

HANDLING MISSINGNESS, OUTLIERS AND MODELING IN LONGITUDINAL DATA ANALYSIS

Maroua BEN GHOU

PhD THESIS

Department of Statistics

Supervisor: Prof. Dr. Berna YAZICI

**Eskişehir
Anadolu University
Graduate School of Science
November 2019**

FINAL APPROVAL FOR THESIS

This thesis titled “**HANDLING MISSINGNESS, OUTLIERS AND MODELING IN LONGITUDINAL DATA ANALYSIS**” has been prepared and submitted by **Maroua BEN GHOUL** in partial fulfilment of the requirements in “Anadolu University Directive on Graduate Education and Examination” for the Degree of Doctor of Philosophy (PhD) in Statistic Department has been examined and approved on 15/11/2019.

Committee Members

Signature

Member (Supervisor)	: Prof. Dr. Berna YAZICI
Member	: Prof. Dr. Alaaddin ŞAMILOV
Member	: Prof. Dr. İlker ERCAN
Member	: Prof. Dr. Özlem ALPU
Member	: Doç. Dr. Betül KAN KILINÇ

Prof. Dr. Murat TANIŞLI
Director of Institute

ABSTRACT

HANDLING MISSINGNESS, OUTLIERS AND MODELING IN LONGITUDINAL DATA ANALYSIS

Maroua BEN GHOU

Department of Statistics
Anadolu University, Graduate School of Science, November 2019

Supervisor: Prof. Dr. Berna YAZICI

Longitudinal data consists in gathering several observations of the same subjects intermittently over time. Attrition, outliers and complexity of modeling are common issues in longitudinal data. Therefore, this dissertation outlines those issues and proposes different approaches to overcome them by following three main pillars. First pillar emphasises the prominence of missingness mechanisms and suggests a novel algorithm to treat missing data via Multilayer Perceptron (MLP) with comparison to the ad hoc methods and Expectation Maximum (EM) algorithm. Second pillar consists in presenting outliers as a friendly subject in statistical data not a misleading dilemma, via proposing two novel algorithms using wavelet decomposition within subjects and across subjects along with applying the winsorisation approach within subjects. Last pillar concentrates on modeling via constructing a semiparametric model that combines parametric and nonparametric features. For the nonparametric part of the model, smoothing approaches are required. This research proposes wavelet analysis to smooth data.

To examine the efficiency of the proposed algorithms, a real longitudinal dataset and a generated one, are utilized. The results revealed that wavelet decomposition has an impressive capacity as a smoothing approach and as a *microscope* figuring out the outliers and handling them without losing the originality of the data features. Also, the novel algorithm related to missing data imputation via the output predictions of MLP showed valuable results better than the ad hoc imputation methods and with very slight difference from the EM algorithm.

Keywords: Longitudinal data, Semiparametric model, Missing data, Missingness mechanisms, Outliers, Wavelet analysis, Neural network.

ÖZET

LONGITUDINAL VERİ ANALİZİNDE EKSİK GÖZLEM, AYKIRI DEĞER VE MODELLEME ÜZERİNE ÇALIŞMA

Maroua BEN GHOU

İstatistik Anabilim Dalı
Anadolu Üniversitesi, Fen Bilimleri Enstitüsü, Kasım 2019

Danışman: Prof. Dr. Berna YAZICI

Longitudinal veriler aynı konulara ait birkaç gözlemin zaman zaman aralıklarla toplanması ile oluşur. Bu verilerde; aşınma, aykırı değerler ve modellemenin karmaşıklığı yaygın görülen sorunlardır. Dolayısıyla, bu tezde bu sorunlar ana hatlarıyla ele alınmaktadır ve bu sorunları ortadan kaldırmak için üç ana bölümde farklı yaklaşımlar önermektedir. İlk bölümde eksik veri mekanizmalarının önemini vurgulanmaktadır ve ad hoc yöntemlere ve Maksimum Beklentı (MB) algoritmasına kıyasla Çok Katmanlı Algılayıcı (ÇKA) ile eksik verileri ele almak için yeni bir algoritma önerilmektedir. İkinci bölüm, aykırı değeri yanıltıcı bir veri değil istatistiksel verilerde, denekler içinde dalgacık ayrıştırması kullanan iki yeni algoritma önererek ve denekler arasında winsorisation yaklaşımını uygulayarak, kullanım kolaylığı olan bir algoritma sunulmasından ibarettir. Son bölüm, parametrik ve parametrik olmayan özellikleri birleştiren bir semiparametrik model inşa ederek modellemeye odaklanmaktadır. Modelin parametrik olmayan kısmı için, düzleştirme yaklaşımları gereklidir. Bu çalışmada, modelin düzeltilme için dalgacık analizi önerilmektedir.

Önerilen algoritmaların etkinliğini araştırmak için gerçek bir longitudinal veri kümesi ve bu kümeden üretilen bir veri seti kullanılmıştır. Sonuçlar, dalgacık ayrışmasının, pürüzsüzleştirici bir yaklaşım, aykırı noktaları belirleyen ve veri özelliklerinin orijinallliğini kaybetmeden bunları kullanma konusunda mikroskop gibi çalışan bir metod olduğunu ortaya koymuştur. Ayrıca, ÇKA'nın çıktı tahminleri üzerinden eksik veri ataması ile ilgili yeni algoritma, ad hoc yöntemlerinden daha iyi ve BM algoritmasından küçük bir farkla daha güçlü sonuçlar göstermiştir.

Anahtar Sözcükler: Longitudinal veriler, Semiparametrik model, Eksik veriler, Eksiklik mekanizmaları, Aykırı değerler, Dalgacık analizi, Sinir ağı.

ACKNOWLEDGMENTS

Praise be to GOD, for all the grace and favour he always gives me.

Primarily, I would like to express my deepest appreciation and my sincere thanks to my honoured supervisor Prof. Dr. Berna YAZICI, for her support, relevant advice, sharing expertise, offering opportunities, constant encouragement and patience along the four years of my study in Turkey. I cannot express how much I am deeply indebted to her. Without her coherent and illuminating instruction, this thesis would not have reached its present form.

Also, I'm extremely grateful to Assoc. Prof. Dr. Ahmet SEZER for suggesting this research study and for supervising it at the beginning.

I would like to extend my sincere thanks to Assoc. Prof. Dr. Betül KAN KILINÇ and Prof. Dr. İlker ERCAN for attending regularly the commission related to this dissertation follow-up and for their constructive comments.

I extend my genuine thanks to the head of Department of Statistics Prof. Dr. Yeliz MERT KANTAR, to all members of the Department of Statistics, who contributed directly or indirectly to the dissertation. Particularly, Prof. Dr. Aladdin SHAMILOV for his support and kindness. Also, Dr. Lecturer Özer ÖZDEMİR who generously shared valuable suggestions related to the neural network section. I also express my gratitude to Prof. Dr. İlhan USTA who kindly helped us to overcome a specific Matlab programming issue. I am thankful to Research Assistant Mustafa ÇAVUŞ for sharing constructive information and for his help.

I sincerely thank the Türkiye Scholarship Program (YTB) and the academic authorities of Eskişehir Technical University that gave me this opportunity to meet new challenges in my career.

I wish also to express my gratitude to many Professors and Doctors that I am honoured to have contact with them during my journey in Turkey and they have faith in me. Mainly, I'm extremely grateful to Prof. Dr. Markus C. KERBER, for his unconditionally support to think out of the box and for providing valuable advice and opportunities. I would like to acknowledge the assistance of Dr. Sadam AL-WADI for providing valuable suggestions regarding the wavelet decomposition and for his encouragement. I greatly appreciate the support of Assoc. Prof. Dr. Magdalena ZIOLO and the opportunities she offered me to contribute to her publications. I would also like to thank Assoc. Prof. Halil İbrahim AYDIN, Prof. Dr. Mustafa ÖZER for their encouragement.

My deep appreciation goes also to my Tunisian Professors Assoc. Prof. Dhafer MALOUCHE and Assoc. Prof. Walid AYADI who I begun with them, 6 years ago, understanding the wavelets analysis and they were ready to support me during my research in Turkey.

I am also very grateful to the CRO dmx, where I started my professional career, and that encouraged me to follow my PhD in Turkey. Despite leaving dmx, I feel that I am indebted to it regarding what I have learnt in SAS programming and the knowledge that I acquired in pharmaceutical industry and I applied within this dissertation. Particularly, I would like to thank Dr. Dominik PFLUGER, Mr. Abdel HMISSI and Mustapha El MOURABIT and all my teammates.

Without the support of all members of my family, I would never finish this thesis and I would never find the courage to overcome all the difficulties during this journey. I am really indebted to my beloved parents Jaloul and Habiba for their love, their patience without complaining, their unconditional confidence, and with this dissertation I am seeking their forgiveness for being far away from them during the last five years. I would like to extend my sincerest gratitude to my dear brothers, the apple of my eye, Assad and Naim and my lovely sister Ameni, her husband and her children Yağmur and Ayan, you are always the source of my strength.

I sincerely express my gratitude to all my classmates and all my friends especially Nadjim MKEDDER, Housseem BELHADJ, Afifa ESSIFI, Christine ESSIFI, Mohammed AL-DABBAGH, Zaineb BENAMARA, Foued JEBALI, Ameni BEJOU, Merve YAVUZ, Sinem AKBUDAK, Aslı KAYA, Zeba KHAN, Huruy ASFHA, Beldine OMONDI. I wish to express my sincere thanks and special gratitude to the friends who left my hand during this journey or I left their hands but still not underestimating that they were important and dear to me once upon a time.

Finally, this dissertation is dedicated to the soul of my uncle Sdeek who I have never forgotten his support and his love and to the soul of our friend from the Department of Statistics Merve AKDEDE.

Maroua BEN GHOU

STATEMENT OF COMPLIANCE WITH ETHICAL PRINCIPLES AND RULES

I hereby truthfully declare that this thesis is an original work prepared by me; that I have behaved in accordance with the scientific ethical principles and rules throughout the stages of preparation, data collection, analysis and presentation of my work; that I have cited the sources of all the data and information that could be obtained within the scope of this study, and included these sources in the references section; and that this study has been scanned for plagiarism with “scientific plagiarism detection program” used by Eskişehir Technical University, and that “it does not have any plagiarism” whatsoever. I also declare that, if a case contrary to my declaration is detected in my work at any time, I hereby express my consent to all the ethical and legal consequences that are involved.

Maroua BEN GHOU
(Signature)

TABLE OF CONTENTS

	<u>Page</u>
TITLE PAGE	i
FINAL APPROVAL FOR THESIS	ii
ABSTRACT	iii
ÖZET	iv
ACKNOWLEDGMENTS	v
STATEMENT OF COMPLIANCE WITH ETHICAL PRINCIPLES AND RULES	vii
TABLE OF CONTENTS	viii
TABLE OF TABLES	xi
TABLE OF FIGURES	xii
LIST OF SYMBOLS AND ABBREVIATIONS	xiii
1. INTRODUCTION	1
2. THEORETICAL FRAMEWORK	8
2.1. Longitudinal Data	8
2.2. Wavelets	9
2.2.1. Historical context: From Fourier to wavelets	9
2.2.1.1. Fourier transformation	10
2.2.1.2. Windowed Fourier transform	11
2.2.1.3. Why wavelets?	12
2.2.2. Wavelets analysis	13
2.2.2.1. Continuous wavelet transform	14
2.2.2.2. Discretization of continuous wavelet transform	16
2.2.2.3. Discrete wavelet transform	17
2.2.2.4. Multiresolution analysis	19
2.2.3. Wavelets application	20
2.3. Missing Data	22
2.3.1. Missing data in longitudinal studies	22
2.3.2. Missing data mechanisms	23
2.3.2.1. Missing completely at random	24
2.3.2.2. Covariate dependent- MCAR	25
2.3.2.3. Missing at random	25
2.3.2.4. Missing not at random	26

	<u>Page</u>
2.3.2.5. Ignorable and nonignorable missingness	26
2.3.3. Handling missing data	27
2.3.3.1. Complete-case analysis	27
2.3.3.2. Ad-Hoc methods	27
2.3.3.3. Imputation	28
2.3.3.4. Maximum likelihood	29
2.3.3.5. Weighting methods	30
2.3.3.6. Bayesian method	30
2.4. Outliers	30
2.4.1.1. What are outliers?	30
2.4.1.2. Handling outliers	31
2.5. Modeling	33
2.5.1. Random versus fixed (variables, effects, coefficients)	33
2.5.1.1. Fixed variable and random variable	33
2.5.1.2. Fixed effect and random effect	34
2.5.1.3. Fixed coefficient and random coefficient	35
2.5.2. Parametric mixed effects models	36
2.5.2.1. Linear mixed effects models	36
2.5.2.2. Non-linear mixed effects models	36
2.5.3. Nonparametric mixed effects models	37
2.5.4. Semiparametric mixed effects models	38
2.6. Smoothing Approaches	39
2.6.1. Kernel smoothing methods	40
2.6.2. Local polynomial smoothing	41
2.6.3. Regression (polynomial) splines	42
2.6.4. Smoothing splines	43
2.6.5. Penalized splines	44
3. NOVEL APPROACHES	46
3.1. Imputation of Missing Observations	46
3.1.1. Neural network theoretical background	47
3.1.2. Multilayer perceptron for imputation of missing data	48
3.2. Handling Outliers	51

	<u>Page</u>
3.2.1. Expectation maximum for imputation of missing data	51
3.2.2. Winsorisation within subjects	51
3.2.3. Wavelets decomposition for outliers detection and handling	51
3.2.4. Outliers detection and handling across subjects	52
3.2.5. Outliers detection and handling within subjects	53
3.3. Wavelets Analysis to Smooth Data	54
4. APPLICATION AND RESULTS	56
4.1. Datasets	56
4.1.1. Generated dataset	56
4.1.2. Real dataset	58
4.1.3. Other datasets	60
4.2. Handling Missing Data	61
4.3. Handling Outliers	64
4.4. Smoothing Approaches	70
4.5. Semiparametric Model Construction	72
5. DISCUSSION, CONCLUSION AND RECOMMENDATION	78
REFERENCES	82
APPENDICES	
CURRICULUM VITAE	

TABLE OF TABLES

	<u>Page</u>
Table 2.1. Signal constructed by successive refinement	19
Table 4.1. Descriptive statistics for the generated data	56
Table 4.2. Number of patients by treatment arm	57
Table 4.3. Descriptive statistics of logrna by timepoints	59
Table 4.4. Other datasets description.....	60
Table 4.5. RMSE and RFPE of the network by activation function	61
Table 4.6. Difference of means	62
Table 4.7. Comparison between the outliers handling methods	63
Table 4.8. Comparison of RMSE between smoothing approaches	71
Table 4.9. Semiparametric modeling results of incident hypertension, LOESS as smoothing approach	72
Table 4.10. Semiparametric modeling results of incident hypertension, wavelet approximation coefficients as smoothed data	73
Table 4.11. Semiparametric modeling results of ACTG, LOESS as smoothing approach	74
Table 4.12. Semiparametric modeling results of ACTG, wavelet approximation coefficients as smoothed data	75

TABLE OF FIGURES

	<u>Page</u>
Figure 1.1. Dissertation diagram	7
Figure 2.1. Handling missing data Taxonomy	27
Figure 2.2. Distribution and Kernel density for Systolic Blood Pressure	40
Figure 2.3. Local linear fitting for hypertension	41
Figure 2.4. Local quadratic fitting for hypertension	41
Figure 2.5. Linear regression fitting for hypertension	42
Figure 2.6. Spline regression fitting for hypertension	42
Figure 2.7. Smoothing splines for hypertension	43
Figure 2.8. Penalized splines for hypertension	44
Figure 3.1. Flowchart of MLPMID algorithm	48
Figure 3.2. Flowchart of suggested methods of handling outliers	50
Figure 3.3. Flowchart for wavelet Decomposition for outliers detection and handling ...	51
Figure 3.4. Smoothing splines for hypertension	54
Figure 4.1. Boxplots of logrna by timepoints and by treatment arm	58
Figure 4.2. Wavelets decomposition of logrna1 across subjects, applying algorithm WDODHAS	64
Figure 4.3. Wavelets decomposition of logrna1 by subjects, applying algorithm WDODHWIS	65
Figure 4.4. Wavelets decomposition of logrna1, applying WDODHWIS	65
Figure 4.5. Wavelets decomposition of logrna3 across subjects, applying WDODHAS	66
Figure 4.6. Wavelets decomposition of logrna3 by subjects, applying WDODHWIS	67
Figure 4.7. Wavelets decomposition of logrna3, applying WDODHWIS	67
Figure 4.8. Wavelets decomposition of logrna5 across subjects, applying WDODHAS	68
Figure 4.9. Wavelets decomposition of logrna5 by subjects, applying WDODHWIS ...	68
Figure 4.10. Wavelets decomposition of logrna5, applying WDODHWIS	69
Figure 4.11. Wavelets decomposition of Diastolic Blood Pressure	70
Figure 4.12. Wavelets decomposition of Systolic Blood Pressure	70

LIST OF SYMBOLS AND ABBREVIATIONS

α	: alpha
β	: Beta
μ	: Mu
λ	: Lambda
ψ	: Psi
Δ	: Delta
ε	: Epsilon
η	: Eta
BMI	: Body Mass Index
CD-MCAR	: Covariate Dependent- MCAR
CWT	: Continuous Wavelet Transform
DB	: Daubechies
DYABP	: Diastolic Blood Pressure
DWT	: Discrete Wavelet Transform
EM	: Expectation Maximum
EMIMID	: Expectation Maximum for Imputation of Missing Data
FT	: Fourier Transform
GLM	: Generalized Linear Model
iid	: Independent and identically distributed
IQR	: Interquartile Range
LME	: Linear Mixed Effects Models
LOESS	: Locally Estimated Scatterplot Smoothing
LOWESS	: Locally Weighted Scatterplot Smoothing
MAR	: Missing at Random
Max	: Maximum

MCAR	: Missing Completely at Random
MCMC	: Markov Chain Monte Carlo
ME	: Mixed Effects
Min	: Minimum
MLP	: Multilayer Perceptron
MLPIMID	: Multilayer Perceptron for Imputation of Missing Data
MNAR	: Missing Not at Random
MSE	: Mean Square Error
NIH	: National Institutes of Health
NLME	: Nonlinear Mixed Effects Models
NN	: Neural Network
NPMEM	: Nonparametric Mixed Effects Models
PI	: Protease Inhibitor
PMEM	: Parametric Mixed Effects Models
P-Splines	: Penalized splines
RFPE	: Root Final Prediction Error
RLOESS	: Robust Locally Estimated Scatterplot Smoothing
RLOWESS	: Robust Locally Weighted Scatterplot Smoothing
RMSE	: Root Mean Square Error
SAS	: Statistical Analysis Software
SYSBP	: Systolic Blood Pressure
SD	: Standard Deviation
SE	: Standard Error
sgola	: Savitzky-Golay
SPMEM	: Semiparametric Mixed Effects Models
WASDA	: Wavelets Analysis to Smooth Data
WINWIS	: Winsorisation Within Subjects

WDODH	: Wavelets Decomposition for Outliers Detection and Handling
WDODHAS	: Wavelets Decomposition for Outliers Detection and Handling Across Subjects
WDODHWIS	: Wavelets Decomposition for Outliers Detection and Handling Within Subjects



1. INTRODUCTION

Longitudinal data consists in gathering several observations of the same subjects intermittently over time (Müller, 1988; Diggle et al., 2002; Fitzmaurice et al., 2004; Hedeker and Gibbons, 2006). This type of data examines the dynamic measure changes over different timepoints, allows the measurement of the duration of events, studies the relationships and models the differences and the heterogeneity among subjects. Attrition or missing observations (due to loss of follow up, death, discontinuation etc.), outliers and complexity of modeling are common issues in longitudinal data. Next sections cover those issues one by one.

Missing data can be considered as a common problem in data analysis that has been largely discussed. Day (1999) defined, in the Dictionary for Clinical Trials, missing data as data value that should have been recorded but, for some reason, was not. It is referred to as non-response or unobserved data and occur in most types of studies. Rubin (1976) introduced the expression of missingness mechanism or the missing data mechanism and set up the common notation that still under use, to explain why data is missing (Hedeker and Gibbons, 2006). Mainly in data analysis, three missingness mechanisms are defined by Rubin (1976): Missing Completely at Random (MCAR), Missing at Random (MAR) and Missing Not at Random (MNAR). Later, despite emphasizing the cruciality of those mechanisms that are required to ensure the efficiency of the imputation methods, many researches still ignoring them (Wood et al., 2004).

After ascertaining the missingness mechanism, handling missing data is required. Approaches to deal with missing observations can be classified into four main categories: complete cases (considering only the observed data), interpolation, imputation and modeling. Handling with complete case can be applied for MCAR data by listwise deletion method. Interpolation is more appropriate for MAR data and when the missingness patterns are monotone (Kreindler and Lumsden, 2012). Imputation method is pertinent with MAR data, it covers weighting methods, single imputation, multiple imputation (Rubin, 1987; White et al., 2011) and Expectation Maximization (Dempster et al., 1977; Howell, 2008; Schafer and Graham, 2002).

Interestingly, plenty research papers have applied Neural Network (NN) as an imputation method for missing observations (Norway, 1997; Nordbotten, 1998; Yoon and Lee, 1999; Sonnberger and Maine, 2000). For instance, Silva-Ramírez et al. (2011) used Multilayer Perceptron (MLP) to impute data on MCAR, they proceed the training in a way that it includes observed and missing data. Parveen and Green (2001), Lipton et al. (2016) and Choi et al.

(2015) utilized the Recurrent Neural Networks (RNNs) by unifying missing entries with the input or by processing with simple imputations. In the same context, Che et al. (2018) developed a novel model called Gated Recurrent Unit (GRU) to better predict the missing patterns via RNN. Smieja et al. (2018) proposed an approach that consists in replacing typical neuron's response in the first hidden layer by the attributed expected value. Zhang and Benveniste (1992) introduced the Wavelet Neural Network (WNN), which is based on the wavelet transform and it is considered as an alternative to feedforward NN for approximating random nonlinear functions. Various studies utilized WNN for different applications (Yu and Chen, 2007; Lung, 2007; Xie and Zhang, 2007; Wang et al., 2013) and particularly Panigrahi et al. (2012) investigated how to remove and interpolate the missing data by WNN.

Thereby, this research highlights the prominence of missingness mechanisms following the work of Wood et al. (2004). Then, based on the researches of (Euredit, 2005; Silva-Ramírez et al., 2011; Panigrahi et al., 2012; Wang et al., 2013), it suggests an algorithm that trains complete data, then fills the data gaps by the obtained predictions of Multilayer Perceptron (MLP). The suggested novel algorithm Multilayer Perceptron for Imputation of Missing Data (MLPIMID) allows the variation of the activation functions (wavelet mother functions such as Gaussian, Morlet and Mexican Hat, and sigmoid). SAS macros, published by Sarle (1994), are updated and utilized to vary the activation functions and to allow the imputation of missing observations for longitudinal and non-longitudinal datasets.

As one keystone of this dissertation pertains the wavelet decomposition to handle outliers and to smooth data, it is indispensable to examine its background. Historically, since the 18th century, Fourier transform has been known as an important mathematical tool, used in the development of telephony, computer science and the audio-visual field. Nevertheless, a major flaw was identified in this tool which is the absence of time information. In fact, Fourier transform gives the information about the number of frequencies contained in the signal but hides the times of the diffusion of these frequencies as if the signal moments are equivalent (Meyer, 1990; Daubechies, 1992; Kaiser, 1994; Burke Hubbard, 1995; Graps, 1995; Quarta, 2001). The wavelet was introduced as the alternative approach that breaks down a signal both in time and in frequency. For instance, in the case of a musical note, the wavelet indicates both the frequencies (notes) and time. Hence, wavelets have become increasingly a popular tool in various fields such as image processing (digital borrowing, medical X-rays, seismic waves etc.), audio processing (voices, musical notes, etc.) and recently the economic and financial areas (Gençay et al., 2002). Arneodo et al. (1995) referred wavelet as a mathematical microscope due to its ability in showing the weak transients and peculiarities in the time series, it utilizes the

optics of the microscope, its amplification varies with the scale factor (Struzik, 2001). Despite the wavelets analysis has been considered as a very powerful mathematical tool, it isn't commonly applied in the longitudinal data. Consequently, this dissertation suggests the wavelet decomposition for two circumstances: outliers' treatment and smoothing approach for nonparametric models.

Several definitions have been attributed to the outliers, demonstrating that, those observations were one of the earliest statistical concern (Che et al., 2018). In 1852, Pierce defined them as the observations which differ from the others and they can mislead and perplex the user. Currently, outliers' definition is remained almost the same: an observation that deviates so much from other observations, as to arouse suspicion it was generated by a different mechanism (Mondal and Percival, 2010). For some researchers, the rule of thumb is that an outlier presents any observation whose removal from the data changes the estimation of the variable of interest at least by 10% (Lung, 2007).

The most common causes of outliers are data entry errors due to human errors, measurement errors, experimental errors during the experiment planning or data extraction, data processing errors and sampling errors or natural showing the originalities in the data (Heaton and Silverman, 2008). Coming back to the earliest definition, outliers created and still creating a confusion and mislead; keeping them or deleting them are two sides of the same coin. The question of what extent removing outliers in hormonal data leads to divergent statistical conclusions was deliberated (Rubin, 1976). Other studies focused on the statistical data preparation and proposed the possible approaches to manage missing values and outliers without deleting them (Nordbotten, 1998).

Some of the popular methods for outlier detection that have been largely examined in the literature are: the parametric approaches (such as Z-Score, probabilistic and statistical modeling), the linear regression models (like PCA, LMS) and the nonparametric approaches (such as proximity-based models). Also, Lipton et al. (2016) developed an approach via overdispersion to detect outliers in the longitudinal data.

Different tactics can be followed to treat outliers. *Trimming* approach consists on excluding the outliers and then examining the data. However, the trimmed estimators such as mean decrease the variance in the data and cause a bias based on under or overestimation. Given that the outliers are also observed values, excluding them from the analysis makes this approach inadequate. To avoid the risk of losing information, once the outliers are identified, different approaches were proposed to treat them notably imputation with mean, median or mode and capping. For the values which lie outside the $1.5 * IQR$ (Interquartile Range) limits, they could

be capped by replacing those observations outside the lower limit with the 5th percentile and the ones greater than the upper limit with the 95th percentile. Another method consists on considering the outliers as Not Available (NA) observations and applying a handling missing approach to treat them.

The weight modification method allows weight modification without discarding or replacing the values of outliers by limiting the influence of the outliers. The value modification method allows the replacement of the values of outliers with the largest or second smallest value in observations excluding outliers. If the nature of the population distributions is known, *robust estimation* method is considered appropriate as it produces estimators robust to outliers, and estimators are consistent. Many studies have proposed a variety of statistical models for robust estimation; however, still not applicable due to the complicated methodological aspects (Nordbotten, 1998).

Some recent studies showed that wavelets analysis allow the data decomposition, treating the outliers and then reconstructing the data without losing much information i.e. reconstructed signal after removing the outliers is very similar to the original signal including outliers (Al-Khazaleh et al., 2015; Al-Wadi, 2015; Sagar et al., 2016).

Accordingly, three approaches will be introduced to treat outliers for longitudinal data: Firstly, Expectation Maximum algorithm will examine the outlier as a missing observation and this modified approach will be addressed as Expectation Maximum for Imputation of Missing Data (EMIMID). Then, winsorisation approach will impute the outlier by modifying the weights of outliers or replacing the values being tested for outliers with the 10th percentile or the 90th percentile by subject and this approach will be named Winsorisation Within Subjects (WINWIS). Finally, the wavelet decomposition will be provided in two ways due to the particularity of the longitudinal data: Primarily, the wavelets decomposition will be applied across the subjects i.e. the whole data will be decomposed via the algorithm Wavelets Decomposition for Outliers Detection and Handling Across Subjects (WDODHAS). Secondly, the wavelets decomposition will be applied within subjects i.e. for each subject a decomposition will take place by utilizing the novel algorithm Wavelets Decomposition for Outliers Detection and Handling Within Subjects (WDODHWIS). WDODHAS algorithm can be applied on any dataset. WDODHWIS algorithm is only applicable on longitudinal data, and it will be highly required that each subject will have at least two repeated measures.

The third keystone of this research is figuring out an appropriate model. Admittedly, over the last decades, mixed models have gained a lot of attention in statistical research studies over the traditional analyses due to their flexibility to handle multilevel or clustered data without

constraints as equal number of observations or non-missing observations. Also, mixed models allow a variety of choices for modeling correlations in the data (Howell, 2008; Maxwell et al., 2003). Mixed models are current in the design that combines random and fixed effects and widely used in several fields such as pharmaceutical industry and economics.

Kreft and De Leeuw (1998), Torres-Reyna (2007) and Newsom (2017) provided clear descriptions to distinguish between random variable and fixed variable, random effect and fixed effect and finally random coefficient and fixed coefficient to avoid confusions.

Mixed effects models can be characterised as parametric and nonparametric; parametric mixed models may take different covariates into account, but they require parametric assumptions that may not be satisfied by non-linear models. Nonparametric mixed models are more flexible to fit longitudinal data and robust against model misspecification, but they involve a few covariates and may be computationally intensive (Maxwell et al., 2003).

For nonparametric features, smoothing approaches are required. Ruppert et al. (2003) categorized four principal smoothing approaches to nonparametric modeling: smoothing splines (Wahba, 1990; Green and Silverman, 1994; Eubank, 1999); series-based smoothers, including wavelets (Tarter and Lock, 1993; Ogden, 1996); kernel methods (Wand and Jones, 1995; Fan et al., 1996); and regression splines (Friedman, 1991; Stone et al., 1997; Hansen and Kooperberg, 2002). Ruppert et al. (2003) focused on the penalized splines, labelled also as p-splines, pseudo splines, and low-rank spline smoothers in the literature.

It is crucial to highlight that the type of dataset is the reference to choose among the smoothing approaches (Ruppert et al., 2003). For instance, moving averages is appropriately used to smooth away seasonality in quarterly data or monthly data, in order to identify trend, LOWESS is commonly applied for non-seasonal series and RLOWESS is utilized if the data contain outliers. Here, wavelets decomposition will be used to smooth longitudinal data via the approach Wavelets Analysis to Smooth Data (WASDA).

Lately, semiparametric mixed models that combine parametric and nonparametric features, have been more applied on longitudinal data (Zhang and Lin, 2003; Ruppert et al., 2003; Ngo and Wand, 2004; Hulin and Jin-Ting, 2007; Szczesniak et al., 2013; Szczesniak et al., 2015) but still under development. In a semiparametric mixed effects model, part of the variations in the response variable can be elucidated by certain parametric models of some covariates in the fixed effect component and/or the random effect component, while the remaining is explained by a nonparametric function of time. In a time, varying coefficient mixed effect model, the coefficients of the fixed effects and random effects covariates are smooth functions of time (Jin-ting, 2007). Ruppert et al. (2003) showed that semiparametric regression

integrates nonlinear functional relationships in regression analyses and semiparametric regression can be applied on several fields that apply regression analysis. Hence, a semiparametric model for longitudinal data will be constructed to test its efficiency and wavelets decomposition will be tested to smooth the nonparametric data.

To sum up, the objective of this dissertation is to examine the common issues in longitudinal data, notably missing data and outliers then construct a semiparametric model, by using some persuasive tools that still not highly accessed.

The contribution of this study is the implementation of a novel algorithm MLPIMID to impute missing data via MLP and two algorithms WDOHDS and WDOHWS to handle outliers without deleting them via wavelet decomposition. The application of the novel algorithms is not limited to pharmaceutical industry. In addition, wavelet tool is largely discussed as a powerful decomposition instrument and as a smoothing approach and this may encourage future researches.

SAS and Matlab have been used as main software in this research. SAS macros and Matlab algorithms that have been used in this dissertation are provided in detail in the appendices.

The remainder of this dissertation is structured as following:

Second chapter titled Theoretical Framework: will be devoted to the theoretical framework. First section will outline the longitudinal data. Second section will summarize the wavelets analysis and their applications in different fields. Third section will examine the missing data as mechanisms and imputation methods. Fourth section will highlight the outliers and the classic methods used to handle them. Fifth section will cover theoretically the modeling part: what are mixed models, the parametric models, the nonparametric models and the semiparametric models. Finally, the smoothing approaches will be scrutinized.

Third chapter titled Novel Approaches: concerns the proposed methods and novel algorithms of this dissertation. Firstly, imputation of missing observations will be concerned and accordingly the winsorisation within subjects WINWS, the EM algorithm EMIMID, Wavelets Decomposition for Outliers Detection and Handling Within Subjects (WDOHWS) and Wavelets Decomposition for Outliers Detection and Handling Across (WDOHDS) will be presented. Secondly, Multilayer Perceptron for Imputation of Missing Data (MLPIMID) algorithm will be introduced. Finally, Wavelets Analysis to Smooth Data (WASDA) will be highlighted.

Fourth chapter titled Application and Results: will present the datasets that will be used to test the proposed approaches. In fact, two principal datasets will be employed: a generated

one based on a previous historical clinical data from National Institutes of Health (NIH), the second one is a real data retrieved from the school of Public Health UTHealth. Then, the empirical results and findings will be summarized.

Fifth chapter titled Discussion, Conclusion and Recommendation: will sum up the findings and the recommendations of this dissertation.

To further empathize the objective and the outlines of this dissertation, see Figure 1.1.

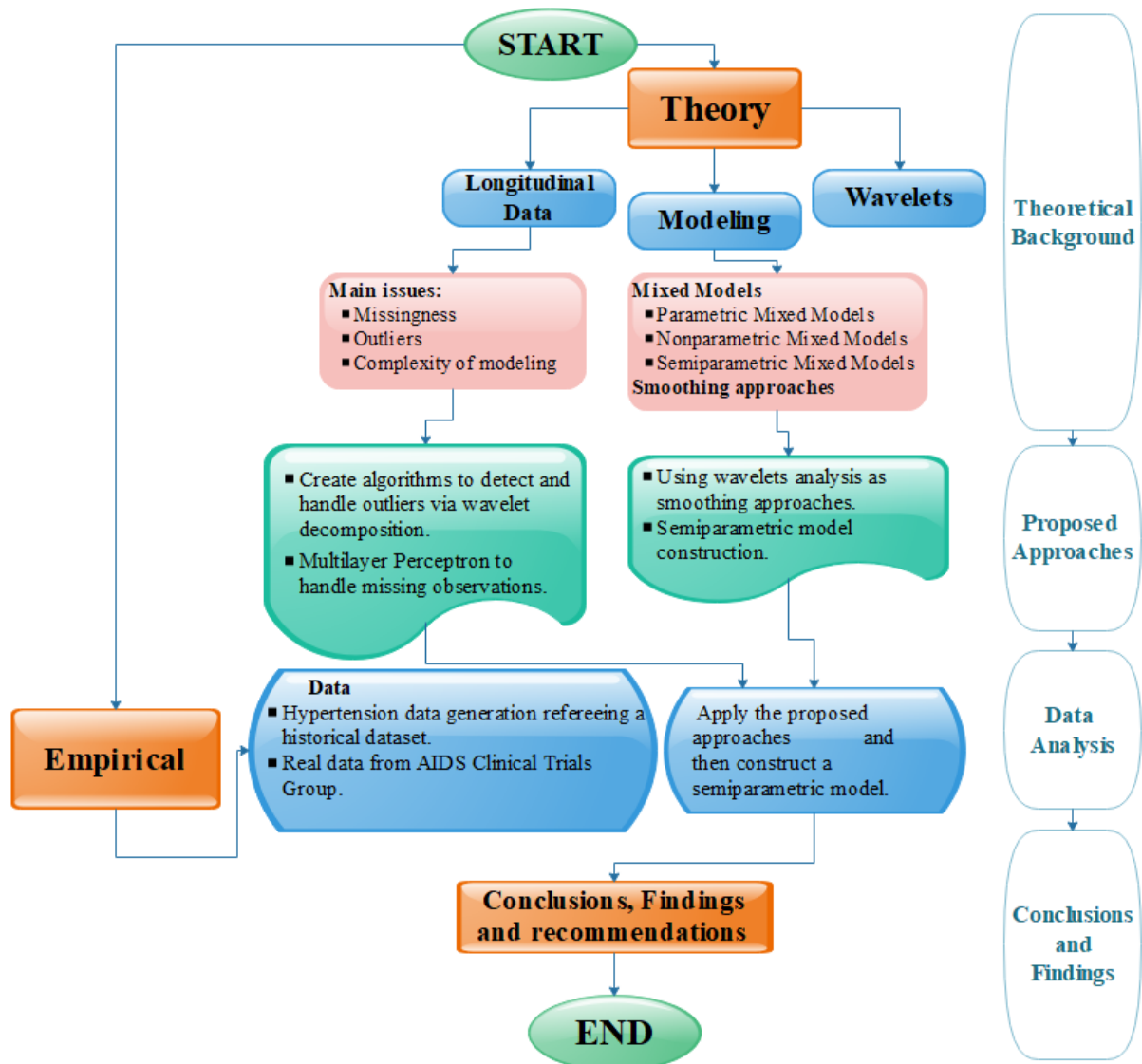


Figure 1.1. Dissertation diagram

2. THEORETICAL FRAMEWORK

This chapter outlines the literature review and the mathematical background of this dissertation. First section of the chapter defines longitudinal data and differentiate the cross-sectional study design from longitudinal study design. Second section depicts the historical context of wavelets and their mathematical background with screening some common applications of wavelets analysis such as denoising the signal, compression and splitting signal components. Third section discusses the missing data in longitudinal data with emphasizing the missingness mechanisms and stating the most known imputation methods of missing observations. Fourth section highlights the dilemma of outliers via stating the common definitions and the common approaches that address those abnormal observations. Fifth section defines the Mixed Effects Models and presents the theoretical framework of parametric, nonparametric and semiparametric mixed effects models. Last section of the theoretical chapter provides some commonly used smoothing approaches that are required in the nonparametric models.

2.1. Longitudinal Data

Cross-sectional dataset, known also as study population's cross section dataset, consists in observing many subjects such as individuals, firms, countries, or regions etc. at the same timepoint or at different timepoints. The analysis of cross-sectional data consists usually in comparing the differences among the defined subjects.

Calder (2012) and Müller (1988) defined longitudinal data as the data that tracks the same type of information on the same subjects at multiple points in time. Frees (2004) described longitudinal data analysis as a “*marriage*” of regression and time series analysis: longitudinal data assemble cross section of subjects (like several regression datasets) and observe subjects over time (unlike regression datasets). Unlike time series data, with longitudinal data many subjects will be tracked. Observing a broad cross section of subjects over time allows studying the data dynamically, as well as cross-sectional. To define a data as “*longitudinal*”, at least two repeated measures should be given and two observations do not construct a longitudinal data (Rogosa, 1991; Bijleveld et al., 1998). Magnusson and Bergman (1990) pointed out that the longitudinal data can be considered as the lifeblood of the study of individual development. In fact, this type of data scrutinizes the dynamic relationships and model the differences and the heterogeneity among subjects (where the measures change). Longitudinal data are largely applicable in the pharmaceutical industry: event data like the timing of deaths, allograft rejection, heart attacks, EEG (electroencephalogram) and ECG (electrocardiogram).

Fitzmaurice et al. (2004) provided very useful examples and case studies of longitudinal data applied in clinical trials and biomedical field.

To clarify the difference between the cross-sectional and longitudinal study design, Fitzmaurice et al. (2004) provided a very understandable example related to Body Mass Index (BMI) measurement before and after a diet to distinguish between the cross-sectional and longitudinal study designs showing that the cross-sectional does not show the effect of the diet as it is based on two distinctive samples, but longitudinal study design allows estimating the change of BMI index by defining the change of BMI index as outcome variable.

Likewise, longitudinal data has drawbacks, the most common ones are difficulty in designing the sampling scheme due to the problem of subjects leaving the study prior to its completion (known as attrition) and the complexity of modeling.

As wavelets will be a lynchpin for the outliers handling and as a smoothing technic, the next section discusses the wavelets analysis theoretical background.

2.2. Wavelets

This section provides the historical context of wavelets and its mathematical foreground with stating some fields' application of wavelet analysis, referring Meyer (1990), Daubechies (1992), Kaiser (1994), Burke Hubbard (1995), Graps (1995) and Polikar (2006).

2.2.1. Historical context: From Fourier to wavelets

In 1807, Joseph Fourier came out with the idea that all periodic functions can be the sum of basic trigonometric functions. In other terms, "*... any continuous signal could be decomposed into the sum of sinusoids. This decomposition is reversible and has other properties knowing the conservation of energy.*"

This idea had several oppositions from mathematicians of that era such as Lagrange, Legendre and Laplace. After fifteen years, Fourier was able to publish the results of his theory and after one hundred and fifty years, Fourier's ideas were generalized and applied to non-periodic functions, and later on the theory was developed by Cooley and Tukey (1965) with the proposal of a fast algorithm to calculate a discrete version.

Fourier transform was widely used in various fields such as medicine with electrocardiography (ECG) measurement (ECG is the graphical representation of the electrical potential that controls the activity of the heart).

2.2.1.1. Fourier transformation

A frequency content of a Mozart sonata piece can be modelled by a function f such that $f(t)$ represents the pressure of the air on the piano at time t .

If the melody is composed by only one note (The associated note of a frequency f is the group of frequencies with the form $N = \{2^k f, k \in \mathbb{Z}\}$), the function f will be periodic, and Fourier series will be the best tool to analyse it. Let f defines a periodic frequency of period T , square integrable over its period. Fourier' coefficients are calculated in t as follows:

$$\forall n \in \mathbb{Z}, \quad c_n = \int_0^T f(t) \exp\left(-\frac{2i\pi}{T}nt\right) dt \quad (2.2.1)$$

Equation 2.2.1 presents the analysis of signal f (as the coefficients allow the continuous analysis of the harmonic content of f). The harmonic of a note is the group of frequencies with the form $N = \{kf, k \in \mathbb{N}\}$. To be noted that: $\exp(iz) = \cos(z) + i \sin(kz)$ and $\exp(niz) = \cos(nz) + i \sin(nz)$.

$$f = \frac{1}{T} \sum_{n \in \mathbb{Z}} c_n \exp\left(-\frac{2i\pi}{T}n\right) \quad (2.2.2)$$

Equation 2.2.2 gives a decomposition of the signal f as series of harmonic modes $e^{2\pi i n t / T}$ with frequency $\omega_n = n/T$. It is called the signal synthesis as it allows the reconstitution of the signal f through its harmonic modes. If the melody is not periodic, another approach of calculation could be used, which is the continuous Fourier transform $\hat{f}$.

$$\forall \omega \in \mathbb{R}, \quad \hat{f}(\omega) = \int_{-\infty}^{+\infty} f(t) \exp(-2i\pi \omega t) dt \quad (2.2.3)$$

It should be emphasised that Equation 2.2.3. is a priori correct only if the functions are integrable $f \in L^1(\mathbb{R})$. A basic characteristic of $\hat{f}$ that it tends to 0 in the infinity and continuous on $\mathbb{R}$.

The signal could be reconstructed from its Fourier transform as demonstrated by the below equation 2.2.4:

$$\forall t \in \mathbb{R}, f(t) = \int_{-\infty}^{+\infty} \hat{f}(\omega) \exp(2i\pi \omega t) d\omega \quad (2.2.4)$$

To determine $\hat{f}$, the function f must be integrated over the entire time interval; therefore, f contains the total amplitude of the frequency in the complete signal (song) rather than the distribution of harmonics in each of the individual notes. So, the signal (song) should be ended to calculate f and the result will be uninteresting.

In summary, in the original signal, the Fourier transform considers the presence or absence of frequencies without considering the moment of this presence: at which moment this frequency appears and then at which moment it disappears of the signal. So, Fourier transform is effective in determining the composite frequencies of a signal, but it is not practical in reconstructing the original signal from these obtained frequencies, due to the inability to define the temporal arrangement of each frequency after the transformation. Thus, there is a loss of this temporal information in the process.

Another approach would be discretising the signal f into small pieces of individual notes and to analyse them separately by the Fourier series. But it is impossible with this approach to know exactly when a given note ends and when the next begins.

Next section aims to present how to overcome the disadvantage that Fourier transform is a global transformation.

2.2.1.2. *Windowed Fourier transform*

Windowed Fourier Transform was introduced to avoid the loss of temporal information of the frequencies due to the global character of Fourier Transform. The idea was inspired by the auditory experience. Indeed, the human ear distinguishes between the changes of the notes of the sound without any partition of the signal. Since the ear analyses the frequency of a signal directly, the new model must give information on the function f simultaneously in time and frequency. Mathematically, the function $\tilde{f}(\omega, t_0)$ represents the distribution of the frequency ω at time $t = t_0$ with t_0 a fixed value, t should satisfy the condition $t \leq t_0$ to determine the corresponding frequency $f(t)$ (based on the idea that the ear can only analyse what it hears). It is also assumed that the ear has a limited memory in time. In other terms, there is a time interval T where only the values $f(t)$ with $t \geq t_0 - T$ has an impact on the calculation of $\tilde{f}(\omega, t_0)$.

Also, it can be expected that the values of $f(t)$ (for t close to the limits of the interval $[t_0 - T, t_0]$) have a less impact on $\tilde{f}(\omega, t_0)$ calculation than the values of t in the centre of this interval.

$g : \mathbb{R} \rightarrow \mathbb{C}$ is a time function with $g(t) = 0$ if $t \notin [-T, 0]$, so $g \subset [-T, 0]$. g is called the analysis window. It is defined that:

$$\forall t \in \mathbb{R}, \quad f_t(u) = \bar{g}(u - t)f(u) \quad (2.2.5)$$

With $\bar{g}(u - t)$ is the conjugate of the complex number $g(u - t)$.

$f_{t_0}(u)$ is a local version of f , which depends only on the $f(t)$ values for $t \in [t_0 - T, t_0]$.

The windowed Fourier transform of f or Gabor transform is defined as the Fourier transform f_t .

$$\tilde{f}(\omega, t) = \hat{f}_t(\omega) = \int_{-\infty}^{+\infty} f(u) \bar{g}(u - t) \exp(-2i\pi \omega u) du \quad (2.2.6)$$

With the function $g_{\omega t}(u) = g(u - t)e^{2\pi i \omega u}$, the above equation 2.2.6. will be in the form:

$$\tilde{f}(\omega, t) = \langle g_{\omega t} | f \rangle \quad (2.2.7)$$

$\langle \cdot | \cdot \rangle$ is the dot product in Hilbert space $L^2(\mathbb{R})$.

The functions $g_{\omega t}$ could be considered as the notes which oscillate to the frequency ω in the defined environment $|g(\cdot - t)|$.

The next step consists in defining the inverse of Fourier transform to reconstruct the original signal f from its windowed Fourier transform $\tilde{f}$.

By applying the inverse Fourier transform, on the above equation 2.2.5., through the variable ω .

$$f_t(u) = \bar{g}(u - t)f(u) = \int_{-\infty}^{+\infty} \tilde{f}(\omega, t) \exp(2i\pi \omega u) d\omega \quad (2.2.8)$$

As the g could be null, the subdivision by $\bar{g}(u - t)$ is not possible to get $f(u)$. The idea then consists in multiplying the both sides of the above equation by $g(u - t)$ then integrate via the variable t .

$$f(u) \int_{-\infty}^{+\infty} |g(u - t)|^2 dt = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \tilde{f}(\omega, t) g_{\omega t}(u) d\omega dt \quad (2.2.9)$$

Or, the value $\int_{-\infty}^{+\infty} |g(u - t)|^2 dt = \|g\|_{L^2}^2$ is null, the construction formula is defined via the equation 2.2.10:

$$f(u) = C \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} g_{\omega t}(u) \tilde{f}(\omega, t) d\omega dt \quad (2.2.10)$$

Equation 2.2.10 presents Gabor Transform that allows the signal reconstruction with C is a constant to be calculated.

2.2.1.3. Why wavelets?

To overcome the limits of Fourier transform and Gabor transform, wavelets tool was developed. Equations 2.4.11. and 2.4.12. present respectively Fourier Transform and Gabor transform.

$$\forall t \in \mathbb{R}, f(t) = \int_{-\infty}^{+\infty} \hat{f}(\omega) \exp(2\pi i \omega t) d\omega \quad (2.2.11)$$

$$f(u) = c \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} g_{\omega t}(u) \tilde{f}(\omega, t) d\omega dt \quad (2.2.12)$$

They can be interpreted as the signal f decomposition on a family of functions presented as a base and the integrals replace the sums.

In Fourier case, the family is known as $e^{2\pi i \omega t}$ and for the case of Gabor it is $g_{\omega t}$. The elements of this family are called the transform atoms or wavelets of Gabor.

Briefly, it can be summarized that the basic functions in the Fourier method are totally concentrated in frequency (in the form of a Dirac pulse) and totally distributed over the axis time because the sinusoid does not reduce and extends from $-\infty$ to $+\infty$, demonstrating that Fourier transform gives the maximum of information related to the distribution of frequencies but completely ignore the details related to time.

Gabor Transform puts a new object called time-frequency plan; by the location of the signal, simultaneously, in time and in frequency by observing it through a window translated in time and consequently in frequency, called also modulated window in time.

These two operations give rise to the notes $g_{\omega t}$. The signal is subsequently reconstructed by the superposition of these notes with the transform $\tilde{f}(\omega, t)$.

Therefore, any signal performance in a specific time interval much be shorter than the length of the window, is sub-localized in time and must be obtained as the result of constructive and destructive interferences between the notes (as Fourier series), which means that many notes will be used and therefore $\tilde{f}(\omega, t)$ must be extended in frequency. This makes the Window Fourier Transform ineffective for the analysis of regular signals in time but very fast or very slow compared to the size of the window.

Since obtaining a time-based response involves introducing a scale of magnitude into the analysis (the size of the window), a reconstruction method (independent of this scale) should be tested.

2.2.2. Wavelets analysis

Previous section demonstrated that Fourier transform is a global transformation and Gabor transform is local, but both are of fixed temporal resolution.

To overcome this issue, a new concept was developed based on Gabor's research, called the wavelet transformation. These methods were developed and extended by authors mainly Grossman and Morlet in 1984, Daubechies in 1990 and 1992 and Meyer in 1992. The success of the wavelets was due to a discretization method introduced in 1989 by Mallat and then developed by Meyer (1990) and Daubechies (1992).

As the name indicates, a wavelet is a small wave. The term "*small*" mainly explains in this context that the wave increases and decreases in a limited time support (Masset, 2008). Thus, the main feature of a wavelet is its compact support, i.e. the wavelet function is limited in both time and frequency domains.

Meyer (1990) defined the wavelets as "... *unlike the Fourier series, the coefficients of a series of wavelets translate the properties of the functions in a simple, precise and reliable way, at least the properties that correspond to a discontinuity, an unplanned event.*"

Theoretically, a wavelet is simply a time function that follows a basic rule, known as the wavelet eligibility condition:

$$C_{\Psi} = \int_0^{\infty} \frac{|\Psi(f)|}{f} df < \infty \quad (2.2.13)$$

$\Psi(f) = \int_{-\infty}^{+\infty} \psi(t) \exp(-2i\pi f t) dt$ defines the Fourier transform.

f is a frequency function for $\Psi(f)$, ψ is called mother wavelet or analysing wavelet.

This condition ensures that $\Psi(f)$ quickly tends to 0 if $f \rightarrow \infty$. (Grossman and Morlet, 1984)

To ensure that $C_{\Psi} < \infty$, the following conditions, related to the mother wavelet, are imposed:

- a. $\psi(0) = 0$, or $\int_{-\infty}^{\infty} \psi(t)dt = 0$
- b. $\int_{-\infty}^{\infty} |\psi(t)|^2 dt = 1$, representing the energy unit.

Some wavelet functions are presented below (will be used in section 3.1.2 related to MLPIMID algorithm):

Gaussian: $\Psi(t) = \frac{t}{\sqrt{2\pi}} \exp(-\frac{t^2}{2})$; Morlet: $\Psi(t) = \cos(1.75t)(-\frac{t^2}{2})$; Meyer (approximate formula) as $\Psi(t) = 35t^4 - 84t^5 + 70t^6 - 20t^7$; Mexican Hat as $\Psi(t) = c(1 - t^2)\exp(-\frac{t^2}{2})$ with $c = \frac{2}{\sqrt{3}}\pi^{-1/4}$.

Generally, wavelet transforms are classified into three categories; Continuous Wavelet Transform (CWT), Discrete Wavelet Transform (DWT) and transformed into wavelets based on multiresolution analysis. These categories will be discussed in the following sections.

2.2.2.1. Continuous wavelet transform

Continuous Wavelet Transform (CWT) was developed as an alternative approach to the Gabor transform to overcome the resolution problem. Indeed, the wavelet analysis is like Gabor transform at the level of the multiplication function (mother wavelet for CWT and window function for windowed Fourier transform). In addition, the transformation is calculated separately for different segments of the signal in the time domain.

Thus, the two concepts allow providing an analysis of the frequency content of a signal in a quasi-instantaneous manner, i.e. break down the signal in a multitude of atoms or coefficients in the time-frequency domain. However, there are some major differences between these two transformations:

- For CWT, the negative frequencies of the FT of windowed signals are not considered, and so a single vertex (summit) will be observed as a spectrum corresponding to a sinusoid.
- For CWT, the width of the window, which is constant for FT regardless of the frequency, will be changed during the transformation and will be calculated for each single spectral component. This is probably the most significant feature of the wavelet transform.
- CWT splitting produces larger atoms in time when the frequency is low and vice versa. Thus, the fast transients of a signal are considered by high frequency atoms that have a very fine temporal resolution, while the slower movements, of low frequency, can benefit from a more precise frequency representation.

CWT consists of creating, from the mother wavelet ψ , a family of wavelets.

$$W(u, s) = \int_{-\infty}^{\infty} f(t) \bar{\psi}_{u,s}(t) dt \quad (2.2.14)$$

With $\psi_{u,s}(t) = \frac{1}{\sqrt{s}} \psi\left(\frac{t-u}{s}\right)$ where u determines the position (the translation by u) and s presents the scale (the dilatation by s). For a signal case, s is the frequency and u presents time. The coefficient $\frac{1}{\sqrt{s}}$ ensures that $\|\psi_{u,s}(t)\| = 1$.

CWT produces a *redundant* analysis because atoms overlap due to the nature of analogue filters that cannot perfectly cut frequencies outside their bandwidth (Edwards, 1999). As a result, it is less appropriate for applications involving data reduction, such as compression and noise suppression.

CWT has been used for the detection of certain EEG (electroencephalogram) signals of epileptic seizures and QRS complex in ECG (electrocardiogram) signal. The inverse wavelet transforms, or the reconstruction of the original signal, is given by the equation 2.2.15.:

$$f(t) = \frac{1}{C_{\psi}} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} W(u, s) \psi\left(\frac{t-u}{a}\right) \frac{1}{a^2} du da \quad (2.2.15)$$

Determining the wavelet transforms for CWT can be challenging sometimes, hence the discretization of CWT can facilitate the calculation.

2.2.2.2. Discretization of continuous wavelet transform

Although the calculations are automatic, it is obvious that it is not evident to determine from the analytical equations or integrals the wavelet transforms. Thus, it is essential to switch to the discretization of these transformations. This section outlines the discretization of CWT, which is more applied, and its characteristics help to develop more applications different to the ones of CWT. To discretize the transform, the most intuitive way is to sample the time-frequency (scale) plane with a uniform sampling rate. For the wavelet transform, scaling can be used to reduce the sampling rate.

At higher scales (at lower frequencies), the sampling rate can be decreased, according to the Nyquist rule (The Nyquist sampling rate is the minimum sampling rate that allows the original continuous time signal to be reconstructed from its discrete samples). In other words, if the time-scale plane is to be sampled with a sampling rate of N_1 at the scale s_1 , the same plane can be sampled with a sampling rate of N_2 , at the scale s_2 , where, $s_1 < s_2$ (or $f_1 > f_2$) and $N_2 < N_1$: $N_2 = \frac{s_1}{s_2} N_1$; $N_2 = \frac{f_1}{f_2} N_1$.

So, the time parameter is discretised with respect to the scale parameter, i.e. a different sampling rate is used for each scale; sampling is done by dyadic sampling. A sample form is dyadic if the daughter wavelets are generated by expanding the wavelet mother by s_0^j and the translator by us_0^j . Dyadic sampling is optimal since the translation variable is sampled at the Nyquist frequency for any given frequency. Consider the area covered by the axes as the whole scale-time plane. The CWT assigns a value to the continuity of points on this plane. Therefore, there is an infinite number of CWT coefficients.

The discretization of the axis of is based on the called logarithmic rule. The base of the logarithm depends on the user; the most common value is 2 because of its convenience. If 2 is chosen, only the scales 2, 4, 8, 16, 32, 64 etc. are calculated. If the value is 3, the scales 3, 9, 27, 81, 243 etc. will be calculated.

The time axis is then discretised according to the discretization of the axis of the scale. Since the discrete scale changes by factors of 2, the sampling rate is reduced for the time axis by a factor of 2 at each scale.

For example, our sample has 64 points;

- For ($s = 2$), only 32 points of the time axis are sampled.
- For the value of the $s = 4$ scale, the sampling rate of the time axis is reduced by a factor of 2 since the scale is increased by a factor of 2 and therefore only 16 points are taken.
- In the next step, $s = 8$ and 8 points are taken in time.

- And so on.

Expressing the above method of discretization in mathematical terms, the discretization of the scale is $s = s_0^j$, and the translation discretization $\tau = k s_0^j \tau_0$, where the integer parameters $s_0 > 1$ and $\tau_0 > 0$. It has been noted that the discretization of the translation depends on the discretization of the scale with s_0 . Hence, the equation related to continuous wavelets, is rewritten as follow:

$$\psi_{u,s}(t) = s_0^{-j/2} \psi(s_0^{-j} t - k \tau_0) \text{ with } j, k \in \mathbb{Z} \quad (2.2.16)$$

ψ is called orthonormal wavelet in the Hilbert space. Hilbert's base is built as a family of functions $\{\psi_{j,k} : j, k \in \mathbb{Z}\}$ by a dyadic translation and a dilatation of ψ .

If $\{\psi_{j,k} : j, k \in \mathbb{Z}\}$ is an orthonormal basis, the series transformed into wavelets (coefficients of the wavelets) is the following one:

$$\Psi^{\psi_{j,k}}(j, k) = \int_{-\infty}^{\infty} f(t) \psi_{j,k}^*(t) dt \quad (2.2.17)$$

$$f(t) = c_{\psi} \sum_j \sum_k \Psi^{\psi_{j,k}}(j, k) \psi_{j,k}^*(t) dt \quad (2.2.18)$$

The family function $\{\psi_{j,k} : j, k \in \mathbb{Z}\}$ should be orthonormal or biorthogonal to define a wavelet series. If $\{\psi_{j,k} : j, k \in \mathbb{Z}\}$ are not orthonormal, then the series transformed into wavelets is written:

$$\Psi^{\psi_{j,k}}(j, k) = \int_{-\infty}^{\infty} f(t) \overline{\psi_{j,k}^*}(t) dt \quad (2.2.19)$$

AL-ANI (2013) mimed, the mathematical concept of discretization of CWT, to the vision concept.

2.2.2.3. Discrete wavelet transform

Along this section, the contribution of Discrete Wavelet Transform (DWT) will be analysed vis-à-vis the CWT and the Discretised Wavelet Transform.

In 1976, Croiser et al. created an approach to break down discrete signals over time. In the same year, Crochiere et al. did similar work on the coding of sound signals; they called their analysis approach as sub-band coding. In 1983, Burt and Adelson defined an approach very similar to sub-band coding and called it pyramid coding which is also known as multiresolution analysis. Later in 1989, Vetterli and Le Gall made some improvements to sub-band coding, removing existing redundancy in pyramid coding.

Although the Discretized Continuous Wavelet Transform allows CWT to compute automatically (by computers), it is not a true discrete transform. In fact, the wavelet series is simply a sampled version of the CWT, and the provided information is highly redundant with respect to signal reconstruction. This redundancy, on the other hand, requires a significant amount of time and computing resources including the rules of Nyquist.

According to Gençay et al. (2002), the major difference between CWT and DWT is mainly related that DWT uses a limited number of translated and expanded versions of the mother wavelet for the original signal decomposition which makes the implementation of the DWT easier than that of the CWT.

CWT is calculated by changing the scale of the analysis window, shifting it in time, multiplying it by the signal, and integrating all the time. In the discrete case, filters of different cut off frequencies are used to analyse the signal at different scales. DWT uses low-pass and high-pass digital filters in cascade and not in a parallel way like the CWT. Indeed, at each level of decomposition, the output signal of the low pass is again filtered and separated into two frequency bands which is sub-sampled, retaining only one sample out of two. Thus, the number of coefficients is constant from one stage to another. The bands denoted by d_i are said to be detailed since they comprise the elements with higher frequency content and the bands f_i are called the approximations.

The below equations presented in Table 2.1 show the decomposition from the discrete signal f_i of the level i into a high frequency part d_{i+1} and a low frequency part f_{i+1} , via the digital filters down ψ and high pass Φ .

DWT is based on two so-called discrete wavelets; mother wavelet $\psi_l = (\psi_0, \dots, \psi_{l-1})$ and father wavelet $\Phi_l = (\Phi_0, \dots, \Phi_{l-1})$.

The mother wavelet is characterized by three properties:

- $\sum_{l=0}^{L-1} \psi_l = 0$; This property shows that the mother wavelet also called differentiation function is a high-pass filter.
- $\sum_{l=0}^{L-1} \psi_l^2 = 1$
- $\sum_{l=0}^{L-1} \psi_l \psi_{l+2n} = 0$ for each integer $n \neq 0$

The three properties ensure that the mother wavelet preserves the variance of the original data and a multiresolution analysis can take place.

The scale wavelet or the father wavelet aims to capture the long scale of the low frequencies, it must respect this property: $\sum_{l=0}^{L-1} \Phi_l = 0$.

The application of the mother and father wavelet allows the separation between the low and high frequencies. In addition, a bandpass can be constructed from a successive recursive application low-pass and high-pass filters. Mathematically, the following table summarizes the above part:

Table 2.1. *Signal constructed by successive refinement*

Signal constructed by successive refinement	
Coefficients + details	
Father $\Phi(t)$	Mother $\psi(t)$
$\phi_{j,k}(t) = 2^{-\frac{j}{2}}\phi(2^{-j}t - k)$ Approximation coefficients at scale j $a(j, k) = \int_{-\infty}^{\infty} x(t) \phi_{j,k}(t) dt$	$\psi_{j,k}(t) = 2^{-\frac{j}{2}}\psi(2^{-j}t - k)$ Detail coefficients at scale j $W(a, b) \cong \int x(t)\psi_{j,k}^*(t)dt = d_{j,k}$
Approximation +	Detail
$f_j(t) = x_j(t) = \sum_{k=-\infty}^{\infty} a(j, k)\phi_{j,k}(t)$	$g_j(t) = \sum_{k=-\infty}^{\infty} d_{j,k}\psi_{j,k}(t)$

The discrete transform can be shifted to an orthogonal basis to obtain a non-redundant representation of the signal, which the CWT cannot provide. Thus, CWT is better suited for detection while the DWT enables more efficient compression of the data.

2.2.2.4. Multiresolution analysis

Although the problems of time resolution (How far two peaks in time can be separated from one another in the domain of transformation) and frequency resolution (How far two spectral components can be separated from one another in the field of transformation) are results of a phenomenon analogous to the Heisenberg uncertainty problem and exist independently of the transformation used, it is possible to analyse any signal using an alternative approach called multiresolution analysis.

A multiresolution analysis allows a good time resolution and a low frequency resolution at high frequencies. Also, a good frequency resolution and a low temporal resolution at low frequencies.

This approach makes sense, notably when the actual signal has high frequency components over short durations and low frequency components over long periods of time. Mathematically, it is often easier to have a signal development on a countable set of wavelets. This can be achieved by varying the scale factor with dyadic way by selecting: $= 2^j, j \in \mathbb{Z} u = k2^j, k \in \mathbb{Z}$.

In this context, a multiresolution analysis is defined as a sequence of closed vector subspaces $(V_j)_{j \in \mathbb{Z}}$ of $L^2(\mathbb{R})$ satisfying the following criteria:

- $\forall j \in \mathbb{Z}, \quad V_{j+1} \subset V_j$
- $\forall j \in \mathbb{Z}, f(t) \in V_j \Leftrightarrow f(\frac{t}{2}) \in V_{j+1}$
- $\lim_{j \rightarrow +\infty} V_j = \bigcup_{j \in \mathbb{Z}} V_j = \{0\}$ and $\lim_{j \rightarrow -\infty} V_j = \overline{\bigcup_{j \in \mathbb{Z}} V_j} = L^2(\mathbb{R})$
- ψ_0 exists such as $\{\psi_0(t - n), n \in \mathbb{Z}\}$ is an orthonormal base of V_0 ; ψ_0 is the scale function.

So, $\left\{2^{-\frac{j}{2}}\psi_0\left(\frac{t}{2^j} - k\right), k \in \mathbb{Z}\right\}$ construct orthonormal base of V_j .

The multiresolution analysis of a signal f consists in making successive orthogonal projections of the signal over the spaces V_j , which leads to increasingly coarse approximations of f as the value of j increases. The difference between two consecutive approximations represents the "detail" information that is lost when moving from one scale to another; this information is contained in the subspace W_j orthogonal to V_j such that: $V_{j-1} = V_j \oplus W_j$.

So, this demonstrates the existence of a wavelet $\psi_1 \in L^2(R)$ such as $\left\{2^{-\frac{j}{2}}\psi_1\left(\frac{t}{2^j} - k\right), k \in \mathbb{Z}\right\}$ is an orthonormal base of W_j .

The orthogonal wavelet decomposition of a signal f can be performed very efficiently by following the algorithm of Mallat (1989). Thus, the approximation coefficients $(c_{j,0}[k])_{k \in \mathbb{Z}}$ and of the details $(c_{j,1}[k])_{k \in \mathbb{Z}}$ are determined at each resolution level and are defined as follows:

$$\forall k \in \mathbb{Z}, \quad c_{j,0}[k] = \langle f, 1/2^{j/2}\psi_0\left(\frac{\cdot}{2^j} - k\right) \rangle \quad (2.2.20)$$

$$\forall k \in \mathbb{Z}, \quad c_{j,1}[k] = \langle f, 1/2^{j/2}\psi_1\left(\frac{\cdot}{2^j} - k\right) \rangle \quad (2.2.21)$$

Where $\langle \cdot, \cdot \rangle$ designed the dot product of $L^2(R)$.

Equitz and Cover (1991) provided a description related to the successive refinement of the information. Andreopoulos and Schaar (2008) worked on the incremental refinement of computation for the DWT.

2.2.3. Wavelets application

Wavelet analysis has been considered as a very powerful tool for compressing, processing and analysing data. At beginning, it was applied to extract information in images and audio

signals in Physics. Then, with the intensive theories' development, the application of wavelets was included in chemistry, biology and recently in Economics and Finance.

The main advantages of wavelets are resumed in Sagar et al. (2016):

Denoising the signal: via using suitable approach called *thresholding*, detail coefficients are discarded when it exceeds certain limit and the denoised signal is reconstructed by using both the coefficients.

Detecting long term evaluation: detecting the overall trend of the signal. As the approximation level increases the trend becomes clearer. Trend represents the slowest part of the signal. In terms of wavelet analysis, as the scale increases, resolution decreases. So, it produces better estimate of unknown trend. With reference to frequency, successive approximations have gradually less high frequency information. With the higher frequencies removed, what is left is the overall trend of the signal.

Splitting signal components: The wavelet transform is used to split the signal in terms of its details and approximate coefficients. The approximate coefficients represent the outlines and the detail coefficients represent detailed information. The detail coefficients are used to detect high frequency components and approximate coefficients are used to detect low frequency components.

Detecting discontinuities and breakdown points: These analyses are used to know at what exact instance the signal change occurs i.e. site of change, type of change, amplitude of change and discontinuities. By using detail coefficients at different levels, the measurement could be identified, and the noise could be stated. So, by using wavelet transform the break down points can be detected.

Multiscale analysis: Multiscale analysis of the signal as a sum of orthogonal signals corresponding to different time scale. So, it is time scale analysis.

Compression: The wavelet transform denoised the signal by applying appropriate thresholding rule. Thresholding means removing the coefficients which are responsible for noise. So, because of reduction in coefficients, the signal is compressed without any original signal degradation.

A lot of recent researches used the wavelets as a tool to analyse the data, and notably the big data. The wavelets analysis overcomes the main limitations of the classic spectral analysis used for times series. Indeed, it is flexible and does not require strong assumption about the data generating process regardless the stationarity assumption. Also, it provides information from both time-domain and frequency-domain. In addition, it is equipped with the ability to

locate discontinuities in the data. He and Nguyen (2015) provided more details about the application of wavelets in Economic and finance.

For the pharmaceutical field, wavelets have been largely used in imaging informatics; to analyse electrocardiogram (ECG) (Crowe et al., 1992; Jones et al., 1992) and in imaging informatics (Wang, 2001). Nguyen (2009) used the wavelet as a method of estimation in longitudinal data to study a random function X that represents patients or subjects observed at randomly distributed points. Pigoli et al. (2012) proposed a wavelet-based method to obtain accurate estimates of curves in more than one dimension and of their derivatives by utilizing Functional Data Analysis (FDA) that was considered very similar to longitudinal data by some researchers like Nguyen (2009).

2.3. Missing Data

This section covers principally three points. The first point defines what missing data are. The second one restates the definitions of missingness mechanisms introduced by Rubin (1976) and reaffirms their cruciality. The last point discusses the different imputation methods to treat missing observations.

2.3.1. Missing data in longitudinal studies

Day (1999) defined in the “Dictionary for Clinical Trials” the Missing data as *a data value that should have been recorded but, for some reason, was not*. It is referred to as non-response or unobserved data and occur in most types of studies. Different reasons can be behind the occurring of Missing values as sited in Dragset (2009) such as failure of measurement, data loss, out of range data and data loading issues, subjects fail to answer all questions, loss of follow-up or other plausible reasons. According to Hedeker and Gibbons (2006), even with well-controlled situations, missing data invariably occur in longitudinal studies. Daniels et al. (2007) discussed also the missing data in longitudinal studies.

Referring Schafer and Graham (2002), the types of missing data will be set as follows:

Unit nonresponse occurs when the entire data collection procedure fails because the sampled person is not at home, refuses to participate, etc.

Item nonresponse means that partial data are available: the person participates but does not respond to certain individual items.

Subjects may attend some stages of data collection and be absents for others. This missingness type can be addressed as *wave nonresponse*. Attrition, or dropout, which is a special case of wave nonresponse, occurs when a subject quits the study and does not return.

Overall, dropout or attrition may be the most common type of wave nonresponse. However, it is not uncommon for participants to be absent from one wave and subsequently reappear (Schafer and Graham, 2002).

As mentioned in Missing data's website of Bartlett and Carpenter, in the presence of missing data, the goal remains making inferences that apply to the population targeted by the complete sample i.e. the goal remains as if the complete data had been seen. However, both making inferences and performing the analysis become more complex. Indeed, setting assumptions in order to draw inferences and then use an appropriate computational approach for the analysis, are highly required. Wood et al. (2004) focused on highlighting the importance of understanding those assumptions or called also "mechanisms". Indeed, they confirmed that missing outcome data are a common problem in randomized controlled trials but are often inadequately handled in the statistical analysis in the top tier medical journals. They emphasized the indispensability of stating the assumptions underlying the handling of the missing outcomes and justify them through data description." (Wood et al., 2004).

Hedeker and Gibbons (2006) showed in detail the "types" of missing data and the appropriated mechanisms. Further details will be shown in the next section.

2.3.2. Missing data mechanisms

Hedeker and Gibbons (2006) defined missing data mechanism as what characterizes the reasons for missing observations i.e. data are missing. This section discusses the mechanisms that Wood et al. (2004) have considered as highly required to be known before handling missing data, as the performance of longitudinal data analysis models can depend critically on the missing data mechanism.

Those mechanisms had not been considered before Rubin (1976) who introduced the notation of the missing data which still in use as common notation. The data that are planned to be observed are denoted X_i for the covariate matrix and Y_i for the dependent variable vector, both for subject i and intentionally observed for n time points. Further the dependent variable can be split into vector Y_i in the observed variable vector Y_i^O and the missing variable vector Y_i^M , thus $Y_i = (Y_i^O ; Y_i^M)$ will be obtained.

R is the missing data distribution that describes the occurrence of missing data in the dataset. R_{ij} is a dichotomous variable (takes two levels of values), equal to one if the value of the dependent variable observed for subject i at time point j , and zero if Y_{ij} is missing. This missing data distribution is important to describe the reasons for missing data, the missing data

mechanism. There are three main missing data mechanisms described in the literature, these are denoted MCAR, MAR and MNAR.

2.3.2.1. *Missing completely at random*

Missing Completely at Random (MCAR) assumes that missing data occur totally, randomly and independently from the other observed or unobserved data. Indeed, a subject is missing at specific timepoint for completely random reason which means that the missing data indicators R_i are independent of both $(Y_i^O$ and $Y_i^M)$, do not depend on the dependent variable values that were observed or those that were not observed.

Pickles (2005) explained the condition of MCAR in a different way that for MCAR the probability of missingness is a constant.

Suppose that data on depression, marital status, family income, and diet will be collected. Data on depression would be considered Missing Completely at Random if the fact that an observation was missing had nothing to do with the person's potential scores on depression, marital status, family income, or diet. If people who eat an unbalanced diet were more likely to have missing data on depression, the data could not be considered MCAR.

It should be noted that the value of the observation, and not its "missingness" that is important. If people who refused to report personal income were also likely to refuse to report their level of depression, the data could still be considered MCAR, so long as neither of these had any relation to the income score, the depression score, or to any other variable. This is an important consideration, because when a dataset consists of responses to several survey instruments, someone who did not complete the Beck Depression Inventory, for example, would be missing all BDI sub scores, but that would not affect whether the data can be classed as MCAR. To be noted also that for a variable to be classed as MCAR, missingness cannot be related either to the level of that variable or to the level of any other variable in the dataset. It can, however, be related to the missingness of other variables.

If the data will include missing observations, then it is preferred that they are MCAR. This makes the situation far easier, though not necessarily easy, to deal with.

Because it is important to include predictors of missingness in a given analysis, Schafer and Graham (2002) recommended including a question like "*How likely is it that the participant will remain in this study through the next measurement period?*" in longitudinal questionnaires. To the extent that this question is related to subsequent missingness, including this variable as a covariate in analyses could convert a non-MCAR situation to one that is essentially MCAR.

Little (1988) provided a statistical test of the MCAR assumption which is the Chi-square test. A significant value indicates that the data are not MCAR. This test is available in SPSS Missing Values Analysis (MVA) and within the procedure PROC MI in SAS. Indeed, Howell (2008) mentioned that this test should be applied whenever there is some question about MCAR.

2.3.2.2. *Covariate dependent- MCAR*

A less stringent case of MCAR is the Covariate Dependent-MCAR (CD-MCAR) proposed by Little (1988). This missing data mechanism is dependent on the fully observed covariates X_i , defined in the previous section.

2.3.2.3. *Missing at random*

Often data are not MCAR, but they may be classifiable as Missing at Random (MAR). For data to be missing completely at random, the probability that X_i is missing is unrelated to the value of X_i or to the value of any other variable. But the data can be considered as missing at random if the data meet the requirement that missingness does not depend on the value of X_i even if it does depend on the value of some other variable. Thus, if the probability that an individual is missing a depression score due to that person's level of family status, income, or diet, the data can still be considered missing at random. However, the data is not MAR if the probability that a depression score is missing depends on the value of that depression score. If depression scores are more likely to be missing for people with poor diets, the data can still be considered missing at random (though not missing completely at random). But if depression scores are more likely to be missing for depressed individuals, the data is not MAR.

Howell (2008) pointed out that data could be missing at random even if there is some relationship between variable Y and missingness caused by the fact that both Y and missingness are correlated with X, so long as the relationship between Y and missingness disappears after controlling X. For example, using family income as the variable of interest, family income is certainly related to person income. If personal income is correlated with missingness, then the correlation between family income and missingness is almost certainly not 0. But if that correlation drops to 0 after controlling the personal income, then the data on family income is missing at random.

2.3.2.4. Missing not at random

If the data are not MAR or MCAR, they are classed as Missing Not at Random (MNAR). It would be found that this is the most difficult situation to deal with. In fact, there are times when it is impossible to deal with it at all realistically.

To summarize this in a nontechnical way, if data are missing for reasons that have nothing to do with the values of any of data variables, the data are missing completely at random. If data are missing for reasons that may have to do with the other variables, but not with this variable, then the variable is missing at random. Finally, if data are missing for reasons that have to do with the value of this variable, then the data are MCAR. To be highlighted that in all this one variable at a time was considered. It is whether depression scores are missing at random, and then speaking about whether family income is MAR, dataset can contain some MAR variables and other variables which are MNAR.

2.3.2.5. Ignorable and nonignorable missingness

MAR has been called noninformative or ignorable dropout, whereas MNAR is called informative (Diggle and Kenward, 1994). Indeed, Howell (2008) considered that having MNAR data facilitates the analysis.

Here, the mechanism controlling missing data in nonignorable which means that it is challenging to solve any model unless, a model that governs missingness, will be written. Modeling missingness is a very difficult thing to do, and most discussions, do not discuss the treatment of data whose missingness is nonignorable. On the other hand, if data are at least MAR, the mechanism for missingness is ignorable. Thus, the procedure can be followed without knowing the model for missingness. Here, it is not required to write a model that gets at missingness and there is no worrying about why the data are missing.

However, Schafer and Graham (2002) pointed out that “*good performance is often achievable through likelihood or Bayesian methods without specifically modeling the probabilities of missingness, because in many psychological research settings the departures from MAR are probably not serious.*”

The question that may be posed here is why the type of mechanism should be known. Daniels et al. (2007) emphasised the importance of following the mechanisms of missingness before choosing the method of handling missing data as the handling method depends on the mechanism. For instance, Multiple Imputation and Maximum Likelihood assume the data are at least missing at random. So, the important distinction here is whether the data are MAR as

opposed to MNAR. Listwise deletion, however, requires the data are MCAR in order to not introduce bias in the results.

2.3.3. Handling missing data

The following classification method was inspired from the paper of Horton and Kleinman (2007) who elaborated a very performant comparison between the different used methods to handle missing data. Further resources, mainly Schafer and Graham (2002) were used to set up the overview of this section.

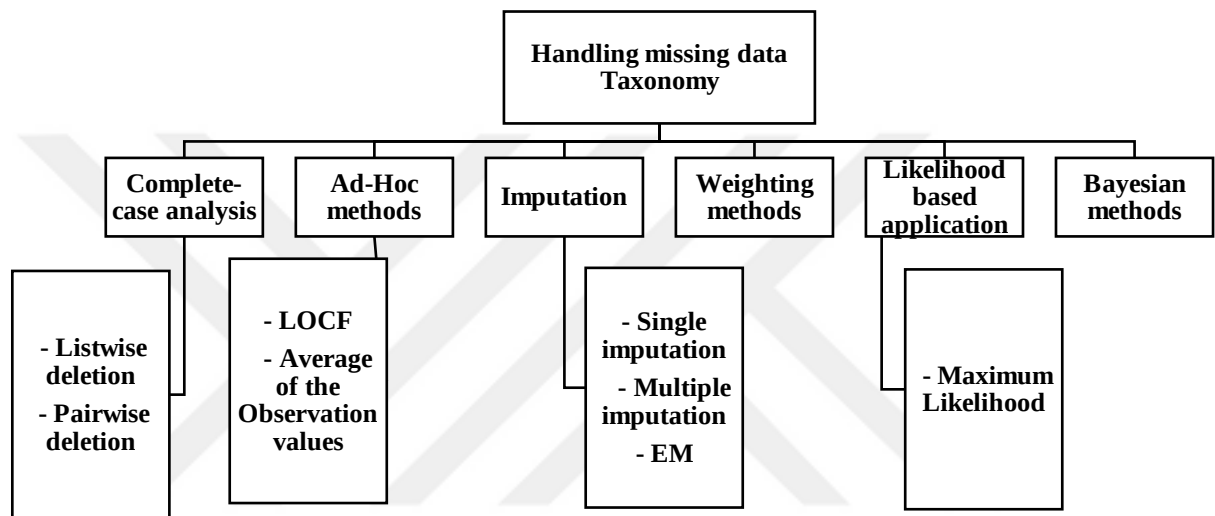


Figure 2.1. Handling missing data Taxonomy

2.3.3.1. Complete-case analysis

Known also as case deletion and *listwise deletion* (LD). Among older methods for missing data, the most popular is to discard units whose information is incomplete. Schafer and Graham (2002) found that LD is unbiased under MCAR. Under MAR and MNAR, the complete cases are unrepresentative of the population.

2.3.3.2. Ad-Hoc methods

The ad-hoc methods are mainly presented by two methods: Last Observation Carried Forward (LOCF) and Average of the Observation values.

Last Observation Carried Forward (LOCF): As the name indicates, the last available observation that a participant gave is entered the empty cells that follow. In the past the Food and Drug Administration (FDA) recommended this approach in clinical trials, but it has been

confirmed that LOCF leads to biased results and underestimates variability across trials (Howell, 2008).

Average of the observation values: Unlike case deletion, this method may introduce bias under MCAR (Schafer and Graham, 2002).

2.3.3.3. *Imputation*

This section mainly outlines three types of imputation: single imputation, multiple imputation and Expectation Maximization (EM).

Single imputation: When a unit provides partial information, it is tempting to replace the missing items with plausible values and proceed with the desired analysis rather than discard the unit entirely. Imputation, the practice of filling in missing items, has several desirable features. It is potentially more efficient than case deletion, because no units are sacrificed; retaining the full sample helps to prevent loss of power resulting from a reduced sample size. Moreover, if the observed data contain useful information for predicting the missing values, an imputation procedure can make use of this information and maintain high precision.

In single imputation, substituting an unknown missing value by a single value and then treat it as if it were a true value is a limitation of this method Rubin (1987). In fact, single imputation does not consider the uncertainty and mostly underestimates the variance. Little (1988) discussed further details related to the limitations of this approach.

Multiple imputation: started with Rubin (1987), the main idea was that certain types of multiple imputation can be viewed as approximations to data augmentation. Before that, some researchers consider that the development of mixture models for handling missing data could be outlined. (Rubin, 1976) described methods for using informative priors in surveys to capture and use subjective information about nonresponse.

Markov chain Monte Carlo (MCMC): in the presence of arbitrary missing data pattern, the MCMC method can be often utilized. MCMC consists in creating multiple imputations by using simulations from a Bayesian prediction distribution for normal data. (Schafer, 1997)

Expectation Maximum: In the late 1970's, Dempster et al. (1977) formalized the EM algorithm, a computational method for efficient estimation from incomplete data.

There are several ways to obtain maximum likelihood estimators, and one of the most common is called the Expectation-Maximization algorithm, abbreviated as the EM algorithm. The basic idea is simple, but the calculation is quite complex. To solve the EM algorithm by hand, firstly the estimates of the variances, covariances and means, should be taken (for example from listwise deletion). Then, those estimates will be used to solve the regression

coefficients, and then estimate missing data based on those regression coefficients. (In fact, whatever the data, $Y = bX + a$ should be estimated and then X will be used to estimate Y wherever it is missing.) This is the “estimation step of the algorithm”. Having filled in missing data with these estimates, then the complete data (including estimated values) to recalculate the regression coefficients will be used. This is the “maximization” step. Having new regression coefficients, the missing data would be re-estimated, new coefficients would be calculated, etc. This process would be completed until the estimates will no longer changed noticeably. At that point maximum likelihood estimates of the parameters have been gotten, and they could be used to make the maximum likelihood estimates of the regression coefficients.

The solution from the EM algorithm is better than coding for missing data, but it will still underestimate the standard errors of the coefficients (Howell, 2008).

There are alternative maximum likelihood estimators that will be better than the ones obtained by the EM algorithm, but they assume that the underlying model (usually the multivariate normal distribution) for the distribution of variables with missing data.

2.3.3.4. Maximum likelihood

This description and example were retrieved from (Howell, 2008). The principle of maximum likelihood is fairly simple, but the actual solution is computationally complex. (Howell, 2008) took an example of estimating a population mean, *because that illustrates the principle without complicating the solution.*

He supposed having the sample data (1, 4, 7, 9), the aim is to estimate the population mean. *It has been probably already known that the best estimate of the population mean is the sample mean* but another assumption will be considered. (Howell, 2008) supposed that the population was normally distributed, *simply because this makes the argument easier.*

Supposing that μ defines the population mean, the probability of obtaining a (1, 4, 7, 9) can be determined for a specific value of μ . This would be the product $p(1)*p(4)*p(7)*p(9)$. *It would probably be guessed that this probability would be very small if the true value of $\mu = 10$ but would be considerably higher if the true value of μ were 4 or 5.* (In fact, it would be at its maximum for $\mu = 5.25$.) For each different value of μ $p(1)$ and so on can be calculate and thus the product. For some value of μ this product will be larger than for any other value of μ . This is the maximum likelihood estimate of μ . It turns out that the maximum likelihood estimator of the population mean is the sample mean, because it is likely to obtain a (1, 4, 7, 9) if μ equals the sample mean than if it equals any other value.

The same principle applies in regression, although it is considerably more complicated. If a multivariate normal distribution is assumed, the maximum likelihood estimators for the means, variances, and covariance given the sample data, can be calculated. These are the values of those parameters that would make the data complete. Once these estimates are determined, they can be used to derive the optimal regression equation.

2.3.3.5. *Weighting methods*

The notion of reducing bias due to non-MCAR missingness by reweighting has a long history in the survey literature (Little, 1988).

Recently, biostatisticians begun to apply this idea in regression modeling with incomplete covariates. Robins, Rotnitzky and Zhao (1994) developed weighted regression that requires an explicit model for the missingness but reduces some of the parametric assumptions in the data model. Their method is an extension of Generalized Estimating Equations (GEE), a popular technique for modeling marginal or population averaged relationships between a response variable and predictors (Zeger et al., 1988). These models are called *semiparametric*, because they require the regression equation to have a specific form (e.g., linear or log-linear) but without specifying any probability distribution for the response variable itself.

2.3.3.6. *Bayesian method*

Schafer (1997) provides an approximation to the fully Bayesian data augmentation procedure.

2.4. Outliers

The following section aims to define the outliers and the main methods that have been used to handle them.

2.4.1.1. *What are outliers?*

Several definitions have been attributed to the outliers demonstrating that outliers were one of the earliest statistical concern (Barnett and Lewis, 1994). In 1852, Pierce defined them as the observations which differ from the others and they can mislead and perplex the user.

Currently, outliers' definition has remained almost the same: an observation which deviates so much from other observations, as to arouse suspicion it was generated by a different mechanism (Hawkins, 1980). For some researchers, the rule of thumb is that an outlier presents

any observation whose removal from the data changes the estimation of the variable of interest at least by 10% (Hansen et al., 1983).

The most common causes of outliers on a data set are data entry errors due to human errors, measurement errors, experimental errors during the experiment planning or data extraction, intentional i.e. dummy outliers could be included to test detection methods, data processing errors for example data manipulation or data set unintended mutations and sampling errors or natural showing the originalities in the data (see also Ghosh and Vogt, 2013).

Some of the popular methods for outlier detection which have been found in the literature are parametric methods and nonparametric ones. For the parametric, there are Z-Score or Extreme Value Analysis, probabilistic and Statistical Modeling, linear regression models (PCA, LMS). For the nonparametric, there are the proximity-based models.

Some procedures for detecting outlying observations in samples were provided in Grubbs (1969). Also, a detection of outliers in longitudinal count data via overdispersion was proposed in Gumedze and Chatora (2014).

The question of deleting outliers or keeping them arises always. It was deliberated in Pollet and Meij (2017), what extent removing outliers in hormonal data leads to divergent statistical conclusions. Statistical data preparation: management of missing values and outliers was discussed also in Kwak and Kim (2017). However, it is indispensable to outline that there is a study research arguing that the presence of outliers may indicate that one or more important variables were omitted from the model and adding additional variables to the model may then account for the outliers and improve the model fit Lieberman (2005).

2.4.1.2. Handling outliers

Hansen et al. (1983) proposed the rule of thumb that “*an outlier is any observation whose removal from the sample changes the estimate of a parameter of interest by 10 percent or more.*” For parameters with values near zero this is obviously suspect. It is also dependent on which parameter is of interest (Ghosh & Vogt, 2012).

Once the outliers are identified, it may be rectified by using one of the following approaches:

Trimming: Under this approach, a data set that excludes outliers is analyzed. The trimmed estimators such as mean decrease the variance in the data and cause a bias based on under or overestimation. Given that the outliers are also observed values, excluding them from the analysis makes this approach inadequate for the treatment of outliers.

Imputation: with mean, median or mode.

Capping/Winsorization: For values that lie outside the $1.5 * IQR$ limits, they are capped by replacing the observations outside the lower limit with 5th percentile and the ones bigger than the upper limit with 95th percentile.

Prediction: the outliers can be replaced with missing values Not available (NA) and then can be predicted by considering them as a response variable.

Robust estimation method: when the nature of the population distributions is known, this approach is considered appropriate because it produces robust estimators to outliers, and estimators are consistent. In the recent years, despite many studies have proposed a variety of statistical models for robust estimation, their applications are not frequent due to complicated methodological aspects. See also the provided description in Kwak and Kim (2017).

Wavelets as outliers handling method: wavelets can detect the discontinuities and breakdown points. Indeed, they have been used to know the exact time when the signal change occurs i.e. site of change, type of change, amplitude of change and discontinuities. Via the detail coefficients at different level, the measurement could be identified, and the noise could be stated. Therefore, by using wavelet transform to detect the break down points is possible.

Li and Reese (2009) applied the wavelets analysis to identify the turning points in economics cycles and to not misidentify them as outliers. Similarly, Bai et al. (2015) used the Discrete Wavelet Transform (DWT) to decompose the internal residue and further extract its middle-frequency signals to detect the turning points, they used stock markets data.

Al-Wadi (2015) adopted the modern Wavelet function known as the Maximum Overlapping Wavelet Transform (MODWT) to detect the outliers in the financial time series Amman Stock Exchange (ASE). The results of their study showed that their updated technique, called the Maximum Overlapping Wavelet Transform - Asymmetric Winsorized Mean (MVAW) was considered the best method in outlier detection comparing to the work of Al-Khazaleh et al. (2015).

Sagar et al. (2016) used the multilevel features of wavelets to detect and to remove the outliers by recomposing the Electricity consumption of France dataset, removing the detail coefficients which present high frequency without modifying the approximate coefficients then they reconstructed the data. They demonstrated that they got the reconstructed data without losing much information i.e. reconstructed signal (after removing the outliers) is very similar to the original signal (with outliers). Recently, the wavelets have received more attention to detect outliers and handle them, however the studies related to this area still under development.

2.5. Modeling

For a design which includes both random and fixed effect, a mixed model is often used. In mixed models, missing observations have no effect on the observed data. In addition, it is not required to be consistent regarding the time, for example, if one subject has a follow-up test at 3 months while another subject has his follow-up test at 4 months, the time of follow-up is simply considered as (3 and 4), there is no issue that both subjects couldn't be tested at the same intervals. Also, with mixed models, the sphericity assumption or the compound symmetry assumption in the model is not required to be tested. The model can be allowed to select its own set of covariances or use covariance patterns that the user supply. (Howell, 2008; Maxwell et al., 2003).

It is crucial to distinguish random/fixed variable from random/fixed coefficient and from random/fixed effects. This will be discussed in the next section.

2.5.1. Random versus fixed (variables, effects, coefficients)

The following section is largely benefited from (Kreft and De Leeuw, 1998; Torres-Reyna, 2007; Newsom, 2017). The main objective consists in providing a clear description to distinct between random variable and fixed variable, random effect and fixed effect and finally random coefficient and fixed coefficient.

2.5.1.1. *Fixed variable and random variable*

It is crucial to differentiate a variable that is varying from a variable that is random. Different values might be attributed to a fixed variable, it is not strictly equal or fixed across groups.

Fixed variable is assumed to be measured without measurement error, desired generalization to population or other studies is set to the same values. The variable used in the study contains all or most of the variable's values in the population. For instance, gender, race and intervention vs. control group could be considered as fixed variables. However, a random variable is assumed to be measured with measurement error. The values come from and are intended to generalize to a much larger population of possible values with a certain probability distribution (e.g., normal distribution). In addition, the number of values in the study is small relative to the values of the variable as it appears in the population it is drawn from. An example to illustrate the random variable is the photographs representing individuals with differing levels of attractiveness manipulated in an experiment.

The following section presents the difference between random effect and fixed effect.

2.5.1.2. *Fixed effect and random effect*

Mostly, the terms Random Effect (RE) and Fixed Effect (FE) are utilized in the context of ANOVA and regression models and refer to a certain type of statistical model. Often, researchers use fixed effect regression or ANOVA and they are rarely faced with a situation involving random effects analyses. Fixed effect ANOVA states assumptions related to the independent variable and the error distribution. Usually, the researcher is interested in only generalizing the results to experimental values used in the study. For instance, a drug study might use 0 mg, 5 mg, and 10 mg of an experimental drug. This is a circumstance when a fixed effect ANOVA would be appropriate. In this example, the extrapolation is to other studies or treatments that might use the same values of the drug (i.e., 0 mg, 5 mg, and 10 mg). However, if the researcher wants to make inferences beyond the values of the independent variable used in the study, a random effect model is used. “...the crucial distinction between fixed and random effects is whether the unobserved individual effect embodies elements that are correlated with the regressors in the model, not whether these effects are stochastic or not” (Greene, 2008)

Fixed Effect (FE) is used whenever there is an interest in analysing the impact of variables that vary over time.

FE explores the relationship between predictor and outcome variables within a subject (country, person, company, etc.). Each subject owns individual features that may or may not impact the predictor variables (for instance, gender could impact the view toward a specific issue, or the political system of a country could have some effect on trade or GDP, or the business practices of a company may impact its stock price). The assumptions of FE are:

1. Assuming something within the individual may impact or bias the predictor or outcome variables and it should be controlled. FE removes the effect of those time-invariant characteristics, so the net effect of the predictors on the outcome variable can be assessed.
2. Time-invariant features are unique by subject and should not be correlated with other subject features. Each subject is different therefore the subject's error term and the constant (which captures individual characteristics) should not be correlated with the others. If the error terms are correlated, then FE is not suitable since inferences may not be correct and that relationship should be modelled (probably using random effects).

Therefore, FE is a statistical model typically used in regression and ANOVA assuming independent variable is fixed, and the generalization of the results applies to similar values of independent variable in the population or in other studies.

If there is a reason to believe that differences across subjects have some influence on the dependent variable, then the random effects (RE) are required to be used. The rationale behind random effects model is that, unlike the fixed effects model, the variation across subjects is assumed to be random and uncorrelated with the predictor or independent variables included in the model. The assumptions of the RE are:

1. The subject's error term is not correlated with the predictors which allows for time-invariant variables to play a role as explanatory variables.
2. In RE, it is needed to specify those individual characteristics that may or may not impact the predictor variables. The problem here is that some variables may not be obtainable which may lead to a bias in the model.

Thereby, RE is a different statistical model of regression or ANOVA model which assumes that an independent variable is random. It is generally used when the levels of the independent variable are thought to be a small subset of the possible values, which one wishes to generalize, but probably produce larger standard errors.

The last paragraph in this section shows the difference between random coefficients and fixed coefficients.

2.5.1.3. *Fixed coefficient and random coefficient*

A fixed coefficient (FC) can be fixed to be non-varying (invariant) across groups by setting the between-group variance to zero such as slopes or intercepts constrained to be equal over different schools. A random coefficient (RC) applies only to Multilevel Regression (MLR) analyses in which intercepts, slopes, and variances can be assumed to be random. MLR mostly assume random coefficients. For example, the level-2 predictor average income, is utilized to predict each school performance. Intercept values for school performance are assumed to be a sample of the intercepts from a larger population of schools.

After discussing fixed and random for the variables, effects and coefficients, the next section aligns the framework of the mixed effects models. Three models will be discussed: parametric mixed effects model, non-parametric mixed effects models and semi-parametric mixed effects models.

2.5.2. Parametric mixed effects models

Linear Mixed Effects Models (LMEM) and Nonlinear Mixed Effects (NLME) models are the two most popular examples. Several books have been published to summarize the achievements in these areas (Diggle et al., 2002; Hulin and Jin-Ting, 2007; Ruppert et al., 2003) among others.

2.5.2.1. Linear mixed effects models

The linear mixed model is defined as following:

$$\mathbf{y}_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta} + \mathbf{z}_{ij}^T \mathbf{b}_i + \boldsymbol{\varepsilon}_{ij} \quad (2.5.1)$$

$\boldsymbol{\beta}$: the parametric fixed effects (population parameters),

$\mathbf{b}_i$: the parametric random effects (individual parameters),

$\boldsymbol{\varepsilon}_{ij}$: the measurement errors,

$\mathbf{x}_{ij}^T$: the associated fixed effects design matrix,

$\mathbf{z}_{ij}^T$: random effects design matrix,

$\mathbf{y}_{ij}$: the vectors of responses.

2.5.2.2. Non-linear mixed effects models

The nonlinear mixed model is defined as following:

$$\mathbf{y}_{ij} = \mathbf{g}(\mathbf{x}_{ij}, \mathbf{z}_{ij}, \boldsymbol{\beta}, \mathbf{b}_i) + \boldsymbol{\varepsilon}_{ij} \quad (2.5.2)$$

$\mathbf{b}_i \sim N(0, D_b)$; $\boldsymbol{\varepsilon}_{ij} \sim N(0, \sigma^2)$ iid

$\mathbf{g}(\cdot)$ is some known nonlinear function

n : is the number of subjects,

n_i : the number of observations for the i^{th} subject,

D : the variance components of the LME model,

$\boldsymbol{\beta}$: the parametric fixed effects (population parameters),

$\mathbf{b}_i$: the parametric random effects (individual parameters),

$\boldsymbol{\varepsilon}_{ij}$: the measurement errors,

$\mathbf{x}_{ij}^T$: the associated fixed effects design matrix,

$\mathbf{z}_{ij}^T$: random effects design matrix,

$\mathbf{y}_{ij}$: the vectors of responses.

The main advantages of the parametric Mixed Effects Model (MEF) are that those models could consider many covariates, they are easy to be analysed via EM algorithm, in addition the methodologies related to parametric MEF are well developed. However, those models need always valid parametric assumptions like the normality as they may mislead the conclusions once they are mis-specified.

2.5.3. Nonparametric mixed effects models

A parametric regression model requires the assumption that the form of the underlying regression function is known except for the values of a finite number of parameters. The selection of such type of model depends always on the proposed problem.

Sometimes the parametric model can be derived from theories behind the scientific problem, whereas at other times the model is based on experience or is simply assumed from scatter plots of the data. A serious drawback of parametric modeling is it may be too restrictive in some applications.

If an inappropriate parametric model is used, it is possible to produce misleading conclusions from the regression analysis. In other situations, a parametric model may not be available to use. Therefore, the nonparametric models will be selected to be used. (Rice and Wu, 2001; Ruppert et al., 2003; Hulin and Jin-Ting, 2007).

Nonparametric Mixed Effects Models (NPMEF) are flexible to fit longitudinal data, furthermore they are robust against model misspecification. But, this type of models involves few covariates (contrary to the parametric MEF), they may be computationally intensive and the corresponding methodologies still always under developing.

Two main types of NPMEF can be defined: Population Mean Mixed Effects Models and Varying Coefficient Mixed Effects Models.

The population mean mixed model is written as:

$$y_{ij} = \eta(t_{ij}) + v_i(t_{ij}) + \varepsilon_{ij}(t_{ij}) \quad (2.5.3)$$

$$v_i(t) \sim N(0, \gamma); \varepsilon_{ij} \sim N(0, \gamma_0)$$

$\eta(t_{ij})$: the nonparametric fixed effect function (the smooth population mean function smooth FE function),

$v_i(t)$: nonparametric random effect function (smooth RE function for the i^{th} subject),

$\varepsilon_{ij}(t)$: the measurement error for the j^{th} measurement for the i^{th} subject,

t_{ij} : denote the design time points (days, weeks, months etc.),

y_{ij} : the responses observed at t_{ij} .

Equation 2.2.4 defines the Varying Coefficient Mixed Effects Models:

$$\mathbf{y}_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta}(\mathbf{t}_{ij}) + \mathbf{v}_i(\mathbf{t}_{ij}) + \boldsymbol{\varepsilon}_{ij}(\mathbf{t}_{ij}) \quad (2.5.4)$$

$\boldsymbol{\beta}(t)$ is the vector of some unknown smooth coefficient functions.

$i = 1, 2, \dots, n; j = 1, 2, \dots, n_i$

n is the number of subjects, n_i present the number of observations for the i^{th} subject.

2.5.4. Semiparametric mixed effects models

As documented in Hulin and Jin-Ting (2007), for Semiparametric Mixed Effects models (SPMEF), part of the variations in the response variable can be explained by given parametric models of some covariates in the fixed effect component and/or the random effect component, while the remaining is explained by a nonparametric function of time. In a time, varying coefficient mixed effects model, the coefficients of the fixed effects and random effects covariates are smooth functions of time. These two kinds of models are very important and useful in practical longitudinal data analysis.

Ruppert et al. (2003) showed that semiparametric regression incorporates nonlinear functional relationships in regression analyses and semiparametric regression can be utilized for several fields which apply regression analysis.

Therefore, the current objective is to construct a SPMEF for longitudinal data, after studying the data (what is missingness mechanism to be chosen, how to handle the missing values and to detect the outliers and handle them).

The following equation 2.5.5. presents the general format of the semiparametric model as presented in Hulin and Jin-Ting (2007).

$$\mathbf{y}_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta}(\mathbf{t}_{ij}) + \boldsymbol{\eta}(\mathbf{t}_{ij}) + \mathbf{z}_{ij}^T \mathbf{b}_i + \mathbf{v}_i(\mathbf{t}_{ij}) + \boldsymbol{\varepsilon}_{ij}(\mathbf{t}_{ij}) \quad (2.5.5)$$

$\boldsymbol{\beta}$: parametric fixed effects (population parameter),

$\mathbf{b}_i$: parametric random effects (individual parameters),

$\mathbf{v}_i(t)$: nonparametric random effect function,

$\boldsymbol{\eta}(t)$ nonparametric fixed effect function,

$\boldsymbol{\varepsilon}_{ij}(t)$ the measurement errors,

$\mathbf{x}_{ij}^T$ and $\mathbf{z}_{ij}^T$ are the associated fixed effects and random effects design matrices,

n is the number of subjects, n_i the number of observations for the i^{th} subject.

$\mathbf{v}_i(t) \sim GP(0, \gamma); \mathbf{b}_i \sim N(0, D_b); \boldsymbol{\varepsilon}_{ij} \sim N(0, \gamma_0)$.

Dan Li (2015) applied the semiparametric models on medical monitoring data.

2.6. Smoothing Approaches

Parametric modeling has restrictions which may not be satisfied by non-linear models. To overcome some of those restrictions, smoothing methods have become largely applied in data analysis and in inferential analysis.

Next paragraph that describes the basic idea behind the smoothing technics, is based on Hulin and Jin-Ting (2007).

The basic idea of these nonparametric approaches is to let the data determine the most suitable form of the functions. There are one or two so-called smoothing parameters in each of these methods for controlling the model complexity and the trade-off between the bias and variance of the estimator. For example, the bandwidth h in local kernel smoothing determines the smoothness of the regression function and the goodness of fit of the model to the data so that when $h = \infty$, the local nonparametric model becomes a global parametric model; whereas when $h = 0$, the resulting estimate essentially interpolates the data points. Thus, the boundary between parametric and nonparametric modeling may not be precise if one takes the smoothing parameter into account.

As described in Hulin and Jin-Ting (2007), kernel smoothing, local polynomial fitting, polynomial regression splines, smoothing splines, penalized splines, Locally Weighted Scatter Plot Smoothing (LOWESS), wavelet-based methods and other orthogonal series-based approaches could be considered as the frequently used smoothing approaches.

Ruppert et al. (2003) determined four major smoothing approaches to nonparametric modeling: *smoothing splines* (Wahba, 1990; Green and Silverman, 1994; Eubank, 1988, 1999); *series-based smoothers*, including wavelets (Tarter and Lock, 1993; Ogden, 1996); *kernel methods*, including local regression (Wand and Jones, 1995; Fan et al., 1996); and *regression splines* (Friedman, 1991; Stone et al., 1997; Hansen and Kooperberg, 2002).

Ruppert et al. (2003) mentioned that the type of data is the reference to choose among those approaches. For instance, LOWESS is more convenient for non-seasonal series, RLOWESS are more used with datasets that contain outliers, and linear smoothers are appropriate for the linear combinations.

Ruppert et al. (2003) focused on the *penalized splines*, labelled also as *P-splines*, *pseudo splines*, and *low-rank spline smoothers* in the literature.

The research related to nonparametric methods applied on longitudinal data had not been largely investigated until 1988. Müller (1988) could be considered as the first reference of introducing nonparametric regression methods in longitudinal data analysis. Indeed, the

concept of the approach of Müller (1988) is estimating distinctly each individual curve, consequently the within-subject correlation of the longitudinal data was not examined in the modeling which make the methodologies of Müller (1988) essentially similar to the nonparametric regression methods for cross-sectional data.

Recently, the development of nonparametric regression methods for longitudinal data analysis, including utilization of kernel-type smoothing methods, have been largely exposed (Hoover et al., 1998; Fan and Zhang, 2000; Wu and Zhang, 2002b; Wang et al., 2005), smoothing spline methods (Brumback and Rice, 1998; Y. Wang, 1998a, 1998b) and regression (polynomial) spline methods (Verbyla et al., 1999, Rice and Wu, 2001; Wu and Zhang, 2002a; Liang et al., 2003).

The following sections outline some principal smoothing methods.

2.6.1. Kernel smoothing methods

Kernel smoothing approach is referred as one of the most known technics for nonparametric estimation of function. As provided in (Wand and Jones, 1995), for the case of simple non-parametric regression or the case of a scatter plot, the kernel smoothing facilitates the estimation of the regression function without the assumption of a parametric model. This approach has been extended to further fields of application such as engineering, medicine and economics. The following graph shows an example of the kernel smoother for the variable Systolic Blood Pressure (SYSBP).

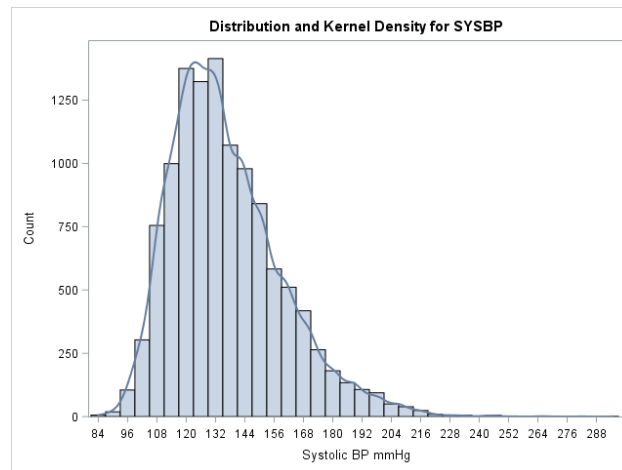


Figure 2.2. *Distribution and Kernel density for Systolic Blood Pressure*

A kernel smoother defines a set of weights $\{W_i(x)\}_{i=1}^n$ for each x and defines the following function 2.6.1.:

$$\hat{f}(x) = \sum_{i=1}^n w_i(x) y_i \quad (2.6.1)$$

In general, the most of smoothers could be referred as kernel smoother. The specificity of the kernel smoother in practice that it has a simple approach to represent the weight sequence $\{w_i(x)\}_{i=1}^n$ through describing the shape of the weight function $w_i(x)$ via a density function with a scale parameter that adjusts the size and the form of the weights near x . This shape function is referred as *Kernel K*. The kernel is a continuous, bounded, and symmetric real function K which integrates to one (Equation 2.6.2.):

$$\int K(u) du = 1 \quad (2.6.2)$$

For a given scale parameter h , the weight sequence is defined by:

$$w_{hi}(x) = K\left(\frac{x-x_i}{h}\right) / \sum_{i=1}^n K\left(\frac{x-x_i}{h}\right), \text{ with } \sum_{i=1}^n w_{hi}(x) = 1 \quad (2.6.3)$$

The kernel smoother is defined for each x as before by:

$$\hat{f}(x) = \sum_{i=1}^n w_{hi}(x) y_i \quad (2.6.4)$$

As it has been considered, that the closer points are similar, a kernel smoother usually defines weights that decrease in a smooth fashion as one moves away from the target point.

Two widely used kernel functions are the gaussian kernel and the uniform kernel. (Silverman, 1982; Ramsay, 1991; Scott, 1992; Wand and Jones, 1995; Hulin and Jin-Ting, 2007; Irizarry, n.d.; OpenTurns, n.d.)

2.6.2. Local polynomial smoothing

Local polynomial kernel estimators are one of the most known class of kernel-type regression estimators. A local polynomial smoother fits a locally weighted regression at each point x to produce the estimate at x . Different types of regression and weight functions could be used in the estimation: a locally weighted mean, a locally weighted regression line and a locally weighted quadratic polynomial regression. The weights are derived from a single function that is independent of the design:

$$W(x, x_i; \lambda_i) = K_0\left(\frac{x - x_i}{\lambda_i}\right) \quad (2.6.5)$$

with K_0 a weight function and λ_i is the local bandwidth at x_i . (Cleveland, 1979; Müller, 1987; Wand and Jones, 1995; Fan et al., 1996; SAS institute, 1999; Littell et al., 2007; Hulin and Jin-Ting, 2007).

The following figures 2.3 and 2.4 show the graphic presentation of the local linear fitting and local quadratic fitting.

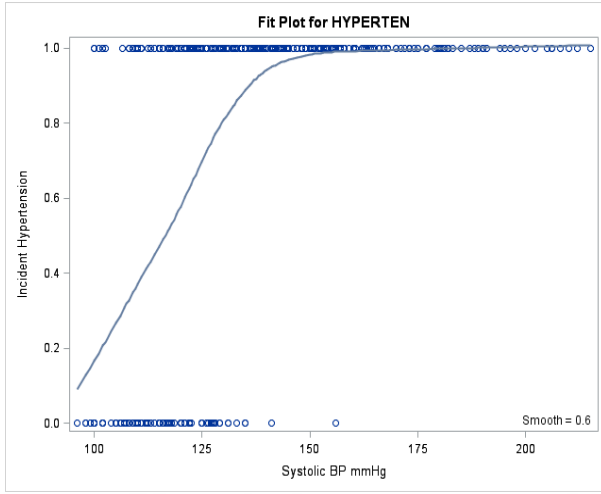


Figure 2.3. Local linear fitting for hypertension

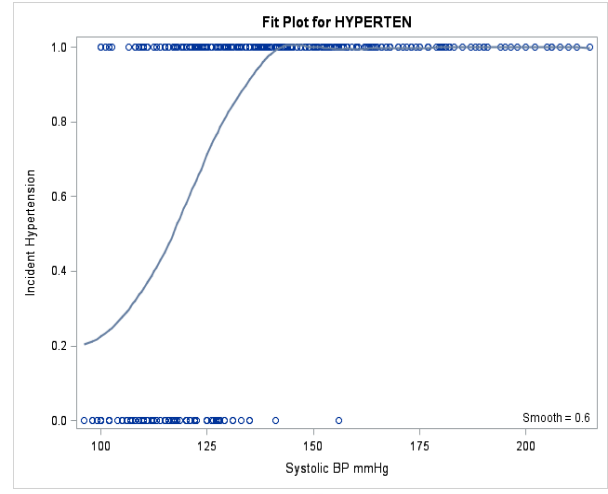


Figure 2.4. Local quadratic fitting for hypertension

It should be noted here that the graphic presentations are not provided for the big data and this is a limitation of this smoothing approach. Indeed, this was confirmed by Hulin and Jin-Ting (2007) that local polynomials are not flexible in their ability to model data across a large range of values.

2.6.3. Regression (polynomial) splines

Regression splines have been considered as one of the most popular nonparametric regression techniques. In local polynomial kernel smoothing, local neighbourhoods were specified by a bandwidth h and a fixed time point t_0 . In regression spline smoothing that local neighbourhoods are specified by a group of locations, $\tau_0, \tau_1, \tau_2, \dots, \tau_K, \tau_{K+1}$. For a defined interval $[a, b]$ with $a = \tau_0 < \tau_1 < \tau_2 < \dots < \tau_K < \tau_{K+1} = b$, the locations are known as knots, and $\tau_r, r = 1, 2, \dots, K$ are called interior knots or simply knots. Those knots divide the interval of interest $[a, b]$ into K subintervals (local neighbourhoods): $[\tau_r, \tau_{r+1}), r = 0, 1, \dots, K$. So, within any enharbouring knots, a Taylor's expansion up to some degree is valid.

The regression spline could be constructed via k^{th} degree truncated power basis defined as:

$$\Phi_p(t) = [1, t, \dots, t^k, (t - \tau_1)_+^k, \dots, (t - \tau_K)_+^k] \quad (2.6.6)$$

With $(w)_+^k$ denotes power k of the positive part of w with $w_+ = \max(0, w)$, $p = K + k + 1$ denotes the number of the basis functions involved. Similarly, the associated coefficients are denoted as: $\beta = [\beta_0, \beta_1, \beta, \dots, \beta_K, \beta_{K+1}]^T$.

Therefore, the regression spline could be re-written as 2.6.7:

$$f(t) = \Phi_p(t)^T \beta \quad (2.6.7)$$

Further details are provided in (Hulin & Jin-Ting, 2007).

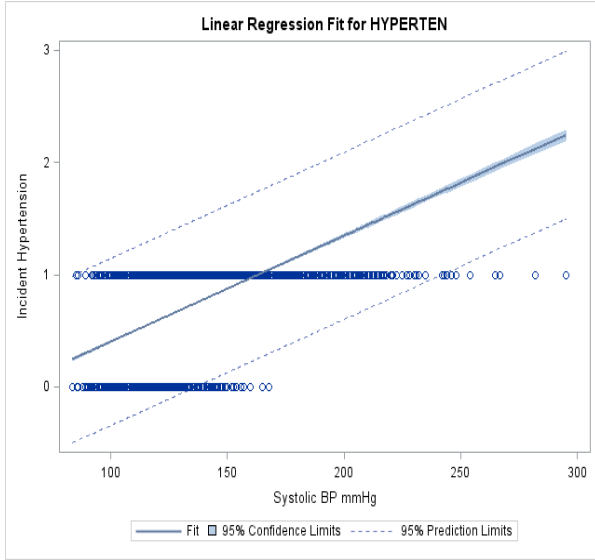


Figure 2.5. Linear regression fitting for hypertension

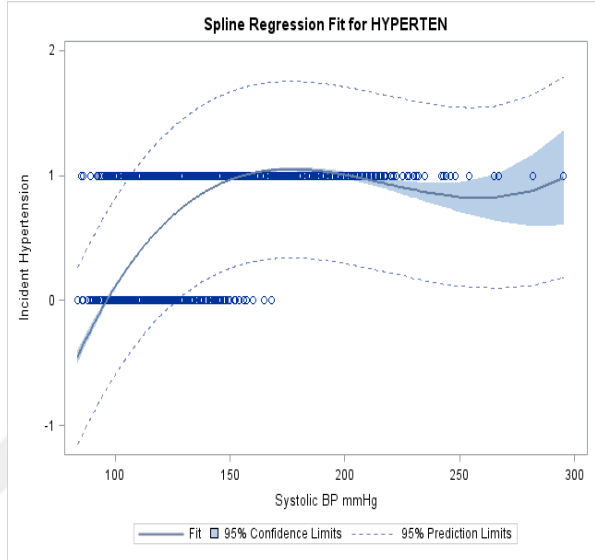


Figure 2.6. Spline regression fitting for hypertension

The above figures 2.5 and 2.6 show an example of linear regression fit versus a spline regression fit.

2.6.4. Smoothing splines

For regression splines, when the knot locating method is specified, the remaining task is to choose the number of knots. In general, K is smaller than the sample size n . Since K must be an integer, the opportunity for adjusting K is limited, and the adjustment is usually rather crude. Alternatively, all the distinct design time points can be used as knots. This will generally result in under smoothing since too many knots are in use. The resulting curve is usually quite rough, showing dramatic curve changes over a short interval.

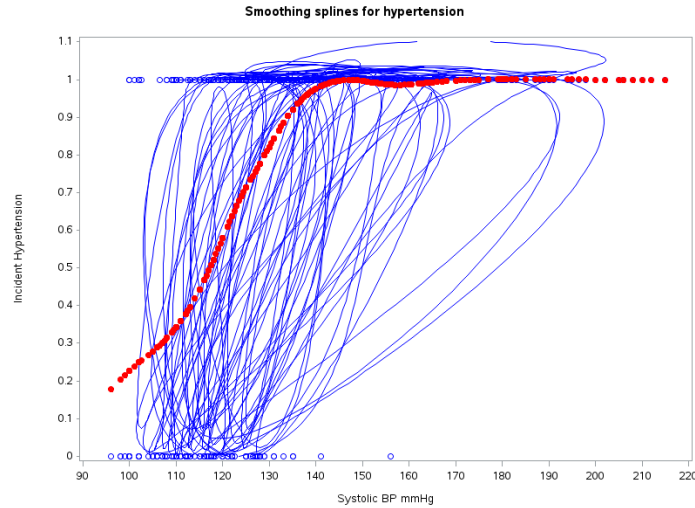


Figure 2.7. *Smoothing splines for hypertension*

Smoothing splines overcome this difficulty by introducing a penalty for the roughness of the curve under consideration. (Hulin & Jin-Ting, 2007)

2.6.5. Penalized splines

Penalized splines are also called as P-splines, pseudo-splines, and low-rank spline smoothers in the literature. Penalized splines are like smoothing splines as they are a generalization of smoothing splines that allow more flexible choices of the spline model, the basic functions for that model, and the penalty. Penalized splines have close ties with ridge regression, mixed models, and Bayesian statistics, ties that were discovered by researchers working on smoothing splines. These ties allow techniques from mixed models, for example, (restricted) maximum likelihood estimation and likelihood ratio tests, to be added to penalized spline methodology. Similarly, Bayesian techniques based on Markov chain Monte Carlo provide what can be considered as the most satisfactory approach to fit complex semiparametric models as well as the direction that semiparametric regression is most likely to take in the future (Ruppert et al., 2003)

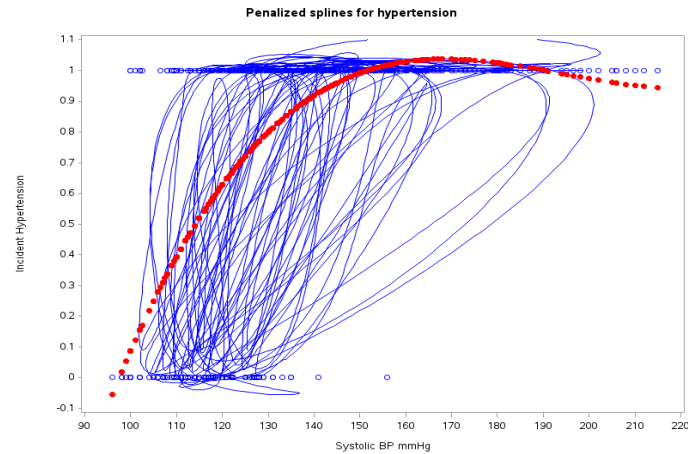


Figure 2.8. *Penalized splines for hypertension*

The performance of the regression splines depends strongly on the location of knots and the number of knots. Smoothing splines overcome these drawbacks via taking all the distinct time points as knots and using a roughness penalty to control the smoothness of the resulting smoothers. When the number of distinct time points is large, the number of the parameters to be estimated will be large. This is one of the reasons why smoothing splines are expensive to compute along with the necessity of computing the integral that defines the roughness of the underlying function. (Hulin & Jin-Ting, 2007)

3. NOVEL APPROACHES

This chapter outlines the novel methods of this dissertation. First section discusses the imputation of missing data via NN and presents the novel algorithm Multilayer Perceptron for Imputation of Missing Data that will be addressed as MLPIMID. Second section frames the novel algorithms to address the outliers by modifying two classic methods: winsorisation to be applied within subjects and presented here as WINWIS and the EM will consider the outlier as missing observation and the method will be called EMIMID. Also, two novel approaches will be introduced to treat the outliers: Wavelets Decomposition for Outliers Detection and Handling Within Subjects (WDODHWIS) and Wavelets Decomposition for Outliers Detection and Handling Across Subjects termed as (WDODHAS). Last section explains how wavelet analysis can be considered as a smoothing approach.

3.1. Imputation of Missing Observations

Several researches such as Bishop (1996), Norway (1997), Yoon and Lee (1999) and Sonnberger and Maine (2000) have studied addressing the missing data via neural network (NN). As noted by Silva-Ramírez et al. (2011), for most of the studies, training the NN is applied only on the not missing and for each variable separately. Indeed, the target value is observed, so the generated NN is proceeded for imputing the missing observations (Norway, 1997; Nordbotten, 1998; Koikkalainen, 2002; Euredit, 2005). Silva-Ramírez et al. (2011) used the Multilayer perceptron (MLP) to impute data on MCAR, they proceed the training subset in a way that it includes observed and missing data. Smieja et al. (2018) addressed the missing data by NN, their proposed approach consists in replacing typical neuron's response in the first hidden layer by its expected value.

Parveen and Green (2001), Lipton et al. (2016), Choi et al. (2015) used Recurrent Neural Networks (RNNs) by combining missing entries with the input or by applying simple imputations. In the same context, Che et al. (2018) developed a novel model called Gated Recurrent Unit (GRU) to better predict the missing patterns via RNN.

As wavelets analysis has gained a lot of attention during the last two decades, it has been applied also for imputation missing data. Mondal and Percival (2010) applied the wavelet variance analysis (defined as a scale-based decomposition of the process variance) for gappy time series. Heaton and Silverman (2008) proposed a new approach of imputation by utilizing the expected sparse representation of a surface in a wavelet or lifting scheme basis.

Rocha, Paredes and Henrique (2010) utilized a wavelet scheme to reconstruct the missing sections in time series signals based on the estimation of the available sections. Altan and

Ustundag (2012) reconstructed the missing observations of a meteorological data via the wavelet transform. Lilly (2017) developed an interesting platform to analyse and visualize time-localized events in noisy time series using wavelets analysis.

It is indispensable to highlight that there are several researches which applied the Wavelet Neural Network (WNN) to impute missing data. WNN concept was invented by Zhang and Benveniste (1992), it is based on the wavelet transform and it is considered as an alternative to feedforward NN for approximating random nonlinear functions. Then, several studies have utilized WNN for different applications such as Yu and Chen (2007) who used the WNN to differentiate six different beat types in ECG signal, Lung (2007) applied WNN for features selection and recognition of text independent speaker by using multi-layered neural network for speaker identification, Xie and Zhang (2007) utilized WNN to predict a short term traffic volume and Dong et al. (2008) who applied fuzzy WNN and irregular sets to forecast fault diagnosis accuracy of power transformers. Panigrahi et al. (2012) investigated how to remove and interpolate the missing data by WNN. Wang et al. (2013) explored the WNN by applying multiple wavelet Functions in target threat assessment.

Next section outlines briefly the theoretical background of Neural Network that will be utilized in MLPIMID.

3.1.1. Neural network theoretical background

Neural Network (NN) is defined as a set of processing series called neurons, forming a highly parallel distributed network like the human cognitive process. It consists in transferring information. Perceptron is the simplest form of a NN that forms one neuron. It is constructed around a nonlinear neuron known as the McCulloch-Pitts model of a neuron.

For SAS algorithm, the multilayer perceptron (MLP) will be utilized. MLP are nonlinear regression or discriminant models, composed by three forms of layers: input, hidden and output.

In several researches notably the one of Sarle (1994), Kangaroo was considered as a key element to explain training a NN. As this approach has been found as simple and clear, it will be summarized in this section. Based on Sarle (1994), the top of the Everest mountain is a global optimum, a less high mountain such as the K2 (known also as the Mount Godwin-Austen is a good local optimum).

As NN is interpreted in terms of minimizing the errors, the analogy of Sarle (1994) is as the attitude increases, the error decreases. Then, the directions represent the weights (for example the north south is one weight, the east west characterizes another weight. Indeed, with different directions, the situation is more complicated (*multidimensional landscape*) and so

training a network with more than 2 weights makes the training more difficult.

Setting the initial weights randomly is a common approach as if a kangaroo is landed randomly in an unknown place before going to the Himalayas. With natural problems such as fog around the Everest, the vision for the Kangaroo is not clear which makes him guess the way with assuming assumptions (such as quadratic and smooth shape of the mountain top). But, most of the cases the assumptions may be violated (the mountain does not have a smooth summit), so the kangaroo would jump repeatedly (iteration) until reaching the peak of the mountain. To avoid jumping toward wrong destination, the kangaroo would follow an altimeter, and this is representing a stabilized *Newton algorithm*.

A further explanation related to the characterization of training a NN as the Kangaroo behaviour to reach the Everest is provided in Plate (1993) and Sarle (1994).

The common expressions for the NN training are:

$$\Delta\omega_j^{(i+1)} = -\eta g_j^{(i)} + \alpha\Delta\omega_j^{(i)} \quad (3.1.1)$$

$$\omega_j^{(i+1)} = \omega_j^{(i)} + \Delta\omega_j^{(i+1)} \quad (3.1.2)$$

$\omega_j^{(i)}$: j^{th} weight on the i^{th} iteration, $g_j^{(i)}$ the partial derivative of the error measure with respect to $\omega_j^{(i)}$, η presents the fixed learning rate and α the control parameter.

The above training method applied to MLP is called generalized delta rule or standard backprop.

3.1.2. Multilayer perceptron for imputation of missing data

The proposed algorithm to impute missing observations Multilayer Perceptron for Imputation of Missing Data (MLPIMID) has two principal pillars, the first one aims to train the neural network as done in the work of (Wang et al., 2013) and the second one concerns the imputation phase.

To allow the reapplication and the reuse of the methodology over different datasets, MLPIMID will be an assembly of SAS macros. SAS has been used in this section to run MLPIMID as there are not plenty and no updated macros related to NN in SAS (SAS Enterprise Miner is mostly used for NN).

The input variables of the global SAS macro MLPIMID are: The raw dataset *DATAIN* that includes the variables to be imputed, *MISS* variable describes the list of missing variables, *CTMISS* presents the count of missing variables, *INDEX* variable refers the id of the outputs.

The outputs of MLPIMID are: The data out *_NN&INDEX* contains the imputed variables and *FIT_&INDEX* summarizes the fit statistics that compare between the classic imputation

methods (MCMC and EM) and MLPIMID.

MLPIMID' steps are summarized in the below flowchart:

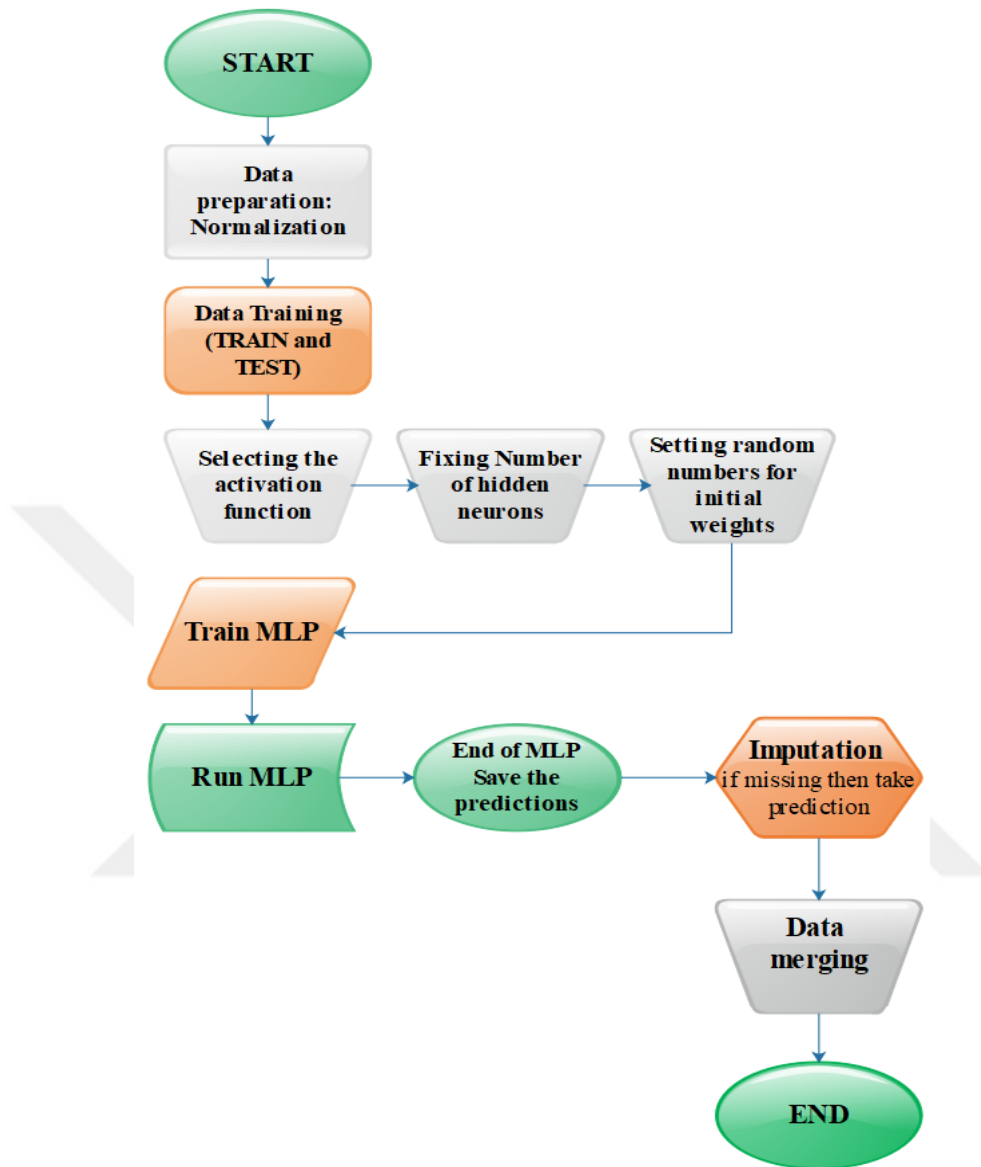


Figure 3.1. Flowchart of MLPIMID algorithm

The following 7 steps provide further details regarding MLPIMID and the corresponding SAS macro is detailed in the appendix 6 as *MLPIMID*.

- (1) **Data preparation:** the raw dataset is quantified and normalized. In fact, the variables to be imputed should be normalized to facilitate the prediction.
- (2) **Data training:** Split the raw data into training dataset (*TRAIN* includes only the observed data) and testing dataset (*TEST*). The training dataset *TRAIN* presents 70% of the raw dataset with ensuring that no missing observation is included. Test dataset *TEST* includes the missing observations. (The percentages of training dataset and test data can be modified).

- (3) **Initializing step:** This step aims to set the initiated values to neural parameters i.e. connection weights, translation factor and scale factor are randomly initialized, and the learning rate is set.
- (4) **Selection of activation function:** Set the activation function. Different activation functions are allowed to be tested in this algorithm with highlighting that each activation function must be tested dependently from the other. MLPIMID includes wavelet functions mainly Gaussian, Mexican hat, Morlet and Meyer along with the classic activation function presented by sigmoid.
- (5) **Training network:** Input the training set, compute network, predict output values, and calculate the error between output and the expected value. This step is based on using the SAS procedure PROC NLP. The Output of this step is the dataset `_TEST&INDEX`.
- (6) **Imputation Phase:** A new variable is created in `_TEST&INDEX`: If the observation is missing, then it is imputed with the attributed prediction value else the raw value is used. The updated `_TEST&INDEX` will be created and will include the imputed variables.
- (7) **Merging:** This step is the trickiest one as dealing with longitudinal dataset is sensitive. It is based on the merge step presented by SAS. Firstly, `_TEST&INDEX` with the imputed values will be merged with the original test dataset TEST (that was used in the PROC NLP procedure and includes missing observations) by the id variables and a flag variable to not lose the order of observations and therefore reform a derived dataset. The final dataset with the original number of observations is recreated by setting the training dataset TRAIN and `_TEST&INDEX`. The final derived dataset with the original number of observations is `_NN&INDEX`.

MLPIMID allows imputation four missing variables, if the data includes more than the fixed number of missing variables, a simple rerun of algorithm is possible. It is not an interpolation by the mean as proposed in several studies that applied NN. To be emphasized also that MLPIMID is not a single imputation method as the imputation of filling the gaps of missing observations is done via predictions that are not a single value.

3.2. Handling Outliers

Four approaches to deal with outliers will be discussed in this section presented by EMIMID, WINWIS, WDODHAS and WDODHWIS as summarized in Figure 3.1.

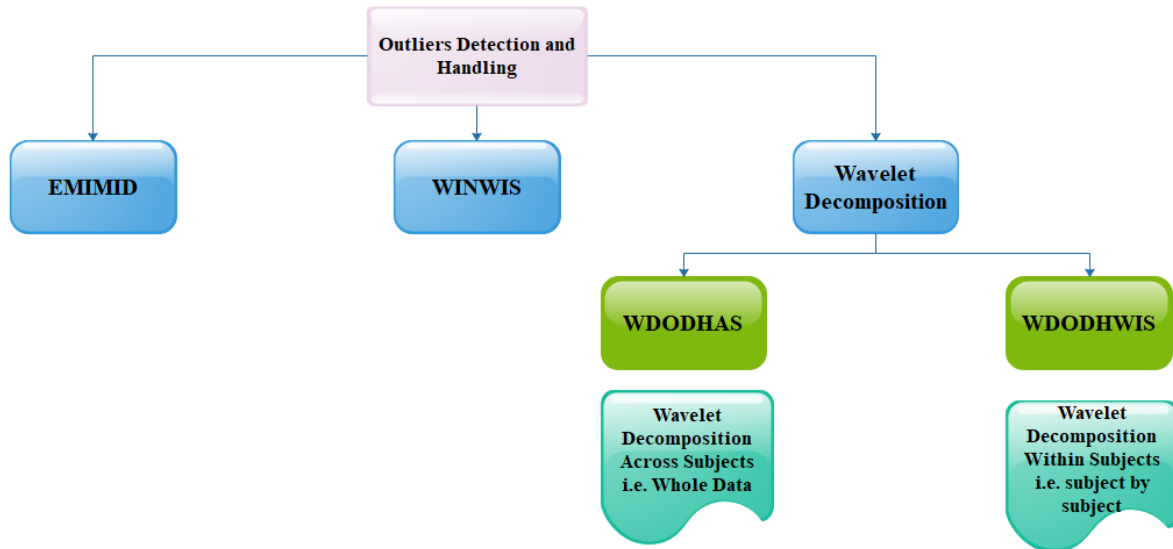


Figure 3.2. Flowchart of suggested methods of handling outliers

3.2.1. Expectation maximum for imputation of missing data

EM algorithm is known as a performant approach to handle the data gaps. The proposed approach Expectation Maximum for Imputation of Missing data (EMIMID) in this study is to consider the outlier, if the field expert confirm that the outliers mislead the analysis, as missing observations. Then, apply the EM algorithm. The corresponding SAS macro is presented in the appendix 7 as *Different methods of Detecting and Handling outliers including EMIMID and WINWIS*.

3.2.2. Winsorisation within subjects

The winsorisation method to handle outliers involves modifying the weights of outliers or replacing the values being tested for outliers with the 10th percentile or the 90th percentile. The proposed approach Winsorisation Within Subjects (WINWIS) is applicable only for longitudinal data, the winsorisation method is applied by subjects. The associated SAS macro is presented in appendix 7 as *Different methods of Detecting and Handling outliers including EMIMID and WINWIS*.

3.2.3. Wavelets decomposition for outliers detection and handling

Due to the particularity of the longitudinal data, two ways of analysis will be provided. Firstly, the wavelets decomposition will be applied across the subjects i.e. the whole data will

be decomposed. For example, if the dataset includes 1000 observations, after the decomposition, 500 observations will present the details and 500 observations will present the coefficients. Later the variable will be reconstructed. Secondly, the wavelets decomposition will be applied within subjects i.e. for each subject, a decomposition will take place. For the second algorithm, it will be highly required that each subject will have at least 2 repeated measures, the case of subject having one measure is quite sensitive as the detail will take zero and if by coincidence this measure is an outlier, wavelets analysis will keep it as outlier even after the variable decomposition and reconstruction. In the next sections, both algorithms will be further explained along with Figure 3.2 that resumes the concept.

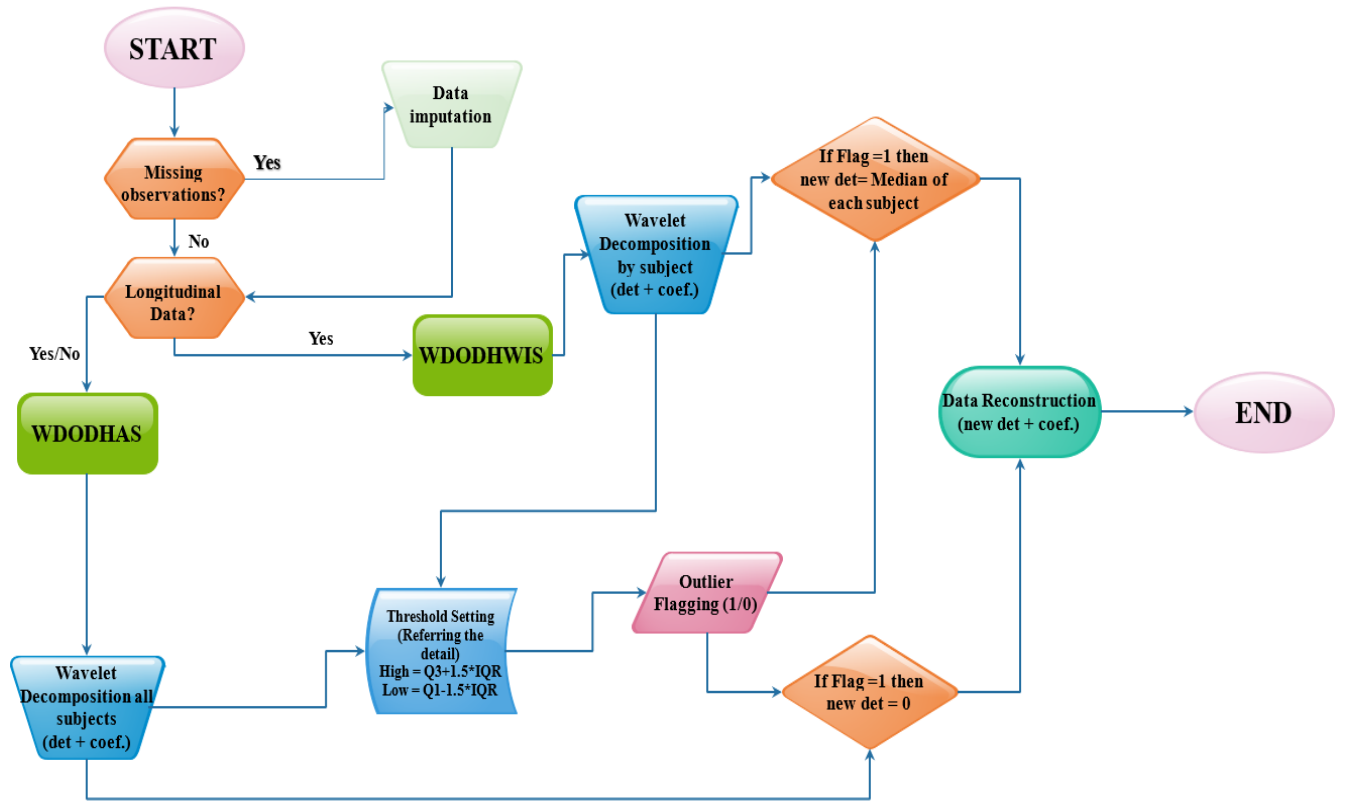


Figure 3.3. Flowchart for wavelet Decomposition for outliers detection and handling

3.2.4. Outliers detection and handling across subjects

The idea of the first algorithm Wavelets Decomposition for Outliers Detection and Handling Across Subjects (WDODHAS) is similar to the work of Grané and Veiga (2009) that handled the outliers in times series with fixing threshold k equal to the 95%-percentile of the distribution of the maximum of the first level detail coefficients (in absolute value) resulting from the DWT of n iid random variables following a standard normal distribution (or a Student's t distribution). Here, the threshold will be changed, and the data will be longitudinal.

Matlab is utilized to construct the algorithm WDODHAS. This algorithm can be reapplied

over longitudinal or non-longitudinal datasets. The inputs of WDODHAS are: The variable that contains outliers and the id variable (or the subject id). The only output is the dataset with the imputed variable and the original variables.

The following steps outline the algorithm WDODHAS and associated Matlab algorithm is described in the appendix 8 as *WDODHAS*.

- (1) **Data Decomposition:** Decompose the variable via wavelet, Daubechies Wavelet Decomposition will be used at the first decomposition level (level 1) (Daubechies ,1992). Details and coefficients of the decomposed variable will have the same observations number ($n/2$), n is the observations number of the raw data.
- (2) **Threshold Setting:** Define the thresholds to detect if there is an outlier in the detail or not. Thresholds will be defined as: High (detail) = $Q3(\text{detail}) + 1.5 * IQR(\text{detail})$ and Low (detail) = $Q1(\text{detail}) - 1.5 * IQR(\text{detail})$. (Those thresholds can be updated or modified).
- (3) **Data Flagging:** A flag variable will be created taking the values in (1, 0): if the detail is greater than High or lower than Low then flag will take 1 i.e. an outlier else it will be 0 (not an outlier).
- (4) **Imputation Phase:** If the flag equals to 1 then the attributed value of detail will be replaced by 0. New detail variable will be created.
- (5) **Data Reconstruction:** Reconstruct the variable by summing (new detail variable) + coefficients. Here, the advantage of the wavelet appears: data decomposition and data reconstruction are possible without losing any information.

3.2.5. Outliers detection and handling within subjects

The second algorithm called Wavelets Decomposition for Outliers Detection and Handling Within Subjects (WDODHWIS) is similar to the first algorithm WDODHAS. The difference consists in decomposing within subjects and not the whole data.

Matlab is utilized to design the algorithm WDODHWIS. This algorithm can be reapplied only over longitudinal datasets. The inputs of WDODHWIS are: The raw dataset, the variable that contains outliers and the id variable (or the subject id). The only output is the dataset with the imputed variable and the original variables.

The following steps sketch the algorithm WDODHWIS and the corresponding Matlab algorithm is stated in the appendix 9 as *WDODHWIS*.

- (1) **Data Decomposition:** Decompose the variable via wavelets, Daubechies will be used at the first decomposition level (level 1), by subject. For example, if the subject has 6

timepoints, then the details will have 3 observations and the coefficients will have 3 observations. If the subject has 5 timepoints, then the details will have 3 observations (the last one is zero) and 3 coefficients etc.

- (2) **Threshold Setting:** Define the thresholds to detect if there is an outlier in the detail or not. Thresholds will be defined as: $\text{High}(\text{detail}) = Q3(\text{detail}) + 1.5 * \text{IQR}(\text{detail})$; $\text{Low}(\text{detail}) = Q1(\text{detail}) - 1.5 * \text{IQR}(\text{detail})$. High and Low are defined for all the subjects. (Those thresholds can be updated or modified).
- (3) **Data Flagging:** A flag variable will be created taking the values in (1,0): if the detail is greater than High or lower than Low then flag will take 1 i.e. an outlier else it will be 0 (not an outlier).
- (4) **Imputation Phase:** If the flag equals to 1 then the attributed value of detail will be replaced by the median of the subject. New detail variable will be created.
- (5) **Data Reconstruction:** Reconstruct the variable by summing the new detail variable and coefficients (new detail variable) + coefficients. Here, the advantage of the wavelet appears: data decomposition and data reconstruction are possible without losing any information.
- (6) **Data Merging:** The algorithm WDODHWIS has a tricky exception that must be highlighted as decomposition is done within subjects. As commonly known, the problem with longitudinal data is attrition i.e. the possibility of having different timepoints by subjects occurs (for example a subject has 6 timepoints, another has 5 timepoints and another has 3 timepoints). To overcome this issue, the maximum of timepoints is considered and wavelets decomposition will consider the missing timepoints with details equal to zero. The final step ensures keeping the same number of observations after the data reconstruction, as getting number of observations greater than the original one is possible. Accordingly, a merge with the original data by subject and by id variable showing the order of timepoints for each subject is highly required to delete the additional observations.

3.3. Wavelets Analysis to Smooth Data

Wavelets analysis have become more used for the nonparametric regression and as a denoising approach. Nason and Silverman (1994) discussed the use of wavelets via application with S to carry out nonlinear regression and image compression. Indeed, they emphasized the use of thresholding of wavelet coefficients as method for attempting to extract signal from noise. Han et al. (2007) created a new threshold algorithm based on wavelet analysis to smooth

noise for a nonlinear time series, by detailing the signals decomposed onto different scales, they smoothed the details by using the updated thresholds to different characters of a noisy nonlinear signal. Their method is an improvement of Donoho's wavelet methods to nonlinear signals. Also, (Horgan, 1999) proposed a wavelet shrinkage for signal smooth by using a generated data. (G.Schimek, 2000; Nason & Silverman, 1994, 1995)

However, the use of the wavelets in the pharmaceutical industry for the nonparametric regression still not highly covered.

The following graph shows the variable Systolic Blood Pressure (SYSBP) before and after the wavelets' analysis, the s presents the signal or the raw variable SYSBP. The variable a_5 presents the smoothing approach or the approximation coefficients, the $a_i, i \in \{1, \dots, 5\}$ are called the details. The wavelet decomposition, as smoothing approach, customize the time and it is not hardly computed even with the big data.

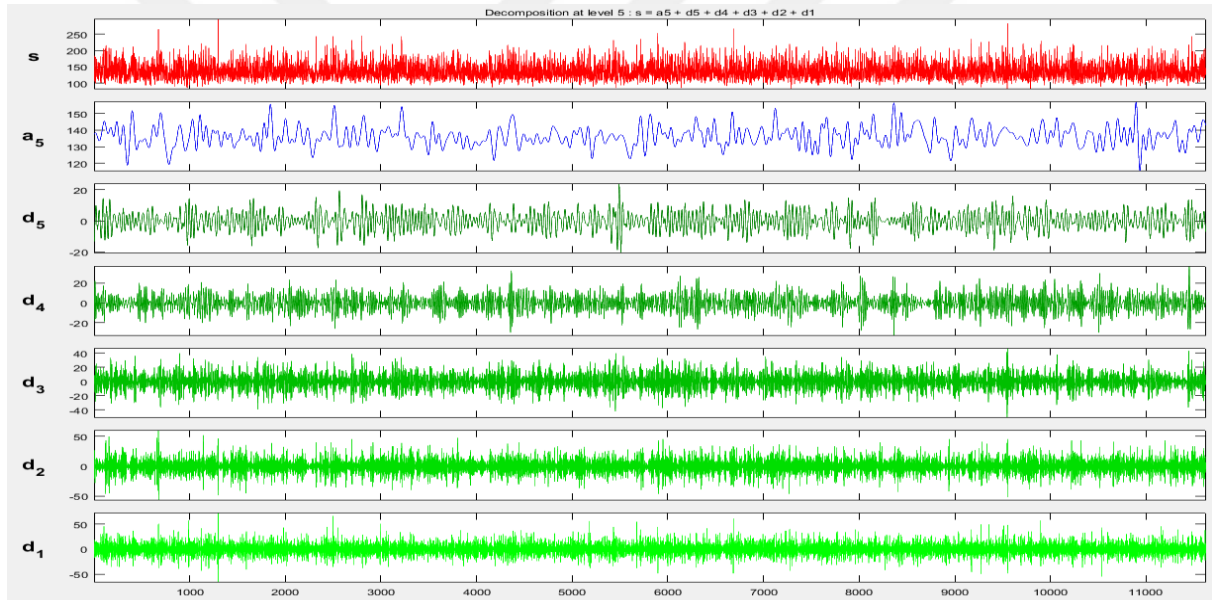


Figure 3.4. *Smoothing splines for hypertension*

Despite the wavelets' decomposition has been mentioned in the literature review as a smoothing approach, there have not been an extensive study researches which used this approach. Therefore, the proposed approach Wavelets Analysis to Smooth Data (WASDA) applies the wavelets' analysis via utilizing the approximation coefficients of the wavelet's decomposition as smoothing coefficients.

4. APPLICATION AND RESULTS

This chapter consists in providing the empirical findings and interpreting the efficiency of the novel algorithms applied on different datasets.

First section defines the longitudinal datasets to be utilised. Second section sketches the results of applying the algorithm MLPIMID and compare the difference of means between MLPIMID and EM, MLPIMID and MCMC. Third section synthesizes the outputs of the proposed approaches WINWIS, EMIMID, WDODHAS and WDOHWHIS to address outliers. The last section compares between the RMSE of smoothing data via wavelet decomposition and other common smoothing approaches such as LOESS, RLOESS and Smoothing spline. Last section demonstrates the results of constructing a semiparametric model for longitudinal data.

4.1. Datasets

Mainly two datasets will be utilized to test the modified approaches and novel algorithms to handle outliers, impute missing observations and finally to construct the semiparametric model. The first dataset is a generated one based on a previous historical clinical data from National Institutes of Health (NIH), the second one is a real data retrieved from the school of Public Health UTHHealth.

4.1.1. Generated dataset

Next to the real dataset, a generated dataset, based on a previous clinical study, will be utilized in the empirical chapter of this dissertation. The concept of generating data by referring previous clinical studies, was adopted from previous research papers (Chen et al., 2008; Psioda, 2012).

Framingham Heart Study provided by National Institutes of Health (NIH) will be considered as the previous study to be used for data generating. It is a long-term cohort study of the etiology of cardiovascular disease among a population of free-living subjects in the community of Framingham, Massachusetts. Framingham Heart Study was a landmark study in epidemiology in that it was the first cohort study of cardiovascular disease and identified the concept of risk factors and their joint effects.

The study began in 1948 and 5,209 subjects were initially enrolled in the study. Participants have been examined biennially since the inception of the study and all subjects are continuously followed through regular surveillance for cardiovascular outcomes. Clinic examination data has included cardiovascular disease risk factors and markers of disease such as blood pressure, blood chemistry, lung function, smoking history, health behaviours, ECG

tracings, Echocardiography, and medication use. Through regular surveillance of area hospitals, participant contact, and death certificates, the Framingham Heart Study reviews and adjudicates events for the occurrence of Angina Pectoris, Myocardial Infarction, Heart Failure, and Cerebrovascular disease.

First step consists in selecting Incident Hypertension as outcome of interest and the predictive variables. In fact, to get the most predictive variables of Incident Hypertension, most of the existing variables in the study were selected and a mixed model was constructed. Consequently, the variables age, Systolic Blood Pressure (SYSBP), Diastolic Blood Pressure (DIABP) and Body Mass Index (BMI) are kept as they have significant relationships with the incident hypertension.

Incident hypertension abbreviated as HYPERTEN, is the first exam treated for high blood pressure or second exam in which either Systolic is ≥ 140 mmHg or Diastolic ≥ 90 mmHg.

Age (in years) age at the examination date, Systolic Blood Pressure (abbreviated as SYSBP), is defined as the mean of last two of three measurements in mmHg, in the range of count: 83.5-295 mmHg, Diastolic Blood Pressure (abbreviated as DIABP), is defined as the mean of last two of three measurements in mmHg with a range of account 30-150 mmHg and finally the Body Mass Index (BMI) is defined as the weight in kilograms divided by the height in meters squared.

Second step aims to generate a longitudinal dataset by using the estimations of HYPERTEN, SYSBP, DIABP, Age and BMI. Thereby, a SAS macro was updated by referring (Chen et al., 2008; Psioda, 2012) to feat the study case of this research, 100 subjects were considered, each subject was generated 10 times (each subject has 10 timepoints).

The following table summarizes the descriptive statistics of the predictive variables.

Table 4.1. Descriptive statistics for the generated data

Variable	Mean (Std)	Min	Max
SYSBP	138.36 (23.383)	69.16	208.69
DIAPB	82.73 (11.203)	51.33	112.80
BMI	25.87 (4.636)	13.55	36.62
Age	54.70 (8.543)	38.00	74.00

The detailed description of the data generation is provided in the appendix 1 *Hypertension Data Generation*.

4.1.2. Real dataset

ACTG data, real dataset in this study, is retrieved from the school of Public Health UTHealth, produced by the Health University AIDS Clinical Trials Group and published by Dr. Hulin Wu (described with more details in Fischl, Ribaud and Collier (2003) and Park and Wu (2006)).

The clinical study is multicentre, randomized, 4-arm trial, double-blind and placebo-controlled for second Protease Inhibitor (PI), taken place between October 1998 and April 2000, for which there was a 24-week primary analysis with extension to 48 weeks. 31 AIDS (Acquired Immunodeficiency Syndrome) Clinical Trials Units in the United States participated in this study and 481 Human Immunodeficiency Virus (HIV)-infected subjects were included. The main findings of the study and further description are provided in Hammer et al. (2002).

The used data has in total 2,626 observations and 7 variables. Along this dissertation, the variable of interest is logrna, defined as the Log10 HIV Ribonucleic Acid (RNA) (it is provided as log10 with the original data). The other variables present the patient id (patid), the Non-Nucleoside Reverse Transcriptase Inhibitor (nnrti), which is a dichotomous variable defined as 1=NNRTI-experienced and 0=NNRTI-naïve. The treatment arms (trtarm), all patients received two NRTIs (abacavir and adefovir), one NNRTI (efavirenz), and at least one PI (amprenavir). In addition, arms A, B, and C received a second PI (saquinavir, indinavir, and nelfinavir, respectively) and arm D received a placebo. The timepoints variables are presented by the calculated scheduled visit week (calwk), the number of days on treatment (txday). The data includes also a censoring indicator (cens) as a dichotomous variable defined 0=actually observed, 1=censored below, 0=censored above (Fischl, Ribaud and Collier, 2003; Park and Wu, 2006). The objective of the study is to assess whether adding a second PI improves antiviral efficacy of a 4-drug combination in patients with virologic failure while taking a PI-containing regimen (Hammer et al., 2002).

The following table 4.2 resumes the number of subjects by treatment arm:

Table 4.2. *Number of patients by treatment arm*

Treatment*	Treatment arm	Number of patid	Dose*
Saquinavir (softgel capsule formulation)	1	116	1600 mg twice daily
indinavir	2	69	1200 mg twice daily
nelfinavir	3	139	1250 mg twice daily
placebo	4	157	
Total		481	

* in combination with Amprenavir 1200 mg twice daily, abacavir 300 mg twice daily, efavirenz 600 mg once daily, and adefovir dipivoxil 60 mg once daily (with L-carnitine supplementation 500 mg once daily).

The following graph shows the boxplots of the variable of interest logrna for the different timepoints.

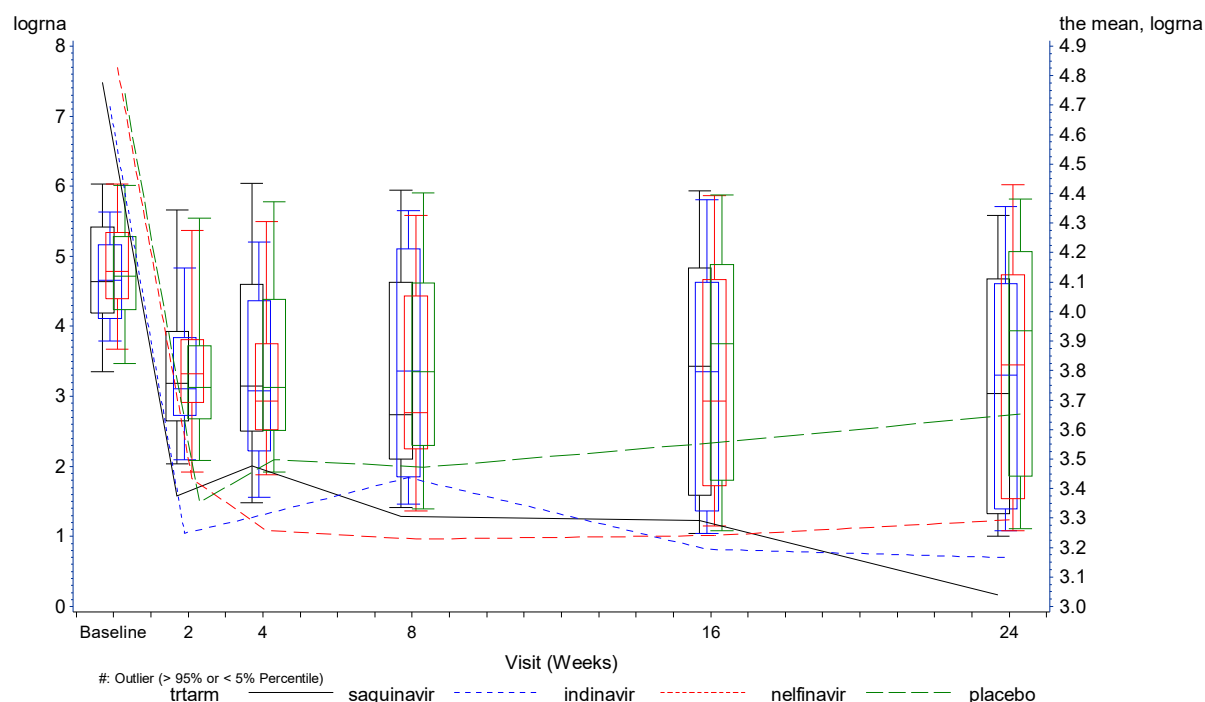


Figure 4.1. Boxplots of logrna by timepoints and by treatment arm

Baseline of logrna has been considered as the latest non missing value of logrna prior to first study drug administration. As treatment start date is not available for this dataset, the timepoint calwk less or equal to 2 is taken as reference (common definition of baseline). The associated program is provided in the appendix 13 *Baseline Calculation*.

The baseline value of logrna is needed to determine logrna change and logrna percentage change. The change from baseline of logrna is defined as $CHG = \text{logrna} - \text{baseline of logrna}$.

The following table resumes the descriptive statistics of logrna, logrna change and logrna percent change for the different timepoints.

Table 4.3. *Descriptive statistics of logrna by timepoints*

Statistic	calwk	logrna	CHG
n	0	479	
Mean	0	4,8	
SD	0	0,76	
Median	0	4,7	
Minimum	0	2,5	
Maximum	0	7,1	
n	2	436	436
Mean	2	3,4	-1,4
SD	2	0,99	0,76
Median	2	3,2	-1,5
Minimum	2	1	-4,2
Maximum	2	6,7	1,4
n	4	442	442
Mean	4	3,4	-1,4
SD	4	1,25	1,03
Median	4	3	-1,6
Minimum	4	1	-3,6
Maximum	4	7	2,1
n	8	441	441
Mean	8	3,4	-1,4
SD	8	1,45	1,25
Median	8	3	-1,7
Minimum	8	0,8	-4,7
Maximum	8	6,7	2,5
n	16	426	426
Mean	16	3,4	-1,4
SD	16	1,68	1,52
Median	16	3,5	-1,1
Minimum	16	0,8	-4,7
Maximum	16	6,8	2,9
n	24	402	402
Mean	24	3,3	-1,4
SD	24	1,7	1,59
Median	24	3,6	-1
Minimum	24	0,8	-4,7
Maximum	24	6,6	3,1

The boxplot and table 4.3 aim to further understand the variables of this study.

4.1.3. Other datasets

Other four clinical datasets, retrieved from University of California, Irvine (UCI) Machine Learning website, will be used to test the efficiency of the novel algorithm MLPIMID to impute missing data.

Table 4.4. *Other datasets description*

Dataset	Number of observations	Number of variables	Missing values?
Breast Cancer Wisconsin	699	10	Yes, one missing variable will be considered in the algorithm.
Heart Disease	294	75	Yes, two missing variables will be considered in the algorithm.
Hepatitis	155	19	Yes, three missing variables will be considered in the algorithm.
Mammographic Mass	961	6	Yes, four missing variables will be considered in the algorithm.

Table 4.4 defines four datasets that include missing variables and to be considered as non-longitudinal datasets to test MLPIMID.

4.2. Handling Missing Data

This section summarizes the results of comparison between the proposed algorithm of imputation of missing observations MLPIMID with EM and MCMC. Also, it provides the Root Mean Squared Error (RMSE) and Root Final Prediction Error (RFPE) of MLPIMID by applying different activation functions. RMSE is an estimate of the prediction error using the unknown optimal weights in the population and RFPE is an estimate of the prediction error using the weights estimated from the sample to make predictions for the population.

Different classic datasets are utilized along with randomly imposing missing values into the variable of interest logrna of the dataset ACTG, six missingness levels will be tested (5%, 10%, 15%, 20%, 25% and 30%). (Appendix 3 *Missing data contamination* provides the program to understand how stricken missing data are created and appendix 10 *Testing MLPIMID algorithm* covers the programming details of this section' tables).

Table 4.5. *RMSE and RFPE of the network by activation function*

	Sigmoid		Gaussian		Meyer		Mexican Hat		Morlet	
	RFPE	RMSE	RFPE	RMSE	RFPE	RMSE	RFPE	RMSE	RFPE	RMSE
Breast Cancer Wisconsin (1 missing variable)	0.3841	0.3626	0.4622	0.4364	0.5943	0.5611	0.7705	0.7274	0.5071	0.4787
Heart Disease (2 missing variables)	0.1591	0.1490	0.2762	0.2587	0.4762	0.4460	0.4835	0.4528	0.2140	0.2004
Hepatitis (3 missing variables)	0.1362	0.1251	0.1357	0.1247	0.2266	0.2083	0.6461	0.5937	0.3287	0.3020
Mammographic Mass (4 missing variables)	0.2646	0.2594	0.4034	0.3955	0.4034	0.3955	0.4181	0.4099	0.2645	0.2601
ACTG (1 variable with 5% missing observations)	0.2148	0.2105	0.3217	0.3153	0.5276	0.5171	0.4904	0.4806	0.2389	0.2341
ACTG (1 variable with 10% missing observations)	0.2152	0.2104	0.3227	0.3156	0.5298	0.5180	0.4926	0.4816	0.2395	0.2342
ACTG (1 variable with 15% missing observations)	0.2087	0.2033	0.3094	0.3013	0.5166	0.5032	0.5075	0.4943	0.2385	0.2323
ACTG (1 variable with 20% missing observations)	0.2132	0.2066	0.3242	0.3140	0.5366	0.5198	0.5012	0.4855	0.2392	0.2317
ACTG (1 variable with 25% missing observations)	0.2204	0.2105	0.3310	0.3162	0.5497	0.5251	0.5192	0.4959	0.2463	0.2352
ACTG (1 variable with 30% missing observations)	0.2552	0.2375	0.3861	0.3595	0.6179	0.5752	0.5361	0.4991	0.2773	0.2582
Hypertension (1 variable with 5% missing observations)	0.1987	0.1838	0.4548	0.4205	0.7241	0.6696	0.3874	0.3582	0.2234	0.2066
Hypertension (1 variable with 10% missing observations)	0.2017	0.1849	0.4613	0.4229	0.7365	0.6751	0.3975	0.3644	0.2258	0.2070
Hypertension (1 variable with 15% missing observations)	0.2073	0.1882	0.4833	0.4388	0.7660	0.6954	0.3970	0.3604	0.2367	0.2149
Hypertension (1 variable with 20% missing observations)	0.2201	0.1958	0.5066	0.4506	0.8044	0.7155	0.4231	0.3763	0.2487	0.2212
Hypertension (1 variable with 25% missing observations)	0.2591	0.2186	0.5859	0.4944	0.9337	0.7879	0.5057	0.4268	0.2894	0.2442
Hypertension (1 variable with 30% missing observations)	0.4028	0.3086	0.9328	0.7147	1.4783	1.1326	0.7679	0.5883	0.4577	0.3507

Referring the RMSE and RFPE of the nets that have been summarized in table 4.5, for most of cases including Morlet as activation function has similar results to including the classic sigmoid function. So, using Morlet activation function can be also advantageous. In addition, Gaussian activation function reflects acceptable results for the datasets Hepatitis and Breast Cancer Wisconsin. Meyer and Mexican Hat wavelet function do not reflect valuable results comparably to Morlet and Sigmoid.

To investigate the performance of the proposed algorithm MLPIMID versus the commonly used imputation methods EM and MCMC, Bland-Altman approach is followed to compare the means of the raw variable and the imputed one (for the different methods) (Bland and Altman; 1986, 1995 and 1996).

Table 4.6. *Difference of means*

Dataset	Difference between	Difference of means* by activation functions ¹			
		Sigmoid	Gaussian	Morlet	Mexican hat
Breast Cancer Wisconsin (1 missing variable)	EM and MLPIMID	3.4307E-17	1.2217E-15	-1.0324E-16	-4.0057E-16
	MCMC and MLPIMID	-1.4612E-17	5.6813E-16	4.0763E-16	8.6665E-16
Heart Disease (2 missing variables)	EM and MLPIMID	-2.8386E-15	-3.1150E-15	-9.8560E-17	2.9152E-16
	MCMC and MLPIMID	-3.8119E-15	-3.3114E-15	-3.6384E-16	5.6814E-16
Hepatitis (3 missing variables)	EM and MLPIMID	-0,0049	-0,0068	-0,0012	-0,0032
	MCMC and MLPIMID	-0,0058	-0,0023	-0,0017	-0,0045
Mammographic Mass (4 missing variables)	EM and MLPIMID	6.3858E-17	-4.2843E-16	6.3858E-17	-3.6977E-16
	MCMC and MLPIMID	4.6040E-16	-1.7433E-16	5.8766E-16	1.1316E-16
ACTG (1 variable with 5% missing observations)	EM and MLPIMID	2.8249E-15	-3.6374E-15	1.5891E-15	-1.5190E-15
	MCMC and MLPIMID	1.0802E-15	-4.6564E-15	-4.4651E-16	-1.7029E-15
ACTG (1 variable with 10% missing observations)	EM and MLPIMID	-6.1386E-15	1.1797E-14	-6.4871E-16	-1.1936E-14
	MCMC and MLPIMID	-3.6476E-15	1.3354E-14	-1.6100E-15	-1.0361E-14
ACTG (1 variable with 15% missing observations)	EM and MLPIMID	-1.9777E-15	1.4850E-14	1.5580E-14	-2.3677E-14
	MCMC and MLPIMID	-1.1313E-14	8.9156E-15	6.1137E-15	-3.1950E-14
ACTG (1 variable with 20% missing observations)	EM and MLPIMID	-2.0765E-14	-3.9626E-14	-1.1400E-13	-6.1194E-14
	MCMC and MLPIMID	-9.5697E-15	-3.1674E-14	-9.7884E-14	-4.7754E-14
ACTG (1 variable with 25% missing observations)	EM and MLPIMID	-3.9235E-14	-8.9990E-14	-2.0224E-13	-1.2069E-13
	MCMC and MLPIMID	-3.8325E-15	-5.4039E-14	-1.6627E-13	-8.0432E-14
ACTG (1 variable with 30% missing observations)	EM and MLPIMID	-6.0211E-14	-3.5541E-14	-2.7474E-13	-1.6216E-13
	MCMC and MLPIMID	-3.5606E-14	-6.2859E-15	-2.4221E-13	-1.3617E-13
Hypertension (1 variable with 5% missing observations)	EM and MLPIMID	4.7245E-15	1.4574E-16	-3.0834E-17	-2.1397E-17
	MCMC and MLPIMID	5.7345E-16	1.1905E-16	-8.9171E-17	7.8724E-18
Hypertension (1 variable with 10% missing observations)	EM and MLPIMID	7.2547E-15	-4.4186E-16	3.7687E-16	1.3464E-16
	MCMC and MLPIMID	6.5879E-16	-4.3924E-16	3.7505E-16	1.1384E-16
Hypertension (1 variable with 15% missing observations)	EM and MLPIMID	7.3698E-15	-1.7755E-15	1.0734E-15	3.2035E-16
	MCMC and MLPIMID	9.4745E-16	-1.7347E-15	9.9536E-16	2.7129E-16
Hypertension (1 variable with 20% missing observations)	EM and MLPIMID	8.7245E-15	1.6713E-16	1.8361E-17	-1.8481E-15
	MCMC and MLPIMID	9.58797E-16	5.9346E-17	-9.411E-17	-1.9594E-15
Hypertension (1 variable with 25% missing observations)	EM and MLPIMID	7.6981E-15	-1.4186E-15	-3.7840E-15	-4.3415E-15
	MCMC and MLPIMID	8.7245E-16	-1.1378E-15	-3.7409E-15	-4.3751E-15
Hypertension (1 variable with 30% missing observations)	EM and MLPIMID	5.7987E-15	1.0247E-14	-1.5313E-14	-1.6593E-14
	MCMC and MLPIMID	6.3257E-16	1.06307E-14	-1.4761E-14	-1.6124E-14

* Only one variable is used to measure the mean difference.

¹ Based the results of Table 4.5, Meyer function does not reflect interesting performance and hence it has not been considered in this table.

The difference between EM, MCMC and the novel algorithm MLPIMID for the different activation functions, is very slight (see Table 4.6). Hence, the novel algorithm demonstrates an interesting ability to impute data for cases of MCAR and this encourage applying MLPIMID in future researches over longitudinal and non-longitudinal data.

4.3. Handling Outliers

To further confirm the utility of the modified approaches EMIMID and WINWIS (based on the classic methods EM and winsorisation) and the novel algorithms WODDHAS and WODHWIS (based on the wavelets decomposition) to deal with outliers, data contamination process is used by imposing three outliers' percentages on the raw variable logrna from ACTG dataset. Thereby, three new variables are created: logrna1 that includes 1% of outliers, logrna3 that includes 3% of outliers and the last one logrna5 that contains 5% of outliers. SAS code of data contamination is provided in the appendix 2 *Outliers contamination*.

Trimming and winsorisation across subjects known as classic methods to deal with outliers are compared with the suggested approaches winsorisation within subjects WINWIS, EMIMID WODDHAS and WODHWIS.

Table 4.5 shows the RMSE (Root Mean Square Error) calculated from the multiple regression analysis (logrna is the dependent variable) and the average of each method.

Table 4.7. Comparison between the outliers handling methods

Outliers Percentage	Method	RMSE	Average
1% outliers (logrna1)	Trimming	1.275	3.619
	Winsorisation across subjects	1.209	3.607
	EMIMID ¹	1.279	3.620
	WINWIS ²	0.408	4.811
	WODHWIS ³	1.387	3.612
	WODHAS ⁴	1.402	3.612
3% outliers (logrna3)	Trimming	1.281	3.611
	Winsorisation across subjects	1.237	3.596
	EMIMID ¹	1.287	3.602
	WINWIS ²	0.416	4.816
	WODHWIS ³	1.530	3.595
	WODHAS ⁴	1.575	3.595
5% outliers (logrna5)	Trimming	1.351	3.611
	Winsorisation across subjects	1.316	3.595
	EMIMID ¹	1.365	3.618
	WINWIS ²	0.438	4.833
	WODHWIS ³	1.894	3.590
	WODHAS ⁴	1.954	3.590

¹ EMIMID: Expectation Maximum for Imputation of Missing Data

² WINWIS: Winsorisation Within Subjects

³ WODHWIS: Wavelets Decomposition for Outliers Detection and Handling Within Subjects

⁴ WODDHAS: Wavelets Decomposition for Outliers Detection and Handling Across Subjects

As shown in the above table 4.5, treating outliers via winsorisation within subjects WINWIS method showed interesting results as it has the least RMSE which encourages applying this modified approach to deal with outliers in longitudinal data.

In addition, the table shows that the RMSE of the two methods WDODHAS and WDODHWIS are almost the same with same averages. Comparing to the other presented handling methods, the RMSE of WDODHAS and WDODHWIS are slightly higher but the advantageous point is that the wavelets decomposition, keep almost the same average whatever the outliers' levels and it highlights some specific cases where the imputation cannot take place. In fact, for 5% missingness, some outliers still present outliers (the cases of the subjects who have only one record or two consecutives outliers).

Hence, it can be concluded that the novel approach WDODHWIS, presented in the second algorithm by segments, improved the accuracy of detecting outliers using the wavelets' decomposition. Furthermore, it emphasizes the efficiency of handling them by substituting the outliers presented in the details by the medians.

The following graphs (Figures 4.2 – 4.10) are presented to further empathize the adequacy of the novel algorithms WDODHAS and WDODHWIS in detecting and handling the outliers,

Figure 4.2 presents the variable logrna1 (that have 1% of outliers) before the decomposition and after the decomposition across the subjects.

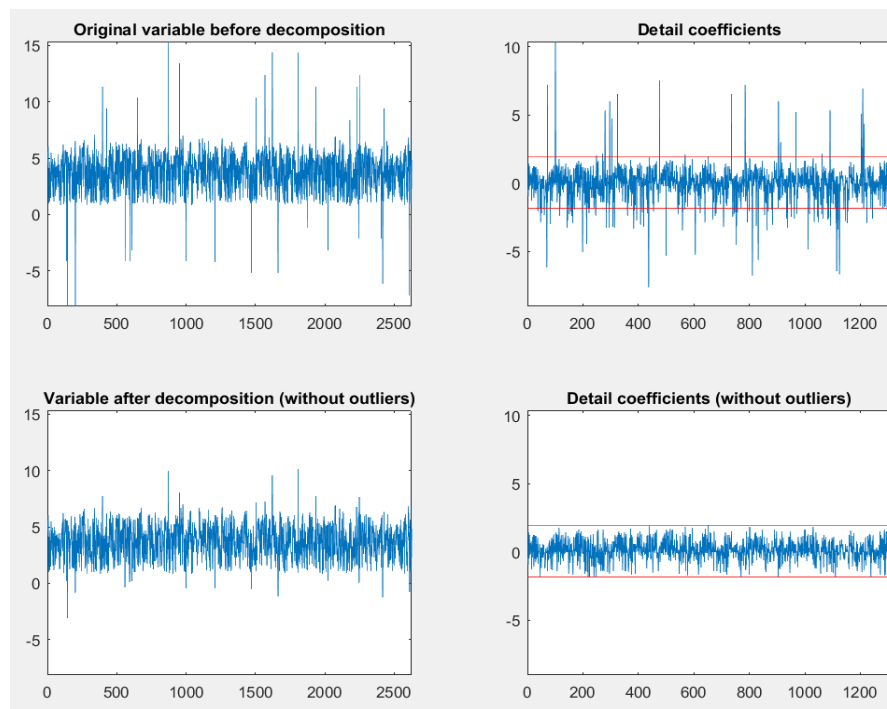


Figure 4.2. Wavelets decomposition of logrna1 across subjects, applying algorithm WDODHAS

As shown in Figure 4.2, the outliers are well handled in the variable logrna1 with keeping the original data features.

Second analysis adopts the second algorithm WDODHWIS which is based on the idea of detecting outliers within subjects and then replace them by the attributed median of each subject. Figure 4.3 shows the logrna1 variable by subjects before the wavelets' decomposition and after the wavelets' decomposition, with and without outliers.

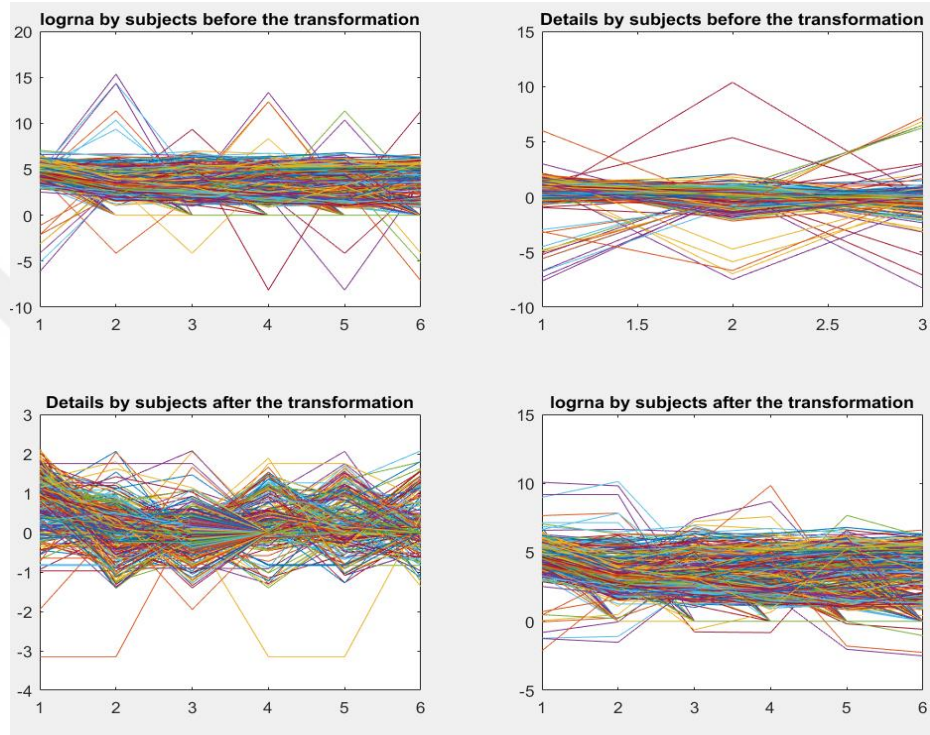


Figure 4.3. Wavelets decomposition of logrna1 by subjects, applying algorithm WDODHWIS

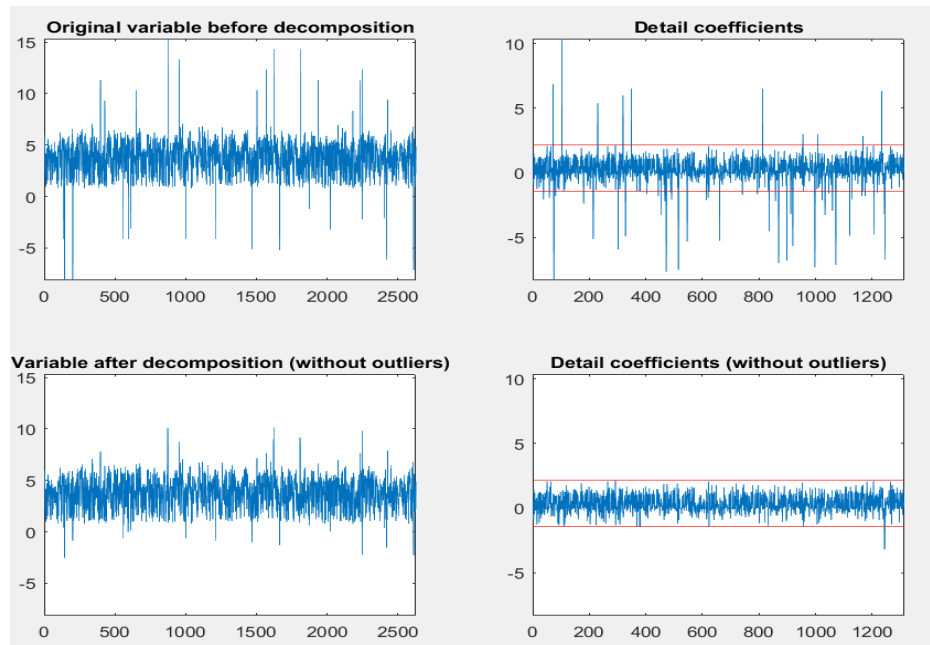


Figure 4.4. Wavelets decomposition of logrna1, applying WDODHWIS

Figure 4.4 presents the logrna1 before decomposition and after substituting the details by the medians and reconstructing the variable. Figure 4.4 demonstrates that the outliers disappeared which proves the efficiency of the wavelets decomposition by segments to handle the outliers without losing the raw information of the variable.

The results presented in Figure 4.2, across subjects, look like the ones presented in Figure 4.4., within subjects. For 1% outliers, the results are quite interesting by applying both algorithms. However, with increasing the outliers' presence in the data notably for 5%, some observations have not been changed and remained as outliers (Figures 4.6 - 4.7 that correspond to logrna3 and 4.9 – 4.10 that correspond to logrna5).

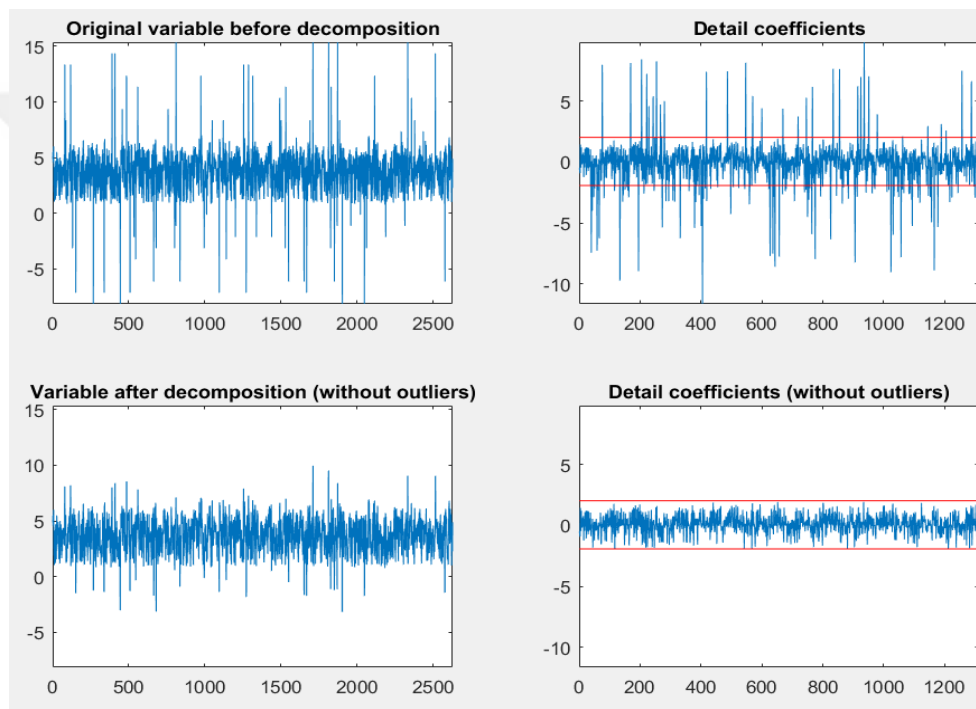


Figure 4.5. Wavelets decomposition of logrna3 across subjects, applying WDODHAS

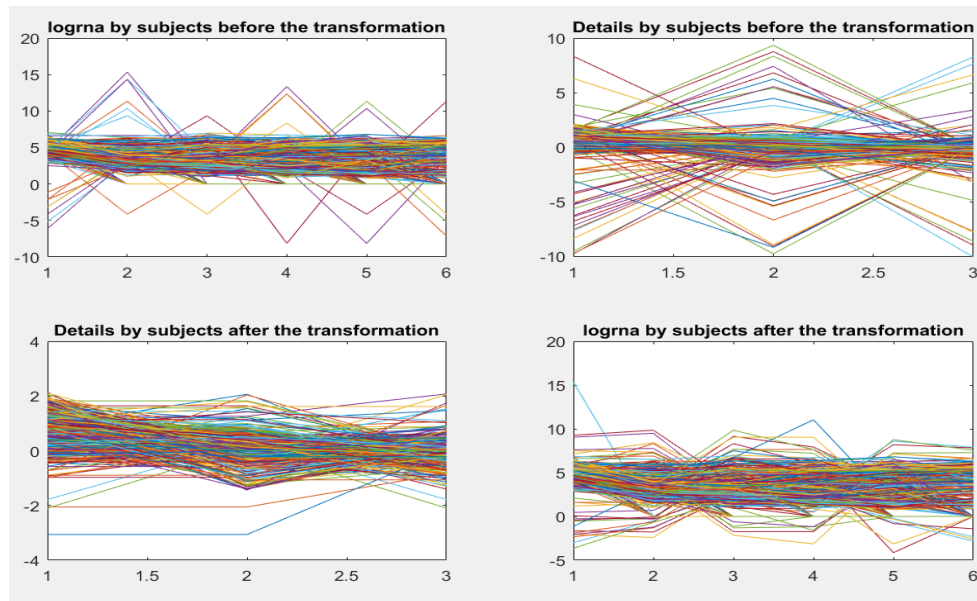


Figure 4.6. Wavelets decomposition of *logrna3* by subjects, applying WDOHWS

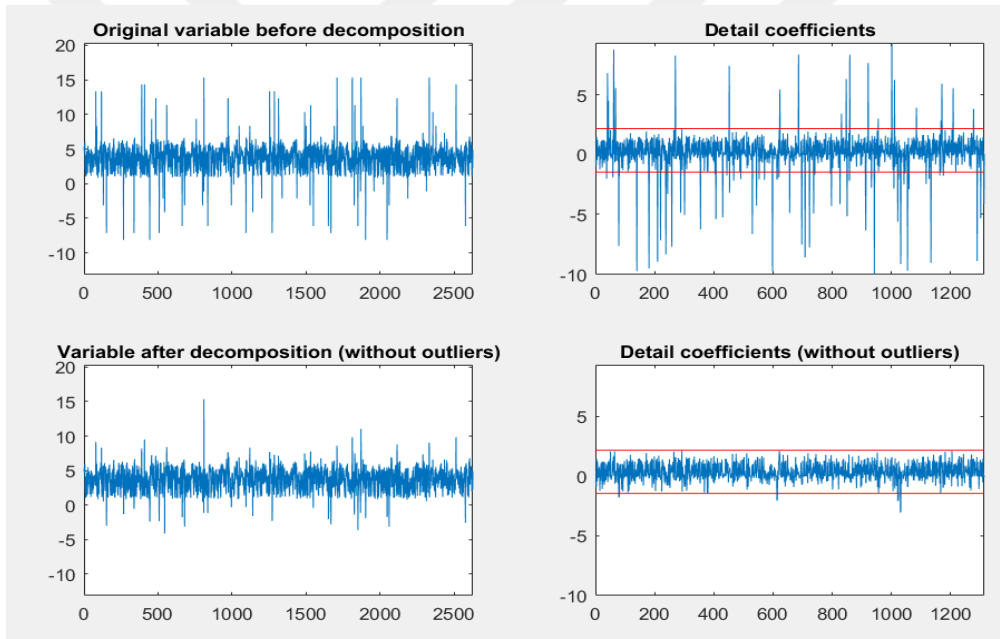


Figure 4.7. Wavelets decomposition of *logrna3*, applying WDOHWS

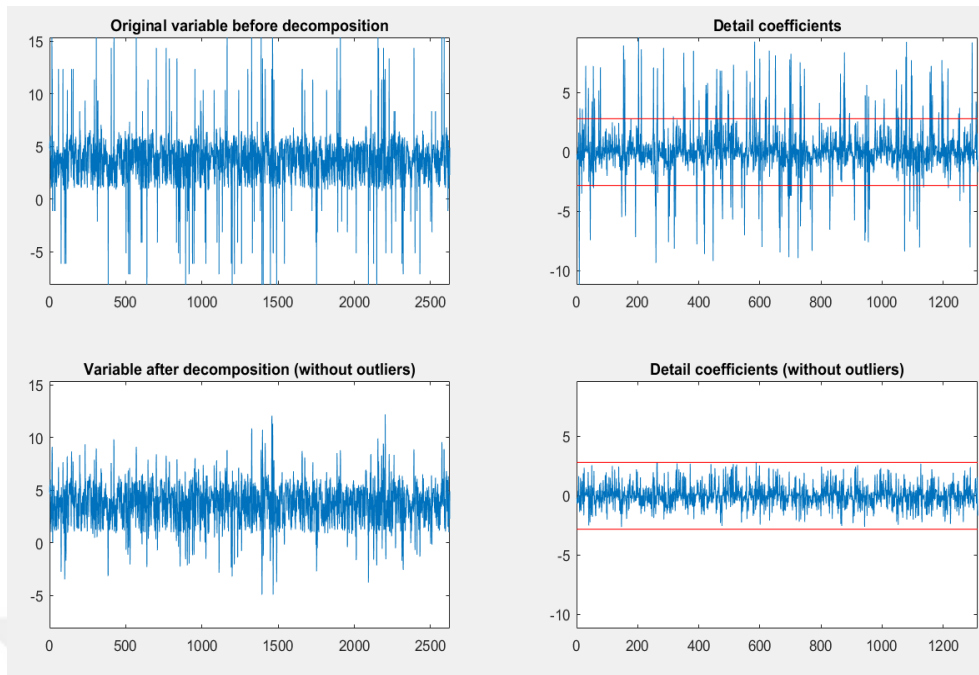


Figure 4.8. Wavelets decomposition of *logrna5* across subjects, applying WDOHDS

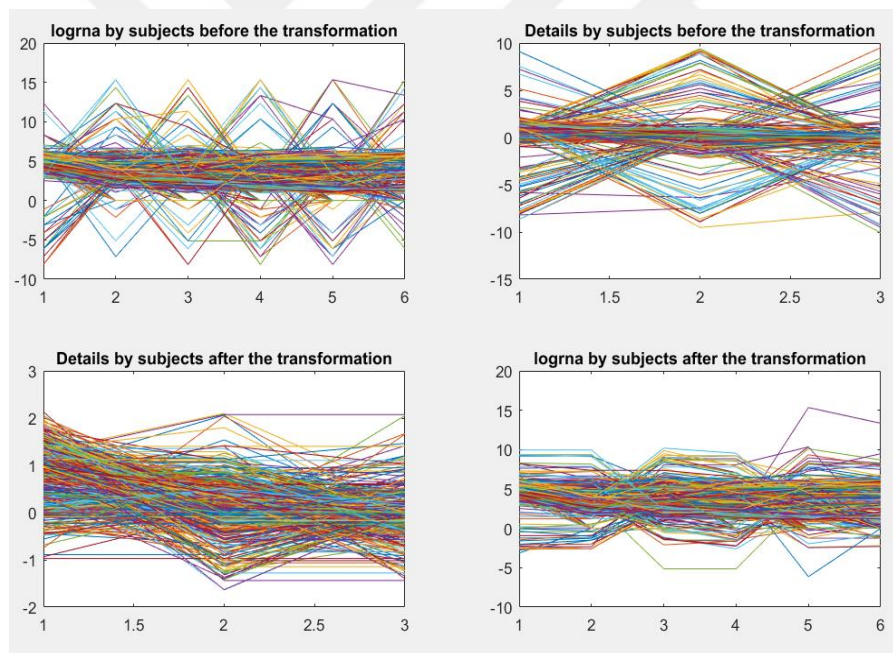


Figure 4.9. Wavelets decomposition of *logrna5* by subjects, applying WDOHWS

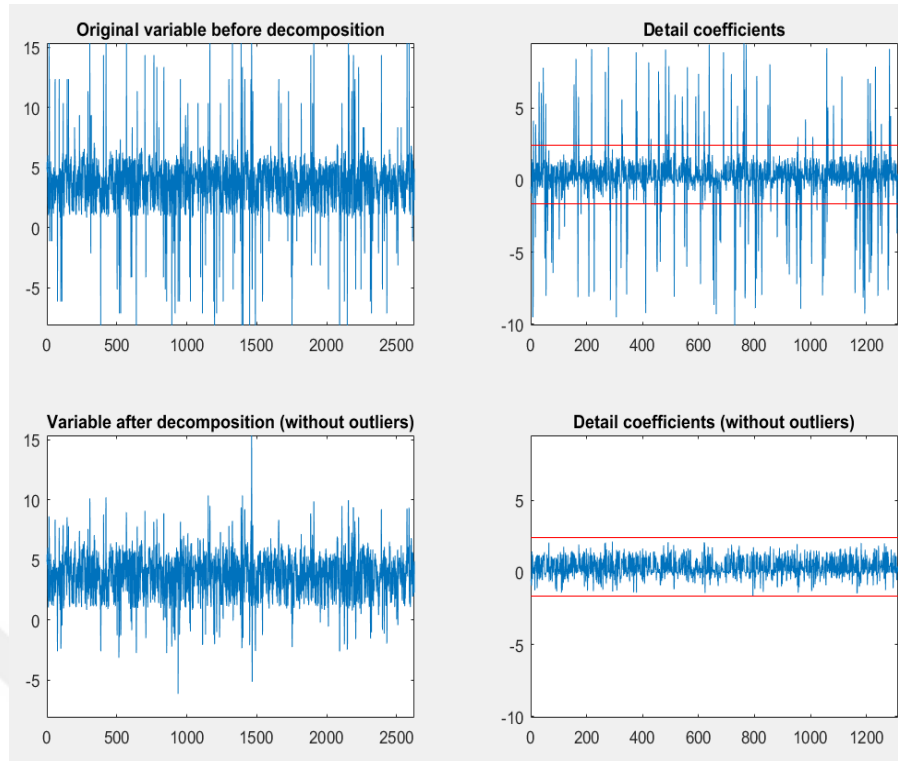


Figure 4.10. Wavelets decomposition of *logrna5*, applying WDOHWHIS

As shown in Figure 4.8, attributed to *logrna5* imputation with WDOHWHAS, through the application of the first algorithm, almost two outliers remained but it is not evident to distinct those subjects. However, through the application of the second algorithm, as presented in Figure 4.9 and Figure 4.10, attributed to *logrna5* imputation with WDOHWHIS, two outliers remain, and as the decomposition was done within subject and the variable was reconstructed by subject, the subject id is easy to be identified. Indeed, this subject had by coincidence two successive outliers. Another case remained as outlier, shows a subject who has only one record which is an outlier.

See appendix 11 *Testing EMIMID and WINWIS* and appendix 12 *Testing WDOHWHAS and WDOHWHIS algorithms* for programming details of this section' tables and graphs.

4.4. Smoothing Approaches

Before the construction of the semiparametric model, the scatter plots of the variables were graphed to examine which variable will be considered in the nonparametric part of the model. Two variables will be used in the nonparametric part of the model, Systolic BP and Diastolic BP from the generated data of incident hypertension.

Figure 4.11 and Figure 4.12 show the wavelet decomposition of the variables Systolic Blood Pressure and Diastolic Blood Pressure respectively. The wavelets analysis allows the decomposition of the series into details (to extract the detailed information of the variable) and

approximation (to see the tendency of the variable). Through the wavelet's decomposition, the coefficients of the details and the approximation could be extracted.

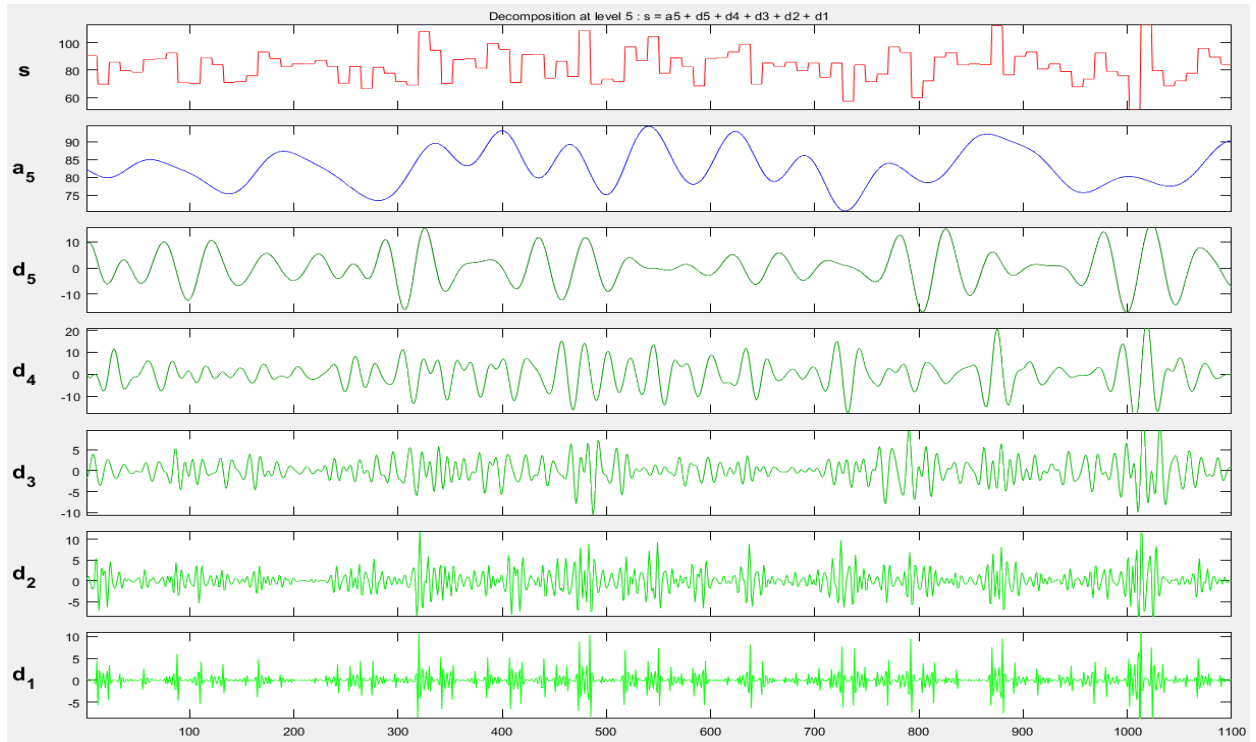


Figure 4.11. Wavelets decomposition of Diastolic Blood Pressure

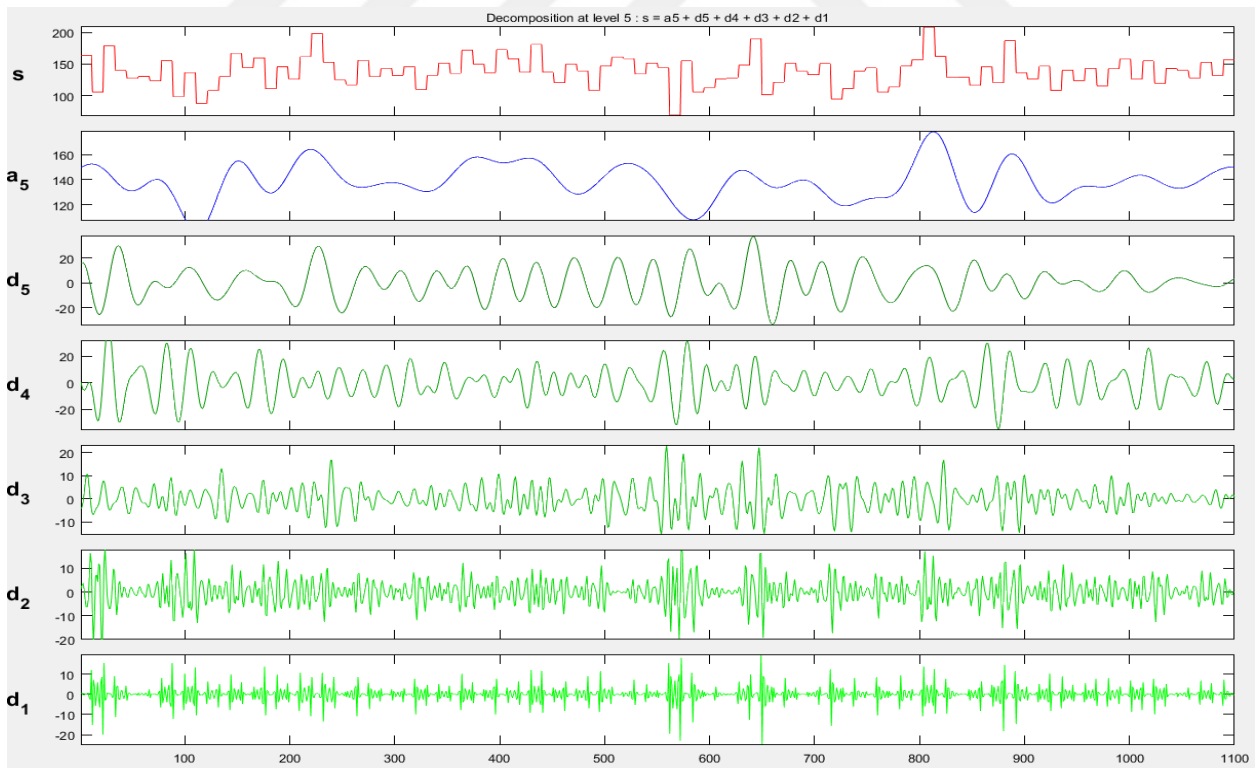


Figure 4.12. Wavelets decomposition of Systolic Blood Pressure

The approximation coefficients after the wavelet's decomposition will be used as smoothed variable.

The following table compare the Root Mean Square Error (RMSE) associated to the smoothing approaches applied for this paper. RMSE was calculated as following: $RMSE = \sqrt{\text{mean}((\text{Raw variable value} - \text{smoothed variable value})^2)}$. Matlab software was used in this section.

Table 4.8. Comparison of RMSE between smoothing approaches

Smoothing approach*	RMSE (DIABP)	RMSE (SYSBP)
Wavelets	1.9583	3.9148
movmedian	2.4942	0
movmean	4.5991	9.4958
loess	2.5075	5.0061
lowess	3.4718	6.9313
rloess	1.8979	4.2089
rloess	3.2293	6.7460
sgolay	3.1262	6.4511
Gaussian	3.2955	6.9424
Smoothing spline	3.049	5.2122

*moving: Moving average (default). A lowpass filter with filter coefficients equal to the reciprocal of the span.
lowess: Local regression using weighted linear least squares and a 1st degree polynomial model
loess: Local regression using weighted linear least squares and a 2nd degree polynomial model
sgolay: Savitzky-Golay filter. A generalized moving average with filter coefficients determined by an unweighted linear least-squares regression and a polynomial model of specified degree (default is 2). The method can accept nonuniform predictor data.
rloess: A robust version of 'lowess' that assigns lower weight to outliers in the regression. The method assigns zero weight to data outside six mean absolute deviations.
rloess: A robust version of 'loess' that assigns lower weight to outliers in the regression. The method assigns zero weight to data outside six mean absolute deviations. (Matlab website)

As shown in Table 4.8 the wavelets decomposition has the least RMSE for both variables Systolic BP and Diastolic BP except the movmedian for the case of SYSBP and rloess for the case of DIABP (no notable difference). Indeed, the wavelets approach showed interesting performance for both variables and consequently may be encourage the researchers to be tested as using the wavelet decomposition in Matlab is not a challenging work.

4.5. Semiparametric Model Construction

This section aims to construct a semiparametric model. The generated dataset hypertension and the real dataset ACTG described in section 4.1.1. and 4.1.2. respectively are used.

For the generated dataset, the incident hypertension is considered as the dependent variable. For the nonparametric part, Systolic Blood Pressure and Diastolic Blood Pressure are

included in the nonparametric part of the model. The variables age and BMI are included in the parametric part as variables with fixed effect.

A generalized mixed model was used to estimate the coefficients, the estimation method is pseudo-likelihood technic. The results of the semiparametric models by applying LOESS and wavelet decomposition as smoothing approaches are presented in the tables 4.9 and 4.10.

Table 4.9. *Semiparametric modeling results of incident hypertension, LOESS as smoothing approach*

Effect	Estimate	SE	DF	t Value	Pr > t
Parametric (Fixed Effects)					
Intercept	-2.2237	0.3043	10	-7.31	<.0001
Age	0.0045	0.0026	1085	1.70	0.0888
BMI	-0.0549	0.0048	1085	-11.30	<.0001
Nonparametric (Fixed Effects, wavelet approximation coefficients as smoothed data)					
Diastolic BP	0.0180	0.0020	1085	8.74	<.0001
Systolic BP	0.0130	0.0009	1085	13.14	<.0001
Fit Statistics					
-2 Res Log Likelihood	2537.07				
AIC (smaller is better)	2539.07				
BIC (smaller is better)	2539.47				
Generalized Chi-Square	619.42				

Table 4.10. *Semiparametric modeling results of incident hypertension, wavelet approximation coefficients as smoothed data*

Effect	Estimate	SE	DF	t Value	Pr > t
Parametric (Fixed Effects)					
Intercept	0.7628	0.4778	10	1.60	0.1415
Age	0.0004	0.0029	1085	0.17	0.8654
BMI	-0.0577	0.0053	1085	-10.80	<.0001
Nonparametric (Fixed Effects, wavelet approximation coefficients as smoothed data)					
Diastolic BP	-0.01056	0.00481	1085	-2.20	0.0283
Systolic BP	0.01068	0.00180	1085	5.93	<.0001
Fit Statistics					
-2 Res Log Likelihood	2717.74				
AIC (smaller is better)	2719.74				
BIC (smaller is better)	2720.14				
Generalized Chi-Square	732.23				

The models' results in tables 4.9 and 4.10 confirm that high Systolic Blood Pressure, high Diastolic Blood Pressure and low BMI cause the incident hypertension. However, the age has not a significant impact on the incident hypertension. There are no major differences between

the semiparametric model that applied LOESS as a smoothing approach and the one that applies wavelet decomposition.

For the real dataset ACTG, the primary objective of the clinical study is to assess whether adding a second PI improves antiviral efficacy of a 4-drug combination in patients with virologic failure while taking a PI-containing regimen. So, the change from baseline of logrna is calculated to examine the efficacy of drugs combination presented by 4 treatment arms (See the appendices 18, 19 and 20 for Spaghetti figures to get a clearer idea about the change from baseline of logrna for all treatments and by treatment). After assuring that ACTG has no missing observations and no outliers, the semiparametric model was constructed.

The dependent variable is the change from baseline of logrna defined as logrna - baseline of logrna. The timepoint calwk has been considered as having nonparametric random effect. The variables trtarm txday nnrti trtarm*nnrti trtarm*calwk baseline patid(trtarm) are included in the parametric part as variables with fixed effect to test the effect of the treatment's combination and the effect of treatment by patient.

A generalized mixed model was used to estimate the coefficients, the estimation method is pseudo-likelihood technic. SAS procedure proc glimmix has been utilized to get the parameters estimation. The results of the semiparametric model are presented in the table 4.11.

Table 4.11. *Semiparametric modeling results of ACTG, LOESS as smoothing approach*

Effect		Estimate	SE	DF	t Value	Pr > t
Parametric (Fixed Effects)						
Intercept		-0.6698	0.2658	2138	-2.52	0.0118
trtarm	1	0.1441	0.2213	2138	0.65	0.5148
trtarm	2	-0.0063	0.2571	2138	-0.02	0.9804
trtarm	3	0.0298	0.2036	2138	0.15	0.8836
trtarm	4	0	.	.	.	.
txday		0.01887	0.0042	2138	4.45	<.0001
nnrti		0.6711	0.1267	2138	5.30	<.0001
baseline		-0.1029	0.0476	2138	-2.16	0.0309
nnrti*trtarm	1	0.00928	0.1974	2138	0.05	0.9625
nnrti*trtarm	2	-0.02294	0.2302	2138	-0.10	0.9206
nnrti*trtarm	3	-0.1992	0.1830	2138	-1.09	0.2764
nnrti*trtarm	4	0	.	.	.	.
calwk*trtarm	1	-0.1713	0.03064	2138	-5.59	<.0001
calwk*trtarm	2	-0.1724	0.03086	2138	-5.59	<.0001
calwk*trtarm	3	-0.1687	0.03055	2138	-5.52	<.0001
calwk*trtarm	4	-0.1502	0.03039	2138	-4.94	<.0001
patid(trtarm)	1	-0.00024	0.00055	2138	-0.43	0.6671
patid(trtarm)	2	0.00017	0.00069	2138	0.25	0.8007
patid(trtarm)	3	-0.00001	0.00045	2138	-0.03	0.9776
patid(trtarm)	4	0.00019	0.00044	2138	0.43	0.6678
Nonparametric (Random Effects, LOESS as smoothing approach)						
Calwk		4.2535	0.4015			
Residual		0.9766	0.0297			
Fit Statistics						
AIC (smaller is better)		8070.87				
BIC (smaller is better)		8074.87				
Generalized Chi-Square		8083.22				
-2 Res Log Likelihood		2546.85				
Patid: Patient id, calwk: calculated scheduled visit week, txday: the number of days on treatment, nnrti: Non-Nucleoside Reverse Transcriptase Inhibitor (1/0)						

Table 4.12 presents the results of applying the wavelets analysis as a smoothing approach.

Table 4.12. Semiparametric modeling results of ACTG, wavelet approximation coefficients as smoothed data

Effect		Estimate	SE	DF	t Value	Pr > t
Parametric (Fixed Effects)						
Intercept		-0.6960	0.2648	2127	-2.63	0.0086
trtarm	1	0.1473	0.2200	2127	0.67	0.5032
trtarm	2	-0.01429	0.2567	2127	-0.06	0.9556
trtarm	3	0.06025	0.2032	2127	0.30	0.7669
trtarm	4	0	.	.	.	.
txday		0.01897	0.00424	2127	4.47	<.0001
nnrti		0.67650	0.12630	2127	5.35	<.0001
baseline		-0.09858	0.04735	2127	-2.08	0.0375
nnrti*trtarm	1	0.01687	0.19620	2127	0.09	0.9315
nnrti*trtarm	2	-0.01928	0.22960	2127	-0.08	0.9331
nnrti*trtarm	3	-0.2185	0.18310	2127	-1.19	0.2329
nnrti*trtarm	4	0	.	.	.	.
calwk*trtarm	1	-0.1720	0.03062	2127	-5.62	<.0001
calwk*trtarm	2	-0.1732	0.03084	2127	-5.62	<.0001
calwk*trtarm	3	-0.1696	0.03052	2127	-5.56	<.0001
calwk*trtarm	4	-0.1510	0.03036	2127	-4.97	<.0001
patid(trtarm)	1	-0.0002	0.00055	2127	-0.43	0.6682
patid(trtarm)	2	0.0001	0.00070	2127	0.21	0.8321
patid(trtarm)	3	-0.00006	0.00045	2127	-0.13	0.8929
patid(trtarm)	4	0.00024	0.00044	2127	0.54	0.5867
Nonparametric (Random Effects, wavelet approximation coefficients as smoothed data)						
calwk		0.005684	0.00054			
Residual		0.9772	0.02981			
Fit Statistics						
-2 Res Log Likelihood		8071.23				
AIC (smaller is better)		8075.23				
BIC (smaller is better)		8083.59				
Generalized Chi-Square		2548.50				
Patid: Patient id, calwk: calculated scheduled visit week, txday: the number of days on treatment, nnrti: Non-Nucleoside Reverse Transcriptase Inhibitor (1/0)						

The model results (table 4.11 and table 4.12) confirm that nnrti and study day (txday) have significant effects on the change of logrna i.e. on reducing the HIV Ribonucleic Acid (RNA) or treatment arm has no significant effect on the reduction of RNA.

Furthermore, there is no significant impact on the modeling fit statistics by using wavelets or LOESS as smoothing approaches except the estimate value of calwk as it is bigger for

LOESS. The normality of residuals and the homoscedasticity of the residuals were rechecked after the models' construction (see appendix 22).

SAS procedure details are provided in the appendix 15 *Semiparametric Model Construction for the real dataset ACTG*.



5. DISCUSSION, CONCLUSION AND RECOMMENDATION

Longitudinal data have been largely discussed and used in several researches notably in the pharmaceutical industry. Modeling this type of data still challenging due to the attrition issue and the violation of some parametric hypothesis. Thereby, one primary objective of this research consists in constructing a semiparametric model for longitudinal data after suggesting some approaches and novel algorithms to treat the common issues. The advantage of the semiparametric model is its capacity to include nonparametric effects and parametric effects, however due to the modeling complexity this type of model still not highly covered notably for longitudinal data. To attain this objective, mainly three pillars have been largely discussed: outliers, missingness and smoothing methods.

First pillar highlights the importance of the missingness mechanisms and not ignoring them to understand why the data is missing. Secondly, it proposes a novel algorithm addressed as Multilayer Perceptron for Imputation of Missing Data (MLPIMID) to impute missing data via a tool which showed a performance in prediction by using non-linear functions and several researches started to adopt it to impute MCAR data. Indeed, the proposed algorithm unifies the advantage of the NN in prediction and the variety of wavelets mother function. Hence, based on previous SAS published macros and the NLP SAS procedure, the multilayer perceptron was utilized to predict the missing observations with training the complete data. Also, the classic sigmoid activation function was compared to wavelets functions (Gaussian, Morlet and Mexican hat). The novel algorithm updates the old neural networks programs to allow changing the activation functions and include the imputation step. Classic known clinical datasets (non-longitudinal datasets) and the longitudinal dataset ACTG (stricken datasets of ACTG were created by randomly imposing different missing missingness levels (5%, 10%, 15%, 20%, 25% and 30%) into the output variable logrna). The results revealed, that the algorithm has an interesting ability to impute the missing data, like the EM and multiple imputation MCMC with very slight deviations. Modifying the activation functions does not impact the results and the RMSE of the nets look alike (except for the Mexican hat function where the RMSE are bit higher than the other functions). More research studies that combine the neural network and the wavelets analysis as WNN approach and the present research are encouraged to better address the missing data.

Second pillar consists in emphasizing that outliers are data components which are not misleading but including a genuine information and highlighting particularities, so deleting them is not always the best option. Secondly, handling outliers can be a challenging work, but

it keeps the originality of the data. Wavelets analysis has started to gain more attention as a powerful mathematical tool in decomposition, smoothing and denoising etc. and notably as a tool showing the discontinuities and the breakdown points, which encourage many researchers to explore it and test it to detect and handle outliers in the financial and economic series.

Therefore, a part of this dissertation applied the wavelets decomposition on longitudinal data to detect the outlier, by introducing two algorithms, the first one Wavelets Decomposition for Outliers Detection and Handling Across Subjects (WDODHAS) applies the wavelets decomposition across subjects (whole data) and the second one Wavelets Decomposition for Outliers Detection and Handling Within Subjects (WDODHWIS) applies the wavelets decomposition for each subject. The concept consists in replacing the outlier, if identified, in the detail and not in the raw observation, by 0 for WDODHAS and by the median (calculated by subject) for WDODHWIS. To prove the efficiency of the wavelets approach, the percentage of outliers presented in the data was posed respectively 1%, 3% and 5% on the variable logrna of ACTG dataset. The results showed that the wavelets algorithms are efficient and wavelets decomposition demonstrated an impressive capacity to figure out the outliers and handling them without losing the originality of the data features. Moreover, with increasing the percentage of outliers in the data, this approach remained powerful but with showing some exceptions: some outliers still there, however the user can find out through them the reason of why those outliers still exist. In fact, those exceptions are by coincidence two successive outliers or a subject with one record which is an outlier, and this cannot be considered as longitudinal observation. Here, it is the role of expert to take the final decision vis a vis those particularities to delete them or to analyze them separately.

Along with WDODHAS and WDODHWIS, EM algorithm was used to handle outliers, by considering the observation, if identified as outlier, as missing and the modified method is addressed as EMIMID. Also, the classic winsorisation method was updated and applied by subjects and not on the whole data, the corresponding algorithm was named WINWIS. Also, WINWIS demonstrated very interesting results with the least RMSE.

Third pillar examines the smoothing methods that are required in the nonparametric part of the model. In fact, different smoothing approaches have been found as performant such as splines, Pseudo spline and kernel. In the literature, one method was noted as smoothing approach but not highly examined, which is the wavelets. This work attempts to figure out the efficiency of the wavelet decomposition as smoothing method. Indeed, Wavelets Analysis to Smooth Data (WASDA) approach suggests using the coefficients outputs of wavelet decomposition as smoothed data. To examine the efficiency of WASDA, a generated

longitudinal dataset, that concerns the incident hypertension, was used. The generated dataset include the variables age, BMI, Diastolic Blood Pressure (SYSBP) and Systolic Blood Pressure (DIABP). SYSBP and DIABP were found to be in the nonparametric part of the model due to the normality assumption violation. Matlab was used to apply a smoothing approach for both variables. Here, different smoothing approaches such as loess, lowess, rloess, rlowess and smoothing spline were called and RMSE was calculated to choose the most performant approach. Although, wavelets were ranked as the second approach having RMSE, but the results were quite interesting for both of variables.

Many research studies constructed semiparametric models to analyze longitudinal data, as this type of model combine the parametric and the nonparametric features. Mostly the timepoint is considered in the random effect part. The modeling results of the real data ACTG showed that the reduction of RNA is mostly related to nnrti, study day and baseline value. SAS procedure proc glimmix was used to model the relationship between the change of logrna and nnrti, baseline, treatment arm, timepoint and study day. This procedure facilitates the estimation method, smoothing method selection, type of effect (random or fixed) selection.

The contribution of this research is the implementation of a novel algorithm to impute missing data via MLP and two algorithms to handle outliers without deleting them via wavelet decomposition. The application of the novel algorithms is not limited to pharmaceutical industry. In addition, wavelet tool is largely discussed as a powerful decomposition tool and as a smoothing approach and this may encourage future researches.

Next to the challenge related to the specificity of longitudinal data and the difficulty to get a real data, the main research challenges can be classified into three points: Firstly, gathering commonly statistic issues for longitudinal data in one research dissertation and trying to present them as friendly subjects in hierarchical way. Secondly, finding a comprehensive methodology to not mislead the reader regarding the diversity of suggested approaches to deal with those issues. Third, introducing the wavelets, that despite its capacity still applied on specific datasets and not longitudinal data, and still unknown, as a main tool to treat some old statistic issues. Finally, the programming part and creating algorithms that will be reused in the future by researchers and statisticians.

Despite the strength of this research is the macro algorithms that can be reused on different datasets, the diversity of using the statistic software is a limitation. In fact, SAS and Matlab have been used, as wavelets still not well developed in SAS.

Further related research can take a place by focusing on the detection of outliers via the clustering approaches in big data that some researchers have already started to test. In addition,

another suggestion is related to the missing data imputation via examining the Wavelets Neural Network (WNN) that still be considered as a complex tool.



REFERENCES

- [1] AL-ANI, T. (2013). Introduction aux ondelettes. Retrieved from <http://www.esiee.fr/~alanit> (Last accessed 11.2018).
- [2] Al-Khazaleh, A.M.H., Al-Wadi, S. and Ababneh, F. (2015). Wavelet transform asymmetric winsorized mean in detecting outlier values. *Far East Journal of Mathematical Sciences*, 96 (3), 339–351. https://doi.org/10.17654/FJMSFeb2015_339_351.
- [3] Altan, N.T and Ustundag, B.B. (2012). Reconstruction of missing meteorological data using Wavelet Transform. *IEEE. First International Conference on Agro- Geoinformatics (Agro- Geoinformatics)*. <https://doi.org/10.1109/Agro-Geoinformatics.2012.6311644>.
- [4] Al-Wadi, S. (2015). Existing outlier values in financial data via wavelet transform. *European Journal of Scientific Research*, 11(22), 245–254.
- [5] Andreopoulos, Y. and Schaar, M.V.D. (2008). Incremental refinement of computation for the discrete wavelet transform. *IEEE Transactions on Signal Processing*, 56(1), 140–157. <https://doi.org/10.1109/ICIP.2007.4379952>.
- [6] Arneodo, A., Bacry, E. and Muzy, J.F. (1995). The thermodynamics of fractals revisited with wavelets. *Physica A: Statistical Mechanics and its Applications*, 213(1), 232-275. [https://doi.org/10.1016/0378-4371\(94\)00163-N](https://doi.org/10.1016/0378-4371(94)00163-N).
- [7] Bai, L., Yan, S., Zheng, X. and Chen, B. M. (2015). Market turning points forecasting using wavelet analysis. *Physica A: Statistical Mechanics and Its Applications*, 437, 184–197. <https://doi.org/10.1016/j.physa.2015.05.027>.
- [8] Barnett, V. and Lewis, T. (1994). Outliers in Statistical Data. Third edition, Chichester, New York, Brisbane, Toronto, Singapore: Wiley Series in Probability and Mathematical Statistics.
- [9] Bartlett and Carpenter (n.d). Missing data. Retrieved from <http://www.missingdata.org.uk/> (Last accessed 17.09.2018).
- [10] Bijleveld, C.C.J.H., van der Kamp, L.J.T. Mooijaart, A., van der Kloot, W.A., Van Der Leeden, R. and Van Der Burg, E. (1998). Longitudinal Data Analysis: Designs, Models and Methods, Thousand Oaks, CA: Sage Publications Ltd.
- [11] Bishop, C. (1996). Neural networks for pattern recognition. First USA, Series: Advanced Texts in Econometrics, Oxford: Oxford University Press.
- [12] Bland, J.M. and Altman, D.G. (1986). Statistical method for assessing agreement between two methods of clinical measurement. *The Lancet*, 1(8476), 307-310. [https://doi.org/10.1016/S0140-6736\(86\)90837-8](https://doi.org/10.1016/S0140-6736(86)90837-8).

- [13] Bland, J.M. and Altman, D.G. (1995). Comparing methods of measurement: why plotting difference against standard method is misleading. *The Lancet*, 346(8982), 1085-1087. [https://doi.org/10.1016/s0140-6736\(95\)91748-9](https://doi.org/10.1016/s0140-6736(95)91748-9).
- [14] Bland, J.M. and Altman, D.G. (1999). Measuring agreement in method comparison studies. *Statistical Methods in Medical Research*, 8(2), 135-160. <https://doi.org/10.1177/096228029900800204>.
- [15] Brumback, B. and Rice, J. (1998). Smoothing spline models for the analysis of nested and crossed samples of curves. *Journal of American Statistical Association*, 93, 961–994. <https://doi.org/10.2307/2669837>.
- [16] Burke Hubbard, B. (1995). Ondes et Ondelettes. La saga d'un outil mathématique. Paris, Pour la Science.
- [17] Burt, P. and E. Adelson (1983). The Laplacian pyramid as a compact image code. *IEEE Transactions on Communications*, 31(4), 532-540.
- [18] Calder. (2012). Retrieved from <https://caldercenter.org/what-are-longitudinal-data>. (Access date 2.03.2018).
- [19] Che, Z., Purushotham, S., Cho, K., Sontag, D. and Liu, Y. (2018). Recurrent Neural Network for Multivariate Time Series with Missing Values. *Scientific Reports*, 8(6085), 1-12. <https://doi.org/10.1038/s41598-018-24271-9>.
- [20] Chen, J., Stock, S. and Deng, C. (2008). Sample size estimation through simulation of a random coefficient model by using SAS abstract. *PharmaSUG*, 3.
- [21] Choi, E. Bahadori, M.T and Sun, J. (2015). Doctor AI: Predicting clinical events via recurrent neural networks. *Cornell University*.
- [22] Cleveland, W. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74(368), 829-836. <https://doi.org/10.2307/2286407>.
- [23] Cooley, J.W. and Tukey, J.W. (1965). An algorithm for the machine calculation of complex Fourier series. *Mathematics of Computation*, 19, 297-301. <https://doi.org/10.1090/S0025-5718-1965-0178586-1>.
- [24] Crochiere, R.E., Weber, S.A. and Flanagan, J. L. (1976). Digital coding of speech in sub-bands. *The Bell System Technical Journal*, 55(8), 1069–1085. <https://doi.org/10.1002/j.1538-7305.1976.tb02929.x>.
- [25] Croiser, A., Esteban, D. and Galand, C. (1976). Perfect channel splitting by use of interpolation/decimation/tree decomposition techniques. *Proceedings of the International Symposium on Information Circuits and Systems*, 443-446.

- [26] Crowe, J., Gibson, N., Woolfson, M. and MG, S. (1992). Wavelet transform as a potential tool for ECG analysis and compression. *Journal of Biomedical Engineering*, 14(3), 268-272. [https://doi.org/10.1016/0141-5425\(92\)90063-Q](https://doi.org/10.1016/0141-5425(92)90063-Q).
- [27] Daniels, M., Hogan, J. W. and Brown, U. (2007). Missing data in longitudinal studies: strategies for bayesian modeling and sensitivity analysis. First edition, Chapman and Hall/CRC Monographs on Statistics and Applied Probability. <https://doi.org/10.1016/10.1201/9781420011180>.
- [28] Day, S. (1999). Dictionary for Clinical Trials, 9. New York: John Wiley and Sons. <https://doi.org/10.1002/0470842563>.
- [29] Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B*, 39(1), 1-38.
- [30] Diggle, P.J. and Kenward, M.G. (1994). Informative dropout in longitudinal data analysis (with Discussion). *Journal of the Royal Statistical Society. Series C (Applied Statistics)*. 43(1), 49–93. <https://doi.org/10.2307/2986113>.
- [31] Diggle, P.J., Heagerty, P., Liang, K.Y. and Zeger, S.L. (2002). Analysis of Longitudinal Data. Second edition, Oxford: Oxford University Press.
- [32] Dong, L., Xiao, D., Liang, Y. and Liu, Y. (2008). Rough set and fuzzy wavelet neural network integrated with least square weighted fusion algorithm-based fault diagnosis research for power transformers. *Electric Power Systems Research*, 78(1), 129–136. <https://doi.org/10.1016/j.epsr.2006.12.013>.
- [33] Daubechies, I (1992). Ten lectures on wavelets. First edition: Society for Industrial and Applied Mathematics.
- [34] Dragset, I.G. (2009). Analysis of longitudinal data with missing values: methods and applications in medical statistics. Norwegian University of Science and Technology.
- [35] Equitz, W.H.R. and Cover, T. M. (1991). Successive refinement of information, *IEEE Transactions on information theory*, 37(2), 269 - 275. <https://doi.org/10.1109/18.75242>.
- [36] Eubank, R. L. (1999). Nonparametric regression and spline smoothing. Second edition, New York: Marcel Dekker Inc.
- [37] Euredit, (2005). Interim report on evaluation criteria for statistical editing and imputation. <http://www.cs.york.ac.uk/euredit>.
- [38] Fan, J. and Gijbels, I. (1996). Local polynomial modeling and its applications. First edition, London: Chapman and Hall/CRC.

- [39] Fan, J., Gijbels, I., Hu, T.C. and Huang, L.S. (1996). A study of variable bandwidth selection for local polynomial regression. *Statistica Sinica*, 6(1), 117–127. Retrieved from <http://www3.stat.sinica.edu.tw/statistica/j6n1/j6n17/j6n17.htm> (Last accessed 05.11.2018).
- [40] Fan, J. and Zhang, J.T. (2000). Two-step estimation of functional linear models with applications to longitudinal data. *Journal of Royal Statistical Society, Series B*, 62(2), 303–322.
- [41] Fischl, M.A., Ribaldo H.J. and Collier, AC. (2003). A randomized trial of 2 different 4-drug antiretroviral regimens versus a 3-drug regimen, in advanced human immunodeficiency virus disease. *Journal of Infectious Diseases*, 188(5), 625-634. <https://doi.org/10.1086/377311>.
- [42] Fitzmaurice, G.M., Laird, N. and Ware, H.J. (2004). Applied longitudinal analysis. First edition, New Jersey: John Wiley and Sons. <https://doi.org/10.1198/jasa.2005.s24>.
- [43] Frees, E.W. (2004). Longitudinal and panel data analysis and applications in the social sciences. USA: Cambridge University Press. <https://doi.org/10.2277/0521828287>.
- [44] Friedman, J.H. (1991). Multivariate adaptive regression splines (with discussion). *Annals of Statistics*, 19(1), 1-67.
- [45] Gabor, D. (1945). Theory of communication. *Institution of Electrical Engineers, Radio and Communication*, 93, 429-457. Retrieved from <http://genesis.eecg.toronto.edu/gabor1946.pdf>.
- [46] Gençay, R., Selçuk, F. and Whitcher, B. (2002). An introduction to wavelets and other filtering methods in finance and economics. First edition, California: Elsevier, Academic Press. <https://doi.org/10.1016/B978-0-12-279670-8.X5000-9>.
- [47] Ghosh, D. and Vogt, A. (2012). Outliers: An Evaluation of Methodologies. *Joint Statistical Meetings*, 3455–3460.
- [48] Grané, A. and Veiga, H. (2009). Wavelet-based detection of outliers in volatility models. *DES - Working Papers, Statistics and Econometrics*, WS 090403, Universidad Carlos III de Madrid. Departamento de Estadística.
- [49] Graps, A. (1995). An Introduction to Wavelets. *IEEE Computational Science and Engineering*, 2(2), 50-61. <https://doi.org/10.2307/2153134>.
- [50] Green, P. J. and Silverman, B.W. (1994). Nonparametric Regression and Generalized Linear Models. First edition, London: Chapman and Hall/CRC.
- [51] Greene, W.H. (2008). Fixed and Random Effects Models for Count Data. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.990012>.

- [52] Grossmann, A. and Morlet, J. (1984). Decomposition of Hardy functions into square integrable wavelets of constant shape. *SIAM Journal of Mathematical Analysis and Applications*, 15(4), 723-736. <https://doi.org/10.1137/0515056>.
- [53] Grubbs, F.E. (1969). Procedures for Detecting Outlying Observations in Samples. *Technometrics*, 11(1), 1–21. <https://doi.org/10.1080/00401706.1969.10490657>.
- [54] Gumedze, F.N. and Chatora, T.D. (2014). Detection of outliers in longitudinal count data via overdispersion. *Computational Statistics and Data Analysis*, 79, 192–202. <https://doi.org/10.1016/j.csda.2014.05.018>.
- [55] Hammer, S.M., Vaida, F. and Bennett, KK (2002). Dual vs single protease inhibitor therapy following antiretroviral treatment failure: a randomized trial. *The Journal of the American Medical Association*, 288(2), 169-180. <https://doi.org/10.1001/jama.288.2.169>.
- [56] Han, M., Liu, Y., Xi, J. and Guo, W. (2007). Noise smoothing for nonlinear time series using wavelet soft threshold. *IEEE Signal Processing Letters*, 14(1), 62–65. <https://doi.org/10.1109/LSP.2006.881518>.
- [57] Hansen, M.H., Madow, W.G. and Tepping, B.J. (1983). An Evaluation of Model-Dependent and Probability-Sampling Inferences in Sample Surveys. *Journal of the American Statistical Association*, 78 (384), 776-793. <https://doi.org/10.1080/01621459.1983.10477018>.
- [58] Hansen, M.H. and Kooperberg, C. (2002). Spline adaptation in extended linear models (with discussion). *Statistical Science*, 17(1), 2–51.
- [59] Hawkins, D.M. (1980). Identification of Outliers. London, New York: Chapman and Hall.
- [60] He, T.X. and Nguyen, T. (2015). Wavelet analysis and applications in economics and finance. *Journal of Statistics and Mathematical Sciences*, 1(1), 22–37.
- [61] Heaton, T.J. and Silverman, B.W. (2008). A wavelet or lifting-scheme-based imputation method. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 70(3), 567-587. <https://doi.org/10.1111/j.1467-9868.2007.00649.x>.
- [62] Hedeker, D. and Gibbons, R. (2006). Longitudinal Data Analysis. First edition, New Jersey: Wiley-Blackwell.
- [63] Hoover, D.R., Rice, J.A., Wu, C.O. and Yang, L.P. (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika*, 85(4), 809–822.
- [64] Horgan, G.W. (1999). Using wavelets for data smoothing: A simulation study. *Journal of Applied Statistics*, 26(8), 923–932. <https://doi.org/10.1080/02664769921936>.

- [65] Horton, N.J. and Kleinman, K.P. (2007). Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. *The American Statistician*, 61(1), 79–90. <https://doi.org/10.1198/000313007X172556>.
- [66] Howell, D.C. (2008). The treatment of missing data. *The SAGE Handbook of Social Science Methodology*, 212–226. <https://doi.org/10.4135/9781848607958.n11>.
- [67] Hulin, W. and Jin-Ting, Z. (2007). Nonparametric regression methods for longitudinal data analysis: Mixed-effects modeling approaches. First edition, New Jersey: Wiley Series in probability and statistics, 229-270.
- [68] Irizarry, R. (n.d.). Kernel Methods. Retrieved from <https://simplystatistics.org/> (Last accessed 26.12.2018)
- [69] Jones, D., Touvannas, J., Lander, P. and Albert, D. (1992). Advanced time-frequency methods for signal averaged ECG analysis. *Journal of Electrocardiology*, 25(Supplement), 188-194. [https://doi.org/10.1016/0022-0736\(92\)90099-L](https://doi.org/10.1016/0022-0736(92)90099-L).
- [70] Kaiser, G. (1994). A Friendly Guide to Wavelets. Boston, Birkhäuser.
- [71] Koikkalainen, P. (2002). Neural networks for editing and imputation. In Data clean 2002 conference. Retrieved from: <http://erin.it.jyu.fi/dataclean/abstracts/node30.html>.
- [72] Kreft, I. and de Leeuw, J. (1998). Introducing statistical methods. Introducing multilevel modeling. Thousand Oaks, CA, US: Sage Publications, Inc. <http://dx.doi.org/10.4135/9781849209366>.
- [73] Kreindler, D.M and Lumsden, C.J. (2012). The effects of the irregular sample and missing data in time series analysis. *Nonlinear Dynamics Psychology and Life Sciences*, 10(2), 187-214.
- [74] Kwak, S. K. and Kim, J. H. (2017). Statistical data preparation: Management of missing values and outliers. *Korean Journal of Anesthesiology*, 70(4), 407–411. <https://doi.org/10.4097/kjae.2017.70.4.407>.
- [75] Lee, T.C.M. (2003). Smoothing parameter selection for smoothing splines: a simulation study. *Computational Statistics and Data Analysis*, 42(1), 139–148. [https://doi.org/10.1016/S0167-9473\(02\)00159-7](https://doi.org/10.1016/S0167-9473(02)00159-7).
- [76] Li, Y. and Reese, S. (2009). Wavelet improvement in turning point detection using a Hidden Markov Model Wavelet improvement in turning point detection using a Hidden Markov Model -from the aspects of cyclical identification and outlier correction. *Computational Statistics, Springer*. 29(6), 1481-1496. <https://doi.org/10.1007/s00180-014-0502-5>.

- [77] Liang, H., Wu, H. and Carroll, R. J. (2003). The relationship between virologic and immunologic responses in AIDS clinical research using mixed-effects varying-coefficient semiparametric models with measurement error. *Biostatistics*, 4(2), 297–312. <https://doi.org/10.1093/biostatistics/4.2.297>.
- [78] Lieberman, E. S. (2005). Nested analysis as a mixed-method strategy for comparative research. *American Political Science Review*, 99(3), 435-452. <https://doi.org/10.1017/S0003055405051762>.
- [79] Lilly, J.M. (2017). Element analysis: a wavelet-based method for analyzing time-localized events in noisy time series. *Proceedings of The Royal Society A Mathematical Physical and Engineering Sciences*, 473(2200). <http://dx.doi.org/10.1098/rspa.2016.0776>.
- [80] Lipton, Z.C, Kale, C.K and Wetzel, R. (2016). Directly modeling missing data in sequences with RNN's: Improved classification of clinical time series. *Proceedings of the 1st Machine Learning for Healthcare Conference*, PMLR 56:253-270.
- [81] Little, R.J.A. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, 83(40), 1198-1202. <https://doi.org/10.2307/2290157>.
- [82] Littell, R.C., Milliken, G. A., Stroup, W.W., Wolfinger, R.D. and Schabenberger, O. (2007). *SAS for Mixed Models*. Second edition, SAS Institute. <https://doi.org/10.1198/tas.2007.s82>.
- [83] Lung, S.Y. (2007). Efficient text independent speaker recognition with wavelet feature selection based multilayered neural network using supervised learning algorithm. *Pattern Recognition*, 40(12), 3616–3620. <https://doi.org/10.1016/j.patcog.2007.05.010>.
- [84] Magnusson, D. and Bergman, L. R. (1990). A pattern approach to the study of pathways from childhood to adulthood. In L.N. Robins & M. Rutter (Eds.), *Straight and devious pathways from childhood to adulthood*. New York, NY, US: Cambridge University Press.
- [85] Mallat, S. (1989). A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7), 674-693.
- [86] Masset, P. (2008). Analysis of financial time-series using Fourier and wavelet methods. *SSRN Electronic Journal*, 1–36. <https://doi.org/10.2139/ssrn.1289420>.
- [87] MathWorks. (n.d.). MathWorks. Retrieved from <https://www.mathworks.com>.
- [88] Maxwell, S.E., Delaney, H. D., and Kelley, K. (2003). *Designing Experiments and Analyzing Data: A Model Comparison Perspective*. Third edition, New York: Routledge.
- [89] Meyer, Y. (1990). *Ondelettes et Opérateurs*, vol. I–III. Paris: Hermann 1990.

- [90] Mondal, D. and Percival, D.B. (2010). Wavelet variance analysis for gappy time series. *Annals of the Institute of Statistical Mathematics*, 62(5), 943–966. <https://doi.org/10.1007/s10463-008-0195-z>.
- [91] Müller, H.G. (1987). Weighted local regression and kernel methods for nonparametric curve fitting. *Journal of the American Statistical Association*, 82(397), 231–238. <https://doi.org/10.2307/2289159>.
- [92] Müller, H.G. (1988). Nonparametric regression analysis of longitudinal data. *Springer-Verlag*. <https://doi.org/10.1007/978-1-4612-5056-2>.
- [93] Nason, G.P., and Silverman, B. W. (1994). The discrete wavelet transform in S. *Journal of Computational and Graphical Statistics*, 3(2), 163–191. <https://doi.org/10.1198/1061860031671>.
- [94] Nason, G.P., and Silverman, B. W. (1995). The stationary wavelet transform and some statistical applications.pdf. *Wavelets and Statistics*, 281–300.
- [95] Newsom, J.T. (2017). Distinguishing between random and fixed: variables, effects, and coefficients. Retrieved from <http://web.pdx.edu/~newsomj/mlrclass/>.
- [96] Ngo, L. and Wand, M. P. (2004). Smoothing with mixed model software. *Journal of Statistical Software*, 9 (1), 1–54. <https://doi.org/10.18637/jss.v009.i01>.
- [97] Nguyen, T.T. (2009). Wavelet estimations for longitudinal data. *Statistica Sinica*, 19, 667–684.
- [98] NIH. (2018). National Institutes of Health. Retrieved from <https://www.nih.gov> (Last accessed 09.05.2019).
- [99] Nordbotten, S. (1998). New methods of editing and imputation. Netherlands: In Institute, W.D.I.S. (Ed.), *Agriculture statistics*.
- [100] Norway, S. (1997). Data editing with artificial neural networks. *Working paper, UN/ECE work session on statistical data editing*, (10). Conference of European statisticians.
- [101] Nyquist, H. (1928). Certain topics in telegraph transmission theory. *Journal of the American Institute of Electrical Engineers*, 47, 617-644. Retrieved from <http://www.loe.ee.upatras.gr/Comes/Notes/Nyquist.pdf>.
- [102] Ogden, R.T. (1997). *Essential Wavelets for Statistical Applications and Data Analysis*. First edition, Boston: Birkhäuser
- [103] OpenTurns. (n.d.). Retrieved from http://openturns.github.io/openturns/master/theory/data_analysis/kernel_smoothing.html (Last accessed 10.2018).

- [104] Panigrahi, L., Das, K, Mishra, D. and Ranjan, R. (2012). Removal and interpolation of missing values using wavelet neural network for heterogeneous data sets. ICACCI '12 Proceedings of the International Conference on Advances in Computing, Communications and Informatics, 1004-1009. <https://doi.org/10.1145/2345396.2345558>.
- [105] Park, J.G. and Wu, H. (2006). Back fitting and local likelihood methods for nonparametric mixed-effects models with longitudinal data. *Journal of Statistical Planning and Inference*, 136(11), 3760-3782. <https://doi.org/10.1016/j.jspi.2005.03.007>.
- [106] Parveen, S. and Green. P. (2001). Speech recognition with missing data using recurrent neural nets. Conference: Advances in Neural Information Processing Systems 14. Neural Information Processing Systems: Natural and Synthetic, NIPS 2001, December 3-8, 2001, Vancouver, British Columbia, Canada.
- [107] Percival, D. and Walden, A. (2000). Wavelet Methods for Time Series Analysis, Cambridge University Press. <https://doi.org/10.1017/CBO9780511841040>.
- [108] Pickles, A. (2005). Missing Data, Problems and Solutions. *Encyclopedia of Social Measurement*. <https://doi.org/10.1016/B0-12-369398-5/00125-0>.
- [109] Pierce, B. (1852). Criterion for the rejection of doubtful observations. *Astronomical Journal*. 2, 161-163.
- [110] Pigoli, D., Sangalli, L. M., Dipartimento, M. O. X. and Milano, P. (2012). Wavelets in Functional Data Analysis: estimation of multidimensional curves and their derivatives. *Computational Statistics and Data Analysis*, 56(6), 1482–1498. <https://doi.org/10.1016/j.csda.2011.12.016>.
- [111] Plate, A.T. (1993). Holographic recurrent networks. In C. L. Giles, S.J. Hanson, and J. D. Cowan, editors, *Advances in Neural Information Processing Systems 5 (NIPS*92)*, 34-41, Morgan Kaufmann, San Mateo, CA, 1993, 8. Retrieved from <http://www.d-reps.org/>.
- [112] Polikar, R. (2006). The wavelet tutorial. Index to series of tutorials to wavelet transform. Second edition, New Jersey: Rowan University.
- [113] Pollet, T.V. and Van Der Meij, L. (2017). To remove or not to remove: The impact of outlier handling on significance testing in testosterone data. *Adaptive Human Behaviour and Physiology*, 3(1), 43–60. <https://doi.org/10.1007/s40750-016-0050-z>.
- [114] Psioda, M. (2012). Random Effects Simulation for Sample Size Calculations Using SAS. First edition, Carolina: University of North Carolina, Chapel Hill NC.
- [115] Quarta, L. (2001). Une introduction (élémentaire) à la théorie des ondelettes. Second edition, Belgium: Mons-Hainaut University.

- [116] Ramsay, J.O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika*, 56(4), 611–630.
- [117] Rice, J.A. and Wu, C. O. (2001). Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrika*, 57(1), 253–259.
- [118] Robins, J.M., Rotnitzky, A., and Zhao, L.P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427), 846-866. <https://doi.org/10.2307/2290910>.
- [119] Rocha, T.R., Paredes, SP. and Henrique, J.H (2010). A wavelet scheme for reconstruction of missing sections in time series signals. *Computing in Cardiology*, 37, 461–464. ISSN 0276–6574.
- [120] Rogosa, D. and Ghandour, G. (1991). Statistical models for behavioral observations. *Journal of Educational and Behavioral Statistics*, 16(3), 157-252. <https://doi.org/10.3102/10769986016003157>.
- [121] Rubin, D.B. (1976). Inference and Missing Data. *Biometrika*. 63(3), 581-592. <https://doi.org/10.2307/2335739>.
- [122] Rubin, D. B. (1987). Multiple Imputation for Nonresponse in Surveys. Wiley John. <https://doi.org/10.1002/9780470316696>.
- [123] Ruppert, D., Wand, M. P. and Carroll, R. J. (2003). Semiparametric Regression. First edition, New York: Cambridge University Press. <https://doi.org/10.1017/CBO9780511755453>.
- [124] Sagar, M. J., Sampat, S. N. and Parmar, S. B. (2016). Electronics and communication engineering outlier removal using multilevel. *Journal of Information, Knowledge and Research in Electronics and Communication Engineering*, 06(02), 671–674.
- [125] Sarle, S.W. (1994). Neural network implementation in SAS software. Proceedings of the Nineteenth Annual SAS Users Group International Conference. SAS Institute Inc., Cary, NC, USA.
- [126] SAS institute. (1999). Nonparametric local polynomial smoother. Last accessed November 2018. Retrieved from <https://v8doc.sas.com/sashtml/insight/chap39/sect45.htm>.
- [127] SAS institute. (n.d.). SAS. Retrieved from <https://support.sas.com/>.
- [128] Schafer, J.L. (1997). Analysis of incomplete multivariate data. First edition: Chapman and Hall/CRC.
- [129] Schafer, J.L. and Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, 7(2), 147–177. <https://doi.org/10.1037//1082-989X.7.2.147>.

- [130] Schimek, M. (2000). Smoothing and Regression Approaches, Computation, and Application. First edition, USA: John Wiley and Son.
- [131] Scott, D.W. (1992). Chapter 19: Multivariate density estimation and visualization. *Handbook of Computational Statistics*, chapter 19, 549-569. https://doi.org/10.1007/978-3-642-21551-3_19.
- [132] Silva-Ramírez, E.L., Pino-Mejías, R., López-Coello, M. and Cubiles-de-la-Vega, M.D. (2011). Missing value imputation on missing completely at random data using multilayer perceptron. *Neural Networks*, 24(1), 121–129. <https://doi.org/10.1016/j.neunet.2010.09.008>.
- [133] Silverman, B.W. (1982). Kernel density estimation using the fast Fourier transform. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 31(1), 93-99.
- [134] Smieja, M., Struski, L., Tabor, J., Zielinski, B. and Spurek, P. (2018). Processing of missing data by neural networks. 32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.
- [135] Sonnberger, H. and Maine, N. (2000). Editing and imputation in Eurostat. Working paper, 21, UN/ECE work session on statistical data editing. Conference of European statisticians.
- [136] Stone, C.J., Hansen, M. H., Kooperberg, C. and Truong, Y.K. (1997). Polynomial splines and their tensor products in extended linear modeling. *Annals of Statistics*, 25, 1371–1425.
- [137] Struzik, Z. (2001). Wavelet methods in financial time-series processing, *Physica A: Statistical Mechanics and its Applications*, 296 (1), 307-319. [https://doi.org/10.1016/S0378-4371\(01\)00101-7](https://doi.org/10.1016/S0378-4371(01)00101-7).
- [138] Szczesniak, R.D., McPhail, G.L., Duan, L.L., Macaluso, M., Raouf, S.A. and Clancy, J.P. (2013). A semiparametric approach to estimate rapid lung function decline in cystic fibrosis. *Annals of Epidemiology*, 23(12), 771–777. <https://doi.org/10.1016/j.annepidem.2013.08.009>.
- [139] Szczesniak, R.D., Li, D. and Raouf, S.A. (2015). Semiparametric mixed models for medical monitoring data: An overview. *Journal of Biometrics and Biostatistics*, 6(2),1-8. <https://doi.org/10.4172/2155-6180.1000234>.
- [140] Tarter, M.E. and Lock, M.D. (1993). Model-free Curve Estimation. First edition, New York: Chapman and Hall/CRC.
- [141] Taylor, J. (n.d.). Statistics 203: Introduction to regression and analysis of variance fixed vs. random effects.
- [142] Torres-Reyna, O. (2007). Panel data analysis fixed and random effects using Stata. Princeton University. Retrieved from <https://dss.princeton.edu/training/>.

- [143] UCI Machine Learning. Retrieved from <https://archive.ics.uci.edu/ml/datasets.php> (Last accessed 22.04.2019).
- [144] Vetterli, M. and Le Gall, D. (1989). Perfect reconstruction FIR filter banks: some properties and factorizations. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(7), 1057 – 1071. <https://doi.org/10.1109/29.32283>.
- [145] Verbyla, A.P., Cullis, B.R., Kenward, M. G. and Welham, S. J. (1999). The analysis of designed experiments and longitudinal data by using smoothing splines. *Journal of the Royal Statistical Society Series C-Applied Statistics*, 48, 269–300. <https://doi.org/10.1111/1467-9876.00154>.
- [146] Wahba, G. (1990). Spline Models for Observational Data. First edition, Philadelphia: CBMS-NSF Regional Conference Series in Applied Mathematics, <https://doi.org/10.1137/1.9781611970128>.
- [147] Wand, M.P. and Jones, M. C. (1995). Kernel Smoothing. First edition, New York: Chapman and Hall/CRC.
- [148] Wang, Y. (1998a). Mixed effects smoothing spline ANOVA. *Journal of Royal Statistical Society, Series B*, 93, 341–348.
- [149] Wang, Y. (1998b). Smoothing spline models with correlated random errors. *Journal of American Statistical Association*, 93(441), 341–348. <https://doi.org/10.2307/2669630>.
- [150] Wang, J.Z. (2001). Wavelets and imaging informatics. *Journal of Biomedical Informatics*, 34(02), 129-141. <https://doi.org/10.1006/jbin.2001.1010>.
- [151] Wang, N., Carroll, R. J. and Lin, X. (2005). Efficient semiparametric marginal estimation for longitudinal /clustered data. *Journal of American Statistical Association*, 100(469), 147–157. <https://doi.org/10.1198/016214504000000629>.
- [152] Wang, G., Guo, L. and Duan, H. (2013). Wavelet neural network using multiple wavelet functions in target threat assessment. *The Scientific World Journal*, 1-7. <http://dx.doi.org/10.1155/2013/632437>.
- [153] White, I.R, Royston, P. and Wood, A.M. (2011). Multiple imputation using chained equations: issues and guidance for practice. *Statistics in medicine*, 30(4), 377–399. <https://doi.org/10.1002/sim.4067>.
- [154] Wood, A.M., White, I.R. and Thompson, S.G. (2004). Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals. *Clinical Trials*, 1(4), 368–376. <https://doi.org/10.1191/1740774504cn032oa>.
- [155] Wu, C.O. and Chiang, C.T. (2000). Kernel smoothing on varying coefficient models with longitudinal dependent variable. *Statistica Sinica*, 10(2), 433-456.

- [156] Wu, H., and Zhang, J.T. (2002a). The study of long-term HIV dynamics using semiparametric nonlinear mixed-effects models. *Statistics in Medicine*, 21(23), 3655–3675. <https://doi.org/10.1002/sim.1317>.
- [157] Wu, H., and Zhang, J.T. (2002b). Local polynomial mixed-effects models for longitudinal data. *Journal of American Statistical Association*, 97(459), 883–897.
- [158] Xie, Y. and Zhang, Y. (2007). A Wavelet Network Model for Short-Term Traffic Volume Forecasting. *Journal of Intelligent Transportation Systems*. 10(3), 141-150. <https://doi.org/10.1080/15472450600798551>.
- [159] Yoon, S.Y. and Lee, S.Y. (1999). Training algorithm with incomplete data for feed-forward neural networks. *Neural Processing Letters*, 10(3), 171–179. <https://doi.org/10.1023/A:1018772122605>.
- [160] Yu, S.N. and Chen, Y.H. (2007). Electrocardiogram beat classification based on wavelet transformation and probabilistic neural network. *Pattern Recognition Letters*, 28(10), 1142–1150. <https://doi.org/10.1016/j.patrec.2007.01.017>.
- [161] Zeger, S.L., Liang, K.Y. and Albert, P.S. (1988). Models for longitudinal data: A generalized estimating equation approach. *Biometrics*, 44(4), 1049-1060. <https://doi.org/10.2307/2531734>.
- [162] Zhang, D. and Lin, X. (2003). Hypothesis testing in semiparametric additive mixed models. *Biostatistics*, 4(1), 57–74. <https://doi.org/10.1093/biostatistics/4.1.57>.
- [163] Zhang, Q. and Benveniste, A. (1992). Wavelet Networks. *IEEE Transactions on Neural Networks*, 3(6), 889-898. <https://doi.org/10.1109/72.165591>.

APPENDICES

Appendix 1 - Hypertension Data Generation

Hypertension Data Generation

```
*****
* PROGRAM NAME :          sim.sas
* DESCRIPTION :          Data Generation based on a historical data
* Assumptions:
* PARAMETERS:
* PROGRAMMER :    Maroua BenGhoul
* DATE WRITTEN : 2018
* Notes      : The idea of data generation based on historical data is found in the works of
*              (Chen et al.. (2008), Psioda, M. (2012))
*****
* INPUT FILES :    Historical data frmgham2
* OUTPUT FILES :   Generated data hypert
*****
* C H A N G E   H I S T O R Y
*****
* DATE      | AUTHOR                | CHANGE
* 2018      | Maroua BenGhoul              | Initial commit
*****
/*Create the libname and import the data*/
%let path = ; /*the path*/
libname hypert "&path.\hypert"; /*Create a libname for the project, the libref must already be assigned before*/

proc sort data = hypert.frmgham2 out = hypert._hypt01 (keep = RANDID SEX TOTCHOL AGE SYSBP DIABP
CURSMOKE
CIGPDAY BMI DIABETES BPMEDS HEARTRTE GLUCOSE
EDUC PREVHYP TIME PERIOD HDLC LDLC DEATH HYPERTEN TIMEHYP); ;

by RANDID;
run;
/*Graphic presentations before the data generation*/
goptions reset=all ftext='calibri' htext=2 gunit=pct;
*symbol1 i=join v='6c'x f='wingdings' w=1 h=2 color=red;
*symbol2 i=join v='6c'x f='wingdings' w=1 h=2 color=blue;
ods graphics on;
proc gplot data= hypert.frmgham2; plot HYPERTEN*SYSBP; run; quit;
ods graphics off;

ods graphics on;
proc capability data= hypert.frmgham2;
var SYSBP; HISTOGRAM SYSBP; PPLOT SYSBP; PROBLOT SYSBP; QQPLOT SYSBP;
run;
```

```
ods graphics on;
proc gplot data= hypert.frmgham2; plot HYPERTEN*DIABP; run; quit;
ods graphics off;
```

```
proc capability data= hypert.frmgham2;
var DIABP; HISTOGRAM DIABP; PPLOT DIABP; PROBLOT
DIABP; QQPLOT DIABP;
run;
```

```
ods graphics on;
proc gplot data= hypert.frmgham2; plot HYPERTEN*age; run; quit;
ods graphics off;
```

```
proc capability data= hypert.frmgham2;
var age; HISTOGRAM age; PPLOT age; PROBLOT age; QQPLOT age;
run;
```

```
ods graphics on;
proc gplot data= hypert.frmgham2; plot HYPERTEN*BMI; run; quit;
ods graphics off;
```

```
proc capability data= hypert.frmgham2;
var BMI; HISTOGRAM BMI; PPLOT BMI; PROBLOT BMI; QQPLOT BMI;
run;
```

```
/*Parameters estimation to be used in the data generation*/
proc glimmix data= hypert._hypt01; /*The default estimation method is
METHOD=RSPL; pseudo-likelihood technic*/
class BPMEDS;
/*The continuous time effect appears in both the MODEL statement (tpoint) and
the RANDOM statement (t2)*/
model HYPERTEN = BPMEDS PERIOD SEX TOTCHOL AGE SYSBP
DIABP CURSMOKE
CIGPDAY BMI DIABETES HDLC LDLC BPMEDS*PERIOD/s
solution;
random PERIOD;
output out=gmxout pred(blup)=pred; /*The OUTPUT statement saves the
predictions of the mean of each observations to the data set gmxout.*/
nloptions tech=newrap;
run;
```

Hypertension Data Generation

```
%let beta0=-1.1; %let beta1=0.0037; %let beta2=0.0081; %let beta3=0.0041; %let beta4=0.007;
%let beta5=-0.05857; %let beta6=0.05; %let beta7=-0.01151; %let betat=-0.2633;
%let beta8=1; %let sigma0=0.7; %let sigma2=0.05; %let sigma=0.001; %let gamma0=0.02;
%let gamma1=0.07; %let delta0=0.001; %let delta1=0.001;
%let n=100;
%let age_av=55; %let age_max=81; %let age_min=32; %let age_sd=9.564;
%let SBP_av=136.32411628; %let SBP_max=295; %let SBP_min=83.5; %let SBP_sd=22.798624823;
%let dbp_av=83.037756945; %let dbp_max=150; %let dbp_min=30;
%let dbp_sd=11.660144287; %let BMI_av=25.877348596;
%let BMI_max=56.8; %let BMI_min=14.43; %let BMI_sd=4.1026399506;

/*Descriptive statistics of the variables to be included in the final model*/
proc univariate data=hypert.frmgham2 noprint;
variable SYSBP; output out=_stat1 mean=mean1 min=min1 max=max1 std=std1;
run;
proc univariate data=hypert.frmgham2 noprint;
variable DIAPB; output out=_stat2 mean=mean2 min=min2 max=max2 std=std2;
run;
proc univariate data=hypert.frmgham2 noprint;
variable age; output out=_stat mean=mean min=min max=max std=std;
run;
proc univariate data=hypert.frmgham2 noprint;
variable BMI; output out=_stat3 mean=mean3 min=min3 max=max3 std=std3;
run;
/*Data generation based on the estimated parameters of the historical data*/
%macro sim(iterations);
%do i=1 %to &iterations;
Data PLACB;
do id=1 to &n;
b0=&sigma0.*rannor(6427);
b2=&sigma2.*rannor(1234);
r0=&delta0.*rannor(4875);
Age=&age_av.+&age_sd.*rannor(4875);
SYSBP=&SBP_av.+&SBP_sd.*rannor(1245);
DIAPB=&dbp_av.+&dbp_sd.*rannor(5484);
BMI=&BMI_av.+&BMI_sd.*rannor(67848);
do CYCLE=0 to 10;
y=(&beta0.+b0)+&beta1.*age+&beta2.*SYSBP+&beta3.*SYSBP+&beta4.*DIAPB
+&beta5.*BMI;
output;
end;
end;
run;

%end;
%mend;
```

```
Options macrogen symbolgen mlogic mprint mfile;
%sim(10000);
```

```
/*Graphic presentations after the data generation*/
ods graphics on;
proc gplot data=PLACB; plot y * SYSBP; run; quit;
ods graphics off;
```

```
ods graphics on;
proc capability data=PLACB;
var SYSBP; HISTOGRAM SYSBP; PPLOT SYSBP; PROBPLOT
SYSBP; QQPLOT SYSBP;
run;
```

```
ods graphics on;
proc gplot data=PLACB; plot y * DIAPB; run; quit;
ods graphics off;
```

```
ods graphics on;
proc capability data=PLACB;
var DIAPB; HISTOGRAM DIAPB; PPLOT DIAPB; PROBPLOT
DIAPB; QQPLOT DIAPB;
run;
ods graphics off;
```

```
ods graphics on;
proc gplot data=PLACB; plot y * BMI; run; quit;
ods graphics off;
```

```
ods graphics on;
proc capability data=PLACB;
var BMI; HISTOGRAM BMI; PPLOT BMI; PROBPLOT BMI; QQPLOT BMI;
run;
ods graphics off;
```

```
ods graphics on;
proc gplot data=PLACB; plot y * age; run; quit;
ods graphics off;
```

```
proc capability data=PLACB;
var age; HISTOGRAM age; PPLOT age; PROBPLOT age; QQPLOT age; run;
```

```
/*Data export*/
data hypert.TRTSIM;
set PLACB;
run;
```

Hypertension Data Generation

```
PROC EXPORT
DATA=hypert.TRTSIM
DBMS=xlsx
OUTFILE="&path.\hypert.xlsx"
REPLACE;
run;
```

Appendix 2 - Outliers Contamination

Outliers contamination

```
*****
* PROGRAM NAME:      outcontam.sas
* DESCRIPTION:       Macro to include random outliers in a data/contamination
* ASSUMPTIONS:       No assumptions
* PARAMETERS:
* PROGRAMMER:        Maroua BenGhoul
* DATE WRITTEN:       2019
* SOFTWARE:          SAS 9.4
*****
* INPUT FILES:       raw data, datain
* OUTPUT FILES:      data with outliers, dataout
*****
* DATE      | AUTHOR          | CHANGE
* 01-2019   | Maroua BenGhoul    | Initial commit
*****/
%let path = ; /*the path*/
/*Create a libname, the libref must already defined before assign it*/
libname hypert "&path.\hypert";

PROC IMPORT OUT= hypert.&datain.
  DATAFILE= "&path.\.." /*Datafile to be defined*/
  DBMS= REPLACE;
  GETNAMES=;
  DATAROW=;
RUN;

%macro outcontam(datain = , dataout=);
/*Calcualte the q1, q3 and IQR*/
proc means data=hypert.&datain. noprint;
  var logrna;
  output out=stat n=n mean=Mean STD=SD median=Median min=Minimum max=Maximum
  qrange=IQR q1=q1 q3=q3;
run;
/* Outliers only in one variable */
data stat (keep = minval maxval Mean SD High Low);
  set stat;
  minval = Mean - 3 * SD; maxval = Mean + 3 * SD;
```

```
/*Checking if outliers exist or not*/
data ACTG_01;
  set hypert.&datain.;
  if _n_ = 1 then set stat;
  if logrna ne . then
  do;
    if logrna > High or logrna < Low then FLG = 1; /*1 means outlier 0 means not outlier*/
    else FLG = 0;
  end;
  else FLG = .;
  _logrna = logrna;

run;
/*the length of a variable*/
%global nobs;
proc sql noprint;
  select count (*) into :nobs from ACTG_01;
quit;
%put &nobs.;
/*SAS macro that duplicates the Excel RANDBETWEEN function*/
%macro RandBetween(min, max);
  (&min + floor((1+&max-&min)*rand("uniform")))
%mend;

data ALL;
  call streaminit(1234);
  set ACTG_01 end=lastline;
  if _n_ = 1 then set stat;

  array x {*} _numeric_ /*1% to be assigned as outliers*/
  array y {*} _numeric_ /*3% to be assigned as outliers*/
  array z {*} _numeric_ /*5% to be assigned as outliers*/

  do i = 5 ;
    if rand("Bern", 0.01) then /* p=0.1 ==> about 1% random outlier*/
      do;
        if x[i] < Mean then x[i] = High + %RandBetween(High, Low);
```

Outliers contamination

```
High = q3 + 1.5*IQR; Low = q1 - 1.5*IQR; run;

if rand("Bern", 0.03) then /* p=0.3 ==> about 3% random outlier */
do;
if y[i] < Mean then y[i] = High + %RandBetween(High, Low);
else if y[i] > Mean then y[i] = Low - %RandBetween(High, Low);

if y[i] ne . then _logrna3 = y[i];
else _logrna3 = _logrna;

end;

if rand("Bern", 0.05) then /* p=0.5 ==> about 5% random outlier */
do;
if z[i] < Mean then z[i] = High + %RandBetween(High, Low);
else if z[i] > Mean then z[i] = Low - %RandBetween(High, Low);

if z[i] ne . then _logrna5 = z[i];
else _logrna5 = _logrna;

end;
end;

if _logrna1 = . then _logrna1 = _logrna;
if _logrna3 = . then _logrna3 = _logrna;
if _logrna5 = . then _logrna5 = _logrna;

run;

else if x[i] > Mean then x[i] = Low - %RandBetween(High, Low);

if x[i] ne . then _logrna1 = x[i];
else _logrna1 = _logrna;

end;
/*Checking the outlier*/
data ALL_01;
set ALL;
if _n_ = 1 then set stat;

if _logrna1 > High or _logrna1 < Low then FLG1 = 1; /* 1 means outlier 0 means not outlier */
else FLG1 = 0;

if _logrna3 > High or _logrna3 < Low then FLG3 = 1; /* 1 means outlier 0 means not outlier */
else FLG3 = 0;

if _logrna5 > High or _logrna5 < Low then FLG5 = 1; /* 1 means outlier 0 means not outlier */
else FLG5 = 0;

run;

data hypert.&dataout.;
set ALL_01;
run;

PROC EXPORT
DATA= hypert.&datain.
DBMS=
OUTFILE="&path.\..." /*outfile to be defined*/
REPLACE;
run;

%mend outcontam;
```

Appendix 3 - Missing Data Contamination

Missing data contamination

```
*****  
* Stricken Missing ACTG dataset  
*****
```

```
data neural.ACTG_Mper; /*Missing data by percentage*/  
call streaminit(1234);  
set neural.ACTG end=lastline;  
_lrna = logma;
```

```
array x {*} _numeric_ /*5% missingness*/  
array y {*} _numeric_ /*10% missingness*/  
array z {*} _numeric_ /*15% missingness*/  
array xx {*} _numeric_ /*20% missingness*/  
array yy {*} _numeric_ /*25% missingness*/  
array zz {*} _numeric_ /*30% missingness*/  
do i = 5;  
    if rand("Bern", 0.05) then x[i]=.; /* p=0.05 ==> about 5% random missing*/  
    if x[i]=. then lma05 = x[i]; else lma05 = _lma;  
end;  
do i = 5;  
    if rand("Bern", 0.1) then y[i]=.; /* p=0.1 ==> about 10% random missing*/  
    if y[i]=. then lma10 = y[i]; else lma10 = _lma;  
end;  
do i = 5;  
    if rand("Bern", 0.15) then z[i]=.; /* p=0.15 ==> about 15% random missing*/  
    if z[i]=. then lma15 = z[i]; else lma15 = _lrna;  
end;  
do i = 5;  
    if rand("Bern", 0.2) then xx[i]=.; /* p=0.2 ==> about 20% random missing*/  
    if xx[i]=. then lma20 = xx[i]; else lma20 = _lma;  
end;  
do i = 5;  
    if rand("Bern", 0.25) then yy[i]=.; /* p=0.25 ==> about 25% random missing*/  
    if yy[i]=. then lma25 = yy[i]; else lma25 = _lma;  
end;  
do i = 5;  
    if rand("Bern", 0.3) then zz[i]=.; /* p=0.3 ==> about 30% random missing*/  
    if zz[i]=. then lma30 = zz[i]; else lma30 = _lrna;  
end;  
run;
```

```
*****  
* Hypertension data set  
*****
```

```
data neural.Hyp_Mper (drop=y _y); /*Missing data by percentage*/  
call streaminit(1234);  
set neural.hypertup end=lastline;  
_outy = y;
```

```
array x {*} _numeric_ /*5% missingness*/  
array beta {*} _numeric_ /*10% missingness*/  
array z {*} _numeric_ /*15% missingness*/  
array xx {*} _numeric_ /*20% missingness*/  
array yy {*} _numeric_ /*25% missingness*/  
array zz {*} _numeric_ /*30% missingness*/  
do i = 5;  
    if rand("Bern", 0.05) then x[i]=.; /* p=0.05 ==> about 5% random missing*/  
    if x[i]=. then outy05 = x[i]; else outy05 = _outy;  
end;  
do i = 5;  
    if rand("Bern", 0.1) then beta[i]=.; /* p=0.1 ==> about 10% random missing*/  
    if beta[i]=. then outy10 = beta[i]; else outy10 = _outy;  
end;  
do i = 5;  
    if rand("Bern", 0.15) then z[i]=.; /* p=0.15 ==> about 15% random missing*/  
    if z[i]=. then outy15 = z[i]; else outy15 = _outy;  
end;  
do i = 5;  
    if rand("Bern", 0.2) then xx[i]=.; /* p=0.2 ==> about 20% random missing*/  
    if xx[i]=. then outy20 = xx[i]; else outy20 = _outy;  
end;  
do i = 5;  
    if rand("Bern", 0.25) then yy[i]=.; /* p=0.25 ==> about 25% random missing*/  
    if yy[i]=. then outy25 = yy[i]; else outy25 = _outy;  
end;  
do i = 5;  
    if rand("Bern", 0.3) then zz[i]=.; /* p=0.3 ==> about 30% random missing*/  
    if zz[i]=. then outy30 = zz[i]; else outy30 = _outy;  
end;  
run;
```

Appendix 4 - MLP Macros (SAS macros programs to be called in MLPIMID)

MLP macros (SAS macros programs to be called in MLPIMID)

```
*****
* PROGRAM NAME :      nlp.sas
* DESCRIPTION :      Neural Network Implementation
* Assumptions:      Activation function could be updated in the macro acthid
* PARAMETERS:      Different macros were included in this program
* PROGRAMMER :
* DATE WRITTEN :
*****
* INPUT FILES :
* OUTPUT FILES :
*****
* C H A N G E H I S T O R Y
* DATE      | AUTHOR      | CHANGE

* 1994      | Warren S.Sarle      | Initial commit
* 04-2019   | Maroua BenGhoul      | Make some parameters as global macro variables,
                        add other activation function
*****/

*****
*      Utility Macros      *;
*****
***** The macros below are not called directly by the user *****;

***** NXY Macro *****;
%macro nxy; %* count variables and store names in macro array;
%global nx ny;

%if %index(&xvar. &yvar.,%str(%)) %then %do;

    %local i token;
    %do i=1 %to 999999;
        %let token=%qscan(&xvar,&i,%str( ));
        %if &token= %then %goto xbreak;
        %global x&i.;
        %let x&i=%unquote(&token);
    %end;
    %xbreak;
    %let nx=%eval(&i-1);

    %do i=1 %to 999999;
        %let token=%qscan(&yvar,&i,%str( ));
        %if &token= %then %goto ybreak;
        %global y&i.;
        %let y&i=%unquote(&token);
    %end;

%end;
```

```
***** CONSTRN Macro *****;
%macro constrn; %* constrain the Euclidean norm of the weights;
%if %bquote(&constrn) ne %then %do;
    array _parm[*] &parms;
    wnorm=0;
    do _i=1 to &nparm;
        wnorm+_parm[_i]**2;
    end;
    wnorm=sqrt(wnorm/&nparm);
    nlincon wnorm < &constrn;
%end;
%mend constrn;
```

```
***** NETLABEL Macro *****;
%macro netlabel; %* labels for the weights in MLPs;
    label
    %let i=0;
    %do ih=1 %to &nh;
        _a&ih="Bias -> Hidden &ih"
    %do ix=1 %to &nxx;
        %let i=%eval(&i+1);
        _b&i="&&x&ix -> Hidden &ih"
    %end;
    %end;
    %let i=0;
    %do iy=1 %to &ny;
        _c&iy="Bias -> &&y&iy"
    %do ih=1 %to &hidden;
        %let i=%eval(&i+1);
        _d&i="Hidden &ih -> &&y&iy"
    %end;
    %end;
    %end;
    ;
%mend netlabel;
```

```
***** NETSTAT Macro *****;
%macro netstat; %* compute fit statistics for the net;
*** compute additive components for graphing;
%let i=0;
%do iy=1 %to &ny;
    %do ih=1 %to &nh;
        %let i=%eval(&i+1);
        _add&i=_h&ih*_d&i;
    %end;
%end;
```

MLP macros (SAS macros programs to be called in MLPIMID)

```

%ybreak;
%let ny=%eval(&i-1);
%end;
%else %do;
data _null_;
    if 0 then set &data;

    array _x[*] &xvar;
    array _y[*] &yvar;

    call symput("nx",trim(left(put(dim(_x),12.))));
    call symput("ny",trim(left(put(dim(_y),12.))));
    length _mname _vname $8;

    do _i=1 to dim(_x);
        _mname='x'||trim(left(put(_i,12.)));
        call execute(%global _||_mname);
        call vname(_x[_i],_vname);
        call symput(_mname,trim(_vname));
    end;

    do _i=1 to dim(_y);
        _mname='y'||trim(left(put(_i,12.)));
        call execute(%global _||_mname);
        call vname(_y[_i],_vname);
        call symput(_mname,trim(_vname));
    end;
stop;

run;
%end;
%mend nxy;

***** NWEIGHT Macro *****;
%macro nweight; %* number of weights (parameters) for MLPs;
    %global na nb nc nd nparm parms;

    %***** size of each parameter array;
    %let na=&nh; /*nh presents the number of hidden neurons*/
    %let nb=%eval(&nx*&nh); /*nx presents the number of input variables*/
    %let nc=&ny; /*ny presents the number of output variables*/
    %let nd=%eval(&nh*&ny);

    %***** number of parameters;
    %let nparm=%eval(&na+&nb+&nc+&nd);

    %if &nh>10000 | &nparm>30000 | &nparm<0 %then %do;
    %put %qcmpres(ERROR: Too many weights (&nparm) or hidden neurons (&nh).);
    %let nparm=0;

```

```

*** compute fit statistics;
_n=n(of _r1-_r&ny);
_nobs+_n;
_sse_
%do iy=1 %to &ny.;
    +_r&iy**2
%end;
;
%mend netstat;

***** PSTAT Macro *****;
%macro pstat(net); %* print fit statistics for the net;
*drop _np_ _nobs_ _df_ _sbc_ _rfpe_ _mse_ _rmse_ _sse_;
%global _np_ _nobs_ _df_ _sbc_ _rfpe_ _mse_ _rmse_ _sse_;

if _end then do;
    _np_=&nparm;
    %if &test. %then %do;
        _df_=_nobs_;
        _sbc_=;
        _rfpe_=;

    %end;
    %else %do;
        _df_=max(0,_nobs_+_np_);
        _sbc_=_nobs_*log(_sse_/_nobs_)+
            _np_*log(_nobs_);
        _rfpe_=sqrt(_sse_*(_nobs_+_np_)/(_nobs_*_df_));

    %end;

    _mse_=_sse_/_df_;
    _rmse_=sqrt(_mse_);

file print;

put "Network. .... &net"
/ "Data .... &data"
/ "Input Variables. .... &xvar"
/ "Output Variables .... &yvar"
/ "Number of Hidden Neurons ... &nh"
/ "Number of Parameters .... " _np_
%if &test %then %do;
/ "Number of Test Data Values .. " _nobs_
%end;
%else %do;
/ "Number of Training Values... " _nobs_
%end;
/ "Degrees of Freedom for Error . " _df_
/ "Sum of Squared Errors. .... " _sse_
/ "Mean Squared Error .... " _mse_

```

MLP macros (SAS macros programs to be called in MLPIMID)

```

%end;

%***** list of all parameters;
%else %let parms=_a1-_a&na _b1-_b&nb _c1-_c&nc _d1-_d&nd;
%mend nweight;
***** NETINIT Macro *****;
%macro netinit; /* random initial values for MLPs;

%if %bquote(&inest.)= %then %do;
%let inest=_INEST_;
data &inest.(type=est);
_type_='PARMS';
retain &parms 0;
array _ab _a: _b;;
do over _ab;
_ab = ranuni(&random.); /*ranuni(&random.)-.5;*/
end;

run;
%end;

%mend netinit;

***** Activation Function Macros *****;

%macro acthid(out,in); /* activation function for hidden alyer;
/*if &in<-45 then &out=0; * avoid overflow;
else if &in>45 then &out=1; * or underflow;
else*/
%let pi=constant("pi");
/*&out.= 1/(1+exp(-&in.)); sigmoid/logistic function;*/
&out = cos(1.75*&in)*(exp(-&in**2)/2);/*Morelet Mother function*/
/*&out. = &in./sqrt(2*&pi)*exp((-&in**2)/2);/*Gaussian wavelet*/
/*&out. = 35*&in**4-84*&in**5+70*&in**6-20*&in**7;/*Meyer wavelet*/
/*&out. = (2/sqrt(3)*(&pi**(-1/4))*(1-&in.**2)*exp((-&in**2)/2));/*Mexican hat wavelet*/

%mend acthid;

%macro actout(out,in); /* activation function for output;
%acthid(&out.,&in.);
%mend actout;

***** NETWORK Macro *****;
%macro network(proc); /* code for 3-layer perceptron using macros;

*** compute hidden layer;
%let i=0;

```

```

/ "Root Mean Squared Error. . . " _rmse_
%if &test=0 %then %do;
/ "Root Final Prediction Error. . " _rfpe_
/ "Schwarz's Bayesian Criterion . " _sbc_
%end;

;

end;

%mend pstat;

*****
* NETNLP Macro *;
*****;

%macro netnlp( /* Train a multilayer perceptron with PROC NPL*/
data=, /* Data set containing inputs and training vlaues */
xvar=, /* List of variables that are inputs */
yvar=, /* List of variables that have training valuse */
hidden=, /* Number of hidden neurons */
inest=, /* Data set containing initial weight estimates. If not specified, random initial estimates are used. */
random=, /* Seed for random numbers for initial weights */
maxiter=1000, /* Maximum number of training iterations alowed */
constrn=, /* optional constraint on norm of weights (requiers 6.10 or later release) */
outest=, /* Data set to contain estimated weights */
nlpopt=SHORT PINIT HISTORY,
TECH= CONGRA);
/* Additional options to be passed to PROC NLP,such as TECH= or ABSCONV=. Do NOT put commas
between the NLP options! */

***** number of variables;
%nxy;
***** number of hidden neurons;
%global nh;
%let nh=&hidden.;
***** number of weights;
%nweight;
%if &nparm=0 %then %goto exit;
***** random initial values;
%netinit;
***** training;
proc nlp start=0 maxfunc=32000 maxiter=&maxiter.
data=&data. inest=&inest. outest=&outest. &nlpopt. TECH=&tech.;
*** run the net;
%network(NLP);
*** define the objective function;
lsq _r1-_r&nny;
*** define the weights;

```

MLP macros (SAS macros programs to be called in MLPIMID)

```
%do ih=1 %to &nh.;
  _sum = _a&ih.
  %do ix=1 %to &nix.;
    %let i=%eval(&i+1);
    +&&x&ix*_b&i
  %end;
;
* apply activation function to hidden node;
%acthid(_h&ih._sum);
%end;

*** compute output;
%let i=0;
%do iy=1 %to &ny.;
  _sum=_c&iy.
  %do ih=1 %to &nh.;
    %let i=%eval(&i+1);
    +_h&ih*_d&i
  %end;
;
* apply activation function to output;
%if &proc=MODEL %then %do;
  %actout(&&y&iy._sum);
%end;

%else %do;
  %actout(_p&iy._sum)
  _r&iy = &&y&iy. - _p&iy.;
%end;

%end;
%mend network;
```

```
parms &parms.;
*** nonlinear constraints;
%constrn;
*** labels for the weights;
%netlabel;

run;
%exit;
/*END OF THE MACRO*/
%mend netnlp;

*****
*      NETRUN Macro      *;
*****
%macro netrun( /* Run a multilayer perceptron with a DATA step*/
  data=, /* Data set containing inputs and training vlaues */
  xvar=, /* List of variables that are inputs */
  yvar=, /* List of variables that have training value */
  hidden=, /* Number of hidden neurons; defaults to value from previous execution of one of the NET... macros */
  test=, /* Specify TEST=1 if these are test data, not training data */
  inest=, /* Data set containing weights */
  out=); /* Data set to contain outputs and residuals */
  ***** number of variables;
  %nxy;
  ***** number of hidden neurons;
  %global nh;
  %if %bquote(&hidden)^= %then %let nh=&hidden;
  ***** number of weights;
  %nweight;
  %if &nparm=0 %then %goto exit;
  ***** run the network;
  data &out./*(keep=&xvar. &yvar. _h1-_h&nh. _p1-_p&ny. _r1-_r&ny. _add1-_add&nh.)*/;
    *** read estimated weights;
    if _n_=1 then set &inest.(where=( _type_='PARMS'));
    *** read input and training values;
    set &data. end=_end;
    *** compute outputs and residuals;
    %network(&data.);
    *** compute SSE and additive components;
    %netstat;
    *** print statistics;
    %pstat(MLP);

run;
%exit;
/*END OF THE MACRO*/
%mend netrun;
```

Appendix 5 - Bland Altman (baplot.sas)

Bland Altman (baplot.sas)

```
/******  
* DESCRIPTION :      Bland Altman type  
* Assumptions:  
* PARAMETERS:  
* CALLED BY :      baplot.sas  
* CALLS TO :  
* PROGRAMMER :  
* DATE WRITTEN : $Date$  
* Source : sas-and-r.blogspot.com/2012/06/example-934-bland-altman-type-plot.html  
*****  
* INPUT FILES :  
* OUTPUT FILES :  
*****  
* CHANGE HISTORY  
*****  
* DATE      | AUTHOR                | CHANGE  
* 05-06-2012 | Ken Kleinman            | Initial commit *****  
* 20-03-2019 | Marwa BenGhoul         | Update the output in proc means, the call symput variables *****  
                        Add the notes in the proc gplot  
*****  
*****/  
  
%macro baplot(dtain=,x=,y=,index=,outd=,title=);  
  
/* proc standard standardizes the variables and saves the results in the  
   same variable names in the output data set. This means we can continue  
   using the input variable names throughout. */  
proc standard data = &datain. out=ba_std&index. mean=0 std=1;  
var &x. &y.;  
run;  
/* calculate differences and averages */  
data ba&index.;  
set ba_std&index.;  
bamean = (&x. + &y.)/2;;  
badiff = &y.-&x.;  
run;  
*ods output summary=basumm;  
*ods select none;  
proc means data = ba&index. mean std;  
var badiff;  
output out=basumm&index. n=n mean=d_mean STD=d_std;  
run;  
ods select all;
```

```
/* In the following, we take values calculated from a data set for the  
confidence limits and store them in macro variables. That's the  
only way to use them later in code.  
The syntax is: call symput('vname', value);  
Note that 'bias' is purely nominal, as the standardization means that  
the mean difference is 0. */
```

```
%global bias hici loci;  
data &outd./*lines&index.*;/  
set basumm&index.;  
call symput('bias',d_mean);  
call symput('hici',d_mean+(1.96 * d_std));  
call symput('loci',d_mean-(1.96 * d_std));  
run;  
%put &bias;  
%put &hici;  
%put &loci;
```

```
/* We use the macro variables just created in the vref= option below;  
vref draws reference line(s) on the vertical axis. lvref specifies  
a line type. */
```

```
symbol1 v=x c=blue h=1 w=1;  
title &title.;  
title2 "&x and &y standardized";  
axis1 rellabel=(j=c h=9pt "1" "B" "c");  
/*footnote1 angle=90 "Mean Difference";  
footnote2 angle=-90 "Bland-Altman Upper limit of agreement (LoA)";  
footnote3 angle=-90 "Bland-Altman Lower limit of agreement (LoA)";*/
```

```
proc gplot data=ba&index.;  
plot badiff * bamean  
/ vref = &hici. &bias. &loci. cv=red lvref=2;;  
/* use the axis defined above as the vertical axis*/  
label badiff = "Difference [&y-&x]" bamean="Mean [&x + &y]/2";  
note m=(60pct,50pct) "Bland-Altman Upper limit d_mean+(1.96 * d_std)";  
note m=(60pct,40pct) "Mean Difference (Bias)";  
note m=(60pct,30pct) "Bland-Altman Lower limit d_mean+(1.96 * d_std)";  
run;  
quit;  
%mend baplot;
```

Appendix 6 - Multilayer Perceptron for Imputation of Missing Data (MLPIMID) (SAS macro program)

Multilayer Perceptron for Imputation of Missing Data (MLPIMID) (SAS macro program)

```
*****
```

```
* PROGRAM NAME :      MLPIMID.sas
* DESCRIPTION :      Impute the missing observations via the predictions of nlp
* Assumptions:      nlp.sas and baplot.sas should be run before this program
* PARAMETERS:
* PROGRAMMER :
* DATE WRITTEN :
```

```
*****
```

```
* INPUT FILES :
```

```
* OUTPUT FILES :
```

```
*****
```

```
* CHANGE HISTORY
```

```
*****
```

```
* DATE      | AUTHOR                                | CHANGE
```

```
* 04-2019   | Maroua BenGhoul      | Initial commit
```

```
*****/
```

```
%macro mlpimid (datain= ,index=
,miss=,miss1=,miss2=,miss3=,miss4=,ctmis=,input=,vardep=,varindp=,ident=,BA=);
/*datain: input data, raw data including the missing variables
```

```
miss: the list of missing variables
```

```
miss1: the first missing variable
```

```
miss2: the second missing variable
```

```
ctmis: Count of missing variables
```

```
input: list of variables to be used in the NN and in the regression
```

```
vardep: dependent variable
```

```
varindp: independent variables
```

```
ident: id variables (like the subject id, the timepoint etc.)
```

```
BA: Dummay variable 1 if the user wants the comparison via Bland Atman, 0 if not
```

```
*/
```

```
*****;
```

```
*      Examining Missing observations in the data      *;
```

```
*****;
```

```
/*1. Number of missing values vs. number of non missing values in each variable*/
```

```
ods exclude all; /* suppress display to open ODS destinations */
```

```
proc means data = &datain. n nmiss STACKODSOUTPUT;
```

```
var &miss. &input.;ods output Summary=miss_&index.;/*write statistics to data set */
```

```
run;quit;
```

```
ods exclude none; /* restore normal display of tables */
```

```
/*proc mi data=&datain.;ods select misspattern;run;*/
```

```
%if "&miss1." ne "" %then
```

```
%do;
```

```
ods exclude all; /* suppress display to open ODS destinations */
```

```
proc means data = &datain. mean std min max STACKODSOUTPUT;
```

```
/* ods exclude all;      suppress display to open ODS destinations
```

```
proc means data =_valid&index. n mean nmiss STACKODSOUTPUT;
```

```
var mp_&miss1.;ods output Summary=m_valid&index.; /* write statistics to data set
```

```
run;quit;
```

```
ods exclude none; */
```

```
data datvf_&index. (drop = idt flag);
```

```
merge datval&index.(in=a) _valid&index.(in=b);
```

```
by idt flag;
```

```
if a and b;
```

```
run;
```

```
data _Finl&index.;
```

```
set Train&index. Test&index. datvf_&index.;
```

```
run;
```

```
data _Final&index. (drop = Min_&miss1. Max_&miss1. m_&miss1. );
```

```
set _Finl&index.;
```

```
%if "&miss1." ne "" %then
```

```
%do;
```

```
if &miss1. =. then &miss1._i=(mp_&miss1.*(max_&miss1.-min_&miss1.))+
min_&miss1.;
```

```
else &miss1._i = &miss1.;
```

```
%end;
```

```
*impvarf=(impVR*StdDev)+mean;
```

```
%if "&miss2." ne "" %then
```

```
%do;
```

```
if &miss2. =. then &miss2._i=(mp_&miss2.*(max_&miss2.- min_&miss2.))+
min_&miss2.;
```

```
else &miss2._i = &miss2.;
```

```
%end;
```

```
%if "&miss3." ne "" %then
```

```
%do;
```

```
if &miss3. =. then &miss3._i=(mp_&miss3.*(max_&miss3.- min_&miss3.))+
min_&miss3.;
```

```
else &miss3._i = &miss3.;
```

```
%end;
```

```
%if "&miss4." ne "" %then
```

```
%do;
```

```
if &miss4. =. then &miss4._i=(mp_&miss4.*(max_&miss4.- min_&miss4.))+
min_&miss4.;
```

Multilayer Perceptron for Imputation of Missing Data (MLPIMID) (SAS macro program)

```
var &miss1.;output out=m_&miss1. mean=m_&miss1. std=sd_&miss1. min=min_&miss1.
max=max_&miss1.; /* write statistics to data set */
run;quit;
ods exclude none;          /* restore normal display of tables */

data data_&index. (drop = _type_ _freq_);
set &datain.;
if _n_ =1 then set m_&miss1.;
if sd_&miss1. not in (0,.) then &miss1._s=(&miss1.-m_&miss1.)/sd_&miss1.;
d_&miss1.=max_&miss1.-min_&miss1.;
if d_&miss1. not in (0,.) then &miss1._n=(&miss1.-min_&miss1.)/(max_&miss1.-min_&miss1.);
run;

%end;

%if "&miss2." ne "" %then
%do;
ods exclude all; /* suppress display to open ODS destinations */
proc means data = &datain. mean std min max STACKODSOUTPUT;
var &miss2.;output out=m_&miss2. mean=m_&miss2. std=sd_&miss2. min=min_&miss2.
max=max_&miss2.; /* write statistics to data set */
run;quit;
ods exclude none;          /* restore normal display of tables */

data data_&index. (drop = _type_ _freq_);
set data_&index.;
if _n_ =1 then set m_&miss2.;
if sd_&miss2. not in (0,.) then &miss2._s=(&miss2.-m_&miss2.)/sd_&miss2.;
d_&miss2.=max_&miss2.-min_&miss2.;
if d_&miss2. not in (0,.) then &miss2._n=(&miss2.-min_&miss2.)/(max_&miss2.-min_&miss2.);
run;

%end;

%if "&miss3." ne "" %then
%do;
ods exclude all; /* suppress display to open ODS destinations */
proc means data = &datain. mean std min max STACKODSOUTPUT;
var &miss3.;output out=m_&miss3. mean=m_&miss3. std=sd_&miss3. min=min_&miss3.
max=max_&miss3.; /* write statistics to data set */
run;quit;
ods exclude none;          /* restore normal display of tables */

data data_&index. (drop = _type_ _freq_);
set data_&index.;
if _n_ =1 then set m_&miss3.;
if sd_&miss3. not in (0,.) then &miss3._s=(&miss3.-m_&miss3.)/sd_&miss3.;
d_&miss3.=max_&miss3.-min_&miss3.;
if d_&miss3. not in (0,.) then &miss3._n=(&miss3.-min_&miss3.)/(max_&miss3.-min_&miss3.);
run;
```

```
else &miss4._i = &miss4.;
%end;

run;

data _NN&index.;
set _Final&index. (drop =&miss.);

if "&miss1." ne "" then &miss1.= &miss1._i;

%if "&miss2." ne "" %then
%do;
&miss2.=&miss2._i;
%end;

%if "&miss3." ne "" %then
%do;
&miss3.=&miss3._i;
%end;

%if "&miss4." ne "" %then
%do;
&miss4.=&miss4._i;
%end;

run;

/*****
* Handling missing observations with other methods */
*****/

/* multiple imputation using EM*/
proc mi DATA = &datain. NIMPUTE=10 noprint;
EM OUT = EM_&index.;
var &miss. &input.;
run;
/*Mean imputation: Use PROC STDIZE to replace missing values with mean*/
proc stdize data=&datain. out=impMean_&index.
oprefix=Orig_ /* prefix for original variables */
reponly /* only replace; do not standardize */
method=MEAN; /* or MEDIAN, MINIMUM, MIDRANGE, etc. */
var &miss. &miss4.; /* you can list multiple variables to impute */
run;
/*Listwise deletion*/
data del_&index. (drop=);
set &datain.;
array x {*} &miss.;
do j = 1 to dim(x);
```

Multilayer Perceptron for Imputation of Missing Data (MLPIMID) (SAS macro program)

```
%end;

%if "&miss4." ne %then
%do;
    ods exclude all; /* suppress display to open ODS destinations */
    proc means data = &datain. mean std min max STACKODSOUTPUT;
        var &miss4.; output out=m_&miss4. mean=m_&miss4. std=sd_&miss4. min=min_&miss4.
max=max_&miss4.; /* write statistics to data set */
    run; quit;
    ods exclude none; /* restore normal display of tables */

    data data_&index. (drop = _type_ _freq_);
    set data_&index.;
    if _n_ = 1 then set m_&miss4.;
    if sd_&miss4. not in (0,.) then &miss4._s=(&miss4.-m_&miss4.)/sd_&miss4.;
    d_&miss4.=max_&miss4.-min_&miss4.;
    if d_&miss4. not in (0,.) then &miss4._n=(&miss4.-min_&miss4.)/(max_&miss4.-min_&miss4.);
    run;

%end;

%let prTrain = 0.7; /* proportion of training data */
%let prValid = 0.1; /* proportion of validation data */
%let prTest = %sysvalf(1 - &prTrain - &prValid); /* remaining are used for testing */

%if &ctmis.=1 %then
%do;
    data Train&index. Test&index. datval&index.;
    array p[2]_temporary_ (&prTrain., &prTest.);
    set data_&index.;
    call streaminit(123); /* set random number seed */
    /*nm = nmiss(of &miss.);*/
    _k = rand("Table", of p[*]);
    if _k = 1 and &miss1. ne . then output Train&index.;
    else if _k = 2 and &miss1. ne . then output Test&index.;
    else output datval&index.;
    drop _k;
run;
%end;

%if &ctmis.=2 %then
%do;
    data Train&index. Test&index. datval&index.;
    array p[2]_temporary_ (&prTrain., &prTest.);
    set data_&index.;
    call streaminit(123); /* set random number seed */
    _k = rand("Table", of p[*]);

    if _k = 1 and &miss1. ne . and &miss2. ne . then output Train&index.;

    if missing(x(j)) THEN DELETE;
end;
run;
/*multiple imputation using MCMC*/
proc mi DATA = &datain. NIMPUTE=10 out= MCMC_&index. noprint;
    MCMC CHAIN = MULTIPLE /*PLOTS = TRACE PLOTS = ACF*/;
    var &miss. &input.;
run;

*****;
* fit statistics: RMSE and R2 *;
*****;
/*proc reg data= MCMC_&index.(where=(Imputation_=5)) outest = _MCMC&index.
ADJRSQ AIC BIC MSE noprint;
    model &vardp.=&varindp.;
run;
quit;

proc reg data= del_&index. outest = _del&index. ADJRSQ AIC BIC MSE noprint;
    model &vardp.=&varindp.;
run;
quit;

proc reg data= impMean_&index. outest = _mean&index. ADJRSQ AIC BIC MSE noprint;
    model &vardp.=&varindp.;
run;
quit;

proc reg data= EM_&index. outest = _EM&index. ADJRSQ AIC BIC MSE noprint;
    model &vardp.=&varindp.;
run;
quit;

proc reg data= _NN&index. outest = _NNF&index. ADJRSQ AIC BIC MSE noprint;
    model &vardp.=&varindp.; /*#### Should be noted that the imputed variable name must be
changed
run;
quit;

data fit_&index.;
    set _MCMC&index. (in=a) _EM&index. (in=b) _del&index. (in=d) _mean&index.(in=e)
_NNF&index.(in=f);
    id = &index.;
    drop _Type_;
    if a then method ="MCMC";
    if b then method ="EM";
    if d then method ="DEL";
```

Multilayer Perceptron for Imputation of Missing Data (MLPIMID) (SAS macro program)

```
    else if _k = 2 and &miss1. ne . and &miss2. ne . then output Test&index.;
    else
        output datval&index.;
    drop _k;

run;

%end;

%if &ctmis.=3 %then
%do;
data Train&index. Test&index. datval&index.;
array p[2]_temporary_ (&prTrain., &prTest.);
set data_&index.;
call streaminit(123); /* set random number seed */
_k = rand("Table", of p[*]);

    if _k = 1 and &miss1. ne . and &miss2. ne . and &miss3. ne . then output Train&index.;
    else if _k = 2 and &miss1. ne . and &miss2. ne . and &miss3. ne . then output Test&index.;
    else
        output datval&index.;
    drop _k;

run;

%end;

%if &ctmis.=4 %then
%do;
data Train&index. Test&index. datval&index.;
array p[2]_temporary_ (&prTrain., &prTest.);
set data_&index.;
call streaminit(123); /* set random number seed */
_k = rand("Table", of p[*]);

    if _k = 1 and &miss1. ne . and &miss2. ne . and &miss3. ne . and &miss4. ne . then output
Train&index.;
    else if _k = 2 and &miss1. ne . and &miss2. ne . and &miss3. ne . and &miss4. ne . then output
Test&index.;
    else
        output datval&index.;
    drop _k;

run;

%end;

*****
* Call the nlp.sas program to use netnlp and netrun *;
*****

%if &ctmis.=1 %then
%do;
```

```
    if e then method ="MEAN";
    if f then method ="NN";
run;*/

/*ods output ParameterEstimates=parms FitStatistics=fitts;*/
ods output FitStatistics=MCMC_F&index.;
proc glm data=MCMC_&index.(where=(Imputation_=5));
    model &vardep.=&varindp.;
    output out=MCMC_E&index. p=yhat r=resid stdr=eresid;
run;
quit;

ods output FitStatistics=EM_F&index.;
proc glm data=EM_&index.;
    model &vardep.=&varindp.;
    output out=EM_E&index. p=yhat r=resid stdr=eresid;
run;
quit;

ods output FitStatistics=Mean_F&index.;
proc glm data=impMean_&index.;
    model &vardep.=&varindp.;
    output out=Mean_E&index. p=yhat r=resid stdr=eresid;
run;
quit;

ods output FitStatistics=del_F&index.;
proc glm data=del_&index.;
    model &vardep.=&varindp.;
    output out=del_E&index. p=yhat r=resid stdr=eresid;
run;
quit;

ods output FitStatistics=NN_F&index.;
proc glm data=_NN&index.;
    model &vardep.=&varindp.;
    output out=NN_E&index. p=yhat r=resid stdr=eresid;
run;
quit;

data fit_&index.;
    set MCMC_F&index.(in=a) EM_F&index.(in=b) Mean_F&index.(in=d)
del_F&index.(in=e) NN_F&index.(in=f);
    id = &index.;
    *drop _Type_;
    if a then method ="MCMC";
    if b then method ="EM";
    if d then method ="DEL";
```

Multilayer Perceptron for Imputation of Missing Data (MLPIMID) (SAS macro program)

```
%netnlp(data=Train&index.,xvar=&input.,yvar = &miss1._n,hidden=3,outest=_EST_, random=12);
%netrun(data=TEST&index.,xvar=&input., yvar = &miss1._n,test=0,inest=_EST_,out=_test&index.);
%netrun(data=datval&index.,xvar=&input.,yvar= &miss1._n,test=1,inest=_EST_,out=_valid&index.);
%end;

%else %if &ctmis.=2 %then
%do;
%netnlp(data=Train&index.,xvar=&input.,yvar = &miss1._n &miss2._n,hidden=3,outest=_EST_, random=12);
%netrun(data=TEST&index.,xvar=&input., yvar = &miss1._n &miss2._n,test=0,inest=_EST_,out=_test&index.);
%netrun(data=datval&index.,xvar=&input.,yvar= &miss1._n &miss2._n,test=1,inest=_EST_,out=_valid&index.);
%end;

%else %if &ctmis.=3 %then
%do;
%netnlp(data=Train&index.,xvar=&input.,yvar = &miss1._n &miss2._n &miss3._n,hidden=3,outest=_EST_,
random=12);
%netrun(data=TEST&index.,xvar=&input., yvar = &miss1._n &miss2._n
&miss3._n,test=0,inest=_EST_,out=_test&index.);
%netrun(data=datval&index.,xvar=&input.,yvar= &miss1._n &miss2._n
&miss3._n,test=1,inest=_EST_,out=_valid&index.);
%end;

%else %if &ctmis.=4 %then
%do;
%netnlp(data=Train&index.,xvar=&input.,yvar = &miss1._n &miss2._n &miss3._n
&miss4._n,hidden=4,outest=_EST_, random=12);
%netrun(data=TEST&index.,xvar=&input., yvar = &miss1._n &miss2._n &miss3._n
&miss4._n,test=0,inest=_EST_,out=_test&index.);
%netrun(data=datval&index.,xvar=&input.,yvar= &miss1._n &miss2._n &miss3._n
&miss4._n,test=1,inest=_EST_,out=_valid&index.);
%end;

data netfit&index.(keep=_rmse_ ind_);
set _test&index. end=lastline;
ind_ = &index.;
if lastline then output;
run;

data datval&index.;
set datval&index.;
idt = _N_;

if &miss1.=. then flag =1;
else flag=0;

run;

if e then method ="MEAN";
if f then method ="NN";
run;

%if &BA. =1 %then
%do;
*****;
* Bland Altman *;
*****;

* Merging the data with different methods to be compared by Bland Altman
*;
proc sort data = MCMC_&index.;
by &ident.;
where _Imputation_=5;
run;
proc sort data = del_&index.;
by &ident.;
run;
proc sort data = impMean_&index.;
by &ident.;
run;
proc sort data = EM_&index.;
by &ident.;
run;
proc sort data = _NN&index.;
by &ident.;
run;

data _Bland&index.;
merge EM_&index.(in=a keep =&miss1. &ident. rename=(&miss1.=&miss1.E))
MCMC_&index.(in=b keep =&miss1. &ident. rename=(&miss1.=&miss1.C))
_NN&index.(in=c keep=&miss1. &ident.);
by &ident.;
if a and b and c;
run;

/*Calling the macro baplot to run the Bland Atman*/
%baplot(datin=_Bland&index., x=&miss1.E,
y=&miss1.,out=DifEM&index.,title="Bland-Altman plot of missing imputation via EM &x
and missing imputation via NN &y");
%baplot(datin=_Bland&index., x=&miss1.C,
y=&miss1.,out=DifMC&index.,title="Bland-Altman plot of missing imputation via MCMC
&x and missing imputation via NN &y");

title;
title2;
title;
```

Multilayer Perceptron for Imputation of Missing Data (MLPIMID) (SAS macro program)

```
data _valid&index. /*(keep=id mp_&miss1. flag)*/;
set _valid&index.;
idt = _N_;

%if "&miss1." ne "" %then
%do;
    if &miss1._n ne . then
    do;
        mp_&miss1. = &miss1._n; /*mpVR represents the imputed variable*/
        flag = 0;
        end;
    else do;
        mp_&miss1.=_p1;
        flag = 1;
    end;
%end;

%if "&miss2." ne "" %then
%do;
    if &miss2._n ne . then mp_&miss2.=&miss2._n; /*mpVR represents the imputed variable*/
    else mp_&miss2.=_p2;
%end;

%if "&miss3." ne "" and "&miss3._n" ne "" %then
%do;
    if &miss3._n ne . then mp_&miss3.=&miss3._n; /*mpVR represents the imputed variable*/
    else mp_&miss3.=_p3;
%end;

%if "&miss4." ne "" and "&miss4._n" ne "" %then
%do;
    if &miss4._n ne . then mp_&miss4.=&miss4._n; /*mpVR represents the imputed variable*/
    else mp_&miss4.=_p4;
%end;

run;
```

```
data Dif&index. (keep=name d_mean d_std);
set DifEM&index.(in=a) DifMC&index.(in=b);
length name $70.;
if a then name="Difference between EM and NN";
if b then name="Difference between MCMC and NN";
run;
```

```
PROC EXPORT
DATA= Dif&index. DBMS=xls REPLACE OUTFILE="&path.\Dif&index..xls" /*output
data*/
REPLACE;
run;
```

%end;

```
*****
* Data export *
*****
```

```
PROC EXPORT
DATA= fit_&index. DBMS=xls REPLACE OUTFILE="&path.\fit_&index..xls" /*output
data*/
REPLACE;
run;
```

%mend mlpimid;

Appendix 7 - Different Methods of Detecting and Handling Outliers Including EMIMID and WINWIS

Different methods of Detecting and Handling outliers including EMIMID and WINWIS

```
*****
* DESCRIPTION : Different outliers handling methods will be used in this program
*               Trimming, Winzorisation and EM algorithm
* Assumptions: %baplot macro should be run before the program
* PARAMETERS:
* CALLED BY : methout.sas
* CALLS TO :
* PROGRAMMER : Marwa BEN GHOU
* DATE WRITTEN :
*****
* INPUT FILES : baplot.sas
* OUTPUT FILES :
*****
* C H A N G E H I S T O R Y
*****
* DATE | AUTHOR | CHANGE

*26MAR2018 | Marwa BEN GHOU | Initial commit *****
*****
/* datain: raw data
dataout: data with the imputed variable
x: variable to be used in the proc MI
Y: Variable to be imputed
vars: list of variables to be used in the proc reg to calculate the fit
id: visits number
sbj: subjects
index: integer number to attribute in the end of the data set
*/
%macro METHOUT(datain=,x=,y=,dataout=,sbj=,id=,var=,index=);
/*Graph*/
ods graphics on;
proc sgplot data = &datain.;
reg y=&y. x = &sbj. / linattr = (color = red thickness=2);
run;

proc gplot data = &datain.;
plot &y.*&id. = &sbj. / nolegend;
run;
quit;

/*Descriptive statistics*/
proc means data=&datain. noprint;
var &y.;
output out=stat n=n mean=Mean STD=SD median=Median
min=Minimum max=Maximum qrange=IQR q1=q1 q3=q3;
run;
```

```
/*Outliers treshold */
data stat (keep = High Low);
set stat;
minval = Mean - 3 * SD;maxval = Mean + 3 * SD;
High = q3 + 1.5*IQR;Low = q1 - 1.5*IQR;
run;
/*Outliers checking and Flagging*/
data ACTG_01;
set &datain.;
if _n_ = 1 then set stat;
if logma ne . then
do;
if &y. > High or &y. < Low then FLG = 1; /* 1 means outlier 0 means not outlier*/
else FLG = 0;
end;
else FLG = .;
run;
/****** Handling outliers by EM; if Flagged as outlier than to be missing*/
data ACTG_02;
set ACTG_01;
if FLG=1 then &y. =.;
run;
/*specifies seed to begin random number generator*/
proc mi data = ACTG_02 out = EM_ACTG seed =1234 NIMPUTE=10;
var &y. &x.;
run;

data EM_ACTG;set EM_ACTG; where _Imputation_=10;run;

proc sgplot data = EM_ACTG; /*without outliers*/
reg y = &y. x = &id. / linattr = (color = red thickness=2);
run;
```

Different methods of Detecting and Handling outliers including EMIMID and WINWIS

```
/*EM*/
%let _timer_start = %sysfunc(datetime());
proc reg data = EM_ACTG plots(only) = fit(nocli stats=(default aic));
model &y. = txday;
by _Imputation_;
run;
/* Stop timer */
data tim_EM;
dur = datetime() - &_timer_start;
put 30*'- ' / 'TOTAL DURATION:' dur time13.2 / 30*'-';
run;
/***** Handling outliers by Trimming; if Flagged as outlier than to be
deleted*/
%let _timer_start = %sysfunc(datetime());
data _TRM;
set ACTG_01;
if FLG=1 then delete;
run;
/* Stop timer */
data tim_TRM;
dur = datetime() - &_timer_start;
put 30*'- ' / 'TOTAL DURATION:' dur time13.2 / 30*'-';
run;
```

```
/*winsorisation by subject data*/
%let _timer_start = %sysfunc(datetime());
%let L=10; /* 10th percentile */
%let H=%eval(100 - &L); /* 90th percentile*/
proc univariate data=ACTG_01 noprint;
var &y.;
by &sbj.;
output out=_winsor pctlpts=&L &H
pctlpre=__y;
run;
data wors2 (drop=__);
set ACTG_01;
by &sbj.;
y=&y.;
if _n_=1 then set _winsor;
array wlo {*} __y&L;
array whi {*} __y&H;
array wval {*} wy;
array val {*} y;
```

```
/*winsorisation whole data*/
/*https://blogs.sas.com/content/iml/2017/02/08/winsorization-good-bad-and-ugly.html*/
%let _timer_start = %sysfunc(datetime());
%let L=10; /* 10th percentile */
%let H=%eval(100 - &L); /* 90th percentile*/
proc univariate data=ACTG_01 noprint;
var &y.;
output out=_winsor pctlpts=&L &H
pctlpre=__y;
run;
data wors (drop=__);
set ACTG_01;
y=&y.;
if _n_=1 then set _winsor;
array wlo {*} __y&L;
array whi {*} __y&H;
array wval {*} wy;
array val {*} y;
do _V=1 to dim(val);
wval[_V]=min(max(val[_V],wlo[_V]),whi[_V]);
end;
run;
/* Stop timer */
data tim_WS;
dur = datetime() - &_timer_start;
put 30*'- ' / 'TOTAL DURATION:' dur time13.2 / 30*'-';
run;
/***** Merge data*/
proc sort data = ACTG_01;
by &sbj. &id.;
run;
proc sort data = WORS out=_Wors (rename=(wy=YWORS) keep = &sbj. &id. wy);
by &sbj. &id.;
run;
proc sort data = WORS2 out=_Worsw (rename=(wy=YWORSW) keep = &sbj. &id. wy);
by &sbj. &id.;
run;
proc sort data = EM_ACTG out=EM_02 (rename=(&y.=YEM) keep = &sbj. &id. &y.);
by &sbj. &id.;
run;
proc sort data = _TRM out=TRM_02 (rename=(&y.=YTR) keep = &sbj. &id. &y. );
```

Different methods of Detecting and Handling outliers including EMIMID and WINWIS

```
do _V=1 to dim(val);
  wval{_V}=min(max(val{_V},wlo{_V}),whi{_V});
end;
run;
/* Stop timer */
data tim_WSW;
  dur = datetime() - &_timer_start;
  put 30*'-' / 'TOTAL DURATION:' dur time13.2 / 30*'-' ;
run;

/*****
/*Calling the macro baplot to run the Bland Atman*/
*****/
%baplot(datain=_Final, x=YWORS, y=YWORSW);
title;
title2;
title;

/*****
/*Graphic presentations*/
*****/
proc gplot data=_Final;
  plot &y.*&id. = &sbj. / nolegend;
  label &y. = "Raw variable logma" &id.="Visits";
run;
quit;

proc gplot data=_Final;
  plot YWORS*&id. = &sbj. / nolegend;
  label YWORS = "Imputed variable by winsorisation across subjects" &id.="Visits";
run;
quit;

proc gplot data=_Final;
  plot YWORSW*&id. = &sbj. / nolegend;
  label YWORSW = "Imputed variable by winsorisation within subjects" &id.="Visits";
run;
quit;

/*EM*/
proc reg data= _Final outest = fit_EM ADJRSQ AIC BIC RMSE MSE noprint;
```

```
by &sbj. &id.;
run;

data _Final;
  merge ACTG_01(in=a) _Wors(in=d) EM_02(in=b) TRM_02(in=c) _Worsw(in=e);
by &sbj. &id.;
if a;
run;

data &dataout;
  set _Final;
run;
proc gplot data=_Final;
  plot YEM*&id. = &sbj. / nolegend;
  label YEM = "Imputed variable by EM" &id.="Visits";
run;
quit;

proc gplot data=_Final;
  plot YTR*&id. = &sbj. / nolegend;
  label YTR = "Imputed variable by trimming" &id.="Visits";
run;
quit;
/*****
/* Mean calculation and MSE, RMSE*/
*****/
/*winsorisation within subjects*/
proc reg data= _Final outest = fit_YWW ADJRSQ AIC BIC RMSE MSE noprint;
  model YWORSW= &var.;
run;
quit;
proc means data = _Final noprint;
  var YWORSW;
  output out=m_YWW mean=m_vr;
run;
quit;
/*winsorisation across subjects*/
proc reg data= _Final outest = fit_YW ADJRSQ AIC BIC RMSE MSE noprint;
  model YWORS= &var.;
run;
quit;
proc means data = _Final noprint;
  var YWORS;
  output out=m_YW mean=m_vr;
run;
quit;

data MEANAL&index.;
```

Different methods of Detecting and Handling outliers including EMIMID and WINWIS

```
model YEM= &var.;
run;
quit;
proc means data = _Final noprint;
var YEM;
output out=m_YEM mean=m_vr;
run;
quit;
/*Trimming*/
proc reg data= _Final outest = fit_YTR ADJR SQ AIC BIC RMSE MSE noprint;
model YTR= &var.;
run;
quit;
proc means data = _Final noprint;
var YTR;
output out=m_YTR mean= m_vr;
run;
quit;
/*****
/* Final data of fitting*/
*****/
data FITALL&index. (keep = _id method _RMSE_ _MSE_);
set fit_YWW(in=a) fit_YW(in=b) fit_EM(in=c) fit_YTR(in=d);
if a then do;
_id = 1;
method = "Winsorisation within subjects";
end;
if b then do;
_id = 2;
method = "Winsorisation across subjects";
end;
if c then do;
_id = 3;
method = "EM";
end;
if d then do;
_id = 4;
method = "Trimming";
end;
run;
```

```
set m_YWW (in= a keep = m_vr) m_YW (in= b keep = m_vr) m_YEM (in= c keep = m_vr) m_YTR (in=
d keep = m_vr);
if a then _id = 1;
if b then _id = 2;
if c then _id = 3;
if d then _id = 4;
run;

data Time&index.;
set Tim_WSW (in= a) Tim_WS (in= b) Tim_EM (in= c) Tim_trm (in= d);
if a then _id = 1;
if b then _id = 2;
if c then _id = 3;
if d then _id = 4;
run;
proc sort data = FITALL&index.;
by _id;
run;

proc sort data = MEANAL&index.;
by _id;
run;

proc sort data = Time&index.;
by _id;
run;

data FITAL&index.;
merge FITALL&index. MEANAL&index. Time&index.;
by _id;
run;

proc datasets library = work;
delete fit_YWW fit_YW fit_EM fit_YTR _Final m_YWW m_YW m_YEM m_YTR Tim_WSW
Tim_WS Tim_EM Tim_trm
ACTG_01 _Wors EM_02 TRM_02 _Worsw EM_ACTG;
run;
quit;
%mend METHOUT;
```

Appendix 8 - Wavelets Decomposition for Outliers Detection and Handling Across Subjects (WDOHDS) (Matlab macro algorithm)

Wavelets Decomposition for Outliers Detection and Handling Across Subjects (WDOHDS) (Matlab macro algorithm)

```
%*****
% File : wdodhas.m *
%*****
% This program wdodhas.m is the abbreviation for outliers across
% subjects. It consists on detecting and handling the outliers in
% longitudinal data across subjects (for the whole data), so it is
% applicable for times series also
% Input:   veri   is the data,
%          var    is the variable to be analyzed,
%          sbj    is the subject variable
% Output:   veriupd is the new variable.

% Assumptions: %be careful to the name of the variable to be used as patid or sbj in actg02
%              before the innerjoin
%*****
% Author : Marwa BenGhoul *
% Note : This program has highly benefited from the work of(Grané & Videia, 2009).
% However, the treshold and the program structure are totally different.
%*****

function veriupd=wdodhas(var,sbj)
u=var; % initialize the variable to be analyzed as u
[cA,cD]=dwt(u,'db1'); % decompose u by Daubechies
% The treshold
q1= quantile(cD,0.25);%define the first quartile
q3= quantile(cD,0.75);%define the third quartile
IQR = iqr (cD);%calculate the inter quartile
High = q3 + 1.5*IQR;Low = q1 - 1.5*IQR; %calculate the tresholds
%Define the flag variable to detect the position of outliers
Flag = [];V = [];
leng= length (cD);

for i=1:leng
    if cD(i)>High | cD(i)< Low
        V(i) = 0;
        Flag (i) = 1;%if outlier than to be replaced in the detail by zero
    else
        V(i) = cD (i);
        Flag (i) = 0;
    end

end;

V2 = [cD transpose(V) transpose(Flag)];
newu = idwt(cA,transpose(V),'db1'); %reconstruction of the variable after the decomposition
veriupd = newu;
%Graph presentation
n = length (sbj);
umin=min(u); umax=max(u); cDmin=Low; cDmax=High;
dmin=min(cD); dmax=max(cD);
subplot(2,2,1); plot(u); title('Original variable before decomposition'); axis([0 n umin umax]);
subplot(2,2,2); plot(cD); hold on
subplot(2,2,2); plot(High*ones(n/2,1),'r-');
subplot(2,2,2); plot(Low*ones(n/2,1),'r-'); title('Detail coefficients'); axis([0 n/2 dmin dmax]);
subplot(2,2,3); plot(newu); title('Variable after decomposition (without outliers)'); axis([0 n umin umax]);
subplot(2,2,4); plot(V); hold on
subplot(2,2,4); plot(High*ones(n/2,1),'r-');
subplot(2,2,4); plot(Low*ones(n/2,1),'r-'); title('Detail coefficients (without outliers)');
axis([0 n/2 dmin dmax]);

end
```

Appendix 9 - Wavelets Decomposition for Outliers Detection and Handling Within Subjects (WDOHWS) (Matlab macro algorithm)

Wavelets Decomposition for Outliers Detection and Handling Within Subjects (WDOHWS) (Matlab macro algorithm)

```
%*****
% File : wdodhws.m *
%*****
% This program outwithsj.m is the abbreviation for outliers within
% subjects (by subjects). It consists on detecting and handling the outliers in
% longitudinal data within subjects, so it is only
% applicable for this type of data
% Input:  veri  is the data,
%         var   is the variable to be analyzed,
%         sbj   is the subject variable
% Output:  veriupd is the updated data including the new variable.

% Assumptions: %be careful to the name of the variable to be used as patid or sbj
%              before the innerjoin
% Notes:      Threshold in this program can be changed
%*****
% Author :   Marwa BenGhoul      *
```

% Decomposition by group subject

```
function veriupd=wdodhws(veri,var,sbj)
```

```
%global matdet0 sbj2 G;
```

```
%persistent sbj2 matdet0;
```

```
[G,sbjt] = findgroups(sbj);
```

```
y = var;%variable to be decomposed
```

```
G = sbj;%group variable to be decomposed by
```

```
yy=1:length(sbjt);%count of the group variables
```

```
for j=1:length(yy)
```

```
    k=0;
```

```
for i=1:length(G)
```

```
    if G(i)==j
```

```
        k=k+1;
```

```
        GG(k,j)=y(i);
```

```
end
```

```
end
```

```
bh(j)=k;
```

```
[coef(1:round(k/2),j),det(1:round(k/2),j)] = dwt(GG(1:k,j),'db1');
```

```
end
```

```
%%% After the decomposition
```

```
%Define the new subjects account after decomposition,
```

```
Countsbj = grpstats(veri,'sbj');%Getting the count of subjects
```

```
Max = max(Countsbj.GroupCount);%get the maximum of count
```

```
rep = round(Max/2); %the frequency of each subject to get the subjects repeated k/2 times
```

```
for i=1:length(sbjt)
```

```
    u (i) = i;
```

```
    sbj02 = repelem(u,rep);%repeat rep times the subject id by using the function repelem
```

```
    sbj2 = transpose (sbj02);%transpose the variable to be used in the merging later
```

```
end
```

```
%Combining the details and the coefficients into one array
```

```
matdet0=det(:);
```

```
matcoef0=coef(:);
```

```
%Combining the subjects and the details, coefficients
```

```
matdet=[sbj2 matdet0];
```

```
matcoef=[sbj2 matcoef0];
```

```
% Getting the stats by subject
```

```
% Median calculation by subject group
```

```
% Assign name to the details data and coeffecient data
```

```
matdet=array2table(matdet, 'VariableNames', {'sbj','det'});
```

```
matcoef=array2table(matcoef, 'VariableNames', {'sbj','coef'});
```

```
matcoef0=array2table(matcoef0, 'VariableNames', {'coef'});
```

```
% Median will be calclated by subject
```

```
MedSbj = grpstats(matdet,'sbj',{@median,@iqr});
```

```
% Full stat data of details
```

```
% The threshold
```

```
q1= quantile(matdet.det,0.25); q3= quantile(matdet.det,0.75);IQR = iqr (matdet.det);
```

```
High = q3 + 1.5*IQR; Low = q1 - 1.5*IQR;
```

```
% Joining data
```

```
matdet03 = join(matdet,MedSbj);
```

```
%Testing if the detail is an outlier
```

```
Flag = [];V2 = [];
```

```
leng= length (matdet03.det);
```

```
for i=1:leng
```

```
    if matdet03.det(i)>High | matdet03.det(i)< Low
```

```
        %Change the outlier by the median of each patient
```

```
        V2(i) = matdet03.median_det(i);%V2 contains the transformed details values
```

```
        Flag2 (i) = 1;
```

```
    else
```

```
        V2(i) = matdet03.det (i);
```

```
        Flag2 (i) = 0;
```

```
    end
```

```
matdet04 = [transpose(V2) transpose(Flag2)];%contains the transformed details values and flags
```

Wavelets Decomposition for Outliers Detection and Handling Within Subjects (WDOHWS) (Matlab macro algorithm)

```

end

matdet05 = [sbj2 matdet03.det matdet04];
matdet05=array2table(matdet05, 'VariableNames', {'sbj','det','trandet','Flag'});
%Final data
Final_d=[matdet05 matcoef00];
%Reconstruction of the variable
newy = idwt(Final_d.coef,Final_d.trandet,'db1');

Countsbj = grpstats(veri,'sbj');%Getting the count of subjects
Max = max(Countsbj.GroupCount);%get the maximum of count
rep1 = round(Max); %the frequency of each subject to get the subjects repeated k/2 times
for i=1:length(sbjt);
    v (i) = i;
    sbj3 = repelem(v,rep1);%repeat rep times the subject id by using the function repelem
    sbj3 = transpose (sbj3);%transpose the variable to be used in the merging later
end
newy02 = [sbj3 newy];
newy03=array2table(newy02, 'VariableNames', {'sbj','newy'});

%Keep only the original number of observation
%Creating an id to join the data by it later
yy=[1:length(sbj)];%count of the group variables
G = sbj;%group variable to be decomposed by
for j=1:length(yy)
    k=0;
for i=1:length(G)
    if G(i)==j
        k=k+1;
        A(i)=k;
    end
end

end
end
FIN=innerjoin(newy04,veri02);
veriupd = FIN;
%Graphic presentation
n = length (FIN.sbj);
umin=min(var); umax=max(var); cDmin=Low; cDmax=High;
dmin=min(matdet0); dmax=max(matdet0);
figure(1)
subplot(2,2,1); plot(var); title('Raw variable before decomposition'); axis([0 n umin umax]);
subplot(2,2,2); plot(matdet0); hold on
subplot(2,2,2); plot(High*ones(n/2,1),'r-');
subplot(2,2,2); plot(Low*ones(n/2,1),'r-'); title('Detail coefficients'); axis([0 n/2 dmin dmax]);
subplot(2,2,3); plot(FIN.newy); title('Variable after decomposition (without outliers)'); axis([0 n umin
umax]);

```

```

sbj4=sbj;B=transpose(A);
id = [B sbj4];
id=array2table(B, 'VariableNames', {'id'});
veri02=[veri id];

[GG,sbj5] = findgroups(newy03.sbj);
[n,p]=size (newy03);
m=size(sbj5);
t=[1:m];%count of the group variables
GG = newy03.sbj;%group variable to be decomposed by
for j=1:length(t)
    l=0;
for i=1:size(GG)
    if GG(i)==j
        l=l+1;
        F(i)=l;
    end
end

end
end
sbj7=newy03.sbj;FF=transpose(F);
id2 = [FF sbj7];
id2=array2table(FF, 'VariableNames', {'id'});
newy04=[newy03 id2];

```

Wavelets Decomposition for Outliers Detection and Handling Within Subjects (WDOHWS) (Matlab macro algorithm)

```
subplot(2,2,4); plot(matdet05.trandet); hold on
subplot(2,2,4); plot(High*ones(n/2,1),'r-');
subplot(2,2,4); plot(Low*ones(n/2,1),'r-'); title('Detail coefficients (without outliers)');
axis([0 n/2 dmin dmax]);
%Graphs
[G,sbjt] = findgroups(sbj);
y1 = var;%variable to be decomposed
G1 = sbj;%group variable to be decomposed by
yy=[1:length(sbjt)];%count of the group variables
for j=1:length(yy)
    k1=0;
for i=1:length(G1)
    if G1(i)==j
        k1=k1+1;
        GG1(k1,j)=y1(i);
end
end
end
```

Appendix 10 - Testing MLPIMID Algorithm

Testing MLPIMID algorithm

```
/******
* PROGRAM NAME :      nlp_test.sas
* DESCRIPTION :      Test nlp_pred_mis program
* Assumptions:      nlp_pred_mis.sas should be run before this program
                    Avoid having the variables names in the data set as Y,_Y,Density or variables having (_)
                    in the name or variable name length greater than 7.
* PARAMETERS:
* PROGRAMMER :
* DATE WRITTEN :
*****
* INPUT FILES :
* OUTPUT FILES :
*****
* C H A N G E H I S T O R Y
*****
* DATE      | AUTHOR      | CHANGE
* 05-2019   | Maroua BenGhoul   | Initial commit
*****/

*****
*      Import some classic data sets      *;
*****;
```

PROC EXPORT

```
DATA= _fitm DBMS=xls REPLACE OUTFILE="&path.\fit_ACTGm.xls" /*output data*/
REPLACE;
run;
```

PROC EXPORT

```
DATA= _diffact DBMS=xls REPLACE OUTFILE="&path.\_diffactgHAT.xls" /*output data*/
REPLACE;
run;
```

data _netfit;

```
set netfit05(in=a) netfit10(in=b) netfit15(in=c) netfit20(in=d) netfit25(in=e) netfit30(in=f);
if a then percent = "5% missingness";
if b then percent = "10% missingness";
if c then percent = "15% missingness";
if d then percent = "20% missingness";
if e then percent = "25% missingness";
if f then percent = "30% missingness";
keep _RMSE_ ind_;
run;
```

PROC EXPORT

```
DATA= _netfit DBMS=xls REPLACE OUTFILE="&path.\_actghatfith.xls" /*output data*/
```

Testing MLPIMID algorithm

```
%let path =; /*the path*/
/*Create a libname for the project, the libref must already before assign it*/
libname neural "&path";

/*Import the data to be considered in the impute macro*/
PROC IMPORT OUT= neural.cancer DATAFILE= "&path.\cancer.xls"
    DBMS=xls REPLACE;GETNAMES=Yes;DATAROW=2;RUN;

PROC IMPORT OUT= neural.heart DATAFILE= "&path.\heart.xls"
    DBMS=xls REPLACE;GETNAMES=Yes;DATAROW=2;RUN;

PROC IMPORT OUT= neural.hepatitis DATAFILE= "&path.\hepatitis.xls"
    DBMS=xls REPLACE;GETNAMES=Yes;DATAROW=2;RUN;

PROC IMPORT OUT= neural.mamo DATAFILE= "&path.\mamo.xls"
    DBMS=xls REPLACE;GETNAMES=Yes;DATAROW=2;RUN;

PROC IMPORT OUT= neural.iris DATAFILE= "&path.\iris.xls"
    DBMS=xls REPLACE;GETNAMES=Yes;DATAROW=2;RUN;

*****
*      Data sets with one missing variable      ;
*****
options symbolgen mprint;
%*mlpimid (datain=neural.cancer,index=1,miss=var7,input=var6);
options symbolgen mprint;
%mlpimid(datain=neural.cancer,index=1,miss=var7,miss1=var7,miss2=,miss3=,miss4=,
ctmis=1,input=var6,wardp=var7,varindp=var6,BA=1);

options symbolgen mprint;
%*mlpimid (datain=neural.Actg_mis,index=2,miss=logma,input=calwk);
%mlpimid(datain=neural.Actg_mis,index=2,miss=logma,miss1=logma,miss2=,miss3=,
miss4=,ctmis=1,input=calwk,wardp=logma,varindp=calwk,BA=1);

data Hyp_mis (drop =_Y Y);
set neural.Hyp_mis;
hyper = y;
run;

options symbolgen mprint;
%*mlpimid (datain=Hyp_mis,index=3,miss=hyper,input=SYSBP DIAPB BMI _age);
%mlpimid(datain=Hyp_mis,index=3,miss=hyper,miss1=hyper,miss2=,miss3=,miss4=,
ctmis=1,input=SYSBP DIAPB BMI _age,wardp=hyper,varindp=SYSBP DIAPB BMI _age);

proc datasets library=work kill noprint;
run;
quit;
```

```
REPLACE;
run;
proc datasets library=work kill noprint;
run;
quit;

*****
*Hypertension data set with different missingness levels*;
*****
options validvarname=v7;
options symbolgen mprint;
%mlpimid(datain=neural.Hyp_Mper,index=05,miss=outy05,miss1=outy05,miss2=,miss3=,
miss4=,ctmis=1,input=SYSBP DIAPB BMI _age,wardp=outy05,varindp=SYSBP DIAPB BMI
_age,ident=id,BA=1);
%mlpimid(datain=neural.Hyp_Mper,index=10,miss=outy10,miss1=outy10,miss2=,miss3=,
miss4=,ctmis=1,input=SYSBP DIAPB BMI _age,wardp=outy10,varindp=SYSBP DIAPB BMI
_age,ident=id,BA=1);
%mlpimid(datain=neural.Hyp_Mper,index=15,miss=outy15,miss1=outy15,miss2=,miss3=,
miss4=,ctmis=1,input=SYSBP DIAPB BMI _age,wardp=outy15,varindp=SYSBP DIAPB BMI
_age,ident=id,BA=1);
%mlpimid(datain=neural.Hyp_Mper,index=20,miss=outy20,miss1=outy20,miss2=,miss3=,
miss4=,ctmis=1,input=SYSBP DIAPB BMI _age,wardp=outy20,varindp=SYSBP DIAPB BMI
_age,ident=id,BA=1);
%mlpimid(datain=neural.Hyp_Mper,index=25,miss=outy25,miss1=outy25,miss2=,miss3=,
miss4=,ctmis=1,input=SYSBP DIAPB BMI _age,wardp=outy25,varindp=SYSBP DIAPB BMI
_age,ident=id,BA=1);
%mlpimid(datain=neural.Hyp_Mper,index=30,miss=outy30,miss1=outy30,miss2=,miss3=,
miss4=,ctmis=1,input=SYSBP DIAPB BMI _age,wardp=outy30,varindp=SYSBP DIAPB BMI
_age,ident=id,BA=1);

data _fithymb;
set fit_05(in=a) fit_10(in=b) fit_15(in=c) fit_20(in=d) fit_25(in=e) fit_30(in=f);
if a then percent = "5% missingness";
if b then percent = "10% missingness";
if c then percent = "15% missingness";
if d then percent = "20% missingness";
if e then percent = "25% missingness";
if f then percent = "30% missingness";
keep Method RootMSE percent;
run;
data _difhyper;
set Dif05(in=a) Dif10(in=b) Dif15(in=c) Dif20(in=d) Dif25(in=e) Dif30(in=f);
if a then percent = "5% missingness";
if b then percent = "10% missingness";
if c then percent = "15% missingness";
if d then percent = "20% missingness";
if e then percent = "25% missingness";
```

Testing MLPIMID algorithm

```
*****;
* Data sets with more than one missing variable *;
*****;
options symbolgen mprint;
%mlpimid(datain=neural.heart,index=4,miss=var5 Var8,miss1=var5,miss2=Var8,miss3=,
miss4=,ctmis=2,input=var10,vardp=var7,varindp=var5 var10 var8,BA=1);

options validvame=v7;
data hepatatis;
set neural.hepatatis;
BILIB = BILIRUBIN;
run;
options symbolgen mprint;
%mlpimid(datain=hepatatis,index=5,miss=BILIB ALK SGOT,miss1=BILIB,miss2=ALK,
miss3=SGOT,miss4=,ctmis=3,input=ANOREXIA Age,vardp=ANOREXIA,varindp=BILIB ALK
SGOT,BA=1);

options validvame=v7;
data mamo(drop=density);
set neural.Mamo;
dent = density;
run;

options symbolgen mprint;
%mlpimid(datain=Mamo,index=6,miss=BIRADS shape dent Margin ,miss1=BIRADS,miss2=shape,
miss3=dent,miss4=Margin,ctmis=4,input=Age,vardp=Severity,varindp=BIRADS shape Margin,BA=1);

*****;
* ACTG data set with different missingness levels *;
*****;
options validvame=v7;
options symbolgen mprint;
%mlpimid(datain=neural.ACTG_Mper,index=05,miss=lma05,miss1=lma05,miss2=,miss3=,
miss4=,ctmis=1,input=calwk,vardp=lma05,varindp=calwk,ident=Patid calwk,BA=1);
%mlpimid(datain=neural.ACTG_Mper,index=10,miss=lma10,miss1=lma10,miss2=,miss3=,
miss4=,ctmis=1,input=calwk,vardp=lma10,varindp=calwk,ident=Patid calwk,BA=1);
%mlpimid(datain=neural.ACTG_Mper,index=15,miss=lma15,miss1=lma15,miss2=,miss3=,
miss4=,ctmis=1,input=calwk,vardp=lma15,varindp=calwk,ident=Patid calwk,BA=1);
%mlpimid(datain=neural.ACTG_Mper,index=20,miss=lma20,miss1=lma20,miss2=,miss3=,
miss4=,ctmis=1,input=calwk,vardp=lma20,varindp=calwk,ident=Patid calwk,BA=1);
%mlpimid(datain=neural.ACTG_Mper,index=25,miss=lma25,miss1=lma25,miss2=,miss3=,
miss4=,ctmis=1,input=calwk,vardp=lma25,varindp=calwk,ident=Patid calwk,BA=1);
%mlpimid(datain=neural.ACTG_Mper,index=30,miss=lma30,miss1=lma30,miss2=,miss3=,
miss4=,ctmis=1,input=calwk,vardp=lma30,varindp=calwk,ident=Patid calwk,BA=1);

proc datasets library=work kill noprint;
```

```
if f then percent = "30% missingness";
run;
```

```
data _netfit;
set netfit05(in=a) netfit10(in=b) netfit15(in=c) netfit20(in=d) netfit25(in=e) netfit30(in=f);
if a then percent = "5% missingness";
if b then percent = "10% missingness";
if c then percent = "15% missingness";
if d then percent = "20% missingness";
if e then percent = "25% missingness";
if f then percent = "30% missingness";
keep _RMSE_ ind_;
run;
```

```
PROC EXPORT
DATA= _fithymh DBMS=xls REPLACE OUTFILE="&path._fithymhh.xls" /*output data*/
REPLACE;
run;
```

```
PROC EXPORT
DATA= _difhyper DBMS=xls REPLACE OUTFILE="&path._difhyperh.xls" /*output data*/
REPLACE;
run;
```

```
PROC EXPORT
DATA= _netfit DBMS=xls REPLACE OUTFILE="&path._netfith.xls" /*output data*/
REPLACE;
run;
/*https://www.hindawi.com/journals/tswj/2013/632437*/
```

```
*****;
*Sinus data set *;
*****;
```

```
proc datasets library=work kill noprint;
run;
quit;
```

```
data sine;
do angle=0 to 6.3 by .05;
sine=sin(angle)/2+.5;
output;
end;
run;
%netnlp(data=sine, xvar=angle, yvar=sine,hidden=3,outest=_EST_, random=12);
%netrun(data=sine, xvar=angle, yvar=sine,test=0,inest=_EST_,out=outsine);
```

```
data _netfit;
```

Testing MLPIMID algorithm

```
run;  
quit;
```

```
/*Fit statistics*/
```

```
data _fitm;  
set fit_05(in=a) fit_10(in=b) fit_15(in=c) fit_20(in=d) fit_25(in=e) fit_30(in=f);  
if a then percent = "5% missingness";  
if b then percent = "10% missingness";  
if c then percent = "15% missingness";  
if d then percent = "20% missingness";  
if e then percent = "25% missingness";  
if f then percent = "30% missingness";  
keep Method RootMSE percent;  
run;
```

```
data _diffact;  
set Dif05(in=a) Dif10(in=b) Dif15(in=c) Dif20(in=d) Dif25(in=e) Dif30(in=f);  
if a then percent = "5% missingness";  
if b then percent = "10% missingness";  
if c then percent = "15% missingness";  
if d then percent = "20% missingness";  
if e then percent = "25% missingness";  
if f then percent = "30% missingness";  
run;
```

```
set netfit1(in=a) netfit4(in=b) netfit5(in=c) netfit6(in=d);  
if a then percent = "Cancer";  
if b then percent = "Heart";  
if c then percent = "Hepatitis";  
if d then percent = "Mamo";  
keep _RMSE_ ind_ percent;  
run;
```

PROC EXPORT

```
DATA= _netfit DBMS=xls REPLACE OUTFILE="&path._fithatsall.xls" /*output data*/  
REPLACE;  
run;
```

```
data _difrm;  
set Dif1(in=a) Dif4(in=b) Dif5(in=c) Dif6(in=d);  
if a then percent = "Cancer";  
if b then percent = "Heart";  
if c then percent = "Hepatitis";  
if d then percent = "Mamo";  
keep d_mean name percent;  
run;
```

PROC EXPORT

```
DATA= _difrm DBMS=xls REPLACE OUTFILE="&path._difhatsall.xls" /*output data*/  
REPLACE;  
run;
```

Appendix 11 - Testing EMIMID and WINWIS

Testing EMIMID and WINWIS

```
%METHOUT(datain=ACTGOUT,x=txday,y=_logrna1,dataout=_data1,sbj=patid,id=id,var=txday  
nnrti cens,index=1);  
%METHOUT(datain=ACTGOUT,x=txday,y=_logrna3,dataout=_data2,sbj=patid,id=id,var=txday  
nnrti cens,index=2);  
%METHOUT(datain=ACTGOUT,x=txday,y=_logrna5,dataout=_data3,sbj=patid,id=id,var=txday  
nnrti cens,index=3);
```

Appendix 12 - Testing WDODHAS and WDODHWIS algorithms

Testing WDODHAS and WDODHWIS algorithms

% WDODHAS%

```
Out1 = wdodhas(Final.logrna1,Final.sbj);  
Out3 = wdodhas(Final.logrna3,Final.sbj);  
Out5 = wdodhas(Final.logrna5,Final.sbj);
```

% WDODHWIS%

```
_Out1 = wdodhwis(Final,Final.logrna1,Final.sbj);  
_Out3 = wdodhwis(Final,Final.logrna3,Final.sbj);  
_Out5 = wdodhwis(Final,Final.logrna5,Final.sbj);
```

Appendix 13 - Baseline Calculation

Baseline calculation

```
proc sort data = semp.actg out=base_01;
  by patid calwk;
  where calwk le 2;
run;
proc sort data = base_01 out=base_02 (keep = patid logma calwk) nodupkey;
  by patid;
run;
data semp.actg_01;
  merge semp.actg(in=a) base_02(in=b rename=(logma=baseline));
  by patid;
  if n(logma,baseline)=2 then chg = logma-baseline;
  if baseline not in (0,.) and calwk ge 2 and n(chg,baseline)=2 then pchg =chg/baseline*100;
run;
```

Appendix 14 - Semiparametric Model Construction for the Generated Dataset

Semiparametric Model Construction for the generated dataset

```
/** Model construction (variables to be updated) */
ods graphics on;
proc glimmix data=hypert.hypert plots=(ResidualPanel(marginal)
  ResidualPanel(unpack conditional)
  StudentPanel(marginal)); /*The default estimation method is METHOD=RSPL;
pseudo-likelihood technic*/
  class CYCLE;
  /*The continuous time effect appears in both the MODEL statement (tpoint) and the RANDOM
statement (t2)*/
  model _Y = age BMI DIAPB SYSBP /s solution ;
  random CYCLE; /*id/ type=rsmooth
  knotmethod=kdtree(bucket=100 knotinfo)subject=id*/ ;
  /* The OUTPUT statement saves the predictions of the mean of each observations to the data set
gmxxout.*/
  output out=gmxxouth1 pred(blup)=pred resid=r STDERR=stderr;
  nloptions tech=newrap;
run;
ods graphics off;
```

Appendix 15 - Semiparametric Model Construction for the Real Dataset ACTG

Semiparametric Model Construction for the real dataset ACTG

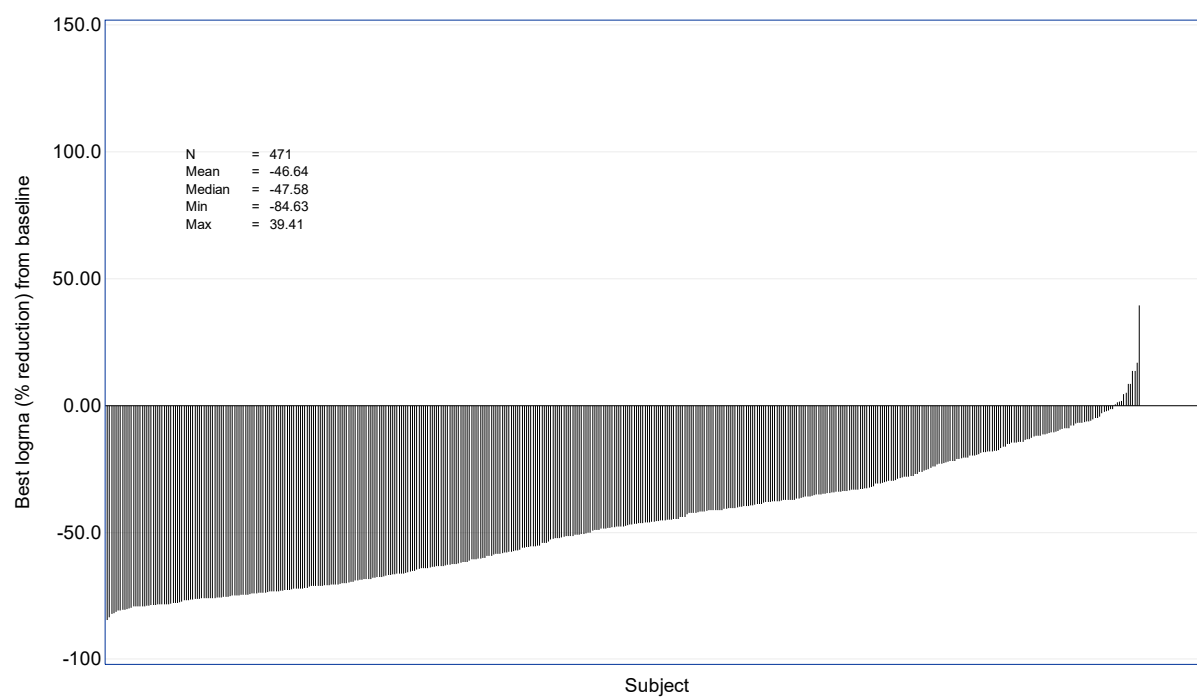
```
ods graphics on;
*ods select FitStatistics CovParms;

proc glimmix data=semp.actg_01 plots=(ResidualPanel(marginal)
    ResidualPanel(unpack conditional)
    StudentPanel(marginal)) /*METHOD=RSPL*/; /*The default estimation method is
METHOD=RSPL; pseudo-likelihood technic*/
class trtarm;
/*The continuous time effect appears in both the MODEL statement (tpoint) and the RANDOM
statement (t2)*/
model chg =trtarm txdays nnrti trtarm*nnrti trtarm*calwk baseline patid(trtarm)/s solution ;
random calwk/ type=rsmooth knotmethod=kdtree(bucket=100 knotinfo);
random intercept / subject=patid;
lsmeans trtarm/pdiff;

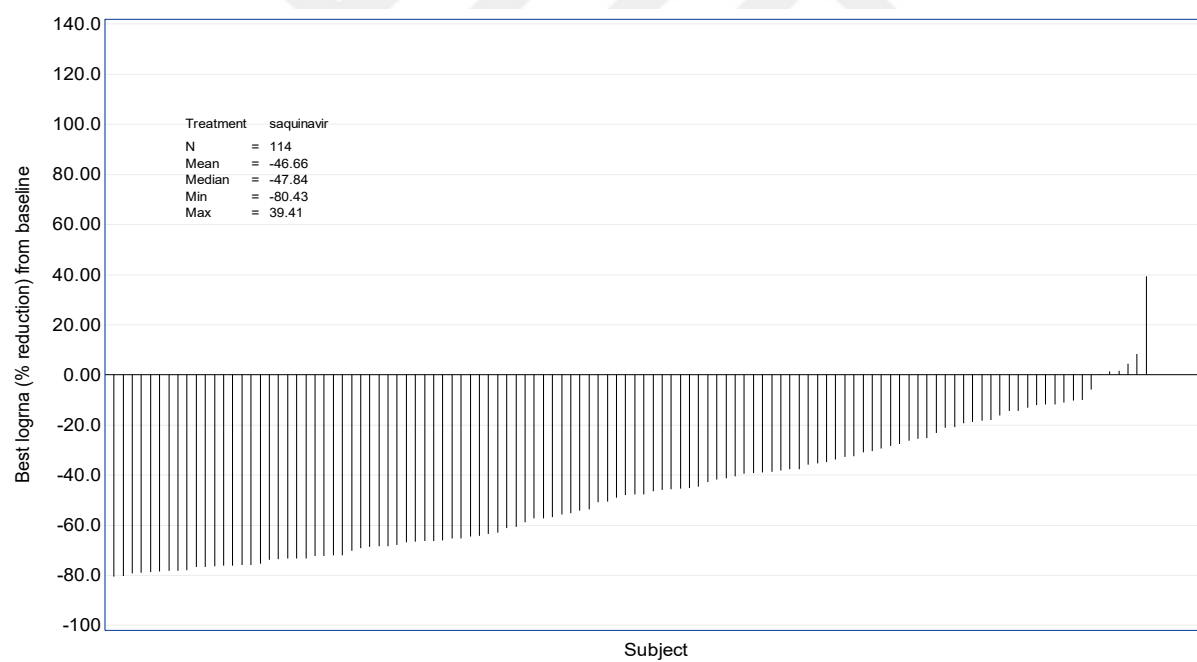
/* The OUTPUT statement saves the predictions of the mean of each observations to the data set
gmxout.*/
output out=gmxouth pred(blup)=pred resid=r STDERR=stderr;
nloptions tech=newrap;
run;

ods graphics off;
```

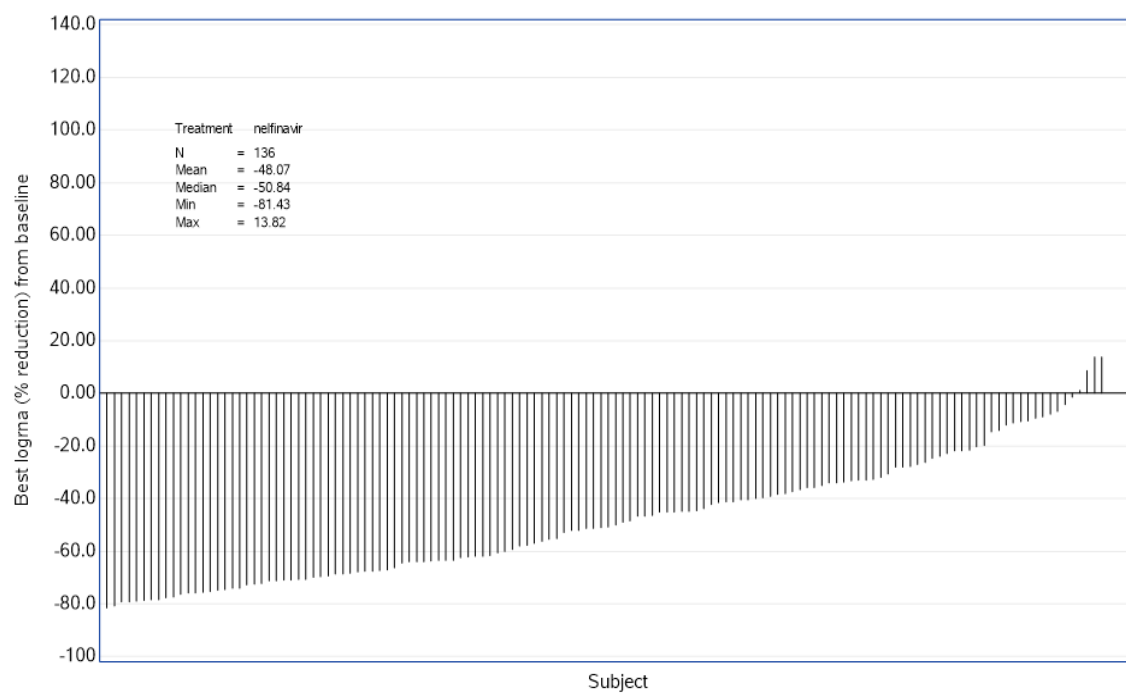
Appendix 16 - Spaghetti graphs for best logrna for all the treatment arms



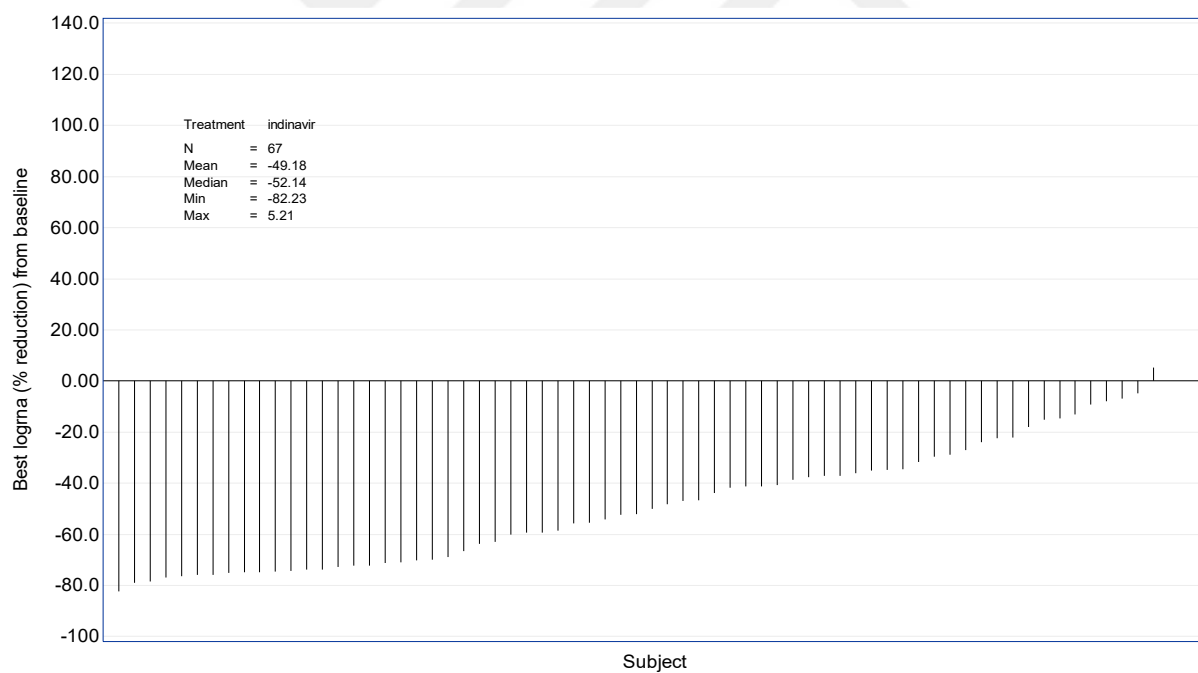
Appendix 17 - Spaghetti graphs for best logrna by treatment arms - Saquinavir



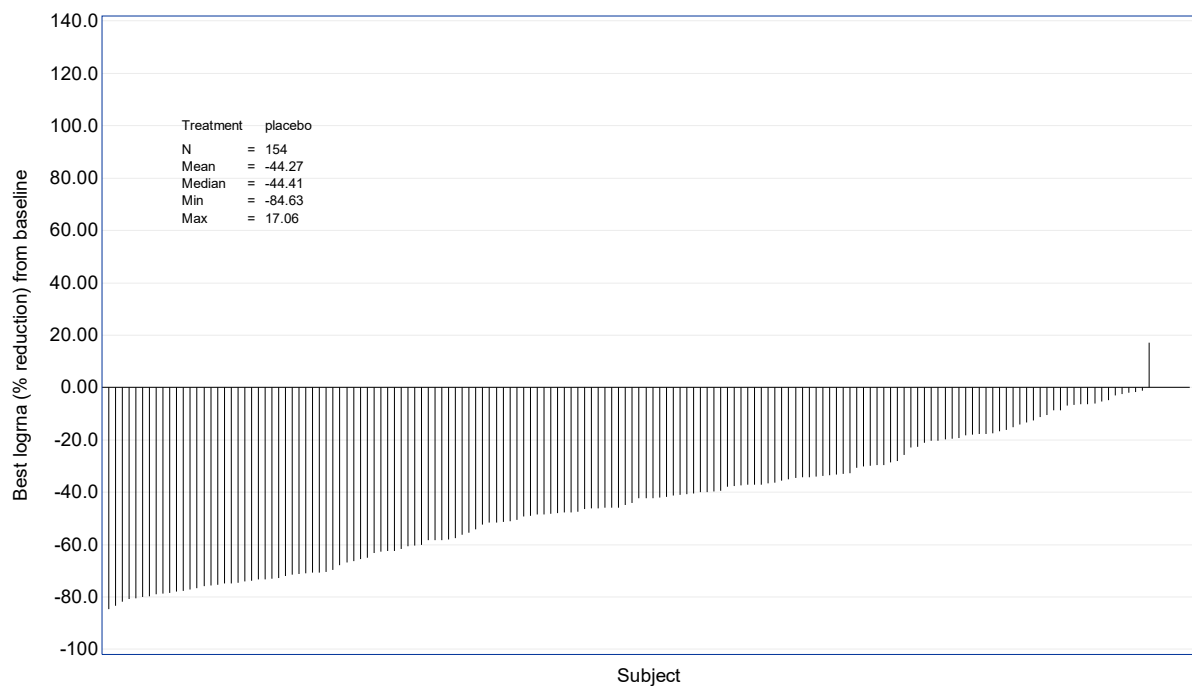
Appendix 18 - Spaghetti graphs for best logrna by treatment arms – Nelfinavir



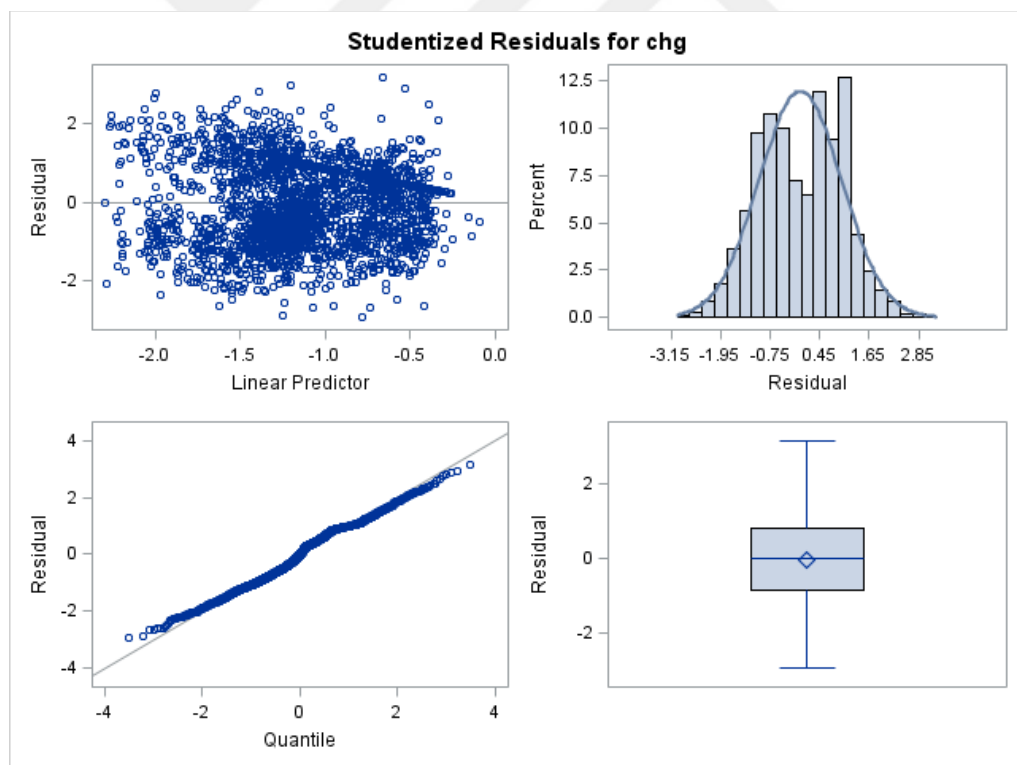
Appendix 19 - Spaghetti graphs for best logrna by treatment arms – Indinavir



Appendix 20 - Spaghetti graphs for best logrna by treatment arms – Placebo



Appendix 22 – Studentized Residuals for change from baseline



CURRICULUM VITAE

Name : Maroua BEN GHOU
Birthplace and Date : Tunisia –December 24, 1986
Email : bengoulmarwa@gmail.com/ marouabenghoul@eskisehir.edu.tr
Foreign Languages : French, English, Turkish, German
Education

- 2015–2019: PhD Statistics, Eskisehir Technical University, Turkey.
- 2011–2013: MSc. Econometrics and Economic Modeling (M2E), Tunisian Polytechnic School (EPT), Tunisia.
- 2008–2011: Engineering cycle, School of Statistics and Analysis of Information (ESSAI), Tunisia.
- 2005–2008: Scientific preparatory cycle in Mathematics and Physics (MP), Preparatory Institute for Engineering Studies of Bizerte (IPEIB), Tunisia.

Professional Experience

- 2017 – 2019: Freelancer, Upwork- platform.
- 10/2014 – 08/2017: Trial statistician (Work Remotely-Part time), dmx.
- 02/2013 – 09/2014: Project Lead/Trial statistician, dmx, Tunisia.
- 12/2011 – 01/2013: Clinical SAS programmer, dmx, Tunisia.

Scientific Activities

- 18/10/2019: Attend the IRSYSC 2019 Conference: 5th International Researchers, Statisticians and Young Statisticians Congress, Kuşadası, Turkey. Lecturer about “*Imputation of Missing Observations in Longitudinal Data via Neural Network*”.
- 28/09/2019: Attend the y-BIS 2019 Conference: Recent Advances in Data Science and Business Analytics, Istanbul, Turkey. Lecturer about “*Detection and Handling Outliers in Longitudinal Data: Can Wavelet Decomposition Be a Solution?*”.
- 22/11/2018 – 24/11/2018: Attend the 4th International Students Social Sciences Congress Nevsehir Hacı Bektas Veli University, Nevsehir, Turkey. Lecturer about “*The impact of refugees on the Turkish workforce*”.
- 08/12/2017: Attend the 10th International Statistics Congress (ISC2017) at Ankara University, Ankara, Turkey. Lecturer about “*Political Risk and Foreign Direct Investment in Tunisia: The Case of the Services Sector*”.

- 14/12/2016 - 15/07/2016: Attend the International conference “Future of Business and Finance: Challenge and Prospects after the crisis at the university of Szczecin, Faculty of Management and Economics of Services, Szczecin, Poland. Lecturer about “*The Role of the Government in Enhancing Sustainable Entrepreneurships. Does A Bigger Government Always Mean Less Entrepreneurship?*”

Publications

- Marwa BenGhoul, Berna Yazıcı and Ahmet Sezer, 2019. *Semiparametric Mixed Models for Longitudinal Data: Wavelets analysis as smoothing approach*. *Turkiye Klinikleri Journal of Biostatistics*. 11(1), 24-35. DOI: 10.5336/biostatic.2019-64979.
- Marwa BenGhoul, 2019. *Political Risk and Foreign Direct Investment in Tunisia: The Case of the Services Sector 2004-2016*. *International Journal of Sustainable Economies Management (IJSEM)*. 8(3), 48-60. DOI: 10.4018/IJSEM.2019070104.
- Marwa BenGhoul, 2019. *The Impact of Refugees on the Turkish Workforce*. 4th International Social Science Congress, 2019, ISBN: 978-605-7824-05-9 (1.c), 301-309. https://www.ilem.org.tr/media/f/uok4_kitap_1_cilt_WEB.pdf.
- Magdalena Ziolo and Marwa BenGhoul. 2019. *The Role of the Government in Enhancing Sustainable Entrepreneurships*. *Int. J. Business and Globalisation*, 22 (49), 595-617. DOI: 10.1504/IJBG.2019.100257.
- Halil İbrahim Aydın, Marwa BenGhoul, Aniela BaLacescu, 2018. *Women's role in economic development: a significant impact in the EU countries?* *International Journal of Sustainable Economies Management (IJSEM)*. 8(1), 29-38. DOI: 10.4018/IJSEM.2019010103.
- Magdalena Ziolo, Marwa Ben Ghoul, Halil İbrahim Aydın, 2018. *Chapter 5: Financial Stability versus Sustainable Development and Its Financing*. Book: *Regaining Global Stability after the Financial Crisis*, Ed. Bruno Sergi, 88-107. DOI: 10.4018/978-1-5225-4026-7.ch005.