UNIVERSITAT DE BARCELONA

Facultat de Matemàtiques
i Informàtica

GRAU DE MATEMÀTIQUES

Treball final de grau

An application of the Mapper algorithm to resynchronization in a model of non-ischaemic cardiomyopathy

Autor: Joan Guich Estevez

Realitzat a: Departament de Matemàtiques i Informàtica

Barcelona, 13 de juny de 2022

Contents

Introduction **iii**

1 Nerve Theorem **1**
1.1 Homotopical Nerve Theorem 1
1.2 Homological Nerve Theorem 6

2 Principal Components Analysis **9**
2.1 Singular Value Decomposition 9
2.2 Principal Components Analysis 10
2.2.1 Definition of Principal Components 10
2.2.2 Principal Components Analysis 14
2.3 *SVD* & *PCA* 14

3 Mapper and its Stability **15**
3.1 Mapper 16
3.2 Nerve Theorem into our Mapper 18
3.3 Structure and Stability of the 1-Dimensional Mapper 20
3.3.1 Extended Persistence 20
3.3.2 Reeb Graphs 23
3.3.3 MultiNerve Mapper 24
3.3.4 Stability in the bottleneck distance 25

4 Results **31**
4.1 Introduction 31
4.2 Dataset 31
4.2.1 Data visualization 33
4.2.2 Conclusions 35
4.3 *PCA* analysis 36
4.4 Applying Mapper algorithm 39
4.4.1 Mapper Results 40

4.5 Quantification in Graphs 41
4.6 Mapper conclusions 43
4.7 Contrast by Statistical Methods 44

5 Conclusions **47**

Bibliography **49**

Annex **51**
A.1 Filter Functions 51
A.2 Clustering Algorithms 53
A.3 Extra PCA plots 54
A.4 Extra Mapper plots 56

Abstract

Mapper is one of the principal tools in topological data analysis (*TDA*) that enables studying topological features of high-dimensional datasets. Many studies from different fields, such as medicine and sports, have recently applied the Mapper algorithm to extract outstanding information from data.

In this work, our goal is to substantiate the conclusions from *Comparison between endocardial and epicardial cardiac resynchronization in an experimental model of non-ischaemic cardiomyopathy* study. In particular, we look for the optimal heart region where cardiac resynchronisation therapy offers a better result. Even though the core of the study is practical, we also profoundly study the theory behind the Mapper algorithm and the statistical methods we apply throughout the process.

Resum

Mapper és un dels mètodes principals dins de la branca de *topological data analysis* (*TDA*) que permet estudiar característiques topològiques sobre conjunts de dades de grans dimensions. Recentment, molts estudis de diferents àrees, com la medicina o els esports, han aplicat l'algorisme de Mapper per extreure informació rellevant de les dades tractades.

En aquest treball, el nostre objectiu és refermar les conclusions que es van obtenir a l'estudi *Comparison between endocardial and epicardial cardiac resynchronization in an experimental model of non-ischaemic cardiomyopathy*. En concret, volem aconseguir la posició del cor òptima on una teràpia de resincronització cardíaca tingui més eficiència. Malgrat que la part principal d'aquest estudi és pràctica, també profunditzem en la teoria que hi ha al darrere de l'algorisme de Mapper i els mètodes estadístics utilitzats durant el procés.

2020 Mathematics Subject Classification. 55N31, 55P10, 62P10

Introduction

These notes were born from the idea to support the results obtained in the paper *Comparison between endocardial and epicardial cardiac resynchronization in an experimental model of non-ischaemic cardiomyopathy* carried out by a team from the Department of Cardiology in Hospital de la Santa Creu i Sant Pau. The study aimed to *"compare the acute response of biventricular pacing from the LV epicardium and endocardium in a swine non-ischaemic cardiomyopathy (NICM) model of dyssynchron"*. In other words, they wanted to find differences, or relevant information, by comparing the obtained heart testings values between a swine with or without a bipolar pacing electrode.

They tested this method on six different swine individuals. The values obtained for each swine were differentiated by the different heart regions and also by endocardial or epicardial pacing. There are three heart regions: basal, mid, and apical. Then, their results reflected that the pacing from basal regions, either from the epicardium or endocardium, produced better responses than mid or apical regions. On the other hand, they could not find any relevant information about the comparison between endocardial and epicardial pacing. After applying traditional statistical methods to their dataset, a whole set of conclusions was abstracted.

Afterwards, the Faculty of Mathematics and Computer Science of the Universitat de Barcelona joined the study in order to apply topological data analysis techniques to the same dataset. The principal aim of this collaboration was to reaffirm the results referred to the distinction between heart regions' responses. On the other hand, we were open to the possibility of finding a differentiation between the endocardium and epicardium.

Topological data analysis is a branch of applied mathematics that uses topological techniques and concepts to analyse data. Topological data analysis, commonly abbreviated by TDA, was born from the necessity to analyse high-dimensional data that traditional statistical methods could not manage.

During the last years, the collection of data in almost every area of our society has been growing exponentially. This vast amount of data plays a fundamental role

in our lives since we can abstract determinant and valuable information from it. However, there was a point in the past when we found ourselves in a situation where we had to deal with high-dimensional datasets that we could not analyse through the traditional methods we used to apply.

The principal aim of TDA is to find relevant information about the studied dataset through quantitative and qualitative topological features (e.g., clusters, branches, holes...). Intuitively, TDA tries to extract information from the shape of data.

This branch of applied mathematics is a continually growing area, and nowadays there is much investment in it. One principal reason for this significant investment is the relevant and valuable results from real-life studies. In particular, there are many examples applied to the medical field; an example could be the paper "*Identification of type 2 diabetes subgroups through topological analysis of patient similarity*" [1]. As the title can tell, they identified three distinct subgroups of Type 2 diabetes (T2D) from topology-based patient-patient networks.

Nowadays, the two principal methods used in TDA are the Mapper algorithm [Singh et al., 2007] and persistent homology.

In this paper, we apply the Mapper algorithm to study the topological features from the Hospital de Sant Pau dataset. We divide this paper into two sections. In the first one, we give a fully detailed explanation of the theoretical concepts behind Mapper and other tools we have used to analyse the data. On the other hand, the second part states the procedure and results from applying the Mapper algorithm to the given dataset.

The idea of the Mapper algorithm is, given a data set X and a well-chosen real-valued function $f : X \to \mathbb{R}^d$, to summarise X through the nerve of the refined pullback of a cover $\mathcal{U}$ of $f(X)$. For well-chosen covers $\mathcal{U}$, this nerve is a graph providing an easy and convenient way to visualise a summary of the data.

Hence, in the first section of these notes (the theoretical one), we state the necessary theory to understand the nerve concept and one of the most important theorems in TDA, the Nerve Theorem. This theorem states the relation between a nerve and the respective topological space through topological features. Furthermore, we also introduce the concepts for principal components analysis (PCA) and give a short comparison between PCA and singular value decomposition. Finally, at the end of this section, we briefly discuss the stability of 1-dimensional Mapper.

In the study of the data, firstly, we briefly introduce a complete description of the dataset we use to facilitate the reader's comprehension of everything implemented and abstracted from it. Then, we discuss the results obtained after submitting our

data to a filter function and PCA analysis and claim that the tools that we used were optimal for our study. After visualizing our point cloud, we apply the Mapper algorithm to it and present the outputs. To conclude the paper, we recapitulate, observe and examine all the results that have been obtained throughout the study, and take out some final conclusions.

Chapter 4

Results

4.1 Introduction

The principal aim of this study is to ascertain the conclusion from *Comparison between endocardial and epicardial cardiac resynchronization in an experimental model of non-ischaemic cardiomyopathy* paper using Topological Data Analysis, *TDA*, tools. In particular, we use the Mapper algorithm to analyze the data set given by the Hospital Sant Pau de Barcelona. Moreover, we are also open to new information that traditional statistics methods could not reflect. Hence, our goal is to contrast the region-dependent response to LV pacing and try to discover other features of the cardiac resynchronization therapy.

We want to remark that this study has been carried out with only a number of six pigs. This fact has its positive and negative aspects. Using statistical methods over a small number of samples can not be really useful, since such a small sample could not reflect the reality. Then, the results obtained applying TDA, in this case the Mapper algorithm, can be much more determinant. On the other hand, with a little amount of studied individuals, there is the possibility of analysing outliers and then the results would also not reflect the reality.

4.2 Dataset

The *Hospital Sant Pau* team analysed the differential effect of endocardial and epicardial pacing on the following variables at each pacing configuration: *LV* peak pressure (*LVP*), *LV dP/dtmax*, *LV dP/dtmin*, mean *ABF*, as well as *QRS* complex width and *QT* interval. The values obtained for each biventricular pacing configuration were compared with those obtained during previous dyssynchronous *RV DDD* pacing at the same *AV* delay.

In other words, they obtained data extracted from six female domestic swine responses, with bipolar pacing electrodes each, against several heart testings. Then, to check the efficiency of the bipolar pacing electrodes, they compared the previous values with those obtained without the bipolar pacing electrodes (*2-basal VD 75 pacing*). These differences were expressed as a percentage of change using the formula:

$$100 \times [(\text{value of variable } X/\text{dyssynchronous value of variable } X) - 1].$$

By applying the previous formula we obtain the variables: ΔLVP, $\Delta LVdP/dtmax$, $\Delta LVdP/dtmin$, ΔABF, ΔQRS and ΔQT. Some of these formulas are the ones we use in our study to check the existence of an improvement using the earlier mentioned method. In particular, we use the variables ΔLVP, $\Delta LVdP/dtmax$, $\Delta LVdP/dtmin$, ΔABF, and notate them as *DPDT+*, *DPDT-*, *LV*, and *FA*, respectively.

Now, we explain the variables we use in the study in detail to facilitate the comprehension throughout the paper for the readers. The four variables compare the results obtained using bipolar pacing electrodes and without them. Hence, considering that the variables represent a comparison, we explain every feature that has been compared.

(a) *DPDT+*. This variable reflects the maximal rate of rising left ventricular pressure (LVP).

(b) *DPDT-*. Samely to *DPDT+*, this one indicates the minimal rate of rising left ventricular pressure (LVP).

(c) *LV*. It represents the left ventricular pressure.

(d) *FA*. This variable stands for arterial blood flow.

Finally, we want to specify that we work with 576 points. There are three regions (*base*, *media*, and *apical*), and each of them has subregions. In particular, base and mid have the same three subzones; *posterior*, *anterior* and *lateral*. On the other hand, the *apical* region has two other subzones (*apical1* and *apical2*). Moreover, endocardial and epicardial pacing is differentiated for these eight regions. Then, we have an amount of 16 labels for every swine.

Additionally, the medical team used six different machine configurations to analyse all the mentioned labels above. Hence, we have a total of 96 points for every swine. So, all the points of all pigs sum 576 points, as we stated at the beginning.

4.2.1 Data visualization

Now, we try to understand and get information about the data by visualizing the point cloud. Hence, we have developed a Python program[1] to plot the mentioned variables in $\mathbb{R}^3$. However, we also study the planes between variables since, in general, it provides a more clear visualization of the distinguished little cluster and variables regression.

In this study, three main labels are used during the process. For every plot, we use colours to differentiate between three labels that can give us some notable information. These labels compare the endocardial and epicardial pacing, different heart regions, and the six individuals. The heart has been divided into basal, mid, and apical zones, so the heart regions' label is made out of them.

Thus, we divide this section into three subsections, one for each label, and we put the most outstanding plots[2]. All three sections ahead follow the same representation pattern. There is just one figure in each section. In every figure, we find 12 planes where all the values are reflected. The axes of every plane are two out of the four we use to analyse the data, so we can also understand the relation between them. This will also be reflected in the third section, where we implement a PCA analysis. However, the represented colours may vary from section to section depending on the reflected label.

Endocardial vs Epicardial

The following plots show the values differentiated by the endocardial and epicardial comparison. We map the epicardial points to blue and the endocardial ones to red colour for this label, and the obtained graphs are represented on the next page:

[1] You can check the complete code in **[POSAR EL NUMERO DEL ANNEX]**

[2] For all the created plots check the annexe.

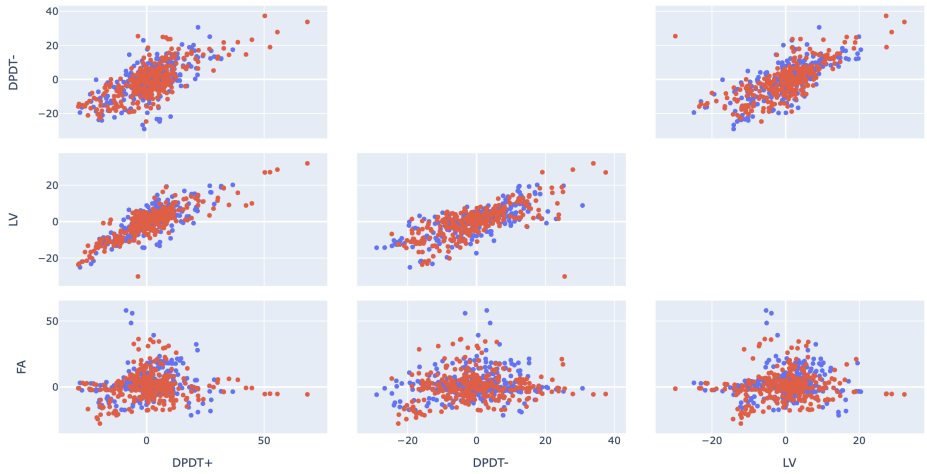

Figure 4.1: *Plot of the data in* $\mathbb{R}^2$ *differentiated by epicardial (blue) and endocardial (red) pacing.*

Regions (Basal, Mid, Apical)

Here, we reflect on the three different heart zones, i.e. the Basal, Mid, and Apical regions. The colours for the next figure are mapped in the following way: blue to basal, red to mid, and green to apical.

Figure 4.2: *Plot of the data in* $\mathbb{R}^2$ *differentiated by the heart's regions; basal (blue), mid (red) and apical (green).*

Swines

Basically, we assign a color to each swine, and the most important plots are:

Figure 4.3: *Plot of the data in* $\mathbb{R}^2$ *differentiated by swines.*

4.2.2 Conclusions

As we can imagine, there is a noticeable linear correlation between the variables *DPDT+*, *DPDT-* and *LV* for all three labels. However, there is nothing we can distinguish from the variable *FA* in that sense.

Note that every plot follows a similar pattern, i.e. there is a big centred cluster where the vast majority of the points lie, and then some other small clusters made out of a few points.

In subsection the regions plots, the small clusters out of the centre only have basal points or a mix of mid and apical. Hence, since one of our goals is to distinguish between basal to mid and apical regions, this fact can be determinant in the implementation of Mapper.

Furthermore, in the epicardial and endocardial figures, we can also find small clusters of just endocardial or epicardial pacing points. Nevertheless, we can not compare the size and frequency of these clusters to those in the regions planes.

In the other section, we can observe a differentiation of the *FIS13* and *FIS6* to the other pigs, although we can also find some separated clusters of *FIS18* and *FIS14*.

Notice that, for the variables *DPDT+*, *DPDT-* and *LV*, all the clusters mentioned

above are aligned with the central one. Then, they don't break the regression, only have considerable higher values, so we are not dealing with outliers.

In conclusion, we have visualised some meaningful results even before filtering the data. Thus, we can be optimistic about obtaining determinant information after applying Mapper with a proper filter function and cluster algorithm, mainly for the heart regions.

4.3 *PCA* analysis

As stated before, we analyse our dataset through PCA in this subsection. Firstly, we show the different plots of our data in the planes where the principal components are the axes. There are three different figures for every label we have, as we did in previous section. Nevertheless, we have decided to show the plots for the data filtered by heart regions since it is the only one we can mention relevant information about[3].

Furthermore, in this section, there are two other figures extracted from the principal components analysis that provide information about the relation of the original variables and the new set of variables, the principal components.

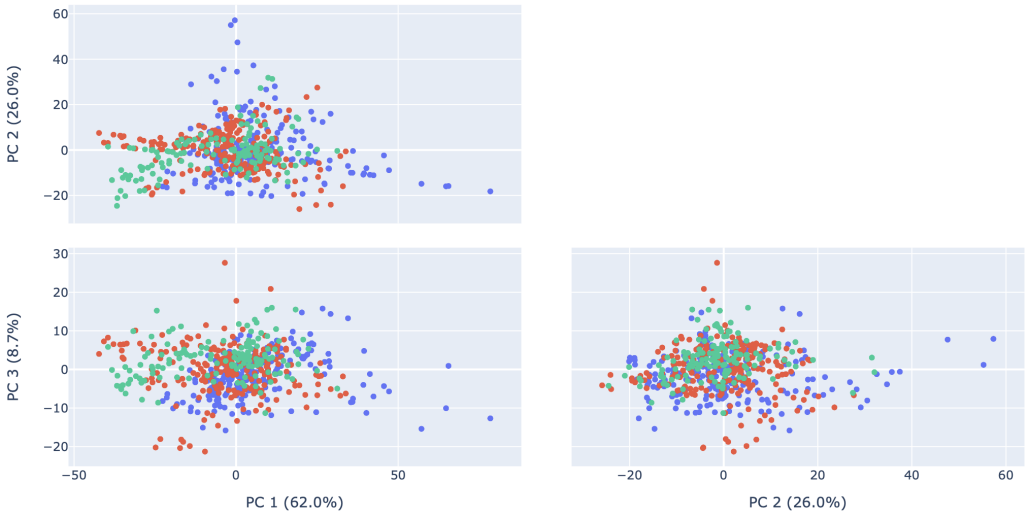

Figure 4.4: *Plot of the first three Principal Components in* $\mathbb{R}^2$ *differentiated by heart's zones; basal(blue), mid (red), and apical (green).*

[3]Check the plots for an epicardial and endocardial comparison, and for the different swines in the Annex.

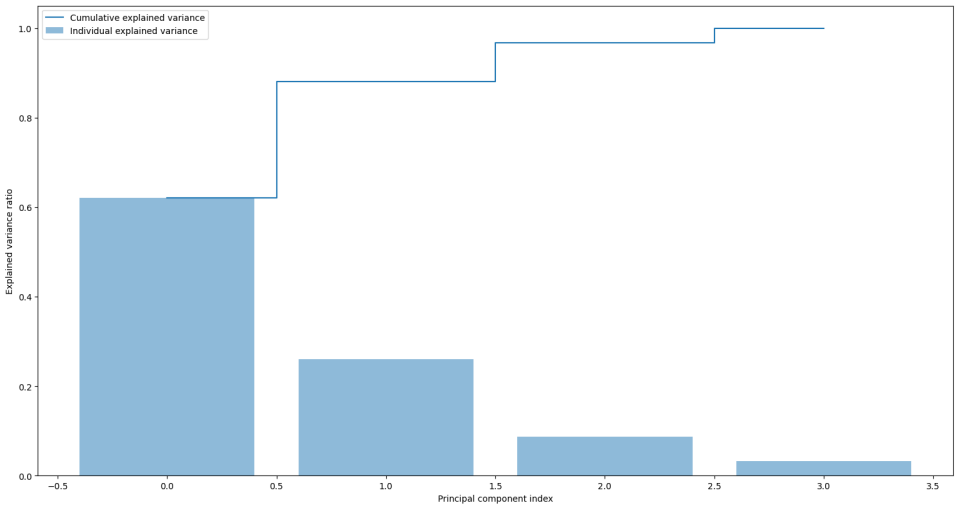

Figure 4.5: *Explained Variance histogram. Each column shows the amount of explained variance of the first kth principal component, and the line represents the cumulative explained variance of the first k principal components.*

By studying the three obtained pictures, we can state that the first and second principal components have an 88% weight over the data information. Hence, since it is representative enough, we use *PCA* with two components as our filter function for Mapper.

From now on, we focus on the planes with the first and second principal components as axes. Thus, in the plot labelled by the heart zones, we can distinguish two clusters, made out of basal points, from the centre, where most of the points lie. Also, there is an apical cluster, even though it is closer to the centre, and some mid-region points are around.

On the other hand, apparently, there are no noticeable distinctions in the epicardial and endocardial plots. Thus, it seems pretty challenging to think about getting new information from this label, although we are still open to finding new information using Mapper. Nevertheless, this ascertains the hypothesis obtained by *Hospital Sant Pau de Barcelona* related to this topic. Specifically, endocardial LV pacing induces similar haemodynamic changes to pacing from the epicardium.

Finally, from the swine differentiated plot, we can state some slight differences between the swines *FIS13* and *FIS6* from the rest. Also, some *FIS15* points are out of the centre, although quite close. Hence, we can not say that the *FIS15* case is not as straightforward as the two mentioned before.

Furhtermore, we can mention that the variables *DPDT+*, *DPDT-* and *LV* have similar directions, but this makes sense because of the earlier linear regressions. By analogous reasoning, the variable FA has a different direction than the rest. Then, as we mentioned before, the first and second principal have a significant weight over the dataset information, particularly 88%.

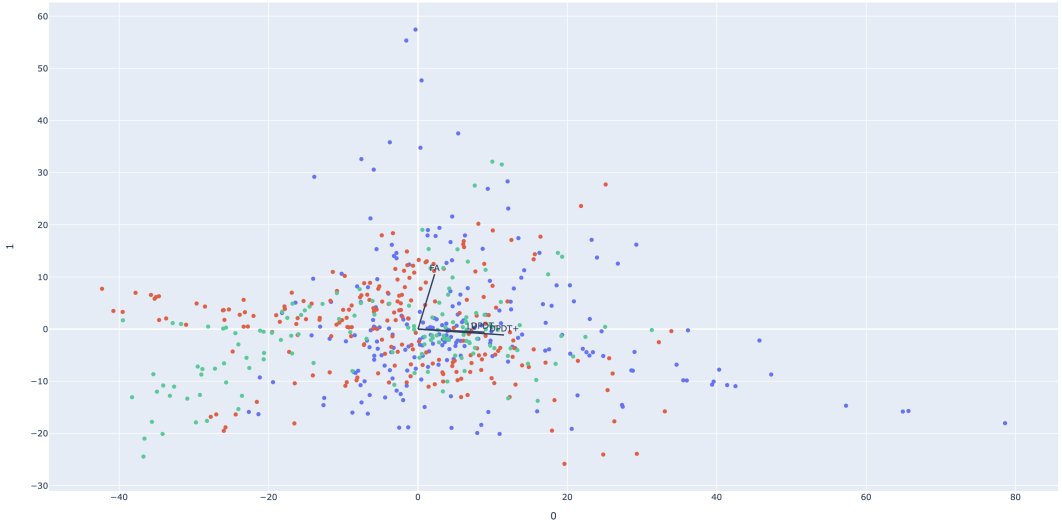

Figure 4.6: *Plot of the filtered values by PCA differentiated by heart's zones, and variables directions over the principal components.*

4.4 Applying Mapper algorithm

Now that we have already decided on the most suitable filter function for our case, we apply the Mapper in this section. However, we still have to determine the clustering algorithm, the intervals and their overlapping. We represent the results for the best clusterer[4] for our case. However, we will keep varying the other two parameters to prove some consistency in our results.

The selected clustering algorithm is the K-Means algorithm from the *sklearn* library. After trying the most common and useful clusterers, and comparing the results, we saw that K-Means was the most suitable option, even though there were repeated patterns in some of them.

It might be fair to remark that K-Means is not entirely stable. The graphs obtained are not exactly the same, but their differences are almost insignificant. Specifically, the values obtained for the variables nodes, total samples, and unique samples do not change from one graph to the other. However, the number of edges in the graph tends to vary a little. Then, even though this change may be pretty negligible, it can slightly affect the visualisation of the final result.

However, even though the clustering algorithm is not entirely stable, it is also fair to remark that the differentiation of the basal region was visible in every graph. Then, it is evident that we have chosen the figures with the most considerable differentiations after running the Python program a few times for each set of parameters[5].

We want to remark that we have identified the basal points with yellow, mid with green, and apical with purple. Then, each node has associated the color of the dominant region. However, if the representation of two colors is the same then it defines a new color in between of the two regions.

Now, we see the plots given by the Mapper algorithm with PCA and K-Means as their lens and clustering algorithm parameters. On the other hand, as we said earlier, we show several figures where we have changed the number of intervals and their overlapping to prove some consistency in the final results.

[4]For all the filter functions analysis check the annexe.

[5]You can check other graphs with the same parameters in the annexe of this paper.

4.4.1 Mapper Results

Figure 4.7: *Mapper graph obtained with seven intervals and* 35% *overlapping percentage between intervals.*

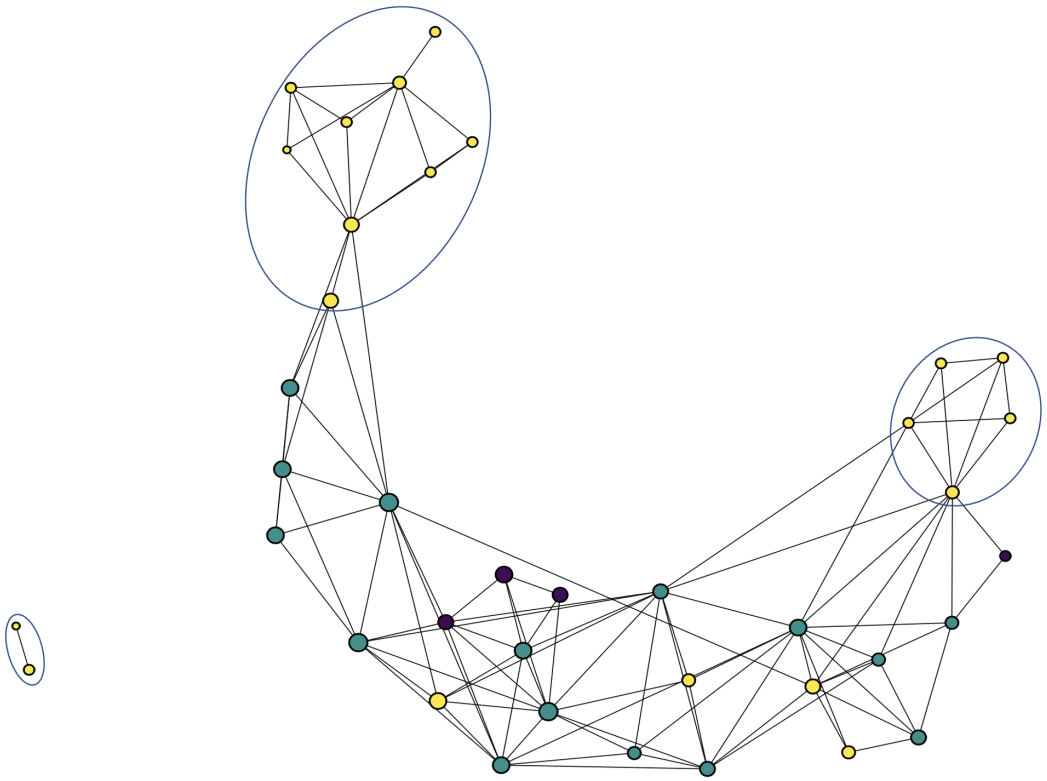

Figure 4.8: *Mapper graph obtained with four intervals and* 42.5% *overlapping percentage between intervals.*

4.5 Quantification in Graphs

In this section, we give arguments through basic graph theory to justify the results obtained from Mapper quantitatively.

Firstly, we want to remark that the information one can extract from a Mapper graph is purely from the connectivity between vertices. Hence, the distance between vertices or the distribution of the graph is relevant. However, getting visually attractive shapes helps to understand the information from the graph more easily.

Let $G = (V, E)$ be a graph, where V is the set of vertices and E is the set of edges. The elements of V, the vertices, are denoted by v_i, and the edges, elements of E, by $e_{i,j} = (v_i, v_j)$ such that $i \neq j$. We want to distinguish between the vertices depending on their colours; hence, we need to define the following vertices subsets:

$$V_Y = \{v_i \in V : v_i \text{ is yellow}\}, V_G = \{v_i \in V : v_i \text{ is green}\},$$
$$\text{and } V_P = \{v_i \in V : v_i \text{ is purple}\}.$$

Furthermore, we also need to define some edges subsets:

$$E_Y = \{e_{i,j} = (v_i, v_j) : v_i \in V_Y, v_j \notin V_y\}, E_G = \{e_{i,j} = (v_i, v_j) : v_i \in V_G, v_j \notin V_G\}$$
$$\text{and } E_P = \{e_{i,j} = (v_i, v_j) : v_i \in V_P, v_j \notin V_P\}.$$

Notice that these sets are defined by the edges that connect different coloured vertices. Then, the first characteristic of our graph that we have studied is given by:

$$\bar{X}_{DC,i} = \frac{|E_i|}{|V_i|}, \text{ for } i \in \{Y, G, P\}.$$

Intuitionally, for a given color i, we calculate the mean of the amount of edges $e_{(i,j)}$ such that $i \neq j$. Now, let us define the following subsets:

$$V_{YY} = \{v_i \in V_Y : \nexists\, e_{i,j} = (v_i, v_j) \text{ s.t. } v_j \in (V \setminus V_Y)\},$$
$$V_{GG} = \{v_i \in V_G : \nexists\, e_{i,j} = (v_i, v_j) \text{ s.t. } v_j \in (V \setminus V_G)\},$$
$$\text{and } V_{PP} = \{v_i \in V_P : \nexists\, e_{i,j} = (v_i, v_j) \text{ s.t. } v_j \in (V \setminus V_P)\}.$$

We have defined the set of vertices that either have edges connecting them to only vertices of the same colour or are not connected. Thus, we have studied the following characteristic:

$$f_{MC,i} = \frac{|V_{ii}|}{|V_i|}, \text{ for } i \in \{Y, G, P\}.$$

We can say that $f_{MC,i}$ defines, for each color, the relative frequency of vertices in V_{ii} over the set V_i, for $i \in \{Y, G, P\}$. In other words, for each colour, $f_{MC,i}$ defines the relative frequency of vertices that are not connected to other coloured vertices over the total amount of vertices of the respective colour.

Now, calculating $\bar{X}_{DC,i}$ and $f_{MC,i}$ over the two illustrated graphs, we obtained the following results:

	Intervals = 4, Overlapping = 42.5%		Intervals = 7, Overlapping = 35%	
	$\bar{X}_{DC}$	f_{MC} (%)	$\bar{X}_{DC}$	f_{MC} (%)
YELLOW	1.5	60%	0.9	66%
GREEN	2.73	13.33%	2.07	20.69%
PURPLE	3	0%	2.33	11.11%

Figure 4.9: *Table with the $\bar{X}_{DC}$ and f_{MC} values for each obtained Mapper graph.*

Observing the table above, we can affirm the following statements:

(a) For both graphs, note $\bar{X}_{DC,Y}$ is the half, or almost half, of $\bar{X}_{DC,G}$ and $\bar{X}_{DC,P}$. Then, the connectivity with yellow vertex with different coloured vertices is lower than the rest.

(b) Moreover, the values for $f_{MC,Y}$ are much higher compared to ones for $f_{MC,G}$ and $f_{MC,P}$. The previous fact implies that the yellow vertices' tendency to connect only to vertices of their same colour is greater than for the green and purple vertices.

4.6 Mapper conclusions

From the above graphs, and their quantification results, we can appreciate a difference between the basal region and the two others (mid and apical). In all shown figures, each with different interval numbers and overlapping percentages, there are some clusters where most of their samples are from the heart basal region. Notice that, in general, this does not happen with any zone in such a clear way.

Hence, we think one of the leading hypotheses we established at the beginning of this study has been accomplished. The hypothesis we are talking about is the following one:

The response of epicardial and endocardial LV pacing was regional dependent and the best response was obtained at the basal regions.

Furthermore, we get more great graphs when filtering the dataset via endocardial values than for epicardial pacing. Even though we can get differences for both of them, the frequency of obtaining highlighting graphs for endocardial is higher. Thus, we can state that it can be a slight improvement in the therapy through endocardial pacing.[6]

[6]Check some Mapper graphs sorted by endocardial and epicardial values in the annex of this paper.

4.7 Contrast by Statistical Methods

We have applied statistical methods to our dataset to compare and contrast the results obtained by the Mapper algorithm. We have computed the mean, standard deviation, standard error, and maximum and minimum value of our data for the entire data. Then, we have studied the mean, standard deviation and standard error differentiating between epicardial and endocardial pacing. The results are the following:

HEART	REGIONS			
	DPDT+	DPDT-	LV	FA
		MEAN ± SEM		
BASAL	7.076567 ± 2.763061	1.840348 ± 2.517819	3.208417 ± 1.633373	2.926700 ± 2.956022
MID	-0.465862 ± 2.686426	-2.110055 ± 2.182487	-0.631245 ± 2.006976	1.339884 ± 2.134980
APICAL	-3.245893 ± 2.579310	-0.889392 ± 2.491564	-1.961357 ± 1.832959	-0.613629 ± 2.333782
		STD		
BASAL	11.722677	10.682203	6.929815	12.541340
MID	11.397541	9.259510	8.514878	9.057954
APICAL	10.943085	10.570810	7.776586	9.901399
		MIN		
BASAL	-10.661261	-19.537381	-11.259651	-17.659084
MID	-29.072220	-28.938618	-29.929461	-21.374188
APICAL	-27.551760	-23.920343	-21.817713	-27.643564
		MAX		
BASAL	68.461892	37.571636	32.269714	58.105418
MID	30.373618	25.653401	16.408424	32.450977
APICAL	28.805269	23.890820	14.757875	34.621425

Figure 4.10: *Table of statistical methods values sorted by heart regions.*

EPI				
	DPDT+	DPDT-	LV	FA
		MEAN ± SEM		
BASAL	5.533626 ± 2.083743	1.634607 ± 2.308975	3.294722 ± 1.379208	4.088982 ± 3.062976
MID	1.281838 ± 2.439489	-0.857306 ± 2.272587	0.402796 ± 1.897390	2.327843 ± 2.197575
APICAL	-3.229637 ± 2.655434	-0.953488 ± 2.388040	-1.876403 ± 1.885841	-0.180911 ± 1.660412
		STD		
BASAL	8.840574	9.796153	5.851482	12.995108
MID	10.349877	9.641771	8.049944	9.323521
APICAL	11.266054	10.131594	8.000947	7.044533

Figure 4.11: *Table of statistical methods values sorted by heart regions via epicardial pacing.*

Analysing the content given in the three tables, we can appreciate some relevant information about the dataset. Observe a noticeable difference between the basal mean and the others. However, notice that there are high standard deviations for all regions and variables. This is consequence of the huge distances between the maximum and minimum values. Then, we mostly base our comparison using the mean and standard error (SEM) together.

ENDO				
	DPDT+	DPDT-	LV	FA
		MEAN ± SEM		
BASAL	8.619508 ± 3.275880	2.046089 ± 2.720655	3.122112 ± 1.859489	1.764419 ± 2.832750
MID	-2.213562 ± 2.864883	-3.362804 ± 2.056784	-1.665286 ± 2.091302	0.351925 ± 2.054283
APICAL	-3.262150 ± 2.519528	-0.825297 ± 2.607685	-2.046312 ± 1.791532	-1.046347 ± 2.862119
		STD		
BASAL	13.898383	11.542760	7.889145	12.018340
MID	12.154670	8.726196	8.872645	8.715585
APICAL	10.689451	11.063470	7.600826	12.142941

Figure 4.12: *Table of statistical methods alues sorted by heart regions via endocardial pacing.*

First, we examine the values from the general table, the one without differentiation between endocardial and epicardial pacing. For the variables DPDT+ and LV, we get that the intersection of the basal interval and the mid interval in this table is empty. Moreover, by comparing the basal and apical, we also obtain that the variables DPDT+ and LV are the only ones with an empty intersection between intervals. Contrarily, there is no empty intersection for any variable comparing mid and apical.

Now, we examine the table with only endocardial pacing values. Comparing the basal and mid intervals, we have empty intersections for the variables DPDT+, DPDT- and LV. Then, we can appreciate empty intersections for DPDT+ and LV for basal and apical. Again, we can not say anything about the comparison between apical and mid regions.

Finally, we analyse the epicardial values table. We can not tell anything about any differentiation (empty intersections) between basal and mid regions for epicardial pacing. However, we get an empty intersection for the variables DPDT+ and LV by comparing basal and apical. Same as the other two tables, there is nothing relevant to highlight comparing mid and apical regions.

Thus, we can conclude that for some variables, there is a noticeable differentiation between basal and mid, or basal and apical. However, there is nothing we can say about comparing the mid and apical regions. These results contrast the conclusions previously obtained by the Mapper algorithm since we can observe some differences from the basal region.

More profoundly, the endocardial pacing has more empty intersections than the epicardial pacing. Then, this reflects the fact that, for endocardial values, the Mapper gets graphs where the basal is differentiated more frequently.

Chapter 5

Conclusions

This study was carried out to complement and enhance the results abstracted by the Department of Cardiology from the *Hospital de la Santa Creu i Sant Pau*. We expected to find apparent differences in the different heart regions by studying the improvement resulting from the heart responses after being submitted to bipolar pacing electrodes. Furthermore, we were also confident about discovering significant results related to other features.

We began the paper by building a theoretical framework that would help the reader to get a fully detailed understanding of the methods and tools used later in our study. Throughout this construction, we have stated results both from topology and statistics. Firstly, an explanation of the Nerve Theorem, the core of the Mapper algorithm, was given with all the topological concepts needed to comprehend it. On the other hand, we also dedicated a section to *PCA* and its differences from *SVD*. Furthermore, we explained Mapper's methodology and offered a proof of the relation between the Nerve Theorem and our Mapper implementation. Finally, we stated some results about the stability of 1-dimensional Mapper via its connection to Reeb graphs.

The practical part started by explaining exhaustively the dataset provided by the Cardiology Department. We described the methods they followed to obtain the data, the displayed variables and their meaning. Moreover, we detailed the three labels we were willing to study and compare to abstract relevant results.

Afterwards, we reflected the data into planes taking two of the four variables we had as their axes. This visualisation allowed us to study the correlation between variables. Hence, a clear correlation was abstracted by comparing the *DPDT+*, *DPDT-* and *LV* variables, although nothing relevant was obtained for *FA*.

Later on, a principal components analysis was implemented into our dataset. From the analysis, we could obtain significant information, but mainly it was

helpful to contrast that the two principal components were the most optimal filter function in our case. We got to this decision after checking that the two principal components had an 88% weight over the data information and visualising the plots of the data filtered by them.
Finally, we provided the most outstanding results of the Mapper application in the initial data. In order to supplement the results obtained through the algorithm, we carried out a study using statistical methods and a quantification of the graphs using some basic graph theory concepts. In particular, we differentiated the basal region from the mid and apical. Hence, we can assert that there is a better improvement of the bipolar pacing electrodes if they are placed in the basal region of the heart. However, we could not get any relevant differentiation from the other labels.

Mainly, we showed the best graphs that reflected the hypothesis and goals we wanted to contrast in the introduction. However, more results can be checked in the Annex of this paper.

Even though we could not provide new information, we still offered more reasons to justify differences in the heart' zones. Since both studies have concluded with the same results, it indeed can be expected that there is a better response in the heart's basal. Hence, we are still motivated to keep studying this data by applying other *TDA* methods, such as persistent homology, to complement these studies.

We hope that all this work has a meaningful impact on the treatment of arrhythmias and is useful to future readers to clarify some aspects of the Mapper Algorithm. Indeed, we encourage these readers to go further in this research and get valuable results for the theory behind Mapper.

Bibliography

[1] Li, L., Cheng, W. Y., Glicksberg, B. S., Gottesman, O., Tamler, R., Chen, R., Bottinger, E. P., & Dudley, J. T., (2015), *Identification of type 2 diabetes subgroups through topological analysis of patient similarity*, Science translational medicine, 7(311), 311ra174. https://doi.org/10.1126/scitranslmed.aaa9364

[2] Chazal, F. and Michel, B., *An Introduction to Topological Data Analysis: Fundamental and Practical Aspects for Data Scientists*, arXiv preprint arXiv:1710.04019, (2017).

[3] Hatcher, Allen, *Algebraic topology*, Cambridge: Cambridge University Press, (2002).

[4] Cavanna, Nicholas J., and Donald R. Sheehy *The generalized persistent nerve theorem*, arXiv preprint arXiv:1807.07920, (2018).

[5] Trefethen, Lloyd Nicholas and Bau, David. Numerical Linear Algebra. Philadelphia: SIAM, (1997).

[6] Jolliffe, I.T., *Principal Component Analysis (2nd ed)*, Springer Verlag, (1986).

[7] Michael, E., *Another Note on Paracompact Spaces*, Proceedings of the American Mathematical Society 8, no. 4 (1957): 822-28, https://doi.org/10.2307/2033306.

[8] Mathieu Carriere, Steve Y. Oudot *Structure and Stability of the 1-Dimensional Mapper*, Foundations of Computational Mathematics, Springer Verlag, (2017), pp.1-64. 10.1007/s10208-017-9370-z , hal- 01633101v2.

[9] Amorós-Figueras G, Jorge E, Raga S, Alonso-Martin C, Rodríguez-Font E, Bazan V, Viñolas X, Cinca J, Guerra JM, *Comparison between endocardial and epicardial cardiac resynchronization in an experimental model of non-ischaemic cardiomyopathy*, Europace, (2018) Jul 1;20(7):1209-1216. doi: 10.1093/europace/eux212. PMID: 29016778.

[10] Kraft, Rami, *Illustrations of Data Analysis Using the Mapper Algorithm and Persistent Homology*, (2016).

[11] Singh, Gurjeet Kaur Chatar, Facundo Mémoli and Gunnar E. Carlsson, *Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition*, PBG@Eurographics (2007).

[12] Chazal, Frédéric, and Bertrand Michel, *Covers and nerves: union of balls, geometric inference and Mapper*, INRIA, Barcelona (2016).

[13] Dey, Tamal K., Facundo Mémoli, and Yusu Wang, *Multiscale mapper: Topological summarization via codomain covers*, Proceedings of the twenty-seventh annual acm-siam symposium on discrete algorithms, Society for Industrial and Applied Mathematics, (2016).

[14] Frédéric Meunier, Luis Montejano, *Different versions of the nerve theorem and colourful simplices*, Journal of Combinatorial Theory, Series A, Elsevier, (2019). hal-03247121

[15] Shlens, Jonathon, *A tutorial on principal component analysis*, arXiv preprint arXiv:1404.1100 (2014).

[16] Mathieu Carriere, Bertrand Michel, Steve Y. Oudot, *Statistical analysis and parameter selection for Mapper*, Journal of Machine Learning Research, Microtome Publishing, (2018), hal-01633106v2

[17] van Veen et al., *Kepler Mapper: A flexible Python implementation of the Mapper algorithm*, Journal of Open Source Software, 4(42), 1315, (2019), https://doi.org/10.21105/joss.01315

[18] Hendrik Jacob van Veen, Nathaniel Saul, David Eargle, & Sam W. Mangham,. *Kepler Mapper: A flexible Python implementation of the Mapper algorithm (v2.0.1)*, Zenodo, (2021), https://doi.org/10.5281/zenodo.4754451

Annex

A.1 Filter Functions

There are many possibilities when choosing our filter function for Mapper. In particular, we can use many projection functions from maths, statistics, econometrics, or machine learning. Moreover, we can also make combinations between them, so we do not have to stay with just 1-dimensional lenses. A list of all the tested filter functions and some visual examples are given below:

(a) **km.KeplerMapper().fit_transform(X, projection='__')**. Projection parameter is either a string, a *Scikit-learn* class with $fit\_transform$, or a list of dimension indices.

(b) **sklearn.manifold.TSNE(n_components=3, init='pca', perplexity = 75, metric = 'euclidean', n_iter = 5000).fit_transform(X)**. The parameter *metric* is the metric to use when calculating distance between instances in a feature array, some examples are: '*braycurtis*', '*canberra*', '*chebyshev*', '*cityblock*', '*correlation*', '*cosine*' and '*euclidean*'.

(c) **sklearn.manifold.MDS(n_components=2, metric = '__').fit_transform(X)**

(d) **sklearn.manifold.SpectralEmbedding(n_components=2, affinity = '__')**. It forms an affinity matrix given by the specified function and applies spectral decomposition to the corresponding graph laplacian. The function to specify can be one of the following ones: '*nearest_neighbors*', '*rbf*', '*precomputed*', '*precomputed_nearest_neighbors*'.

(e) **sklearn.manifold.LocallyLinearEmbedding(n_components=2).fit_transform(X)**

(f) **sklearn.manifold.Isomap(n_components=2).fit_transform(X)**

(g) **mapper.filters.Gauss_density(X, sigma = 10, metricpar=, callback=None)**

(h) **mapper.filters.eccentricity(X, exponent=1.0, metricpar=, callback=None)**

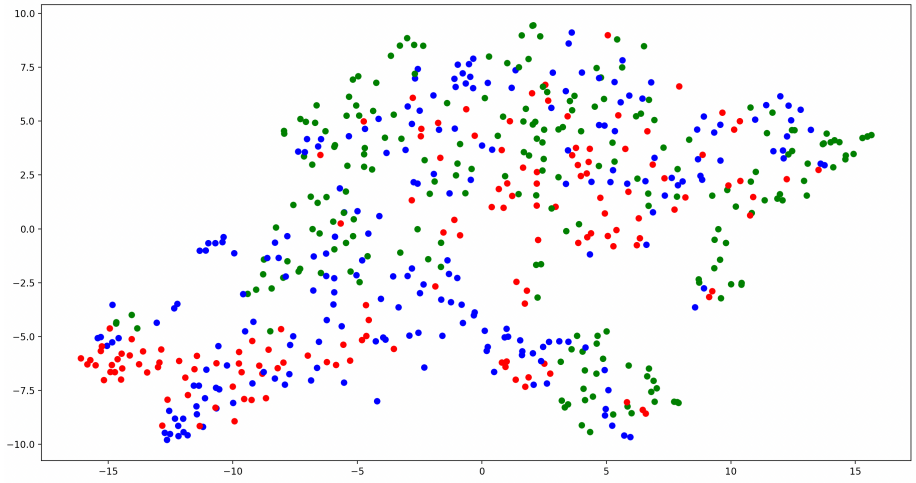

Figure 1: Data filtered by *sklearn.manifold.TSNE*

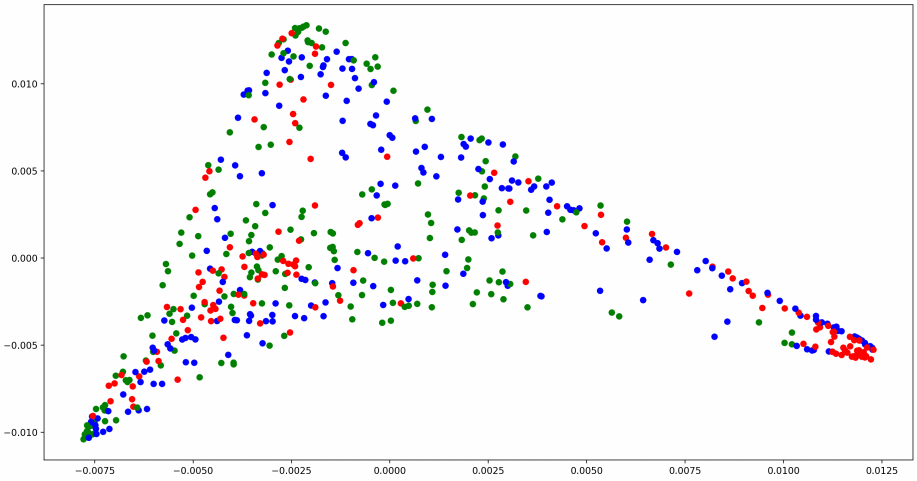

Figure 2: Data filtered by *sklearn.manifold.SpectralEmbedding*

A.2 Clustering Algorithms

Similarly to the filter functions section, we will also list the clustering algorihtms we have tested when applying Mapper to our dataset. The list is as follows:

(a) **sklearn.cluster.DBSCAN(eps=1.4, min samples=3)**.

(b) **sklearn.cluster.AgglomerativeClustering(n clusters=3)**.

(c) **sklearn.cluster.AffinityPropagation(damping = 0.8665)**

(d) **sklearn.cluster.Birch(threshold=0.000001, n clusters=3)**.

(e) **sklearn.cluster.MiniBatchKMeans(n clusters = 3)**

(f) **sklearn.cluster.MeanShift()**

(g) **sklearn.cluster.OPTICS(eps=4.5, min samples=4)**

(h) **sklearn.cluster.SpectralClustering(n clusters=3)**

After comparing the results obtained, with several parameter configurations, by this clustering algorithms with the *K-Means* algorithm we concluded that using the last was the optimal selection in our case.

A.3 Extra PCA plots

Now, we show the plots where we illustrate our dataset in a plane that takes the principal components as its axes. The first figure in this section corresponds to the dataset filtered by epicardial and endocardial pacing. As stated previously in the paper, the epicardial points are mapped to blue and the endocardial to red. Then, in the other figure, we filter the data by the different swines, each with a different colour.

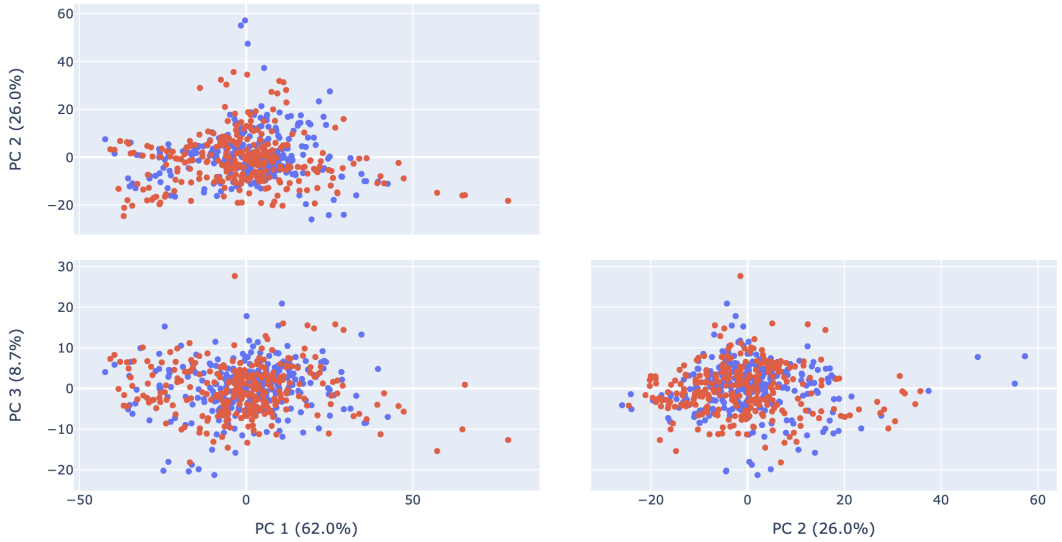

Figure 3: *Plot of the first three Principal Components in* $\mathbb{R}^2$ *differentiated by epicardial (blue) and endocardial (red).*

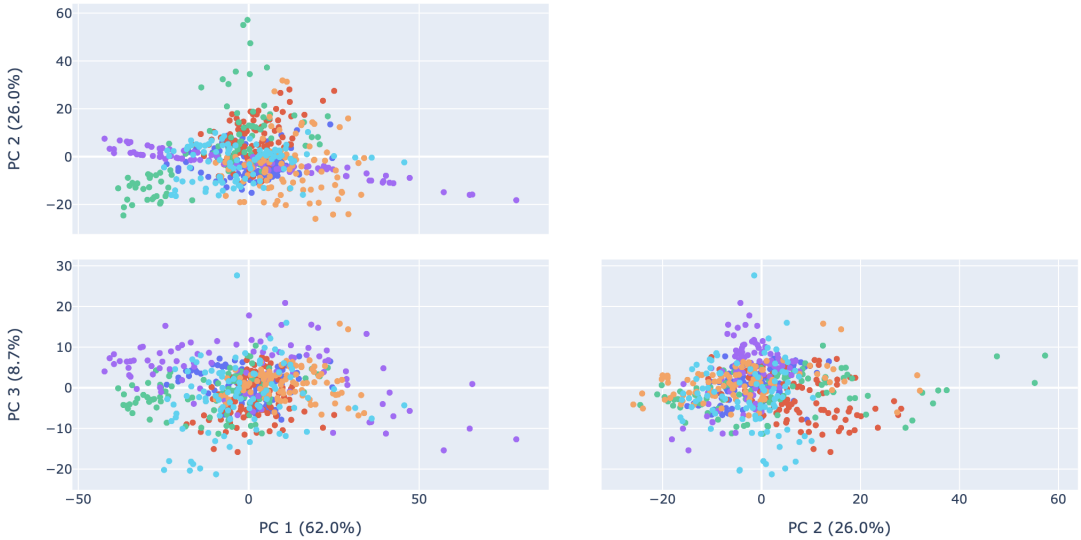

Figure 4: *Plot of the first three Principal Components in* $\mathbb{R}^2$ *differentiated by swines.*

A.4 Extra Mapper plots

In this section, we illustrate more plots where it can be possible to differentiate between basal values from the rest. The first two figures are the obtained results for the whole dataset, but with different values for the number of intervals and overlapping percentages. Moreover, we will see an extra example for endocardial pacing with different parameters. Finally, even though it is not as straightforward as in general or endocardial values, we show the most illustrative graph for epicardial pacing.

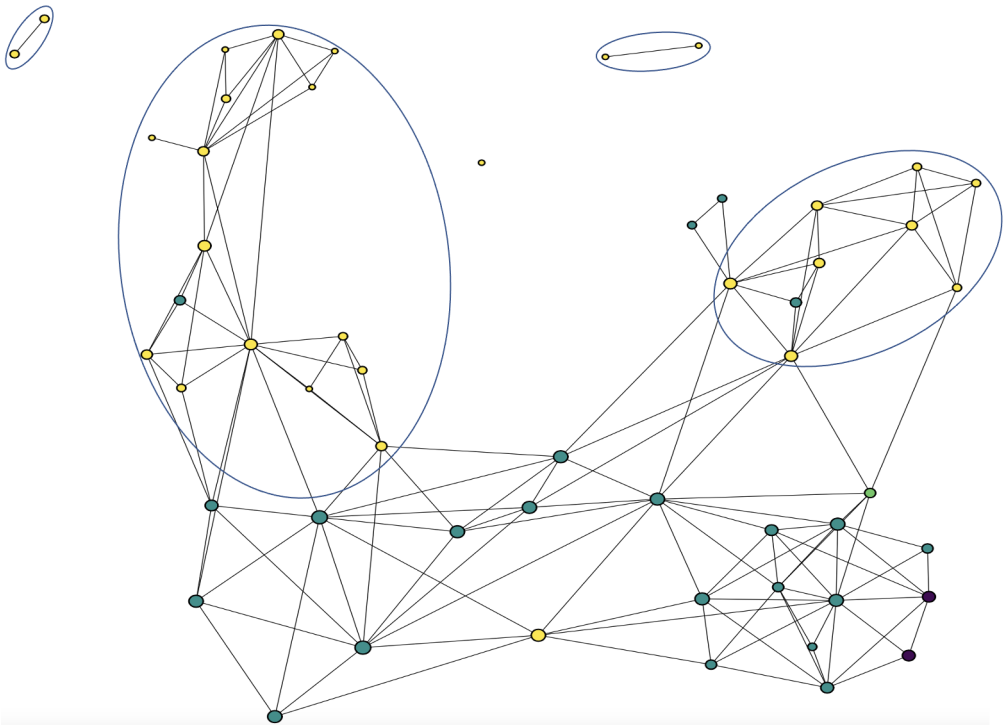

Figure 5: *Mapper graph obtained with six intervals and* 42.5% *overlapping percentage between intervals.*

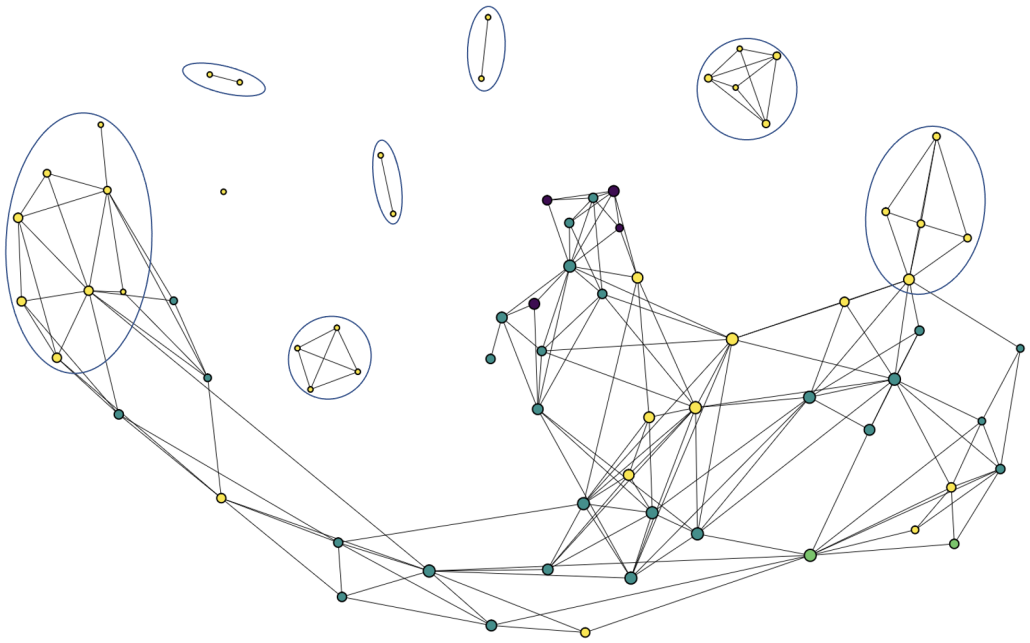

Figure 6: *Mapper graph obtained with six intervals and* 40% *overlapping percentage between intervals.*

The following figures show outstanding Mapper results applied to our dataset but previously filtered by endocardial and epicardial values. There are two plots for endocardial and just one for epicardial since, as mentioned earlier, the frequency is great graphs is higher for the first one.

Figure 7: *Mapper graph obtained with seven intervals and* 40% *overlapping percentage between intervals applied to data filterd by endocardial values.*

Figure 8: *Mapper graph obtained with six intervals and* 45% *overlapping percentage between intervals applied to data filterd by endocardial values.*

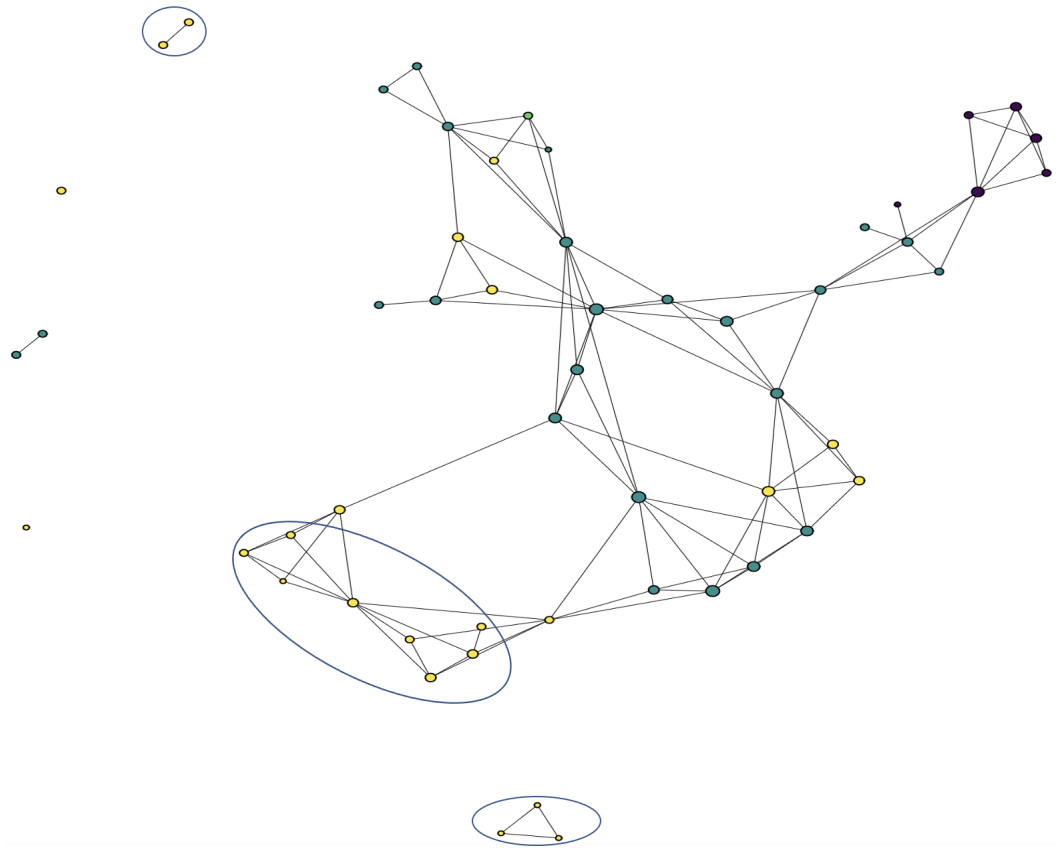

Figure 9: *Mapper graph obtained with five intervals and* 40% *overlapping percentage between intervals applied to data filterd by epicardial values.*