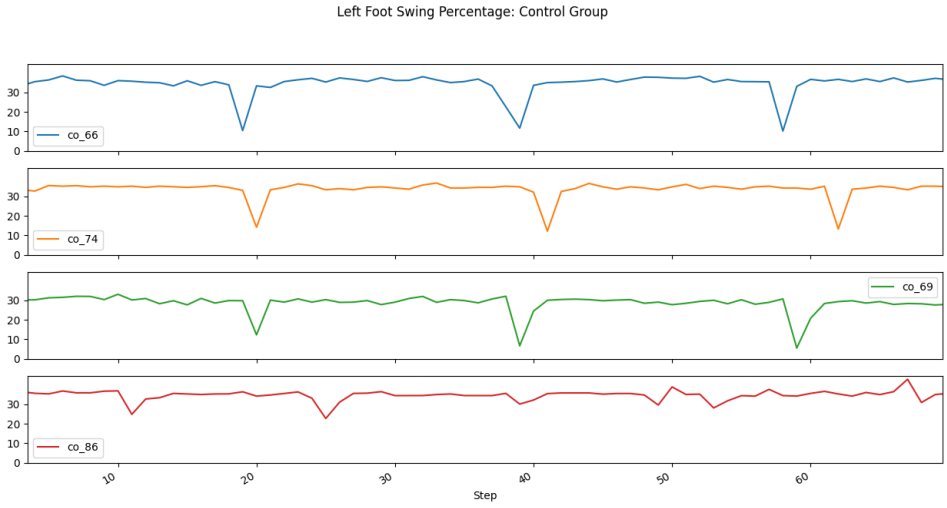


Figure 4.15: Swing Percentage: Control Group

4.6 Classification models

The time series used to apply the clustering algorithm were the swing percentages of each individual's gait.

The Hierarchical Clustering strategy chosen was agglomerative clustering ("bottom-up"). The number of clusters introduced was ten in order to see how the algorithm aggregates each element.

In order to compare the results, it is explored four distance measures: Euclidean, Manhattan, DTW and Fourier distances. The Euclidean, the Manhattan and the Fourier distance find similarity in time. The DTW distance group time series by similariy on shape. Also, it were chosen seven aggregation measures: Average, Centroid, Complete, Median, Single, Ward and Weighted.

To make it simple to see the clustering results, dendrograms of each distance measure were made, combined with each aggregation measure.

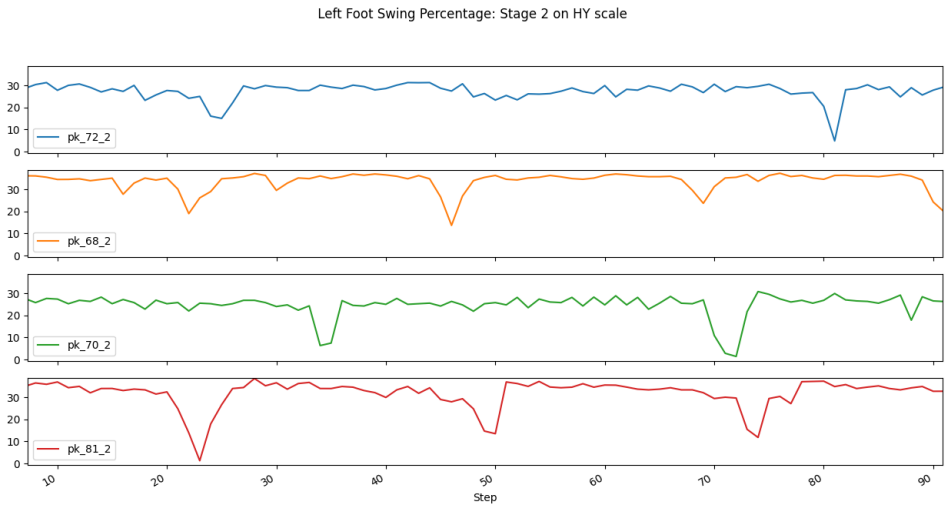


Figure 4.16: Swing Percentage: Parkinson 2 HY

Figure 4.17: Swing Percentage: Parkinson 2.5 HY

The clusters of each aggregation and distance measurements are represented in the tables below.

The gait patterns were divided into three categories. This categories classification was made with the base of the 2 section about the importance of the age in gait patterns. It is proved that a high percentage of people older than 65 years present differences in gait patterns compared to younger ones because they have a lower number of dopamine-producing cells.

- The first one is named good mobility group: includes people with Parkinson with age inferior to 60 years and in an early stage of the disease, and healthy individuals younger than 70.

Figure 4.18: Swing Percentage: Parkinson 3 HY

We can consider that this group having an almost regular quantity of dopamine-producing cells.

- The second category is named median mobility group, and we consider that the elements from this group have median dopamine-producing cells. This includes healthy people with age superior to 60 years and subjects with Parkinson younger than 70 years and in a stage of the disease lower than 3. We can admit that this group having less than usual regular quantity of dopamine-producing cells.

- The third category is named severe difficulty in the walking group, and we admit that the elements have little dopamine-producing cells. It includes subjects with Parkinson in advanced age (older than 60 years) or/and advanced stage of the disease. It also includes healthy people older than 65 years old. We can consider that this group having the lowest quantity of dopamine-producing cells.

Here we can see the dendrograms obtained for Euclidean distance, with different aggregation measures: Average, Centroid, Complete, Median, Single, Ward and Weighted. Each dendrogram is complemented with a table that represents the clusters.

4.6.1 Euclidean distance- Average

By looking at the dendrogram 4.19 and the table 4.20, we can see that 112 of the individuals are together in a big cluster and the others are in smaller clusters. There are four clusters with just one individual each.

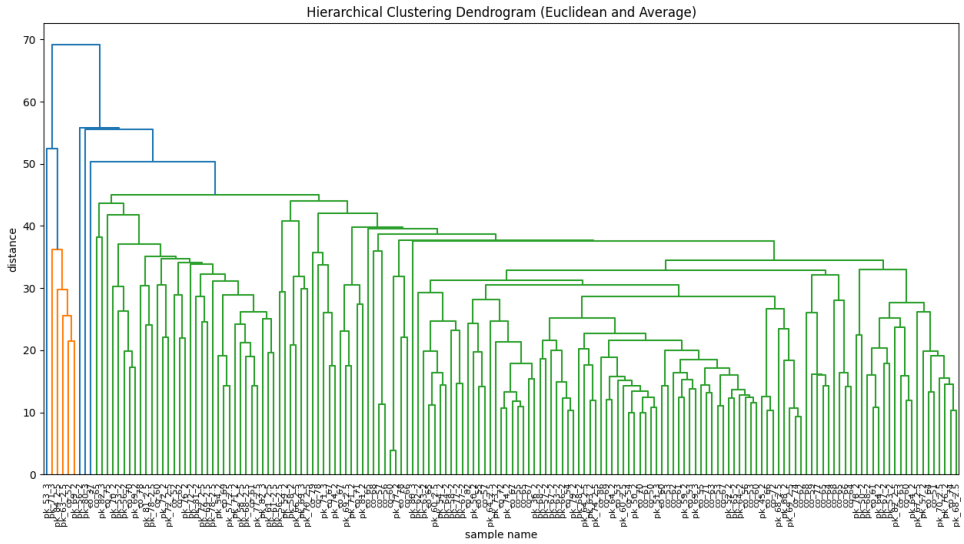


Figure 4.19: Euclidean: Average

Cluster 8 number of entries 112 co_66 co_74 co_86 co_82 co_78 co_69 co_65 co_68 co_67 co_63 co_67 co_71 co_66 co_68 co_55 co_74 co_62 co_77 co_65 co_72 co_72 co_61 co_55 co_53 co_66 co_65 co_64 co_59 co_67 co_60 co_62 co_68 co_64 co_53 co_60 co_53 co_65 co_71 co_55 co_70 co_57 co_65 co_56 co_50 co_60 co_61 co_54 co_64 co_61 co_51 co_56 co_53 co_54 co_57 co_60 co_67 co_53 co_63 pk_82_2.5 pk_68_2 pk_59_2 pk_81_2 pk_79_2 pk_76_2 pk_80_3 pk_78_2 pk_68_2.5 pk_68_2 pk_67_2.5 pk_63_2 pk_77_2 pk_74_2.5 pk_74_2 pk_51_3 pk_69_3 pk_67_2 pk_64_2.5 pk_64_2 pk_64_2.5 pk_68_2 pk_64_2 pk_69_2.5 pk_68_2 pk_68_2.5 pk_45_2 pk_66_2 pk_58_2 pk_69_2.5 pk_73_2 pk_54_2 pk_67_2 pk_63_2 pk_77_2 pk_64_2 pk_52_2 pk_60_2.5 pk_62_2 pk_50_2 pk_64_2 pk_57_2 pk_53_2 pk_50_2 pk_36_2 pk_63_2 pk_50_2 pk_70_2.5 pk_60_2.5 pk_64_2 pk_67_2 pk_71_2 pk_57_2 pk_65_2	Cluster 4 number of entries 1 co_75
Cluster 5 number of entries 2 pk_56_2.5 pk_59_2	Cluster 7 number of entries 5 co_79 co_78 co_67 pk_71_3 pk_74_2
	Cluster 2 number of entries 2 co_65 pk_82_3
	Cluster 1 number of entries 5 co_52 pk_71_3 pk_69_2 pk_63_2.5 pk_84_2.5
	Cluster 6 number of entries 4 pk_69_3 pk_70_2.5 pk_58_2 pk_66_2.5
	Cluster 10 number of entries 1 pk_56_2
	Cluster 9 number of entries 1 pk_80_3
	Cluster 3 number of entries 30 co_69 co_70 co_78 co_60 co_62 co_52 co_61 pk_68_2.5 pk_72_2 pk_70_2 pk_76_2 pk_81_2 pk_81_2.5 pk_78_2.5 pk_72_2 pk_61_2.5 pk_72_2.5 pk_78_2.5 pk_54_2 pk_61_2.5 pk_58_2.5 pk_82_3 pk_57_2.5 pk_77_2.5 pk_71_2 pk_72_2.5 pk_56_2 pk_53_2 pk_69_2 pk_69_2.5

Figure 4.20: Euclidean: Average

4.6.2 Euclidean distance- Centroid

By looking at the dendrogram 4.21 and the table 4.22 of the Euclidean distance using the Centroid method it is seen that the algorithm has trouble separating the individuals, and it agglomerates almost the individuals into one big cluster. There are seven clusters with just one individual each.

4.6.3 Euclidean distance- Single

By looking at the dendrogram 4.23 and the table 4.24 it is seen that the Euclidean distance using the aggregation measure Single results in a big cluster with almost all the elements inside and 9 clusters with one element each.

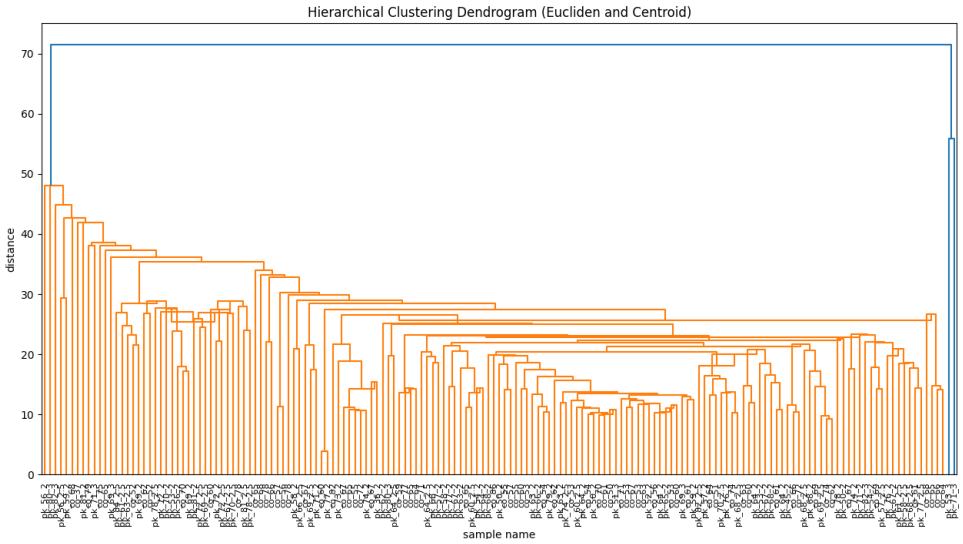


Figure 4.21: Euclidean: Centroid

Cluster 3 number of entries 152	Cluster 4 number of entries 1
co_66 co_74 co_69 co_86 co_82 co_78 co_78 co_69 co_70 co_65	co_75
co_67 co_67 co_63 co_67 co_65 co_78 co_71 co_66 co_68 co_55	Cluster 2 number of entries 2
co_74 co_62 co_77 co_65 co_72 co_72 co_61 co_55 co_53 co_66	co_79 pk_71_3
co_65 co_64 co_59 co_67 co_60 co_62 co_68 co_64 co_53 co_60	Cluster 6 number of entries 1
co_53 co_65 co_71 co_55 co_70 co_57 co_65 co_56 co_50 co_60	co_68
co_61 co_52 co_54 co_60 co_64 co_61 co_51 co_62 co_56 co_52	Cluster 7 number of entries 1
co_53 co_54 co_57 co_60 co_67 co_61 co_53 co_63 pk_68_2.5	pk_82_3
pk_82_2.5 pk_72_2 pk_68_2 pk_69_3 pk_59_2 pk_70_2 pk_79_2	Cluster 9 number of entries 1
pk_76_2 pk_76_2 pk_81_2 pk_80_3 pk_78_2 pk_68_2.5 pk_81_2.5	pk_56_2
pk_78_2.5 pk_72_2 pk_61_2.5 pk_68_2 pk_69_2 pk_67_2.5 pk_63_2	Cluster 5 number of entries 1
pk_63_2.5 pk_77_2 pk_72_2.5 pk_74_2.5 pk_70_2.5 pk_74_2	pk_81_2
pk_78_2.5 pk_51_3 pk_54_2 pk_61_2.5 pk_69_3 pk_74_2 pk_58_2.5	Cluster 10 number of entries 1
pk_67_2 pk_64_2.5 pk_64_2 pk_64_2.5 pk_82_3 pk_68_2 pk_64_2	pk_71_3
pk_84_2.5 pk_69_2.5 pk_68_2 pk_68_2.5 pk_45_2 pk_66_2 pk_58_2	Cluster 8 number of entries 1
pk_57_2.5 pk_69_2.5 pk_77_2.5 pk_73_2 pk_54_2 pk_67_2 pk_63_2	pk_80_3
pk_77_2 pk_64_2 pk_52_2 pk_60_2.5 pk_62_2 pk_50_2 pk_64_2	Cluster 1 number of entries 2
pk_57_2 pk_53_2 pk_58_2 pk_71_2 pk_50_2 pk_36_2 pk_63_2	pk_56_2.5 pk_59_2
pk_50_2 pk_70_2.5 pk_72_2.5 pk_60_2.5 pk_56_2 pk_64_2 pk_67_2	
pk_71_2 pk_57_2 pk_53_2 pk_66_2.5 pk_65_2 pk_69_2 pk_69_2.5	

Figure 4.22: Euclidean: Centroid

4.6.4 Euclidean distance- Complete

Looking at the table 4.26, we can see that the complete method does not concentrate all the individuals into one big cluster, unlikely the first two aggregation measures. In this method, we have two clusters, almost the same size, six clusters with less elements and two clusters with one element each.

Cluster 1 is majority Parkinson. It has only one control individual with 52 years. This subject is the outlier here because he is very young to be aggregated with subjects with Parkinson and older than 60. All the other elements have Parkinson and age superior to 63 years. We consider that there is one element wrong classified in five entries.

Cluster 2 has two elements, both with Parkinson and almost the same age. One is in stage 2

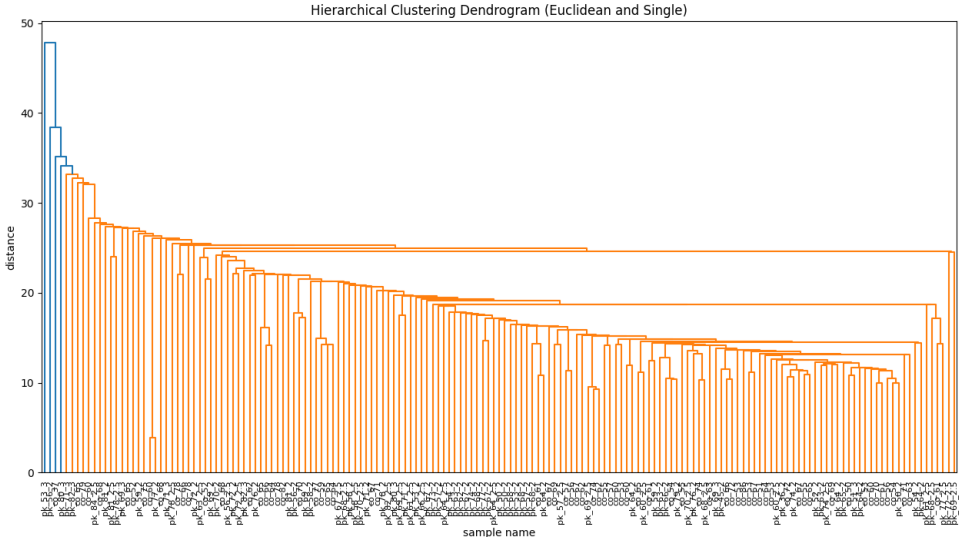


Figure 4.23: Euclidean: Single

Cluster 1 number of entries 156	Cluster 3 number of entries 1
co_66 co_74 co_69 co_86 co_75 co_82 co_78 co_78 co_69 co_70	co_79
co_65 co_67 co_68 co_67 co_63 co_67 co_78 co_71 co_66 co_68	Cluster 4 number of entries 1
co_55 co_74 co_62 co_77 co_65 co_72 co_72 co_61 co_55 co_53	co_65
co_66 co_65 co_64 co_59 co_67 co_60 co_62 co_68 co_64 co_53	Cluster 8 number of entries 1
co_60 co_53 co_65 co_71 co_55 co_70 co_57 co_65 co_56 co_50	co_37
co_60 co_61 co_52 co_54 co_64 co_61 co_51 co_62 co_56 co_52	Cluster 2 number of entries 1
co_53 co_54 co_57 co_60 co_67 co_61 co_53 co_63 pk_68_2.5	co_60
pk_82_2.5 pk_72_2 pk_68_2 pk_69_3 pk_59_2 pk_70_2 pk_81_2	Cluster 5 number of entries 1
pk_79_2 pk_71_3 pk_76_2 pk_76_2 pk_81_2 pk_80_3 pk_78_2	pk_82_3
pk_68_2.5 pk_81_2.5 pk_78_2.5 pk_72_2 pk_61_2.5 pk_68_2 pk_69_2	Cluster 10 number of entries 1
pk_67_2.5 pk_63_2 pk_63_2.5 pk_77_2 pk_72_2.5 pk_74_2.5	pk_53_3
pk_70_2.5 pk_74_2 pk_78_2.5 pk_51_3 pk_54_2 pk_61_2.5 pk_69_3	Cluster 9 number of entries 1
pk_74_2 pk_58_2.5 pk_67_2 pk_64_2.5 pk_64_2 pk_64_2.5 pk_82_3	pk_56_2
pk_68_2 pk_64_2 pk_84_2.5 pk_69_2.5 pk_68_2 pk_68_2.5 pk_45_2	Cluster 6 number of entries 1
pk_66_2 pk_58_2 pk_57_2.5 pk_69_2.5 pk_77_2.5 pk_73_2 pk_56_2.5	pk_71_3
pk_54_2 pk_67_2 pk_63_2 pk_77_2 pk_64_2 pk_52_2 pk_60_2.5	Cluster 7 number of entries 1
pk_62_2 pk_50_2 pk_64_2 pk_57_2 pk_53_2 pk_59_2 pk_58_2	pk_80_3
pk_71_2 pk_50_2 pk_36_2 pk_63_2 pk_50_2 pk_70_2.5 pk_72_2.5	
pk_60_2.5 pk_56_2 pk_64_2 pk_67_2 pk_71_2 pk_57_2 pk_53_2	
pk_66_2.5 pk_65_2 pk_69_2 pk_69_2.5	

Figure 4.24: Euclidean: Single

and the other on stage 2.5 of Parkinson disease. We consider that there is none element wrong classified in two entries.

Cluster 3 has subjects with median difficulty in walking. Healthy individuals with age inferior to 60 and Parkinson individuals with age superior to 70 are wrong classified. The outliers are pk-82-2.5, pk-76-2, pk-70-2.5 and pk-70-2.5. We consider that there are four elements wrong classified in nineteen entries.

Cluster 4 reflects people with severe difficulty in walking and little dopamine-producing cells. The only healthy person is 67 years, so he fits in this group due to his age. We consider that there is none element wrong classified in four entries.

Cluster 5 has the same criteria characteristics as cluster 3, so individual pk-53-2 and co-60 are wrong classified. We consider that there are two elements wrong classified in five entries.

Cluster 6 is the biggest cluster. It represents subjects with median difficulty in walking. The elements that do not belong here are control younger than 60 and Parkinson older than 70. We consider that there are eighteen elements wrong classified in seventy-six entries.

Cluster 7 shows individuals with very difficulty in walking. The youngest element in the control group is 67, and the youngest element in the Parkinson group is 71, so we consider that there is no element wrong classified.

Cluster 9 has five elements and just one control. The control element has 70 years, so the probability of having gait problems is high. In this way, the individual wrong classified is pk-56-2 because he is young and in an early stage of the disease. So we consider that there is one element wrong classified in 5 entries.

Cluster 10 has elements with Parkinson with a high average age. This cluster reflects people with serious gait problems and little dopamine-producing cells. All the control elements with age inferior to 65 do not belong here: co-60, co-60, co-52, co-53, co-61, co-64, co-54, co-51, co-63, co-55, co-65, co-62. All Parkinson elements with age inferior to 60 do not fit here either: pk-50-2, pk-54-2. In this way, we consider that there are fourteen elements wrong classified in forty-two entries.

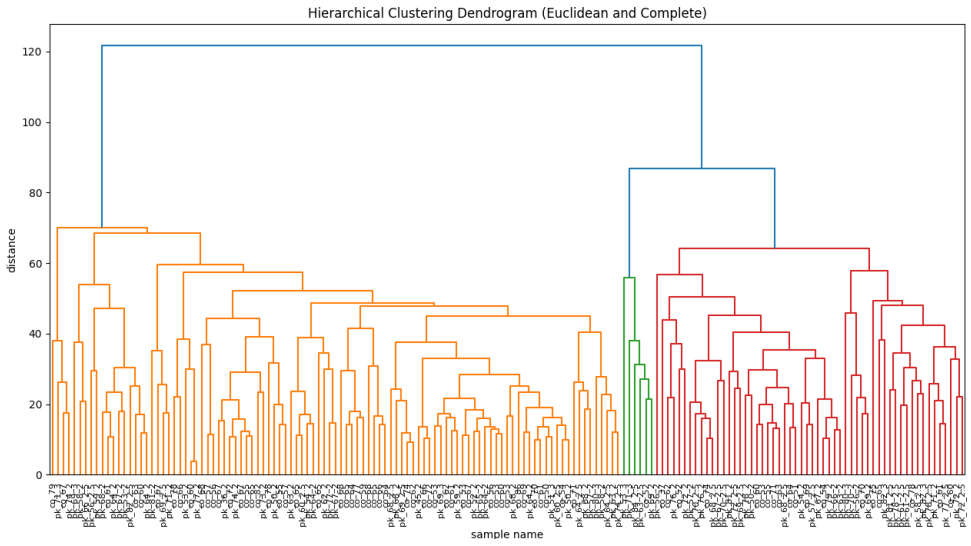


Figure 4.25: Euclidean: Complete

4.6.5 Euclidean distance- Median

By looking at the table 4.28 we can see that the majority elements are aggregated in just one big cluster, so the algorithm does not have an adequate aggregation capacity. There are five clusters with just one individual each. In the big cluster, we consider forty-two elements wrong classified in one hundred and thirty-seven entries.

Cluster 6 number of entries 76 co_66 co_74 co_86 co_82 co_78 co_69 co_65 co_68 co_67 co_67 co_66 co_68 co_55 co_62 co_77 co_65 co_72 co_72 co_61 co_53 co_65 co_64 co_59 co_62 co_68 co_64 co_53 co_65 co_71 co_55 co_70 co_57 co_65 co_56 co_50 co_61 co_56 co_53 co_54 co_57 co_60 co_67 co_53 pk_68_2 pk_59_2 pk_80_3 pk_68_2.5 pk_63_2 pk_74_2.5 pk_74_2 pk_51_3 pk_69_3 pk_64_2.5 pk_64_2.5 pk_68_2 pk_64_2 pk_69_2.5 pk_68_2 pk_45_2 pk_58_2 pk_73_2 pk_54_2 pk_77_2 pk_64_2 pk_52_2 pk_60_2.5 pk_62_2 pk_50_2 pk_50_2 pk_36_2 pk_63_2 pk_60_2.5 pk_64_2 pk_67_2 pk_57_2 pk_65_2	Cluster 4 number of entries 4 co_67 pk_81_2 pk_69_2.5 pk_71_2 Cluster 1 number of entries 5 co_52 pk_71_3 pk_69_2 pk_63_2.5 pk_84_2.5 Cluster 8 number of entries 1 pk_56_2 Cluster 2 number of entries 2 pk_56_2.5 pk_59_2 Cluster 9 number of entries 5 co_70 pk_70_2 pk_80_3 pk_56_2 pk_69_2
Cluster 10 number of entries 42 co_69 co_75 co_65 co_78 co_71 co_55 co_60 co_53 co_54 co_60 co_64 co_51 co_62 co_52 co_61 co_63 pk_82_3 pk_68_2.5 pk_72_2 pk_79_2 pk_76_2 pk_81_2 pk_78_2 pk_81_2.5 pk_78_2.5 pk_72_2 pk_61_2.5 pk_72_2.5 pk_78_2.5 pk_54_2 pk_61_2.5 pk_58_2.5 pk_67_2 pk_82_3 pk_66_2 pk_57_2.5 pk_77_2.5 pk_63_2 pk_71_2 pk_50_2 pk_72_2.5 pk_69_2.5	Cluster 7 number of entries 4 co_79 co_67 pk_71_3 pk_74_2 Cluster 5 number of entries 5 co_78 co_66 co_60 pk_77_2 pk_53_2 Cluster 3 number of entries 19 co_63 co_74 co_60 co_61 pk_82_2.5 pk_69_3 pk_76_2 pk_68_2 pk_67_2.5 pk_70_2.5 pk_64_2 pk_68_2.5 pk_67_2 pk_64_2 pk_57_2 pk_53_2 pk_58_2 pk_70_2.5 pk_66_2.5

Figure 4.26: Euclidean: Complete

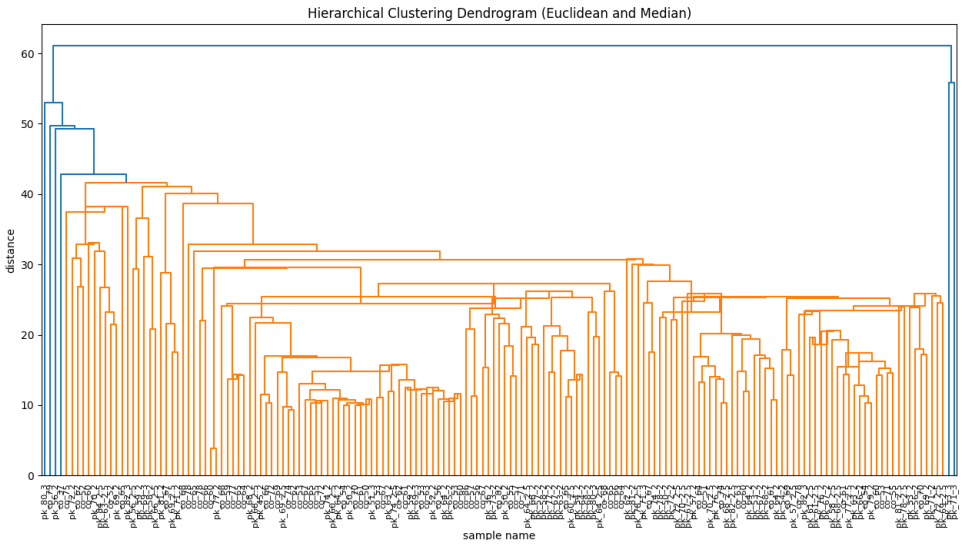


Figure 4.27: Euclidean: Median

4.6.6 Euclidean distance- Ward

Cluster number 1 was characterized by people with a movement disorder. So, the only individual that does not fit is the co-52 because it is young, and the probability of having movement disorder is low.

Cluster 2 also characterize people with movement disorders. This problem can be for many reasons, but we can only considerate the ones provided: the stage of Parkinson and the age. So, we exclude everyone control with less than 70 because the probability of having trouble walking because of the age is low. Moreover, we exclude everyone with Parkinson but is too young and is an early stage of the disease to have gait problems. So, the individuals that does not fit this cluster are co-67 and pk-56-2.

Cluster 3 is a group with a majority of individuals with Parkinson. The only Control element

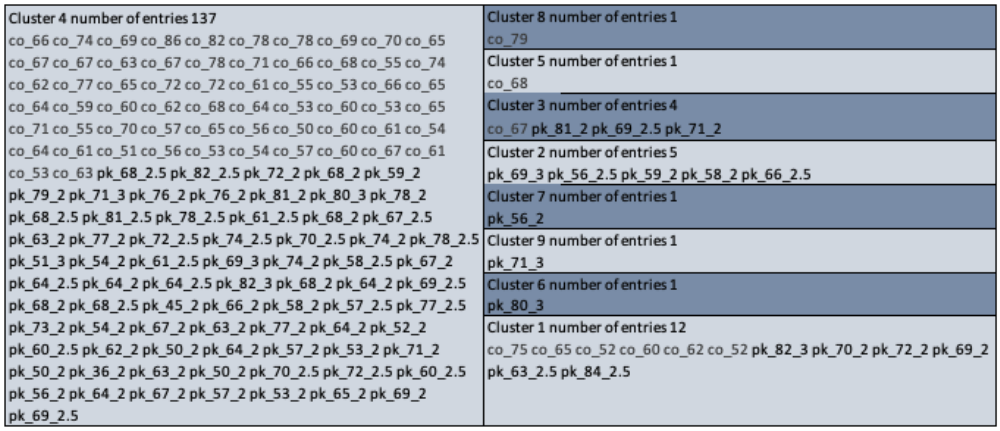


Figure 4.28: Euclidean: Median

as 70 year old so it can be considered part of this cluster. The element that does not fit this cluster is pk-56-2 because is young and is in an early stage of the disease.

Cluster 4 has the majority of individuals with Parkinson disease, so it is a group characterized by movement disorders. Everyone that belongs to the Control group and has age lower than 70 years must be considered outliers. This is because it is not probable having gait problems. People young and in an early stage of Parkinson Disease should be considered outliers too. The individuals wrong classification are: pk-54-2, co-65, co-52,co-62, co-60 and co-61.

Cluster 5 has individuals with Parkinson disease in an early stage of the disease, and there is none with more than 70 years. So, we consider that this cluster is well done.

Cluster 6 should have individuals control with high age and Parkinson in an early stage of the disease and not older than 70. In this way, we consider that the outliers are co-51, pk-82-2.5 and pk-76-2.

Cluster 7 is characterized by individuals with median mobility and less than usual dopamine-producing cells. Individuals Control who are too young probably do not have gait problems, so they do not belong in this cluster. Individuals with a high stage of Parkinson disease and with an old age do not belong here. We consider that the outliers are pk-74-2, pk-73-2, co-50 and co-53.

Cluster 8 is characterized by individuals with median mobility, as the cluster before. We consider that the outliers are the individual control too young, and the Parkinson is too old or in a high disease stage. In this way, the individuals that are wrong allocated are: pk-57-2, pk-58-2 and co-54.

Cluster 9 should have individuals with Parkinson disease without ages too high. The control individuals have 65 years, so they have less than usual dopamine-producing cells and consequently, some gait problems. We consider that Pk-77-2 does not fit in this cluster.

Lastly, Cluster 10 has a majority of Control individuals. The Control individuals that are too young do not fit here. The Parkinson too old does not fit here either. We consider that the outliers are: co-53, co-53, pk-78-2, co-53, co-53 and pk-77-2.

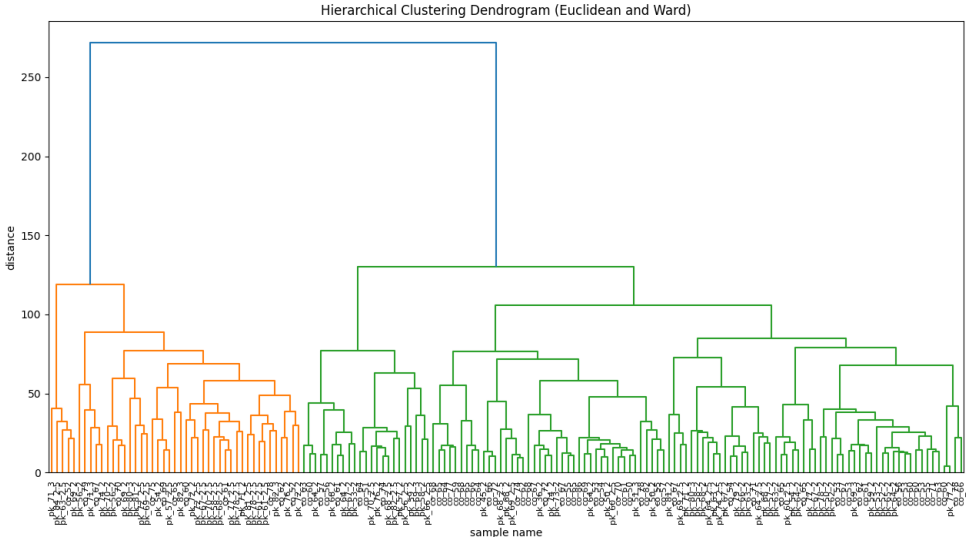


Figure 4.29: Euclidean: Ward

Cluster 7 number of entries 42 co_66 co_74 co_86 co_82 co_78 co_69 co_65 co_68 co_67 co_66 co_66 co_68 co_55 co_77 co_65 co_72 co_72 co_64 co_59 co_62 co_68 co_64 co_65 co_55 co_70 co_50 co_61 co_54 co_57 co_53 **pk_68_2.5 pk_74_2 pk_51_3 pk_68_2 pk_64_2 pk_69_2.5 pk_45_2** **pk_73_2 pk_60_2.5 pk_50_2 pk_50_2 pk_36_2**	**Cluster 5 number of entries 10** co_63 co_57 co_56 co_60 co_61 **pk_68_2 pk_64_2 pk_67_2 pk_64_2** **pk_53_2**
	Cluster 6 number of entries 13 co_74 co_64 co_51 pk_82_2.5 pk_69_3 pk_76_2 pk_68_2.5 pk_56_2.5 pk_57_2 pk_59_2 pk_58_2 pk_70_2.5 pk_66_2.5
Cluster 4 number of entries 27 co_69 co_75 co_65 co_78 co_60 co_62 co_52 co_61 **pk_82_3** **pk_68_2.5 pk_72_2 pk_76_2 pk_81_2.5 pk_78_2.5 pk_72_2 pk_61_2.5** **pk_67_2.5 pk_72_2.5 pk_70_2.5 pk_78_2.5 pk_54_2 pk_61_2.5** **pk_58_2.5 pk_82_3 pk_57_2.5 pk_77_2.5 pk_71_2**	**Cluster 9 number of entries 8** co_65 co_65 **pk_54_2 pk_77_2 pk_64_2 pk_63_2 pk_60_2.5 pk_67_2**
	Cluster 8 number of entries 19 co_67 co_71 co_54 **pk_68_2 pk_81_2 pk_79_2 pk_80_3 pk_63_2 pk_74_2.5** **pk_67_2 pk_64_2.5 pk_64_2.5 pk_68_2 pk_66_2 pk_58_2 pk_69_2.5** **pk_63_2 pk_71_2 pk_57_2**
Cluster 10 number of entries 26 co_78 co_71 co_62 co_61 co_55 co_53 co_66 co_60 co_53 co_60 co_53 co_56 co_53 co_60 co_67 co_63 **pk_59_2 pk_78_2 pk_77_2** **pk_69_3 pk_52_2 pk_62_2 pk_50_2 pk_64_2 pk_53_2 pk_65_2**	**Cluster 1 number of entries 5** co_52 **pk_71_3 pk_69_2 pk_63_2.5 pk_84_2.5**
Cluster 3 number of entries 8 co_70 **pk_70_2 pk_81_2 pk_80_3 pk_72_2.5 pk_56_2 pk_69_2** **pk_69_2.5**	**Cluster 4 number of entries 1** co_68

Figure 4.30: Euclidean: Ward

4.6.7 Euclidean distance- Weighted

Cluster 1 is majority Parkinson, and it has only one control individual with 52 years. This cluster is very similar to Cluster 1 from the previous aggregation measure. We consider that co-52 is wrong classified and do not belong in this cluster because of his age.

Cluster 2 has subjects with median difficulty in walking. We consider that healthy individuals with age inferior to 60 and Parkinson individuals with age superior to 70 or in stage 3 of the disease are wrong classified. In this way, this elements are wrong classified: co-51, pk-76-2, pk-69-3, pk-82-2,5 and pk-70-2,5. We admit that there are five elements wrong classified in fifteen entries.

Cluster 3 is very random, and it was supposed to have subjects with median difficulty in walking. We consider that healthy individuals with age inferior to 60 and Parkinson individuals

with age superior to 70 are wrong classified. In this way, there are twenty-nine elements wrong classified in one hundred and ten entries.

Cluster 5 shows individuals with very difficulty in walking. The control group's younger element has 67 years, and in the Parkinson's group has 71 years. We consider that the cluster has none element wrong classified.

Cluster 7 has majority elements with Parkinson. The Parkinson group's average age is high (71 years), so it translates a group with serious gait problems. Control elements younger than 65 do not fit in this group, such as co-60, co-62 and co-52. Parkinson elements younger than 60 do not fit here either, such as pk-53-2 and pk-56-2. We consider that there are five elements wrong classified in twenty-two entries.

Cluster 4, Cluster 6, Cluster 8 and Cluster 10 only have one element each.

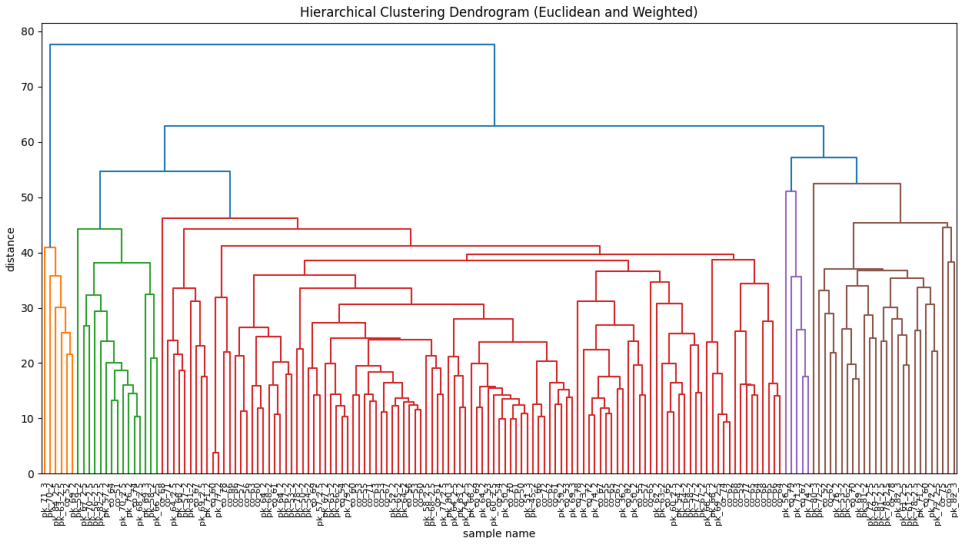


Figure 4.31: Euclidean: Weighted

Cluster 3 number of entries 110 co_66 co_74 co_69 co_86 co_82 co_78 co_78 co_69 co_65 co_67 co_63 co_67 co_71 co_66 co_68 co_55 co_62 co_77 co_65 co_72 co_72 co_61 co_55 co_53 co_66 co_65 co_64 co_59 co_67 co_60 co_62 co_68 co_64 co_53 co_60 co_53 co_65 co_71 co_55 co_70 co_57 co_65 co_56 co_50 co_60 co_61 co_54 co_61 co_56 co_53 co_54 co_57 co_60 co_67 co_61 co_53 co_63 pk_68_2.5 pk_68_2 pk_59_2 pk_81_2 pk_79_2 pk_80_3 pk_78_2 pk_68_2.5 pk_68_2 pk_63_2 pk_77_2 pk_74_2.5 pk_74_2 pk_51_3 pk_54_2 pk_69_3 pk_58_2.5 pk_67_2 pk_64_2.5 pk_64_2 pk_64_2.5 pk_68_2 pk_64_2 pk_69_2.5 pk_68_2 pk_45_2 pk_66_2 pk_58_2 pk_57_2.5 pk_69_2.5 pk_77_2.5 pk_73_2 pk_54_2 pk_67_2 pk_63_2 pk_77_2 pk_64_2 pk_52_2 pk_60_2.5 pk_62_2 pk_50_2 pk_64_2 pk_53_2 pk_50_2 pk_36_2 pk_63_2 pk_50_2 pk_60_2.5 pk_64_2 pk_67_2 pk_71_2 pk_57_2 pk_65_2	Cluster 8 number of entries 2 co_65 pk_82_3 Cluster 1 number of entries 6 co_52 pk_70_2 pk_71_3 pk_69_2 pk_63_2.5 pk_84_2.5 Cluster 6 number of entries 1 pk_56_2 Cluster 10 number of entries 1 pk_80_3 Cluster 2 number of entries 15 co_74 co_64 co_51 pk_82_2.5 pk_69_3 pk_76_2 pk_67_2.5 pk_70_2.5 pk_68_2.5 pk_56_2.5 pk_57_2 pk_59_2 pk_58_2 pk_70_2.5 pk_66_2.5 Cluster 7 number of entries 22 co_70 co_78 co_60 co_62 co_52 pk_72_2 pk_76_2 pk_81_2 pk_81_2.5 pk_78_2.5 pk_72_2 pk_61_2.5 pk_72_2.5 pk_78_2.5 pk_61_2.5 pk_82_3 pk_71_2 pk_72_2.5 pk_56_2 pk_53_2 pk_69_2 pk_69_2.5
Cluster 9 number of entries 1 co_75	Cluster 4 number of entries 1 co_68
Cluster 5 number of entries 4 co_79 co_67 pk_71_3 pk_74_2	

Figure 4.32: Euclidean: Weighted

Below we can see the dendrograms obtained for the DTW distance, with different measures of aggregation.

4.6.8 DTW distance- Average

By looking at the table 4.34 and dendogram A.3, we can see that the data is separated in almost just two clusters: the second and the fifth. The cluster 2 represents people with median mobility problems, so healthy people younger than 60 years and Parkinson subjects older than 70 years and stage 3 of the disease are wrong classified. The cluster 5 represents people with low mobility, so healthy people younger than 65 and Parkinson subjects older younger than 60 years are wrong classified. In this way, there are 17 elements wrong classified in cluster 2 and 16 elements wrong classified in cluster 5. In cluster 1, 4 and 9 we have three, three and one elements wrong classified, respectively.

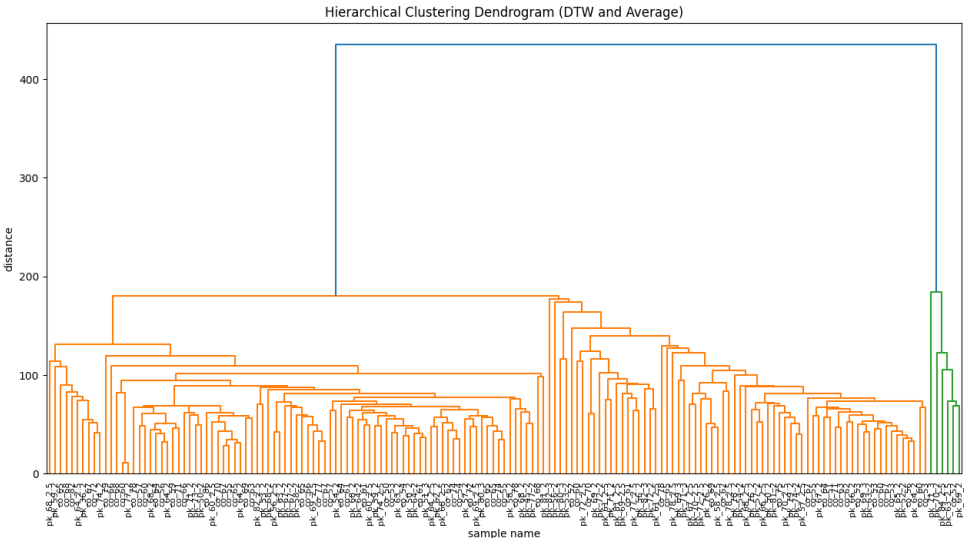


Figure 4.33: DTW: Average

4.6.9 DTW distance- Centroid

By looking at the table 4.36 we can see that this method has lots of clusters with just with one element. The bigger cluster is the cluster four and aggregates almost every element. It represents individuals with median mobility and has 28 elements wrong classified.

4.6.10 DTW distance- Complete

Cluster 3 has subjects with median difficulty in walking. Healthy individuals with age inferior to 60 and Parkinson individuals with age superior to 70 are wrong classified. In this way, we consider that the outliers are: co-57, co-56, pk-73-2, pk-74-2 and pk-71-2. Five wrong in 21.

Cluster 2 number of entries 79 co_66 co_74 co_86 co_79 co_78 co_78 co_69 co_65 co_67 co_63 co_66 co_68 co_55 co_74 co_77 co_65 co_72 co_61 co_66 co_64 co_59 co_67 co_62 co_68 co_64 co_60 co_65 co_71 co_55 co_70 co_57 co_65 co_56 co_50 co_60 co_61 co_54 co_61 co_54 co_57 co_53 pk_82_2.5 pk_68_2 pk_59_2 pk_81_2 pk_80_3 pk_68_2 pk_63_2 pk_77_2 pk_74_2.5 pk_51_3 pk_64_2.5 pk_64_2 pk_68_2 pk_64_2 pk_69_2.5 pk_68_2 pk_68_2.5 pk_45_2 pk_58_2 pk_69_2.5 pk_73_2 pk_56_2.5 pk_54_2 pk_67_2 pk_77_2 pk_64_2 pk_60_2.5 pk_62_2 pk_50_2 pk_64_2 pk_53_2 pk_58_2 pk_50_2 pk_63_2 pk_60_2.5 pk_67_2 pk_71_2 pk_57_2	Cluster 5 number of entries 48 co_69 co_75 co_67 co_65 co_78 co_71 co_62 co_55 co_53 co_60 co_53 co_53 co_64 co_51 co_62 co_56 co_53 co_60 co_67 co_63 pk_68_2.5 pk_69_3 pk_79_2 pk_71_3 pk_76_2 pk_76_2 pk_81_2 pk_78_2 pk_78_2.5 pk_67_2.5 pk_70_2.5 pk_78_2.5 pk_54_2 pk_69_3 pk_74_2 pk_58_2.5 pk_67_2 pk_66_2 pk_57_2.5 pk_63_2 pk_52_2 pk_57_2 pk_50_2 pk_70_2.5 pk_72_2.5 pk_64_2 pk_66_2.5 pk_65_2
	Cluster 9 number of entries 5 co_52 pk_70_2 pk_69_2 pk_63_2.5 pk_84_2.5
	Cluster 6 number of entries 1 co_52
Cluster 3 number of entries 2 pk_80_3 pk_53_2	Cluster 8 number of entries 1 pk_82_3
Cluster 4 number of entries 15 co_70 co_60 co_61 pk_72_2 pk_81_2.5 pk_72_2 pk_61_2.5 pk_72_2.5 pk_61_2.5 pk_82_3 pk_77_2.5 pk_71_2 pk_56_2 pk_69_2 pk_69_2.5	Cluster 7 number of entries 1 pk_56_2
	Cluster 10 number of entries 1 pk_71_3
Cluster 1 number of entries 10 co_82 co_68 co_67 co_72 co_65 pk_68_2.5 pk_74_2 pk_64_2.5 pk_59_2 pk_36_2	Cluster 4 number of entries 15 co_70 co_60 co_61 pk_72_2 pk_81_2.5 pk_72_2 pk_61_2.5 pk_72_2.5 pk_61_2.5 pk_82_3 pk_77_2.5 pk_71_2 pk_56_2 pk_69_2 pk_69_2.5

Figure 4.34: DTW: Average

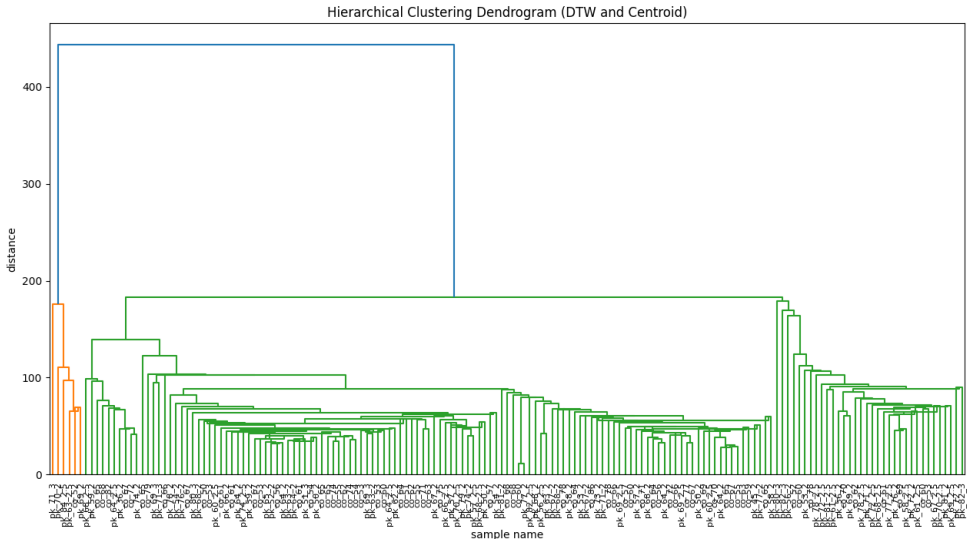


Figure 4.35: Distância DTW - Centroid

Note that individuals pk-73-2 pk-74-2 pk-71-2 are also wrong classified in cluster 1 from DTW distance- Ward.

Cluster 4 has subjects with median difficulty in walking as Cluster 3. Moreover, we consider that the individuals control co-55, co-59, co-55 , co-50, co-54, co-57 and co-53 do not belong here. The Parkinson subjects older than 70 do not belong here either pk-82-2,5, pk-77-2, pk-81-2 and pk-80-3. Note that the individuals pk-82-2,5, pk-77-2, pk-81-2 and pk-80-3 are the same as the wrong classified in Cluster 3 from DTW distance- Ward.

Cluster 5 has individuals with severe movement disorders. The average age of Parkinson's subjects is high. We consider that control individuals with age inferior to 65 do not belong in this cluster – co-60 and co-62. Pk-56-2 is not consider well classified because he is too young to be

Cluster 4 number of entries 112 co_66 co_74 co_86 co_75 co_79 co_78 co_78 co_69 co_65 co_67 co_67 co_63 co_71 co_66 co_68 co_55 co_74 co_62 co_77 co_65 co_72 co_61 co_55 co_53 co_66 co_64 co_59 co_67 co_62 co_68 co_64 co_60 co_53 co_65 co_71 co_55 co_70 co_57 co_65 co_56 co_50 co_60 co_61 co_54 co_64 co_61 co_51 co_56 co_53 co_54 co_57 co_60 co_67 co_53 co_63 pk_82_2.5 pk_68_2 pk_69_3 pk_59_2 pk_81_2 pk_79_2 pk_71_3 pk_76_2 pk_80_3 pk_78_2 pk_68_2 pk_63_2 pk_77_2 pk_74_2.5 pk_51_3 pk_54_2 pk_69_3 pk_74_2 pk_67_2 pk_64_2.5 pk_64_2 pk_68_2 pk_64_2 pk_69_2.5 pk_68_2 pk_68_2.5 pk_45_2 pk_66_2 pk_58_2 pk_57_2.5 pk_69_2.5 pk_73_2 pk_56_2.5 pk_54_2 pk_67_2 pk_63_2 pk_77_2 pk_64_2 pk_52_2 pk_60_2.5 pk_62_2 pk_50_2 pk_64_2 pk_57_2 pk_53_2 pk_58_2 pk_50_2 pk_63_2 pk_50_2 pk_70_2.5 pk_60_2.5 pk_64_2 pk_67_2 pk_71_2 pk_57_2 pk_66_2.5 pk_65_2	Cluster 7 number of entries 1 co_52 Cluster 9 number of entries 1 pk_82_3 Cluster 8 number of entries 1 pk_56_2 Cluster 2 number of entries 1 pk_71_3 Cluster 10 number of entries 1 pk_80_3 Cluster 5 number of entries 1 co_65 Cluster 1 number of entries 5 co_52 pk_70_2 pk_69_2 pk_63_2.5 pk_84_2.5 Cluster 3 number of entries 10 co_82 co_68 co_67 co_72 co_65 pk_68_2.5 pk_74_2 pk_64_2.5 pk_59_2 pk_36_2
Cluster 6 number of entries 30 co_69 co_70 co_78 co_60 co_53 co_60 co_62 pk_72_2 pk_61_2.5 pk_67_2.5 pk_72_2.5 pk_70_2.5 pk_78_2.5 pk_61_2.5 pk_58_2.5 pk_82_3 pk_77_2.5 pk_71_2 pk_72_2.5 pk_56_2 pk_53_2 pk_69_2 pk_69_2.5	

Figure 4.36: DTW: Centroid

aggregated with elderly Parkinson. We admit three elements wrong classified in fifteen entries.

Cluster 6 has only two individuals, both with similar age, but one has Parkinson disease, and the other is healthy. So, we consider that one is wrongly classified.

Cluster 7 has only two individuals, both with Parkinson.

Cluster 8 has subjects with median difficulty in walking. Healthy individuals with age inferior to 60 and Parkinson individuals with age superior to 70 are wrong classified. So, we consider that the outliers are: co-53, co-54, co-53, co-56, co-53, pk-79-2 and pk-74-2,5. Seven wrong in 25.

Cluster 9 presents individuals with the same difficulty in walking as Cluster 8. We consider that the outliers are Control with age inferior to 60: co-55, co-51 and co-53.

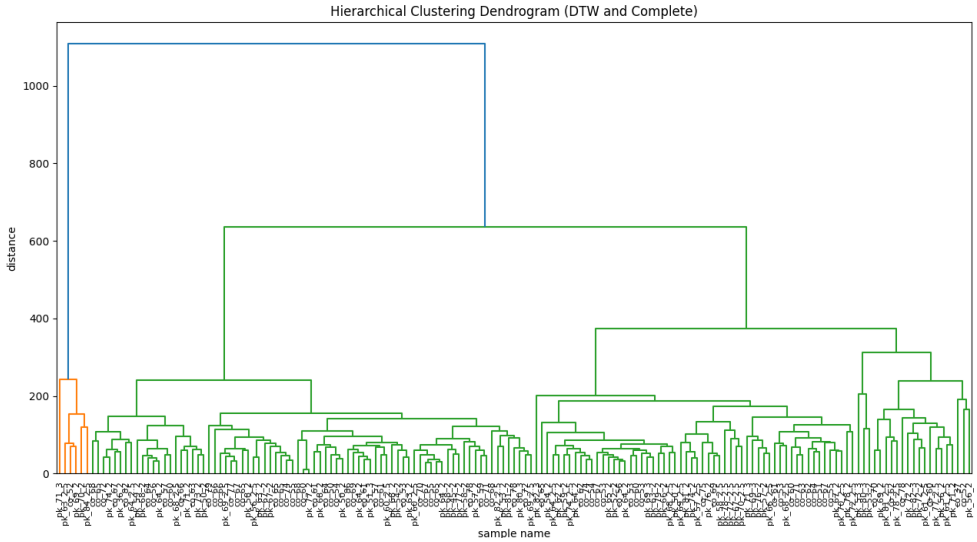


Figure 4.37: DTW: Complete

Cluster 4 number of entries 59 co_66 co_74 co_86 co_79 co_78 co_78 co_69 co_65 co_67 co_68 co_55 co_77 co_65 co_72 co_61 co_66 co_64 co_59 co_67 co_68 co_60 co_65 co_71 co_55 co_70 co_65 co_50 co_61 co_61 co_54 co_57 co_53 pk_82_2.5 pk_68_2 pk_81_2 pk_80_3 pk_68_2 pk_63_2 pk_77_2 pk_51_3 pk_64_2 pk_69_2.5 pk_68_2 pk_45_2 pk_58_2 pk_69_2.5 pk_56_2.5 pk_54_2 pk_67_2 pk_77_2 pk_64_2 pk_60_2.5 pk_50_2 pk_53_2 pk_58_2 pk_63_2 pk_60_2.5 pk_67_2 pk_57_2	Cluster 5 number of entries 57 co_69 co_75 co_67 co_65 co_71 co_74 co_62 co_55 co_53 co_60 co_62 co_53 co_53 co_54 co_64 co_51 co_56 co_53 co_60 co_67 co_61 co_63 pk_68_2.5 pk_69_3 pk_59_2 pk_79_2 pk_71_3 pk_76_2 pk_76_2 pk_81_2 pk_78_2 pk_67_2.5 pk_72_2.5 pk_74_2.5 pk_70_2.5 pk_78_2.5 pk_54_2 pk_69_3 pk_74_2 pk_58_2.5 pk_67_2 pk_64_2.5 pk_68_2.5 pk_66_2 pk_57_2.5 pk_63_2 pk_52_2 pk_62_2 pk_64_2 pk_57_2 pk_50_2 pk_70_2.5 pk_72_2.5 pk_64_2 pk_66_2.5 pk_65_2 pk_69_2.5
Cluster 3 number of entries 21 co_82 co_68 co_63 co_67 co_66 co_72 co_65 co_64 co_57 co_56 co_60 pk_68_2.5 pk_74_2 pk_64_2 pk_64_2.5 pk_68_2 pk_73_2 pk_59_2 pk_50_2 pk_36_2 pk_71_2	Cluster 9 number of entries 15 co_70 co_78 co_60 co_62 pk_72_2 pk_81_2.5 pk_78_2.5 pk_72_2 pk_61_2.5 pk_61_2.5 pk_82_3 pk_77_2.5 pk_71_2 pk_56_2 pk_69_2
Cluster 6 number of entries 1 pk_82_3	Cluster 10 number of entries 3 co_37 co_52 pk_56_2
Cluster 8 number of entries 1 pk_53_3	Cluster 1 number of entries 5 co_52 pk_70_2 pk_69_2 pk_63_2.5 pk_84_2.5
Cluster 2 number of entries 1 pk_71_3	Cluster 7 number of entries 2 pk_80_3 pk_53_2

Figure 4.38: DTW: Complete

4.6.11 DTW distance- Median

By looking at the table 4.40 we can see that this method creates two bigger clusters and the rest of the elements are in individuals clusters. Both bigger clusters represent people with median difficulty in walking. In this way, we consider that cluster 1 has twenty-one elements wrong classified and cluster three has twenty-eight. There are five clusters with only one element.

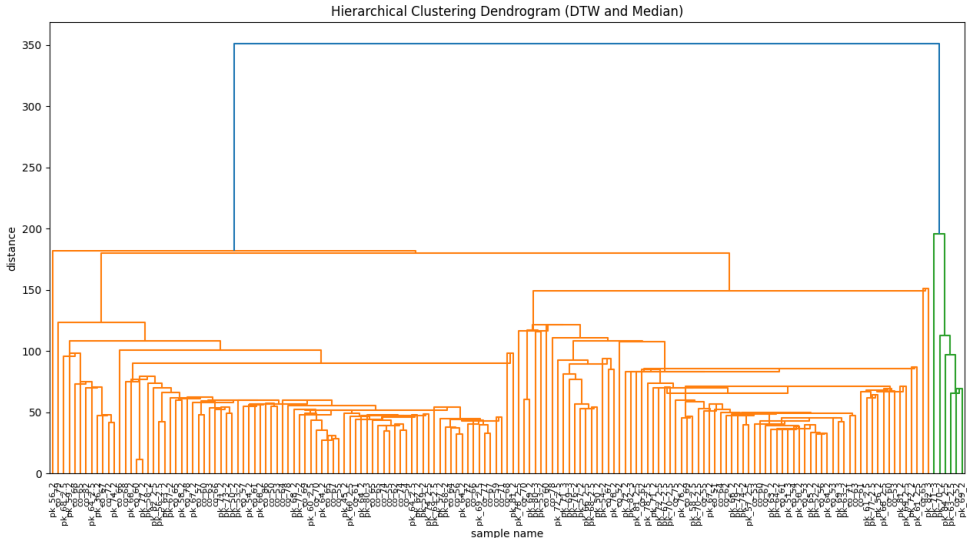


Figure 4.39: DTW: Median

4.6.12 DTW distance- Single

By looking at the dendrogram 4.41 and table 4.42, we can see that this distance metric with this aggregation measure is not able to separate different gait patterns. All the elements are aggregated in one only cluster, and the other nine clusters just have one element each.

Cluster 1 number of entries 81 co_66 co_74 co_86 co_82 co_78 co_78 co_69 co_65 co_68 co_67 co_63 co_67 co_66 co_68 co_55 co_74 co_77 co_65 co_72 co_72 co_61 co_66 co_65 co_64 co_59 co_67 co_62 co_68 co_64 co_60 co_65 co_71 co_55 co_70 co_57 co_65 co_56 co_50 co_60 co_54 co_61 co_57 co_53 pk_82_2.5 pk_68_2 pk_59_2 pk_81_2 pk_80_3 pk_68_2.5 pk_68_2 pk_77_2 pk_74_2.5 pk_74_2 pk_64_2.5 pk_64_2 pk_64_2.5 pk_68_2 pk_64_2 pk_69_2.5 pk_68_2 pk_45_2 pk_58_2 pk_69_2.5 pk_73_2 pk_56_2.5 pk_54_2 pk_67_2 pk_77_2 pk_60_2.5 pk_62_2 pk_64_2 pk_53_2 pk_59_2 pk_58_2 pk_50_2 pk_36_2 pk_63_2 pk_60_2.5 pk_67_2 pk_71_2 pk_57_2	Cluster 3 number of entries 72 co_69 co_75 co_70 co_67 co_78 co_71 co_62 co_55 co_53 co_60 co_53 co_53 co_61 co_60 co_64 co_51 co_62 co_56 co_52 co_53 co_54 co_60 co_67 co_61 co_63 pk_68_2.5 pk_72_2 pk_69_3 pk_79_2 pk_71_3 pk_76_2 pk_76_2 pk_81_2 pk_78_2 pk_81_2.5 pk_78_2.5 pk_72_2 pk_61_2.5 pk_67_2.5 pk_63_2 pk_72_2.5 pk_70_2.5 pk_78_2.5 pk_51_3 pk_54_2 pk_61_2.5 pk_69_3 pk_74_2 pk_58_2.5 pk_80_3 pk_67_2 pk_82_3 pk_68_2.5 pk_66_2 pk_57_2.5 pk_77_2.5 pk_63_2 pk_64_2 pk_52_2 pk_50_2 pk_57_2 pk_71_2 pk_50_2 pk_70_2.5 pk_72_2.5 pk_56_2 pk_64_2 pk_53_2 pk_66_2.5 pk_65_2 pk_69_2 pk_69_2.5
Cluster 6 number of entries 1 pk_56_2	Cluster 2 number of entries 1 co_79
Cluster 8 number of entries 1 pk_71_3	Cluster 4 number of entries 1 co_65
Cluster 5 number of entries 1 pk_82_3	Cluster 7 number of entries 5 co_52 pk_70_2 pk_69_2 pk_63_2.5 pk_84_2.5

Figure 4.40: DTW: Median

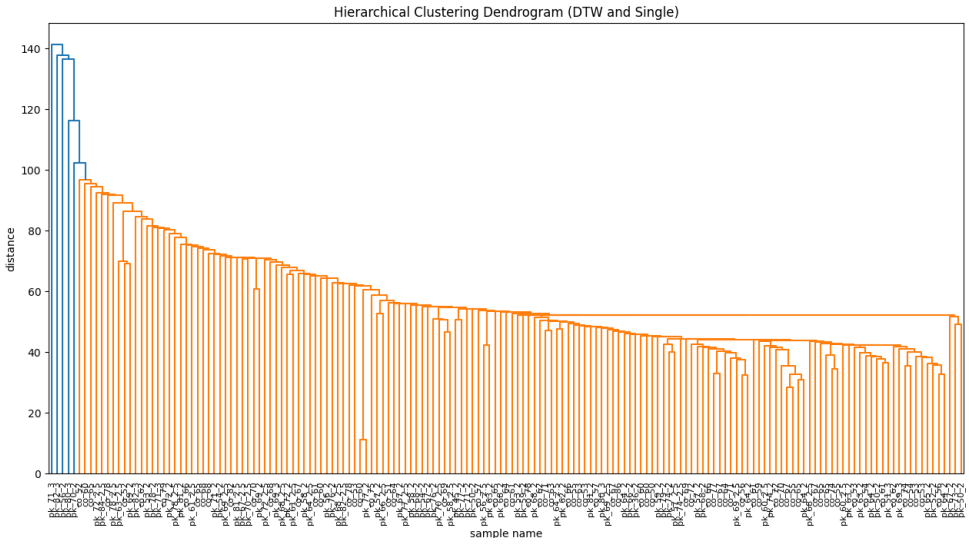


Figure 4.41: DTW: Single

4.6.13 DTW distance- Ward

The majority of the element of the cluster 1 are control individuals. The individuals allocated in this cluster have median difficult in walking, so control subjects with age inferior to 60 do not belong here: co-55, co-59 and co-57. Parkinson's subjects with age superior to 70 do not belong here either because they have gait problems due to Parkinson and due to age, so they have very difficulties walking. So, we consider that pk-73-2, pk-74-2 and pk-71-2 are outliers. Here we have six elements wrong classified in twenty-seven elements.

Cluster 2 also has subjects with median difficulty walking, so we have to use the same criteria as the first cluster. We consider that the outliers are: co-50, co-54, co-54, co-57, co-53, pk-77-2 and pk-74-2.5. In this way, we have seven wrong elements in twenty-eight entries.

Cluster 3 also has subjects with median difficulty in walking. Healthy individuals with age inferior to 60 and Parkinson individuals with age superior to 70 are consider wrong classified. So,

Cluster 1 number of entries 154 co_66 co_74 co_69 co_86 co_75 co_82 co_79 co_78 co_78 co_69 co_70 co_65 co_67 co_68 co_67 co_63 co_67 co_78 co_71 co_66 co_68 co_55 co_74 co_62 co_77 co_65 co_72 co_72 co_61 co_55 co_53 co_66 co_65 co_64 co_59 co_67 co_60 co_62 co_68 co_64 co_53 co_60 co_53 co_65 co_71 co_55 co_70 co_57 co_65 co_56 co_50 co_60 co_61 co_52 co_54 co_64 co_61 co_51 co_62 co_56 co_53 co_54 co_57 co_60 co_67 co_61 co_53 co_63 pk_68_2.5 pk_82_2.5 pk_72_2 pk_68_2 pk_69_3 pk_59_2 pk_81_2 pk_79_2 pk_71_3 pk_76_2 pk_76_2 pk_81_2 pk_80_3 pk_78_2 pk_68_2.5 pk_81_2.5 pk_78_2.5 pk_72_2 pk_61_2.5 pk_68_2 pk_69_2 pk_67_2.5 pk_63_2 pk_63_2.5 pk_77_2 pk_74_2.5 pk_70_2.5 pk_74_2 pk_78_2.5 pk_51_3 pk_54_2 pk_61_2.5 pk_69_3 pk_74_2 pk_58_2.5 pk_67_2 pk_64_2.5 pk_64_2 pk_64_2.5 pk_82_3 pk_68_2 pk_64_2 pk_84_2.5 pk_69_2.5 pk_68_2 pk_68_2.5 pk_45_2 pk_66_2 pk_58_2 pk_57_2.5 pk_69_2.5 pk_77_2.5 pk_73_2 pk_56_2.5 pk_54_2 pk_67_2 pk_63_2 pk_77_2 pk_64_2 pk_52_2 pk_60_2.5 pk_62_2 pk_50_2 pk_64_2 pk_57_2 pk_53_2 pk_59_2 pk_58_2 pk_71_2 pk_50_2 pk_36_2 pk_63_2 pk_50_2 pk_70_2.5 pk_72_2.5 pk_60_2.5 pk_56_2 pk_64_2 pk_67_2 pk_71_2 pk_57_2 pk_53_2 pk_66_2.5 pk_65_2 pk_69_2 pk_69_2.5	Cluster 3 number of entries 1 co_65 Cluster 4 number of entries 1 co_60 Cluster 5 number of entries 1 co_52 Cluster 9 number of entries 1 pk_82_3 Cluster 6 number of entries 1 pk_70_2 Cluster 8 number of entries 1 pk_56_2 Cluster 10 number of entries 1 pk_71_3 Cluster 2 number of entries 1 pk_72_2.5 Cluster 7 number of entries 1 pk_80_3

Figure 4.42: DTW: Single

we consider that the outliers are: co-55, co-56, pk-82-2.5, pk-77-2, pk-81-2 and pk-80-3. We have six elements wrong classified in thirty-four.

Cluster 4 shows individuals with very difficulty in walking. All the Parkinson subjects have more than 60 years. The only control subject is very young, is 52 years, so he does not belong in this group. We consider one element wrong classified in 5 total elements.

Cluster 5 has only one element that is from the Parkinson group and has 71 years old.

Cluster 6 has elements with Parkinson with a high average age. This cluster reflects people with serious difficulty in walking. The only control subject is too young to belong here. The Parkinson element with 56 years is not consider well classified either because he is too young to be compared with Parkinson people that are old. Therefore, we consider to have two elements wrong classified in eleven entries.

Cluster 7 has only a healthy subject, and he is 70 years old. Therefore, it makes sense that he is aggregated with Parkinson people. This cluster reflects people with serious difficulty walking, so we consider that the element wrong classified is the pk-53-2.

Cluster 8 has subjects with Parkinson disease with a high average age. So, it reflects individuals with severe difficulties in walking with little dopamine-producing cells. Healthy people with age inferior to 65 are considered wrong classified: as co-55, co-53, co-52, co-62 and co-61. Parkinson individuals with age inferior to 60, and the disease's stage inferior to 2 are not considered well classified either. However, no individual has these characteristics because the only Parkinson's subject has age inferior to 60 has 2.5 stage in HY scale. We consider five wrong classified in twenty-one.

Cluster 9 has the majority of healthy individuals. It shows people with median difficulty in walking. Parkinson individuals with age superior to 70 or high stage of the disease and healthy individuals with age inferior to 60 are considered wrong classified. We consider that the outliers are: pk-79-2, co-51, co-53, co-53, pk-69-3, co-56 and co-53. We admit seven elements wrong

classified in twenty entries.

Cluster 10 has 100% Parkinson subjects, so we admit that they are all well classified. We consider zero elements wrong classified in twelve entries.

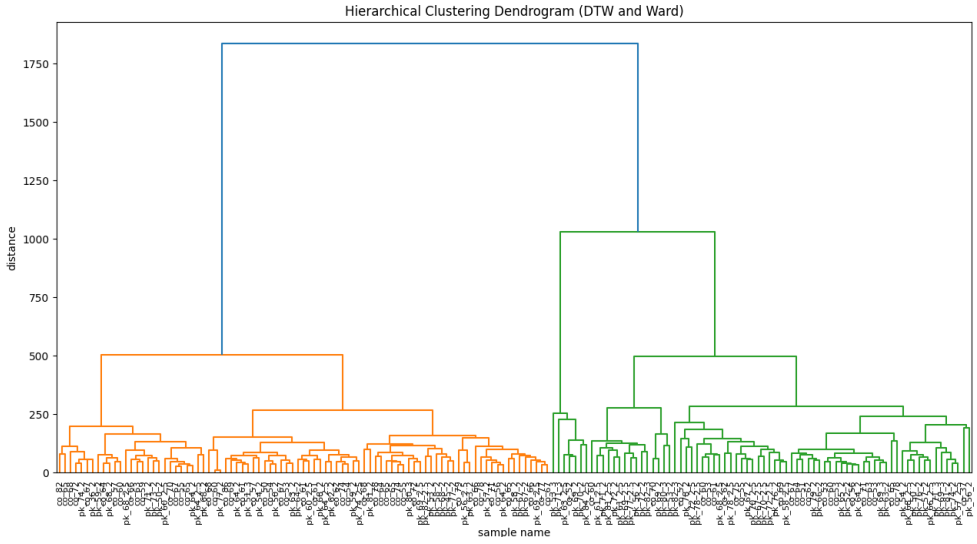


Figure 4.43: DTW: Ward

Cluster 3 number of entries 34 co_66 co_74 co_79 co_78 co_78 co_65 co_68 co_77 co_65 co_72 co_66 co_64 co_67 co_71 co_55 co_65 co_56 pk_82_2.5 pk_68_2 pk_81_2 pk_80_3 pk_64_2 pk_69_2.5 pk_45_2 pk_58_2 pk_69_2.5 pk_56_2.5 pk_67_2 pk_77_2 pk_53_2 pk_58_2 pk_63_2 pk_67_2 pk_57_2	Cluster 7 number of entries 4 co_70 pk_80_3 pk_53_2 pk_69_2
	Cluster 9 number of entries 20 co_67 co_65 co_78 co_71 co_62 co_53 co_53 co_64 co_51 co_56 co_53 co_60 co_63 pk_79_2 pk_69_3 pk_66_2 pk_63_2 pk_52_2 pk_64_2 pk_65_2
Cluster 8 number of entries 21 co_69 co_75 co_55 co_60 co_53 co_62 co_52 co_61 pk_82_3 pk_68_2.5 pk_76_2 pk_78_2 pk_78_2.5 pk_67_2.5 pk_72_2.5 pk_70_2.5 pk_78_2.5 pk_58_2.5 pk_67_2 pk_70_2.5 pk_72_2.5	Cluster 4 number of entries 5 co_52 pk_70_2 pk_69_2 pk_63_2.5 pk_84_2.5
	Cluster 5 number of entries 1 pk_71_3
Cluster 2 number of entries 28 co_86 co_69 co_74 co_61 co_62 co_68 co_60 co_50 co_61 co_54 co_61 co_54 co_57 co_67 co_53 pk_59_2 pk_68_2 pk_63_2 pk_77_2 pk_74_2.5 pk_51_3 pk_64_2.5 pk_54_2 pk_64_2 pk_62_2 pk_50_2 pk_64_2 pk_60_2.5	Cluster 6 number of entries 11 co_60 pk_72_2 pk_81_2.5 pk_72_2 pk_61_2.5 pk_61_2.5 pk_82_3 pk_77_2.5 pk_71_2 pk_56_2 pk_69_2.5
	Cluster 10 number of entries 12 pk_69_3 pk_56_2 pk_71_3 pk_76_2 pk_81_2 pk_54_2 pk_74_2 pk_68_2.5 pk_57_2.5 pk_57_2 pk_50_2 pk_66_2.5
Cluster 1 number of entries 27 co_82 co_68 co_67 co_63 co_67 co_66 co_55 co_72 co_65 co_59 co_64 co_65 co_70 co_57 co_60 pk_68_2.5 pk_74_2 pk_64_2.5 pk_68_2 pk_64_2 pk_68_2 pk_73_2 pk_60_2.5 pk_59_2 pk_50_2 pk_36_2 pk_71_2	

Figure 4.44: DTW: Ward

4.6.14 DTW distance- Weighted

Now, by looking at the Figures 5.18 and 5.19 we can conclude somethings about the clusters.

Cluster 1 has one control element with 52 years. The Parkinson elements are old, so co-52 is consider not well classified.

Cluster 4 has control individuals with a high average age and reflects subjects with high movement disorders. The member that does not belong in this cluster is pk-36-2. So we consider one element wrong classified in eight.

Cluster 5 shows individuals with median movement disorders, so control subjects with age inferior to 60 do not belong here. Parkinson's subjects with age superior to 70 do not belong here either. So, the outliers are: co-55, co-59, co-55, co-57, co-59, co-56, co-50, co-54, co-54, co-53, co-57, pk-82-2.5, pk-77-2, pk-81-2, pk-80-3, pk-74-2.5, pk-71-2 and pk-77-2. We consider eighteen elements wrong classified in eighty-one elements.

Cluster 6 has one control element and one element with Parkinson with 82 years and in a advanced stage of the disease. We consider that there is one element wrong classified in two entries.

Cluster 7 has one only control element with 70 years. This cluster is similar to cluster 7 of DTW distance and ward method. The outlier is the same as the cluster 7: pk-53-2. So we consider one element wrong classified in five.

Cluster 8 has individuals with severe movement disorders. Control individuals with age inferior to 65 and Parkinson individuals younger than 60 years do not belong in this cluster. co-55, co-53, co-51, co-52, co-53, co-53, co-56, co-53, pk-52-2, pk-57-2, pk-54-2, pk-50-2, pk-56-2. We consider this cluster has thirteen elements wrong classified in sixty total elements.

Cluster 2, cluster 3, cluster 9 and cluster 10 have one element each.

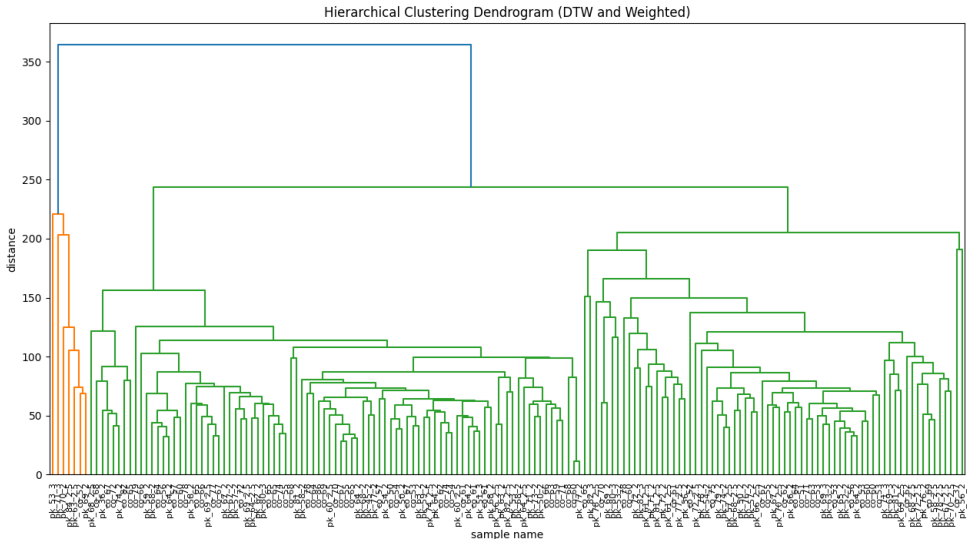


Figure 4.45: DTW: Weighted

4.6.15 Manhattan distance- Average

By looking at Figures 4.47 and 4.48 we can see that this method has one big cluster with almost all the elements. We can assume that this cluster represents people with median difficulty walking,

Cluster 5 number of entries 81 co_66 co_74 co_86 co_79 co_78 co_78 co_69 co_65 co_67 co_63 co_66 co_68 co_55 co_74 co_77 co_65 co_72 co_61 co_66 co_64 co_59 co_67 co_62 co_68 co_64 co_60 co_65 co_71 co_55 co_70 co_57 co_65 co_56 co_50 co_60 co_61 co_54 co_61 co_54 co_57 co_67 co_53 pk_82_2.5 pk_68_2 pk_59_2 pk_81_2 pk_80_3 pk_68_2 pk_63_2 pk_77_2 pk_74_2.5 pk_51_3 pk_64_2.5 pk_64_2 pk_64_2.5 pk_68_2 pk_64_2 pk_69_2.5 pk_68_2 pk_45_2 pk_58_2 pk_69_2.5 pk_73_2 pk_56_2.5 pk_54_2 pk_67_2 pk_77_2 pk_64_2 pk_60_2.5 pk_62_2 pk_50_2 pk_64_2 pk_53_2 pk_59_2 pk_58_2 pk_50_2 pk_63_2 pk_60_2.5 pk_67_2 pk_71_2 pk_57_2	Cluster 8 number of entries 60 co_69 co_75 co_67 co_78 co_71 co_62 co_55 co_53 co_60 co_53 co_53 co_60 co_64 co_51 co_62 co_56 co_52 co_53 co_60 co_61 co_63 pk_68_2.5 pk_72_2 pk_69_3 pk_79_2 pk_71_3 pk_76_2 pk_76_2 pk_81_2 pk_78_2 pk_81_2.5 pk_78_2.5 pk_72_2 pk_61_2.5 pk_67_2.5 pk_72_2.5 pk_70_2.5 pk_54_2 pk_61_2.5 pk_69_3 pk_74_2 pk_58_2.5 pk_67_2 pk_82_3 pk_68_2.5 pk_66_2 pk_57_2.5 pk_77_2.5 pk_63_2 pk_52_2 pk_57_2 pk_71_2 pk_50_2 pk_70_2.5 pk_72_2.5 pk_56_2 pk_64_2 pk_66_2.5 pk_65_2 pk_69_2.5
Cluster 4 number of entries 8 co_82 co_68 co_67 co_72 co_65 pk_68_2.5 pk_74_2 pk_36_2	Cluster 6 number of entries 2 co_65 pk_82_3
Cluster 7 number of entries 5 co_70 pk_78_2.5 pk_80_3 pk_53_2 pk_69_2	Cluster 9 number of entries 1 co_37
Cluster 2 number of entries 1 pk_71_3	Cluster 1 number of entries 5 co_52 pk_70_2 pk_69_2 pk_63_2.5 pk_84_2.5
	Cluster 3 number of entries 1 pk_53_3
	Cluster 10 number of entries 1 pk_56_2

Figure 4.46: DTW: Weighted

so we consider it has thirty-six elements wrong classified in a hundred and twenty entries. There are five clusters with just one element each.

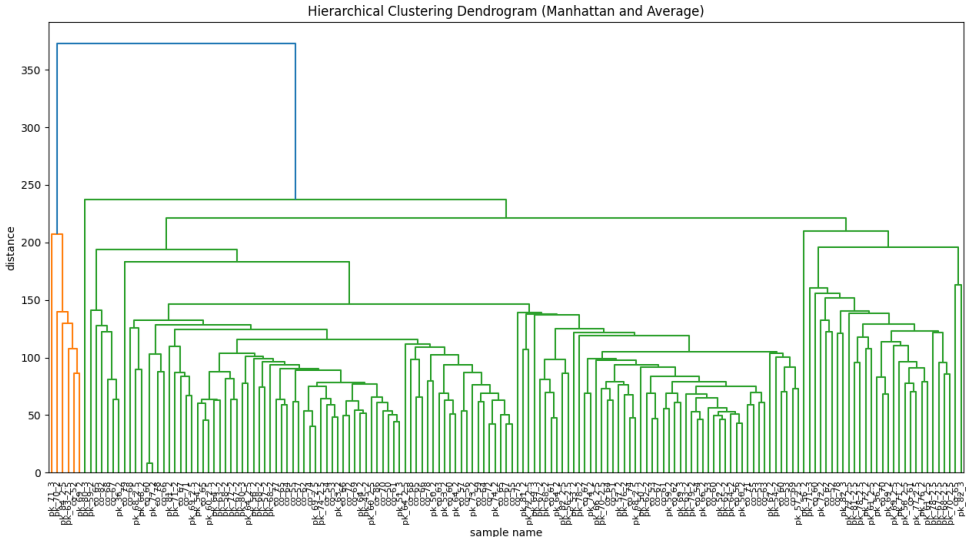


Figure 4.47: Manhattan: Average

4.6.16 Manhattan distance- Centroid

The Manhattan distance used with the centroid measure shows two significant clusters. The cluster four has 107 elements and represents people with median difficulty in walking. We consider this cluster has thirty-two elements wrong classified. The other cluster has thirty-seven elements and represents people with severe difficulty in moving. We consider it has eight people wrong classified. Almost the rest of the clusters have one element each.

Cluster 4 number of entries 120 co_66 co_74 co_69 co_86 co_75 co_78 co_78 co_69 co_65 co_67 co_67 co_63 co_71 co_66 co_68 co_55 co_74 co_62 co_77 co_65 co_72 co_72 co_61 co_55 co_53 co_66 co_64 co_59 co_67 co_60 co_62 co_68 co_64 co_53 co_60 co_53 co_65 co_71 co_55 co_70 co_57 co_65 co_56 co_50 co_60 co_61 co_54 co_64 co_61 co_51 co_56 co_53 co_54 co_57 co_60 co_67 co_53 co_63 pk_82_2.5 pk_68_2 pk_69_3 pk_59_2 pk_81_2 pk_79_2 pk_76_2 pk_81_2 pk_80_3 pk_78_2 pk_68_2.5 pk_68_2 pk_63_2 pk_77_2 pk_74_2.5 pk_74_2 pk_51_3 pk_54_2 pk_69_3 pk_74_2 pk_67_2 pk_64_2.5 pk_64_2 pk_64_2.5 pk_68_2 pk_64_2 pk_69_2.5 pk_68_2 pk_68_2.5 pk_45_2 pk_66_2 pk_58_2 pk_57_2.5 pk_69_2.5 pk_73_2 pk_56_2.5 pk_54_2 pk_67_2 pk_63_2 pk_77_2 pk_64_2 pk_52_2 pk_60_2.5 pk_62_2 pk_50_2 pk_64_2 pk_57_2 pk_53_2 pk_58_2 pk_50_2 pk_63_2 pk_50_2 pk_70_2.5 pk_72_2.5 pk_60_2.5 pk_64_2 pk_67_2 pk_71_2 pk_57_2 pk_53_2 pk_66_2.5 pk_65_2	Cluster 7 number of entries 1 co_65
Cluster 3 number of entries 6 co_82 co_68 co_67 co_65 pk_59_2 pk_36_2	Cluster 1 number of entries 5 co_52 pk_70_2 pk_69_2 pk_63_2.5 pk_84_2.5
Cluster 5 number of entries 1 co_79	Cluster 8 number of entries 1 pk_82_3
	Cluster 9 number of entries 1 pk_56_2
	Cluster 2 number of entries 1 pk_71_3
	Cluster 10 number of entries 1 pk_80_3
	Cluster 6 number of entries 26 co_70 co_78 co_60 co_62 co_52 co_61 pk_68_2.5 pk_72_2 pk_71_3 pk_76_2 pk_81_2.5 pk_78_2.5 pk_72_2 pk_61_2.5 pk_67_2.5 pk_72_2.5 pk_70_2.5 pk_78_2.5 pk_61_2.5 pk_58_2.5 pk_82_3 pk_77_2.5 pk_71_2 pk_56_2 pk_69_2 pk_69_2.5

Figure 4.48: Manhattan: Average

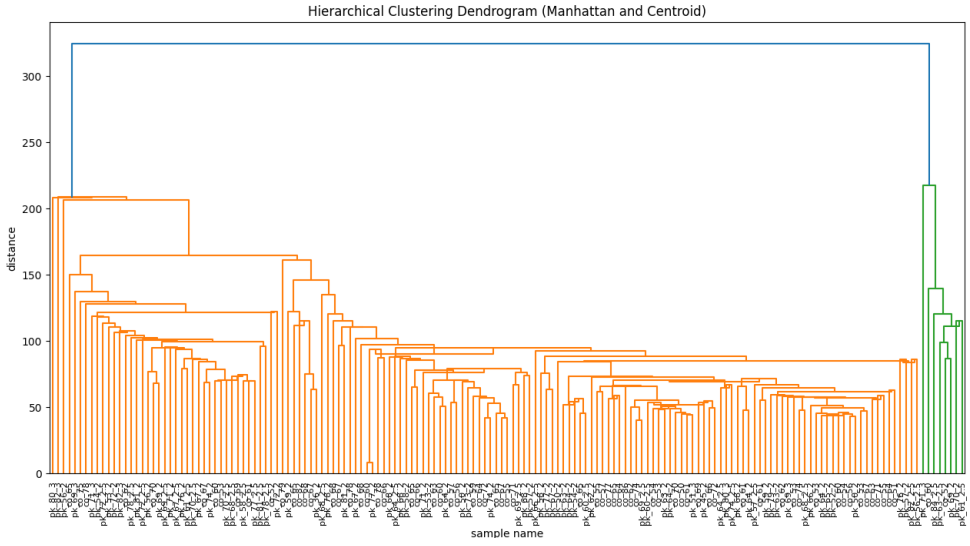


Figure 4.49: Manhattan: Centroid

4.6.17 Manhattan distance- Complete

Cluster 1 has a high average age for the Parkinson group, so has elements with severe difficulty walking. The two elements Control are young, so the probability of having gait disorders is low, so we consider that they do not belong in this cluster. All the elements with Parkinson are older than 60. We consider two elements wrong classified in six entries.

Cluster 2 has only one element. The element has Parkinson disease in stage 3 and is 71 years old.

Cluster 3 represents individuals with median difficulty in walking. The outliers are Control subjects younger than 60 (co-56 and co-57) and Parkinson subjects older than 70 years or in stage

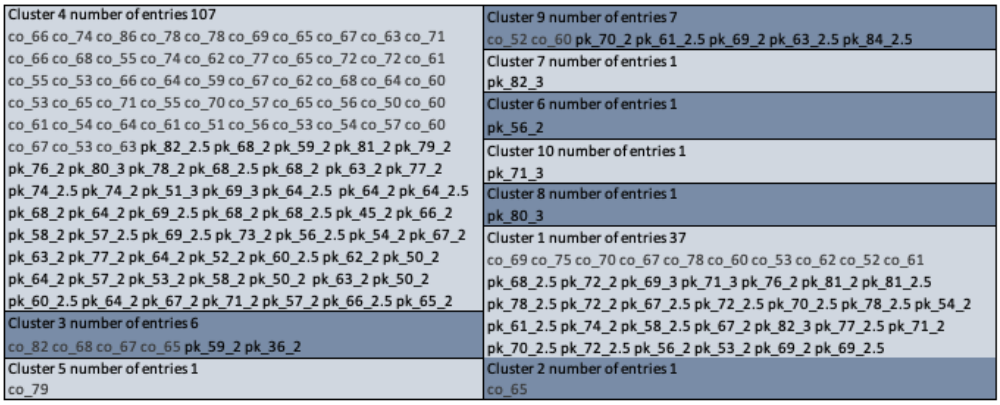


Figure 4.50: Manhattan: Centroid

3 of the disease (pk-82-2.5). We consider two elements wrong classified in eleven entries.

Cluster 4 was supposed to have elements from the Control group with age superior to 60 and elements from the Parkinson group with age inferior to 70. In this way, the elements wrong classified are co-55, co-59, pk-73-2 and pk-74-2. We consider four elements wrong classified in eighteen.

Cluster 5 shows individuals with median difficulty walking, so this cluster should have people control over 60 and Parkinson subjects with age inferior to 70 years. co-55, co-50, co-54, co-57, co-53, co-59, pk-77-2, pk-81-2, pk-77-2 and pk-71-2 . We consider nine elements wrong classified in forty-seven.

Cluster 6 has majority Parkinson with a high average age. Hence, it characterizes as a group with elements with severe movement disorders. Control people with age inferior to 65 do not belong in this cluster: co-62, co-52 and co-61. People with Parkinson and younger than 60 years do not fit here either: pk-56-2 and pk-53-2. Like in cluster 10 from the Manhattan distance and Ward method, the pk-58-2.5 is considered well classified in this cluster. We consider five elements wrong classified in twenty-eight.

Cluster 7 only has one element, and it is a healthy individual with 79 years.

Cluster 8 only has two subjects with Parkinson with similar age and the same stage of the disease. This cluster is equal to cluster 9 from Manhattan distance and Ward method.

Cluster 9 characterized individuals with median difficulty walking, so this cluster should have Control individuals over 60 and Parkinson subjects with age inferior to 70 years and/or in an advanced stage of the disease. The outliers are: co-55, co-53, co-53, co-53, co-54, co-51, co-56, co-53, pk-74-2.5, pk-79-2, pk-76-2, pk-80-3, pk-74-2, pk-78-2, pk-70-2.5 and pk-69-3. We consider sixteen elements wrong classified in forty-eight.

Cluster 10 only has one element. He is healthy and has 79 years old.

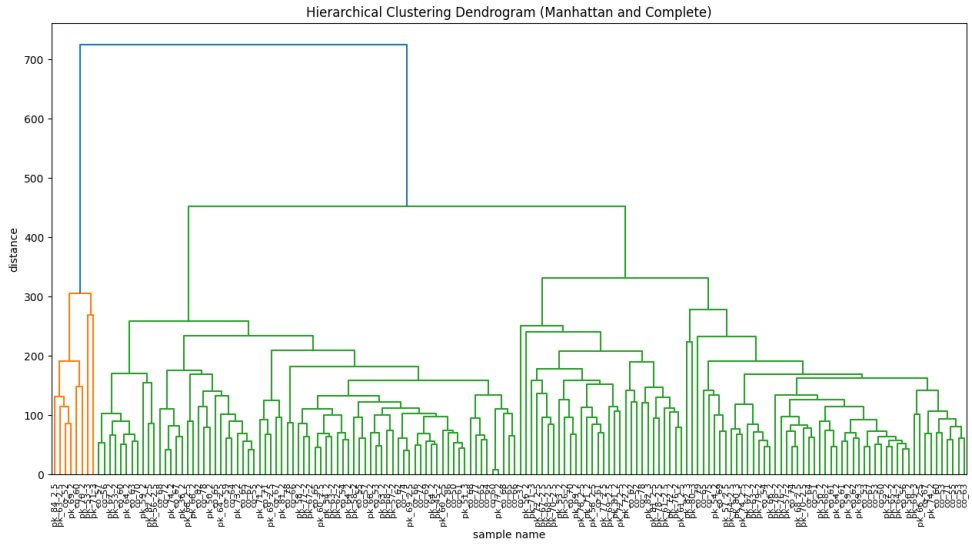


Figure 4.51: Manhattan: Complete

Cluster 5 number of entries 65	Cluster 10 number of entries 49
co_66 co_74 co_86 co_82 co_78 co_78 co_69	co_69 co_75 co_79 co_67 co_65 co_71 co_74
co_65 co_68 co_67 co_67 co_66 co_68 co_55	co_62 co_61 co_55 co_53 co_60 co_53 co_53
co_77 co_65 co_72 co_72 co_66 co_65 co_64	co_54 co_64 co_61 co_51 co_56 co_53 co_60
co_59 co_67 co_62 co_68 co_64 co_60 co_65	co_67 co_63 pk_69_3 pk_59_2 pk_79_2 pk_76_2 pk_80_3
co_71 co_55 co_65 co_50 co_61 co_54 co_57	pk_78_2 pk_68_2 pk_74_2.5 pk_54_2 pk_69_3 pk_74_2 pk_67_2
co_53 pk_68_2 pk_81_2 pk_68_2.5 pk_63_2 pk_77_2 pk_74_2	pk_64_2.5 pk_68_2.5 pk_66_2 pk_57_2.5 pk_63_2 pk_52_2 pk_50_2
pk_51_3 pk_64_2.5 pk_68_2 pk_64_2 pk_69_2.5 pk_68_2 pk_45_2	pk_64_2 pk_57_2 pk_50_2 pk_70_2.5 pk_64_2 pk_66_2.5 pk_65_2
pk_58_2 pk_69_2.5 pk_73_2 pk_54_2 pk_77_2 pk_64_2 pk_60_2.5	Cluster 6 number of entries 28
pk_62_2 pk_58_2 pk_50_2 pk_36_2 pk_63_2 pk_60_2.5 pk_67_2	co_70 co_78 co_62 co_52 co_61 pk_68_2.5 pk_72_2
pk_71_2 pk_57_2	pk_71_3 pk_76_2 pk_81_2 pk_81_2.5 pk_78_2.5 pk_72_2 pk_61_2.5
Cluster 4 number of entries 11	pk_67_2.5 pk_72_2.5 pk_70_2.5 pk_78_2.5 pk_61_2.5 pk_58_2.5 pk_82_3
co_63 co_70 co_57 co_56 co_60 pk_82_2.5 pk_64_2	pk_77_2.5 pk_71_2 pk_72_2.5 pk_56_2 pk_53_2 pk_69_2 pk_69_2.5
pk_56_2.5 pk_67_2 pk_53_2 pk_59_2	Cluster 1 number of entries 6
Cluster 8 number of entries 1	co_52 co_60 pk_70_2 pk_69_2 pk_63_2.5 pk_84_2.5
co_37	Cluster 2 number of entries 1
Cluster 9 number of entries 2	pk_53_3
pk_82_3 pk_80_3	Cluster 7 number of entries 1
Cluster 3 number of entries 1	pk_56_2
pk_71_3	

Figure 4.52: Manhattan: Complete

4.6.18 Manhattan distance- Median

By looking at the table 4.54 we can see that the majority elements are aggregated in two big clusters, Cluster 5 has elements with median difficulty in walking and we consider it presents thirty-eight elements wrong classified. Cluster 1 has fifty-one elements and we consider thirteen are wrong classified.

4.6.19 Manhattan distance- Ward

Cluster 1 is majority control. The difficulty in walking is median, so this cluster should have people control over 60 and Parkinson subjects with age inferior to 70 years. In this way, the outliers are co-55, co-59, pk-74-2 and pk-73-2. Here we consider to have four elements wrong classified in nineteen.

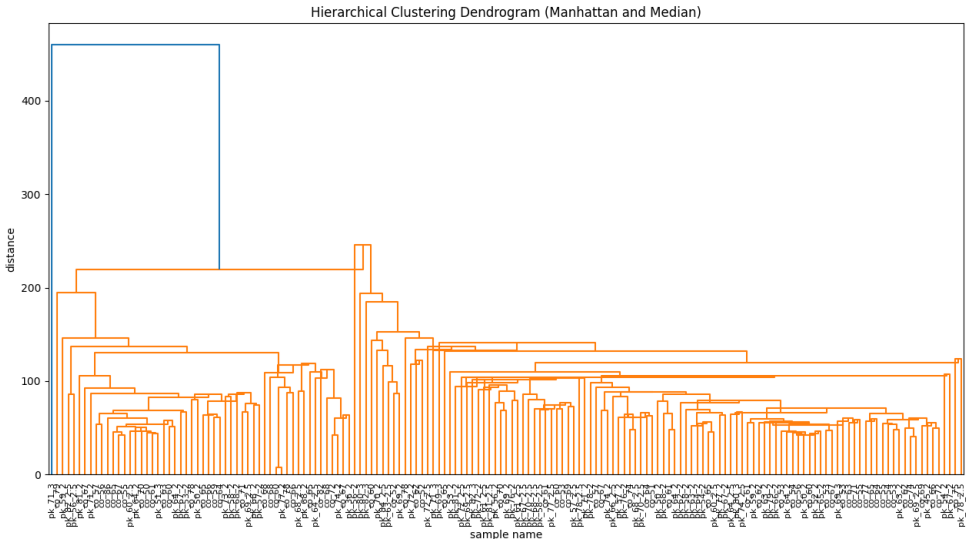


Figure 4.53: Manhattan: Median

Cluster 5 number of entries 99 co_66 co_74 co_69 co_75 co_69 co_70 co_67 co_65 co_71 co_74 co_62 co_77 co_65 co_72 co_61 co_55 co_53 co_64 co_60 co_62 co_53 co_53 co_55 co_65 co_54 co_64 co_61 co_51 co_62 co_56 co_52 co_53 co_54 co_57 co_60 co_67 co_61 co_53 co_63 pk_68_2.5 pk_72_2 pk_69_3 pk_59_2 pk_79_2 pk_71_3 pk_76_2 pk_76_2 pk_81_2 pk_80_3 pk_78_2 pk_81_2.5 pk_78_2.5 pk_72_2 pk_61_2.5 pk_68_2 pk_67_2.5 pk_63_2 pk_72_2.5 pk_74_2.5 pk_70_2.5 pk_78_2.5 pk_54_2 pk_61_2.5 pk_69_3 pk_74_2 pk_58_2.5 pk_67_2 pk_64_2.5 pk_82_3 pk_69_2.5 pk_68_2.5 pk_45_2 pk_66_2 pk_58_2 pk_57_2.5 pk_77_2.5 pk_54_2 pk_63_2 pk_77_2 pk_64_2 pk_52_2 pk_62_2 pk_50_2 pk_64_2 pk_57_2 pk_71_2 pk_63_2 pk_50_2 pk_70_2.5 pk_72_2.5 pk_60_2.5 pk_56_2 pk_64_2 pk_67_2 pk_53_2 pk_66_2.5 pk_65_2 pk_69_2 pk_69_2.5	Cluster 4 number of entries 6 co_52 co_60 pk_70_2 pk_69_2 pk_63_2.5 pk_84_2.5
Cluster 3 number of entries 1 co_79	Cluster 7 number of entries 1 pk_82_3
Cluster 6 number of entries 1 co_78	Cluster 9 number of entries 1 pk_56_2
	Cluster 10 number of entries 1 pk_71_3
	Cluster 8 number of entries 1 pk_80_3
	Cluster 2 number of entries 1 pk_59_2
	Cluster 1 number of entries 51 co_86 co_82 co_78 co_78 co_65 co_68 co_67 co_63 co_67 co_66 co_68 co_55 co_72 co_66 co_65 co_59 co_67 co_68 co_64 co_60 co_65 co_71 co_70 co_57 co_56 co_50 co_60 co_61 pk_82_2.5 pk_68_2 pk_81_2 pk_68_2.5 pk_77_2 pk_74_2 pk_51_3 pk_64_2 pk_64_2.5 pk_68_2 pk_64_2 pk_68_2 pk_69_2.5 pk_73_2 pk_56_2.5 pk_67_2 pk_60_2.5 pk_53_2 pk_58_2 pk_50_2 pk_36_2 pk_71_2 pk_57_2

Figure 4.54: Manhattan: Median

Cluster 2 is characterized as a group with median difficulty in gait, just like cluster 1. So, the outliers are: co-55, co-53, co-53, co-54, co-57, pk-74-2.5, pk-77-2, pk-73-2 and pk-80-3. We consider nine elements wrong classified in thirty-nine

Cluster 3 is majority Parkinson with a high average age. In this way, this group represents individuals with severe movement disorders. The younger element with Parkinson disease is 57 and is in an early stage of the disease (stage 2 on HY scale), so he is wrong classified. Both control elements are older than 65 years; therefore, they fit in this group. We consider one element wrong classified in seven.

Cluster 4 follows the same criteria as the first cluster. So, the elements that are wrong classified are co-50, co-57, co-56 and pk-82-2.5. We consider four elements wrong classified in nineteen elements.

Cluster 5 has Parkinson elements with a high average age. The two elements from the Control group are too young to be aggregated with these Parkinson elements with severe gait problems. We consider the outliers are co-60 and co-52. We admit two elements wrong classified in six.

Cluster 6 is a group with two individuals, both with Parkinson and stage 3 of the disease. The standard deviation of the age is high. However, as both are in the same stage of the disease, we can consider that the group is homogeneous.

Cluster 7 is characterized as a group with elements with severe difficulty in walking. The two control individuals have 67 and 79 years, so they fit in this cluster. The younger Parkinson element does not fit because he is too young and in an early disease stage.

Cluster 8 has an early average age for the Control group. The difficulty in walking is median. We consider the Control elements younger than 60 years are excluded: co-55, co-53, co-56, co-37, co-54, co-51, co-5 and co-53. Parkinson individuals with age inferior to 70 are also excluded: pk-76-2, pk-79-2 and pk-78-2.

Cluster 9 is a uniform group. Only has two elements, both with Parkinson, both with similar age and both on stage 3 of the HY scale.

Cluster 10 has majority Parkinson with a high average age. Therefore, it characterizes as a group with elements with severe difficulty in walking. Control individuals with age inferior to 65 do not belong in this cluster - co-61, co-62 and co-52. Parkinson individuals with age inferior to 60 and/or in an early stage of Parkinson disease do not belong here either – pk-56-2. Note that there is another individual with Parkinson with age inferior to 60. This subject has stage 2.5 on HY scale and 58 years, so he is considered right classified. We consider four elements wrong classified in twenty-six.

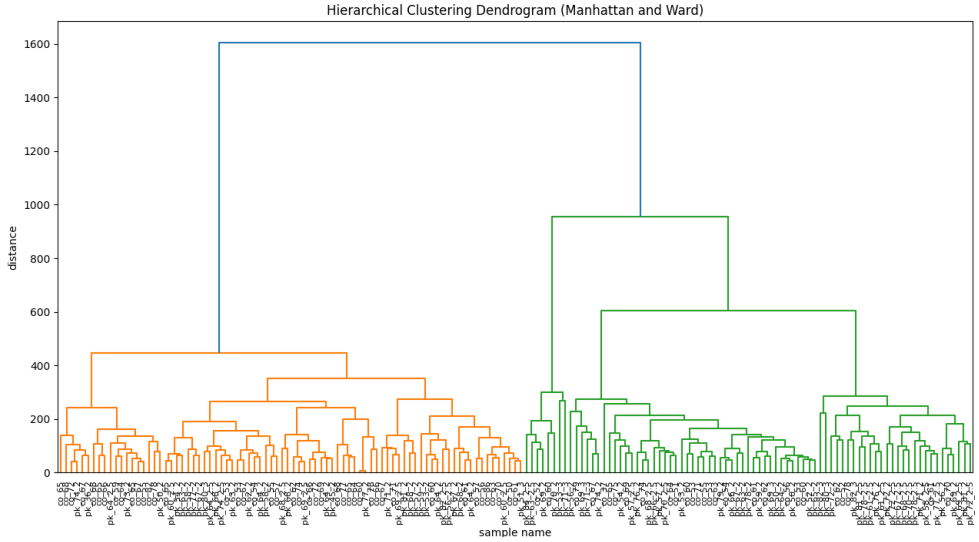


Figure 4.55: Manhattan Distance: Ward

Cluster 2 number of entries 39 co_66 co_74 co_78 co_69 co_68 co_77 co_65 co_72 co_66 co_64 co_62 co_60 co_53 co_55 co_65 co_54 co_57 co_67 co_53 pk_80_3 pk_68_2.5 pk_63_2 pk_77_2 pk_74_2.5 pk_64_2.5 pk_68_2 pk_64_2 pk_69_2. pk_68_2 pk_45_2 pk_58_2 pk_54_2 pk_77_2 pk_64_2 pk_62_2 pk_58_2 pk_63_2 pk_60_2.5 pk_67_2	Cluster 8 number of entries 39 co_69 co_75 co_65 co_71 co_74 co_62 co_61 co_55 co_53 co_60 co_53 co_37 co_54 co_64 co_51 co_56 co_53 co_60 co_63 pk_59_2 pk_79_2 pk_76_2 pk_78_2 pk_54_2 pk_69_3 pk_67_2 pk_68_2.5 pk_66_2 pk_57_2.5 pk_63_2 pk_52_2 pk_50_2 pk_57_2 pk_50_2 pk_70_2.5 pk_64_2 pk_53_2 pk_66_2.5 pk_65_2
Cluster 4 number of entries 19 co_86 co_63 co_70 co_57 co_56 co_50 co_60 co_61 co_61 pk_82_2.5 pk_68_2 pk_51_3 pk_64_2 pk_56_2.5 pk_67_2 pk_60_2.5 pk_64_2 pk_53_2 pk_59_2	Cluster 1 number of entries 19 co_82 co_78 co_65 co_68 co_67 co_67 co_66 co_55 co_72 co_65 co_59 co_68 co_64 co_65 pk_74_2 pk_64_2.5 pk_73_2 pk_50_2 pk_36_2
Cluster 10 number of entries 26 co_70 co_78 co_62 co_52 co_61 pk_68_2.5 pk_72_2 pk_76_2 pk_81_2 pk_81_2.5 pk_78_2.5 pk_72_2 pk_61_2.5 pk_67_ pk_72_2.5 pk_70_2.5 pk_78_2.5 pk_61_2.5 pk_58_2.5 pk_82_3 pk_ pk_71_2 pk_72_2.5 pk_56_2 pk_69_2 pk_69_2.5	Cluster 7 number of entries 6 co_79 co_67 pk_69_3 pk_56_2 pk_71_3 pk_74_2
Cluster 9 number of entries 2 pk_82_3 pk_80_3	Cluster 3 number of entries 7 co_67 co_71 pk_68_2 pk_81_2 pk_69_2.5 pk_71_2 pk_57_2
Cluster 6 number of entries 2 pk_53_3 pk_71_3	Cluster 5 number of entries 6 co_52 co_60 pk_70_2 pk_69_2 pk_63_2.5 pk_84_2.5

Figure 4.56: Manhattan Distance: Ward

4.6.20 Manhattan distance- Single

By looking at the table 4.58 we can see a big cluster with almost all the elements inside. This means that the method do not can distinguish different gait patterns.

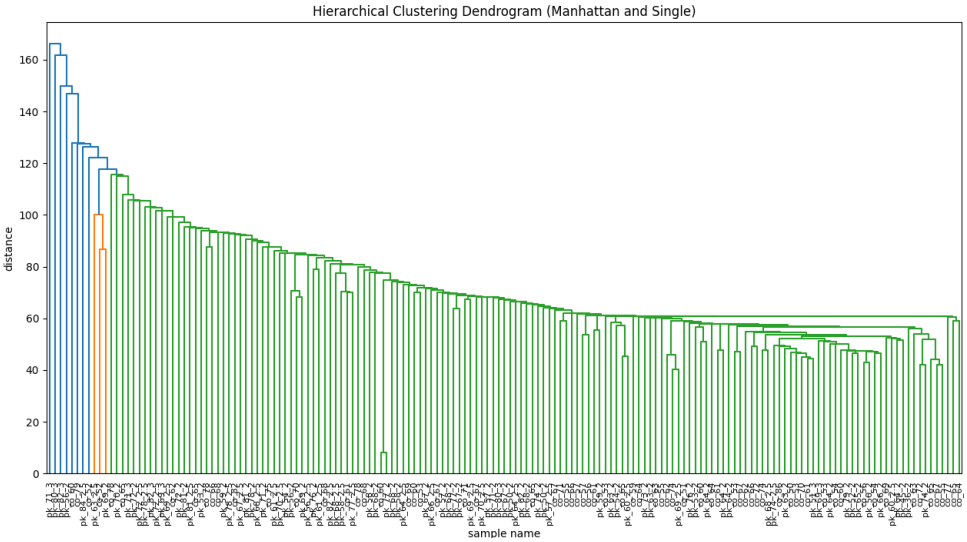


Figure 4.57: Manhattan: Single

4.6.21 Manhattan distance- Weighted

Cluster 1 has a high average age for the Parkinson group, so it has elements with severe difficulty walking. The two control individuals are younger than 65, so they do not fit in this group: co-56 and co-52. We consider two elements wrong classified in six entries.

Cluster 2 only has one element, and he has Parkinson in stage 3. He is 71 years old.

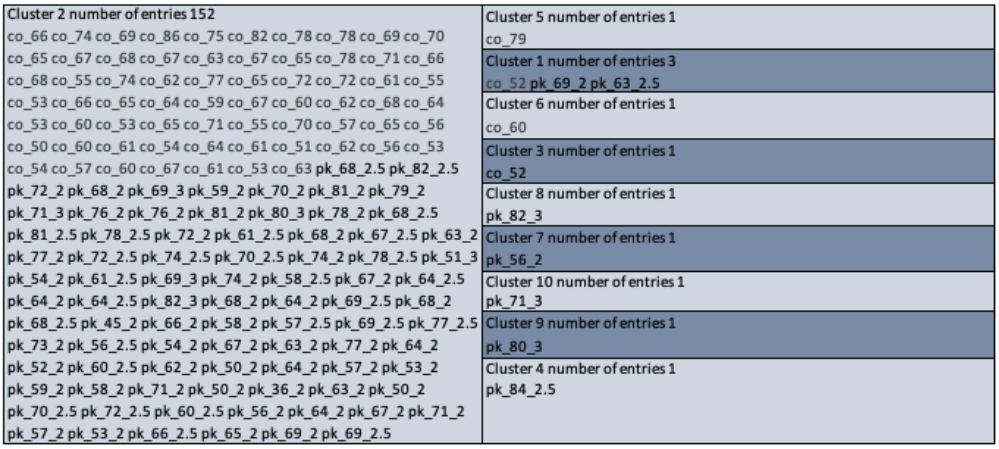


Figure 4.58: Manhattan: Single

Cluster 3 has one element. The individual belongs to the Parkinson group, is in stage 3 of disease, and 53 years old.

Cluster 4 has 84 individuals. Represents subjects with median difficulty walking, so Control younger than 60 and Parkinson older than 70 do not belong here and are wrongly classified. The outliers are: co-55, co-53, co-56, co-59, co-50, co-55, co-57, co-53, co-54, co-57, pk-82-2.5, pk-81-2, pk-77-2, pk-73-2, pk-80-3, pk-74-2, pk-74-2.5, pk-77-2 and pk-71-2. We consider there are nineteen elements wrong classified in a total of eighty-four.

Cluster 5 has one only element, the same as in cluster 10 from Manhattan distance and complete method.

Cluster 6 has 68 elements, most of them with Parkinson disease. In this way, it characterizes as a group with elements with severe movement disorders. Healthy people with age inferior to 65 do not fit in this cluster, and people with Parkinson and younger than 60 years do not fit here either. So we consider the outliers are: co-54, co-55, co-51, co-52, co-53, co-56, co-53, co-53, co-60, co-62, co-64, co-61, co-61, co-62 and co-60.

Cluster 7, 8, 9 ans 10 have only one element each.

Lastly, we have the dendograms regarding the Fourier distance.

4.6.22 Fourier distance- Average

The Fourier distance used with the average measure shows one big cluster. The cluster nine has 117 elements and represents people with median difficulty in walking. This cluster has 33 elements are consider wrong classified. There is another cluster with a reasonable number of elements, cluster 5. Cluster 5 has 27 entries and 7 wrong classified. Almost the rest of the clusters have one element each.

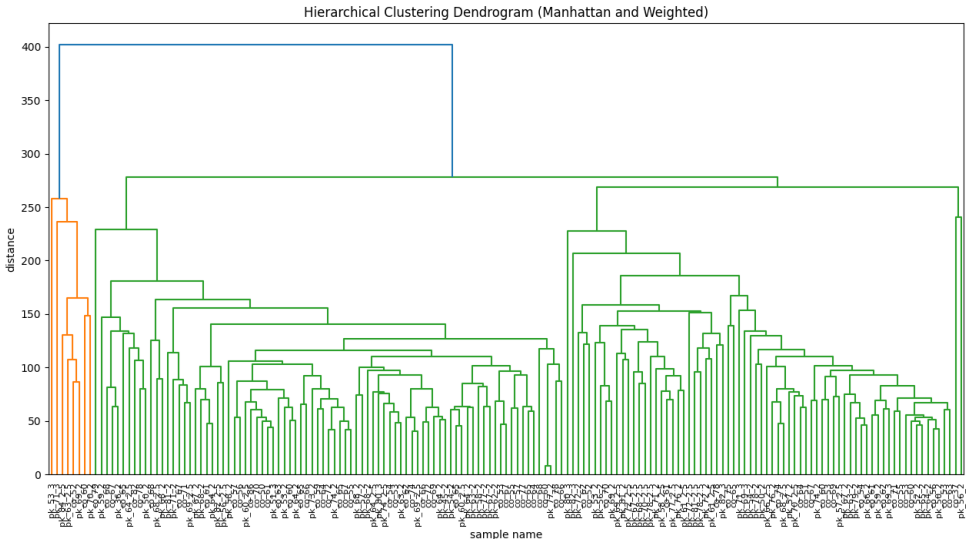


Figure 4.59: Manhattan: Weighted

Cluster 4 number of entries 84	Cluster 6 number of entries 68
co_66 co_74 co_86 co_82 co_78 co_78 co_69	co_69 co_75 co_70 co_67 co_65 co_78 co_71
co_65 co_68 co_67 co_63 co_67 co_66 co_68	co_74 co_62 co_61 co_55 co_53 co_60 co_53
co_55 co_77 co_65 co_72 co_72 co_66 co_65	co_54 co_64 co_51 co_62 co_56 co_52 co_53
co_64 co_59 co_67 co_62 co_68 co_64 co_60	co_60 co_61 co_63 pk_68_2.5 pk_72_2 pk_69_3 pk_59_2
co_53 co_65 co_71 co_55 co_70 co_57 co_65	pk_79_2 pk_71_3 pk_76_2 pk_76_2 pk_81_2 pk_78_2 pk_81_2.5
co_56 co_50 co_60 co_61 co_61 co_54 co_57	pk_78_2.5 pk_72_2 pk_61_2.5 pk_67_2.5 pk_72_2.5 pk_70_2.5 pk_78_2.5
co_67 co_53 pk_82_2.5 pk_68_2 pk_81_2 pk_80_3 pk_68_2.5	pk_54_2 pk_61_2.5 pk_69_3 pk_74_2 pk_58_2.5 pk_67_2 pk_82_3
pk_68_2 pk_63_2 pk_77_2 pk_74_2.5 pk_74_2 pk_51_3 pk_64_2.5	pk_68_2.5 pk_66_2 pk_57_2.5 pk_77_2.5 pk_63_2 pk_52_2 pk_50_2
pk_64_2 pk_64_2.5 pk_68_2 pk_64_2 pk_69_2.5 pk_68_2 pk_45_2	pk_57_2 pk_71_2 pk_50_2 pk_70_2.5 pk_72_2.5 pk_56_2 pk_64_2
pk_58_2 pk_69_2.5 pk_73_2 pk_56_2.5 pk_54_2 pk_67_2 pk_77_2	pk_53_2 pk_66_2.5 pk_65_2 pk_69_2 pk_69_2.5
pk_64_2 pk_60_2.5 pk_62_2 pk_64_2 pk_53_2 pk_59_2 pk_58_2	
pk_50_2 pk_36_2 pk_63_2 pk_60_2.5 pk_67_2 pk_71_2 pk_57_2	Cluster 7 number of entries 1
Cluster 5 number of entries 1	pk_82_3
co_79	Cluster 3 number of entries 1
Cluster 9 number of entries 1	pk_53_3
co_37	Cluster 10 number of entries 1
Cluster 1 number of entries 6	pk_56_2
co_52 co_60 pk_70_2 pk_69_2 pk_63_2.5 pk_84_2.5	Cluster 2 number of entries 1
Cluster 8 number of entries 1	pk_71_3
pk_80_3	

Figure 4.60: Manhattan: Weighted

4.6.23 Fourier distance- Single

The Fourier distance combined with single, do not present good results, as we can see by looking at the dendrogram 4.27. By looking at the table 4.28 we can see that the majority elements are aggregated in just one big cluster, so the algorithm does not have an adequate aggregation capacity. In this big cluster, there are 154 elements and 53 elements we consider wrong classified.

4.6.24 Fourier distance- Centroid

We can see by looking at the dendrogram 4.65 and the table 4.66 that the majority elements are aggregated in just one big cluster, so the algorithm does not have an adequate aggregation capacity. In this big cluster, there are 152 elements and 53 elements we consider wrong classified.

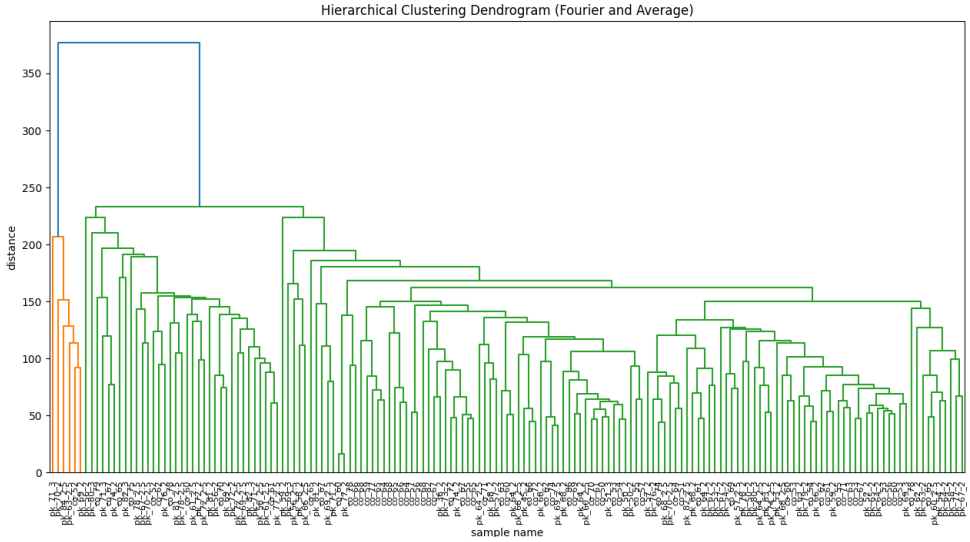


Figure 4.61: Fourier: Average

Cluster 9 number of entries 117 co_66 co_74 co_69 co_86 co_82 co_78 co_78 co_69 co_65 co_68 co_67 co_63 co_67 co_71 co_66 co_68 co_55 co_74 co_62 co_77 co_65 co_72 co_72 co_61 co_55 co_53 co_66 co_65 co_64 co_59 co_67 co_60 co_62 co_68 co_64 co_53 co_60 co_53 co_65 co_71 co_55 co_70 co_57 co_65 co_56 co_50 co_60 co_61 co_54 co_64 co_61 co_51 co_56 co_53 co_54 co_57 co_60 co_67 co_53 co_63 pk_68_2.5 pk_82_2.5 pk_68_2 pk_59_2 pk_81_2 pk_79_2 pk_76_2 pk_80_3 pk_78_2 pk_68_2.5 pk_68_2 pk_63_2 pk_77_2 pk_74_2.5 pk_74_2 pk_51_3 pk_54_2 pk_69_3 pk_67_2 pk_64_2.5 pk_64_2 pk_64_2.5 pk_68_2 pk_64_2 pk_69_2.5 pk_68_2 pk_68_2.5 pk_45_2 pk_66_2 pk_58_2 pk_57_2.5 pk_69_2.5 pk_73_2 pk_54_2 pk_67_2 pk_63_2 pk_77_2 pk_64_2 pk_52_2 pk_60_2.5 pk_62_2 pk_50_2 pk_64_2 pk_57_2 pk_53_2 pk_50_2 pk_36_2 pk_63_2 pk_50_2 pk_70_2.5 pk_60_2.5 pk_64_2 pk_67_2 pk_71_2 pk_57_2 pk_53_2 pk_65_2	Cluster 4 number of entries 2 co_65 pk_82_3
Cluster 3 number of entries 4 co_79 co_67 pk_71_3 pk_74_2	Cluster 1 number of entries 5 co_52 pk_70_2 pk_69_2 pk_63_2.5 pk_84_2.5
	Cluster 8 number of entries 4 pk_69_3 pk_56_2.5 pk_58_2 pk_66_2.5
	Cluster 7 number of entries 1 pk_56_2
	Cluster 2 number of entries 1 pk_71_3
	Cluster 6 number of entries 1 pk_80_3
	Cluster 10 number of entries 1 pk_59_2
	Cluster 5 number of entries 27 co_75 co_70 co_78 co_60 co_62 co_52 co_61 pk_72_2 pk_76_2 pk_81_2 pk_81_2.5 pk_78_2.5 pk_72_2 pk_61_2.5 pk_67_2.5 pk_72_2.5 pk_70_2.5 pk_78_2.5 pk_61_2.5 pk_58_2.5 pk_82_3 pk_77_2.5 pk_71_2 pk_72_2.5 pk_56_2 pk_69_2 pk_69_2.5

Figure 4.62: Fourier: Average

4.6.25 Fourier distance- Complete

Cluster 1 only has one healthy subject with 52 years old. The cluster reflects people with median movement disorders, so we consider the outlier is this Control individual: co-52. We admit one element wrong classified in five.

Cluster 2 has people with median difficulties in walking. The average age of the control group is 61, and the Parkinson group is 64. All the control individuals are older than 60 and the only Parkinson individual with more than 70 years and 2.5 on HY scale is the outlier of this group: pk-82-2.5. We consider one element wrong classified in thirteen

In both clusters 1 and 3, there are people with Parkinson and only one healthy individual. In this case, the control individual is older than 65; therefore, he can be aggregated with Parkinson. Hence we consider all four individuals are well classified.

Cluster 4 has people almost without any movement disorder. There are two people with Parkinson, one that has 36 years and other with 74. The pk-36-2 is perfectly integrated into this cluster.

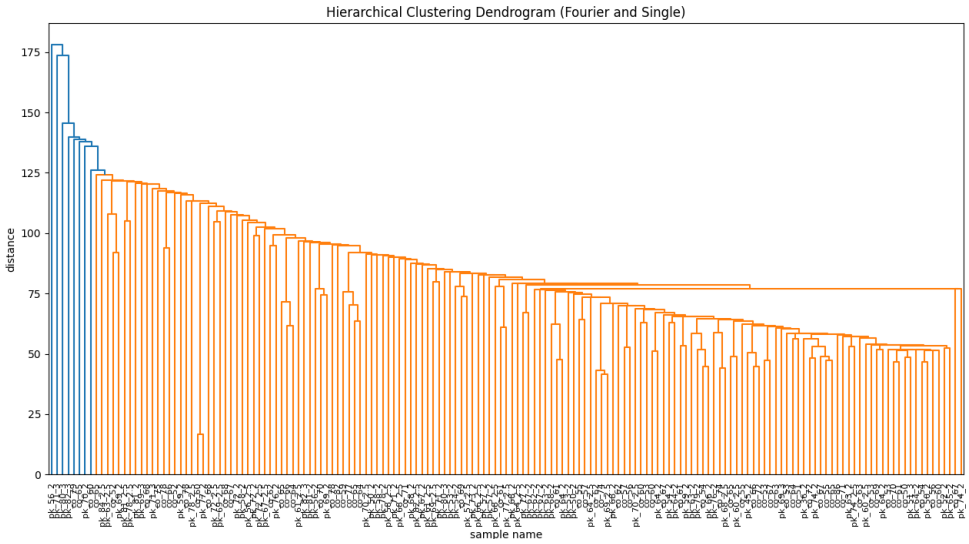


Figure 4.63: Fourier: Single

Cluster 1 number of entries 154	Cluster 6 number of entries 1
co_66 co_74 co_69 co_86 co_75 co_82 co_78 co_78 co_69 co_70	co_79
co_65 co_67 co_68 co_67 co_63 co_67 co_78 co_71 co_66 co_68	Cluster 5 number of entries 1
co_55 co_74 co_62 co_77 co_65 co_72 co_72 co_61 co_55 co_53	co_65
co_66 co_64 co_59 co_67 co_60 co_62 co_68 co_64 co_53 co_60	Cluster 2 number of entries 1
co_53 co_65 co_71 co_55 co_70 co_57 co_65 co_56 co_50 co_60	co_65
co_61 co_52 co_54 co_64 co_61 co_51 co_62 co_56 co_52 co_53	Cluster 3 number of entries 1
co_54 co_57 co_60 co_67 co_61 co_53 co_63 pk_68_2.5 pk_82_2.5	co_60
pk_72_2 pk_68_2 pk_69_3 pk_59_2 pk_81_2 pk_79_2 pk_71_3	Cluster 7 number of entries 1
pk_76_2 pk_76_2 pk_81_2 pk_80_3 pk_78_2 pk_68_2.5 pk_81_2.5	pk_82_3
pk_78_2.5 pk_72_2 pk_61_2.5 pk_68_2 pk_69_2 pk_67_2.5 pk_63_2	Cluster 4 number of entries 1
pk_63_2.5 pk_77_2 pk_72_2.5 pk_74_2.5 pk_70_2.5 pk_74_2	pk_70_2
pk_78_2.5 pk_51_3 pk_54_2 pk_61_2.5 pk_69_3 pk_74_2 pk_58_2.5	Cluster 10 number of entries 1
pk_67_2 pk_64_2.5 pk_64_2 pk_64_2.5 pk_82_3 pk_68_2 pk_64_2	pk_56_2
pk_84_2.5 pk_69_2.5 pk_68_2 pk_68_2.5 pk_45_2 pk_66_2 pk_58_2	Cluster 9 number of entries 1
pk_57_2.5 pk_69_2.5 pk_77_2.5 pk_73_2 pk_56_2.5 pk_54_2 pk_67_2	pk_71_3
pk_63_2 pk_77_2 pk_64_2 pk_52_2 pk_60_2.5 pk_62_2 pk_50_2	Cluster 8 number of entries 1
pk_64_2 pk_57_2 pk_53_2 pk_59_2 pk_58_2 pk_71_2 pk_50_2	pk_80_3
pk_36_2 pk_63_2 pk_50_2 pk_70_2.5 pk_72_2.5 pk_60_2.5 pk_56_2	
pk_64_2 pk_67_2 pk_71_2 pk_57_2 pk_53_2 pk_66_2.5 pk_65_2	
pk_69_2 pk_69_2.5	

Figure 4.64: Fourier: Single

So, the outlier is pk-74-2. We consider one element wrong classified in seven.

Cluster 5 has subjects with median difficulty in walking. Healthy individuals with age inferior to 60 and Parkinson individuals with age superior to 70 are wrong classified. The individual with 51 years and three on HY scale is also considered wrong classified because he is in an advanced stage of the disease. The outliers are: pk-77-2, pk-74-2.5, pk-74-2, pk-73-2, pk-77-2, pk-80-3, pk-51-3, co-50, co-54, co-57, co-55, co-59, co, 55 and co-53. We consider thirteen element wrong classified in sixty entries.

Cluster 6 has one only healthy subject, and he is 70 years old. The outlier seams to be pk-56-2 because he is young and in an early stage of the disease. We consider one element wrong classified in five.

Cluster 7 has a high average age in general and represents a group with severe movement disorders. The younger in the control group has 67 years. We consider ll of the individuals are well classified.

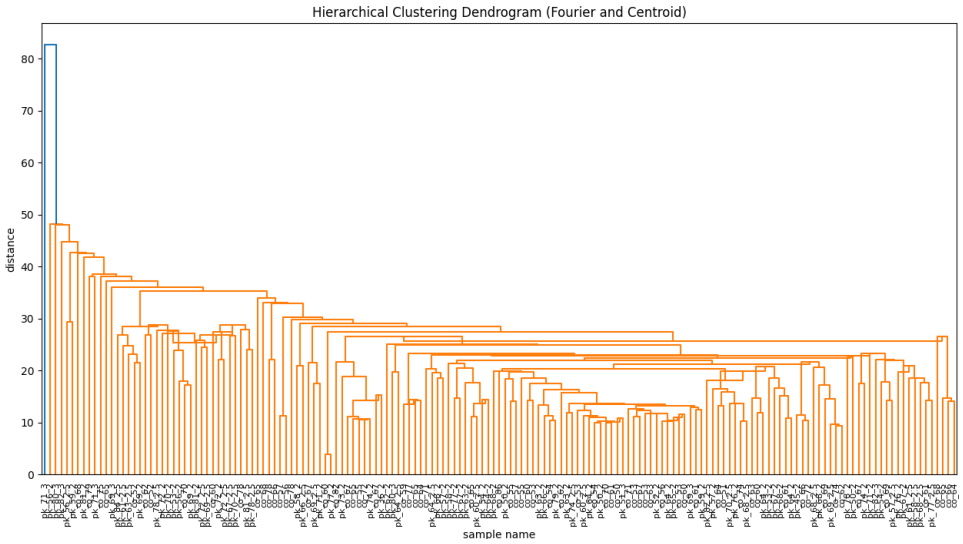


Figure 4.65: Fourier: Centroid

Cluster 3 number of entries 152	Cluster 4 number of entries 1
co_66 co_74 co_69 co_86 co_82 co_78 co_78 co_69 co_70 co_65	co_75
co_67 co_67 co_63 co_67 co_65 co_78 co_71 co_66 co_68 co_55	Cluster 2 number of entries 2
co_74 co_62 co_77 co_65 co_72 co_72 co_61 co_55 co_53 co_66	co_79 pk_71_3
co_65 co_64 co_59 co_67 co_60 co_62 co_68 co_64 co_53 co_60	Cluster 6 number of entries 1
co_53 co_65 co_71 co_55 co_70 co_57 co_65 co_56 co_50 co_60	co_68
co_61 co_52 co_54 co_60 co_64 co_61 co_51 co_62 co_56 co_52	Cluster 7 number of entries 1
co_53 co_54 co_57 co_60 co_67 co_61 co_53 co_63 pk_68_2.5	pk_82_3
pk_82_2.5 pk_72_2 pk_68_2 pk_69_3 pk_59_2 pk_70_2 pk_79_2	Cluster 9 number of entries 1
pk_76_2 pk_76_2 pk_81_2 pk_80_3 pk_78_2 pk_68_2.5 pk_81_2.5	pk_56_2
pk_78_2.5 pk_72_2 pk_61_2.5 pk_68_2 pk_69_2 pk_67_2.5 pk_63_2	Cluster 5 number of entries 1
pk_63_2.5 pk_77_2 pk_72_2.5 pk_74_2.5 pk_70_2.5 pk_74_2	pk_81_2
pk_78_2.5 pk_51_3 pk_54_2 pk_61_2.5 pk_69_3 pk_74_2 pk_58_2.5	Cluster 10 number of entries 1
pk_67_2 pk_64_2.5 pk_64_2 pk_64_2.5 pk_82_3 pk_68_2 pk_64_2	pk_71_3
pk_84_2.5 pk_69_2.5 pk_68_2 pk_68_2.5 pk_45_2 pk_66_2 pk_58_2	Cluster 8 number of entries 1
pk_57_2.5 pk_69_2.5 pk_77_2.5 pk_73_2 pk_54_2 pk_67_2 pk_63_2	pk_80_3
pk_77_2 pk_64_2 pk_52_2 pk_60_2.5 pk_62_2 pk_50_2 pk_64_2	Cluster 1 number of entries 2
pk_57_2 pk_53_2 pk_58_2 pk_71_2 pk_50_2 pk_36_2 pk_63_2	pk_56_2.5 pk_59_2
pk_50_2 pk_70_2.5 pk_72_2.5 pk_60_2.5 pk_56_2 pk_64_2 pk_67_2	
pk_71_2 pk_57_2 pk_53_2 pk_66_2.5 pk_65_2 pk_69_2 pk_69_2.5	

Figure 4.66: Fourier: Centroid

Cluster 8 only has one element, an individual with Parkinson with 56 years and an early stage of the disease.

Cluster 9 has subjects with median difficulty in walking. It was considered wrong classified healthy individuals with age inferior to 60 and Parkinson individuals with age superior to 70 or in an advanced disease stage. The outliers are: pk-79-2, pk-76-2, pk-72-2.5, pk-81-2, pk-78-2, pk-69-3, pk-69-3, pk-70-2.5, co-53, co-56, co-53, co-53, co-53, co-54, co-55 and co-51. We consider sixteen elements wrong classified in forty-two.

Cluster 10 has a high average age for the Parkinson group. In this way, it represents a group with severe difficulties in walking. The individuals with Parkinson younger than 60 and in an early stage do not fit in this cluster. The younger individual with Parkinson has 58 and is in stage 2.5 of the disease, so he is not wrong classified. Control subjects with age inferior to 65 do not belong here, such as co-65, co-60, co-62, co-52 and co-61. We consider five elements wrong classified in twenty-two.

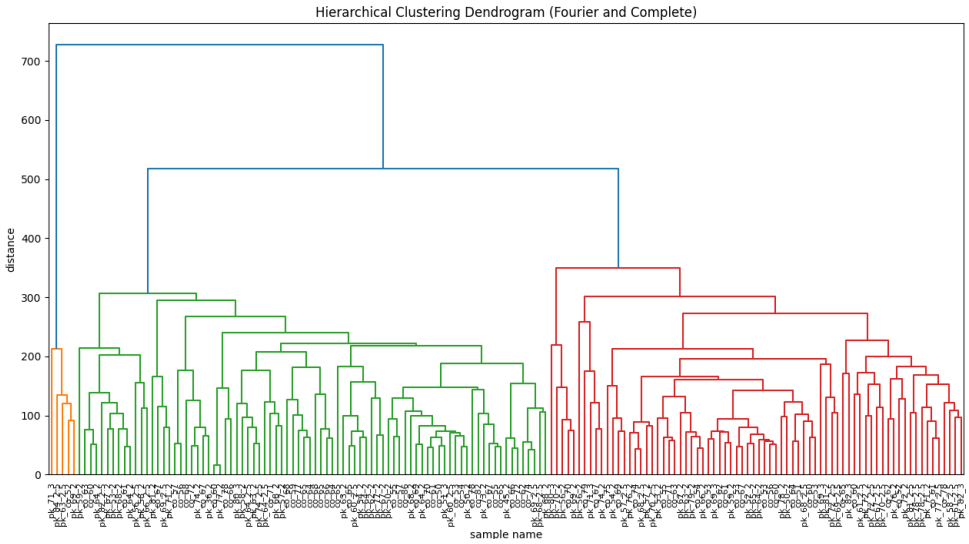


Figure 4.67: Fourier: Complete

Cluster 5 number of entries 60	Cluster 9 number of entries 42
co_66 co_74 co_86 co_82 co_78 co_78 co_69	co_69 co_75 co_71 co_74 co_62 co_61 co_55
co_65 co_67 co_66 co_68 co_55 co_77 co_65	co_53 co_60 co_53 co_53 co_54 co_64 co_51
co_72 co_66 co_65 co_64 co_59 co_62 co_68	co_56 co_53 co_60 co_67 co_63 pk_68_2.5 pk_69_3
co_64 co_60 co_65 co_71 co_55 co_70 co_65	pk_59_2 pk_79_2 pk_76_2 pk_81_2 pk_78_2 pk_54_2 pk_69_3
co_50 co_61 co_54 co_57 co_53 pk_68_2 pk_80_3	pk_67_2 pk_68_2.5 pk_66_2 pk_57_2.5 pk_63_2 pk_52_2 pk_57_2
pk_68_2.5 pk_63_2 pk_77_2 pk_74_2.5 pk_51_3 pk_64_2.5 pk_64_2.5	pk_50_2 pk_70_2.5 pk_72_2.5 pk_64_2 pk_53_2 pk_65_2 pk_69_2.5
pk_68_2 pk_64_2 pk_69_2.5 pk_68_2 pk_45_2 pk_58_2 pk_73_2	Cluster 10 number of entries 22
pk_54_2 pk_77_2 pk_64_2 pk_60_2.5 pk_62_2 pk_50_2 pk_50_2	co_65 co_78 co_60 co_62 co_52 co_61 pk_82_3 pk_71_2
pk_63_2 pk_60_2.5 pk_67_2 pk_57_2	pk_72_2 pk_76_2 pk_81_2.5 pk_78_2.5 pk_72_2 pk_61_2.5 pk_67_2.5
Cluster 2 number of entries 13	pk_72_2.5 pk_70_2.5 pk_78_2.5 pk_61_2.5 pk_58_2.5 pk_82_3 pk_77_2.5
co_63 co_60 co_61 pk_82_2.5 pk_68_2 pk_64_2 pk_56_2.5	Cluster 7 number of entries 4
pk_67_2 pk_64_2 pk_53_2 pk_59_2 pk_58_2 pk_66_2.5	co_79 co_67 pk_71_3 pk_74_2
	Cluster 6 number of entries 5
Cluster 3 number of entries 4	co_70 pk_70_2 pk_80_3 pk_56_2 pk_69_2
co_67 pk_81_2 pk_69_2.5 pk_71_2	Cluster 4 number of entries 7
Cluster 1 number of entries 5	co_68 co_67 co_72 co_57 co_56 pk_74_2 pk_36_2
co_52 pk_71_3 pk_69_2 pk_63_2.5 pk_84_2.5	Cluster 8 number of entries 1
	pk_56_2

Figure 4.68: Fourier: Complete

4.6.26 Fourier distance- Median

Cluster 1 has majority Parkinson subjects. In this way, it represents a group with severe difficulties in walking. We consider that the outliers are Control elements with age inferior to 65. We admit six elements wrong classified in twenty-four entries.

Cluster 3 represents a group with severe movement disorders, and all the elements are consider well classified.

Cluster 4 has majority Control subjects. The element with Parkinson and in stage 3 of the disease and 80 years, do not fit in this group. Neither the other elements with Parkinson older than 70. However the element with Parkinson with 51 years and stage 3 of the diseases can fit this group because of his age. We consider six elements wrong classified in sixty-seven.

Cluster 6 represents majority Parkinson subjects. It was considered wrong classified healthy individuals with age inferior to 60 and Parkinson individuals with age superior to 70 or in an

advanced disease stage.In this way, we admit twenty-two elements wrong classified in sixty-three entries.

The other clusters have one element each.

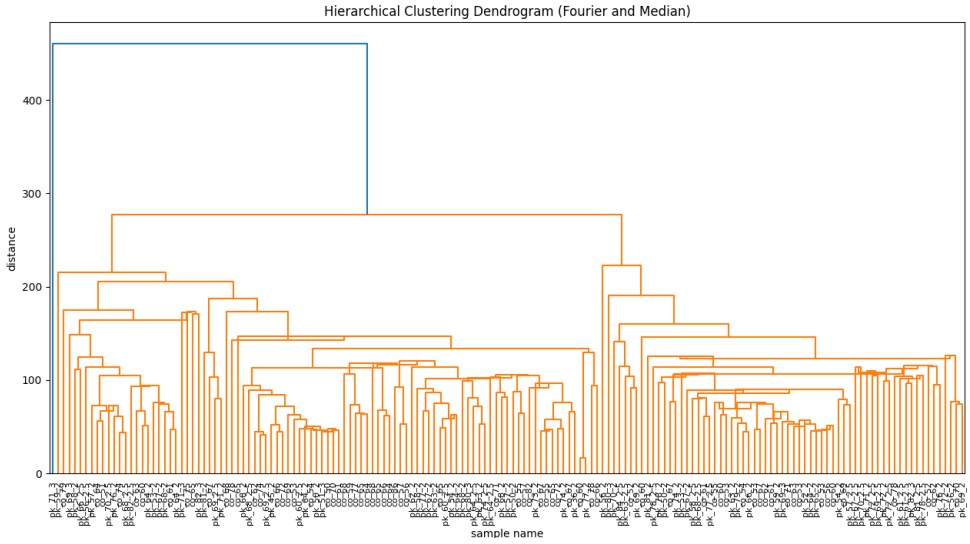


Figure 4.69: Fourier: Median

Cluster 4 number of entries 67	Cluster 6 number of entries 63
co_66 co_74 co_86 co_82 co_78 co_78 co_69	co_69 co_70 co_67 co_78 co_71 co_62 co_61
co_65 co_68 co_67 co_67 co_66 co_68 co_55	co_55 co_53 co_60 co_53 co_53 co_52 co_54
co_77 co_65 co_72 co_72 co_66 co_65 co_64	co_60 co_62 co_56 co_52 co_53 co_60 co_67
co_59 co_62 co_68 co_64 co_60 co_65 co_71	co_61 co_63 pk_68_2.5 pk_72_2 pk_59_2 pk_70_2 pk_79_2
co_55 co_70 co_57 co_65 co_56 co_50 co_61	pk_76_2 pk_81_2 pk_78_2 pk_81_2.5 pk_78_2.5 pk_72_2 pk_61_2.5
co_54 co_57 co_53 pk_68_2 pk_80_3 pk_68_2.5 pk_63_2	pk_69_2 pk_67_2.5 pk_63_2.5 pk_72_2.5 pk_70_2.5 pk_78_2.5 pk_54_2
pk_77_2 pk_74_2.5 pk_74_2 pk_51_3 pk_64_2.5 pk_64_2.5 pk_68_2	pk_61_2.5 pk_69_3 pk_74_2 pk_58_2.5 pk_67_2 pk_82_3 pk_84_2.5
pk_64_2 pk_69_2.5 pk_68_2 pk_45_2 pk_58_2 pk_73_2 pk_54_2	pk_66_2 pk_57_2.5 pk_77_2.5 pk_63_2 pk_52_2 pk_71_2 pk_50_2
pk_77_2 pk_64_2 pk_60_2.5 pk_62_2 pk_50_2 pk_50_2 pk_36_2	pk_72_2.5 pk_56_2 pk_64_2 pk_53_2 pk_65_2 pk_69_2 pk_69_2.5
pk_63_2 pk_60_2.5 pk_67_2 pk_57_2	Cluster 1 number of entries 24
Cluster 2 number of entries 1	co_75 co_63 co_65 co_74 co_60 co_64 co_61
co_79	co_51 pk_82_3 pk_82_2.5 pk_69_3 pk_71_3 pk_76_2 pk_68_2
Cluster 3 number of entries 4	pk_64_2 pk_68_2.5 pk_56_2.5 pk_67_2 pk_64_2 pk_57_2 pk_53_2
co_67 pk_81_2 pk_69_2.5 pk_71_2	pk_58_2 pk_70_2.5 pk_66_2.5
Cluster 8 number of entries 1	Cluster 9 number of entries 1
pk_56_2	pk_71_3
Cluster 5 number of entries 1	Cluster 7 number of entries 1
pk_59_2	pk_80_3

Figure 4.70: Fourier: Median

4.6.27 Fourier distance- Ward

Cluster 1 has majority Parkinson subjects. It represents a group with severe difficulties in walking, though, the only healthy subject with 52 years do not belong in this group. We consider one element wrong classified in five elements.

Cluster 2, like cluster 6 from Fourier distance-Complete, has just one healthy subject with 70 years old. The outlier is the same as this cluster: pk-56-2. We consider one element wrong classified in five entries.

Cluster 3 has a high average age for the Parkinson group, presenting individuals with severe difficulties in walking. The outliers are controls with age inferior to 65:co-60, co-62, co-52, co-61; and Parkinson with age inferior to 60 and/or in an early disease stage: pk-54-2. Pk-58-2.5 and pk-57-2.5 are younger than 60 years but have stage 2.5 on HY scale, so they are not considered wrong classified. We consider five elements wrong classified in thrity-one.

Cluster 4 has healthy majority subjects but with a high average age, so represents gait problems. Control individuals with age inferior to 65 and Parkinson individuals younger than 60 years do not belong in this cluster. The outliers are co-55 and pk-36-2. We consider two elements wrong classified in twelve.

Cluster 5 represents subjects with severe problems in gait. The only control individual is 67 (>65), and the youngest Parkinson individual has 69 years. So, there is no wrong classified in this cluster.

Cluster 6 has the majority of healthy subjects. It represents people with median difficulties in walking. All the control individuals older than 60 and all Parkinson individual younger than 70 years fit in this group. So we admit that the outliers are: pk-80-3, pk-77-2, pk-74-2.5, co-59, co57, co-54, co-57, co-55 and co- 53. We consider nine elements wrong classified in forty-four entries.

Cluster 7 has subjects with median difficulty in walking. It was considered wrong classified control individuals with age inferior to 60 and Parkinson individuals with age superior to 70 or in an advanced disease stage. The outliers are: co-50, co-37, co-51, pk-82-2.5 and pk-76-2. We consider five elements wrong classified in twenty-two entries.

Cluster 8 represents a group with severe difficulties in walking, so controls with age inferior to 65 and Parkinson with age inferior to 60 and/or in an early disease stage are not right classified. So we consider that the outlier is pk-56-2.

Cluster 9 has almost only Parkinson subjects. The outlier is the co-56 because he is too young to be aggregated with Parkinson with average age 63. On the other hand, pk-77- 2 is too old to have median gait problems, and he should have severe. We consider that the outliers are co-56 and pk-77-2. We admit two elements wrong classified in eight total elements.

Cluster 10 has people with median difficulties in gait. Control younger than 60 years are too young and Parkinson people with more than 70 years are too old to be a part of this cluster. The outliers are: co-54, co-56, co-55, co-53, co-53, pk-79-2 and pk-78-2. We consider seven elements wrong classified in twenty-seven entries.

4.6.28 Fourier distance- Weighted

Cluster 1 has Parkinson individuals with a high average age. The only control is very young and does not fit the other's gait problems. So, co-52 is wrong classified. We consider one element wrong classified in five.

Cluster 2 has one element: a Parkinson individual, old and in an advanced stage of Parkinson disease. (pk-71-3)

Cluster 3 has the majority of healthy individuals. The control group has a high average age, so these clusters have elements with severe gait problems. The younger control is 36 years so is too

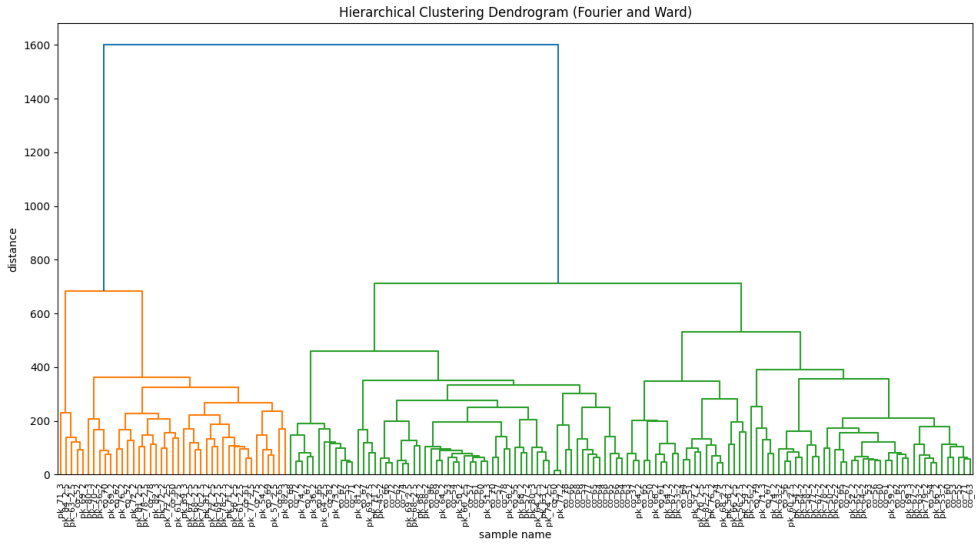


Figure 4.71: Fourier: Ward

Cluster 6 number of entries 44 co_66 co_74 co_86 co_78 co_78 co_69 co_65 co_66 co_68 co_77 co_65 co_72 co_66 co_64 co_59 co_62 co_68 co_64 co_60 co_55 co_70 co_57 co_60 co_61 co_54 co_57 co_53 pk_68_2 pk_80_3 pk_68_2.5 pk_63_2 pk_77_2 pk_74_2.5 pk_51_3 pk_64_2.5 pk_68_2 pk_64_2 pk_69_2.5 pk_68_2 pk_45_2 pk_60_2.5 pk_50_2 pk_50_2 pk_57_2	Cluster 3 number of entries 31 co_69 co_75 co_65 co_78 co_60 co_62 co_52 co_61 pk_82_3 pk_68_2.5 pk_72_2 pk_69_3 pk_76_2 pk_81_2 pk_81_2.5 pk_78_2.5 pk_72_2 pk_61_2.5 pk_67_2.5 pk_72_2.5 pk_70_2.5 pk_78_2.5 pk_54_2 pk_61_2.5 pk_58_2.5 pk_82_3 pk_57_2.5 pk_77_2.5 pk_71_2 pk_72_2.5 pk_69_2.5
Cluster 4 number of entries 12 co_82 co_68 co_67 co_67 co_55 co_72 co_65 co_71 pk_74_2 pk_64_2.5 pk_73_2 pk_36_2	Cluster 10 number of entries 27 co_71 co_62 co_61 co_55 co_53 co_60 co_53 co_65 co_54 co_56 co_53 co_60 co_67 co_63 pk_59_2 pk_79_2 pk_78_2 pk_69_3 pk_67_2 pk_66_2 pk_63_2 pk_52_2 pk_62_2 pk_50_2 pk_64_2 pk_53_2 pk_65_2
Cluster 7 number of entries 22 co_63 co_74 co_65 co_50 co_37 co_64 co_61 co_51 pk_82_2.5 pk_76_2 pk_68_2 pk_64_2 pk_68_2.5 pk_56_2.5 pk_67_2 pk_64_2 pk_57_2 pk_53_2 pk_59_2 pk_58_2 pk_70_2.5 pk_66_2.5	Cluster 8 number of entries 5 co_79 co_67 pk_56_2 pk_71_3 pk_74_2
Cluster 2 number of entries 5 co_70 pk_70_2 pk_80_3 pk_56_2 pk_69_2	Cluster 5 number of entries 4 co_67 pk_81_2 pk_69_2.5 pk_71_2
Cluster 1 number of entries 5 co_52 pk_71_3 pk_69_2 pk_63_2.5 pk_84_2.5	Cluster 9 number of entries 8 co_56 pk_58_2 pk_54_2 pk_77_2 pk_64_2 pk_63_2 pk_60_2.5 pk_67_2

Figure 4.72: Fourier: Ward

young to be compared to a control group with an average age 68, so he does not belong here. The younger on the control group do not belong here either because he has only 55 (<65). We consider two elements wrong classified in eleven.

Cluster 4 has a high average age in general and represents a group with severe movement disorders. The younger in the control group has 67 years. We consider that all of the individuals are well classified. Note that it is very similar to cluster 7 on DTW complete.

Cluster 5 has subjects with median difficulty in walking. Healthy individuals with age inferior to 60 and Parkinson individuals with age superior to 70 are wrong classified. So, we consider the outliers are co-59,co-53, co-50, co-57, pk-73-2, pk-80-3, pk-77-2, co-55, co-55, co-54, co-53, co-53, pk-78-2, pk-77-2, co-53, co-57, co-56, co-53, co-56, co-54 and pk-74-2.5. We admit this cluster has twenty-one elements wrong classified in eighty-eight.

Cluster 6 has people with median difficulties in walking, just like cluster 5. So we consider the outliers are: co-51, pk-82-2.5, pk-76-2 and pk-70-2.5. We consider this cluster has four elements wrong classified in fifteen.

Cluster 7 represents a group with severe difficulties in walking, so controls with age inferior to 65 and Parkinson with age inferior to 60 and/or in the early stage of the disease are not right classified. So we consider the outlier is pk-58-2. We consider this cluster has one element wrong classified in seven.

Cluster 8 has a Parkinson group with high average age, and it is a group majority constituent by Parkinson elements. This group represents a severe gait disorder. So, the outliers are co-52, co-61, co-62 and pk-56-2. We consider this cluster has four elements wrong classified in twenty-eight.

Cluster 9 represents a group with severe difficulties in walking. The younger control is older than 65, so it is considered right classified. We consider this cluster has no wrongs classified elements.

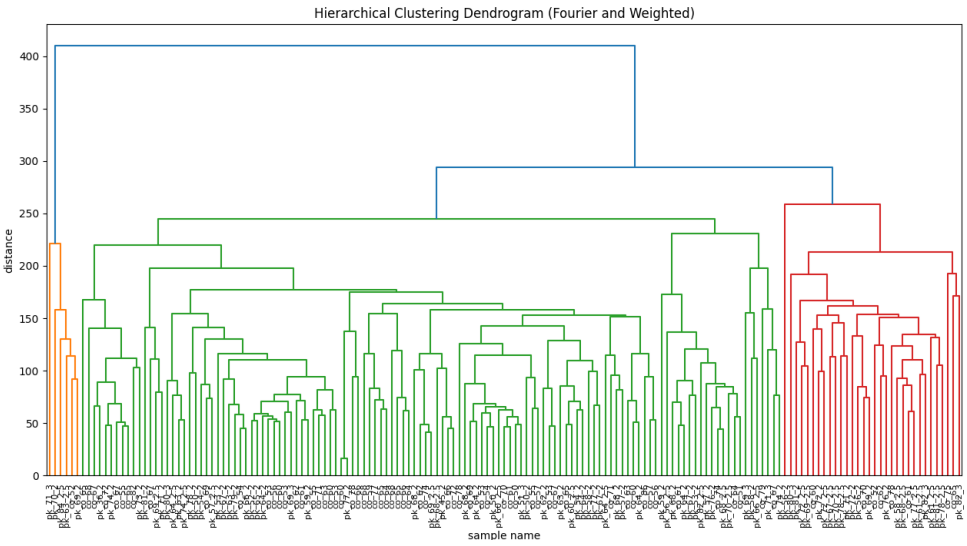


Figure 4.73: Fourier: Weighted

Cluster 5 number of entries 88	Cluster 8 number of entries 28
co_66 co_74 co_69 co_86 co_78 co_78 co_69	co_70 co_78 co_60 co_62 co_52 co_61 pk_68_2.5
co_65 co_63 co_71 co_66 co_68 co_62 co_77	pk_72_2 pk_76_2 pk_81_2 pk_81_2.5 pk_78_2.5 pk_72_2 pk_61_2.5
co_65 co_72 co_61 co_55 co_53 co_66 co_64	pk_67_2.5 pk_72_2.5 pk_70_2.5 pk_78_2.5 pk_61_2.5 pk_58_2.5 pk_80_3
co_59 co_60 co_62 co_68 co_64 co_53 co_60	pk_82_3 pk_77_2.5 pk_71_2 pk_72_2.5 pk_56_2 pk_69_2 pk_69_2.5
co_53 co_71 co_55 co_70 co_57 co_65 co_56	Cluster 1 number of entries 5
co_50 co_60 co_61 co_54 co_56 co_53 co_54	co_52 pk_70_2 pk_69_2 pk_63_2.5 pk_84_2.5
co_57 co_60 co_67 co_53 co_63 pk_68_2 pk_59_2	Cluster 6 number of entries 15
pk_79_2 pk_80_3 pk_78_2 pk_68_2.5 pk_63_2 pk_77_2 pk_74_2.5	co_74 co_64 co_61 co_51 pk_82_2.5 pk_76_2 pk_68_2
pk_51_3 pk_54_2 pk_69_3 pk_67_2 pk_64_2.5 pk_64_2 pk_64_2.5	pk_68_2.5 pk_56_2.5 pk_67_2 pk_64_2 pk_57_2 pk_53_2 pk_59_2
pk_68_2 pk_64_2 pk_69_2.5 pk_68_2 pk_45_2 pk_66_2 pk_58_2	pk_70_2.5
pk_57_2.5 pk_54_2 pk_63_2 pk_77_2 pk_64_2 pk_52_2 pk_60_2.5	Cluster 7 number of entries 7
pk_62_2 pk_50_2 pk_50_2 pk_63_2 pk_50_2 pk_60_2.5 pk_64_2	co_79 co_67 pk_69_3 pk_71_3 pk_74_2 pk_58_2 pk_66_2.5
pk_67_2 pk_57_2 pk_53_2 pk_65_2	Cluster 4 number of entries 4
Cluster 9 number of entries 3	co_67 pk_81_2 pk_69_2.5 pk_71_2
co_75 co_65 pk_82_3	Cluster 10 number of entries 1
Cluster 3 number of entries 11	pk_56_2
co_82 co_68 co_67 co_67 co_55 co_72 co_65	Cluster 2 number of entries 1
co_65 pk_74_2 pk_73_2 pk_36_2	pk_71_3

Figure 4.74: Fourier: Weighted

4.7 Cluster Evaluation

As seen on 3, there are different evaluation methods. The most frequently used are the external criteria, such as purity. Purity has a value close to 0 when the clustering is worse and a value of 1 when it is perfect.

In order to evaluate the clusters, purity is the measure chosen to use As we can see, some elements are alone in a cluster. So this specific cluster is considered pure. Nevertheless, we cannot consider that the algorithm is perfect if each element is allocated to a different cluster. In this way, to control this problem, we have chosen to remove the clusters with just one element from the purity counting.

To complement the purity, we also used the dendrograms to evaluate the quality of each distance and aggregation measure. If the dendrograms show lots of elements alone in a cluster, the method is worse.

By looking at the table 4.1 and the graphic 4.75, we can see that the method that achieves the best purity is the Manhattan distance using Weighted aggregation measure, with a purity percentage of 82,802%. However, the distance that presents the best results in general is the DTW distance, with a purity mean of 74,5%. The aggregation measure that presents the best results is the Weighted, with a purity mean of 78,4%.

The distance that do not have any cluster with just one element are the Manhattan with the Ward aggregation measure and the Euclidean with the Ward aggregation measure. This distances combined with the ward aggregation measures shows promising results.

Table 4.1: Purity Percentage of *Gait in Parkinson's Disease Dataset*

	Euclidean	**DTW**	**Manhattan**	**Fourier**
Average	75,2%	75,0%	71,5%	72,5%
Centroid	69,4%	75,3%	72,2%	64,3%
Complete	75,9%	76,5%	75,2%	75,6%
Median	70,4%	67,9%	65,0%	77,4%
Single	69,0%	69,0%	64,1%	65,20%
Ward	77,3%	79,1%	78,7%	78,7%
Weighted	73,1%	78,8%	82,8%	76,2%

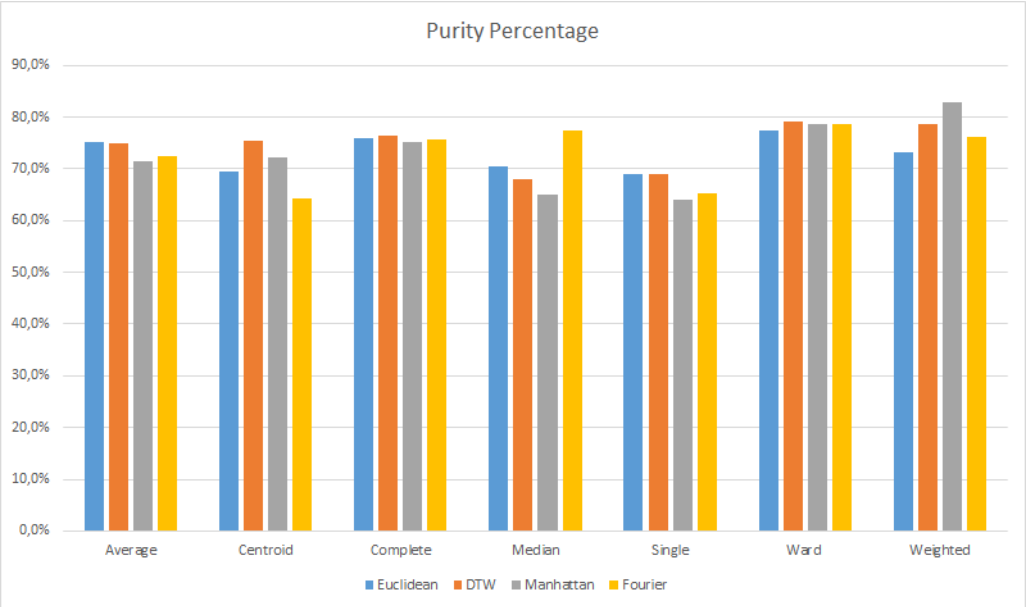


Figure 4.75: Purity Percentage

Chapter 5

Gait in Neurodegenerative Disease

In this chapter, the Dataset *Gait in Neurodegenerative Disease* is analysed. In the first section, a description of the data is done, and it is explained how it is obtained. The second section specifies the dataset's demographic data, such as age, gender, and body weight. The third section analyses the scales that were chosen to measure the diseases' severity. In the final section, gait patterns of individuals with different neurodegenerative diseases are analysed.

5.1 Gait in Neurodegenerative Disease Dataset

The second dataset [54], named *Gait in Neurodegenerative Disease dataset*, has time series obtained during a walking test of 15 patients with Parkinson's disease, 20 with Huntington's disease, 13 with amyotrophic lateral sclerosis and 16 healthy patients for control. Each patient was asked to walk 77 meters for 5 minutes non-stop. In this way, force-sensitive resistors (FSR) were introduced into the patients' shoes. This sensors change their resistance when subjected to pressure; thus, we obtained the time series to be analysed.

It was provided the demographic data of each individual, such as age, gender, height, weight, walking speed, and a measure of disease severity or duration.

In order to make the individuals characteristics clear, we change subjects names to this format:

1. If is Control, the format is control - age; for example, control-64 is a healthy individual with 64 years.

2. If has ALS, the format is als - age - time in months since the diagnosis; for example, als-63-14 is an individual with ALS, with 63 years with 14 months since the diagnosis.

3. If has Huntington, the format is hunt - age - total functional capacity; for example, hunt-90-12 is an individual with Huntington disease, with 90 years and 12 of total functional capacity.

4. If has Parkinson, the format is park - age - stage of disease on HY scale; for example, park-70-2 is an individual with Parkinson disease, with 70 years and stage 2 on HY scale.

5.2 Demographic data

Figure 5.1 shows the gender of the individuals in the study. The control group has 14 female and two male subjects. The group with Huntington disease has the same number of females as the control group and has six males. Parkinson's group and ALS's group have fewer females than males. Parkinson's group has five females, and ten males and ALS's group has three females and ten males.

The figure 5.2 represents the age of each group. The median of the control group is 35 years. The maximum age is 74, and the minimum is 20 years. By looking at the box plot for Parkinson's group, we can see that the median is 68 years. The maximum age is 80, and the minimum is 44. The group of individuals with Huntington disease has a median age of 44,5, and its minimum value and maximum are 29 and 71, respectively. Lastly, ALS's group has a median age of 62. The minimum and maximum values are 36 and 70 years.

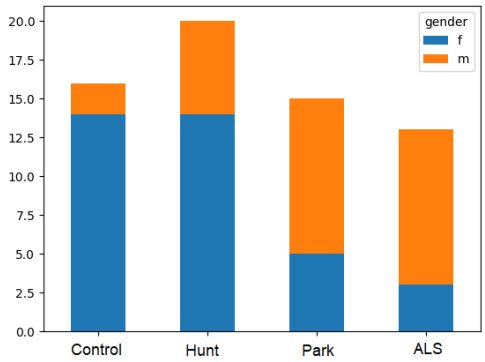


Figure 5.1: Gender of each group

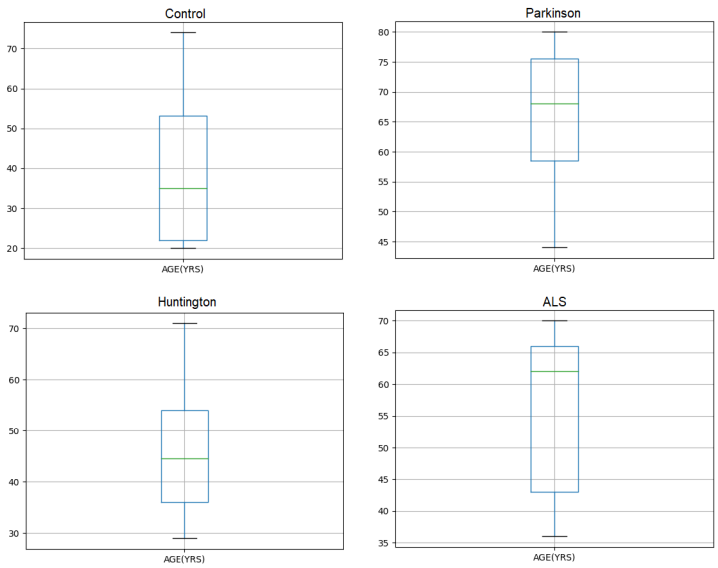


Figure 5.2: Age of each group

5.3 Severity of the diseases

For Parkinson individuals, we have the same scale used before, the Hohn and Yahr scale. In this case, a higher value indicates a more severe stage of the disease.

On the other hand, to measure ALS individuals' severity, it is used the time in months since the diagnosis of the disease.

Finally, to measure Huntington individuals' severity, it was used the total functional capacity. In this way, a lower number means more advanced functional impairment. Figure 5.3 shows the distribution of individuals in different stages of the diseases.

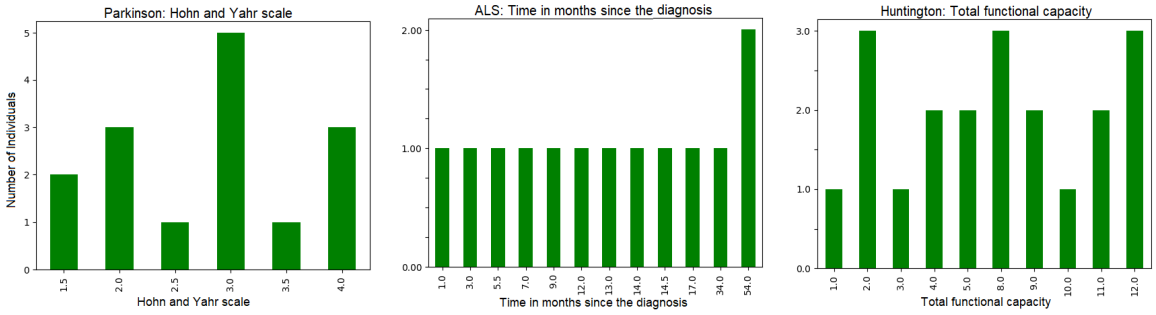


Figure 5.3: Severity of the disease

5.4 Gait Data

Unlike the first dataset, this one has already calculated the swing percentage. So, the data used to clustering was the Swing Percentage.

The Figures 5.4, 5.5, 5.6 and 5.7 shows the swing percentage of random individuals of each group. The group that shows more irregularity is the Huntington Group. The control group is very regular, as expected.

5.5 Classification models

The Hierarchical clustering was applied to this dataset.

For this dataset, it was tested the same distance measures as before. However, as seen on the Chapter 4, the aggregation measures Average, Centroid, Median and Complete present lots of clusters with just one element and achieve less purity percentage comparing to the other methods. In this way, it was chosen to only present Weighted and Ward aggregation measures. The other aggregation measures are present in the Appendix A.

The gait patterns were divided into three categories:

- The first one is good mobility. This includes people with Parkinson with age inferior to 60 years and in an early stage of disease; healthy individuals younger than 65 years; ALS

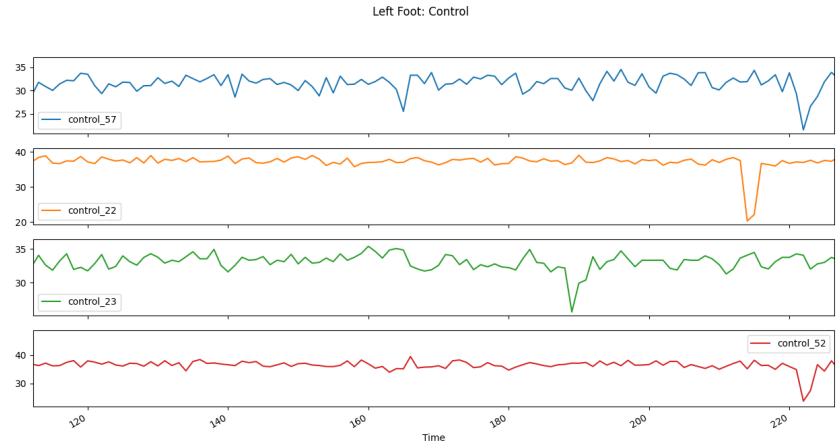


Figure 5.4: Swing Percentage: Control Group

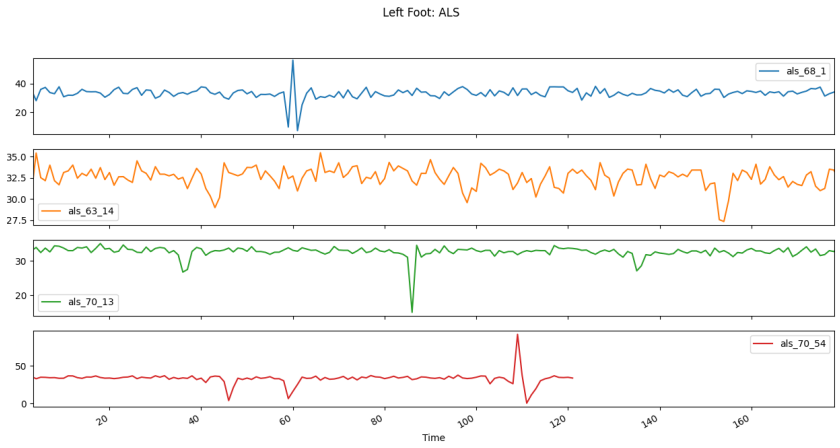


Figure 5.5: Swing Percentage: ALS group

individuals with less time since diagnosis; and subjects with Huntington motor function high.

- The second category is median mobility. This includes healthy people with age superior to 60 years; subjects with Parkinson not too old and in a not very high stage of the disease; ALS individuals with more than six months since diagnosis and not very old; and subjects with Huntington with median motor function and not too old.
- The third category involves people with the most severe difficulty in walking. It includes subjects with Parkinson in advanced age (older than 60 years) or/and advanced stage of the disease; healthy people older than 65, ALS individuals with more time since diagnosis; and subjects with Huntington with low motor function or old.

The ALS has not a very linear scale, so it is not easy to understand these elements' gait pattern.

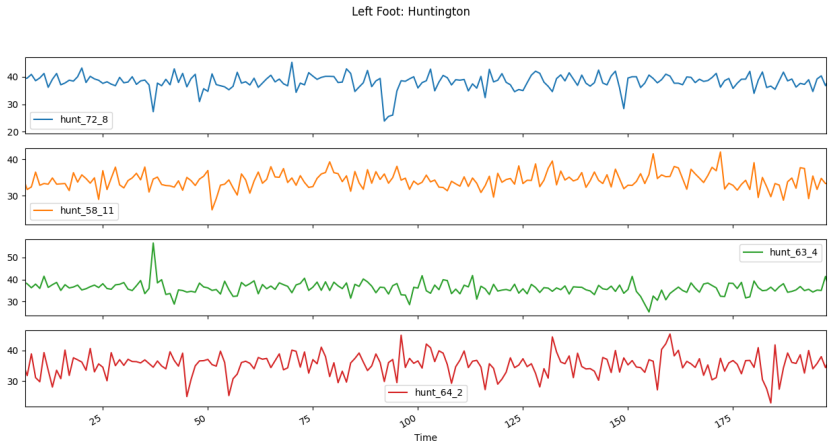


Figure 5.6: Swing Percentage: Huntington Group

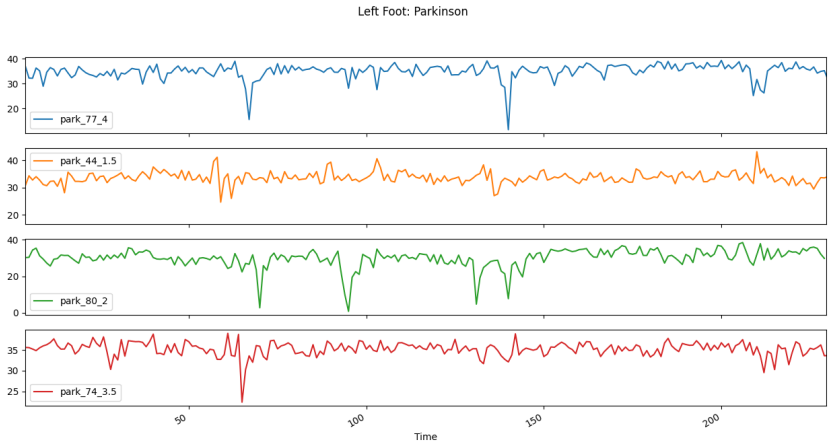


Figure 5.7: Swing Percentage: Parkinson Group

5.5.1 Manhattan distance- Ward

The dendrogram that illustrates this method is in the Figure 5.9. First, we have the cluster 9 that has seventeen entries, and we consider it has five elements wrong classified. These elements are: a people with ALS detect in a few ago months and young; a healthy young subject; a young patient with Huntington disease and with good motor function; and two elements with Parkinson in an early stage and young.

Cluster 8 has 11 entries and people from all the groups. The sick individuals are all well classified because they are in an advanced stage or have advanced age. The element wrong classified is the control element because he is too young.

Cluster 5 has 20 entries and 11 elements wrong classified. This cluster has 12 control elements, and we consider that the younger do not belong in this cluster. We consider that the two patients

with Huntington who are young or have a good motor function are wrong classified.

Cluster 4 has elements with Huntington and a control subject. We admit this Control is wrong classified.

Cluster 1 has people with ALS and Parkinson disease. All of them in advanced age or advanced stage of the disease, so we admit they are well classified.

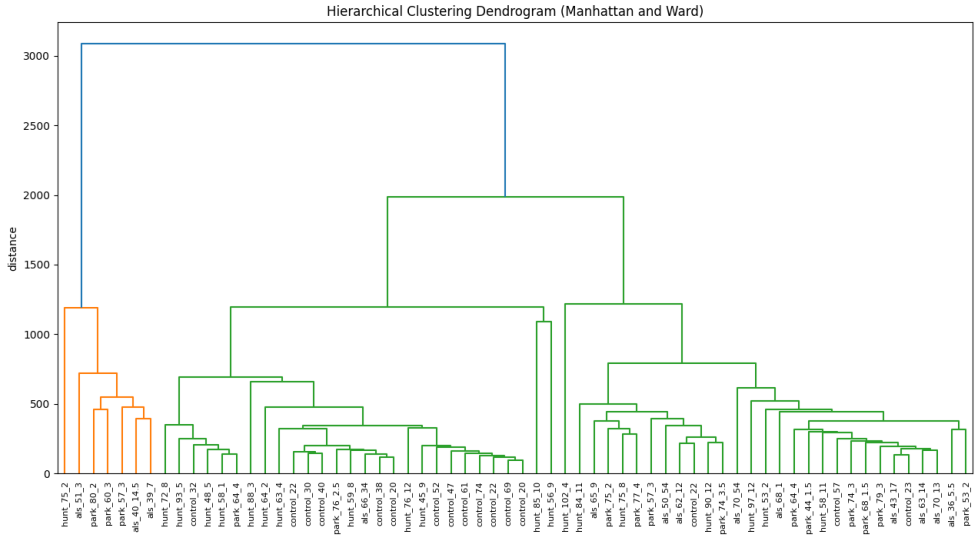


Figure 5.8: Manhattan: Ward

Cluster 9 number of entries 17 als_68_1 als_63_14 als_70_13 als_70_54 als_36_5.5 als_43_17 control_57 control_23 hunt_58_11 hunt_97_12 hunt_53_2 park_44_1.5 park_53_2 park_64_4 park_68_1.5 park_74_3 park_79_3	Cluster 5 number of entries 20 als_66_34 control_22 control_52 control_47 control_30 control_22 control_38 control_69 control_74 control_61 control_20 control_20 control_40 hunt_63_4 hunt_64_2 hunt_59_8 hunt_88_3 hunt_76_12 hunt_45_9 park_76_2.5
Cluster 8 number of entries 11 als_65_9 als_50_54 als_62_12 control_22 hunt_90_12 hunt_84_11 hunt_75_8 park_77_4 park_74_3.5 park_75_2 park_57_3	Cluster 4 number of entries 6 control_32 hunt_72_8 hunt_48_5 hunt_93_5 hunt_58_1 park_64_4
Cluster 2 number of entries 1 als_51_3	Cluster 6 number of entries 1 hunt_85_10
Cluster 1 number of entries 5 als_40_14.5 als_39_7 park_80_2 park_60_3 park_57_3	Cluster 3 number of entries 1 hunt_75_2
Cluster 10 number of entries 1 hunt_102_4	Cluster 7 number of entries 1 hunt_56_9

Figure 5.9: Manhattan: Ward

5.5.2 Manhattan distance- Weighted

Cluster 4 has 16 entries. The individual with ALS with 36 years and time since diagnosis equal to 5.5 months and the two Parkinson patients in an early stage of the disease are not well classified. The Control individual with 23 years and the individual with Huntington young and 11 motor capacity is not well classified either. So, we admit to have five elements wrong classified in sixteen.

Cluster 5 has ten elements control too young to be compared to people with neurodegenerative diseases. There is an individual with Huntington disease that is young and with an acceptable motor capacity that is wrong classified too. We consider eleven elements wrong classified in thirty-six entries.

Cluster 1 from weighted aggregation measure has the exact elements as the cluster 1 of the ward aggregation measure (Manhattan distance). We consider they are all well classified.

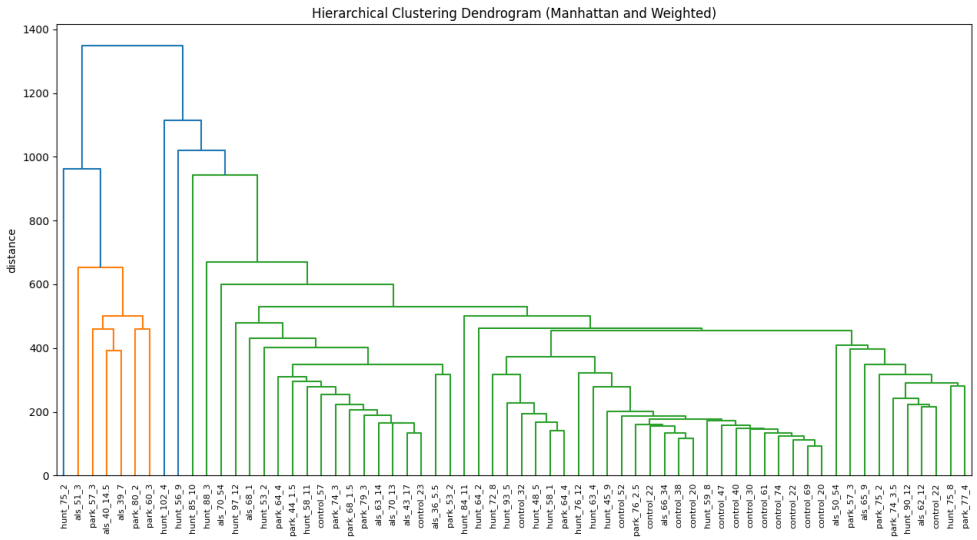


Figure 5.10: Manhattan: Weighted

Cluster 4 number of entries 16 als_68_1 als_63_14 als_70_13 als_36_5.5 als_43_17 control_57 control_23 hunt_58_11 hunt_97_12 hunt_53_2 park_44_1.5 park_53_2 park_64_4 park_68_1.5 park_74_3 park_79_3	Cluster 1 number of entries 5 als_40_14.5 als_39_7 park_80_2 park_60_3 park_57_3
Cluster 5 number of entries 36 als_65_9 als_50_54 als_62_12 als_66_34 control_22 control_52 control_47 control_30 control_22 control_22 control_32 control_38 control_69 control_74 control_61 control_20 control_20 control_40 hunt_72_8 hunt_63_4 hunt_64_2 hunt_59_8 hunt_90_12 hunt_84_11 hunt_75_8 hunt_48_5 hunt_93_5 hunt_76_12 hunt_58_1 hunt_45_9 park_77_4 park_74_3.5 park_75_2 park_64_4 park_57_3 park_76_2.5	Cluster 6 number of entries 1 als_70_54 Cluster 2 number of entries 1 als_51_3 Cluster 8 number of entries 1 hunt_85_10 Cluster 3 number of entries 1 hunt_75_2 Cluster 9 number of entries 1 hunt_56_9 Cluster 10 number of entries 1 hunt_102_4
Cluster 7 number of entries 1 hunt_88_3	

Figure 5.11: Manhattan: Weighted

5.5.3 Euclidean distance- Ward

Cluster 7 has almost the same elements the cluster 9 from the Manhattan distance using ward method. We have 14 entries and three elements wrong classified, one young subject with less time since the diagnosis, one young, healthy subject and one Parkinson element young and in stage 2 of the disease.

Cluster 6 has almost the same elements the cluster 8 from the Manhattan distance using ward method. The only element wrong classified is the control with 22 years.

Cluster 1 has no wrong classified elements and is equal to cluster 1 from the Manhattan distance using ward method.

Cluster 4 has 27 entries and ten control elements that are wrong classified because of their age. There is a subject with Huntington disease that is also wrong classified.

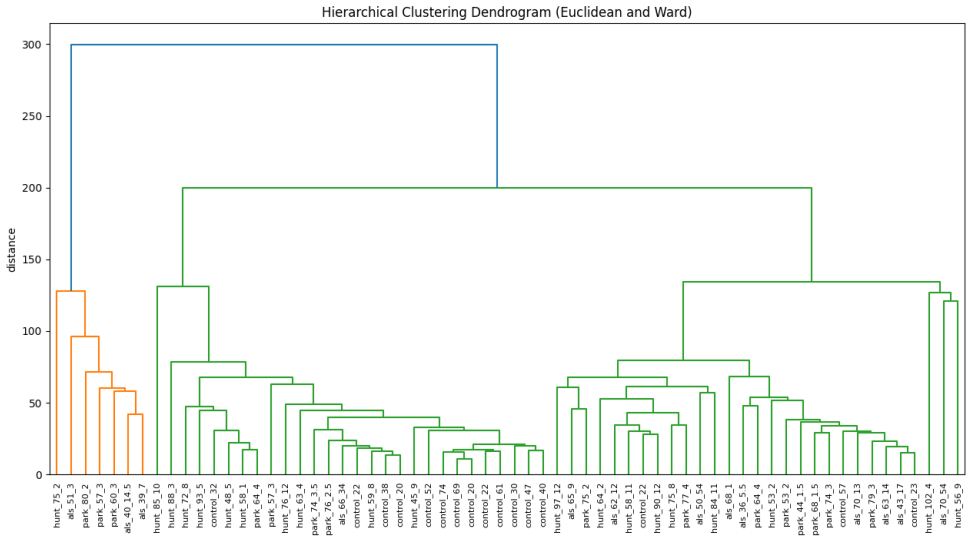


Figure 5.12: Euclidean: Ward

Cluster 4 number of entries 27 als_66_34 control_22 control_52 control_47 control_30 control_22 control_32 control_38 control_69 control_74 control_61 control_20 control_20 control_40 hunt_72_8 hunt_63_4 hunt_59_8 hunt_48_5 hunt_88_3 hunt_93_5 hunt_76_12 hunt_58_1 hunt_45_9 park_74_3.5 park_64_4 park_57_3 park_76_2.5	Cluster 1 number of entries 5 als_40_14.5 als_39_7 park_80_2 park_60_3 park_57_3
	Cluster 5 number of entries 1 hunt_85_10
	Cluster 3 number of entries 1 hunt_75_2
	Cluster 9 number of entries 1 hunt_56_9
Cluster 8 number of entries 1 als_70_54	Cluster 10 number of entries 1 hunt_102_4
Cluster 2 number of entries 1 als_51_3	Cluster 6 number of entries 12 als_65_9 als_50_54 als_62_12 control_22 hunt_58_11 hunt_64_2 hunt_90_12 hunt_84_11 hunt_75_8 hunt_97_12 park_77_4 park_75_2
Cluster 1 number of entries 5 als_40_14.5 als_39_7 park_80_2 park_60_3 park_57_3	
Cluster 7 number of entries 14 als_68_1 als_63_14 als_70_13 als_36_5.5 als_43_17 control_57 control_23 hunt_53_2 park_44_1.5 park_53_2 park_64_4 park_68_1.5 park_74_3 park_79_3	Cluster 5 number of entries 1 hunt_85_10

Figure 5.13: Euclidean: Ward

5.5.4 Euclidean distance- Weighted

This method has very similar results as the method using the DTW distance aggregated with Weighted.

The giant cluster is five and is almost equal to cluster 4 in DTW-weighted method. It has elements from all the groups, and the young, healthy subjects do not belong here. The elements in

the early stages of the disease or too young do not fit in this cluster either. So, we consider to have fifteen elements wrong classified in fifty-two.

Cluster 1 has elements with ALS and Parkinson. We consider there is no element is wrong classified.

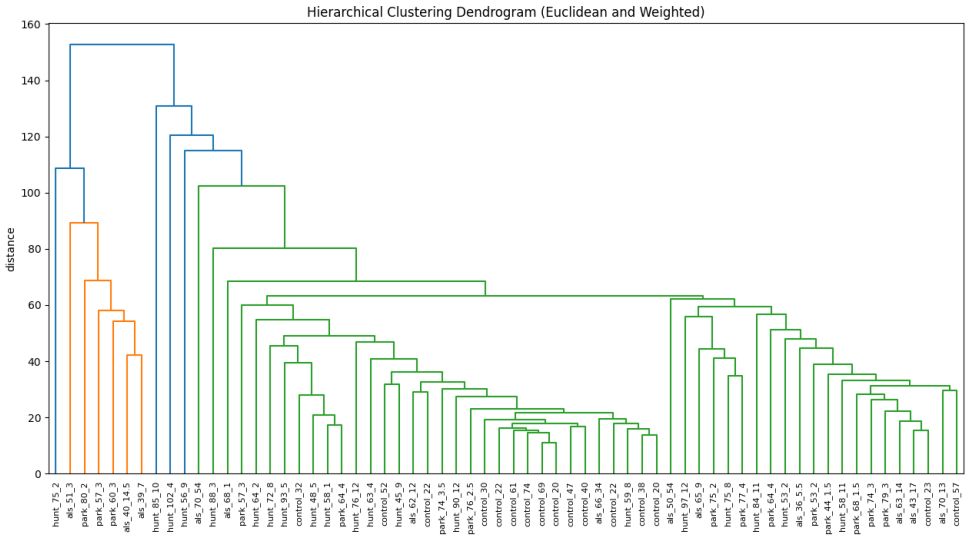


Figure 5.14: Euclidean: Weighted

Cluster 5 number of entries 52 als_68_1 als_63_14 als_70_13 als_36_5.5 als_43_17 als_65_9 als_50_54 als_62_12 als_66_34 control_57 control_22 control_23 control_52 control_47 control_30 control_22 control_22 control_32 control_38 control_69 control_74 control_61 control_20 control_20 control_40 hunt_72_8 hunt_58_11 hunt_63_4 hunt_64_2 hunt_59_8 hunt_90_12 hunt_84_11 hunt_75_8 hunt_48_5 hunt_97_12 hunt_93_5 hunt_76_12 hunt_53_2 hunt_58_1 hunt_45_9 park_77_4 park_44_1.5 park_74_3.5 park_75_2 park_53_2 park_64_4 park_64_4 park_68_1.5 park_74_3 park_57_3 park_79_3 park_76_2.5	Cluster 3 number of entries 1 als_51_3
Cluster 7 number of entries 1 als_70_54	Cluster 1 number of entries 4 als_40_14.5 als_39_7 park_60_3 park_57_3
Cluster 2 number of entries 1 park_80_2	Cluster 10 number of entries 1 hunt_85_10
	Cluster 4 number of entries 1 hunt_75_2
	Cluster 8 number of entries 1 hunt_56_9
	Cluster 9 number of entries 1 hunt_102_4
	Cluster 6 number of entries 1 hunt_88_3

Figure 5.15: Euclidean: Weighted

5.5.5 DTW distance- Ward

By looking at the tables, we can see that DTW distance with Ward aggregation measure has almost the same elements in each cluster as the Manhattan and Euclidean distance using the same aggregation measure.

Cluster 1 has people with ALS and Parkinson disease. Two of the three people with Parkinson's disease are in stage 3 of the HY scale, and the other is in stage 2 of HY but is very old. The individuals with ALS were diagnosis less than 15 months, so they are the elements wrong classified. We consider two elements wrong classified in five entries.

Cluster 4 has control elements and elements with Huntington. It also has one element with Parkinson in stage 4 of the disease. This cluster represents elements from the third category, so we have seven control elements wrong classified. The elements with Huntington are all right classified because they are old or have motor capacity lower than 5. We consider seven elements wrong classified in nineteen.

Cluster 5 has elements from all the groups and has characteristics of the second category. Individuals with ALS with six months since the diagnosis, do not belong here, and Control elements younger than 60 years. People with Parkinson very old or in a high stage of the disease are not well classified. Individuals with Huntington disease are well classified. We consider six elements wrong classified in nineteen.

Cluster 6 also has elements from all the groups and represents individuals with serious difficulty in walking. We admit that the control elements are wrong classified. In this way, we consider four elements wrong classified in eleven.

Cluster 7 has only Huntington disease elements, so we assume that all of them are well classified.

The other clusters have one element each, so they are not relevant.

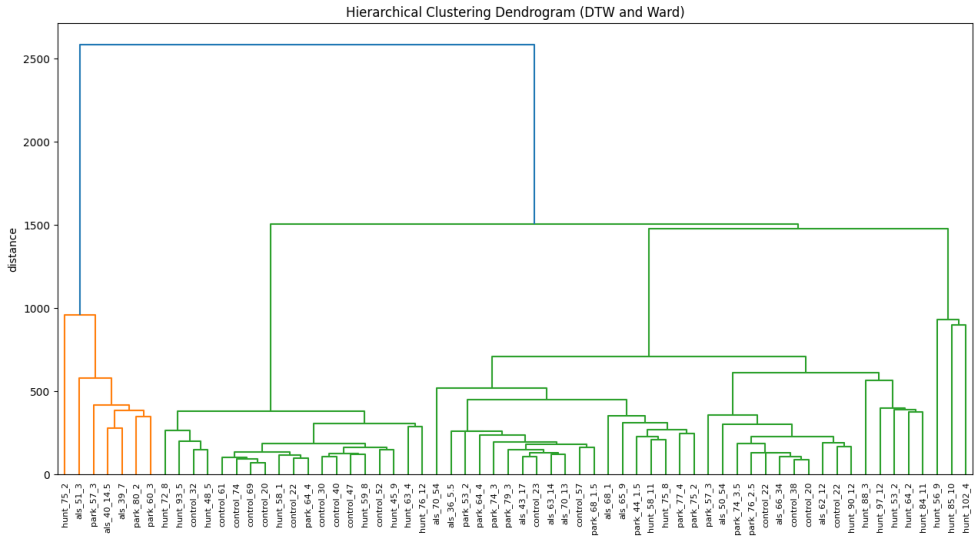


Figure 5.16: DTW: Ward

5.5.6 DTW distance- Weighted

Cluster 1 aggregates elements with ALS with elements with Parkinson. The elements with Parkinson are in an advanced stage of the disease or are elderly. The elements with ALS have discovered the diagnosis for less than 15 months. In this way, they are not well aggregated with this Parkinson elements. We consider two elements wrong classified in five entries.

Cluster 4 is a giant cluster. It has elements from all the groups and represents subjects with median problems in gait. The elements from the control group that are younger than 65 do not

Cluster 5 number of entries 19 als_68_1 als_63_14 als_70_13 als_70_54 als_36_5.5 als_43_17 als_65_9 control_57 control_23 hunt_58_11 hunt_75_8 park_77_4 park_44_1.5 park_75_2 park_53_2 park_64_4 park_68_1.5 park_74_3 park_79_3	Cluster 4 number of entries 19 control_22 control_52 control_47 control_30 control_32 control_69 control_74 control_61 control_20 control_40 hunt_72_8 hunt_63_4 hunt_59_8 hunt_48_5 hunt_93_5 hunt_76_12 hunt_58_1 hunt_45_9 park_64_4
Cluster 2 number of entries 1 als_51_3	Cluster 7 number of entries 5 hunt_64_2 hunt_84_11 hunt_97_12 hunt_88_3 hunt_53_2
Cluster 6 number of entries 11 als_50_54 als_62_12 als_66_34 control_22 control_22 control_38 control_20 hunt_90_12 park_74_3.5 park_57_3 park_76_2.5	Cluster 8 number of entries 1 hunt_85_10
Cluster 1 number of entries 5 als_40_14.5 als_39_7 park_80_2 park_60_3 park_57_3	Cluster 3 number of entries 1 hunt_75_2
	Cluster 10 number of entries 1 hunt_56_9
	Cluster 9 number of entries 1 hunt_102_4

Figure 5.17: DTW: Ward

belong here. ALS patients with the recent diagnosis do not belong here either. There is one element wrong classified from the Huntington group because he is young and has a high motor function. There are two elements with Parkinson too young and in an early stage of the disease that is also wrong classified. We consider seventeen elements wrong classified in thirty-nine entries.

Cluster 5 has one only control element and very young; he can not be aggregated with people with the disease. The ALS element with just nine months since diagnosis is not well classified. The elements with Parkinson and Hungtinton are well classified; the age compensated the early stages of the disease. We consider two elements wrong classified in thirteen entries.

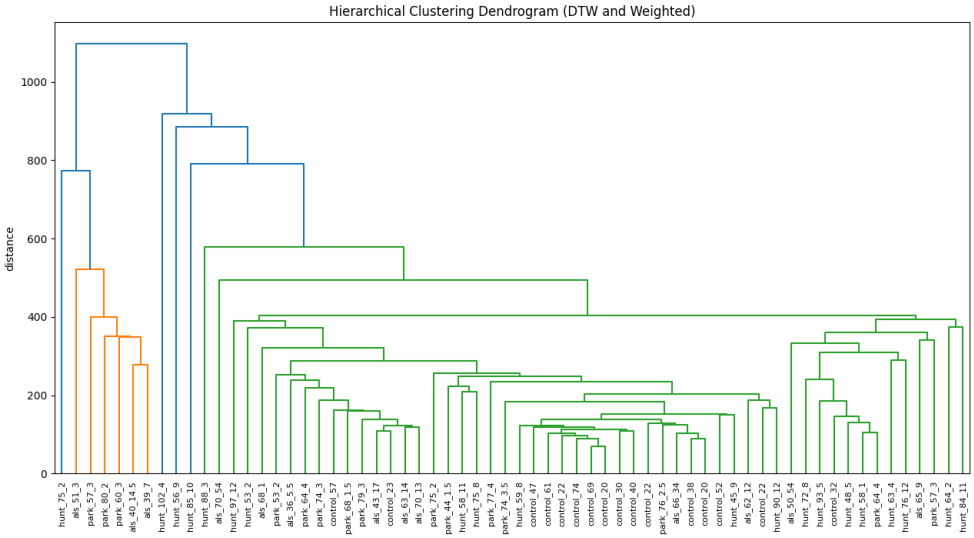


Figure 5.18: DTW: Weighted

5.5.7 Fourier distance- Ward

The clusters seen on this Figure 5.20 are exactly the same as the Euclidean Ward.

5.5.8 Fourier distance- Weighted

Cluster 7 has 14 entries and we consider it has four elements wrong classified: one young element with a recent ALS diagnosis; one healthy individual with 23 years and two elements with

Cluster 4 number of entries 39 als_68_1 als_63_14 als_70_13 als_36_5.5 als_43_17 als_62_12 als_66_34 control_57 control_22 control_23 control_52 control_47 control_30 control_22 control_22 control_38 control_69 control_74 control_61 control_20 control_20 control_40 hunt_58_11 hunt_59_8 hunt_90_12 hunt_75_8 hunt_97_12 hunt_53_2 hunt_45_9 park_77_4 park_44_1.5 park_74_3.5 park_75_2 park_53_2 park_64_4 park_68_1.5 park_74_3 park_79_3 park_76_2.5	Cluster 6 number of entries 1 als_70_54
Cluster 5 number of entries 13 als_65_9 als_50_54 control_32 hunt_72_8 hunt_63_4 hunt_64_2 hunt_84_11 hunt_48_5 hunt_93_5 hunt_76_12 hunt_58_1 park_64_4 park_57_3	Cluster 2 number of entries 1 als_51_3
Cluster 7 number of entries 1 hunt_88_3	Cluster 1 number of entries 5 als_40_14.5 als_39_7 park_80_2 park_60_3 park_57_3
	Cluster 8 number of entries 1 hunt_85_10
	Cluster 3 number of entries 1 hunt_75_2
	Cluster 9 number of entries 1 hunt_56_9
	Cluster 10 number of entries 1 hunt_102_4

Figure 5.19: DTW: Weighted

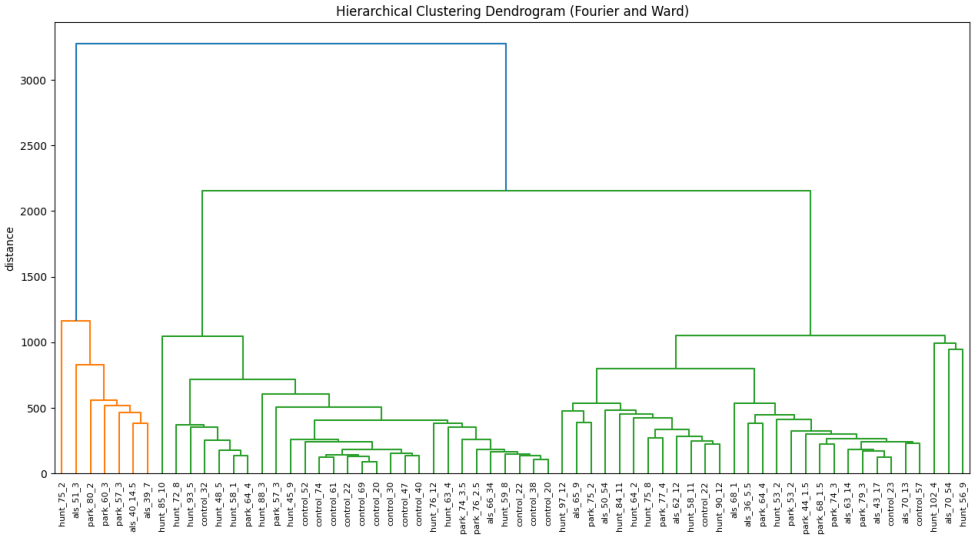


Figure 5.20: Fourier: Ward

Cluster 4 number of entries 27 als_66_34 control_22 control_52 control_47 control_30 control_22 control_32 control_38 control_69 control_74 control_61 control_20 control_20 control_40 hunt_72_8 hunt_63_4 hunt_59_8 hunt_48_5 hunt_88_3 hunt_93_5 hunt_76_12 hunt_58_1 hunt_45_9 park_74_3.5 park_64_4 park_57_3 park_76_2.5	Cluster 1 number of entries 5 als_40_14.5 als_39_7 park_80_2 park_60_3 park_57_3
Cluster 8 number of entries 1 als_70_54	Cluster 5 number of entries 1 hunt_85_10
Cluster 2 number of entries 1 als_51_3	Cluster 3 number of entries 1 hunt_75_2
Cluster 1 number of entries 5 als_40_14.5 als_39_7 park_80_2 park_60_3 park_57_3	Cluster 9 number of entries 1 hunt_56_9
Cluster 7 number of entries 14 als_68_1 als_63_14 als_70_13 als_36_5.5 als_43_17 control_57 control_23 hunt_53_2 park_44_1.5 park_53_2 park_64_4 park_68_1.5 park_74_3 park_79_3	Cluster 10 number of entries 1 hunt_102_4
	Cluster 6 number of entries 12 als_65_9 als_50_54 als_62_12 control_22 hunt_58_11 hunt_64_2 hunt_90_12 hunt_84_11 hunt_75_8 hunt_97_12 park_77_4 park_75_2
	Cluster 5 number of entries 1 hunt_85_10

Figure 5.21: Fourier: Ward

Parkinson young and in an early stage of the disease.

Cluster 6 has elements from all the four groups. The control element has only 22 years, so he is too young to be compared to an individual with neurodegenerative diseases. The element with Huntington disease with motor function of 11 and age 58 years is also too young to fit in this cluster.

Cluster 1 has all the elements with neurodegenerative disease and in an advanced stage.

Cluster 4 has 27 entries and has elements from the control group, elements with Huntington disease, and Parkinson's elements. Three healthy subjects are too young to fit in this cluster. The element with Huntington disease with 59 years and eight motor function we consider to be also wrong classified.

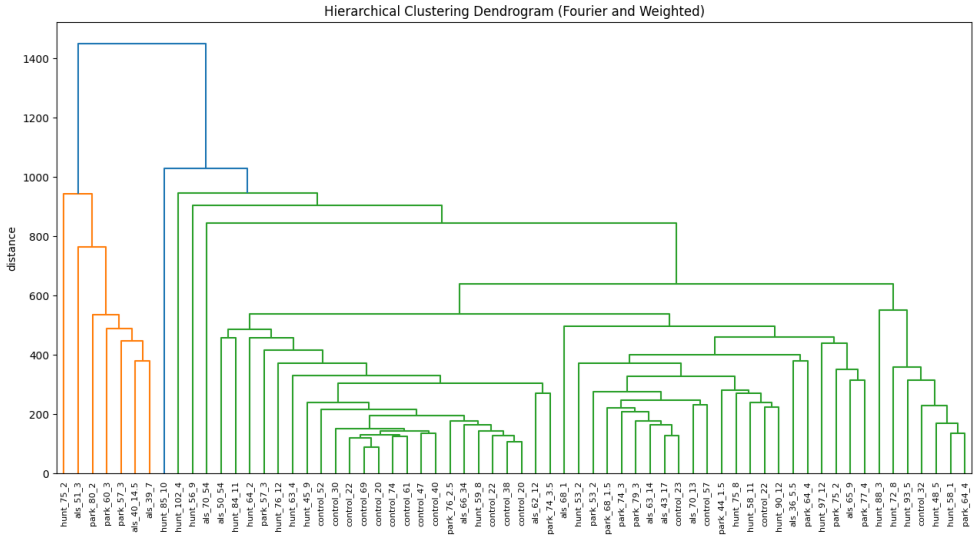


Figure 5.22: Fourier: Weighted

Cluster 4 number of entries 27 control_38 control_69 control_74 control_61 control_20 control_20 control_40 hunt_72_8 hunt_63_4 hunt_59_8 hunt_48_5 hunt_88_3 hunt_93_5 hunt_76_12 hunt_58_1 hunt_45_9 park_74_3.5 park_64_4 park_57_3 park_76_2.5	Cluster 5 number of entries 1 hunt_85_10
Cluster 2 number of entries 1 als_51_3	Cluster 3 number of entries 1 hunt_75_2
Cluster 1 number of entries 5 als_40_14.5 als_39_7 park_80_2 park_60_3 park_57_3	Cluster 9 number of entries 1 hunt_56_9
Cluster 6 number of entries 12 als_65_9 als_50_54 als_62_12 control_22 hunt_58_11 hunt_64_2 hunt_90_12 hunt_84_11 hunt_75_8 hunt_97_12 park_77_4 park_75_2	Cluster 10 number of entries 1 hunt_102_4
	Cluster 8 number of entries 1 als_70_54
	Cluster 7 number of entries 14 als_68_1 als_63_14 als_70_13 als_36_5.5 als_43_17 control_57 control_23 hunt_53_2 park_44_1.5 park_53_2 park_64_4 park_68_1.5 park_74_3 park_79_3

Figure 5.23: Fourier: Weighted

5.6 Cluster Evaluation

In order to evaluate the algorithm results, purity is the measure chosen to use, as it was in the first dataset.

The aggregation measure with the best purity is the Weighted, the same as the first dataset. However, the distance measure that present the best purity percentage is the Fourier distance.

The table 5.1 and the graphic 5.24 shows the purity obtained for each distance measure combined with aggregation methods. The distance measure that presents the lowest purity percentage is the DTW distance. The distance measure that achieves high purity in both aggregation measures

is the Fourier distance. In this way, the Fourier distance combined with the weighted aggregation measure has the higher purity percentage: 82,758%.

Table 5.1: Purity Percentage of *Gait in Neurodegenerative Disease Dataset*

	Euclidean	DTW	Manhattan	Fourier
Ward	74,1%	67,8%	69,5%	74,1%
Weighted	73,2%	63,2%	71,9%	82,8%

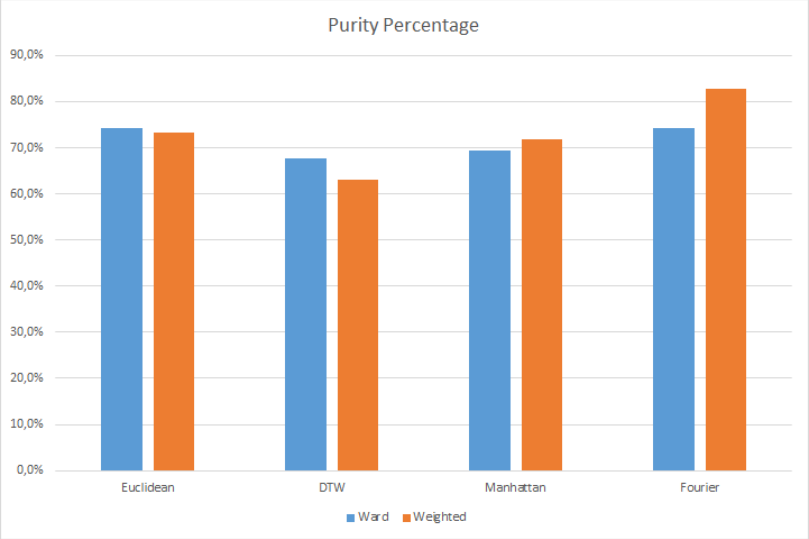


Figure 5.24: Purity Percentage

Chapter 6

Discussions and Conclusions

In this study, a hierarchical clustering method has been applied to classify patients' gait with neurodegenerative diseases from people with a normal gait. The goal was to compare and aggregate gait patterns from different individuals.

It were chosen two datasets: one with data of the gait of patients with Parkinson's disease and control patients; and the other with data of the gait of the three diseases understudy and control patients.

The input variable to the algorithm was the swing percentage, as it is a variable that is expected to present distinguishing characteristics between gaits.

A preprocessing of the data from *Gait in Parkinson's Disease Dataset* was done and the swing percentage was calculated through the data coming from the pressure sensors.

The hierarchical clustering was applied to both datasets using four distance measures: Euclidean, DTW, Manhattan and Fourier. Moreover, these distance measures were combined with seven aggregation methods for the first dataset. For the second dataset, the same distance measure were used combined with the aggregation methods with higher purity observed in the previous dataset.

In order to observe the results, it were done dendrograms. And, in order to evaluate the algorithm, the purity percentage was calculated.

The distance and aggregation measures that present the highest purity percentage for the *Gait in Parkinson's Disease Dataset* is the Manhattan distance combined with the Weighted method. The purity percentage achieved is 82,802%.

However, for this dataset, the distance measure that presents the best results in general is the DTW distance, with a purity mean of 74,5%. The aggregation measure that presents the best results is the Weighted, with a purity mean of 78,4%.

Furthermore, for the *Gait in Neurodegenerative Disease Dataset*, the Fourier distance combined with the weighted achieves the purity percentage of 82,758%, the highest compared to the other methods.

Unlike the previous dataset, in this dataset the measure of distance and aggregation that present the highest purity results in mean are the same ones that present the best combined result. The

Fourier distance presents a purity mean of 78,4% and the Weighted method presents a purity mean of 72,8%.

In conclusion, there are two different distances that present the better results in each dataset. However, the best aggregation method is common in both datasets, the Weighted method.

6.1 Limitations of the study

Regarding the *Gait in Parkinson's Disease Dataset*, the individuals chosen to study were at very early stages of the disease. Most individuals were in stage two and a few in stage three. These individuals were chosen because one of this study's variables was to study the ability of dual-tasking and patients in stage 4 or 5 of the disease might have great difficulties. However, with the progress of the disease, the difficulties walking are higher. Due to this, if the chosen individuals were in an advanced stage of the diseases, the gait differences were significant.

Another limitation of both studies is that individuals with neurodegenerative diseases were taking medication to reduce disease's symptoms. Furthermore, we do not have information about each individual's physical condition and if he has a prosthesis or not.

The age of the subjects used for this study should be more similar to reduce the variability of gait depending on age.

For both datasets, we only have data from ground force pressure sensors. A way to improve the results was to have more data like accelerometer or electromyography.

The sensors used have some limitations too because sometimes they failed.

In future work, the data should be selective, and people should not be taken any medication to control the motor symptoms. The individuals chosen to participate in the study should be the same age approximately. Besides, the physical condition should be considered, and another varies that can compromise the results.

Appendix A

Appendix

Gait in Neurodegenerative Disease Dataset

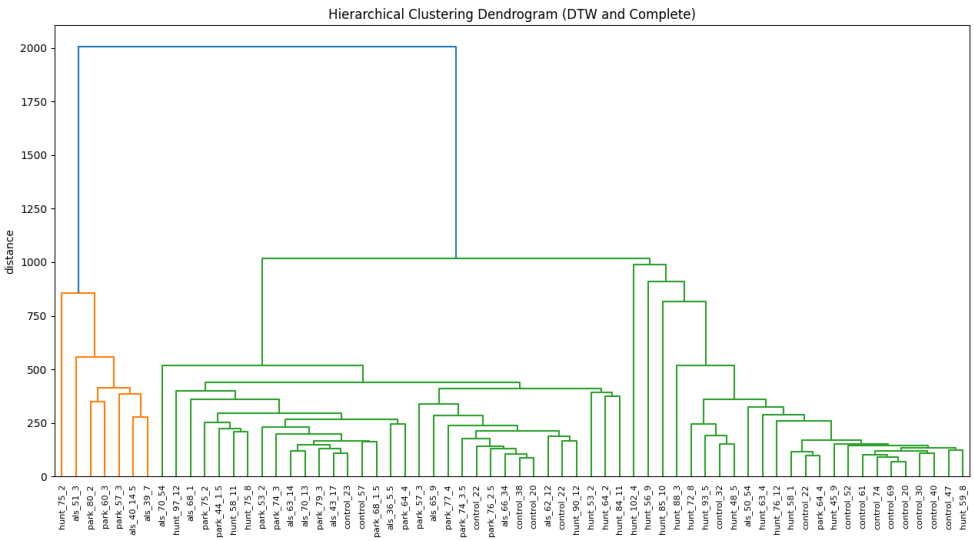


Figure A.1: DTW: Complete

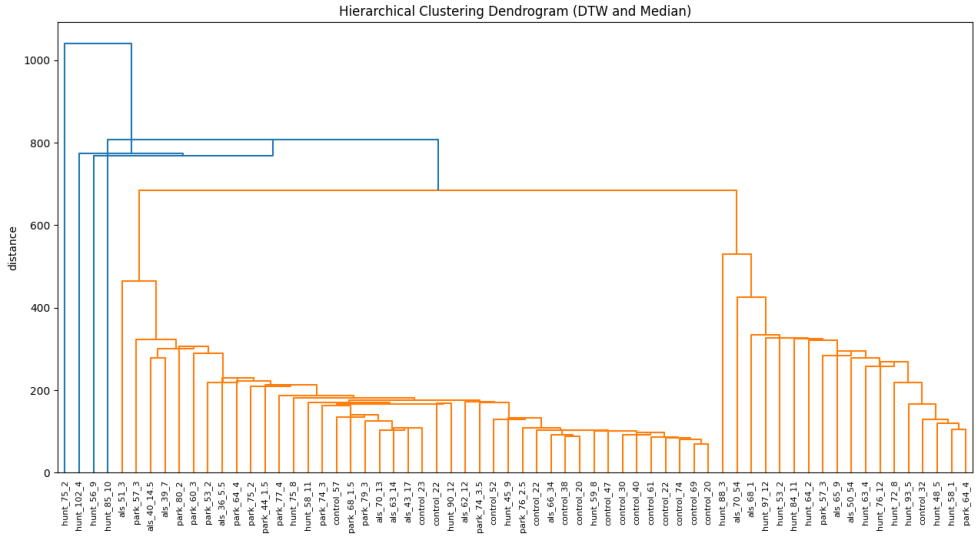


Figure A.2: DTW: Median

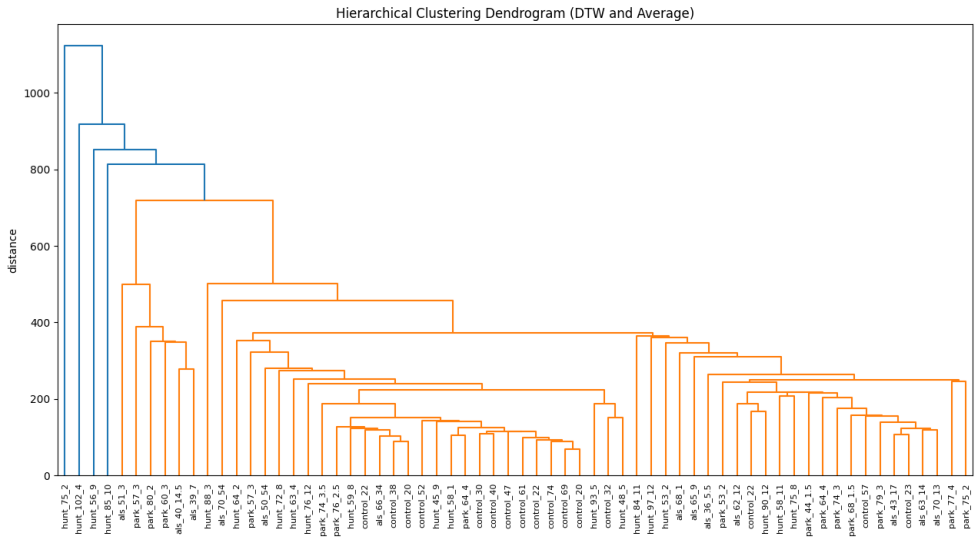


Figure A.3: DTW: Average

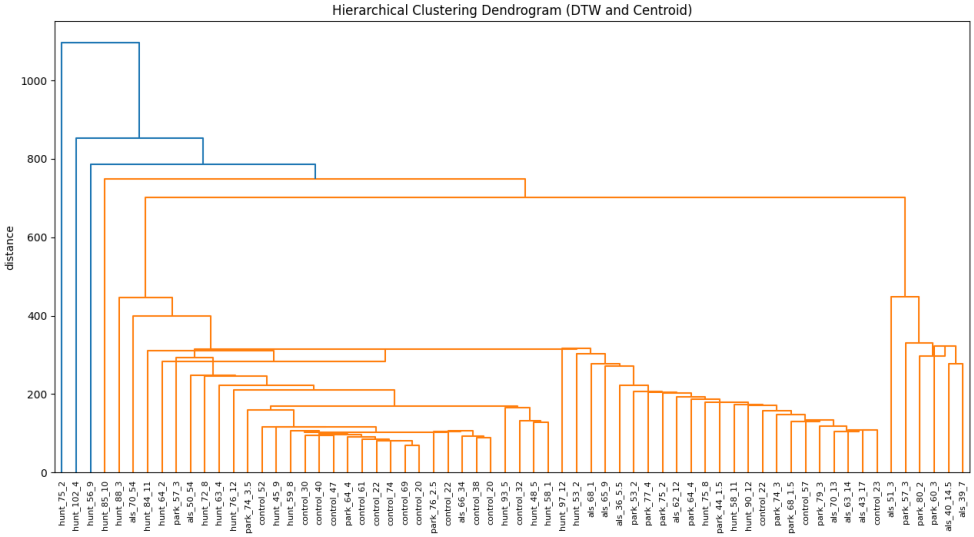


Figure A.4: DTW: Centroid

References

[1] Weijun Tao, Tao Liu, Rencheng Zheng, and Hutian Feng. Gait analysis using wearable sensors. *Sensors*, 12(2):2255–2283, 2012.

[2] Jacquelin Perry, Jon R Davids, et al. Gait analysis: normal and pathological function. *Journal of Pediatric Orthopaedics*, 12(6):815, 1992.

[3] A. Bonnefoy-Mazure and Stéphane Armand. *Normal gait*. 01 2015.

[4] Ary L Goldberger, Luis A N Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-kang Peng, and H Eugene Stanley. Components of a New Research Resource for Complex Physiologic Signals. 2000.

[5] Saeed Aghabozorgi, Ali Seyed Shirkhorshidi, and Teh Ying Wah. Time-series clustering - A decade review. *Information Systems*, 53:16–38, 2015. URL: http://dx.doi.org/10.1016/j.is.2015.04.007, doi:10.1016/j.is.2015.04.007.

[6] Xiaohang Zhang, Jiaqi Liu, Yu Du, and Tingjie Lv. A novel clustering method on time series data. *Expert Systems with Applications*, 38(9):11891–11900, 2011.

[7] Sadiq Hussain and Gopal Hazarika. Enhanced hierarchical clustering for genome databases. *International Journal of Computer Science Issues (IJCSI)*, 8(4):245, 2011.

[8] Ali Seyed Shirkhorshidi, Saeed Aghabozorgi, and Teh Ying Wah. A comparison study on similarity and dissimilarity measures in clustering continuous data. *PloS one*, 10(12):e0144059, 2015.

[9] Wei Zeng, Fenglin Liu, Qinghui Wang, Ying Wang, Limin Ma, and Yu Zhang. Parkinson's disease classification using gait analysis via deterministic learning. *Neuroscience Letters*, 633:268–278, 2016. doi:10.1016/j.neulet.2016.09.043.

[10] World Health Organization. *Neurological disorders: public health challenges*. World Health Organization, 2006.

[11] Adriano Chiò, G Logroscino, BJ Traynor, J Collins, JC Simeone, LA Goldstein, and LA White. Global epidemiology of amyotrophic lateral sclerosis: a systematic review of the published literature. *Neuroepidemiology*, 41(2):118–130, 2013.

[12] J Douglas Mitchell and Gian Domenico Borasio. Amyotrophic lateral sclerosis. *The lancet*, 369(9578):2031–2041, 2007.

[13] Tamara Pringsheim, Katie Wiltshire, Lundy Day, Jonathan Dykeman, Thomas Steeves, and Nathalie Jette. The incidence and prevalence of huntington's disease: a systematic review and meta-analysis. *Movement Disorders*, 27(9):1083–1091, 2012.

[14] J Jankovic. Parkinson's disease: clinical features and diagnosis. *Journal of Neurology, Neurosurgery & Psychiatry*, 79(4):368–376, 2008. URL: https://jnnp.bmj.com/content/79/4/368, arXiv:https://jnnp.bmj.com/content/79/4/368.full.pdf, doi:10.1136/jnnp.2007.131045.

[15] Barbara S Connolly and Anthony E Lang. Pharmacological treatment of parkinson disease: a review. *Jama*, 311(16):1670–1683, 2014.

[16] Tuan D. Pham. Texture Classification and Visualization of Time Series of Gait Dynamics in Patients with Neuro-Degenerative Diseases. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 26(1):188–196, 2018. doi:10.1109/TNSRE.2017.2732448.

[17] Joseph Jankovic. Parkinson's disease: clinical features and diagnosis. *Journal of neurology, neurosurgery & psychiatry*, 79(4):368–376, 2008.

[18] EuropeanParkinsonAssociation. [Online; accessed 01-02-2020]. URL: https://www.epda.eu.com/.

[19] Joe R Nocera, Elizabeth L Stegemöller, Irene A Malaty, Michael S Okun, Michael Marsiske, Chris J Hass, National Parkinson Foundation Quality Improvement Initiative Investigators, et al. Using the timed up & go test in a clinical setting to predict falling in parkinson's disease. *Archives of physical medicine and rehabilitation*, 94(7):1300–1305, 2013.

[20] Deepak Joshi, Aayushi Khajuria, and Pradeep Joshi. An automatic non-invasive method for Parkinson's disease classification. *Computer Methods and Programs in Biomedicine*, 145:135–145, 2017. doi:10.1016/j.cmpb.2017.04.007.

[21] Gillian Bates, Peter S Harper, and Lesley Jones. *Huntington's disease*. Number 45. Oxford University Press, USA, 2002.

[22] National Organization for Rare Disorders. Huntington's disease. *NORD (National Organization for Rare Disorders)*. URL: https://rarediseases.org/rare-diseases/huntingtons-disease/.

[23] Raymund AC Roos. Huntington's disease: a clinical review. *Orphanet journal of rare diseases*, 5(1):1–8, 2010.

[24] Darrell Hulisz. Disease state overview. *The American journal of managed care*, 24(15):S320–S326, 2018.

[25] Björn Oskarsson, Tania F Gendron, and Nathan P Staff. Amyotrophic lateral sclerosis: an update for 2018. In *Mayo Clinic Proceedings*, volume 93, pages 1617–1628. Elsevier, 2018.

[26] Silvi Frenkel-Toledo, Nir Giladi, Chava Peretz, Talia Herman, Leor Gruendlinger, and Jeffrey M Hausdorff. Treadmill walking as an external pacemaker to improve gait rhythm and stability in parkinson's disease. *Movement disorders: official journal of the Movement Disorder Society*, 20(9):1109–1114, 2005.

[27] Katherine A Boyer, Russell T Johnson, Jacob J Banks, Carl Jewell, and Jocelyn F Hafer. Systematic review and meta-analysis of gait mechanics in young and older adults. *Experimental gerontology*, 95:63–70, 2017.

[28] François Prince, Héléne Corriveau, Réjean Hebert, and David A Winter. Gait in the elderly. *Gait & posture*, 5(2):128–135, 1997.

[29] Suzanne E Halliday, David A Winter, James S Frank, Aftab E Patla, and Francois Prince. The initiation of gait in young, elderly, and parkinson's disease subjects. *Gait & posture*, 8(1):8–14, 1998.

[30] Karen M Ostrosky, Jessie M VanSwearingen, Ray G Burdett, and Zena Gee. A comparison of gait characteristics in young and old subjects. *Physical therapy*, 74(7):637–644, 1994.

[31] Bento Murteira, Daniel Muller, and Kamil Feridun Turkman. Analise de sucessoes cronologicas. 2000.

[32] T Warren Liao. Clustering of time series data—a survey. *Pattern recognition*, 38(11):1857–1874, 2005.

[33] A Kianimajd, MG Ruano, P Carvalho, J Henriques, T Rocha, S Paredes, and AE Ruano. Comparison of different methods of measuring similarity in physiologic time series. *IFAC-PapersOnLine*, 50(1):11005–11010, 2017.

[34] Félix Iglesias and Wolfgang Kastner. Analysis of similarity measures in times series clustering for the discovery of building energy patterns. *Energies*, 6(2):579–597, 2013.

[35] Chotirat Ann Ratanamahatana and Eamonn Keogh. Making time-series classification more accurate using learned constraints. In *Proceedings of the 2004 SIAM international conference on data mining*, pages 11–22. SIAM, 2004.

[36] Dan Gusfield. Algorithms on stings, trees, and sequences: Computer science and computational biology. *Acm Sigact News*, 28(4):41–60, 1997.

[37] Seyedjamal Zolhavarieh, Saeed Aghabozorgi, and Ying Wah Teh. A Review of Subsequence Time Series Clustering. 2014, 2014. doi:10.1155/2014/312521.

[38] J Swarndeep Saket and Sharnil Pandya. An overview of partitioning algorithms in clustering techniques. *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, 5(6):1943–1946, 2016.

[39] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. 96(34):226–231, 1996.

[40] Deepti Sisodia, Lokesh Singh, Sheetal Sisodia, and Khushboo Saxena. Clustering techniques: a brief survey of different clustering algorithms. *International Journal of Latest Trends in Engineering and Technology (IJLTET)*, 1(3):82–87, 2012.

[41] Wei Wang, Jiong Yang, Richard Muntz, et al. Sting: A statistical information grid approach to spatial data mining. 97:186–195, 1997.

[42] Rashi Chauhan. Clustering techniques: A comprehensive study of various clustering techniques. *International Journal of Advanced Research in Computer Science*, 5(5), 2014.

[43] Yinglin Xia. Correlation and association analyses in microbiome study integrating multiomics in health and disease. *Progress in Molecular Biology and Translational Science*, 171:309–491, 2020.

[44] Angur Mahmud Jarman. Hierarchical cluster analysis: Comparison of single linkage, complete linkage, average linkage and centroid linkage method.

[45] Frédéric Ros and Serge Guillaume. A hierarchical clustering algorithm and an improvement of the single linkage criterion to deal with noise. *Expert Systems with Applications*, 128:96–108, 2019.

[46] Daniel Müllner. Modern hierarchical, agglomerative clustering algorithms. *arXiv preprint arXiv:1109.2378*, 2011.

[47] Eamonn Keogh and Shruti Kasetty. On the need for time series data mining benchmarks: a survey and empirical demonstration. *Data Mining and knowledge discovery*, 7(4):349–371, 2003.

[48] Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to information retrieval*. Cambridge university press, 2008.

[49] Monica Chiş, Soumya Banerjee, and Aboul Ella Hassanien. Clustering time series data: an evolutionary approach. In *Foundations of Computational, IntelligenceVolume 6*, pages 193–207. Springer, 2009.

[50] Francesco Gullo, Giovanni Ponti, Andrea Tagarelli, Giuseppe Tradigo, and Pierangelo Veltri. A time series approach for clustering mass spectrometry data. *Journal of Computational Science*, 3(5):344–355, 2012.

[51] Yunfeng Wu and Sridhar Krishnan. Statistical analysis of gait rhythm in patients with parkinson's disease. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 18(2):150–158, 2010. doi:10.1109/TNSRE.2009.2033062.

[52] Abed Khorasani and Mohammad Reza Daliri. HMM for Classification of Parkinson's Disease Based on the Raw Gait Data. *Journal of Medical Systems*, 38(12), 2014. doi:10.1007/s10916-014-0147-5.

[53] Cristina Linares, Pablo Martinez-Martin, Carmen Rodríguez-Blázquez, Maria João Forjaz, Rocío Carmona, and Julio Díaz. Effect of heat waves on morbidity and mortality due to Parkinson's disease in Madrid: A time-series analysis. *Environment International*, 89-90:1–6, 2016. URL: http://dx.doi.org/10.1016/j.envint.2016.01.017, doi:10.1016/j.envint.2016.01.017.

[54] Jeffrey M Hausdorff, Apinya Lertratanakul, Merit E Cudkowicz, Amie L Peterson, David Kaliton, A R Y L Goldberger, M Jeffrey, Apinya Lertratanakul, E Cudkowicz, Amie L Peterson, and David Kaliton. Dynamic markers of altered gait rhythm in amyotrophic lateral sclerosis. pages 2045–2053, 2019.