Circadian Variability and Discrimination in Day-Night periods based on Morphological Characteristics of P & T waves

MASTER THESIS

submitted to the designated

by the General Assembly Composition of the Department

of Computer Science & Engineering inquiry committee

from

Dimitrios Zavantis

in fulfillment of the requirements for the

MASTER'S DEGREE IN COMPUTER SCIENCE

WITH EXPERTISE IN

TECHNOLOGIES - APPLICATIONS

September 2015

Κιρκάδια Μεταβλητότητα και Διάκριση Ημέρας-Νύχτας στα χαρακτηριστικά της μορφολογίας των P & T κυμάτων

ΜΕΤΑΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ ΕΞΕΙΔΙΚΕΥΣΗΣ

υποβάλλεται στην

ορισθείσα από την Γενική Συνέλευση Ειδικής Σύνθεσης

του Τμήματος Μηχανικών Η/Υ & Πληροφορικής Εξεταστική Επιτροπή

από τον

Δημήτριο Ζαβαντή

ως μέρος των υποχρεώσεων για την λήψη του

ΜΕΤΑΠΤΥΧΙΑΚΟΥ ΔΙΠΛΩΜΑΤΟΣ ΣΤΗΝ ΠΛΗΡΟΦΟΡΙΚΗ

ΜΕ ΕΞΕΙΔΙΚΕΥΣΗ

ΣΤΙΣ ΤΕΧΝΟΛΟΓΙΕΣ - ΕΦΑΡΜΟΓΕΣ

Σεπτέμβριος 2015

Dedication

Everybody deserves somebody who
makes them look forward to tommorow.

To my love ...

Acknowledgements

I would like to express my gratitude to my supervisor, **Assistant Prof. Georgios Manis** for his guidance, advise and constant support throughout my thesis work.

Next, I want to mention the names of Paris Tsantarliotis, Prokopis Kontogiannis and especially Ermioni Mastora all the Msc students for helping me a lot during my thesis period as well for the manuall detection of the waves.

I want to thank all of my friends and mostly Athina Thoma and Theodosia Salika for all the thoughtful and motivating discussions we had, which encouraged me to think beyond the observable.

I would like to thank all faculty members and staff of the Department of of Computer Science & Engineering, University of Ioannina, Greece for their generous help in various ways for the completion of this thesis.

I am especially grateful to my parents for their love and support and would like to thank my parents for raising me in a way to believe that I can achieve anything in life with hard work and dedication.

Dimitrios Christou Zavantis
September 2015

Table of Contents

List of Figures

List of Tables

Abstract

Dimitrios C. Zavantis: MsC, Department of Computer Science & Engineering, University of Ioannina, Greece; Graduation September, 2015.

MsC thesis: Circadian Variability and Discrimination in Day-Night periods based on Morphological Characteristics of P & T waves

Supervisor: George Manis, Assistant Professor.

One of the most important research issues of recent years in the section of biomedical informatics is the electrocardiogram (ECG) since it represents a non-invasive method which provides information about the heart rate. The provided information is significant for comprehending the function of human heart and its influencing factors. Apart from the most visible peak (the R one), there also exist the T one and the less examined, the P wave.

The aim of this thesis is an extensive study of P & T waves in ECGβέΰs of healthy people. Using a multitude of exported features, which describe the morphology of these waves, we enhanced the analysis of a potential differentiation between day and night periods and a suspected occurrence of circadian rhythm in the waves.

The export of P & T waves initially implemented in a manual mode. However, the need for a rapid and effective detection algorithm prompted the creation of two automated algorithms: the *percentile based Automatic Detection* and the *Graphical based Automatic Detection*. Both algorithms use statistical and probabilistic concepts (functions) to achieve adequate delineation and detection of the waves. The first of them, uses percentile to define and restrict the location of these waves, while the second one takes advantage of the existing monotonicity and slope of an ECG and creates a collection of waves which then imports to a probabilistic graphical model for classifying them to P or T.

Finally, the extraction of the provided information in combination with the number of our waves, extends our research area to the study of PP and TT intervals in comparison to the already well-known RR interval by examining the potential of their relationship. The analysis of these intervalsβέΰ behavior in terms of morphological characteristics in two time windows (30 and 60 minutes) is becoming the focal point of our thesis for the recognition of periodic behavior (throughout the 24- hour period) which will confirm the existence of circadian rhythm.

ΠΕΡΙΛΗΨΗ

Δημήτριος Χ. Ζαβαντής MsC, Τμήμα Μηχανικών Η/Υ & Πληροφορικής, Πανεπιστήμιο Ιωαννίνων. Αποφοίτηση, Σεπτέμβριος 2015.

Μεταπτυχιακή Διατριβή: Κιρκάδια Μεταβλητότητα και Διάκριση Ημέρας-Νύχτας στα χαρακτηριστικά της μορφολογίας των P & T κυμάτων

Επιβλέπων: Γεώργιος Μανής, Επίκουρος Καθηγητής.

Ένα από τα σημαντικότερα αντικείμενα - έρευνας, των τελευταίων ετών, στον χώρο της βιοϊατρικής πληροφορικής αποτελεί το ηλεκτροκαρδιογράφημα (ΗΚΓ) καθώς εκπροσωπεί μία μη επεμβατική μέθοδο που δίνει πληροφορίες, μεταξύ άλλων, και για τον καρδιακό ρυθμό. Η πληροφορία η οποία παρέχει είναι σημαντική για την κατανόηση της λειτουργία της καρδιάς όπως και των παραγόντων που την επηρεάζουν.

Στόχος αυτής εργασίας είναι μία εκτεταμένη μελέτη σε υγιείς ανθρώπους των P & T κυμάτων, τα οποία αποτελούν σημαντικά επάρματα του ηλεκτροκαρδιογραφήματος. Η χρήση πληθώρας χαρακτηριστικών που εξήχθησαν, τα οποία περιγράφουν την μορφολογία αυτών των κυμάτων, ενισχύει την ανάλυση για την πιθανή διαφοροποίηση τους ανάμεσα σε δύο χρονικές περιόδους ημέρας-νύχτας, όπως επίσης και της υπόνοια εμφάνισης κιρκάδιου ρυθμού στα κύματα. Η εξαγωγή των P & T κυμάτων έγινε με χειρωνακτικό τρόπο αρχικά. Ωστόσο, η ανάγκη για ένα γρήγορο και αποτελεσματικό αλγόριθμο ανίχνευσης τέτοιων κυμάτων μας ώθησε στην δημιουργία δύο αυτοματοποιημένων αλγορίθμων: του *percentile based Automatic Detection*, και του *Graphical based Automatic Detection*. Οι δύο αλγόριθμοι χρησιμοποιούν στατιστικές και πιθανοτικές έννοιες (συναρτήσεις) για την οριοθέτηση και ανίχνευση των κυμάτων. Ο πρώτος εξ αυτών χρησιμοποιεί το εκατοστημόριο για να καθορίσει και περιορίσει την περιοχή που βρίσκονται τα κύματα ενώ

ο δεύτερος εκμεταλλευόμενος την μονοτονία και την κλίση του ηλεκτροκαρδιογραφήματος δημιουργεί μία συλλογή κυμάτων τα οποία τα εισάγει στην συνέχεια σε ένα πιθανοθεωρητικό γραφικό μοντέλο για την ταξινόμηση τους σε P & T.

Τέλος, η εξόρυξη όλης αυτής της πληροφορίας και του πλήθους των κυμάτων εντείνει την έρευνα μας για τη μελέτη των χρονικών διαστήματών PP και TT έναντι του ήδη γνωστού RR και του δυναμικού της σχέσης τους. Ο έλεγχος της συμπεριφοράς των παραπάνω διαστημάτων και των μορφολογικών χαρακτηριστικών σε δύο χρονικά παράθυρα (30 και 60 λεπτών) γίνεται το επίκεντρο της παρούσας εργασίας για την αναγνώριση περιοδικής συμπεριφοράς (σε όλο το 24-ωρο) η οποία θα επιβεβαιώσει την ύπαρξη του κιρκάδιου ρυθμού.

Chapter 1

Introduction

1.1 Chronobiology

Chronobiology [1] is known to be a field of biology that investigates periodic phenomena in living organisms and their adjustment to solar- and lunar-related rhythms, known as

biological rhythms. The major study in chronobiology includes the research of biological clocks mechanisms and their relationships with the environment. This knowledge has been adapted to other scientific field (e.g. genetics, comparative anatomy, molecular biology, ecology, physiology, neuroscience, and much more).

1.1.1 Biological Rhythms

Biological rhythms [1] are related to some physiological functions or activities of the body and usually run on a diurnal cycle. These rhythms can be internal or external depending on the influence factor. The first is associated with body functions and necessary activities (e.g. body temperature, daily performance, alertness, sleep schedules and endocrine activity) while the latter is related with environmental time cues (e.g. sunlight, noise, food, drugs, caffeine).

The period time with reference point 24-hour can categorize the biological rhythms as follows:

- Circadian rhythms (24-hour cycle)
- Diurnal rhytms (circadian day/night cycle)
- Ultradian rhythms (shorter period times than circadian)
- Infradian rhythms (more than 24 hours period times)

In our research we will focus only on circadian rhythms.

1.1.2 Circadian Rhythms

Circadian[1] originates from a Latin phrase meaning "about a day" (circa + diem). These rhythms are part of a 24-hour cycle and can be physiological and behavioral rhythms. These occur in sleep/wakefulness cycle, blood pressure, body temperature, hormone secretions, digestive secretions, etc. In general, the brain controls and helps maintain the "internal clock" for these rhythms.

[1] https://en.wikipedia.org/wiki/Circadian_ rhythm

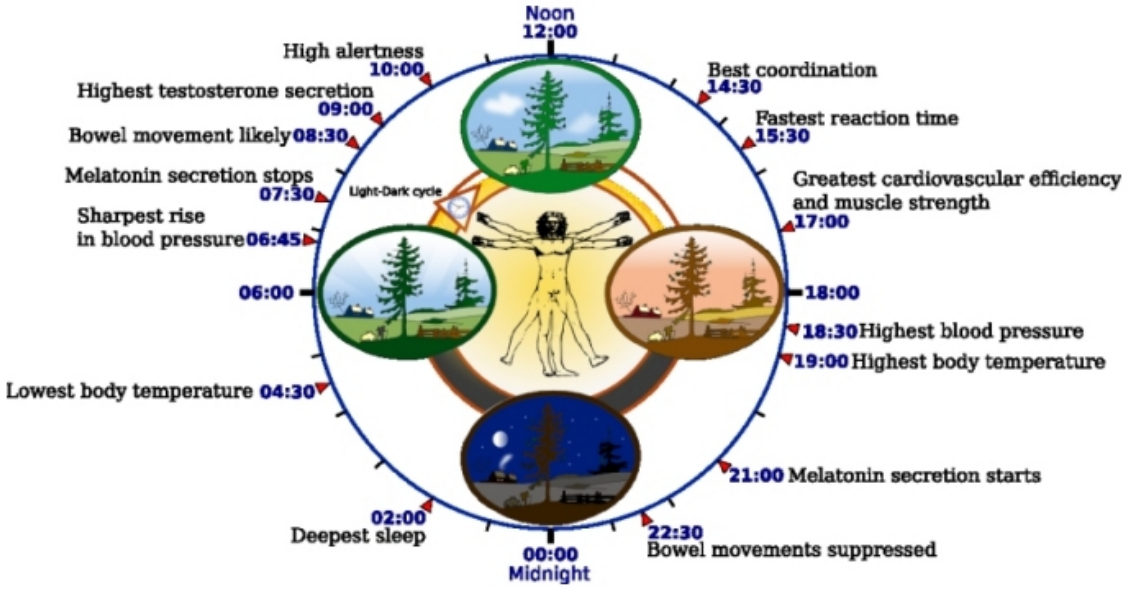

Figure 1.1: Human circadian biological clock.

The circadian rhythm is a 24-hour cycle that tells our bodies when to sleep and arranges many other physiological processes as displayed in fig. 1.1[2]. When one's circadian rhythm is disrupted, some patterns can run amok like sleeping and eating. A lot of research is examining if adverse health results can disrupt circadian rhythm, like increasing the chances of cardiovascular events and a correlation with neurological problems like bipolar disorder and depression.

The circadian rhythm decreases and increases at different times of the day. For instance, adults' strongest sleep drive occurs between 2:00-4:00 in the morning and between 1:00-3:00 in the afternoon, even though this varies according to the type of person we are ("morning or evening person"). During these circadian decreases, the sleepiness we experience will be less intense after a sufficient sleep, and more intense with lack of sleep. The circadian rhythm keeps us in alert at fixed points of the day, even if we have been awake for hours and our sleep/wake stimulative process would either-way make us feel

[2]"Biological clock human" by NoNameGYassineMrabetTalk fixed by Addicted04 - The work was done with Inkscape by YassineMrabet. Informations were provided from "The Body Clock Guide to Better Health" by Michael Smolensky and Lynne Lamberg; Henry Holt and Company, Publishers (2000). Landscape was sampled from Open Clip Art Library (Ryan, Public domain). Vitruvian Man and the clock were sampled from Image:P human body.svg (GNU licence) and Image:Nuvola apps clock.png, respectively. Licensed under CC BY-SA 3.0 via Wikimedia Commons - https://commons.wikimedia.org/wiki/File:Biological_clock_human.svg # /media/File:Biological_clock_human.svg

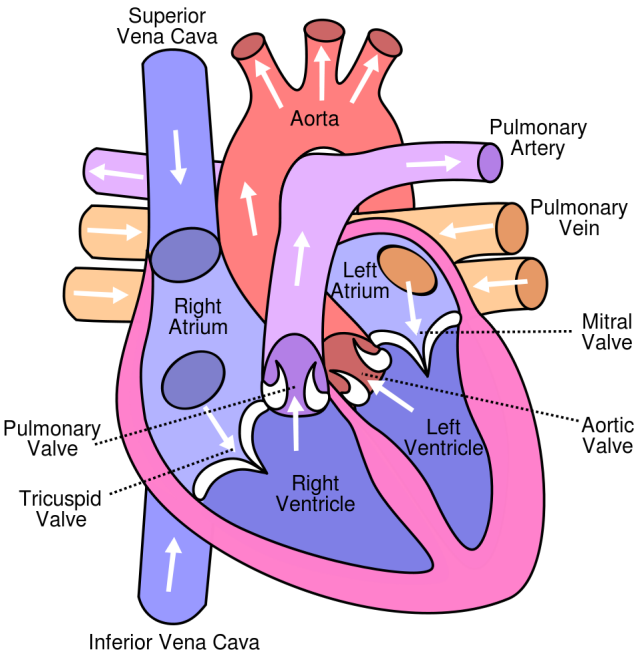

Figure 1.2: The heart, showing valves, arteries and veins. The white arrows shows the normal direction of blood flow.

more sleepy.

1.2 The heart anatomy

Our heart[3] has 4 chambers as shown in fig. 1.2[4]. The two upper chambers defined as the left and right atria, and the two lower chambers defined as the left and right ventricles. The septum is a wall of muscle that divides the left and right atria and the left and right ventricles. The largest and strongest chamber in our heart is the left ventricle, while the left ventricle's chamber walls are about a half-inch thick. The latter despite its thickness is able to push blood into our body through the aortic valve.

Four valves regulate blood flow through our heart:

- The blood flow between the right atrium and right ventricle is controlled by the tricuspid valve.

[3] https://en.wikipedia.org/wiki/Heart

[4] "Diagram of the human heart (cropped)" by Own work. Licensed under CC BY-SA 3.0 via Commons - https://commons.wikimedia.org/wiki/File:Diagram_ of_ the_ human_ heart_ (cropped).svg# /media/File:Diagram_ of_the_ human_ heart_ (cropped).svg

- The blood flow from the right ventricle inside the pulmonary arteries is regulated by the pulmonary valve, carrying blood to our lungs to pick up oxygen.
- The oxygen-rich blood from our lungs passes through the mitral valve, from the left atrium into the left ventricle.
- The oxygen-rich blood passes through the aortic valve, from the left ventricle into the aorta, which represents our body's largest artery.

1.3 Electrocardiogram

Electrocardiogram[5] (abbreviated ECG) is a tool mostly used in the clinical practice due to its excellent benefit-cost relationship, as well for diagnosis that indicates the electrical activity of heart by skin electrode recordings. The cardiac health of human heart beat is expressed by the morphology and heart rate. It is a noninvasive technique, meaning the surface of human body is used for the measurement of the signal, which helps in identification of the heart diseases. Any change in the morphological pattern or disorder of heart rate or rhythm, is an evidence of cardiac arrhythmia and could be recognized using waveform analysis on the recorded ECG. The ECG waveform is well characterized by the waves: P, QRS, and T. The most significant wave is the QRS complex, but there is much interest in examining T-waves and, especially lately, P-waves. The duration and amplitude of the P-QRS-T wave expresses useful information about the nature of the heart disorders. Atrial and ventricular depolarization and repolarization of Na^+ and K^- ions in the blood are the origin of the electrical wave. The ECG signal provides the following information of a human heart:

- heart rhythm and conduction disturbances
- heart position and its relative chamber size

[5] https://en.wikipedia.org/wiki/Electrocardiography

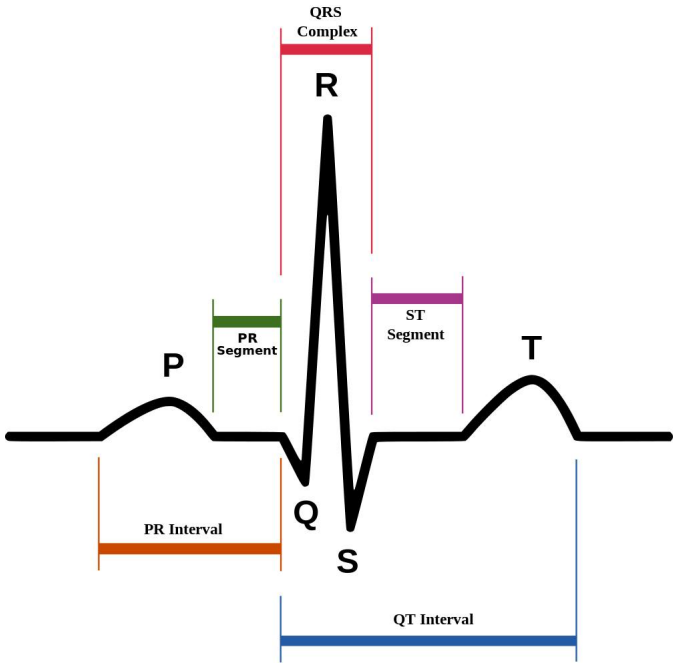

Figure 1.3: Schematic representation of normal ECG waveform.

- changes in electrolyte concentrations
- extent and location of myocardial ischemia
- drug effects on the heart
- impulse origin and propagation

1.4 ECG waves and intervals

As described in the previous section the ECG (see fig. 1.3[6]) consists of:

- The P wave is associated with right and left atrial depolarization. The wave of atrial repolarization is invisible because of low amplitude. A clear P wave before the QRS complex represents sinus rhythm whereas absence of P waves may suggest junctional rhythm or ventricular rhythm, atrial fibrillation. It is very difficult to

[6]"SinusRhythmLabels" by Created by Agateller (Anthony Atkielski), converted to svg by atom. - en:Image:SinusRhythmLabels.png. Licensed under Public Domain via Commons - https://commons.wikimedia.org/wiki/File:SinusRhythmLabels.svg# /media/File:SinusRhythmLabels.svg

analyze P waves with a high signal-to-noise ratio in ECG signal. Normal P wave is less than 2.5 mm (two-and-a half 1-mm-divisions) tall and no more than 120 ms (three 1-mm-divisions) in width in any lead.

- The second wave is the QRS complex. Typically this complex has a series of 3 deflections that represent the current related with right and left ventricular depolarization. By convention the first negative deflection in the complex is called a Q wave, whereas, the R wave is the first positive deflection in the complex. Finally, a negative deflection after an R wave is defined as S wave. A second positive deflection after the S wave is called the R wave, if there is one. Some QRS complexes do not have all three deflections. But irrelevant of the present number of waves, they are all QRS complexes. Duration of the QRS complex expresses the time for the ventricles to depolarize and can give information about conduction problems in the ventricles such as bundle branch block. QRS duration is the width of that complex from beginning to end, irrespective of the number of deflections. Normally it lasts less than 120 ms (three 1-mm-divisions).

- The T wave represents the current of rapid phase 3 ventricular repolarization. The polarity of this wave normally follows that of the main QRS deflection in any lead. During that period of repolarization the ventricles are observed to be electrically unstable extending from the peak of the T wave to its primary downslope. A stimulus (e.g. a premature beat) falling on this vulnerable period has the power to trigger ventricular fibrillation: the so call R-on-T phenomenon. Large T waves may represent ischemia, and Hyperkalaemia.

1.5 Holter Monitoring

A Holter monitor [2] is a portable device operated with battery that measures and records our heart's activity (ECG) continuously for 24 to 48 hours or longer according to the

monitor used. The device is the size of a small camera. Wired silver dollar-sized electrodes attach to our skin to tape records. The Holter monitor and other devices that record our ECG as we go about our daily activities are called ambulatory electrocardiograms. As a result of extended recording period, the observation of occasional cardiac arrhythmia or epileptic events is possible. On the other hand, such disorders would be difficult to identify in a shorter period of time. In transient symptoms patients have to wear for a month or more a holter to monitor a cardiac event.

1.6 ECG Database

In this study the MIT-BIH Normal Sinus Rhythm Database was used. The dataset includes 18 long-term (24-hour) ECG recordings. The subjects had no significant arrhythmia. They include 5 men, aged 26 to 45, and 13 women, aged 20 to 50. The database is sampled at 128 Hz and the data is available at uniform intervals of 7.8125 msec [3]. From this database we selected day and night periods from one to three o'clock in order to study the discrimination in diametrically opposed time intervals.

1.7 Motivation and Objectives

Generally, the shape of heart rate and ECG waveform reflects the state of cardiac heart. It is obtained by a non invasive way which can provide a lot of information directly or indirectly. In this thesis we focus mostly in the detection of P and T waves from healthy people.

The first objective is the discrimination of day-night periods using exclusively P & T Waves. In particular we will investigate several features extracted from P and T waves that have been detected manually or automatic in diametrically opposed time intervals. One manual detection has been implemented. The manual detection is initially created

to show that there is a significant difference between those two time periods.

However, the need for a fast and sufficient detection algorithm for these waves motivated us to create two new automatic detections. The first automatic detection named as *Percentile based Automatic Detection*; it uses the percentile in order to find the waves within a bounded area. The second automatic detection named as *Graphical based Automatic Detection*; it uses a Graphical Probabilistic model (HCRF) which classifies the two categories of the waves (P or T).

Hence, the waves that have been extracted from the methods described above give us the opportunity to investigate more about these waves. The big amount of collected waves makes possible the study for the PP and TT intervals against RR intervals and their potential relationship with the Circadian Rhythm. Finally, an additional inspection in the behavior of wave's feature in the sleep/wake cycle (24-hour) opposite to circadian's behavior becomes the objective of this work.

1.8 Thesis Structure

This master thesis includes 8 chapters:

The Chapter 1 of this thesis explains the chronobiology with two rhythms, biological and circadian. The basic of ECG and ECG morphology as well as the MIT-BIH Normal Sinus Rhythm Database is discussed. At last, the motivation and the objectives of this work are described.

In Chapter 2, a manual detection of P & T waves is described. This algorithm uses visual inspection every 2000 values in ECG signal.

Chapter 3 defines two novel automatic detections with their results for the waves that have been selected.

A variety of characteristic features of ECG are extracted, which consist of morphological features and their ratios for each wave. In Chapter 4 feature extraction methodology of above features are discussed.

The Discrimination between day and night using paired t-test and its results are performed in Chapter 5.

In Chapter 6 several classification methods are applied to strengthen the discrimination hypothesis using the extracted features.

Chapter 7 investigates the Circadian behavior of PP and TT intervals as well as the existence of this rhythm in features.

Finally, in Chapter 8 the discussion of all above results takes place with future work improvements for the described methods.

Chapter 2

Manual Detection

2.1 Introduction

Manual detection was used to provide a known information for the research. The aim of gathering waves manually was the collection of reliable results independent of potential limitations of an automatic method. Two time periods were selected knowing the influences between day and night in human physiology and behavior as shown in fig. 2.1. According to that figure and with the knowledge of dips and rises of the circadian rhythm two time periods were selected for the detection and further analysis described in next Chapters.

This kind of detection can be characterized as time-consuming, exhaustive and subjective task. The waves correctness depends on quality of the signals, the experience of the researcher and ECG lead. In order to convert the nature of this task to an objective

one and to achieve better and reliable results, we assembled a team of four people. The team was familiar with the concepts of heart rate, ECG's components.

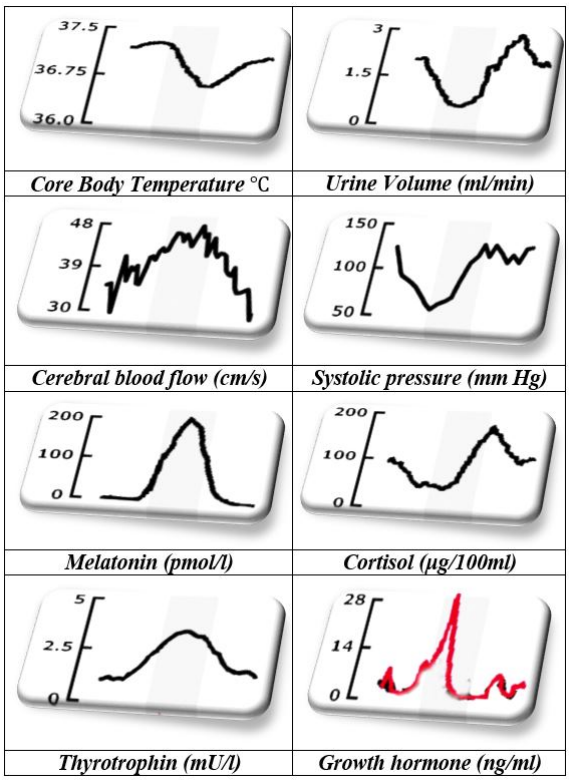

Figure 2.1: Rhythmic changes in human physiology and behavior from 2PM to 2PM

2.2 Methodology

A manual selection was implemented using an auxiliary visual inspection tool to assist fast and accurate selection, uniformly per period. The ECG signal is very large, 24-hour long. However, we focused on two specific time periods during the 24-hour period, daytime and nighttime. Daytime and nighttime periods are defined as the periods from 1 PM to 3 PM and from 1 AM to 3 AM respectively. Thus, we want to manually detect P-waves and T-waves during these periods. In order to minimize the risk of error, the detection is done in two phases.

At first, the ECG was cut into buckets of 2000 values (fig. 2.2). Each value represents $7.8125 msec$, so every bucket describes a $15.625 sec$ part of ECG. This tactic provides

the opportunity for a better visual inspection using local information in a different scale ($1/\#ofBuckets$) for the detection of the waves. In the procedure of manual detection the user reads each bucket of ECG and marks the beginning and end of every P-wave and T-wave as shown in fig. 2.3. The number of waves which would be selected was determined by the user for each bucket. For the rejection of waves common guidelines were set among the team members. Rejection was decided is presence of :

- noise,
- variability
- baseline drift
- double peak.

In the second phase, a second crosscheck was performed on the waves selected during the previous phase, to avoid any malfunction during the first selection. Furthermore, this check helped us to achieve more accuracy on the onset and offset points of the waves, improving their quality.

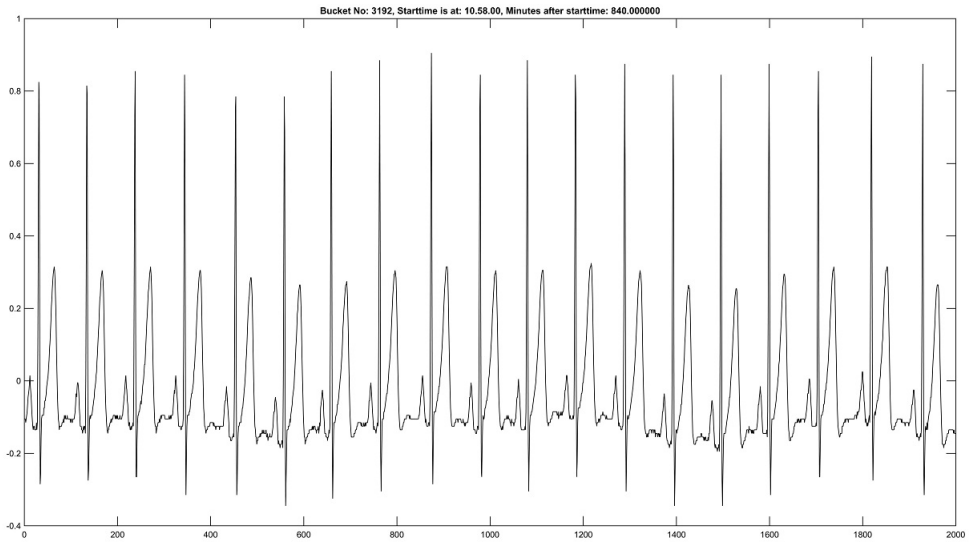

Figure 2.2: Sample of a bucket

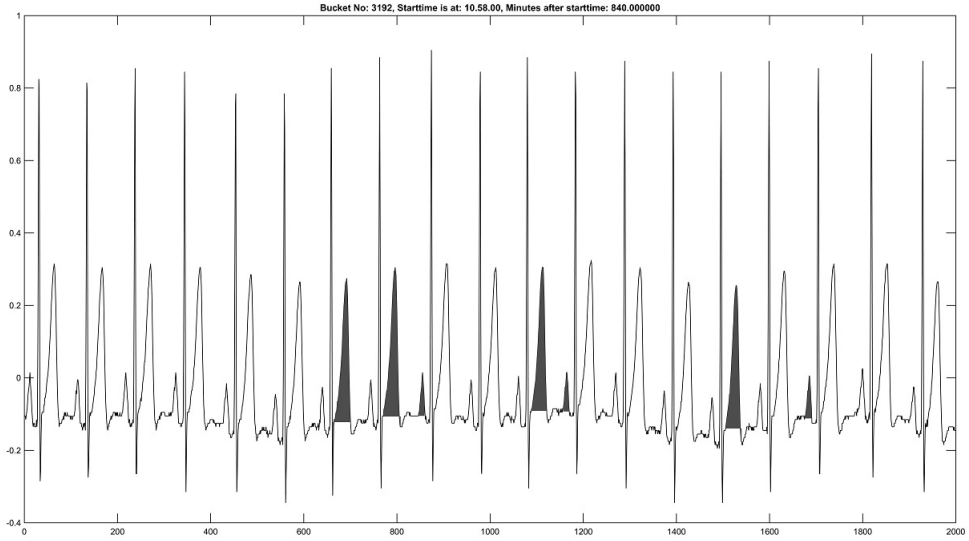

Figure 2.3: Sample of a bucket with manual selection

2.3 Results

We manually detected P-waves and T-waves in the MIT-BIH Normal Sinus Rhythm Database, which was described in Chapter 1. The workload was distributed equivalently to the members of the team. The total number extracted from those two time periods was 2700 P and 2700 T waves from all patients. The next step was an additional research of the waves. For that purpose several features were extracted according to their morphology that will be discussed below (see Chapter 4).

Chapter 3

Automatic Detection

3.1 Introduction

Over the last few years, many sophisticated methods have been proposed for the detection of P & T waves. Trahanias and Skordalakis [4] applied a syntactic approach to ECG pattern recognition and parameter measurement for the detection of P, QRS and T waves. Murthy and Prasad [5] used the discrete cosine transform (DCT) for delineation of P waves, whereas Murthy & Niranjan [6] the discrete Fourier transform (DFT). Thakor and Zhu [7] used their own adaptive filters focusing on P waves.

The detection and furthermore the annotation of P and T waves is not as simple compared to the QRS complex for a number of reasons which include a low signal-to-noise ratio (SNR), morphological variability, low amplitude and amplitude variability and the chance of overlapping of the P or T wave with the QRS complex. The P wave may not even be present in some ECG recordings. In the majority of these methods P and T waves are detected by their relativity to the position of the R peak by applying the appropriate threshold. The primary problems of the thresholding methods are the acute sensitivity to noise and their poor efficiency when dealing odd morphologies (e.g. negative or biphasic waves).

The initial aim of this work is to propose automated detection methods with the knowledge that the subjects are healthy with normal waves. Hence, we present two methods; The *percentile Automatic Detection*, and the *Graphical based Automatic Detection*. Both of these two methods are used to increase the number of the sample of the waves, for better and reliable statistical analysis.

The *percentile Automatic Detection* (pAD) detects T and P waves using the percentile. If we suppose that the values of the ECG are sorted in ascending then it is noticeable that most frequent values are close to baseline or near it. pAD takes advantage of this observation and detects the peaks of the waves that belong in a certain area determined by specific percentile values of the signal, preventing interaction with the values of QRS complex. The buckets of 2000 values are also used in pAD so that the baseline drift of the signal to be eliminated. As the set of peaks of the waves is determined, two

phases are following: the delineation and acceptation. The delineation of the wave is the determination of onset and offset point of the wave based on its monotonicity left and right to the peak. The acceptation of the candidate waves is decided by comparison with R peaks, indicated as P waves those before R peak and as T waves those after R peak. pAD can be categorized in methods using threshold but percentile is changing from bucket to bucket taking into consideration the wave. As a result, the threshold is not fixed for the entire ECG but changing dynamic according to the wave.

Generally, the annotation of R peak is not always provided in online databases. The need for a further analysis in ECGs' waves led us in creating a second approach for the detection of the waves which can handle more loose demands independent of R peaks. The second method of automatic detection *Graphical based Automatic Detection* (GlAD) is using a "feature wave-bank" to learn features of the waves of each patient and using a graphical probabilistic model named as Hidden Conditional Random Field (HCRF) to categorize the candidate waves in P and T waves online. In brief, a random subset of selected waves from pAD is used to extract some morphology features for each patient. These features are used to train HCRF. The next phase is a procedure for the selection of candidate waves and their feature extraction. Finally, HCRF classifies candidates in P or T waves.

3.2 Hidden Conditional Random Fields

Hidden Markov model (HMM) [8] is one of the most widely used tools that includes hidden-state structure instead of fully observable construction. HMM is a finite automaton containing individual-valued states H broadcasting a data vector X at each time point, depending on the current state of the distribution of the data at each time point. This kind of models are generative requiring evaluation of the joint probability density function (PDF) over multiple time points in the observed data samples. In order to make the problem's assumption manageable, we should make some affairs conditioned on the

states about independence of the data at each time point. Although such hypotheses are disrupted in many practical schemes.

Conditional Random Fields (CRFs) were first introduced by Lafferty *etal*. [9] overcome the independence assumptions, since these are discriminative models that prevent the need to model the data distribution. Despite the fact they deal with independence issues, they propose a label assignment for each observation which is not so easy task with big data. Moreover, CRFs do not use hidden states nor precisely provide a way for the estimation of the conditional probability of a class label. CRF training for wave recognition requires an appropriate assignment of part labels to the local features in the training data. In public datasets this kind of information is not always available, and manual annotation (e.g. P or T wave) is a time-consuming task.

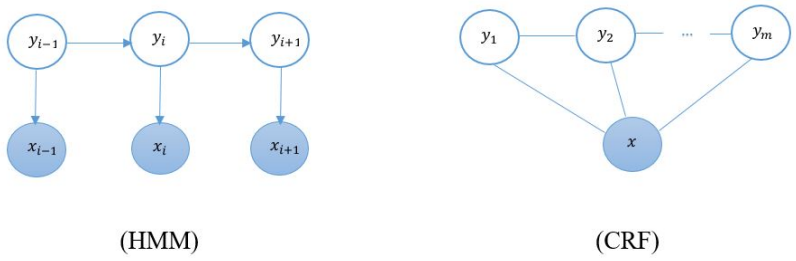

Figure 3.1: Graphical representation of HMM and CRF

Hidden Conditional Random Fields (HCRFs) [10] are based on CRFs as a natural extension of them. HCRFs directly output the action label and do not require the assignment of part labels in the training data. They model the spatial and temporal structures by introducing an additional layer of structured hidden variables with dependencies among them. HCRFs decide the joint distribution of a class label and hidden state labels conditioned on observations and expressed using an undirected graph.

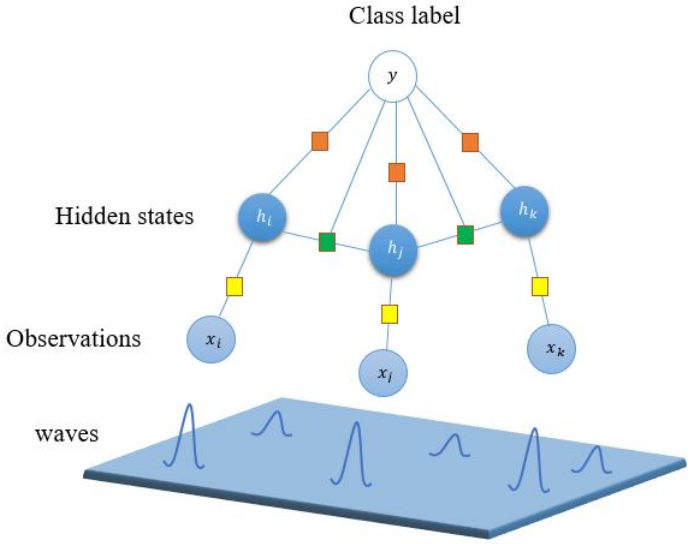

Figure 3.2: Graphical representation of HCRF

The difference between HMMs, CRFs and HCRFs can be shown using graphical representations in Fig. 3.1,3.2. The colored rectangles in HCRF denote the relationship between observation and hidden states (yellow), hidden states and labels (red), as well as the compatibility of class label with a transition from one hidden state to another (green).

Our target is a mapping of observations X to class labels $y \in \mathcal{Y}$, where x is a vector of m local observations, $X = \{x_1, x_2, ...x_m\}$, and each local observation x_j is a feature vector $\phi(x_j) \in \mathcal{R}^d$, where d is the dimensionality of the representation as described in Bousmalis et al. [11]. An HCRF models the conditional probability of a class label given an observation sequence by:

$$P(y|x,\theta) = \sum_h P(y,h|x,\theta) = \frac{\sum_h \exp \Psi(y,h,x;\theta)}{\sum_{y' \in \mathcal{Y}} \sum_h \exp \Psi(y',h,x;\theta)} \tag{3.1}$$

where $h = \{h_1, h_2, ..., h_T\}$, are the hidden variables $h_i \in \mathcal{H}$. If we assume that there is a single class label y and that h is observed then the conditional probability of h given x turns into a regular CRF. The potential function $\Psi(y,h,x;\theta) \in \mathcal{R}$, is parameterized by θ, which measures the compatibility between a set of observations, a label and a configuration of the hidden states. The graph of our model is a chain where each node corresponds

to a latent variable h_j. Our parameter vector θ is made up of three components: $\theta = [\theta_{k_1}^T \theta_{k_2}^T \theta_{k_3}^T]^T$. Parameter vector θ_{k_1} models the relationship between features $\phi(x_j)$ and hidden states $h_j \in H$ and is typically of length $(d \times |\mathcal{H}|)$. θ_{k_2} models the relationship of the hidden states $h_j \in \mathcal{H}$ and labels $y \in \mathcal{Y}$ and is of length $(|\mathcal{Y}| \times |\mathcal{H}|)$. θ_{k_3} represents the links between hidden states. It is equivalent to the transition matrix in a HMM, with a significant difference that a HCRF keeps a matrix of "transition" weights for each label and θ_{k_3} is of length $(|\mathcal{Y}| \times |\mathcal{H}| \times |\mathcal{H}|)$. We define potential functions for each of these relationships, and our Ψ as their product along the chain

$$\Psi(y, h, x; \theta) = \sum_{j \in V} \sum_{k_1 \in \mathcal{K}_1} \theta_{k_1} \cdot \phi_{k_1}(x_j, h_j) + \sum_{j \in V} \sum_{k_2 \in \mathcal{K}_2} \theta_{k_2} \cdot \phi_{k_2}(y, h_j) + \sum_{(i,j) \in E} \sum_{k_3 \in \mathcal{K}_3} \theta_{k_3} \cdot \phi_{k_3}(y, h_i, h_j) \tag{3.2}$$

Note that this function is a general form and it is formulated in this way for simplicity. The first two terms are node terms and the third is one edge term.

We use the notation $\theta_{k_1} \cdot \phi_{k_1}(x_j, h_j)$ to the weight or potential that estimates the compatibility between the feature indexed by state $h_j \in \mathcal{H}$ and x_j. Similarly,$\theta_{k_2} \cdot \phi_{k_2}(y, h_j)$ stand for weights or potentials that correspond to class y and state h_j , whereas $\theta_{k_3} \cdot \phi_{k_3}(y, h_i, h_j)$ measure the compatibility of the label y with a transition from h_i to h_j.

From Eq.(3.1) and $P(y, h|x, \theta)$, we can use Bayes' rule to derive the joint probability of assigning a set of part labels h when its features x, class label y and weight parameters θ are known:

$$P(h|y, x, \theta) = \frac{P(y, h|x, \theta)}{P(y|x, \theta)} = \frac{e^{\Psi(y,h,x;\theta)}}{\sum_h e^{\Psi(y,h,x;\theta)}}. \tag{3.3}$$

The training of HCRF model is the same as the ordinary CRF model except the sum of hidden variables. Following previous work on CRF, we want to maximize the joint conditional probability $P(y|x, \theta)$ for all training examples. The objective function used for training parameters θ is defined as:

$$\mathcal{L}(\mathcal{T}, \theta) = \sum_{(x,y) \in \mathcal{T}} \mathcal{L}(\theta|y, x) - \frac{1}{2\sigma^2}\|\theta\|^2 = \sum_{(x,y) \in \mathcal{T}} \log P(y|x; \theta) - \frac{1}{2\sigma^2}\|\theta\|^2. \tag{3.4}$$

The first term in Eq.(3.4) is the conditional log-likelihood on the training waves. The second term is a penalized term to prevent the L_2 norm of the model parameter $\|\theta\|$ becoming too big. It is the log of a Gaussian prior with variance σ^2 . That is, we assume the model parameter follows a normal distribution $P(\theta) \sim N(0, \sigma^2)$ to constrain $\|\theta\|$. The optimal $\theta^\star$ is learned by maximizing the objective function in Eq.(3.4), thus

$$\theta^\star = \arg\max_\theta \mathcal{L}(\theta). \tag{3.5}$$

The optimal $\theta^\star$ which maximize $\mathcal{L}$ can not be computed analytically; instead we need to employ iterative methods to estimate it.

For the evaluation of the optimal weight described by Eq.(3.5) from a set of training samples we use an iterative gradient-based optimization method. Broyden-Fletcher-Goldfarb-Shanno (BFGS) is nowadays considered the most efficient and is indisputably the most popular quasi-Newton update formula. However, if the number of the variables is very large it becomes too expensive method. For that reason a less computationally intensive method has been proposed. Liu et al. [12] first introduced Limited-memory BFGS (LBFGS) as a method that requires repeated estimations of objective function $\mathcal{L}$ and its derivatives with respect to each model parameter in θ. LBFGS method instead of storing and updating the entire inverse Hessian matrix (next search direction) it stores only the information from the past m iterations using implicitly this information for the inverse Hessian matrix requirements.

However, likewise with other hidden states models (e.g. HMM) the addendum of hidden states leads to a non convex objective function $\mathcal{L}(\theta)$ implying not always a global optimum point. Therefore we search for parameters by initializing from random start points and searching for the best local optimum.

The next step is to describe an effective way to calculate the gradient of $\mathcal{L}(\theta)$. Denote the log-likelihood of the training set as

$$\mathcal{L}(\theta) = \log P(y|x, \theta) = \log \frac{\sum_h \exp \Psi(y, h, x; \theta)}{\sum_{y' \in \mathcal{Y}} \sum_h \exp \Psi(y', h, x; \theta)}. \tag{3.6}$$

The calculation of the derivatives would be really time consuming because if we have m features there are $|\mathcal{H}|^m$ possible h. For the avoidance of such situation we use a belief propagation (BP) algorithm to calculate marginal probabilities and their normalization term efficiently.

$$\forall y \in \mathcal{Y}, \quad Z(y|x,\theta) = \sum_s e^{\Psi(y,s,x;\theta)}, \tag{3.7}$$

$$\forall y \in \mathcal{Y}, \forall j \in V, \forall \alpha \in \mathcal{H}, \quad P(h_j = \alpha|y,x,\theta) = \sum_{\mathbf{h}:h_j=\alpha} P(\mathbf{h}|y,x,\theta), \tag{3.8}$$

$$\forall y \in \mathcal{Y}, \forall (j,k) \in E, \forall \alpha \in \mathcal{H}, \forall b \in \mathcal{H}, \quad P(h_j = \alpha, h_k = b|y,x,\theta) = \sum_{\mathbf{h}:h_j=\alpha,h_k=b} P(\mathbf{h}|y,x,\theta). \tag{3.9}$$

Eq.(3.7) defines a normalization term $Z(y|x,\theta)$ that sums over all possible h. Eq.(3.8) defines a marginal probability over an individual variable h_j . Eq.(3.9) defines a marginal probability over pairs of variables h_j and h_k, which correspond to edges in graph G.

The first derivatives of Eq.(3.6) with respect to each parameter θ_{k_1}, θ_{k_2} and $\theta_{k_3}$with the use of BP algorithm are:

$$\begin{aligned}
\frac{\partial \mathcal{L}(\theta|y,x)}{\partial \theta_{k_1}} &= \sum_{h\in\mathcal{H}} \left\{ \frac{\exp \Psi(y,h,x;\theta)}{\sum_t \exp \Psi(y,s_t = h_k,x;\theta)} \frac{\partial \Psi(y,h,x;\theta)}{\partial \theta_{k_1}} \right\} \\
&\quad - \sum_{y'\in\mathcal{Y}} \sum_{h\in\mathcal{H}} \left\{ \frac{\exp \Psi(y',h,x;\theta)}{\sum_{y'\in\mathcal{Y}} \sum_{h\in\mathcal{H}} \exp \Psi(y',h,x;\theta)} \frac{\partial \Psi(y',h,x;\theta)}{\partial \theta_{k_1}} \right\} \\
&= \sum_{h\in\mathcal{H}} \sum_{j\in V} P(y,h_j = s|x;\theta) \cdot \phi_{k_1}(x_j,h_j) \\
&\quad - \sum_{y'\in\mathcal{Y}} \sum_{h\in\mathcal{H}} \sum_{j\in V} P(h_j = s|y',x;\theta) \cdot \phi_{k_1}(x_j,h_j) \\
&= g_{k_1}(y,h,x;\theta).
\end{aligned} \tag{3.10}$$

$$\begin{aligned}
\frac{\partial \mathcal{L}(\theta|y,x)}{\partial \theta_{k_2}} &= \sum_{s\in\mathcal{H}} \sum_{j\in V} P(y,h_j = s|x;\theta) \cdot \phi_{k_2}(y,h_j = s) \\
&\quad - \sum_{y'\in\mathcal{Y}} \sum_{s\in\mathcal{H}} \sum_{j\in V} P(h_j = s|y',x;\theta) \cdot \phi_{k_2}(y',h_j = s) \\
&= g_{k_2}(y,h,x;\theta)
\end{aligned} \tag{3.11}$$

$$\frac{\partial \mathcal{L}(\theta|y,x)}{\partial \theta_{k_3}} = \sum_{s\in\mathcal{H}}\sum_{s'\in\mathcal{H}}\sum_{(i,j)\in E} P(y, h_j = s, h_i = s'|x;\theta)\cdot\phi_{k_3}(y, h_j = s, h_i = s')$$
$$- \sum_{y'\in\mathcal{Y}}\sum_{s\in\mathcal{H}}\sum_{s'\in\mathcal{H}}\sum_{(i,j)\in E} P(h_j = s, h_i = s'|y', x;\theta)\cdot\phi_{k_3}(y', h_j = s, h_i = s')$$
$$= g_{k_3}(y, h, x;\theta) \tag{3.12}$$

Here $s \in \mathcal{H}$ is a hidden state and the $\sum_{s\in\mathcal{H}}$ is the summation of all possible states of h_j at site j, $j \in V$.

Using Eq. (3.8) and (3.9) in gradients above (Eq. (3.10),(3.11),(3.12)) we can say that all four probabilities $P(y, h_j = s|x;\theta), P(h_j = s|y', x;\theta), P(y, h_j = s, h_i = s'|x;\theta), P(h_j = s, h_i = s'|y', x;\theta)$ can be calculated in a time that grows only linearly with the number of part labels.

Several works using HCRFs have been implemented in speech or action recognition, in ECGs classification of different types of heart beats.

3.3 Methodology for percentile Automatic Detection

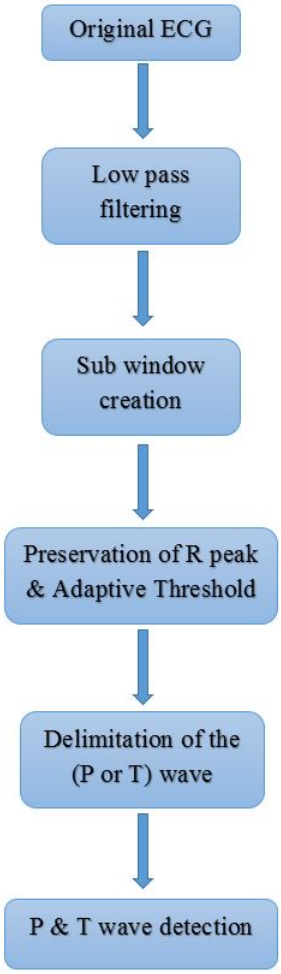

Figure 3.3: Block diagram representation of the pAD method for P & T wave detection

A first approach in percentile based Automatic Detection for P & T wave is proposed. The block diagram of the proposed method is shown in the fig. 3.3. The detailed description of the proposed method is given bellow.

3.3.1 Filtering

Initially, the signals are pre-processed to eliminate the undesirable frequencies (parasitic, noises). To achieve that, a moving average filter was implemented to smooth data by replacing each data point with the average of the N neighboring data points. This process

is equivalent to lowpass filtering and is given by the difference equation:

$$y_s(i) = \frac{1}{2N+1}(y(i+N) + y(i+N-1) + .. + y(i-N)) \tag{3.13}$$

where $y_s(i)$ is the smoothed value for the i_{th} data point and N is the number of neighboring data points (in our case $N = 5$). Fig. 3.4 points out the original ECG and the result obtained after filtering which is a smooth response to the original data.

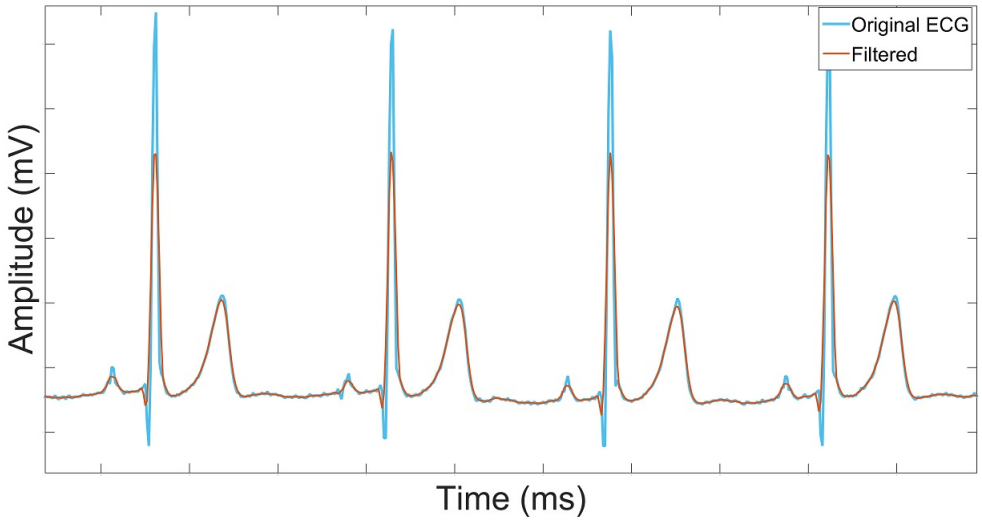

Figure 3.4: Original & Filtered ECG

3.3.2 Preservation of R peak & Adaptive Threshold

The signal has been divided into buckets of 2000 values as has been discussed in Manual Detection (see Chapter 2) to reduce baseline drift. The detection of P and T waves, with respect to R peak, would be easier if R peaks were preserved unaffected by the filter. Thus, the location and the value of R were used from the original signal and not from the filtered one.

The peak of each wave has been detected using as threshold the percentile. The k_{th} percentile of the signal is the value below which k percent of the observations may be found. It has been noticed that the intervals $I_T = [90th\ percentile \pm 0.3\ mV]$ and $I_P = [65th\ percentile \pm 0.2\ mV]$ contain a set of values and among them the peaks of T

and P-wave respectively. Hence, the values that are inside those limits of percentile will be selected including some other values of noise or QRS complex.

3.3.3 Delimitation of the wave

For the detachment of the peaks, the set S must be found:

$$S = \{x : f(x) \in I_T \ (or \ I_P)\}, \tag{3.14}$$

where $f(x)$ are ECG values and x their locations. The set S has to be divided into subsets S_i, where each contains consecutive elements. Next, maximum values for subsets, $M = \{maxS_i\}$ were found which are acceptable only if they were not the first or last element of S_i, in order to avoid points of QRS complex. For each R, the peak of M which is $(35 * 7.8125)msec = 0.2734sec$ after R or $(25 * 7.8125)msec = 0.1953sec$ before R was kept. The first interval denotes the T wave while the latter the P wave respectively. Those intervals preserve the known P and T waves time interval.Finally, the set $M' = \{maxS_i\}$ consists of the location of the peaks.

For each peak we move to the right until the gradient stops to be negative (assuming there is no noise) and we consider this point the end (offset) of the wave. The start (onset) is the point on the left part of the wave having the closest ordinate with the ending point.

This approximation has the opportunity to select the waves that are inside an area defined by percentile values with an excellent result for both P and T waves. However, in most cases an ECG signal has a shifted baseline across the time (change in mV). For that, a new method has been created in order to achieve better results defying those shifts among other things that will be discussed below.

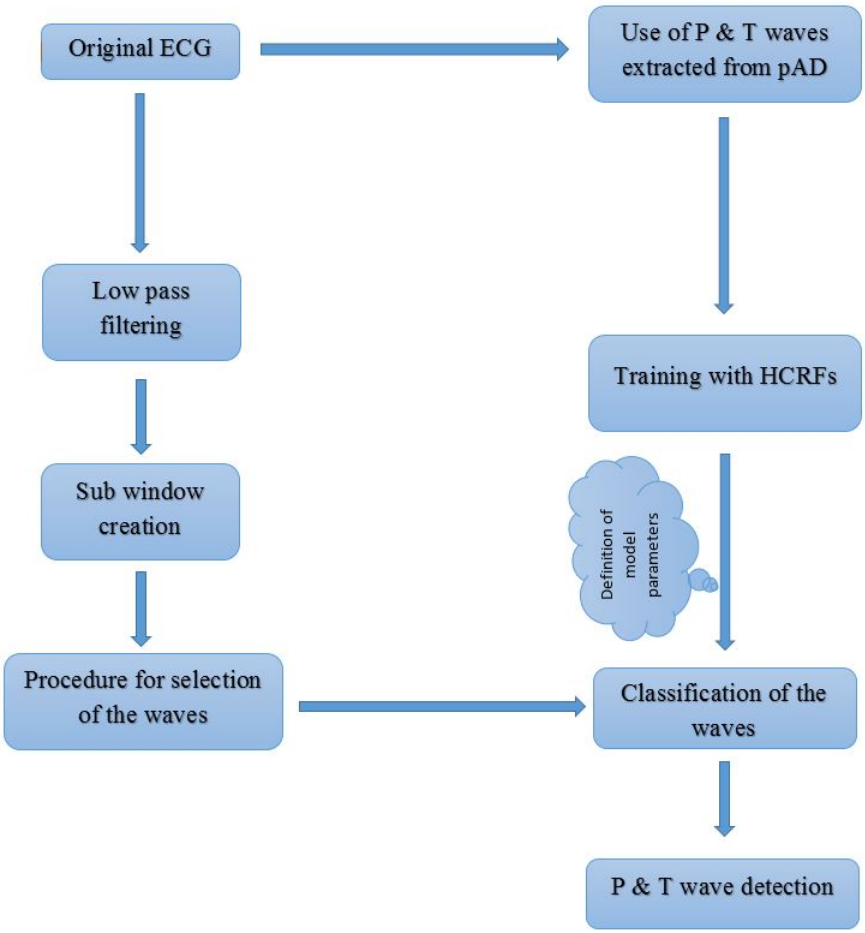

Figure 3.5: Block diagram representation of the proposed method for P & T wave detection

3.4 Methodology for Graphical based Automatic Detection

From the previous process several waves have been extracted both P and T. Due to the thresholds that have been used in the first automatic we had the chance to collect an adequate number of waves. However, this was the initial step for our research. The next goal is to isolate as much as possible waves from the ECG.

The way that the Graphical based Automatic Detection has been implemented is shown in fig. 3.5.

3.4.1 Use P & T waves extracted from pAD

The main idea was to use a set of 200 P waves and 200 T waves randomly selected from every patient separately. The waves as have been mentioned before are extracted from pAD algorithm. Morphology of the waves provides a lot of information that seems to be a good separator. Hence, four features (area, height, left slope and right slope) are estimated based on those waves. Further details for the features will be discussed in Chapter 4.

3.4.2 Training with HCRF

Given a training set $(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)$, where x_i denotes a vector with the four features described above and y_i is the class label for every wave (P or T), the HCRF model trains the model parameter θ. Several numbers of hidden states h were used in order to find the lower possible without disrupting the success rate of the algorithm. In our case 4 is the ideal number for hidden states. The goal of the model targets in keeping weights of edges (θ) higher for correct assignments of labels (possessing higher probabilities), and lower for the incorrect assignments (possessing lower probabilities). In brief, only the important edges will determine the output (label) for the model. Therefore, even if we use more hidden states the final estimated θ can reject all the unnecessary states.

For the estimation of θ the objective function is defined in Eq. 3.4 as the difference of conditional log-likelihood term and a penalty term to avoid overfitting. The maximization of this objective function will output the desirable θ. Generally, the maximization of a function is achieved by the following steps:

- denote the partial derivatives of the function and set it equal to 0
- solve the above equations in order to find critical points
- those critical points (belong in domain) for which the second derivative is negative, are the local maximum.

In our case the calculation of these derivatives are so hard to be calculated analytically. So, the The Belief Propagation (BP) is used (Eq. 3.10, 3.11, 3.12) as an approach. In order to find the optimal parameters $\theta^{\star}$ (see Eq. 3.5) the L-BFGS method is used, which requires the gradient of the first term (Eq. 3.4) with respect to each parameter.

As soon as the training phase is over the optimal parameters $\theta^{\star}$ will form the model which will be used in testing phase of classification.

3.4.3 Procedure for selection of the waves

The first two steps before recording the waves is the same as discussed in previous sections. Thus, a moving average filter is applied all over the ECG signal which is separated into buckets.

In the proposed method the difference between adjacent values is calculated in order to find all the ascending and descending values as an approximation of first derivative. The objective is to select both P and T waves. For this purpose three stages are implemented in the algorithm.

First stage

Determination of the start-to-peak of the waves. The rise on the values of differences does not represent only the waves start points but also noise. For that reason a second statement is implemented to dispose every noise. If the rise differences exceed 4 continuous values and until there are three declines in differences (a.k.a. variability), we denote these values as the first half of the wave (onset to peak). Fig. 3.6 denotes the results from the First Stage.

Second stage

It is obvious that in First stage QRS complex is also selected. Hence, to reject these errors the left fitting slope is calculated and used as a constrain. The range of the slope will be in a specific area. The minimum threshold value is defined as the minimum of left fitting slope from both P and T waves for all patients, whereas the maximum threshold value as the maximum left fitting slope. Fig. 3.7 depicts

the results from the Second Stage.

Third stage

To fill the other half in the remaining waves an addendum with the half length of the current wave is implemented to locate the maximum value. From this value we will move to the right until variability is less than three values. Fig. 3.8 shows the results from the Third Stage.

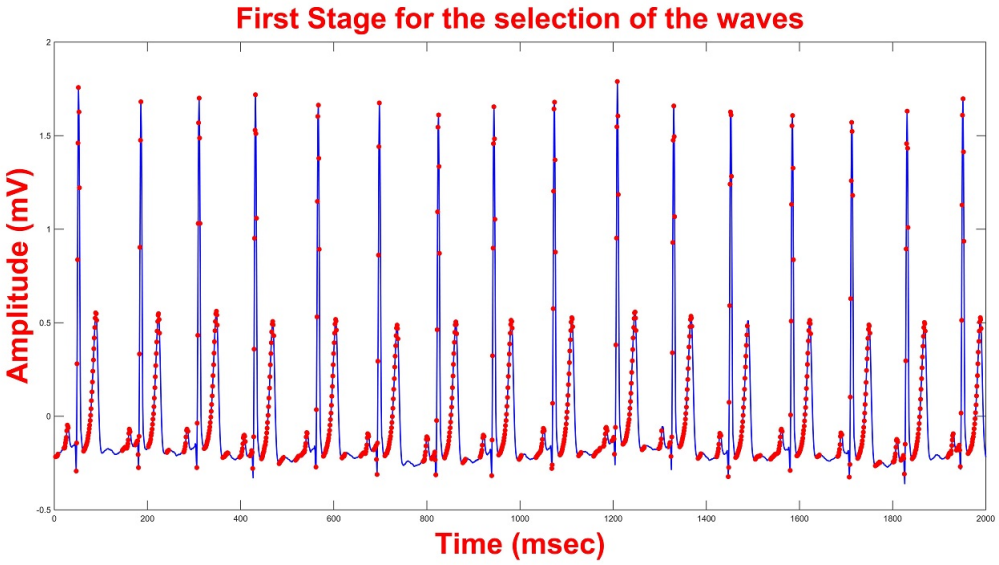

Figure 3.6: The First stage in the procedure for the selection of the waves

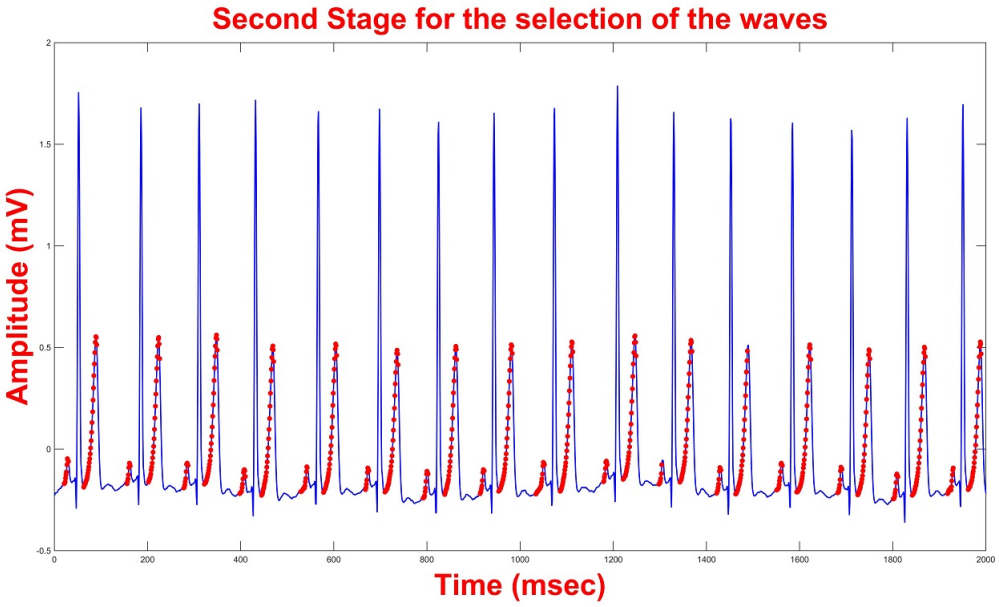

Figure 3.7: The Second stage in the procedure for the selection of the waves

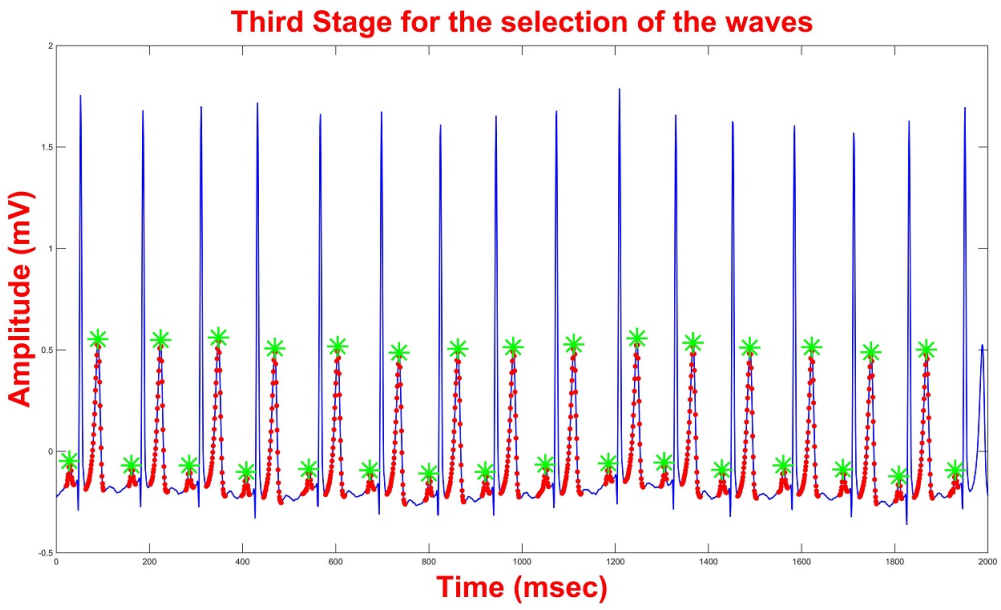

Figure 3.8: The Third stage in the procedure for the selection of the waves

3.4.4 Classification

In the classification phase from each wave the same four features are calculated as denoted in training phase. Those features are imported in the HCRF algorithm with the corresponding weights extracted from learning phase and the output denotes the class label where they belong (P or T waves).

3.4.5 Check for false records

As we may consider there is a possibility where the GlAD will enter a false wave or will miss in the classification stage. Therefore, a check for false records is implemented in every patient respectively. The criterion for this is the R peak annotation. The results show a $85\% - 97\%$ success rate in subjects with low noise ratio whereas $70\% - 80\%$ success rate in more noisy signals almost equally for P and T waves. The intuition of these percentages (e.g. in 97%) is to declare that among $100K$ R peaks the $97K$ are detected correctly. The other $3K$ may not be even present nor can be detected (e.g. variability, odd morphology).

3.5 Results

In order to evaluate the performance, the proposed algorithm was tested using MIT-BIH Normal Sinus Rhythm database. The pAD algorithm is able to detect both P and T waves as shown in Fig. 3.9 and 3.10 respectively.

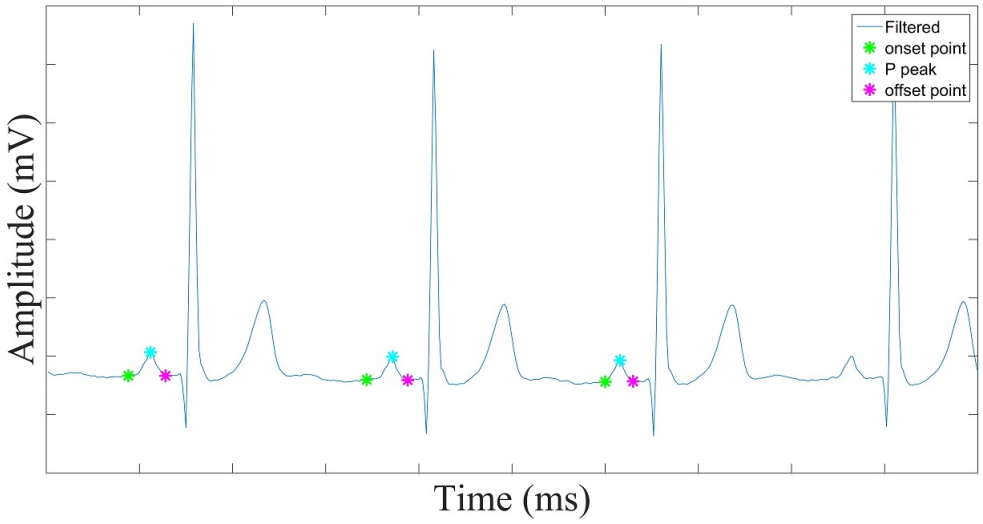

Figure 3.9: P wave Detection using pAD

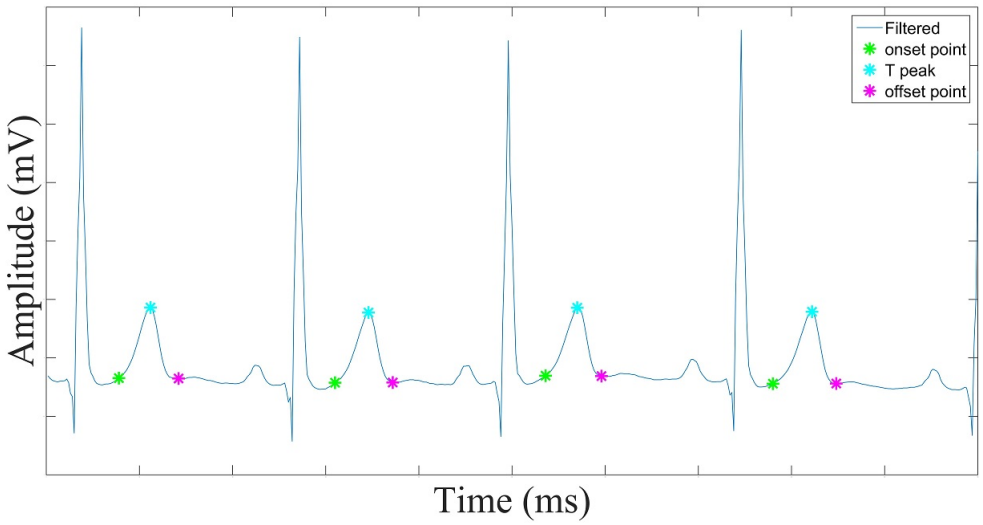

Figure 3.10: T wave Detection using pAD

For the GlAD algorithm an example of the results is shown in fig. 3.11.

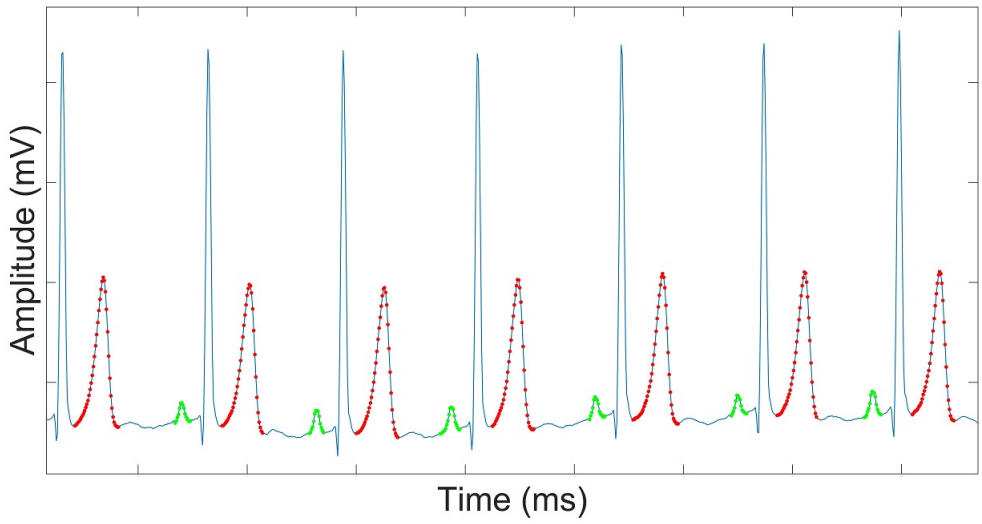

Figure 3.11: P and T wave Detection using GlAD

The total number of selected waves for both methods are shown in Table 3.1. As we may notice all P waves from pAd algorithm are less than P waves from GlAD algorithm. In the majority of T waves the results are also better in GlAD than in pAD. The cases where T waves are less in second method are because P and T waves don't differ much and that leads to miss-classification. It is worth to mention that the results of this table show only the correct waves.

Table 3.1: Total number of selected waves

Patients	pAD		GlAD	
	P-wave	T-wave	P-wave	T-wave
16265	40750	45735	77226	88728
16272	34529	47859	38397	32450
16273	34473	31194	69669	78351
16420	10256	48185	69695	66254
16483	55872	57013	97105	95334
16539	23185	51229	63752	61663
16773	40964	39231	44572	42628
16786	50609	60555	94747	99008
16795	20912	47167	66253	65458
17052	39109	142	59290	3173
17453	22945	26530	67457	69313
18177	39405	44993	86504	23555
18184	25583	36931	68329	76334
19088	3681	45140	58787	28827
19090	24060	60745	76291	77974
19093	53507	59771	60255	63873
19140	9316	41690	77213	86136
19830	10880	11949	24708	76440
Mean	30002	42003	66681	63083

Chapter 4

Feature Extraction

The selection of features aims to illustrate that sufficient information can be obtained, not only from the entire wave, but also from part of it. Arsenos et al. [13] and Zeraatkar et al. [14] have used some conventional ones before. A significant advantage of this approach is the independence between the accuracy of metrics and the accuracy of onset and offset points.

The detection of P and T waves from ECG signal prepares the field to extract necessary descriptors or features from these parts of signal. In this work 30 features have been

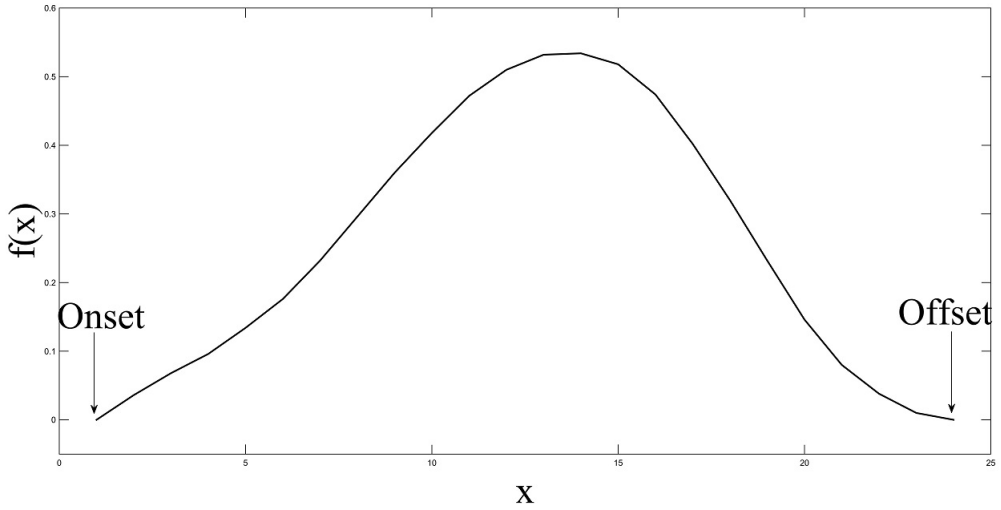

Figure 4.1: Sample of T wave

considered and extracted, the same for both waves. There are 15 features characterizing the morphology of the wave and 15 referring to their ratio. By morphology the author means the peak (maximum value), minimum value, height or duration of the wave, as well as, the global area and the semi-areas such as left or right area, upper or down area and upper left or upper right area. There are also left or right slope given from onset point to peak and from peak to offset point and the slope that fits better to data points.

For simplicity, the set $wave$ was defined

$$wave = \{f(x_i) : x_i \in \mathscr{D}\}, i = 1, ..., n \tag{4.1}$$

where $P_i = (x_i, f(x_i))$ corresponds to i_{th} point and $\mathscr{D} = [onset\ point,\ offset\ point]$ (fig. 4.1). Each feature is described analytically bellow:

4.1 Peak - Height - Duration

- Feature 1: Maximum value or Peak (fig. 4.2) of the wave

$$Max = max(f(x_i)) \tag{4.2}$$

- Feature 2: Minimum value of the wave

$$Min = min(f(x_i)) \tag{4.3}$$

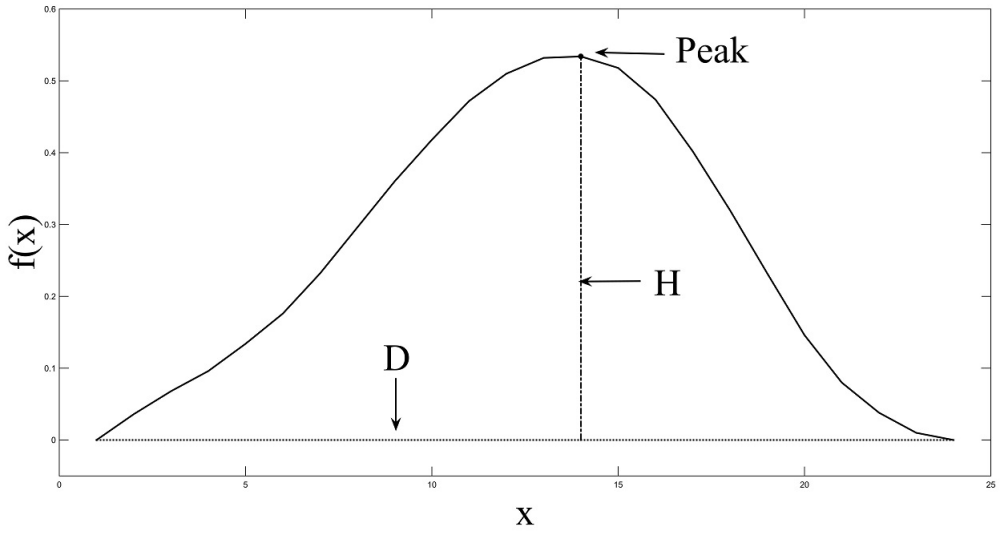

Figure 4.2: Peak, height & Duration

- Feature 3: Amplitude or Height (fig. 4.2) is the distance from minimum to Peak

$$H = Max - Min \tag{4.4}$$

- Feature 4: Duration (fig. 4.2) of the wave

$$D = x_n - x_1 \tag{4.5}$$

4.2 Global Area

In order to calculate areas, the wave must be shifted, in such way that the minimum value of the wave falls on the x axis.

- Feature 5: Global Area (fig. 4.3) is defined as the area under the curve formed from

the onset to the offset point of the wave

$$A_G = \sum_{i=1}^{n-1} \int_{x_i}^{x_{i+1}} (\alpha(x - x_i) + f(x_i))\, dx$$
$$and\ \alpha = \frac{f(x_{i+1}) - f(x_i)}{x_{i+1} - x_i}, \tag{4.6}$$

where the amount in the integral is the line formed between two consecutive points P_i and P_{i+1}. The α is the slope of this line.

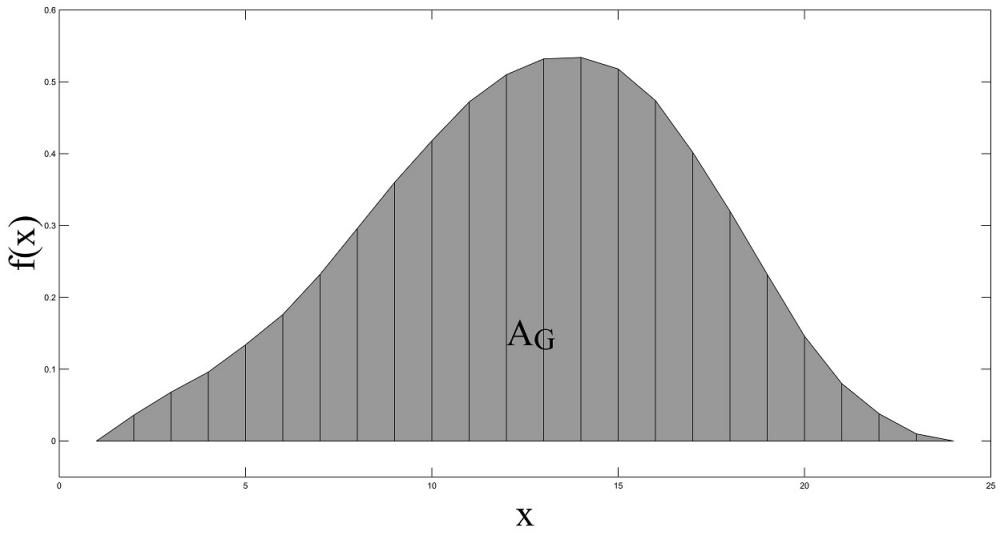

Figure 4.3: Global Area

4.3 Left & Right Area

- Feature 6: Left Area (fig. 4.4) is defined as the area under the curve formed from the onset to the Peak of the wave

$$A_L = \sum_{i=1}^{p-1} \int_{x_i}^{x_{i+1}} (\alpha(x - x_i) + f(x_i))\, dx, \tag{4.7}$$

and P_p are the coordinates of the Peak.

- Feature 7: Right Area (fig. 4.4) is defined as the area under the curve formed from

the Peak to the offset of the wave

$$A_R = \sum_{i=p}^{n-1} \int_{x_i}^{x_{i+1}} (\alpha(x - x_i) + f(x_i))\, dx, \tag{4.8}$$

and P_p are the coordinates of the Peak.

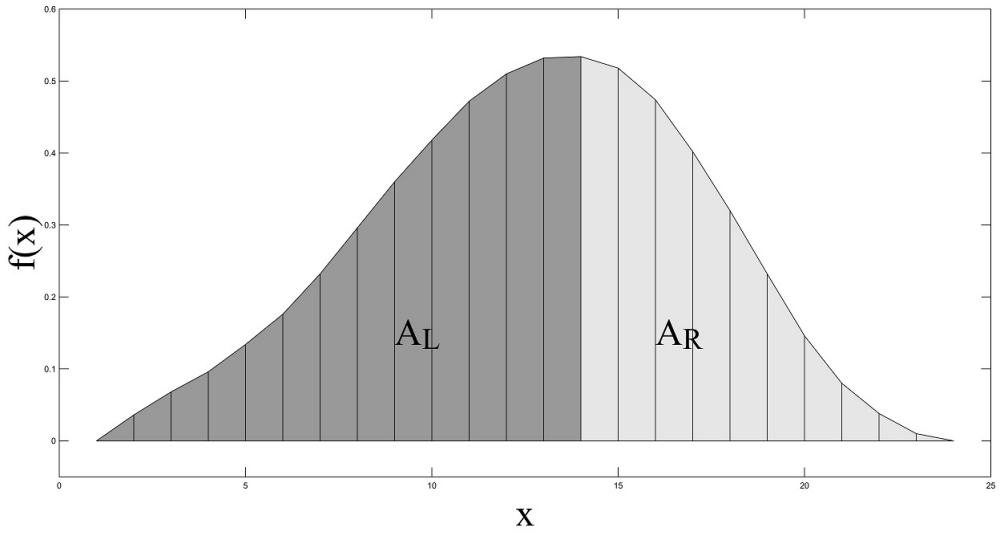

Figure 4.4: Left & Right Area

4.4 Upper & Down Area

For the next four features the wave has to be divided horizontally by a line. This line is defined as the horizontal line vertical to the height (vector) and passing through the middle (of the height). The intersection points between the line and the wave were determined as points with the shortest distance from the line, P_l and P_r (left and right) with respect to the Peak.

- Feature 8: Upper Area (fig. 4.5) is defined as the area under the curve formed from the P_l point to the P_r point

$$A_U = \sum_{i=l}^{r} \int_{x_i}^{x_{i+1}} (\alpha(x - x_i) + f(x_i))\, dx, \tag{4.9}$$

shifted according to the $min(f(x_l), f(x_r))$.

- Feature 9: Down Area (fig. 4.5) is defined as the area under the curve formed from the onset to the P_l point, from the P_l to the P_r point (below) and from the P_r to the offset point

$$
\begin{aligned}
A_D &= A_1 + A_2 + A_3, \\
A_1 &= \sum_{i=1}^{l-1} \int_{x_i}^{x_{i+1}} (\alpha(x - x_i) + f(x_i))\, dx, \\
A_2 &= \int_{x_l}^{x_r} \left(\frac{f(x_r) - f(x_l)}{x_r - x_l}(x - x_l) + f(x_l) \right) dx, \\
A_3 &= \sum_{i=r}^{n-1} \int_{x_i}^{x_{i+1}} (\alpha(x - x_i) + f(x_i))\, dx
\end{aligned}
\tag{4.10}
$$

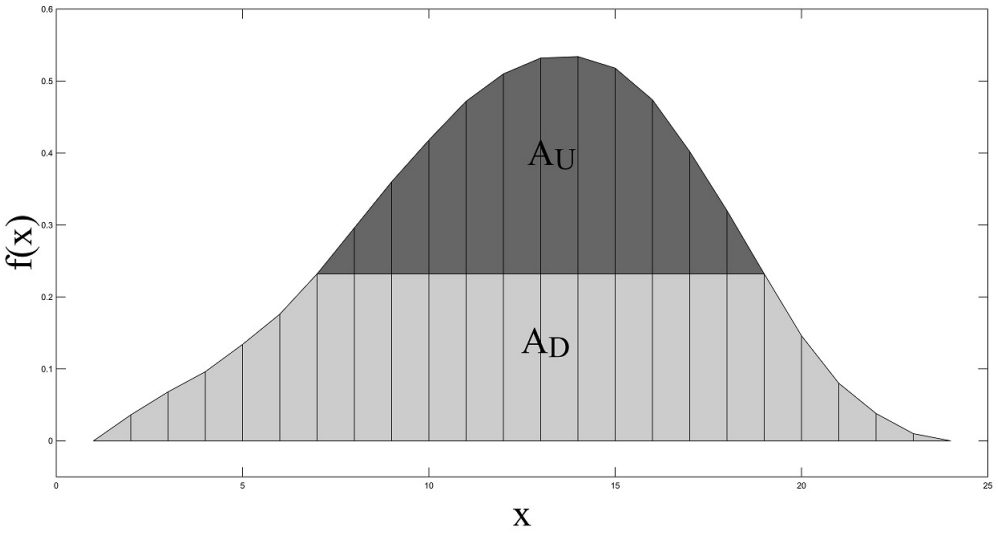

Figure 4.5: Upper & Down Area

4.5 Upper Left & Right Area

- Feature 10: Upper Left Area (fig. 4.6) is defined as the area under the curve formed from the P_l to the Peak

$$A_{UL} = \sum_{i=l}^{p-1} \int_{x_i}^{x_{i+1}} (\alpha(x - x_i) + f(x_i))\, dx, \tag{4.11}$$

  shifted according to the $f(x_l)$.

- Feature 11: Upper Right Area (fig. 4.6) is defined as the area under the curve formed from the Peak to the P_r point

$$A_{UR} = \sum_{i=p}^{r-1} \int_{x_i}^{x_{i+1}} (\alpha(x - x_i) + f(x_i))\, dx, \tag{4.12}$$

  shifted according to the $f(x_r)$.

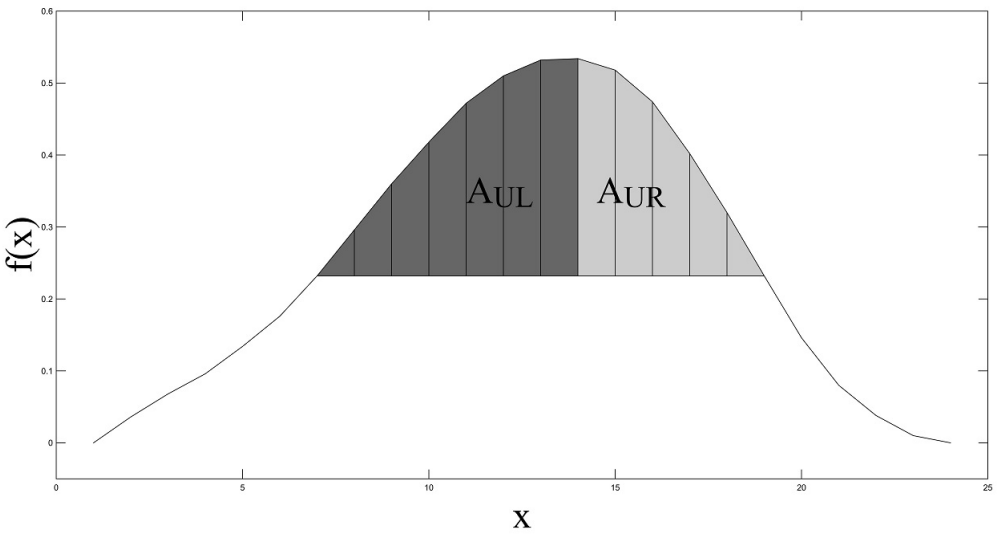

Figure 4.6: Upper Left & Right Area

4.6 Left & Right Slope

- Feature 12: Left Slope (fig. 4.7) is defined as the slope of the line formed from P_1 and the Peak

$$S_L = \frac{f(x_p) - f(x_1)}{x_p - x_1} \tag{4.13}$$

- Feature 13: Right Slope (fig. 4.7) is defined as the slope of the line formed from Peak and the P_n

$$S_R = \frac{f(x_n) - f(x_p)}{x_n - x_p} \tag{4.14}$$

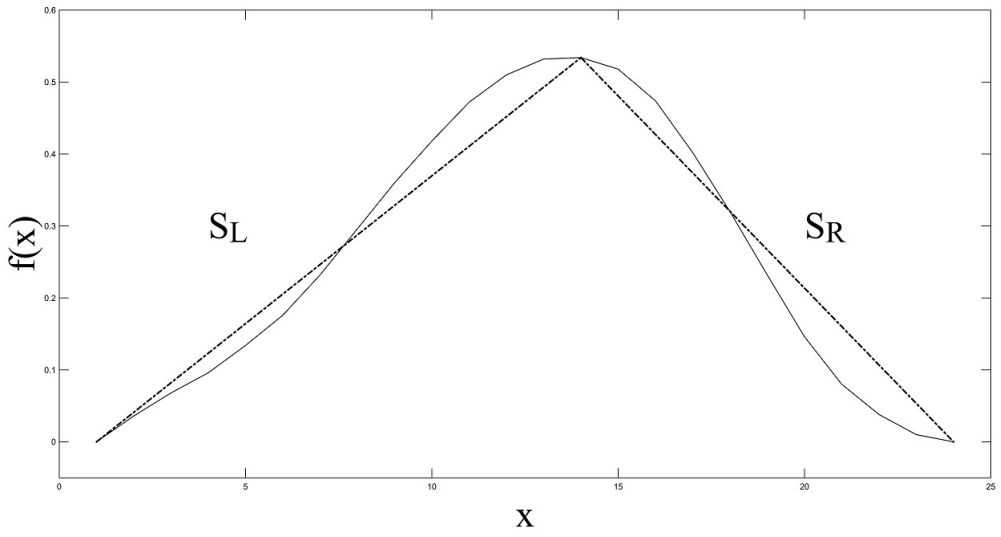

Figure 4.7: Left & Right Slope

4.7 Left & Right Fitting Slope

- Feature 14: Fitting Left Slope (fig. 4.8) is defined as the slope of the straight line which would provide a *best* fit for the data points $P_i,\ i = 1, ..., p$. This line has to minimize the sum of squared residuals of the linear regression model $min_{\alpha,\beta} Q(\alpha, \beta)$

$$for\ \ Q(\alpha, \beta) = \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \sum_{i=1}^{n} (f(x_i) - \beta - \alpha x_i)^2 \tag{4.15}$$

where β is the y-intercept and α is the slope F_L.

- Feature 15: Fitting Right Slope (fig. 4.8) is defined as the slope of the straight line which would provide a *best* fit for the data points $P_i,\ i = p, ..., n$. The α is the slope F_R.

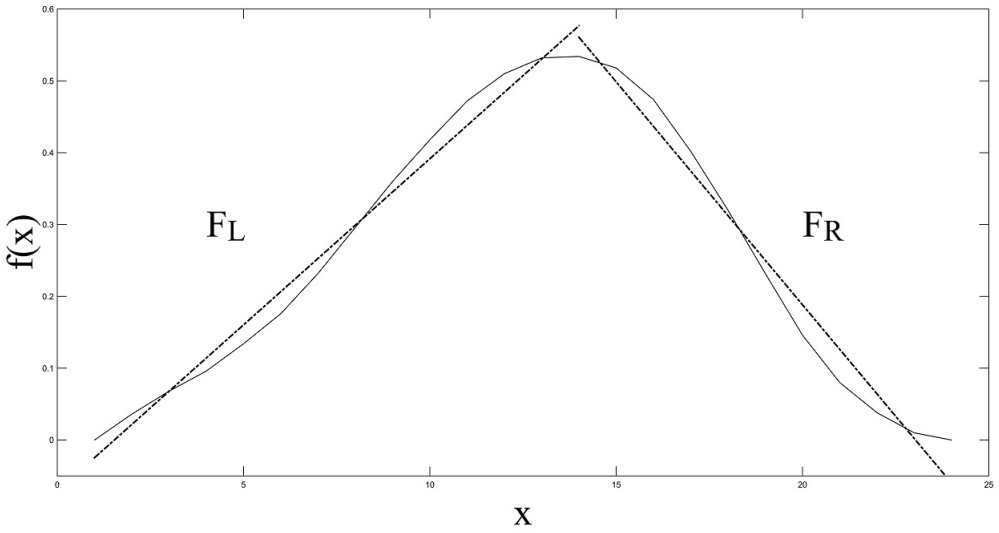

Figure 4.8: Left & Right Fitting Slope

4.8 Ratios of Features

The other 15 features are some of the ratios:

$$\frac{A_L}{A_G},\ \frac{A_R}{A_G},\ \frac{A_L}{A_R},\ \frac{A_U}{A_G},\ \frac{A_D}{A_G},\ \frac{A_U}{A_D},\ \frac{A_{UL}}{A_L},\ \frac{A_{UR}}{A_R},$$

$$\frac{A_{UL}}{A_{UR}},\ \frac{A_{UL}}{A_U},\ \frac{A_{UR}}{A_U},\ \frac{A_{UL}}{A_G},\ \frac{A_{UR}}{A_G},\ \frac{S_L}{S_R},\ \frac{F_L}{F_R}.$$

Chapter 5

Discrimination of day-night periods

5.1 Introduction

Many researchers studied P or T waves for certain time periods or for the whole 24 hour recording. Neyroud et al. [15] concluded that there is an influence of day and night in ventricular repolarization behavior. Braga et al. [16] and Ramirez et al. [17] studied QT/RR differences between day and night for both genders. Dilaveris et al. [18] presented the diurnal pattern of P-wave duration, P area, and PR interval.

The first objective of this thesis is to examine the discrimination of day-night periods based on P and T waves. New features have been employed, which have been discussed in Chapter 4, for examining P and T wave's morphology.

Paired t-test is used for the evaluation of the values extracted from features; the significance level is set at $\alpha = 0.05 = 5\%$. A lower value for $p - value$ expresses the

discrimination of the waves between the two opposing periods that will be discussed in next section.

5.2 Procedure for paired t-test

The paired t-test is used to compare two population means or medians. In our case, we have two samples day and night in which observations in first sample can be paired with observations in the second sample. Suppose a sample of n waves during the day (1:00-3:00 AM) and n waves during the night (1:00-3:00 PM) similarly for P or T waves.

Let x =daytime waves and y = nighttime waves. We calculate the difference between the two observations on each pair as $d_i = y_i - x_i$, making sure we distinguish between positive and negative differences.

There are three assumptions must be checked:

- The sample is randomly selected.
- The percentage of extreme outliers in available observations d_i is no more than 10%.
- The paired t-test does not assume that observations within each group are normal, only that the differences are normally distributed; the Shapiro Wilk test is used in order to test the normality of our data.

The null hypothesis is that the mean difference between paired observations is zero. When the mean difference is zero, the means of the two groups must also be equal.

We can use the results from our sample of waves to draw conclusions about the impact of this discrimination in general.

5.3 Results

After implementing the paired t-test the p-values of manual selection, for both waves, are all significantly lower for all features ($p < 0.00001$) except A_{UL}/A_G. Also, p-values of pAD give for every feature significantly lower value ($p < 0.00001$) except Min ($with\ p = 0.0047\ for\ P\ waves$) and appeared to be better, compared to manual selection, due to larger amount of data analyzed. The majority of investigated features for both manual and pAD have been proved rich descriptors for heart performance.

The correlation coefficient (CC) and the corresponding p-value for T and P waves collected by the pAD, are presented in Table 5.1. In this thesis the dependent variable is each feature of the wave of daytime period and the independent variable is the same feature of nighttime period. The results have shown significant dependence for these variables ($p < 0.00001$) except A_U/A_D for P waves. It is important to be mentioned that 10 features of T waves (Max, H, A_G, A_L, A_R, A_D, S_L, S_R, F_L, F_R) had approximately CC greater than 0.7, which indicate a strong positive linear relationship via a firm linear rule and predicts 50% of the variance in the independent variable.

Table 5.1: Results for Association

Features	T-wave	P-wave	Features	T-wave	P-wave
	CC	CC		CC	CC
Max	0.6956	0.4144	A_L/A_G	0.2793	0.3777
Min	0.5185	0.3134	A_R/A_G	0.2767	0.3762
H	0.7895	0.4140	A_L/A_R	0.3343	0.3944
D	0.2572	0.2449	A_U/A_G	0.3920	0.2474
A_G	0.8408	0.4253	A_D/A_G	0.4043	0.2508
A_L	0.8252	0.4249	A_U/A_D	0.2890	0.0036
A_R	0.8321	0.3946	A_{UL}/A_L	0.3144	0.2423
A_U	0.5384	0.2011	A_{UR}/A_R	0.3700	0.2328
A_D	0.8630	0.4064	A_{UL}/A_{UR}	0.0607	0.3567
A_{UL}	0.5265	0.2048	A_{UL}/A_U	0.0721	0.3424
A_{UR}	0.5058	0.2437	A_{UR}/A_U	0.0728	0.1313
S_L	0.7286	0.2762	A_{UL}/A_G	0.3426	0.1615
S_R	0.7431	0.3803	A_{UR}/A_G	0.3453	0.3102
F_L	0.6957	0.2652	S_L/S_R	0.2476	0.2036
F_R	0.7495	0.3783	F_L/F_R	0.3510	0.1714

Chapter 6

Classification

6.1 Introduction

For the discrimination capability of day-night periods for P or T wave several methods of classification have been tested.

The naive Bayes classifier or simple Bayesian classifier [19], is a classifier built upon the Bayes' theorem. It is essentially a simple Bayesian Network (BN) and particularly suitable for the case when the dimensionality of the inputs is high.

K-Nearest Neighbor (KNN) method [20] has been used in applications such as recognition of handwriting, data mining, statistical pattern recognition, ECG disease classification and image processing. This work is primarily motivated by the desire to create an algorithm for accurate and precise delineation of P & T waves between two time periods.

Decision trees [21] are among the most livable medical decision support models, which are already successfully used for many medical decision making purposes. The goal is to learn how to classify objects by analyzing a set of instances whose classes are known.

The technique of SVM [22] is a powerful, widely used technique for solving supervised classification problems due to its generalization ability. Basically, SVM classifiers maximize the margin between training data and the decision boundary, hence the optimal hyperplane that separates them. Maximization problem can be formulated as a quadratic optimization problem in a feature space. Support vectors are called subset of examples (patterns) which are closest to the decision boundary.

6.2 Naive Bayes

The naive Bayes algorithm [19] depends on a strong hypothesis; the value of any feature is independent of the existence of any other feature. Generally, in most of the real life examples, the naive Bayes hypothesis is never satisfied but the algorithm predicts with a good enough accuracy the classes.

Assume that a set of samples $x_i,\ i = 1, ..., K$ is given with their associated class labels $c_{x_i} \in \Omega = \{c_1, c_2, ..., c_L\}$. Further assume that the samples have n features denoted as $z_j, j = 1, ..., n$. The task is to use the samples to learn a naive Bayes model that will predict the label c_x for any future sample x.

A general BN classifier, which uses the Bayes rule to compute the posterior of classification variable c based on the feature variables $z_j, j = 1, ..., n$, can be described as follows:

$$p(c|z_1, z_2, ..., z_n) = \frac{p(z_1, z_2, ..., z_n|c)p(c)}{p(z_1, z_2, ..., z_n)} \tag{6.1}$$

By default, the MATALB software models the predictor distribution within each class using a Gaussian distribution having some mean and standard deviation.

6.3 K Nearest Neighbor

KNN [20] is a classification method based on closest training samples. It is an instance-based learning algorithm that, instead of performing explicit generalization, compares new problem instances with stored in memory instances seen in training. It is called instance-based because it constructs hypotheses directly from the training instances themselves. The classification is performed by finding the minimum distance from a data set which contains the training data and data set which contains the reference values. During the training phase no actual model or learning is performed, although a training dataset is required. It is used solely to fill with instances whose class is known from a sample of the search space, for this reason, this algorithm is also known as lazy learning algorithm. KNN do not perform any generalization and during the testing phase all the training data is needed. The decision of an instance, whose class is unknown, is made by computing its K closest neighbors. The majority of votes among those neighbors assign the class of that instance.

The KNN algorithm consists of two phases. The first one is training while the other is testing phase. In training phase, the training examples are vectors in a multidimensional feature space. In this phase, the feature vectors (mean of every feature from Chapter 4) and class labels of training samples are stored. In the testing phase, K is a user-defined constant. The label assigned to a test point (unlabeled vector) for the classification is the most frequent among the K training samples nearest to that query point. In other words, the KNN method compares the library of reference vectors with the input feature vector. For labeling the query point the nearest class of library feature vector is used. This way of categorizing query points based on their distance to points in a training data set is an effective and a simple way of classifying new points.

One of the advantages of the KNN method for an object classification is that it requires only few parameters to be defined: K and the distance metric. By default (in Matlab) K was assigned equal to 1 and Euclidean distance metric was used for distance. In our experiments several K values were assigned but the most preferable was 1. Euclidean distance is defined as the root of square differences between coordinates of a pair of objects and is defined as follows:

$$dist = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{6.2}$$

6.4 Decision Tree

A decision tree (DT) [21] is a tree shaped classifier which consists of nodes and edges. Root node is called the node without any incoming edge. Leaves are called the nodes which do not possess any outcoming edges. The remaining nodes are called internal nodes. To each leaf a class or a probability class is assigned. Every non-leave node represents a split regarding the input space.

Growing a classification tree faces the task of recursive partitioning the input space. The input space is commonly represented by a learning set consisting of N instances which are represented by a feature vector x and its belonging class y. For the classification of an instance x trained by a decision tree, i.e. making hypothesis on the class membership of x, the instance is spread through the tree and a class is assigned to which the leaf belongs where the instance ends up.

For learning and classification by a decision tree the author used the ECG features (see Chapter 4). More precisely, since ECG records consist of rows of each wave; several values were extracted from the attributes. The most common use is averages of each attribute in a single ECG record.

The experiment was performed with classification into two classes, where one class is daytime and the other class is nighttime. The motivation for this experiment was the

question whether it is possible to discriminate dependably between the P & T waves during the day and night. The best values were acquired with attribute average values using all features for decision tree generation.

6.5 Support Vector Machine

SVMs [22] revolve around the notion of a "margin"βέὐeither side of a hyperplane that separates two data classes. Creating the largest possible distance between the separating hyperplane and the instances on either side of it and thereby maximizing the margin has been proven to decrease the expected generalization error.

If the training data is linearly separable, then a pair, (w, b) exists such that

$$\begin{cases} w^T x_i + b \geq 1 & \text{when } y_i = +1 \\ w^T x_i + b \leq -1 & \text{when } y_i = -1 \end{cases} \tag{6.3}$$

where x_i is input, w is weight vector, b is a bias and y_i are class labels $\{-1, 1\}$ with the decision rule given by $f_{w,b} = y_i(w^T x_i + b)$.

It is easy to show that, when it is possible separating two classes linearly, an optimum separating hyperplane can be found by minimizing the squared norm of the separating hyperplane. In many practical situations, a separating hyperplane does not exist. The minimization can be achieved as a convex quadratic programming (QP) problem:

$$\underset{w,b}{\text{Minimize}}\ \Phi(x) = \frac{1}{2}\|w\|^2 \tag{6.4}$$

The solution induces to an objective function of the form:

$$f(x) = sgn\left[\sum_{i=1}^{l}(y_i\alpha_i(x.x_i) + b)\right] \tag{6.5}$$

where α_i are Lagrange multipliers. Only a small fraction of the α_i coefficients are nonzero.

The objective function is defined by support vectors which are corresponding pairs of x_i entries.

By applying the kernel function $K(x, x_i)$ instead of inner product $(x.x_i)$ the input data are mapped to a higher dimensional space. Hence, the seperating hyperplane constructed to maximize the margin is in this higher dimensional space. As long as the kernel function can be applied, the SVM will operate correctly even if the designer does not know exactly what kind of features are being used in the kernel-induced transformed feature space from the training data.

In this thesis several Kernel functions were tested and the most suitable for the features was Radial-basis function (RBF):

$$K(x, y) = \exp\left(-\frac{\|x - y\|^2}{2\sigma^2}\right) \tag{6.6}$$

6.6 10-fold Cross Validation

Cross-validation [23] is a technique to evaluate predictive models by partitioning the original sample into a train set that helps in the creation of a model and a test set to evaluate it. In K-fold cross-validation, the sample is randomly separated into K equal size subsamples. Of the K subsamples, a single subsample is kept as the validation data for testing the model, and the remaining $K - 1$ subsamples are used as training data. The process is repeated K times (with respect to the number of folds), with each of the K subsamples used exactly once as the validation data. The estimation is formed by averaging (or otherwise combining) the K results from the folds. The advantage of this method is that all observations are used for both training and validation, and each observation is used for validation one time only.

For classification problems, one typically uses stratified 10-fold cross-validation, in which the selected folds contain roughly the same proportions of class labels. We also used 2-fold cross-validation.

In repeated cross-validation, the procedure is repeated n times, yielding n random partitions of the original sample. Hence, the estimation is formed by averaging the n results.

6.7 Classification Performance

The classification performance are analyzed on 18 records of the MIT/BIH Normal Sinus Rhythm database, which includes approximately 100,000 P and 100,000 T waves of pAD algorithm to be classified into two types, day and night. The extracted features are taken for the calculation of averages which are given to the four classification methods discussed above. Two different ways are used in order to show the discrimination: 2-Fold Cross-Validation and 10-Fold Cross-Validation for every method.

Table 6.1: Results for Classification

	T-wave		P-wave	
	C_M	C_A	C_M	C_A
2-fold NB	91.805	90.485	82.924	80.696
10-fold NB	91.009	89.658	83.774	80.736
2-fold KNN	91.671	94.500	81.934	87.784
10-fold KNN	91.973	94.987	84.863	87.801
2-fold DT	90.152	93.383	84.865	85.456
10-fold DT	91.681	93.128	83.127	85.943
2-fold SVM	95.011	95.542	85.389	89.451
10-Fold SVM	93.592	94.881	89.039	89.153

Chapter 7

Circadian Rhythm

7.1 Introduction

Several circadian rhythm studies can be found in the literature. Many of them [24–27] reported differences in RR intervals and Heart Rate Variability (HRV). They all showed a significant circadian variation in healthy subjects. In order to show that the annotation of P and T peaks in the GlAD method worked well, the PP and TT intervals are created. In literature the only interval have been used is the RR due to the visible peaks that differ all over the signal. The annotation in P and T peaks is not an easy task as it is described in previous Chapter.

The next goal of this thesis is to determine that the waves extracted from every patient and by extension the peaks of P and T waves are more than enough for the research. So, we want to check if the PP and TT intervals are following the same Circadian rhythm as RR interval does. Thus, PP or TT interval is defined the difference between two adjacent

peaks. The expected result is to show that the GlAD is a well characterized method for detecting P and T waves that can perform in the same way the recognition of a Circadian rhythm through those intervals.

After showing that the intervals revealed a Circadian rhythm, the same task has been implemented in features extracted from all over the signal.

7.2 Procedure on PP and TT intervals and features

Initially, as it has been mentioned in section 3.4.5 a check for false waves has been preceded. The sets of waves are seperated in 2 time windows: every 30 minutes and every 1 hour. From every set the mean is computed as well the standard deviation (SDNN). Even if the standard deviation of any interval is usually a single number over the whole signal, we calculate SDNN for each time separately using the Eq. 7.1.

$$SDNN = \sqrt{\frac{1}{N-1}\sum_{n=2}^{N}[x_i - \bar{x}]^2} \tag{7.1}$$

For the avoidance of the outliers the median value is also calculated.

The start of the recordings in every patient is in different time. The classification is done using the information given from the database. A 24-hour period is produced starting at 11:00 AM.

The process followed for the features is the same as mentioned for the intervals. Two time windows are selected here too. The mean and the median for every feature is calculated as well the SDNN.

7.3 Results

The experimental results for the PP and TT intervals show a Circadian behavior in the same way as RR operates. The SDNN shows no significant change in the time domain. The fig. 7.1 indicates the mean values of each interval per hour while fig. 7.2 shows the median for all intervals per hour. It is worth to notice that the difference (d) between the graphs of intervals (e.g. 0.9 sec in TT interval and 0.92 sec for PP interval) is represented as the $d * 7.8125$ msec.

A polyfit curve is also implemented in every result. The polyfit function of Matlab R2014b is used which returns the coefficients for a polynomial $p(x)$ of degree n that is a best fit (in a least-squares sense) for the data in y (mean, median or SDNN values). The coefficients in 7.2 are in descending powers, and the length of p is $n + 1$.

$$p(x) = p_1 x^n + p_2 x^{n-1} + \cdots + p_n x + p_{n+1} \tag{7.2}$$

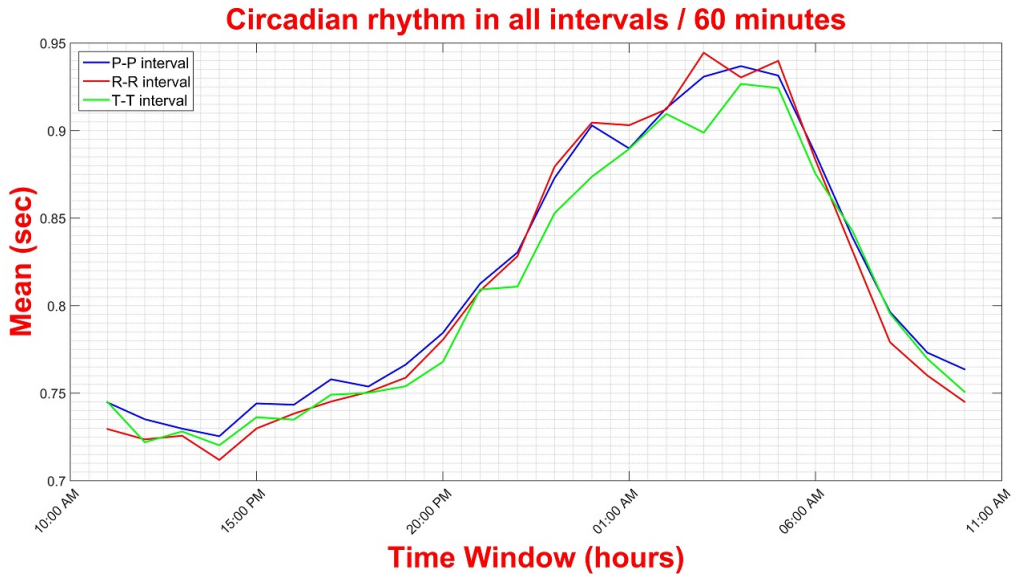

Figure 7.1: Mean values for all intervals per hour

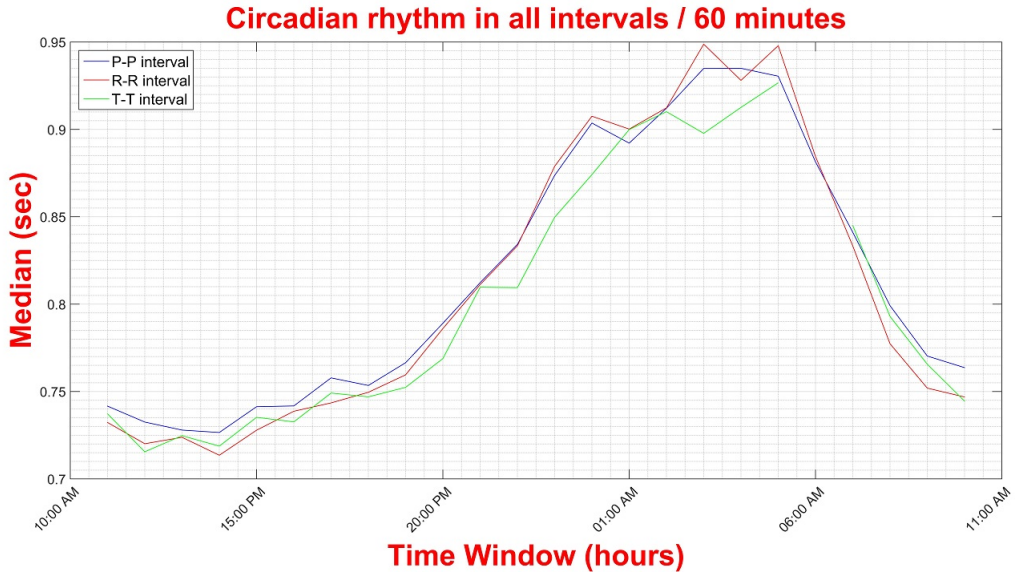

Figure 7.2: Median values for all intervals per hour

The SDNN across the 24-hours is shown in fig. 7.3.

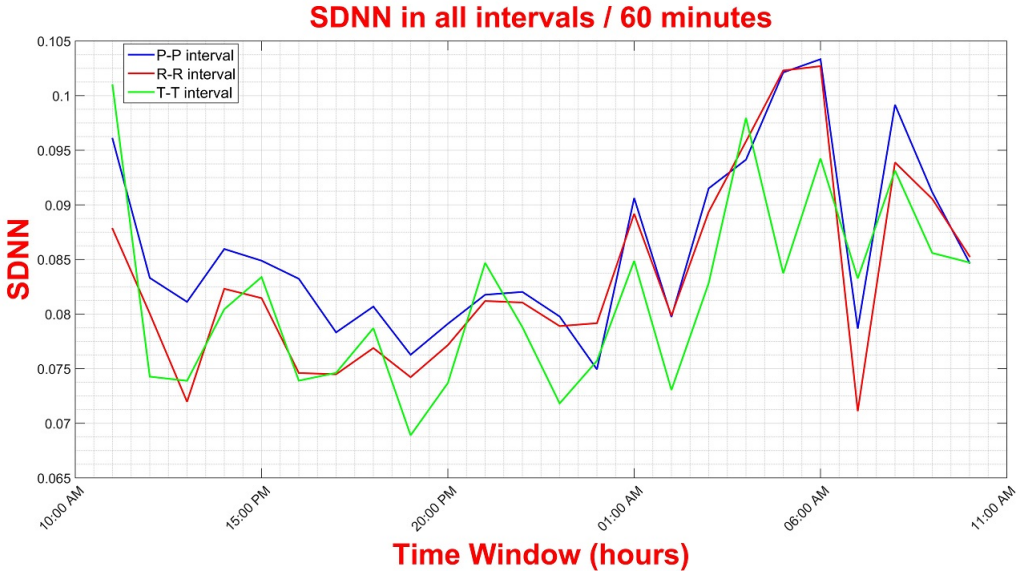

Figure 7.3: SDNN values for all intervals per hour

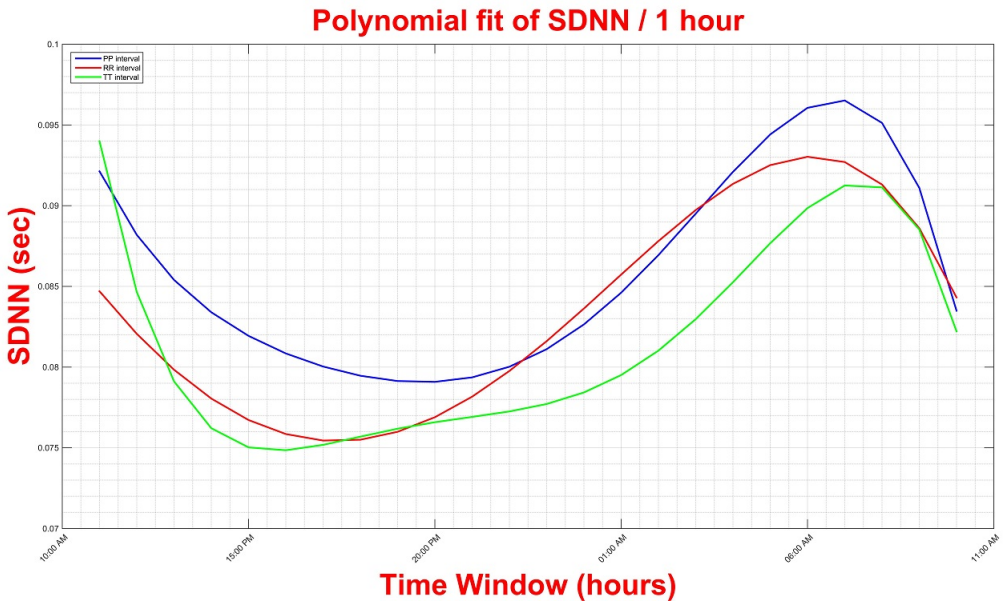

Figure 7.4: SDNN values for all intervals per hour using 5th degree polynomial curve

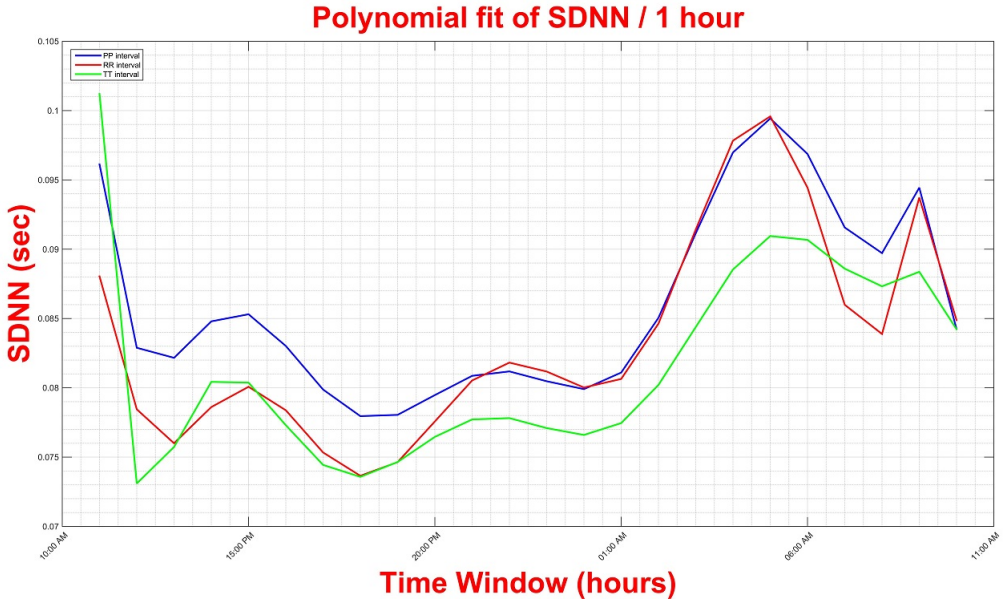

Figure 7.5: SDNN values for all intervals per hour using 10th degree polynomial curve

Bellow in Fig. 7.6, 7.7, 7.8 we display the results computed every 30 minutes.

Figure 7.6: Mean values for all intervals per 30 minutes

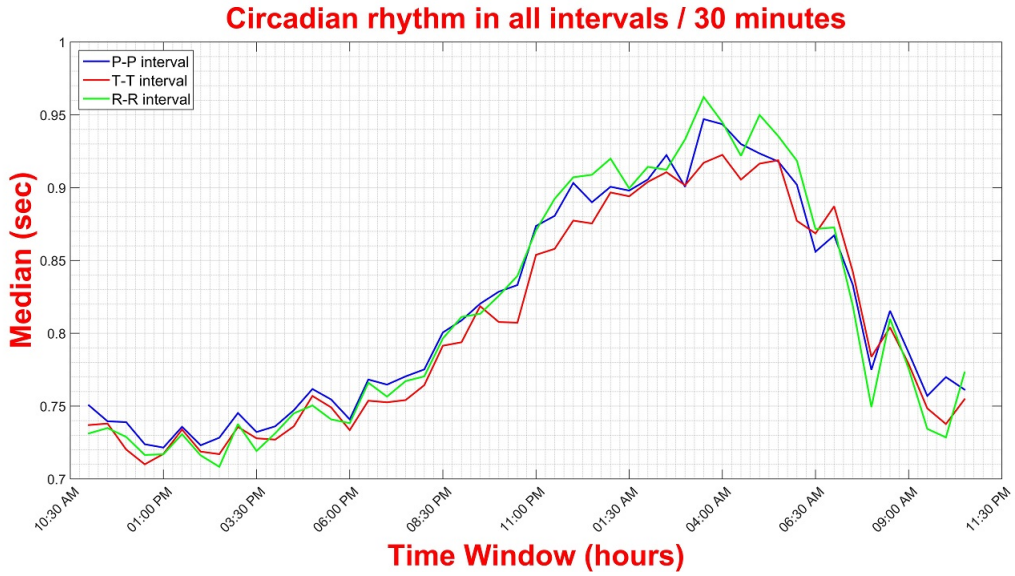

Figure 7.7: Median values for all intervals per 30 minutes

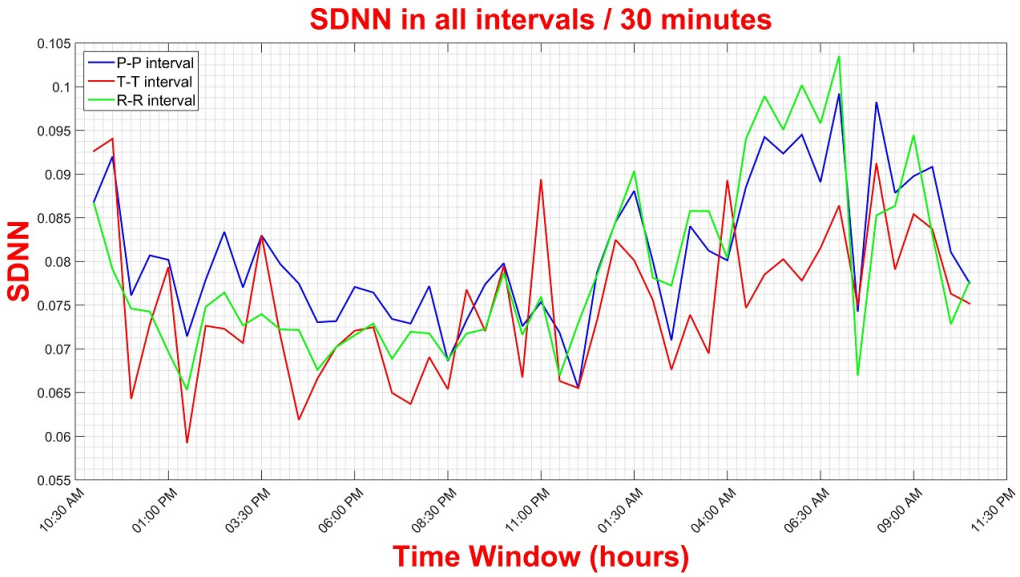

Figure 7.8: SDNN values for all intervals per 30 minutes

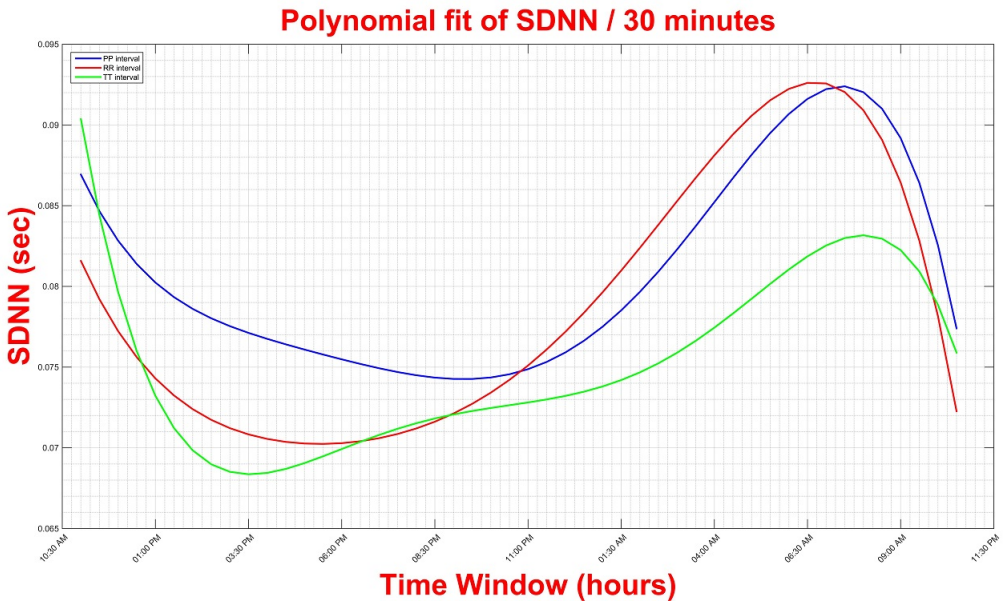

Figure 7.9: SDNN values for all intervals per 30 minutes using 5th degree polynomial curve

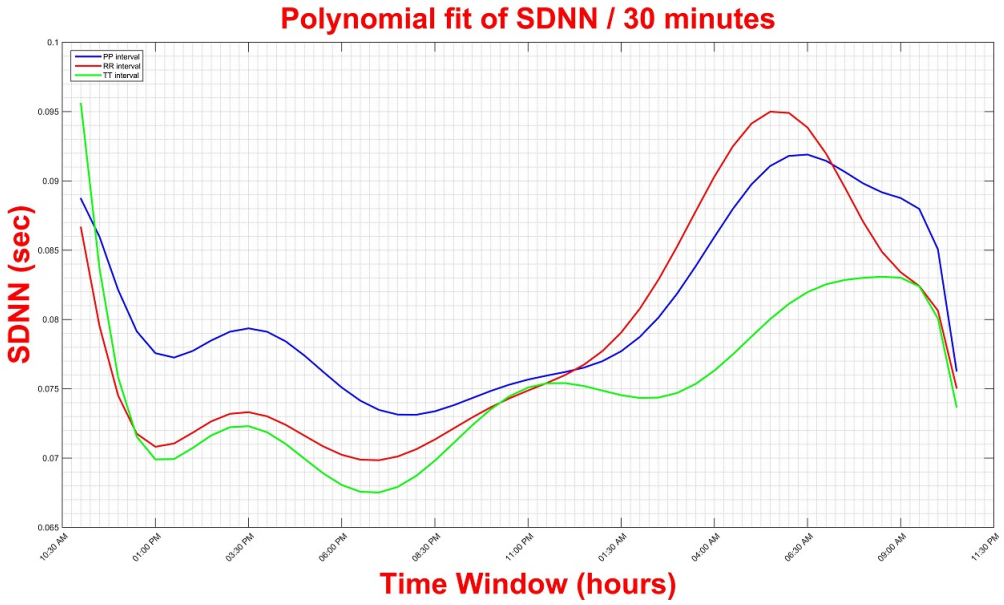

Figure 7.10: SDNN values for all intervals per 30 minutes using 10th degree polynomial curve

The fig. 7.4, 7.5, 7.9, 7.10 represent a polynomial fit as described above in eq. 7.3 of the SDNN (5 and 10 degree polynomial function). It seems there is a difference in PP against RR and TT intervals during the day as well in TT against PP and RR intervals during the night. But, we will try to indicate a significant linear relationship between them.

In order to show that kind of relationship between RR and PP intervals, RR and TT intervals and also PP and TT intervals a linear regression is implemented. Fig. 7.11 depicts this relationship in every case.

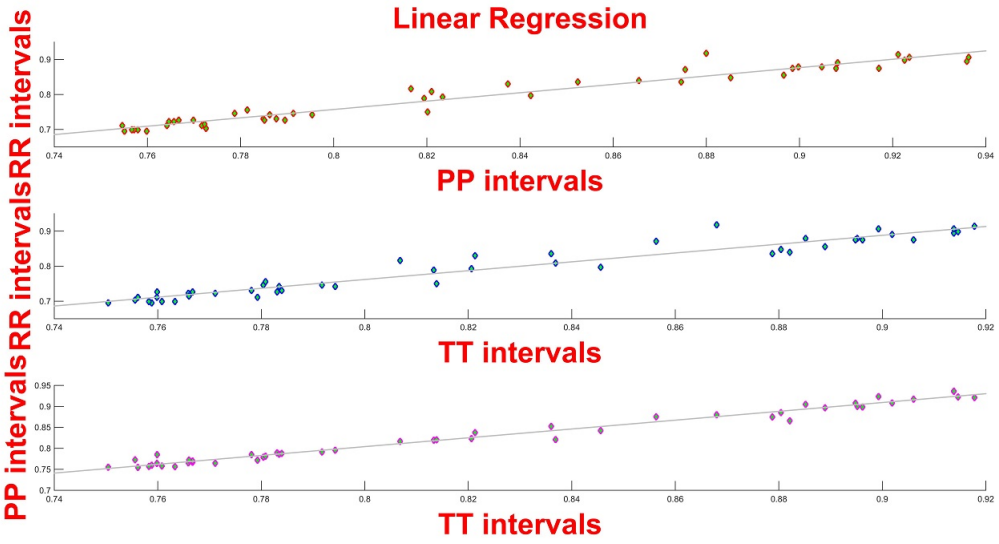

Figure 7.11: Linear Regression of the intervals

Table 7.1: Results from Linear Regression in all intervals

	β_0	β_1	SE	MSE	$t-statistic$	$p-value$
RR-PP	-0.2001	1.1966	0.0418	0.0003124	28.6446	$\ll 0.0001$
RR-TT	-0.2471	1.2613	0.0501	0.0003986	25.160	$\ll 0.0001$
PP-TT	-0.0367	1.0510	0.0228	0.0000826	46.0567	$\ll 0.0001$

Table 7.1 indicates the results in each case. The β_0 and β_1 declare the y-intercept and the slope respectively of the eq 7.3. We want to reject the null hypothesis (see eq. 7.4); observed test statistic (value bigger than 2) with a low p-value can conclude that there is a significant linear relationship between those intervals.

$$y = \beta_0 + \beta_1 x + \epsilon, \epsilon = noise \tag{7.3}$$

$$H_0 : \beta_1 = 0,$$

$$H_a : \beta_1 \neq 0$$

$$use \ \frac{\widehat{\beta_1}}{SE(\widehat{\beta_1})} \tag{7.4}$$

Fig. 7.12 indicates the results of meadian for all intervals while fig. 7.13 shows the means respectively.

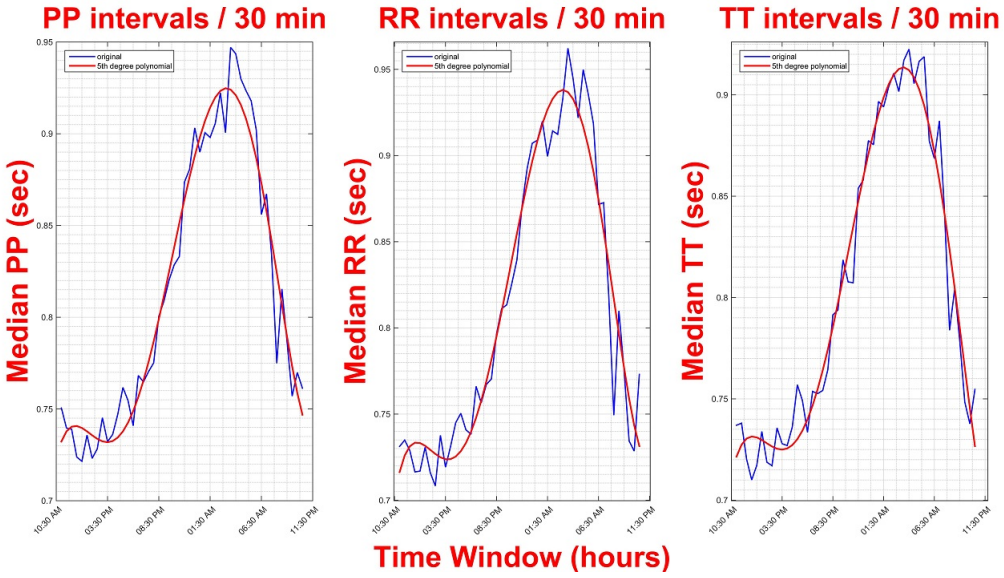

Figure 7.12: Median values for all intervals per 30 minutes using polynomial fit

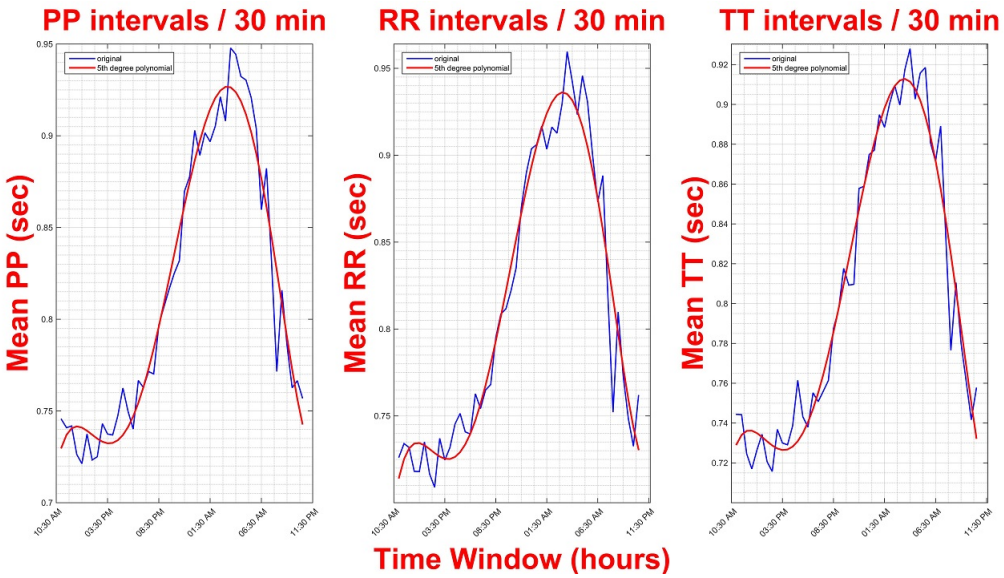

Figure 7.13: Mean values for all intervals per 30 minutes using polynomial fit

Fig. 7.14 depicts the SDNN for the PP, RR and TT intervals. We can see here the difference between the intervals as it is expected. All the intervals tend to be higher during the nighttime whereas during the daytime exactly the opposite.

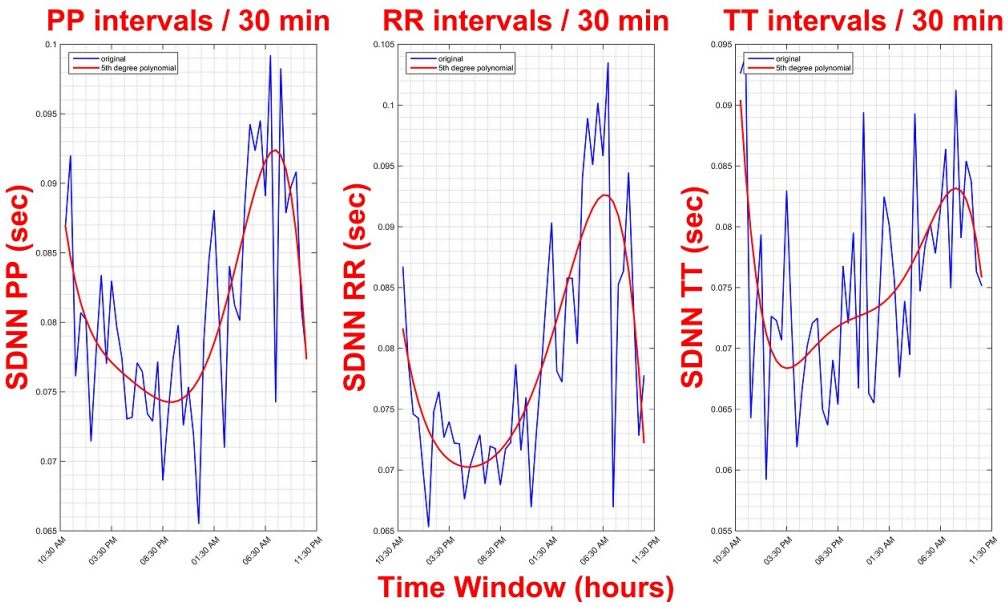

Figure 7.14: SDNN values for all intervals per 30 minutes using polynomial fit

The results for the mean, median and SDNN are also displayed in a different time frame (1 hour) in fig. 7.16, 7.15 and 7.17 respectively.

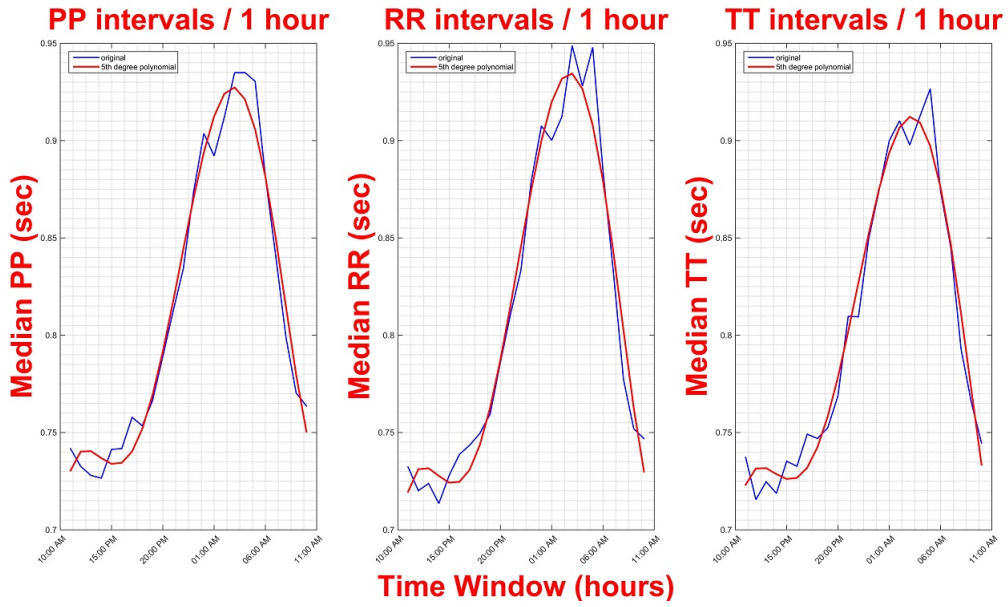

Figure 7.15: Median values for all intervals per hour using polynomial fit

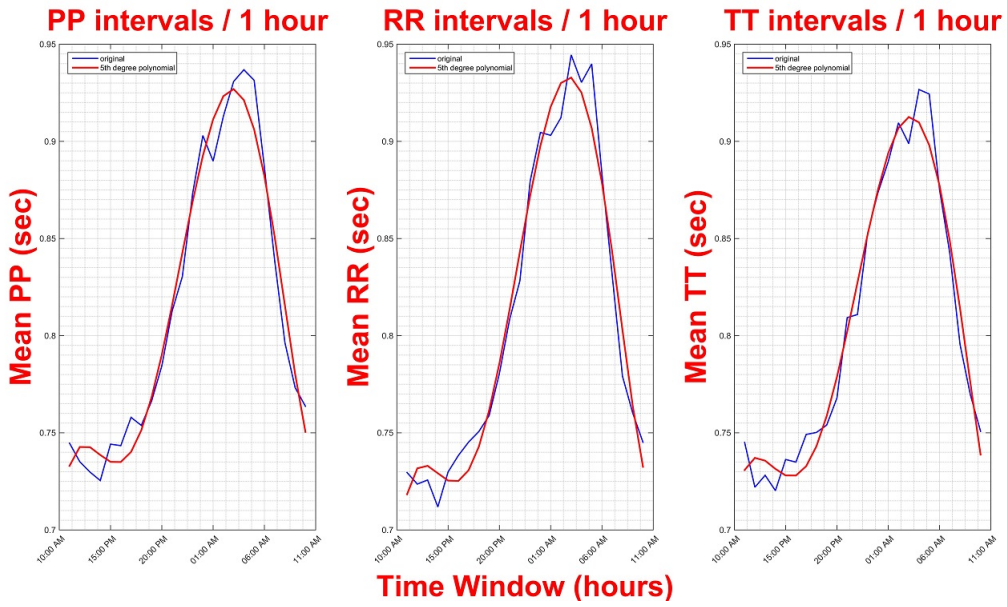

Figure 7.16: Mean values for all intervals per hour using polynomial fit

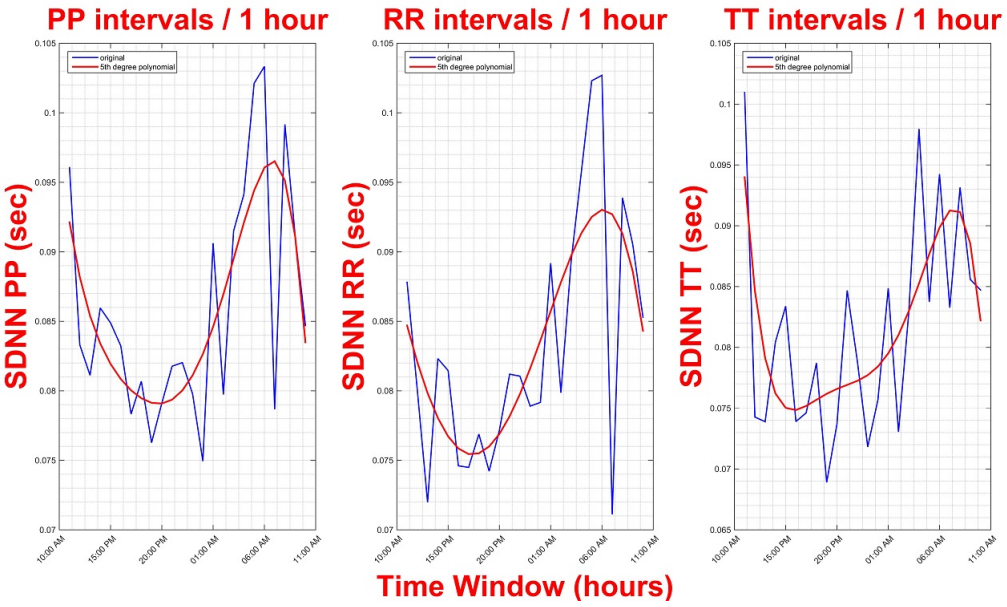

Figure 7.17: SDNN values for all intervals per hour using polynomial fit

As we may notice in fig. 7.13, 7.12, 7.16, 7.15 the polynomial line in the first hour of every interval starts from a lower point. This is due to the fact of low information at the start time of the records.

What is common for every interval is the low values during the day $(0,7 - 0.75sec)$ and higher values during the night $(0,9 - 0.95sec)$ with an increase in the afternoon and a decrease early in the morning. It is obvious that these values have to be in that way

because as we are awake our heartbeat works faster than when we sleep. According to Circadian rhythm the alertness of body in the morning rises so the interval decreases On the contrary at the first night hours the rhythm falls so the intervals increases.

After seeing the behavior in the PP, RR and TT intervals with respect to Circadian rhythm, the results on features will be discussed below compared with circadian behavior also.

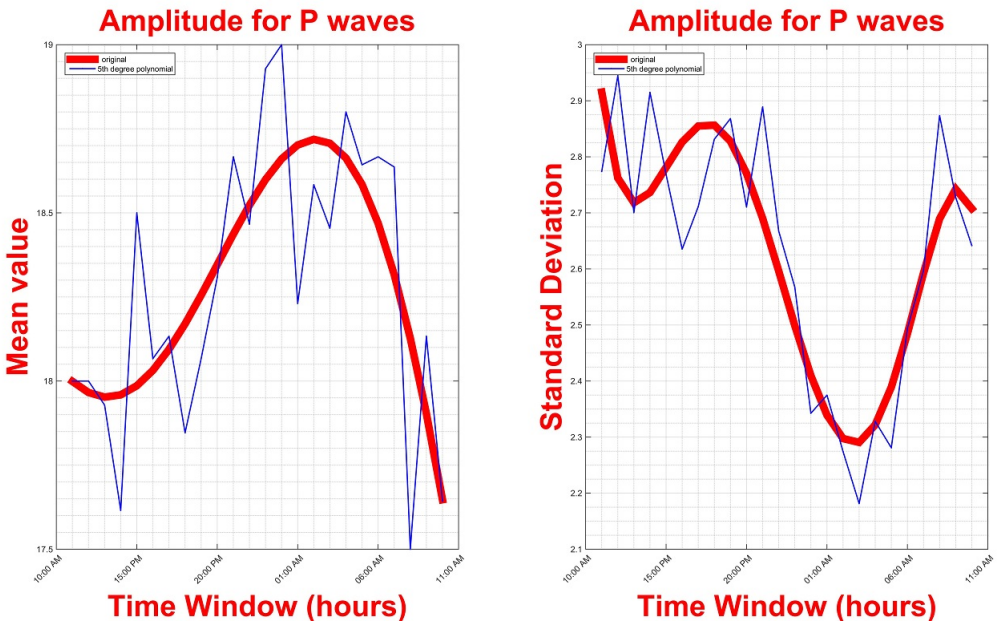

Figure 7.18: Mean and Standard Deviation values per hour for Amplitude feature with polynomial fit

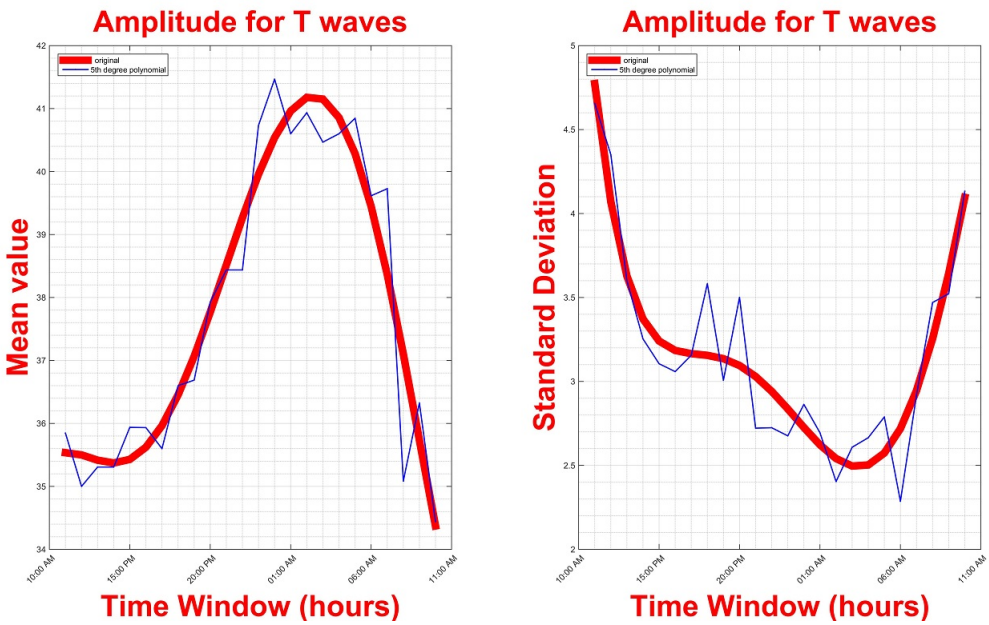

Figure 7.19: Mean and Standard Deviation values per hour for Amplitude feature with polynomial fit

Fig. 7.18 and 7.19 depict the amplitude (duration) of the 24-hour pattern of mean and the standard deviation in P and T waves respectively. We can notice that in P and T waves the amplitude has lower values in the daytime than in the nighttime. This is obvious since RR intervals show similar behavior in those time periods.

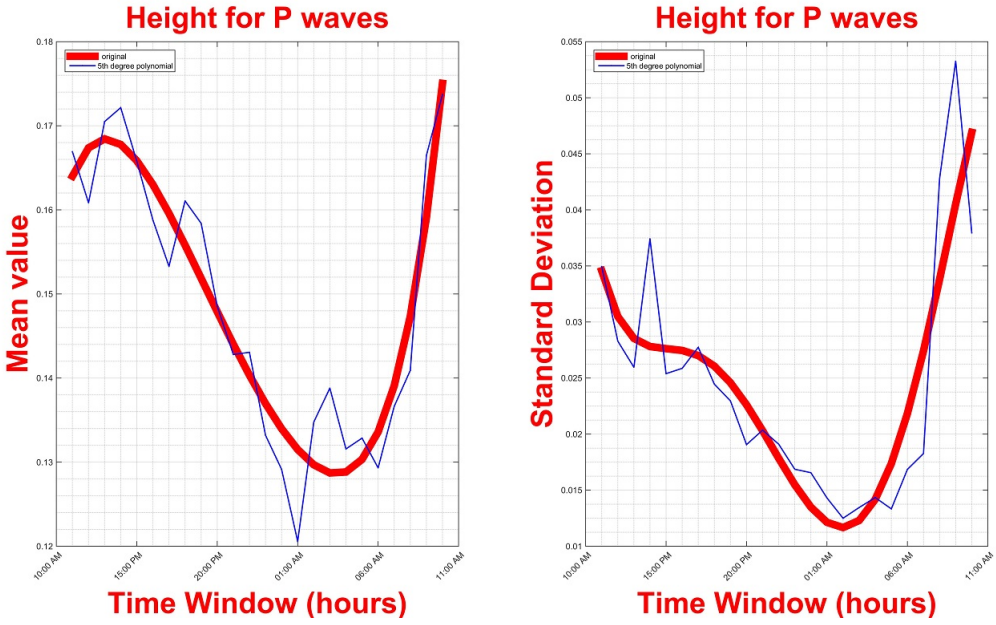

Figure 7.20: Mean and Standard Deviation values per hour for Height feature with polynomial fit

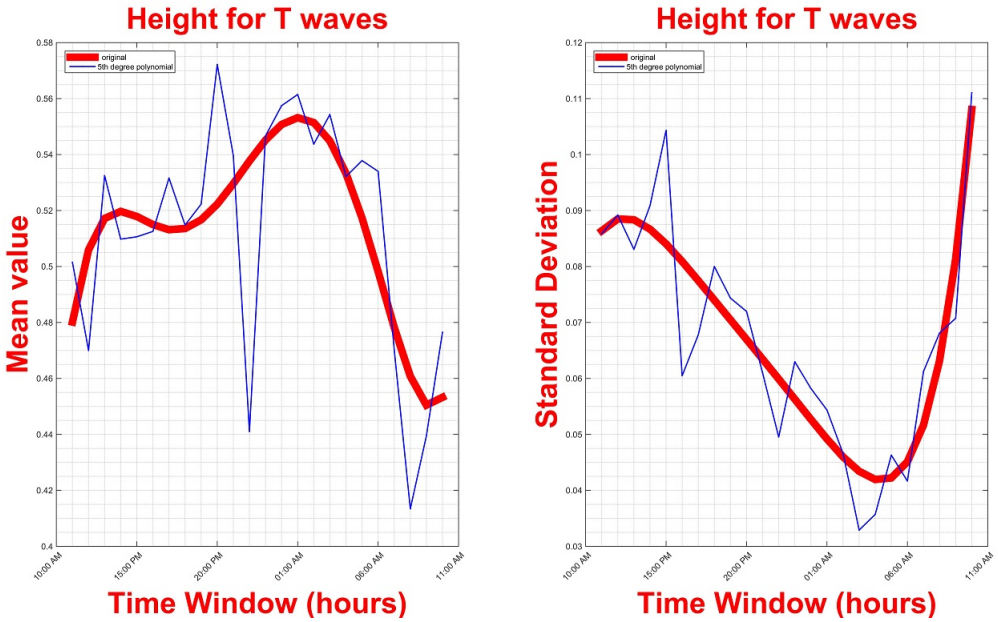

Figure 7.21: Mean and Standard Deviation values per hour for Height feature with polynomial fit

Fig. 7.20 and 7.21 show the alternations in height curve of P and T waves over the 24-hour. Here the height of P waves seems to have an opposite behavior with those of T waves. So, when the mean height values of P waves decrease the correspond values of T waves increase and vice versa.

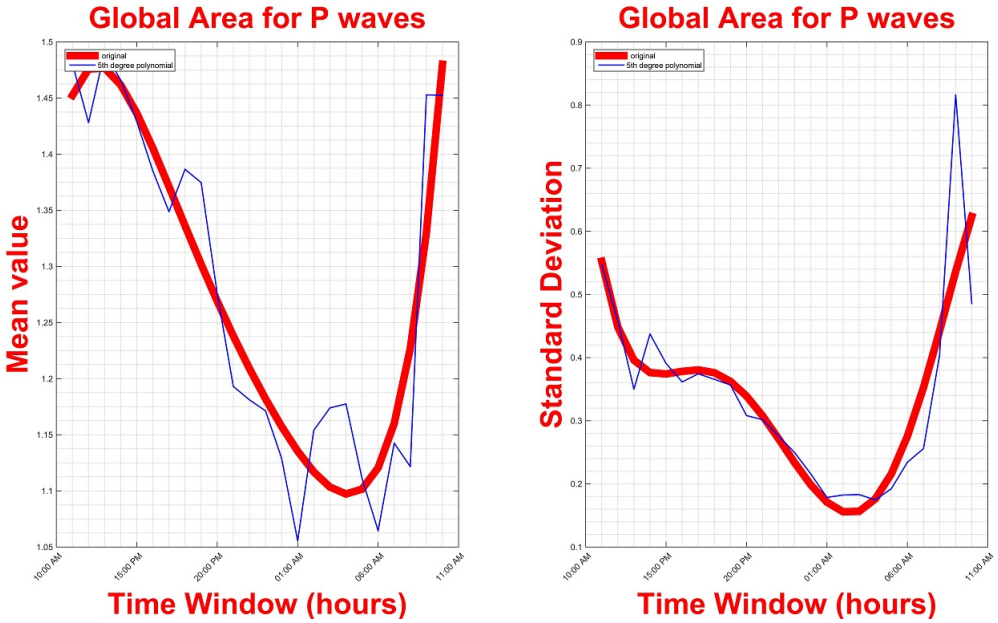

Figure 7.22: Mean and Standard Deviation values per hour for Global Area feature with polynomial fit

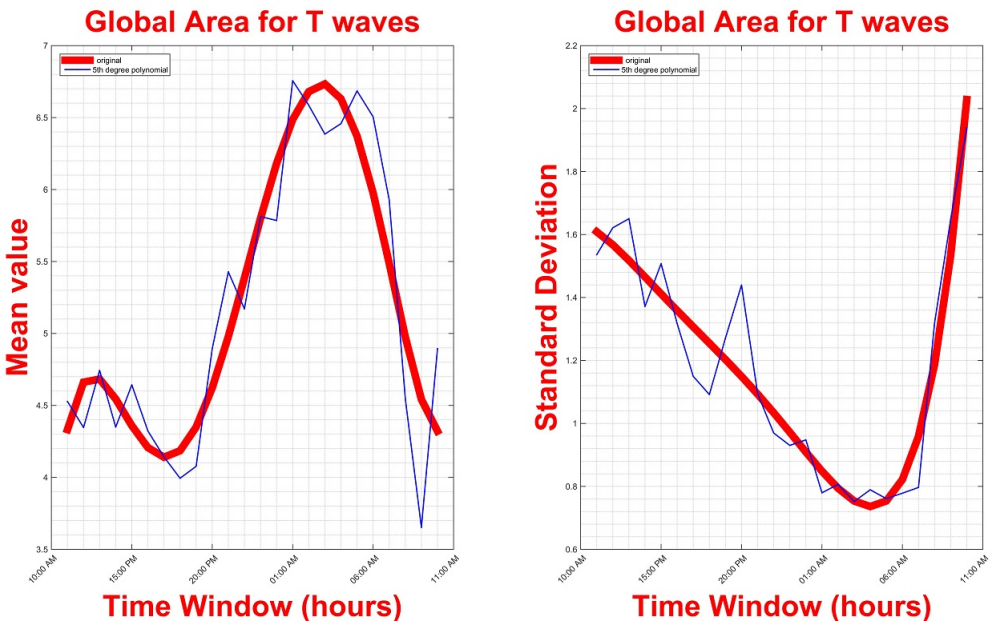

Figure 7.23: Mean and Standard Deviation values per hour for Global Area feature with polynomial fit

Fig. 7.22 and 7.23 above present the diurnal pattern of mean and the standard deviation of the global area in P and T waves respectively. We can observe that in P waves the global area has greater values in the daytime than in the nighttime, while the global area in T waves has exactly the opposite behavior. This pattern was expected according to amplitude and height ones. While the amplitude in both waves have the same behavior, the height of these waves will determine their area (lower height implies lower area) as is shown in above figures.

However, the standard deviation seems to have the same pattern in those features (amplitude, height, global area). This indicates high variation or dispersion in data points during the day and low during the night.

As we can mention, there is a usual circadian pattern in all features discussed above in their mean value and standard deviation. Similar results are shown in the rest features (see Appendix).

Chapter 8

Conclusion and Future work

8.1 Conclusion

Summary of the thesis

In this thesis, we presented new methods and we examined new features for the discrimination of day and night periods, by implementing one manual and two automatic P & T wave detection. We presented the comparison between the results of all features and for every method. Also we achieved to indicate that there is a circadian behavior in P and T waves.

We introduced the processing of our manual detection algorithm, a simple idea that triggered the creation of the automatic algorithms in order to extend our results and generalize the idea of discrimination. After that, we presented these new automatic algorithms and explains their characteristics; pAD was first implemented by using dynamic threshold (percentile) for the limitation and the new version GlAD in which a probabilistic model named as HCRF was used and the selection of the waves was done online.

We analyze in detail every feature that we implemented in order to examine the ability of discrimination between day and night periods, we presented the ratios of them. We pair-tested the hypothesis of the referring discrimination that gave us remarkable results. Next we discussed the classification that came from the implementation of our features and we analyze the main famous classification methods. Finally, we brought to light the dependencies and relationships between consecutive intervals of ECG; PP and TT intervals are following the same behavior of RR intervals in terms of circadian rhythm. This fact can be extremely important especially in some cases where the QRS complex shows defects in terms of morphology. Several features showed a possible association with circadian behavior.

Contribution of the thesis

This thesis has developed a manual and two automatic algorithms for the detection of P and T waves in an ECG signal. The results are used to extract many new features of these waves that seem to have the ability of discriminating between day and night periods.

This work reveals that by examining the P and T waves of an ECG signal one can decide if they refer to a day or to a night period. Although the QRS complex is usually the central and most visually obvious part of the tracing, since it corresponds to the depolarization of the ventricles of the human heart, it is also most affected by some dysfunctions of the heart and thus, in some cases, may be more defected than P or T wave. For this reason, our ability to discriminate day and night periods through the next two more significant waves of an ECG signal, could give more information.

Because R wave can be more defected than P or T waves in some cases, such us tachycardia or Wolf Parkinson White Syndrome, the examination of RR intervals may be less significant than PP or TT intervals and thus there is an obvious danger of misleading to wrong results in terms of circadian rhythm. Our thesis supports that the PP and TT intervals should be reset and re-analyzed in order to decide about circadian rhythm in cases of specific human disorders.

Another new approach of our thesis, is the GlAD algorithm, where the annotation of P and T waves is happening with no respect to R waves. This is also an advantage of

detecting these waves when several dysfunctions do not allow the R peak to be the most visible and significant point of a beat tracing.

Limitations of the current work

The database we examined in our thesis, corresponds to healthy volunteers; the importance of P and T detection might be even more remarkable in cases where some patients are suffering from certain diseases. A comparison between the classification rates of healthy and unhealthy patients as will be mentioned on Future Work section. In our thesis we presented only a certain 2-hour timing of each period, 13:00 -15:00 and 01:00 βἐŏ 03:00. For some people it is possible that none of these two periods are working or resting period.

8.2 Future work

While our thesis has demonstrated novel methodology for detecting the P and T waves and discriminating between day and night periods of an ECG signal, many opportunities for extending the scope of this thesis remain.

A crucial parameter for evaluating the contribution of our thesis to the medical community is the implementation of our methods to ECG databases that refer to unhealthy patients. That work, could reveal the importance of P and T waves annotation, if it indicates better classification results from the current thesis, or in comparison to defected R waves.

Another comparison that could be done, is to implement our methods and features to all the sequential 2-hour periods during all day and compare between each other. This method could reveal the best classification rate for each period per patient or examine the heart behavior during or after lunch or even analyze waves' behavior between patients of different biological clock.

In addition, the implementation of our thesis according to the schedule of each patient so as to distinguish the 2-hours sampling intervals that one is sleeping or working, could

give better results and thus enforce our hypothesis of circadian rhythm annotation through P and T waves.

The most obvious stage is to improve the GlAD algorithm in order to achieve better results. The four features used for the training phase from HCRF may be increased learning more details in the waves' condition for a better classification.

Bibliography

[1] M. Hastings, "The brain, circadian rhythms, and clock genes," *BMJ*, vol. 317, no. 7174, pp. 1704–1707, 1998.

[2] T. Hilbel, T. M. Helms, H. A. Katus, P.-D. D. C. Zugck *et al.*, "Telemetrie," *Herzschrittmachertherapie+ Elektrophysiologie*, vol. 19, no. 3, pp. 146–154, 2008.

[3] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "Physiobank, physiotoolkit, and physionet components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000.

[4] P. Trahanias and E. Skordalakis, "Syntactic pattern recognition of the ecg," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 12, no. 7, pp. 648–657, 1990.

[5] I. Murthy and G. Prasad, "Analysis of ecg from pole-zero models." *IEEE transactions on bio-medical engineering*, vol. 39, no. 7, pp. 741–751, 1992.

[6] I. Murthy and U. Niranjan, "Component wave delineation of ecg by filtering in the fourier domain," *Medical and Biological Engineering and Computing*, vol. 30, no. 2, pp. 169–176, 1992.

[7] N. V. Thakor and Y.-S. Zhu, "Applications of adaptive filtering to ecg analysis: noise cancellation and arrhythmia detection," *Biomedical Engineering, IEEE Transactions on*, vol. 38, no. 8, pp. 785–794, 1991.

[8] J. Yamato, J. Ohya, and K. Ishii, "Recognizing human action in time-sequential images using hidden markov model," in *Computer Vision and Pattern Recognition, 1992. Proceedings CVPR'92., 1992 IEEE Computer Society Conference on.* IEEE, 1992, pp. 379–385.

[9] J. Lafferty, A. McCallum, and F. C. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," 2001.

[10] A. Quattoni, S. Wang, L.-P. Morency, M. Collins, and T. Darrell, "Hidden conditional random fields," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 10, pp. 1848–1852, 2007.

[11] K. Bousmalis, S. Zafeiriou, L.-P. Morency, and M. Pantic, "Infinite hidden conditional random fields for human behavior analysis," *Neural Networks and Learning Systems, IEEE Transactions on*, vol. 24, no. 1, pp. 170–177, 2013.

[12] D. C. Liu and J. Nocedal, "On the limited memory bfgs method for large scale optimization," *Mathematical programming*, vol. 45, no. 1-3, pp. 503–528, 1989.

[13] P. Arsenos and G. Manis, "The variability of the t-wave shape can discriminate young and elderly subjects," in *Information Technology and Applications in Biomedicine (ITAB), 2010 10th IEEE International Conference on.* IEEE, 2010, pp. 1–3.

[14] E. Zeraatkar, S. Kermani, A. Mehridehnavi, A. Aminzadeh, E. Zeraatkar, and H. Sanei, "Arrhythmia detection based on morphological and time-frequency features of t-wave in electrocardiogram," *Journal of medical signals and sensors*, vol. 1, no. 2, p. 99, 2011.

[15] N. Neyroud, P. Maison-Blanche, I. Denjoy, S. Chevret, C. Donger, E. Dausse, J. Fayn, F. Badilini, N. Menhabi, K. Schwartz *et al.*, "Diagnostic performance of qt interval variables from 24-h electrocardiography in the long qt syndrome," *European heart journal*, vol. 19, no. 1, pp. 158–165, 1998.

[16] F. Braga, E. Caiani, E. Locati, and S. Cerutti, "Automated qt/rr analysis based on selective beat averaging applied to electrocardiographic holter 24 h," in *Computers in Cardiology*, 2004, pp. 9–12.

[17] J. Ramırez, I. Cygankiewicz, P. Laguna, M. Malik, and E. Pueyo, "Circadian pattern and sex differences of qt/rr and t-peak-to-end/rr curvatures and slopes in chronic heart failure patients," in *Computing in Cardiology Conference (CinC)*.

[18] P. E. Dilaveris, P. Färbom, V. Batchvarov, A. Ghuran, and M. Malik, "Circadian behavior of p-wave duration, p-wave area, and pr interval in healthy subjects," *Annals of noninvasive electrocardiology*, vol. 6, no. 2, pp. 92–97, 2001.

[19] S. J. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 2nd ed. Pearson Education, 2003.

[20] N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression," *The American Statistician*, vol. 46, no. 3, pp. 175–185, 1992. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/00031305.1992.10475879

[21] L. Rokach and O. Maimon, *Data Mining with Decision Trees: Theroy and Applications*. River Edge, NJ, USA: World Scientific Publishing Co., Inc., 2008.

[22] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Sep. 1995. [Online]. Available: http://dx.doi.org/10.1023/A:1022627411411

[23] S. Geisser, *Predictive Inference*, ser. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, 1993. [Online]. Available: https://books.google.gr/books?id=wfdlBZ_iwZoC

[24] M. M. Massin, K. Maeyns, N. Withofs, F. Ravet, and P. Gérard, "Circadian rhythm of heart rate and heart rate variability," *Archives of Disease in Childhood*, vol. 83, no. 2, pp. 179–182, 2000.

[25] M. Nakagawa, T. Iwao, S. Ishida, H. Yonemochi, T. Fujino, T. Saikawa, and M. Ito, "Circadian rhythm of the signal averaged electrocardiogram and its relation to heart rate variability in healthy subjects," *Heart*, vol. 79, no. 5, pp. 493–496, 1998.

[26] P. Boudreau, G. Dumont, N. Kin, C.-D. Walker, and D. B. Boivin, "Correlation of heart rate variability and circadian markers in humans," in *Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE.* IEEE, 2011, pp. 681–682.

[27] E. M. Ekholm, J. Hartiala, and H. V. Huikuri, "Circadian rhythm of frequency-domain measures of heart rate variability in pregnancy," *BJOG: An International Journal of Obstetrics & Gynaecology*, vol. 104, no. 7, pp. 825–828, 1997.

Appendix

This appendix containts the results from all features extracted either from P or T waves, as described in Chapter 7. Thus, all the figures below denote the results of the mean value and the standard deviation all over the 24-hour in a time window of 1 hour. A polynomial of 5th degree is computed in order to show the circadian behavior in those features.

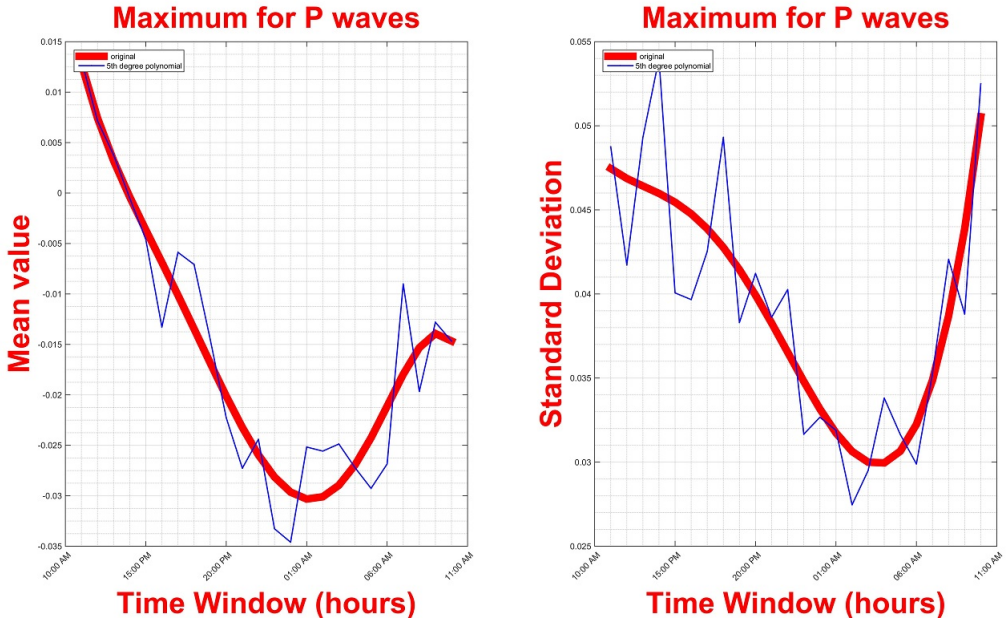

Figure 8.1: Mean and Standard Deviation values per hour for Maximum feature with polynomial fit

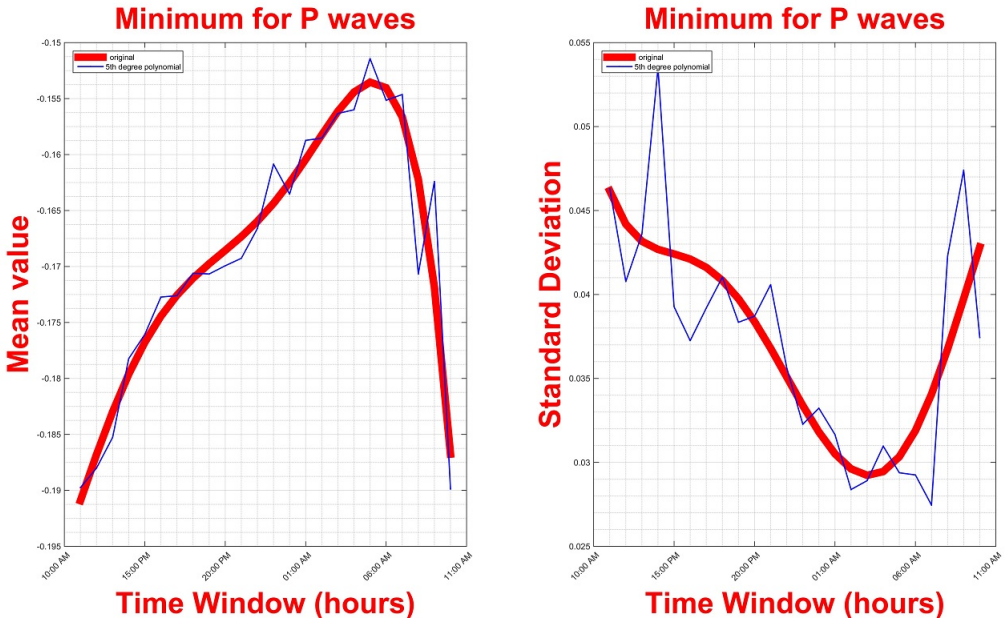

Figure 8.2: Mean and Standard Deviation values per hour for Minimum feature with polynomial fit

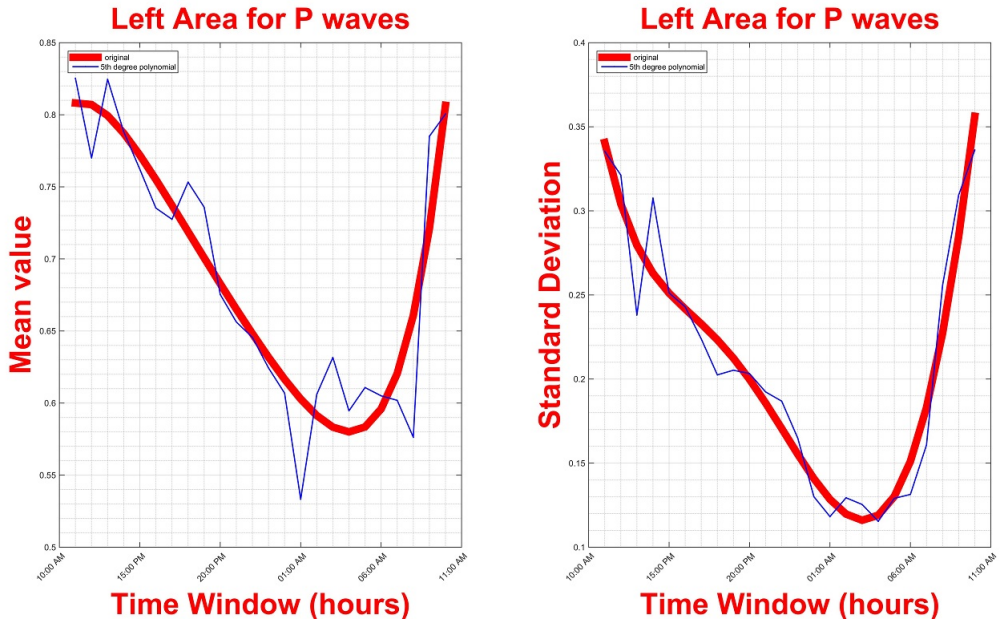

Figure 8.3: Mean and Standard Deviation values per hour for Left Area feature with polynomial fit

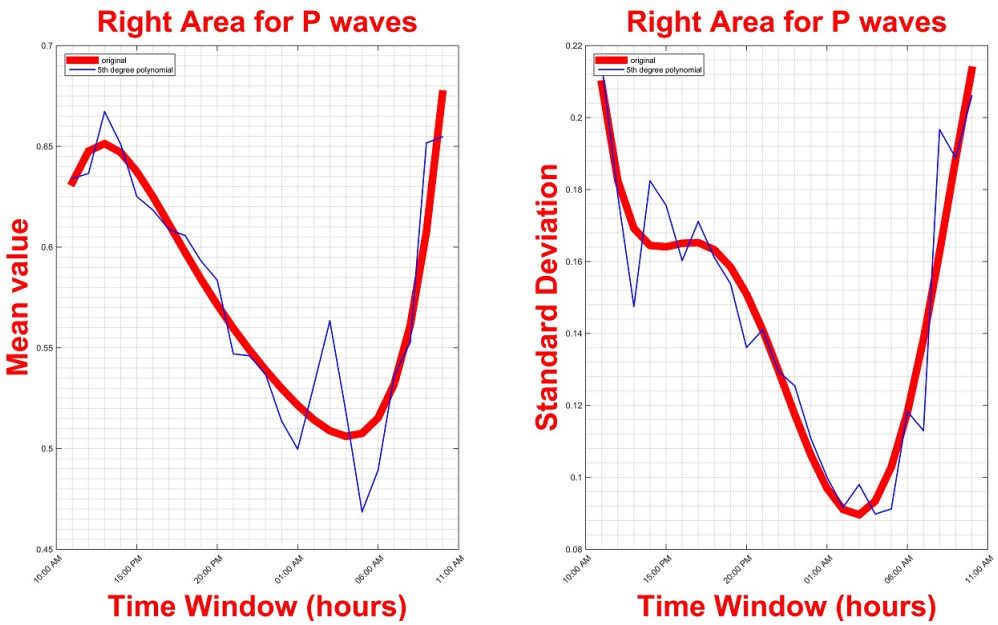

Figure 8.4: Mean and Standard Deviation values per hour for Right Area feature with polynomial fit

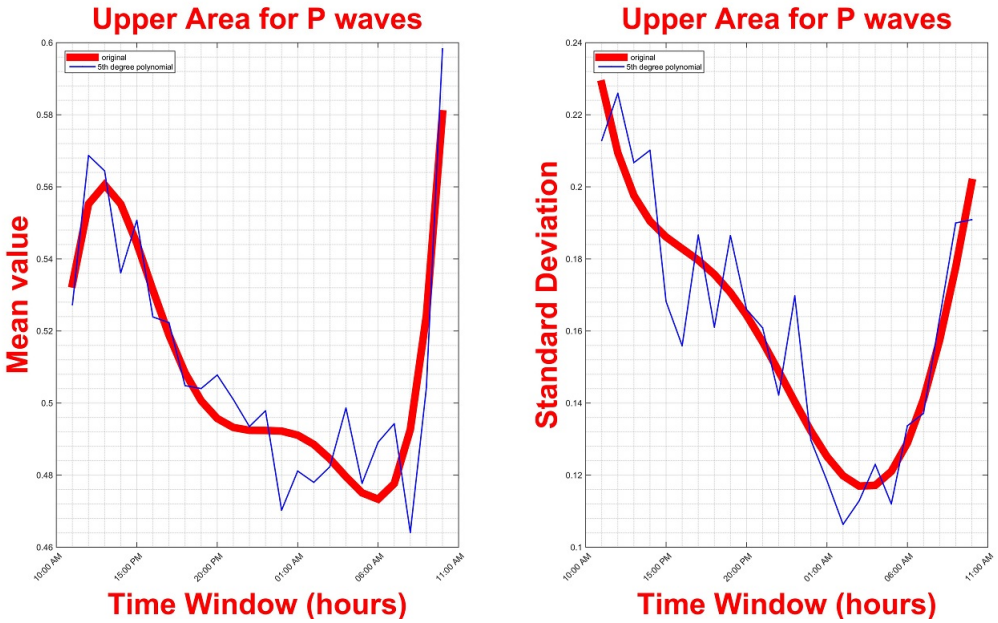

Figure 8.5: Mean and Standard Deviation values per hour for Upper Area feature with polynomial fit

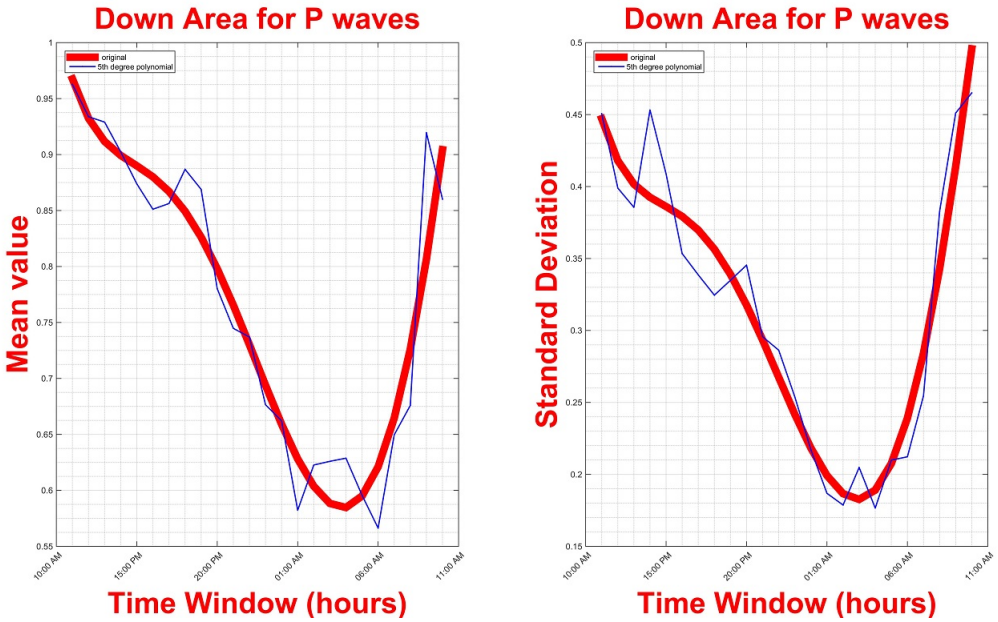

Figure 8.6: Mean and Standard Deviation values per hour for Down Area feature with polynomial fit

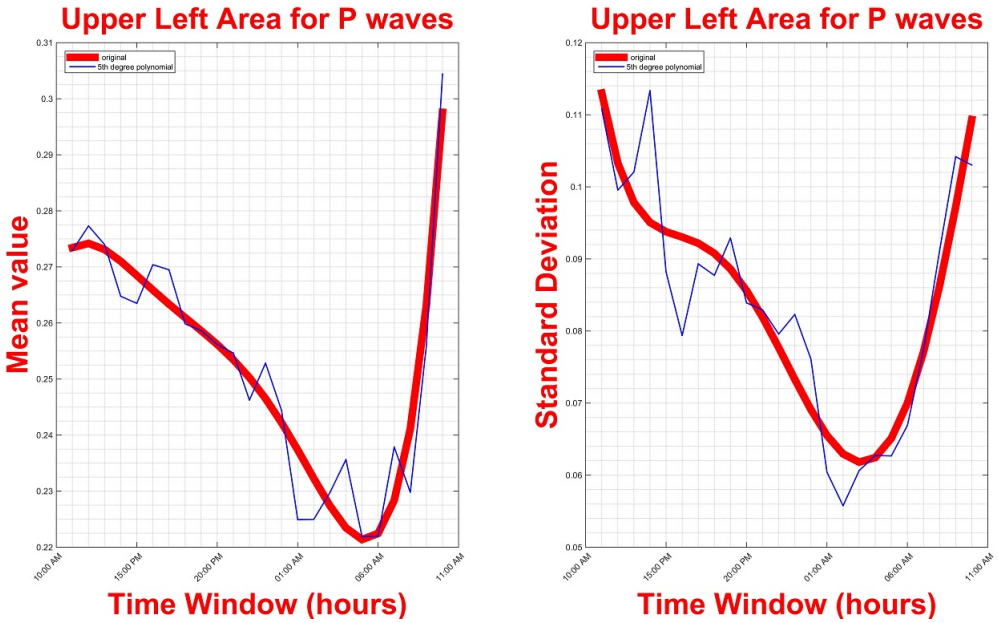

Figure 8.7: Mean and Standard Deviation values per hour for Upper Left Area feature with polynomial fit

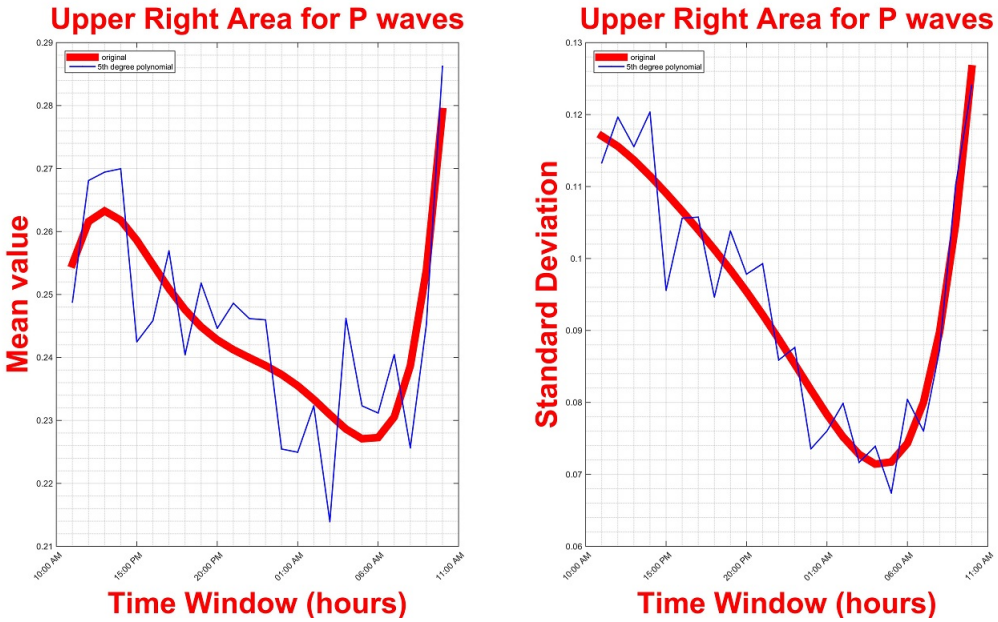

Figure 8.8: Mean and Standard Deviation values per hour for Upper Right Area feature with polynomial fit

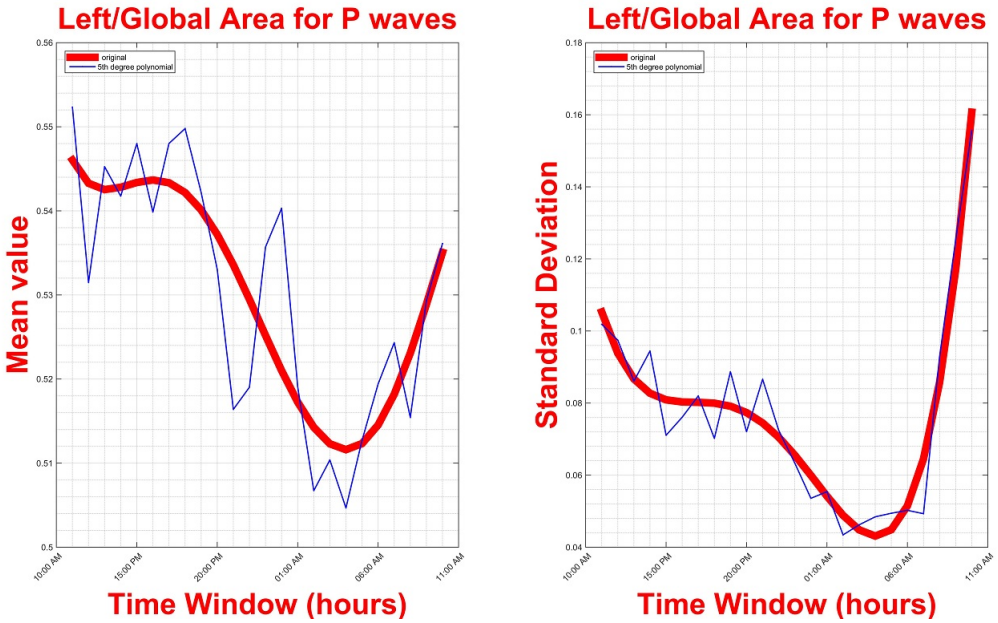

Figure 8.9: Mean and Standard Deviation values per hour for Left/Global Area feature with polynomial fit

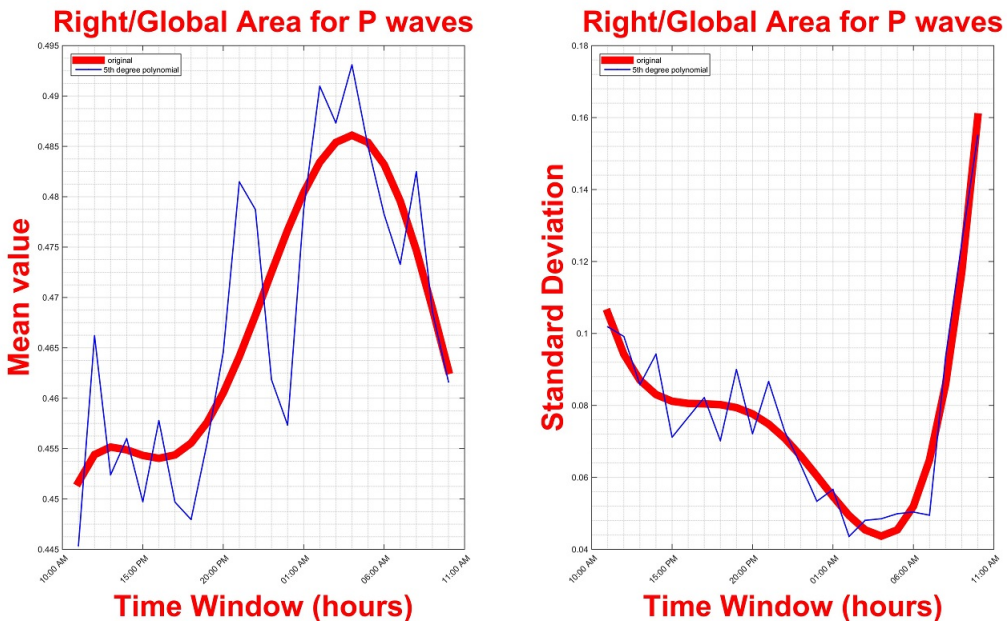

Figure 8.10: Mean and Standard Deviation values per hour for Right/Global Area feature with polynomial fit

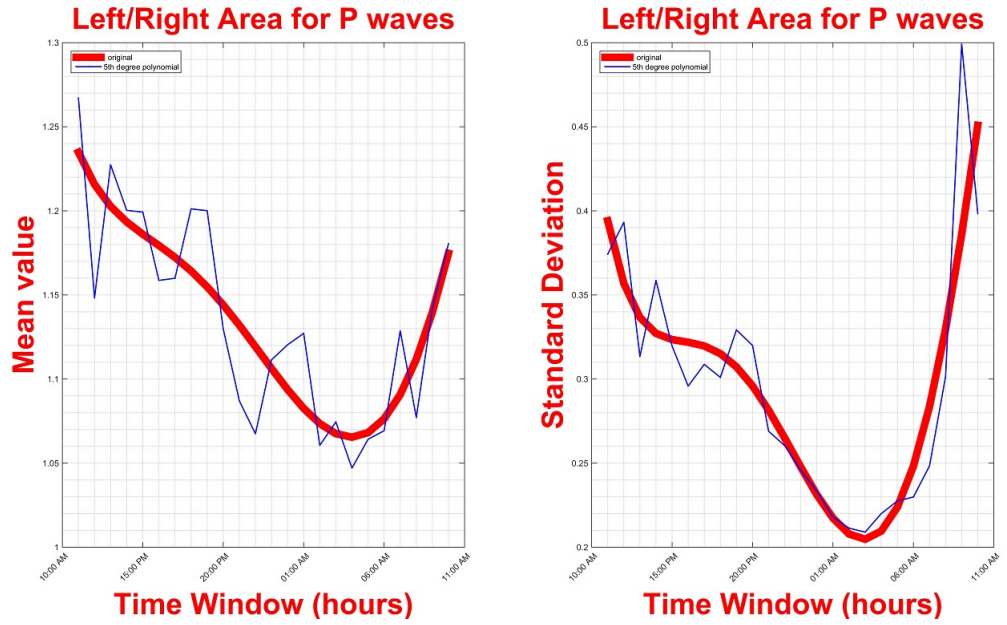

Figure 8.11: Mean and Standard Deviation values per hour for Left/Right Area feature with polynomial fit

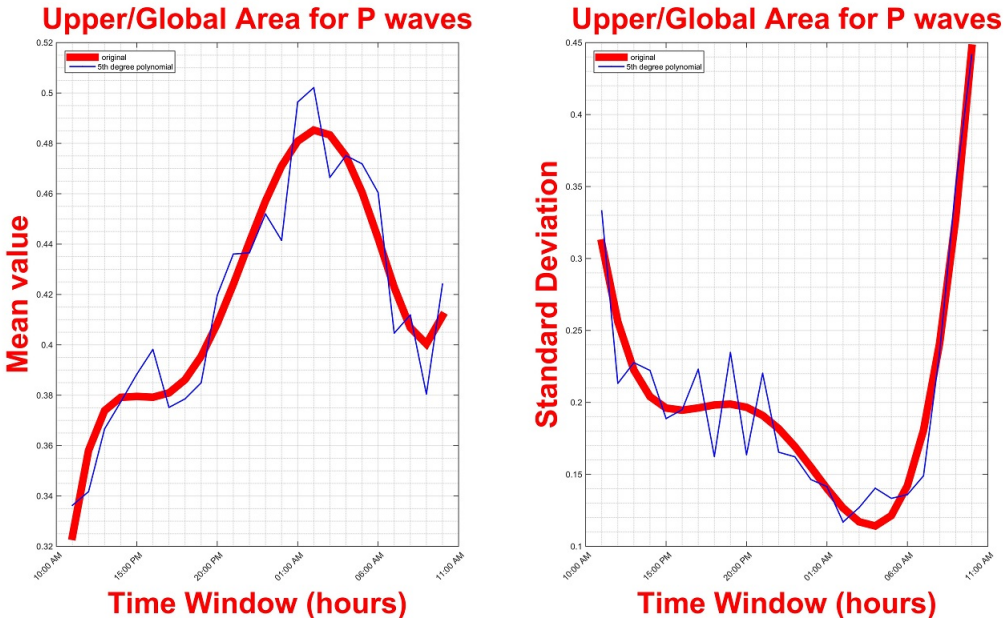

Figure 8.12: Mean and Standard Deviation values per hour for Upper/Global Area feature with polynomial fit

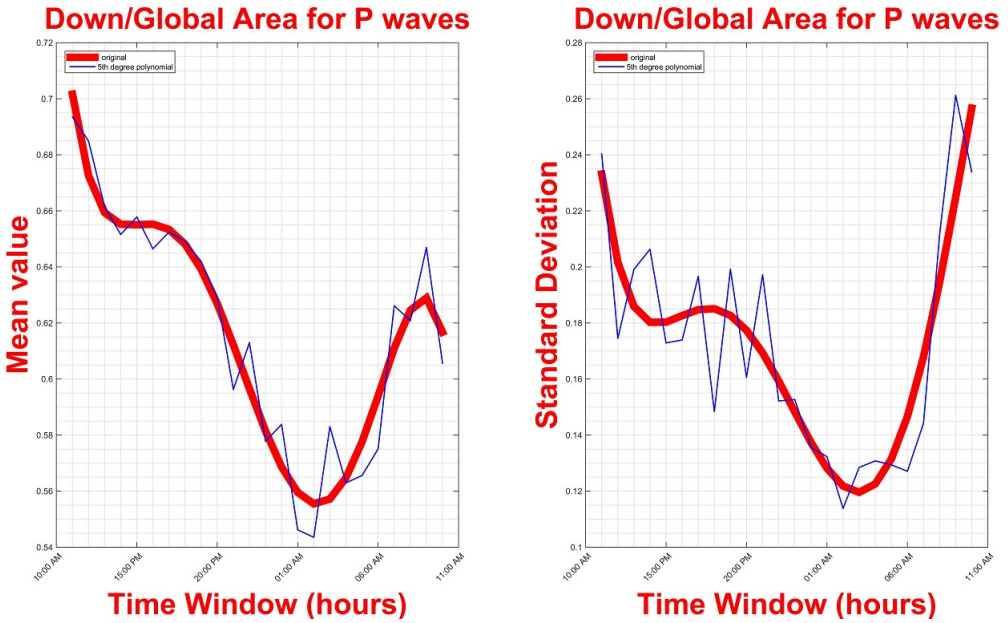

Figure 8.13: Mean and Standard Deviation values per hour for Down/Global Area feature with polynomial fit

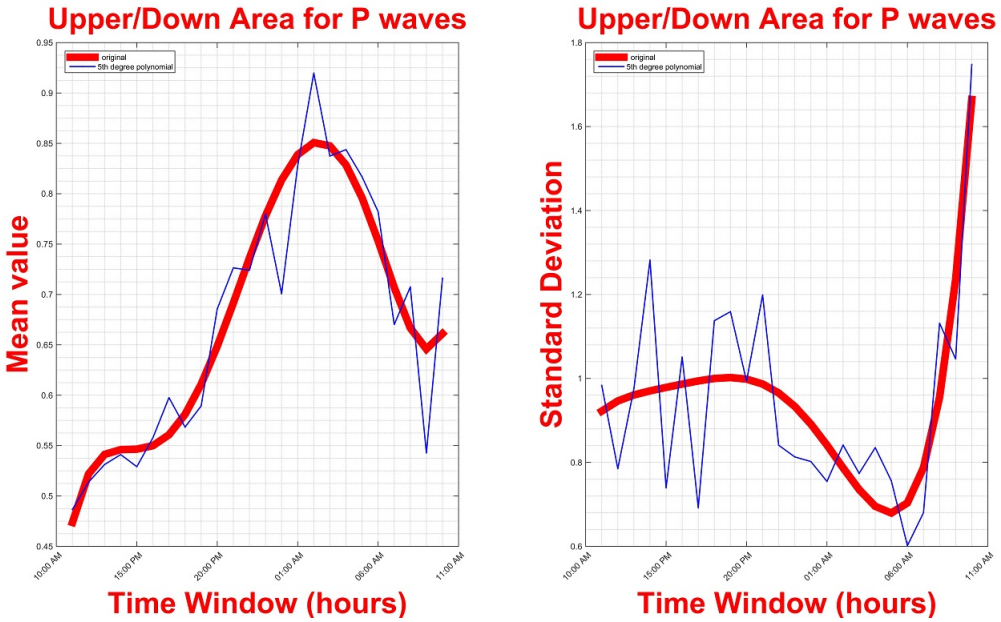

Figure 8.14: Mean and Standard Deviation values per hour for Upper/Down Area feature with polynomial fit

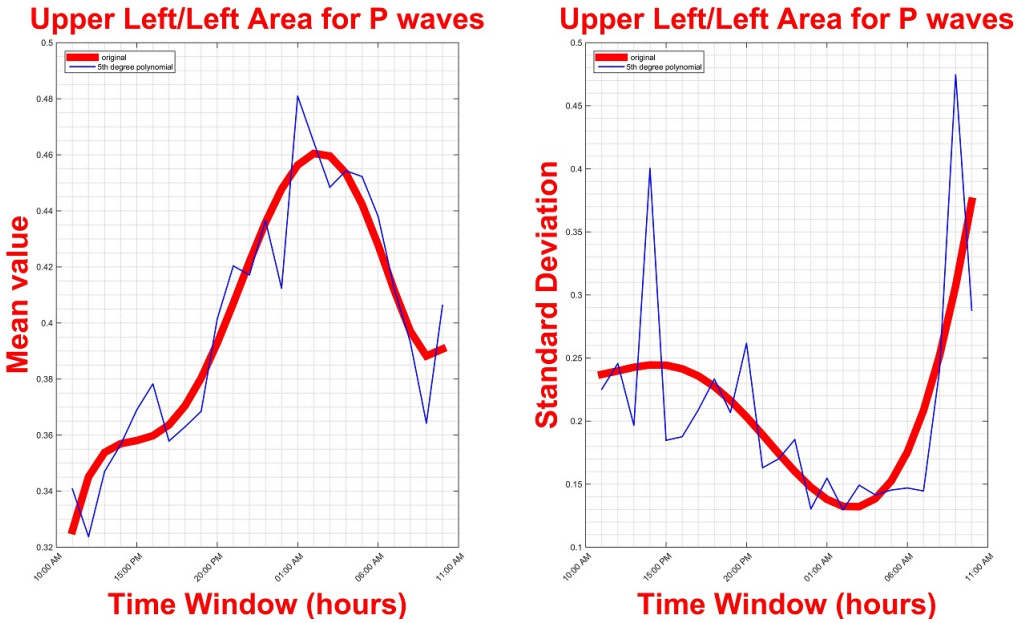

Figure 8.15: Mean and Standard Deviation values per hour for Upper Left/Left Area feature with polynomial fit

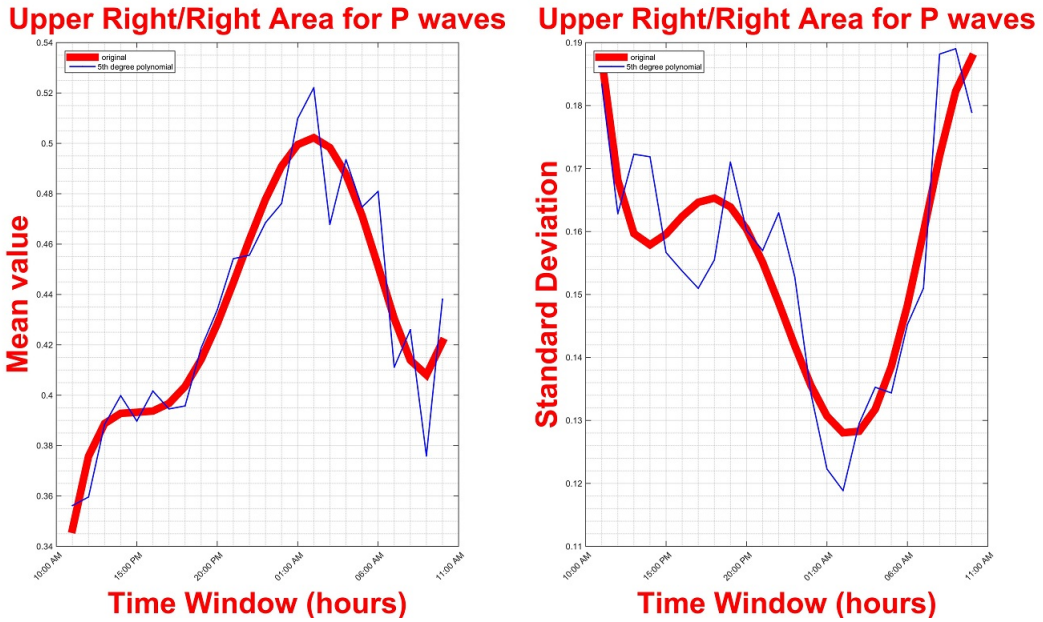

Figure 8.16: Mean and Standard Deviation values per hour for Upper Right/Right Area feature with polynomial fit

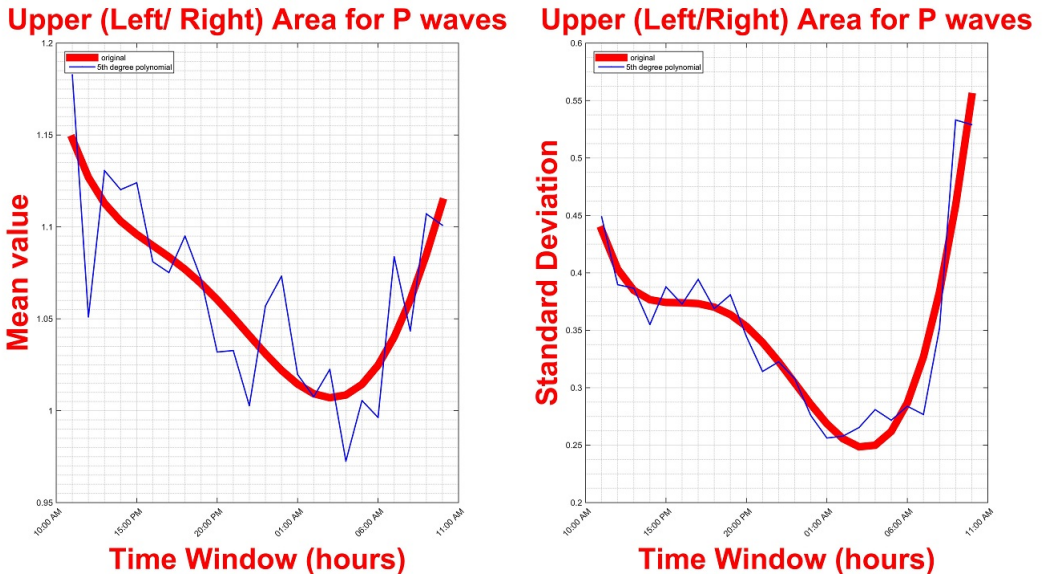

Figure 8.17: Mean and Standard Deviation values per hour for Upper Left/Upper Right Area feature with polynomial fit

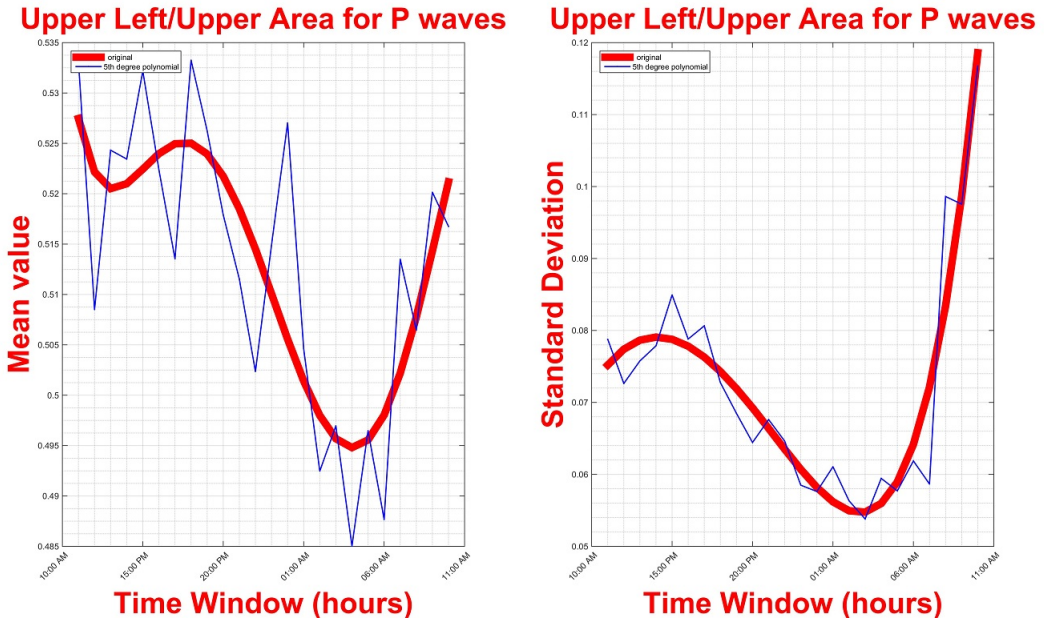

Figure 8.18: Mean and Standard Deviation values per hour for Upper Left/Upper Area feature with polynomial fit

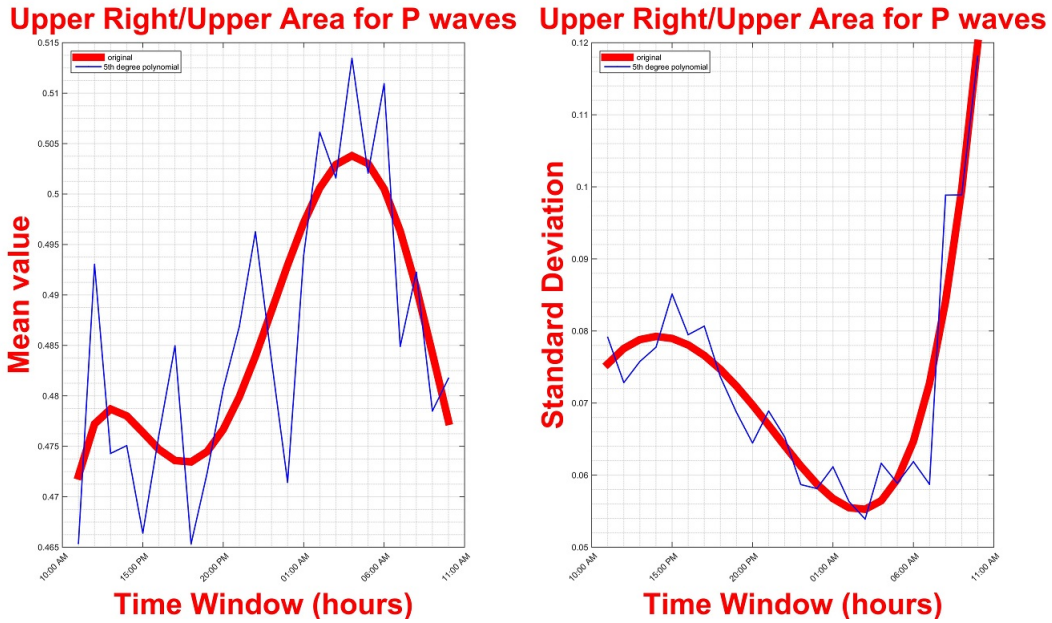

Figure 8.19: Mean and Standard Deviation values per hour for Upper Right/Upper Area feature with polynomial fit

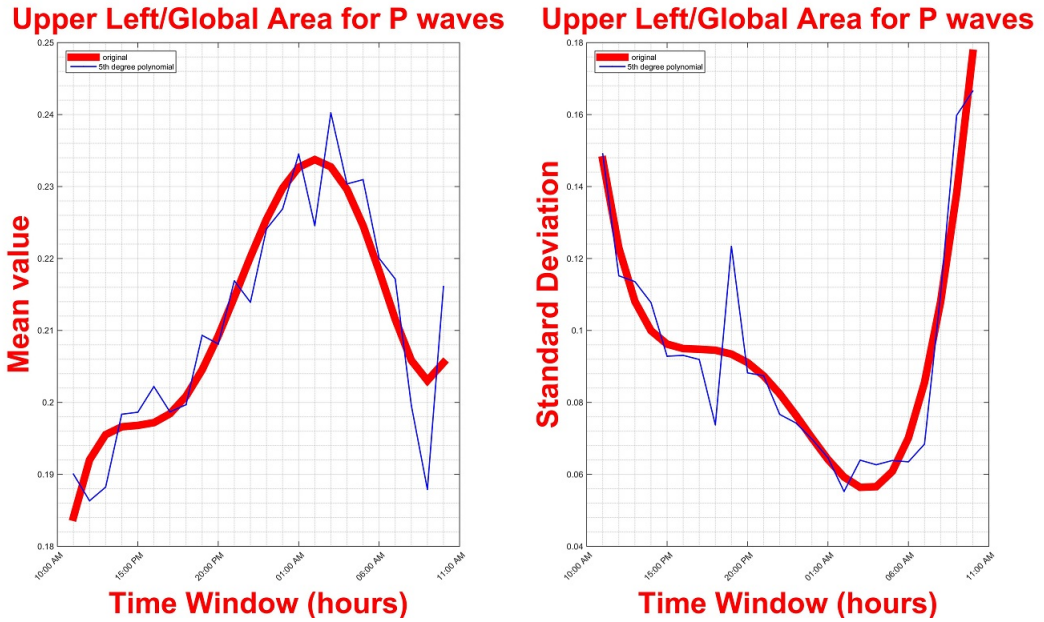

Figure 8.20: Mean and Standard Deviation values per hour for Upper Left/Global Area feature with polynomial fit

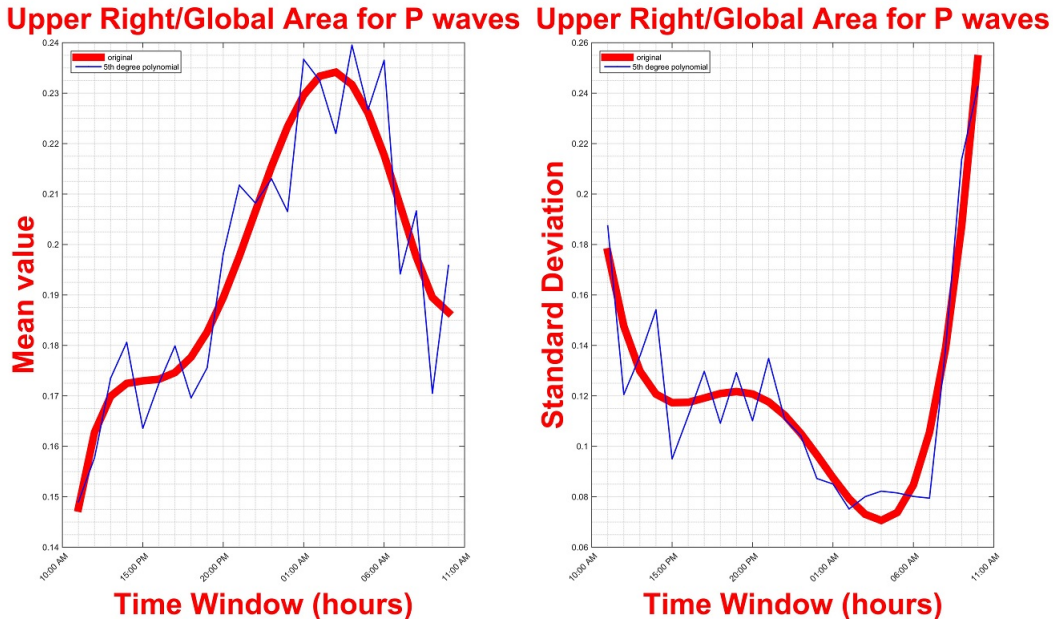

Figure 8.21: Mean and Standard Deviation values per hour for Upper Right/Global Area feature with polynomial fit

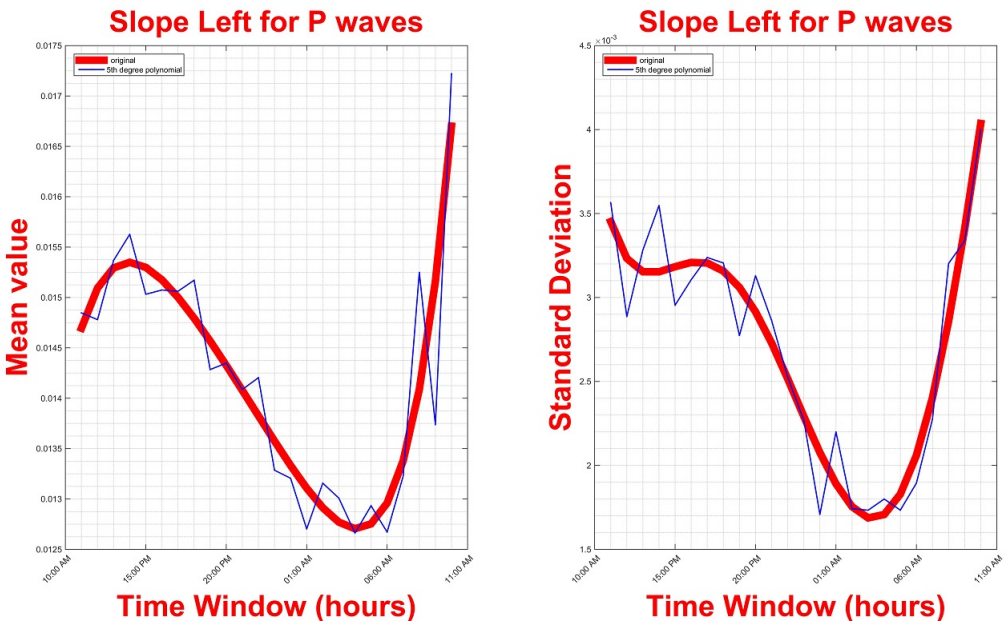

Figure 8.22: Mean and Standard Deviation values per hour for Left Slope feature with polynomial fit

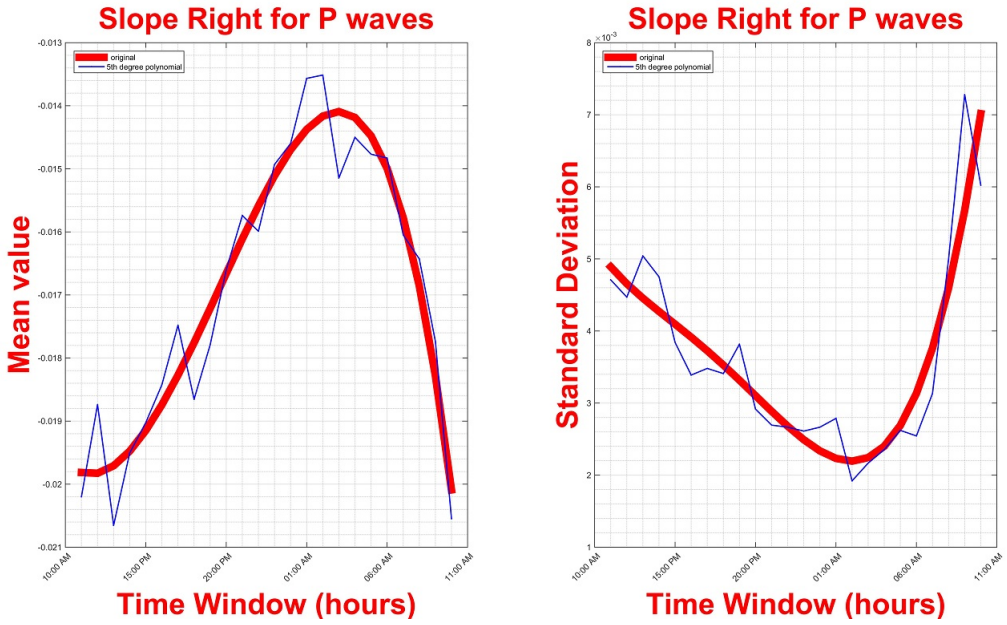

Figure 8.23: Mean and Standard Deviation values per hour for Right Slope feature with polynomial fit

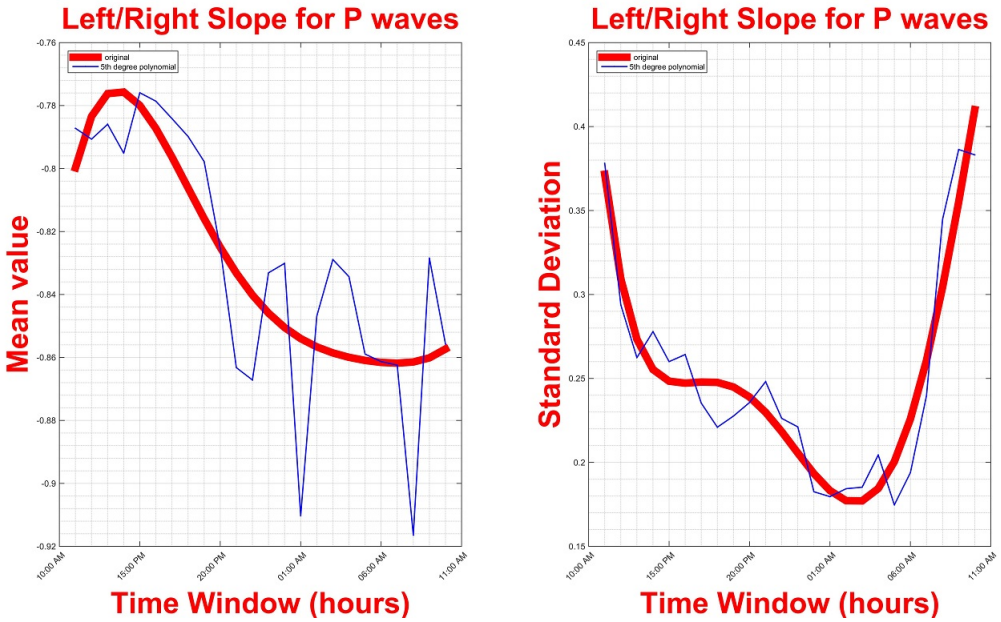

Figure 8.24: Mean and Standard Deviation values per hour for Left Slope/Right Slope feature with polynomial fit

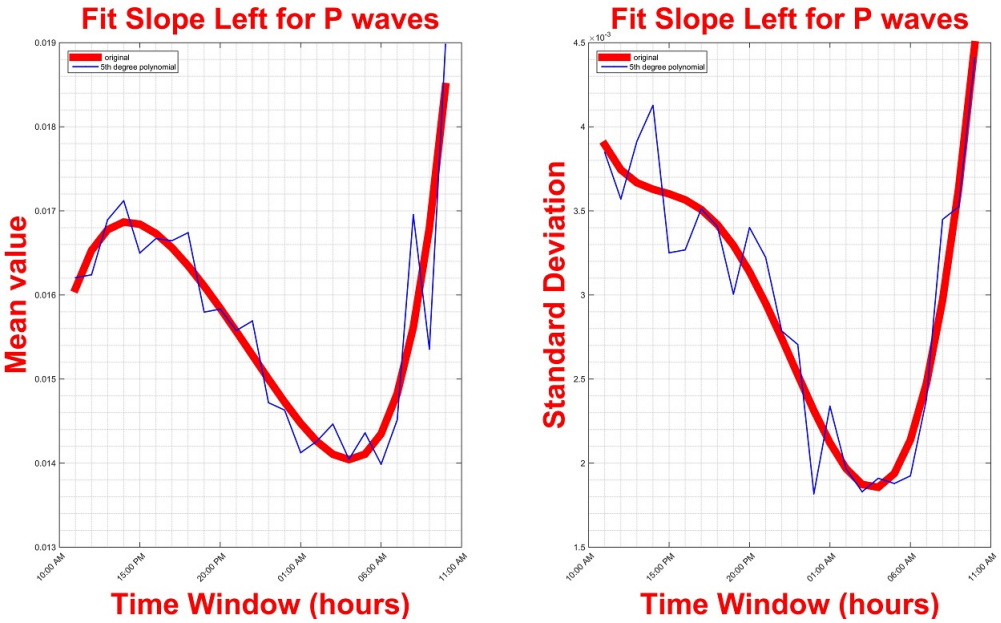

Figure 8.25: Mean and Standard Deviation values per hour for Fitting Left Slope feature with polynomial fit

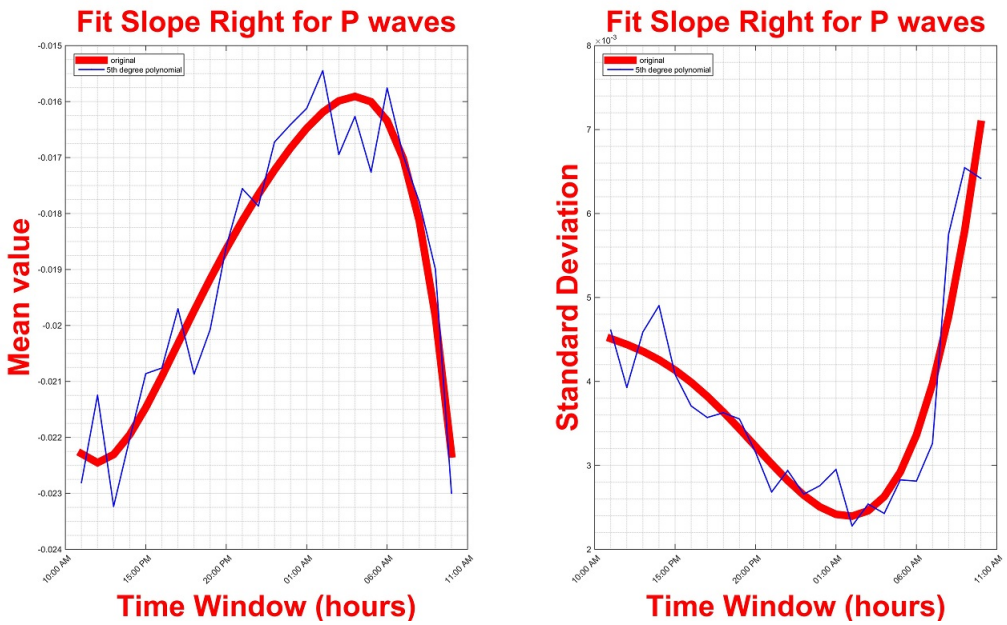

Figure 8.26: Mean and Standard Deviation values per hour for Fitting Right Slope feature with polynomial fit

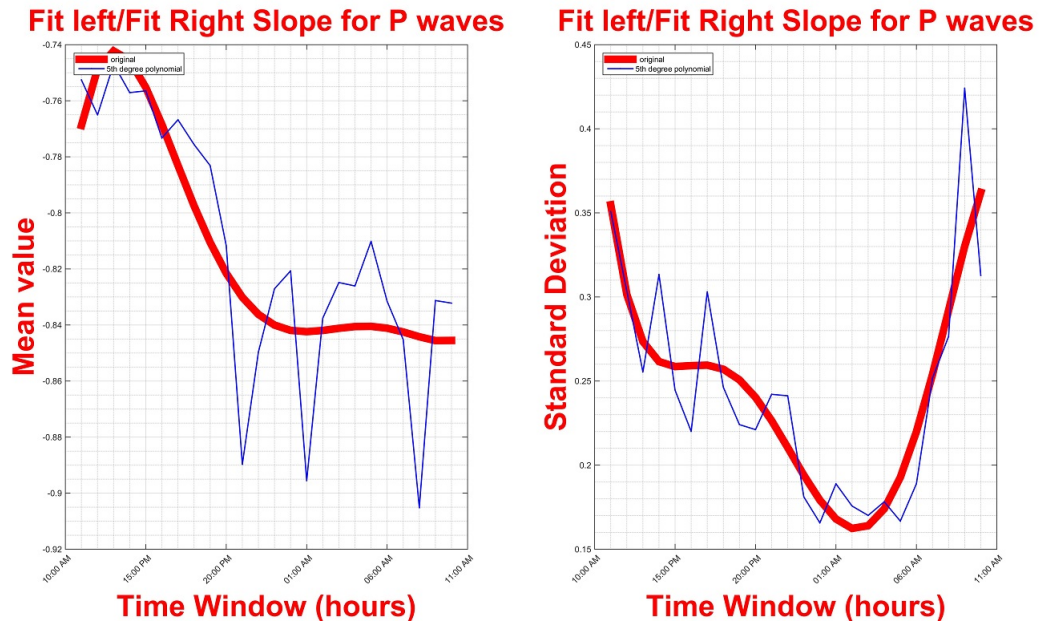

Figure 8.27: Mean and Standard Deviation values per hour for Fitting Left / Fitting Right Slope feature with polynomial fit

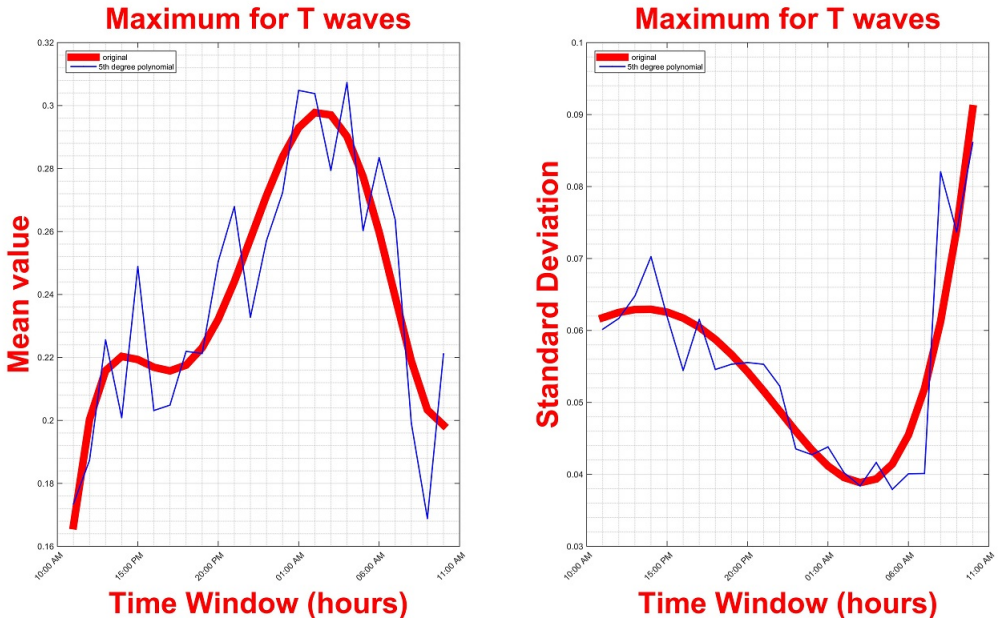

Figure 8.28: Mean and Standard Deviation values per hour for Maximum feature with polynomial fit

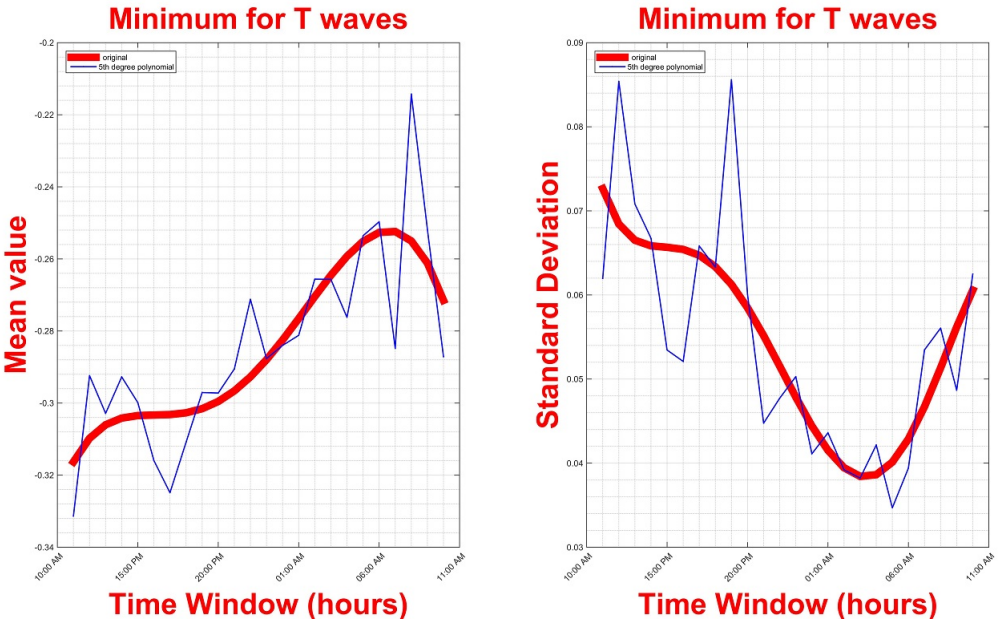

Figure 8.29: Mean and Standard Deviation values per hour for Minimum feature with polynomial fit

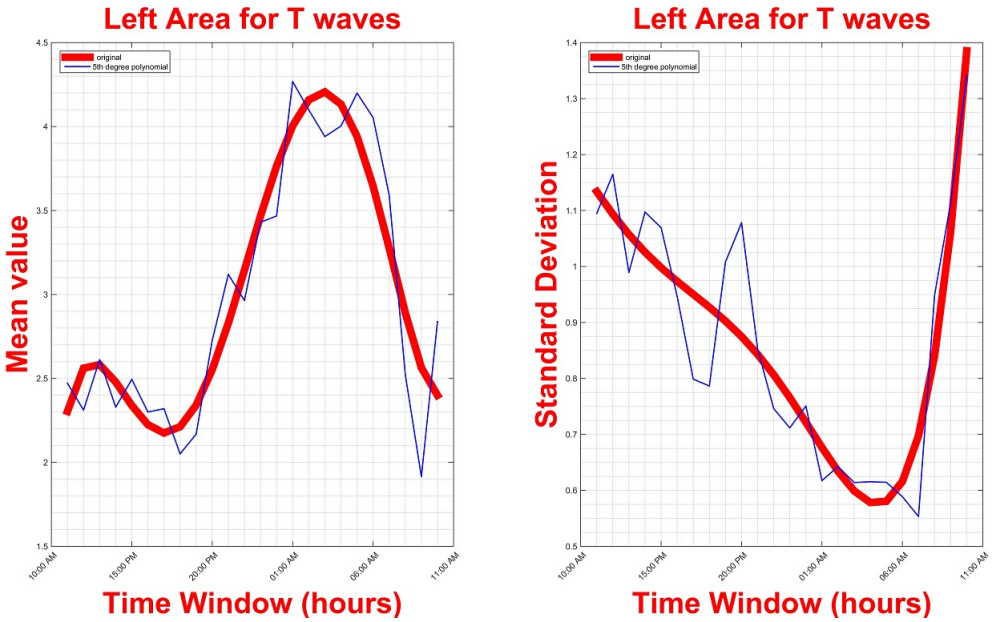

Figure 8.30: Mean and Standard Deviation values per hour for Left Area feature with polynomial fit

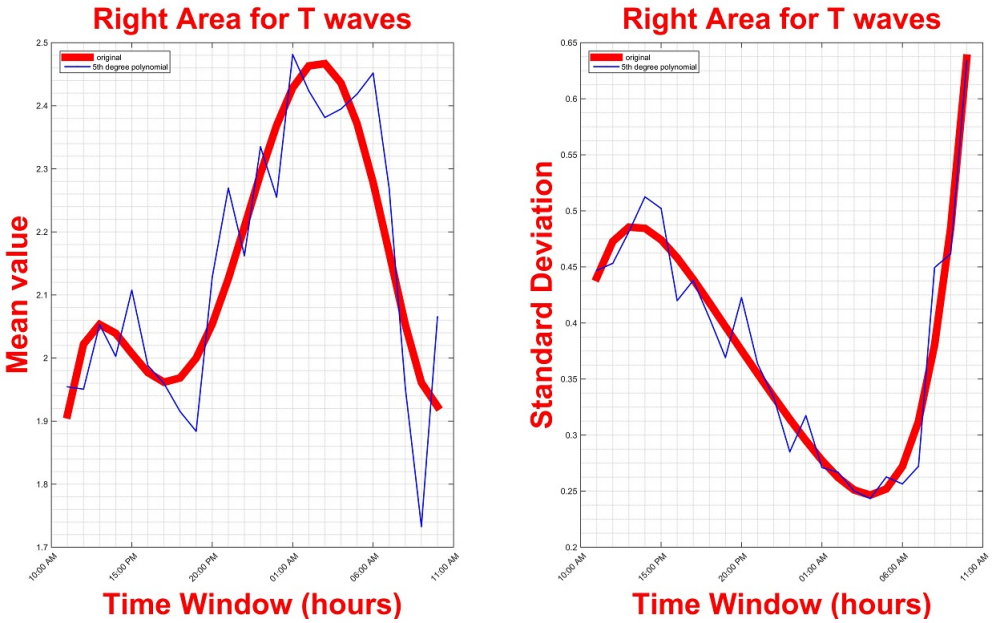

Figure 8.31: Mean and Standard Deviation values per hour for Right Area feature with polynomial fit

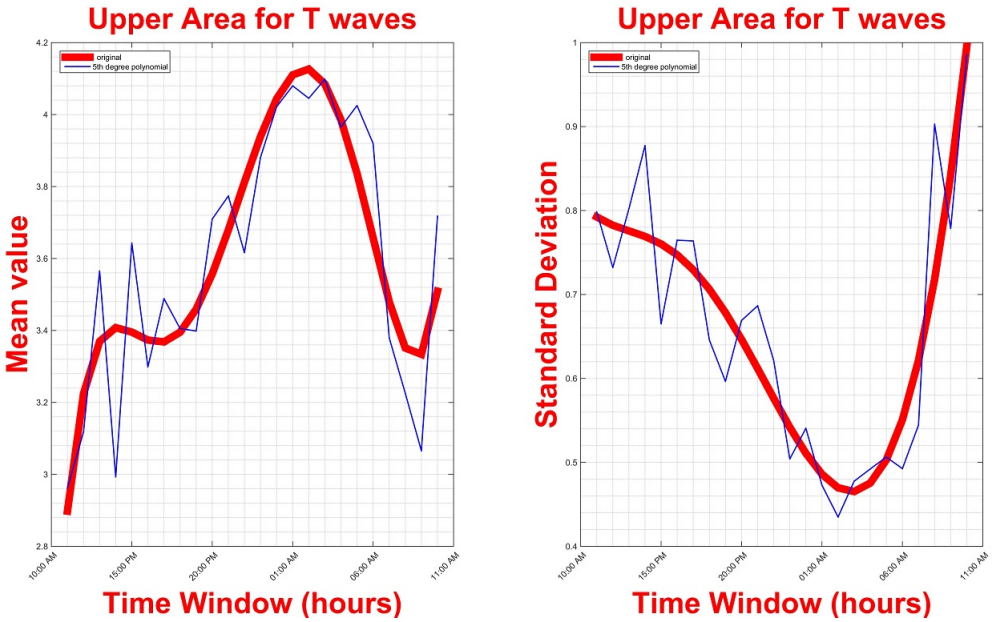

Figure 8.32: Mean and Standard Deviation values per hour for Upper Area feature with polynomial fit

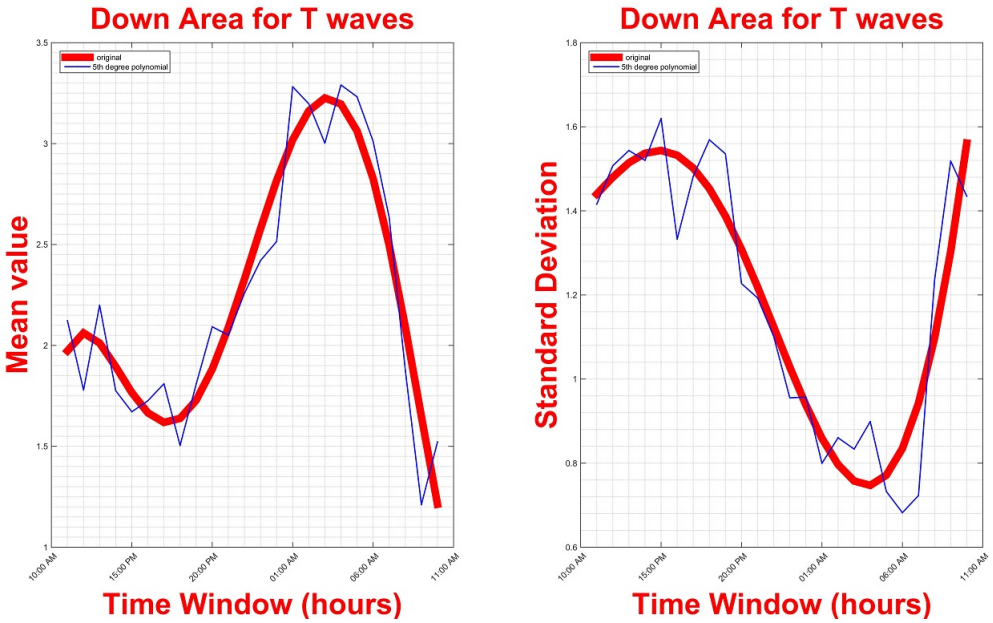

Figure 8.33: Mean and Standard Deviation values per hour for Down Area feature with polynomial fit

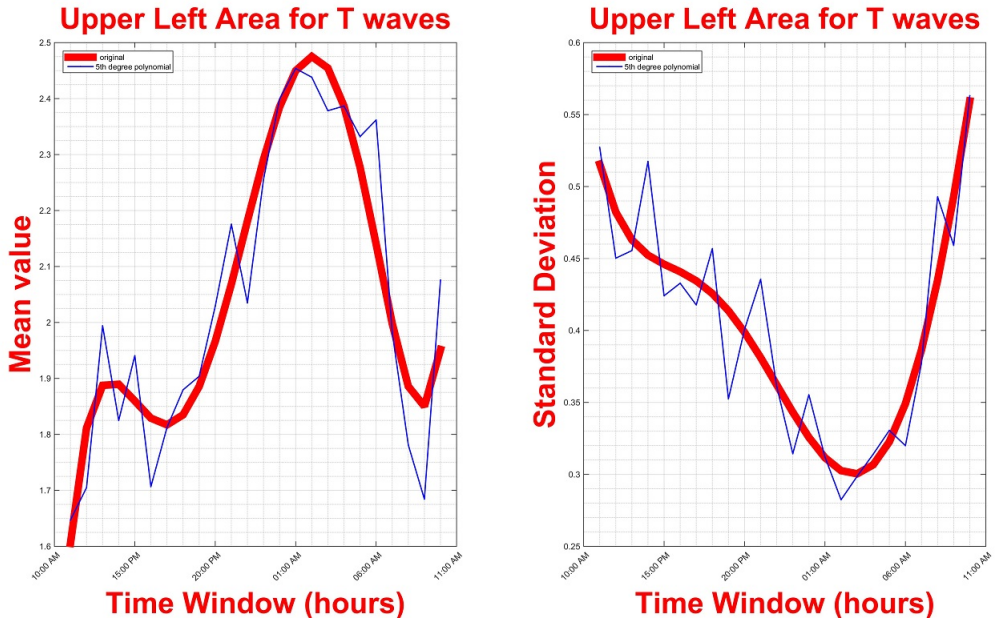

Figure 8.34: Mean and Standard Deviation values per hour for Upper Left Area feature with polynomial fit

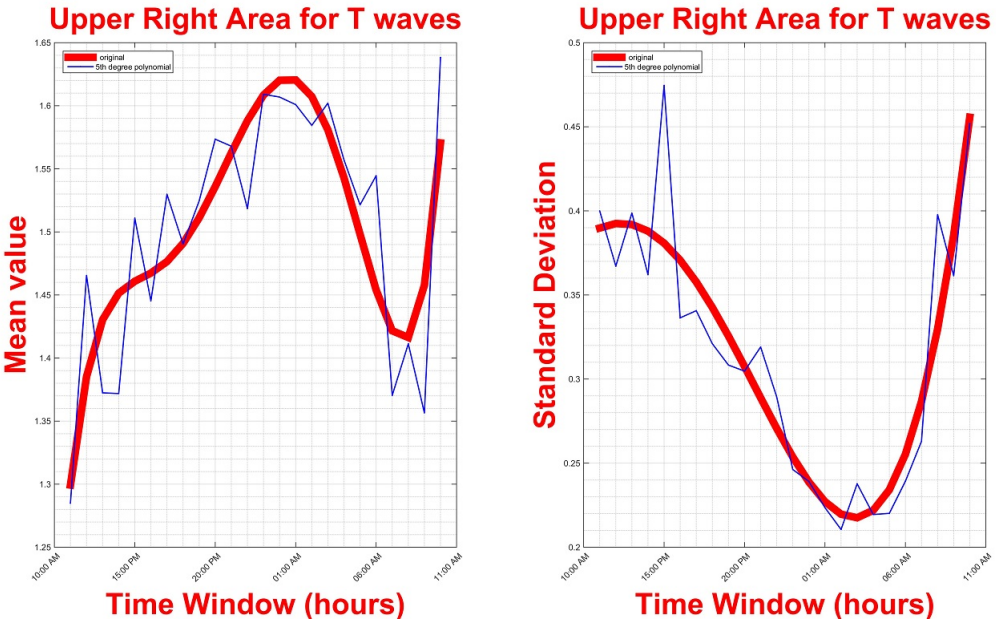

Figure 8.35: Mean and Standard Deviation values per hour for Upper Right Area feature with polynomial fit

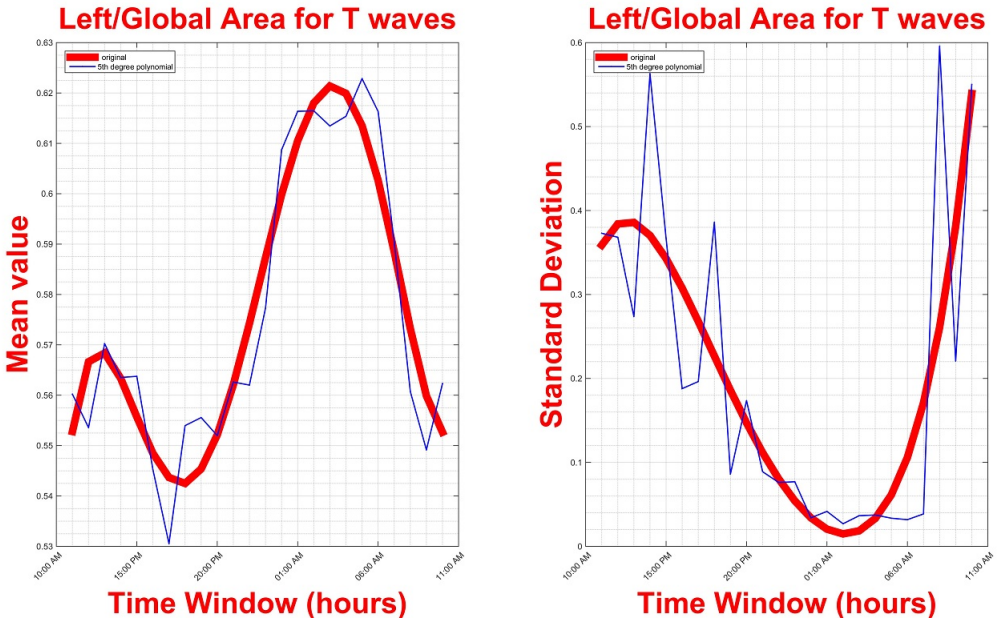

Figure 8.36: Mean and Standard Deviation values per hour for Left/Global Area feature with polynomial fit

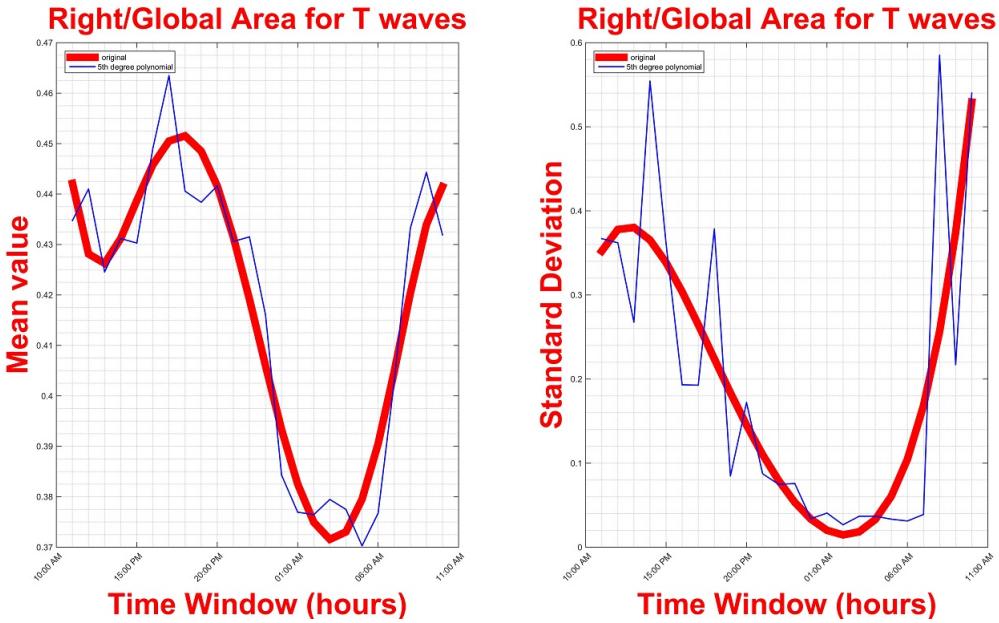

Figure 8.37: Mean and Standard Deviation values per hour for Right/Global Area feature with polynomial fit

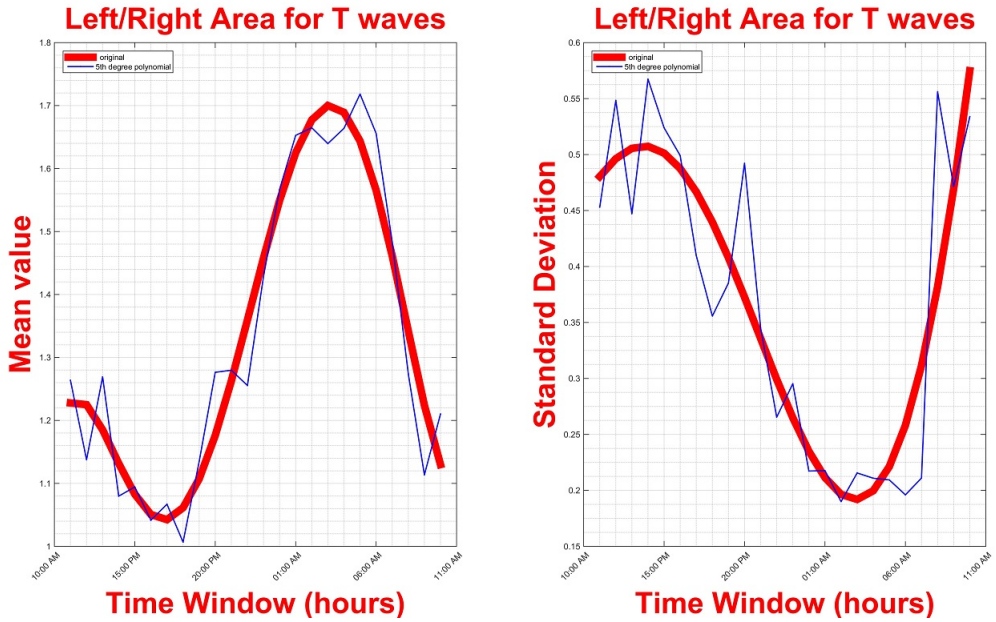

Figure 8.38: Mean and Standard Deviation values per hour for Left/Right Area feature with polynomial fit

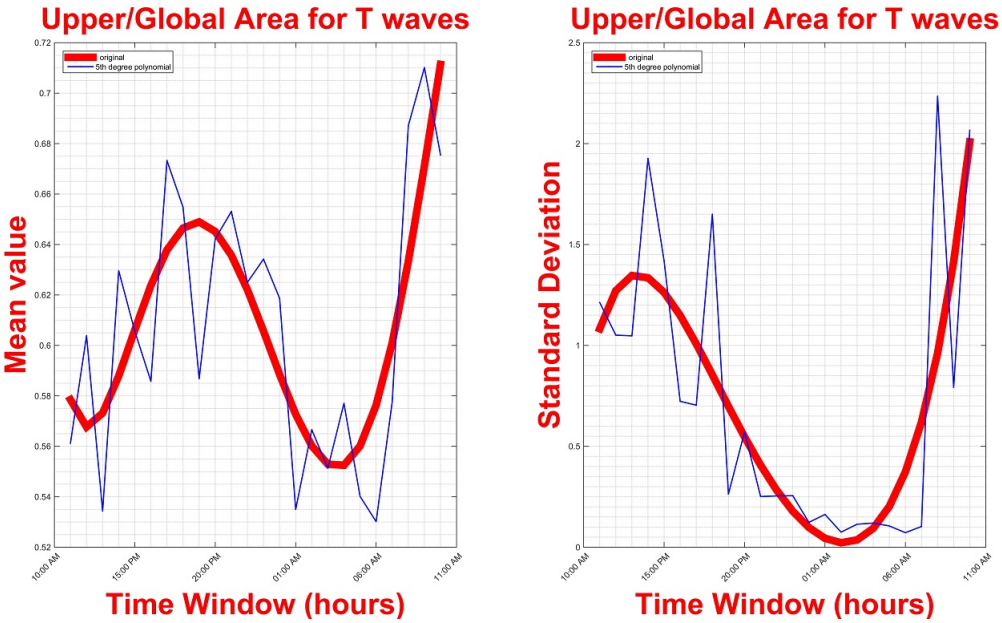

Figure 8.39: Mean and Standard Deviation values per hour for Upper/Global Area feature with polynomial fit

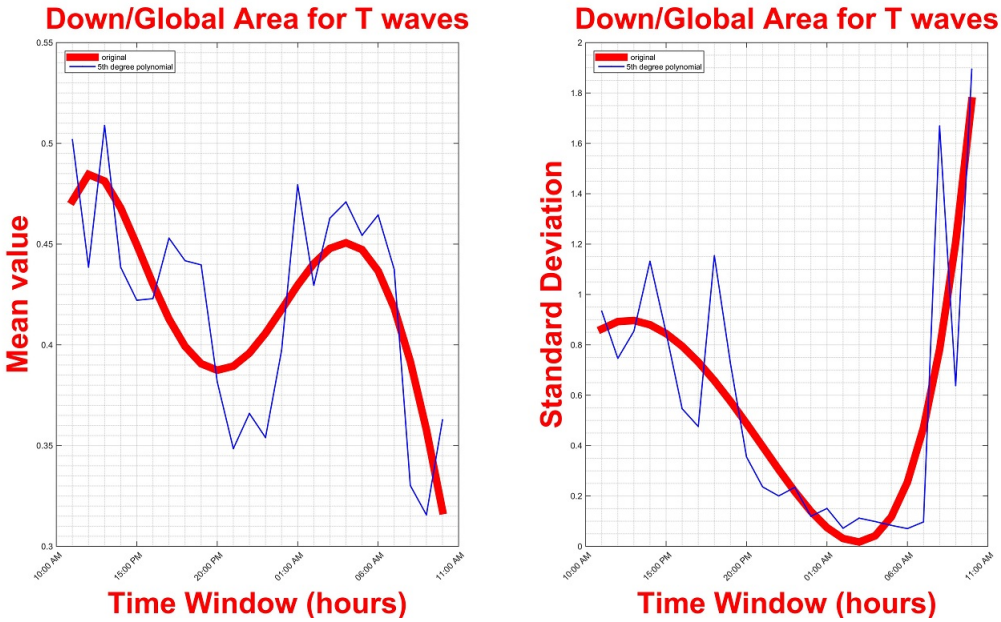

Figure 8.40: Mean and Standard Deviation values per hour for Down/Global Area feature with polynomial fit

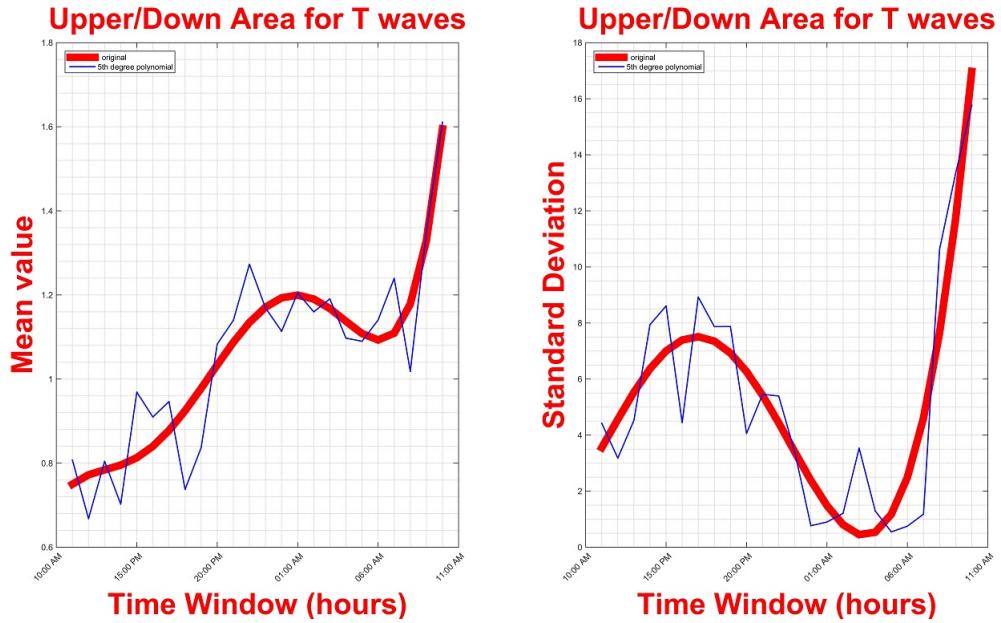

Figure 8.41: Mean and Standard Deviation values per hour for Upper/Down Area feature with polynomial fit

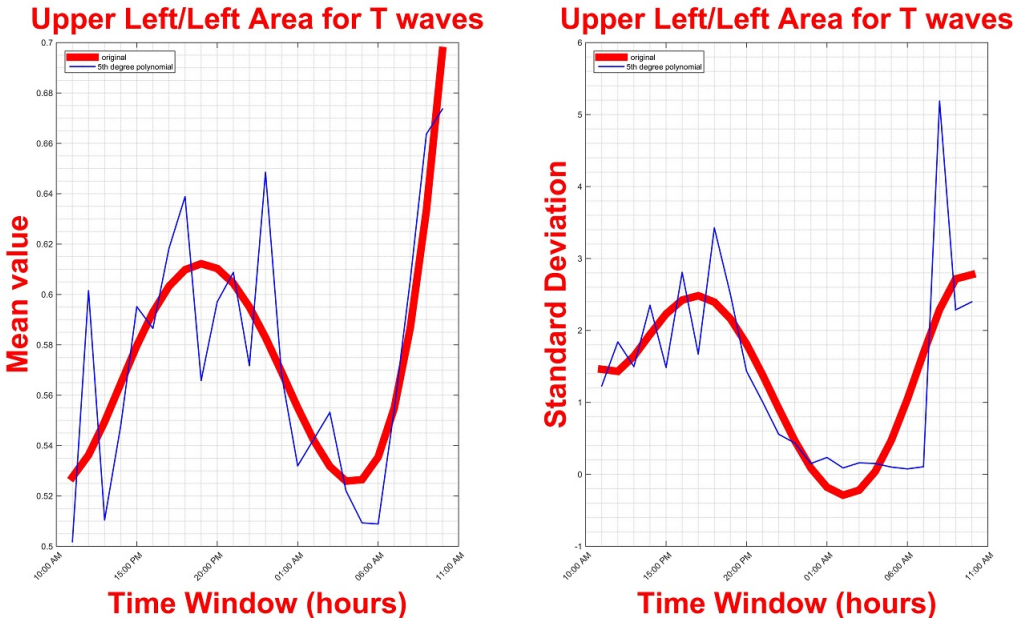

Figure 8.42: Mean and Standard Deviation values per hour for Upper Left/Left Area feature with polynomial fit

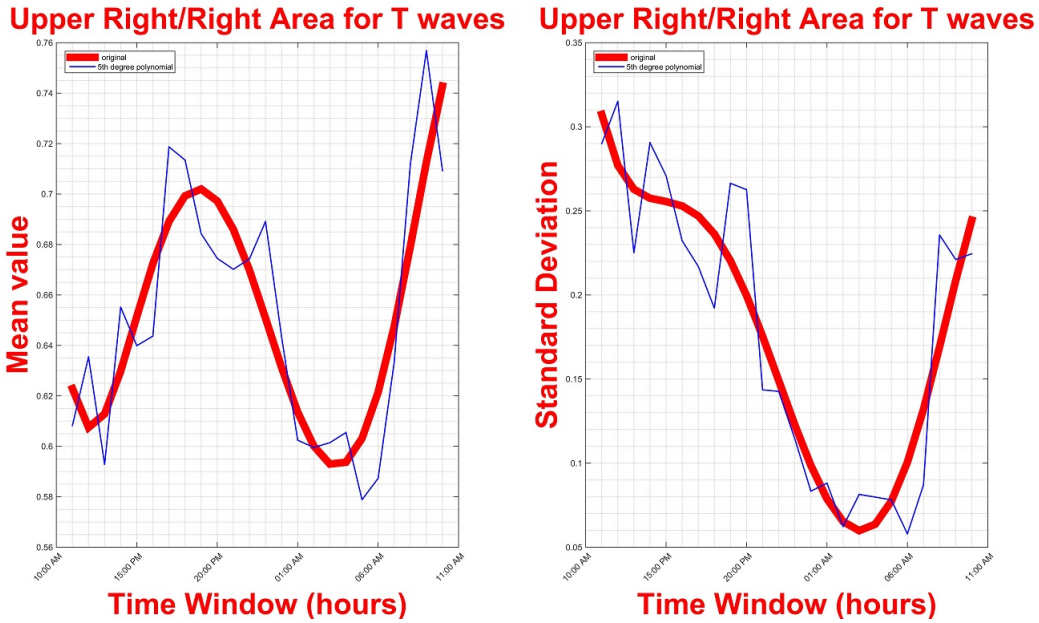

Figure 8.43: Mean and Standard Deviation values per hour for Upper Right/Right Area feature with polynomial fit

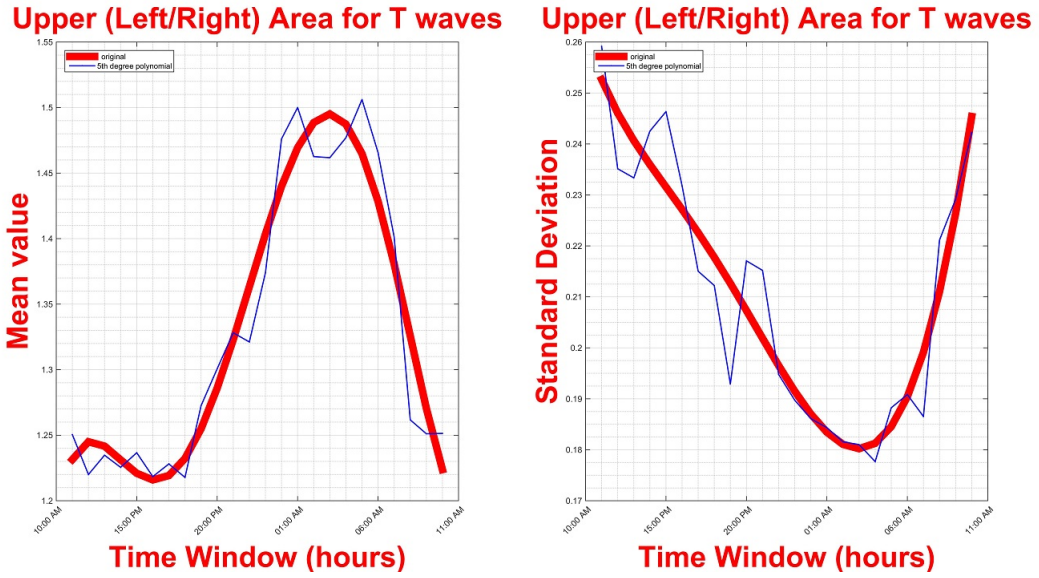

Figure 8.44: Mean and Standard Deviation values per hour for Upper Left/Upper Right Area feature with polynomial fit

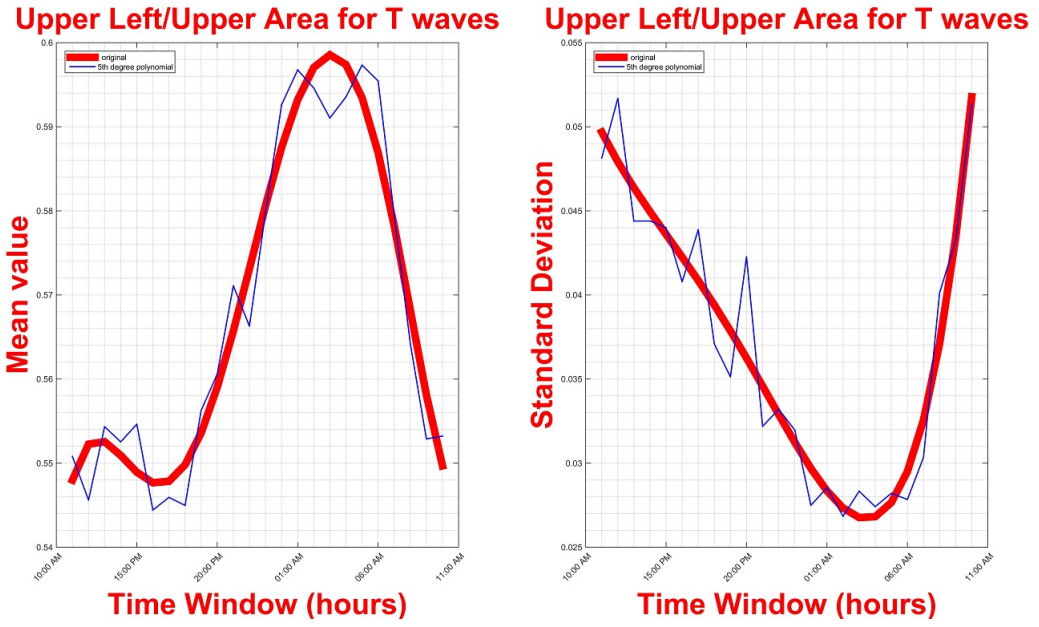

Figure 8.45: Mean and Standard Deviation values per hour for Upper Left/Upper Area feature with polynomial fit

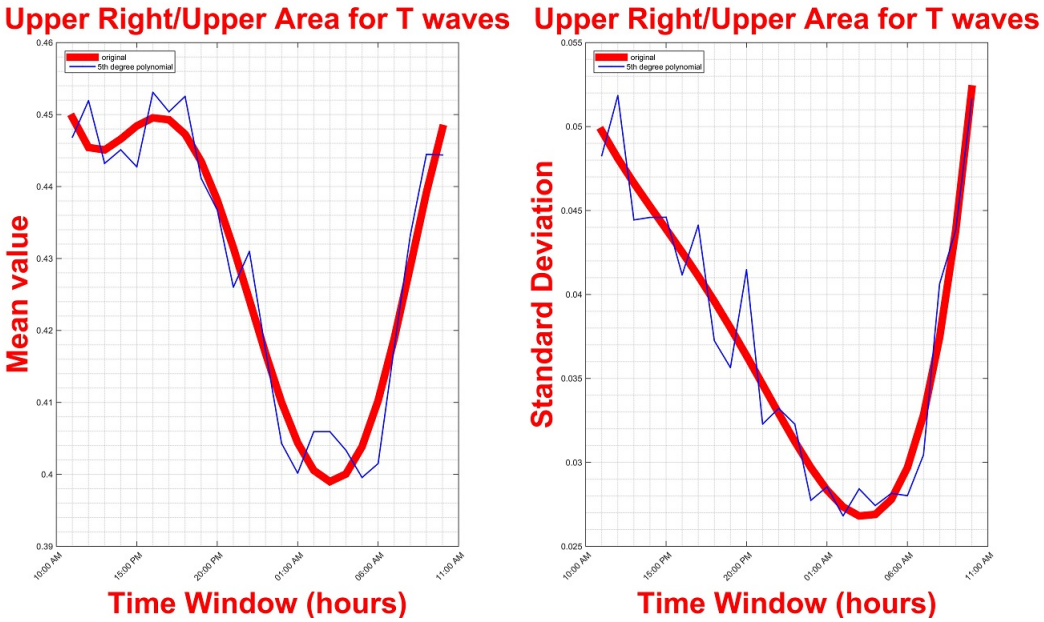

Figure 8.46: Mean and Standard Deviation values per hour for Upper Right/Upper Area feature with polynomial fit

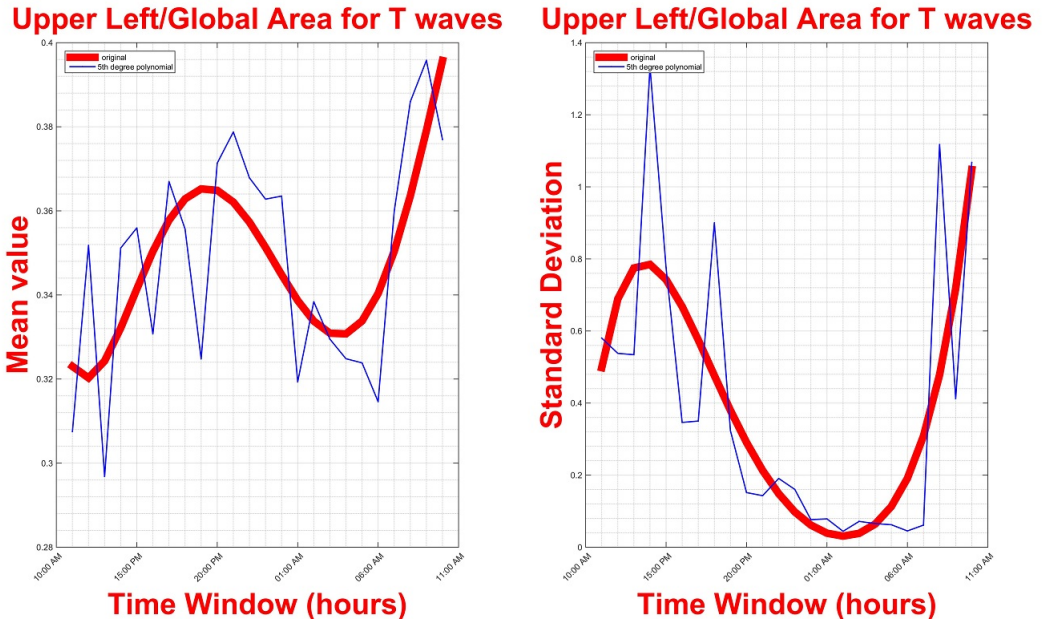

Figure 8.47: Mean and Standard Deviation values per hour for Upper Left/Global Area feature with polynomial fit

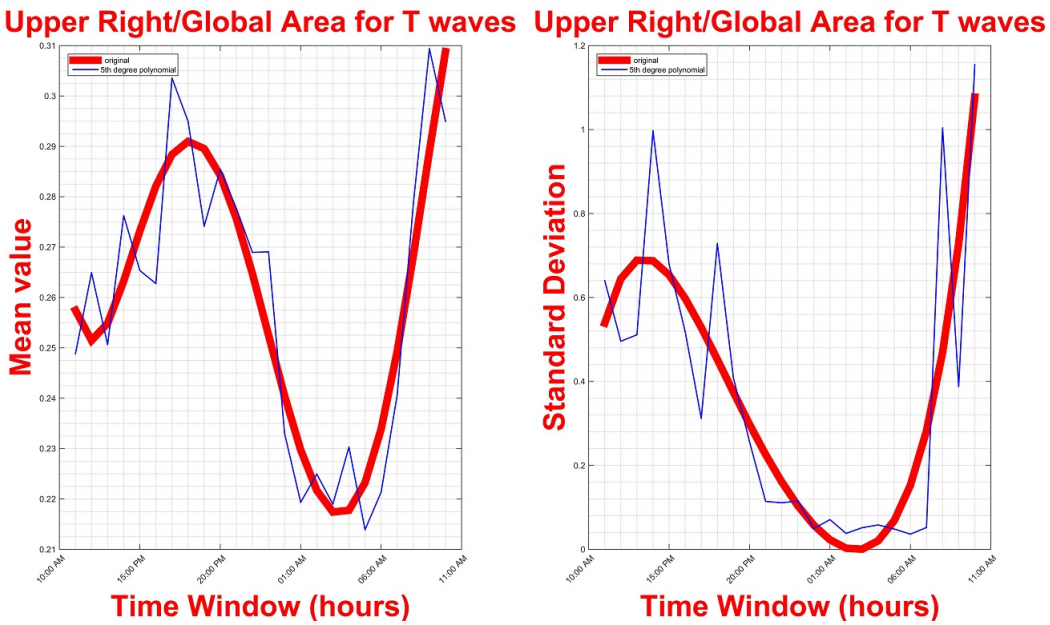

Figure 8.48: Mean and Standard Deviation values per hour for Upper Right/Global Area feature with polynomial fit

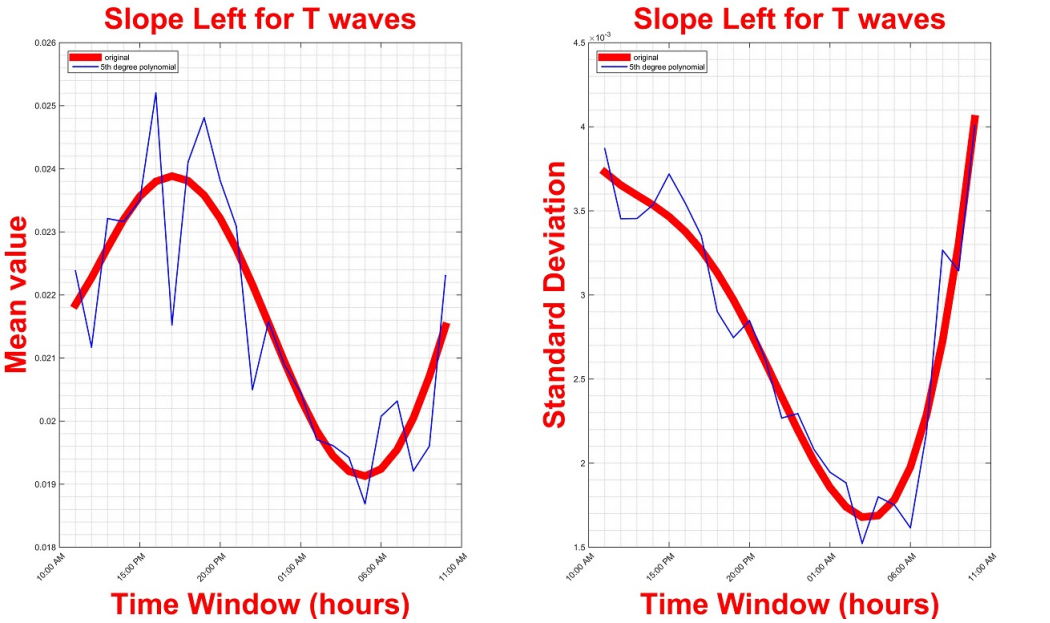

Figure 8.49: Mean and Standard Deviation values per hour for Left Slope feature with polynomial fit

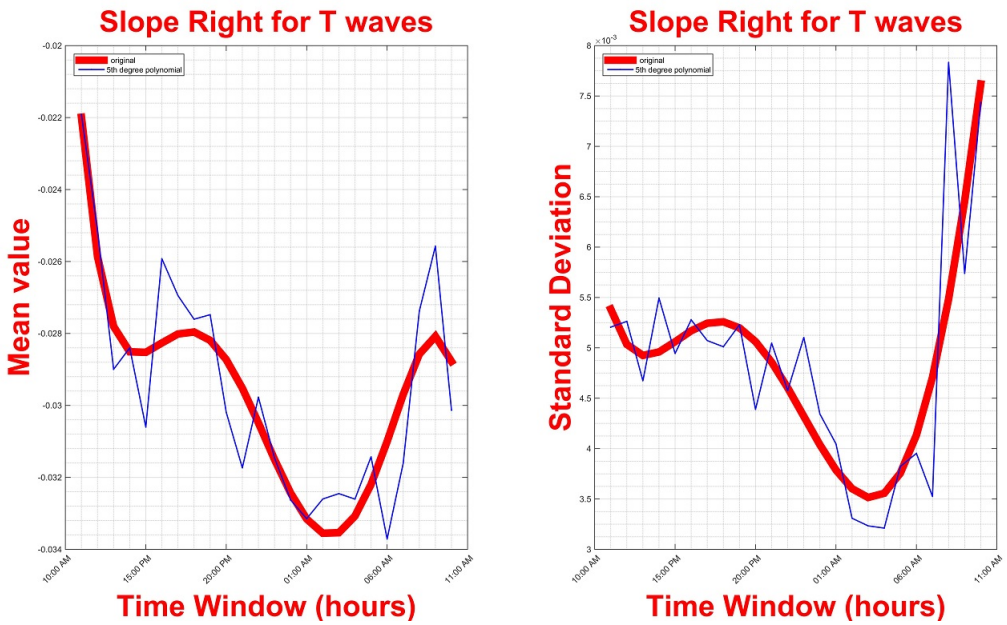

Figure 8.50: Mean and Standard Deviation values per hour for Right Slope feature with polynomial fit

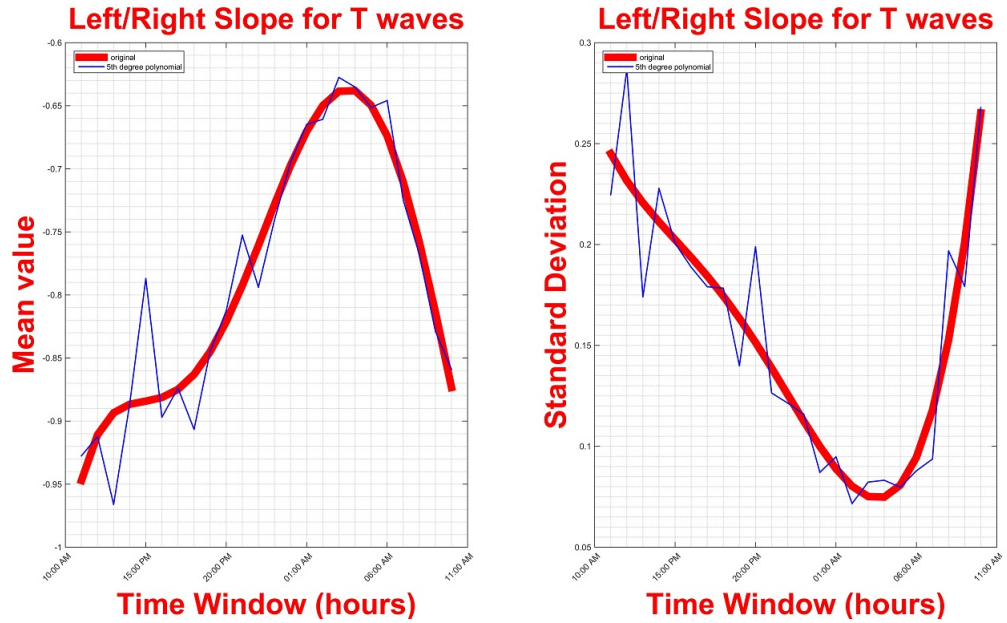

Figure 8.51: Mean and Standard Deviation values per hour for Left Slope/Right Slope feature with polynomial fit

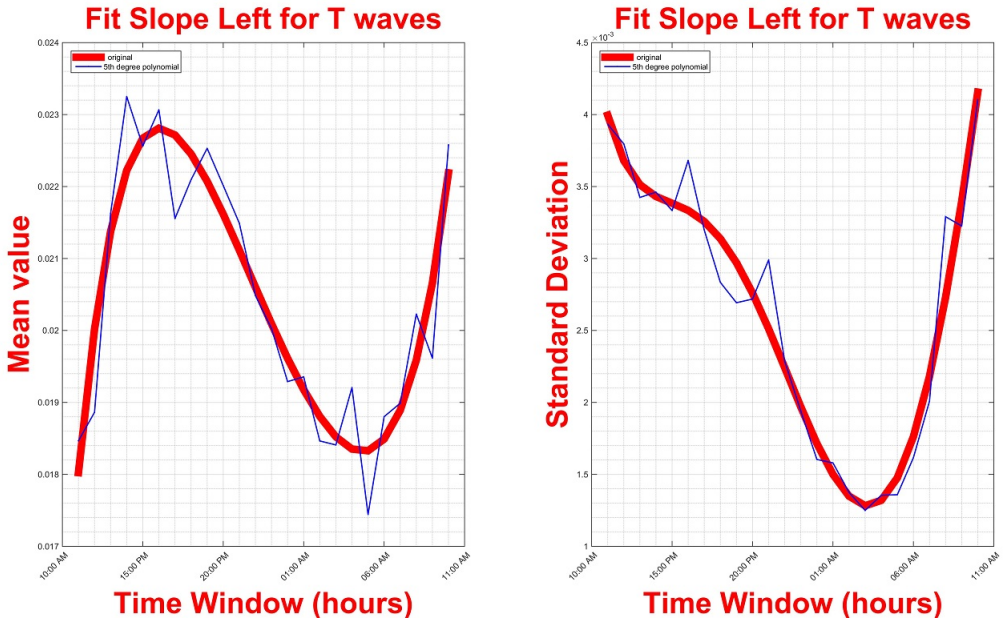

Figure 8.52: Mean and Standard Deviation values per hour for Fitting Left Slope feature with polynomial fit

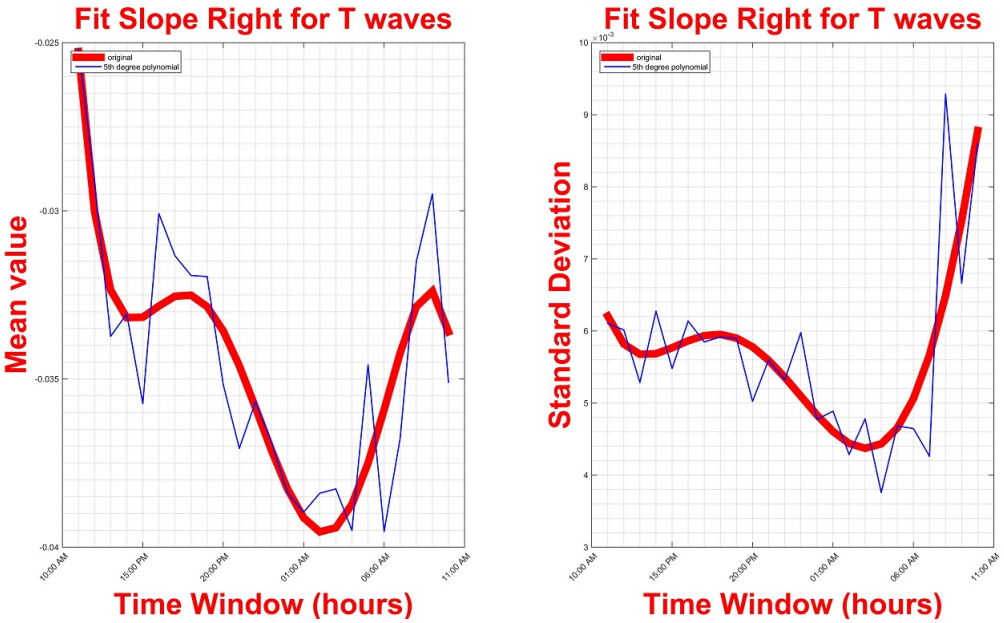

Figure 8.53: Mean and Standard Deviation values per hour for Fitting Right Slope feature with polynomial fit

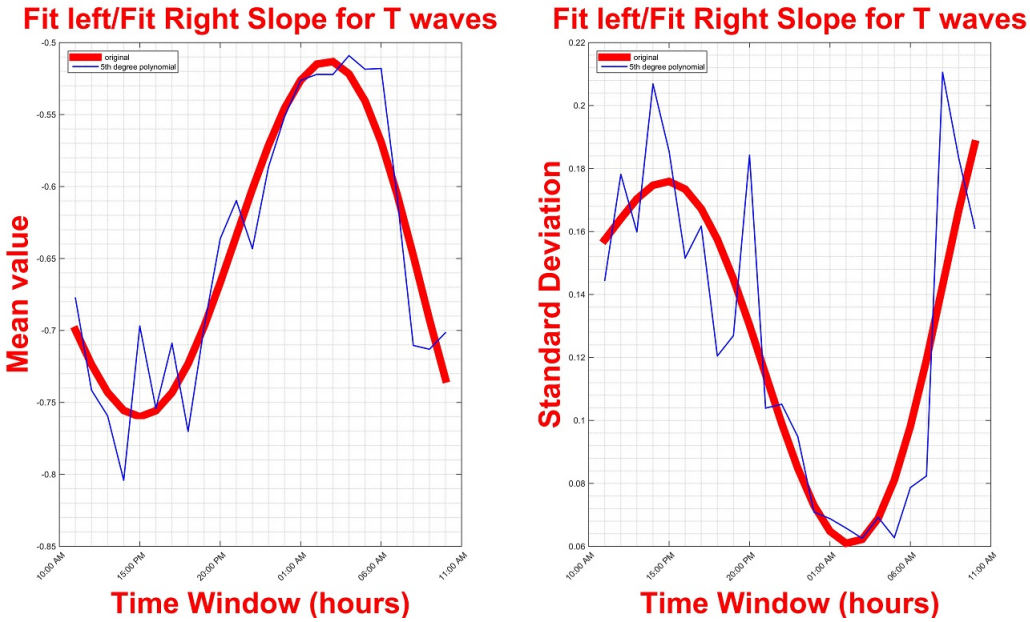

Figure 8.54: Mean and Standard Deviation values per hour for Fitting Left / Fitting Right Slope feature with polynomial fit

Short Vita

Dimitrios Zavantis was born in Larisa, Greece in 1989. He was admitted at Mathematics Department of the University of Ioannina in 2007. He recieved his BSc degree in Mathematics (minor Computer Science) in 2012. Currently a Postgraduate student at the Department of Computer Science & Engineering of the University of Ioannina and a member of the Information Processing & Analysis Research Group. His main research interests include Heart Rate Variability, Circadian Rhythms and Signal processing.