Time Series Analysis of Biomarkers of Progression in Neurodegenerative Diseases

Íñigo Sanz Ilundain

Degree in Computer Science
Faculty of Computer Science and Engineering
Complutense University of Madrid

Degree Thesis in Computer Science

May 30 2022

Tutor/s and/or colaborator/s:

Laura Hernández Lorenzo
José Luis Ayala Rodrigo

Resumen en castellano

Las enfermedades neurodegenerativas son enfermedades complejas y temporalmente dependientes. Entre ellas, la más común es la enfermedad de Alzheimer, en la que los pacientes atraviesan una serie de estados sintomáticos antes de llegar al diagnóstico de demencia causada por el Alzheimer. Debido a sus características, es necesario estudiar los biomarcadores asociados a esta enfermedad desde el punto de vista de las series temporales. En este trabajo, exploramos exhaustivamente los biomarcadores encontrados en la cohorte de Alzheimer's Disease Neuroimaging Initiative (ADNI), aplicando una colección de algoritmos de clustering, tanto temporales como no temporales. Se han aplicado estos modelos a varios conjuntos de datos: considerando un solo biomarcador, combinándolos por pares y utilizando las regiones MRI. Los resultados obtenidos con estos conjuntos de datos fueron evaluados desde un punto de vista computacional, al igual que uno clínico, este último correspondiendo a los resultados esperados. Estos resultados muestran que, desde un punto de vista computacional, los modelos no temporales suelen obtener mejores resultados que los modelos temporales. Adicionalmente, los conjuntos de datos combinando biomarcadores con características similares incrementan el valor del resultado. Mientras tanto, desde un punto de vista clínico, el resultado influye mucho del algoritmo y del conjunto de datos utilizado: los modelos no temporales suelen obtener mejores resultados en conjuntos de datos que contengan biomarcadores como AV45 o ABETA, pero obtienen peores cuando se usan otros biomarcadores como TAU o PTAU, un límite que se puede superar usando modelos temporales. Este trabajo pone de manifiesto un enorme potencial en el clustering de series temporales en el conocimiento de las enfermedades dependientes del tiempo, como las neurodegenerativas.

Palabras clave

Enfermedad de Alzheimer, biomarcadores, deformación dinámica del tiempo, clustering

Abstract

Neurodegenerative diseases are complex and highly time-dependent diseases. Among them, the most common is Alzheimer's Disease (AD), in which the patient goes through a series of symptomatic stages before receiving the diagnosis of dementia caused by AD. Due to its temporal characteristics, it is necessary to study the biomarkers associated with the AD from a time series point of view. In this work, we have exhaustively explored the biomarkers found on the Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort applying a wide array of clustering algorithms both temporal and non-temporal. These models were applied to several biomarker datasets: implying one biomarker only, combining them in pairs as well as making use of the MRI regions. The results obtained with these dataset types were evaluated using a computational and a clinical standpoint, where the latter corresponded very clearly with its expected outcomes. Results show that from a computational perspective, non-temporal models generally obtain greater outcomes than temporal models. Moreover, datasets combining biomarkers with similar properties also increase the resulting score. Contrarily, from a clinical standpoint, the outcome depends greatly on the algorithm and the dataset used: non-temporal models obtain better results in datasets containing AV45 or ABETA but struggle when used with TAU or PTAU, a limitation that can be surpassed by making use of temporal models. The present work raises enormous potential found in time series clustering to discover knowledge in time-dependent diseases such as the neurodegenerative ones.

Keywords

Alzheimer's Disease, biomarkers, dynamic time warping, clustering

Contents

Índice i

Agradecimientos ii

Dedicatoria iii

1 Introduction 1

2 Materials and Methods 6
- 2.1 Database description . . . 6
- 2.2 Data filtering . . . 10
- 2.3 Data preprocessing . . . 13
  - 2.3.1 Missing values . . . 13
  - 2.3.2 Data transformations . . . 15
- 2.4 Unsupervised classification . . . 15
  - 2.4.1 Agglomerative Hierarchical Clustering . . . 16
  - 2.4.2 K-Means . . . 19
  - 2.4.3 Density-Based Spatial Clustering of Application (DBSCAN) . . . 20
  - 2.4.4 Self Organizing Maps (SOM) . . . 21
  - 2.4.5 Dynamic Time Warping with Hierarchical Clustering . . . 23
  - 2.4.6 Dynamic Time Warping with DBSCAN . . . 26
- 2.5 Evaluation Metrics . . . 27

3 Results 29
- 3.1 Non-Temporal models . . . 29
  - 3.1.1 Non-Temporal models with separated biomarkers . . . 30
  - 3.1.2 Non-Temporal models with pairs of biomarkers . . . 33
  - 3.1.3 Non-Temporal models with MRI regions . . . 34
- 3.2 Temporal models . . . 35
  - 3.2.1 Temporal models with separated biomarkers . . . 36
  - 3.2.2 Temporal models with pairs of biomarkers . . . 39
  - 3.2.3 Temporal models with MRI regions . . . 40

4 Discussion 42

5 Conclusion 53

Bibliography 56

Agradecimientos

Quiero agradecerle a toda mi familia el apoyo incondicional que me han dado. A mis padres, a mi tía Piluca y a mi tío Carlos por apoyarme en los momentos más complicados.

Muchas gracias a Berta, a Víctor, y al grupo de teatro de la Complutense, Triaca, por haberme acompañado tan de cerca durante todos estos años.

Estoy especialmente agradecido a mis tutores Laura y José, no solo por haberme enseñado tanto sobre las enfermedades neurodegenerativas y guiado inmensamente a lo largo del año, sino también por su invaluable apoyo durante la elaboración de este trabajo.

No habría sido posible realizar este trabajo sin vosotros.

Dedicatoria

A mi abuela, a mis padres y a mis hermanos.

Chapter 1

Introduction

Neurodegenerative diseases are complex diseases that take place in the patient's brain, causing the progressive loss of neurons and its structure, a biological process known as neurodegeneration. A wide variety of neurodegenerative diseases exist, and the interceptions and treatments differ greatly from one disease to another. These diseases cause a great sanitary cost, as the affected patients need to be in constant monitorization both prior and after neurodegeneration begins in order to slow down the cell loss rate. Neurodegenerative diseases also have an economic impact, since the treatments employed are complex and thus expensive. Lastly and more importantly, the social impact of these diseases. Because of the cognitive impairment, patients with neurodegeneration cannot live their life in a normal way, depend on caretakers to perform their daily life activities and can ultimately end up developing dementia, something that impacts in their psychology, reducing their chances to slow cognitive decline, and also their social environment. Therefore it is essential to understand how neurodegeneration takes place and develop accurate techniques to diagnose and prevent it. Unfortunately, most neurodegenerative diseases are extremely complex and anticipating and curing their biological mechanisms is generally unknown. Examples of these neurodegenerative diseases are Parkinson's or Huntington's diseases, although the most common one is Alzheimer's Disease, where one in 10 Americans are affected by it.[1]

Alzheimer's Disease (AD) is a brain disease that mainly affects cognitive functions, beginning with mild memory loss, worsening to the extent of being unable to respond to

the environment. As stated, these symptoms are caused by neurodegeneration meaning a loss of neurons and its inter-connections, resulting in brain atrophy. This particular disease is characterized by three main clinical stages: preclinical AD, mild cognitive impairment (MCI), and finally dementia due to AD[2]. Depending on the current stage, diverse changes take place on the brain, with no or varying degrees of cognitive symptoms. Figure 1.1 shows these stages as well as the cognitive symptoms for each one of them.

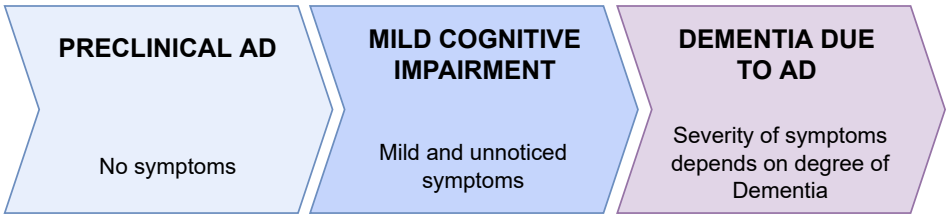


Figure 1.1: *Three main stages and symptoms in Alzheimer's disease continuum. Note that the preclinical stage is the ideal stage to anticipate Dementia before the biological mechanisms take place.*

These stages and their particularities can be studied with the use of certain biomarkers present among the affected individuals as well as other assessments. Biomarkers are substances that indicate a biological status which are used to detect diseases before they take place as well as evaluate their progression once they have. Said biological status can be static measures such as the genotype of a person, or a varying measure, for instance molecular tests. The biological features of AD are described by the accumulation of two proteins in their pathogenic forms: beta-amyloid and Tau, forming plaques and neurofibrillary tangles, respectively. Therefore, one example of a very common biomarker is the progression of increased beta-amyloid levels in Positron Emission Tomography (PET) scans. However, similarly to the other neurodegenerative diseases, AD's main challenge is to foresee when these mechanisms will take place and their evolution in time. All these mechanisms already trigger in the preclinical stage, long before any symptoms start taking place. Consequently, it is essential to study these mechanisms and anticipate their progression over time and implication at the clinical level, for instance the diagnosis transitions and their potential

relation with other important biomarkers such as in individuals' genetics.

While these diseases are extremely complex, the medicine field has also made advances such as reseaching new possible ramifications of medicine. It is the case of "P4 medicine"[3], an approach to make medicine more predictive, preventive, personalized and participatory, with the major objective of improving wellness and anticipating the appearance of a disease's symptoms. This has benefited the treatment of neurodegenerative diseases, especially the personalized approach.

Due to the complexity of AD, placing each patient in one of three specific stages in the Alzheimer's Disease continuum is limiting at best, as a considerable amount of information is lost and the differences between patients at a given stage are significant. Therefore, instead of visualizing stages, a more personalized perspective is the use of a gradient, where patients with similar biomarkers are placed close to each other. Because of the complexity in patient diagnosis and the amount of people affected by Alzheimer's Disease, a very efficient tool to compare patients together is the use of Machine Learning techniques. Machine Learning is a discipline of Artificial Intelligence that allows systems to make observations and learn from patterns in data without any kind of human intervention or assistance. Its many applications range from data processing software in fields such as finance and data security to robotics capable of performing operations on patients with the use of sensors.

Machine Learning has many uses due to the three types of classifications it is capable of, these being unsupervised, supervised and reinforcement learning. Unsupervised learning or clustering is a technique that is based on grouping together samples with similar characteristics. Therefore it is a useful tool in clinical applications because it permits to find new related groups of patients with the disease, which ultimately offers new knowledge about how the disease occurs without biasing the result towards previous knowledge about it. When treating with AD, clustering is the most interesting technique as it learns from the input data without the use of labels, which in our application field would represent the patients' diagnoses. Moreover, this technique is also beneficial as it avoids taking into

consideration possible misclassifications such as preclinical AD classified as control or other dementias classified as AD.

In clustering, data can be represented in several ways, but for the present work, since we want to consider the evolution of the patient, a useful tool is through the study of time series. A time series is represented by two or more points in time, where each indicates a certain value for that given instant. Therefore, including the temporal component allows us to study the disease according to its development, since AD, being a neurodegenerative disease, it is characterized by the progressive loss of neurons, thus studying AD from a static point of view would be erratic. Therefore, for this work, we have used one of the most significant studies in this regard: the Alzheimer's Disease Neuroimaging Initiative (ADNI).

Due to the effectiveness of Machine Learning techniques and the evolution of medicine during the last few years, multiple works of similar characteristics have been carried out. For example, many studies apply the ADNI database to further comprehend the characteristics of AD, for instance one[4] which was published in 2011, makes use of said database, specifically the MRI regions to classify patients in different stages of the Alzheimer's disease continuum making comparisons between pairs of stages. Concerning time series analysis, there has been a considerable improvement in the last decade, where works such as Elizabeth Bradley's and Holger Kantz's[5] explain how to analyze nonlinear time series, a complex type of time series that does not necessarily have the same length but still carries useful data. Time series have been constantly used in works regarding neurodegenerative diseases, for instance a work[6] published in 2016, which studies the fatal effect of heat waves on patients with Parkinson's disease through time series analysis. Regarding algorithms that employ temporality, a recent work[7] published in 2018 employs Dynamic Time Warping (DTW), a distance metric used specifically for time series clustering, where similar temporal patterns are identified regardless of the instant at which they appear. In general, these works demonstrate that time series analysis is an interesting tool that can be employed in any field where temporality is taken into account.

This work presents an exhaustive exploration of several unsupervised learning techniques (KMeans, Hierarchical clustering, Self Organizing Maps, and Dynamic Time Warping) on different biomarkers both analyzed separately as well as jointly (pairs of biomarkers). Moreover these biomarkers were in some occasions modified in order to incorporate temporality.

The main objective of this Degree Thesis was to explore and select the best clustering techniques to group patients together based on the time series data collected for several biomarkers related with Alzheimer's Disease. Secondary objectives of this work were: (**i**) to evaluate from the analysis or computational point of view the obtained results and analyze the similarities and differences between clustering results, i.e. which clustering algorithms and biomarkers work best and why; and (**ii**) to evaluate from a clinical point of view the obtained results through the use of the available diagnoses and several sociodemographic features of the patients.

This manuscript is organized as follows. In Chapter 2, we explain the different materials and methods used to carry out this work, these being the treatment of the input data, as well as the algorithms and evaluation metrics used. In Chapter 3, we display and describe the results we've obtained by applying the aforementioned methods to the ADNI dataset, in both non-temporal and temporal models. In Chapter 4, we discuss the results for each kind of model and give interpretations on both the computational and clinical points of view. Finally, Chapter 5 provides the main conclusions extracted in the present work.

Chapter 2

Materials and Methods

2.1 Database description

The data used for this work originates from the Alzheimer's Disease Neuroimaging Initiative (ADNI) (adni.loni.usc.edu). The main goal of this database is to test whether several neuroimaging techniques, biological markers, and clinical and neuropsychological evaluations can be combined to define the progression of Alzheimer's Disease (AD). This database contains data from several clinical assessments such as Magnetic Resonance Imaging (MRI) and Positron Emission Tomography (PET) neuroimaging, genetics, cognitive tests as well as cerebrospinal fluid (CSF) biomarker measurements in order to be used as predictors for the disease. From the ADNI database, we were able to obtain the ADNIMERGE key table, which contains the most important clinical information, as well as some genetic information for each patient.

At the time we collected the data (November 2021), this database contained 2.353 subjects distributed in 15.637 medical visits with a total of 115 variables per visit. Each patient has several time points in which the above-mentioned biomarkers were measured. The first visit (t = 0) is named 'baseline'. Clinical visits are identified by the number of months since the baseline visit (3, 6, 12, etc.). This number is represented by the variable named "VISCODE". Figure 2.1 shows the number of datapoints available for each VISCODE. As it can be observed in this Figure, the number of available datapoints varies greatly depending

on the VISCODE, denoting that there will be a limit to the length of our timeseries. We will return to this issue in the next section, dedicated to data filtering.

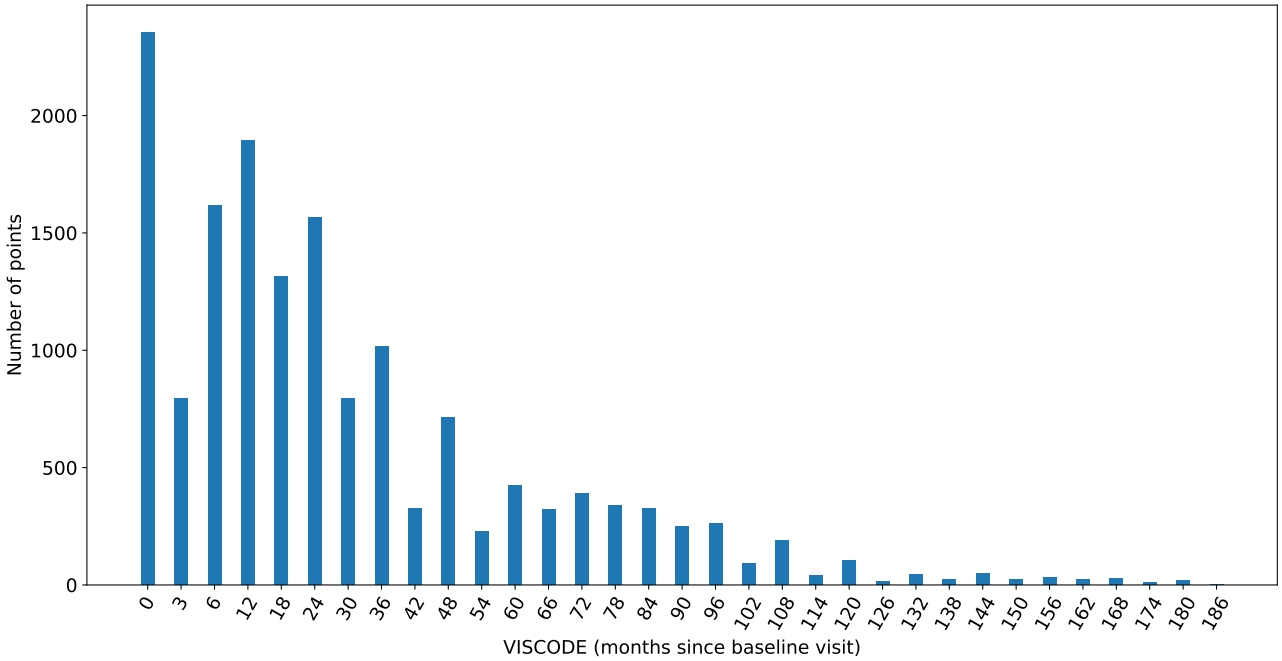


Figure 2.1: *Distribution of the amount of data points in function of the VISCODE.*

We used several features coming from the ADNIMERGE key table, which represent different Alzheimer's Disease biomarkers. These variables were as named in ADNI: ABETA, TAU, PTAU, AV45, FDG and MRI, the latter being a grouping of six different brain region volumes. Table 2.1 is a summary of all the features used in this work.

Firstly, TAU, PTAU and ABETA are proteins measured on cerebrospinal fluid (CSF). TAU and PTAU, the latter being much more exclusive to Alzheimer's disease, measure the amount of TAU protein released during neurodegeneration. ABETA represents the quantity of beta-amyloid in the CSF. Secondly, AV45 and FDG are two neuroimaging biomarkers, both obtained through PET scans, employing different radiotracers (Florbetapir and Fluorodeoxyglucose, respectively). AV45 measures the average beta-amyloid burden in several regions of the brain, while FDG measures the glucose metabolism of the brain. Lastly, we obtained the volume of six regions of the brain: the ventricles, the hippocampus, the

Table 2.1: *Summary of the variables from ADNIMERGE key table used in this work.*

Variable	Type of variable	Description	Worst diagnosis
AV45	Biomarker Predictor variable	Measures beta-amyloid in several brain regions (average) through PET scans	Higher values
ABETA	Biomarker Predictor variable	Measures beta-amyloid in CSF	Lower values
TAU	Biomarker Predictor variable	Measures Tau protein in CSF	Higher values
PTAU	Biomarker Predictor variable	Measures phosphorylated-Tau protein in CSF	Higher values
FDG	Biomarker Predictor variable	Measures glucose metabolism in several brain regions (average) through PET scans	Lower values
MRI	Biomarker Predictor variable	Several variables: *Ventricles*, *Hippocampus*, *Entorhinal*, *Fusiform*, *Midtemp*, *WholeBrain* Measures the volume of each region through MRI	Lower values, except for *Ventricles* (higher values)
PTGENDER	Sociodemographic Descriptive variable	Sex of patient (Male/Female/Unknown)	Non-applicable
PTEDUCAT	Sociodemographic Descriptive variable	Number of education years	Non-applicable
MMSE	Cognitive decline Descriptive variable	Score obtained in MMSE cognitive test	Non-applicable
APOE4	Genetics Descriptive variable	Genotype e4 of APOE gene	Non-applicable
DX	Diagnosis Descriptive variable	Diagnosis in the AD continuum given at each patient: CN, MCI or Dementia	Non-applicable

entorhinal, fusiform, the middle temporal gyrus, and the whole brain, all of them measured by MRI. We have also obtained one last biomarker called ICV (Intracranial Volume), which measures the total volume of the patient's brain. These biomarkers are clinically interpreted in different ways. Regarding TAU, PTAU and AV45, the higher their values, the closer the patient is to being affected by neurodegeneration. Meanwhile, ABETA, FDG and MRI (excluding the ventricles region) work in the opposite way: lower values represent a more pathological outcome. These mechanisms can also be seen in Table 2.1.

Finally, to describe our results we obtained different features representing other demographic and clinical data of the patients: the education years (PTEDUCAT), their gender (PTGENDER), their age at baseline (AGE), their Mini-Mental State Examination score (MMSE) and the APOE e4 genotype (APOE4). MMSE is a widely used test of cognitive function, which scores out of 30. The MMSE variable represents the score that each patient obtained for each visit and measures how much they are affected by cognitive decline, where

the lower the resulting score, the higher the patient's neurodegeneration. Although it is less accurate than biomarkers, it still is an effective tool for learning the progression of the patient and using it as demographic information. Moreover, the APOE4 variable represents the presence or absence of a genetic mutation highly associated with the development of dementia due to AD.

Lastly, we queried the diagnosis of the patient (DX). This variable indicates the diagnosis of a given patient at a given visit. Their possible values are: Cognitively Normal (CN), Mild Cognitive Impairment (MCI), or Dementia; each representing a stage in the so-called 'Alzheimer's disease continuum'. This concept is represented in Figure 2.2, where each patient goes through the three different diagnoses mentioned before showing increased cognitive symptoms over time due to increased neurodegeneration.

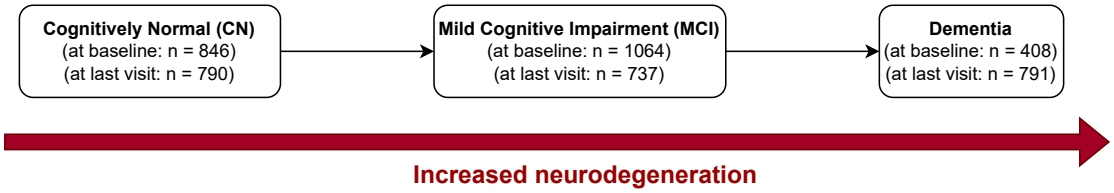


Figure 2.2: *Description of the different stages in Alzheimer's Disease continuum as well as the possible transitions among them. At each stage, the amount of patients for that given diagnosis is displayed.*

Patients that are in the CN stage usually have healthy biomarkers far from reaching dangerous values as well as a high MMSE score, contrary to the Dementia stage. Furthermore, these diagnosis transitions are one-way, which means that once a patient enters an advanced stage such as MCI or Dementia, they cannot be moved back to a healthier stage. Figure 2.2 also shows the number of patients with each diagnosis made at both the beginning (baseline visit) and the end (last visit) of the timeline. At baseline, there were 846 (35.95%) CN subjects, 1064 (45.22%) patients with MCI diagnosis, and 408 (17.34%) patients diagnosed with dementia. The remaining 35 (1.49%) did not have a diagnosis assigned to them. At the last visit for each patient, 790 (33.57%) patients were diagnosed with CN, 737 (31.32%)

patients had a MCI diagnosis and 791 (33.62%) suffered from Dementia, with the same 35 (1.49%) previous patients that did not have a set diagnosis. As we can see, the number of patients with Dementia has increased since the baseline, therefore this database is suitable to study the progression of AD among these patients.

2.2 Data filtering

We divided the entire dataset into smaller subsets, each being biomarker specific to meet with all the filtering criteria, since a patient that might have all their datapoints available for a given biomarker, may also contain only a few for the rest of the biomarkers. Hence, each patient's time series in the dataset has to go through a filtering process in order to make sure it is accurate and does not show ambiguities. Figure 2.3 shows the filtering steps carried out in this work.

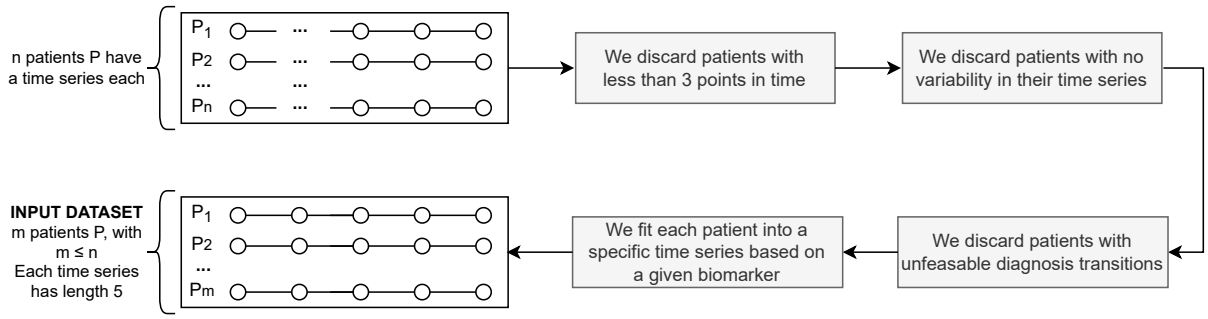


Figure 2.3: *Description of each stage in the filtering process.*

Regarding this biomarker division, and the distribution of datapoints shown in the previous Section, we plotted in Figure 2.4 the number of datapoints available for each of the biomarkers used in this work as predictor variables. As it can be seen in Figure 2.4, for each biomarker, the number of available points has a different frequency of measurement over time.

Therefore, we selected for each biomarker dataset the patients that had at least three measurements over time for that given biomarker. The threshold value of three visits was

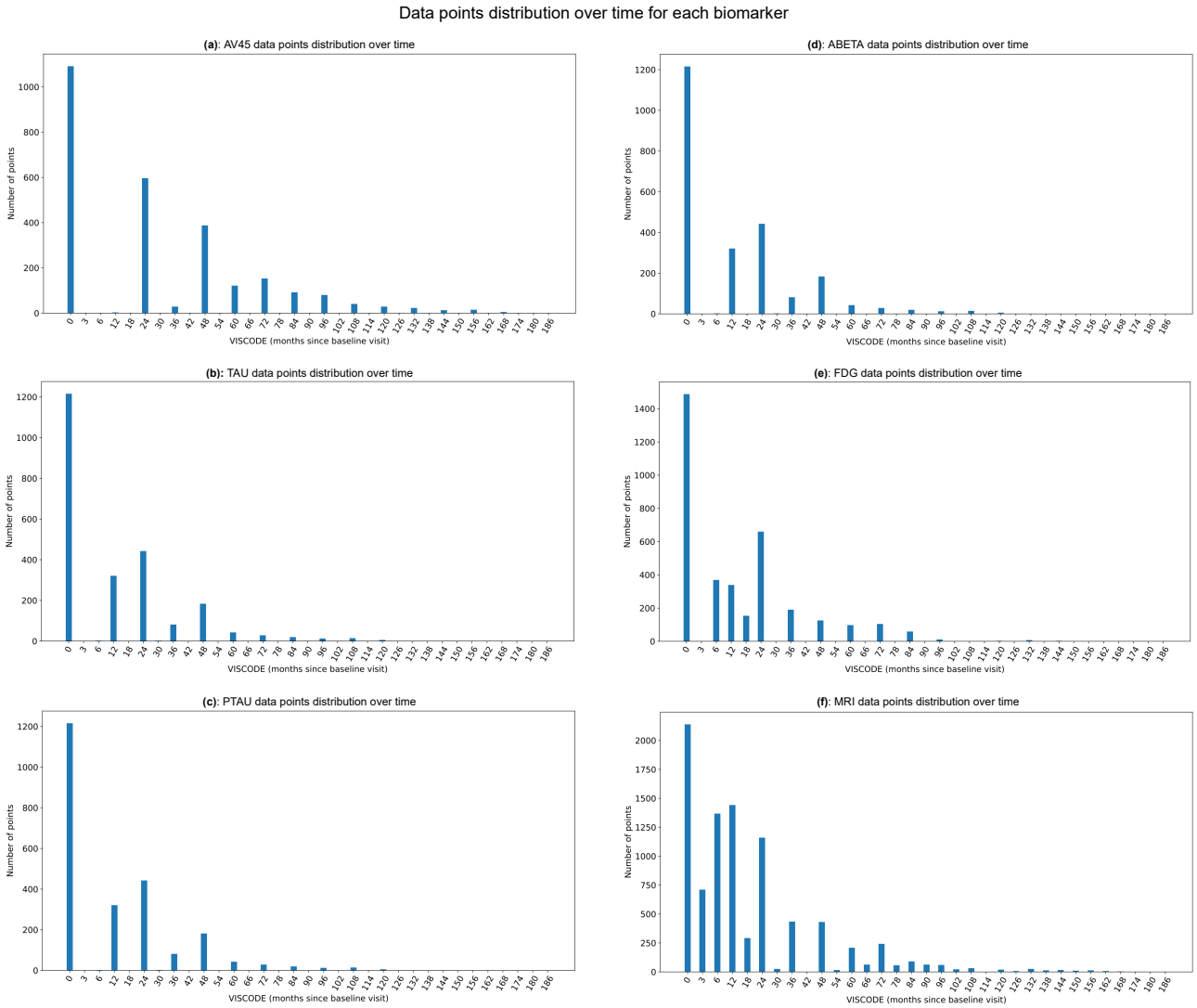


Figure 2.4: *Distribution of the amount of data points in function of the VISCODE for each biomarker. Note that only patients with at least 3 points over time have been taken into account, since those with less points are not considered when defining the timeseries.*

decided due to two main reasons. In the first place, selecting a smaller number of needed measurements (1 or 2) would make our data more inaccurate. Moreover, selecting a bigger number as the threshold (e.g. at least 4 points) would have left us with barely any patients when combining biomarker datasets.

Secondly, we removed patients with no variability in their biomarker measurements, which kept their progression constant.

Thirdly, we discarded patients depending on their diagnosis transitions. The diagnosis transition is defined by their diagnosis at baseline and last medical visits. We removed subjects that stayed with a normal cognitive profile during all visits (CN to CN). We also removed patients that presented unfeasible diagnosis transitions, different from those shown in Figure 2.2, such as going from Dementia to MCI, from Dementia to CN, or transitioning between diagnoses more than twice during their whole timeline.

Finally, we selected a relevant time series for each biomarker; e.g, measurements taken every six months. This time series was defined taking into account how frequently the patients' measurements are updated for each biomarker (Figure 2.4) and as described in ADNI's protocol in their webpage[8]. Datapoints outside this time series were not considered since they would only be available in a small proportion of the total amount of patients, and it is crucial that all patient time series among a given biomarker have the same timeline. As a result of the filtering step, Table 2.2 shows the number of datapoints before and after the filtering process described above as well as the corresponding time series (in months) for each biomarker.

Table 2.2: *Number of data points before and after the filtering process and corresponding timeseries for each biomarker*

Biomarker	**Before filt.**	**After filt.**	**Timeseries**
TAU	2370	1180	(bl, 12, 24, 36, 48)
PTAU	2369	1175	(bl, 12, 24, 36, 48)
ABETA	2370	1175	(bl, 12, 24, 36, 48)
AV45	2678	1645	(bl, 24, 48, 72, 96)
FDG	3605	1760	(bl, 6, 12, 18, 24)
Ventricles (MRI)	8955	6805	(bl, 12, 24, 36, 48)
Hippocampus (MRI)	8317	6335	(bl, 12, 24, 36, 48)
Midtemp (MRI)	7907	5975	(bl, 12, 24, 36, 48)
Entorhinal (MRI)	7907	5975	(bl, 12, 24, 36, 48)
Fusiform (MRI)	7907	5975	(bl, 12, 24, 36, 48)
WholeBrain (MRI)	9210	6930	(bl, 12, 24, 36, 48)

2.3 Data preprocessing

We then performed multiple preprocessing steps over all datasets meaning missing values processing and several data transformations.

2.3.1 Missing values

As described in Section 2.2, by fitting each biomarker into a given time series specified by ADNI[8] and described by the dataset we considerably reduced the number of points without measurements, known as missing values. However, even with the recommended time series for each biomarker, some measurements over time were still missing. We input these missing values depending on the time point they were found.

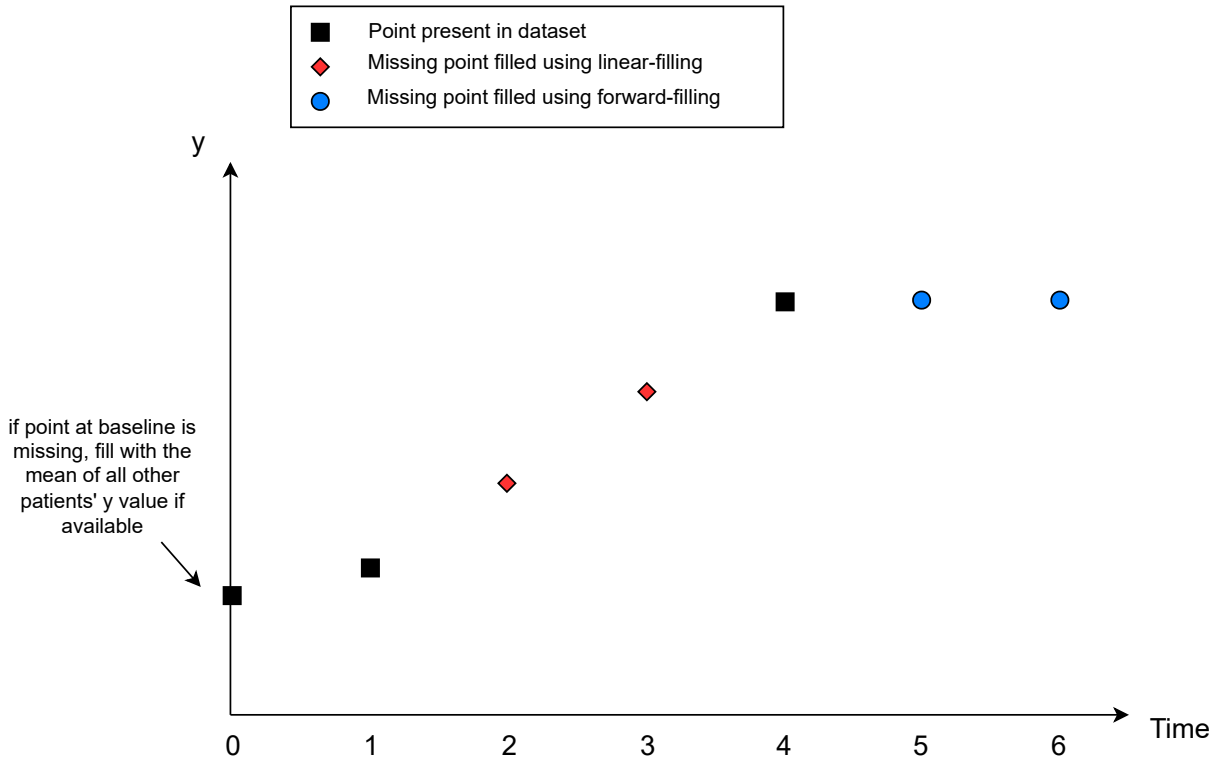
For missing values at the baseline time point ($t = 0$) for a given biomarker, we input the mean of all available patients' values in baseline for that biomarker. This way, the variance of the dataset is kept constant, which will be useful when we begin using unsupervised classification.

Missing values found at non-baseline and last time points were filled using the methods known as 'forward filling' and 'linear filling'[9], respectively. Figure 2.5 summarizes both of these methods.

Missing values appearing at non-baseline and non-last time points were filled using 'linear-filling'[9]. This method uses linear interpolation in order to keep the progression linear, which works as follows. Let y be the missing value and x its position in the time series. Moreover, let y_0, x_0 be the same parameters for a known point before x; and y_1, x_1 the same parameters for a known point after x. As long as there are two known points in the time series, all other points in between can be computed via interpolation with the formula:

$$y = y_0 + (x - x_0) * \frac{(y_1 - y_0)}{(x_1 - x_0)}$$

This method can be used for any time series in our dataset since the baseline is always present (if it is not, it is filled in as explained above) and there are at least two other known



Time	0	1	2	3	4	5	6
Value of feature y	y0	y1	N/A	N/A	y4	N/A	N/A

Figure 2.5: *Explanation of the methods 'forward filling' and 'linear filling' used to fill in missing values.*

values in the time series.

Lastly, missing values appearing at the end of the time series (last visit) were filled using 'forward-filling'[9], in order to keep the remainder of the time series constant. Let y_i be the last known value in our time series of length n and x_i its position in the series, with $i < n$. Let also $[y_{i+1}, \ldots, y_{n-1}, y_n]$ be missing points after y_i. With 'forward-filling', $y_k = y_i$ where $k \in [i+1, n]$.

2.3.2 Data transformations

We performed several data transformations regarding the clinical adjustements needed for some measures and the type of datasets we wanted to build.

For the datasets involving the MRI variables, the brain regions' volumes in a patient are all dependent on their intracranial volume (ICV). This means that while two patients may have an identical volume on a certain brain region, they in fact depend on their ICV. Therefore, we normalized at each point in the timeline, the patients' brain region volumes by their own ICV.

Because we wanted to take into account the temporality of the data while keeping them in tabular format, we carried out the following transformation for AV45, ABETA, TAU, PTAU, FDG and MRI biomarkers. We calculated their polynomial coefficients to perform a feature extraction. For this, we fit each time series into the equation $ax^2 + bx + c$, with a, b and c polynomial coefficients.

Regarding datasets combining different biomarkers, since each biomarker has its own range, we normalized the values of each biomarker in order to keep them in the interval [0,1], therefore biomarkers with higher values will not add greater weight during the clustering decisions than biomarkers with a smaller range of values. Lastly, when the unsupervised classification algorithm applied needed it, we standardized the data so that it had a mean of 0 and a standard deviation of 1.

2.4 Unsupervised classification

The aim of this study is to group these patients together by means of unsupervised classification or clustering. As such we used the following algorithms with tabular-formatted datasets: (**i**) Agglomerative Hierarchical Clustering, (**ii**) K-Means, (**iii**) Density-Based Spatial Clustering of Application with Noise (DBSCAN) and (**iv**) Self-Organizing Maps (SOM). Regarding time series datasets, we used Dynamic Time Warping combined with (**v**) Hierarchical or (**vi**) DBSCAN clustering methods.

Regarding all Section 2.3, we built six different types of datasets which will be evaluated in Chapter 3 using the described unsupervised classification algorithms. The datasets names and the unsupervised classification methods applied to each of them are shown in Table 2.3.

Table 2.3: *Datasets and unsupervised algorithms applied in this work.*

Biomarker type	Biomarkers employed	Algorithm
Tabular One biomarker	AV45, ABETA, TAU, PTAU, FDG, MRI	KMeans Hierarchical SOM
Tabular Pair of biomarkers	AV45-ABETA, AV45-TAU, AV45-PTAU, AV45-FDG, ABETA-TAU, ABETA-PTAU, ABETA-FDG, TAU-PTAU, TAU-FDG, PTAU-FDG	KMeans Hierarchical SOM
Tabular Several biomarkers	MRI regions: Ventricles, Hippocampus, Entorhinal, Fusiform, Midtemp, WholeBrain	KMeans Hierarchical SOM
Polynomial transformation One biomarker	AV45, ABETA, TAU, PTAU, FDG, MRI	KMeans Hierarchical SOM
Polynomial transformation Pair of biomarkers	AV45-ABETA, AV45-TAU, AV45-PTAU, AV45-FDG, ABETA-TAU, ABETA-PTAU, ABETA-FDG, TAU-PTAU, TAU-FDG, PTAU-FDG	KMeans Hierarchical SOM
Polynomial transformation Several biomarkers	MRI regions: Ventricles, Hippocampus, Entorhinal, Fusiform, Midtemp, WholeBrain	KMeans Hierarchical SOM
Timeseries One biomarker	AV45, ABETA, TAU, PTAU, FDG, MRI	DTW + Hierarchical DTW + DBSCAN
Timeseries Pair of biomarkers	AV45-ABETA, AV45-TAU, AV45-PTAU, AV45-FDG, ABETA-TAU, ABETA-PTAU, ABETA-FDG, TAU-PTAU, TAU-FDG, PTAU-FDG	DTW + Hierarchical DTW + DBSCAN
Timeseries Several biomarkers	MRI regions: Ventricles, Hippocampus, Entorhinal, Fusiform, Midtemp, WholeBrain	DTW + Hierarchical DTW + DBSCAN

The next subsections describe all of the mentioned clustering algorithms.

2.4.1 Agglomerative Hierarchical Clustering

Hierarchical clustering is a clustering technique that, depending on its type, merges data points into one cluster or divides a cluster into multiple clusters. Both types can be represented as a dendrogram to know which clusters were merged together or split up. Hierarchical clustering can use an agglomerative or bottom-up approach, where each data point is its own cluster and according to their distance with other clusters they are merged together or left alone. Meanwhile, there is the divisive or top-down approach, where all data points

are considered one unique cluster and based on their distance with each other they are split up into smaller clusters. While both types of hierarchical clustering yield the same results, we chose the agglomerative type over the divisive one since a bottom-up approach seemed easier to analyze over a top-down one. This explanation can be further visualized in Figure 2.6.

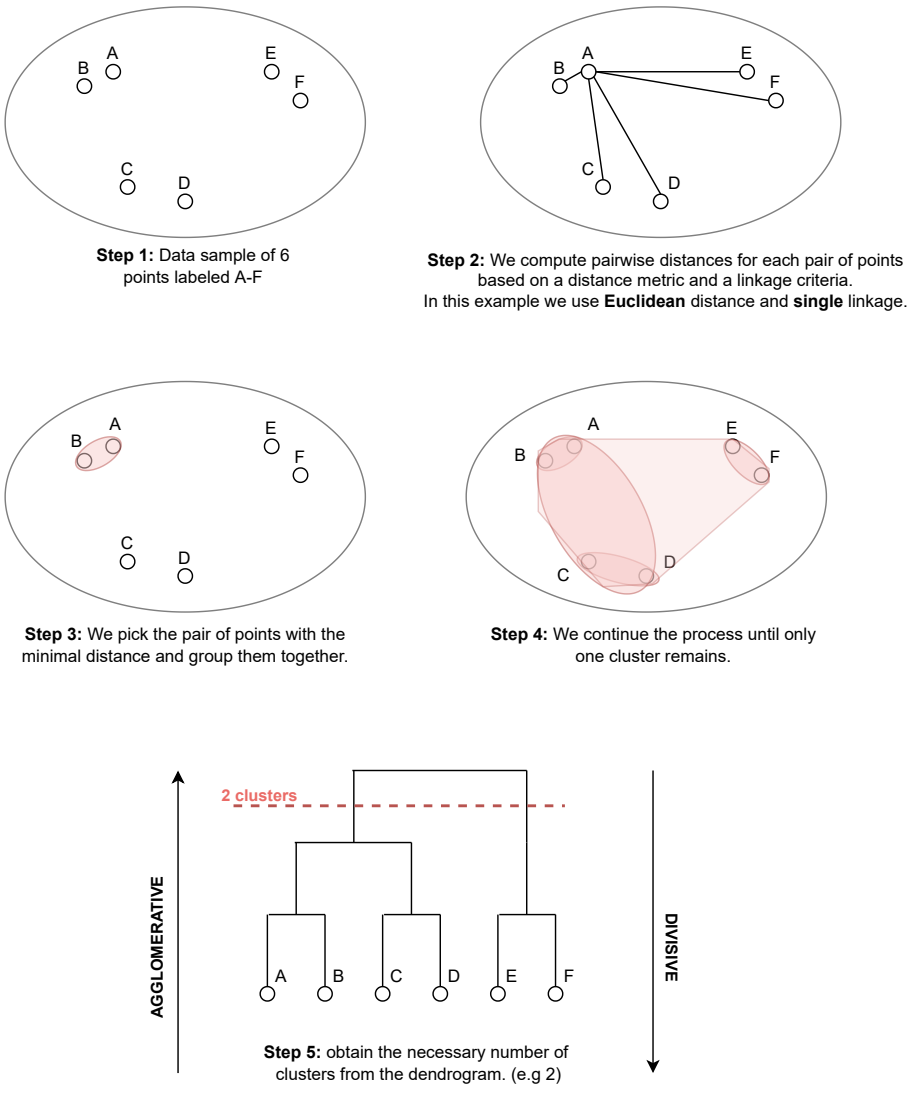


Figure 2.6: *Explanation of the agglomerative hierarchical clustering method.*

As mentioned previously, the agglomerative hierarchical clustering first considers all data points as their own clusters and groups them together based on two parameters: the distance

metric and the linkage criteria. The distance metric is used to determine the separation between each cluster using one of the available metrics such as the Euclidean distance[10], which measures the linear distance between two points, or the Manhattan distance[10], which is the sum of the absolute difference between the measures in all dimensions of two points, among many others. The linkage criteria is the basis by which each pair of clusters is grouped together based on their distance. These criteria can be single-linkage -where we group together pairs of clusters that are closest to each other-, complete-linkage -where we group together pairs of clusters that are the furthest from each other-, centroid-linkage -where pairs of clusters are grouped together based on the closest centroid distance- or average-linkage -where we group together pairs of clusters that have the shortest mean distance between pairs of elements from each of the clusters-, among others. The explained linkages are visually detailed in Figure 2.7.

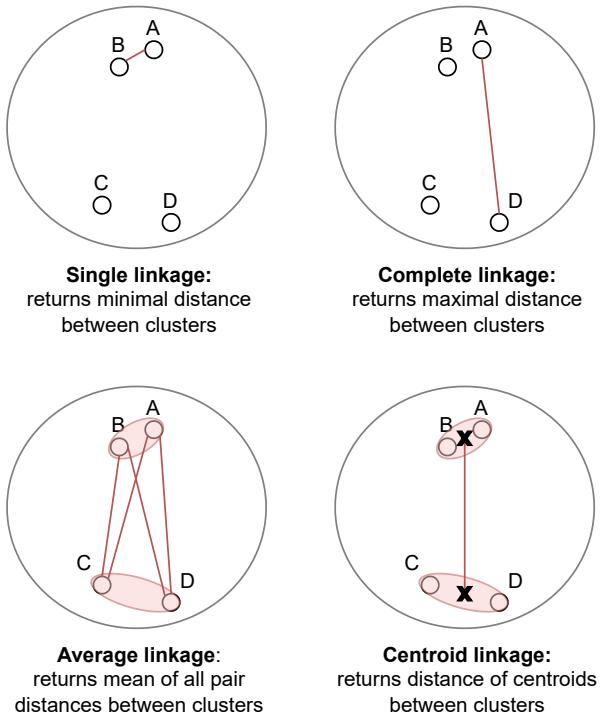


Figure 2.7: *Explanation of some of the different types of linkage that can be used in hierarchical clustering.*

Since there are multiple metrics and linkages that can be used to cluster our data, we

performed a grid-search considering each combination metric-linkage. In section 3, we kept for each biomarker the combination that returned the best outcomes based on the evaluation metrics and analyzed the results.

2.4.2 K-Means

K-Means[11] is a clustering algorithm that clusters data into K clusters while keeping each cluster's variance as minimal as possible. The methodology of this algorithm can be visualized in Figure 2.8.

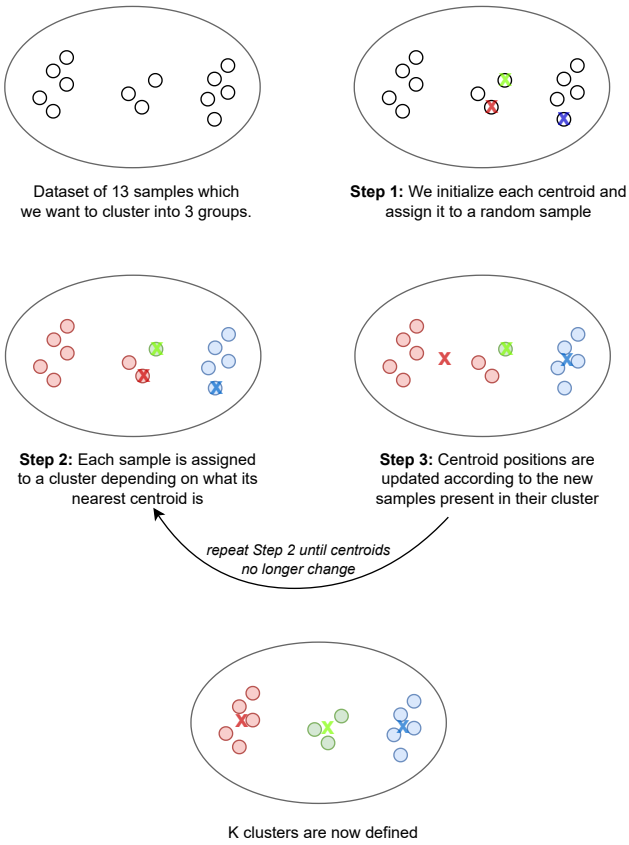


Figure 2.8: *Explanation of the K-Means clustering method.*

It functions by separating N samples from a dataset into K disjoint clusters. Each cluster is defined by a centroid, which is the mean u_k of all the samples contained among a cluster. At first, K centroids are placed either in K different samples or initialized randomly in the

data space. Then the euclidean distance between each sample to each centroid is computed, and those samples that are closest to one of the centroids now belong to that centroid's cluster. By the definition of a centroid, since now the samples contained in the cluster have changed, it has to be recomputed. We then recalculate the distances from each sample to each of the newly positioned centroids and repeat the steps until the centroids no longer move. Once the algorithm finishes, we can visually see the data as well as study the position of each centroid to know whether a cluster favors a group of features.

When performing K-Means, it is important to standardize the features used, since a variable such as ABETA that is much larger than AV45 would carry far more weight. Moreover, since the centroids are initialized at random, it is best to transform the data so it has a zero mean and unit variance.

We will use K-Means with different K parameters to know which amount of clusters suits the dataset best.

2.4.3 Density-Based Spatial Clustering of Application (DBSCAN)

The Density-Based Spatial Clustering of Application with Noise (DBSCAN) [12] is a unsupervised classification technique that groups together samples located in high density regions in the sample space. This algorithm needs two parameters that greatly change the outcome of the clusters: the eps value and the minimum samples required to form a cluster. Let us define three types of nodes in the sample space as shown in Figure 2.9.

First there is the core node which represents a cluster and how far it reaches. Then there is the basic node which is part of a cluster and finally an outlier node, which is not part of a cluster yet. At first, all nodes are to be considered outlier nodes. Each node has a "neighborhood" assigned to it defined by the eps value. When two nodes, no matter their type, are separated by a distance lower than the eps value then they are part of each other's neighborhood. Moreover, a node is always part of its own neighborhood. Once a basic node or an outlier node has inside their neighborhood at least a minimum amount of samples

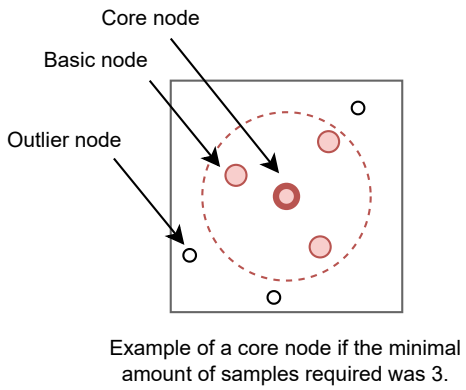


Figure 2.9: *Explanation of a core node and how it affects other types of nodes during DBSCAN.*

defined by the user, that node becomes a core node. The properties of a core node are that they represent a cluster, and every node inside that core node's neighborhood is also part of the cluster. If two core nodes are in each other's neighborhood, then they represent the same cluster. If an outlier node is inside a core node's neighborhood, then the outlier node will become a basic node and it will also be part of the core node's cluster. If an outlier node is not part of a core node's neighborhood by the end of the execution, then it will remain as noise in the dataset. Finally, if a basic node is in the neighborhood of two different core nodes representing different clusters, then the basic node will be part of the core node's cluster that integrated it first.

Due to the amount of values possible for the parameters, we decided to do a grid search that combined different values of eps and the minimal samples required.

2.4.4 Self Organizing Maps (SOM)

Self Organizing Maps[13] is a unsupervised machine learning technique used to cluster our data. The main difference between this technique and the aforementioned is that SOM is a type of neural network, which is unique in its own way since it is trained using competitive learning rather than backpropagation. Competitive training is what makes the nature of this neural network unsupervised, as neurons compete with one another to represent a subset of

the dataset. SOM's objective is to perform a feature extraction, where a high-dimensional dataset is transformed into a low-dimensional one, as well as creating a discretized representation of the dataset, where we can gather the clusters from. This machine learning algorithm initializes a map of $A \times B$ neurons specified by the user (Figure 2.10).

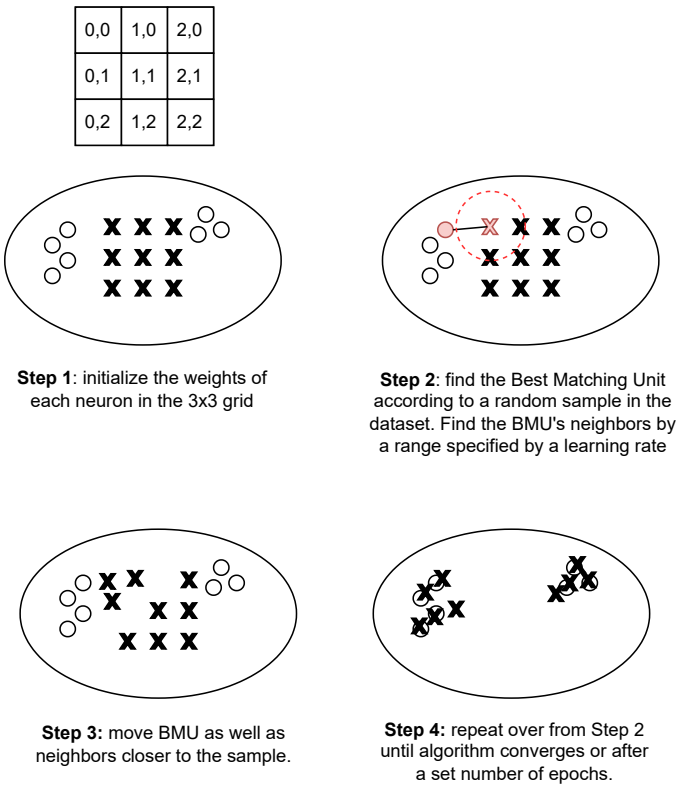


Figure 2.10: *Overview of the functionality of a Self Organising Map with its map of neurons.*

At first, the weight of each neuron can be precalculated or randomly initialized, the former allowing the algorithm to converge before the latter if done correctly. Let X be our dataset of dimension M and size N, where each node contains a set of weights $w_{i,j}$ with $i \in [0, N]$ and $j \in [0, M]$. The weight vector for each node is the position of the node in the input space. Therefore, the goal of each node is to have their respective weights as close as possible to a sample in the input space.

The algorithm goes as follows. The weights for each node in the map are initialized into the sample space. A data point from the sample space is chosen randomly and the

closest neuron to that data point is denominated as the Best Matching Unit (BMU). We then move the BMU's weights towards the data point as well as the BMU's neighbors, where the neighbors closest to the BMU move more than the ones farthest away. The distance at which the BMU and its neighbors are moved, as well as the BMU's range to consider its neighbors depends on a learning rate which decreases with each iteration so that the algorithm can converge. We perform these steps either until the SOM converges or a certain amount of epochs have passed.

After the algorithm converges, each neuron in the map moves towards a sample from our data space, although more than one sample can be defined by a neuron. Thus, we calculated the SOM nodes for each datapoint, meaning each datapoint is defined in the map by at least one neuron, which does not have to be unique. We completed the dimensionality reduction by applying K-Means on the new dataset. (Figure 2.11)

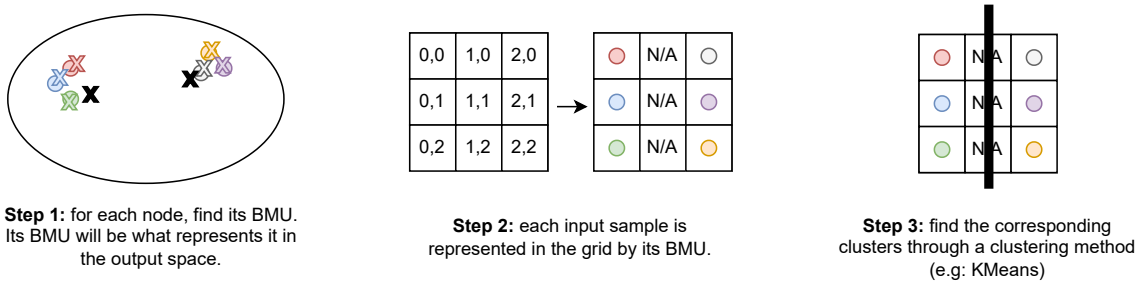


Figure 2.11: *Overview of the procedure that takes place after performing a feature extraction using SOM and using K-Means to obtain the clusters.*

For our study, we performed a grid search that varied on the number of neurons on the grid for both the length and the width as well as the number of clusters for the K-Means at the end.

2.4.5 Dynamic Time Warping with Hierarchical Clustering

Dynamic Time Warping (DTW) is an algorithm used to obtain a metric that measures the disimilarity between two time series, taking into account the speed at which they progress, differentiating it greatly of the other metrics like the Euclidean distance. It returns a positive

value that indicates how similar both temporal sequences are. An output of 0 indicates that they are identical while the larger the value, the more different they are. Until now, we have compared each time series with linear metrics, however we now aim to use DTW as a distance measure to compare the ADNI time series and instead of interpreting whether a biomarker's value is high or low, we are also interested in the progression of the time series itself throughout the visits. Figure 2.12 shows schematically the steps followed in DTW to process uni- and multidimensional datasets.

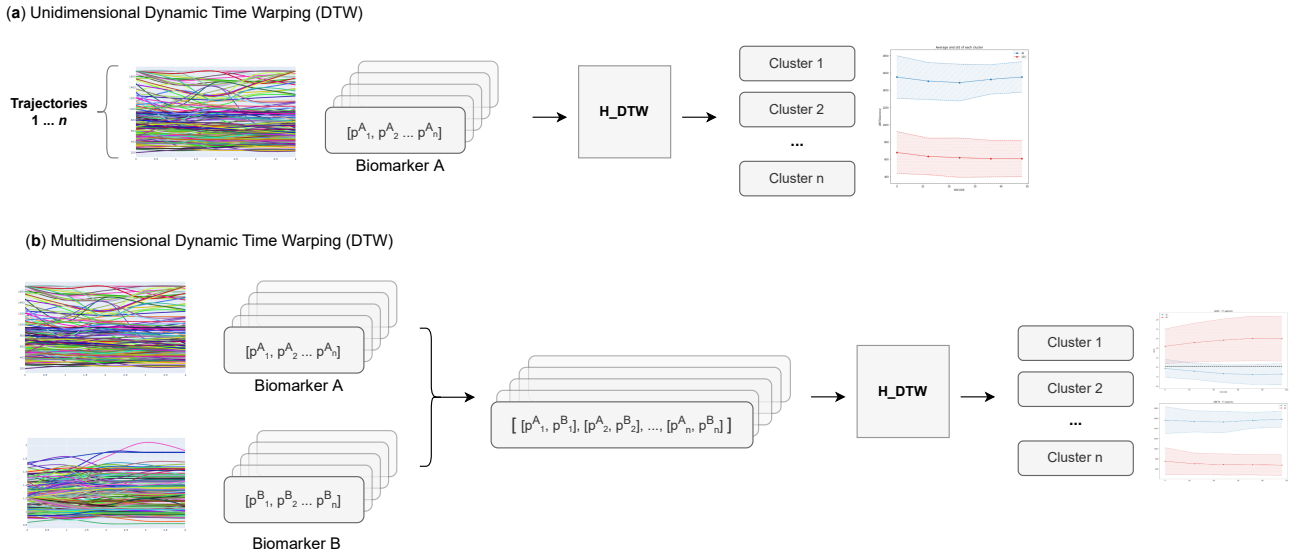


Figure 2.12: *Different steps in Dynamic Time Warping (DTW) in order to process (a) unidimensional and (b) multidimensional data.*

While this algorithm can be used for both uni- and multidimensional datasets, we will first go through its usability with the unidimensional ones. Firstly, we begin with a list of time series that we will use as input, where each element is one of the patient's measurements given by one of the biomarkers. Since we are working with unidimensional datasets, each time series will have a length of 5 measurements. The aim of the pipeline is to compute consecutively the DTW distances between each pair of trajectories, so that, using hierarchical clustering, we can group together those trajectories that are the least different, ending up with groups containing more than one time series which we will call clusters.

Let us call a "prototype cluster" a group containing more than one timeseries. Moreover, we define as "trajectory" both individual time series as well as prototype clusters. A prototype cluster has the possibility of becoming a cluster once all the algorithm's iterations are over, as long as it contains at least 10% of the total timeseries. Those trajectories that do not reach a minimum of 10% will be considered noise. Thus, with each iteration, we will group pairs of trajectories together into prototype clusters.

The DTW + Hierarchical clustering algorithm consists of multiple iterations that perform the following steps, as shown in Figure 2.13.

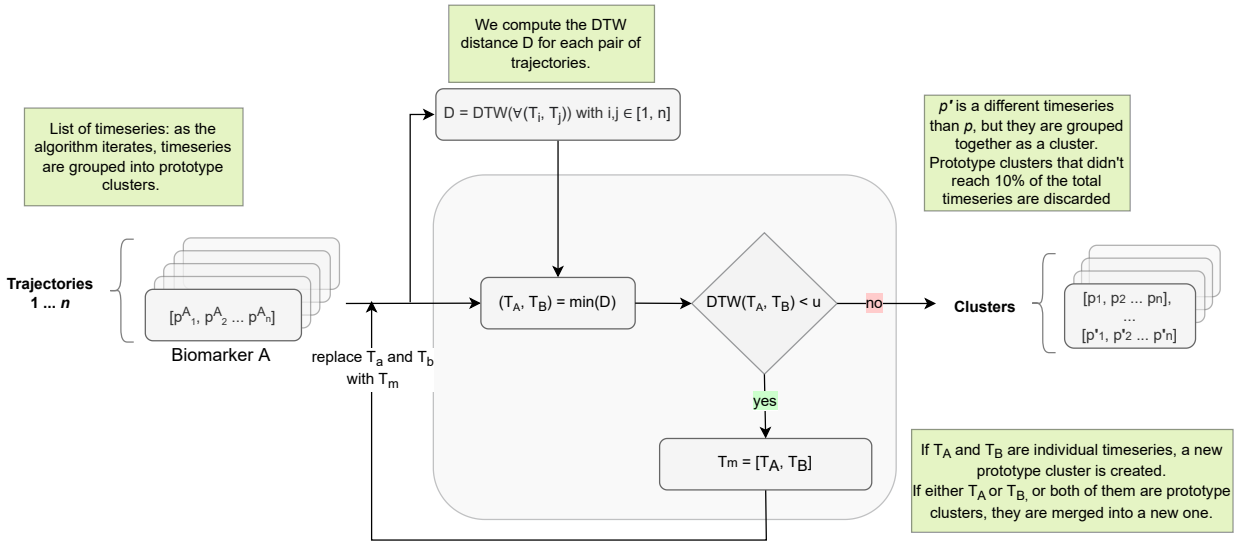


Figure 2.13: *Overview of the algorithm that combines DTW and hierarchical clustering.*

First, we obtain the minimal DTW distance between two trajectories T_i and T_j in our dataset and group them together into the same prototype cluster only if the DTW distance between them is lower than a user-specified threshold u.

$$DTW(T_i, T_j) < u$$

This threshold u allows us to keep a defined distance between the prototype clusters, and the lower the threshold, the more bounded and defined the groups of clusters will turn out to be. We repeat this process until there is only one cluster left -in which case the algorithm has

failed because the threshold was set too high-, or the minimal distance between trajectories is bigger than the threshold. Since we group trajectories together, and a trajectory can either be a time series or a prototype cluster made of two or more timeseries, we must define a specific formula to compute the DTW distance in these specific cases.

Let A and B be trajectories of one or more timeseries:

$$DTW(A, B) = op(DTW(\forall A, \forall B))$$

As hierarchical clustering has multiple linkage criteria that need to be considered, the operator *op* in the formula above depends on the chosen criteria. This operator can be replaced as follows. When we perform complete-linkage, we want to get the maximal possible distance between a pair of points of the two timeseries, therefore we can use max instead. When performing single-linkage, we seek the minimal distance between a pair of points of the timeseries, thus we replace *op* with min. We also performed average-linkage, by calculating the mean of all the computed distances, by using $mean$. Finally, we used median-linkage, finding the median of all the computed distances using $median$ instead.

Each linkage criteria has been used with each kind of dataset both in uni- and multidimensional ones. Regarding multidimensional datasets, for our algorithm to function correctly we modified the dimensionality of the input and used a different method for DTW computation.[14]

2.4.6 Dynamic Time Warping with DBSCAN

We also combined DTW with the DBSCAN algorithm since it excels in datasets with regions of high density and DTW seemed to be a promising lead. We first computed for each pair of trajectories their DTW distance and created a matrix of $N \times N$ with i,j corresponding to the distance between the trajectory of patient i and the trajectory of patient j. We then used this distance matrix as input data instead of the patients' timeseries. In order to obtain the best results, we performed a grid search to find the hyperparameters that worked best

with the algorithm, taking into consideration certain filtering criteria, such as a maximum of 20% of patients being considered as outliers, as well as each cluster having a minimum of 10% of total patients.

2.5 Evaluation Metrics

To evaluate how well each of our algorithms cluster the data, we have used two metrics: the Silhouette Score (SI), very often used in clustering applications, as well as the Progression Accuracy (PA), a custom metric that indicates the differences among each cluster from a clinical standpoint.

The Silhouette Score[15] is a useful metric when considering how dense and well defined the resulting clusters are. It ranges from -1 to 1, and the higher the score, the more defined the clusters will turn out to be. If the score is in the negative range, it usually considers the clustering to be incorrect, while a score close to 0 signifies that the clusters are overlapping each other.

Since this score only determines how well the clustering algorithm is performed from a computational viewpoint, it is also necessary to evaluate the clustering performance from a clinical point of view.

The Progression Accuracy is a metric that aims to compute the quality of the diagnoses separation between clusters. We consider a patient that progresses further in the Alzheimer's Disease continuum as one that started with a certain diagnosis and has progressed towards a more advanced stage by their final diagnosis. These patients will be considered as active. Contrarily, a patient that has a static diagnosis transition between the baseline and their last visit is considered to have a slower progression towards Dementia. These other patients will be considered as static. Therefore this metric computes the proportion between active and static patients, favoring clustering results that show considerable difference in diagnosis progression between clusters.

We want to find the PA value of N clusters. Let S be the combinations of all pairs of

clusters without repetitions, such as:

$$S = (i, j) | i \in [0, N], j \in [0, N], i < j$$

To calculate this metric, we first need to compute the progression ratio (PR) between all pairs of clusters in S.

Let C_i and C_j be two clusters, with A_i and S_i being the amount of active and static patients for C_i respectively, and A_j and S_j the amount of active and static patients for C_j. Then, the PR between clusters C_i and C_j can be considered as:

$$PR(C_i, C_j) = \frac{max(A_i + A_j, S_i + S_j)}{A_i + A_j + S_i + S_j}$$

Thus, in order to find the PA between N clusters:

$$PA(C_1, \ldots, C_n) = mean(PR(C_i, C_j)) \; with \; (i, j) \in S$$

We will use these metrics to consider for each clustering both the computational and the clinical standpoints and compare each clustering algorithm with another.

When comparing how each dataset worked with each type of algorithm, we will describe the best clustering result in terms of the Progression Accuracy. This description is done from the point of view of diagnosis change and demographic variables as previously described. Additionally, when working with temporal models, we computed for each biomaker the regression line $y = Ax + B$ of each cluster so that we could measure more precisely how each cluster progressed over time, depending on the value of A. Finally, we analyzed each cluster using the demographical data at our disposal. We performed a t-test for each variable so that, by obtaining its p-value, we could conclude whether there was a significant difference between clusters regarding these demographic variables. A result is considered significant if its p-value between clusters is less than 0.05.

Chapter 3

Results

As introduced in Section 2.4, we applied a wide array of clustering algorithms to try on our data. Because said array is lengthy, we have separated the following description of the results into two main sections. Firstly, in Section 3.1, we describe algorithms that do not take into account the temporality of the dataset, meaning models that take as input normal tabular datasets where the features represent different timestamps for when the biomarker was measured. Secondly and contrarily, Section 3.2 describes results with algorithms that do take into account the temporality of the dataset.

For each type of model, we describe differences in the evaluation metrics between the various clustering algorithms. Next, we report both the biomarker dataset and clustering algorithm that obtained the best progression accuracy score described in Section 2.5. Lastly, we describe the clusters obtained in the best results using the sociodemographic variables we mentioned in Section 2.1. All the percentages shown in this section are calculated over the total number of patients in each cluster.

3.1 Non-Temporal models

We applied several canonical clustering algorithms such as KMeans, Hierarchical Clustering and Self Organizing Maps, giving them as input two different datasets: (**i**) only one feature corresponding to the first biomarkers' measurement (i.e. baseline visit) available in the timeseries, and (**ii**) all features available in a patient's timeseries, meaning each feature

represents a patients' visit to the doctor. The first kind of models could serve us as a reference, showing what we could expect from the very first visit and measurement taken of each biomarker both in terms of clustering and the diagnosis progression. Moreover, we evaluated these two kinds of input datasets using information from (**i**) only one biomarker (e.g. AV45), (**ii**) a pair of biomarkers (e.g. AV45+ABETA) or (**iii**) several biomarkers (MRI). Therefore resulting in six different kinds of datasets evaluated with the abovementioned clustering algorithms. For the sake of understandability we have divided these results in categories involving non-tempral and temporal models. Figure 3.1 displays the proportion of diagnosis transitions between clusters in the non-temporal models that will be described. Meanwhile, Figure 3.2 shows the distribution of the sociodemographic variables for said models.

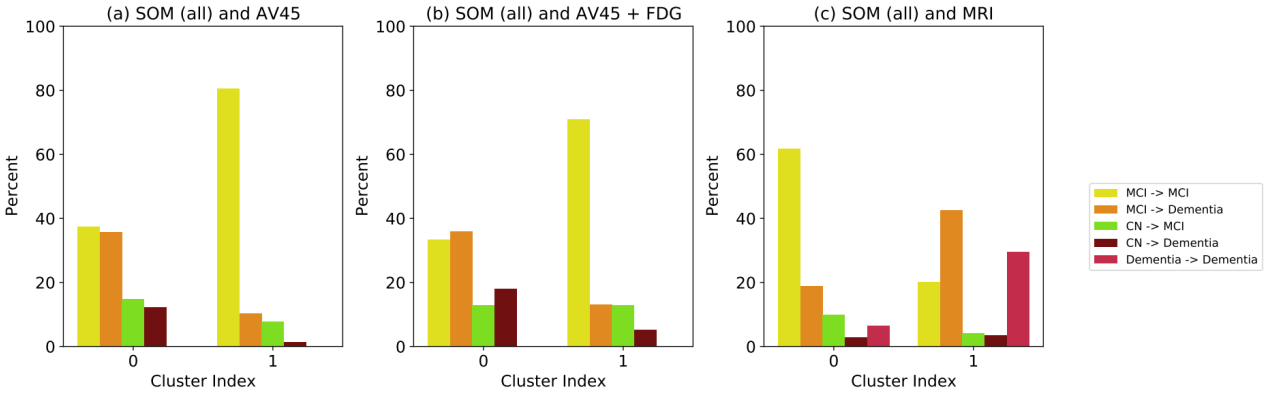


Figure 3.1: *Barplots showing the diagnosis transitions for each cluster in the three best clustering results for non-temporal models: (a) best non-temporal model for separated biomarkers (SOM with AV45, all features), (b) best non-temporal model for paired biomarkers (SOM with AV45+FDG, all features) , (c) best non-temporal model for the MRI dataset. (SOM, all features)*

3.1.1 Non-Temporal models with separated biomarkers

We applied KMeans, Hierarchical and SOM clustering to datasets containing only information of one biomarker using (**i**) only the baseline visit measurement, and (**ii**) all the visits' measurements. Figure 3.3 shows the results obtained in terms of SI score and accuracy.

As it can be seen in Figure 3.3a, most of the clustering models obtained a SI score greater

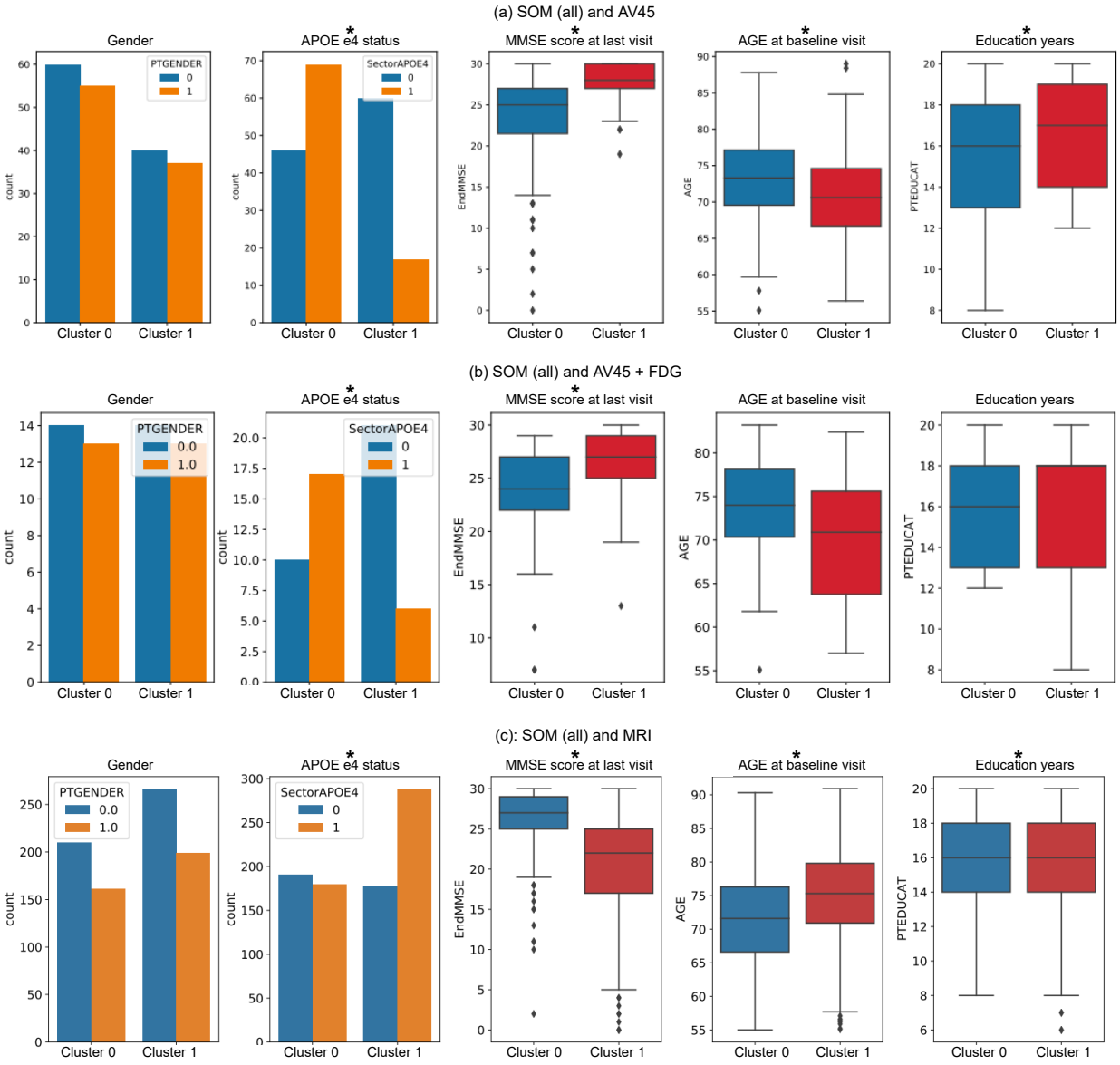


Figure 3.2: *Distribution of five different sociodemographical features by clusters: gender, APOE e4 genotype, MMSE score at the last visit, age at baseline visit, and education years in non-temporal models. Plots that are labeled with an asterisk on top indicate that the sociodemographic variable considered is significantly different (p-value < 0.05) between clusters.*

than 0.5, except for the majority of FDG datasets and algorithms combinations. In three out of five evaluated biomarkers (AV45, ABETA and FDG), K-Means was the algorithm that obtained the best results in terms of SI score, both using only the baseline visit (KMeans

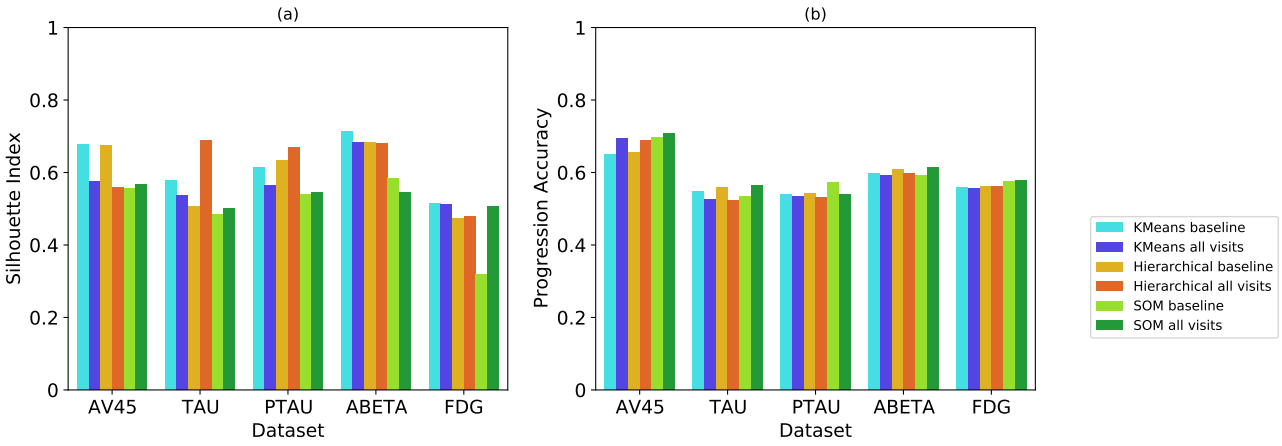


Figure 3.3: *Evaluation metrics obtained for each non-temporal model. Figure (a) shows the Silhouette Index Score and Figure (b) shows the Progression Accuracy achieved with unidimensional datasets.*

baseline, see Figure 3.3a) and all visits (KMeans all visits, see Figure 3.3a). Therefore, K-Means makes more defined clusters in terms of distance regarding non-temporal models. In fact, the highest SI score is obtained with K-Means using the baseline ABETA measurement (SI score of 0.7143). Regarding the progression accuracy defined in Section 2.5, as seen in Figure 3.3b, only datasets using AV45 and ABETA obtained an accuracy greater than 0.60. Overall, AV45 datasets obtained the best progression accuracy values compared to the other biomarkers. In fact, the combination that obtained the best accuracy results in AV45 datasets was the SOM model using all the available features (0.7083).

SOM model using all the available AV45 features gave two clusters, containing 115 and 77 patients, respectively. In the smaller cluster, 80.52% (62/77) of the subjects were diagnosed both at the baseline and last visit with MCI, while only 19.48 % (15/77) advanced towards the Alzheimer's Disease continuum between their initial and last visit. Unlike the smaller group, only 37.39% (43/115) in the bigger cluster stayed on the very same diagnosis, meanwhile a total of 62.61% (72/115) progressed towards a worse diagnosis. Thus, we can clearly see two very different groups, one of them containing mainly patients that stayed on the same stage of AD, and another group with patients that continued to a more advanced stage. These results also correlate with the demographic variables: we obtained a significant

difference regarding the number of patients or their mean values in each cluster for four out of the five sociodemographic variables used. Figure 3.2 shows the distribution of these variables in each cluster. The variables that obtained a significant p-value, meaning a significant difference between clusters were (p-values between parentheses): APOE4 ($9.30e-08$), the MMSE score ($7.35e-09$), the patients' age at baseline (AGE, $1.7e-02$) and finally the years of education of the patients (PTEDUCAT, $2.5e-02$).

3.1.2 Non-Temporal models with pairs of biomarkers

The same algorithms were applied in the very same datasets (baseline visits and all visits measurements) combined by pairs, in order to gain a better understanding whether a certain combination of biomarkers favors higher evaluation results.

In Figure 3.4a, the Silhouette Index scores obtained with combinations of biomarkers range from scores starting from 0.193 to 0.5743, with the majority above 0.4. The best score is obtained when combining the TAU and PTAU datasets, with a maximum value of 0.5743 when used with KMeans. As for the Progression Accuracy score, in Figure 3.4b, it seems that those pairs of biomarkers that were combined with AV45 obtained a remarkable result when used with Self Organizing Maps, compared to the other two algorithms, with a maximum of 0.741 in AV45 + FDG when using all available features. With most of the other pairs, it also seems like Self Organizing Maps obtains a better result than the other algorithms.

When applying SOM in the dataset combining AV45 and FDG, we obtained two clusters, each one containing 27 patients, as such we will name them C0 and C1. C0 only contains a total of 22.2 % (6/27) patients that do not transition towards a more advanced stage. Meanwhile the other 77.8 % (21/27) did suffer from more neurodegeneration. Contrarily, C1 has a total of 70.4 % (19/27) of patients that stay in the same diagnosis, and the remaining 29.6 % (8/27) that do progress towards a more advanced stage. As for the demographic variables, both the APOE4 gene ($1.9e-03$) and the MMSE score ($1.4e-02$) are significant.

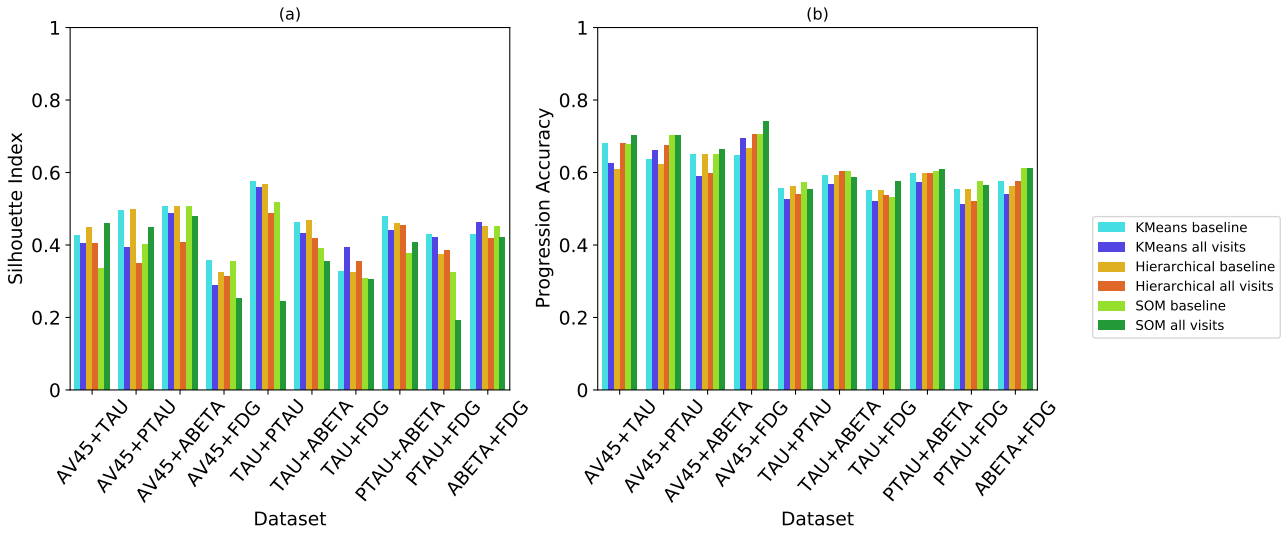


Figure 3.4: *Evaluation metrics obtained for each non-temporal model. Figure (a) shows the Silhouette Index score and Figure (b) shows the Progression Accuracy achieved with multidimensional datasets.*

3.1.3 Non-Temporal models with MRI regions

Finally, we applied the aforementioned clustering algorithms to a dataset containing all of the cerebral regions described in Section 2.1, with the same variations that the other datasets had, these being (**i**) only baseline measurements, and (**ii**) all visits' measurements. Figure 3.5 shows the results obtained regarding SI scores and Progression Accuracies.

As seen in Figure 3.5a, all clustering models obtained a Silhouette Index score similar to 0.3, with a maximum of 0.322 when using K-Means and all available features in the time series. Regarding the algorithm variants, it seems that the algorithms that use all of the available features in the dataset obtain a better score than those that do not. Regarding the accuracies obtained, we can gather from Figure 3.5b that all models obtain an accuracy below 0.6, with Self Organizing Maps and the variant that uses all features obtaining a maximum accuracy of 0.593. As with the Silhouette Index scores, the algorithms that use all of the available features obtain a better accuracy than those that only use the baseline visit measurement.

When combining the variant of SOM that uses all the available features with the MRI

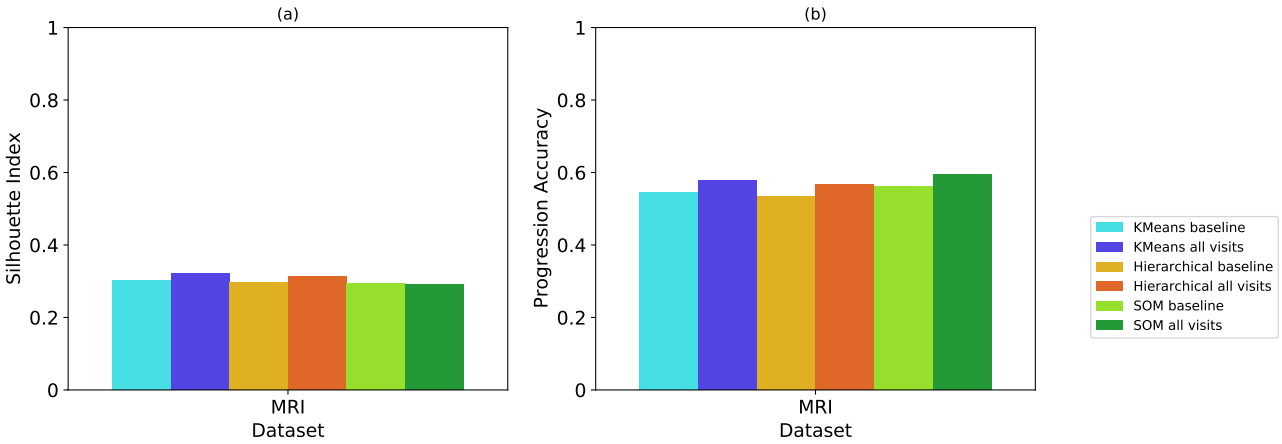


Figure 3.5: *Evaluation metrics obtained for each non-temporal model. Figure (a) shows the Silhouette Index Score and Figure (b) shows the Progression Accuracy achieved with the MRI dataset.*

regions, we obtain two clusters, a smaller one containing 374 patients and a bigger one with 462 patients. In the smaller cluster, a total of 61.8 % (255/374) had a static diagnosis transition, while the remaining 38.2 % (119/374) moved to a more advanced stage. Meanwhile, in the bigger cluster, 49.8 % (230/462) of patients stayed in the same diagnosis while the other 50.2 % (232/462) did not. Regarding the sociodemographic variables, four variables were considered significant, these being the APOE4 ($2.61e-06$), the MMSE score ($2.50e-49$), the patients' age (AGE, $4.46e-15$) and the patients' years of education (PTEDUCAT, $4.9e-02$)

3.2 Temporal models

After applying canonical clustering algorithms to the datasets, we applied other clustering techniques that consider temporality (a) within the datasets, or (b) within the algorithm. Its main purpose is to compare the difference w.r.t the non-temporal models evaluated previously (Section 3.1) to assess whether including any kind of temporal structure to the data could give better results.

Firstly, as described in Section 2.4, for each of the five biomarkers, we applied a poly-

nomial regression to the original tabular datasets including all visits. We extracted the polynomial coefficients of each timeseries and used them as features with KMeans, Hierarchical and SOM clustering algorithms. Secondly, Dynamic Time Warping was merged with other types of clustering algorithms to obtain results using mainly the temporality of the data. These models include DTW combined with Hierarchical clustering, as well as DTW and DBSCAN.

Figure 3.6 displays the proportion of diagnosis transitions between clusters in the following temporal models that will be described. Meanwhile, Figure 3.7 shows the distribution of the sociodemographic variables for said models.

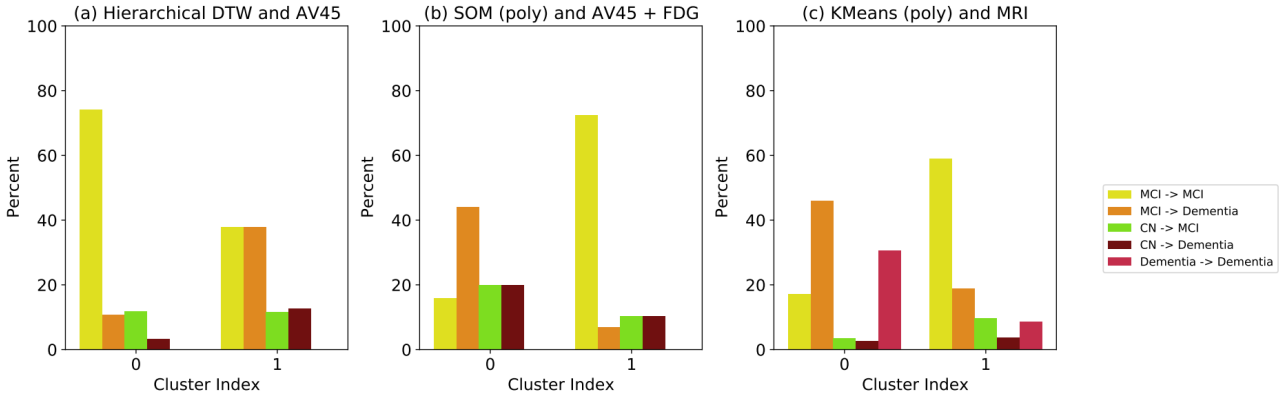


Figure 3.6: *Barplots showing the diagnosis transitions for each cluster in the three best clustering results for temporal models: (a) best temporal model for separated biomarkers (Hierarchical DTW with AV45), (b) best temporal model for paired biomarkers (polynomial variant of SOM with AV45+FDG) , (c) best temporal model for the MRI dataset. (polynomial variant of KMeans).*

3.2.1 Temporal models with separated biomarkers

We firstly evaluated these temporal models using only one biomarker measurements per dataset, as done in Section 3.1 with non-temporal models. Figure 3.8 shows the SI scores and progression accuracy values obtained from all of the temporal models (using polynomial datasets and DTW timeseries clustering).

Regarding results using the **polynomial datasets**, the SI scores vary greatly (Figure

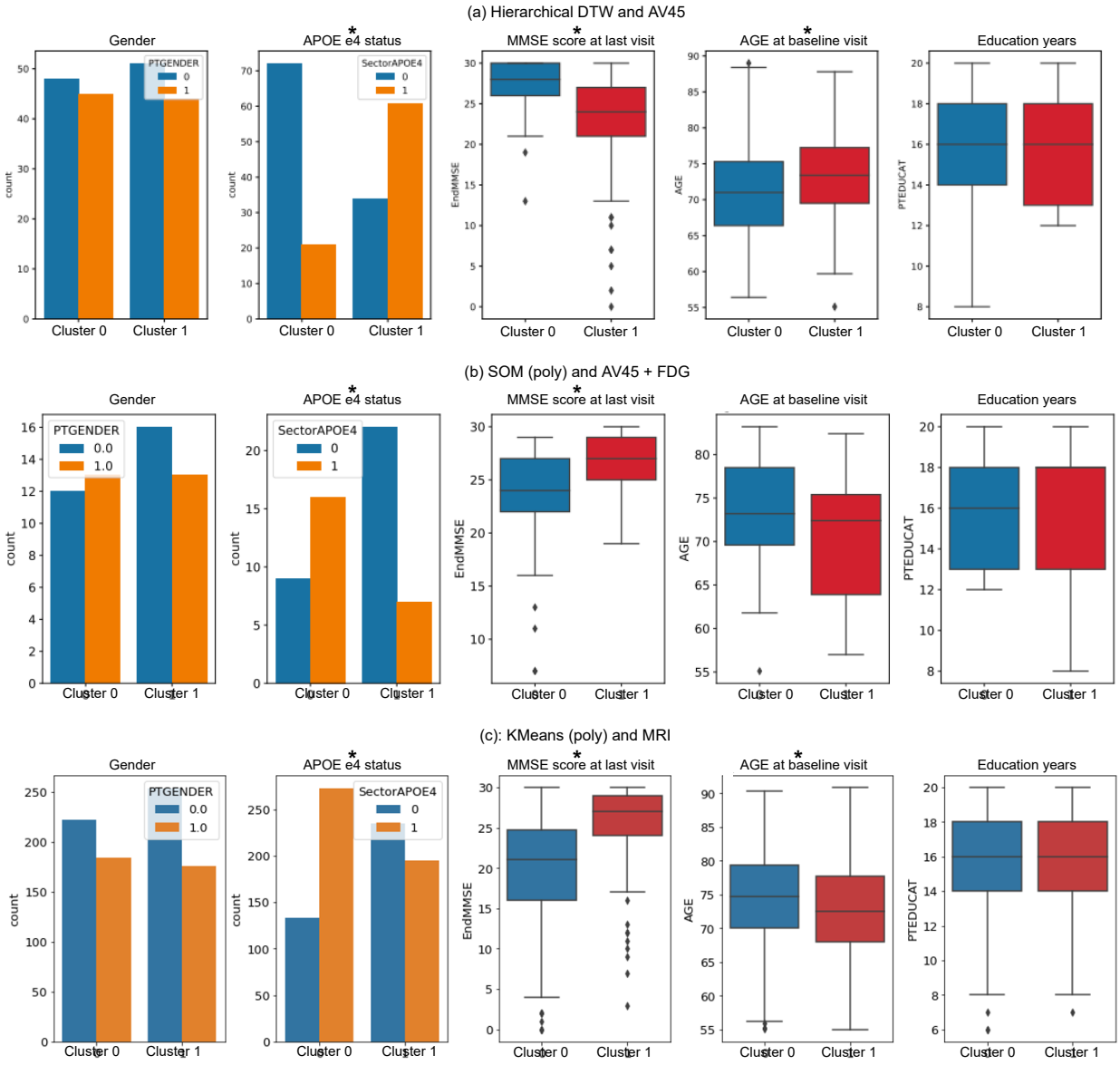


Figure 3.7: *Distribution of five different sociodemographical features by clusters: gender, APOE e4 genotype, MMSE score at the last visit, age at baseline visit, and education years in temporal models. Plots that are labeled with an asterisk on top indicate that the sociodemographic variable considered is significantly different (p-value < 0.05) between clusters.*

3.8a), with K-Means and Hierarchical clustering usually obtaining scores over 0.4, contrarily to SOM, always scoring less. The maximum score obtained in this case is for the PTAU biomarker (0.626) when combined with KMeans. Progression accuracy values (Figure 3.8b) show that most datasets are close to 0.50, excluding the AV45 dataset which, regardless

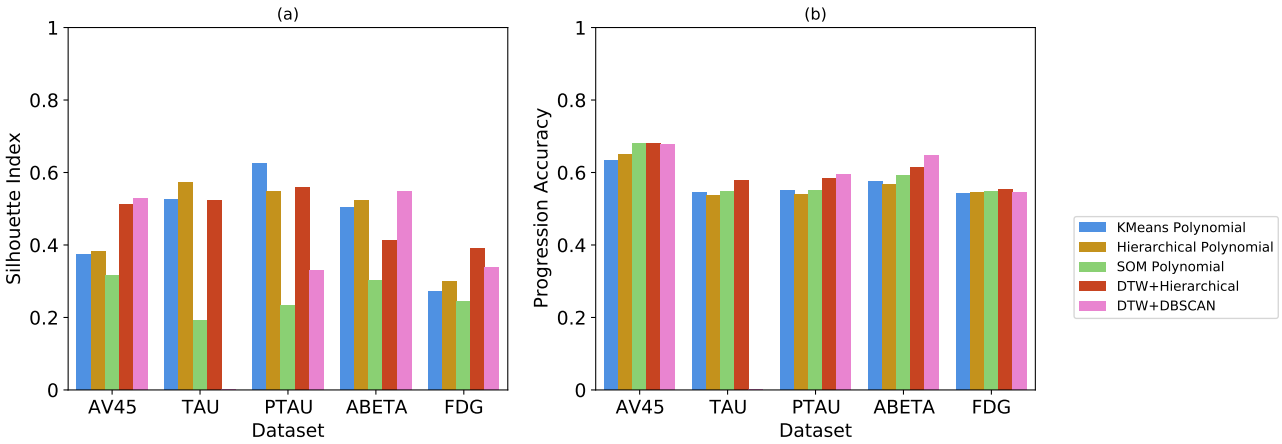


Figure 3.8: *Evaluation metrics obtained for each temporal model. Figure (a) shows the Silhouette Index score and Figure (b) shows the Progression Accuracy achieved with unidimensional datasets.*

of the algorithm, scores over 0.60, topping at 0.6809 with the model combining the AV45 dataset and SOM.

Regarding results using **DTW**, the SI scores obtained are mostly close to 0.50, except for the FDG dataset where both models obtain a score around 0.4. The PTAU dataset with DTW+Hierarchical clustering was the combination that obtained the maximum SI score (0.560). As for the accuracy values, most models performed over 0.60. As it was the case with the SI score, the FDG dataset obtained the lowest accuracies compared to other biomarkers (less than 0.60). The maximum progression accuracy obtained was 0.6813 using the AV45 dataset by means of DTW + Hierarchical clustering.

Comparing **DTW and polynomial transformation**, the best accuracy was obtained with DTW + Hierarchical clustering using AV45 dataset. This model yielded 2 clusters, one consisting of 93 patients and the other of 95 patients. Since we are now working with temporal models, each patient could be represented by a regression line. Therefore we computed the regression line coefficient of each of the obtained clusters to measure the average speed and its direction to describe them. The smaller cluster had a coefficient of $-4.40e-04$, while the bigger cluster had a coefficient of $8.72e-04$. As we can see, the average of patients in the bigger cluster have had their AV45 value increased, contrarily to the

smaller one. This correlates to the following diagnosis transitions. The smaller cluster had 74.2 % (69/93) of patients that did not transition to an advanced stage, while the other 25.8 % (24/93) continued towards Dementia at a faster rate. As for the bigger cluster, only 37.9 % (36/95) had a static transition, with the remaining 62.1 % (59/95) progressing towards neurodegeneration. The sociodemographic variables that showed a significant difference between clusters were: APOE4 ($2.02e-09$), MMSE score ($1.49e-09$) and the patients' age at baseline (AGE, $1.7e-02$).

3.2.2 Temporal models with pairs of biomarkers

Secondly, we evaluated the temporal models above-mentioned using datasets formed by pairs of biomarkers. Figure 3.9 shows the results obtained in terms of SI scores and progression accuracy values.

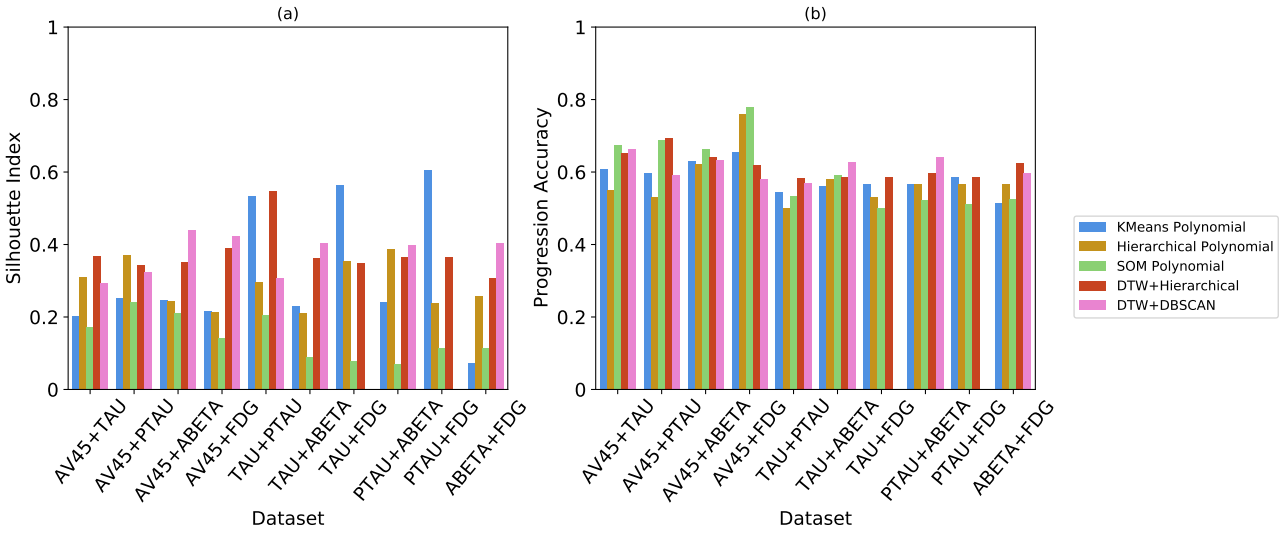


Figure 3.9: *Evaluation metrics obtained for each temporal model. Figure (a) shows the Silhouette Index score and Figure (b) shows the Progression Accuracy achieved with multi-dimensional datasets.*

Regarding **polynomial models**, the majority of the models scored a SI metric below 0.40, with very few exceptions, e.g. K-Means algorithm with the pairs TAU+PTAU, TAU+FDG, PTAU+FDG. The worst SI scores obtained were from SOM (lower than 0.2)

suggesting a highly incorrect clustering. The maximum SI score obtained was 0.606 when using the pair PTAU+FDG with the K-Means algorithm. As for the progression accuracies obtained, most of the datasets containing the AV45 biomarker reached above 0.60, especially the pair AV45+FDG, which scored the maximum value (0.778). Moreover, for datasets containing the AV45 biomarker, SOM models obtained a better score than the other algorithms, contrarily to the rest of the biomarkers, where SOM usually obtained a worse score.

Regarding **DTW models**, there seemed to be a considerable variability regarding the SI scores, as they range from 0.2947 to 0.5468, with TAU+PTAU the pair obtaining the maximum value. Most of the accuracies obtained were over 0.60, with the maximum being 0.6943 in the pair combining the AV45 and PTAU biomarkers.

Lastly, when comparing **DTW and polynomial transformation**, the model that obtained the best progression accuracy was the pair combining AV45 and FDG by means of SOM and their polynomial coefficients. This clustering yielded 2 clusters, a smaller one with 25 patients and a bigger one with 29. In the smaller cluster, only 16% (4/25) of patients did not transition to a more advanced stage, while a total of 84% (21/25) of patients suffered from more neurodegeneration. However, for the bigger cluster, a total of 72.4% (21/29) of patients stayed in the same diagnosis, with the remaining 27.6% (8/29) transitioned to a more advanced stage. A total of two demographic variables were considered as significantly different between these two clusters: APOE4 ($2.51e-03$) and the MMSE score ($9.27e-04$).

3.2.3 Temporal models with MRI regions

Lastly, we applied these algorithms to the MRI dataset. Figure 3.10 shows the results obtained in terms of scores and accuracies.

As for the **polynomial models**, they all obtained scores lower than 0.2, with the majority not reaching 0.1. The maximum score obtained is 0.134, applying Hierarchical clustering in its polynomial variant. Regarding the accuracies obtained, only K-Means in its polynomial variant obtained a score above 0.6, concretely 0.6017.

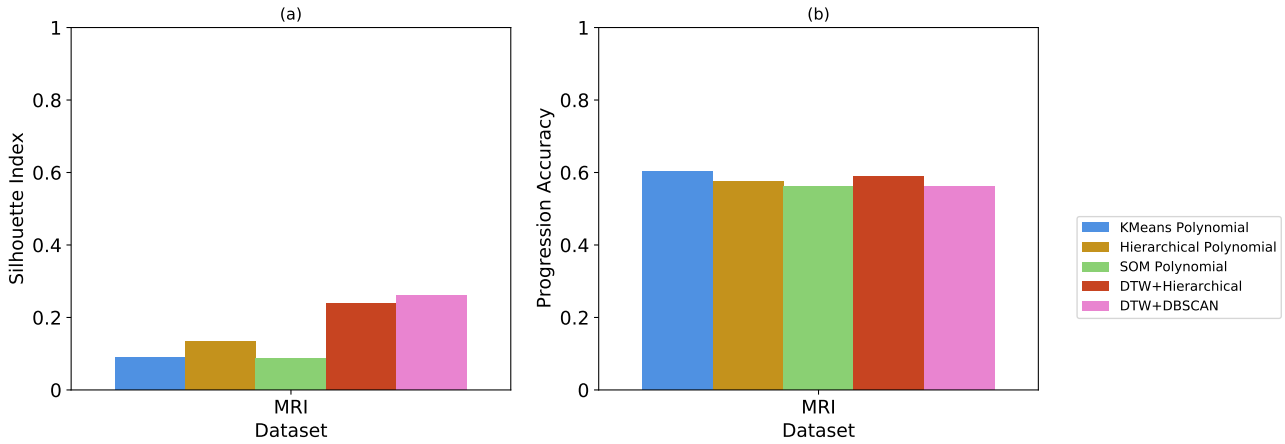


Figure 3.10: *Evaluation metrics obtained for temporal models with the MRI dataset. Figure (a) shows the Silhouette Index Score and Figure (b) shows the Progression Accuracy.*

Regarding **DTW models**, they obtained scores above 0.2, with the maximum being 0.262 when applying DBSCAN with DTW distances. As for its accuracies, they all fall below 0.6, with the maximum value being 0.588 when applying DTW and Hierarchical clustering.

Finally, when comparing **DTW and polynomial transformation**, the algorithm that obtained the best progression accuracy with the MRI dataset was K-Means in its polynomial variant. This model obtained two clusters, a smaller one consisting of 406 patients, while the other contains 430 patients. The smaller cluster has a total of 47.7 % (194/406) of patients with a static diagnosis transition, while the remaining 52.3 % (212/406) transitioned to a more advanced stage. Meanwhile, 67.7 % (291/430) of patients had a static diagnosis transition while the other 32.3 % (139/430) did not. Finally, regarding the sociodemographic variables, three were considered significant, these being APOE4 ($1.17e-10$), the MMSE score ($4.93e-60$) and finally the patients' age (AGE, $1.97e-04$).

Chapter 4

Discussion

We have performed an extensive research of combinations between multiple clustering algorithms and biomarkers datasets for the discovery of new clusters related to the progression of AD. Regarding the clustering algorithms, we evaluated several algorithms ranging from those that do not take into account the temporality of the data (Section 3.1) to other algorithms that do use time series (Section 3.2) for grouping patients. Moreover, the array of datasets used started with those that do not have any temporal components to others that do integrate them and also combinated different biomarkers alone and in pairs, as well as the MRI regions. Therefore, we will analyze in this section whether there is a correlation between integrating temporality to the model and obtaining a better result in each of the biomarkers, using the evaluation metrics proposed. In the following paragraphs we employ the term "model" as the application of a certain clustering algorithm with a specific dataset, e.g. K-Means with AV45 dataset.

In general, SI scores seemed to vary between models, as some scores reach up to 0.7 and others barely make it to 0.2. Because of the uneven SI scores between models, the following discussion is organized to give an interpretation from different perspectives, differentiating the results obtained between datasets and the algorithms used.

Regarding non-temporal models that used individual biomarkers datasets, from a dataset perspective, there is a difference of results between linear algorithms such as K-Means and Hierarchical clustering, and non-linear algorithms like SOM. AV45 and ABETA datasets

worked better with linear algorithms, obtaining SI scores higher than 0.6, contrarily to non-linear algorithms. However, FDG dataset worked better with SOM, obtaining a SI score as good as the ones previously described with linear algorithms. From an algorithmic point of view, both K-Means and Hierarchical clustering algorithms obtained better results than SOM. These results make sense because the former algorithms use an approach completely based on distances between samples, the principle in which SI score is based. Specially, the K-Means algorithm as it is highly distance-based, since it is an algorithm that minimizes cluster variance and therefore favors a good SI score. Meanwhile, SOM integrates the underlying structure of the data and favors a higher distance between neurons, but these neurons do not necessarily imitate the variance of the data, specially because we have performed a feature extraction and some information about the original data will be lost. Moreover, when using AV45 and ABETA datasets, we obtained a better SI score when giving as input only the baseline visit measure. On the contrary, we obtained better SI results when using all features in the TAU or PTAU datasets.

As for pairs of biomarkers, from a dataset perspective, the SI scores obtained are generally worse than those acquired when using separated biomarkers. This is probably due to the increased dimensionality when adding another biomarker into the dataset: with the datasets that only focus on the first feature, the dimensionality of each sample becomes 2 since we add the first feature of the other biomarker, meanwhile for those that use all 5 components, the dimensionality increases to 10. This dimensionality increases the distance between samples, and the clusters obtained are less dense, therefore obtained a lower SI score. The dataset that obtained the best score was the TAU+PTAU pair as most of the algorithms that use this dataset obtained a score close to 0.6. A dataset that also performed well is the pair combining the AV45+ABETA biomarkers, which seemingly obtains the second best score out of the other datasets. The reason for obtaining denser clusters with the pairs TAU+PTAU, as well as AV45+ABETA make sense accordingly to the processes that these biomarkers represent. Both TAU and PTAU measure the amount of Tau

protein in their normal and phosporilated form, respectively. Moreover, both AV45 and ABETA measure the amount of accumulated beta-amyloid in the brain or in the CSF, respectively. Therefore, there will be a higher correlation between biomarker measurements in these pairs, than between those that are less similar. For example, the pair AV45 and FDG performs the worst between all the datasets, as all of the algorithms that use it obtained a score lower than 0.40. Although both AV45 and FDG are measures obtained through PET scans and measure neurodegeneration, AV45 measures the quantity of beta-amyloid while FDG measures the amount of glucose consumed in the brain, thus they correspond to very different neurodegeneration biological processes. Therefore, it seems that the SI score is highly affected by the nature of the paired biomarkers. Algorithmically-wise, as it happened and was explained before with the separated biomarkers' datasets, linear algorithms usually obtained a higher score than non-linear algorithms in each of the datasets. Moreover, for each algorithm, the variant that only uses the first feature obtains a slightly higher SI score than the variant that uses all of the available features. This is explained for the reason that variants that only use the first feature have lower dimensionality in their datasets than the other variant, therefore clusters have lower intra-cluster distance.

As for MRI, from an algorithmic perspective, all models obtained a very poor SI score when compared to the prior models. However, since we have introduced 6 new dimensions for our algorithm to work with, and where the *Ventricles* region has a different direction than the other regions in terms of increased neurodegeneration, it is logical that the distance between samples will increase considerably, difficulting the possibility of obtaining high scores. Nevertheless, models that use all of the available features obtain higher scores than the models that do not, as evidenced by Figure 3.5. Therefore, a plausible hypothesis could be that in multidimensional datasets, variants that use all features obtain higher scores as opposed to unidimensional datasets, where variants using only the baseline visit feature obtain better scores. Moreover, in both variants, the K-Means algorithm obtains higher scores than SOM, which supports the idea that linear models obtain better scores than

non-linear ones.

In conclusion, for non-temporal models, we have seen that in the majority of cases, an algorithm that is based on linear regression obtains a higher SI score than non-linear algorithms. Moreover, unidimensional datasets seem to obtain better scores when only using the baseline visit whereas multidimensional datasets obtain higher scores when using all the available features, as evidenced by the MRI regions which obtained greater scores with the variants that make use of all features, unlike the separated biomarkers. Finally, we have seen how pairs of similar biomarkers generally obtained higher scores. We wanted to confirm further these results with temporal models.

Among temporal models, we differentiate between those (a) including the temporality within the dataset through a polynomial regression transformation; and (b) those including the temporality within the algorithm therefore the dataset is built as a time series and the algorithm (DTW) works with that kind of data. Regarding polynomial regression models, when using individual biomarkers, it is clear that these models obtained a worse score (the majority falling under a SI score of 0.40) than the non-temporal models (SI score higher than 0.60). The main reason for this could be related with the polynomial transformation performed. This transformation works as a feature extraction technique: we change the features in which the data is represented. However polynomial regression could not accurately represent the original time series progression, therefore information is lost. This can be observed when applying the polynomial variant of SOM, where we firstly perform a feature extraction from 5 to 3 features, and a second feature extraction from 3 to 2 features. Regarding that last statement, linearity in models obtains a higher score than non-linear models, as evidenced by K-Means and Hierarchical clustering obtaining a higher SI score than SOM. Moreover, a key difference between these models and the non-temporal models is that AV45 no longer obtained the highest score, instead, TAU and PTAU, which previously obtained worse scores in non-temporal models, now give better results. Furthermore, regarding DTW models from the dataset perspective, we obtained uniform results for each dataset as they

all obtained a score over 0.50, except for FDG dataset. In order to calculate the SI score for these models, we have considered DTW distance instead of Euclidean distance since clusters will be made based on the former instead of the latter. From an algorithmic point of view, it seems that the combination of DTW and DBSCAN struggles with biomarkers such as TAU, where it does not obtain a result based on the filtering criteria, or PTAU, where it obtained a very low SI score. However, DTW+DBSCAN models obtained better results when using AV45 or ABETA datasets.

Temporal models were also tested using pairs of biomarkers, as done with non-temporal models. Regarding the polynomial regression model, from a dataset point of view, all pairs that contained either AV45 or ABETA features performed worse compared to other datasets that contained either TAU or FDG. Therefore, when temporality is taken into account, it seems the AV45 dataset is no longer the dataset that obtains the best score, as it previously happened with non-temporal models. Moreover, from an algorithmic point of view, once again linear algorithms (KMeans and Hierarchical clustering) obtained considerably better results than the non-linear ones (SOM). Finally, regarding DTW models, we obtained uniform results between datasets, generally SI scores around 0.4 with a few exceptions. In the case of DTW, it makes sense for the SI score to be lower compared to the DTW models with biomarkers alone, as we have added dimensionality. Moreover we obtained the same results than one-biomarker models: pairs with AV45 obtained worse scores than those with TAU or PTAU, contrarily to the non-temporal models. Finally, combining DTW+DBSCAN generally returns better results than DTW+Hierarchical clustering. This is due to the fact that DBSCAN by definition is a clustering algorithm that creates clusters out of initial samples and builds around them, therefore favoring a better SI score.

Regarding the MRI models from an algorithmic point of view, there is a clear difference between the scores obtained when using polynomial regression and those obtained using DTW, as the former score close to 0.1 whereas the latter score above 0.2. Similarly to the non-temporal models, the scores are lower when compared to separated biomarkers or

when combined by pairs as the dimensionality has increased greatly, distancing each sample further. Finally, as for the algorithms using the polynomial variant, once again linear models obtain better scores. Regarding algorithms that use DTW, DBSCAN obtains a better SI score because of its nature as explained above.

In conclusion, for temporal models, the SI scores obtained are generally worse than the non-temporal models since we use feature selection like in DTW. Moreover, when considering temporality, more complex biomarkers such as TAU or PTAU obtained better scores than biomarkers like AV45 or ABETA. Finally, datasets with higher dimensionality, such as MRI datasets, will generally obtain lower scores.

Although the results obtained with SI scores yielded interesting discussions about which biomarkers work better and what clustering algorithms are preferred from the point of view of SI score, we also wanted to see if the obtained clusters were making sense with the known diagnosis transitions and the sociodemographic information of the subjects. One of the greatest potentialities of clustering is the discovering of new knowledge, and therefore it is important to take into account clinical information, to properly label these new patients' groups obtained.

Therefore, we developed a metric based on accuracy score to measure the reliability of these new groups from a clinical point of view. We named this metric "progression accuracy" (see Section 2.5). This metric measures the overrepresentation of patients that progress worse in diagnosis (towards MCI or Dementia) against patients that did not progress, staying in their diagnosis. Therefore in the following lines we described what happened regarding the previously described models.

When using separated biomarker datasets, from a dataset perspective, AV45 and ABETA obtained considerably better accuracies than the rest of the biomarkers regardless of the algorithm used, as all of their accuracy scores are over 0.60, contrarily to the rest of the datasets. Moreover, TAU and PTAU obtain similar accuracies as well as the worst ones between all the datasets. AV45 is a biomarker that is very considered when giving a patient

diagnosis[16], as it measures amyloid burden in the brain, therefore it differentiates best between patients in the Alzheimer's disease continuum. Algorithmically-wise, in general Self Organizing Maps seem to obtain higher accuracies than the other algorithms, specially in AV45. Moreover, in AV45 regarding the features used, considering all five features seems to obtain a higher accuracy, specially in linear algorithms as a correlation[17] higher than 0.8 between features can be observed. However, Self Organizing Maps with only one feature obtains as high an accuracy as linear algorithms with all features, which suggests that non-linear models are capable of inferring further than linear algorithms. Finally, as for the other datasets, using all features in a time series does not necessarily improve the accuracy.

As for pairs of biomarkers, from a dataset standpoint, pairs that consider the AV45 biomarker always obtain a higher accuracy than other pairs, specially the combination of the AV45 and FDG datasets which obtains the highest result when used with Self Organizing Maps. Moreover, the FDG dataset, which obtains a very good accuracy with AV45, does not obtain as good a result when combined with other datasets, which indicates that the biomarkers in the pair have a considerable influence on the accuracy obtained. From an algorithmic perspective, generally the algorithm that works best is Self Organizing Maps. Moreover, linear algorithms obtain better accuracies either when using the first component or all of them, depending on the pair of biomarkers used.

Regarding the MRI dataset, the accuracies obtained are low, as all of the models' accuracies are lower than 0.6. However, as aforementioned, non-linear models such as SOM performed better than linear models. It is important to note that, while these models obtained poor results regarding the accuracies, they have an extremely high potential to describe the clusters formed, as SOM's obtained clusters with MRI biomakers resulted in four significantly different sociodemographic variables between them (with very low p-values, in the order of $e - 49$ for the MMSE score). This result proves that the amount of significant variables in a model does not depend of the progression accuracy obtained, but rather of the biomarker used: in this case the MRI regions. Furthermore, the AV45 biomarker is an

average of all the cerebral regions, and while the former usually obtains very high accuracies, the latter seems to describe clusters better. Therefore using biomarkers that make up the average of other biomarkers will eventually omit information about its patients and poorly describe the clusters.

Therefore with non-temporal models we can surmise that, first and foremost, non-linear models work better than linear algorithms, contrary to the computational perspective described above. Moreover, the AV45 biomarker seems to favor a better accuracy than other datasets. Finally, the combination of biomarkers when used in multidimensional datasets influences the Progression Accuracy as well as the significant variables obtained for that model.

From a clinical perspective in temporal models, when using separated biomarkers, and taking into account polynomial regression, from a dataset perspective we can find better accuracies when using the AV45 dataset, as found in non-temporal models. However, the rest of the datasets obtain a very similar accuracy, especifically TAU and PTAU, which now obtain a very high accuracy. Regarding the algorithms employed, SOM obtained the best accuracies, specially when applied to AV45 dataset. Meanwhile, K-Means and Hierarchical clustering obtain similar results, due to how similarly they both function. Moreover, when considering the DTW models, from a dataset perspective, models that include AV45 no longer obtain as good an accuracy as with non-temporal models. Moreover, TAU and PTAU models improved their accuracies, reaching an accuracy higher than 0.60. Therefore, it seemed that temporality, specially algorithms that exploit temporality like DTW, favor better accuracy while also worsening it in datasets like AV45. From an algorithmic standpoint, both algorithms implementing DTW obtained similar results, although one overcomes the other depending on the dataset: Hierarchical clustering with DTW obtains better accuracies in AV45 and FDG datasets, meanwhile the DBSCAN variant obtains better results in PTAU and ABETA datasets, specially struggling in finding a suitable result in the TAU dataset.

As for pairs of biomarkers when using polynomial regression, regarding the datasets, the best accuracies are obtained from the pairs containing the AV45 biomarker, specially when it is paired with FDG. Models using other datasets have increased too, for example the pair TAU+PTAU. This pair now yields an accuracy closer to the other models, reducing the distance between them. An interpretation similar to the one we gave with non-temporal models can be made: certain combinations of biomarkers favor a higher accuracy. Moreover, temporality seems to increase the accuracy in datasets like TAU or PTAU. As for the algorithms, SOM models obtained the highest accuracy in AV45 datasets, specially in the pair AV45+FDG. While FDG on its own generally obtains low accuracies, when paired with other biomarkers, especifically AV45, the resulting accuracy is much higher. As for K-Means and Hierarchical clustering, they both obtained similar accuracies, but depending on the dataset one overcomes the other. Finally, DTW models combining biomarkers in pairs, from a dataset standpoint, worked best when using the pair AV45+PTAU. The majority of the scores obtained with DTW models in the case of pairs of biomarkers were close to 0.60. As for the algorithms used, both DTW+Hierarchical clustering DTW+DBSCAN obtained similar scores and overcame each other depending on the biomarkers.

Regarding the MRI dataset, all models obtain an accuracy close to 0.6. However, when compared to the non-temporal models, it seems that these models perform better as a few go beyond 0.6. This might be explained by the fact that we are no longer using only AV45, which represents the average of all regions, but rather six complex biomarkers that do not necessarily follow the same direction and therefore, adding temporality to the model improves the accuracy. As for the algorithms used, the polynomial variant of K-Means seems to do better than its other polynomial counterparts, meanwhile Hierarchical clustering and DTW do better than DBSCAN combined with DTW. K-Means obtaining a better accuracy than Self Organizing Maps might be due to the possibility that due to the high temporality, the latter is not able to morph its grid of neurons to represent accurately the data, therefore obtaining a poor result. Finally, this model obtained three significant variables, and not

only is it a bigger amount than other models when used with their polynomial variants, but also the descriptive power of the MRI dataset returns very low p-values (MMSE, $e-60$).

Therefore, temporal models seemed to work exceptionally well combining non-linear models with the AV45 biomarker. Moreover, increasing the temporality in the model increases the accuracies obtained in TAU and PTAU datasets, although it does not favor the AV45 biomarker. Lastly, as for the MRI dataset, while the accuracies obtained are low, it is more descriptive than other biomarkers. Finally, we have seen how SI scores and progression accuracies (PA) are not interrelated, since the SI score is ultimately based on outer- and intra-cluster distance while the PA considers diagnosis transitions. The Alzheimer's Disease continuum does not follow a linear progression for a patient to be placed in one of the three stages therefore the clusters obtained in a perfect PA setting will inevitably overlap eachother.

Lastly, we evaluated the differences between clusters regarding the sociodemographic variables of the patients on the models that obtained the best progression accuracies. We observed that a significant difference in one of these variables between clusters is not necessarily linked to the Progression Accuracy obtained. For instance, the temporal model that obtained an accuracy of 0.778 when using SOM and the pair of biomarkers AV45 and FDG only had two significant variables: the APOE4 genotype and MMSE score. Meanwhile, in the non-temporal model that obtained an accuracy of 0.7083 using SOM and the AV45 dataset, there were a total of four significant variables: APOE4 genotype, MMSE score, the patients' age at baseline visit, and the education years of the patient. Additionally, the pair AV45+PTAU obtained a total of three out of five significant variables. Therefore the significant difference between clusters regarding a sociodemographic variable could be linked to the biomarkers in question: AV45 -it being an average of the volume of six brain regions- is more descriptive of Dementia, and when combined with other biomarkers, said descriptive potential is lost.

Moreover, from a clinical perspective and for the sake of interpretability, we can gather

from the described results that certain demographic variables have a considerable correlation with Alzheimer's Disease while others do not. In all the described models, both the MMSE score as well as the APOE e4 genotype were significantly different between the obtained clusters. MMSE score results are logical since it is a test specifically taken by patients to measure cognitive decline. Therefore, if we have a model with a high progression accuracy, one cluster is going to be overrepresented with cognitively impaired subjects (patients progressing towards worse diagnoses) while the other is going to be overrepresenting less cognitively impaired subjects (patients that stayed in the same diagnosis). APOE e4 results are also interesting. This variable represents a specific mutation in the APOE gene confirmed to be very linked to Alzheimer's Disease[18], specially to beta-amyloid deposition. Although it is confirmed to be related to beta-amyloid deposition (measured by AV45 and ABETA biomarkers) it is also interesting to see that it is linked to other biomarkers progression such as TAU, PTAU or FDG. Furthermore, the patients' age at baseline has also been considered significant on several occasions. This also makes sense because the older the patient, the bigger the chances of suffering from neurodegeneration. In fact, early-onset Alzheimer's Disease (developed before 65 years old) is a rare-condition, being the late-onset form of this disease much more common.[19] The education years of a subject was only considered significant once, when the model combining the AV45 biomarker with all its features and the SOM algorithm was used, and it is logical since it is proven that higher learning results in less cognitive disorders associated with AD[20]. Regarding the patient's gender, although being female has been determined as a relative risk to develop Dementia due to Alzheimer's Disease[21], in our study it appears that it is not as determinant as the other variables used, since the patient's gender was never considered significant.

Chapter 5

Conclusion

In this work we have explored different types of clustering and how they can be considered a reliable technique to analyze time series in a cohort of patients with Dementia. We have extensively explored the different biomarkers collected for the ADNI study, organizing them in multiple ways (separated and combined by pairs) to assess their performance w.r.t. two different metrics (SI score and progression accuracy), giving rise to clinically meaningful clusters. In addition, we have evaluated all obtained clusters in terms of known demographic characteristics related to Alzheimer's Disease, as well as diagnostic changes occurring during the time periods described for each biomarker.

The use of clustering techniques, mainly non-linear and temporal ones, can be considered a potential method to further analyze time series describing neurodegenerative diseases such as Alzheimer's, since the most crucial part to study is the evolution of the patient and not only the fact that the patient receives the diagnosis.

The following main lines of work may be considered. Firstly, to carry out a more extensive exploration of the algorithms and the datasets, since some biomarkers were lacking considerable information and a heavy timeseries filtering process had to be applied. Secondly, explore other time series clustering tools that consider non-linearity but do not train their data through competitive learning, such as other neural networks other than SOM. And finally, obtain a larger amount of data by employing non-linear timeseries.

Bibliography

[1] Alzheimer's Association. 2019 alzheimer's disease facts and figures. *Alzheimer's & dementia*, 15(3):321–387, 2019.

[2] John Wiley. Alzheimer's disease facts and figures. *Alzheimer's Dementia*, 17:327–406, 2021.

[3] Patricia Sobradillo, Francisco Pozo, and Álvar Agustí. P4 medicine: the future around the corner. *Archivos de Bronconeumología ((English Edition))*, 47(1):35–40, 2011.

[4] Rémi Cuingnet, Emilie Gerardin, Jérôme Tessieras, Guillaume Auzias, Stéphane Lehéricy, Marie-Odile Habert, Marie Chupin, Habib Benali, Olivier Colliot, Alzheimer's Disease Neuroimaging Initiative, et al. Automatic classification of patients with alzheimer's disease from structural mri: a comparison of ten methods using the adni database. *neuroimage*, 56(2):766–781, 2011.

[5] Elizabeth Bradley and Holger Kantz. Nonlinear time-series analysis revisited. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 25(9):097610, 2015.

[6] 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*, 89:1–6, 2016.

[7] Alexia Giannoula, Alba Gutierrez-Sacristán, Álex Bravo, Ferran Sanz, and Laura I Furlong. Identifying temporal patterns in patient disease trajectories using dynamic time warping: A population-based study. *Scientific Reports*, 8, 2018.

[8] Adni documentation time series. https://adni.loni.usc.edu/data-samples/data-types/pet/. Accessed: 2022-03-31.

[9] Minh Nguyen, Tong He, Lijun An, Daniel C. Alexander, Jiashi Feng, and B.T. Thomas Yeo. Predicting alzheimer's disease progression using deep recurrent neural networks. *NeuroImage*, 222:117203, 2020.

[10] Steven Salzberg. Distance metrics for instance-based learning. In *International Symposium on Methodologies for Intelligent Systems*, pages 399–408. Springer, 1991.

[11] Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982.

[12] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pages 226–231, 1996.

[13] Felix M Riese and Sina Keller. Susi: Supervised self-organizing maps for regression and classification in python. *arXiv preprint arXiv:1903.11114*, 2019.

[14] Mohammad Shokoohi-Yekta, Bing Hu, Hongxia Jin, Jun Wang, and Eamonn Keogh. Generalizing dtw to the multi-dimensional case requires an adaptive approach. *Data mining and knowledge discovery*, 31(1):1–31, 2017.

[15] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

[16] Keith A Johnson, Reisa A Sperling, Christopher M Gidicsin, Jeremy S Carmasin, Jacqueline E Maye, Ralph E Coleman, Eric M Reiman, Marwan N Sabbagh, Carl H Sadowsky, Adam S Fleisher, et al. Florbetapir (f18-av-45) pet to assess amyloid burden in alzheimer's disease dementia, mild cognitive impairment, and normal aging. *Alzheimer's & Dementia*, 9(5):S72–S83, 2013.

[17] David Freedman, Robert Pisani, and Roger Purves. Statistics (international student edition). *Pisani, R. Purves, 4th edn. WW Norton & Company, New York*, 2007.

[18] Md Uddin, Md Kabir, Abdullah Al Mamun, Mohamed M Abdel-Daim, George E Barreto, Ghulam Md Ashraf, et al. Apoe and alzheimer's disease: evidence mounts that targeting apoe4 may combat alzheimer's pathogenesis. *Molecular neurobiology*, 56(4):2450–2465, 2019.

[19] 2021 alzheimer's disease facts and figures. *Alzheimer's & Dementia*, 17(3):327–406, 2021.

[20] Robert S Wilson, Y Li, NT Aggarwal, LL Barnes, JJ McCann, DW Gilley, and DA Evans. Education and the course of cognitive decline in alzheimer disease. *Neurology*, 63(7):1198–1202, 2004.

[21] Jessica L Podcasy and C Neill Epperson. Considering sex and gender in alzheimer disease and other dementias. *Dialogues in clinical neuroscience*, 2022.