UCL Université catholique de Louvain

École polytechnique de Louvain (EPL)

Heart Rate Variability

Using several techniques to detect heart rate anomalies

Dissertation presented by
Maxence LARDINOIS

for obtaining the Master's degree in
Mathematical Engineering

Supervisor(s)
Pierre-Antoine ABSIL, Sébastien MARCHANDISE

Reader(s)
Emilie RENARD, Dimitri DE SMET D'OLBECKE

Academic year 2017-2018

First of all, I would like to warmly thank my supervisor Professor Pierre-Antoine Absil and Mrs Emilie Renard, for letting me realize a thesis in such a fascinating field, and for accompanying me throughout the year. Their availability and valuable advice have helped me a lot to carry out my research.

I would also like to thank very much Doctor Sébastien Marchandise, cardiologist at Saint-Luc (UCL) hospital in Brussels, for devoting time during his busy days in order to let me collect useful information.

Finally, many thanks to my family and friends, for their tremendous support and encouragement.

Contents

Introduction

The concept of heart rate variability is very old. The first writings about heart rhythm can be traced back to quotations of Herophilus (ca. 335–280 BC), who didn't only observe arteries and veins and notice their differences, but also described the arteries as pulsing rhythmically. Even before that, scientific researchers observed variations in heart frequency, but only in the last 150 years, more specific methods and ideas emerged.

After the Second World War, heart rate variability (HRV) became a clinical matter, when Lee and Hon observed for the first time in 1965 HRV fetal electrocardiography (ECG). They noted that reduced beat-to-beat variation of the fetal heart was associated with distress before other detectable symptoms (Hon and Lee 1965), a principle still used today. In cardiology, Wolf was the first in 1967 to pay attention to the relation between heart rate variability and nervous system status. Shortly thereafter, it was observed that HRV changes for patients with brain injury. [1]

It is hardly surprising that heart rate variability has strong links with the mathematical world. In 1987, Kleiger demonstrated a possible role of standard deviation in inter-heartbeat duration for predicting mortality after acute myocardial infarction, which was a starting point for several important mathematical studies in cardiology.

In this thesis work, several mathematical techniques will be implemented and applied to clinical cases. Some of these techniques, such as the Poincaré plots, have already been widely used clinically. We will implement and apply them to some groups of patients to see if it can enable us to distinguish them. Other techniques and methods will be tried out to specific clinical data with a prediction objective. None of these techniques could have been usefully implemented in this work without the valuable data obtained from the hospital *Saint-Luc* (UCL), Brussels. We are grateful to them for that.

Cardiovascular system and electrocardiography

The cardiovascular system is the organ system that allows blood to flow in the body. It consists in a blood vessels network and the heart, that makes the bloodstream possible by pumping continuously. The blood transports nutrients and oxygen necessary for the organism to the cells everywhere in the body. It also carries waste materials away from all body tissues. There are three types of blood vessels providing blood transport : the arteries, the veins and the capillaries. The rhythmic contrations of the heart propel the red liquid in the arteries. Those arteries bring the blood to all the regions of the organism. The capillaries, which are minuscule vessels, make an exchange possible between blood and cells thanks to their extremely thin wall. The blood is finally rerouted to the heart through the veins.

The heart is a vital organ. It propels blood and makes it circulate in the blood vessels network of the body. Located at the left center of the rib cage, between the lungs, the heart contracts on an average of 70 times a minute, propelling some 7000 liters of blood in the cardiovascular system. This organ is essentially composed of a muscle, the myocardium, which defines a limit between four cavities : two atria and two ventricles. The atria receive blood while the ventricles, that are larger, expel it. The ventricles are delimited by cardiac valves, which are thin elastic structures that open up and shut to let the blood pass without flowing back.

The blood vessels form a network whose total length reaches 150 000 km ! They are divided into two circuits : the pulmonary circulation and the systemic circulation. The pulmonary circulation is responsible for the re-oxygenation of the venous blood. The pulmonary artery leaves the heart's right ventricle, bringing the blood towards lungs, where it comes into contact with air. The re-oxygenated blood returns to the left atrium of the heart via the pulmonary vein. The systemic circulation is responsible for the blood supply of all the organs and tissues, through the contraction of the heart's left ventricle. When the heart is contracting, both ventricles are able to eject blood simultaneously in both circuits.

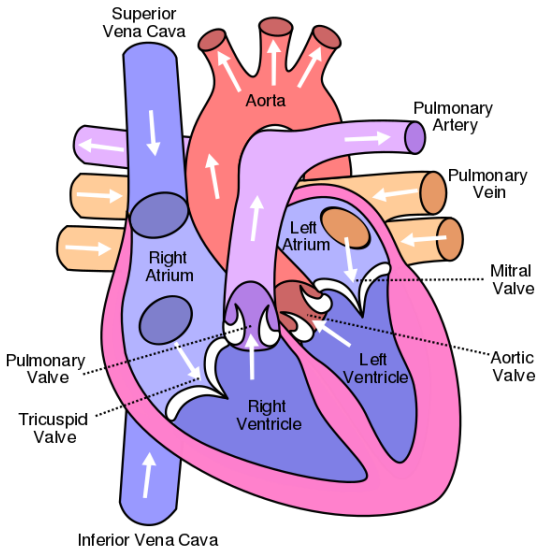


Figure 1: Front view of heart showing the atria, ventricles and valves. [2]

The cardiac rhythm contains two components :

- a mechanical component : the cardiac cycle, which is a succession of contraction (systole) phases, ejecting the blood from the left ventricle, and relaxation (diastole) phases, allowing the filling of the cardiac cavity.
- an electrical component, generating the mechanical phase which it is perfectly synchronized with.

The electric current takes its origin at a specific point of the heart, called the sinus node, having a diameter of a few millimetres, and located at the top of the right atrium.

This electric source is made of a cluster of cells capable of creating an electric current of a few millivolts. Starting from the sinus node, the current spreads throughout the two atria until their basis, causing their contraction. From this basis, it converges to the atrio-ventricular node (AV node), a node separating atria and ventricles.

From this AV node, the influx advances simultaneously towards the two (right and left) ventricles, until the apex of the heart, causing the contraction of both ventricles. This whole electric current propagation is very rapid. The cycle's average duration is 14 *milliseconds*.

Electrocardiography is the recording of the electrical activity, i.e. recording of potential variation during the cycle, thanks to electrodes placed at specific places on the skin. A typical representation of a cardiac cycle is the following :

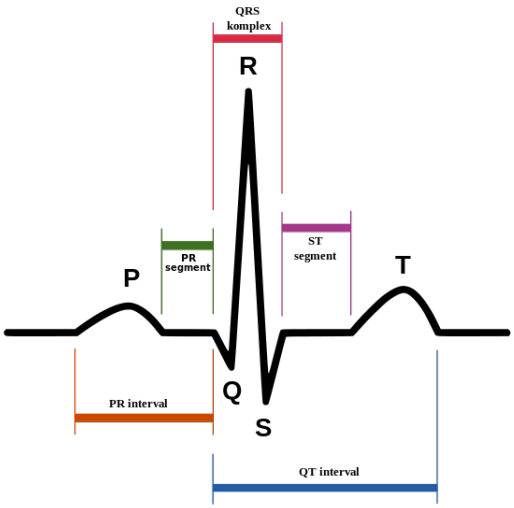


Figure 2: Typical ECG representation.
(Source : https://www.futura-sciences.com/sante/definitions/medecine-electrocardiogramme-3353/)

We know that specific parts of this diagram correspond to following events happening in the cycle:

- The P-wave corresponds to the de-polarization (which causes the contraction) of the left and right atria. It's duration is typically 0.08 to 0.1 second.
- The QRS complex corresponds to the rapid de-polarization of the ventricles, once again causing the contraction. The duration is often below 0.08 s and the amplitude is larger, which is due to the larger muscle mass, compared to the atria. During this phase, the re-polarization of the atria takes also place, which causes the relaxation of those. This hidden event is also called the atrial T-wave.
- Finally, the T-wave represents the re-polarization (and relaxation) of the ventricles.

From this diagram, it is sometimes possible to detect some heart anomalies of patients. For example, if the P-wave is of unusually long duration, it may represent atrial enlargement. Also, if the QRS complex is wide (longer than 120 ms), it can suggest ventricular rhythm anomalies such as ventricular tachycardia.

RR time series

An RR time series is a series containing the successive R-R interval durations, i.e. the time between two R-waves as depicted in Figure [2], for a patient over the period of recording. It can be obtained thanks to the ECG recordings. There are different ways of representing the series on a plot. The most common way is to plot the R-R interval duration on the Y-axis, and the interval number on the X-axis. Another way could be to use the X-axis as the time at which the specific interval occurs.

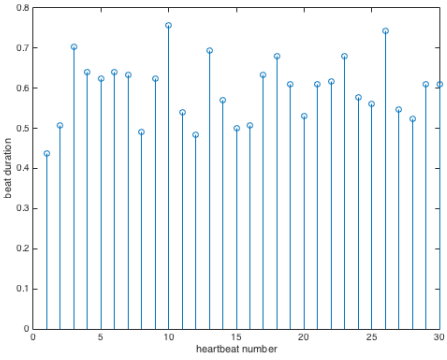


Figure 3: RR times series plot, using the X-axis as the beat number.

The plotted RR times series can sometimes reveal some anomalies for a patient, and distinguish a healthy case from some heart failure cases. For example, plotting the RR time series of a healthy patient and a patient suffering from congestive heart failure side by side gives following figure :

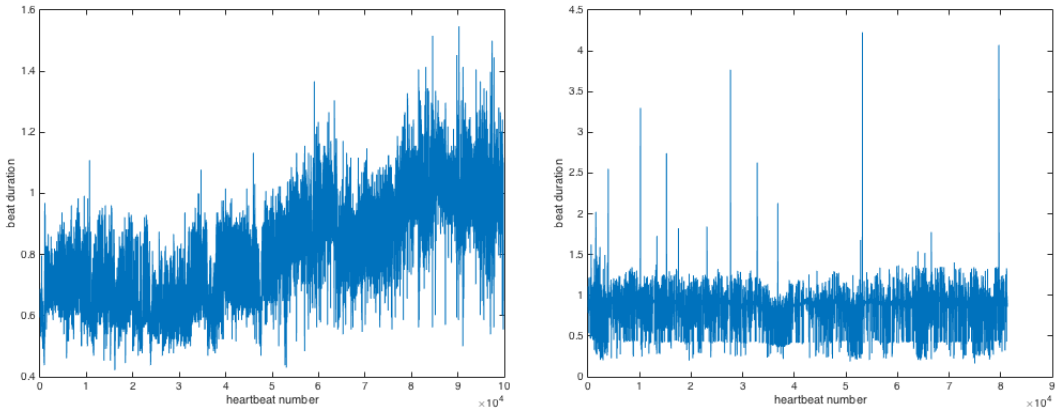


Figure 4: RR time series over 24 hours for healthy (left) versus heart failure (right) case.

Both plots represent RR time series of patients over 24 hours. We clearly observe two very different plots. For the healthy patient, we see a slight increase at the end, which might mean that the patient is sleeping or more passive. For the congestive heart failure case, we don't see any evolution of the mean interval duration over the day. This might indicate a failure of the heart to react to exterior factors, such as an active occupation. We can't even guess the sleeping hours on the plot.

Plenty of methods aim to analyze the time series of the patients. Before presenting the Poincaré plots and the power-law approximation methods in the next chapters, we can firstly state time-domain methods, for example statistical methods such as :

- SDDR : the standard deviation of the RR intervals, which is a measure of total variability of the RR intervals. This index is one of the most used value in the literature.
- RMSSD : the Root mean square of the successive differences of the RR intervals. It measures parasympathetic modulation of heart rate.

Other methods are in the area of frequency domain and exist for estimating the spectrum from the RR intervals, which include both parametric and non-parametric techniques. Finally, nonlinear methods include :

- Fractal measures (power-law correlation, detrended fluctuation analysis, multifractal analysis)
- Entropy measures
- Symbolic dynamics measures
- Poincaré plots

Methods treating nonlinear dynamics can provide additional prognostic information and complement traditional time and frequency-domain analyses of HRV.

Congestive heart failure

As will be described in the next chapter, the data used in this thesis work will consist of three patient types : healthy patients, congestive heart failure cases and finally atrial fibrillation cases. It might therefore be interesting to take a look at those two last heart affections, in order to get some intuition about future results.

Firstly, the congestive heart failure (CHF) occurs when the heart is not able to pump blood normally. Consequently, there is no longer enough blood to provide oxygen and nutrients to the organs of the body. The expression 'heart failure' doesn't mean that the heart stops beating, it means that the heart doesn't work as efficiently as it should. CHF is one of the most common causes of hospitalization for people over 65 years old. Men are globally more exposed than women.

Congestive heart failure mainly leads to two problems : the systolic dysfunction, which occurs when the heart doesn't pump enough blood to satisfy the needs of the organism, and the diastolic dysfunction, occurring when the heart can't receive all the blood coming in.

Generally, other health disorders can explain a congestive heart failure :

- A coronary disease (which leads to narrowing of the arteries, providing blood to the heart) can affect certain parts of the heart.
- A valvular disease.
- Persistant high blood pressure, which forces the heart to pump blood against a higher pressure in the blood vessels, ending up weakening the heart.
- Heart attacks damaging the cardiac muscle. Patients having suffered from a heart attack are 5 times more likely to suffer from congestive heart failure.
- Diabetes also increases the risks.
- Heart arrhythmia, which can prevent the heart from pumping blood efficiently, especially if the beatings are too fast.
- Hypertrophy, which is a thickening of the walls of a ventricle of the heart. It can prevent the heart from pumping normally.
- Kidney diseases can lead to a higher arterial pressure.
- Drugs consumption, like cocaine, or excessive alcohol consumption can considerably weaken the heart condition.

The best way to avoid congestive heart failure is to maintain a good cardiac health, which contributes avoiding heart attacks, vascular accidents and coronary troubles. This can be favoured by:

- healthy nutrition
- physical exercises
- keeping the sugar concentration in the blood at an appropriate level
- keeping the blood cholesterol low
- reducing alcohol consumption and avoid smoking

Atrial Fibrillation

Atrial fibrillation is the most common cardiac disorder, especially for subjects older than 60 years old. This pathology is estimated to concern 1% of the global population, and 10% of people aged 80 and over in Belgium and France. [3]

Atrial fibrillation (AF) corresponds to an anarchic contraction of the atria, leading to chaotic and irregular contractions of the ventricles. It occurs when the muscle fibers are all contracting at different times, resulting in quivering or twitching movement. Normally, an electrical signal is sent out from the sinus node in the right atrium. It then rapidly spreads through both atria rapidly, allowing them to depolarize at about the same time. The signal then moves out to the ventricles and causes them to contract shortly after. With atrial fibrillation, the signals move around in the atria in a completely disorganized way, that tends to override the sinus node. Figure 5 depicts this electrical disorganization. Instead of one 'big' contraction, we get mini-contractions as if atria are just quivering. On a ECG, this quivering can nicely be observed :

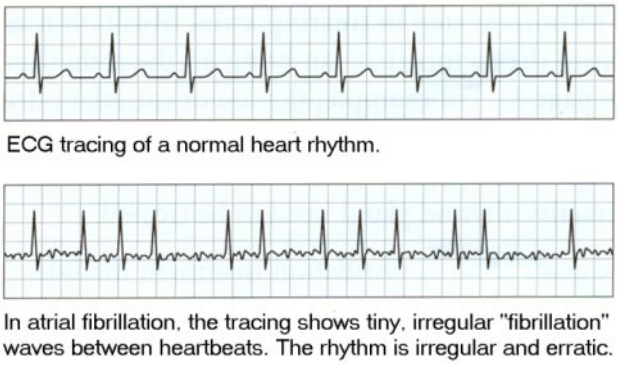


Figure 5: ECG of a normal case versus an atrial fibrillation case.
(Source : https://www.chss.org.uk/heart-information-and-support/about-your-heart-condition/common-heart-conditions/heart-arrhythmias-2/medical-treatment-atrial-fibrillation/)

Indeed, we cannot distinguish the classic P-wave from this ECG, as many little peaks appear corresponding to this "quivering'. Sometimes, a signal reaches the ventricles and causes the ventricular contraction. These contractions and QRS complexes are interspersed at irregular intervals though, and usually at fairly high rates between 100 and 175 beats per minute.

There are many risk factors that predispose someone to developing AF, and the exact mechanisms aren't always perfectly understood. AF often happens alongside other cardiovascular diseases, like high blood pressure, coronary artery disease, valvular diseases. These pathologies can all create an inflammatory situation or create stress on the atria and potentially damage the cells in those. All those risk factors can stress the cells in the atria, which can lead to tissue heterogeneity, meaning that cells begin to take on different electrical properties. This can ultimately cause the conduction in the atria to become unpredictable.

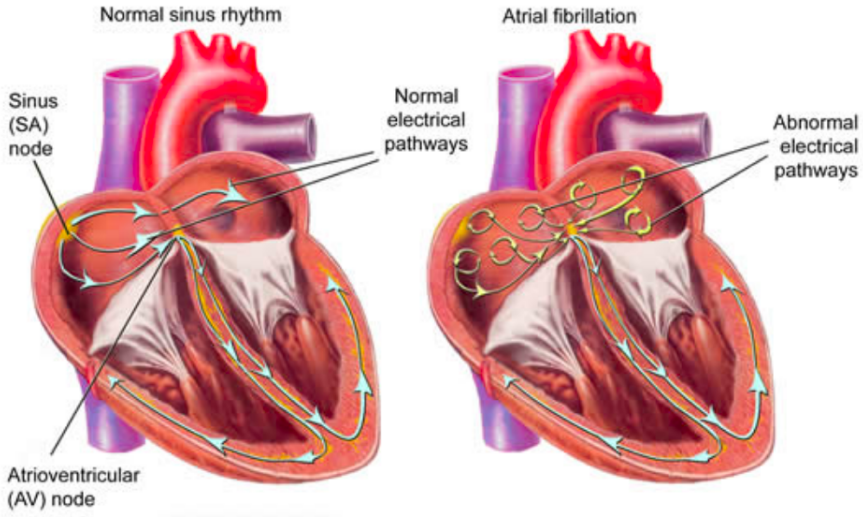


Figure 6: Abnormal electrical conduction of the cells in atria.
(Source : http://www.drholdright.co.uk/dynamicpage.php?pg=cardiac-conditionspageid=NDI=)

There are several forms of atrial fibrillation :

- paroxysmal AF : the AF episodes last a few seconds up to a few days. In general, they arise and then cease spontaneously.
- persistent AF, which can last weeks up to months. Usually, a treatment can end it.
- permanent AF : the AF is constant. This can occur when different treatments cannot correct the anomaly and it is not likely that the cardiac rhythm gets back to normal.

A common symptom of AF is a feeling of general tiredness, since the heart rate is no longer guided by the sinus node, and contractions come at irregular intervals, delivering blood less effectively to the tissues. Other related symptoms include dizziness, shortness of breath and weakness. Patients might also feel palpitations or sensations of "thumping" heart.

Diagnosis of persistent AF is done by ECG. If the episodes are paroxysmal, but they look suspicious, it might be advised to have a Holter monitor, which is a portable device placed on the chest monitoring the rhythm over longer periods of time and recording potential AF events.

Since atrial fibrillation is caused by a wide variety of issues, treatment is generally different from one patient to another. Medications helping heart rate control might be given, or medications that reduce the likelihood of blood clot (a clump of blood that has changed from a liquid to a gel-like or semisolid state) formation and therefore prevent stroke. Also, patients might receive an implantable cardiac pacemaker, which can reduce the chance of an AF episode by constantly pacing the atrium. Finally, some patients may have a radiofrequency catheter ablation, a medical procedure in which certain areas of tissues are destroyed, so that the electrical signal no longer spreads from these areas.

The data used in the second part of this work was collected at the hospital *Saint-Luc (UCL)* in Brussels. This data consists of a set of patients having undergone a catheter ablation procedure. It might therefore be interesting to have a few words on this procedure.

Catheter ablation for atrial fibrillation

A catheter ablation is a non-surgical intervention which uses long flexibel tubes, catheters, in order to get into the inside of the heart. The intervention generally requires general anesthesia. This technique is more and more used, thanks to the emergence of new technologies. In 2010, the number of catheter ablations doubled in Belgium (2100 ablations). [3]

In order to realize this intervention, one or more catheters are threaded through blood vessels and routed towards the heart by using x-ray fluoroscopy (a 'moving radiography'). The catheters are used to study abnormal cardiac beatings and detect spots where a problem is occurring. Once the defective tissues have been located, an ablation catheter is routed and positioned close to the defective area. The catheter tip then emits an electric energy at high frequency, which destroys the defective tissues and creates a scar. Afterwards, the scar tissues are neutralized and unable to emit an electrical signal causing arrhythmia. In other words, the short circuit disappears.

20 years ago, the basis for the development of catheter ablation was established when it was observed that the ectopic foci (abnormal pacemaker sites within the heart that display automaticity) originating from the pulmonary veins were capable to trigger AF. Since then, circumferential ablation around the orifices of the pulmonary veins leading to electrical disconnection of the pulmonary vein from the left atrium has become the cornerstone of catheter ablation. The ablation consists of a series of point-by-point radiofrequency lesions encircling each or both orifices of the pulmonary veins :

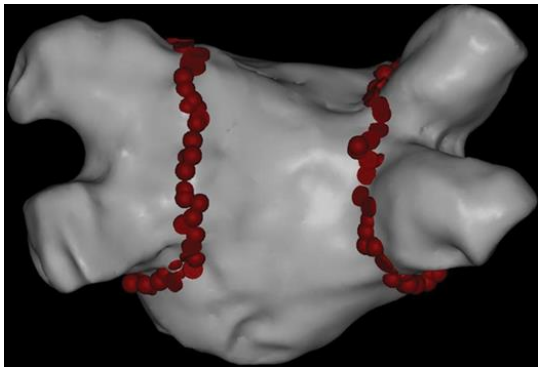


Figure 7: Circumferential ablation lesions (red points) around ipsilateral pulmonary veins [4].

An other approach often used today, is to encircle both pulmonary vein's ostia (orifices) on one side, with one single wider elliptical line. Nevertheless, the surgeon must always ensure that each pulmonary vein is disconnected.

Just like any other medical intervention, there are some benefits and risks associated to the catheter ablation. An important benefit of a successful ablation is the reduced occurrence of the symptoms such as the shortness of breath, weakness feeling or tiredness. Some risks include cerebrovascular accidents, pericardial tamponade (when fluid builds up in the pericardium and results in compression of the heart), or narrowing the pulmonary veins and irritation.

Chapter 1

Data description

1.1 Physionet challenge and open-source data provided

In 2008, the editors of Chaos [5] announced a new column for their journal, *"Controversial Topics in Nonlinear Dynamics"*, whose first topic was *"Is the Normal Heart Rate Chaotic?"*. In order to help researches to be made, the open-source website `Physionet` has provided a set of 15 heart beat (RR-interval) time series from healthy people and patients with disease (congestive heart failure and atrial fibrillation). All of those time series were gotten by continuous ambulatory (Holter) electrocardiograms (ECG), during a 24 hours period, provided as a text file for analysis. Along with the RR-interval duration, the text file indicates the type of heart beat that ended the RR interval (N is normal, and anything else is abnormal).

Unfortunately, those file don't contain any information about the patients themselves, such as the age, gender, tabac information etc. Also, no information is given as to when patients sleep or are doing sport for example. Nevertheless, is can be guessed when a subject is sleeping during a period of several hours, as the average of inter-heartbeat duration significantly increases (slower beating heart).

It might be useful to apply a filter on the signals, as the recordings can contain outliers or noise. For example, a Savitzky-Golay filter aims at smoothing the data, which can be done by fitting successive sub-sets of adjacent data points with a polynomial. In order to visually distinguish sleeping hours from more active hours, one can apply such a filter with an long frame-length parameter, such that the data is extremely smoothed :

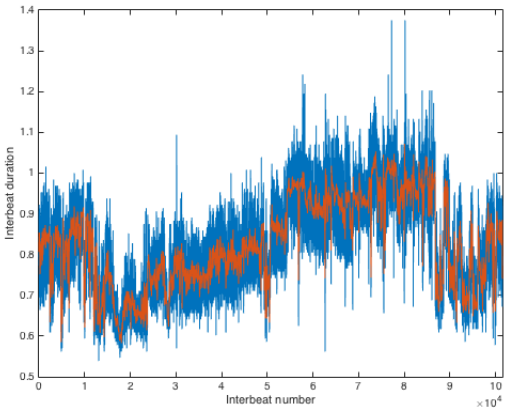


Figure 1.1: RR-interval time series along with the Savitzky-Golay filter in orange.

This figure represents the electrocardiogram recording of a healthy patient, over 24 hours. It can be visually observed that the patient is probably sleeping or passive in any case on the interbeats range $5.5 \times 10^4 \rightarrow 8.5 \times 10^4$. The same reasoning can bring us to the conclusion that this patient is more active on the interbeats range $8.5 \times 10^4 \rightarrow 9.5 \times 10^4$.

1.2 Saint-Luc (UCL) hospital data

The second part of this work is dedicated to classification of patients whose data were collected at the Hospital *Cliniques universitaires Saint-Luc*, in Brussels. Unlike the data introduced in the previous section, this data concerns a set of 68 patients having, at least once, undergone an atrial fibrillation ablation. The data from the hospital can be divided into two parts : one set of text files and one global information sheet. The text files are the recordings of the atrial rate of each patient during the ablation. For some of those patients, we know precisely the disconnection time for each pulmonary vein (as explained earlier), while it remains unclear for others at what time those disconnections happen. The other part of the data is the global sheet, that contains information on each patient of the dataset, such as gender, age, and useful medical information, that can be used for classification.

1.2.1 Atrial rates in text files

For all 68 patients, we had access to the recordings of the atrial rate from a time before the beginning of the whole ablation procedure until a period after the intervention. This allows to see how the rate evolves through the procedure. The frequency of sampling for the recording is known, and this recording consists in potential measured at each sampling time step. For example, this is a 30-seconds recording of one patient (reference : `A38284P`) right after a first pulmonary vein disconnection :

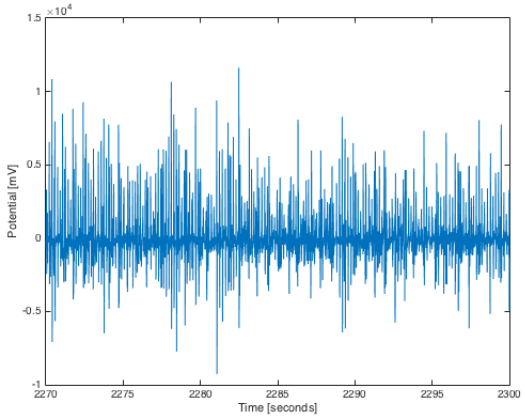


Figure 1.2: Atrial rate for patient 'A38284P' right after first pulmonary vein disconnection.

Again, those recordings might contain noise and outliers, which can encourage us to apply a filter, such as the Savitzky-Golay filter introduced earlier. Applying the filter and zooming in, in order to better notice the peaks, produces following figure for the same patient :

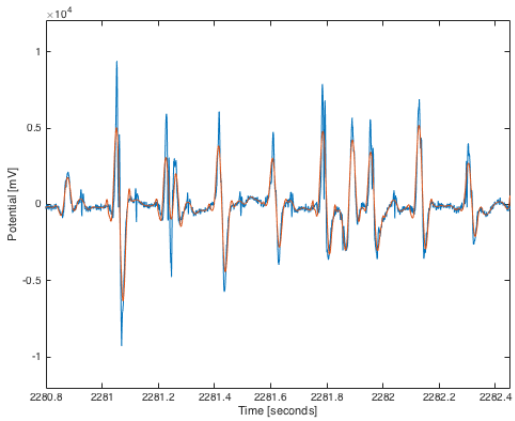


Figure 1.3: Atrial rate for patient 'A38284P' right after first pulmonary vein disconnection, along with Savitzky-Golay filter in orange.

We can see the effects of the Savitzky-Golay, as the signal is smoothed. The first part when analyzing such signals will be to detect the peaks of the atrial rate, that we just filtered. Once we have all the peaks, we will have the inter-peak time series on which we can extract useful information such as the maximum inter-peak duration or the number of beats on a time interval. Unfortunately we could not get any disconnection times for 14 patients, which reduces the database to 54 patients for which we have complete information.

1.2.2 Global repertory sheet

The text files presented here above contain the recordings during the ablation procedure. A valuable complement to those files is an `EXCEL` sheet containing global information on each of the patients. Out of a total of 68 patients, 19 of them have relapsed and had to undergo a new ablation procedure. An important aspect to be considered is the gender and the age of the patients. Following chart represents the gender and the relapsing percentage in the dataset :

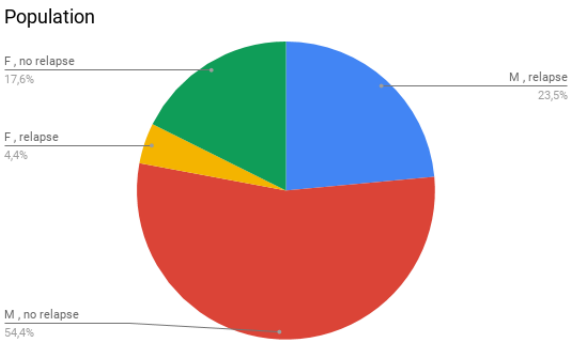
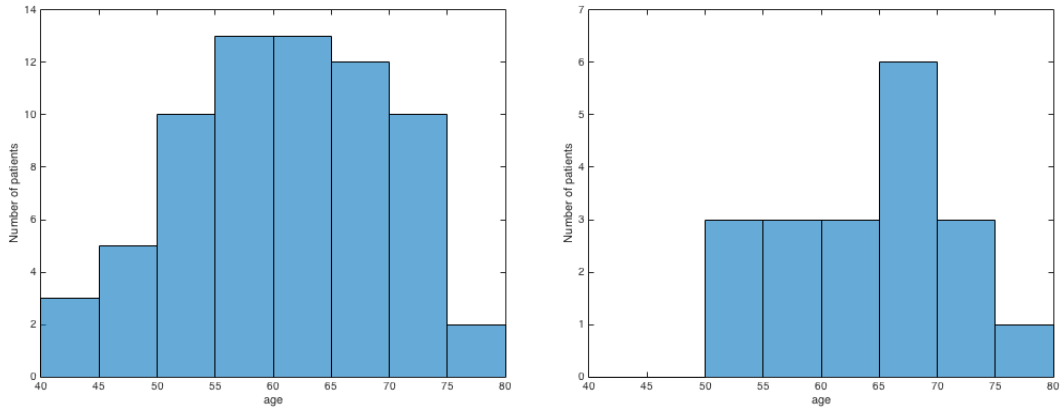


Figure 1.4: Gender and relapsing percentages in the dataset from (UCL) Saint-Luc Hospital. F stands for the female patients while M stands for the male patients.

In other words, 15 women and 53 men are present in the database. 30% of the men have relapsed while only 20% of the women have relapsed.

If we consider the age of the patients and the age of the relapsing patients only, we can observe following distribution :



(a) Histogram showing the age distribution of the total population in the dataset.

(b) Histogram showing the age distribution of the relapsing population.

Figure 1.5: Age distribution of the total population versus relapsing population.

Observing this last figure, one can notice that, not surprisingly, the relapsing population subset is older than the total population undergoing catheter ablation. The peaks of the histogram for the whole population is at the age range $55 \rightarrow 65$, while for the relapsing population the peak is at $65 \rightarrow 70$ years old.

In the repertory sheet, we also have more medical information on the patients such as whether the patient is a smoker or not (15 patients), has diabetes (8 patients) etc. All of this available information can help to create a classification model, where the ultimate goal would be to be able to predict whether a patient would relapse or not after an ablation procedure for atrial fibrillation, based on a set of criteria he would satisfy or not. By doing so, one has to stay careful about the information he is using, as some criteria might be unhelpful for the model, or might even mathematically influence the model's results, while the interpretation of these criteria would mean something different medically. If we take for example tobacco information we have on the patients in the database, we see that :

- A third (33.33 %) of smoking patients in the population relapse after ablation.
- While only 26.42% of the non smoking patients in the population relapse after ablation.

This observation encourages us to use the smoking criterion on the patients in a model, as it could seem logical that smoking might worsen the heart condition of patients. If we now take a look at diabetes information, we observe that :

- Exactly 25% of the patients suffering from diabetes relapse after ablation procedure.
- While 28.33% of the patients non affected by diabetes relapse after ablation.

The difference between patients with and without diabetes is not blatant, nevertheless this statistic shows that patients in the database with diabetes relapse less than the others. This might for example be explained by a more careful and healthy way of living for those patients. Nevertheless, the difference is small and we want to avoid letting the model think that that patients with diabetes are less likely to relapse.

A final part of this database that has to be discussed is the information on possibly deceased patients. This concerns two of the patients. Both of them have had a relapse after ablation procedure, which eases the modelling part. As described later, these patients were not distinguished from the others, and we simply considered them as relapsing patients.

To conclude this section, a classification has been made possible thanks to a set of information we have at our disposal for each of the 68 patients. For 54 of them, we have recording files during the ablation with the needed disconnection times for each pulmonary vein during this ablation. This will enable us to add useful parameters, extracted from those file, to the set of information, in order to get a more accurate classification for these 54 patients.

Chapter 2

Poincaré plots

A wide variety of factors influence the heart rate, such as respiration or mental load. Many techniques suggested by nonlinear dynamics have therefore been applied to the classification of heart rate variability (HRV), motivated by the high complexity and the nonlinear interactions between different physiological subsystems. Several of those techniques have been proven to be of diagnostic relevance.

Amongst them, following nonlinear methods are quite renowned in cardiology :

- Fractal measures (such as the detrended fluctuation analysis, power-law correlation, ...)
- Entropy measures (sample entropy, compression entropy, ...)
- Symbolic dynamics measures
- Poincaré plot

The Poincaré plot was first used in 1992 as a qualitative tool (Woo et al. [6]) and later as a geometrical analysis by fitting an ellipse to the shape of the plot in order to calculate HRB indices (Tulppo et al. [7]). The Poincaré plot analysis is a geometrical and nonlinear method to assess the dynamics of HRV. The plot is a representation of a time series into a phase space, where the values of each pair of successive elements of the time series define a point in the plot. The great advantage of this method is that the plots are easy to understand and interpret, the drawback being the lack of temporal information about the 24h EGC recordings used as data.

2.1 Construction of the plot

The construction of Poincaré plots is quite direct and intuitive, given the time series of the RR time intervals. Skimming through the RR time series, and taking each successive pair one by one, we let the first RR interval, $RR(i)$, represent the x-coordinate, and the second interval, $RR(i+1)$, represent the y-coordinate of the point to be added.
The typical shape of a Poincaré plot is an elongated cloud of points around the line of identity.

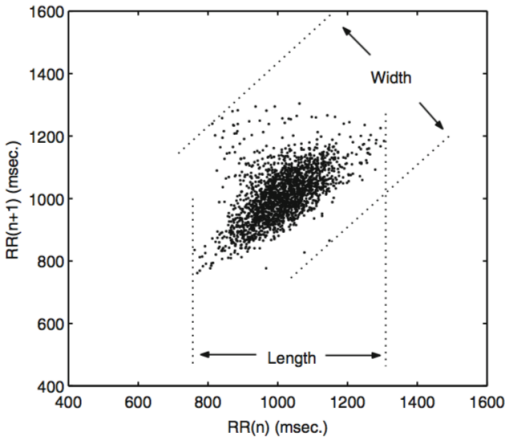


Figure 2.1: Construction of a Poincaré plot

A quantitative analysis can be made by adjusting it to an ellipse. The evaluation parameters are then two different standard deviations, $SD1$ and $SD2$, and the area of the ellipse.

2.2 Ellipse-fitting technique

Most researchers adopt the ellipse-fitting technique to characterize the Poincaré plot mathematically. The ellipse's major axis is aligned with the line of identity (passing through the origin with a slope of 45°), and the minor axis is perpendicular to this major axis. Finally, the intersection point is the centroid of the plot.

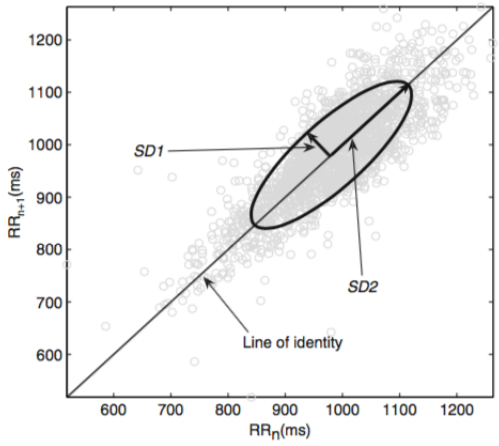


Figure 2.2: Construction of a Poincaré plot

The standard deviation $SD1$ corresponds to the minor axis of the ellipse and is the standard deviation of the instantaneous (short term) beat-to-beat RR interval variability. This Poincaré plot's width is considered as a pure measure of parasympathetic activity. It is computed as follows :

$$SD1 = \sqrt{var\Big(\frac{\overrightarrow{RR_t} - \overrightarrow{RR_{t+1}}}{\sqrt{2}}\Big)}$$

where $\overrightarrow{RR_t} = (RR_1, RR_2, ..., RR_{N-1})$ and $\overrightarrow{RR_{t+1}} = (RR_2, RR_3, ..., RR_N)$.

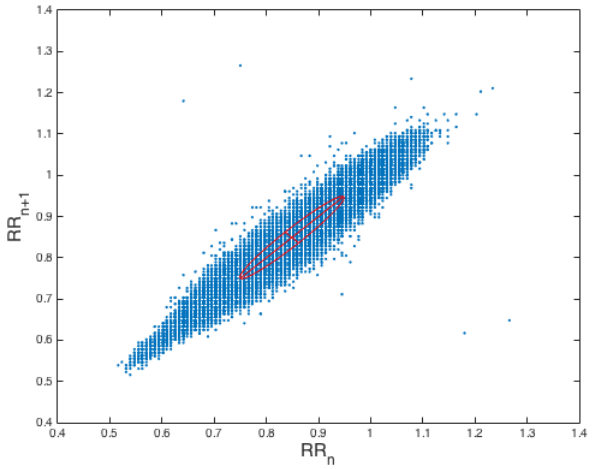
Equivalently, the standard deviation $SD2$ corresponds to the major axis of the ellipse. It represents the long-term RR interval variability and is given by :

$$SD2 = \sqrt{var\Big(\frac{\overrightarrow{RR_t} + \overrightarrow{RR_{t+1}}}{\sqrt{2}}\Big)}$$

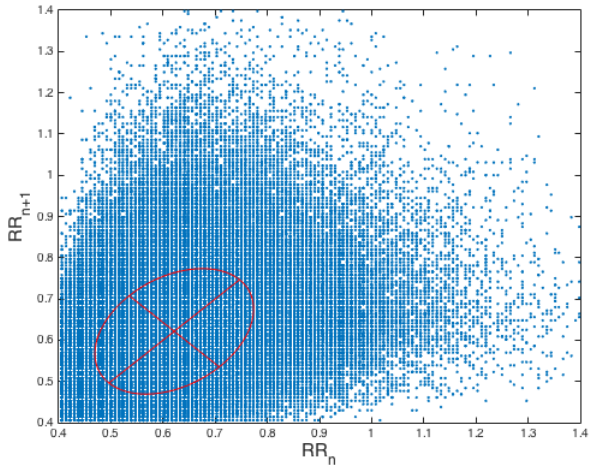
Finally, the area of the ellipse is easily obtained : $S = \pi.SD_1.SD_2$

2.3 Implementation tests and observations

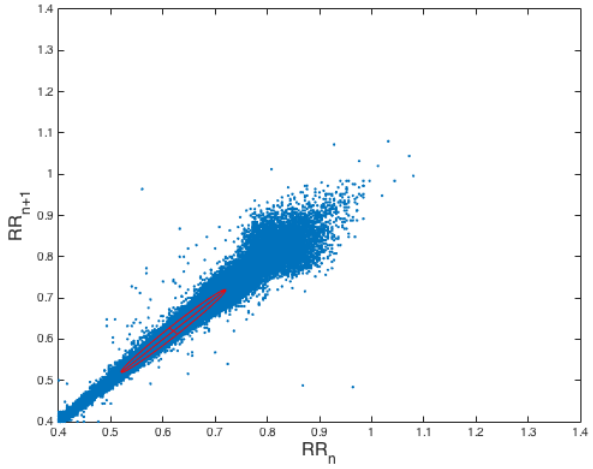
Many researches have shown that healthy patients are likely to produce bigger values for the ellipse area than patients with cancer or heart diseases such as congestive heart failure, reflecting in lower heart rate variability. To verify this, 15 ECG recordings from `PhysioNet` ([8]) over 24 hours were used, classified into three groups. 5 of those correspond to healthy patients, 5 others to congestive heart failure and finally 5 to patients suffering from atrial fibrillation. Constructing Poincaré plots for each of those ECG recordings, one can have a nice visual idea of which group each patient belongs to.



(a) Poincaré plot for a healthy patient



(b) Poincaré plot for a patient suffering from atrial fibrillation



(c) Poincaré plot for a patient suffering from congestive heart failure

Figure 2.3: Poincaré plots typically observed for the three different groups. The same scale is used for all three plots.

Before tackling the quantitative, mathematical results, one can already see that, compared to healthy normal cases, the points constituting the Poincaré plots related to congestive heart failure are much more dense. Using the same scale on all three Poincaré plots, it's easy to see that, unlike the healthy case, it's hard to find any white space between any point for the heart failure. This observation is quite intuitive, as we expect a congestive heart failure patient to have a smaller heart rate variability, having a heart less robust to exterior factors such as stress or astonishment, where the heart rate should increase for a healthy patient. More precisely, this translates into a smaller value for the standard deviation $SD1$ (the width). No immediate relevant observation can visually be made about the length $SD2$.

For the atrial fibrillation cases, the observation is on the opposite side. Here, the points of the Poincaré plots are much more spaced all around the phase space. The plot is a bit chaotic in fact, which is also quite intuitive, as atrial fibrillation implies irregular heartbeats (arrhythmia) that can lead to blood clots, stroke, and other heart-related complications. This irregular heartbeat is nicely represented by the chaos on all the Poincaré plots for those cases. This time, it translates mathematically into a greater value for the standard deviation $SD1$ while no immediate relevant observation can visually be made about the length $SD2$.

Following values are the computed standard deviations for each of the plots here above :

Case	$SD1$ (in ms)	$SD2$ (in ms)	Ellipse area S
(a) healthy	18.31	140.02	0.0081
(b) atrial fibrillation	122.05	177.09	0.0679
(c) congestive heart failure	10.9403	141.6140	0.0049

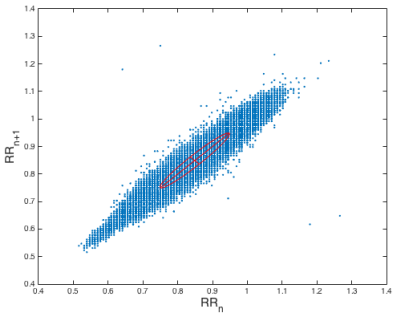
This confirms quite well the visual expectations made before. Globally, those observations hold on well to all of the other 15 patients for which an ECG recording over 24 hours was obtained. All the results are to be found in the APPENDIX.

To finish this section let's note that Poincaré plots are a nice visual tool, along with quantitative statistical analysis that can accurately help to detect some heart rate anomalies. Nevertheless, it has some limitations, the main one being the lack of temporal information. The two parameters $SD1$ and $SD2$ represent the signal in two-dimensional space, information about width and length of the graph, but the temporal dynamics are not included at all. This means that two Poincaré plots could be quite similar, although having completely different underlying temporal variations.

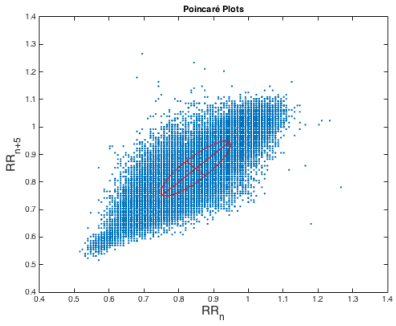
2.4 What about another shift ?

As explained here above, Poincaré plots constist in plotting two successive RR interval side-by-side. One could wonder what the result would be if instead of plotting successive intervals, we would plot interval i along with $(i + D)$, where D would be the chosen shifting parameter.

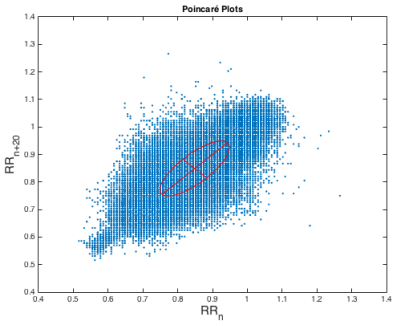
Let us see how a plot would evolve for an individual patient by increasing the shift parameter :



(a) Initial Poincaré plot of a healthy patient with a shift $D = 1$
(successive interbeats considered)



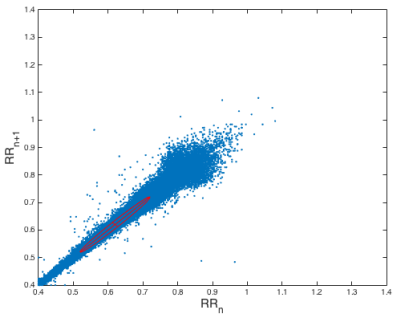
(b) Poincaré plot of the same patient with a shift $D = 5$



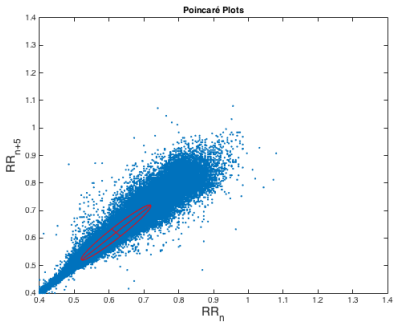
(c) Poincaré plot of the same patient with a shift $D = 20$

As we could expect, the length of the ellipse will remain exactly the same. Indeed, the range of values for the inter-beating intervals will not change. What happens is that the width will become larger and larger, letting the plot become more and more chaotic.

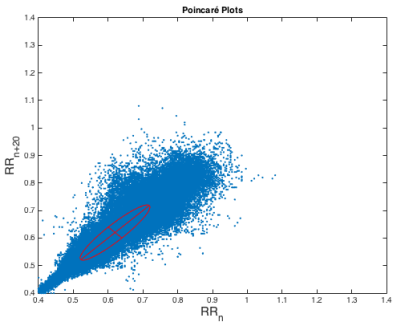
For a congestive heart failure case, the same obervations can be made :



(a) Initial Poincaré plot of a patient suffering from congestive heart failure with a shift $D = 1$
(successive interbeats considered)

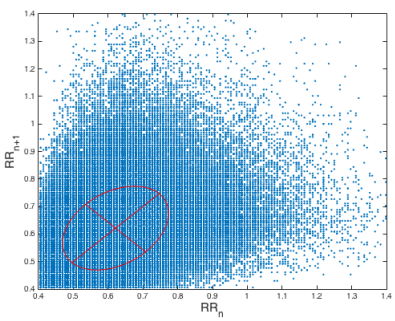


(b) Poincaré plot of the same patient with a shift $D = 5$

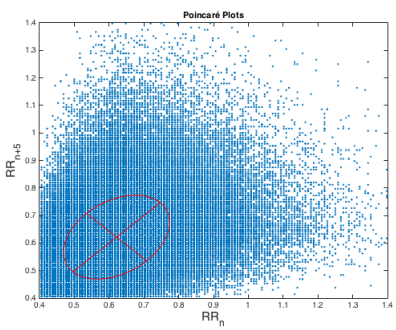


(c) Poincaré plot of the same patient with a shift $D = 20$

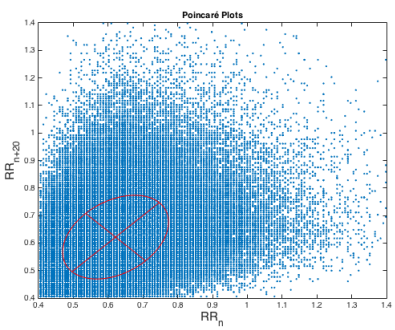
For an atrial fibrillation case, the evolution might look like this :



(a) Initial Poincaré plot of a patient suffering from atrial fibrillation with a shift $D = 1$
(successive interbeats considered)



(b) Poincaré plot of the same patient with a shift $D = 5$

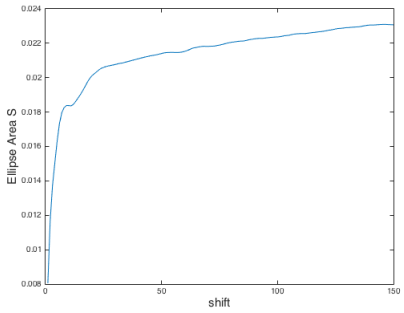


(c) Poincaré plot of the same patient with a shift $D = 20$

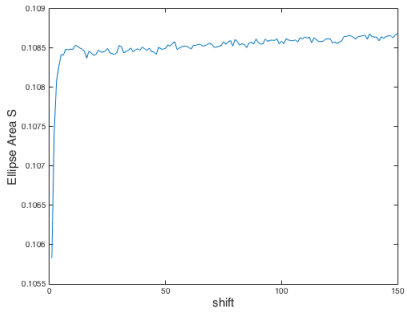
Here, we see that, globally, the ellipse area will increase, but the points being initially very spaced

on the plots, the area won't always increase that much when increasing by 1 the shifting parameter.

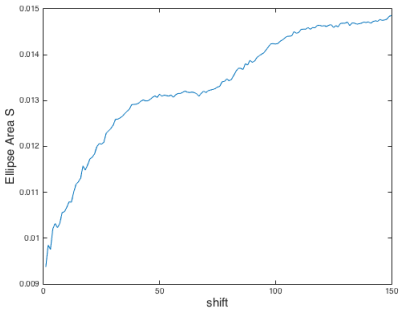
To conclude, by recording the ellipse area at each chosen shifting parameter over a range of 1 to 150 shifts, we can observe following plots :



(a) Evolution of the ellipse area S with the shifting parameter D for healthy case.



(b) Evolution of the ellipse area S with the shifting parameter D for atrial fibrillation case.



(c) Evolution of the ellipse area S with the shifting parameter D for congestive heart failure case.

We see that the evolution of the area is a bit different from one case to another. For healthy cases, the evolution will follow approximately a logistic growth. For the atrial fibrillation cases too, even though the growth is oscillating while it never oscillates for any the healthy cases. Finally, the evolution for the congestive heart failure cases is more complex. It looks like a logistic growth with two different 'carrying capacities'.

Chapter 3

Power-law distributed approximations

Power-law distributions are quite often used in several fields of scientific interest and can sometimes bring a totally new understanding of natural and man-made phenomena. For example, the intensities of earthquakes or sizes of power outages are thought to follow power-law distributions. Those distributions attracted special attention over the years for the mathematical properties they contain, such as their scale invariance.

In 2007, Aaron Clauset et al. [9] proposed a method to create a power-law model from empirical data, based on maximum-likelihood fitting methods. In this chapter, we will use again the `PhysioNet` database [8]. We will see more in details how the models are constructed, as well as how well they will fit or not the ECG recordings of the patients, and finally if we can find an interesting difference between healthy patients, patients suffering from atrial fibrillation and from congestive heart failure.

3.1 Definitions

Mathematically, we say a quantity x obeys a power-law if its probability distribution is given by :

$$p(x) \propto x^{-\alpha}$$

However, in practice, few empirical data from some phenomena follow a precise power-law for all values of x. To this end, the method proposed by Clauset et al. only applies the approximation to the values that are greater than or equal to a minimum x_{min}, chosen by testing several possible values and selecting the one that will minimize some distance we will chose. We'll say in that case that the tail of the distribution obeys a power law. In other words, the method here presented will aim at finding the parameters α and x_{min} that make the best power-law model approximation for the given patient's data.

Provided that $\alpha > 1$, and knowing that $P(X \geq x_{min}) = 1$ (the probability of any x to be greater than x_{min} is by definition 1), we get the normalized probability function :

$$p(x) = \frac{\alpha - 1}{x_{min}} \Big(\frac{x}{x_{min}} \Big)^{-\alpha}$$

One noteworthy property of this power-law probability function is that it is scale-invariant, i.e. scaling the argument x by a factor c gives :

$$p(cx) = \frac{(\alpha - 1)\ c^{-\alpha}}{x_{min}} \Big(\frac{x}{x_{min}} \Big)^{-\alpha} = c^{-\alpha}\ p(x)$$

which is simply multiplying the original power-law relation by a constant $c^{-\alpha}$

3.2 Estimation of the exponent factor α

Let us first introduce how the method computes the parameter α, which uses the other parameter x_{min}, that we initially will consider as fixed. Taking the dataset containing n observations $x_i \geq x_{min}$, the α to be found is the one that is most likely to generate the data. The probability that this data would have been generated from the model is :

$$p(x|\alpha) = \prod_{i=1}^{n} \frac{\alpha - 1}{x_{min}} \Big(\frac{x_i}{x_{min}} \Big)^{-\alpha}$$

This is the likelihood of the given data from the model. The most likely data to be generated with this model, goes along with the α that maximizes this latest function, or the logarithm of this function :

$$\begin{aligned} ln\ p(x|\alpha) &= ln \prod_{i=1}^{n} \frac{\alpha - 1}{x_{min}} \Big(\frac{x_i}{x_{min}} \Big)^{-\alpha} \\ &= \sum_{i=1}^{n} ln(\alpha - 1) - ln(x_{min}) - \alpha\ ln\Big(\frac{x_i}{x_{min}} \Big) \\ &= n\ ln(\alpha - 1) - n\ ln(x_{min}) - \alpha \sum_{i=1}^{n} ln\Big(\frac{x_i}{x_{min}} \Big) \end{aligned}$$

deriving this last equation with respect to α and setting it to zero, in order to get the maximizer, we get :

$$(ln\ p(x|\alpha))' = 0 = \frac{n}{\alpha - 1} - \sum_{i=1}^{n} ln\Big(\frac{x_i}{x_{min}} \Big)$$

which finally produces our MLE (maximum likelihood estimate) :

$$\hat{\alpha} = 1 + n \Big[\sum_{i=1}^{n} ln\Big(\frac{x_i}{x_{min}} \Big) \Big]^{-1}$$

Finally, from [9], we know that this estimator is consistent.

3.3 Choosing the second parameter x_{min}

The idea behind the method of choosing x_{min} is to test for all possible values of x_{min} (i.e. all unique values of x) which one produces the best fitting model. The fitting criterion that we will use is the Kolmogorov-Smirnov or KS test, i.e. the maximum distance between the cumulative distribution functions (CDF) of the fitted model and the data :

$$D = max_{x \geq x_{min}} |P(x) - S(x)|$$

where $S(x)$ is the CDF of the data.

When testing all possibles values for x_{min}, all the computed distances $D(x_{min})$ are saved and the value for x_{min} producing the best distance (i.e. the minimal one) is finally considered as the best x_{min} parameter for the model. We now have both parameters necessary for our approximation model.

3.4 Results and goodness-of-fit for the patients

The objective here will be to observe results for the three categories of patients (healthy, atrial fibrillation and congestive heart failure) and see if there is any difference between them in terms of scaling parameter α or goodness-of-fit.

This goodness-of-fit can actually be defined in many ways, but here the Mean Squared Error (MSE) has been chosen as goodness criterion :

$$\begin{aligned} MSE &= \frac{1}{N}\sum_{i=0}^{N-1}(P(z_i) - S(z_i))^2 \\ &= \frac{1}{N}\sum_{i=0}^{N-1}\left[\left(\frac{z_i}{x_{min}}\right)^{1-\alpha} - \frac{N-i}{N}\right]^2 \\ \text{where} \qquad z &= (x \geq x_{min}) \end{aligned}$$

and x is the vector of recorded RR intervals of a patient. N is the length of this vector z, containing all elements of x that are greater than or equal to x_{min}.

Visually, following figures were obtained by plotting both approximated and experimental cumulative distribution functions on a logarithmic plot :

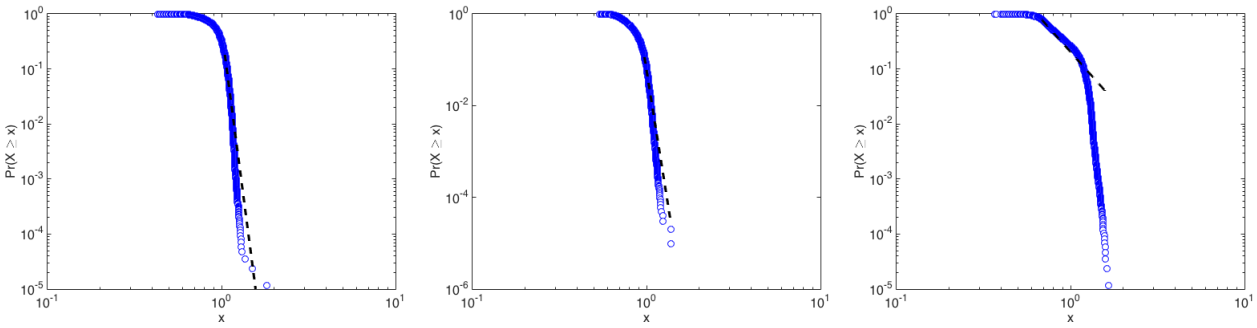


Figure 3.1: Three of the five healthy cases in the dataset. The x-axis represents the values of the RR interval duration (in seconds), while the y-axis represents the probability for an RR interval to be greater than x.

The approximated model is depicted by the black lines, while the blue bubbles represent the experimental values of the patient. Let us now see what the plots look like for the congestive heart failure cases :

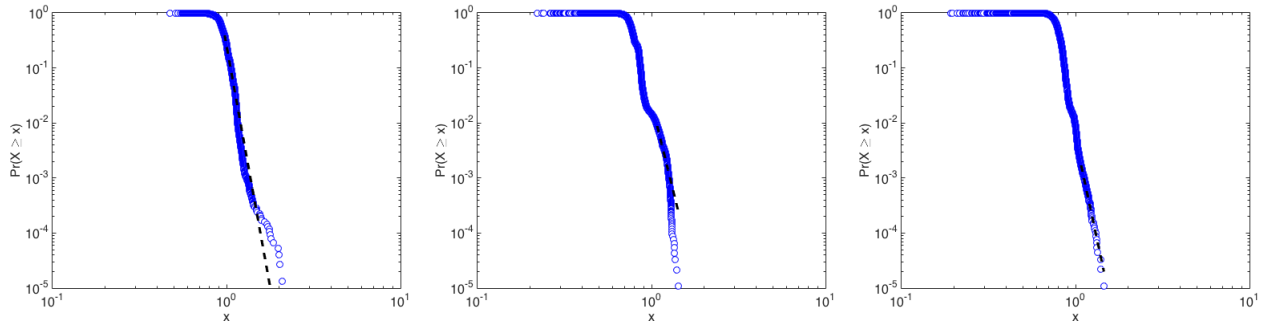


Figure 3.2: Three of the five congestive heart failure cases in the dataset. The x-axis represents the values of the RR interval duration (in seconds), while the y-axis represents the probability for an RR interval to be greater than x.

Purely visually, it seems that congestive heart failure cases are slightly better approximated, but we will see more in details how they compare by computing the Mean Squared Error, as well as the scaling parameter α. Finally, the plots obtained for the atrial fibrillation cases are the following :

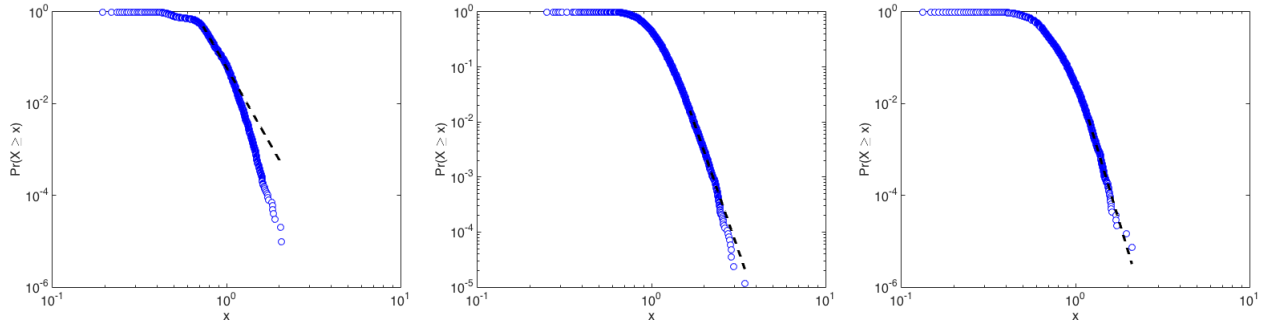


Figure 3.3: Three of the five atrial fibrillation cases in the dataset. The x-axis represents the values of the RR interval duration (in seconds), while the y-axis represents the probability for an RR interval to be greater than x.

All the plots, along with the mathematical results, can be found in the appendix.
Let us now observe the results, beginning with the scaling parameter :

Patient number	Healthy cases	CHF cases	AF cases
1	25.7511	18.4980	7.6533
2	25.1216	10.2667	10.3830
3	24.8008	13.8127	10.1268
4	4.5462	15.7783	13.6937
5	24.2342	15.3075	6.6308

On this set of fifteen patients, it's hard to make any conclusion regarding the scaling factor α returned by Clauset's method. Four of the healthy cases seem to have very similar values, around $\bar{\alpha} = 25.0$, while congestive heart failure cases take values around a mean 14.7. and atrial fibrillation cases have a mean value of $\bar{\alpha} = 9.7$.

Regarding the Mean Square Error, following results are obtained for the respective patients :

Patient number	Healthy cases	CHF cases	AF cases
1	9.1856×10^{-4}	4.4746×10^{-4}	3.8179×10^{-4}
2	26.5326×10^{-4}	7.9855×10^{-4}	9.3379×10^{-4}
3	25.4520×10^{-4}	22.2006×10^{-4}	1.2524×10^{-4}
4	28.7250×10^{-4}	4.5012×10^{-4}	3.2396×10^{-4}
5	26.7117×10^{-4}	12.7609×10^{-4}	8.8512×10^{-4}

Ideally, we would have expected a smaller error on the congestive heart failure, as many papers have already shown that a small deviation in the RR time series represent a risk factor, which would mean they could maybe be better approximated than RR times series of healthy patients. Unfortunately, we cannot really assert that it is the case here, as we would require more patients for each group in order to take any consistent conclusion. Nevertheless, we can observe that the average error is $\bar{MSE} = 23.3214 \times 10^{-4}$ for the healthy group, while the average is $\bar{MSE} = 10.3846 \times 10^{-4}$ for the congestive heart failure cases, which is indeed smaller. The best approximated group is the atrial fibrillation cases, with an average $\bar{MSE} = 5.2998 \times 10^{-4}$ error.

Finally, we can observe following values for x_{min}, i.e. the minimum value for the RR times series, still for the same respective patients :

Patient number	Healthy cases	CHF cases	AF cases
1	1.3120	0.9760	0.7340
2	1.0390	1.0200	0.8520
3	0.9610	1.0680	1.6640
4	0.6720	1.0760	1.1720
5	0.9380	0.7169	0.6090

Once again, no global conclusion can be made on the minimal x_{min} value, as the mean of each of the three groups are quite comparable. We can conclude this section by saying that the small number of patients considered for each group doesn't make it easy to make any conclusion, especially when noticing that the power-law approximations proposed by Clauset's method didn't produce remarkably different results from one group to another.

Chapter 4

Random Forests

There are two types of supervised learning : regression, where the target output is some (set of) real numbers, and classification, where the target output is a class label. This last type, classification, is the supervised learning we will explore in this chapter. In the medical field, when considering a patients database, classification methods can be a huge attraction, as it might sometimes provide a solid mathematical prognostic test.

As explained in the database description, we have collected valuable information about a group of patients having undergone a catheter ablation for atrial fibrillation at the hospital *Saint-Luc (UCL)* in Brussels. For some of them, we even have full information on recordings during the ablation procedure. We also know which of these patients have relapsed. With such a database, one inevitable question is : "how can we model a prognostic test for the patients, based on the information we already have ? In other words, can we tell if a new AF patient will relapse ? What error percentage will this model have ?". In this chapter we will present and implement a classification method, called Random Forests, that will try to respond to those questions.

4.1 Idea behind the algorithm.

A forest is an ensemble of trees. A decision tree is a popular method for various machine learning tasks. It is a flowchart-like structure in which each internal node represents a "test" on an attribute, each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes by browsing the tree). The paths from root to leaf represent classification rules. [10] A classic famous example where decision tree is used is known as Play Tennis:

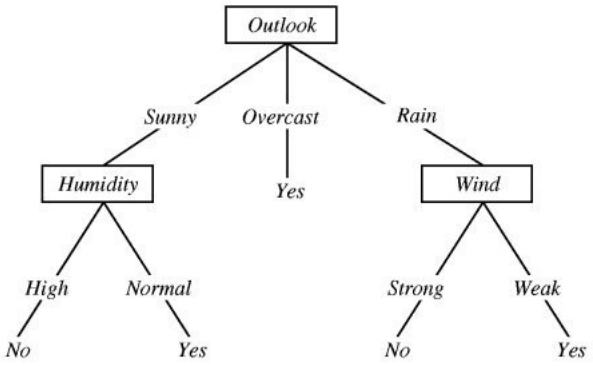


Figure 4.1: Famous example of a decision tree to decide whether or not play tennis.
(Source : https://nullpointerexception1.wordpress.com/2017/12/16/a-tutorial-to-understand-decision-tree-id3-learning-algorithm/)

The tricky part is to build the tree and decide which decision nodes (tests on the attributes) to build, in order to split the data. This is done by an optimization problem, working as follows :

$$\theta_j = argmax_{\theta \in \tau_i} \ I$$
$$I = I(S_j, \theta)$$
$$\text{where } I \text{ is the information gain : } \quad I(S,\theta) = H(S) - \sum_{i \in \{L,R\}} \frac{|S^i|}{|S|} H(S^i)$$
$$\text{and where H is Shannon's entropy : } \quad H(S) = -\sum_{c \in C} p(c) log(p(c))$$

Intuitively, the Shannon's entropy is a mathematical function which corresponds to the information quantity contained by an information source. This means that the nodes of a decision tree are chosen, based on the data, so that the split causes a maximum increased information gain.

Given the training set $X = \{x_1, ..., x_N\}$ of size $N = 68$ (the number of patients in our database) and the outcomes vector Y, representing an eventual relapse for the patients, we can appeal to bagging. Bagging (or Bootstrap aggregating) was proposed by Leo Breiman in 1994 to improve classification by combining classifications of randomly generated training sets. [11]

Bagging uses the training set X with the outcomes $Y = \{y_1, ..., y_N\}$, in order to repeatedly select a random sample of X, with replacement. For each of the selected random samples, we can fit a classification decision tree, creating a set of different trees. A first design choice concerns the size of the random samples we make. If this size equals the global training set size N, which is known as the bootstrap sample, then the randomly sampled set is expected to have a fraction 63.2% of the unique data of X. So in summary, the Random Forest's algorithm can be described as follows :

1. We have at our disposal a training set $X = \{x_1, ..., x_N\}$ of size N with outcomes $Y = y_1, ..., y_N$
2. For $t = 1, ...T$, sample N training examples from X, that we will call X_t, with respective samples Y_t.
3. Train a classification tree F_t on $\{X_t, Y_t\}$ for each $t = 1, ..., T$.
4. After training, we have a tree forest, and predictions for unseen samples X' can be made by a majority vote of the trees.

In order to reduce potential strong correlations between estimators, the final Random Forest's algorithm uses feature bagging, by training them on random samples of features instead of the entire feature set.
A second design choice concerns the parameter T, the number of trees to grow in the forest. Theoretically, increasing the number of trees can only improve the model, as it can only add a new classification tree to the vote. Of course, at some point, it is no longer necessary to grow new trees, as we would use redundant information. The final parameter T used in the models will be experimentally tested.

4.2 A first implementation on the whole dataset.

The features considered for this first classification model are the features contained in the repertory sheet, introduced in the data description. This sheet contains basic medical information on 68 patients. After discussion with Dr *Marchandise*, following criteria were used for the training set of our classification model, as they are thought of as influencing factors for atrial fibrillation :

(a) Gender

(b) Age

(c) Body Mass Index

(d) Diabetes (binary variable, patient having it or not)

(e) Tobacco

(f) High blood pressure

(g) Dyslipidemia

(h) Family history

(i) Congestive heart failure

(j) Stroke

(k) Vascular history

(l) Atrial Flutter

(m) Ischemic heart disease

(n) Respiratory pathology

(o) Left Ventricular ejection fraction

(p) Left Atrium size

(q) Converting Enzyme Inhibitor (CEI)

(r) Transient Ischemic Attack (TIA)

(s) Anti-hypertension diuretic

(t) Statins

This is a total of $N = 20$ features we have information about for each of the 68 patients. We can now start implementing the model, having a training set and outcomes for the patients (relapse or not after surgery).

The implementation on `MATLAB` of a Random Forest is quite implicit, as there exists a function *TreeBagger*, which creates a bag of decision trees. As explained in the documentation, *TreeBagger* grows the decision trees in the ensemble using bootstrap samples of the data. Also, *TreeBagger* selects a random subset of predictors to use at each decision split as in the Random Forest's algorithm.

In the arguments of this function, in addition to the training data set with the outcomes, we have to fix the number of trees we want to grow, and we can also use an a priori probability for the outcome classes. This is clearly a relevant information we can give to the model, as only a few percentage of the patients will relapse after catheter ablation. Finally, a useful argument we can also use is a cost associated to wrong predictions in a certain class of outcomes. For example, if we want to avoid making false negatives (wrong predictions stating that the patient won't relapse while he will relapse), we can set a higher cost associated to such mistakes. The higher the cost for a certain class, the more 'unbalanced' the predictions will be, causing a higher error percentage.

A first parameter whose value is to be seeked out is the number of trees in the forest to grow. Experimentally, we can observe the out-of-bag error at each possible number of trees. The out-of-bag (OOB) error of a Random Forest's model is a method of measuring the error in predictions. It is the mean prediction error on each training sample x_i, using only the trees that didn't use x_i in their bootstrap sample.

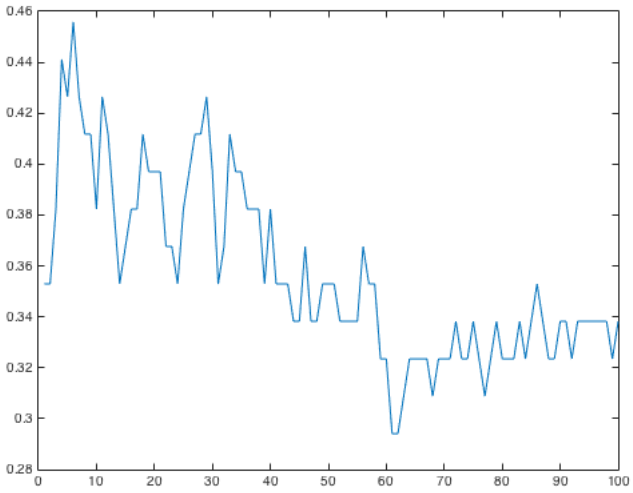


Figure 4.2: OOB error versus number of trees grown.

This figure is obtained by taking the mean on 100 different Random Forests created, for a fixed number of trees. We can see that, although a mean is taken over a hundred different samples, the OOB error is quite unstable from one value to another. Nevertheless, the global error percentage still decreases nicely as the number of trees increases, and this error converges from $T = 60$ on. For the rest of the tests that will be made, we will use a standard parameter of 60 trees to grow for a Random Forest.

Let us now take an example of decision tree grown in those Random Forests :

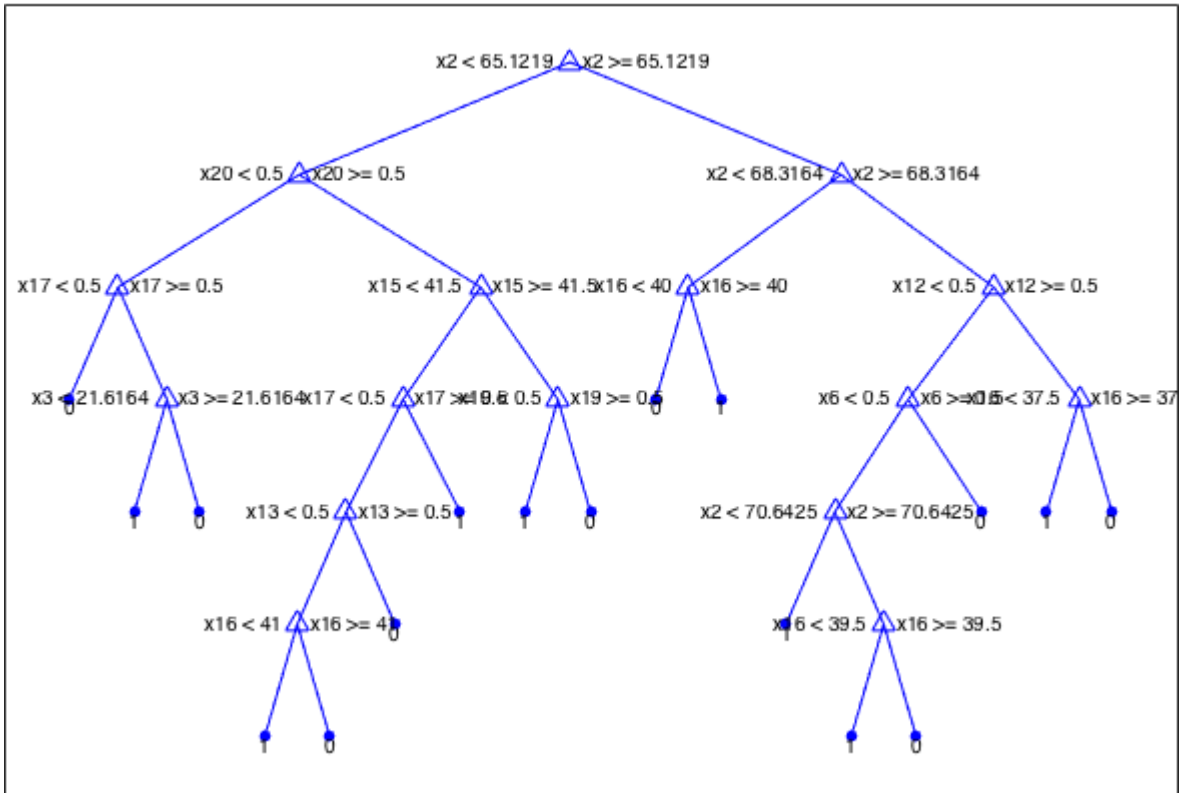


Figure 4.3: Sample of a decision tree grown in a Random Forest.

For example, if we want to predict if a new patient will relapse or not after surgery, this tree will vote for a non-relapsing if :

- The patient is younger than 65.1219 years old.
- The patient hasn't taken statins (medication).
- The patient hasn't taken CEI.

This corresponds to the extreme left path of the tree, and the leaf ending up at 0, the tree will therefore vote for 'no relapse' if those three attribute tests are satisfied.

A question that must now be asked is what fraction of the errors made are false positives and what fraction of errors are false negatives. We could also observe how those fractions change as we introduce a cost function as presented earlier, when we set a cost to a certain class of errors.

To study the number of false positives/negatives made by the model, we can create a Random Forest using the whole dataset except for one patient. Predicting the outcome for this patient, we can then observe if the prognosis is correct or which type of error is made.

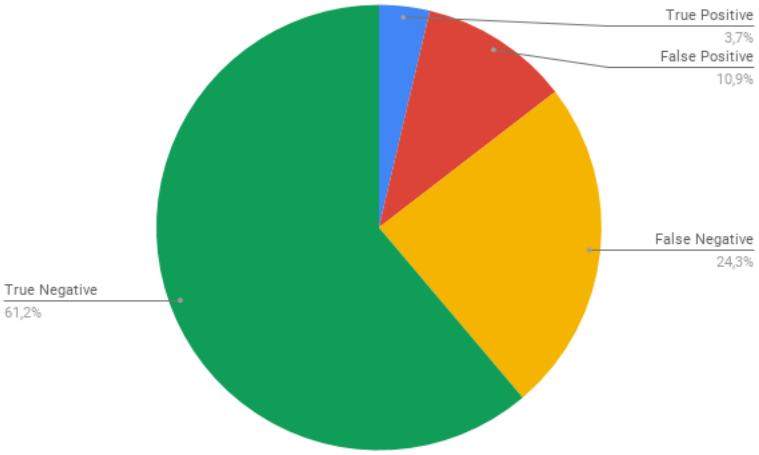


Figure 4.4: Statistical hypothesis testing.

From this chart, we can see that the model tends to predict a lot more negative outcomes ('no relapse'). This is due to the fact that the non-relapsing population is much more larger than the relapsing population. This prediction statistic can have dramatic consequences, as false negatives (patients who will relapse but are predicted not to) can endanger patients lives. For this purpose, we might want to set a cost function associated to this type of wrong predictions. Following table shows the evolution of the prediction error types with the cost function :

Cost associated to a false negative	True negatives	False negatives	True positives	False positives	Error percentage of the samples
1	40	16	3	9	0.3676
2	35	15	4	14	0.4265
3	32	15	4	17	0.3823
4	28	12	7	21	0.4853
5	27	10	9	22	0.4706
6	25	10	9	24	0.5000
7	23	11	8	26	0.5441
8	20	9	10	29	0.5588
9	14	8	11	35	0.6324
10	12	7	12	37	0.6470
11	14	9	10	35	0.6471
12	12	5	14	37	0.6176
13	11	8	11	38	0.6765
14	11	7	12	38	0.6618
15	7	5	14	42	0.6912
16	7	3	16	42	0.6618
17	9	3	16	40	0.6324
18	4	2	17	45	0.6912
19	4	4	15	45	0.7206
20	5	2	17	44	0.6765

The cost is defined as a matrix 2×2, where the element $\{i, j\}$ is the cost of predicting a patient of class i to the class j. The cost associated to false positives has been fixed to 1. Clearly, there is a trade-off to make between the false negatives we want to limit, and the error that we don't want to increase too much.

This concludes the first approach regarding Random Forests implemented on the repertory sheet containing information about the 68 patients. Now, we can wonder how we could improve this model by using additional information we have concerning the recordings of atrial rate and disconnection time for 54 of the 68 patients. Let us first look at which information we can extract from those recordings.

4.3 Recordings from patients and resulting boxplots

From the recordings of the 54 patients with complete PV disconnection times information, we can extract new parameters to add to the features in the database of our Random Forest's model. Let's see how we can, in this section, extract those new parameters.

The atrial rates recordings of the patients are realized during the catheter ablation they undergo. At the beginning of each recording *.txt* file, we have information concerning the sampling rate (which is 1000 Hz for each recording in the database) and the starting time of the recording. The measures are obtained by placing a dipole in the coronary sinus. As a result, we have a voltage value time series, sampled at the given constant frequency. In order to reduce noise and remove outliers in the signal, a smoothing filter is a valuable tool that can improve the signal's reliability. The Savitzky-Golay filter was used in order to smooth the signals.

Now that we have filtered the voltage signals, the goal is to obtain the atrial rate, which is done by detecting the peaks in the voltage time series. For that purpose, the `MATLAB` function *findpeaks* has been used, which finds the local maxima of the times series (i.e. the samples larger than their two neighbours). Additional arguments are used, such as a minimum peak height or a minimum peak prominence.

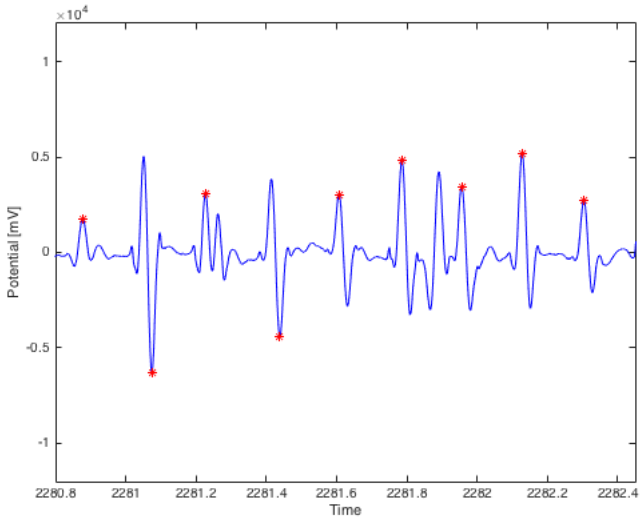


Figure 4.5: Filtered atrial rate signal and peak detection.

With all the peaks in the signal discovered, we have an atrial rate interval time series for each patient. Parts in the times series that we could be very interested in are those right after a pulmonary vein disconnection during the ablation procedure. Indeed, by observing the reaction of the atria to the ablation procedure after disconnection of each pulmonary vein (right superior, right inferior, left superior, left inferior), we could maybe see some difference between patients who will relapse and those who won't.

In order to respond to this latter question, following parameters will be extracted from the atrial rates :

- the maximum inter-peak duration in the 30 seconds following each pulmonary vein's disconnection : AA_{max}
- the total number of atrial beatings (peaks) during the 30 seconds following each pulmonary vein's disconnection : AR

This established a total of 8 parameters for each patient of the 54 patients in the database. The complete results of the computations are given by patient in the appendix. Visually, we can get an first idea on these parameters for both classes of patients (relapsing and non-relapsing class) by creating for each parameter boxplots showing the evolution from the first pulmonary vein's disconnection to the fourth one.

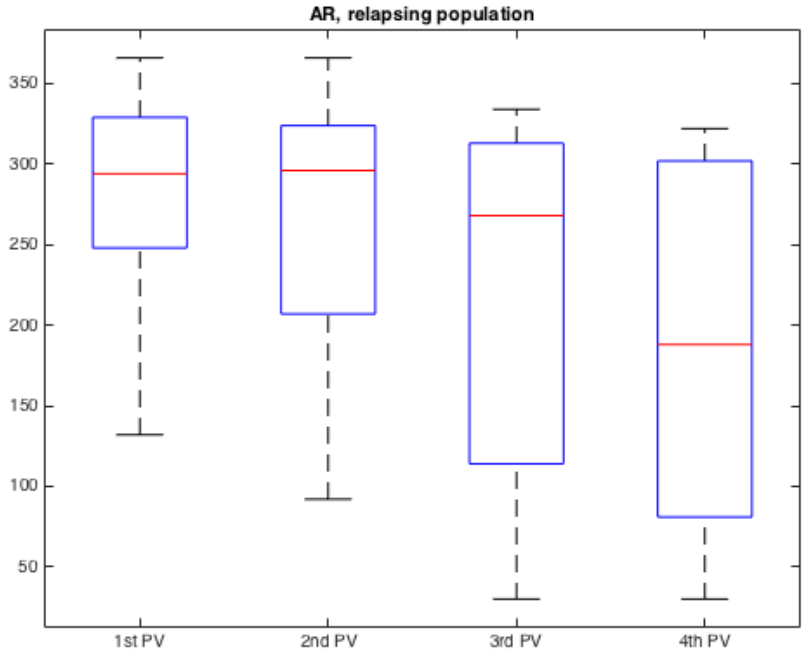


Figure 4.6: The average number of peaks per minute in the atrial rate's signal following the respective pulmonary vein's disconnections for the relapsing population (16 patients).

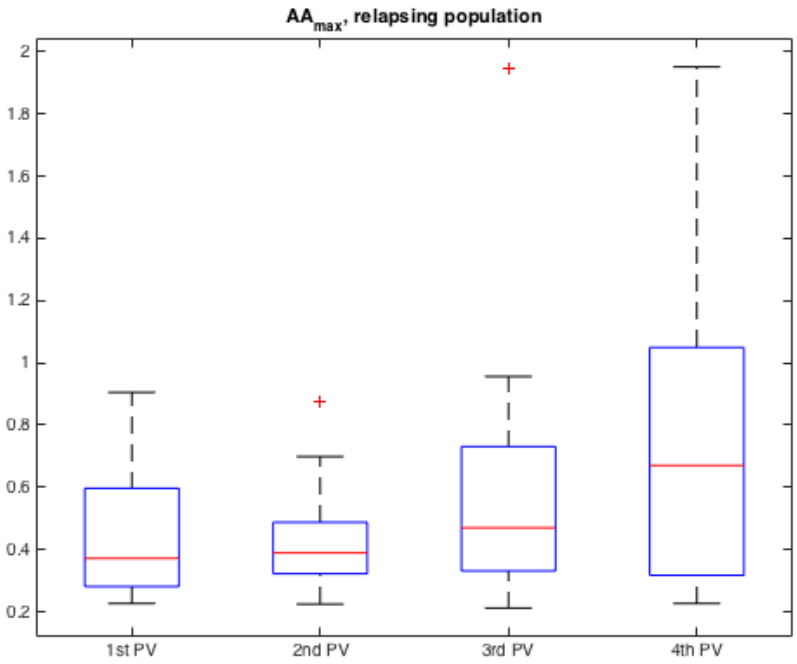


Figure 4.7: The maximum peak duration following the respective pulmonary vein's disconnections for the relapsing population (16 patients).

From those figure, we can deduct a global observation for the relapsing population : the average number of peaks in the atrial rate signal seem to decrease as the ablation procedure advances. This logically comes with an increase of the average peak duration. It is therefore not surprising to see the maximum peak duration increase too. Let's see how this decrease-increase pair compares with the non-relapsing population :

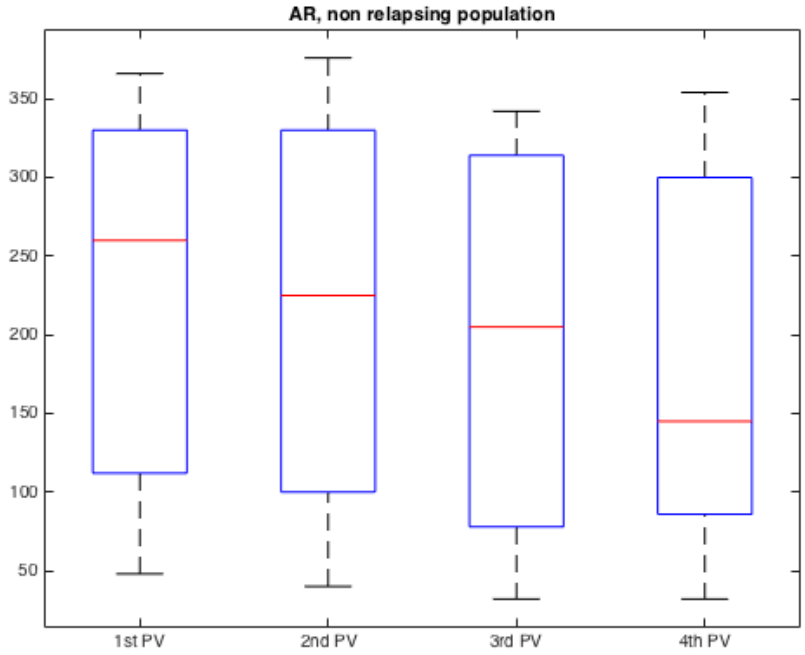


Figure 4.8: The average number of peaks per minute in the atrial rate's signal following the respective pulmonary vein's disconnections for the non-relapsing population (38 patients).

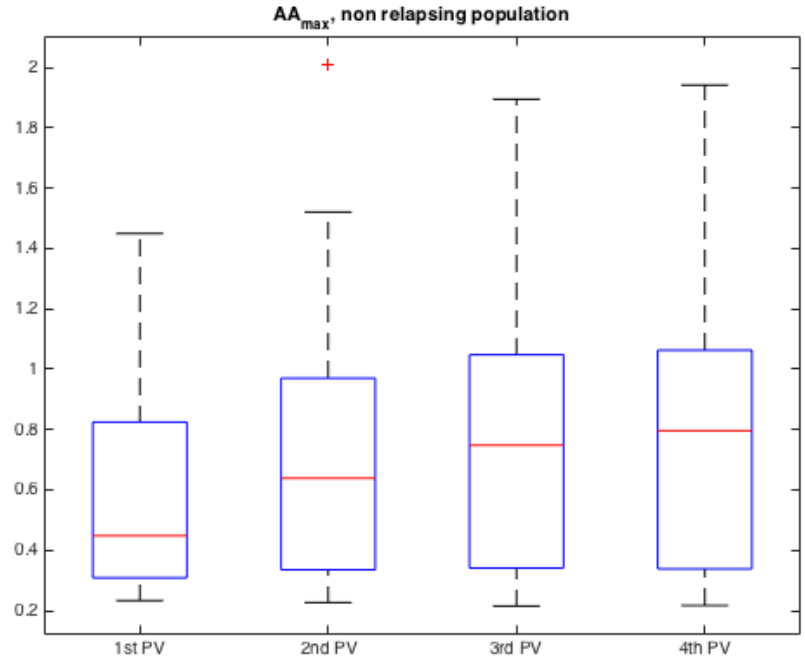


Figure 4.9: The maximum peak duration following the respective pulmonary vein's disconnections for the non-relapsing population (38 patients).

From the boxplots, there is no obvious difference to report between the relapsing to the non-relapsing population. The boxes have larger areas for the non-relapsing population, corresponding to first and third quartiles being more spaced. Nevertheless, this can easily be explained by the fact that the non-relapsing population is larger (more than two times the relapsing population), causing the range of values to be larger too. Let us observe the mean and variance for both populations for both the AA_{max} and AR parameters :

Population	PV1 mean AA_{max}	PV2 mean AA_{max}	PV3 mean AA_{max}	PV4 mean AA_{max}
relapsing	0.4386	0.4345	0.5822	0.7526
non-relapsing	0.6072	0.7132	0.7713	0.7899
	PV1 var AA_{max}	PV2 var AA_{max}	PV3 var AA_{max}	PV4 var AA_{max}
relapsing	0.0364	0.0281	0.1788	0.2168
non-relapsing	0.1295	0.1934	0.2176	0.2066

The values obtained in the table are given in *seconds*. From those results, we can see that the average maximum inter-beat duration is higher for the non-relapsing population, even considering the higher variance. Also, the increase of this maximum interval from one disconnection to another seem to happen more quickly for the non-relapsing population. Indeed, from $PV1$ to $PV2$, there is no increase observed for the relapsing population, while the increase is flagrant for the non-relapsing population.

Let us now consider the AR parameter, i.e. the average number of beats per minute following the respective PV disconnections :

Population	PV1 mean AR	PV2 mean AR	PV3 mean AR	PV4 mean AR
relapsing	276.62	264.25	226.75	190.50
non-relapsing	226.10	215.31	197.31	184.52
	PV1 var AR ($\times 10^4$)	PV2 var AR ($\times 10^4$)	PV3 var AR ($\times 10^4$)	PV4 var AR ($\times 10^4$)
relapsing	0.4536	0.7644	1.1155	1.2369
non-relapsing	1.2407	1.3300	1.3026	1.2610

Going along with the higher average maximum peak interval, we can see that the average number of peaks per minute is lower for the non-relapsing population. We also notice that the decrease from $PV1$ to $PV4$ for the relapsing population is more impressive than for the non-relapsing population. The variance is also lower, even if it increases from one disconnection to another, while it surprisingly decreases for the non-relapsing patients.

A question we should ask ourselves is whether there are a few patients strongly influencing the boxplot results. By picturing and linking the patients parameters together, we get following boxplots :

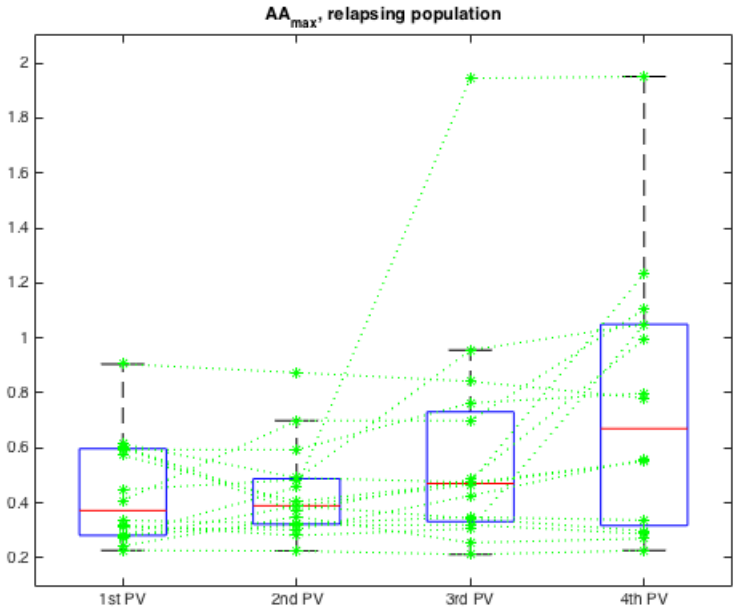


Figure 4.10: The maximum peak duration following the respective pulmonary vein's disconnections for the relapsing population (16 patients).

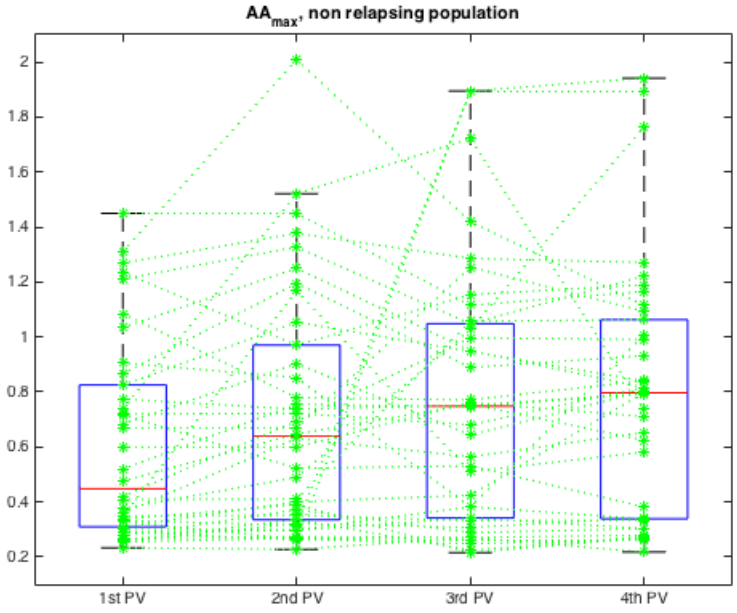


Figure 4.11: The maximum peak duration following the respective pulmonary vein's disconnections for the non-relapsing population (38 patients).

It is hard to make any decisive conclusion from those figures, but one reassuring observation is that the boxplots don't seem to be too much influenced by a few eventual outliers. Both populations contain a few patients having a maximum interval duration not really increasing from one disconnection time to another, but globally we can observe that the increase is present for most of the patients. The best way to see if the patients parameters collected here are really relevant, is to see how the model's prediction error percentage is improved or not when adding those last parameters to the database.

In the following section, the parameters AR and AA_{max} will be added to the training set used for the Random Forests, and we will see if this helps to improve our prediction model.

4.4 Final Random Forest's model using atrial rate information

The final model implementation completely compares with the initial model described here above. The number of patients has been reduced to patients for whom we have complete information about the atrial rate during the catheter ablation. Also, the two parameters at each of the four disconnection times for those patients have been added to the training dataset. Even though the number of patients decreases compared with our initial model, we still expect the OOB error to increase thanks to the eight added features in the dataset. Comparing the Random Forests growing from both datasets, side by side, we have :

training set	mean out-of-bag error on a total of 300 model samples	out-of-bag error of the best fitting model found
initial set of 68 patients	0.3581	0.27941
final set of 54 patients	0.31426	0.22222

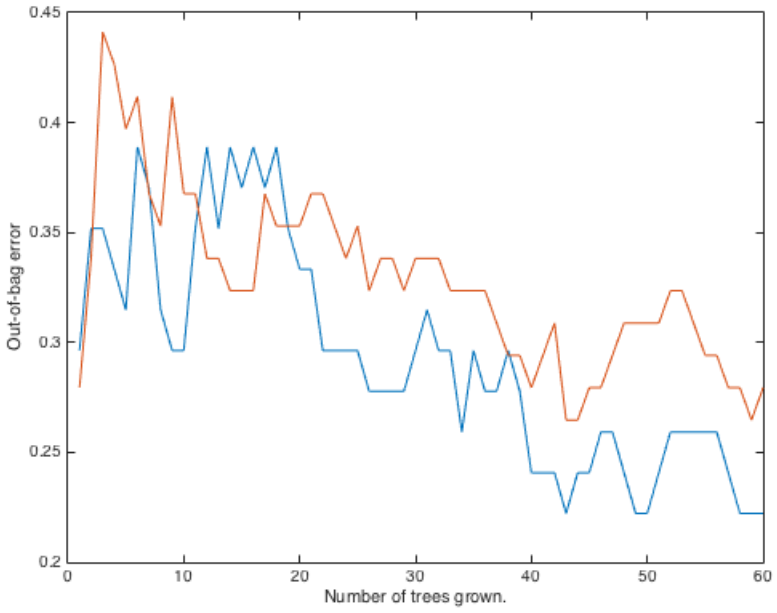


Figure 4.12: Out-of-bag error of the best fitting models for the final versus the initial dataset, in blue and orange respectively.

This result is quite satisfying, as it shows that the information obtained from the atrial rates of the patients makes a positive contribution to the prediction model. Once again, we can get statistical results by growing a Random Forest from the database excluding a patient, then predict this latter patient's outcome, and see how it differs from the true outcome. By doing so, we can compose following pie chart :

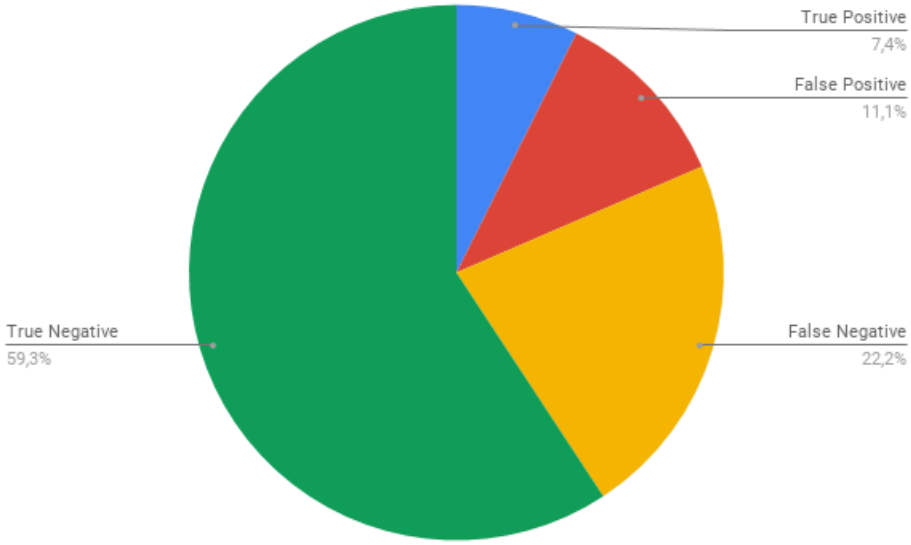


Figure 4.13: Pie chart depicting the number of false positives and negatives for the new Random Forest's model.

From this chart, we observe that, although the number of true negatives doesn't change much comparing with the initial model, the number of true positives increases considerably.

A few final words on the final dataset

It has to be added that information extracted from the dataset, consisting of the atrial rates of the patients, has some limitations that we must stay careful about.

First of all, from the recordings that can last up to four hours, only four small periods of 30 seconds are considered (the 30 seconds following each pulmonary vein disconnection). This small period duration has been chosen as it is in order to observe the immediate atrial rate's reaction to the catheter ablations. No longer period of time has been tested, as the information content in the 30 seconds periods seemed to have satisfying impact on the error reduction.

Secondly, the chosen parameters (AA_{max} and AR) alone don't completely describe the atrial rate of a patient following a pulmonary vein disconnection. Filtering the signal before analyzing it allows to remove possible outliers, but still the maximum interval between two peaks in a 30 seconds period can vary a lot from one patient to another, whatever the ultimate outcome might be. External factors as breathing and stress, among others, can explain this variation.

Finally, the disconnection time noted for each pulmonary vein is not always known with an extreme accuracy. As those disconnection times are manually written during the procedure, there might be some small delay between the actual and written disconnection time. Moreover, for some of the patients in the database, no disconnection time could be found, so that the transition time between two pulmonary veins was considered as the disconnection time.

Conclusion

This thesis work is composed of two main parts. The first part consists in the analysis of two methods, Poincaré plots and power-law approximation, in order to distinguish patients classified in three different groups : healthy cases, atrial fibrillation cases and congestive heart failure cases. For each of those patients, the time series of the inter-heartbeat duration over a 24-hour period were build up. Poincaré plots are a very visual tool plotting successive inter-beat durations side by side. By fitting an ellipse on the plots, we could clearly observe how cases from the different groups differed. The power-law approximations made on those same time series aim to approximate the probability distribution for the inter-beat durations. The analysis of the results for this method implementation was harder to make as no group of patients was better fitted by the approximations, or with a more characteristic and common exponent factor. The fact that only fifteen patients composed the database didn't help either to detect any common characteristic within a group of patients.

In the second part of the project, a totally new database was used, which is composed of data collected at the hospital *Saint-Luc (UCL)*, in Brussels. Here, all the patients considered underwent a catheter ablation procedure for atrial fibrillation. Some of them had a relapse after this procedure. Based on medical information we had on these patients, such as gender or age, as well as atrial rate recordings during their ablation procedure, a learning method, Random Forests, was used to make a classification. By implementing this method, the goal was to make a prediction on an possible relapse for patients. At the end, an error percentage of 22.22 % could be achieved on the predictions. Given some inevitable inaccuracies due to exterior factors commanding the heart-beat, like breathing or stress, among others, and due to the fact that the state of the heart after surgery is not foreseeable, this error percentage doesn't seem excessive.

An interesting complement for this work could be a rigorous analysis of the statistical dependence between the group of parameters computed from the atrial rates of patients in the database collected at the hospital *(UCL) Saint-Luc*. This could statistically support the consistency of the data used in the models. An *ANOVA* test for example, is a way to find out if experiment results are significant. Also, a new learning method could be tested on the same patients data, and may be combined with the Random Forest model in place in order to further decrease the prediction error percentage.

Appendices

1. Abbreviations

HRV	heart rate variability
ECG	electrocardiogram (or electrocardiography)
UCL	Université Catholique de Louvain
AV node	atrio-ventricular node
CHF	congestive heart failure
AF	atrial fibrillation
SD	standard deviation
var	variance
MLE	maximum likelihood estimation (or maximum likelihood estimator)
CDF	cumulative distribution function
MSE	mean squared error
OOB error	out-of-bag error
PV	pulmonary vein

2. *MATLAB* code inventory

2.1. Codes for the Poincaré plots

name of *.m* file	arguments (if any)	result(s)
read_txt	*.txt* file representing the inter-beat durations recorded on a patient	creates an RR time series
poincare_plots_ellipse	shift between successive points to be considered	based on the time series, computes the $SD1$,$SD2$ and S parameters of the fitted ellipse
compute_param		browses all the patients files and computes the parameters by calling poincare_plots_ellipse
draw_ellipse	RR times series and SD parameters	creates a figure showing the Poincaré plot of a patient with the fitted ellipse
test_shift_param	patient's *.txt* file	observes the evolution of the plot and ellipse area when the shifting parameter increases

2.2. Codes for the power-law approximations

name of *.m* file	arguments (if any)	result(s)
fit_power_law	patient's RR time series	computes α and x_{min} for the power low approximation on the RR interval duration distribution
goodness_of_fit	patient's RR time series and fitting parameters α and x_{min}	mean squared error made by the approximation
plplot	patient's RR time series and fitting parameters α and x_{min}	plots on log axes the the data contained in x and the power-law distribution
run_fit_all		runs the approximations and their respective error

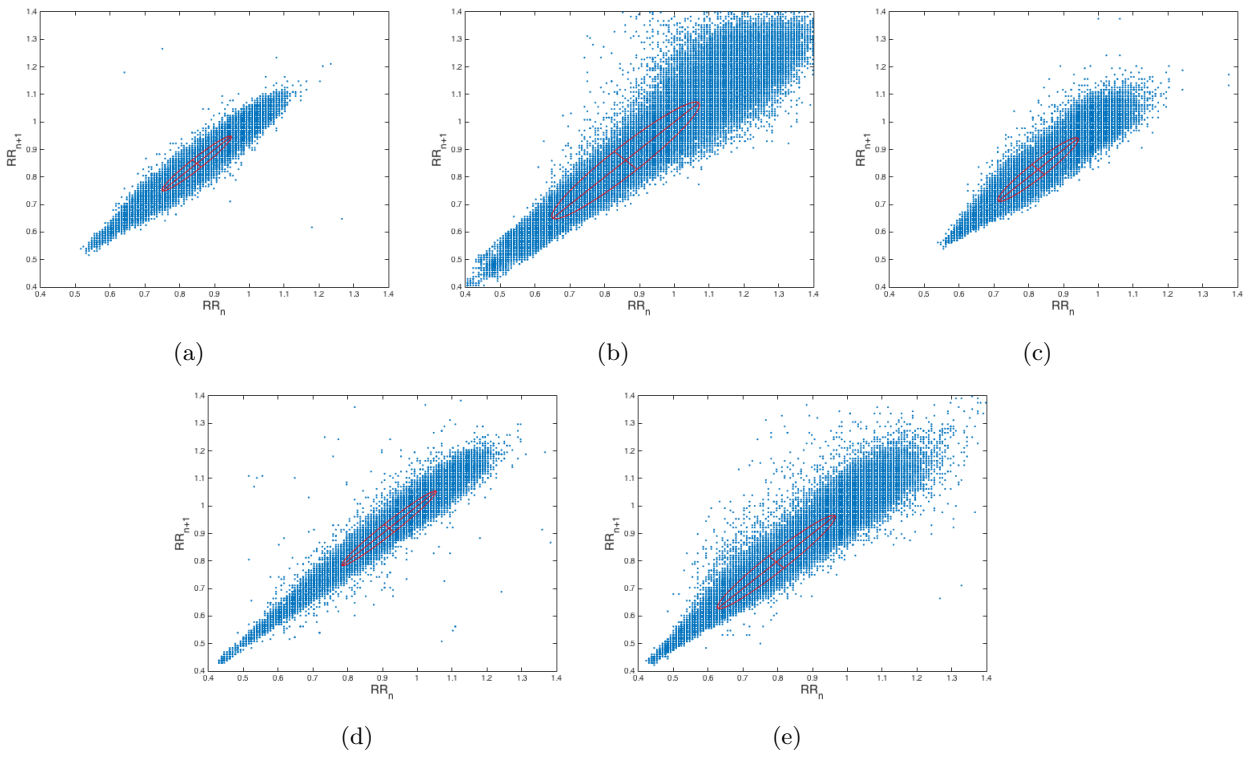
2.3. Codes for the Random Forests

name of *.m* file	arguments (if any)	result(s)
load_data_rep	path to the repertory sheet	creates the training data set along with outcome vector
filter_data_atrial	sheet containing computed AA_{max} and AR parameters	adds to training set atrial rate's information and removes patients with missing information
mean_best_error		realizes Random Forests on the training set and computes the mean and best OOB error
statistical_test		computes separate Random Forests to count the number of false positives/negatives
numtrees_test		tests different values for the number of trees and observes the OOB error evolution
cost_vs_error		tests different values for the cost function and observes the OOB error evolution

3. Poincaré plots results

As explained in the chapter on the Poincaré plots, fifteen patients ECG recordings were used as data for the Poincaré plot analysis : 5 healthy patients, 5 suffering from atrial fibrillation and 5 suffering from congestive heart failure. All the results (the plots along with the parameters considered during the analysis) are displayed here.

First of all, let us consider the normal cases :



(a) (b) (c)

(d) (e)

Figure 4.14: Poincaré plots of the ECG recordings for all the healthy patients.

We see that, apart for the second plot (b), which has a higher length (long-term variability), each ellipse computed is more or less similar. The parameters for each of those graphs respectively are :

Case	$SD1$ (in ms)	$SD2$ (in ms)	Ellipse area S
a	18.3126	140.0155	0.0081
b	46.2926	296.6494	0.0431
c	25.5800	161.9540	0.0130
d	19.7157	191.1383	0.0118
e	31.1171	238.3101	0.0233

Let us now consider the congestive heart failure cases and see how they differ from the normal cases :

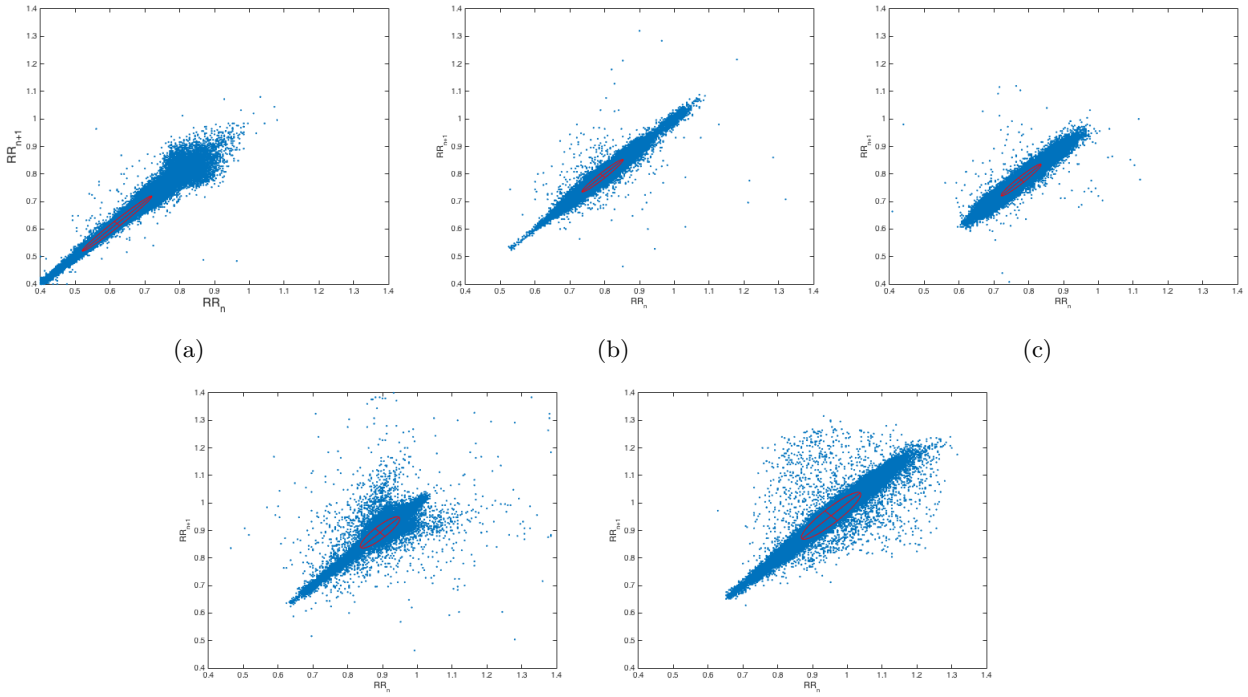


Figure 4.15: Poincaré plots of the ECG recordings for all the patients suffering from congestive heart failure.

with the respective parameters :

Case	$SD1$ (in ms)	$SD2$ (in ms)	Ellipse area S
a	10.8403	141.6140	0.0049
b	9.9656	83.4300	0.0026
c	10.7261	81.2955	0.0027
d	20.2255	77.8668	0.0049
e	25.2302	118.2687	0.0094

In this case, we see globally that the ellipse area is indeed reduced and is of order 10^{-3} instead of 10^{-2} for the healthy cases, due to a smaller width of the ellipse and a higher density of points along the line of identity.

Finally, let us observe the results for all the atrial fibrillation cases :

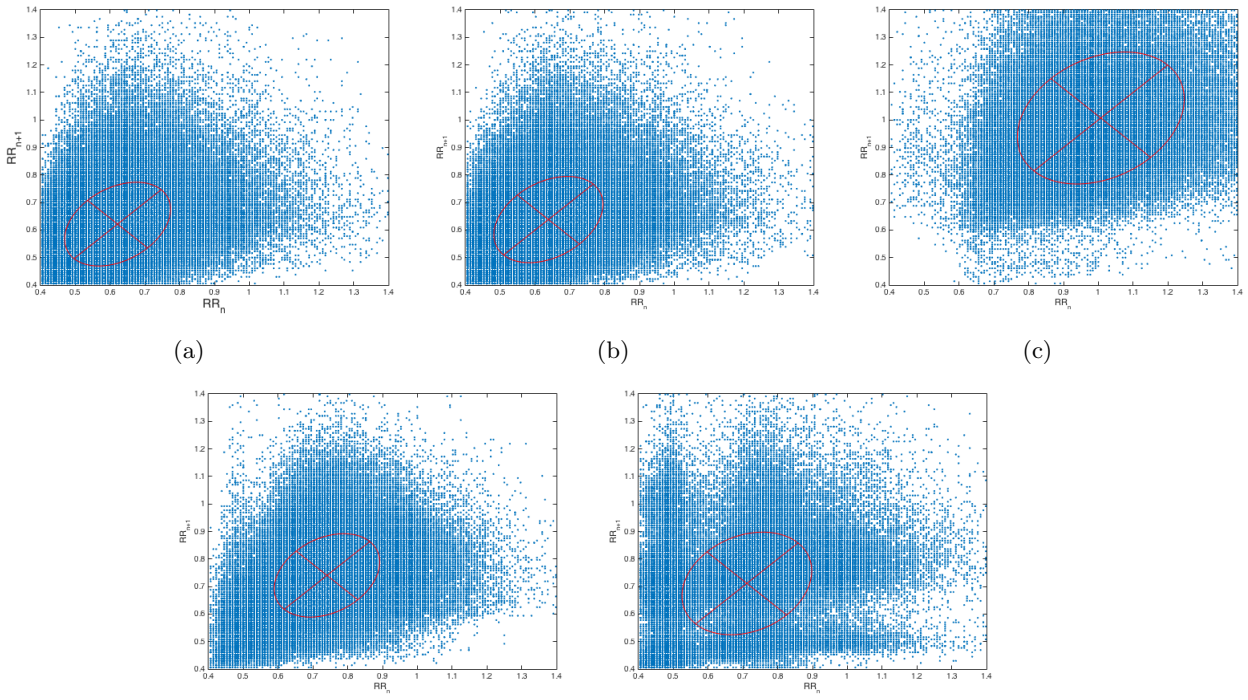


Figure 4.16: Poincaré plots of the ECG recordings for all the patients suffering from atrial fibrillation.

The observation is here very visual. Some chaos is present in the plots, representing the irregular heartbeat occurring when atrial fibrillation is at stake. The parameters computed along with the plots are respectively :

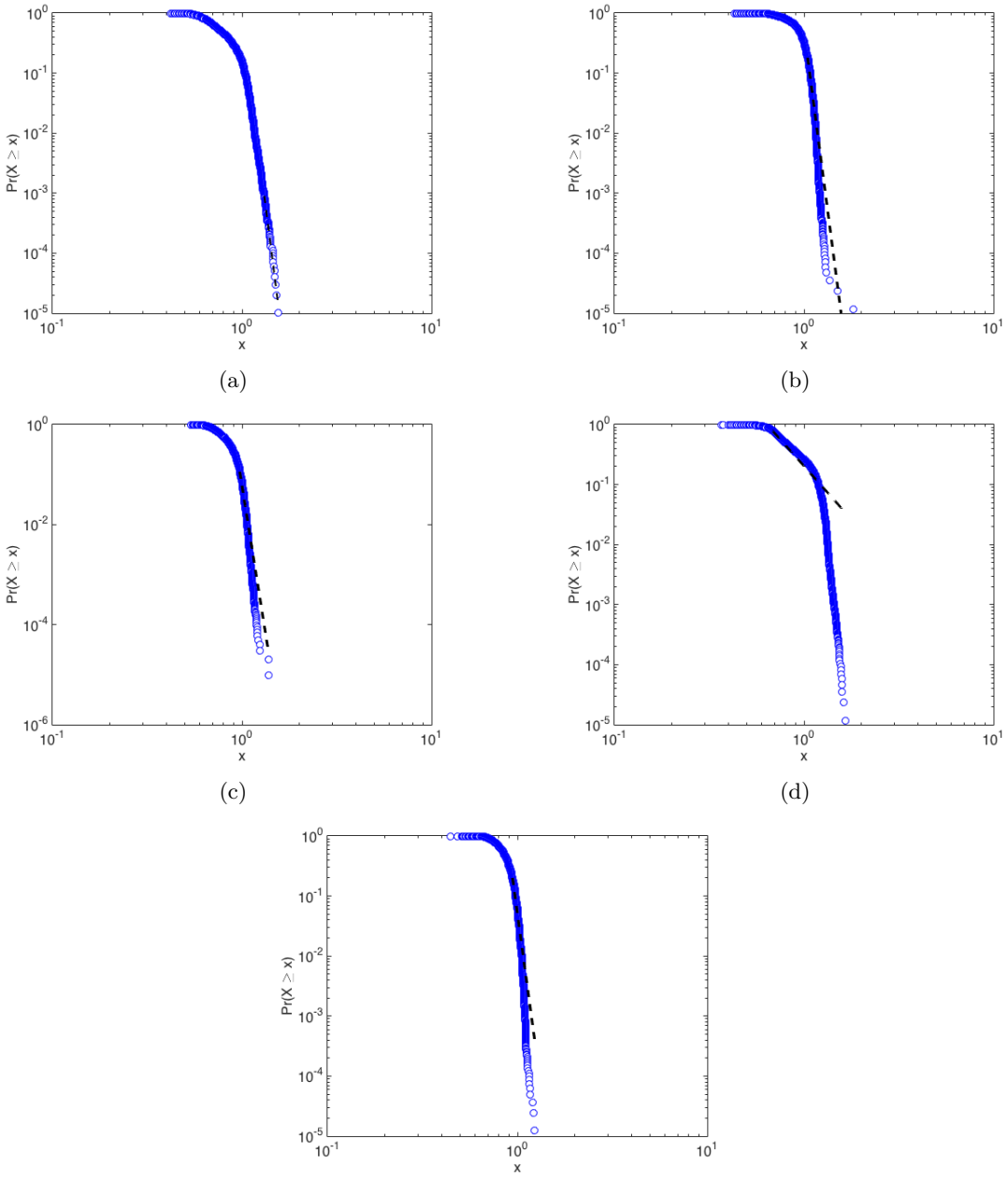
Case	$SD1$ (in ms)	$SD2$ (in ms)	Ellipse area S
a	122.0510	177.0883	0.0669
b	126.0803	182.1317	0.0721
c	202.5362	271.8955	0.1730
d	125.0066	173.9390	0.0683
e	161.9367	208.0224	0.1058

The most eye-catching factor here is the width of the ellipse, $SD1$, which is a lot larger than for the other cases. This results in a greater ellipse area too.

4. Power-law approximation results

As explained in the chapter on the power-law approximations, fifteen patients ECG recordings were used as data for the Poincaré plot analysis : 5 healthy patients, 5 suffering from atrial fibrillation and 5 suffering from congestive heart failure. All the results (the plots along with the parameters considered during the analysis) are displayed here.

Let us start by the five healthy patients :



(a) (b) (c) (d) (e)

Figure 4.17: Logarithmic plots of the cumulative distribution functions (approximated and experimental) for the healthy cases.

The respective parameters obtained in the approximated model are :

Case	Scaling parameter α	x_{min} (in seconds)	Mean Squared Error ($\times 10^{-4}$)
a	25.7511	1.3120	9.1856
b	25.1216	1.0390	26.5326
c	24.8008	0.9610	25.4520
d	4.5662	0.6729	28.7250
e	24.2342	0.9380	26.7117
mean	20.8908	0.9844	23.3214

The plots for the congestive heart failure cases are given in Figure 4.18. The respective parameters in the approximation models are now :

Case	Scaling parameter α	x_{min} (in seconds)	Mean Squared Error ($\times 10^{-4}$)
a	18.4980	0.9760	4.4746
b	10.2667	1.0200	7.9855
c	13.8127	1.0680	22.2006
d	15.7783	1.0760	4.5012
e	15.3075	0.7160	12.7609
mean	14.7327	0.9712	10.3846

Finally, the plots for the atrial fibrillation cases are given in Figure 4.19 and the parameters for the respective patients are :

Case	Scaling parameter α	x_{min} (in seconds)	Mean Squared Error ($\times 10^{-4}$)
a	7.6533	0.7340	3.8179
b	10.3830	0.8520	9.3379
c	10.1268	1.6640	1.2524
d	13.6937	1.1720	3.2396
e	6.6308	0.6090	8.8512
mean	9.6975	1.0062	5.2998

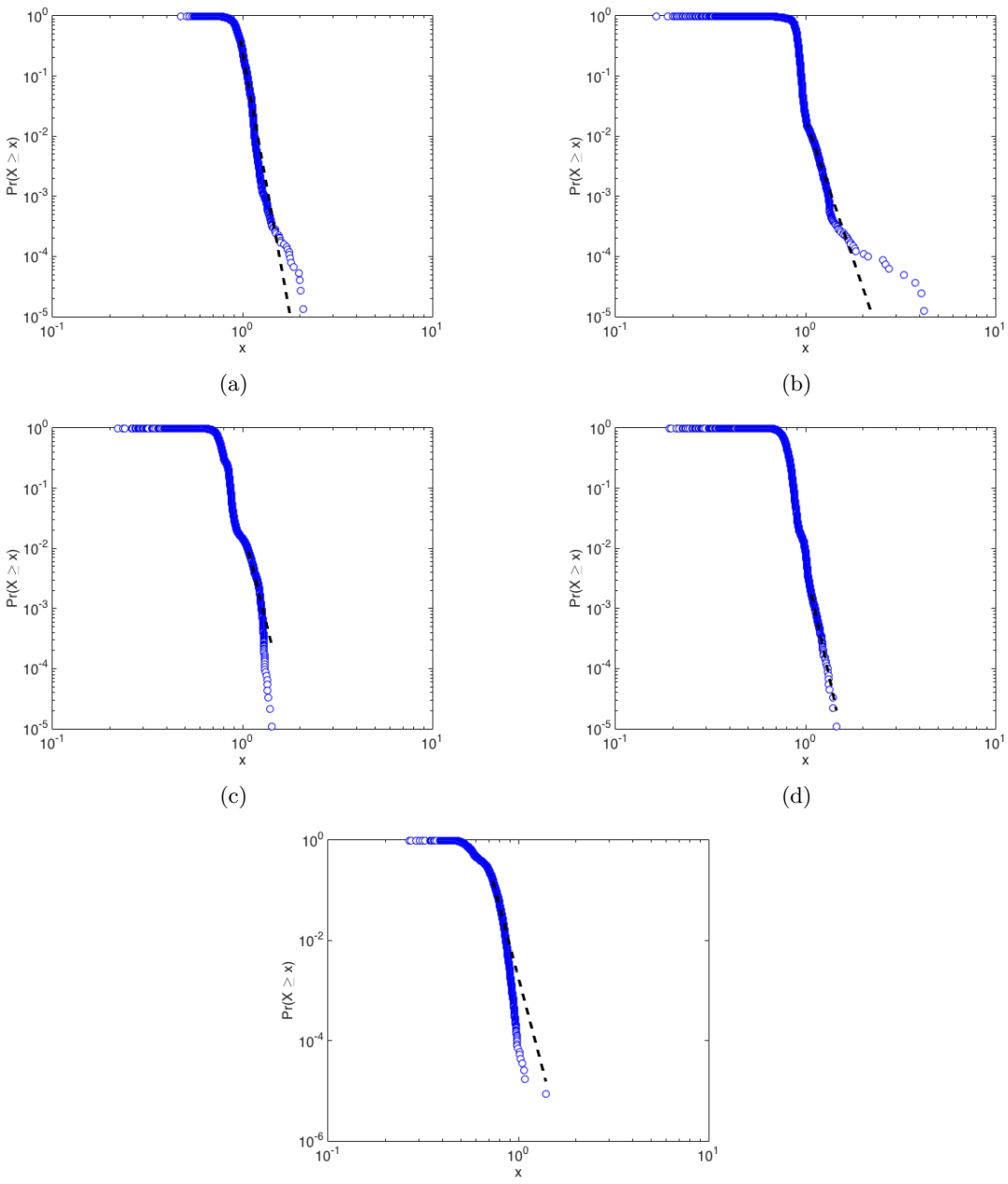


Figure 4.18: Logarithmic plots of the cumulative distribution functions (approximated and experimental) for the congestive heart failure cases.

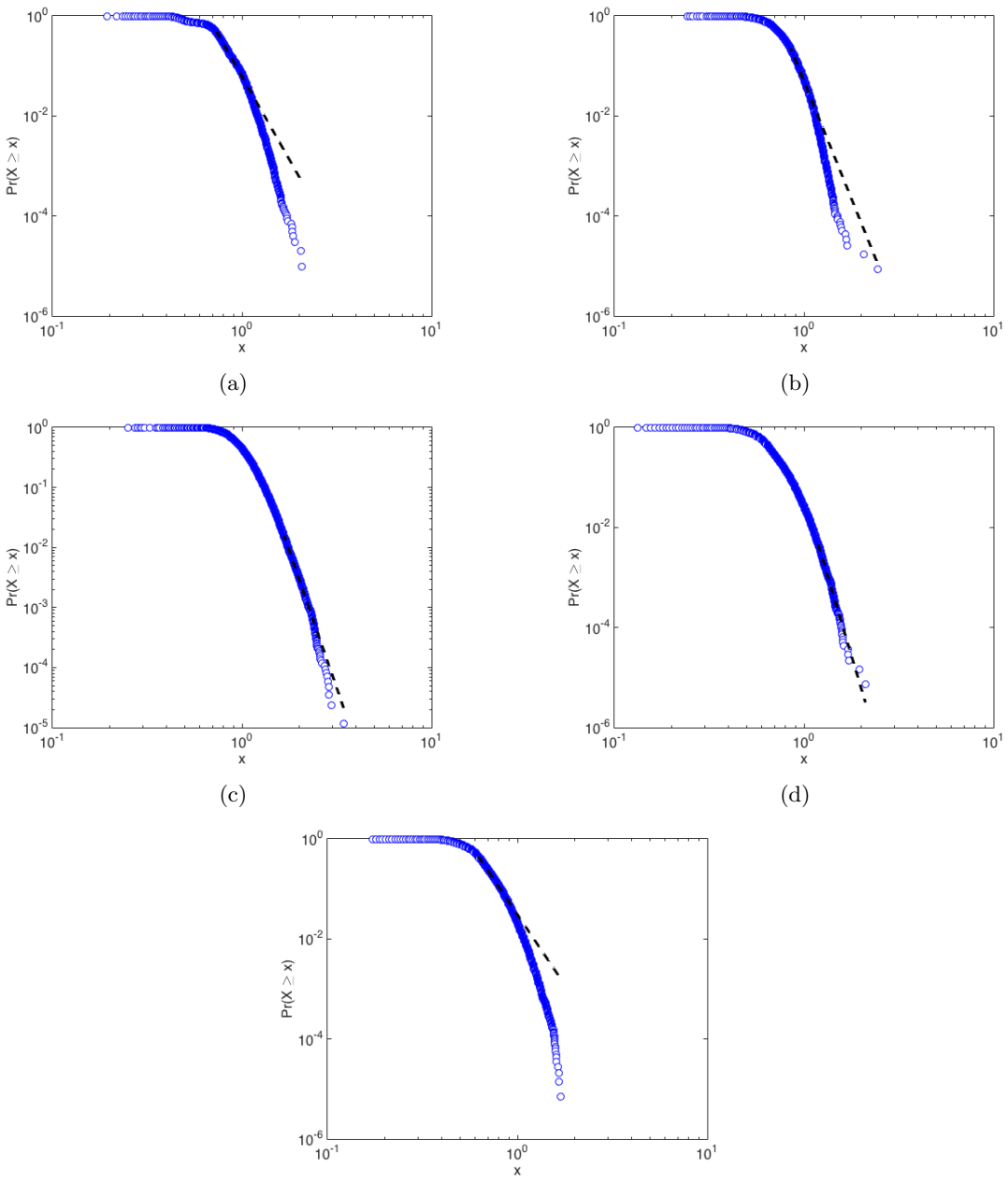


Figure 4.19: Logarithmic plots of the cumulative distribution functions (approximated and experimental) for the congestive heart failure cases.

5. Boxplot's detailled parameter results

The boxplots for the patients realized in the chapter on Random Forests were made on two classes of information extracted from the atrial rates :

- the maximum inter-peak duration in the 30 seconds following each pulmonary vein's disconnection : AA_{max}
- the total number of atrial beatings during the 30 seconds following each pulmonary vein's disconnection : AR

The resulting parameters for each patient are the following :

Table 4.1: Results AR and AA_{max} for the 54 patients in the training set.

patient	AA_{max}				AR				relapse
	PV1	PV2	PV3	PV4	PV1	PV2	PV3	PV4	
B14332X	0,866	1	0,889	0,932	228	222	218	236	0
R46370C	0,319	0,281	0,306	1,049	334	334	334	80	1
P29505W	0,448	0,484	0,956	1,05	184	178	68	64	1
CT8725W	0,347	1,168	0,51	0,581	326	116	314	280	0
C49701D	1,211	1,327	1,048	1,222	102	110	150	124	0
L75759J	0,337	0,39	0,215	0,269	336	338	342	336	0
DK1632B	0,288	0,296	0,341	0,303	326	324	302	302	0
H69269B	0,280	0,457	1,945	1,951	366	366	30	30	1
M97330X	0,376	0,489	1,028	1,765	330	318	228	156	0
A87586T	0,292	0,261	1,895	1,942	366	376	66	32	0
K46124K	0,476	0,647	0,755	0,624	136	142	130	134	0
A38284P	0,332	0,315	1,893	1,893	330	334	32	32	0
M72763Z	0,336	0,325	0,348	0,337	320	320	322	322	1
A32254S	0,573	0,404	0,474	0,551	286	318	298	288	1
DJ8974M	6,097	1,107			212	48			0
P73276K	0,283	0,319	0,32	0,287	328	320	316	316	1
R75211E	1,235	0,97	1,154	1,19	50	62	60	62	0
DB4032P	0,260	0,335	0,224	0,273	336	328	328	342	0
AR2652J	0,408	0,699	0,699	1,104	302	92	90	120	1
A94671C	0,233	0,227	0,27	0,217	354	350	340	354	0

patient	AA_{max}				AR				relapse
	PV1	PV2	PV3	PV4	PV1	PV2	PV3	PV4	
B95557X	0,594	0,593	0,763	0,799	160	110	102	82	1
E32945B		1,644				44			1
BR6881P	0,774	0,737	0,751	0,795	124	88	90	84	0
K10100H	0,615	0,491	0,475	1,233	252	236	238	68	1
CU4954T									0
K53536W	0,683	0,903	0,643	0,803	204	170	208	96	0
CY9456W	0,717	0,522	0,531	0,381	250	346	342	342	0
BR9461M	0,825	1,054	0,947	0,84	116	98	108	120	0
AL5400A	0,669	0,669	1,254	1,093	112	104	56	70	0
AR5091P	0,272	0,389	0,343	0,298	280	282	282	280	1
C58301R	0,311	0,302	0,426	0,559	330	342	276	246	1
CY6956W	0,275	0,371	0,288	0,302	334	338	340	338	0
DK4405J	0,254	0,321	0,307	0,341	330	334	332	320	0
BH5097U	0,243	0,350	0,255	0,27	308	306	310	320	1
D36149L	0,227	0,225	0,212	0,227	330	328	334	320	1
BM1634T	0,905	0,874	0,842	0,781	132	138	126	130	1
DL1804M	0,309	0,400	0,424	0,798	320	304	306	158	0
J81952L	0,519	0,755	0,745	0,793	244	236	238	192	0
E97063F	1,085	0,779	0,771	0,841	130	76	78	102	0
DH4373T	0,359	0,631	0,564	0,653	360	340	322	286	0
J17655D	0,719	0,722	0,68	0,817	282	224	202	218	0
CL3972F									1
DM4871F	0,266	0,274	0,270	0,278	300	302	300	288	0
CR5550B	0,600	0,369	0,486	0,995	270	272	260	124	1
K94099P	0,599	0,599	0,762	0,71	100	136	92	86	0
U16379E	0,905	0,852	0,381	1,008	72	72	202	88	0
D96217K	0,604	0,390	0,466	0,551	244	286	242	258	1
AH8080M	1,034	1,250	0,950	0,816	100	100	96	102	0
DC8092R	0,299	0,298	0,382	0,338	330	324	312	300	0
DL5404Z	0,254	0,268	0,260	0,261	352	344	336	348	0
L54584C	0,409	0,410	1,057	1,063	172	172	132	134	0
CR7432H	1,311	2,009	1,419	1,115	48	48	58	54	0
E16857K	0,738	1,521	1,722	0,738	82	40	42	96	0
A97512D	0,349	0,351	0,332	0,333	282	282	282	282	0
J22815C	1,336	1,224	1,953		78	152	30		0
B79166R	0,420	0,694	0,997	0,991	270	226	60	64	0
CT0562S	0,329	0,264	0,248	0,262	316	330	316	322	0
P11659F	1,269	1,377	1,286	1,271	78	64	68	66	0
U01678Z	1,450	1,450	1,115	1,166	64	64	70	66	0
D52113M									0

Some patients in the table have missing information. Those patients were removed and not considered in the training set of 54 patients used for the final Random Forests model.

Bibliography

[1] Gernot Ernst, 2014
Heart Rate Variability
Springer-Verlag London 2014

[2] Wikipedia page about the atria of the heart, which includes a clear diagram.
https://en.wikipedia.org/wiki/Atrium_(heart)

[3] KCE, Centre d'expertise des soins de santé, 2012
Ablation par cathéter pour troubles du rythme cardiaque : moins efficace qu'espéré.
https://kce.fgov.be/fr/ablation-par-cathéter-pour-troubles-du-rythme-cardiaque-moins-efficace-qu'espéré

[4] Laurent M. Haegeli, Hugh Calkins, 2014.
Catheter ablation of atrial fibrillation: an update.
European Heart Journal, Volume 35, Issue 36, Pages 2454–2459.

[5] L. Glass, 2009
Introduction to controversial topics in nonlinear science: Is the normal heart rate chaotic?
Journal Chaos 19, 028501.

[6] Woo, M.A. Stevenson, W.G. Moser, D.K. Trelease, R.B. Harper, R.M., 1992.
Patterns of beat-to-beat heart rate variability in advanced heart failure.
American Heart Journal 123 (3), p.704-710.

[7] Tulppo, M.P., Makikallio, T.H., Takala, T.E., Seppanen T., Laukkanen, R.T., Huikuri, H.V., 1996.
Quantitative beat-to-beat analysis of heart rate dynamics during exercise.
American Journal of Physiology 274 (1), H244-H252

[8] PhysioNet, *'Is the Normal Heart Rate Chaotic ?'*, challenge proposed by `PhysioNet` website, 2009.
Data containing ECG recordings over 24h from fifteen patients.
https://physionet.org/challenge/chaos/

[9] Aaron Clauset, Cosma Rohilla Shalizi and M.E.J Newman, 2007
Power-law distributions in empirical data.
Society for Industrial and Applied Mathematics Review 51, 661-703 (2009).

[10] Wikipedia page about the decision trees.
https://en.wikipedia.org/wiki/Decision_tree

[11] Leo Breiman, September 1994
Bagging Predictors, Technical Report No. 421.
Berkeley, California 94720.

Rue Archimède, 1 bte L6.11.01, 1348 Louvain-la-Neuve **www.uclouvain.be/epl**