Preprocessing and analysis of environmental data : Application to the water quality assessment of Mexican rivers

Eva Carmina Serrano Balderas

THÈSE

Pour obtenir le grade de

Docteur

Délivré par l'**UNIVERSITÉ DE MONTPELLIER**

Préparée au sein de l'école doctorale **I2S**
Et de l'unité de recherche **UMR - Espace DEV**

Spécialité: **Informatique**

Présentée par **Eva Carmina Serrano Balderas**

Preprocessing and analysis of environmental data: Application to the water quality assessment of Mexican rivers

Soutenue le 31 Janvier 2017 devant le jury composé de

Mme Hélène Fenet	Pr	Université de Montpellier	Rapporteur
Mme Karina Gibert	Pr	Univ. Polytecnica Barcelona	Rapporteur
Mme Nathalie Villa-Vilianeix	CR (HDR)	INRA Toulouse	Rapporteur
M. François Pinet	DR (HDR)	IRSTEA, Clermont-Ferrand	Président du jury
Mme Maria Aurora Armienta Hernandez	Pr	UNAM, Instituto de Geofisica	Co-directrice
Mme Laure Berti-Equille	DR (HDR)	IRD, UMR ESPACE DEV	Co-directrice
Mme Corinne Grac	Ingénieur	ENGEES Strasbourg	Co-encadrante
M. Jean-Christophe Desconnets	Ingénieur	IRD, UMR ESPACE DEV	Co-encadrant

If you torture the data long
enough, it will confess

Ronald H. Coase

Acknowledgments - *Remerciements - Agradecimientos*

Et puis...il y a ceux que l'on croise, que l'on conaît à peine, qui vous disent un mot, une phrase, vous accordent une minute, une demi-heure et changent le cours de votre vie. Katherine Pancol.

I have the pleasure to take some lines to acknowledge the people who have been with me all along these three years.

First, I thank my mentors Dr. Laure Berti-Equille and Dr. Ma. Aurora Armienta Hernandez who believe in this project and who always know how to encourage me and give me the best advises to improve my work. Thank you for your support, guidance and cheering, but mainly to have accepted to be my mentors in this project that take us extra-job to « translate » English to Spanish, Spanish to French, chemistry to computing science.

I thank Dr. Hélène Fenet, Dr. Karina Giber, Dr. François Pinet and Dr. Nathalie Villa-Vialaneix for have accepted to evaluate my thesis work. The experience of defending my PhD was very pleasant despite the stress and many hair left on my shower sink.

Je tiens sincèrement à remercier à Corinne Grac et Jean-Christophe Desconnets mes encadrants. Je suis vraiment ravi d'avoir été encadré et d'avoir travaillé avec vous. A la fois pour vos inestimables conseils, vos relectures précieuses (surtout dans les deadlines très fermés) et vos encouragements dans mes moments de faiblesse. Je remercie spécialement à Corinne Grac pour son support dans la recherche de financement, des solutions pour mes multiple et variées problèmes administratives, je n'aurai pas pu commencer sans ton précieuse aide.

Mes remerciements vont aussi à Isabelle Mougenot et Florence Le Ber pour avoir fait partie de mon comité de suivi de thèse en évaluant l'avancement de mes travaux.

Cette thèse a été effectué sur plusieurs sites en commençant par la belle alsace en france puis en partant à mon bien aimé (et chaotique) Mexique et en finissant pour le jovial Montpellier. Tout ça n'as pas pu être possible sans l'accueil des responsables d'équipes et directeurs de laboratoires. Je remercie donc Frédérique Seyler (ESPACE-DEV, IRD, Montpellier) et Jean-François Quéré (ENGEES) pour m'avoir accueilli et m'avoir offert un cadre de travail idéal.

Également, je remercie à tout l'équipe de ESPACE-DEV pour leur conseils, leur précieuses commentaires envers mon travail de thèse et surtout pour l'ambiance familiale qui fait que l'on a vraiment mal à quiter le labo.

Un grand merci à l'équipe MICADO qui m'a accueilli très chaleureusement, pour tout

ces moments des échanges et discussions. Pour notre café matinale et les délicieux croissants ou sucreries. Mais aussi pour tout ces moments que l'on a partagé sans climatisation en été ou sans chauffage en hiver ça a donnée plus d'animation dans nos journées.

Tout d'abord merci à mes collègues doctorants, aux spicy girls Minh et Yi. Il n'y a que vous pour partager ma passion pour la nourriture épicé ou les petit piments avec des fruits ;). Minh pour nos après midi-culturelles et pour m'avoir fait partie de l'arrivé d'Emma, Yi ma accompagnante des dernière jours de souffrance et pour avoir resté avec moi "enfermé dehors" au labo. Moejdeh et Phuong, vous me manqué énormément quand vous êtes pas là. A l'équipe portugais Claudio, Alexandre, Eudes, Leandro et Savio, vous m'avez donné encore plus d'envie pour apprendre le portugais. Les vrais caipiriña c'est que avec vous. A Christian et Hatim pour leur bonne humeur, et les pose café.

Un remerciement particulier à ma chère coloc et copine Cléo, qui m'a supporté, soutenu et nourri. Les petit plats, les soirée clignotage et tous les moments de complicité ont nourri pas seulement mon estomac mais ils m'ont permis d'arriver vivante jusqu'à la fin. Merci mamacita.

Agradezco, al Instituto de Geofísica por haberme permitido trabajar en sus instalaciones. Gracias a las muchachas que hacen vivir el laboratorio. A la maestra Alejandra Aguayo y a las químicas Olivia Cruz, Nora Ceniceros y Aurelia Juárez por su ayuda, consejo estimulo paciencia y compañía. Gracias a mis colegas Chuy, Esther, Juan, Israel, Azu. El trabajo en el laboratorio se vuelve una tarea agradable y ligera en su compañía. Gracias a mi estudiante Atzin por haber compartido el trabajo de conteo de cuerpos desmembrados, cabezas volantes y patas destazadas de nuestros bichitos. Gracias a la doctora Claudia A. Ponce de Leon Hill por haberme permitido trabajar y hacer uso de las instalaciones en su laboratorio. Agradezco especialmente al maestro Manuel Hernández Quiroz por su disponibilidad, su tiempo y dedicación para introducirme en el mundo de los plaguicidas. Gracias a todos los chicos del laboratorio de ecologia, trabajar en su labo siempre fue muy placentero.

Agradezco a mis padres, hermanas y mis sobrinos, quienes siempre han creído en mi y que a pesar de la distancia siempre han estado presentes para apoyarme y ayudarme. Estar lejos de ustedes siempre ha sido el mas grande desafío, todas nuestras sesiones de skype dominicales me dieron la fuerza para continuar en este proyecto personal que se volvió nuestro proyecto. Los quiero mucho.

Agradezco a mis muy queridos amigos Miguel Angel, Carlos, Jorge, Paulina y Margarita por su ayuda incondicional, sus consejos, nuestras platicas existenciales, nuestros chistes de barrio. Gracias por estar siempre presentes en forma física, electrónica o espiritual.

Enfin, je termine avec un remerciement espécial pour la personne qui a partagé avec moi dès le début le stress d'une réponse de bourse, pour me soutenir dans mon bourdel administrative, pour les jours épluchantes (avec et sans marques sur le front). Merci Batu.

To all of you THANK YOU, you will never know the extent of my gratitude.

Este trabajo a sido realizado bajo la ayuda de una beca CONACyT (Beca para estudiantes mexicanos en el extranjero)

Abstract

Data obtained from environmental surveys may be prone to have different anomalies (i.e., incomplete, inconsistent, inaccurate or outlying data). These anomalies affect the quality of environmental data and can have considerable consequences when assessing environmental ecosystems. Selection of data preprocessing procedures is crucial to validate the results of statistical analysis however, such selection is badly defined. To address this question, the thesis focused on data acquisition and data preprocessing protocols in order to ensure the validity of the results of data analysis mainly, to recommend the most suitable sequence of preprocessing tasks. We propose to control every step in the data production process, from their collection on the field to their analysis. In the case of water quality assessment, it comes to the steps of chemical and hydrobiological analysis of samples producing data that were subsequently analyzed by a set of statistical and data mining methods. The multidisciplinary contributions of the thesis are: (1) in environmental chemistry: a methodological procedure to determine the content of organochlorine pesticides in water samples using the SPE-GC-ECD (Solid Phase Extraction – Gas Chromatography – Electron Capture Detector) techniques; (2) in hydrobiology: a methodological procedure to assess the quality of water on four Mexican rivers using macroinvertebrates-based biological indices; (3) in data sciences: a method to assess and guide on the selection of preprocessing procedures for data produced from the two previous steps as well as their analysis; and (4) the development of a fully integrated analytics environment in R for statistical analysis of environmental data in general, and for water quality data analytics, in particular. Finally, within the context of this thesis that was developed between Mexico and France, we have applied our methodological approaches on the specific case of water quality assessment of the Mexican rivers Tula, Tamazula, Humaya and Culiacan.

Keywords : Data preprocessing, Data analysis, Environmental Data, Water Quality Assessment, water pollution.

Résumé

Les données acquises lors des surveillances environnementales peuvent être sujettes à différents types d'anomalies (i.e., données incomplètes, inconsistantes, inexactes ou aberrantes). Ces anomalies qui entachent la qualité des données environnementales peuvent avoir de graves conséquences lors de l'interprétation des résultats et l'évaluation des écosystèmes. Le choix des méthodes de prétraitement des données est alors crucial pour la validité des résultats d'analyses statistiques et il est assez mal défini. Pour étudier cette question, la thèse s'est concentrée sur l'acquisition des données et sur les protocoles de prétraitement des données afin de garantir la validité des résultats d'analyse des données, notamment dans le but de recommander la séquence de tâches de prétraitement la plus adaptée. Nous proposons de maîtriser l'intégralité du processus de production des données, de leur collecte sur le terrain et à leur analyse, et dans le cas de l'évaluation de la qualité de l'eau, il s'agit des étapes d'analyse chimique et hydrobiologique des échantillons produisant ainsi les données qui ont été par la suite analysées par un ensemble de méthodes statistiques et de fouille de données. En particulier, les contributions multidisciplinaires de la thèse sont : (1) en chimie de l'eau : une procédure méthodologique permettant de déterminer les quantités de pesticides organochlorés dans des échantillons d'eau collectés sur le terrain en utilisant les techniques SPE–GC-ECD (Solid Phase Extraction - Gas Chromatography - Electron Capture Detector) ; (2) en hydrobiologie : une procédure méthodologique pour évaluer la qualité de l'eau dans quatre rivières Mexicaines en utilisant des indicateurs biologiques basés sur des macroinvertébrés ; (3) en science des données : une méthode pour évaluer et guider le choix des procédures de prétraitement des données produites lors des deux précédentes étapes ainsi que leur analyse ; et enfin, (4) le développement d'un environnement analytique intégré sous la forme d'une application développée en R pour l'analyse statistique des données environnementales en général et l'analyse de la qualité de l'eau en particulier. Enfin, nous avons appliqué nos propositions sur le cas spécifique de l'évaluation de la qualité de l'eau des rivières Mexicaines Tula, Tamazula, Humaya et Culiacan dans le cadre de cette thèse qui a été menée en partie au Mexique et en France.

Mots clés : prétraitement des données, analyse des données, données environnementales, évaluation de qualité de l'eau, pollution de l'Eau.

Resumen

Los datos obtenidos de monitoreos ambientales pueden estar sujetos a diferentes tipos de anomalías (i.e., datos incompletos, inconsistencias, inexactitudes, o valores extremos). Estas anomalías que afectan la calidad de los datos ambientales pueden tener consecuencias graves al ser utilizados en la interpretación de resultados y en la evaluación de los ecosistemas. La elección de métodos de pre-tratamiento de datos resulta crucial en la validación de resultados de análisis estadísticos sin embargo, esta esta mal definida. Para estudiar esta problemática, esta tesis se concentro en la adquisición y en los protocolos de pre-tratamiento de datos para garantizar la validez de los resultados de análisis, con la finalidad principal de dar recomendaciones sobre las secuencias de pre-tratamiento mas adecuadas. Nosotros proponemos un control integral en el proceso de producción de datos, desde la colecta en el terreno de estudio hasta su análisis. En el caso de la evaluación de la calidad del agua, se trata de las etapas de análisis químico e hidrobiológico de muestras, cuyos datos producidos han sido posteriormente analizados a través de un una serie de métodos estadísticos y de exploración de datos. Las contribuciones multidisciplinarias de la tesis son: (1) en química ambiental: un procedimiento metodológico que permite determinar el contenido de plaguicidas organoclorados en muestras de agua utilizando las técnicas de SPE-GC-ECD (Solid Phase Extraction – Gas Chromatography -Electron Capture Detector); (2) en hidrobiología: un procedimiento metodológico para evaluar la calidad del agua en cuatro ríos Mexicanos, utilizando indicadores biológicos basados en macro-invertebrados; (3) en ciencia de los datos: un método para evaluar y guiar la elección de procedimientos de pre-tratamiento de datos producidos luego de las dos etapas precedentes, así como su análisis; y (4) el desarrollo de un ambiente analítico integrado en forma de aplicación desarrollada en R para el análisis estadístico de datos ambientales en general y el análisis de la calidad del agua en particular. Finalmente, dentro del contexto de esta tesis desarrollada entre México y Francia, hemos aplicado nuestras propuestas en el caso especifico de la evaluación de la calidad del agua de los rios mexicanos Tula, Tamazula, Humaya y Culiacan.

Palabras clave: pre-tratamiento de datos, análisis de datos, datos ambientales, evaluación de la calidad del agua, contaminación del agua.

Contents

1 Introduction **1**

1.1 Environmental informatics 2

1.2 Motivations 2

1.3 Objectives 4

1.4 Outline 5

2 Environmental data analysis: the case of water quality assessment with a multidisciplinary survey **7**

2.1 Introduction 8

2.2 Data collection in physico-chemical and chemical water quality assessment . 8

2.2.1 Water quality assessment of rivers 8

2.2.2 Collection of data about emerging pollutants in rivers 11

2.2.3 Chemical analysis in water 12

2.3 Data collection from biomonitoring in water quality assessment 17

2.3.1 Biological organisms in biomonitoring 17

2.3.2 Ecological assessment of rivers using macroinvertebrates 18

2.3.3 Biomonitoring data: general characteristics, common uncertainties and anomalies 23

2.4 Preprocessing and analysis of environmental data 25

2.4.1 Data anomalies and their detection 26

2.4.1.1 Data anomalies 26

2.4.1.2 Detection of data anomalies 28

2.4.2 Dealing with data anomalies and main preprocessing procedures . . 31

2.4.3 Impact of data preprocessing procedures on statistical analysis results 37

2.5 Summary 38

3 Acquisition of Environmental Data **41**

3.1 Introduction 42

3.2 Description of the study sites 42

3.3 Physico-chemical and chemical data acquisition 45

3.3.1 Sampling of water samples 45

3.3.2 Analysis of major elements, heavy metals, and arsenic in water samples 46
3.3.3 Analysis of organochlorine pesticides and PPCPs in water samples . 48
3.4 Hydrobiological data acquisition 53
3.4.1 Sampling of macroinvertebrates 53
3.4.2 Analysis of macroinvertebrates 53

4 Preprocessing and Analysis of Environmental Data 59
4.1 Introduction 60
4.2 Data preprocessing procedures 62
4.2.1 Feature selection 62
4.2.2 Normalization 62
4.2.3 Imputation methods 63
4.2.4 Outlier detection methods 66
4.3 Synthetic data description 69
4.4 Semi-synthetic data description 73
4.5 Robustness study of data preprocessing procedures 75
4.5.1 Selection of features 75
4.5.2 Normalization of data 76
4.5.3 Handling missing data 76
4.5.4 Handling outlying data 77
4.5.5 Results and discussion on robustness of data preprocessing procedures 77
4.5.5.1 Imputation methods robustness 77
4.5.5.2 Outlier detection assessment 78
4.6 Study of the impact of data preprocessing procedures on statistical results . 80
4.6.1 Impact on regression results 80
4.6.2 Impact on classification results 81
4.6.3 Impact on clustering results 82
4.6.4 Results and discussion about preprocessing procedures on statistical results 85
4.6.4.1 Regression analysis 85
4.6.4.2 Classification analysis 95
4.6.4.3 Clustering results 106
4.7 Summary and concluding remarks 114

5 Development and experimental results 119
5.1 Introduction 120
5.2 Related work on scientific workflow systems 120
5.3 EvDA: Development of R Shiny application for environmental data preprocessing and analysis 121

5.3.1 Main workflow . . . 121
5.3.2 Overview of the application . . . 123
5.4 Case study: Water pollution of the Tula, Culiacan, Tamazula, and Humaya rivers . . . 128
5.4.1 Data description . . . 128
5.4.1.1 Physico-chemical and chemical data . . . 128
5.4.1.2 Biological data . . . 129
5.4.2 Data preprocessing . . . 129
5.4.3 Results of the analysis . . . 131
5.5 Conclusion and future development . . . 137

6 Conclusions **139**
6.1 Contributions . . . 140
6.2 Future work . . . 141

7 Résumé étendue **145**
7.1 Motivations . . . 146
7.2 Objectifs . . . 146
7.3 Acquisition des données . . . 147
7.3.1 Sites d'étude . . . 147
7.3.2 Acquisiton de données physico-chimiques et chimiques . . . 149
7.3.3 Analyse des pesticides organochlorés et PPCPs . . . 150
7.3.4 Acquisition des données hydrobiologiques . . . 150
7.4 Pré-traitement et analyse des données . . . 151
7.4.1 Données synthétiques et semi-synthétiques . . . 152
7.4.2 Pré-traitement de données . . . 154
7.4.3 Analyse statistique . . . 154
7.5 Développement . . . 156
7.6 Conclusions . . . 158

A Biotic indices **159**

B Synthetic and semi-synthetic datasets **167**
B.1 Datasets for assessment of feature selection . . . 167
B.2 Datasets for assessment of normalization . . . 169
B.3 Datasets for assessment of imputation of missing values . . . 171
B.4 Datasets for assessment of outliers processing . . . 173
B.5 Semi-synthetic datasets . . . 174

Bibliography **175**

List of Figures

2.1 General water quality monitoring workflow. 9

2.2 Basic chromatographic process. (a) The different components (C: glass column, SP: stationary phase, MP: mobile phase, S: sample); (b) introduction of the sample; (c) elution of constituents of the mixture; (d) recovering of compounds after separation. Figure taken from Rouessac and Rouessac (2008). 14

2.3 Schematic representation of chromatographic instrumentation 14

2.4 Structures of missingness: (a) univariate pattern, (b) monotone pattern, and (c) artibtrary pattern. Rows correspond to observations and columns to variables (Cottrell et al., 2009). 26

2.5 Representation of (a) missing completely at random (MCAR), (b) missing at random (MAR), and (c) missing not at random (MNAR) in a univariate missing-data pattern. X represents completely observed variables, Y denotes partly missing variable, Z represents the component of the causes of missingness unrelated to X and Y, and R represents the missingness (Schafer and Graham, 2002). 28

3.1 Sampling sites in the Tula river (Hidalgo state, Mexico) 43

3.2 Sampling sites in the Tamazula, Humaya, and Culiacan rivers (Sinaloa state, Mexico). 45

3.3 Flow diagram of the SPE method for analysis of OCPs in water. 50

3.4 Example of macroinvertebrates counting process to determine the number of runs in a sample. Each form represents a different organism. The number of runs is used to compute the Seqential Compraison index (SCI). 56

4.1 General procedure for assessment of data preprocessing procedures on statistical results. 61

4.2 Flow diagram of the generation of synthetic datasets for the assessment of preprocessing procedures on statistical results. FT denotes highly correlated data, AL denotes non-normalized data, NA stands for not available/missing values and OUT denotes outlying data. 69

4.3 Flow diagram of the generation of semi-synthetic datasets for the assessment of preprocessing procedures on statistical results. 73

4.4 Extract of the general workflow for assessment of data preprocessing procedures. Figure shows the workflow diagram of data preprocessing assessment. 75

4.5 Experimental procedure for assessment of imputation methods. 76

4.6 Experimental procedure for assessment of outlier detection and processing. 77

4.7 NRMSE results of the imputation of missing values using the Hot-deck (hd), IRMI (ir), k-NN (kn) and Mice (mi) imputation methods on datasets N21 (A), N600 (B), N4000 (C) and N20000 (D). 78

4.8 Precision and detection rate results of outliers detection methods. Adjusted quantile (adj.quan), Inter Quartile Range (iqr), Local Outlier Factor (lof) and Principal Components decomposition approach (pcout). 79

4.9 Experimental procedure for assessment of the impact of data preprocessing procedures on statistical analysis results. 80

4.10 Experimental procedure for assessment of the impact of data preprocessing procedures on results from regression analysis. 85

4.11 RMSE results of the analysis of regression after feature selection processing on synthetic datasets. Linear correlation-based feature selection (FI), Correlation-based Feature selection (HC) and Wrapper Subset evaluator (WR). 88

4.12 RMSE results of the analysis of regression after feature selection processing on semi-synthetic datasets. Linear correlation-based feature selection (FI), Correlation-based Feature selection (HC) and Wrapper Subset evaluator (WR). 88

4.13 RMSE results of the analysis of regression after normalization on synthetic datasets. Decimal scale noramlization (DS), Logistic sigmoidal normalization (SS) and Sigmoidal normalization using the hyperbolic tangent function (SM). 89

4.14 RMSE results of the analysis of regression after normalization on semi-synthetic datasets. Decimal scale normalization (DS), Logistic sigmoidal normalization (SS) and Sigmoidal normalization using the hyperbolic tangent function (SM). 89

4.15 Preprocessing errors of RMSE of the analysis of regression after imputation of missing values on synthetic dataset. Hot-deck (HD), Iterative Stepwise Regression Imputation (IR), K-Nearest Neighbour (KN) and Multiple Imputation by Chained Equations (MI) 90

4.16 Preprocessing errors of RMSE of the analysis of regression after imputation of missing values on semi-synthetic datasets. Hot-deck (HD), Iterative Stepwise Regression Imputation (IR), K-Nearest Neighbour (KN) and Multiple Imputation by Chained Equations (MI) 91

4.17 Preprocessing errors of RMSE of the analysis of regression after outlier detection followed by imputation of outlying data on synthetic datasets. Detection methods: Inter Quartile Range (1), Adjusted Quantile (2), PCOUT (3) and LOF (4). Imputation methods: Hot-deck (HD), Iterative Stepwise Regression Imputation (IR), K-Nearest Neighbour (KN) and Multiple Imputation by Chained Equations (MI). 93

4.18 Preprocessing errors of RMSE of the analysis of regression after outlier detection followed by imputation of outlying data on semi-synthetic datasets. Detection methods: Inter Quartile Range (1), Adjusted Quantile (2), PCOUT (3) and LOF (4). Imputation methods: Hot-deck (HD), Iterative Stepwise Regression Imputation (IR), K-Nearest Neighbour (KN) and Multiple Imputation by Chained Equations (MI). 94

4.19 Experimental procedure for assessment of the impact of data preprocessing procedures on results from classification. 95

4.20 Accuracy and Kappa results of the analysis of classification after feature selection on synthetic datasets. Linear Correlation-based Feature Selection (FI), Correlation-based Feature selection (HC) and Wrapper Subset evaluator (WR). 97

4.21 Accuracy and Kappa results of the analysis of classification after feature selection on semi-synthetic datasets. Linear Correlation-based Feature Selection (FI), Correlation-based Feature selection (HC) and Wrapper Subset evaluator (WR). 98

4.22 Accuracy and Kappa results of the analysis of classification after normalization on synthetic datasets. Min-Max normalization (MM), Decimal scale normalizaton (DS) and Z-score normalization (ZS). 99

4.23 Accuracy and Kappa results of the analysis of classification after normalization on semi-synthetic datasets. Min-Max normalization (MM), Decimal scale normalizaton (DS) and Z-score normalization (ZS). 100

4.24 Accuracy and Kappa preprocessing error from the analysis of classification after imputation of missing data on synthetic datasets. Imputation methods: Hot-deck (HD), Iterative Stepwise Regression Imputation (IR), K-Nearest Neighbour (KN) and Multiple Imputation by Chained Equations (MI). 102

4.25 Accuracy and Kappa preprocessing error from the analysis of classification after imputation of missing data on semi-synthetic datasets. Imputation methods: Hot-deck (HD), Iterative Stepwise Regression Imputation (IR), K-Nearest Neighbour (KN) and Multiple Imputation by Chained Equations (MI). 103

4.26 Preprocessing errors of Accuracy and Kappa of the analysis of classification after outlier detection followed by imputation of outlying data on synthetic datasets. Detection methods: Inter Quartile Range (1), Adjusted Quantile (2), PCOUT (3) and LOF (4). Imputation methods: Hot-deck (HD), Iterative Stepwise Regression Imputation (IR), K-Nearest Neighbour (KN) and Multiple Imputation by Chained Equations (MI). 104

4.27 Preprocessing errors of Accuracy and Kappa of the analysis of classification after outlier detection followed by imputation of outlying data on semi-synthetic datasets. Detection methods: Inter Quartile Range (1), Adjusted Quantile (2), PCOUT (3) and LOF (4). Imputation methods: Hot-deck (HD), Iterative Stepwise Regression Imputation (IR), K-Nearest Neighbour (KN) and Multiple Imputation by Chained Equations (MI). 105

4.28 Experimental procedure for assessment of the impact of data preprocessing procedures on results from clustering. 106

4.29 Rand index, Adjusted Rand index, Jaccard index and Cophenetic coefficient of Clustering methods after feature selection processing on synthetic datasets. Feature selection methods: Correlation-based Feature Selection (HC), Linear Correlation-based Feature Selection (FI) and Wrapper Subset Evaluator (WR). 108

4.30 Rand index, Adjusted Rand index, Jaccard index and Cophenetic coefficient of Clustering methods after feature selection processing on semi-synthetic datasets. Feature selection methods: Correlation-based Feature Selection (HC), Linear Correlation-based Feature Selection (FI) and Wrapper Subset Evaluator (WR). 108

4.31 Rand index, Adjusted Rand index, Jaccard index and Cophenetic coefficient of Clustering methods after imputation of missing data on synthetic datasets. Imputation methods: Hot-deck (HD), Iterative Stepwise Regression Imputation (IR), K-Nearest Neighbour (KN) and Multiple Imputation by Chained Equations (MI). 109

4.32 Rand index, Adjusted Rand index, Jaccard index and Cophenetic coefficient of Clustering methods after imputation of missing data on semi-synthetic datasets. Imputation methods: Hot-deck (HD), Iterative Stepwise Regression Imputation (IR), K-Nearest Neighbour (KN) and Multiple Imputation by Chained Equations (MI). 110

4.33 Rand index, Adjusted Rand index, Jaccard index and Cophenetic coefficient of Clustering methods after outlier detection followed by imputation of outlying data on synthetic datasets. Detection methods: Inter Quartile Range (1), Adjusted Quantile (2), PCOUT (3), and LOF (4). Imputation methods: Hot-deck (HD), Iterative Stepwise Regression Imputation (IR), K-Nearest Neighbour (KN) and Multiple Imputation by Chained Equations (MI). 112

4.34 Rand index, Adjusted Rand index, Jaccard index and Cophenetic coefficient of Clustering methods after outlier detection followed by imputation of outlying data on semi-synthetic datasets. Detection methods: Inter Quartile Range (1), Adjusted Quantile (2), PCOUT (3), and LOF (4). Imputation methods: Hot-deck (HD), Iterative Stepwise Regression Imputation (IR), K-Nearest Neighbour (KN) and Multiple Imputation by Chained Equations (MI). 113

4.35 Results of Naïves Bayesian classification on datasets preprocessed by: (A) normalization followed by imputation of missing values and (B) imputation of missing values followed by normalization. Min-max normalization (MM).Imputation methods: Hot-deck (HD), IRMI (IR), K-NN (KN) and MICE (MI). 116

5.1 Scientific workflow of EvDA 122

5.2 Screenshot of EvDA application. Display of the Data input tab. 124

5.3 Screenshot of EvDA application. Display of the Data inspection tab. 125

5.4 Screenshot of EvDA application. Display of the Data Preprocessing tab. . . 126

5.5 Screenshot of EvDA application. Example of display of the analysis by K-means clustering. 127

5.6 Google Earth image of the sampling sites C1 to C11 situated on the rivers Tamazula, Humaya and Culiacan. Distance between agricultural areas and the sampling sites are defined by a red line, sampling sites are highligted by a yellow icon. *Source:* "Culiacan." 24°48'00"N 107°23'00"O. **Google Earth** ©. August 5, 2016. November 22, 2016. 132

5.7 Concentrations of Pesticides on sites C1 to C11 situated on the rivers Tamazula, Humaya, and Culiacan. 133

5.8 Lower level correlation matrix of the Mexican dataset. Name of variables are abbreviated, see Table 5.1 for details. 134

5.9 PCA biplot of the two principal components for the macro-pollutants (a), micro-pollutants (b), and metals (c). Biomonitoring metrics are coloured in blue. 135

7.1 Sites de prélèvement dans la rivière Tula (État d'Hidalgo, Mexique). 148

7.2 Sites de prélèvement des rivières Tamazula, Humaya et Culiacn (État de Sinaloa, Mexique). 149

7.3 Procédure générale d'évaluation des procédures de pré-traitement des données sur les résultats statistiques. 153

7.4 Capture d'écran de l'application EvDA. Affichage de l'onglet inspection de données. 157

List of Tables

2.1 Water quality variables. 10

2.2 List of some Emerging Pollutants that are frequently found in aquatic environment. 12

2.3 Measures of richness (Barbour et al., 1999) (Resh and Jackson, 1993). 19

2.4 Enumeration metrics. 20

2.5 Similarity and Diversity indices. 21

2.6 Biotic indices. 21

3.1 Location of sampling sites of the Tula river. 43

3.2 Location of sampling sites of Tamazula, Humaya and Culiacan rivers. 44

3.3 Example of data from the physochemical and chemical values obtained from the analysis of water samples of the rivers Tula. Tamazula. Humaya and Culiacan. 47

3.4 Example of the data from the concentration of arsenic and heavy metals on water samples of the rivers Tula, Tamazula, Humaya and Culiacan. 47

3.5 SPE-GC-ECD limit of detection (LOD), limit of quantification (LOQ) and recovery values for water analysis. 48

3.6 SPME-GC-MS limit of detection (LOD), limit of quantification (LOQ) and recovery values for water analysis. 49

3.7 Example of data from the analysis of organochlorine pesticides on water samples from the Tula river. Only ten out of the six-teen compounds are shown. 52

3.8 Example of data from the analysis of PPCPs on water samples from the Tula river. 52

3.9 Biomonitoring metrics for the biological assessment of Mexican rivers (Serrano Balderas et al., 2016). 54

3.10 Example of data from the biomonitoring metrics computed for the rivers Tula, Tamazula, Humaya and Culican. 57

4.1 Synthetic datasets used to assess feature selection methods. Only the μ and σ^2 values of the first ten variables are shown. Highly correlated variables are denoted as Y_k. 71

4.2 Synthetic datasets used to assess normalization methods. Only the μ and σ^2 values of the first ten variables are shown. 71

4.3 Synthetic datasets used to assess imputation methods. Only the μ and σ^2 values of the first ten variables are shown. 72

4.4 Synthetic datasets used to assess outlying preprocessing. Only the μ and σ^2 values of the first ten variables are shown. 72

4.5 Contingency table obtained from the results of two classifiers. 82

4.6 Extract of summary results from the study of data preprocessing procedures on regression, classification and clustering methods. FS (Feature Selection), N (Normalization), I (Imputation method), O (Outlier processing). 117

5.1 List of variables that were used for the assessment of water quality of the Mexican rivers Tula, Humaya, Tamazual and Culiacan. 130

5.2 Summary of results about the linear model for the regression of distance between sampling sites and agricultural areas on the concentration of pesticides of the Mexican dataset. 131

5.3 Results from the linear regression analysis on the Mexican dataset. Variables used for the analysis were chosen using the Linear correlation-based feature selection method. 132

5.4 Results of the linear correlation-based feature selection on the data subsets *macro*, *micro* and *bio*. 136

5.5 Results of the linear regression on the macro, micro and bio data subsets before and after feature selection. 137

7.1 Emplacement des sites d'échantillonnage de la rivière Tula 148

7.2 Emplacement des sites d'échantillonnage des rivières Tamazula, Humaya et Culiacan. 149

A.1 Trent Biotic Index (TBI). Key groups for the estimation of the Trent River Board Biotic Index and Biotic Index values related to the total number of groups present in a sample (Table taken from Metcalfe, 1989). 159

A.2 Table to calculate Extended Biotic Index (EBI) values and conversion table to transform EBI values into Quality Classes (Ghetti, 1997). SU: Number of Systematic Units observed of the taxonomic group. 160

A.3 Benthic macroinvertebrates clased according to Beck's biotic Index (BBI) Classes (Beck, 1955). 161

A.4 Evaluation of water quality using the Family Biotic Index (FBI) and the tolerance values for families of stream arthropods (Hilsenhoff, 1988). . . . 162

A.5 Biomonitoring Working Party Score System (BMWP) (National Water Council; 1981) 164

A.6 Biological Monitoring Working Party (BMWP) and Average Score Per Taxon (ASPT) scores and their related quality index (Armitage et al., 1983; Friedrich et al., 1996; National Water Council, 1981). 165

A.7 Functional Feeding Groups (FFG): Categorization and food resources. Coarse Particulate Organic Matter (CPOM); Fine Particulate Organic Matter (FPOM) (Merrit et al., 2008). 165

A.8 Scoring criteria of the Macroinvertebrate-based Index of Biotic Integrity (IBI designed for west-central Mexico streams) (Weigel et al., 2002). 166

A.9 Macroinvertebrates-based Index of Biotic Integrity (IBI) quality values and their related biological responses to environmental conditions (Weigel et al., 2002). 166

B.1 N21 synthetic dataset ($n = 21$, $p = 8$) used to assess feature selection methods. The μ and σ^2 values of all variables are shown. Highly correlated variables are denoted as Y_k 167

B.2 N600 synthetic dataset ($n = 600$, $p = 30$) used to assess feature selection methods. The μ and σ^2 values of all variables are shown. Highly correlated variables are denoted as Y_k 167

B.3 N4000 synthetic dataset ($n = 4000$, $p = 53$) used to assess feature selection methods. The μ and σ^2 values of all variables are shown. Highly correlated variables are denoted as Y_k. 167

B.4 N20000 synthetic dataset ($n = 20000$, $p = 98$) used to assess feature selection methods. The μ and σ^2 values of all variables are shown. Highly correlated variables are denoted as Y_k. 168

B.5 N21 synthetic dataset ($n = 21$, $p = 8$) used to assess normalization methods. The μ and σ^2 values of all variables are shown. 169

B.6 N600 synthetic dataset ($n = 600$, $p = 30$) used to assess normalization methods. The μ and σ^2 values of all variables are shown. 169

B.7 N4000 synthetic dataset ($n = 4000$, $p = 53$) used to assess normalization methods. The μ and σ^2 values of all variables are shown. 169

B.8 N20000 synthetic dataset ($n = 20000$, $p = 98$) used to assess normalization methods. The μ and σ^2 values of all variables are shown. 169

B.9 N21 synthetic dataset ($n = 21$, $p = 8$) used to assess imputation methods. The μ and σ^2 values of all variables are shown. 171

B.10 N600 synthetic dataset ($n = 600$, $p = 30$) used to assess imputation methods. The μ and σ^2 values of all variables are shown. 171

B.11 N4000 synthetic dataset ($n = 4000$, $p = 53$) used to assess imputation methods. The μ and σ^2 values of all variables are shown. 171

B.12 N20000 synthetic dataset ($n = 20000$, $p = 98$) used to assess imputation methods. The μ and σ^2 values of all variables are shown. 172

B.13 N21 synthetic dataset ($n = 21$, $p = 8$) used to assess outliers processing methods. The μ and σ^2 values of all variables are shown. 173

B.14 N600 synthetic dataset ($n = 600$, $p = 30$) used to assess outliers processing methods. The μ and σ^2 values of all variables are shown. 173

B.15 N4000 synthetic dataset ($n = 4000$, $p = 53$) used to assess outliers processing methods. The μ and σ^2 values of all variables are shown. 173

B.16 FQ16 semi-synthetic dataset ($n = 16$, $p = 13$). The μ and σ^2 values of all variables are shown. 174

B.17 FQ1000 semi-synthetic dataset ($n = 1504$, $p = 26$). The μ and σ^2 values of all variables are shown. 174

B.18 FQ7000 semi-synthetic dataset ($n = 7520$, $p = 26$). The μ and σ^2 values of all variables are shown. 174

Nomenclature

Acronyms

AAS Atomic Absorption Spectroscopy

ANN Artificial Neural Networks

BBI Beck Biotic Index

BMWP Biological Monitoring Working Party

BMWP-ASPT BMWP - Average Score Per Taxon

BOD Biochemical Oxygen Demand

COD Chemical Oxygen Demand

DM Data Management

DO Dissolved Oxygen

EBI Extended Biotic Index

EDTA Ethylenediaminetetraacetic acid

EI Environmental Informatics

EMB Expectation-Maximization with bootstrapping

EP Emerging Pollutant

EPT Ephemeroptera Plecoptera Trichoptera

FFG Functional Feeding Groups

GC-ECD Gas Chromatography-Electron Capture Detector

GC-MS Gas Chromatography-Mass Spectrometry

HBI Hilsenhoff's Biotic Index

HPLC High Performance Liquid Chromatography

IBI Index of Biotic Integrity

IBI-Mex Index of Biotic Integrity for west-central Mexico

ICP-MS Inductively Coupled Plasma Mass Spectrometry

IRD Institut de Recherche pour le Développement

IRMI Iterative Robust Model-based Imputation

IT Information Technology

LC-MS Liquid Chromatography-Mass Spectrometry

LC-MS-MS Liquid Chromatography-tandem Mass Spectrometry

LC-PAD Liquid Chromatography with a Photodiode Array Detector

MAR Missing at Random

MCAR Missing Completely at Random

MICE Multiple Imputation by Chained Equations

MNAR Missing Not at Random

OCPs Organochlorine Pesticides

PBDEs PolyBrominated DiphenylEthers

PPCPs Pharmaceutical and Personal Care Products

SCI Sequential Comparison Index

SEA Strategic Environmental Assessment

SFC-DAD Supercritical-Fluid Chromatography - Diode-Array Detections

SPE Solid Phase Extraction

SPME-GC-MS Solid Phase Micro Extraction-Gas Chromatography-Mass Spectrometry

TBI Trent Biotic Index

UNAM Universidad Nacional Autónoma de México

WHO World Health Organization

WQ Water Quality

WQA Water Quality Assessment

WQMP Water Quality Monitoring Program

Symbols

B Some definition of B

IQR Inter Quartil Range

As Arsenic

B Boron

Ca Calcium

Cd Cadmium

Cl Chloride

CO_3 Carbonate

Cu Copper

Eh Reduction potential

Fe Iron

F Fluoride

HCl Hydrogen chloride

HCO_3 Bicarbonate

K Potassium

Mg Magnesium

Mn Manganese

Na Sodium

NO_3 Nitrate

Pb Lead

SiO_2 Silicon dioxide

SO_4 Sulfate

Zn Zinc

CHAPTER 1

Introduction

According to UNESCO, rivers, lakes and coastal areas of developing countries are polluted by more than 80% of untreated sewage. In 2012, the World Health Organization (WHO) estimated that 871,000 human deaths were caused by contaminated drinking water and soil (WHO, 2016). Regarding freshwater biodiversity, globally, more than 20% of vertebrates species have been threatened by pollution (Assessment, 2005; Vörösmarty et al., 2010), 32% of the world's amphibian species are threatened with extinction (Dudgeon et al., 2006) and it is estimated that rates of species loss have a tendency to increase with higher values in tropical latitudes (Groombridge and Jenkins, 2000).

Faced with the need to reduce water pollution, intensive studies to monitor the quality of water are continuously implemented. Monitoring activities consist in measuring various water quality components at many locations during numerous time periods. Though data from environmental surveys may be prone to have different anomalies (i.e., incomplete, inconsistent, inaccurate or outlying data). Anomalies on data are ubiquitous, they arise due to experimental problems, human errors or system failures.

Bad data quality, may be significant costly (Haug et al., 2011) and have considerable consequences when assessing environmental ecosystems (Wahlin and Grimvall, 2008). To generate quality data and reduce data anomalies, it is necessary both: to acquire quality data and preprocess data (Han and Kamber, 2000; Berti-Équille, 2007a).

The acquisition of quality data can be obtained by following standardized sampling and analytical protocols, and using advanced analytical tools (i.e., Inductively Coupled Plasma Mass Spectrometry (ICP-MS) , Gas Chromatography Mass Spectrometry (GC-MS), Liquid Chromatography Mass Spectrometry (LC-MS) or Liquid Chromatography tandem Mass Spectrometry (LC-MS-MS)) (UNEP, 2004). However, in developing countries access to advanced analytical tools may be limited or standardized protocols are difficult to implement. It is thus, necessary to adapt tools that can be easy and cheap to use without compromising the quality of data.

Concerning data preprocessing, it consists of all actions necessary to generate quality data. These actions include but are not limited to:

1. Cleaning data (i.e., removing outliers) and inconsistencies;
2. Resolving data conflicts (i.e., settling discrepancy);
3. Recovering incomplete data (i.e, filling missing values);
4. Reducing data (i.e., selecting relevant attributes) or creating new aggregate data.

Although anomaly detection has been extensively studied in the recent years, few works have been published and specifically applied in the domain of environmental science. Considering that anomalies on data may appear all along an environmental study, it is essential to identify, control (Berti-Équille, 2007a), and if necessary design well-defined data acquisition and data preprocessing protocols in order to guarantee accurate results. In this work we focused in both data acquisition and data preprocessing, specifically for the analysis of environmental data from a developing country: Mexico.

1.1 Environmental informatics

Pollution problems have boosted environmental research activities going from general to more detailed measurements and involving various disciplines such as Geology, Chemistry, Ecology, Biology or Statistics. The large and complex production of environmental research findings urge the need to use tools to analyse data. Environmental Informatics science emerged from this need.

Environmental Informatics (EI) is a relatively new multidisciplinary, the first EI community was founded in the 80's (Pillmann et al., 2006). It aims to carry out research and develop IT tools focused on environmental sciences related to the creation, collection, processing, modelling, analysis and diffusion of environmental data, information and findings. Environmental Informatics incites an aperture of the environmental sciences to the Computer Science in order to exploit the concepts, methods, and tools to advance knowledge in Environmental Sciences.

The work presented here is positioned on Environmental Informatics since it has a multidisciplinary approach between Environmental Chemistry, Hydrobiology, Statistics, Data science and Computer Science. The study was focused on two main challenges: acquisition of quality data and definition of adequate data preprocessing procedures to finally analyse data.

The importance of this work is highlighted by the definition of a generic, flexible and extensible methodological framework dedicated to data acquisition, specifically water quality data, and data preprocessing practices for analysis of environmental data.

This project was developed in collaboration with the Institute of Geophysics at The National Autonomous University of Mexico (UNAM) in Mexico and the Research Institute for Development (IRD) in France within a scientific collaboration between the two countries for the environmental preservation. In order to provide methods and tools for the acquisition of data, different activities in environmental chemistry and hydrobiology were designed. The activities related to environmental chemistry were developed at the Institute of Geophysics at UNAM in Mexico and those related to hydrobiology were developed at ENGEES in Strasbourg, France. The activities in the Statistics and Computer Science that define the methods to preprocess and analyse data were developed at IRD in Montpellier, France.

1.2 Motivations

During environmental analysis, it is essential to only consider quality data that provide valid information and discard results from deformed or distrust data. It is clear that to provide accurate predictive and analytical results we need good quality data. In data science, for instance, preprocessing procedures are proposed to reduce anomalies in data.

In fact, a vast number of data preprocessing procedures are available rending the selection of an optimal procedure a very arduous task.

There are many inherent challenges in this respect, for instance:

- Conventional approaches treat each data anomaly as isolated cases however, data anomalies co-occur in data. Therefore it is necessary to develop preprocessing procedures that deal with multiple data anomalies within a dataset;
- The ordering in which preprocessing procedures should be executed is under studied. It is necessary to define a data preprocessing ordering to produce the least biased results;
- The impact of data preprocessing on statistical analysis results has not been widely studied, it is necessary to conduct studies in this area to produce accurate results;
- An extensive work that provides information about the accuracy on the application of preprocessing procedures is needed.

Data preprocessing, will definitely improve the quality of data. But, good data acquisition practices are also necessary to prevent data anomalies. An important problem on the acquisition of quality environmental data is the lack of standardized sampling and analytical protocols particularly on laboratories where the access to advanced analytical instruments is limited. Thus, it is necessary to adapt analytical procedures that are cost-effective and easy to implement in order to get quality data.

We are interested in developing a new approach that combines both; good data acquisition and data preprocessing practices. Our main motivation is to provide tools and methodological approaches to the scientific community for the acquisition, preprocessing and analysis of environmental data. In this manner we were concentrated in two main challenges: acquire quality data and preprocess data. Our purposes are: (1) to provide methodological approaches and tools for the acquisition of water quality data and (2) to provide a general overview of the advantages and disadvantages of preprocessing procedures on statistical results.

Concerning the acquisition of data we focused on the development of tools in Environmental Chemistry and Hydrobiology. We were especially interested in: (1) water pollution problems caused by heavy metals, pesticides, pharmaceutical and personal care products, and (2) the use of biomonitoring metrics using macroinvertebrates for water quality assessment of Mexican rivers.
In Mexico, the monitoring of these type of pollutants is limited regardless their toxic effects on ecosystems and human health. Such limitation is due to the restricted access to advanced analytical instruments for their quantification. By adapting analytical techniques we will be able to: quantify pesticides, pharmaceuticals and personal care products in water; acquire quality data by reducing the presence of anomalies such as missing values, censored or outlying data; and identify sources of pollution that may affect ecosystems and human health.
Use of biomonitoring metrics based on macroinvertebrates is an interesting approach for the assessment of water quality in rivers. It is a low-cost and easy to implement tool that provides valuable information about the ecological state of aquatic ecosystems. Nonetheless, in Mexico biomonitoring metrics are scarcely used. In fact, well-described sampling and analytical protocols for their implementation do not exist yet. We were interested in developing sampling and analytical protocol to use biomonitoring metrics based on macroinvertebrates for the water quality assessment of Mexican rivers.

Regarding data preprocessing, we were interested in highlighting the importance of

having quality data by implementing appropriate data preprocessing procedures. In previous works (Serrano Balderas, 2012; Serrano Balderas, 2013) we observed that the quality of data has a significant impact on analytical results. These works conducted us to perform a study where we demonstrated the importance of having quality data in an environmental application framework (Berrahou et al., 2015). Data quality is undoubtedly important in data analysis, we are concerned now about the procedures to get quality data namely, preprocessing procedures.

1.3 Objectives

Our principal objective is to provide to the scientific community the methodological approaches and tools for the acquisition, preprocessing and analysis of environmental data by guaranteeing the quality of data and analysis results. We aim at applying our methodological approaches for a specific case: water quality assessment of four Mexican rivers (Tula, Tamazula, Humaya and Culiacan).

Our purpose on providing an integrated approach that combines both; good data acquisition and data preprocessing practices, is to control the entire pipeline this is, from the production and to the analysis of data. To achieve our aim, we have defined specific objectives related to the three areas of our study. The objectives are:

Environmental Chemistry

- Acquire data by deploying reliable and low-cost methods of analysis for the quantification of organochlorine pesticides, pharmaceuticals and personal care products in water. These methods aimed at acquiring quality data for water quality assessment in four Mexican rivers (Tula, Humaya, Tamazula and Culiacan);
- Specify the sampling and methodological protocols for the analysis of organochlorine pesticides, pharmaceuticals and personal care products;
- Conduct the sampling campaign and analytical procedures for the analysis of water samples from the four Mexican rivers (Tula, Humaya, Tamazula and Culiacan).

Hydrobiology

- Acquire data by defining a methodological approach using macroinvertebrates-based biomonitoring metrics as new complementary tools for water quality assessment of Mexican rivers;
- Conduct the sampling campaign and analytical procedures for the acquisition of hydrobiological data.

Data science

- Define a methodological approach for the selection of data preprocessing procedures to treat the most common data anomalies and data problems (missing values, outliers, feature selection and normalization);
- Evaluate the impact of data preprocessing procedures on subsequent statistical analysis;
- Determine the most appropriate data preprocessing procedures to get the less biased analytical results;
- Specify the procedures to preprocess and analyse data that are necessary to ensure the reliability of results on environmental studies in general and for water quality data analysis in particular.

1.4 Outline

The document is organized as follows:
Chapter 2 presents an overview of related work including: a) the acquisition of water quality data for physico-chemical, chemical and biological parameters and b) preprocessing and analysis of environmental data. In Chapter 3 we describe our methodological approaches for the acquisition of chemical and biological data. We also describe the study sites and the methods followed to collect and analyse the samples. Our methodological approach for the assessment of data preprocessing procedures is detailed in Chapter 4. In Chapter 5 we present *EvDa*, a Shiny/R application designed to preprocess and analyse environmental data. Results of *EvDa* application to water quality data from the Mexican rivers: Tula Tamazula, Humaya and Culiacan are also described in Chapter 5.
Finally, conclusions and future work associated to our contributions are given in Chapter 6.

CHAPTER 2

Environmental data analysis: the case of water quality assessment with a multidisciplinary survey

Contents

2.1 Introduction 8

2.2 Data collection in physico-chemical and chemical water quality assessment 8

2.2.1 Water quality assessment of rivers 8

2.2.2 Collection of data about emerging pollutants in rivers 11

2.2.3 Chemical analysis in water 12

2.3 Data collection from biomonitoring in water quality assessment 17

2.3.1 Biological organisms in biomonitoring 17

2.3.2 Ecological assessment of rivers using macroinvertebrates 18

2.3.3 Biomonitoring data: general characteristics, common uncertainties and anomalies 23

2.4 Preprocessing and analysis of environmental data 25

2.4.1 Data anomalies and their detection 26

2.4.2 Dealing with data anomalies and main preprocessing procedures 31

2.4.3 Impact of data preprocessing procedures on statistical analysis results 37

2.5 Summary 38

2.1 Introduction

The main tasks presented in this manuscript are: data acquisition, data preprocessing and data analysis. Concerning data acquisition we aimed at developing and applying tools in environmental chemistry and hydrobiology to acquire water quality data. For data preprocessing, we aimed at developing a methodological framework to prepare the data and analyse them. As a consequence our bibliographic study ranges from environmental chemistry (Section 2.2), hydrobiology (Section 2.3), computer science and statistics (Section 2.4).

2.2 Data collection in physico-chemical and chemical water quality assessment

Collection of water quality data involves a set of tasks that include but are not limited to: design of experiment, selection of sites of study, sample collection, development and/or adaptation of analytical methods for analysis of pollutants and analysis of data (Bartram and Ballance, 1996). For a better understanding of the characteristics of water quality data and an appropriate subsequent analysis, we provide a general overview of methods and tasks that are implemented to get water quality data. We focused mainly on methods to acquire data related to pesticides, pharmaceuticals and beauty products as pollutants.

2.2.1 Water quality assessment of rivers

Water quality monitoring programs (WQMP) are conducted, in general, to get information about the characteristics of water resources, identify pollution problems, evaluate and describe water management actions (WMO, 2013). Assessment of water quality is commonly carried out following standardized methods and protocols. Globally, the WHO have proposed a practical guide for designing and implementing freshwater quality studies (Bartram and Ballance, 1996). Today, each country have established their own protocols, water quality standard and permissible limits.

Water Quality Monitoring Programs refer to the acquisition of quantitative and representative information on the existing conditions (physical, chemical and biological characteristics) of a water body. They are designed following, in general, twelve steps according to the diagram shown in Figure 2.1.

Identification and definition of objectives are the first two steps in monitoring programs (Smith et al., 2014). Then the necessary experiments are designed, these need to be structured to meet specific needs according to the objectives defined. The experiments are composed of subsequent steps including: selection of scale, determination of frequency, location and design of stations, selection of variables, selection of sample type, definition of sample collection, definition of analysis methods and definition of land use monitoring.

Scale is defined whether at local, regional or national level. Frequency of sampling varies according to the purpose of the monitoring, the characteristics of the water body and the importance of the sampling station location. In general, monthly and seasonal samplings are made for characterizing water quality over long time periods (e.g., over a year) whereas weekly samplings are done for control purposes. On cases where significant differences are suspected or detected, samples may be collected on a continuous basis.

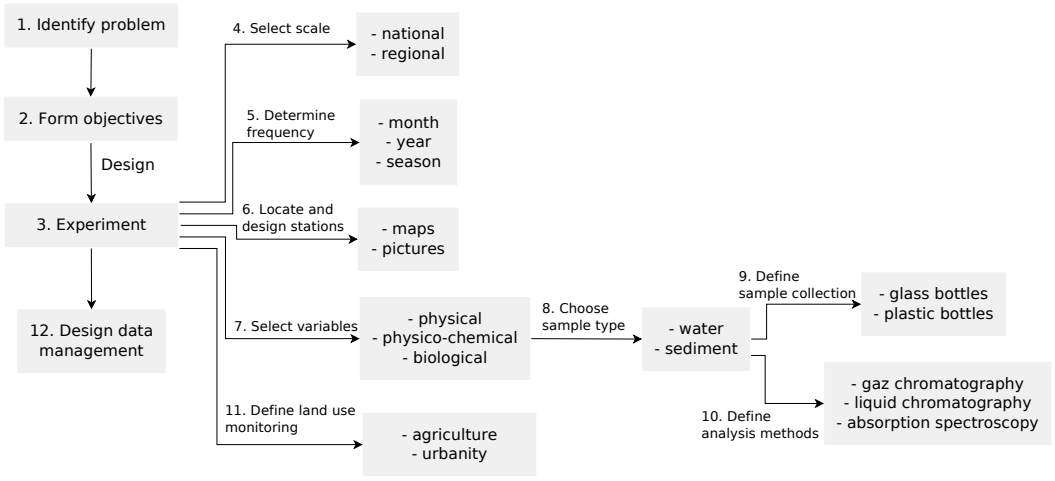


Figure 2.1 – General water quality monitoring workflow.

Traditionally, the design of stations is done subdividing the locations into macro-location and micro-locations. Macro-locations are river reaches sampled within a watershed and micro-locations are sampling points belonging to a macro-location that represent specific points (e.g., discharge of a river, source of pollution) (Harmancioglu et al., 1999). Today, design of stations is done using remote sensing techniques. By integrating remotely sensed data, GPS or GIS technologies natural resource managers are able to geographically locate a monitoring site (Ritchie et al., 2003).

Selection of variables depend directly to the objectives established on the monitoring program and their importance vary with the type of water body and uses. Commonly the variables that characterize water quality are classed into: general water quality parameters, dissolved salts, organic matter, microbial pollution, nutrients and inorganic variables. The development of new analytical technologies have allowed the identification of chemical compounds that were not possible to identify with traditional techniques. These compounds are classed as emergent pollutants and are monitored whether in surface waters (Geissen et al., 2015b) or ground waters (Lamastra et al., 2016). Such organic compounds include among others: pesticides, fungicides, oils, hydrocarbons, organic solvents, phenols, surfactants, pharmaceuticals, fats and hormones (c.f. Table 2.1).

In addition to the aforementioned variables, ecological indicators (e.g., fish, macroinvertebrates, macrophytes, diatoms, phytoplankton.) are integrated as part of biological monitoring. Though the use of ecological indicators is advantageous (c.f. Section 2.3.1), they only provide information about indirect causes for a given change therefore it is recommended to use them as complementary tools in WQ monitoring activities (Sponseller et al., 2001; Serrano Balderas et al., 2016).

Concerning sample type and sample collection, they depend on the variables to be measured (Pegram et al., 2013). Commonly; water, particular matter and living organisms are collected. Quality of water and particular matter are analysed by measuring physical and chemical parameters whereas living organisms can be used in different ways, for instance: as ecological indicators studying communities of organisms or as toxicological indicators by studying the physiology, morphology or chemical composition of tissues on specific organisms (Dolédec and Statzner, 2010).

Analytical methods are determined according to the type of sample, the quality parameter to be analysed, amount of sample and whether the sample is analysed in situ

Table 2.1 – Water quality variables.

General water quality	**Organic matter**	**Other Inorganic variables**
Water discharge/level[a]	Organic carbon (OC)[b]	Boron[b]
Total suspended solids[b]	Chemical Oxygen Demand (COD)[b]	Cyanide[b]
Temperature[d]	Biochemical Oxygen Demand (BOD)[b]	Heavy metals[b]
pH	Chlorophyll *a*	Arsenic and selenium[b]
Conductance[e]		Sulphide[b]
Redox potential (Eh)[f]	**Microbial pollution**	
Hardness[g]	Faecal coliforms[h]	**Emergent pollutants**
Disolved Oxygen (DO)[b]	Pathogens[h]	Oil and hydrocarbons[b]
Turbidity[k]	Bacteria[h]	Organic solvents[b]
Colour[i]		Phenols[b]
Odour[j]	**Nutrients**	Pesticides[c]
Taste[j]	Nitrate plus nitrite[b]	Surfactants[b]
Particle size	Ammonia[b]	Pharmaceuticals[c]
	Organic nitrogen[b]	Fats[b]
Dissolved salts	Phosphorus[b]	Hormones[b]
Calcium[b]	Silica[b]	Fungicides[b]
Magnesium[b]		
Sodium[b]		
Potassium[b]		
Chloride[b]		
Fluoride[b]		
Sulphate[b]		
Alkalinity[g]		

[a] Units are in $m^3 s^{-1}$.
[b] Units are in $mg L^{-1}$.
[c] Units are in $\mu g L^{-1}$.
[d] Units are in °C.
[e] Units are in $\mu S cm^{-1}$.
[f] Units are in mV.
[g] Units are in $mg/CaCO_3$.
[h] Units are in number/100 mL.
[i] Units are in $mg L^{-1}$ Pt/Co [Hazen units].
[j] Units are in dilution rate at stated temperature.
[k] Units are in FTU (Formazin Turbidity Units) or NTU (Nephelometric Turbidity Units) or JTU (Jackson Turbidity Units).

or at laboratory. Generally, water sampling operations include in situ measurements, sample pre-treatment, conservation, identification and transportation. There are numerous methods for field monitoring they range from simple devices to sophisticated equipment (WMO, 2013)(UNEP, 2004)(Meyers, 2000)(APHA/AWWA/WPCF, 2005). The selection of methods and devices to measure the selected variables depend on the availability of equipment and reagents, degree of expertise and the accuracy needed according to the objectives of the program.

Data management (DM) is an integral part of a monitoring program: it aims at converting data into information that will meet monitoring objectives. DM in WQMP includes analysis, distribution and storage of data. As an important part, data analysis allows to have a basic understanding of data and to determine data quality and quantity (EPA, 2006). A first insight to understand data can be obtained through summary statistics. Some of the characteristics that can be described include: normality, variability, symmetry of data distribution, missing values or outliers. This kind of information can be useful in order to define the suitability of data to support the intended use and, in most of the cases, to identify problems on data that may affect conclusions on the environmental assessment program (Helsel and Hirsch, 2002).

2.2.2 Collection of data about emerging pollutants in rivers

Today, there are numerous synthetic organic compounds that are used for industrial, agricultural or domestic purposes. They can enter to natural effluents through infiltration, draining, natural deposition or by direct discharge. Except for specific removal by wastewater treatment processes, they may be released into running waters at trace level concentrations (in the nanogram (ng) or microgram per liter ($\mu g l^{-1}$) range). Some trace pollutants are referred as Emerging Pollutants (EP). Emerging pollutants (EPs) are synthetic or naturally occurring chemicals that are rarely part of environmental monitoring practices but which enter the environment and cause known or suspected adverse ecological and (or) human health effects (Geissen et al., 2015a). Common wastewater processes are not able to treat EPs.

Though release of EPs have been occurred for a long time, the interest to include them in environmental assessment programs has increased over the last decades. Some of the EPs that are found in aquatic environment have been recognized as hazardous to ecosystems (von der Ohe et al., 2011). The persistence of EPs, their metabolites and transformation products in natural waters has increased environmental and human health concern, since little is known about their fate, behaviour and ecotoxicological effects (Petrie et al., 2015). The EPs most constantly studied in the freshwater environments include: industrials, pesticides, pharmaceuticals and personal care products (PPCPs) (Murray et al., 2010). In Table 2.2 we provide a list of some of these compounds.

Industrials are organic compounds used in manufacturing and production processes, pesticides are substances or mix of substances used to destroy, suppress or alter the life cycle of any pest[1] and PPCPs are compounds administrated internally or externally to the bodies of humans and domestic animals. PPCPs are part of the compounds of human health concern because they can exhibit multixenobiotic resistance and show resistance to microbial degradation (Smital et al., 2004).

As well as in a traditional water quality monitoring program (c.f. Subsection 2.2.1), collection of data for assessment of EPs includes twelve steps (c.f. Figure 2.1). Due

[1] http://www.epa.nsw.gov.au/pesticides/pestwhatrhow.htm

Table 2.2 – List of some Emerging Pollutants that are frequently found in aquatic environment.

Industrials	**Pesticides**	**PPCPs**
antioxidants	carbamates	analgesics
perfluorates	chloro-acetanilides	anti-epileptic drugs (AEDs)
phenols	chlorophenoxy acids	anticonvulsants
phthalates	organochlorines	lipid regulators
poly-brominated diphenylethers (PBDEs)	organophosphates	antimicrobials
triazoles	pyrethroids	polycyclic musks (PCMs)
	triazines	non-steroidal anti-inflammatory drugs (NSAIDs)
	diuron	synthetic hormones
	isoproturon	fragances
	mecoprop	insect repellant
	prometon	

Units are given in $\mu g\,L^{-1}$.

to the physico-chemical properties of EPs (e.g., adsorption, volatility, polarity), sample collection and analytical methods are the critical steps. Water quality data that includes EPs is numerical, volume of data may depend on the objectives and scale of study (i.e., local, regional, national) thus, it can vary from 5 to over thousand of individuals (or sampling sites) and from one to hundred of compounds.

Uncertainties may occur all along data collection and could be linked mainly to: field sampling problems, transportation of samples, bad sample manipulation or problems with analytical instrumentation. Data anomalies such as missing values, outliers and non-detected data are frequently found when EPs are included in a water quality monitoring program. Missing values and outlying data may occur by different reasons (e.g., difficulties on field sample collection, lost of sample, equipment failure, anomalous conditions). Non-detected concentrations or ND are frequently reported when a concentration of a compound in a sample is less than a specified limit of detection (LOD), the resulting left-censored data preclude subsequent statistical analysis (Clarke, 1998). For the case of missing values and outlying data different data preprocessing procedures could be implemented to improve the quality of data and facilitate statistical analysis (c.f. Subsection 2.4). Concerning non-detected data, use of advanced analytical techniques could be advantageous, since advanced techniques allow the detection of EPs at very low concentrations (µg and $ng\,l^{-1}$).

2.2.3 Chemical analysis in water

Samples collected for environmental assessment constitute a representative fragment of the site under study. Such sample contains the target specie called the *analyte*, which in general is mixed with a number of other compounds constituting the *matrix*. To identify or quantitatively determine an analyte, separation methods can be used. Such methods are achieved by differences in physical or chemical properties between the constituents of a mixture (Wilson et al., 2000).

There are different separation methods including: chemical precipitation, distillation, extraction by solvents, ion exchange, electrolysis, and chromatography (Skoog Douglas et al., 2014; Harvey, 2000).

- *Chemical precipitation* is based on the solubility of the compounds. It requires to have big differences of solubility between the analyte and the other compounds. The main drawback is that undesired co-precipitates can be formed making the separation process more complex and slow.
- *Distillation* is used to separate volatile analytes from non-volatile interferences. Big

differences of volatility are required to separate the target compound therefore, the use of this method is limited when compounds in a mixture have similar or very close volatility characteristics.

- *Extraction by solvents* is based on solubility differences between two immiscible liquids. This method involves the selective transfer of an analyte from one liquid phase to another by using a solvent with a polarity similar to the analyte. In water samples, non-polar solvents are used to extract the analyte from water.
  Compared to precipitation, the extraction method is less complex and faster however, large quantities of organic solvents are needed, emulsions may be formed and it is not suitable for thermally unstable compounds.
- *Ion exchange*. Water is composed of electrically charged atoms or molecules named *ions*. Ions can be charged positively or negatively, named *cations* or *anions* respectively. Ion exchange is used to separate undesired cations and anions within a water matrix such undesired cations are replaced with other similarly charged ions. The process occurs when a water matrix is in contact with an ion exchange resin which contains charged ions. The ions of the matrix are then replaced by the ions of the resin. Resins are composed of organic polymer chains which contain charged functional groups with either positive or negative charge. In a matrix containing multiple compounds the ion exchange separation will not be highly efficient for different reasons: 1) the sample need to be passed through different resins in order to separate the analyte from the matrix, causing a lost of sample and analyte; 2) it is necessary to know the physico-chemical characteristics of the compounds present in a matrix in order to use specific resins; and 3) resins composed of mixed polymer chains are indispensable but they can be costly and there is not guarantee of getting good separation results.
- *Electrolysis* occurs when an electric current is passed through a mixture containing ions. The electric current induce chemical reactions and separates compounds of the matrix. The process occurs in an electrolytic cell which is a device with positive and negative electrodes. The electrodes are then introduced in a solution containing charged ions. When the electric current is passed through the solution, cations travel to the electrode charged negatively (cathode) and are transformed to neutral molecules. While anions travel to the electrode charged positively (anode) to be transformed into neutral molecules. Separation by electrolysis is widely used in metallurgical processes.
- *Chromatography* is a separation process where substances are distributed between two phases named stationary and mobile phase. The mobile phase is a sample-free phase that is passed over a second sample-free phase that remains stationary (stationary phase). The chromatography is achieved by placing a sample into the stationary phase and putting it in contact with the mobile phase, as the mobile phase moves the components of the sample are partitioned between the mobile and the stationary phases.
  The basic chromatographic process can be described in the following steps (c.f. Figure 2.2):

  1. First, the stationary phase, which is a finely powdered solid, is placed into a vertical hollow glass tube. Then a sample matrix containing the components to be separated is placed at the top of the glass tube;
  2. Then, the mobile phase is added continuously to the glass tube. In this manner, the different constituents of the mixture are carried on along with the mobile phase. This process is known as elution;
  3. Finally, the separated components of the mixture can be recovered.

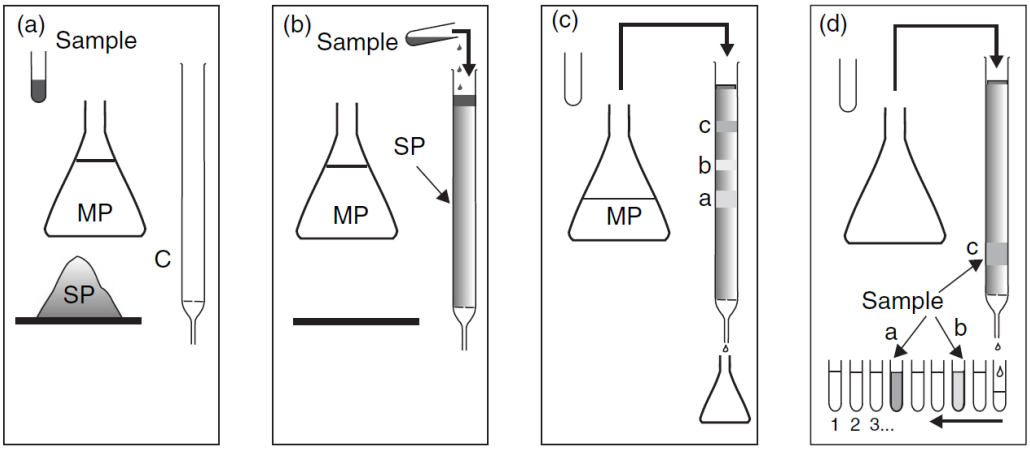


Figure 2.2 – Basic chromatographic process. (a) The different components (C: glass column, SP: stationary phase, MP: mobile phase, S: sample); (b) introduction of the sample; (c) elution of constituents of the mixture; (d) recovering of compounds after separation. Figure taken from Rouessac and Rouessac (2008).

The basic chromatographic process has been largely used for the separation and purification of divers compounds. This technique has been improved greatly. Nowadays chromatographic techniques are composed of divers accessories designed to automatize, control and assure reproducibility of the entire separation process.
In general, the chromatography instrumentation is composed by a mobile phase supply system, a chromatographic column, a detector and an equipment for data acquisition and processing (Figure 2.3).

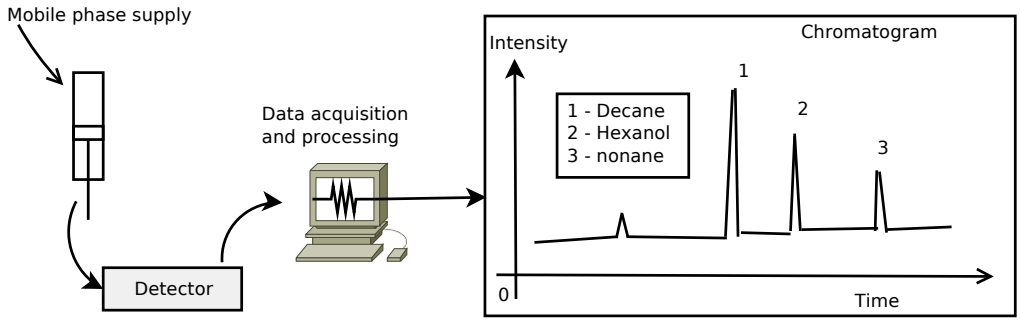


Figure 2.3 – Schematic representation of chromatographic instrumentation

The mobile phase supply systems is designed to transfer the mobile phase through the chromatographic column. To achieve the separation of the compounds present in a mixture, the mobile phase should be composed of divers solvents or solvent mixtures that will interact with the molecules of the sample to proceed with the elution of compounds. With the help of the mobile phase the components of a mixture are carried out to the chromatographic column. The column contains the stationary phase that allows the separation of compounds. Once the different compounds are eluted from the column they are identified by a detector.

The detector along with the column is an important part of the chromatographic system, it identifies the components separated by the column and allows their quantification. Detectors provide whether a single information (e.g., retention time[2] or combined information (e.g., retention time and structure of the analyte). For this reason, chromatographic systems can be equipped with two or three detectors. The detectors must frequently used include: Thermal conductivity detector (TCD), flame ionization detector (FID), Nitrogen phosphorus detector (NPD), Electron capture detector (ECD), Photo-ionization detector (PID), Mass spectrometry detector (MSD), Infrared detector (IR) and ultraviolet detector (UV).
The compounds identified by the detector are represented in a chromatogram. The chromatogram shows the variation of the amount of the analyte in the mobile phase in specific period of time. The chromatogram consist of a peak for each of the separated compounds identified by the detector (Rouessac and Rouessac, 2008).
Considering the physical nature of the two phases involved, chromatographic techniques can be classed into Liquid phase chromatography (LC) and Gas phase chromatography (GC).
The liquid phase chromatography uses a liquid as the mobile phase. Depending on the retention phenomenon this category can be sub-divided into liquid/solid chromatography (or adsorption chromatography), ion chromatography (IC), liquid/liquid chromatography (or partition chromatography, LLC), and liquid/bound chromatography.
In gas chromatography the mobile phase is an inert gas[3] (i.e., helium, hydrogen or nitrogen). Gas chromatography can be sub-divided into gas/liquid/chromatography (GLC) and gas/solid chromatography (GSC).

When measuring an analyte at trace level[4] within a matrix, interferences between analyte and other compounds may occur. Therefore it is necessary to meticulously prepare a sample before its analysis. Some methodologies for sample pretreatment include: solid phase extraction (SPE), immunoaffinity extraction, microextraction procedures, gas extraction on a cartridge or a disc, headspace, supercritical phase extraction and microwave reactors. Among these methodologies the SPE is the most frequently used.

SPE is used to purify or concentrate an extract prior to the analysis of its constituents (Pawliszyn, 2010). It is implemented using a plastic cartridge similar to a syringe which contains a solid sorbent upon which a known volume of liquid sample is passed.

In the initial step of the SPE method the target compound is retained by the sorbent, then undesired substances are eliminated by rising them. Finally a small volume of a strongly enriched solution is recovered using an appropriate solvent. This solution contains the analyte. The SPE procedure isolates and pre-concentrates the analyte, this is particular useful when analysing traces.
An analogous methodology to SPE is the liquid/liquid extraction, where large volume of solvent is used to dilute the analyte. The main drawbacks of the liquid/liquid extraction are: 1) large volume of solvents are necessary and 2) impurities along with the analyte are dissolved by solvent so the extracted solution contains both the analyte and impurities which makes the method unacceptable.

Today, numerous analytical techniques are available, most of them include the use of

[2]The time a solute takes to move from the point of injection to the detector (Harvey, 2000).

[3]Inert gas is a gas that do not undergo chemical reactions, they are used to avoid unwanted chemical reactions

[4]A compound is considered to be in trace amount when its concentration in the environment is less than 1000 ppm ($1\,\mu L\,L^{-1}$ or $1\,mg\,kg^{-1}$)

Liquid Chromatography tandem Mass Spectrometry (LC-MS-MS) or Gas Chromatography Mass Spectrometry (GC-MS). They have been used for the analysis of various compounds in wastewater, surface and ground waters (Robles-Molina et al. 2014; Bruzzoniti et al. 2006; Ma et al. 2009; Nebot et al. 2007; Mansilha et al. 2010).

In Mexico, different efforts to advance on the study of EPs have been made. Some of the analytical methods that have been proposed for the analysis of EPs in Mexico include: Supercritical-Fluid Chromatography coupled with Diode-Array Detection (SFC-DAD) (Salvatierra-Stamp et al., 2015a), Selective Elution and analysis by Gas Chromatography-Mass Spectrometry (GC-MS) (Gibson et al., 2007), Liquid Chromatography with a Photodiode Array Detector (LC-PAD) (Salvatierra-Stamp et al., 2015b), Solid Phase Micro Extraction followed by Gas Chromatography-Mass Spectrometry (SPME-GC-MS) (Peña-Álvarez and Castillo-Alanís, 2015), Liquid Chromatography (LC) with post-column fluorescence detection (García de Llasera and Bernal-González, 2000a) and Liquid Chromatography diode array UV detection coupled on-line to a Solid Phase Extraction (SPE) (García de Llasera and Bernal-González, 2000a).
These methods can be used to analyse different EPs in water samples including: pharmaceuticals (e.g., carbamazepine, carbofuran, glyburide), endocrine disruptors (e.g., 17α-ethinyl estradiol, bisphenol A, 17β-estradiol), bactericides (i.e., triclosan) and pesticides (carbamates, organochlorine, organophosphorus).
In addition to the above mentioned methods, some others were also developed for the analysis of EPs in soils (García de Llasera and Bernal-González, 2000b) and sediments (Alcántara-Concepción et al., 2013).

The works published on environmental assessment of EPs in Mexico are still limited, but the already published work gives evidences of the importance to conduct studies on this topic, and most importantly, to include EPs in the monitoring programs to assess the quality of Mexican surface waters.

2.3 Data collection from biomonitoring in water quality assessment

The growing need for controlling good surface water quality and healthy aquatic ecosystems demands an increasing effort to monitor the quality of aquatic ecosystems. This section presents the importance of using macroinvertebrates-based monitoring metrics as a rapid, low cost, and reliable alternative to monitor the quality of surface waters. In the following subsections a state-of-the art on biomonitoring and collection of data from biomonitoring practices using macroinvertebrates is presented.

2.3.1 Biological organisms in biomonitoring

Aquatic ecosystems are affected by a number of interrelated physical, chemical, and biological factors. The continuous degradation of surface waters due to natural or anthropogenic causes prompts an interest to the monitoring and management from a regulatory approach to a holistic ecosystem approach (Pinto et al., 2009). In this manner, biological monitoring has become an effective tool for the assessment of water quality and aquatic ecosystem integrity (Barbour et al. 1999; Markert et al. 1999; Masese et al. 2013; Thorne and Williams 1997).

Biomonitoring or biological monitoring is an environmental survey using biological variables. Biotic communities present changes in their structure, composition and behaviour when the physical, chemical or biological attributes of a river are altered. Therefore, aquatic ecosystems are increasingly monitored by observing the overall responses of a biocenosis[5] able to integrate over space and time all the stressors undergone by the ecosystem.

Usually biomonitoring tools concern only a part of biocenosis, e.g., fish. Their application supposes to know biological organisms currently present in a geographic site and includes a listing of these organisms. Today, different aquatic organisms are used to assess biotic integrity of streams including: bacteria, protozoans, diatoms, algae, macrophytes, macroinvertebrates, and amphibians (Amaral Pereira et al. 2012; Barbour et al. 1999; Kane et al. 2009; Lane and Brown 2007). Each group responds to disturbances differently, providing diverse information. Thus, the group of assemblage is chosen depending on the objective of the study and the characteristics of the ecosystem to be assessed.

Biomonitoring of lotic systems; i.e., rivers, is done using metrics based on periphyton, benthic macroinvertebrates and fish communities (Li et al., 2010). Metrics based on periphyton and fish communities are good indicators of habitat/hydro-morphological alterations. While metrics based on macroinvertebrates have a better response when enrichment of pollutants (like NO_3) is the main stressor affecting streams integrity (Johnson et al., 2006).

Use of macroinvertebrates on bioassessment practices present different advantages (Bonada et al. 2006; Rosenberg and Resh 1993) including:

1. their benthic nature, which allows a spatial analysis of pollutants;
2. their relatively long-live cycles, that allows to follow environmental changes over long periods of time (they provide evidence of conditions of about 6 last months), contrary to diatoms that can reflect environmental changes only for the last 2 months;

[5] a self-sufficient community of naturally occurring organisms occupying and interacting within a specific biotope.

3. their sensitivity to different types of chemical and/or hydro-morphological alterations;
4. their abundance and diversity: a large number of species produces a range of responses to a wide variety of disturbances;
5. they are easy to identify to a family level (at this level of identification, they provide in general, enough information of alteration level), compared to periphyton;
6. sampling is easy to implement, requires few people, and equipment is relatively inexpensive; compared to fishes;
7. and their ubiquitous occurrence.

2.3.2 Ecological assessment of rivers using macroinvertebrates

Assessment of environmental condition using biological organisms has been historically practised in developed countries (Cairns and Pratt, 1993). Today, analysis of macroinvertebrates communities in rivers is carried out using mainly structural and taxonomical approaches, this means relying on the presence/absence, sensitivity, richness, abundance, and diversity of particular taxa. All this information is then converted into numerical values such as indices and scores.

The main steps on rivers assessment using macroinvertebrates include: (1) field sampling, (2) sample treatment, and (3) interpretation of data.

1. *Field sampling*: Sampling can be implemented in a representative fragment of the water body. To select a representative fragment, it is necessary first to record the specific locality, topography, and drainage characteristics of the aquatic ecosystem habitats. Locality can be specified by latitude, longitude, section number, habitat location (country, state, province or township), and elevation of the study area above the sea level. The topographic description of the study area includes: the type of water body (creek, river, pond, lake or reservoir); surface features (slope and form of the surrounding terrain and shoreline, form of stream channel and formations such as riffles, rapids, falls or islands); size of the water body, and its approximate centre and average depths; current velocity and sediment structure. For drainage characteristics, the major drainage system should be identified along with the name of the water body as well as the smaller watershed that drains the body water (Brower et al., 1997).
   The characteristics of the aquatic habitats are then used to select the sampling methodology and tools. All habitats or specific habitats (e.g., habitats with uniform and clearly defined conditions) can be sampled (Jáimez-Cuéllar et al., 2004). Within the studied fragment of body water, samples are collected generally from mineral and vegetable (alive or dead) substrates and from slow and fast water's course speed. Samples can be collected with the help of different devices (APHA 1998; Ghetti 1997; Rosenberg and Resh 1982). The most commonly-used ones include: nets (e.g., handnet, surber net, kick net, D-frame net, dredges), grabs, core samplers, artificial substrates and drift. Important for the nets is the mesh size which should be between 250 µm and 500 µm. Sampling can be completed by hand, picking macroinvertebrates that may be situated on hard substrates (stones or plants). Biological organisms are frequently preserved in 70% ethanol for posterior examination.

2. *Sample treatment*: Examination of macroinvertebrate samples can be performed in situ or soon after capture, or most frequently later in a laboratory, (Bervoets et al., 1989). Samples are sorted using a pile of sieves of 250, 500, 750, 1000 or 1250 µm

in order to separate different size of substrates and of macroinvertebrates, to make easier the visualization of organisms. The sample can be cleaned with running tap water. Sieve contents should be placed in white trays to sort the organisms. The organisms can be examined under a stereoscope and finally be conserved in glass or plastic containers with 70% ethanol (APHA, 1998)(De Pauwn and Vanhooren, 1993)(Ghetti, 1997)(U.S.E.P.A, 2013).
Family or genus are the level of taxa identification the most adopted in many biomonitoring programs (Thorne and Williams, 1997) where family is considered the minimal identification level recommended (Hilsenhoff, 1988). Taxa is identified by the use of keys of local taxa.

3. *Interpretation of data*: All the information obtained from the identification and quantification of taxa is then treated to compute various metrics (i.e., richness, diversity, similarity, biotic indices, functional traits, multimetric approaches). Finally, analysis of data is performed applying different statistical methods. The selection of the statistical method depends mainly on the purpose of the study and may eventually combine the biological metrics with environmental metrics (e.g. chemical or physicochemical parameters). For instance, graphical techniques are used to provide rapidly accessible information of a selected number of variables while multivariate statistical techniques can be used to test the degree of similarities in a larger number of variables. Some differences on the performance of graphical, multivariate and classification methods are described in Güler et al. (2002) and Walley and S.Džeroski (1996).

The idea of using aquatic organisms as biological indicators of environmental condition was first introduced by Kolkwitz and Marsson (1908). They introduced the concept of saprobity which refers to a measure of organic pollution in rivers and the associated decrease in dissolved oxygen. Since then, different approaches have been developed. Among them, Metcalfe (1989) has distinguished three major approaches: the saprobic, diversity, and biotic approaches. In recent years, however, new approaches have been developed, e.g., multiple biological traits, multimetric and predictive approaches. Below we describe the most common approaches. They are classed into five groups (Resh and Jackson, 1993) which are: (1) measures of richness; (2) enumerations; (3) diversity and similarity indices; (4) biotic indices, and (5) functional traits. In addition to these groups, there are multiple biological traits, multimetric, and predictive approaches.

1. *Measures of richness*: There are as many measures of richness as there are macroinvertebrates orders and levels of identification: the most frequently-used for macroinvertebrates are family, genus and species. Orders of EPT taxa are considered to be sensible to anthropogenic stressors. Therefore, they have been widely used as indicators of environmental disturbances (Wallace et al., 1996). Five measures of richness are the most commonly used (c.f. Table 2.3)

Table 2.3 – Measures of richness (Barbour et al., 1999) (Resh and Jackson, 1993).

Measure name	**Common expression**
Number of total taxa	Num. Total Taxa
Number of EPT taxa (Ephemeroptera, Plecoptera, Trichoptera)	Num. EPT
Number of Ephemeroptera taxa	Num. Ephemeroptera
Number of Plecoptera taxa	Num. Pleocoptera
Number of Trichoptera taxa	Num. Trichoptera

2. *Enumerations*: Concerning enumeration metrics, it is considered that some stresses increase or decrease the total numbers of individuals of some taxa. Some measures

are based on the identification of pollution-sensible groups and their relative abundances to the total abundance of macroinvertebrates (e.g., Ratio of EPT abundance to Chironomidae abundance). The most commonly used include five metrics (c.f. Table 2.4) (Merrit et al., 2008b). They simple consist on counting and classifying individuals, to finally compute the corresponding ration of each metric.

Table 2.4 – Enumeration metrics.

Metric name	Common expression
Ratio of EPT individuals to total individuals	% EPT
Ratio of Chironomidae individuals to total number of individuals	% Chironomidae
Ratio of EPT abundance to Chironomidae abundance	EPT:Chironomidae
Ratio of individuals in numerically dominant family to total number of individuals	% of most dominant genera
Ratio of individuals in numerically dominant taxa to total number of individuals	% of dominant taxa

3. *Diversity and similarity indices:* Diversity approaches use richness, evenness, and abundance to evaluate the community structure with respect to the occurrence of species (Li et al. 2010; Merrit et al. 2008b). These approaches are based on the principle that stressed or polluted water will lead to a reduction in diversity on the community. The main problems of diversity approaches is the unclear reference level and the considerable variations on the diversity of natural undisturbed waters. The use of diversity on water quality monitoring relies on the idea that balanced, stable communities will be represented by high diversity index values. Similarity indices compare community structure between sites. Thus, communities at disturbed and undisturbed sites, will present remarkable dissimilarities. Diversity and similarity indices have been strongly criticized when employed separately in assessment of river systems (Metcalfe, 1989) and today, they are preferably used jointly with other metrics (e.g., multimetric approaches) in order to integrate the behaviour of the elements and processes of biological systems (Karr, 1999). In Table 2.5, the most commonly used Similarity and Diversity metrics are listed. It must be noticed that Bray-Curtis distance is a dissimilarity index, its interpretation is teh opposite of teh results obtained with the other similarity indices but provides relevant information.
4. *Biotic indices:* Biotic approaches combine measures of diversity with the ecological sensitivity of individual taxa into a single numerical expression. The principle of biotic approaches is that macroinvertebrate groups disappear as pollution increases and a reduction of the number of taxonomic groups is observed when pollution increases. The main disadvantage of these approaches are the difficulty to determine representative reference communities and the uncertainties that can be generated when biotic approaches are adopted in geographic regions different from which they were originally designed.
   Biotic indices are calculated as follows: first, different values are assigned to macroinvertebrate taxa according to their sensitivity or tolerance: then, all individual values of the taxa present in the studied sample are summed up to finally generate a single index or score. The values assigned to macroinvertebrates differ among taxa according to their sensitivity and tolerance to pollutants. Those values are constructed more commonly taking into account the tolerance of macroinvertebrates to: pH, organic pollution, eutrophication, heavy metals, and pesticides.
   From the numerous biotic indexes that have been developed (Cairns and Pratt 1993; Metcalfe 1989) six are the most widely used (c.f. Table 2.6). Among them, the Biological Monitoring Working Party (BMWP) has been adopted and modified for effective use in several countries, such as Spain (Alba-Tercedor and Sánchez-

Table 2.5 – Similarity and Diversity indices.

	Index name	Definition	Reference
Diversity indices	Shannon's Index	$H = \sum_{i=1}^{i=S} p_i \log p_i$	(Shannon, 1948)
	Simpson's Index	$D = \frac{\sum_{i=1}^{i=S} n_i(n_i-1)}{N(N-1)}$	(Simpson, 1949)
	Sequential Comparison Index (SCI)	$SCI = \frac{number\ of\ runs}{number\ of\ specimens}$	(Cairns et al., 1968)
	Margalef Index	$D = S - \frac{1}{\log N}$	(Margalef, 1951)
Similarity indices	Jaccard's Coefficient	$C_j = \frac{j}{a+b-j}$	(Jaccard, 1908)
	Sørensen Coefficient	$C_s = \frac{2j}{a+b}$	(Sørensen, 1948)
	Bray-Curtis distance	$BC = 1 - \frac{2C_{ij}}{S_i+S_j}$	(Bray and Curtis, 1957)
	Community loss index	$CLI = \frac{(A-B)}{C}$	(Courtemanch and Davies, 1987)

p_i = proportion of individuals of the ith taxon $p_i = n_i/N$.
n_i = total number of individuals of the ith taxon.
N = total number of individuals for all ith taxa.
S = total number of taxa.
j = number of taxa in common between the stations A and B.
a = number of all taxa in station A and b = number of all taxa in station B.
A = Number of taxa at the reference site.
B = Number of taxa at the study site.
C = The taxa common to both sites.

Ortega, 1988), Australia (Chessman, 1995), India (De Zwart and Trivedi, 1994), Costa Rica (Mafla, 2005), and Colombia (Roldán Pérez, 2003).

Table 2.6 – Biotic indices.

Index name	Common expression	Reference
Trent biotic Index	TBI	(Woodiwiss, 1964)
Extended Biotic Index	EBI	(Ghetti, 1997)(Hellawell, 1978)
Beck Biotic Index	BBI	(Beck, 1955)
Hilsenhoff's Biotic Index	HBI	(Hilsenhoff, 1982)
Biological Monitoring Working Party Score System	BMWP	(Armitage et al., 1983)
BMWP - Average Score Per Taxon	BMWP-ASPT	(Armitage et al., 1983)

Each biotic index contains a table listing the tolerance values assigned for each taxa. Computation of the biotic indexes is done using the corresponding tables for each metric.

5. *Functional traits:* Macroinvertebrate traits are characteristics that have been adopted as a complementary and indirect approach to reflect ecological integrity (Statzner et al., 2001b). This approach is based on the habitat characteristics and the biological and ecological functions of species. Thus, by relating species traits to habitat characteristics, important insights into the structure and functioning of streams communities can be observed (Bremner et al., 2006; Kilbane and Holomuzki, 2004; Poff et al., 2006; Statzner et al., 2001a).
   Traits are classified in two categories mainly: biological and ecological. Biological traits (i.e., size, body form, life cycle, food and feeding habits, reproductive strategies, mobility, etc.) describe the biology of the species. Ecological traits (i.e., pH and temperature tolerances, bio-geographic distribution, tolerance to organic pollution, etc.) are related to habitat preferences.Usseglio-Polatera et al. (2000) have proposed twenty-two (eleven biological and eleven ecological traits).

Functional traits measures organize species into functional feeding froups (FFGs) according to their morphological behaviour, food-gathering mechanisms or locomotion-attachment adaptation (Merrit et al., 2008b; Tachet et al., 2010). Each of these FFGs are expected to vary significantly, they can increase or decrease upon accumulation or loss of particular food sources, presence of pollutants or with particular habitat types. FFGs indicate perturbation of the community when deviations of expected abundances of FFGs or habit groups occur (Merrit et al., 2008b).
FFGs have been applied by several authors to assess perturbations on the aquatic ecosystem such as: land use effects (Dolédec et al., 2006), high-flow disturbance (Holomuzki and Biggs, 2000), anthropogenic effects (Usseglio-Polatera and Beisel, 2002), organic pollution (Lafont et al., 2006) or monitoring of water quality (Bady et al., 2005; Charvet et al., 2000; Statzner et al., 2005) and, in recent years, they have been applied together with other metrics (i.e., multimetric approaches). The most commonly used are: the percentage of filtering collectors, the percentage of scrapers, the percentage of gathering collectors, the percentage of predators and the ratio of scrapers to filtering collectors (scrapers/filtering collectors).

6. *Multimetric approaches*: Multimetric approaches provide robust and integral responses of an assemblage to natural and anthropogenic stressors. They combine various measures of richness, enumerations, pollution-tolerance values, functional feeding groups, dominance, life cycle and density.
The Index of Biotic Integrity (IBI), introduced by Karr (1981) was the first multimetric approach, which was designed to assess fish assemblages. Inspired by the IBI index, different multimetric approaches have been proposed and adopted in different countries including: France (I2M2 (Mondy et al., 2012)), Germany (Vlek et al., 2004), India (Sivaramakrishnan et al., 1996), East Africa (IBI-LVB (Masese et al., 2013)), Brazil (Multimetric approach for Central Amazonia region (Couceiro et al., 2012)), Panama (NLSMI (Helson and Williams, 2013)) and Mexico (IBI-west-central-Mx (Weigel et al., 2002)). Multimetric indices are potential and efficient assessment tools because they integer individual metrics that consider different attributes of communities in order to respond to different types of pressure. Nevertheless, these multimetric approaches are region-specific and cannot be implemented globally.

7. *Predictive approaches*: Predictive models assess river health by comparing reference sites with altered ones. The assumption on these models is that the least-impacted sites with similar environmental characteristics should have similar fauna patterns in the absence of anthropogenic impact. In this context, predictive models have been proven to be useful in biomonitoring activities.
Both multimetric and predictive approaches use biotic indices. The predictive model approaches most frequently used are the Biological Monitoring Working Party score (BMWP) and the Average Score Per Taxon (ASPT) (Norris and Hawkins, 2000). The multivariate predictive models that are widespread used include:

   - RIVPACS (River Invertebrate Prediction and Classification System (Wright et al., 1984));
   - AusRivAS (Australian Rivers Assessment System (Smith et al., 1999));
   - BEAST (Benthic Assessment Sediment (Reynoldson et al., 1995)) and
   - ANNA (Assessment by Nearest Neighbor Analysis (Linke et al., 2005))

   Inspired by RIVPACS, some other models have also been developed such as: Mondego (Feio et al., 2007), Medpacs (Poquet et al., 2009) or the predictive model

for Bolivian streams (Moya et al., 2011). But so far, due mainly to the absence of reference sites and incomplete information on rivers functions, there is no equivalent model developed in Mexico.

Some of the foreign biomonitoring metrics that have been used to assess Mexican aquatic ecosystems using macroinvertebrates include: the Hinselhoff Family-level Biotic Index (Henne et al., 2002)(Barba-Álvarez et al., 2013), the Beck Biotic Index (Rosas et al., 1984) and the Extended Biotic Index (López-Hernández et al., 2007). Nonetheless, they have been used only on isolated cases or for a particular study case. In fact, the use of macroinvertebrates for bioassessment of lotic systems in Mexico is still scarce (Mathuriau et al., 2011). This is due mainly to: untrained personnel, unknown information of pristine communities, absence of sampling material, lack of identification keys, absence of local experts and incomplete information on macroinvertebrates functions on Mexican rivers (Resh, 2007).

Applicability of biological monitoring methods in developing countries can be always debated. For instance, in a study published by Damanik-Ambarita et al., 2016a carried out in different tropical river basins (Ecuador, Ethiopia and Vietnam), it was observed that different biological metrics, presence and absence of sensitive/tolerant taxa as well as relative abundance of macroinvertebrates families could be used to assess the quality of the studied rivers. They particularly observed that functional traits, like FFG (Functional Feeding Groups) were promising. In other study published by Damanik-Ambarita et al., 2016b different biomonitoring metrics were used in Middle and South America Rivers. To study the ecological water quality in Ecuador (Guayas river basin), they selected physico-chemical data (temperature, conductivity, turbidity, Chlorophyll, nutrients, organic matters) and two macroinvertebrates indices: the BMWP-CO (Roldán Pérez, 2003) and the Neotropical Low-land Stream Multimetric Index (NLSMI) (Helson and Williams, 2013). The NLSMI's metrics are : % of scrapers, % of shredders, Margalef's index, % of chironomidae and % of Diptera, % of Trichoptera and Shannon Index, which are among the metrics we recommend. They used these metrics because macroinvertebrates' responses towards environmental changes might vary across sites and habitats (García-de la Parra et al., 2012; Helson and Williams, 2013). However in the above mentioned publications they did not study micro-pollutants. In our study we plan to study macro-pollutants and micro-pollutants as chemical variables, we choose to keep all the 35 biotic metrics and their potential correlation with abiotic data because, 1) as showed by Damanik-Ambarita et al., 2016a the relationship between macroinvertebrates communities and habitat disturbance are still lacking and poorly understood in South America, and 2) as explained previously, bioassessment of freshwaters in Mexico is scarce.

2.3.3 Biomonitoring data: general characteristics, common uncertainties and anomalies

Biomonitoring data consist of a list of fauna living in the rivers. The term taxa is used to define the family, sub-family or species from which a giving individual belongs. Fishes, oligochaetes or macroinvertebrates could be part of the fauna on a biomonitoring study. The resulting list of taxa is used to compute biological indices that provide a numerical value. This numerical value is then used as an indicator of the water quality characteristics of river. In the previous section (c.f. 2.3.2), we described the different indices that can be computed. Each index provides different information that scientist can integer to have a general overview of the characteristics of a river.

Uncertainties in biomonitoring data are ubiquitous. They may occur from collection to

analysis of data. Uncertainties may occur due to differences on: sampling protocols, number of volume of samples, type of substrates, analysis of samples by inexperienced personal, sorting method, level of taxa identification or rare specimens consideration (Wiederkehr, 2015).

Misclassification of taxa, is one of the most common errors found on biomonitoring data. It has an important impact on the boundary between good and moderate rivers water quality status (Wiederkehr et al. 2016; Haase et al. 2006; Metzeling et al. 2003). Another common error is missing data, that may occur mainly to lost of samples.

2.4 Preprocessing and analysis of environmental data

Data exploration and analysis are the basis to discover knowledge, understand studied phenomena, solve problems, and make decisions. Data quality is an important issue on data analysis, bad data quality may be significant costly (Haug et al., 2011) and induce misleading conclusions (Wahlin and Grimvall, 2008). Quality data is mandatory particularly on environmental data because erroneous data may lead to faulty conclusions leading expensive decisions and causing dramatic consequences on environmental systems. Aiming at improving the quality of data for the particular case of water quality, Rieger et al. (2010), Alferes et al. (2013), Alferes and Vanrolleghem (2014) have proposed to evaluate and validate data from *in situ* measurements. Such evaluation allows to detect data anomalies and implement corrective actions for on-line monitoring systems however, most of the monitoring systems are not on-line. Thus, different approaches are necessary in order to correct data anomalies and improve the quality of data.

Anomalies in data are ubiquitous and may have dramatic consequences in data analysis (Eppler and Helfert, 2004), limit performance of statistical procedures (Cortes et al., 1995; Coussement et al., 2014) and produce misleading analytical results (Wahlin and Grimvall, 2008). To mitigate the impacts of data anomalies such as missing values, inconsistencies, outlying data, duplicates, etc., it is a first necessary step to preprocess data (Famili et al., 1997; Gibert et al., 2008).

Data preprocessing is a fundamental and critical step in data analysis, it consists of all actions necessary to prepare the data before data analysis. In general it is performed to: (1) solve data problems, (2) understand the nature of data and perform more meaningful data analysis, and (3) extract meaningful knowledge from a given dataset and application domain (Famili et al., 1997).

Although, numerous procedures to preprocess data anomalies are available, there are many inherent challenges in this area, for instance:

1. Conventional approaches usually treat data anomalies as independent types of anomaly and handle them in isolation. Preprocessing approaches that take into consideration jointly different types of data anomalies still need to be developed;
2. Anomalies co-occur in data. It is crucial to define a data preprocessing ordering that produce the results with the least bias. This means that preprocessing procedures should not impact statistical results;
3. The impact of data preprocessing procedures on statistical analysis results is under studied. Thus, it is important to conduct this type of studies;
4. It is necessary to define a systematic framework that will allow users to perform the most appropriate preprocessing procedure according to a specific statistical analysis task.
5. Very few research is focused on environmental data preprocessing.

In this thesis, we focus on the different strategies to preprocess data anomalies and particularly, on their impact on statistical analysis results. We have conducted a bibliographic study related to data anomalies, data preprocessing procedures, and their impact on statistical analysis.
In the following section, we briefly describe data anomalies and related detection methods. Then, we describe data preprocessing procedures for missing values, outliers, and duplicates.

2.4.1 Data anomalies and their detection

2.4.1.1 Data anomalies

Abnormal patterns in data (or *data glitches*) are errors in the measurement and recording of data that negatively impact analysis. They are ubiquitous and occur due to a variety of reasons from human errors (e.g., typos) to software and hardware problems (e.g., inconsistencies due to failures of automated equipment) (Berti-Equille et al., 2015). Data may be distorted during the collection step, or when data is transcribed, merged, transferred or copied (De Veaux and Hand, 2005).

Different types of data glitches may be found. According to the taxonomy of data distortion proposed by Kim et al. (2003), we can identify three main classes: missing data, not missing but wrong data, and not missing and not wrong but unusable.

We focus on data glitches related to missing data and not missing but wrong data. Hereafter, we refer only to numerical data problems. Data glitches, detection methods, and preprocessing procedures related to non-numerical data are out-of-the scope of this work and will not be addressed. Hereafter, we describe the data glitches most commonly encountered.

Missing data: Missing data are instances where no data is stored for a given variable in the current observation. They may occur due to failure in equipment or human errors. Let us consider a dataset which composed by a table with rows (observations) and columns(variables). A dataset is considered incomplete if there are at least one variable with at least one missing data. A complete dataset is then defined as

$$Y_{com} = (Y_{obs}, Y_{mis}) \tag{2.1}$$

where Y_{obs} corresponds to the "observed" part and Y_{mis} to the missing part.

To better understand the behaviour of missing data it is necessary to identify them within a dataset Y_{com}. In the dataset indicator variables R are defined, they identify what is known ($R = 1$) and what is missing ($R = 0$). R is known as the *missingness* which is represented by a matrix composed by the indicator variables linked to the variables of Y_{com}. The pattern of missingness depends on the structure of missing values of an incomplete dataset. It can be: (a) univariate if only one variable has missing values, (b) monotone if the variables in the dataset can be classified from left to right so that if a value is missing from Y_{mis_i} so it si also for $Y_{mis_{i+1}}$, and (c) arbitrary (c.f. Figure 2.4).

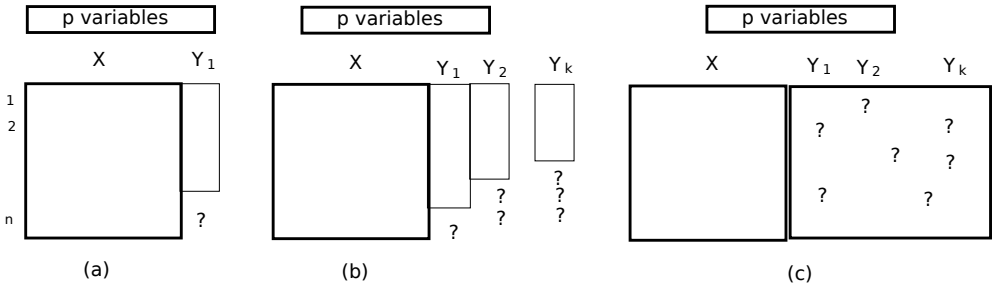


Figure 2.4 – Structures of missingness: (a) univariate pattern, (b) monotone pattern, and (c) artibtrary pattern. Rows correspond to observations and columns to variables (Cottrell et al., 2009).

.

According to Rubin (1976) three types of missing data can be distinguished: Missing at Random (MAR), Missing Not at Random (MNAR) and Missing Completely at Random (MCAR).

By denoting a complete data as Y_{com} and partition it as $Y_{com} = (Y_{obs}, Y_{mis})$ where Y_{obs} and Y_{mis} are the observed and missing parts respectively.

Missing data are considered to be MAR when the distribution of missingness does not depend on the missing parts. This can be expressed as

$$P(R/Y_{com}) = P(R/Y_{obs}) \tag{2.2}$$

This equation indicates that all the information on the missing part (Y_{mis}) of the data is contained in the observed part (Y_{obs}).

An special case of MAR is MCAR, which occurs when the distribution of missingness does not depend on both Y_{mis} and Y_{obs}.

$$P(R/Y_{com}) = P(R) \tag{2.3}$$

Missing data is considered as MNAR when the distribution of missingness depends on Y_{mis} (Schafer and Graham, 2002).

Considering for example the structure of univariate data in Figure 2.4, where the variable X is known for all individuals but the variable Y is missing for some individuals. MCAR means that the probability that Y is missing for an individual it does not depend of values in X or Y. MAR means that the probability that Y is missing may depend on X but not on Y, and MNAR means that the probability of the existence of the missing values depend on Y.

The previous definitions describe the statistical relationships between a set of data and their missing values while ignoring the causes of the gaps. It is assumed now that these causes can be encoded by a group of variables Z. These group of variables may have, for example, variables that will explain the reasons for which some individuals have missing values. It is possible that the variables that cause missing values are not present in a dataset, however some of these variables may be related to X and Y and so, relation between X, Y and R can be deduced.

By setting Z as a variable of cause that is unrelated to X and Y, MCAR, MAR and MNAR data can therefore be represented by the graphical relationship of Figure 2.5. On MCAR it is necessary to have the causes of missing data entirely contained in Z. MAR allows some causes of being related to X. On MNAR, it is required to have some causes related to Y once the relationship between X and R have been taken into account (Schafer and Graham, 2002).

In general a dataset is necessarily either MNAR or MAR.

Inconsistent and faulty data: Inconsistent data are considered as data that do not respect certain constraints (e.g., domain expert or business constraint). Outliers for instance may be considered as inconsistencies in the case that data do not respect a given statistical model as a constraint such as distribution or density. And our focus is on outlying values as a type of inconsistencies that are model-based.
Inconsistencies may arise from misreading instruments or misrecording values at any level. For instance, when "John F. Kennedy" is recorded as "Jhon F. Kenedy", this may be easy identified as a typo error. There may be more complex cases for instance: "400-02-06-2005" that is entered as "400-20-06-2005". Another example may be when subjects do not follow questionnaire instructions by answering questions that they were supposed to

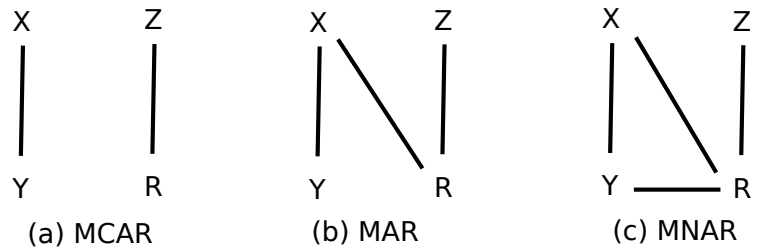


Figure 2.5 – Representation of (a) missing completely at random (MCAR), (b) missing at random (MAR), and (c) missing not at random (MNAR) in a univariate missing-data pattern. X represents completely observed variables, Y denotes partly missing variable, Z represents the component of the causes of missingness unrelated to X and Y, and R represents the missingness (Schafer and Graham, 2002).

.

skip (Arlanturk et al., 2016). In these cases, constraint satisfaction methods are used (Berti-Équille, 2007b; Berti-Équille and Dasu, 2009).

Outliers: These are unexpected data that do not conform to a given model. According to Chen et al. (2010), an outlier is considered as a pattern in data that does not comply with the general behaviour of the data. Park et al. (2003) identify two types of outliers: (a) outliers caused by equipment failure or measurement errors and (b) outliers reflecting ground truth. Outliers occur due to human errors, instrument errors, natural deviations in populations, fraudulent behaviours and changes or faults in the systems (Hodge and Austin, 2004)

Duplicates: Duplicates are repeated entries. They may not share common identities (Sarawagi and Bhamidipaty, 2002; Berti-Équille, 2007b). These kind of glitches occur frequently in data integration from different databases, they are defined as *data heterogeneity* problems (Chatterjee and Segev, 1991). Two types of data heterogeneity can be distinguished: *structural* and *lexical* (Elmagarmid et al., 2007). Structural heterogeneity arise when databases are structured differently. For instance, user name might be recorded in one field *Name*, while, in another database the same information might be recorded in multiple fields, say, *First Name*, *Last Name*. Lexical heterogeneity occurs when different representations are used to refer to the same object in multiple databases (e.g., *George Walker Bush* versus *G.W. Bush*).

Undocumented data: Data may be *unusable* when proper data documentation is lacking. Undocumented data is a very important problem because, without an appropriate data dictionary, many suppositions can be made. For instance, a record can be composed of a set of numbers as follows: 123 456 789. One can suppose that is a telephone number, or a membership number, or many other possibilities. However, this is still a guess: without proper data documentation this type of data is *unusable* (Berti-Équille and Dasu, 2009; Dasu, 2013).

2.4.1.2 Detection of data anomalies

There are different tools in the fields of statistics, data mining, and data management to detect problems in data. Detection methods are mainly classified into two categories (Dasu, 2013): (1) quantitative methods based on statistical and data mining approaches, and (2)

methods based on constraints developed from data properties and functional dependencies, or manually defined by experts.

We focused mainly on statistical methods and particularly on methods to detect missing and outlying data. We review techniques for the detection of such anomalies. We also briefly review some approaches for detection of inconsistent and duplicated data.

Data profiling is performed by using statistical methods to get standard characteristics of data (i.e., data types, granularity, format patterns, content patterns, value sets, and cross-column relationships) (Sebastian-Coleman, 2012; Abedjan et al., 2016). This first exploration of data allows to detect mistakes, check assumptions, preselect appropriate models and determine relationships among the explanatory variables. The statistical methods used in data profiling are either univariate or multivariate.

Missing data. Number, pattern of missing observations (MAR, MNAR, MCAR) and plausible reasons for missing data are useful information for an appropriate treatment and analysis (Pigott, 2001). Number of missing observations is an important information. Lower percentage of missing observations may imply the use of simple methods to process them though, there is not a consistent definition of "low percentage of missing observations" they may go to 20% to lower values to be processed (Little and Rubin, 2002).

Currently numerous statistical software's allow exploratory data analysis (e.g., SAS, ADaMSoft, IMB SPSS Modeler, JASP, R, RapidMiner, STATISTICA, MATLAB, WEKA). Must of them include packages/modules that allow the detection of missing values. We can mention for instance the package VIM available in the R environment for statistical computing which allows the exploration and analysis of missing observations (Templ et al., 2011a).

Outliers detection. Considering the definition given by Hawkins (1980) *"An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism"*. Thus, the purpose is to detect observations that fit into this description. There are many methods to detect outliers, some of them include: statistics-based (proximity-based, parametric, non-parametric and semi-parametric), distance-based, density-based, deepness-based, deviation-based, clustering-based, neural networks (supervised and unsupervised), machine learning, etc. (Huang et al., 2006; Kriegel et al., 2010). We will just describe three outlier detection approaches in detail. A more detailed description of the different methods can be consulted on the survey of Kriegel et al., 2010.

- *Statistic-based outlier detection approach.* This approach assumes a distribution or probability model for the given dataset then a discordancy test is used to identify outliers with respect to the model. A discordancy test examines a *null* or *working hypothesis* and an *alternative hypothesis*. The working hypothesis, $H_0 : x_i \in F, i = 1, 2, ..., n$, assumes the entire dataset of n records comes from the same distribution model F. In the alternative hypothesis, H_1 it is assumed that data comes from another distribution model G, $H_1 : x_i \in G, i = 1, 2, ..., n$. The test of discordancy verifies whether an object x_i is significantly large or small in relation to the distribution F. Many tests have been proposed. A review of the most frequently used has been done by Barbato et al. (2011). Briefly, they include: Interquartile range (IQR), modified IQR, Pierce, Chauvenet, Grubbs and Dixon's methods.
  Some of the major disadvantage of the test of discordancy are: (1) they assume that data follows a normal distribution. This may be misleading, particularly when data does not follow a normal distribution; (2) results depend on chosen model F. Thus, x_i may be an outlier under one model and a valid value under another and; (3) they

are appropriate for one-dimensional data but not applicable to multi-dimensional data (Barnett and Lewis, 1994; Barbato et al., 2011).
Detection of outliers in multivariate data is more complex than univariate data because outliers may be masked in the bulk of multi-dimensional data. Multivariate outlier detection methods consider the variations of multiple variables, they depend frequently on a distance measure. Such distance measure attempts to measure the distance between an observation and the centre of the data distribution, by considering the inherent variability and correlation in the data. Multivariate outlier detection methods are frequently based on Mahalanobis distance (Filzmoser, 2004; Riani et al., 2009; Berti-Équille and Dasu, 2009; Kriegel et al., 2010).

- *Distance-based outlier detection approach.* Distance-based approaches measure similarity between two objects with the help of distance between them in the data space. When this distance exceeds a particular threshold, the object is considered as an outlier. In a given dataset X, an outlier $DB(p, d)$ with parameters p and d, will be considered as such if a fraction of p of the object x in X lies at a distance greater than d from x. This approach is suitable for datasets that do not fit any standard distribution model, it discovers outliers effectively and it can be used on multivariate data. The k-NN approaches such as (Knorr and Ng, 1998; Knorr et al., 2000; Ramaswamy et al., 2000; Byers and Raftery, 1998) are the most popular.

- *Density-based outlier detection approach.* In this approach, the degree of an object to be an outlier is measured with respect to the density of the local neighbourhood. This degree named local outlier factor (LOF) is assigned to each object (Breunig et al., 2000). Compared to other approaches, the density-based approach assigns an outlier factor, which is the degree the object is being outlier. Different density-based outlier mining algorithms have been proposed, for instance: the local correlation integral (LOCI)(Papadimitriou et al., 2003), the relative density factor (RDF)(Ren et al., 2004) or the density-similarity-neighbour based outlier factor (DSNOF)(Cao et al., 2010) to mention some. The singularity of the density-based outlier detection and its ability to detect outliers has become an attractive approach to detect outliers (Chen et al., 2010).

Although, the selection of appropriate outlier detection methods is an important task when determining relationships among the studied variables (Alameddine et al., 2010; Filzmoser et al., 2005) an universal outlier detection method has not been identified, rather, different aspects may be considered for the selection of a particular method including: characteristics of the data (e.g., dimension, type, sample size), scalability, speed, model accuracy, algorithm efficiency and performance. A detailed insight of the pros and cons of different outlier detection techniques has been provided by different authors (Reimann et al., 2005; Pasha, 2013; Chen et al., 2010; Filzmoser, 2005; Zimek et al., 2012; Barbato et al., 2011; Ghosh and Vogt, 2012; Cousineau and Chartier, 2015).

For the case of the detection of **inconsistent** and **duplicate** data, various techniques based mainly, on constraints have been developed, some of them are described below.

Detection of inconsistencies. There are various approaches for the resolution of inconsistencies (Dongilli et al., 2005; Kolev et al., 2015). Data inconsistency occurs at the level of data representation. This kind of conflict is based generally on the content and not the schema and so it can be resolved by simple conversion in the case of numerical values.

Detection of duplicated data. Typically, detection of duplicated data relies on

string comparison techniques to deal with typographical variations. A detailed survey of methods to detect typographical variations has been made by Elmagarmid et al. (2007). Briefly, they are classed into: character-based similarity metrics, token-based similarity metrics, phonetic similarity metrics and numeric similarity metrics. Some other approaches to detect duplicates in data streams have been proposed by Deng and Rafiei (2006), Shen and Zhang (2008), Gopalan and Radhakrishnan (2009) and Wei et al. (2011).

For duplicated data where multi-instance tuples are versions of each other, a *tuple similarity measure* is adopted. The *tuple similarity measure* is used to map each tuple to a binary vector in a semantic vector space, then the similarity between tuples of the multi-instance is computed as one of the *vector coefficients* (e.g., matching coefficient, Dice coefficient, overlap coefficient or Jaccard coefficient). Two tuples can have high probabilities of being versions of each other if they are the most similar (Anokhin et al., 2001).

In data from water quality assessment, the anomalies most frequently found are missing values, outlying data, duplicate data and inconsistent data.

2.4.2 Dealing with data anomalies and main preprocessing procedures

The best way to improve the quality of data is to prevent errors since the data collection step. Though, even if good data collection practices are implemented, data glitches may inevitable occur. When data glitches occur, different procedures can be implemented to process them. It is important to notice that the preprocessing may differ depending mainly on the purpose of the study and the statistical method to be used. Below we give an overview of features selection and normalization procedures then, we continue presenting an overview of relevant methods to process data anomalies, particularly missing values and outliers.

Data preprocessing is a mandatory task needed to convert raw data into new data that serve as input for a certain data mining (DM) algorithm. Data preprocessing involves different subtasks mainly: data integration, data cleaning, data reduction and data transformation (García et al., 2015; Hellerstein, 2008).

In *data integration*, multiple data sources (e.g., databases, data cubes, flat files) are combined to construct new tuples and values.
Data cleaning involves all operations necessary to correct anomalous data (e.g., treating missing or outlying data, removing noise, resolving inconsistencies).
Data reduction include techniques (e.g., dimension reduction, data compression, discretization and concept hierarchy generation) to reduce the volume or dimensions (number of attributes) in a dataset without compromising the integrity of the original data and yet producing quality knowledge.
In *Data transformation*, data is converted or consolidated into forms appropriate for mining. Subtasks in data transformation include: normalization, smoothing, aggregation, feature construction and generalization of the data.

We focus our overview on the procedures for data cleaning, particularly to solve missing and outlying data on data transformation specifically normalization and feature selection for data reduction.

Missing values. Today various imputation methods are available (Olinsky et al., 2003). Most of them use indirect approaches to replace the missing values by an imputed form. In general, the imputation methods are focused on MAR data. We can identify two

main types: univariate (*single imputation*) and multivariate (*multiple imputation*) methods (Olinsky et al., 2003; Donders et al., 2006; Farhangfar et al., 2008). One important difference between univariate and multivariate methods is that univariate methods are based on the estimation of parameters in order to maximize the similarity between variables when replacing the value, whereas multivariate imputation methods are based on similarities among the objects and/or variables.

The univariate imputation methods most commonly used include: the mean, hot-deck and regression (Olinsky et al., 2003; Schafer and Graham, 2002). Multivariate imputation methods are based either on the *k*-NN, the Expectation-Maximization (EM) algorithms or they are multiple imputations.

k-NN methods are based on the use of a distance measure in order to find the nearest neighbour, subsequently, the average weighted value found is used to replace the missing value. In general, the average weight value used for numerical data is the mean (Troyanskaya et al., 2001; Jerez et al., 2010; García-Laencina et al., 2009; Hron et al., 2010).

The main advantages of *k*-NN are: (1) it predicts both quantitative and qualitative features and (2) the training dataset is used as a 'lazy' model so explicit predictive models are not created. The major disadvantage is that the algorithm searches through all the dataset to look for the most similar instances. This is a big limitation for large databases. When applying the *k*-NN approach different distance metrics are used to define similarities between target and reference records. The most frequently used are based on absolute differences, Euclidean or Mahalanobis distance functions (Eskelson et al., 2009). Absolute differences are computed as:

The Expectation-Maximization approach iterates through two main steps (Expectation, E, and Maximization, M). An incomplete data matrix have observed data Y_{obs}, missing data Y_{mis} and a vector of parameters, θ. Complete data Y_{com} is then defined as $Y_{com} = (Y_{obs}, Y_{mis})$. The expected complete data log-likelihood function is defined as $Q(\theta|\theta^{'}) = E\{ln[f(Y_{com}|\theta)]Y_{obs}, \theta^{'}\}$. Where the complete data log-likelihood function and the observed data log-likelihood function are defined as: $L(\theta) = f(Y_{com}|\theta)$ and $L(\theta) = f(Y_{obs|\theta})$ respectively. The EM algorithm alternates between the following steps, where θ is initialized at some value:

1. Expectation step (E step): Computing $Q(\theta|\theta^{(t)})$ as a function of θ;
2. Maximization step (M step): Find $\theta^{(r+1)}$ that maximizes $Q(\theta|\theta^{(t)})$

In the M step, the maximum log-likelihood estimation is performed as if there were no missing data. Then, in the E step, the conditional expectation of the missing values is found, such value is predicted given the observed data and current estimated parameters. The values obtained from the expectation are used to replace the missing values. The approach iterates until there is convergence in the parameters estimates (Ho et al., 2001; Olinsky et al., 2003; Palarea-Albaladejo and Martín-Fernández, 2008). The EM is a simple and easy to implement approach. The algorithm is numerically stable and it provides fitted values for the complete data during the E step which do not demand further computation. The main disadvantage is its slow linear convergence in some cases in addition the EM approach tends to underestimate the variability of the estimate (Couvreur, 1997).

Multiple imputation (MI) involves three steps: the generation of completed datasets, the computation of overall estimates and standard errors, and the combination of results. Each missing value is replaced by a list of simulated values, which generates m possible versions of the complete datasets then, each m possibility is analysed by standard proce-

dures which are specific for handling complete data. Finally, the results are combined to obtain overall estimates and standard errors that will provide information about missing-data uncertainty and finite-sample variation. According to (Rubin, 1976) the three steps of MI are:

Step 1 - Imputation: An appropriate model that includes random variation is used to impute missing values. 'Completed' datasets are created by setting plausible values for missing observations. The set of values is used M times to create M 'completed' datasets.

Step 2 - Analysis: The M datasets are analysed using standard complete-data methods.

Step 3 - Combination: In the final step, results are combined in order to take into consideration the uncertainty of the imputation. The following estimations are involved in the procedure:

- produce a single-point estimate (i.e., $\hat{\theta} = (1/M)\sum_{m=1}^{M}\hat{\theta}^{(m)}$) by averaging the value of the parameter estimates across the M samples;
- compute standard errors by (a) getting the average of the squared standard errors of the M estimates, (b) computing the variance of the M parameter estimates, and (c) include the uncertainty due to imputation by combining the two quantities.

The purpose of the imputation is to perform statistical analysis taking into account the uncertainty due to the presence of missing values present in the incomplete dataset and preserving the main characteristics and structure of the data (Farhangfar et al., 2008; Abayomi et al., 2008; Cottrell et al., 2009). As the EM approach, MI is advantageous because it incorporates uncertainty associated to the imputation, it is simpler to adapt however, when the percentage of missing information is large, several imputations are required to get precise estimates, and the results are very sensitive to a misspecification of the model (Heitjan, 1997).

In the selection of an imputation method to replace missing values, different aspects can be considered such as: the characteristics and structure of data (e.g., percentage of missing values, distribution, pattern of missing values), robustness of method, speed and suitability (univariate or multivariate). Aiming at identifying the most performing imputation method, different comparative studies have been done (Kadengye et al., 2012; Hron et al., 2010; Zhang et al., 2009; Borgoni and Berrington, 2013; Olinsky et al., 2003; Engels and Diehr, 2003; Mercer et al., 2011), results of such studies indicate that multiple imputation methods perform better than single imputation methods.

Outliers: As previously mention, outliers are patterns in data that are not necessarily incorrect but likely suspicious. Once an outlier has been detected, the way to proceed for its treatment should be taken cautiously. Outlying data can be whether eliminated from data, retained with an appropriate label or replaced by an imputed form (e.g., mean, median and mode), using single imputation methods. The way in which an outlier is treated will mainly depend on the application domain and expert knowledge (Hodge and Austin, 2004).

Feature selection is performed to find the minimum set of attributes that are necessary to extract the same knowledge as the one obtained using all attributes. It allows to speed the learning stages and facilitates the understanding of data patterns. Feature selection techniques are classified into three categories: Filter, Wrapper and Hybrid (Liu and Motoda, 2012; Sutha and Tamilselvi, 2015).

- *Filter methods* look at the intrinsic properties of the data to evaluate the relevance of features. In general, the score of feature relevance is computed so low-scoring features are identified and removed. Filter methods are advantageous because they can be

used in high-dimensional datasets, are computationally simple, fast and independent of the classification algorithm. However, most of the available methods are univariate and ignore feature dependencies.

- Contrary to filter techniques where relevant features are found independently of the model selection step, *wrapper methods* embed the model selection within the feature subset search. In wrapper techniques, different possible feature subsets are searched, evaluated, and compared. A specific classification model is trained and tested to evaluate a specific subset of features. To search for possible feature subsets, a search algorithm is 'wrapped' around the classification model. To guide the search of an optimal subset and thus avoid the generation of exponential feature subsets, heuristic search methods are used. The main advantages of wrapper techniques is that they take into account feature dependencies and that they integer feature subsets search with model selection. However, they are computationally expensive and have high probabilities of overfitting.
- In *Hybrid methods*, the optimal subset of features is search while the classifier is constructed. As in wrapper methods, hybrid methods are specific to a given learning algorithm; however, hybrid methods are less computationally expensive.

Interesting results of a comparative study on feature selection methods has been published by Bolón-Canedo et al. (2013). In accordance to their results, Bolón-Canedo et al. (2013) suggest to use filtering methods over wrapper and hybrid methods. However, when selecting a method, Murtaugh (2009) suggest to base the selection on the appropriateness for the task at hand. Some other aspects can also be considered such as: characteristics of data, computational costs and robustness of the method (Murtaugh, 2009; Saeys et al., 2007; Sutha and Tamilselvi, 2015).

In addition to the above mentioned techniques, there are some approaches that automatically perform feature selection or variable selection e.g., within a regression model. We can distinguish three important classes of methods: (1) Subset selection, (2) Shrinkage, and (3) Dimension reduction.

- *Subset selection* In this approach a subset of p predictors that we believe to be related to the response are identified. A model using least squares is then fitted on the reduced subset. Subset selection methods include best subset and stepwise model selection.
  In the *best subset selection* a least squares regression is fitted for each possible combination of the p predictors. Then the resulting models are compared in order to identify the one that fit the *best*. Its algorithm is described as follows:

Algorithm 2.1. Best subset selection

1. Let M_0 denote the *null model*, which contains no predictors. This model simply predicts the sample mean for each observation.
2. For $k = 1, 2, ...p$:
   (a) Fit all $\binom{p}{k}$ models that contain exactly k predictors
   (b) Pick the best among these $\binom{p}{k}$ models, and call it M_k. Here *best* is defined as having the smallest Residual sum of squares (RSS), or equivalently largest R^2.
3. Select a single bet model from among $M_0, ..., M_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

The main disadvantage with best subset selection is that it can not be applied with very large p. When p is large, there are higher chances to overfitting and high variance of the coefficient estimates. An alternative to best subset selection, *stepwise*

methods can be used. They explore a more restricted set of models and include the forward stepwise selection and backward stepwise selection.
Forward stepwise selection starts with a model with no predictors and add predictors one-at-a-time until all predictors are included in the model. Contrary to forward stepwise selection *Backward stepwise selection* begins the least squares model containing all p predictors and removes the least useful predictor one-at-a-time. Below we provide the algorithm for Forward stepwise selection, more details on the the backward stepwise selection can be consulted at (Harrell, 2015).

Algorithm 2.2. Forward stepwise selection

1. Let M_0 denote the *null model*, which contains no predictors.
2. For $k = 0, ...p-1$:
   (a) Consider all $p-k$ models that augment the predictors in M_k with one additional predictor.
   (b) Choose the *best* among these $p-k$ models, and call it M_{k+1}. Here the *best* is defined as having smallest RSS or highest R^2.
3. Select a single bet model from among $M_0, ..., M_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

Forward stepwise can be used when $n < p$, while backward selection requires that the number of observations n to be larger than the number of variables p.
Subset selection, forward selection and backward selection provide a set of models each one containing a subset of p predictors. To select the best model with respect to test error the *C_p, Akaike information criterion (AIC), Bayesian information criterion (BIC)* and *Adjusted R^2* can be used. From a standard linear model

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p + \epsilon \tag{2.4}$$

The C_p estimate of test MSE is computed as follows:

$$C_p = \frac{1}{n}(RSS + 2d\hat{\sigma}^2) \tag{2.5}$$

where $\hat{\sigma}^2$ is an estimate of the variance of the error ϵ associated with each response measurement in 2.4. In the selection of the best model, the model with the lowest C_p value is chosen. The AIC criterion is defined as:

$$AIC = \frac{1}{n\hat{\sigma}^2}(RSS + 2d\hat{\sigma}^2) \tag{2.6}$$

C_p and AIC are proportional to each other. The BIC is given by

$$BIC = \frac{1}{n}(RSS + log(n)d\hat{\sigma}^2) \tag{2.7}$$

Similarly to C_p the model that has the lowest BIC values is selected.
For a least square model with d variables the adjusted R^2 is defined as:

$$Adjusted\, R^2 = 1 - \frac{RSS/(n-d-1)}{TSS/(n-1)} \tag{2.8}$$

where $TSS = \sum (y_i - \bar{y})^2$. Contrary to C_p, AIC and BIC where small values indicate a model with low test error, large values of adjusted R^2 indicate a model with a low test error. The above mentioned methods involve the use of least squares linear models that contains a subset of predictor variables. As an alternative shrinkage methods can be used to fit a model that contains all p predictors.

- *Shrinkage.* Shrinkage methods constrains the coefficient estimates towards zero. By shrinking the coefficient estimates the variance of the model is reduced. The two best-known techniques for shrinking the regression coefficients are *ridge regression* and the *lasso.*
  The fitting procedure in least squares estimates $\beta_0, \beta_1, \beta p$ uses values that minimize

$$RSS = \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 \tag{2.9}$$

  In *ridge regression* the coefficients are estimated by minimizing

$$\sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} \beta_j^2 = RSS + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{2.10}$$

  The objective in *ridge regression* is to make the RSS small in order to get coefficient estimates that fit the data well. The term $\lambda \sum_j \beta_j^2$ named *shrinkage penalty* will be small when $\beta_1, ..., \beta_p$ are close to zero, therefore it will shrink the estimates of β_j towards zero. The advantage of ridge regression over least squares is based on the *bias-variance trade-off.* In fact as λ increases ridge regression fit decreases, this leads to decreased variance but increased bias. In addition, ridge regression is advantageous over best subset selection computationally speaking, because ridge regression only fits a single model for any fixed value of λ while best subset selection requires searching through 2^p models. One important disadvantage of ridge regression is that it includes all p predictors in the final model and so $\lambda \sum \beta_j^2$ will shrink all the coefficients towards zero, but any of them will be set exactly to zero (unless $\lambda = \infty$). This is problematic particularly in model interpretation where the number of variables p is large. An alternative to ridge regression that overcomes this disadvantages is *lasso.* The lasso coefficients estimates are the values that minimize

$$\sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} |\beta_j| = RSS + \lambda \sum_{j=1}^{p} |\beta_j| \tag{2.11}$$

  The difference with respect to lasso is that the β_j^2 term in the ridge regression has been replaced by $|\beta_j|$. Similarly to ridge regression, lasso shrinks the coefficient estimates towards zero. In the case of lasso some of the coefficient estimates are forced to be exactly equal to zero when λ is sufficiently large. Lasso performs *variable selection* which generated models are easier to interpret compared to those produced by ridge regression. It is expected that lasso perform better when a relatively small number of predictors have substantial coefficients, and the remaining predictors equal zero. Ridge regression performs better when the response variable is a function of many predictors with coefficients of equal size. The main problem is that in real-datasets the number of predictors that are related to the response is a priori unknown.

  Subset selection and shrinkage methods control variance either by using a subset of variables or by shrinking the coefficients toward zero. An alternative to these methods are the dimension reduction methods which transform the predictors to then fit a least square model with the transformed variables.
- *Dimension Reduction* By considering $Z_1, Z_2, ..., Z_M$ represent $M < p$ linear combinations of original p predictors, that is

$$Z_m = \sum_{j=1}^{p} \phi_{jm} X_j \tag{2.12}$$

for some constants ϕ_{1m}, ϕ_{2m}, ...,ϕ_{pm}, $m = 1, ..., M$. Using least squares, the linear model is then fitted as

$$y_i = \theta_0 + \sum_{m=1}^{M} \theta_m z_{im} + \epsilon_i, i = 1, ..., n \tag{2.13}$$

The *dimension reduction* approach reduces the problem of estimating the $p+1$ coefficients $\beta_0, \beta_1, ..., \beta_p$ to estimating the $M+1$ coefficients $\theta_0, \theta_1, ..., \theta_M$ where $M < p$. Dimension reduction methods include two steps. In the first step the transformed predictors $Z_1, Z_2, ...Z_M$ are obtained and in the second step, the model is fitted using the M predictors previously obtained. Principal components and partial least squares are two dimension reduction approaches.

Subset selection, shrinkage and dimension reduction are approaches to perform feature selection or variable selection from a regression model. Within the purpose of this manuscript we assess approaches that are not automatically performed in statistical methods, except for lasso, they were not used in this study.

Normalization Transformation and normalization is performed when attributes values refer to different measurement units or have different scales. It is carried out to give all attributes equal weight and also to smooth the outliers. To this aim, attributes are scaled to range within a specific scale (e.g. -1.0 to 1.0, or 0 to 1.0). Different methods for data normalization are available, the most commonly used are: $min - max$ *normalization*, $z - score$ *normalization*, and *normalization* by *decimal scaling* (García et al., 2015).

2.4.3 Impact of data preprocessing procedures on statistical analysis results

Data preprocessing has been acknowledged as a primary task to counteract the negative effects that data glitches may produce on analytical results (Van den Broeck et al., 2005; Dasu, 2013; Gibert et al., 2008). Indeed, various publications indicate that more performing results can be obtained when preprocessing procedures are performed previous to statistical analysis (Crone et al., 2006b; Crone et al., 2006a; Nawi et al., 2013; Kotsiantis et al., 2006; Praveena et al., 2012).

Different published works have demonstrated that results from statistical analysis can be improved when procedures such as imputation of missing values (Farhangfar et al., 2008), outlier processing (Ramezani and Fatemizadeh, 2010), feature selection (Kotsiantis et al., 2006) and normalization (Nawi et al., 2013; Crone et al., 2006a) are applied to preprocess data.

Except for the interesting study published by Farhangfar et al. (2008) most of the studies have assessed a particular preprocessing method for a particular statistical analysis. In the comparative study of Farhangfar et al. (2008), the effect of different missing data imputation methods on six classifiers was assessed. Their results indicate that improvements on classification may vary according to the imputation method used to replace missing values. Their results highlight the importance of conducting comparative studies to define the specific methods necessary to preprocess data for a particular statistical analysis.

Another aspect that is important to mention when preprocessing data is the absence of tools to: (1) measure the accuracy of preprocessing procedures and (2) frame a methodological workflow to preprocess data according to specific statistical analyses.
Aiming at providing tools to measure the impact of data preprocessing procedures Dasu

and Loh (2012) have developed the *statistical distortion* metric. The *statistical distortion* was proposed to measure the effectiveness of data cleaning strategies, it includes three dimensions: glitch improvement, statistical distortion and cost-related criteria. Another interesting approach proposed by Berti-Equille et al. (2015), is the *masking index* which is an indicator for quantifying hidden glitches. However, these metrics do not assess the impact on statistical methods.
Concerning a methodological data preprocessing framework, to our knowledge, there are still no publication on this topic. This remains an interesting topic to be explored which actually motivated my work. Finally, few research has been focused on preprocessing of environmental data. Some approaches that has been published to assess and control the quality of data have been published by Rieger et al. (2010); Alferes et al. (2013); Alferes and Vanrolleghem (2014). And Gibert et al. (2016) have provided guide lines to non-expert user for the selection of pre-processing techniques. Though a robust methodological approach to guide on the selection of preprocessing techniques remains a topic to be explored.

As previously mentioned, presence of glitches is an ubiquitous problem, their detection and processing is an important task to obtain reliable results on data analysis. However there are still many open questions:

- Among the different detection methods, which one is the most suitable one according to the statistical analysis to be applied?
- Which data preprocessing procedure should be applied for a specific statistical analysis and dataset?
- How different are the analysis results when different preprocessing procedures are applied?
- Do preprocessing procedures adversely impact analytical results and conclusions?

In the data preprocessing study that we present in Chapter 3 we intend to answer these questions in general, and also in the particular application use cases of water quality assesment and environmental data analysis.

2.5 Summary

In this chapter, we have presented an overview of the state-of-the art related to three disciplines of my work. The state-of-the art that we present refers to two main topics: data acquisition and data preprocessing. For data acquisition we were interested in the different tools in environmental chemistry and hydrobiology to acquire data specifically for water quality assessment. While in data preprocessing, we were interested in data anomalies that occur in environmental datasets, their detection, and processing.

Related to environmental chemistry, Section 2.2.1 gives a general overview of water quality assessment, later in section 2.2.2 we present important aspects on pollution by emerging pollutants and relevant works related to the assessment of pollution in water. We put an emphasis on the pollution of Mexican rivers by emerging pollutants because they are introduced to the natural effluent without following an specific treatment and they are not included in national monitoring programs. The restricted access to advanced analytical instrumentation is one of the main causes of the limited analysis of emerging pollutants. We present in Section 2.2.2, our bibliographic study related to analytical methods for the quantification of emerging pollutants.

Concerning hydrobiology, we present a state-of-the art on biomonitoring practices (Section 2.3.1), usefulness of macroinvertebrates in bio-assessment practices (Section 2.3.2)

and uncertainties on biomonitoring data (Section 2.3.3). In fact, as we described in our bibliographic study, the use of macroinvertebrates-based biomonitoring metrics are very pertinent for the assessment of Mexican rivers. However, their use is limited and the challenges are vast: description of sampling and analytical protocols, development of local or regional metrics, development of specific taxonomical keys. It is necessary to solve this issues in order to acquire quality biomonitoring data.

Finally, to complete our bibliographic survey in Section 2.4 we present different data aspects related to data anomalies, their detection (Section 2.4.1), procedures to process data anomalies (Section 2.4.2) and impact of data preprocessing on statistical analysis (Section 2.4.3).

CHAPTER 3

Acquisition of Environmental Data

Contents

3.1 Introduction 42
3.2 Description of the study sites 42
3.3 Physico-chemical and chemical data acquisition 45
3.3.1 Sampling of water samples 45
3.3.2 Analysis of major elements, heavy metals, and arsenic in water samples 46
3.3.3 Analysis of organochlorine pesticides and PPCPs in water samples 48
3.4 Hydrobiological data acquisition 53
3.4.1 Sampling of macroinvertebrates 53
3.4.2 Analysis of macroinvertebrates 53

3.1 Introduction

In environmental studies it is essential to have quality analytical results to provide accurate responses or solutions to an specific problem. The quality of analytical results relies mainly on the quality of data therefore, good laboratory and field practices are necessary to ensure good data quality.
The assessment of water quality with respect to the content of emerging pollutants such as pesticides, pharmaceuticals, and personal care products is at its early stage in Mexico. One important problem when analysing these type of pollutants is the elevated cost and the difficulty to analyse them specially in laboratories with limited access to specialized analytical instrumentation. Therefore, development or adaptation of tools as well as the definition of protocols to analyse these type of pollutants are necessary.

Data from biomonitoring water quality assessment is a regular practice in developed countries however, in Mexico the use of biomonitoring metrics for the water quality assessment of rivers is scarce despite their advantages (i.e., easy to implement, cheap). An essential problem is the absence of biomonitoring protocols to sample and analyse biological samples. To acquire quality biomonitoring data is indispensable to define such protocols.

Within the interest to acquire quality data, we propose to control the entire pipeline. This means from the collection of samples on the field and until the analysis of data on the laboratory. We focused on the acquisition of water quality data with respect to emerging pollutants and biomonitoring data. Our aim is to provide a methodological approach for good laboratory practices to acquire quality data.

In this chapter, we describe our methodological approach for the acquisition of water quality data in Mexico with respect to: pollution by pesticides, pharmaceuticals, and personal care products, a biomonitoring using macroinvertebrates. To our purpose we have collected and analysed water and biological (macroinvertebrates) samples from four Mexican rivers (Tula, Culiacan, Humaya and Tamazula). In addition, we have acquired physico-chemical and chemical data.

This chapter is organized as follows: we present the sites of our study in Section 3.2. In Section 3.3, we describe the sampling and analytical methods used for the acquisition of physico-chemical and chemical data. In Section 3.3.3, we present our analytical method for the analysis of organochlorine pesticides. Finally, in Section 3.4, we present the sampling and analytical protocol for the bio-assessment of Mexican river waters using macroinvertebrates.

3.2 Description of the study sites

We have studied four rivers in Mexico the Tula, Culiacan, Humaya, and Tamazula.

The sub-watershed of Tula River is located at the sud-west of Hidalgo State, Mexico. This sub-watershed has 330 km length, semi-arid climate, mean annual temperature of 16 °C, between 110 mm and 1270 mm of rainfall annually (between May and October in storms) and the mean evaporation rate is 42.8 mm^3 (INEGI, 2013).

The Tula river is a water flow system that runs from the State of Mexico to the south-central part of Hidalgo state in Mexico. The Tula river receives the untreated wastewater from Mexico City and from different municipalities in the State of Hidalgo, including south of Zimapán, South-East of Tasquillo, South-West of Ixmiquilpan, Progreso de Obregón,

Mixquihuala de Juárez, Tezontepec de Aldama, Tula de Allende, and Tepeji del Río de Ocampo. As a result of wastewater irrigation, the flow of the Tula river increased from 1.6 to 12.7m^3 s^{-1} in the last decades (DFID, 1998).

Water from the Tula River is used mainly for agricultural purposes. Municipalities of Hidalgo state that use a vast soil extension for agriculture are Mixquihaula de Juárez (81.8%), Tezonepec de Aldama (62.4%), Progreso de Obregón (62.2%), Ixmiquilpan (56%), Tula de Allende (53.2%), Tasquillo (34.9%), and Zimapán (16.5%) (INEGI, 2013).

Five sampling sites (named H1 to H5) placed along the river were selected to collect samples (c.f. Figure 3.1). Sites H2 and H3 are located closed to agricultural fields and were selected to assess the quantities of organochlorine pesticides in the water of the Tula River. Sites H2, H3, and H5 were located near urbanized areas, they were selected to assess the levels of pollutants, mainly PPCPs. Details of sampling site location is given in Table 3.1.

Table 3.1 – Location of sampling sites of the Tula river.

Name of sampling site	Municipality	Latitude	Longitude	Altitude (m)	Focus of the study
H1	Tasquillo	20° 33.703'	099° 18.581'	1761.134	domestic waste pollution
H2	Ixmiquilpan	20° 28.829'	099° 13.277'	1730.654	pesticides from
H3	Ixmiquilpan	20° 29.585'	099° 13.310'	1706.270	agriculture
H4	Tlacotlapilco	20° 22.451	099° 13.414'	1703.222	domestic waste
H5	Progreso	20° 14.696'	099° 12.344'	1702.917	pollution

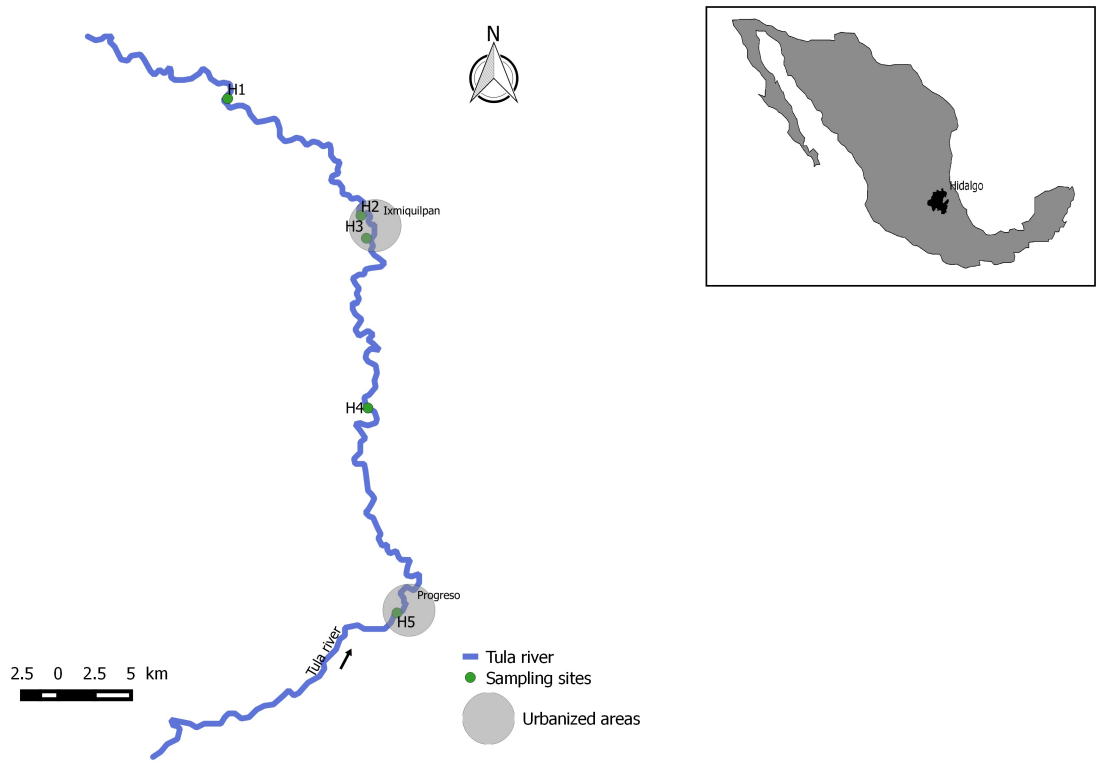


Figure 3.1 – Sampling sites in the Tula river (Hidalgo state, Mexico)

The rivers Tamazula, Humaya, and Culiacan are located in the Pacific coast plain

province of Mexico in the state of Sinaloa, Mexico. Its a semiarid region with seasonal rainfall from June to September (mean annual precipitation 678 mm) and dryness during the rest of the year. This region has a mean annual temperature of 25.7 °C and mean annual evapotranspiration of 2100 mm (INEGI, 2014).

Water flow of Tamazula River starts at the South-East of Durango state and ends at the Culiacan city, it has a length of 152 km and a drainage basin of 3,307 km².
The Humaya River is considered as the main watercourse forming the Culiacan river. It has a total length of 179 km and a drainage basin of 10,770 km².

The Culiacan River begins at the Culiacan city after the union of the rivers Tamazula and Humaya. It runs 82.8 km until it reaches the Gulf of California and on its way, it increases its volume by groundwater recharge (INEGI, 1990). The Culiacan river is the most important perennial river in the region it has a drainage basin of 17,195 km², it receives the treated water from the sewage of Culiacan city and surrounding communities (approximately 855,000 inhabitants) at a rate of 160,206 m^3 day^{-1} (INEGI, 1995). Until 1958, the discharge rate of the Culiacan river used to be 99.5 m^3s^{-1}; however, currently its discharge rate has diminished due, mainly, to its use to irrigate agricultural fields.

Water from the Culiacan river is used for agricultural, industrial, domestic and livestock activities. Though, irrigation of agricultural fields is the main use. The Culiacan valley is the biggest and the most important agricultural zone of Mexico, it comprises more than 4000 Km² of agricultural fields which represents 42% of Sinaloa state. The Culiacan agricultural area has ≈ 140,000 ha with 60% of the land irrigated (INEGI, 1990).

We have selected a total of eleven sampling sites to assess the water quality along the rivers Tamazula, Humaya, and Culiacan (c.f. Figure 3.2). The location of each sampling site is detailed in Table 3.2. We selected four sampling sites along the longitudinal axis of the Humaya river (C1 to C4) to assess the levels of pollutants mainly from domestic wastes. Three sampling sites (C5-C7) were selected on the Tamazula River to represent the least polluted state. Sampling sites C8 and C9 were selected to assess the levels of domestic wastes and anthropogenic activities in the Culiacan River and sites C10 and C11 were chosen to assess pollution by organochlorine pesticides.

Table 3.2 – Location of sampling sites of Tamazula, Humaya and Culiacan rivers.

Name of Sampling site	Municipality	Latitude	Longitude	Altitude (m)	Focus of the study
Humaya river					
C1	Mojolo	24° 50.316'	107° 24.775'	45.110	
C2	Mojolo	24° 50.395'	107° 24.305'	44.805	domestic waste
C3	Mojolo	24° 49.934'	107° 24.227'	46.024	pollution
C4	Humaya	24° 49.045'	107° 24.199'	41.757	
Tamazula river					
C5	Culiacan	24°48.796'	107° 23.755'	62.788	least polluted
C6	Tamazula	24° 49.308'	107° 22.758'	46.329	state
C7	La Limita de Itaje	24° 49.005'	107° 21.630'	50.292	
Culiacan river					
C8	Culiacan	24° 48.551'	107° 24.767'	45.415	anthropogenic
C9	Culiacan	24° 47.524'	107° 26.873'	33.528	activities
C10	Bacurimi	24° 47.848'	107° 30.462'	33.223	pesticides from
C11	Culiacancito	24° 48.409'	107° 31.883'	25.298	agriculture

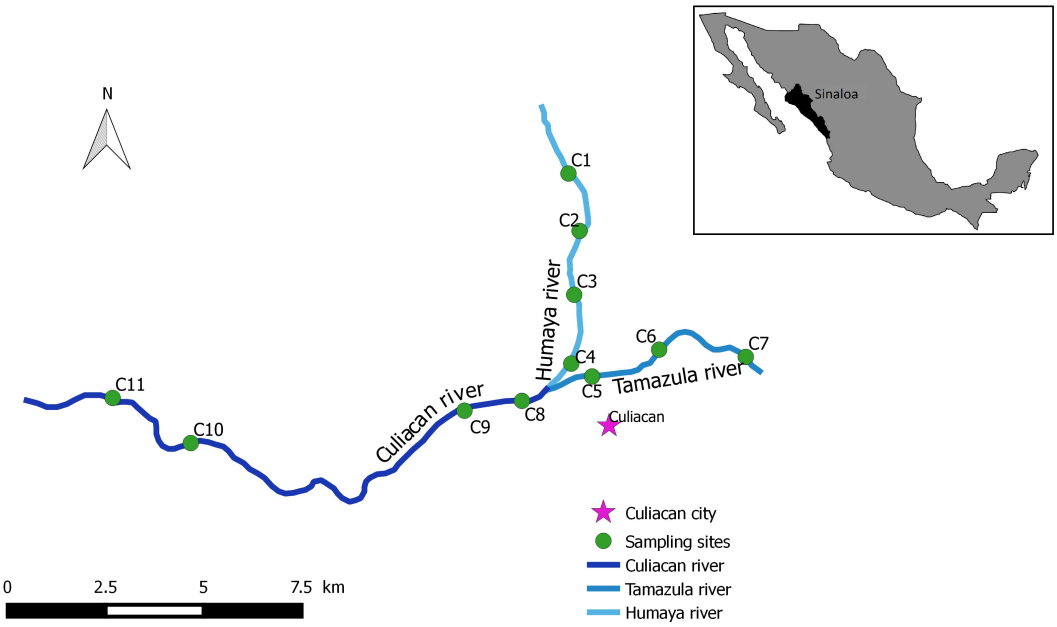


Figure 3.2 – Sampling sites in the Tamazula, Humaya, and Culiacan rivers (Sinaloa state, Mexico).

3.3 Physico-chemical and chemical data acquisition

3.3.1 Sampling of water samples

Due to a limited budget and time we did the sampling campaign only for one year and two seasons (dry and post-rainy). Five sampling stations were sampled on April 2015 (dry season) and October 2014 (post-rainy season) at the Tula River. While 11 sampling stations were sampled on January 2015 (dry season) at the Tamazula, Humaya, and Culiacan rivers. The sampling sites described in the previous section were chosen according to accessibility, ease of recognition in the field and ecological relevance.

Water samples were collected to determine contents of: major elements pH, conductivity, carbonate (CO_3), bicarbonate (HCO_3), sulfate (SO_4), chloride (Cl), fluoride(F), sodium (Na), potassium (K), calcium (Ca), magnesium (Mg), boron (B), silicon dioxide (SiO_2), nitrate (NO_3) arsenic (As), heavy metals cadmium (Cd), lead (Pb), iron (Fe), copper (Cu), manganese (Mn), zinc (Zn), eighteen organochlorine pesticides, namely: I-BHC, II-BHC, III-BHC, IV-BHC, heptachlor, aldrin, heptachlor epoxide, I-endosulfan, II-endosulfan, dieldrin, DDE, endrin, DDD, endrin aldehyde, endosulfan sulfate, DDT, endrin ketone and methoxychlor, and eight PPCPs (ibuprofen, 2-benzyl-4-chlorophenol, naproxen, triclosan, ketoprofen, diclofenac, bisphenol A and estrone).

For analysis of major elements, one liter of water was collected and kept at 4°C until analysis. For analysis of heavy metals and As 500 mL of water were collected and added HNO_3. Polypropilene bottles used were previously rinsed first with 30% HNO_3 and then three times with deionized water. Samples were taken by duplicates.

One liter of water was collected by duplicate in amber glass bottles for analysis of pesticides, while 500 mL were collected for analysis of PPCPs. Samples were kept at 4°C and analysed right after collection. Glass bottles were washed and rinsed 5 times with

deionized water then, they were placed in a container with HCl 10% during 2 hours. After that time, bottles were rinsed with deionized water and dried. Finally, dried glass bottles were rinsed first with acetone and then with hexane. The above mentioned conditions of cleaning were carry out to reduce pollution effects produced by the presence of external compounds that may not be present in the sampling sites and to get the best recovery of all pesticides under study.

3.3.2 Analysis of major elements, heavy metals, and arsenic in water samples

Physico-Chemical Parameters

Temperature, pH, conductivity and redox potential (Eh) parameters were determined at each sampling station using an ORION handheld Multiparameter for the determination of temperature, pH, and Eh. A portable conductometer conductronic PC-18 was used to determine conductivity.

Determination of major elements, heavy metals and arsenic concentrations

Determination of major species was done following the protocol of Armienta et al. (1987). Alkalinity (CO_3, HCO_3) was determined by titration, boron by carmine method, Ca and Mg by titration with EDTA, Cl and F by potentiometry, SiO_2 through the molybdosilicate method, sulfate by turbidimetric method, Na, and K by Atomic Absorption Spectrometry (AAS), and NO_3 by HPLC.

Analysis of arsenic was performed using atomic absorption hydride generation with a Perkin Elmer AAnalyst 100 Spectrometer and FIAS 100. Analysis of lead was done using an Atomic Absorption Spectrometer Perkin Elmer AAnalyst 100 with graphite furnace. And analysis of Cd, Fe, Cu, Mn, and Zn was performed by Atomic Absorption Spectrometry with a Perkin Elmer AAnalyst 2000 Spectrometer.

The physico-chemical and chemical data that we acquired is composed of 16 individuals (sampling sites) and 21 variables (physicochemical and chemical parameters) of numerical values. Only the sites from the Tula river where sampled for two seasons (dry and rainy). An example of data obtained from these analysis is presented on Tables 3.3 and 3.4.

Table 3.3 – Example of data from the physochemical and chemical values obtained from the analysis of water samples of the rivers Tula. Tamazula. Humaya and Culiacan.

Sample	Season	pH	Ω μS/cm	CO_3	HCO_3	SO_4	Cl	F
H1		7.82 ± 0.06	1777 ± 11	44.96 ± 1.55	447.65 ± 6.30	204.07 ± 0.375	206 ± 2.5	0.76 ± 0.0095
H2		7.67 ± 0.06	1576 ± 12	–	554.83 ± 0	146.1 ± 1.205	175.5 ± 5	0.749 ± 0.009
H3		7.73 ± 0	1578 ± 5	37.21 ± 0	472.87 ± 0.005	137.1 ± 4.04	163.25 ± 0.25	0.773 ± 0.003
H4		7.79 ± 0.07	1567.5 ± 12.02	41.86 ± 6.58	471.29 ± 15.60	156.4 ± 3.260	166.75 ± 5.30	0.766 ± 0.0064
H5		7.7 ± 0.01	1564.5 ± 0.5	40.31 ± 0	488.63 ± 3.15	130.68 ± 1.4	169.5 ± 1	0.758 ± 0.002

ND = Non detected. Concentrations are given in mg/L. Each sample was analyzed by triplicates for each parameter.

Table 3.4 – Example of the data from the concentration of arsenic and heavy metals on water samples of the rivers Tula, Tamazula, Humaya and Culiacan.

Sample	Saison	Cd	As	Fe	Cu	Mn	Zn	Pb
H1	Dry	0.017 ± 0.0003	0.00949 ± 0.00079	0.14 ± 0.005	ND	0.12 ± 0	ND	0.028 ± 0.0005
H2		0.015 ± 0.0005	0.01009 ± 0.00039	0.08 ± 0	ND	0.11 ± 0	ND	0.024 ± 0
H3		0.016 ± 0	0.00907 ± 0.00004	ND	ND	0.11 ± 0	ND	0.026 ± 0.0015
H4		0.016 ± 0.014	0.00973 ± 0.00070	0.13 ± 0	ND	0.12 ± 0.005	ND	0.031 ± 0.0071
H5		0.015 ± 0.00005	0.00954 ± 0.00037	ND	ND	0.1 ± 0	ND	0.024 ± 0.001

ND = Non detected. Concentrations are given in mg/L. Each sample was analyzed by triplicates for each element.

3.3.3 Analysis of organochlorine pesticides and PPCPs in water samples

Quantification of pollutants in water at low concentrations (in the range of $\mu g\,L^{-1}$ and $ng\,L^{-1}$), has been a challenging problem in water quality assessment studies. A vast number of techniques have been developed to overcome this problem however, most of them include the use of advanced analytical tools (i.e., ICP-MS, GC-MS, LC-MS or LC-MS-MS). In laboratories where such tools are not available it is necessary to adapt or develop new tools that are cheaper and easy to use without compromising the quality of data. In Mexico monitoring of pesticides is scarce partially due to the limited access to specialized analytical instrumentation. To provide an accessible tool for the analysis of pesticides in water we have adapted an analytical method based on Solid Phase Extraction (SPE) followed by Gas-Chromatography with an Electron Capture Detector (GC-ECD) for the quantification of organochlorine pesticides.

We adapted the method described by Guardia Rubio et al. (2007). Analysis was performed using a solid phase extraction (SPE) and gas chromatography with electron capture detection (GC-ECD). Contrary to the method proposed by Guardia Rubio et al. (2007), we used ethyl acetate-hexane 25:75%, v/v to dissolve the extract obtained from the solid phase extraction. We have performed an analysis using different combinations of the solvents hexane, cyclohexane and ethyl acetate-hexane, and the ethyl acetate-hexane provided the best results. The solid phase extraction step that we proposed in our method was compared to a liquid-liquid extraction (LLE) method which was previously used in the laboratory. With our methodological approach we obtained better recoveries than with the LLE (recovery ranging from 15% to 40%). Details of the analytical performance of the proposed methodology are given in Table 3.5.

Table 3.5 – SPE-GC-ECD limit of detection (LOD), limit of quantification (LOQ) and recovery values for water analysis.

No.	**Pesticide**	**RT(min)**	**SD**	**LOD** ($ng\,L^{-1}$)	**LOQ** ($ng\,L^{-1}$)	**Recovery (% ± RSD)**
1	I-BHC	9.01	0.15	0.43	1.54	109.07 ± 0.55
2	II-BHC	9.86	0.30	0.82	2.96	143.42 ± 0.40
3	III-BHC	10.02	0.12	0.32	1.16	128.18 ± 0.49
4	IV-BHC	10.85	0.17	0.48	1.73	177.67 ± 0.22
5	Heptachlor	12.22	0.71	1.97	7.08	48.53 ± 0.63
6	Aldrin	13.34	0.34	0.95	3.42	24.93 ± 0.69
7	Heptachlor epoxide	14.64	0.85	2.36	8.48	64.80 ± 0.38
8	I-Endosulfan	15.83	1.18	3.27	11.80	62.74 ± 0.41
9	Dieldrin	16.75	0.54	1.51	5.44	17.70 ± 0.67
10	DDE	16.86	0.54	1.51	5.44	17.70 ± 0.67
11	Endrin	16.86	0.76	2.12	7.63	50.38 ± 0.40
12	II-Endosulfan	17.68	0.70	1.95	7.02	89.66 ± 0.48
13	DDD	18.01	0.01	0.03	0.11	79.38 ± 0.42
14	Endrin aldehyde	18.23	1.46	4.04	14.56	35.27 ± 0.47
15	Endosulfan sulfate	18.56	0.22	0.61	2.20	120.75 ± 0.29
16	DDT	19.18	0.52	1.46	5.24	105.47 ± 0.61
17	Endrin ketone	20.07	0.10	0.28	1.01	132.72 ± 0.39
18	Methoxychlor	20.24	0.74	2.05	7.37	91.82 ± 0.90

Spiked level of water $500\,ng\,L^{-1}$.
Replicates $n = 3$.
RT=Retention time.
SD = Standard deviation.
RSD = Residual standard deviation.

Analysis of PPCPs was performed using a new method based on SPME and GC-MS. The method used for PPCPs analysis was tested and developed in collaboration with a

group of researchers of the Faculty of Chemistry at UNAM, Mexico under the direction of Dr.Araceli Peña A. The new SPME-GC-MS that we proposed is a simple, rapid, efficient, and sensitive alternative for the simultaneous analysis of PPCPS (ibuprofen, 2-benzyl-4-chlorophenol, naproxen, triclosan, ketoprofen, diclofenac, bisphenol A and estrone) in water at trace levels ($ng\,L^{-1}$).
In this manuscript, we only describe the procedure followed for the analysis of organochlorine pesticides. Details about the SPME-GC-MS method for PPCPs analysis on water is presented by Diaz-Flores et al.(under revision, Determination of pharmaceuticals and personal care products (PPCPs) in river water and sediment by solid phase extraction followed by gas chromatography-mass spectrometry (SPME-GC-MS), submitted to *Analytical Chemistry*). In Table 3.6 we provide details of the analytical performance of the SPME-GC-MS method for the analysis of PPCPs.

Table 3.6 – SPME-GC-MS limit of detection (LOD), limit of quantification (LOQ) and recovery values for water analysis.

No.	**Pesticide**	**LOD** ($ng\,L^{-1}$)	**LOQ** ($ng\,L^{-1}$)	**Recovery (% ± RSD)**	**Recovery (% ± RSD)**
1	Ibuprofen	0.4	1.1	94.1 ± 3.4	90.6 ± 3.0
2	2-benzyl-4-chlorophenol	0.4	1.1	91.2 ± 2.3	96.1 ± 2.4
3	Naproxen	0.7	2.0	77.4 ± 2.1	88.9 ± 1.7
4	Triclosan	0.5	1.6	90.6 ± 3.7	97.2 ± 3.8
5	Ketoprofec	0.4	1.2	85.6 ± 4.1	96.2 ± 8.7
6	Diclofenac	4.2	12.8	57.5 ± 0.4	54.4 ± 2.5
7	Bisphenol A	1.4	4.1	93.2 ± 1.4	84.0 ± 3.5
8	Estrone	1.5	4.6	67.5 ± 1.1	76.0 ± 3.6

Spiked level of water $20\,ng\,L^{-1}$.
Spiked level of water $90\,ng\,L^{-1}$.
Replicates $n = 3$.
RSD = Residual standard deviation.

The new methodological approaches that we proposed allowed to: (1) simultaneously detect eighteen organochlorine pesticides (SPE-GC-ECD) and eight PPCPs (SPME-GC-MS), (2) perform analysis in a shorter period of time, (3) reduce sample manipulation which reduces errors on data and, (4) reduce volume of solvents to be used.

Solid Phase Extraction

Solid Phase Extraction was performed for the extraction of pesticides in water samples, the procedure followed is as follows. Samples were first filtered under vacuum through a 0.45 μm Millipore filter to eliminate particulated material until the sample remained transparent. C18 cartridge were placed in a 12-port Visiprep SPE vacuum manifold and conditioned by passing 5 mL of dichloromethane, 5 mL of methanol and 5 mL of Milli-Q water at a flow rate of $1\,mL\,min^{-1}$. Then, 500 mL of water sample was passed through the cartridge at a flow rate of $1\,mL\,min^{-1}$. The solid phase in the cartridge was not allowed to become dry at any moment. After loading the sample, the SPE cartridge was dried for 20 min under vacuum. Pesticides were eluted using 2 mL of dichloromethane by gravity and finally under vacuum. The extract was brought to dryness in a nitrogen stream and redissolved in 1 mL of ethyl acetate-hexane (25 : 75%, v/v). Analyses of a spiked blank sample and a blank sample were performed together with every set of samples. Figure 3.3 schematizes the SPE procedure for extraction of OCPs in water samples.

Solid Phase Micro Extraction

Solid Phase Micro Extraction (SPME) was performed for the extraction of PPCPs in water samples, SPME was carried out using 85 μm poliacrylate fibers (PA) supported on a manual device for SPME (Bellefonte, PA, USA). PA fibers were previously conditioned

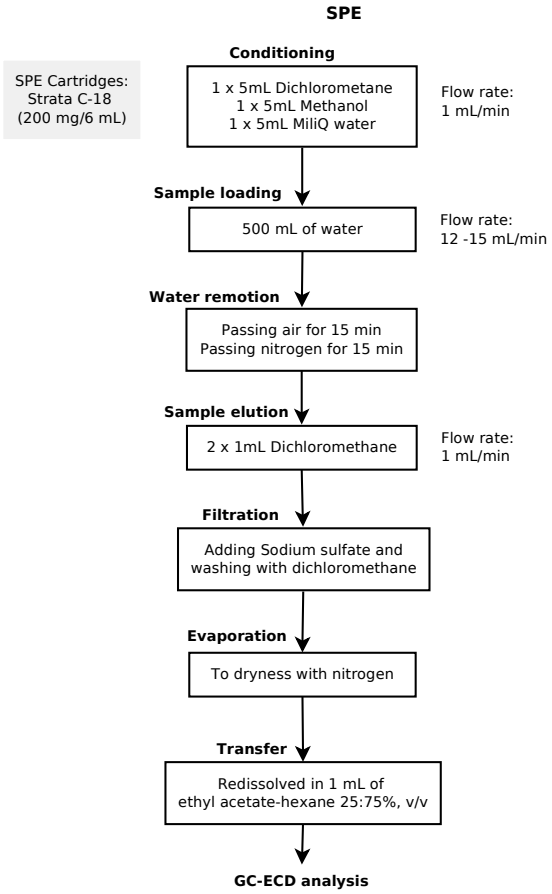


Figure 3.3 – Flow diagram of the SPE method for analysis of OCPs in water.

directly on the GC injector at 280 °C for 1 h.

Water samples were firstly filtered under vacuum through a 0.45 μm Millipore filter. Then, samples were fortified with analytes at (1000 $ng l^{-1}$), 5 mL of water were transferred to a glass vial and adjusted to pH 3 with 0.1% concentrated formic acid(89%), 0.54 g sodium chloride were added and a stir bar was introduced. The glass vial was closed and placed on a water bath at 40°C and maintained 5 min for temperature equilibrium. The analytes were extracted by SPME in immersion mode for 45 min at 40 °C and 1200 rpm. Following SPME, the analytes were derivatized on head-space mode by exposing the fiber to the vapors of 100 µL of N,N-tertbutyldimethylsilil-N-methyltrifluoroacetamide (MTBSTFA) in order to obtain the volatile derivatives. After derivatization, the fiber was retracted and inserted directly on the GC injector. Analytes desorption was carried out at 250 °C for 10 min.

Analytical instrumentation and operating conditions

An Agilent Technology 6890N gas chromatograph equipped with an ECD detector was used for pesticide quantification. The chromatograph was equipped with an autosampler and split/splitless injector. The column used was a capillary column (Phenyl Methyl Siloxane 30m×250μm×0.25μm i.d.), with nitrogen as carrier gas at a constant flow of 1 $mL min^{-1}$.
A 1 µL aliquot of the extract was injected in a splitless mode into the gas chromatograph. After each injection, syringe was first rinsed with acetone and then with ethyl acetate-hexane (25:75%, v/v). Temperature programme was the following: initial temperature 80 °C, held for 1 min, 30 °C min^{-1} ramp to 175 °C then 6 °C min^{-1} to 215 °C, finally by 15 °C min^{-1} to 290 °C and held for 5 min.

GC-MS: A Hewlett Packard HP 5890A gas chromatograph equipped with a 5973 mass spectrometer (Agilent Technologies, U.S.A.) was used for PPCPs analysis. The chromatographic column used was a ZB-5 (30 m x 0.25 mm i.d. x 0.25 µm film thickness), with helium as carrier gas at a constant flow of 1 $mL min^{-1}$ and injector temperature of 250°C.
Column temperature program was 70 °C (1 min), then increased at 10 °C min-1 to 230 °C (4 min), then 14 °C min-1 to 300 °C (6 min); total running time 33 min. The sample was injected in splitless mode (1 min). Mass spectrometric detection was carried out by single ion monitoring (SIM), selecting two fragment ions for identification and quantification of each analyte: ibuprofen 263 y 320 m/z; naproxen 327 y 324 m/z; diclofenac 214 y 352 m/z; ketoprofen 295 y 311 m/z; 2-benzyl-4-chlorophenol 275 y 332 m/z; triclosan 200 y 375 m/z; bisphenol A 442 y 456 m/z; estrone 328 y 384 m/z. Quantification was carried out by external standardization by integrating area of the chromatographic peaks. Peak identity was confirmed with retention time and mass spectrum of each analyte, where characteristic fragment ions were selected.

With the use of analytical methods described above we were able to analyse for the first time the content of organochlorine pesticides and PPCPs in the Mexican rivers (Tula, Tamazula, Humaya and Culiacan). Data from the analysis of PPCPs and organochlorine pesticides is composed of 16 individuals (sampling sites) and 23 variables (5 PPCPs and 18 Oragnochlorine Pesticides) of numerical values. An example of these data is presented on Tables 3.7 and 3.8.

Table 3.7 – Example of data from the analysis of organochlorine pesticides on water samples from the Tula river. Only ten out of the six-teen compounds are shown.

Season	Sample	II-Endosulfan	DDD	Endrin aldehyde	Endosulfan sulfate	DDT
Dry	H1	0.777 ± 0.002	3.775 ± 2.436	0.797 ± 0.564	1.072 ± 0.033	0.228 ± 0.126
	H2	0.554 ± 0.202	0 ± 0	0.680 ± 0.126	1.103 ± 0.069	0.453 ± 0.138
	H3	0.282 ± 0.250	1.240± 1.23	0 ± 0	0.963 ± 0.127	0.492 ± 0.243
	H4	0.276 ± 0.172	0 ± 0	0.480 ± 0.043	0.943 ± 0.099	0.168 ± 0.180
	H5	0.088 ± 0.098	0 ± 0	0.608 ± 0	1.040 ± 0.026	0.052 ± 0.088

ND = Non detected. Concentrations are given in ng/L. Each sample was analysed by duplicates for each compound.

Table 3.8 – Example of data from the analysis of PPCPs on water samples from the Tula river.

Sample	Season	Ibuprofen	Naproxen	Triclosan	Diclofenac	Bisphenol A
H1	dry	59±2.1	160.1±8.4	31.4±1.1	–	–
H2		79.7±3.8	240.3±10.9	29.8±1.1	–	–
H3		76.1±5.0	212.5±10.7	31.1±0.6	–	–
H4		79.6±0.3	246±1.0	28.6±1.3	–	–
H5		100.3±1.0	101.8±2.6	29.7±0.6	–	–

ND = Non detected. Concentrations are given in ng/L. Each sample was analysed by triplicates for each compound.

3.4 Hydrobiological data acquisition

As explained in Section 2.3 currently, in Mexico, the use of macroinvertebrates for ecological assessment of aquatic systems is scarce due partially to the absence of biomonitoring protocols and undefined metrics. Aiming at providing guidelines and practical tools for the biomonitoring of aquatic systems, we have done a study whose objectives were: 1) identify the macroinvertebrate-based monitoring approaches with potential application to the ecological assessment of Mexican streams, and 2) describe sampling and analytical procedures necessary to implement such approaches.
Results of this study have been published (Serrano Balderas et al., 2016) and are part of the work of this manuscript. Briefly, thirty-five biomonitoring metrics were identified as potential approaches to be applied for the ecological assessment of Mexican streams. A description of the sampling and analytical procedures to compute the selected metrics is also given in the above mentioned publication.
We have applied for the first time the guidelines described by Serrano Balderas et al. (2016) in order to acquire biological data for Tula, Tamazula, Humaya, and Culiacan rivers. Below we described the sampling and analytical procedures followed for this purpose.

3.4.1 Sampling of macroinvertebrates

Macroinvertebrate sampling was performed first by a visual inspection of a 100 m stretch of the river channel to identify the different habitats and substrates. Samples were collected from multi-habitats as suggested by Jáimez-Cuéllar et al. (2004), using a surber net (surface 500 cm^2, mesh size 500μm). For each sampling point, 2 replicates were obtained. Samples were stored in plastic bags and preserved in 70% ethanol.

3.4.2 Analysis of macroinvertebrates

In laboratory, macroinvertebrate samples were separated from vegetable and mineral substrates and sorted using a 250 μm mesh sieve. Macroinvertebrates were examined under a stereoscope and identified at family level using different taxonomic keys (Heckman, 2006; Heckman, 2008; Heckman, 2011; Merrit et al., 2008a; Tachet et al., 2010; Novelo-Gutiérrez, 1997b; Novelo-Gutiérrez, 1997a). Thirty-five biomonitoring metrics were calculated. They include: 5 metrics of richness, 11 enumeration metrics, 6 diversity and similarity indices, 7 biotic indices , 5 functional feeding metrics and 1 multimetric approach. For the case of similarity indices, they were computed by comparing two stations. In fact, during the sampling campaign one station was selected to represent the least polluted state of the river, this station was located far from anthropogenic and urban activities and upstream. This station considered as "non-polluted" or "the least polluted" was used to compute similarity indices by comparing it against the other stations.

Details concerning the taxonomic resolution, definition, and expected response to increasing perturbation for each metric are given in Table 3.9.

Table 3.9 – Biomonitoring metrics for the biological assessment of Mexican rivers (Serrano Balderas et al., 2016).

Metric		Taxonomic resolution	Definition	Expected response to increasing perturbation	Reference
Measures of richness	Number of total taxa	Family	All different taxa at a site.	Decrease	Resh and Jackson, 1993;Barbour et al., 1999
	Number of EPT taxa		Total number of taxa of the orders Ephemeroptera (mayflies), Plecoptera (stoneflies) and Trichoptera (caddisflies).		
	Number of Ephemeroptera taxa	Genus or Species	Total number of taxa of the order Ephemeroptera (mayflies).		
	Number of Plecoptera taxa		Total number of taxa of the order Plecoptera (stoneflies).		
	Number of Trichoptera taxa		Total number of taxa of the order of Trichoptera (caddisflies).		
Enumerations	Number of families in common	Family	Number of families in common between 2 samples.		
	%EPT		Ratio of EPT abundance.		
	%Ephemeroptera		Ratio of mayflies to total number of individuals.		
	%Plecoptera		Ratio of stoneflies to total number of individuals.		
	%Trichoptera		Ratio of caddisflies to total number of individuals.		
	%Coleoptera		Ratio of individuals of the order of Coleoptera to total number of individuals.		
	%Diptera		Ratio of individuals of the order of Diptera to total number of individuals.	Increase	
	%Chironomidae		Ratio of chironomidae individuals to total number of individuals.		
	EPT / Chironomidae		Ratio of EPT abundance to chironomidae abundance.	Decrease	
	% of most dominant genus	Genus	Ratio of individuals in numerically dominant genus to total numbers of individuals.	Increase	
	% of dominant taxa	Family	Ratio of individuals in numerically dominant taxa to total number of individuals.		
Diversity and Similarity indices	Shannon's Index		Description of community structure (Diversity).	Decrease	Shannon, 1948
	Simpson's Index		Diversity index commonly used in Ecology and Biology.	Increase	Simpson, 1949
	Margalef Index		Diversity index, where the number of species in the sample and the total number of organisms in the sampling are considered.	Decrease	Margalef, 1951
	Sequential Comparison Index (SCI)	Order or Family	Description of stream quality method, based upon distinguishing the number of different types of organisms and the number of "runs".		Cairns et al., 1968
	Jaccard's Coefficient	Family	Description of the similarity between two samples.		Jaccard, 1901
	Sørensen Coefficient		Description of the similarity between two samples.		Sørensen, 1948
Biotic indices	Trent Biotic Index (TBI)	Family (genus for Plecoptera and Ephemeroptera nymphs and species-level for Annelida, Mollusca, Crustacea and Megaloptera)	Description of the pollution level of streams, according to the sensitivity of key groups to pollution.		Woodiwiss, 1964; Metcalfe, 1989
	Extended Biotic Index (EBI)	Family or Genus	Description of the pollution level of streams, according to the sensitivity of key groups to pollution. Aquatic ecosystems are described using "quality classes".		Ghetti, 1997
	Beck Biotic Index (BI)	Genus	Classification of streams according to their organic pollution. Organisms are classed according to their tolerance to organic pollution.		Beck, 1955
	Family Biotic Index (FBI)	Family	Uses tolerance values to weight abundance in an estimate of overall pollution. Originally designed to evaluate organic pollution.	Increase	Hilsenhoff, 1982; Hilsenhoff, 1988; Plafkin et al., 1989

Table 3.9 – Biomonitoring metrics for the biological assessment of Mexican rivers (Serrano Balderas et al., 2016). (continued...)

	Metric	Taxonomic resolution	Definition	Expected response to increasing perturbation	Reference
	Biological Monitoring Working Party (BMWP)		Score system using benthic macroinvertebrate classed according to their pollution tolerance for Britain rivers.	Decrease	ISO, 1979; National Water Council, 1981
	Biological Monitoring Working Party Costa Rica (BMWP CR)		Score system based on the BMWP modified to include macroinvertebrate communities of Costa Rica		Gutiérrez-Fonseca and Lorion, 2014
	Biological Monitoring Working Party Average Score per Taxon (BMWP ASPT)		Average Score Per Taxon, added to the BMWP Score System.		Armitage et al., 1983
Functional Feeding measures	%Filterer collectors		Percent of the macrobenthos that filter fine organic material.	Increase	Merrit et al., 2008b
	% Scrapers		Percent of the macrobenthos that feed on epiphytes.	Decrease	
	%Shredder		Percent of the macrobenthos that feed on leaf litter.		
	%Predators		Percent of the predator functional feeding group. They are carnivores-scavangers, engulf or pierce prey.		
	% gathering collectors		Percent of the macrobenthos that collect fine deposited organic material.	Variable	
Multimetric approach	IBI-west central Mexico		Computation of multimetric approaches which is summarized in a single value.	Decrease	Weigel et al., 2002

The taxonomic level indicated for the diversity and similarity indices is the level preconceived by the author of each metric. However, they can be computed using some other taxonomic level (e.g., species-level).

Metric of richness is the number of specimens found in each sample site, the richness metrics calculated includes: number of total taxa, number of EPT taxa, number of Ephemeroptera taxa, number of Plecoptera taxa, and number of Trichoptera taxa.

Enumerations were computed by estimating the relative abundance of certain pollution-sensible groups to the total abundance of macroinvertebrates they consisted of: number of families in common, %EPT, %Ephemeroptera, %Plecoptera, %Trichoptera, %Coleoptera, %Diptera, %Chironomidae, EPT:Chironomidae, % of most dominant genera, and % of dominant taxa.

Four diversity indices (Shannon's Index, Simpson's Index, Margalef Index, and Sequential Comparison Index) and two similarity indices (Jaccard's Coefficient and Sørensen Coefficient) were computed. The formulaes of three diversity indices are given below:

$$Shanon\_Index : H = \sum_{i=1}^{i=S} p_i \log p_i \tag{3.1}$$

$$Simpson\_Index : D = \frac{\sum_{i=1}^{i=S} n_i(n_i - 1)}{N(N-1)} \tag{3.2}$$

$$Margalef\_Index : D = S - \frac{1}{\log N} \tag{3.3}$$

Where p_i is the proportion of individuals of the ith taxon ($p_i = n_i/N$), n_i is the total number of individuals of the ith taxon, N is the total number of individuals for all ith taxa and S is the total number of taxa.

Sequential Comparison Index (SCI) was computed using Equation 3.4 (Cairns et al., 1968). The procedure to estimate the SCI is as follows: Let us consider a sample containing

macroinvertebrates. These macroinvertebrates are placed in a container for examination (c.f., Figure 3.4). Counting from left to right in row A, we can observe that the first four organisms are similar and, therefore, are part of the same run. Organisms six and seven are similar but both are different compared to organism five and, therefore, the last two organisms are part of a new run. In row A there are 3 runs for 7 organisms. Compared to row A, row B presents 5 runs for 10 organisms.

$$SCI = \frac{number\ of\ runs}{number\ of\ organisms} \tag{3.4}$$

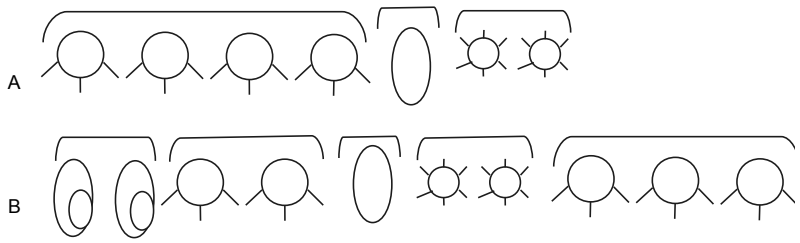


Figure 3.4 – Example of macroinvertebrates counting process to determine the number of runs in a sample. Each form represents a different organism. The number of runs is used to compute the Seqential Compraison index (SCI).

The Jaccard's (Jaccard, 1901) and (Sørensen, 1948) coefficients are similarity indices used to evaluate the mean differences between sampling stations. They use presence-absence data and were computed using the following equations:

$$Jaccard\_Coefficient : C_j = \frac{j}{a+b-j} \tag{3.5}$$

$$Sørensen\_Coefficient : C_s = \frac{2j}{a+b} \tag{3.6}$$

where j is the number of taxa in common between the stations A and B, a is the number of all taxa in station A and b is the number of all taxa in station B.

Biotic indices and the Macroinvertebrate-based Index of Biotic Integrity (IBI) for west-central Mexican streams were computed according to authors description. Biotic indices include: Trent Biotic Index (Metcalfe, 1989), Extended Biotic Index (Ghetti, 1997), Beck Biotic Index (Beck, 1955), Family Biotic Index(Hilsenhoff, 1988), Biological Monitoring Working Party (National Water Council, 1981) and its associated Average Score per Taxon (Armitage et al., 1983). Details on the computation of each metric are given in Appendix A.

Six Functional Feeding Groups were computed including: filtering collectors, scrapers, shredders, predators, and gathering collectors. Each organism was counted and grouped in feeding groups according to Merrit et al. (2008a).

Hydrobiological data from the rivers Tula, Tamazula, Humaya, and Culiacan is composed of 16 individuals (sampling sites) and 35 numeric parameters (35 biomonitoring metrics). Only the Tula river contains information related to two sampling periods (dry and rainy). Below we provide an example of data from biomonitoring of the rivers Tula, Tamazula, Humaya and Culiacan.

Table 3.10 – Example of data from the biomonitoring metrics computed for the rivers Tula, Tamazula, Humaya and Culican.

Metric		Sampling sites				
		H1	H2	H3	H4	H5
Measures of richness	Number of total taxa	710	2684	11419	26636	9762
	Number of EPT taxa	0	0	0	0	0
	Number of Ephemeroptera taxa	0	0	0	0	0
	Number of Plecoptera taxa	0	0	0	0	0
	Number of Trichoptera taxa	0	0	0	0	0
	Number of families in common	5	6	5	4	8
Enumerations	%EPT	0	0	0	0	0
	%Ephemeroptera	0	0	0	0	0
	%Plecoptera	0	0	0	0	0
	%Trichoptera	0	0	0	0	0
	%Coleoptera	0	0	0	0	0,0102
	%Diptera	6,0563	4,8808	0,7531	0,1840	1,9975
	%Chironomidae	5,9155	4,8435	0,7531	0,1727	1,9975
	EPT : Chironomidae	0	0	0	0	0
Diversity and Similarity indices	% of most dominant genera	81,55	91,62	98,65	89,43	95,16
	% of dominant taxa	81,41	90,91	98,64	89,43	95,16
	Shannon's Index	0,6686	0,4097	0,0855	0,3504	0,2600
	Simpson's Index	0,3227	0,1704	0,0269	0,1896	0,0937
	Margalef Index	1,5232	1,2666	1,0703	0,9814	1,0886
	Sequential Comparison Index (SCI)	0,0070	0,0022	0,0005	0,0002	0,0009
	Jaccard's Coefficient	0,5556	0,6667	0,6250	0,4000	1
	Sørensen Coefficient	0,7143	0,8000	0,7692	0,5714	1
Biotic indices	Trent Biotic Index (TBI)	3	3	3	3	3
	Extended Biotic Index (EBI)	5	5	4	5	5
	Beck Biotic Index (BI)	1	2	1	2	3
	Family Biotic Index (FBI)	4,5775	4,2973	4,0375	7,5823	7,9045
	Biological Monitoring Working Party (BMWP)	15	23	15	17	21
	Biological Monitoring Working Party (BMWP-CR)	22	23	16	21	24
	Biological Monitoring Working Party- Average Score per Taxon (BMWP- ASPT)	3,00	3,29	3,00	3,40	3,00
Functional feeding measures	%Filterer collectors	0,0000	0,0373	0	0,0038	0
	% Scrapers	0	0,0373	0	0	0,0819
	%Shredder	81,5493	91,6170	98,6513	10,3394	1,9461
	%Predators	2,1127	0,8197	0,2102	0,0563	0,8194
	% gathering collectors	16,3380	7,4888	1,1384	89,6005	97,1525
Multimetric approach	IBI-west central Mexico	30	35	40	10	10

CHAPTER 4

Preprocessing and Analysis of Environmental Data

Contents

4.1 Introduction . . . 60
4.2 Data preprocessing procedures . . . 62
4.2.1 Feature selection . . . 62
4.2.2 Normalization . . . 62
4.2.3 Imputation methods . . . 63
4.2.4 Outlier detection methods . . . 66
4.3 Synthetic data description . . . 69
4.4 Semi-synthetic data description . . . 73
4.5 Robustness study of data preprocessing procedures . . . 75
4.5.1 Selection of features . . . 75
4.5.2 Normalization of data . . . 76
4.5.3 Handling missing data . . . 76
4.5.4 Handling outlying data . . . 77
4.5.5 Results and discussion on robustness of data preprocessing procedures . . . 77
4.6 Study of the impact of data preprocessing procedures on statistical results . . . 80
4.6.1 Impact on regression results . . . 80
4.6.2 Impact on classification results . . . 81
4.6.3 Impact on clustering results . . . 82
4.6.4 Results and discussion about preprocessing procedures on statistical results . . . 85
4.7 Summary and concluding remarks . . . 114

4.1 Introduction

Data Mining (DM) methods include non supervised methods (i.e., clustering, PCA, classifiers) and supervised methods (i.e., regression, classification), through the use of such methods, we can discover relevant patterns in the data and gain more knowledge. To use DM methods it is necessary to provide input data in the amount, structure and format that suits DM methods perfectly. However, real-world data such as data from environmental surveys are highly affected by anomalies such as missing values, inconsistencies, inaccuracies, outlying data, etc. To provide high quality data and thus quality analytical results it is necessary to preprocess data.
Data preprocessing is a critical step in data mining processes. Results from data preprocessed inappropriately may lead to misleading conclusions and, in the worst case, dramatic consequences and wrong decision making that can affect the survival of certain environmental ecosystems in our application domain.
We aim at providing to data and environmental scientists a guide to inspect, preprocess and analyse environmental data. We focused our study on the best practices to preprocess data with respect to an specific statistical analysis in order to get quality analytical results.

To our aim, we have developed a comprehensive study to assess the impact of preprocessing procedures on accuracy of different statistical methods namely regression, classification, and clustering. We focused our study on procedures to select features, normalize data, and deal with missing values and outliers because these are the anomalies most frequently find in water quality data.

The main objectives of the approach are:

1. Assess the robustness of methods to process missing and outlying data;
2. Evaluate the effect of feature selection, normalization, missing data imputation, and outlying processing on the accuracy of subsequent classification, clustering, and regression analysis;
3. Identify the best data preprocessing procedures for a particular statistical analysis.

We have structured our study in three parts: (1) generation of synthetic datasets, (2) data preprocessing, and (3) statistical analysis (c.f. Figure 4.1).

In the first part, we construct synthetic data to assess preprocessing procedures then, in step two we generated preprocessed data by executing different preprocessing procedures on the synthetic data of step one, and we assessed the robustness of preprocessing procedures. In the third step we used the preprocessed data to assess the impact of preprocessing procedures on the results of statistical analysis. In the following sections we describe each part of our study.

The chapter is organized as follows: in Section 4.2 we briefly describe the preprocessing procedures used in our study, in Section 4.3 we describe our synthetic datasets. Then, in Section 4.5 we detail the methodological procedure followed to study the robustness of the preprocessing procedures and, in Section 4.6 we describe our analysis of the impact of data preprocessing procedures on the results of the main statistical methods. Finally, our results and concluding remarks are given in Section 4.7 respectively.

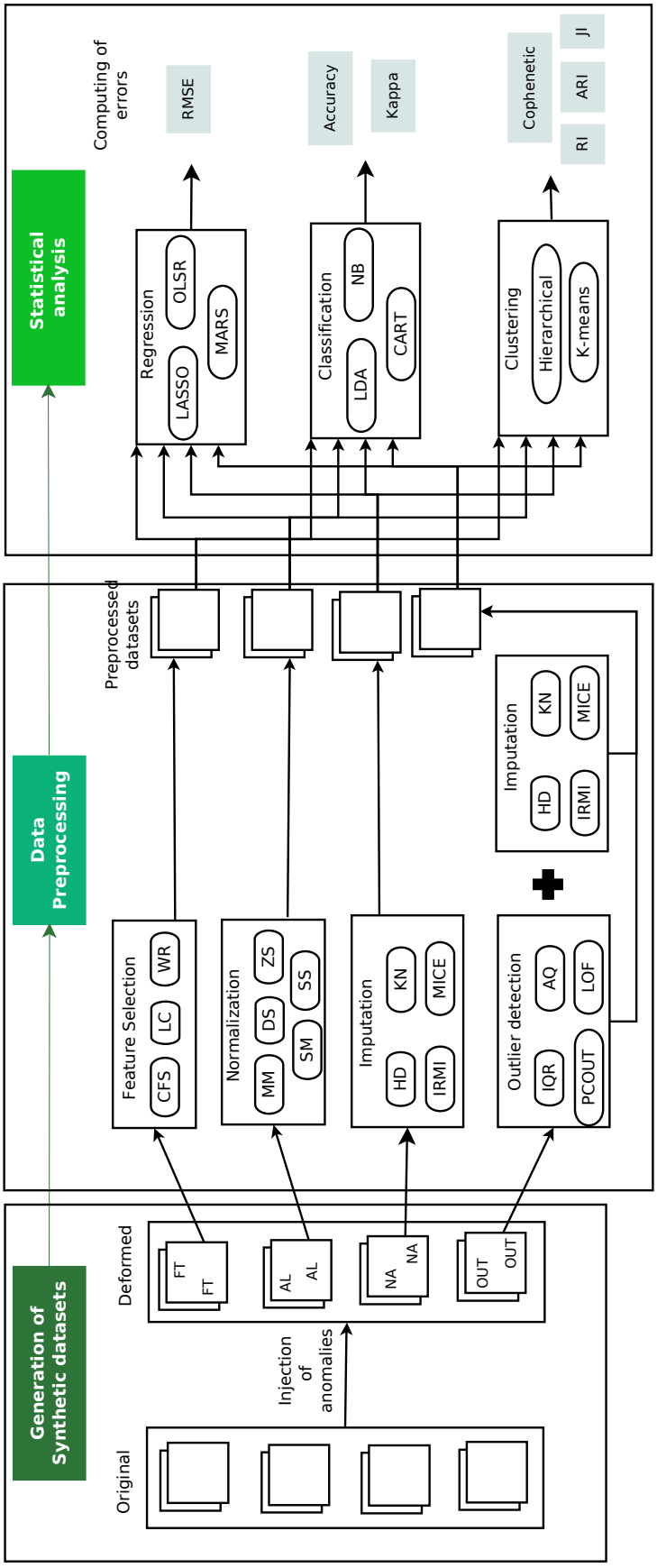


Figure 4.1 – General procedure for assessment of data preprocessing procedures on statistical results.

4.2 Data preprocessing procedures

4.2.1 Feature selection

Feature selection is performed to reduce the dataset by removing redundant or irrelevant features (or variables). It increases the speed of the learning methods and facilitates the understanding of patterns. We have studied three feature selection methods, two *filter* and one *wrapper* method. They were chosen because they belong to two different families of techniques that allow us to have an heterogeneous suite of methods to carry out our experiments. The methods are the following:

- *Correlation-based Feature Selection (CFS)* is a filter algorithm that ranks features according to a correlation-based evaluation (Hall, 1999; Liu et al., 2002). Relevant features are selected according to their correlation among classes and with each other. Features are considered relevant when they present a strong correlation with the class and they are not correlated with each other. Acceptance of a feature will depend on its efficiency to predict classes in the instance space that are not predicted by other features.
- *Linear Correlation* is an attribute evaluation method. This univariate filter method provides a ranking of all features for a certain threshold. Here, we set up the threshold to select features with strong correlation values (Eid et al., 2013).
- *Wrapper Subset Evaluator* evaluates usefulness of a feature set by using a learning algorithm. Cross-validation is used to estimate accuracy of the subsets (Kohavi and John, 1997). Finally, the feature subset which provides the best learning performance is chosen. In this work, we have used a simple regression model.

4.2.2 Normalization

Normalization is performed in order to give all variables equal weight. By normalizing data, all variables are expressed in the same measurement units, therefore measurement units can not affect data analysis. We have performed three normalization methods on numerical data: min-max, z-score and decimal-scale normalization. They were chosen due to their differences on the computation and their popularity.

- In *Min-max normalization*. Let us define min_A and max_A be the minimum and maximum values of an attribute A. In min-max normalization a value v_i of A is transformed to $v_i^{'}$ in the range $[new\_min_A, new\_max_A]$ by computing

$$v_i^{'} = \frac{v_i - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A \tag{4.1}$$

- *Z-score normalization* (or *zero-mean normalization*) is performed based on the mean and standard deviation of an attribute A. A value v_i of A is normalized to v_i^i as follows:

$$v_i^{'} = \frac{v_i - \bar{A}}{\sigma_A} \tag{4.2}$$

  where $\bar{A}$ is the mean and σ_A the standard deviation of A.
- In *decimal scale normalization*, the values of an attribute v_i is normalized to $v_i^{'}$ by moving the decimal point which depends on the maximum absolute value of $v_i^{'}$. Decimal scale normalization is computed using the following formula:

$$v_i^{'} = \frac{v_i}{10^j} \tag{4.3}$$

where j is the smallest integer such that $Max(|v_i^{'}|) < 1$.

4.2.3 Imputation methods

Imputation methods, are used to replace missing values on the dataset. This alternative can be used when an statistical method can not be performed under the presence of missing values. To our purpose, we have tested four imputation methods: two distance-based imputations: Hot-deck and K-NN, and two model-based imputation methods: Multiple Imputation by Chained Equations (MICE) and Iterative Robust Model-based Imputation (IRMI). They were chosen because they represent different techniques of imputation that allow us to have an heterogeneous comparison of methods.

- *Hot-deck* imputation method uses an actual value from a dissimilar case in the current dataset to replace the missing value. To determine the similar case, the user selects classification variables. The cases that agree with the case under consideration on these classification variables are placed into a pool from which one case is chosen randomly (Olinsky et al., 2003). The values used to impute in Hot-deck preserve the distributional characteristics of the data. This approach is effective essentially when the data are MAR. However, there is little empirical work which can determine the accuracy of this imputation method, it also can provide large bias on the estimates of the error variances.
- *K-nearest neighbour* imputation (K-NN) is based on the application of a distance measure in order to find the closest neighbours. Once the k-nearest neighbours have been found, a replacement value is estimated and used as a substitute for the missing value. The replacement value is calculated depending on the type of data, the mode is used for qualitative data and the mean for continuous data (Troyanskaya et al., 2001). One important issue when applying the K-NN approach is the selection of the appropriate distance metric. The distance between two observations was defined as

$$d_{i,j} = \frac{\sum_{k=1}^{p} w_k \delta_{i,j,k}}{\sum_{k=1}^{p} w_k} \tag{4.4}$$

where w_k is the weighted mean of the contributions of each variable and $\delta_{i,j,k}$ is the contribution of the kth variable. For continuous variables the absolute distance was computed as follows:

$$\delta_{i,j,k} = |x_{i,k} - x_{j,k}|/r_k \tag{4.5}$$

where $x_{i,k}$ is the value of the kth variable of the ith observation, and r_k the range of the kth variable (Templ et al., 2011a).
The advantage of K-NN is that it provides a robust procedure for missing data estimation, it can make prediction for both discrete and continuous attributes, it can treat multiple missing values easily because the creation of predictive models are not needed however, the K-NN algorithm is particularly problematic for large datasets because the algorithm search on the entire dataset to find the most similar instances (Batista et al., 2002).
- The *Multiple Imputation by Chained Equations (MICE)* is based on the assumption that missing data is MAR. The procedure consists of the computation of a set of regression models where each variable with missing values is modelled conditionally

upon the other variables in the data (Buuren and Groothuis-Oudshoorn, 2011). The algorithm can be summarized into the following steps:

Algorithm 4.1. Multiple Imputation by Chained Equations (MICE)

- Step 1: Initialization of a temporal imputation using a simple imputation method (e.g., mean);
- Step 2: For one of the variables y_i set back the temporal imputation into missing values;
- Step 3: Regression based on the observed values from the variable y_i in Step 2 on the other variables in the imputation model. This means that in a regression model y_i will be considered as the dependent variable and the other variables as independent;
- Step 4: Replace missing values of y_i by predictions drawn from the regression model. When y_i is subsequently used as independent variable in the regression models for other variables, both these imputed values and the observed will be used;
- Step 5: Steps 2-4 are repeated for each variable with missing values.
- Step 6: The loop through each variables where steps 2-4 are repeated constitutes one iteration. At the end of each iteration, all missing values have been replaced. Steps 2-4 are repeated for a number of iterations defined by user. The final imputations at the end of all iterations are kept as the final result for one version of imputed dataset.

- The *Iterative Robust Model-based Imputation* (IRMI) (Templ et al., 2011b) is based on the EM algorithm, where, in the "Expectation" step, the regression method is applied iteratively. For each iterative step, one variable is used as the response variable and the other variables are used as regressors. In this way, "all" the information of the variables is used for the imputation of the responding variables. The algorithm is summarized into seven steps as follows:

The main differences between MICE and IRMI is that IRMI includes a robust regression which reduces the influence of outlying observations and protects against poorly initialized missing values. Another difference is that in each step of MICE the predictive values are used to update former missing values while IRMI uses predictive values to update expected values, this allow to keep track of convergence of sequential methods. IRMI also provides (co-)variances that are included in the final iteration.

The four selected imputation methods are available in R. Scripts involving the studied methods and their respective packages, namely: `VIM` for hot-deck, K-NN and IRMI, and `mice` for MICE have been implemented for testing and assessing the methods.

Algorithm 4.2. Iterative Robust Model-based Imputation (IRMI)

- Step 1: Initialization of a temporal imputation using a simple imputation method (e.g., mean imputation);
- Step 2: Sort each variable x according to the amount of missing values. Set I = {1,...,p};
- Step 3: Set l = 1;
- Step 4: Define $X^{o_l}_{I/\{l\}}$ and $X^{m_l}_{I/\{l\}}$ the matrices with the variables of the observed and missing cells of x_l, respectively, where $m_l \subset \{1,...,n\}$ denotes the indices of missing observations in variable x_l and $o_l = \{1,...,n\}/m_l$ the indices of the observed cells of x_l. The first column of $X^{o_l}_{I/\{l\}}$ and $X^{m_l}_{I/\{l\}}$ will consist of ones, and an intercept term $x^{o_l}_l = X^{o_l}_{I/l}\beta + \epsilon$ in the regression problem should be considered. Where β is the regression coefficient and ϵ is the error term. In each regression the distribution of the response $x^{o_l}_l$ is considered to fit if the response is:
  - *continuous*, the link is μ and a robust regression method is applied;
  - *categorical*, a generalized linear regression is applied;
  - *binary*, a logistic linear regression is applied when the link is $log(\frac{\mu_i}{1-\mu_i})$ for i = 1,...,n;
  - *semi-continuous*, a two-stage approach is applied, in the first stage a logistic regression is applied and in the second stage a robust regression based on the continuous (non-constant) part of the response is used to impute;
  - *count*, a robust generalized linear regression of family Poisson is applied and the link is $log(\mu_i)$, for i = 1,...,n.
- Step 5: Evaluate β with the model in Step 4. Replace the missing parts $x^{m_l}_l$ by $\widehat{x}^{m_l}_l = X^{m_l}_k \widehat{\beta}$;
- Step 6: Perform Step 4-5 for each l=2,...,p;
- Step 7: Repeat Steps 3-6 until imputed values stabilize, i.e. until

$$\sum_i (x^{\hat{m}_l}_{l,i} - x^{\bar{m}_l}_{l,i})^2 < \delta, for\ all\ i \epsilon m_l\ and\ l \epsilon I \tag{4.6}$$

where $x^{\hat{m}_l}_{l,i}$ is the i-th imputed value of the current iteration, and $x^{\bar{m}_l}_{l,i}$ is the i-th imputed value from the previous iteration.

4.2.4 Outlier detection methods

Four detection methods based on different approaches were implemented including; an statistic-based approach (Inter quartile Range, IQR), two multivariate outlier detection approaches (Adjusted-Quantile), and an algorithm that uses Principal Components decomposition (PCOUT) (Filzmoser et al., 2008), and a density-based approach (Local Outlier Factor, LOF; (Breunig et al., 2000)). They were chosen because they belong to different families of methods that allow us to have an heterogeneous comparison of methods.

- Inter Quartile Range (IQR) covers the central 50% of data, it is determined as the range between the 25 and 75% of the quantiles Q_1 (lower quantile value) and Q_3 (upper quartile value), $IQR = Q_3 - Q_1$. When an observation falls below $Q_1 - 1.5(IQR)$ or above $Q_3 + 1.5(IQR)$ it is then considered as an outlier.
- Adjusted-Quantile (AQ). Outliers are detected through the adjusted quantile as follows (Filzmoser et al., 2005): First, a chi-squared plot is used to visualize the deviation of the data distribution from multivariate normality in the tails. Then the tails of the empirical distribution function $G_n(u)$ of the squared robust distances RD_i^2 and the distribution function $G(u)$ of χ^2_ρ are compared to detect outliers.
  Let us denote the empirical distribution function of the squared robust distances RD_i^2 as $G_n(u)$ and $G(u)$ as the distribution function of χ^2_ρ. G_n will converge to G for multivariate normally distributed samples. The tails of G_n and G are compared to detect outliers.
  The tails are defined by $\delta = \chi^2_\rho; 1-\alpha$ for small α (e.g., $\alpha = 0.02$), and

$$\rho_n(\delta) = \underset{u \geq \delta}{sup}(G(u) - G_n(u))^+ \tag{4.7}$$

  where + indicates the positive differences. $p_n(\delta)$ is not used directly as a measure of outliers. Instead a critical value P_{crit} is introduced to distinguish between outliers and extremes where extremes of the distribution are considered as observations with a large RD. The measure of outliers in the sample is defined as

$$\alpha_n(\delta) = \left\{ \begin{array}{cc} 0 \quad \text{if} \quad p_n(\delta) \leq p_{crit}(\delta, n, p), \\ p_n(\delta) \quad \text{if} \quad p_n(\delta) > p_{crit}(\delta, n, p) \end{array} \right\} \tag{4.8}$$

  $c_n(\delta) = G_n^{-1}(1 - \alpha_n(\delta))$ determines the threshold value.
- Principal Component decomposition for outlier detection (PCOUT) consist of two parts: a step to detect location outliers and a step to detect scatter outliers. Location outliers are described by different location parameter while scatter outliers are differentiated due to their scatter matrix which is different from the rest of the data (Filzmoser et al., 2008). The algorithm of PCOUT that includes the two parts is summarized bellow in Algorithm 4.3.
- Local Outlier Factor (LOF) assigns an outlier factor to each observation, which is the degree the observation is being outlier. This degree is measured with respect to the density of the local neighbourhood (Breunig et al., 2000; Kriegel et al., 2010). LOF is computed as follows:
  For any positive integer k, k-distance(p) is defined as the distance $d(p,o)$ between object p and $o\ \epsilon D$ such that:
  1. $o^{'} \epsilon D \setminus \{p\}$ holds that $d(p, o^{'}) \leq d(p,o)$, for at least k objects, and
  2. $o^{'} \epsilon D \setminus \{p\}$ holds that $d(p, o^{'} < d(p,o))$ for at most $k-1$ objects

  The reachability distance between object p and object o is defined as:

$$reach - dist_k(p,o) = max\{k - distance(o), dist(p,o)\} \tag{4.14}$$

Algorithm 4.3. Principal component decomposition for outlier detection (PCOUT; (Filzmoser et al., 2008))

- **Step 1** Transform the data by subtracting the median and dividing by the median absolute deviation (MAD), in each dimension. Compute the covariance matrix of the transformed data. The MAD is defined for a sample $\{x_1,...,x_n\} \subset \mathbb{R}$ as

$$MAD(x_1,...,x_n) = 1.4826 \cdot \underset{j}{med}|x_j - \underset{j}{med}x_i| \tag{4.9}$$

- **Step 2** Calculate the principal component decomposition of the semi-robust covariance matrix from Step 1. Retain the p^* values/eigenvectors that contribute to at least 99% of the total variance. Sphered the transformed data through the median and MAD.
- **Step 3** Compute the robust kurtosis weights for each components and weighted norms for the sphered data as follows

$$w_j = \left| \frac{1}{n} \sum_{i=1}^{n} \frac{(z_{ij} - med_i z_{ij})^4}{MAD(z_{1j,...,z_{nj}})^4} - 3 \right| \tag{4.10}$$

  By scaling data through MAD, the Euclidean norms in principal component space are equivalent to Mahalanobis distances. Transform these distances using the following equation

$$d_i = RD_i \cdot \frac{\sqrt{\chi^2_{p*,0.5}}}{median\{RD_i\}} \tag{4.11}$$

  where $\chi^2_{p*,0.5}$ is the χ^2_{p*} 0.5 quantile.
- **Step 4** Determine weights w_{1i} for each robust distances as follows:

$$w_{1i} = \begin{cases} 0, & d_i \geq c \\ \left(1 - (\frac{d_i - M}{c - M})^2\right)^2, & M < d_i < c \\ 1, & d_i \leq M \end{cases} \tag{4.12}$$

  where $i = 1,...,n$ and M is the 0.33 quantile of the distances $\{d_1,...,d_n\}$ and $c = median\{d_1,...,d_n\} + 2.5 \cdot MAD\{d_1,...,d_n\}$.
- **Step 5**Compute the (unweighted) Euclidean norms of the data using the semi-robust principal component decomposition used in Step 2. Transform using Eq. (4.11) to get a set of distances.
- **Step 6** For each robust distance, determine weights w_{2i} according to Eq. (4.12) with $c^2 = \chi^2_{p*}0.99$ quantile and $M^2 = \chi^2_{p*}0.25$ quantile.
- **Step 7** Determine final weights for all observations according to Eq. (4.13) using weights of Steps 4 and 6

$$w_i = \frac{(w_{1i} + s)(w_{2i} + s)}{(1 + s)^2} \tag{4.13}$$

The local reachability density (lrd) of object p is the inverse of the average reach-dist of the kNNs of p which is defined as:

$$lrd_k(p) = l/\left(\frac{\sum_{o \epsilon kNN(p)} reach - dist_k(p,o)}{Card(kNN(p))}\right) \tag{4.15}$$

The local outlier factor (LOF) of object p is the average ratio of local reachability distance of neighbours of p and lrd of o

$$LOF_k(p) = \frac{\sum_{o \epsilon kNN(p)} \frac{lrd_k(o)}{lrd_k(p)}}{Card(kNN(p))} \tag{4.16}$$

4.3 Synthetic data description

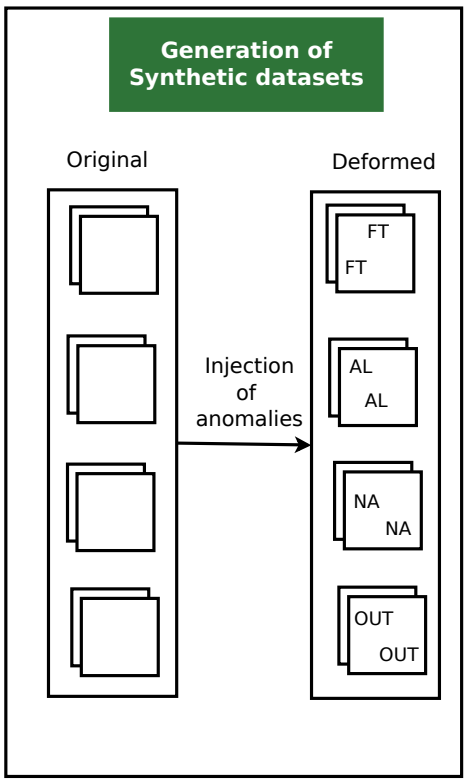


Figure 4.2 – Flow diagram of the generation of synthetic datasets for the assessment of preprocessing procedures on statistical results. FT denotes highly correlated data, AL denotes non-normalized data, NA stands for not available/missing values and OUT denotes outlying data.

The first step of our comprehensive study for the assessment of preprocessing procedures is the construction of synthetic datasets. Our experiments can not rely only on our dataset produced using the methods described in the previous chapter (c.f. Section 3) because the results would not be meaningful. Thus, we needed to construct synthetic datasets. Our synthetic datasets were constructed in order to be the most similar to environmental data in terms of distribution, correlation, etc., specifically, water quality data. As illustrated in Figure 4.2, we created non-deformed datasets that we named "original" then, we deformed the original datasets by injecting anomalies (i.e., missing values and outlying data) or by deforming the original characteristics of the datasets. For instance, creating a new dataset with a non-normal distribution or by generating high correlated variables. We created different datasets for a better control on our experimentations. The resulting deformed datasets were subsequently used on the second part of our study (c.f. Section 4.5). Below we detail the characteristics of our datasets.

Synthetic data for Feature Selection

Four synthetic datasets were used to assess the impact of feature selection procedures on subsequent regression, classification, and clustering analysis. Each dataset follows a normal distribution and is composed of different number of observations n, numerical variables p, and one categorical variable of five classes with an uniform distribution, variables include irrelevant features (c.f. Table 4.1). These irrelevant features are highly correlated variables.

Synthetic datasets for Normalization

Experiments to assess the impact of normalization were performed using four synthetic datasets. Each dataset is composed of different number of observations n, numerical variables p from a Weibull distribution and one categorical variable of five classes from an uniform distribution (c.f. Table 4.2).

Synthetic datasets for missing data

Assessment of imputation methods was carried out using four synthetic datasets. Each dataset is composed of different number of observations n, numerical variables p from a normal distribution and one categorical variable of five classes from an uniform distribution. Missing data were introduced randomly in numerical variables, using the MCAR mechanism, into each of the datasets. The missing values were introduced into all variables in all datasets in the following six amounts: 5%, 10%, 15%, 20%, 25% and 30%. For each missing ratio, we generated ten different missing datasets from the original complete dataset to ensure that experimental results were statistically acceptable. Table 4.3 summarizes the characteristics of our experimental datasets.

Synthetic datasets for outlying data

The experiments were performed using three synthetic datasets. Each dataset is composed of different number of observations n, numerical variables p from a normal distribution and one categorical variable of five classes from an uniform distribution. Outlying data were introduced randomly into each of the datasets in the following five amounts: 1.5%, 2.5%, 5%, 10% and 15%. A detailed description of the synthetic datasets used for assessment of outlier detection methods is given in Table 4.4.

Hereafter we provide the characteristics of the first ten variables of the datasets used for our experimentations.

Table 4.1 – Synthetic datasets used to assess feature selection methods. Only the μ and σ^2 values of the first ten variables are shown. Highly correlated variables are denoted as Y_k.

Dataset			Variables									
FS21	$n = 21$	$p = 8$	X1	X2	X3	X4	X5	Y1	Y2	Y3		
μ			321.72	181.80	518.91	734.27	714.80	775.90	619.09	753.99		
σ^2			1.11	1.12	0.87	1.43	1.20	0.84	0.86	0.74		
FS600	$n = 600$	$p = 30$	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
μ			511.01	32.00	69.99	238.07	159.08	125.98	350.98	578.01	281.04	445.00
σ^2			0.95	1.06	1.06	1.02	1.01	1.03	0.91	1.05	0.96	0.97
FS4000	$n = 4000$	$p = 53$	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
μ			769.98	155.98	663.00	19.01	720.99	684.00	467.02	599.98	477.00	326.01
σ^2			0.98	0.99	1.01	0.97	0.99	1.02	1.00	0.99	0.99	1.00
FS20000	$n = 20000$	$p = 98$	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
μ			23.00	142.00	746.01	351.00	459.99	270.00	55.99	132.00	804.99	783.01
σ^2			0.99	0.99	0.99	1.00	1.00	1.00	1.00	0.99	1.01	1.00

Table 4.2 – Synthetic datasets used to assess normalization methods. Only the μ and σ^2 values of the first ten variables are shown.

Dataset			Variables									
AL21	$n = 21$	$p = 8$	X1	X2	X3	X4	X5	X6	X7	X8		
μ			213.98	255.75	761.98	562.19	35.16	411.48	348.67	393.36		
σ^2			41343.79	62385.58	679029.21	129683.04	1010.23	221852.32	175850.52	122081.17		
AL600	$n = 600$	$p = 30$	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
μ			171.95	193.49	72.89	107.32	633.29	310.19	244.98	558.62	443.10	156.39
σ^2			29166.08	40522.14	5041.46	12452.19	443577.50	82307.87	53645.94	333400.05	195958.17	20831.51
AL4000	$n = 4000$	$p = 53$	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
μ			259.17	227.73	838.73	67.93	165.33	323.62	741.48	531.30	620.14	789.61
σ^2			66203.82	51371.52	700756.56	4462.29	28778.14	99923.59	529437.41	271397.61	392837.24	605888.40
AL20000	$n = 20000$	$p = 98$	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
μ			256.25	225.65	839.71	67.18	168.80	319.50	710.35	532.69	632.25	779.89
σ^2			66029.69	50712.29	683255.96	4518.53	27972.67	102515.49	500528.68	283402.55	407322.39	602853.08

Table 4.3 – Synthetic datasets used to assess imputation methods. Only the μ and σ^2 values of the first ten variables are shown.

Dataset			Variables									
N21	$n = 21$	$p = 8$	X1	X2	X3	X4	X5	Y1	Y2	Y3		
μ			191.25	156.16	536.92	544.79	693.06	766.90	744.81	269.08		
σ^2			1.18	1.29	1.19	0.83	0.80	0.55	0.79	0.78		
N600	$n = 600$	$p = 30$	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
μ			216.99	512.02	244.01	333.99	172.01	241.02	211.04	310.99	158.94	402.06
σ^2			0.92	1.00	0.99	1.03	1.01	1.01	0.96	1.05	0.96	0.91
N4000	$n = 4000$	$p = 53$	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
μ			97.02	513.00	517.00	487.99	362.98	893.01	355.99	688.98	661.01	817.00
σ^2			1.01	1.02	1.00	0.97	1.01	1.01	1.01	1.02	0.98	0.98
N20000	$n = 20000$	$p = 98$	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
μ			311.99	338.00	154.00	272.99	513.98	391.99	417.00	578.00	386.00	560.01
σ^2			1.00	1.00	1.01	0.99	1.00	0.99	1.00	0.99	0.98	0.99

Table 4.4 – Synthetic datasets used to assess outlying preprocessing. Only the μ and σ^2 values of the first ten variables are shown.

Dataset			Variables									
N21	$n = 21$	$p = 8$	X1	X2	X3	X4	X5	X6	X7	X8		
μ			190.88	155.89	536.80	544.77	693.12	766.84	744.70	268.90		
σ^2			4.43	5.43	3.05	0.76	1.10	1.43	1.52	3.31		
N600	$n = 600$	$p = 30$	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
μ			216.79	511.82	243.71	333.76	171.73	240.75	210.81	310.80	158.73	401.81
σ^2			2.74	3.20	3.75	3.35	3.95	3.36	3.30	2.96	2.86	3.45
N4000	$n = 4000$	$p = 53$	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
μ			96.81	512.78	516.77	487.76	362.76	892.80	355.78	688.74	660.76	816.75
σ^2			3.04	3.21	3.31	3.21	3.24	3.07	3.09	3.40	3.29	3.48

4.4 Semi-synthetic data description

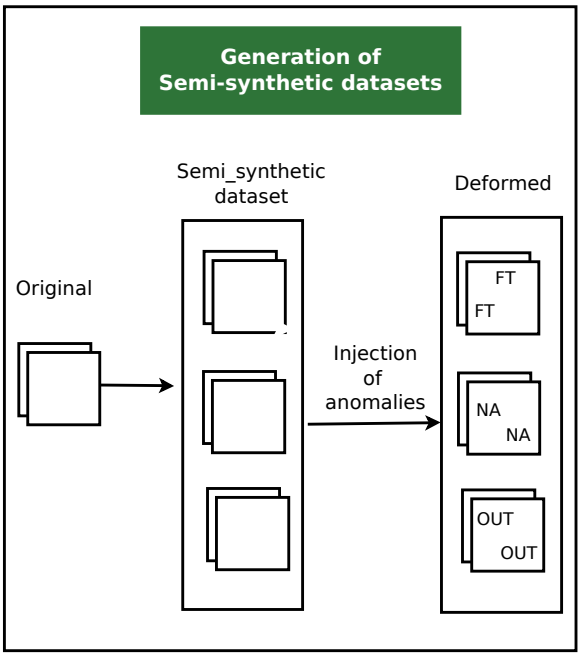


Figure 4.3 – Flow diagram of the generation of semi-synthetic datasets for the assessment of preprocessing procedures on statistical results.

In the need to implement useful tools to evaluate the effectiveness of restoration and/or protection actions of French rivers, the FRESQUEAU[1] was conceived. The project brought together experts on computational intelligence and data mining methods (from the four laboratory involved, LHYGES[2], TETIS[3], LSIIT[4] and LIRMM[5].) and experts on hydro-ecology: LHYGES, TETIS as well as two engineering consultants, AQUASCOP and AQUABIO. The principal goals of this project were: to develop a tool that will allow the assessment of rivers global function by considering the different parts of the ecosystem and to broaden the knowledge of data mining from complex, heterogeneous and large databases with temporal and spatial variations. Within the objectives of the project, physical, chemical and biological data produced by the French Water Agencies DREAL[6] and ONEMA[7] from the Rhin-Meuse (RM) and Rhône-Méditerranée-Corse (RMC) basins were collected.

[1]The FRESQUEAU project (« Fouille de données pour l'évaluation et le suivi de la qualité hydro-biologique des cours d'eau ») which was a French multidisciplinary project financially supported by the Research National Agency (ANR) 2011-2013

[2]Laboratory of Hydrology and Geochemistry of Strasbourg

[3]A joint research centre of Territories, Environment Remote Sensing satellite and Spatial Information (TETIS from its French acronym).

[4]LSIIT from its French abbreviation is the Laboratory of Science of Images, Informatics and Remote Sensing Satellite of Strasbourg.

[5]LIRMM from its French abbreviation is the Laboratory of Informatics, Robotics and Microelectronics of Montpellier

[6]The Regional Direction for the Environment, Development and Housing (DREAL from its French abbreviation).

[7]The National Office of Water and Aquatic Environment (ONEMA from its French abbreviation).

The FRESQUEAU dataset was composed of 234 monitoring stations of the RM area dating from 2002 to 2010. It concerned physical and physico-chemical information. The data contained the annual statistical summary values of each parameter (arithmetic mean, median, percentile 10, percentile 90) obtained from streams water quality monitoring done every month or every two months.

The physico-chemical data were formed by the monitoring of 644 parameters analysed for each station. Out of these 644 parameters, 33 were considered as classic physico-chemical parameters called now with the Water Framework Directive (DCE) "physico-chemical parameters supporting biology" (in the text those parameters will be mentioned as macro pollutants) and 611 as micro pollutants. The list of macro pollutants is made up of the parameters describing the content of oxygen (e.g. DBO_5, Dissolved Oxygen O_2, saturated oxygen O%), organic matter (e.g., Organic Carbon, ${CO_3}^{2-}$) phosphorus (total phosphorus, phosphates ${PO_4}^{3-}$), nitrogen (e.g., Kjendhal Nitrogen, ${NH_4}^{-}$, ${NO_2}^{-}$, ${NO_3}^{-}$), minerals (e.g., Na, K, Ca, Cl) or eutrophication (e.g. chlorophyl a, pheopigments).

The FRESQUEAU dataset contained the couple station-year (which was considered as an independent station) and its corresponding summary statistic value for each parameter.

We have used the data which is composed of 1565 stations-year with 33 variables for macro pollutants related to the mean summary statistic value. This dataset was treated in order to remove noisy data. Therefore outliers, duplicate data and missing values were removed. The cleaned dataset was composed of 1504 observations (stations-year) and 26 variables that we named FQ1000. From this dataset, two semi-synthetic datasets were generated. The first semi-synthetic dataset, named FQ16 is a reduced presentation of FQ1000, it is composed of 16 observations and 13 variables. Observations and variables were selected randomly from the FQ1000 dataset. The second semi-synthetic dataset, FQ7000 was constructed by introducing new observations following the distribution of the FQ1000 dataset. FQ7000 dataset is composed of 7520 observations and 26 variables.

The cleaned semi-synthetic dataset (FQ16, FQ1000 and FQ7000) were subsequently deformed by injecting anomalies (i.e., missing values, outlying data) in the same fashion as described in Section 4.3 for the synthetic datasets. Details of the μ and σ^2 of the semi-synthetic datasets are given in Appendix B.

4.5 Robustness study of data preprocessing procedures

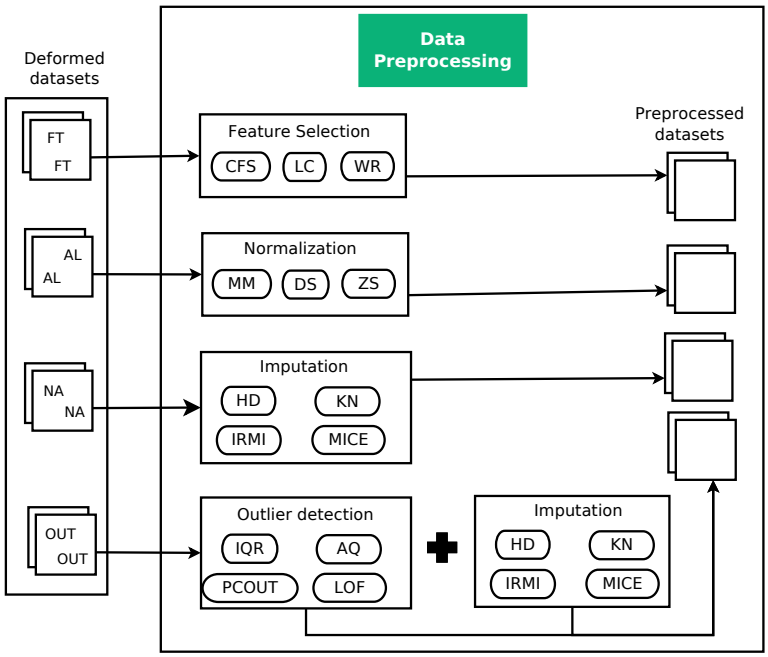


Figure 4.4 – Extract of the general workflow for assessment of data preprocessing procedures. Figure shows the workflow diagram of data preprocessing assessment.

In the second part of our comprehensive study, we aimed at assessing the robustness of methods to process missing and outlying data and preprocess data by feature selection and normalization. To our purpose, we have used our deformed data constructed in the previous step (c.f. Section 4.3), next we have executed four data preprocessing procedures including: feature selection, normalization, imputation of missing values and outlying processing (c.f. Figure 4.4). We were interested in these procedures because they treat the data anomalies most frequently found in environmental data. The resulting preprocessed datasets were subsequently used on the third part of our study (c.f. Section 4.6).

Additional to the construction of preprocessed datasets, we have compared the performance of the selected imputation and outlier detection methods. We did a comparative study for the selected methods because, we did not find in the literature a study that compares them. Concerning feature selection and normalization, previous studies have compared the performance of different feature selection (Bolón-Canedo et al., 2013) and normalization methods (Mustaffa and Yusof, 2011; Grimvall et al., 2001) and thus we did not studied them in here. Below we detail our experimental procedure.

4.5.1 Selection of features

Feature selection is a process to select an optimal subset of features. An optimal subset could be a subset that gives the best estimate on a predictive model. It is implemented for different purposes including: (1) to improve performance of a model in terms of simplicity of the model, speed, etc.; (2) to better visualize data; (3) to remove noise and reduce dimensionality.

We have performed three feature selection methods a Correlation-based feature selection (CFS), a linear-based correlation (LC) and a Wrapper subset evaluator (WR) to select an optimal data subset. Selection of features on each deformed dataset was performed as follows: First, we have aleatory chosen one variable that we used as independent variable then, by using the selected feature selection methods we have identified the best data subset with respect to our independent variable. The resulting data subsets were subsequently used to assess the impact of feature selection procedures on statistical results (c.f. Section 4.5).

4.5.2 Normalization of data

Water quality data is represented by a set of variables with different dimensions and units. To give all variables equal weight and obtain accurate predictive model we need to normalize data. We have performed three normalization methods: min-max (MM), Z-score (ZS) and decimal scale normalization (DS). Normalization was applied to numerical data. The resulting normalized datasets were subsequently used for our study described in Section 4.5.5.

4.5.3 Handling missing data

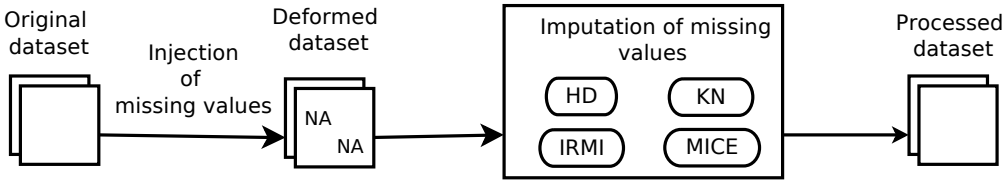


Figure 4.5 – Experimental procedure for assessment of imputation methods.

Experimental procedure to assess robustness of missing data imputation is presented in Figure 4.5. First, the 'original' synthetic datasets were deformed by injecting six different amounts of missing values. Next, the missing values in the 'deformed' datasets were imputed using five imputation methods. The imputation accuracy was evaluated by computing the normalized root mean squared error (NRMSE) (Oba et al., 2003) as follows:

$$NRMSE = \sqrt{\frac{mean[(y_{imputed} - y_{complete})^2]}{variance[y_{complete}]}} \tag{4.17}$$

where the mean and variance are calculated over missing entries in the whole matrix. $y_{complete}$ is known because the missing entries are artificial. The larger the NRMSE is, the less is the prediction accuracy.

The experiments, which include 4 datasets, 10 times 6 amounts of missing values, 4 imputation methods gives us a total of $4 \times 10 \times 6 \times 4 = 960$ preprocessed datasets. All the experiments were performed using the R environment for statistical computing (version 3.2.2; R Core Team (2016)).

4.5.4 Handling outlying data

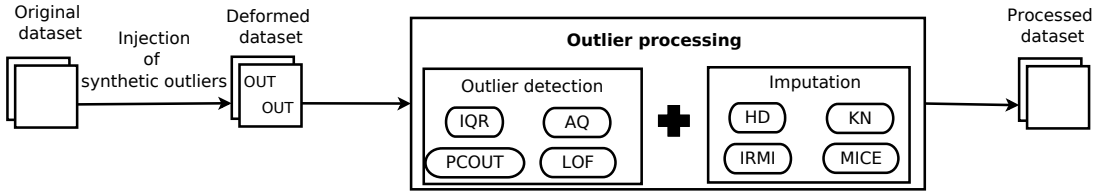


Figure 4.6 – Experimental procedure for assessment of outlier detection and processing.

Experimental procedure to assess robustness of outlying processing is presented in Figure 4.6. First, 'original' synthetic datasets were constructed, then outlying datasets were generated by introducing the five different amounts of outlying data. Next, outliers were detected using four methods to detect outliers. Finally, outliers were replaced using the imputation methods *Hot-deck*, *k-NN*, *mice* and *IRMI*. The resulting imputed datasets were subsequently used to study the impact of outlier preprocessing on statistical results (c.f. Subsection 4.6). Outlier detection methods were assessed by computing detection rate and precision of detection as follows (Chen et al., 2010),

$$detection\ rate = \frac{the\ number\ of\ outliers\ correctly\ detected}{total\ number\ of\ true\ outliers} \times 100 \tag{4.18}$$

$$precision = \frac{t_p}{t_p + f_p} \times 100 \tag{4.19}$$

where t_p is the number of points correctly labelled as outlier and f_p is the number of points wrongly labelled as outlier.

The experiments, which include three datasets, five amounts of outlying data, four outlier detection methods and four methods of imputation gives us a total of $3 \times 5 \times 4 \times 4 = 240$ preprocessed datasets. All the experiments were performed using different functions available in the R environment for statistical computing (version 3.2.2; R Core Team (2016)).

4.5.5 Results and discussion on robustness of data preprocessing procedures

4.5.5.1 Imputation methods robustness

Results of our assessment on imputation methods are graphically illustrated in Figure 4.7. We observed that, from the two distance-based imputation methods, the K-NN is more accurate than the Hot-deck method. While for the two model-based imputation methods, IRMI showed up to be more accurate than MICE. To our surprise, the K-NN method has the lowest NRMSE values. We consider that this results is due to the characteristics of the datasets. More experiments that include heterogeneous set of datasets (e.g., non normal distribution, correlated variables) will be necessary to have a more general view of the behaviour of the tested methods. Except for K-NN, our results are congruent to previous studies (Cottrell et al., 2009; Templ et al., 2011b; Hron et al., 2010) in that multivariate imputation methods (here: MICE and IRMI) are more robust when

comparing with more simple imputation methods (here: Hot-deck). It must be noticed that the results that we obtained depend on the characteristics of the data, in our case data with multivariate normal distribution. It must be necessary to perform other experiments with a more heterogeneous set of data to get a more complete vision of the behaviour of the tested imputation methods. For our subsequent comparative study (c.f.4.6), we decided to continue using the imputation methods described in this section.

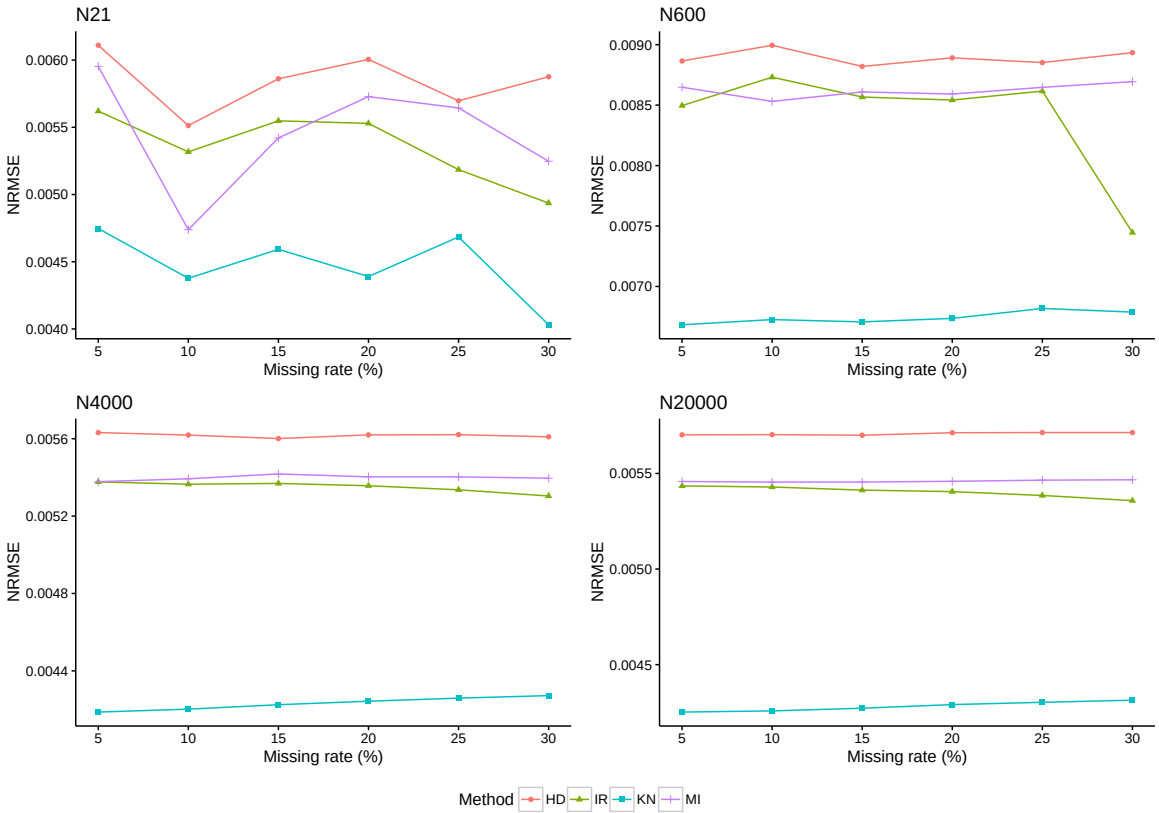


Figure 4.7 – NRMSE results of the imputation of missing values using the Hot-deck (hd), IRMI (ir), k-NN (kn) and Mice (mi) imputation methods on datasets N21 (A), N600 (B), N4000 (C) and N20000 (D).

4.5.5.2 Outlier detection assessment

Results of our outlier detection assessment are as follows: As shown in Figure 4.8, we observed that precision of outlier detection methods was in the following ordering: PCOUT > Adjusted quantile > LOF > IQR for $N4000$ and $N600$ datasets. For $N21$ dataset the LOF detection method show the best results. We observed that the IQR method has negative values, this indicated that IQR is precise but the rate of outliers detected is low. This assumption was confirmed with the results of detection rate where the IQR method show, in general, the lowest values.

Concerning outlier detection rates, detection performance is observed in the following order: LOF > PCOUT > Adjusted quantile > IQR. In general, LOF method provides the best results for datasets N600 and N4000 for the five outlying data rates. For dataset N21, we observed that at low outlying data rates (1.5% and 2.5%) methods Adjusted quantile and PCOUT perform well but at higher outlying data rates (5% and 15%) method LOF has the best detection results. For our synthetic datasets that have multivariate

normal distribution and individual outliers LOF shows to be the best method. Additional experiments that may include heterogeneous set of datasets will be necessary. We want to emphasize that the results that we observed were characteristic for our datasets (multivariate normal distribution and individual outliers injected at random). And further analysis may be necessary to observe the behaviour of the tested methods on datasets with different distribution (e.g., Weibul, Cauchy) and with other type of outliers (e.g., cluttered, patterns).

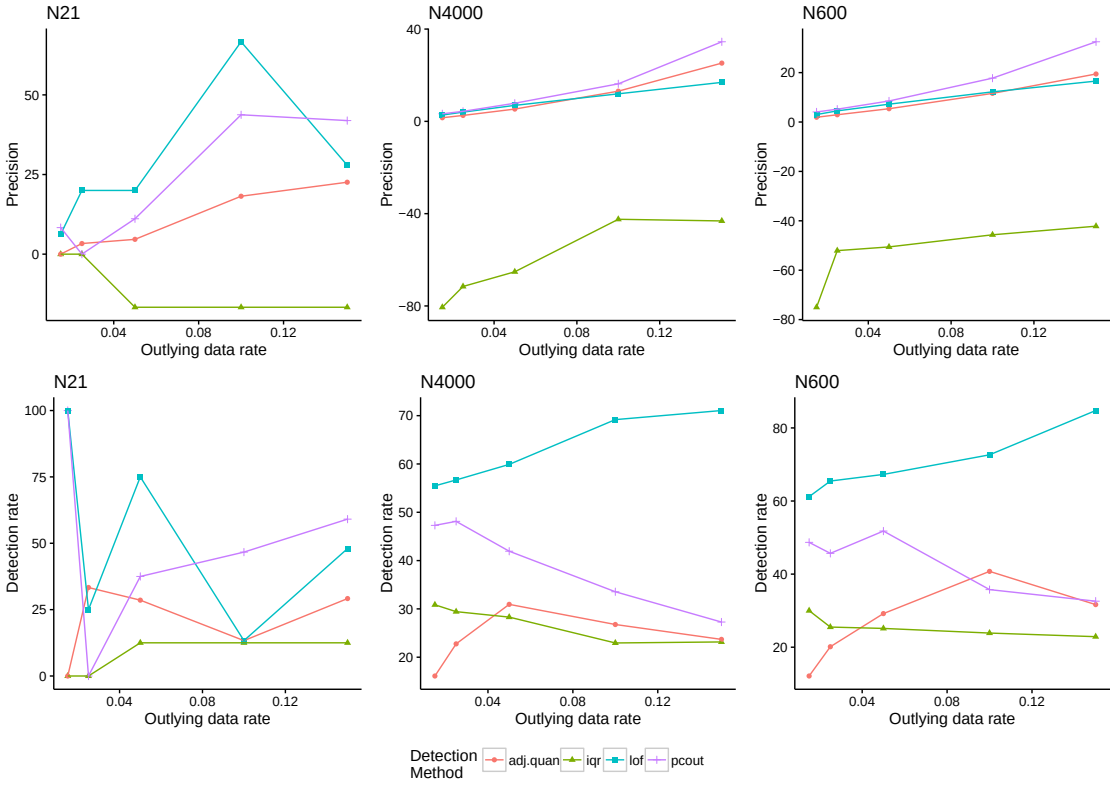


Figure 4.8 – Precision and detection rate results of outliers detection methods. Adjusted quantile (adj.quan), Inter Quartile Range (iqr), Local Outlier Factor (lof) and Principal Components decomposition approach (pcout).

4.6 Study of the impact of data preprocessing procedures on statistical results

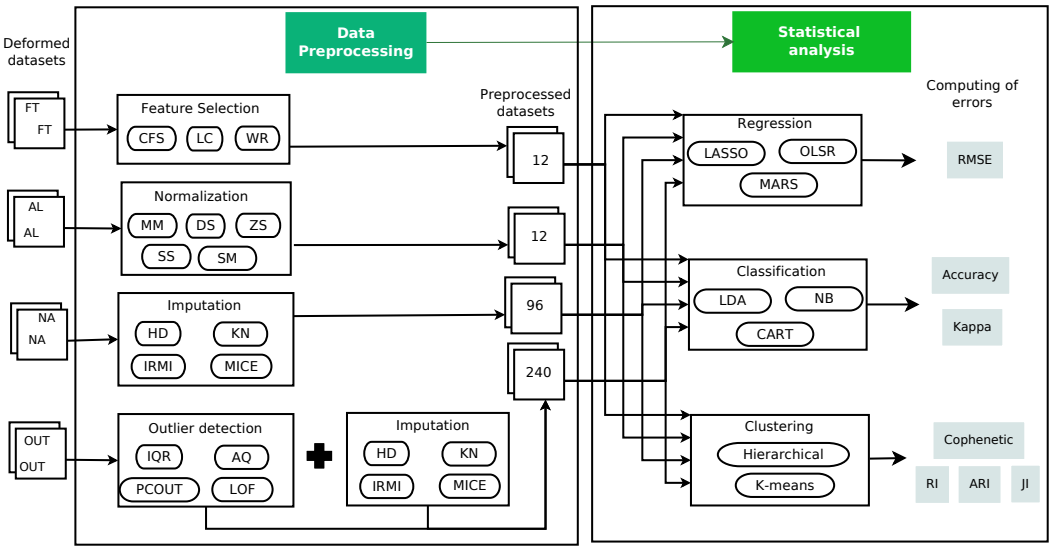


Figure 4.9 – Experimental procedure for assessment of the impact of data preprocessing procedures on statistical analysis results.

Our experimental procedure to assess the impact of data preprocessing procedures on statistical analysis results is shown in Figure 4.9. First, we have used synthetic deformed datasets (c.f. Section 4.3) that were preprocessed using: feature selection, normalization, imputation and outlier processing (c.f. Section 4.5). Then, the resulting preprocessed datasets were used to execute regression, classification and clustering methods. The selected statistical methods were chosen because they are frequently used on environmental studies to predict phenomena, find patterns and discover relationships among variables and observations (Wiseman, 2006). Finally, we have compared the statistical results of each preprocessed data by computing different statistical errors. We have performed our experiments using a combination that, seems the most simple to us. This is, only one type of data anomaly was studied at a time the resulting preprocessed dataset was then used to implement the selected statistical methods. For example, we have four deformed datasets for feature selection, they were preprocessed using three feature selection methods. Finally the resulting twelve preprocessed datasets were used to implement three regression methods, three classifiers and two clustering methods.

In the following sections we detail the procedures followed to implement the statistical methods and the computation of the errors.

4.6.1 Impact on regression results

We have performed three regression methods on our preprocessed data that resulted from feature selection (c.f. Subsection 4.5.1), normalization (c.f. Subsection 4.5.2), imputation of missing values (c.f. Subsection 4.5.3), and outliers processing (c.f. Subsection 4.5.4). Regression methods were chosen for their diversity of representation and learning style. They include:

- Penalized regression: Least Absolute Shrinkage and Selection Operator (LASSO);
- Linear regression: Ordinary Least Squares Regression (OLSR);
- Non-Linear regression: Multivariate Adaptive Regression Splines (MARS).

Regression analysis for each dataset was implemented as follows:

1. We begin by randomly splitting the observations into a training and test set where 66% of data was used for training and 34% for testing,
2. Regression model was fit on the training set, and the fitted model was used to predict the responses for the observations in the testing set,
3. Resulting validation was assessed using the RMSE as in Equation 4.20.

$$RMSE = \sqrt{mean[(y - \hat{y})]^2} \tag{4.20}$$

where $\hat{y}$ is the predicted value and y is the actual value.

4. Finally, we estimated the preprocessing error rate by comparing the RMSE value of the original non-deformed dataset and the preprocessed dataset as follows:

$$Error_{RMSE_{processing}} = \frac{RMSE_{Preprocessed} - RMSE_{original}}{RMSE_{original}} * 100\% \tag{4.21}$$

Low values of $Error_{RMSE_{processing}}$ will be indicative of small data preprocessing impact on regression results.

4.6.2 Impact on classification results

Three classification methods based on different learning style were chosen. they are:

- Linear classification: Linear Discriminant Analysis (LDA);
- Non-Linear classification: Naïve Bayes (NB);
- Non-Linear classification with Regression Trees: Classification and Regression Trees (CART)

Similarly to regression analysis (c.f. Section 4.6.1) we have applied the classification method to different alternatives of data preprocessing outputs. Classification analysis was performed in four steps:

1. We randomly split the observations into a training and test sets where 66% of data was used for training and 34% for testing;
2. Classification models were fit on the training set, and the fitted model was used to predict the responses for the observations in the testing set;
3. Resulting validation of the previous step was assessed through computation of accuracy, and Cohen's Kappa coefficient (Cohen, 1960).
   Accuracy is the number of correct predictions made divided by the total number of predictions

$$Accuracy = \frac{number\ of\ correct\ classifications}{total\ number\ of\ classifications} \tag{4.22}$$

The kappa coefficient (Cohen, 1960) is commonly used as a measure of agreement between two classifications, it ranges from -1 to $+1$. A value of 1 indicates perfect agreement.
Let us consider the following results from two classifiers;

Example. Classification results of two classifiers for 20 observations:
Classifier 1 : b, b, c, a, c, c, c, a, a, b, c, b, b, a, c, a, b, c, c, a.
Classifier 2 : b, b, b, a, c, c, b, a, a, c, c, b, b, a, c, b, c, c, c, a.

The resulting classifications can be assimilated into two qualitative variables of p terms (here $p = 3$) giving the following contingency table:

Table 4.5 – Contingency table obtained from the results of two classifiers.

	A	**B**	**C**
A	5	1	0
B	0	4	2
C	0	2	6

In the case of perfect agreement between the two classifiers, the contingency table would have been zero out of the diagonal. By denoting the contingency table as $N = (n_{ij})_{i,j=1,\dots,p}$ and n the total observations, the observed proportional agreement is defined as:

$$p_0 = \frac{1}{n}\sum_{i=1}^{p} n_{ii} \tag{4.23}$$

In the case where the two variables are independent (i.e., if the agreement between the two classifiers was perfectly random), the theoretical proportion of observed agreements is estimated by :

$$p_e = \frac{1}{n^2}\sum_{i=1}^{p} n_{i.} n_{.i} \tag{4.24}$$

kappa is then defined as:

$$\kappa = \frac{p_0 - p_e}{1 - p_e} \tag{4.25}$$

4. Lastly, preprocessing error was computed using Eq (4.26) and Eq (4.27) as follows:

$$Absolute\ Error_{Accuracy_{processing}} = |Accuracy_{Preprocessed} - Accuracy_{original}| \tag{4.26}$$

$$Absolute\ Error_{Kappa_{processing}} = |Kappa_{Preprocessed} - Kappa_{original}| \tag{4.27}$$

4.6.3 Impact on clustering results

Clustering analysis was performed using the K-means (KM) and Hierarchical clustering (HC) methods on our preprocessed data that resulted from feature selection (c.f. Subsection 4.2.1), normalization (c.f. Subsection 4.2.2), imputation of missing values (c.f. Section 4.5.3) and outliers processing (c.f. Section 4.5.4). The two clustering methods were chosen for their algorithmic differences.

To perform K-means clustering, we first specified a number of clusters K using the elbow method (Thorndike, 1953). The elbow method was applied first to the original dataset and, the K number of clusters found was then used on the processed dataset. Then, we performed the K-means algorithm with the previously specified number of clusters. Hierarchical clustering was computed using the Ward's agglomeration method (Ward Jr, 1963).

Clustering results of K-means on preprocessed data were compared against the clustering results of the original non-deformed datasets using the Rand Index (RI, Equation 4.30), Adjusted Rand Index (AR, Equation 4.31) and Jaccard Index (JI, Equation 4.32) (Meilă, 2007).

From resulting cluster C_k of the original non-preprocessed dataset D, the number of data points in D and in C_k are defined as n and n_k, respectively thus,

$$n = \sum_{k=1}^{K} n_k \tag{4.28}$$

A second cluster of the preprocessed dataset D is defined as $C' = \{C'_1, C'_2, ..., C'_{k'}\}$, and the cluster sizes as $n'_{k'}$.

To compare clustering from the pair (C, C'), we use a *confusion matrix*. Within the *confusion matrix* (which is a $K \times K^{'}$ matrix) the $kk^{'}$th element is the number of points in the intersection of clusters C_k of C and $C'_{k'}$ of C'. Thus $n_{kk'} =\mid C_k \cap C'_{k'} \mid$.

We compare clusters by counting the pairs of points on which two clusterings agree/disagree. They were obtained from the contingency table $[n_{kk'}]$. The four cases that can be found are:
N_{11} the number of pairs that are in the same cluster under both C and C';
N_{00} the number of pairs in different clusters under both C and C';
N_{10} the number of pairs in the same cluster under C but not under C';
N_{01} the number of pairs in the same cluster under C' but not under C;

The four counts satisfy

$$N_{11} + N_{00} + N_{10} + N_{01} = n(n-1)/2 \tag{4.29}$$

We use the following indices to represent the probability that a pair of points belonging to a cluster under C is also in the same cluster C'.

$$Rand\ Index : RI(C,C') = \frac{N_{11} + N_{00}}{n(n-1)/2} \tag{4.30}$$

Adjusted Rand Index :

$$\begin{aligned} AR(C,C') &= \frac{R(C,C') - E[R]}{1 - E[R]} \\ &= \frac{\sum_{k=1}^{k}\sum_{k'=1}^{K'}\binom{n_{kk'}}{2} - [\sum_{k=1}^{K}\binom{n_k}{2}][\sum_{k'=1}^{K'}\binom{n'_{k'}}{2}]/\binom{n}{2}}{[\sum_{k=1}^{K} K\binom{n_k}{2} + \sum_{k'=1}^{K'}\binom{n'_{k'}}{2}]/2 - [\sum_{k=1}^{K}\binom{n_k}{2}][\sum_{k'=1}^{K'}\binom{n'_{k'}}{2}]/\binom{n}{2}} \end{aligned} \tag{4.31}$$

$$Jaccard\ index\ : J(C,C') = \frac{N_{11}}{N_{11} + N_{01} + N_{10}} \tag{4.32}$$

For Hierarchical clustering we have compared the clustering results of preprocessed dataset against the clustering results of the original non-preprocessed dataset using the Cophenetic correlation coefficient (Sokal and Rohlf, 1962). Datasets D were modelled using the hierarchical clustering method to produce a dendrogram T which distance measures are defined as:
$x(i,j) =\mid D_i - D_j \mid$, the Euclidean distance between the ith and jth observations;
$t(i,j) =$ the distance between the model points T_i and T_j;

and the Cophenetic correlation coefficient is defined as:

$$c = \frac{\sum_{i<j}(x(i,j) - \bar{x})(t(i,j) - \bar{t})}{\sqrt{[\sum_{i<j}(x(i,j) - \bar{x})^2][\sum_{i<j}(t(i,j) - \bar{t})^2]}} \tag{4.33}$$

where $\bar{x}$ is the average of $x(i,j)$ and $\bar{t}$ the average of $t(i,j)$. The Cophenetic correlation coefficient allow us to estimate the similarities between dendograms of preprocessed and original datasets. Where Cophenetic values close to 1 indicate high similarity between two dendrograms.

4.6.4 Results and discussion about preprocessing procedures on statistical results

4.6.4.1 Regression analysis

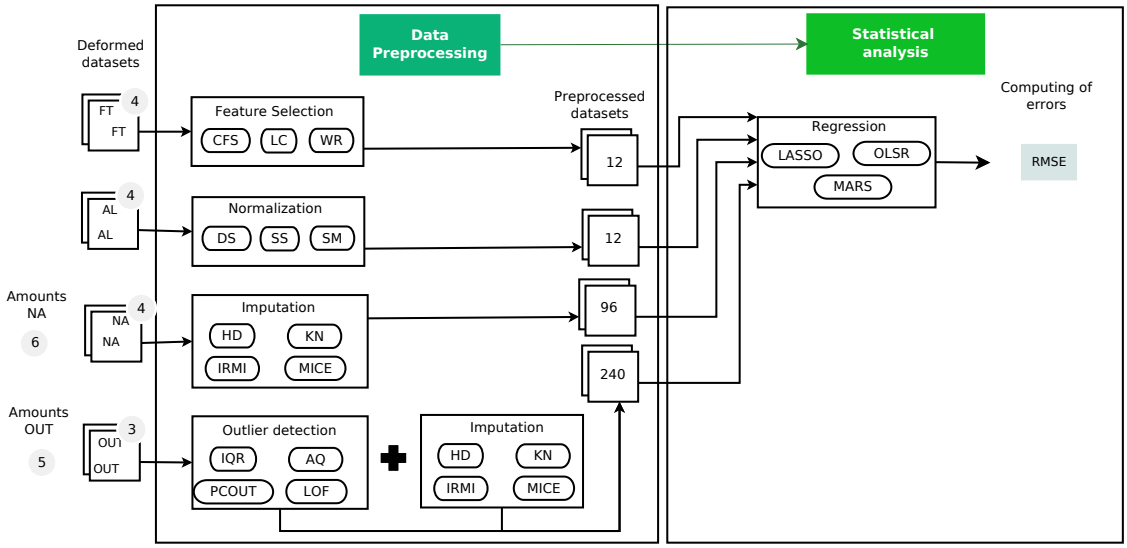


Figure 4.10 – Experimental procedure for assessment of the impact of data preprocessing procedures on results from regression analysis.

Our results report the preprocessing error values for three different regression methods (LASSO, OLSR and MARS) against (c.f. Figure 4.10):

- *For feature selection*
  - three feature selection methods; Correlation-based Feature Selection (CFS), Linear Correlation-based Feature Selection, and Wrapper Subset Evaluator.
- *For normalization preprocessing*
  - three normalization methods; Decimal scale (DS), Sigmoidal normalization using logistic sigmoid function (SS) and Sigmoidal normalization using the hyperbolic tangent funciton (SM);
- *For missing values preprocessing*
  - six amounts of missing values (5%, 10%, 15%, 20%, 25% and 30%)
  - four imputation methods; Hot-Deck (HD), K-Nearest Neighbour (KN), Iterative Stepwise Regression Imputation (IRMI) and Multiple Imputation by Chained Equation (MICE)
- *For outliers preprocessing*
  - five amounts of outlying data (1.5%, 2.5%, 5%, 10% and 15%);
  - four outlier detection methods; (1) Inter Quartile Range (IQR), (2) Adjusted Quantile (AQ), (3) Principal Components decomposition for multivariate outlier detection approach (PCOUT) and (4) Local Outlier Factor (LOF);
  - subsequent outlier imputation by four imputation methods; Hot-Deck (HD), K-Nearest Neighbour (KN), Iterative Stepwise Regression Imputation (IRMI) and Multiple Imputation by Chained Equations (MICE)

Results for feature selection

Results obtained from regression analysis on datasets processed by feature selection

are represented in Figures 4.11 and 4.12. They present the RMSE results on a side by side comparison between different feature selection methods, respective of the regression algorithms. It is known that noise features that are not truly associated with the response will lead to a deterioration of the fitted model and will increase the test-set error and the risk of over-fitting. This is important in the analysis of highly dimensional data which is known as "curse of dimensionality". This was the main reason to have performed feature selection before analysis on regression methods. However, it must be noticed that LASSO already includes feature selection. To make our analysis comparable we have performed feature selection before analysis on LASSO but this step should be omitted on a regression analysis by LASSO.

For our synthetic datasets we observed that in general, Linear correlation-based feature selection show the lowest RMSE values for LASSO and OLSR regression methods. For MARS regression, none of the three feature selection methods stand up as the best for $FS21$, $FS4000$ and $FS20000$ datasets. For $FS600$ dataset the Linear correlation-based method show the best results. Concerning our semi-synthetic datasets the linear correlation-based feature selection show the lowest RMSE values for the three regression methods.

Results for normalization

Predictive variables can be normalized in order to obtain a numerical stability that enables one to compare effects across multiple explanatory variables. In other words, by normalizing predictive variables one assure to give the same weight to all the predictors. One important aspect in linear regression is that linear transformations are not expected to affect regression results because the model coefficients (intercept and slope terms in the linear model) are estimated in order to convert the units of predictive variables into the units of response variable. In this manner, linear transformation of predictive variables does not seem to be meaningful in linear regression. Our goal in this section was to assess the variability induced by different non-linear transformations on the final results of regression analysis. To our purpose we have performed three non-linear normalizations they are: decimal scale and two sigmoidal normalizations. They were computed using the following formulas:

- Decimal scale normalization:

$$v_i^{'} = \frac{v_i}{10^j} \tag{4.34}$$

- Sigmoidal normalization using logistic sigmoid function:

$$v_i^{'} = \frac{1}{1+e^{-\frac{v_i-\mu_i}{\sigma_i}}} \tag{4.35}$$

- Sigmoidal normalization using the hyperbolic tangent function:

$$v_i^{'} = \frac{1-e^{-\frac{v_i-\mu_i}{\sigma_i}}}{1+e^{-\frac{v_i-\mu i}{\sigma_i}}} \tag{4.36}$$

  where v_i denotes the values of an attribute, $v_i^{'}$ is the normalization of v_i and j is the smallest integer such that $Max(|v_i^{'}|) < 1$.

We have selected one response variable and the remaining variables were used as predictors. Only predictor variables were transformed (non Gaussian initially). Figures 4.11 and 4.12 show that there is not a significant difference when non-linear transformations are applied to predictor variables on LASSO, OLSR and MARS regressions.

Our results indicate that, similarly to linear transformations, non-linear transformations do not affect linear regression results. Such behaviour can be explained through the estimation of the model coefficients because they produce estimates in order to transform the units of each predictor variable into units of the response variable appropriately. Our results suggest that transformations on data are not relevant when regression by LASSO, OLSR and MARS are performed.

Results for missing values preprocessing

A side-by-side comparison between different imputation methods, respective of the regression algorithms, is given in Figures 4.15 and 4.16. It shows the preprocessing error rates computed for RMSE over the three regressions methods and the four datasets, resulting from imputation against different amounts of missing values. The preprocessing error rates are provided for the four imputation methods.

Preprocessing error results on our synthetic datasets show that the impact of imputation varies for different regression methods. In general, at high amounts of missing values (25% and 30%) imputation methods Hot-Deck and MICE give the lowest error values. The lowest error rates were observed for LASSO and OLSR regression methods (with 7% and 9% on average respectively). In general, the highest RMSE values were observed for the dataset of small size (N21) and the highest error values were observed on dataset N21 for MARS regression. High preprocessing error values indicate high probabilities to get inaccurate regression results. From this observation, we could conclude that inaccurate results may be obtained when regression by MARS is performed after imputing missing values particularly for small datasets (<100 observations) and missing data percentage superior to 5%. Imputation is required for LASSO and OLSR and it was observed that Hot-Deck and MICE imputation methods impact the least LASSO and OLSR regression results. LASSO and OLSR methods are the least impacted by imputation methods when the dataset size is greater than 100 observations and the percentage of missing values is lower than 20%.

Whit respect to the semi-synthetic data (c.f., Figure 4.16) we observed that error values increase with respect to the increased missing data rate. We observed that the imputation methods Hot-Deck and K-Nearest Neighbour have the highest error for the three datasets. Results about the imputation methods are not the same between the two groups of dataset (synthetic and semi-synthetic) such behaviour is justified by the fact that datasets have different characteristics (different number of observations and variables) however, we noticed that both groups of datasets have the same tendency. This means that error values increase with respect to the increased missing rates.

Results on outliers preprocessing

A side-by-side comparison between different outlier detection and imputation methods, respective of the regression algorithms is given in Figures 4.17 and 4.18. It shows the preprocessing error rates of RMSE over the three regression methods and the synthetic and semi-synthetic datasets, resulting from outlier detection-imputation with respect to different amounts of outlying data. The preprocessing error rates are provided for the sixteen combinations of outlier detection and imputation methods.

In general, for our synthetic datasets we observed that at small amounts of outlying data (1.5% and 2.5%) and small length dataset (N21), the combined methods PCOUT-Hot-Deck and PCOUT-IRMI give the lowest preprocessing error values. While at high amounts of outlying data (5%, 10%, or 15%) and large datasets (e.g., N600 and N4000), detection methods PCOUT and LOF combined with imputation methods MICE and IRMI give the lowest preprocessing error values. We noticed that, in general, at high amounts of

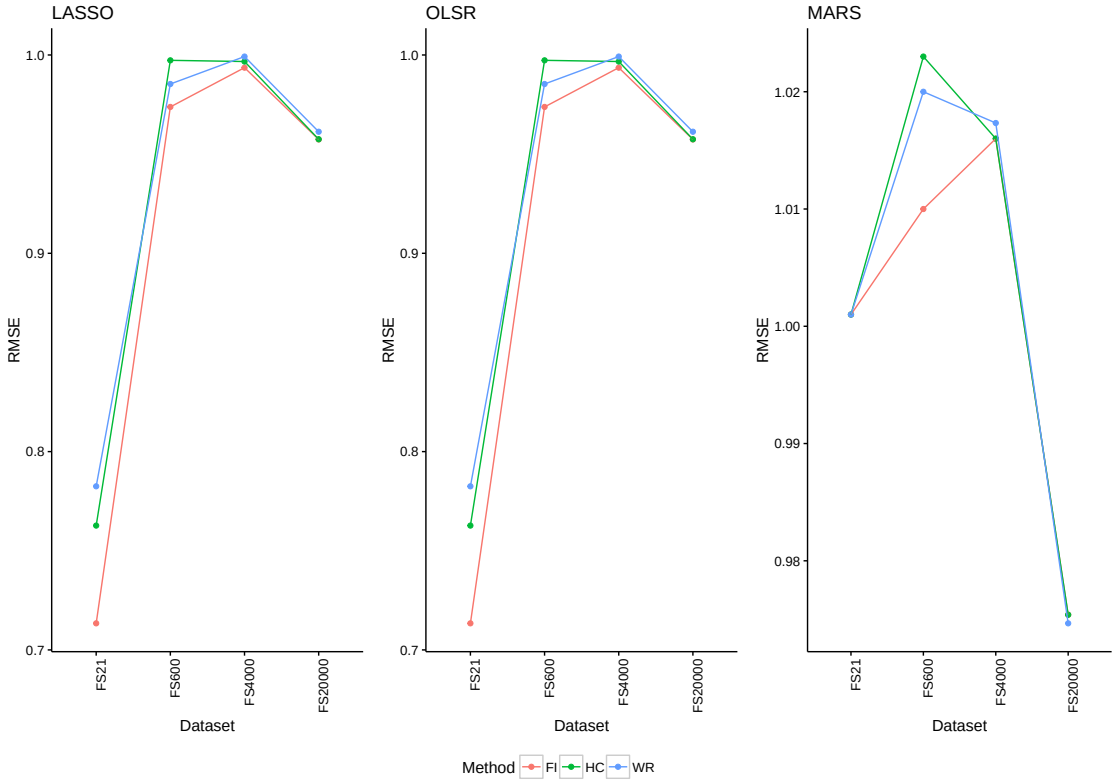


Figure 4.11 – RMSE results of the analysis of regression after feature selection processing on synthetic datasets. Linear correlation-based feature selection (FI), Correlation-based Feature selection (HC) and Wrapper Subset evaluator (WR).

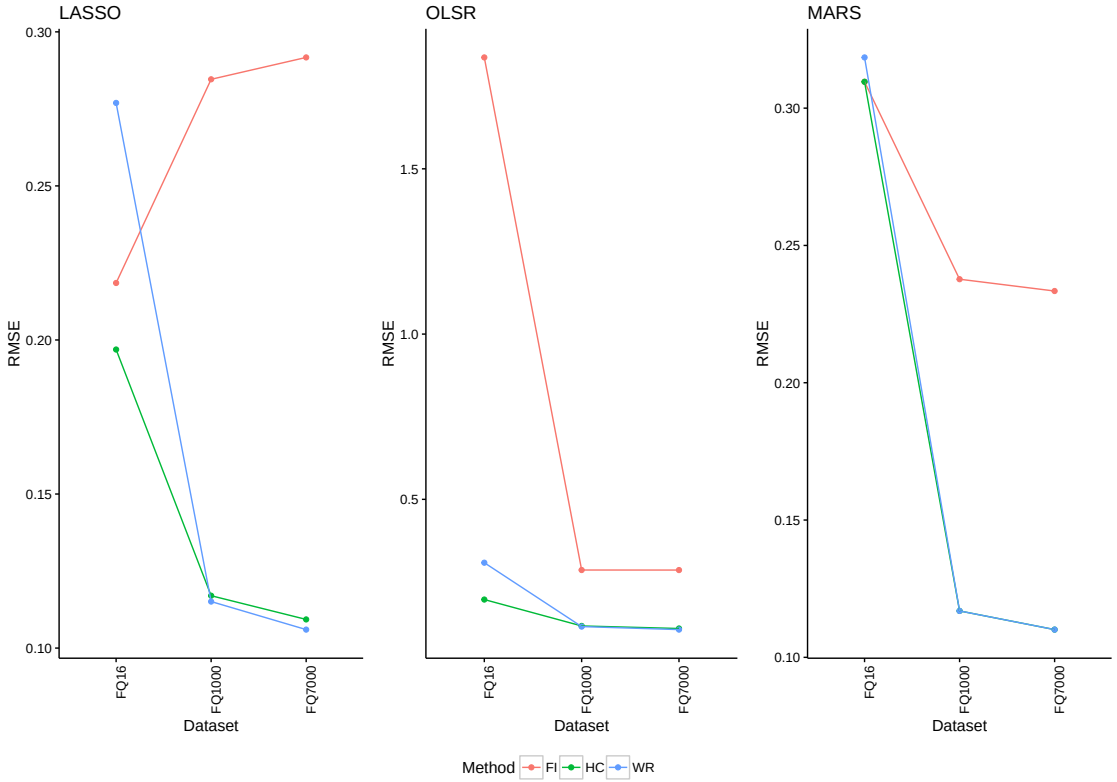


Figure 4.12 – RMSE results of the analysis of regression after feature selection processing on semi-synthetic datasets. Linear correlation-based feature selection (FI), Correlation-based Feature selection (HC) and Wrapper Subset evaluator (WR).

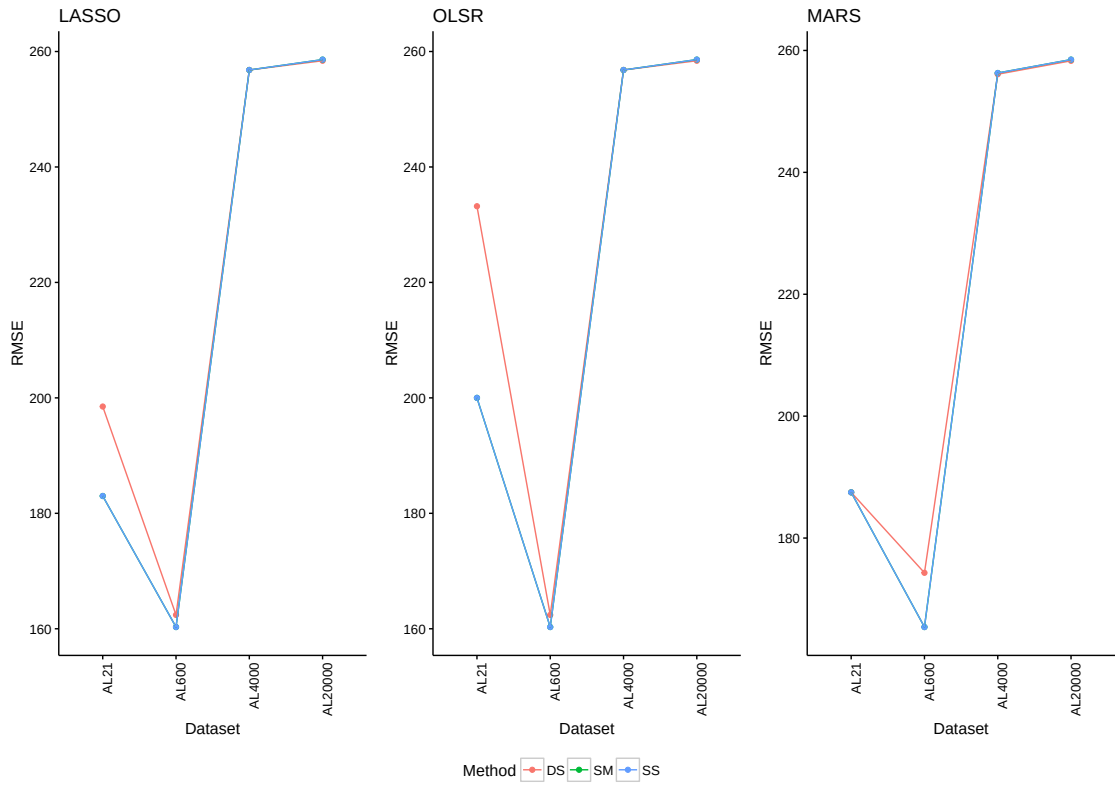


Figure 4.13 – RMSE results of the analysis of regression after normalization on synthetic datasets. Decimal scale noramlization (DS), Logistic sigmoidal normalization (SS) and Sigmoidal normalization using the hyperbolic tangent function (SM).

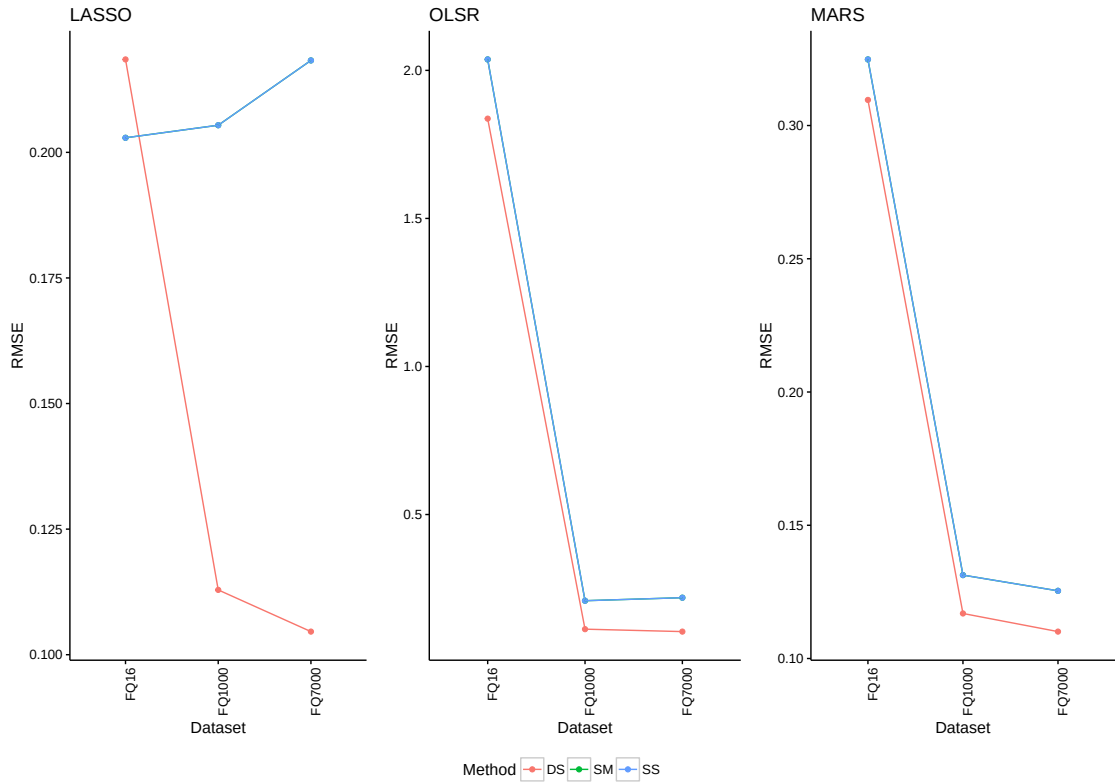


Figure 4.14 – RMSE results of the analysis of regression after normalization on semi-synthetic datasets. Decimal scale normalization (DS), Logistic sigmoidal normalization (SS) and Sigmoidal normalization using the hyperbolic tangent function (SM).

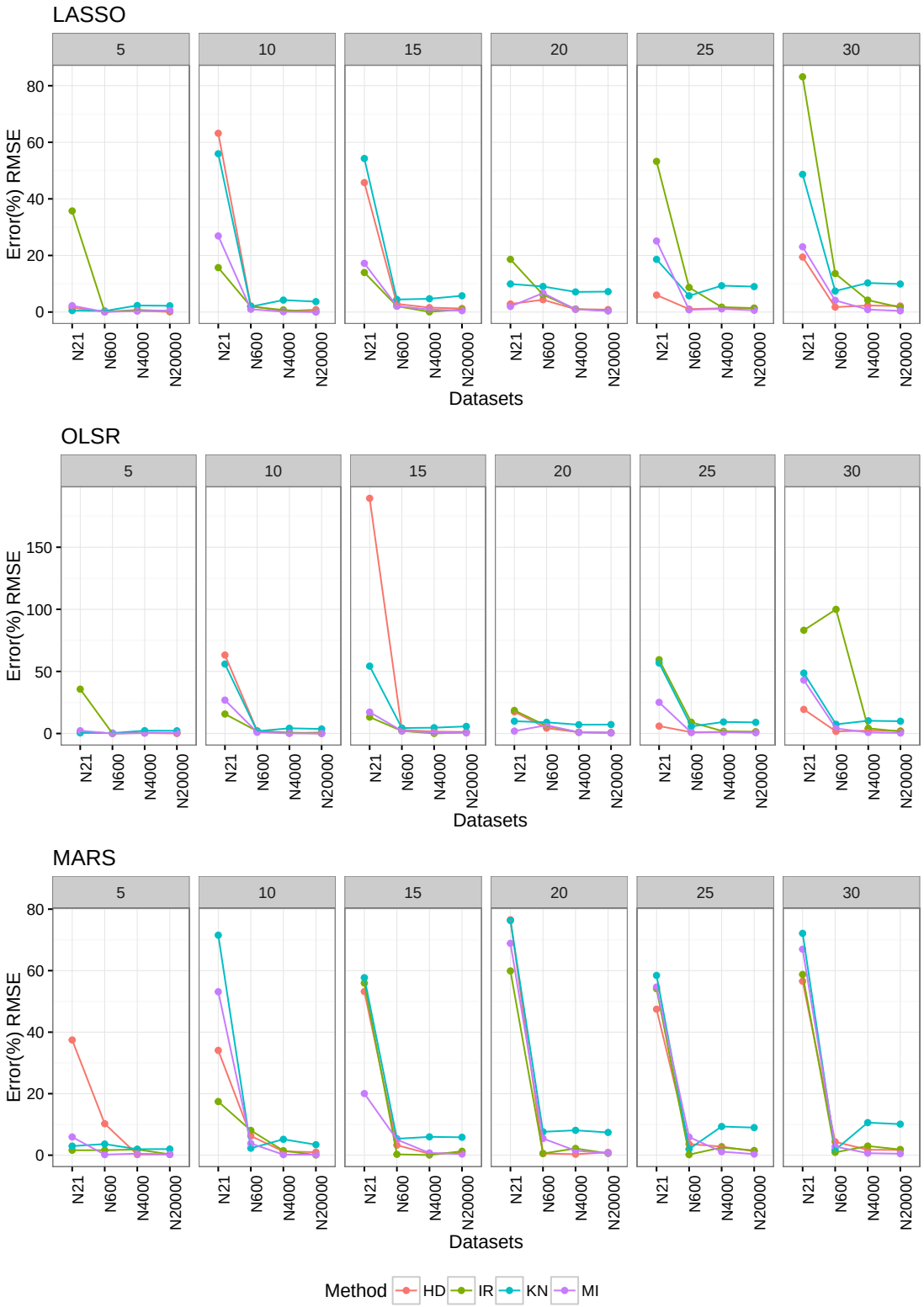


Figure 4.15 – Preprocessing errors of RMSE of the analysis of regression after imputation of missing values on synthetic dataset. Hot-deck (HD), Iterative Stepwise Regression Imputation (IR), K-Nearest Neighbour (KN) and Multiple Imputation by Chained Equations (MI)

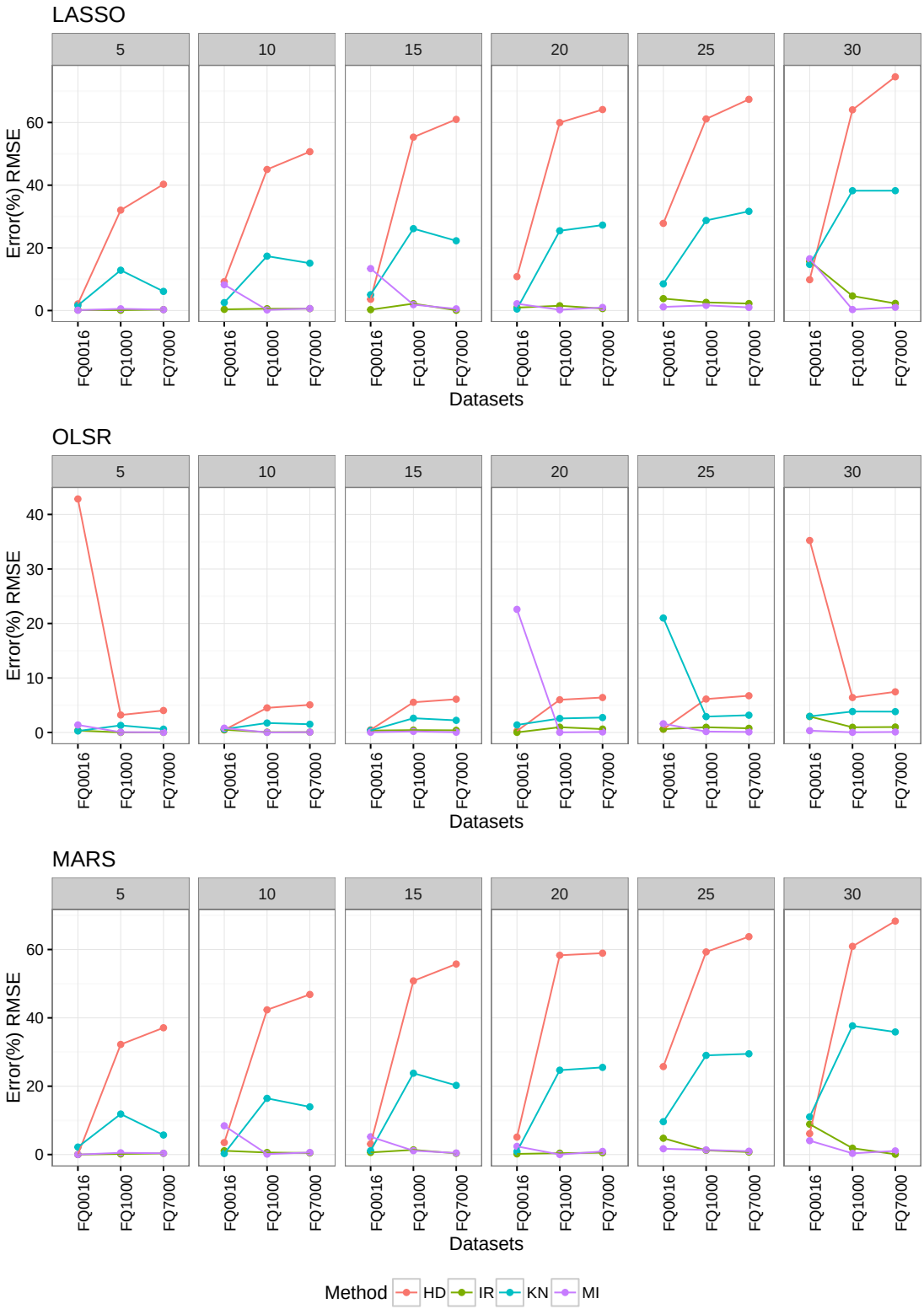


Figure 4.16 – Preprocessing errors of RMSE of the analysis of regression after imputation of missing values on semi-synthetic datasets. Hot-deck (HD), Iterative Stepwise Regression Imputation (IR), K-Nearest Neighbour (KN) and Multiple Imputation by Chained Equations (MI)

outlying data (10% and 15%) preprocessing error rates were higher. For our semi-synthetic datasets (c.f., Figure 4.18), we observed that error rates increase at higher amounts of outlying data. For LASSO the detection method IQR combined with IRMI and MICE show the lowest error values. While for OLSR and MARS we did not observe remarkable differences among the outlying preprocessing methods.

It has been widely suggested to process outliers carefully because, depending of the observer they may provide different information according to the domain of study and could be considered either as noise or as signal (Huang et al., 2006). According to our results, and if the observer considers to preprocess outlying data, we could suggest the use of multivariate methods (i.e., PCOUT and LOF combined with MICE and IRMI).

For the specific characteristics of our synthetic datasets, we observed that, in general, multivariate methods give the best results and particularly on large datasets with high amounts of missing values and outlying data. We interestingly observed that, simple methods (e.g., K-NN imputation method or Inter Quartile Range) provide the best results on our small datasets with small amounts of data anomalies.

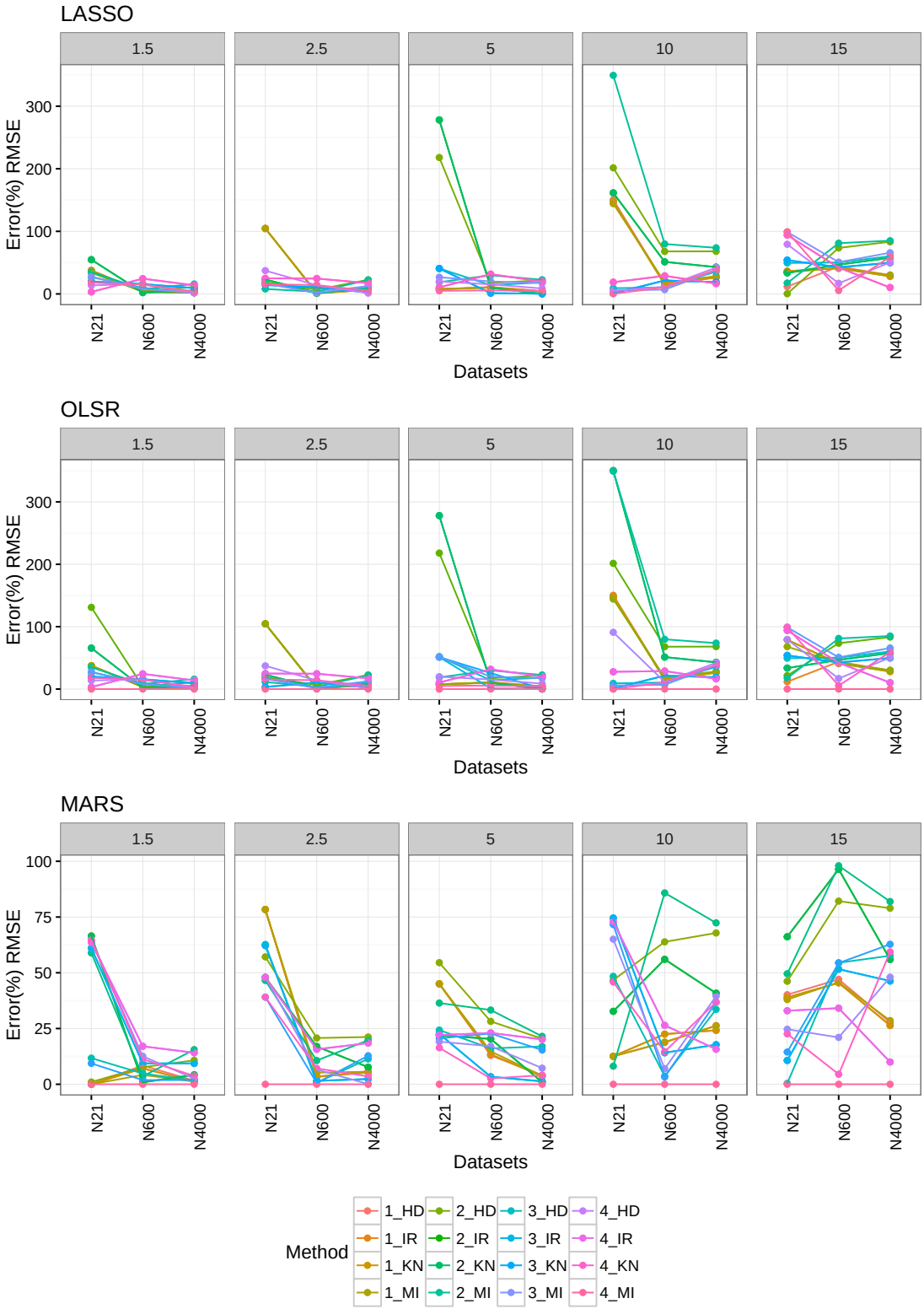


Figure 4.17 – Preprocessing errors of RMSE of the analysis of regression after outlier detection followed by imputation of outlying data on synthetic datasets. Detection methods: Inter Quartile Range (1), Adjusted Quantile (2), PCOUT (3) and LOF (4). Imputation methods: Hot-deck (HD), Iterative Stepwise Regression Imputation (IR), K-Nearest Neighbour (KN) and Multiple Imputation by Chained Equations (MI).

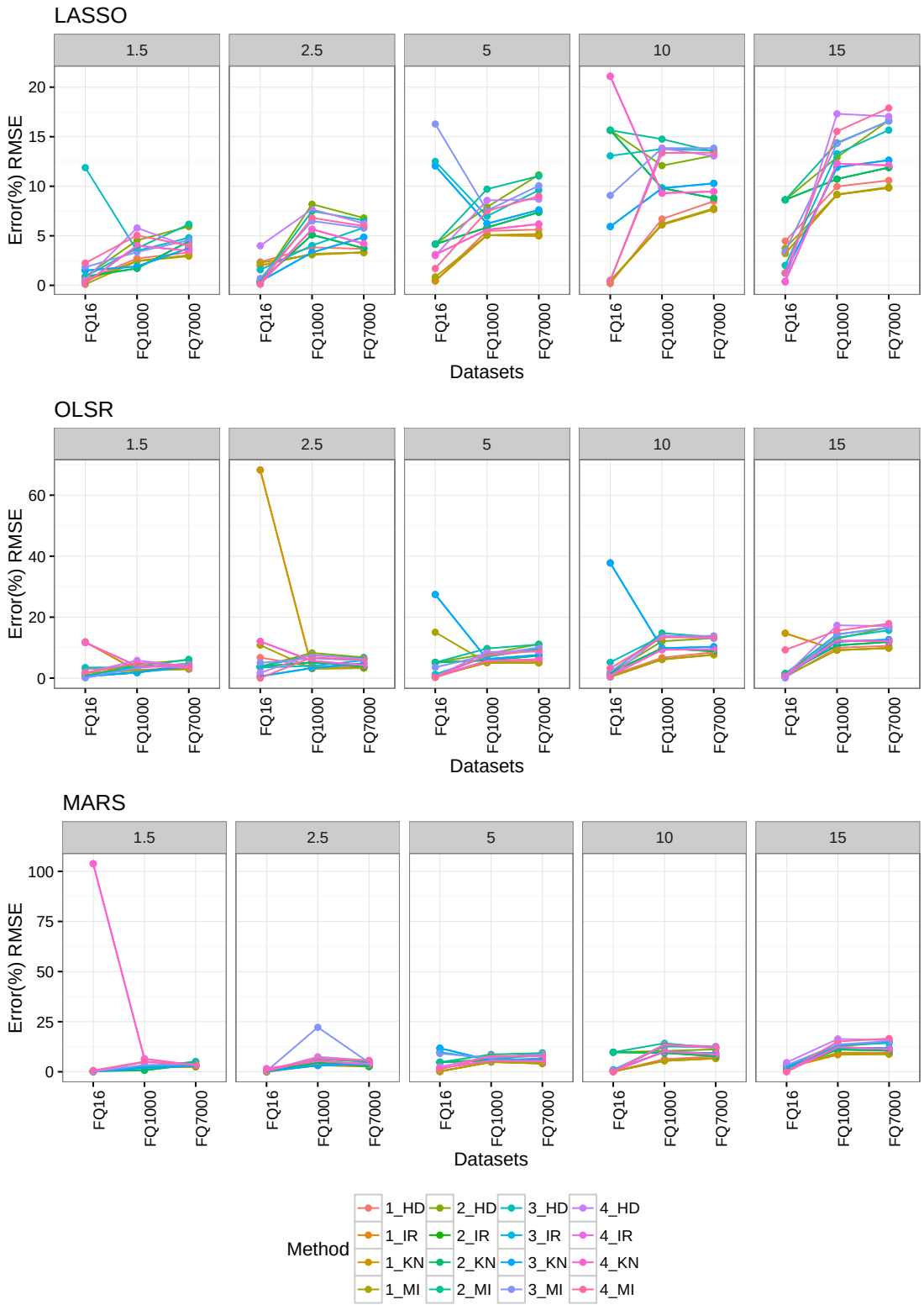


Figure 4.18 – Preprocessing errors of RMSE of the analysis of regression after outlier detection followed by imputation of outlying data on semi-synthetic datasets. Detection methods: Inter Quartile Range (1), Adjusted Quantile (2), PCOUT (3) and LOF (4). Imputation methods: Hot-deck (HD), Iterative Stepwise Regression Imputation (IR), K-Nearest Neighbour (KN) and Multiple Imputation by Chained Equations (MI).

4.6.4.2 Classification analysis

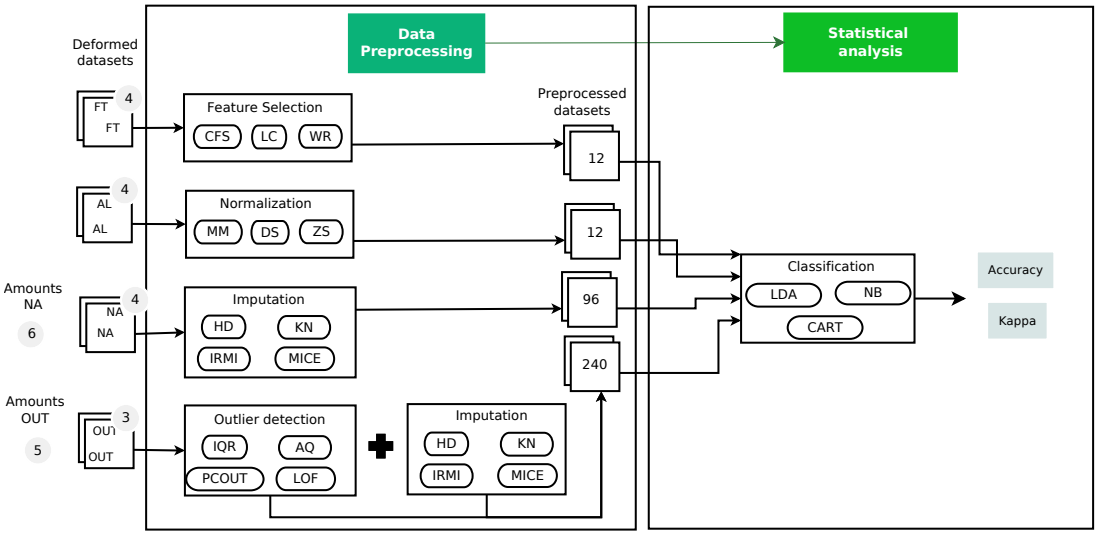


Figure 4.19 – Experimental procedure for assessment of the impact of data preprocessing procedures on results from classification.

In these experiments, we report the preprocessing error values for three different classification methods: Classification and Regression Trees (CART), Linear Discriminant analysis (LDA) and Naïve Bayes (NB) with respect to (c.f. Figure 4.19):

- *For feature selection*
  - three feature selection methods; Correlation-based Feature Selection (CFS), Linear Correlation-based Feature Selection and Wrapper Subset Evaluators.
- *For normalization preprocessing*
  - three normalization methods; min-max (MM), decimal scale (DS) and z-score (ZS)
- *For missing values preprocessing*
  - six amounts of missing values (5%, 10%, 15%, 20%, 25% and 30%)
  - four imputation methods; Hot-Deck (HD), K-Nearest Neighbour (KN), Iterative Stepwise Regression Imputation (IRMI), and Multiple Imputation by Chained Equations (MICE)
- *For outlier preprocessing*
  - five amounts of outlying data (1.5%, 2.5%, 5%, 10% and 15%)
  - four outlier detection methods; (1) Inter Quartile Range (IQR), (2) Adjusted Quantile, (3) a Principal Components decomposition for multivariate outlier detection approach (PCOUT), and (4) Local Outlier Factor (LOF);
  - subsequent outlier imputation by four imputation methods; Hot-Deck (HD), K-Nearest Neighbour (KN), Iterative Stepwise Regression Imputation (IRMI), and Multiple Imputation by Chained Equations (MICE)

Results for feature selection

Classification results on data preprocessed by feature selection are illustrated in Figures 4.20 and 4.21. It shows a side-by-side comparison of Accuracy and Kappa results over the three classification methods and the synthetic and semi-synthetic datasets resulting from feature selection.

Results on our synthetic datasets for accuracy show that the filter method Correlation-based feature selection give the most accurate results on the four datasets for the three classifiers. Concerning Kappa, the best results for LDA were observed on the Correlation-based feature selector. While for NB classification, the Linear Correlation-based feature selection show the best results. For CART classifier none of the feature selection methods stand-up as the best one.

For our semi-synthetic datasets the wrapper subset evaluator give the best results for LDA, while the linear correlation-based feature selection give best results for NB. Finally, non of the feature selection methods stand-up as the best for CART classifier.

CART is known to be highly non robust this explains its behaviour in our results. We consider that the correlation-based feature selection methods show the best results on our synthetic dataset due to the characteristics of our datasets (e.g., variables with a correlation > 0.7, multivariate distribution).

Results for normalization

Figures 4.22 and 4.23 shows a side-by-side comparison of the results of accuracy and Kappa values over the three classification methods and the four datasets resulting from normalization.

The results on our synthetic datasets show that, decimal-scale normalization provides the best results for $AL21$ and $AL20000$ datasets on the LDA and NB classifiers, while for $AL600$ and $AL4000$ datasets, the min-max and z-score normalizations stand up as the best for LDA and NB respectively. Except for $AL4000$ dataset the z-score normalization gives the best results for CART classifier. For our semi-synthetic datasets the z-score and min-max normalitzations give the best results for LDA and NB respectively. For CART, none of the normalization strategies stand up as the best.

From Kappa results we observed that, for small datasets decimal-scale and z-score give the best results for synthetic and semi-synthetic datasets respectively. While for long datasets (e.g., $AL4000$, $AL20000$ and $FQ1000$) none of the methods stand up as the best. Our results suggest that the selection of normalization method will provide different classification results. We assume that the difference in our synthetic and semi-synthetic datasets is due to the characteristics of our datasets (e.g., distribution, size) and on the differences of the learning style of the three classifiers.

Results for missing values preprocessing

A side-by-side comparison between different imputation methods, with respect to the regression algorithms, is given in Figures 4.24 and 4.25. They show the preprocessing absolute errors computed for accuracy and Kappa over the three classification methods and the synthetic and semi-synthetic datasets, resulting from imputation against different amount of missing values. The preprocessing absolute errors are provided for the four imputation methods.

Results show that, in general, for small datasets ($N21$, $FQ16$), imputation methods Hot-Deck and k-Nearest Neighbour give low preprocessing error rates at small amounts of missing values (5%, 10%, and 15%). For large datasets (N4000, N20000, FQ1000 and FQ7000) imputation methods IRMI and MICE show the lowest preprocessing error values in both, accuracy and Kappa, for the six amounts of missing values. For the three classification methods the highest preprocessing error values of accuracy were observed on N21 and FQ16 datasets for the four imputation methods.

Concerning the preprocessing error results for Kappa, we observed that small datasets ($N21$, $FQ16$) and small amounts of missing values (5%, 10%, and 15%) the imputation

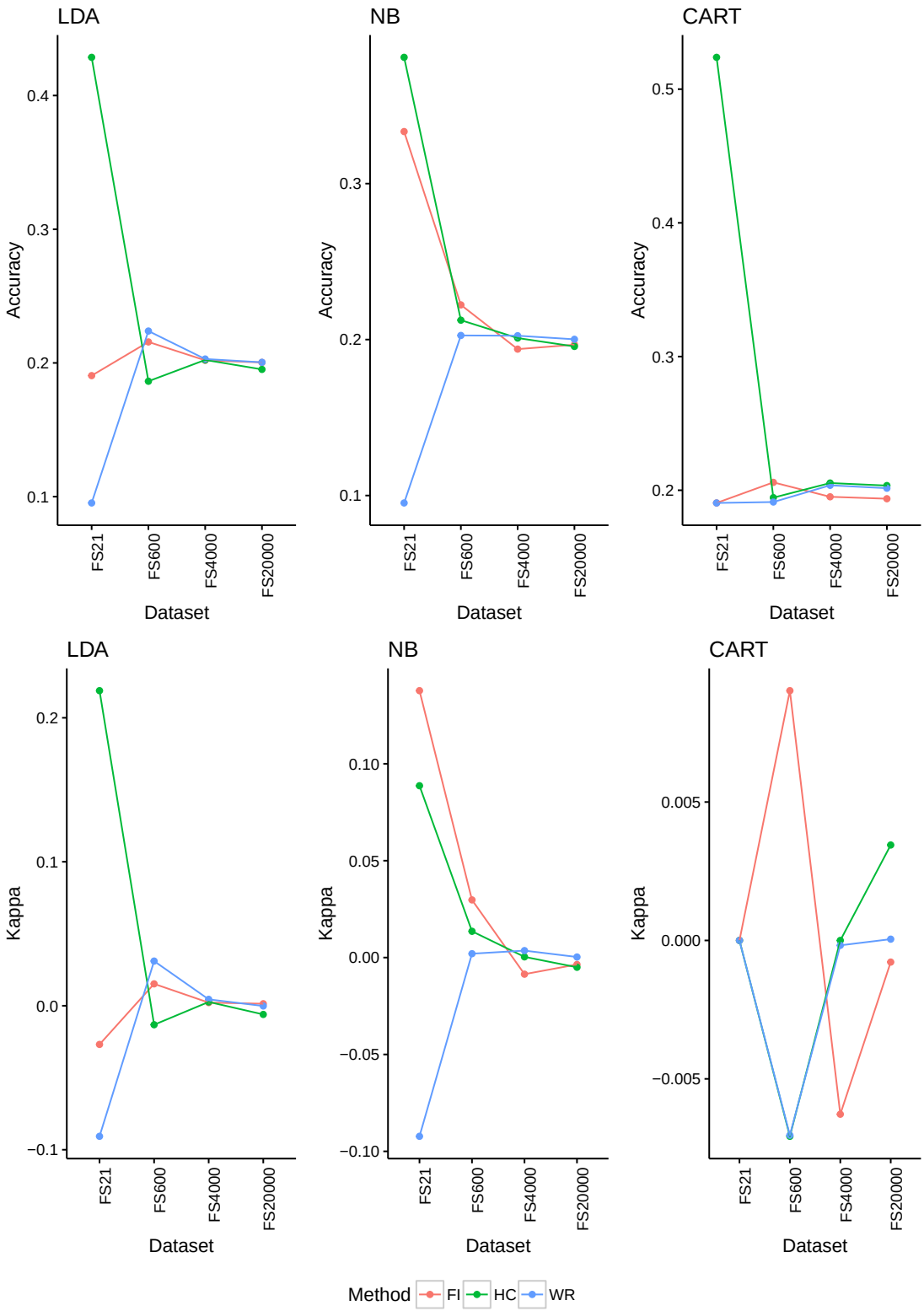


Figure 4.20 – Accuracy and Kappa results of the analysis of classification after feature selection on synthetic datasets. Linear Correlation-based Feature Selection (FI), Correlation-based Feature selection (HC) and Wrapper Subset evaluator (WR).

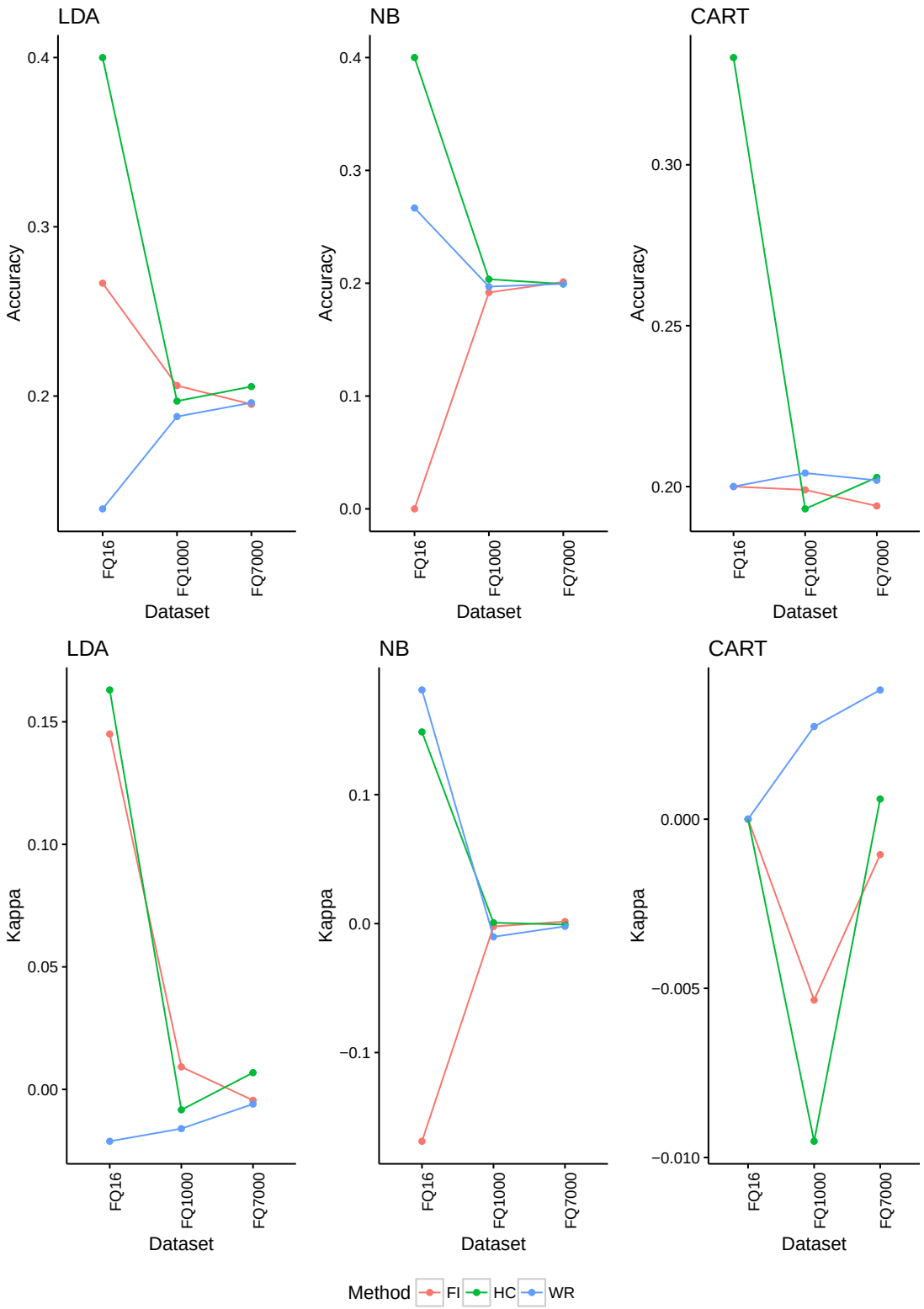


Figure 4.21 – Accuracy and Kappa results of the analysis of classification after feature selection on semi-synthetic datasets. Linear Correlation-based Feature Selection (FI), Correlation-based Feature selection (HC) and Wrapper Subset evaluator (WR).

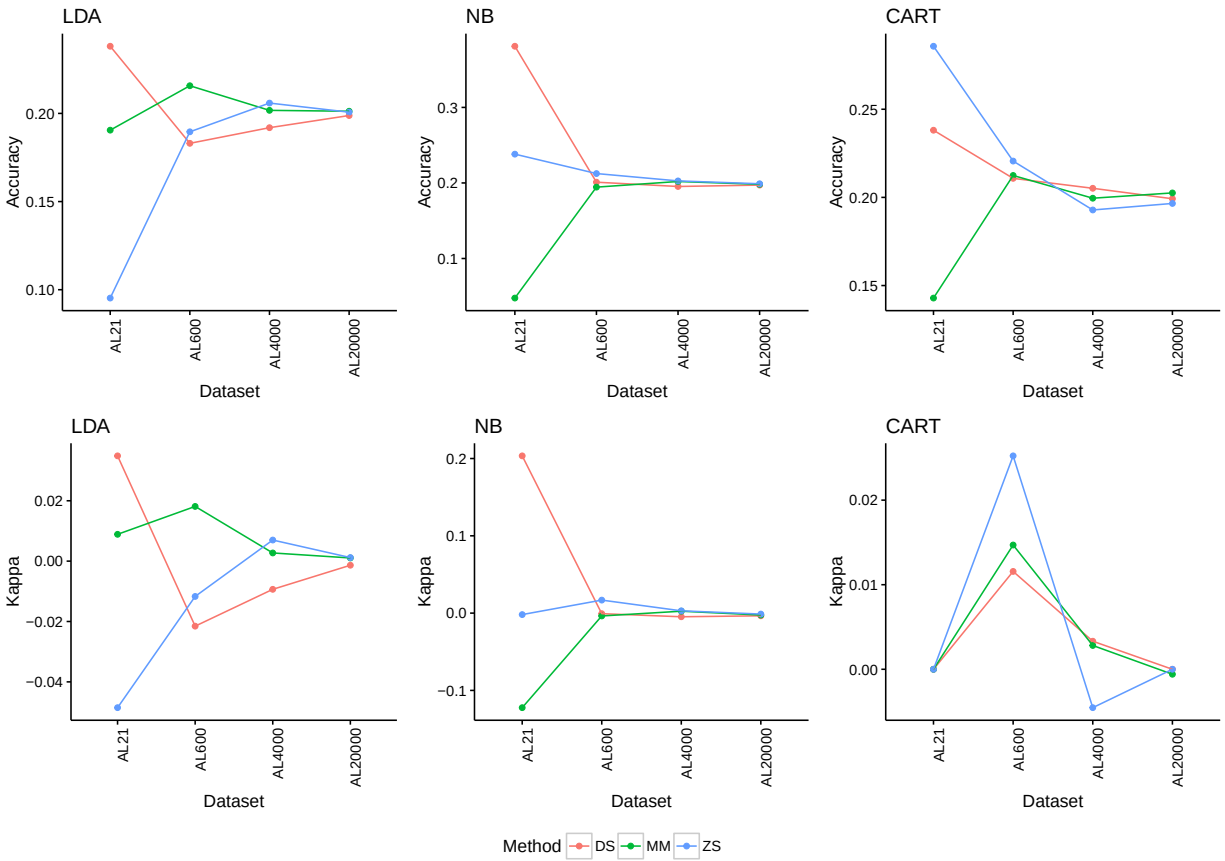


Figure 4.22 – Accuracy and Kappa results of the analysis of classification after normalization on synthetic datasets. Min-Max normalization (MM), Decimal scale normalizaton (DS) and Z-score normalization (ZS).

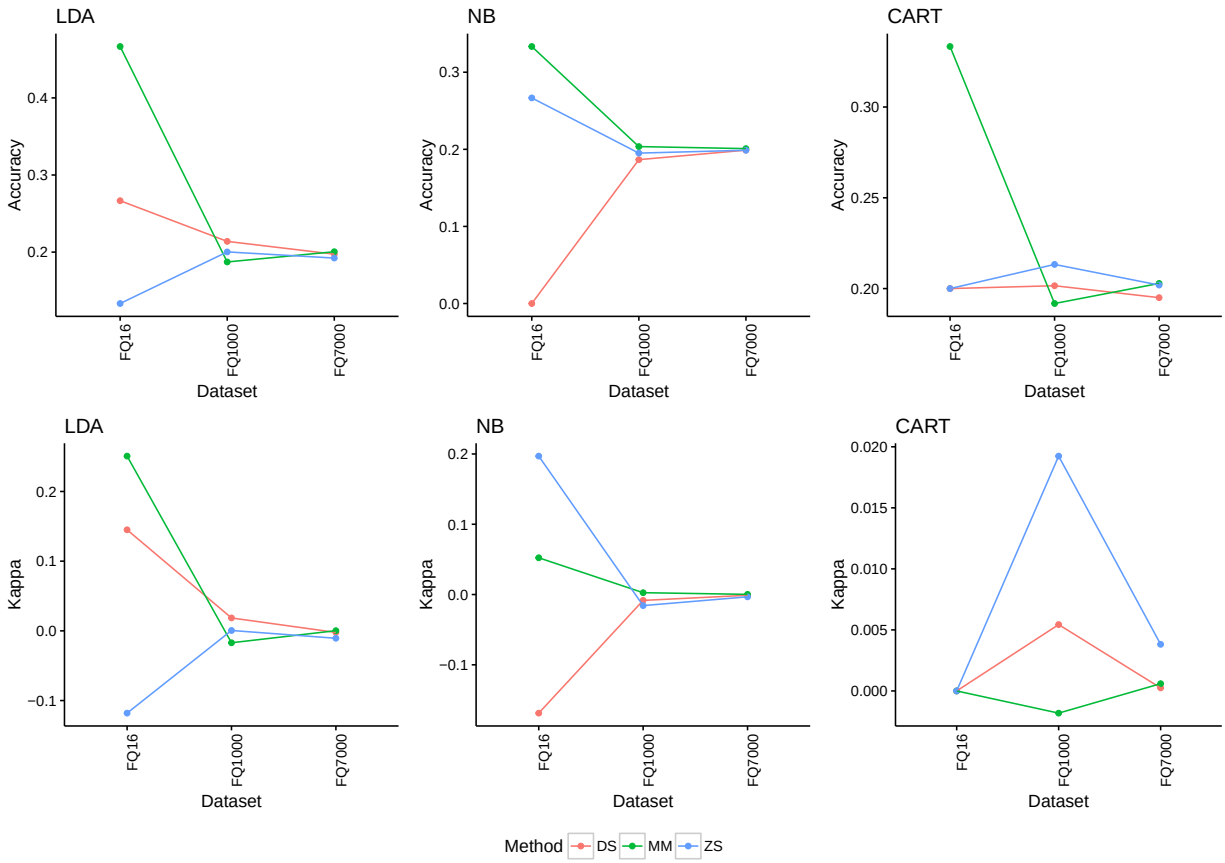


Figure 4.23 – Accuracy and Kappa results of the analysis of classification after normalization on semi-synthetic datasets. Min-Max normalization (MM), Decimal scale normalizaton (DS) and Z-score normalization (ZS).

method Hot-Deck have the lowest error values on the three classification methods for the seven datasets. From these results, we could conclude that for datasets with similar characteristics as ours and small amounts of missing values (5%, 10%, or 15%), the imputation by Hot-Deck will have the lowest impact on CART, LDA, and NB classification methods. While for large datasets (more than 600 observations and 30 variables), IRMI and MICE imputation methods should be preferred. We inferred that this behaviour depends directly on the characteristics of the datasets that we use for our study. Therefore, results may change if datasets have different characteristics than ours.

Results on outlying data preprocessing

In Figures 4.26 and 4.27, we present a side-by-side comparison between different outlier detection and imputation methods, with respect to the classification algorithms. It shows the preprocessing absolute errors computed for accuracy and Kappa over the three classification methods and the synthetic and semi-synthetic datasets, resulting from outlier detection-imputation against different amounts of outlying data. Sixteen combinations of outlier detection and imputation methods are shown.

Results show that, on average, there is no universally best method to detect and impute outliers. The impact of detection-imputation of outliers varies for different classifiers. In general, we observed that classification results were the most impacted on the small dataset (N21 and FQ16), where preprocessing error values of accuracy were, on average, over 0.4. CART classifier seems to be the most affected on N21, F16, N600 and FQ1000 datasets for the five amounts of outlying data.

Concerning the preprocessing error values for Kappa, in general, small datasets (N21 and FQ16) gives the highest errors for LDA and NB classifiers. While for CART, the highest errors were observed on N600 and FQ1000 dataset for the five amounts of outlying data. Among the sixteen combinations assessed, we noticed that outlier detection methods PCOUT and LOF combined with IRMI and MICE imputation methods provide the lowest error values on the three classifiers and on the five amounts of outlying data. For synthetic datasets, the detection method PCOUT combined with Hot-Deck imputation also shows low preprocessing error values on the three classifiers particularly for N21 and N600 datasets at outlying data rates of 2.5%, 5%, and 10%.

Our experiments on synthetic and semi-synthetic datasets show that results on outliers preprocessing depend on the characteristics of the datasets. We can see that multivariate methods (e.g., PCOUT and MICE) give the best results for our datasets which have a multivariate distribution with variables somehow correlated. The results observed for CART classifier, indicate us that this method is the most impacted by the preprocessing procedure which is explained due to its poor robustness.

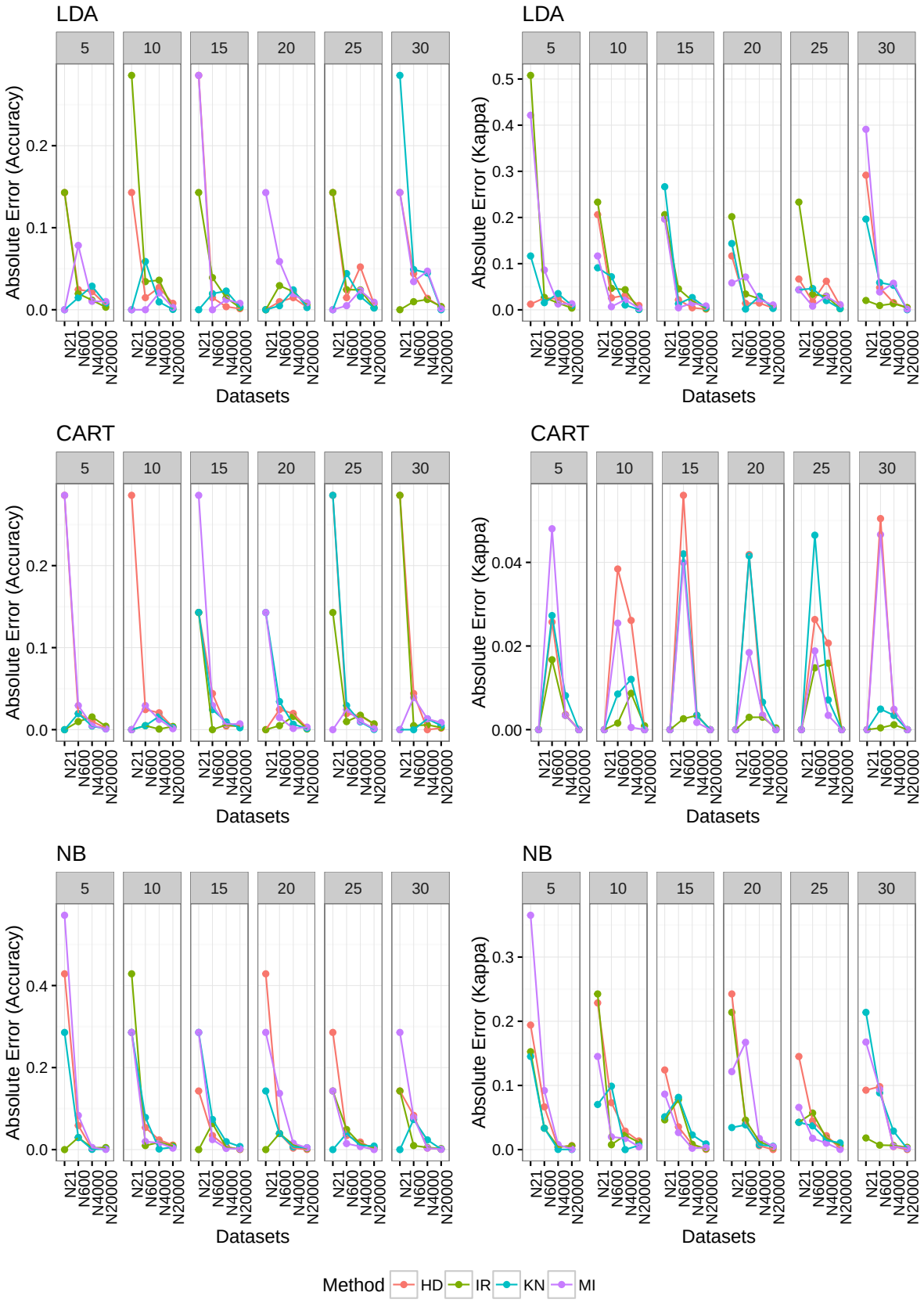


Figure 4.24 – Accuracy and Kappa preprocessing error from the analysis of classification after imputation of missing data on synthetic datasets. Imputation methods: Hot-deck (HD), Iterative Stepwise Regression Imputation (IR), K-Nearest Neighbour (KN) and Multiple Imputation by Chained Equations (MI).

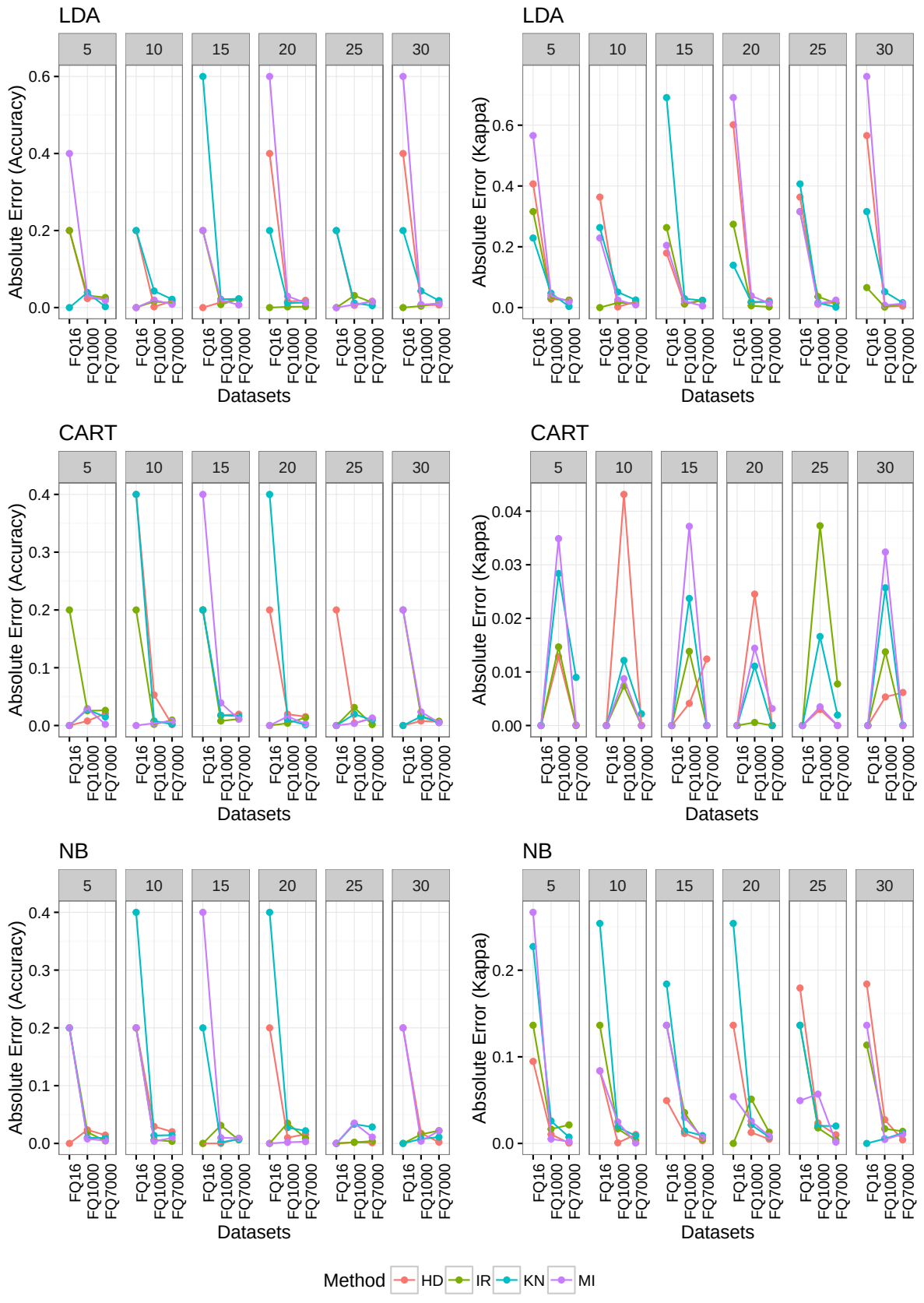


Figure 4.25 – Accuracy and Kappa preprocessing error from the analysis of classification after imputation of missing data on semi-synthetic datasets. Imputation methods: Hot-deck (HD), Iterative Stepwise Regression Imputation (IR), K-Nearest Neighbour (KN) and Multiple Imputation by Chained Equations (MI).

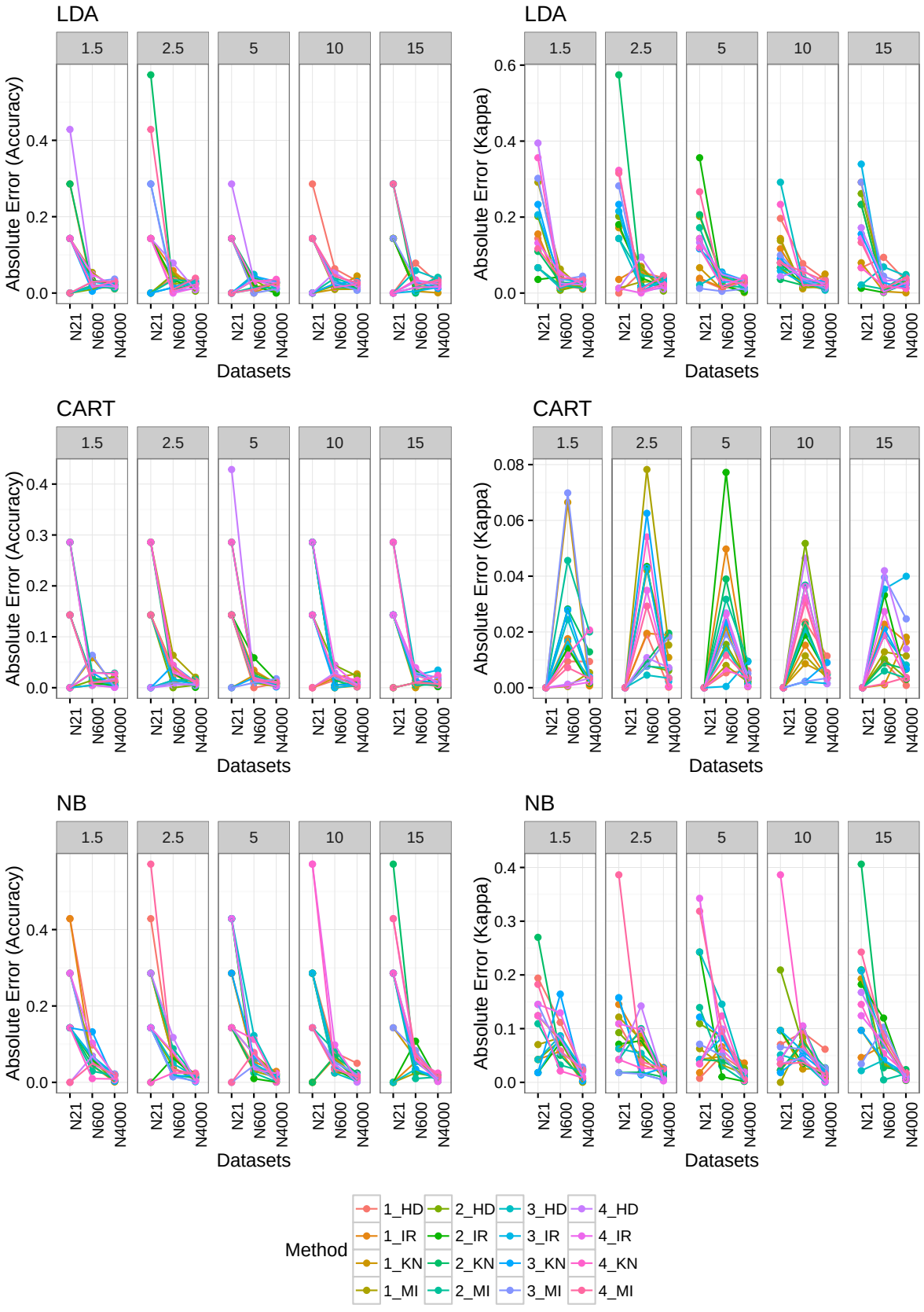


Figure 4.26 – Preprocessing errors of Accuracy and Kappa of the analysis of classification after outlier detection followed by imputation of outlying data on synthetic datasets. Detection methods: Inter Quartile Range (1), Adjusted Quantile (2), PCOUT (3) and LOF (4). Imputation methods: Hot-deck (HD), Iterative Stepwise Regression Imputation (IR), K-Nearest Neighbour (KN) and Multiple Imputation by Chained Equations (MI).

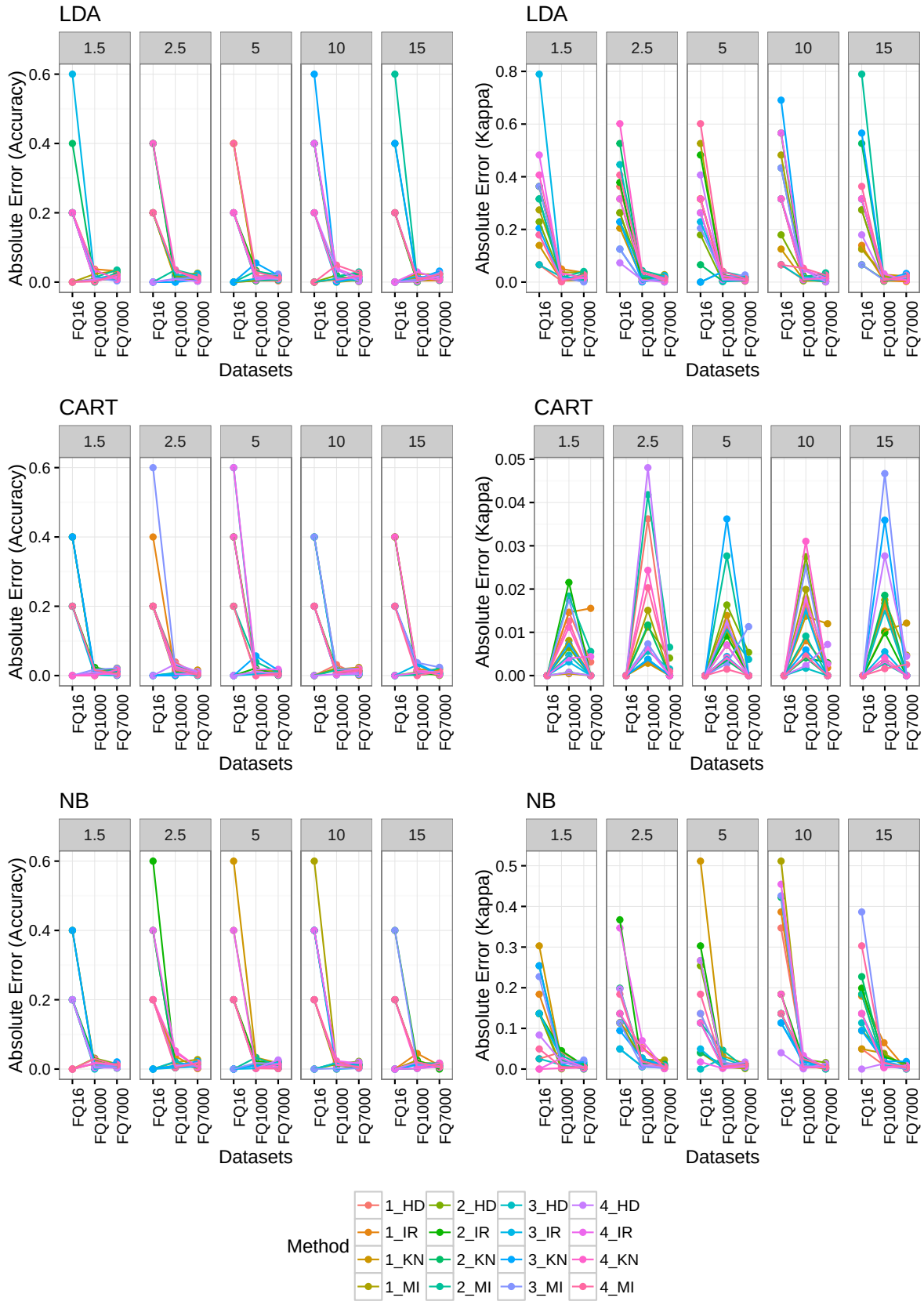


Figure 4.27 – Preprocessing errors of Accuracy and Kappa of the analysis of classification after outlier detection followed by imputation of outlying data on semi-synthetic datasets. Detection methods: Inter Quartile Range (1), Adjusted Quantile (2), PCOUT (3) and LOF (4). Imputation methods: Hot-deck (HD), Iterative Stepwise Regression Imputation (IR), K-Nearest Neighbour (KN) and Multiple Imputation by Chained Equations (MI).

4.6.4.3 Clustering results

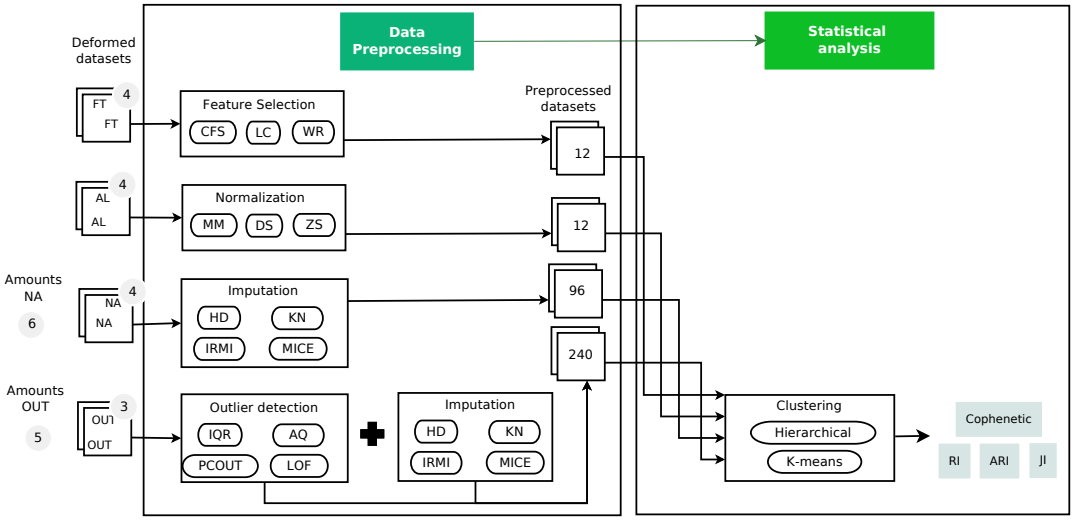


Figure 4.28 – Experimental procedure for assessment of the impact of data preprocessing procedures on results from clustering.

Our experiments report the Rand Index (RI), Adjusted Rand Index (AR), Jaccard Index (JI) for K-means clustering and Cophenetic coefficient for Hierarchical clustering with respect to (c.f. Figure 4.28):

- *For feature selection preprocessing*
  - three feature selection methods; Correlation-based Feature Selection (CFS), Linear Correlation-based Feature Selection (FI) and Wrapper Subset Evaluator (WR).
- *For missing values preprocessing*
  - six amounts of missing values (5%, 10%, 15%, 20%, 25%, and 30%)
  - four imputation methods; Hot-Deck (HD), K-Nearest Neighbour (KN), Iterative Stepwise Regression Imputation (IRMI) and Multiple Imputation by Chained Equations (MICE).
- For outliers preprocessing
  - five amounts of outlying data (1.5%, 2.5%, 5%, 10%, and 15%)
  - four outlier detection methods; (1) Inter Quartile Range (IQR), (2) Adjusted Quantile (AQ), (3) Principal Components decomposition for multivariate outlier detection approach (PCOUT), and (4) Local Outlier Factor (LOF);
  - subsequent outlier imputation by four imputation methods; Hot-Deck (HD), K-Nearest Neighbour (KN), Iterative Stepwise Regression Imputation (IRMI) and Multiple Imputation by Chained Equations (MICE)

Clustering by K-means was performed for all synthetic datasets and for FQ16 and FQ1000 semi-synthetic datasets. Hierarchical Clustering is not robust for datasets which number of observations are larger than 4000 therefore, only the datasets N21, N600, N400, FQ16 and FQ1000 were used for this method.

Results for feature selection

A side by side comparison between different feature selection methods, with respect

to the clustering algorithms is given in Figures 4.29 and 4.30. It shows the Rand index, Adjusted Rand index and Jaccard index over the K-means on the synthetic and semi-synthetic datasets, and the Cophenetic coefficient over Hierarchical clustering on datasets FS21, FS600, FS4000, FQ16 and FQ1000.

We observed that Wrapper Subset evaluator show to have the best results on both clustering methods for datasets FS600, FS4000, and FS20000. While for datasets FQ16 and FQ1000 the Linear Correlation-based feature selection show the best results on both clustering methods. It can be observed that clustering results on the studied datasets after feature selection are different. It was assumed that such differences are due mainly to the characteristics of the datasets (e.g., distribution, size).

Results for missing values preprocessing

A side-by-side comparison between the different imputation methods, respective of the clustering algorithms, is given in Figures 4.31 and 4.32. It shows the RI, AR, JI, and Cophenetic coefficient values over the two clustering methods and the synthetic and semi-synthetic datasets resulting from imputation against different amounts of missing values.

In general, it was observed that the precision of the clustering methods reduced with an increasing amount of missing data. Our results show that the Hot-Deck and MICE imputation methods give the best results on both clustering methods on small datasets (N21 and FQ16) and low missing data rates (5% and 10%). K-Nearest Neighbour imputation method shows, in general, the best results on K-means clustering for the synthetic datasets and the six amounts of missing data. While for $FQ16$ and $FQ1000$ datasets the MICE and IRMI imputation methods show the best results on K-means clustering.
Except for N21 dataset at 5% missing data, MICE and IRMI imputation methods provide the best results on Hierarchical clustering for the three dataset and the six amounts of missing data. For N21 dataset and 5% missing data, Hot-Deck imputation shows the best results. We consider that the observed results depend on the characteristics of the data. We observed that the multivariate imputation methods MICE and IRMI stand up as the best since our datasets are multivariate. On the particular case of N21 and FQ16 datasets, the Hot-Deck method provide the best results, we consider that this is due in part to the type of missing values (Missing at Random) and the characteristics of our datasets.

Results for outlying data preprocessing

A side-by-side comparison between different outlier detection and imputation methods, respective of the regression algorithms is given in Figures 4.33 and 4.34. It shows the Rand index, Adjusted Rand index, Jaccard index, and Cophenetic coefficient over the two clustering methods and the three datasets resulting from outlier detection-imputation against different amounts of outlying data. The three indices and coefficient values are shown for the sixteen combinations of outlier detection and imputation methods.

Results show that, on average, there is no universally best method to detect and impute outliers. The impact of detection-imputation of outliers varies for the two clustering methods and for the synthetic and semi-synthetic datasets. Rand index, Adjusted Rand index, and Jaccard index results show that, Inter Quartile Range (IQR) combined with K-Nearest Neighbour and MICE imputation give the best results on the two clustering methods for N21 dataset at 1.5% of outlying data. While for dataset FQ16, Adjusted Quantile (AQ) combined with IRMI imputation give the best results at 1.5% and 2.5% of outlying data.

For N600 dataset and the five amounts of outlying data, IQR detection method com-

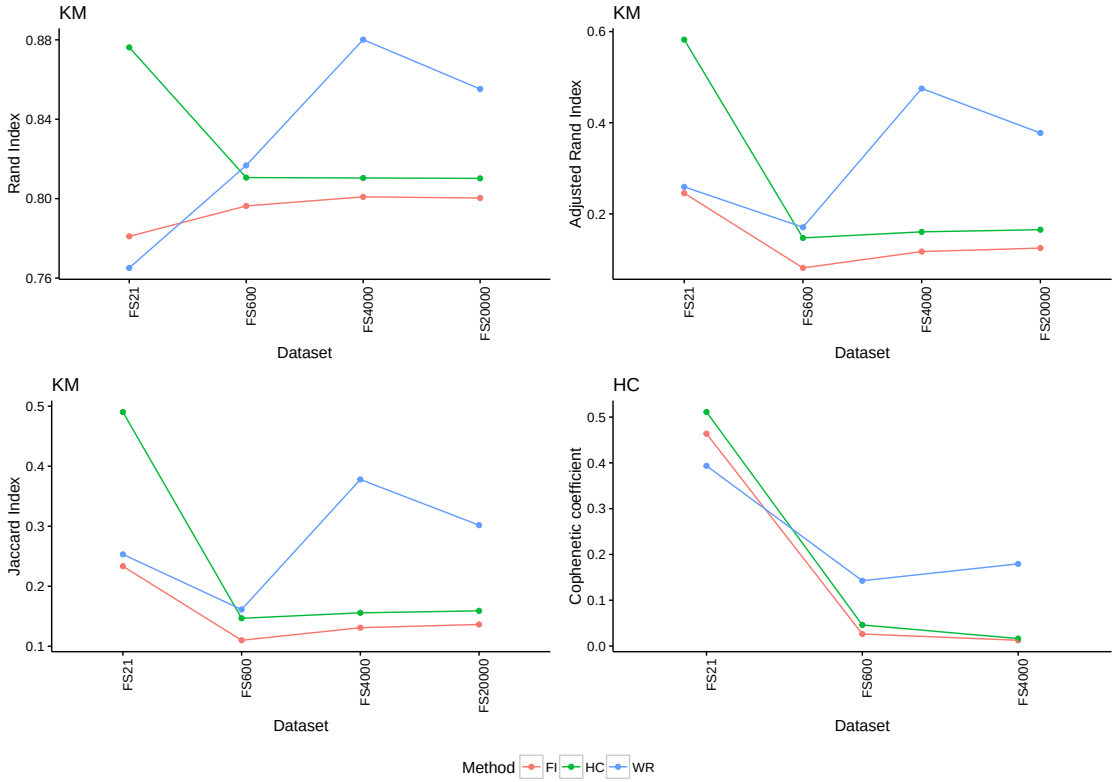


Figure 4.29 – Rand index, Adjusted Rand index, Jaccard index and Cophenetic coefficient of Clustering methods after feature selection processing on synthetic datasets. Feature selection methods: Correlation-based Feature Selection (HC), Linear Correlation-based Feature Selection (FI) and Wrapper Subset Evaluator (WR).

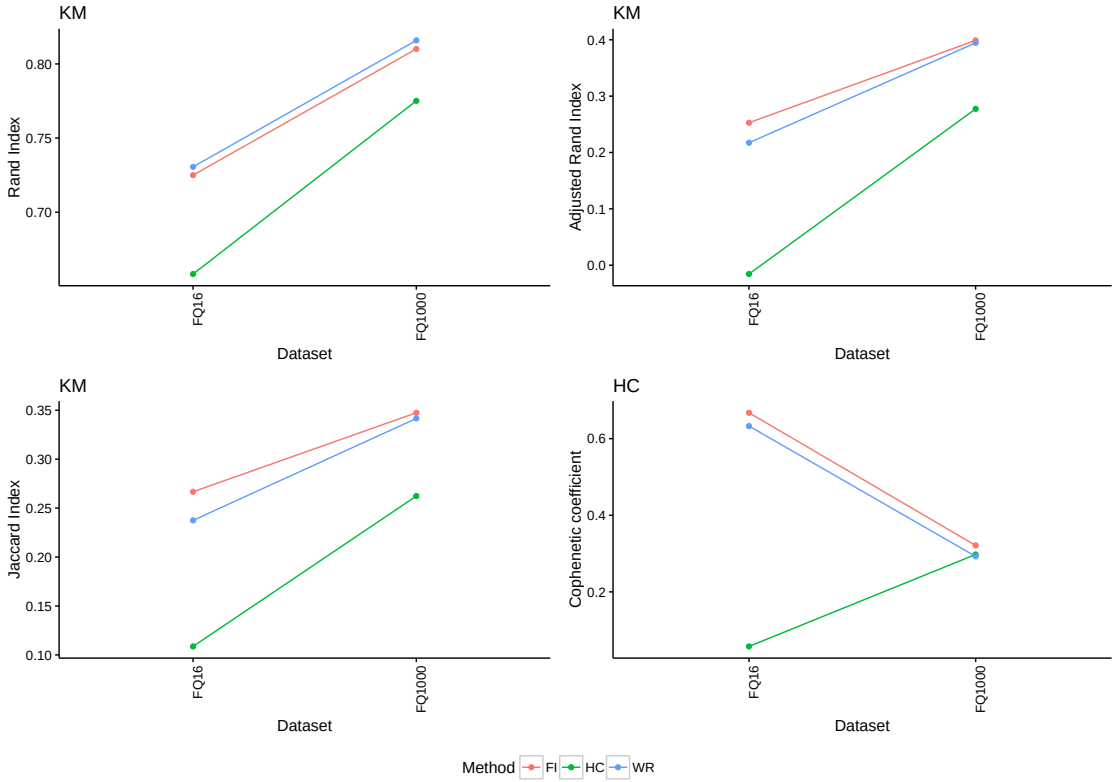


Figure 4.30 – Rand index, Adjusted Rand index, Jaccard index and Cophenetic coefficient of Clustering methods after feature selection processing on semi-synthetic datasets. Feature selection methods: Correlation-based Feature Selection (HC), Linear Correlation-based Feature Selection (FI) and Wrapper Subset Evaluator (WR).

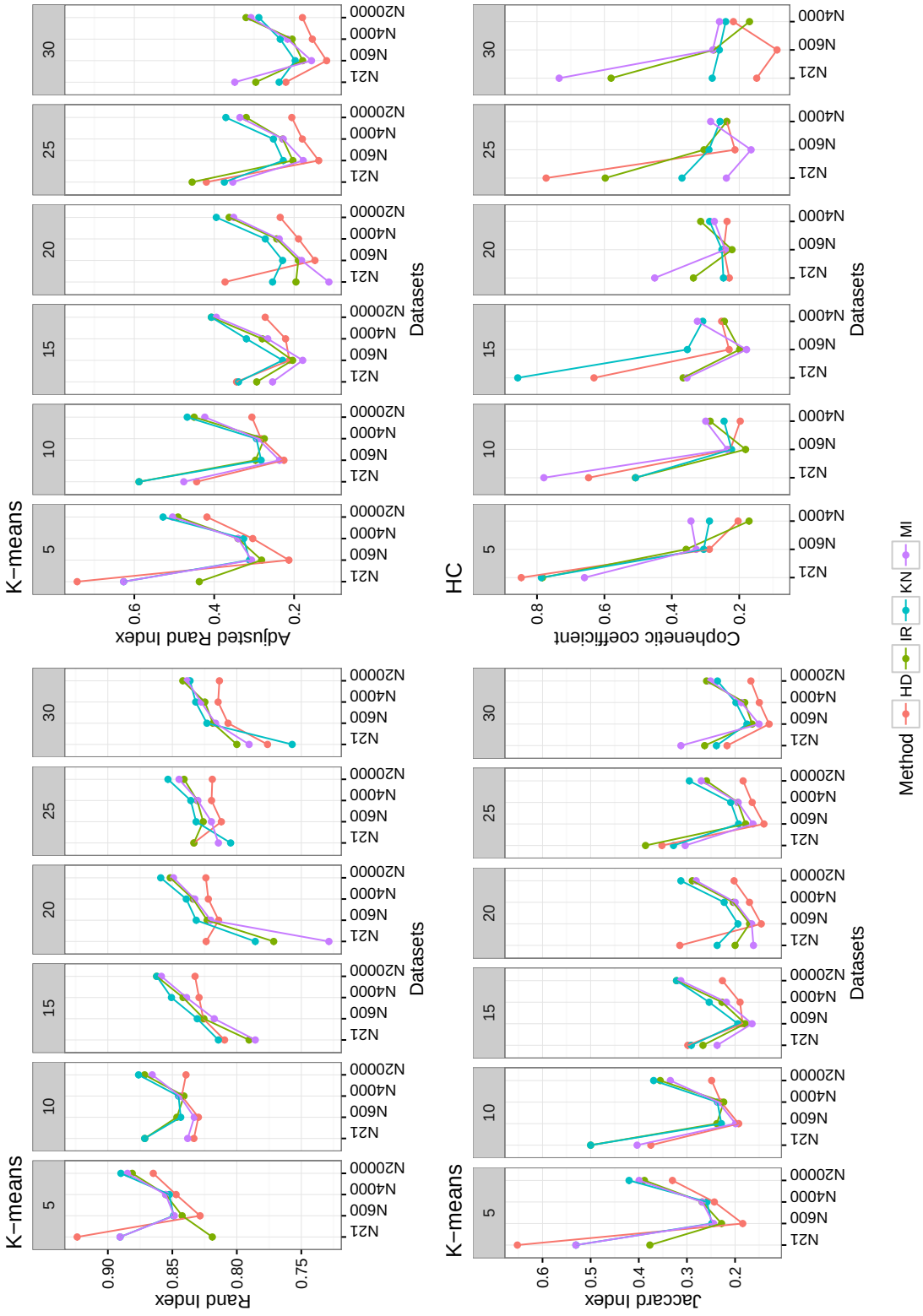


Figure 4.31 – Rand index, Adjusted Rand index, Jaccard index and Cophenetic coefficient of Clustering methods after imputation of missing data on synthetic datasets. Imputation methods: Hot-deck (HD), Iterative Stepwise Regression Imputation (IR), K-Nearest Neighbour (KN) and Multiple Imputation by Chained Equations (MI).

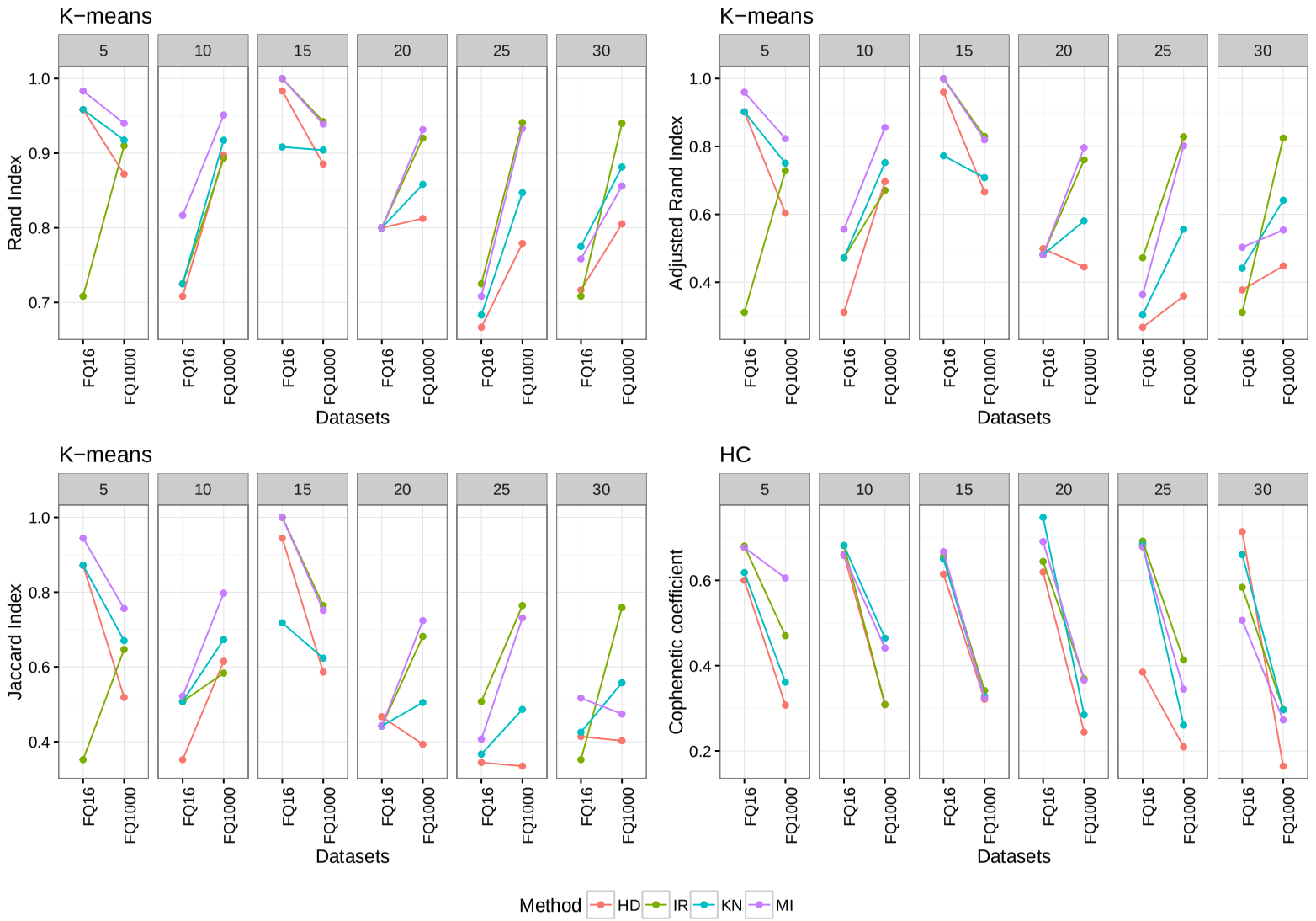


Figure 4.32 – Rand index, Adjusted Rand index, Jaccard index and Cophenetic coefficient of Clustering methods after imputation of missing data on semi-synthetic datasets. Imputation methods: Hot-deck (HD), Iterative Stepwise Regression Imputation (IR), K-Nearest Neighbour (KN) and Multiple Imputation by Chained Equations (MI).

bined with IRMI, MICE and Hot-Deck shows the best results for K-means clustering. While for N4000 and FQ1000 datasets for the five amounts of outlying data, IQR combined with K-Nearest Neighbour and MICE are the best outlier preprocessing procedures for K-means clustering.
Concerning Hierarchical clustering, the IQR detection method combined with K-Nearest Neighbour and Hot-Deck imputation methods show the best Cophenetic coefficient values for N21 and N600 datasets for the five amounts of outlying data. For dataset N4000, IQR combined with Hot-Deck or IRMI imputation methods show to be the best outlying data preprocessing option. The combined methods AQ-MICE and IQR-MICE provide the best results for datasets FQ16 and FQ1000 respectively.

According to our results on Section 4.4.5, the IQR method, provided a very low detection rate, but a high precision. This means that the probability that the detected point using IQR were a real outlier was high. When comparing our results from Section 4.4.5 and 4.5.4.3 we observe that in general the IQR outlier detection method stands as the best when combined with imputation methods such as MICE and IRMI. We consider that these results depend to the characteristics of the datasets and also on the preprocessing methods. It seems that a precise outlier detection method such as the IQR combined with multivariate imputation methods will be a good option to preprocess outliers. However, it is probable that different results may be obtained on dataset with characteristics different from ours. Thus, our results can not be generalized for all type of datasets but are useful to visualize the impact of the preprocessing and statistical methods that we studied on our synthetic and semi-synthetic datasets.

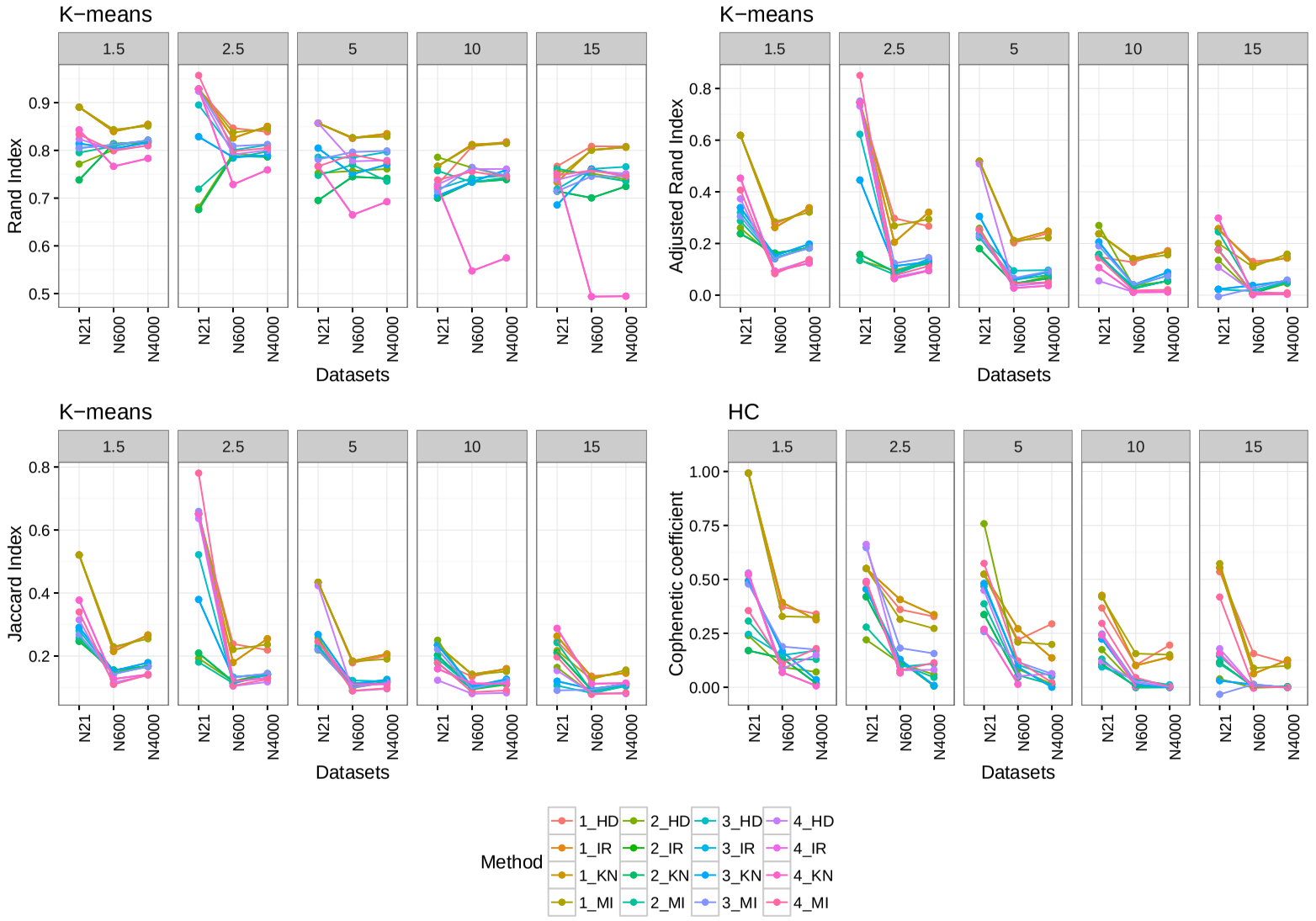


Figure 4.33 – Rand index, Adjusted Rand index, Jaccard index and Cophenetic coefficient of Clustering methods after outlier detection followed by imputation of outlying data on synthetic datasets. Detection methods: Inter Quartile Range (1), Adjusted Quantile (2), PCOUT (3), and LOF (4). Imputation methods: Hot-deck (HD), Iterative Stepwise Regression Imputation (IR), K-Nearest Neighbour (KN) and Multiple Imputation by Chained Equations (MI).

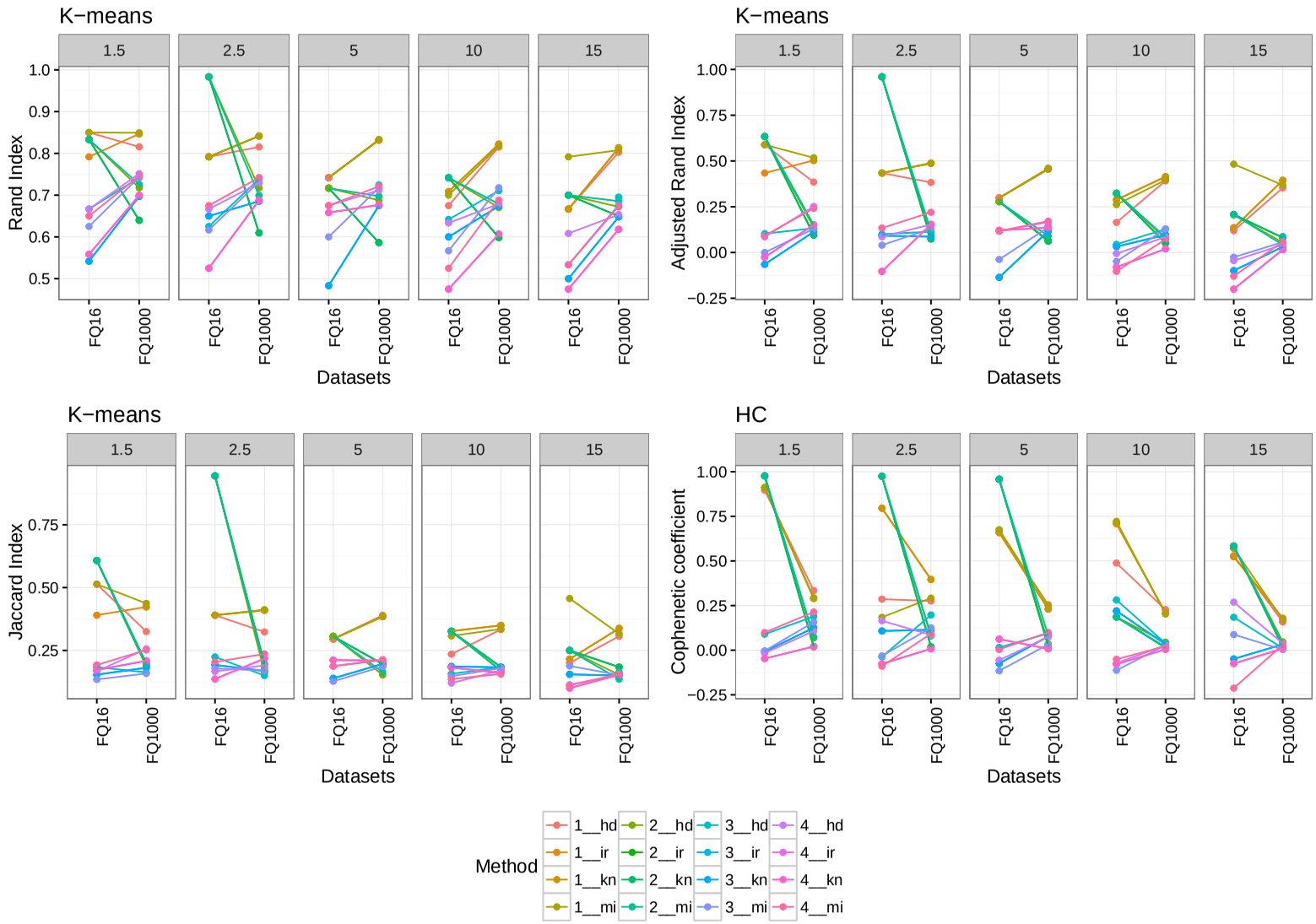


Figure 4.34 – Rand index, Adjusted Rand index, Jaccard index and Cophenetic coefficient of Clustering methods after outlier detection followed by imputation of outlying data on semi-synthetic datasets. Detection methods: Inter Quartile Range (1), Adjusted Quantile (2), PCOUT (3), and LOF (4). Imputation methods: Hot-deck (HD), Iterative Stepwise Regression Imputation (IR), K-Nearest Neighbour (KN) and Multiple Imputation by Chained Equations (MI).

4.7 Summary and concluding remarks

A comprehensive study to assess the impact of preprocessing procedures on accuracy of three statistical methods (regression, classification and clustering) was presented. Our methodological approach allowed us to identify the best data preprocessing strategies for each statistical method for our synthetic and semi-synthetic datasets. We discovered that there is not a universal data preprocessing procedure since it depends on the characteristics of the dataset in terms of size, distribution, Skewness, Kurtosis, etc.

Results obtained from this study were collected and summarized in a matrix. It contains the main characteristics of our datasets and the preprocessing procedures that provides the best results for each statistical method. With the help of such matrix we could construct a set of rules for combining optimally data preprocessing and data analysis. This rules were subsequently used on the construction of a fully integrated analytics environment in R for statistical analysis of environmental data in general, and for water quality data analytics, in particular (c.f. Chapter 5). Table 4.6 gives an extract of our summary table. For simplicity purposes, only results for N21 dataset on one regression, one classification and one clustering methods are shown. The set of rules that we obtained can be used as the basis to extract knowledge related to data preprocessing and most importantly, it can be completed and extended to study: other type of datasets, data anomalies and data preprocessing procedures.

It should be noticed that our study was performed using datasets with a multivariate normal distribution and on four procedures to prepare data (feature selection and normalization) and process anomalies (missing values and outliers). Further studies will be necessary to cover a wider spectrum of datasets including temporal or spatial series, different distribution along with other types of data preprocessing and data mining techniques. Moreover, the approach that we presented was performed studying each data anomaly as isolated cases. For instance, datasets including missing values have been preprocessed only using imputation methods. However, multiple anomalies may be present in a real-world dataset thus requiring the application of divers preprocessing strategies. An important issue when multiple anomalies are present in a dataset is the order in which preprocessing strategies need to be applied. Below we provide an example to illustrate the importance of data preprocessing ordering.

Non-normalized datasets with missing values were preprocessed by different order. The synthetic datasets have different length (ALNA21; $n = 21$ $p = 8$; ALNA600; $n = 600$ $p = 30$) with a Weibul distribution and different amounts of missing at random values (i.e., 5%, 10%, 15%, 20%, 25% and 30%). These datasets were preprocessed by (1) normalization followed by imputation of missing values and, (2) imputation of missing values followed by normalization. Min-max normalization was used. Imputation of missing values were done using the IRMI, MICE, Hot-deck and k-NN methods. To compare the results of preprocessing strategies a classification by Naïves Bayes was performed following the same methodology described in Section 4.6.2.

Results of classification by Naïves Bayes on the different ordering of preprocessing strategies is shown in Figure 4.35. It was observed that the differences between the two preprocessing ordering are small but, in general, lower error rates are obtained when normalization is performed after imputation of missing values. Such differences may be due to divers aspects including: distribution of data before and after the the application of the first preprocessing procedure, general assumption of preprocessing and statistical procedures, etc. Therefore an extended study in this respect is necessary in order to have a better understanding of preprocessing ordering.

Though ordering of preprocessing strategies is out of the scope of this thesis work, with this example we want to highlight the importance of conducting studies in this topic.

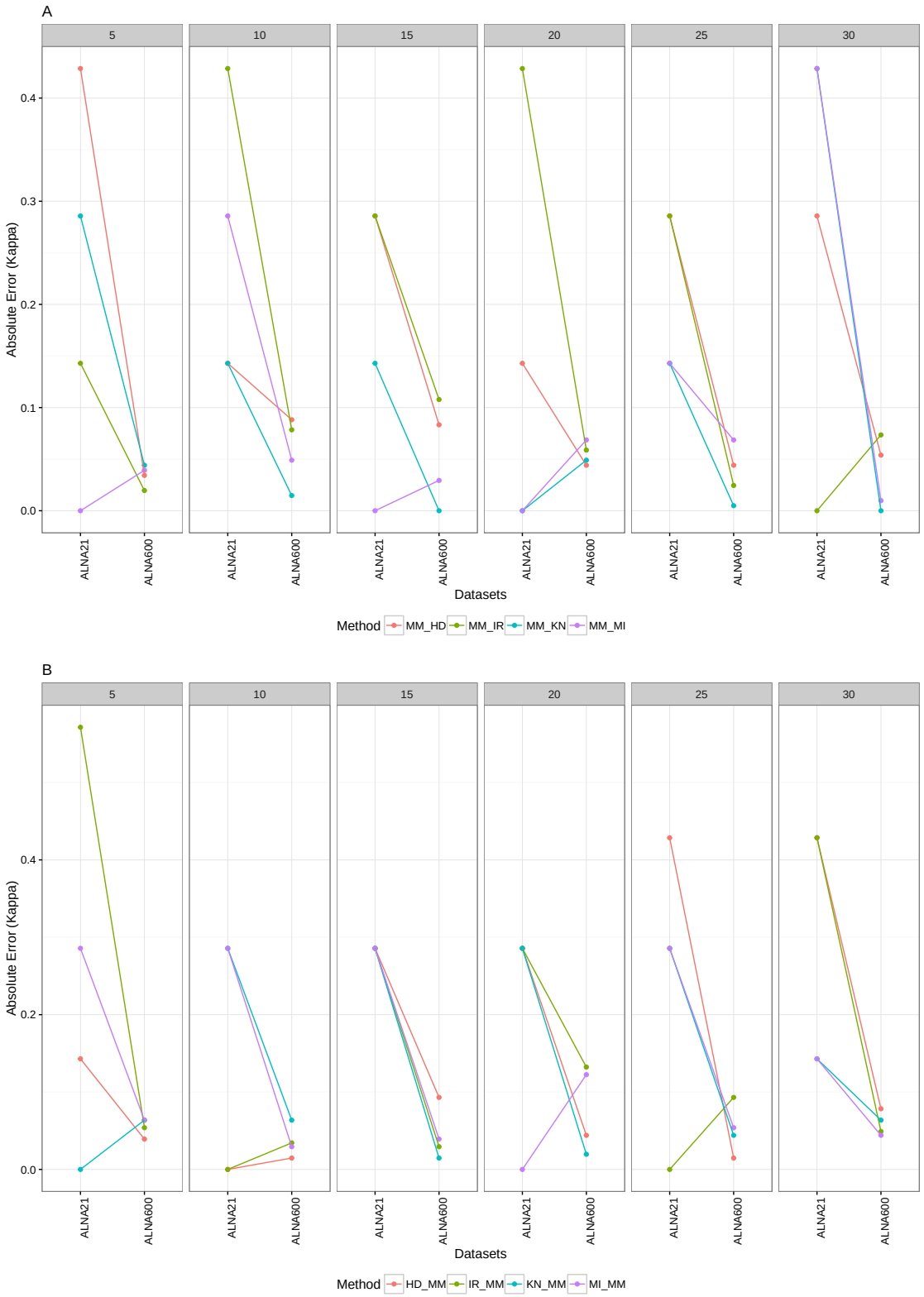


Figure 4.35 – Results of Naïves Bayesian classification on datasets preprocessed by: (A) normalization followed by imputation of missing values and (B) imputation of missing values followed by normalization. Min-max normalization (MM).Imputation methods: Hot-deck (HD), IRMI (IR), K-NN (KN) and MICE (MI).

Table 4.6 – Extract of summary results from the study of data preprocessing procedures on regression, classification and clustering methods. FS (Feature Selection), N (Normalization), I (Imputation method), O (Outlier processing).

				Regression				Classification				Clustering			
				LASSO				Naives Bayes				K-means			
Num Observations	Num Numerical variables	Num. Categorical variables	Distribution	FS	N	I	O	FS	N	I	O	FS	N	I	O
21	8	1	Normal	fi	ds	kn (5%)	PCOUT-MICE	wr	ds	kn (5%)	PCOUT-MICE	fi	-	hd	IQR-MICE
21	8	1	Normal	fi	ds	kn (10%)	LOF-MICE	wr	zs	mi (10%)	LOF-HD	fi	-	hd	PCOUT-KN
21	8	1	Normal	fi	ds	mi (15%)	IQR-MICE	wr	zs	kn (15%)	LOF-MICE	fi	-	mi	PCOUT-MICE

CHAPTER 5

Development and experimental results

Contents

5.1 Introduction 120
5.2 Related work on scientific workflow systems 120
5.3 EvDA: Development of R Shiny application for environmental data preprocessing and analysis 121
5.3.1 Main workflow 121
5.3.2 Overview of the application 123
5.4 Case study: Water pollution of the Tula, Culiacan, Tamazula, and Humaya rivers 128
5.4.1 Data description 128
5.4.2 Data preprocessing 129
5.4.3 Results of the analysis 131
5.5 Conclusion and future development 137

5.1 Introduction

In the previous chapter, we have presented our methodology to collect water quality data as well as our approach to assess data preprocessing quality.
In order to provide to the scientific community the tools to better preprocess and analyse environmental data, we have developed a fully integrated environment in R that integrates our data preprocessing approach.

The dataset collected for assessing water quality in Mexican rivers was used as input to the tool that we describe in this chapter. In this chapter, we present our results in our application domain of Water Quality Assessment.

This chapter is organized as follows: in Section 5.2 we present related work on scientific workflow systems as our tool fits in the same line. Then, in Section 5.3 we present the main functionalities of our prototype. In Section 5.4, we present the results of the analysis of water quality data of the Mexican rivers by using our prototype. Finally, we provide some conclusions of this chapter in Section 5.5

5.2 Related work on scientific workflow systems

Advances in scientific computation allow scientist to conduct their analysis by using workflow systems. Scientific workflow systems (SWFS) provide an infrastructure to perform and monitor a set of data manipulation steps in a goal-oriented scientific application.

A scientific workflow system is composed of a set of tasks linked together to create a final product in therms of derivative/aggregate data, or hypothesis validation. In computing sciences, workflows are derived from programming models that allow the integration of software routines, datasets and services for the scientific discovery process. SWFS are generally applied to: (1) describe complex scientific procedures, (2) automate data derivation processes, (3) improve through put and performance on high performance computing and (4) manage and query of provenance (Zhao et al., 2008).

There are several software environments that provide tools to define, compose, map, and execute workflows. They have been grouped into two classes: script-like systems and graphical-based systems (Neubauer et al., 2006).

Script-like systems describe workflows using a textual programming languages such as Java, Perl or C^{++}. Tasks, parameters, constraints and data dependencies are established in order to build up a workflow. Examples of these type of systems are GridAnt (von Laszewski and Hategan, 2005), Karajan (Von Laszewski et al., 2007) or Askalon (Fahringer et al., 2007) to mention some. Some of their advantages include:

- Configuration of specialized process model;
- Control of sequences, constraints, loops and parallel tasks;
- Parallel execution of workflows;
- Monitoring of the execution.

Though, Script-like systems often have complex semantics, require a deep knowledge of the workflow engine functionality and knowledge on programming languages.

Graphical-based systems specify workflows through graphical elements (i.e., nodes and edges) that correspond to the graph component. Workflow tasks are often represented by nodes while data dependencies or communications between tasks are represented by links. Some of the most popular graph-based systems are: Kepler (Altintas et al., 2004),

Weka4WS (Talia et al., 2005) and Triana (Taylor et al., 2003). Their advantages include:

- Easy and intuitive use for unskilled user;
- Model of workflows through the use of graphical interface;
- Limited use of programming languages;

One important drawback of graphical-based systems, is that complex workflows can not be represented due to the limited expressiveness on the directed acyclic graph-based languages.

Scientific workflow systems aim at supporting research by providing a framework that encompass all the necessary tasks and algorithms to access, collect, process and analyse scientific data. Numerous workflow systems are available (Talia, 2013; Liu et al., 2015), often require user to have a deep knowledge of the workflow engine functionality or ability on programming languages. They also require a comprehensive knowledge of the characteristics of data, thus the designed scientific workflow can include the necessary data preprocessing tasks to manage anomalies on data. Notwithstanding, as we have mentioned in our previous chapters, the impact of data preprocessing task on subsequent data analysis is under study and as such, the integration of data preprocessing procedures on scientific workflow systems may be taken cautiously. Additional to this problem, it must be noticed that environmental scientist are not computer experts, they may not be able nor want to cope to the complexity of the use and deployment of these systems for their scientific purposes. To our knowledge, there are not a scientific workflow system specifically for environmental sciences that includes data preprocessing procedures and that provides a user-friendly interface for non computer experts.
We aim to cover this need by providing a simple scientific workflow system. Our system integers the necessary tasks to inspect, preprocess and analyse environmental data. Data preprocessing procedures can be assessed and if necessary be modified in order to get valid results according to user criteria. It also take advantages of an user-friendly interface that allows users with little knowledge on R to analyse environmental data.

5.3 EvDA: Development of R Shiny application for environmental data preprocessing and analysis

5.3.1 Main workflow

Data collected by environmental scientists cover highly diverse topics of study going from public health studies to atmospheric phenomena. Moreover, the laborious tasks of collecting environmental data increase the probability of having anomalies in the data. As a result, the analysis of environmental data has itself become an arduous effort. The methodological approaches developed in this work aim at providing tools to support and assist the scientific community in the preprocessing and analysis of data. We have focused on procedures for preprocessing the environmental data specifically features selection, normalization, imputation of missing values, and outlying data as well as regression, classification, and clustering methods since they are frequently used to analyse environmental data.

We present a prototype named EvDA that aims to inspect, preprocess, and analyse environmental data easier. The user can upload his/her own data. Preprocessing procedures can be executed sequentially and data analysis such as regression, classification, clustering or PCA analysis produce meaningful results for the domain experts using an user-friendly interface.

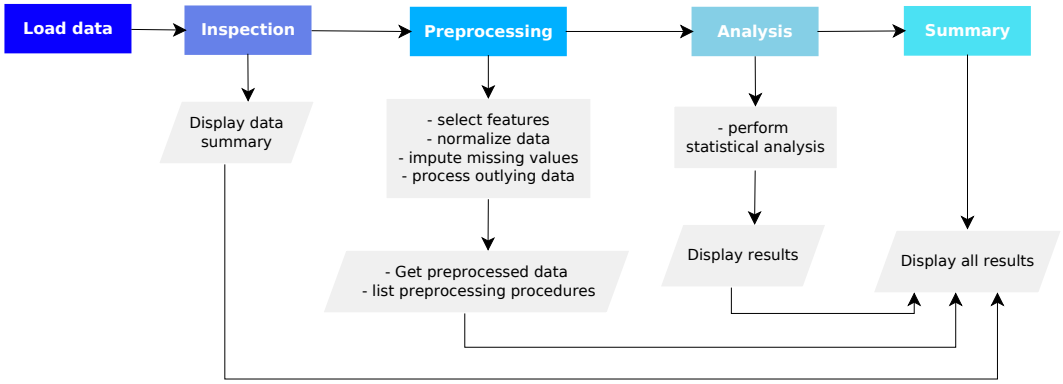


Figure 5.1 – Scientific workflow of EvDA

Implementation To this aim, we have designed and developed a scientific workflow system prototype in EvDA, based on our methodological approach described in the Section 4.2. It was constructed to execute a set of data manipulation tasks for environmental data analysis in general and for water quality data analytics in particular. Our prototype consists of graphical user interface from which user is able to load, inspect, preprocess and analyse data (c.f. Figure 5.1). EvDA was implemented using Shiny web application framework (R package version 3.3.0) for R statistics software (R Core Team, 2016). It uses several R packages and functions internally. In EvDA, the input is an original dataset. After inspection, a set of preprocessing tasks are executed to perform statistical analysis on the preprocessed data. Finally, the outputs include visualization and analytical results of the preprocessed data.

Data inspection. EvDA provides results of summary statistics such as mean, median, minimum and maximum values, as well as Skewness and Kurtosis coefficients to characterize the data distribution. Information about normality, missing values, and outlying data is also provided. Results are displayed in a table and can be graphically visualized. In the data inspection step, the user can have a general overview of the characteristics of data. This information is helpful to make decisions about data preprocessing procedures that may be necessary to execute.

Data preprocessing. Data preprocessing procedures such as: feature selection, normalization, imputation of missing data, and outlying data preprocessing can be performed. The most optimal data preprocessing procedures are automatically determined in EvDA: a message listing optimal data preprocessing procedures are displayed. The user could choose between the suggested options and the procedures available in the application.
Three feature selection methods are available: wrapper subset evaluator, Correlation-based and linear correlation-based feature selectors, they are calculated using the `caret` and the `mlr` R packages. Feature selection outputs are: (1) a list containing pertinent features and (2) a new dataset containing only the selected features.
Six normalization methods can be performed: decimal scale, min-max, z-score, 0-1 range normalization, sigmoidal and softmax normalization; they were adapted from the `dprep` R package.
Five methods to impute missing values are proposed in EvDA including: mean, Hot-Deck, K-Nearest Neighbour (K-NN), Iterative Robust Model-based Imputation (IRMI), and Multiple Imputation by Chained Equations (MICE). Imputation methods are computed using their respective packages, namely `ForImp` for mean, `VIM` for Hot-Deck, K-NN and IRMI and, `mice` for MICE.

Outlying data are processed as follows: first outliers are detected using one of the outlier detection methods, then they could be removed from the dataset or imputed using imputation methods. The outlier detection methods available in EvDA are: inter quartile range (IQR), adjusted-quantile (AQ), Principal Components decomposition (PCOUT) and Local Outlier Factor (LOF). They are computed using the `mvoutlier` and `Rlof` R packages. And the imputation methods are the Hot-Deck, K-Nearest Neighbour (K-NN), the Iterative Robust Model-based Imputation (IRMI), and the Multiple Imputation by Chained Equations (MICE).

Data analysis. Regression, classification, clustering, and PCA analysis can be performed in EvDA. Analytical methods are computed using various R packages.

Nineteen regression methods are available including: four linear regression (ordinary least square regression, stepwise linear regression, principal component regression and partial least squares regression), three penalized regression (Ridge regression, least absolute shrinkage and selection operator and, elastic net), four non-linear regression (multivariate adaptive regression splines, support vector machine, k-Nearest Neighbour and neural network) and eight decision trees (Classification and Regression Trees (CART), conditional decision trees, model trees, rule system, bagging CART, random forest, gradient boosted machine and cubist).

Sixteen classification methods can be computed they are: two linear classifiers (logistic regression and linear discriminant analysis), eight non linear classifiers (Mixture discriminant analysis, quadratic discriminant analysis, regularized discriminant analysis, neural network classification, flexible discriminant analysis, support vector machine classification, K-Nearest Neighbour classification and Naïves Bayes) and six non-linear classifies with decision trees (Classification and regression trees, bagging CART classification, Random Forest Classification, C 4.5, C 5.0, and PART).

The hierarchical clustering, K-means, correlation matrix and PCA can also be computed in EvDA. For clustering methods, information such as number of clusters (i.e., K-means) or the height of the cut to the dendrogram for hierarchical clustering need to be specified by user. Principal Component Analysis (PCA) and correlation matrix are performed using the `FactoMineR` and `Hmisc` R packages respectively.

5.3.2 Overview of the application

EvDA has a user-friendly interface. It consists of four main steps including: loading, inspection, preprocessing, and analysis of data. The user can click on different tabs to move between each step. Parameters are inserted through the use of widgets such as drop-down menus, sliders, check-box, etc. At each step, the user can define the settings and the results are calculated and shown automatically.

- *Data input.* The user's data can be uploaded by uploading a file in .csv, .txt or .xls format. Delimiters are detected automatically from the data (c.f. Figure 5.2) but, it is also possible to specify it manually. Sometimes the user may be interested in studying a subset of the dataset. In this case the user can select the desired rows and columns manually and use the subset as a new dataset. EvDA uses only numerical and categorical data since the methods available in EvDA are implemented mainly on these type of data.
- *Data inspection.* Inspection of data is done through: descriptive statistics, tests of normality, analysis of missing data, and detection of outlying data. The inspection tasks are accessible by selecting the corresponding tab. The users can choose which

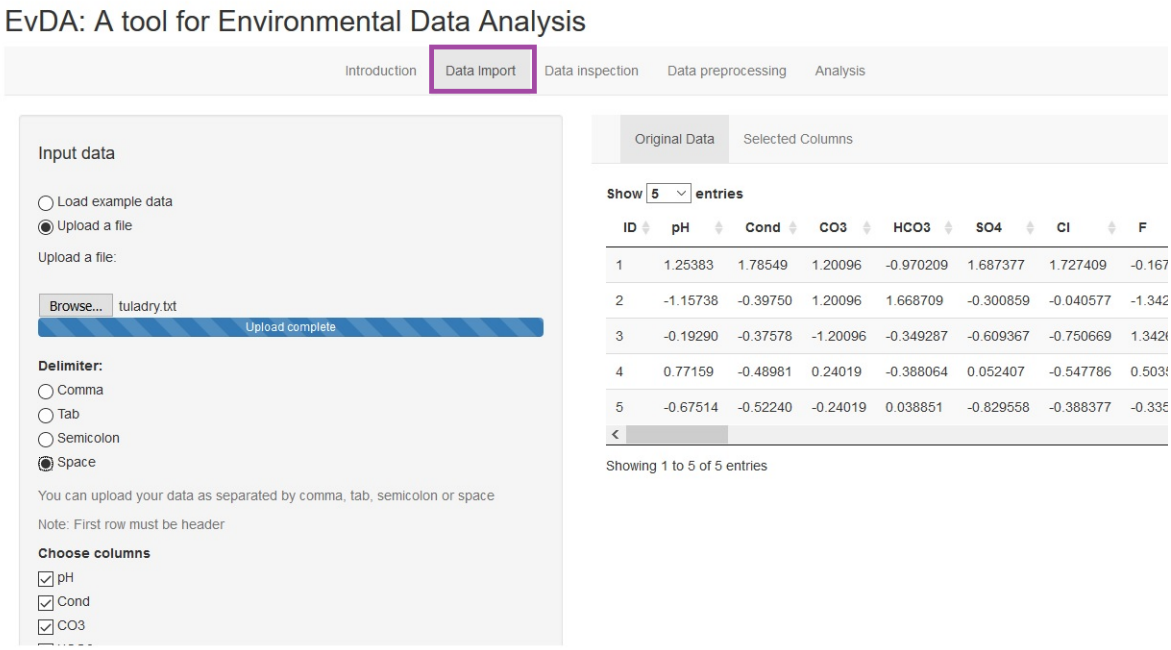
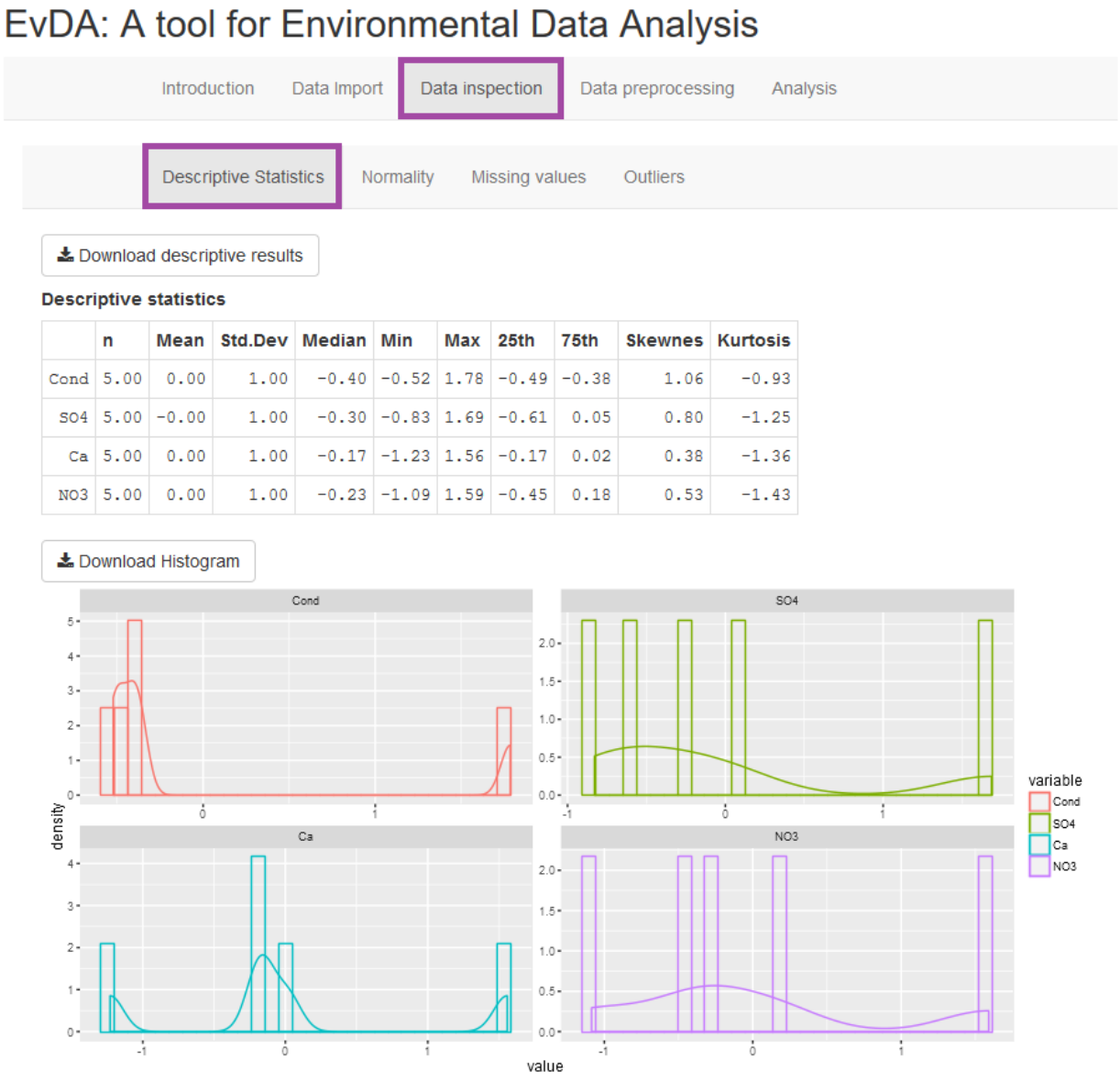


Figure 5.2 – Screenshot of EvDA application. Display of the Data input tab.

normality test and outlier detection method to use. Results are displayed in tables and graphs (see Figure 5.3) and can be downloaded by users for future use.

- *Data preprocessing.* Many real-world data may contain different type of anomalies. To improve the quality of data and perform statistical analysis, data need to be preprocessed. EvDA provides a set of preprocessing tasks including: feature selection, normalization, imputation of missing values, and outliers processing. In order to guide the user, a panel indicating the most optimal data preprocessing procedures is shown. The user can choose one among the different methods that are available for each preprocessing task. They can be performed whether in the original input dataset or in a dataset resulted from a previous preprocessing task. An example of the information displayed on the data preprocessing tab is given in Figure 5.4. It shows the different methods to impute missing values on a given dataset. Preprocessed data can be used for subsequent statistical analysis or downloaded for future use.
- *Analysis.* Our prototype implements some of the functionalities that we considered the most frequently used on environmental studies.
  For regression analysis, independent and dependent variables can be specified by the user. The models are constructed by splitting the observations into a training and test sets. Percentage of splitting can be manually defined. Accuracy of the model is then computed and the results can be saved.
  Similarly to regression, classification analysis is performed using splitted data. Prediction can be made using the resulting model and the user can obtain accuracy and Kappa measure results.
  The construction of clustering by K-means requires to specify the number of cluster k. For easy selection of k, EvDA performs the elbow method. A graphic representing the number of clusters against the total within-clusters sum of squares is displayed. With the help of this visualization, the user can select the number of cluster k. Hierarchical clustering does not require to pre-specify the number of clusters. In our prototype, Hierarchical clustering is performed and the results are represented in a



	n	Mean	Std.Dev	Median	Min	Max	25th	75th	Skewnes	Kurtosis
Cond	5.00	0.00	1.00	-0.40	-0.52	1.78	-0.49	-0.38	1.06	-0.93
SO4	5.00	-0.00	1.00	-0.30	-0.83	1.69	-0.61	0.05	0.80	-1.25
Ca	5.00	0.00	1.00	-0.17	-1.23	1.56	-0.17	0.02	0.38	-1.36
NO3	5.00	0.00	1.00	-0.23	-1.09	1.59	-0.45	0.18	0.53	-1.43

Figure 5.3 – Screenshot of EvDA application. Display of the Data inspection tab.

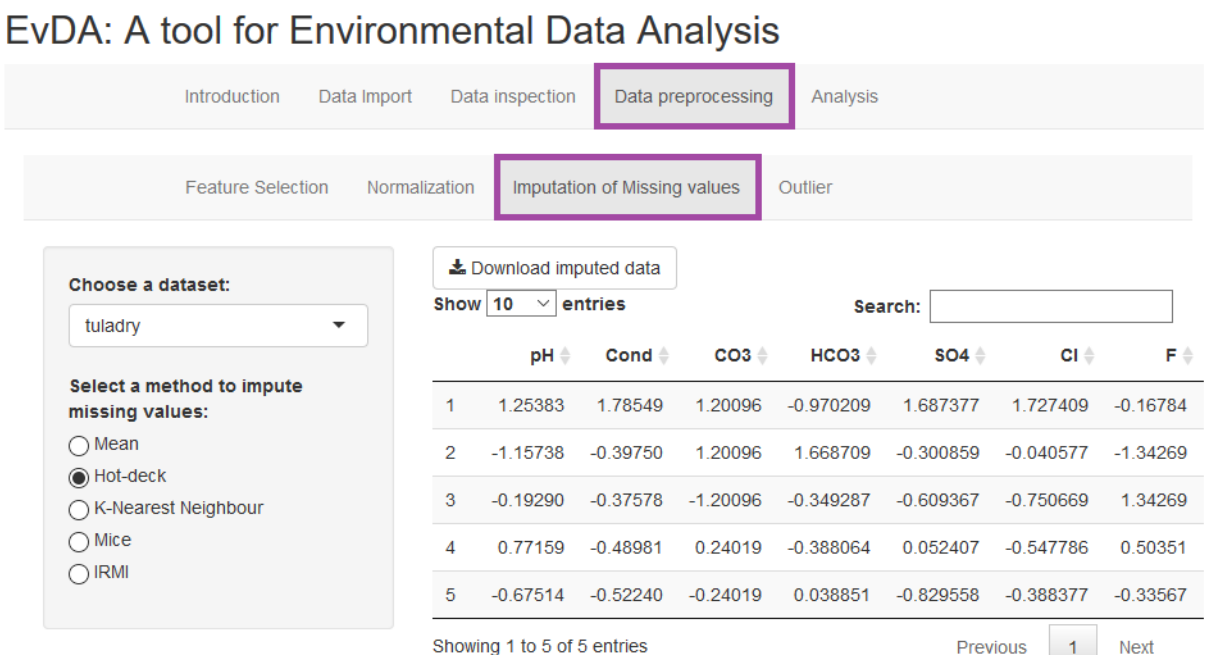


Figure 5.4 – Screenshot of EvDA application. Display of the Data Preprocessing tab.

dendrogram which is a tree-based representation of the observations.
Correlation matrix is performed using all variables, the output is composed of a correlation matrix, a list of the most correlated variables and a graphical display of the correlation matrix.
PCA is constructed only using numerical data. PCA output consists of a table that summarizes PCA results and, PCA plots of the observations and variables. PCA visualization is a scatterplot of the principal components corresponding to axes 1 and 2 and PCA plot of variables shows the projection of the variables within the first two principal components. Optionally, the user can add supplementary variables, they can be whether continuous or categorical. These variables can be used to get additional information about the variability. In the resulting plots, they are coloured differently for better visualization. In Figure 5.5, we show an example of the information displayed when the K-means clustering analysis is performed.

In the next section we present results obtained from the application of EvDA. In the following section, we aimed at demonstrating the utility of our prototype and providing useful information to specialists regarding the water pollution of the four Mexican rivers.

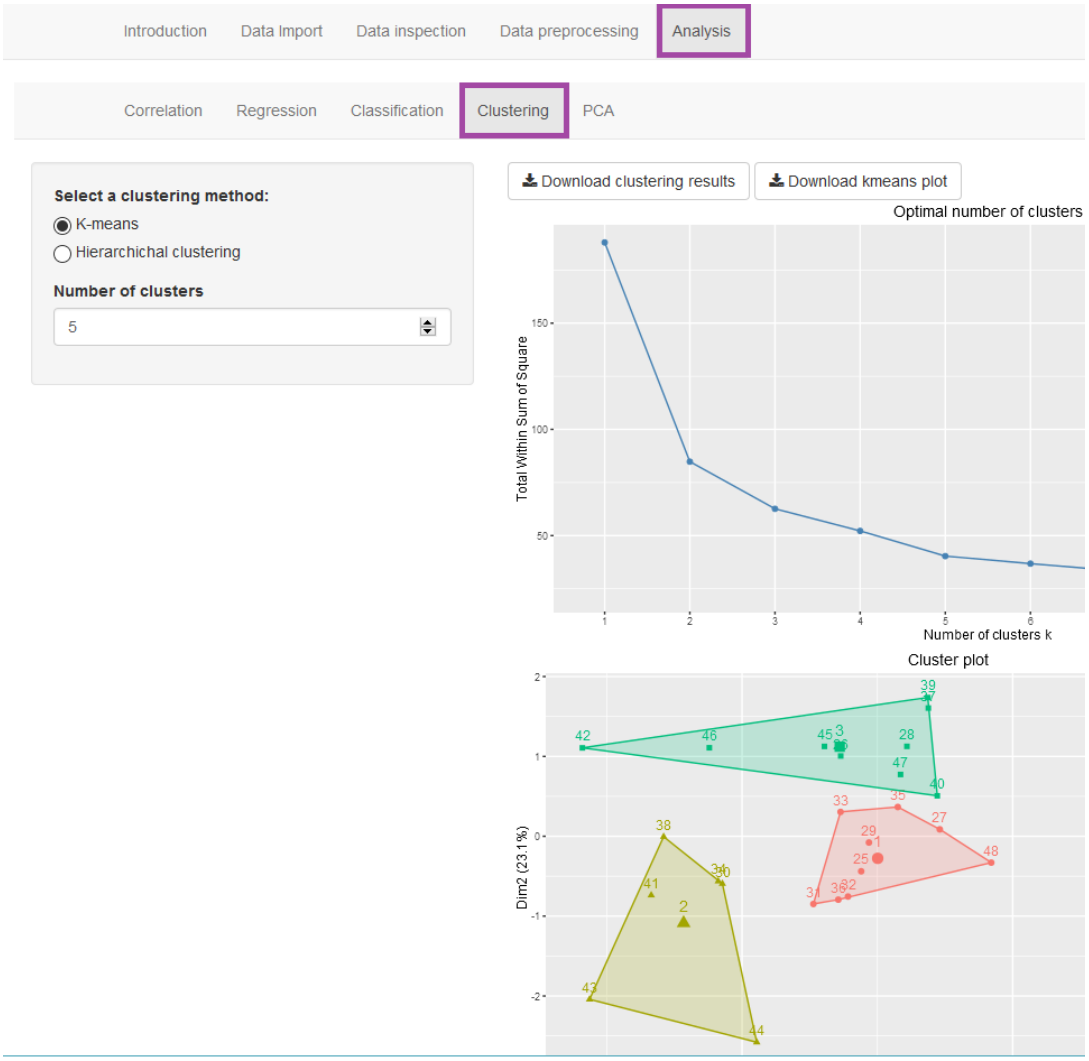


Figure 5.5 – Screenshot of EvDA application. Example of display of the analysis by K-means clustering.

5.4 Case study: Water pollution of the Tula, Culiacan, Tamazula, and Humaya rivers

By applying our methodological approaches, we aimed now at responding to a concrete applicative need in the context of water quality assessment. Precisely, we aimed at providing answers to the following specialist-driven questions:

1. Considering the different sources of pollution from PPCPs and pesticides, are they found in the same place or are they found with the same concentrations in the different sampling sites?
2. What is the correlation between heavy metals, pesticides, PPCPs, and biomonitoring metrics?
3. What are the most representative or pertinent pollutants to focus on in order to assess the impact of specific anthropological activities (i.e., agricultural activities)?
4. Is it appropriate to analyse all the pollutants if the conclusions are similar for one type of pollutant?

5.4.1 Data description

5.4.1.1 Physico-chemical and chemical data

To assess the quality of water of the Tula, Humaya, Tamazula, and Culiacan Mexican rivers, the methodological approach we designed to collect water quality data was described in Chapter 3. By applying our methodological approach, we have collected data that describe the physico-chemical, chemical and biological characteristics of the rivers. The physico-chemistry includes a set of physical and chemical parameters that characterize the river. Such parameters could be necessary either for aquatic biodiversity or for a pollution source. We have organized the different physico-chemical and chemical parameters into two groups: macro-pollutants and micro-pollutants.

- **Macro-pollutants.** They include a set of compounds that are naturally present in the river, which are necessary for good functioning of the aquatic ecosystem. We can mention for instance: organic carbon, nitrate, phosphates, sulphate or metals which are necessary for the good functioning of aquatic ecosystems. However, when their concentrations exceed normal values, pollution problems may emerge such as eutrophication[1] or anoxia[2].
- **Micro-pollutants.** They are compounds that are not naturally present in the river; they include oil, hydrocarbons, organic solvents, surfactants, pesticides, pharmaceuticals, and personal care products. They could be introduced due to filtration, runoff or direct discharge to the watercourse. A tiny dose of these compounds will poison the aquatic environment.
  We have analysed water samples to determine the content of organochlorine pesticides and PPCPs. The analytical methods that we adapted for the analysis of pesticides and PPCPs are under submission.

[1] Process in which a body of water acquires high concentrations of nutrients, specifically phosphates and nitrates. It also causes a massive development of plants.

[2] Lack of oxygen in a body of water produced by oxidation of high concentrations of organic matter.

5.4.1.2 Biological data

It concerns the biological organisms (flora and fauna) living in rivers. In hydrobiology the biological organisms are grouped into: macrophytes, diatoms, fishes, macroinvertebrates, and oligochaetes. Macrophytes and diatoms are representative of the aquatic flora while fishes, macroinvertebrates, and oligochaetes are representative of the aquatic fauna. We have chosen macroinvertebrates in order to compute biological indices to assess the quality of the aquatic ecosystems of the Mexican rivers.
Macroinvertebrates are biological organisms living at the bottom of rivers and lakes. Their size is superior to 0.5mm and they are visible without the help of special lens. Macroinvertebrates include different species: for example; molluscs, crustaceans or larvae of insects. Macroinvertebrates are inventoried and the resulting list of taxa has been then used to compute different indices that provide information about the diversity, fauna richness, and quality characteristics of the aquatic ecosystem.
In our study case, we have collected and analysed macroinvertebrates from each sampling site. A total of 35 indices were computed as suggested by Serrano Balderas et al. (2016).

Our dataset is composed of 16 observations and a total of 78 numerical variables that include 20 macro-pollutants, 23 micro-pollutants, and 35 biological indices. Collection of samples was done in two seasons (rainy and dry). For the sake of simplicity, we present only the results of the dry season. In Table 5.1, we list the different parameters that compose our dataset. Below we will refer to this dataset as the *Mexican dataset*.

5.4.2 Data preprocessing

Data collected from the analysis of samples can not be used in its original form. We need to perform a set of preprocessing tasks in order to initialize data properly to serve as input for subsequent statistical analysis. The preprocessing task that were implemented are (1) feature selection, (2) normalization, (3) imputation of missing values, and (4) outlying data processing. The preprocessing procedures that we applied were those proposed by EvDa.

1. **Normalization.** Our dataset contains variables with different measurement units; to scale all numerical variables to an specific range, we have performed z-score normalization. By applying this transformation, values of variables have a mean equal to 0 and a standard deviation of 1.
2. **Missing values.** Our data subsets contained an amount of 9.70% of missing data. Missing data were due mainly, to some technical accidents during the manipulation of samples. To obtain a complete dataset, we have imputed missing data using the Multiple Imputation by Chained Equations (MICE) imputation method.
3. **Outlying data processing.** Detection of outliers was performed using the Inter Quartile Range (IQR) method. A total amount of 7.61% of outlying data was detected. We decided to process outliers by imputing them using the K-Nearest Neighbour imputation method. However, they will be studied conscientiously for an extended future analysis.

Table 5.1 – List of variables that were used for the assessment of water quality of the Mexican rivers Tula, Humaya, Tamazual and Culiacan.

Variables				
Macro-pollutants	**Micro-pollutants**			**Biological indices**
General chemistry	**Heavy metals**	**Organochlorine pesticides**	**PPCPs**	
pH	As	I-BHC	Ibuprofen	Number of total taxa
conductivity	Cd	II-BHC	Naproxen	Number of EPT taxa
CO_3	Fe	III-BHC	Triclosan	Number of Ephemeroptera taxa
HCO_3	Mn	IV-BHC	Diclofenac	Number of Plecoptera taxa
SO_4	Pb	Heptachlor	Bisphenol A	Number of Trichoptera taxa
Cl	Zn	Aldrin		Number of families in common
F		Heptachlor epoxide		% EPT
Na		I-Endosulfan		% Ephemeroptera
K		Dieldrin		% Plecoptera
Ca		DDE		% Trichoptera
Mg		Endrin		% Coleoptera
B		II-Endosulfan		% Diptera
SiO_2		DDD		% Chironomidae
NO_3		Endrin aldehyde		EPT/ Chironomidae
		DDT		% of most dominant genus
		Endrin ketone		% of dominant taxa
		Methoxychlor		Shannon's Index
				Simpson's Index
				Margalef Index
				Sequantial Comparison Index
				Jaccard's coefficient
				Sørensen coefficient
				Trent Biotic index
				Extended Biotic index
				Beck Biotic index
				Family Biotic index
				BMWP
				BMWP-CR
				BMWP-ASPT
				% Filterer collectors
				% Scrapers
				% Shredder
				% Predators
				% gathering collectors
				% IBI-west central Mexico

BMWP: Biological Monitoring Working Party.
BMWP-CR: Biological Monitoring Working Party-Costa Rica.
BMWP-ASPT: Biological Monitoring Working Party-Average Score per Taxon.

5.4.3 Results of the analysis

In order to respond to the questions listed at the beginning of Section 5.4, we have applied a set of statistical analyses to our preprocessed data. Below, we present the results.

1. **Considering the different sources of pollution from PPCPs and pesticides, are they found in the same place or are they found with the same concentrations in the different sampling sites?**
   For the sake of simplicity, we show exclusively the results obtained for the analysis of pesticides. We have computed a linear regression. First we have done a selection of variables by performing a Linear correlation-based feature selection. Feature selection was performed to obtain a regression model that represented the best our phenomena. A total of 9 variables over the 18 corresponding to the concentrations of organochlorine pesticides were used as predictors and a variable named *distance* was used as response. We have created the variable *distance* using **Google Earth**, it measures the distance between agricultural areas and each sampling site. In Figure 5.6, we illustrate the distance taken between agricultural areas and the sampling sites C1 to C11.

   We have graphically represented the concentrations of the pesticides for each sampling site (c.f. Figure 5.7). In Tables 5.2 and 5.3, we provide details of the linear model for the regression of distance between sampling sites and agricultural areas on the concentration of pesticides for the Mexican data. The R^2 value indicates that our model fits well our data, the F-statistics value suggests that at least one of the pesticides must be related to the distance between sampling sites and the agricultural areas. However, the p-values of the variables used for the analysis (c.f. Table 5.3) indicate that changes on the distance are not related to changes in the pesticides variables. The linear regression model using only the distance variable and the concentration of pesticides only provides us information about the relationship between the proximity of agricultural fields and the content of pesticides in the samples. This means that, the closest an agricultural field is from a sampling site, the content of pesticides in a water sample will be higher. In Figure 5.7, we observed that site C11 has the highest content of the analyzed pesticides. We also observed that site C11 collects the water discharges of the flow and is the closest of an agricultural area (c.f. Figure 5.6). From these observations, we can deduce that the quantities of pesticides depend in part, on the proximity of pollution sources in this case from the agricultural fields thus, high concentrations of pesticides will be found at sampling stations closer to agricultural fields.
   It must be noticed that the regression model built using only the distance variable as the response and the concentrations of organochlorine pesticides as predictors does not provide complete information to model pesticides fate in surface waters. To this purpose additional parameters (i.e., organic matter content, particle size of sediment, concentration of pesticide, polarity, etc.) and complex models should be further considered (Nowell et al. (2009); Holvoet et al. (2007)).

Table 5.2 – Summary of results about the linear model for the regression of distance between sampling sites and agricultural areas on the concentration of pesticides of the Mexican dataset.

Quantity	Value
Residual standard error	467
R^2	0.9117
F-statistic	13.91

Table 5.3 – Results from the linear regression analysis on the Mexican dataset. Variables used for the analysis were chosen using the Linear correlation-based feature selection method.

	Coefficient	**Std. error**	**t-statistics**	**p-value**
Intercept	46980.5403	19359.5346	2.4267	0.1360
I_BHC	-804097.7114	2540280.9059	-0.3165	0.7815
I_Endosulfan	-1785134.3199	2313596.7093	-0.7715	0.5210
Dieldrin	4833030.3504	1352013.2102	3.5746	0.0701
Endrin	-10029972.3708	3403391.7747	-2.9470	0.0984
Endrin_aldehyde	2688714.3977	1645159.3280	1.6343	0.2438
Endosulfan_sulfate	-12548183.7978	5488622.0854	-2.2862	0.1495
Endrin_ketone	-1008695.0461	387653.0960	-2.6020	0.1213
Methoxychlor	4606679.6139	1603210.0739	2.8734	0.1027

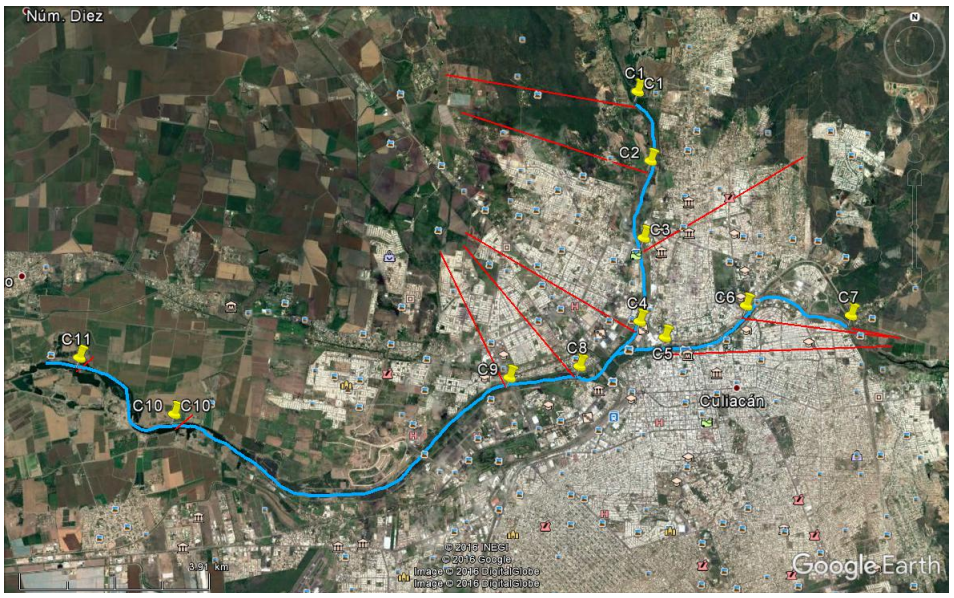


Figure 5.6 – Google Earth image of the sampling sites C1 to C11 situated on the rivers Tamazula, Humaya and Culiacan. Distance between agricultural areas and the sampling sites are defined by a red line, sampling sites are highligted by a yellow icon. *Source:* "Culiacan." 24°48'00"N 107°23'00"O. **Google Earth** ©. August 5, 2016. November 22, 2016.

.

2. **Which is the correlation between heavy metals, pesticides, PPCPs, and biomonitoring metrics?**
   To distinguish relationships among the different variables, we have subsetted the original dataset into three datasets named *macro*, *micro*, and *metals*. The *macro* subset contains the general chemistry variables related to the macro-pollutants listed on Table 5.1, while micro and metals datasets contain the variables of Micro-pollutants and Heavy metals respectively. A correlation matrix and a PCA analysis were performed on each data subset. We have constructed the PCA using the biomonitoring metrics as supplementary variables.
   From the correlation matrix (c.f. Figure 5.8), we observed that, out of the 35 biomonitoring metrics, only 13 are positively correlated to the macro-pollutants, 8 to the micro-pollutants, and 19 to metals.
   A better visualization of the correlation between the different variables was obtained from a PCA analysis. Figure 5.9 shows the first two principal components for the macro-pollutants, micro-pollutants, and metals subsets respectively.
   For macro-pollutants and micro-pollutants, we observed that, some biomonitoring

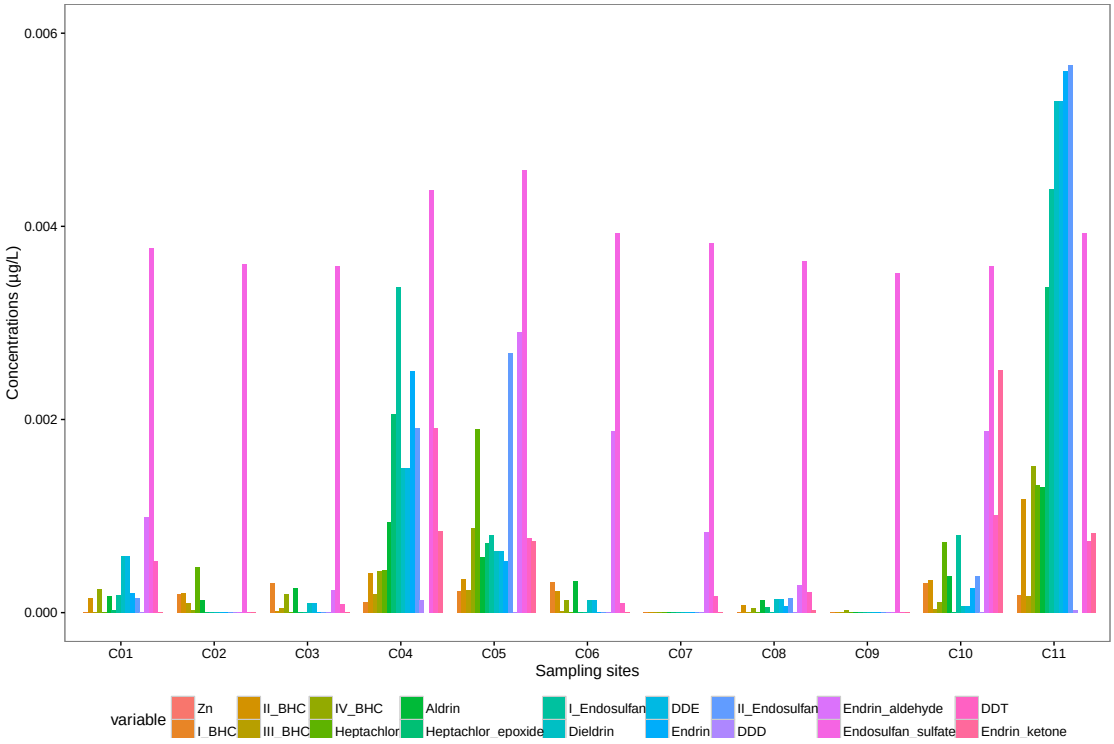


Figure 5.7 – Concentrations of Pesticides on sites C1 to C11 situated on the rivers Tamazula, Humaya, and Culiacan.

.

metrics show a negative correlation with these pollutants. This observation indicates that at high concentrations of pollutants, low values of biomonitoring metrics will be observed. Such observation is consistent with biomonitoring metrics themselves indeed, certain biomonitoring metrics (i.e., BMWP, EBI, Shannon or Simpson's indices) show low values when poor aquatic biodiversity is observed therefore a polluted aquatic system. The Family Biotic Index (FBI) show a similar behaviour compared to the other metrics. In fact, FBI has high values at high concentration levels of pollutants and, as we can observe on Figure 5.9 (a) and (b) this metric is positively correlated to the macro- and micro-pollutants.

For the case of metals, our results do not provide enough information about the relationship between the metals and the biomonitoring metrics. As we observed on the PCA the percentage of explained variance provided by the two axes is 55% which is lower compared to the axes for macro and micro pollutants (78% and 69% respectively).

3. **What are the most representative or pertinent pollutants to study in order to assess the impact of specific anthropological activities (i.e., agricultural activities)?**

   For this question, we focused on the impact of agricultural activities. To this, we have performed a linear correlation-based feature selection to identify the variables that are the most pertinent to assess this activity. We have used our variable *distance* that we previously described in Question 1 (Considering the different sources of pollution between PPCPs and pesticides, are they found in the same place or have the same trends?).

   For a detailed analysis, we have subsetted the Mexican dataset into three subsets

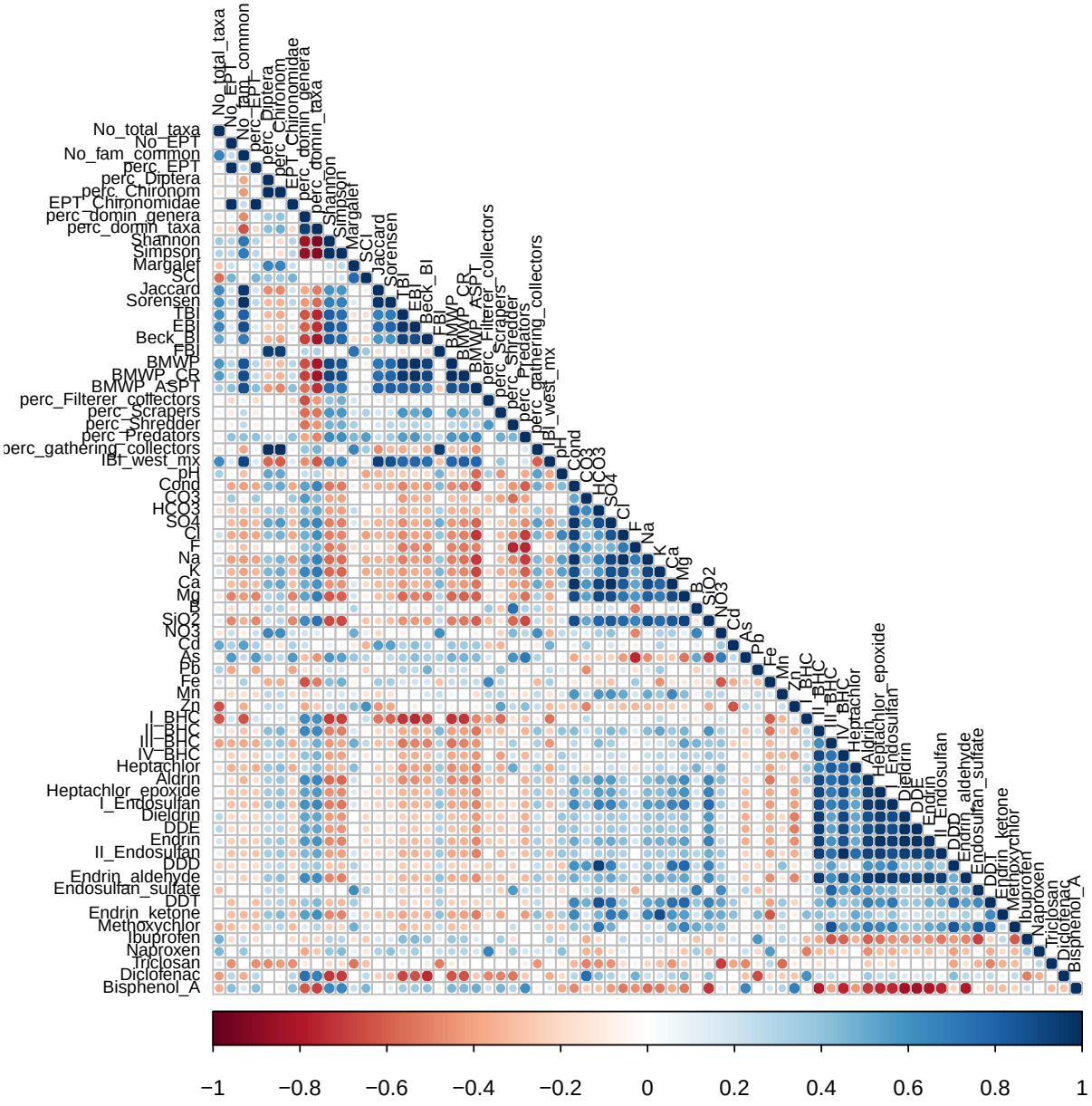


Figure 5.8 – Lower level correlation matrix of the Mexican dataset. Name of variables are abbreviated, see Table 5.1 for details.

.

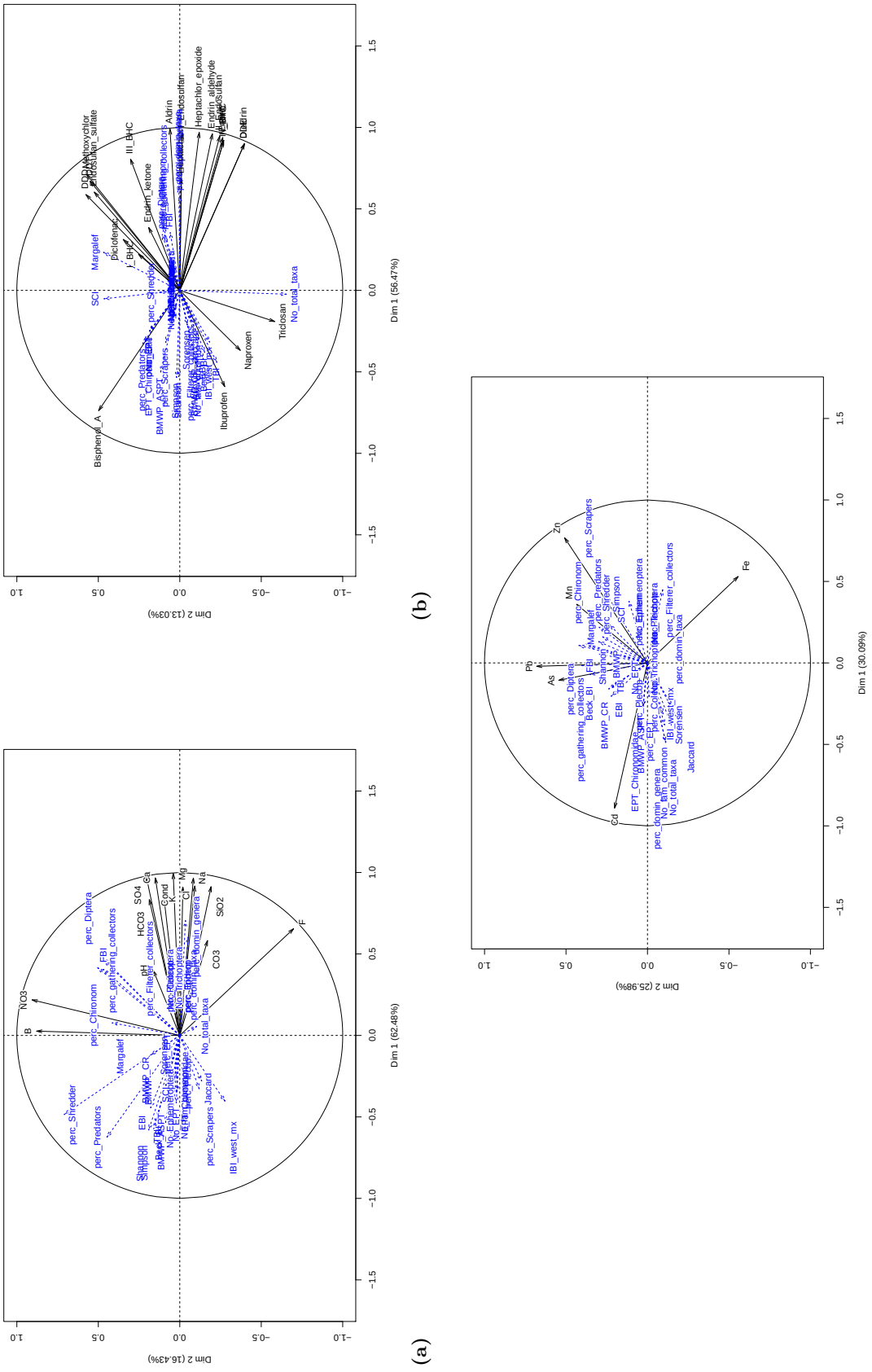


Figure 5.9 – PCA biplot of the two principal components for the macro-pollutants (a), micro-pollutants (b), and metals (c). Biomonitoring metrics are coloured in blue.

named *macro*, *micro*, and *bio*. They are composed of the variables *Macro-pollutants*, *Micro-pollutants* and *Biological indices* as detailed in Table 5.1.
Results from the feature selection on each subset is given in Table 5.4. For the assessment of the impact of agricultural activities, it appear that only 8 macro-pollutants, 12 micro-pollutants, and 12 biological indices are pertinent to use.

Table 5.4 – Results of the linear correlation-based feature selection on the data subsets *macro*, *micro* and *bio*.

Name of subset	Number of original variables	Number of selected variables	Name of selected variables
macro	15	8	pH, conductivity, SO_4, Cl, Na, K, Ca, SiO_2
micro	24	12	II-BHC, I-Endosulfan, Dieldrin, DDE, Endrin, Endrin-aldehyde, Endosulfan-sulfate, Endrin-ketone, Methoxychlor, Ibuprofen, Bisphenol-A
bio	36	9	percentage of diptera, EPT-Chironomidae, Shannon's index, Simpson's index, Margalef, SCI, Sørensen index, TBI, BMWP

4. **Is it appropriate to analyse all the pollutants if the conclusions are similar for each type of pollutant?**
Certainly, each type of pollutant provide different information and scientist may be interested to study most of them. However, as we will illustrate, not all pollutants can necessarily be included to study the impact of specific anthropogenic activity.
Lets us continue using the agricultural activity as the subject of the next study. To assess the impact of agricultural activities on the quality of water, scientist may be interested in determining the content of different pollutants such as metals, minerals, and pesticides. Then, the complete set of variables could be used to predict the pollution due to the proximity of agricultural activities. However, this action is not necessarily the most efficient. To exemplify this, we have performed a linear regression model.
We have used the three data subsets named *macro*, *micro*, and *bio* as previously described in Question 3 (*What are the most representative or pertinent pollutants to study in order to assess the impact of specific anthropological activities (i.e., agricultural activities)?*). We have subsetted the data in order to study separately each type of pollutant as they provide different information.
A linear correlation-based feature selection was computed for the selection of variables. The regression model was performed on the data subsets that contain the totality of variables and on the subsets after the selection of features.
In Table 5.5, we present the results of the linear model for the regression of the distance between the sampling sites and the agricultural areas on the variables, for the *macro*, *micro*, and *bio* data subsets. The construction of the regression model using all variables seems to be inadequate: the results indicate that the number of observations n is not large enough to construct a model with 15 (for *macro*), 24 (*micro*) and 36 (*bio*) variables. However, we observed that the model fits better

with a reduced number of features. The most promising results were observed on the *micro* and *bio* subsets.

Our result suggest that a limited number of variables will be sufficient to construct a linear regression model to predict the impact of agricultural activities on the quality of water.
It must be noted that a linear model will perform well on test observations if n, the number of observations, is much larger than p, the number of variables. However, a large variability of the model exists if n is not much larger than p, which results in over-fitting thus on poor predictions. It is frequently observed that some or many of the variables used in a model are not associated with the response. Such irrelevant variables lead to complex resulting models. A model that can be easily interpreted can be obtained by removing irrelevant variables. Some of the approaches that include feature selection or variable selection include stepwise model selection procedures (e.g., Forward Stepwise selection, Backward Stepwise selection), shrinkage methods (e.g., Ridge regression, LASSO), and dimension reduction methods (e.g., Principal Component Regression, Partial Least Squares) (James et al., 2013).

Table 5.5 – Results of the linear regression on the macro, micro and bio data subsets before and after feature selection.

	Name of data subset	**Number of variables**	**Quantity**		
			Residual standard error	R^2	**F-statistic**
After feature selection	macro	8	999.7	0.5955	2.84
	micro	12	266.4	0.9713	43.25
	bio	9	364.9	0.9461	20.5

5.5 Conclusion and future development

The analysis of environmental data demands the application of various data preprocessing tasks in order to obtain a dataset in conformance with the requirements of data analysis methods and useful for further data mining. Therefore, scientists may be confronted to two main challenges: (1) the efficient application of data preprocessing procedures and (2) the appropriate selection of the methods to preprocess data. To cover these challenges, we have developed a fully integrated analytics environment in R.

The originality of our prototype, is that it includes a set of data preprocessing procedures and statistical analysis for environmental data in general and for water quality data analytics in particular with user-friendly interface. The users with little knowledge on R can inspect, preprocess, and analyse their data by simply clicking on the desired action then, EvDA recommends and executes the necessary scripts to analyse data and finally returns results that can be saved. Concerning preprocessing procedures, EvDA analyses the characteristics of data and guides the user for the selection of the most appropriate preprocessing methods.

EvDA provides the following functionalities:

- Inspection of the statistical characteristics of data;
- Data preprocessing procedures;

- Visual representation of analytical steps;
- Display of original and preprocessed data;
- Recording of the preprocessing procedures that have been executed.

We have presented our prototype EvDA for preprocessing and analysis of environmental data and its application on our case study: Water pollution of the Tula, Culiacan, Tamazula, and Humaya river. Currently our prototype has some limitations and some of the adaptations that will be implemented in a short-term future include:

- Adaptation of data preprocessing procedures for categorical data. Currently, the data preprocessing procedures are suitable only for quantitative data;
- Addition of supplementary data preprocessing procedures: EvDa includes a limited number of methods to normalize, select features, impute missing values, and process outliers;
- Inclusion of other data preprocessing procedures such as data transformation, discretization, instance selection, etc;
- Development of more interactive options: our prototype does not allow the setting of the method parameters for most of the statistical analysis;
- Development of interactive options related to visualization of data and presentation of data that may be interesting for the user;
- Inclusion of spatial spatio-temporal data analysis methods: the user may also be interested in representing statistical information on the map particularly to study the dynamic phenomena.

CHAPTER 6

Conclusions

Data obtained from environmental surveys may be prone to have different anomalies (i.e., incomplete, inconsistent, inaccurate, or outlying data). These anomalies affect the quality of environmental data and may have severe consequences when assessing environmental ecosystems. To get correct and useful results from data mining and statistical analysis tasks, it is necessary to acquire quality data and preprocessed data. Different shortcomings are associated to the acquisition and preprocessing of environmental data. In chemical data acquisition for instance, there are analytical techniques with limited performance, non standardized methods or ambiguous data collection protocols. In data science, some of the shortcomings include: the sequence of data preprocessing procedures is not handed and depends on the dataset characteristics; the impact of data preprocessing procedures on statistical results has been understudied. Within the context of this thesis, we were interested on a particular shortcoming, the impact of data preprocessing procedures on subsequent statistical analysis and in one specific applicative domain: water quality assessment. Although we have applied our approaches in this domain, they can be adapted for analysis and preprocessing of other type of environmental data.

Collection, analysis, and preprocessing of environmental data is a complex subject and it requires specialized knowledge of each domain. Environmental scientists are interested among others in developing methods and/or protocols to analyse and collect environmental samples while data scientist are interested in developing algorithms and information systems to better analyse, preprocess, and visualize any data. The goal is to provide scientific solutions for understanding environmental phenomena that involve both environmental and data science. The challenges are vast and, in this work, we focused on one main problem: **How to analyse environmental data and provide accurate results?** To answer this question we propose a suite of methodological approaches to collect, preprocess, and analyse environmental data in order to control the entire process and guarantee accurate analysis results.

We were interested on the development of approaches and tools that will allow scientists to: (1) acquire quality data and (2) perform the most appropriate data preprocessing procedures to finally obtain accurate results in the context of water quality assessment.

Our approaches are multidisciplinary and thus related to both environmental sciences and data science, they include:

- a method for the collection of water quality data, specifically for the quantification of pollutants at low concentrations, which reduces incompleteness of data;

- a methodological procedure for the collection of biological data, that diminishes data errors due to outlying data, missing values or inconsistencies;
- an approach to identify the optimal data preprocessing strategies for specific statistical methods.

6.1 Contributions

We have provided contributions related to Environmental Chemistry, Hydrobiology, and Data Science. Hereafter, we detail each one.

Environmental Chemistry:

- *The analysis of pesticides in samples of water:* The existing analytical methods to quantify emerging pollutants on environmental samples use advanced instrumentation such as Liquid Chromatography tandem Mass Spectrometry (LC-MS-MS) or Gas Chromatography Mass Spectrometry (GC-MS). However, in developing countries access to advanced analytical instrumentation may be limited as consequence, analysis of emerging pollutants is rarely done. We adapted an analytical method for the simultaneous analysis of eighteen organochlorine pesticides in samples of water including: I-BHC, II-BHC, III-BHC, IV-BHC, heptachlor, aldrin, heptachlor epoxide, I-endosulfan, II-endosulfan, dieldrin, DDE, endrin, DDD, endrin aldehyde, endosulfan sulfate, DDT, endrin ketone and methoxychlor. The proposed analytical method uses the Solid Phase Extraction (SPE) to extract analytes followed by Gas Chromatography-Electron Capture Detector (GC-ECD) for their quantification. The Liquid-Liquid extraction method (LLE) has been used for the analysis of organochlorine pesticides (Wu et al., 2010; Diaz et al., 2008) however, it is a laborious method, use large amount of solvent and require intensive cleaning procedures to assure contaminant free material to determine pesticides at trace levels. The analytical method that we proposed requires small amount of solvent, is simple, it provides better recovery for the majority of the pesticides (recovery ranging from 72% to 92% on average) compared to LLE (recovery ranging from 15% to 40%) and reduces incertitudes and errors that may appear due to intensive manipulation of samples.

- *Analysis of Pharmaceutical and Personal Care Products (PPCPs) in samples of water:* A collaborative work between the Institute of Geophysics and the Faculty of Chemistry at UNAM has also been done to develop a new Solid Phase Micro Extraction-Gas Chromatography-Mass Spectrometry (SPME-GC-MS) method for the simultaneous analysis of PPCPs (ibuprofen, 2-benzyl-4-chlorophenol, naproxen, triclosan, ketoprofen, diclofenac, bisphenol A and estrone) in river water. Compared to other methods (Huang et al., 2015; Yu et al., 2012), ours is simple, rapid and efficient. It allows to extract target analytes from a miniaturized system where volume samples reach few millilitres and where manual intervention of the analyst is reduced to the minimum. This is advantageous to reduce anomalies on data.

The analytical methods that we propose allow us to quantify pesticides at a concentration range of $\mu g L^{-1}$ and $ng L^{-1}$ for pesticides and PPCPs respectively. We propose to use our analytical methods to reduce the presence of data anomalies such as missing values, outliers, and censored data.

Hydrobiology:

Use of biomonitoring metrics for the assessment of water quality in Mexico is scarce

despite their advantages (Mathuriau et al., 2011). Their limited used is due to: absence of sampling and analytical protocols, lack of identification keys, incomplete information of local communities and absence of monitoring metrics. These shortcomings increase the problems on data quality.
We have designed a new methodological approach for the acquisition of data using biomonitoring metrics in Mexico. We have done a bibliographic study where we identified thirty five macroinvertebrate-based biomonitoring metrics of potential use for the ecological assessment of surface waters in Mexico. Metrics were selected considering the following characteristics: sensitivity, ecological relevance, representative, feasibility, metric interpretation, performance, and geographical suitability. Sampling and analytical procedures to compute the selected metrics are also described. We suggest the use of our methodological approach to increase the quality of data, reduce uncertainties, incompleteness or inconsistencies in the dataset due to undefined sampling and analytical protocols for biomonitoring assessment in Mexico.

Data Science:

When analysing environmental data, various data preprocessing tasks need to be applied to obtain a dataset in conformance with the requirement of data analysis methods, and useful for further data analysis. Numerous data preprocessing procedures are available (c.f. Chapter 2.4) and their appropriate selection is necessary to get the less biased statistical analysis results. An important aspect when analysing environmental data is that contrary to data scientists, environmental scientists are not familiar with the tasks to preprocess data such as: data cleaning, normalization, data transformation, feature selection, etc. Therefore, data preprocessing and data analysis may be a very critical step. We aim at providing to environmental scientists a guide to inspect, preprocess and analyse environmental data. Our contributions are:

- *Selection of data preprocessing procedures:* We proposed an optimal selection of data preprocessing procedures to treat common data anomalies and data problems (feature selection, normalization, missing values and outliers). They consist of procedures already available on the R environment for statistical computing. The selected preprocessing procedures respond to the need to treat data anomalies of datasets with different dimensions, providing accurate statistical results.
- *Assessment of data preprocessing procedures:* We evaluated the impact of data preprocessing procedures on subsequent statistical analysis. From this evaluation we suggested the data preprocessing procedures that are the most appropriate to get the least biased analytical results.
- *Development of an integrated analytics environment in R:* An integrated analytics environment, named *EvDA*, for statistical analysis of environmental data, was proposed. EvDA is a user-friendly Shiny/R application relying on statistics to guarantee data quality and quality results. It allows users with little knowledge of R to inspect, preprocess, and analyse their data. An stable version of our prototype will be available on Github in a short-term future.

6.2 Future work

The approaches developed in this work are, to our-point-of-view, an important step on environmental informatics. However, there are still more work to be done including:

For our approaches in Environmental Chemistry and Hydrobiology:

- Adapt of our analytical technique (SPE GC-ECD) for the analysis of other types of

emerging pollutants and types of samples (i.e., soils and sediments). We were focused on the development of an analytical method that allow us analyse specifically organochlorine pesticides and some PPCPs but, many other pollutants affect the aquatic environment (e.g., oils, surfactants, industrial wastes,etc.) and the development or adaptation of methods to analyse them is necessary;
- Development of local biomonitoring metrics. We have demonstrated the usefulness of biomonitoring metrics using macroinvertebrates for the water quality assessment of Mexican rivers. However, it will be favourable to develop a biomonitoring metric suitable to the ecological, climatological and geographic characteristics of Mexico. Because in Mexico the diversity of biological species is vast, there are endemic species, and the climate and geographic characteristics are heterogeneous and complex.

For our approaches in Data science:

We have designed our experimentations to study the most frequent data anomalies (non normalized data, irrelevant features, missing values, outlying data) using the most general case (datasets with multivariate normal distribution). However, other data anomalies may appear on data (e.g., censored data, uncertainties), and real-world dataset may have different characteristics to those conducted in our study. Due to our limited time, we have decided to focused on the above mentioned cases but it most be noticed that further studies need to be done in order to cover the full spectrum of possibilities in order to appropriately preprocess data.

In addition, our experimentations were developed under the assumption that data anomalies appear as isolated cases. However, data anomalies may co-occur (Berti-Equille et al., 2015) thus, the results of data preprocessing may be different to those observed in our study. We insist on the need to continue studying in this topic and some future work that we suggest include:

- Enlarge the study related to data anomalies and data preprocessing. In this study we were focused on a limited number of data anomalies and data preprocessing procedures. To continue contributing to this area of study, other types of data anomalies (e.g., censored data, duplicates, data redundancies, inconsistencies) and preprocessing procedures (e.g., transformation, discretization, integration) should be studied;
- Adjust our approach to other environmental data such as ecological biodiversity, geological studies, air or soil pollution;
- Study the effect of preprocessing procedures under the presence of multiple data anomalies;
- Adjust our approach to datasets of varied characteristics.

We also have proposed a prototype to inspect, preprocess and analyse environmental data. It can be used as a basis to create robust-user-friendly tools that allow preprocess and analyse environmental data. Anomalies on data are not exclusive of environmental studies, they may occur also in biology, astrophysics, geology or engineering, to mention some. It is necessary to continue producing knowledge on this exiting and interesting area.

Related publications

- Diaz-Flores, L., Peña, A., Armienta-Hernandez, M., and Serrano-Balderas, E. (under revision). Determination of pharmaceuticals and personal care products (PPCPs) in river water and sediment by solid phase extraction followed by gas chromatography mass spectrometry (SPME-GC-MS). Submitted to Journal of Analytical Chemistry.

- Serrano Balderas, E.C., Berti-Equille, L., Armienta Hernandez, M.A., Desconnets, J.-C. 4238. Water Quality Data Aanlytics. In: Sauvage, S., Sánchez-Pérez, J.M., Rizzoli, A.E. (Eds.), 4238. Proceedings of the 8th International Congress on Environmental Modelling and Software, July 32-36, (2016). Toulouse, FRANCE. ISBN : 978-88-9035-745-9.

- Serrano Balderas, E., Grac, C., Berti-Equille, L., and Armienta Hernandez, M.A. (2016). Potential application of macroinvertebrates indices in bioassessment of mexican streams. Ecological Indicators, 61:558-567.

- Berrahou L., Lalande N., Serrano E. Molla G. Berti-Equille L. Bimobnte S., Bringay S. Cernesson F., Grac C., Ienco D., Le Ber F., Teisseire M. (2015). A quality-aware spatial data warehouse for querying hydroecological data. Computers & Geosciences, :7: 348-357.

Oral communications

- Serrano Balderas, E. C., Berti-Equille, L., Grac, C., and Armienta Hernandez, M.A. (2014). Data processing for controlling data quality on surface water quality assessment. In Inforsid - Atelier SI et Environnement, LYON, France.

- Serrano Balderas E.C., Berti-Equille L., Armienta Hernandez M.A., Grac C. (2014). Impacts of Data Quality on Environmental Analysis: Application to Mexican Rivers Pollution. WomENcourage 1st Annual Meeting, Manchester, UK.

CHAPTER 7

Résumé étendue

Confrontés à la nécessité de réduire la pollution d'eau, des programmes de surveillance de la qualité de l'eau sont fréquemment mis en place. De tels programmes consistent à quantifier divers composants physiques et chimiques dans différents sites et pendant plusieurs périodes. Or les données collectés lors des programmes de surveillance sont sujets aux différents anomalies (i.e., incomplets, inconsistantes, imprécis ou aberrants). Des anomalies dans les données sont omniprésents et peuvent être présents dû aux problèmes dans les expériences, par des erreurs humains ou défaillances du système.

La mauvaise qualité de données peut être significativement coûteuse (Haug et al., 2011) et avoir des sévères conséquences dans l'évaluation des écosystèmes environnementaux (Wahlin and Grimvall, 2008). Pour produire des données de qualité et réduire des anomalies de données il est nécessaire : acquérir des données de qualité et pré-traiter des données (Han and Kamber, 2000 ; Berti-Equille, 2007a).

L'acquisition des données de qualité peut être obtenue 1) en suivant des protocoles d'échantillonnage et d'analyse normalisés et 2) en utilisant des outils analytiques avancées (i.e. spectrométrie de masse couplée à un plasma inductif (SF-ICP-MS), chromatographie en phase gazeuse – spectrométrie de masse (CG-MS) ou chromatographie liquide couplée à la spectrométrie de masse en tandem (LC-MS/MS)). Cependant, dans les pays en voie de développement l'accès aux tels outils analytiques est limité ou des protocoles normalisés sont difficiles à mettre en place. Il est donc nécessaire d'adapter des outils qui peuvent être faciles à utiliser et bon marché sans compromettre la qualité des données.

Concernant le pré-traitement des données il s'agit de tout procédure qui est nécessaire afin de générer des données de qualité. Ces procédures incluent :

- Nettoyage des données (i.e., suppression de valeurs aberrantes et des incohérences) ;
- Élimination des inconsistances des données (i.e. réglage des différences);
- Complétude des données manquantes (i.e., remplissage des valeurs manquantes);
- Réduction des données (i.e., sélection d'attributs relevants) ou création des données agrégées.

Bien que la détection des anomalies dans les données ait été largement étudiée, très peu des travaux ont été publié en particulier dans le domaine de sciences de l'environnement. En considérant que les anomalies sur des données peuvent apparaître tout au long d'une étude environnementale il est nécessaire d'identifier, contrôler et dans le cas nécessaire

designer des protocoles d'acquisition et pré-traitement des données pour garantir des résultats précis.

Le travail de thèse présenté ici est positionné sur l'informatique environnementale puisqu'il a une approche pluridisciplinaire entre chimie de l'environnement, hydrobiologie, statistique, science des données et informatique. Ce travail a été centré dans deux défis principaux : l'acquisition des données de qualité et la définition des procédures de pré-traitement des données appropriés pour finalement analyser des données.

7.1 Motivations

Lors des études environnementaux il est important de considérer uniquement des données de qualité qui fournissent des informations valables. Il est évident que pour fournir des résultats précis il est nécessaire d'avoir des données de bonne qualité. Dans la science des données les procédures de pré-traitement des données sont proposées pour réduire des anomalies et ainsi améliorer la qualité des données. Le choix des méthodes de pré-traitement des données est crucial pour la validité des résultats d'analyses statistiques et il est assez mal défini.

Les procédures de pré-traitement de données amélioreront certainement la qualité de données cependant, des bonnes pratiques d'acquisition des données sont aussi nécessaires pour éviter des anomalies de données. Un problème important sur l'acquisition de données environnementales est l'absence des protocoles standardisés pour l'échantillonnage et l'analyse en particulière dans des laboratoires où l'accès aux instrumentations analytiques est limité. Il est donc nécessaire d'adapter des procédures analytiques qui sont rentables et faciles de mettre en œuvre pour obtenir des données de qualité.

Dans ce travail de thèse un intérêt particulier a été porté sur le développement d'une nouvelle approche qui combine à la fois des bonnes pratiques d'acquisition et de pré-traitement des données.

7.2 Objectifs

L'objectif principal de ce travail est de fournir à la communauté scientifique des approches méthodologiques et des outils pour l'acquisition, le pré-traitement et l'analyse des données environnementales en garantissant la qualité de résultats d'analyse et de données. Les approches méthodologiques ont été appliqués dans un cas d'étude : évaluation de la qualité de l'eau dans quatre rivières mexicaines (Tula, Tamazula, Humaya et Culiacan).

Le but de fournir une approche intégrée qui combine à la fois la bonne acquisition de données et des bonnes pratiques de pré-traitement de données est de contrôler tout la chaîne de traitement de données dès leur acquisition et jusqu'à leur analyse. Pour attendre l'objectif principal des objectifs spécifiques ont été définis. Ils sont liés aux trois domaines d'étude.

Chimie de l'environnement

- Acquérir des données en déployant des méthodes fiables et bon marché pour la quantification des pesticides organochlorés, des produits pharmaceutiques et de soin dans l'eau;
- Définir des protocoles pour l'échantillonnage et l'analyse des pesticides organochlorés, des produits pharmaceutiques et de soin;

- Mettre en place la campagne d'échantillonnage et les procédures analytiques pour l'analyse des échantillons d'eau des rivières mexicaines Tula, Humaya, Tamazula et Culiacan.

Hydrobiologie

- Acquérir des données en définissant une approche méthodologique à l'aide de métriques de biosurveillance (macroinvertébrés) en tant que nouveaux outils complémentaires pour la surveillance de la qualité d'eau des rivières mexicaines;
- Mettre en place la campagne d'échantillonnage et les procédures analytiques pour l'acquisition des données hydrobiologiques.

Sciences des données

- Définir une approche méthodologique pour la sélection des procédures de pré-traitement des données afin de traiter des anomalies sur les données et des problèmes sur les données (i.e., valeurs manquantes, valeurs aberrantes, sélection des variables, normalisation);
- Évaluer l'impact des procédures de pré-traitement des données sur des ultérieure analyses statistique;
- Déterminer les procédures de pré-traitement de données les plus appropriés pour obtenir les résultats des analyses moins biaisées;
- Préciser les procédures pour pré-traiter et analyser des données qui sont nécessaires pour garantir la fiabilité des résultats sur des études environnementaux en général et pour l'analyse de données de qualité de l'eau en particulier.

7.3 Acquisition des données

Dans le but d'acquérir des données de bonne qualité des approches méthodologiques en chimie de l'environnement et en hydrobiologie ont été proposées. Il s'agit des approches pour l'acquisition des données issue de la surveillance de la qualité de l'eau par rapport à la biosurveillance à l'aide des macroinvertébrés et à la pollution par : pesticides, produits pharmaceutiques et de soin. Pour attendre les objectifs préalablement décrit, des échantillons d'eau et des macroinvertébrés ont été prélevés dans quatre rivières mexicaines (Tula, Tamazula, Humaya et Culiacan).

7.3.1 Sites d'étude

Cinq sites (H1 – H5) situés au long de la rivière Tula ont été sélectionnés pour prélever des échantillons (c.f. Figure 7.1). Les sites H2 et H3 sont situés proche aux champs agricoles, ils ont été choisis pour évaluer les quantités des pesticides organochlorés dans l'eau de la rivière Tula. Les sites H1, H4 et H5 sont localisés près des zones urbanisées ils ont été sélectionnés pour évaluer les niveaux des polluants principalement PPCPS (Produits pharmaceutiques et de soin). Les détails sur l'emplacement des sites d'échantillonnage sont données dans la table 7.1

Pour évaluer la qualité de l'eau des rivières Tamazula, Humaya et Culiacan once sites ont été sélectionnés (c.f. Figure 7.2) Les détails sur l'emplacement de sites d'échantillonnage sont données dans la table 7.2. Quatre sites localisés au long de la rivière Humaya (C1 - C4) ont été choisis pour évaluer les concentrations de polluants principalement de déchets domestiques. Trois sites (C5 – C7) ont été pris dans la rivière Tamazula pour représenter l'état le moins pollué. Les sites C8 et C9 ont été choisis pour évaluer le niveaux de

Table 7.1 – Emplacement des sites d'échantillonnage de la rivière Tula

Sites	Commune	Latitude	Longitude	Altitude (m)
H1	Tasquillo	20° 33.703'	099° 18.581'	1761.134
H2	Ixmiquilpan	20° 28.829'	099° 13.277'	1730.654
H3	Ixmiquilpan	20° 29.585'	099° 13.310'	1706.270
H4	Tlacotlapilco	20° 22.451	099° 13.414'	1703.222
H5	Progreso	20° 14.696'	099° 12.344'	1702.917

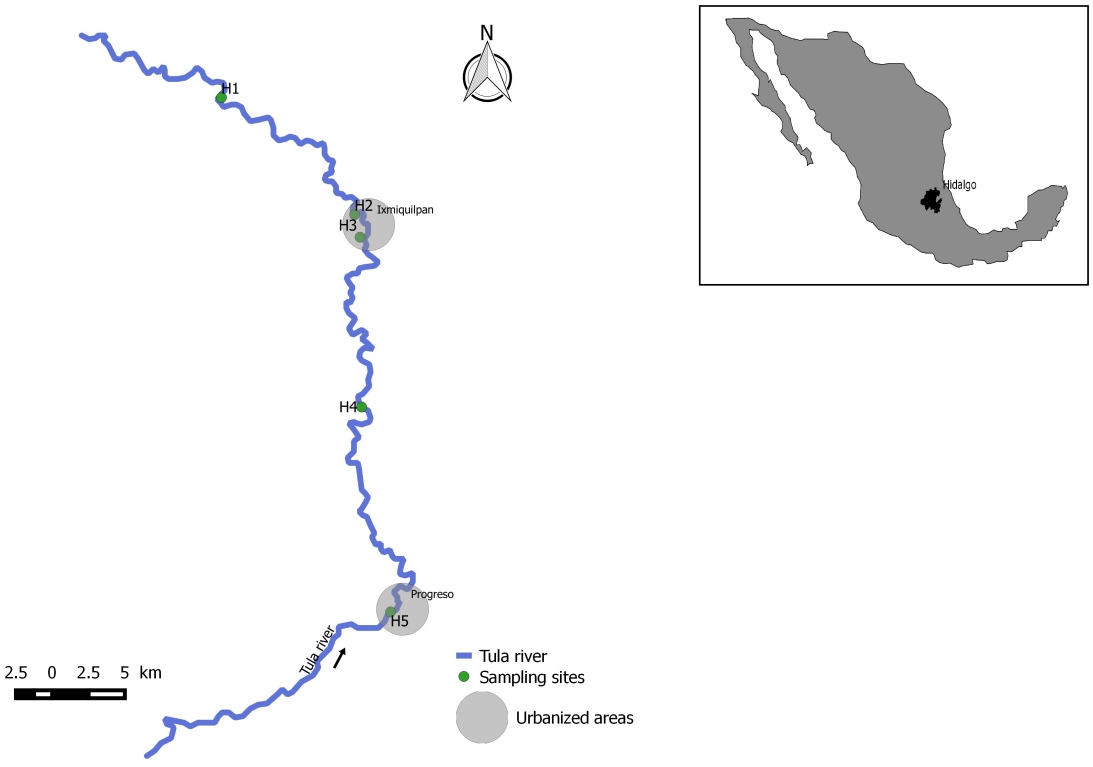


Figure 7.1 – Sites de prélèvement dans la rivière Tula (État d'Hidalgo, Mexique).

pollution par déchets domestiques et activités anthropiques dans la rivière Culiacan. En fin, les sites C10 et C11 ont été sélectionnées pour évaluer la pollution par des pesticides organochlorés.

Table 7.2 – Emplacement des sites d'échantillonnage des rivières Tamazula, Humaya et Culiacan.

Site	Commune	Latitude	Longitude	Altitude (m)
Humaya river				
C1	Mojolo	24° 50.316'	107° 24.775'	45.110
C2	Mojolo	24° 50.395'	107° 24.305'	44.805
C3	Mojolo	24° 49.934'	107° 24.227'	46.024
C4	Humaya	24° 49.045'	107° 24.199'	41.757
Tamazula river				
C5	Culiacan	24°48.796'	107° 23.755'	62.788
C6	Tamazula	24° 49.308'	107° 22.758'	46.329
C7	La Limita de Itaje	24° 49.005'	107° 21.630'	50.292
Culiacan river				
C8	Culiacan	24° 48.551'	107° 24.767'	45.415
C9	Culiacan	24° 47.524'	107° 26.873'	33.528
C10	Bacurimi	24° 47.848'	107° 30.462'	33.223
C11	Culiacancito	24° 48.409'	107° 31.883'	25.298

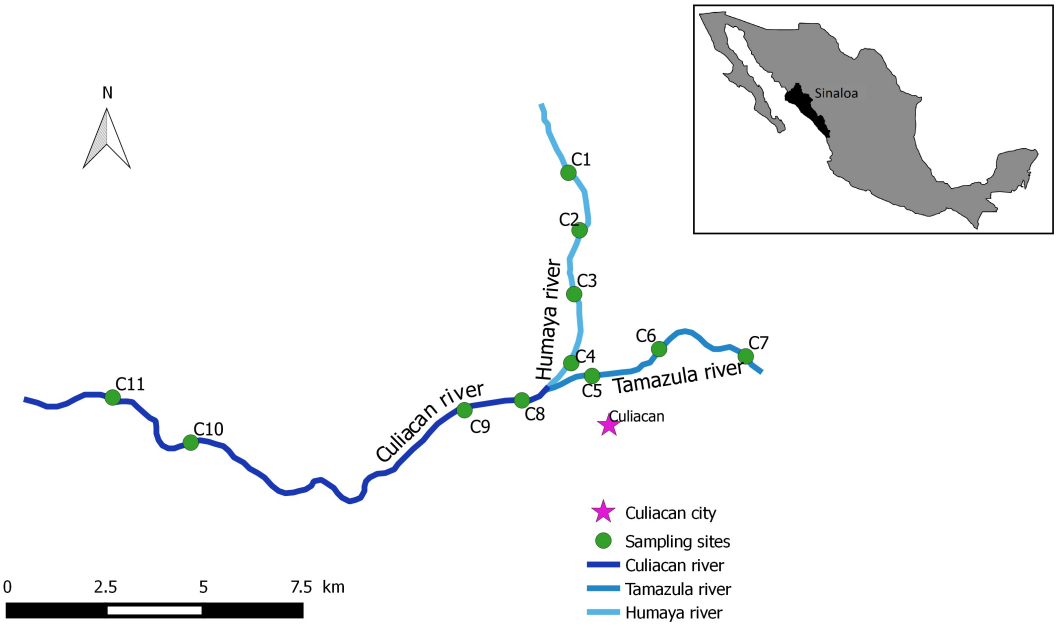


Figure 7.2 – Sites de prélèvement des rivières Tamazula, Humaya et Culiacn (État de Sinaloa, Mexique).

7.3.2 Acquisiton de données physico-chimiques et chimiques

Des échantillons d'eau ont été collectés pour déterminer le contenu de : pH, conductivité, carbonate (CO_3), bicarbonate (HCO_3), sulphate (SO_4), chlore (Cl), fluor (F), sodium (Na), potassium (K), calcium (Ca), magnésium (Mg), bore (B), dioxyde de silicium (SiO_2), nitrate (NO_3) arsenic (As), cadmium (Cd), plomb (Pb), fer (Fe), cuivre (Cu), manganèse (Mn), zinc (Zn), dix-huit pesticides organochlorés: I-BHC, II-BHC, III-BHC, IV-BHC, heptachlor, aldrin, heptachlor epoxide, I-endosulfan, II-endosulfan, dieldrin, DDE, endrin,

DDD, endrin aldehyde, endosulfan sulfate, DDT, endrin ketone et methoxychlor, et huit PPCPs (ibuprofen, 2-benzyl-4-chlorophenol, naproxen, triclosan, ketoprofen, diclofenac, bisphenol A et estrone).

L'analyse des éléments majeurs et métaux à été effectué en suivant le protocole de Armienta et al. (1987). Les données physico-chimiques et chimiques qui a été acquis est composé de 16 individus (sites d'échantillonnage) et 21 variables des valeurs numériques. Les sites de la rivière Tula ont été échantillonnées dans deux saisons (pluie et étiage) tandis que les rivières Tamazula, Humaya et Culiacan ont été échantillonnées uniquement dans la saison d'étiage.

7.3.3 Analyse des pesticides organochlorés et PPCPs

Une méthode analytique pour l'analyse des pesticides organochlorés a été adapté en modifiant la méthode décrit par Guardia Rubio et al. (2007). L'analyse a été réalisé en utilisant une extraction en phase solide (SPE) et la chromatographie en phase gazeuse avec détection par capture d'électrons (GC-ECD). L'analyse de PPCPs a été effectué en utilisant une nouvelle méthode basée sur la micro-extraction en phase solide suivi de la chromatographie en phase gazeuse couplé à un spectromètre de masse (SPME – GC – MS). La nouvelle méthode permet la quantification de PPCPs (ibuprofen, 2benzyl-4-chlorophenol, naproxen, triclosan, ketoprofen, diclofenac, bisphenol A et estrone) dans des échantillons d'eau à niveau de traces (ng/L). Les détails sur la méthode SPME-GC-MS sont présentés par Diaz-Flores et al. (en révision, Determination of pharmaceuticals and personal care products (PPCPs) in river water and sediment by solid phase extraction followed by gas chromatography-mass spectrometry (SPME-GC-MS) soumis à *Analytical Chemistry*) et donc il ne sera pas détaillé ici.

L'extraction en phase solide a été réalisée pour extraire des pesticides dans des échantillons d'eau. La première phase de la procédure consiste à filtrer les échantillons à travers un filtre Millipore 0.45 µm. Ensuite l'échantillon est déposé dans des cartouches C18 qui sont installés sur un système manifold. Les cartouches sont premièrement conditionnées en faisant passer 5mL de cichloromethane, 5mL de methanol et 5 mL d'eau distillé avec un flux de 1mL/min. Ensuite 500 mL d'échantillons d'eau sont versées à travers les cartouches (flux : 1mL/min). La cartouche a été séchée pendant 20 min sous vide. En fin l'extrait a été mis à sec dans un courant d'azote et redissous dans 1 mL d'ethyl acetate-hexane 25:75%, v/v.

Les données obtenus lors de l'analyse de pesticides et PPCPs est constitué de 16 individus (sites d'échantillonnage) et 23 variables (5 PPCPs et 18 pesticides organochlorés) des valeurs numériqes.

7.3.4 Acquisition des données hydrobiologiques

Afin de fournir des lignes directrices et des outils pratiques pour la biosurveillance des systèmes aquatiques une étude à été menée dont les objectives ont été : 1) identifier des méthodes de biosurveillance basées sur des macroinvertébrés qui peuvent être potentiellement utilisées dans l'évaluation écologique des rivières au Mexique, et 2) décrire les procédures d'échantillonnage et d'analyse pour mettre en place les méthodes identifiées en 1. Les résultats de cette étude ont été publiés (Serrano Balderas et al., 2016) et font partie de ce manuscrit. De manière synthétique trente-cinq métriques ont été identifiées comme des outils de potentiel application pour la surveillance écologique des rivières mexicaines.

Les procédures d'échantillonnage et d'analyse pour calculer les métriques sélectionnées sont décrits dans la même publication.

Le prélèvement des macroinvertébrés a été effectué tout d'abord par une inspection visuelle d'un tronçon de la rivière afin d'identifier les différents habitats et substrats. Des échantillons ont été prélevés dans des habitats multiples à l'aide d'un filet surber (surface 500 m^2, maillage 500 µm). Pour chape point d'échantillonnage, des prélèvements ont été répliqués deux fois. Les échantillons ont été stockés dans des sacs en plastique et conservés dans de l'éthanol 70%.

Les échantillons ont été analysés dans le laboratoire. Les macroinvertébrés ont été séparés des substrats et triés en utilisant une maille de 250 µm. Ensuite, ils ont été examinées sous un stéréoscope et identifiés au niveau des familles à l'aide des différents clés taxonomiques (Heckman, 2006; Heckman, 2008; Heckman, 2011; Merrit et al., 2008a; Tachet et al., 2010; Novelo-Gutiérrez, 1997b; Novelo-Gutiérrez, 1997a). Finalement, trente-cinq métriques (indicateurs biologiques) ont été calculé, ils comprennent:

- 5 indicateurs de richesse (nombre de taxons totales, nombre des taxons EPT, nombre des taxons Ephemeroptera, nombre des taxons Plecoptera et nombre des taxons Trichoptera) ;
- 11 indicateurs d'énumération (nombre des familles en commun, %EPT, %Ephemeroptera, % Plecoptera, % Trichoptera, % Coleoptera, % Diptera, % Chironomidae, % EPT: Chironomidae, % taxons plus dominants, % genres plus dominantes);
- 6 indices de diversité et similitude (Shanon, Simpson, Margalef, Sequential Comparison Index – SCI, Jaccard et Sørensen);
- 7 indices biotiques (Trent Biotic index, Extended Biotic index, Beck Biotic index, Family biotic index, Biological Monitoring Working Party et Average Score per Taxon);
- 5 groupes fonctionnelles (les filtreurs, les prédateurs, les détritivores, les herbivores et les omnivores) et;
- 1 indice multi-métrique (IBI-west central Mexico).

Le jeux des données hydrobiologiques collectées des rivières Tula, Tamazuala, Humaya et Culiacan est composé de 16 individus (sites d'échantillonnage) et 35 paramètres numériques (35 métriques).

7.4 Pré-traitement et analyse des données

Afin de fournir aux spécialiste de l'environnement un guide pour inspecter, pré-traiter et analyser des données environnementales une étude sur des meilleurs pratiques pour pré-traiter les données a été menée. Cette étude a été centré sur des procédures de pré-traitement des données (sélection des variables, normalisation des données, imputation des valeurs manquantes et traitement des valeurs aberrants) et leur impact sur des subséquent analyses statistique principalement les méthodes de régression, classification et groupement. Les objectives de l'approche sont les suivants :

- Évaluer la robustesse des méthodes pour traiter des valeurs manquantes et aberrants ;
- Examiner l'effet des procédures de pré-traitement de données sur la précision dans des analyses de classification, groupement et régression ;
- Identifier les meilleurs procédures de pré-traitement des données pour une analyse statistique particulière.

L'étude a été structurée en trois parties : (1) génération des données synthétiques, (2)

pré-traitement des données et (3) analyse statistique (c.f. Figure 7.3). Dans la première partie, des données synthétiques ont été construites afin d'évaluer les procédures de pré-traitement de données puis, à l'étape deux, des données pré-traités ont été construites en utilisant différents procédures de pré-traitement sur les données synthétiques qui ont été construites dans la première partie. Dans la troisième étape, les données pré-traitées ont été utilisées pour évaluer l'impact des procédures de pré-traitement sur les résultats des analyses statistique. Une description détaillée de chaque partie de l'étude est donnée ci-dessous.

7.4.1 Données synthétiques et semi-synthétiques

Quatre jeux de données synthétiques ont été utilisés pour évaluer l'impact des procédures de sélection de variables sur des subséquentes analyses de régression, classification et groupement. Chaque jeu suit une loi normale et est composé d'un nombre différent d'observations (i.e., $n = 21$, $n = 600$, $n = 4000$, $n = 20000$), variables numériques (i.e., $p = 8$, $p = 30$, $p = 53$, $p = 98$) et une variable catégorielle de cinq classes suivant une distribution uniforme. Les variables incluent des variables non pertinentes.

Les tests pour évaluer l'impact de pré-traitement par normalisation des données ont été exécutes en utilisant quatre jeux de données synthétiques composés d'un nombre différent d'observations (i.e., $n = 21$, $n = 600$, $n = 4000$, $n = 20000$), variables numériques (i.e., $p = 8$, $p = 30$, $p = 53$, $p = 98$) et, une variable catégorielle de cinq classes. Chaque jeu de données suit une distribution Weibull.

L'évaluation des méthodes d'imputation des valeurs manquantes a été effectué en utilisant quatre jeu de données synthétiques. Chaque jeu a été composé d'un nombre différent d'observations (i.e., $n = 21$, $n = 600$, $n = 4000$, $n = 20000$), variables numériques (i.e., $p = 8$, $p = 30$, $p = 53$, $p = 98$) suivant une loi normale et une variable catégorielle de cinq classes suivant une distribution uniforme. Des valeurs manquantes on été injectés de manière aléatoire sur les variables numériques sur chaque jeu de données dans les proportions suivantes : 5%, 10%, 15%, 20%, 25% et 30%. Pour chaque injection de valeurs manquantes dix répétitions ont été effectués pour s'assurer que les résultats expérimentaux ont été statistiquement acceptables.

En fin, des tests pour évaluer le pré-traitement de valeurs aberrants a été exécuté en utilisant trois jeux de données synthétiques composées d'un nombre différent des observations (i.e., $n = 21$, $n = 600$, $n = 4000$), variables numériques (i.e., $p = 8$, $p = 30$, $p = 53$) suivant une loi normale et une variable catégorielle de cinq classes suivant une distribution uniforme. Des valeurs aberrantes ont été injectés de manière aléatoire sur chaque jeu de données dans les proportions suivantes: 1.5%, 2.5%, 5%, 10% et 15%.

Nous avons utilisé également des données semi-synthétiques. Tout d'abord nous avons collecté des données réelles issues de campagnes de surveillance des eaux de la région Rhin-Meuse, France. Ce jeu de données se composent de 3787 individus (stations d'échantillonnage) et 55 variables. Cet ensemble de données a été traitée afin d'éliminer les données bruitées. Le jeu de données nettoyé se composait de 3726 individus et 48 variables. A partir de ce jeu de données deux ensembles de données semi-synthétiques ont été générés. Le premier est une présentation réduite, il est composé de 38 individus et 35 variables sélectionnés de manière aléatoire. Le deuxième jeu de données a été construit en introduisant de nouvelles observations. Ces observations suivent la même répartition que le jeu de données nettoyé. Le deuxième jeu de données est composé de 9742 individus et 48 variables.

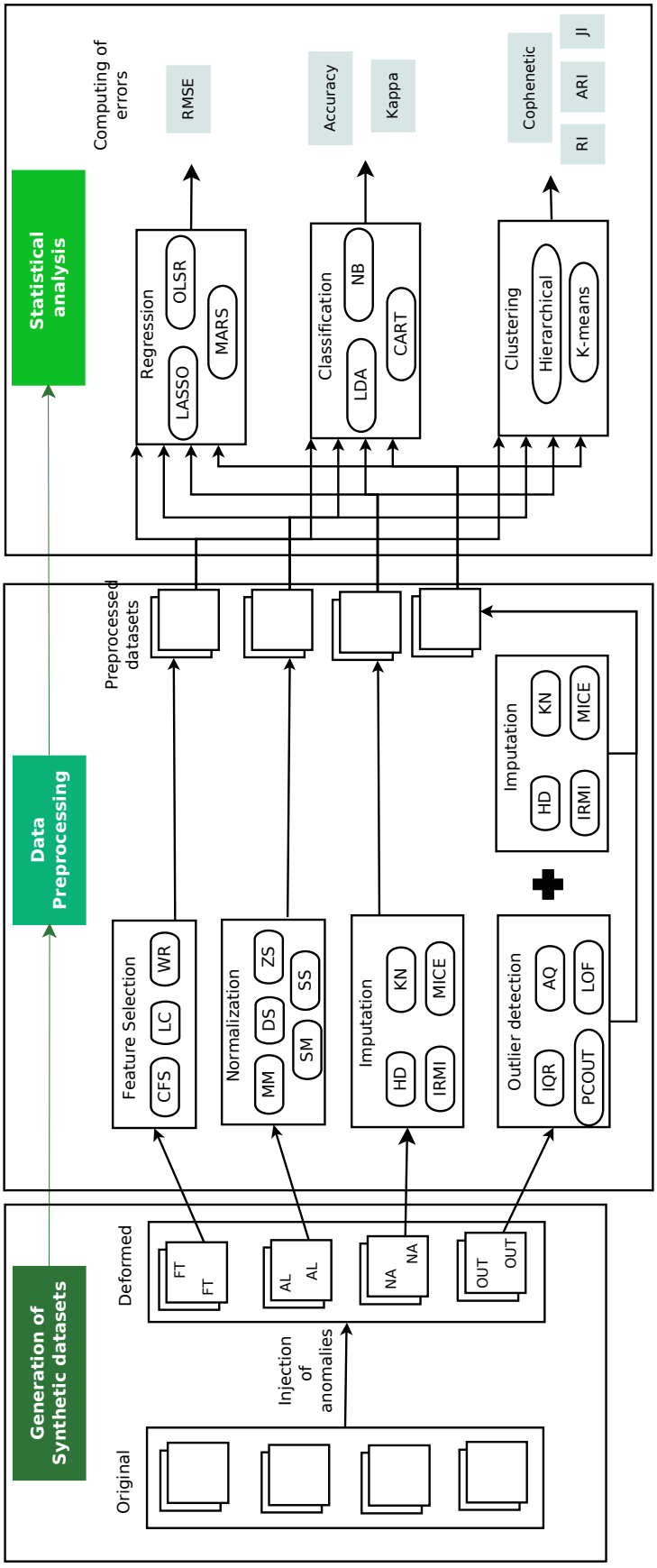


Figure 7.3 – Procédure générale d'évaluation des procédures de pré-traitement des données sur les résultats statistiques.

7.4.2 Pré-traitement de données

Dans la deuxième partie de notre étude, nous avons cherché à évaluer la robustesse des méthodes pour traiter les valeurs manquantes et aberrantes et, pré-traiter les données par sélection des variables et normalisation. Pour attendre les objectifs de cette étude, les données déformés, construites dans l'étape précédente, ont été pré-traités en utilisant différents méthodes (imputation des valeurs manquantes, traitement des valeurs aberrants, sélection des variables et normalisation de données).

Trois méthodes de sélection de variables ont été exécutés (Correlation-based feature selection (CFS), linear-based correlation (LC) et Wrapper subset evaluator (WR)) afin de sélectionner un sous-ensemble de données optimale. La sélection de variables sur chaque jeu de données déformée a été implémenté comme suit: tout d'abord, une variable a été utilisé comme variable indépendante, puis, en utilisant les méthodes de sélection de variables et par rapport à la variable indépendante, le meilleur sous-ensemble de données a été choisi. Le sous-ensembles de données qui en résultent ont par la suite servis à évaluer l'impact des méthodes de sélection de variables sur les résultats des analyses statistiques.

Trois méthodes de normalisation ont été implémentés (min-max, Z-score et decimal scale) sur les données numériques. Les données qui en résultent ont par la suite été utilisés pour évaluer l'impact des méthodes de normalisation sur les résultats des analyses statistiques.

Les valeurs manquantes dans chaque jeux de données déformés ont été traités en utilisant quatre différents méthodes d'imputation (*Hot-deck, K-NN, mice* et *IRMI*). Les jeux de données imputés ont par la suite été utilisés pour évaluer l'impact des méthodes d'imputation sur les résultats des analyses statistiques.

Les données déformés par des valeurs aberrantes ont été traités tout d'abord, en utilisant quatre différents méthodes de détection de valeurs aberrantes (Inter Quartil Range (IQR), Adjusted-Quantile, Principal Components decomposition (PCOUT) et Local Outlier Factor (LOF)). Par la suite, les valeurs aberrantes ont été remplacés en utilisant des méthodes d'imputation (*Hot-deck, K-NN, mice* et *IRMI*). Les données pré-traitées ont été utilisées pour évaluer l'impact du traitement des valeurs aberrants sur les résultats des analyses statistiques.

7.4.3 Analyse statistique

Pour évaluer l'impact des procédures de pré-traitement des données différents méthodes statistiques ont été appliquées aux données pré-traitées. Les méthodes statistiques utilisées ont été choisis car ils sont fréquemment utilisées dans des études environnementaux. La comparaison des résultats statistiques sur chaque jeu de donnée pré-traitée a été effectuée en calculant différents erreurs statistiques. Par simplicité, chaque anomalie de données a été étudiées en les considérant comme cas isolées. Les anomalies sur les données qui occurrent simultanément dans un jeu de données n'as pas fait l'objet d'étude dans cette thèse.

Trois méthodes de régression ont été implémentés sur les jeux de données pré-traitées. Ils incluent :

- Penalized regression : Leas Absolute Shrinkage and Selection Operator (LASSO) ;
- Linear regression : Ordinary Least Squares Regression (OLSR) ;
- Non-Linear regression : Multivariate Adaptive Regression Splines (MARS).

L'analyse par régression sur chaque jeu de donnée a été implémentée de la manière suivante :

1. Les données ont été divisées de manière aléatoire en deux sous-ensemble de données nommées training et test où 66% des données ont été utilisé pour training et 34% pour test ;
2. Le modèle de régression a été calculé sur le sous-ensemble de données training. Par la suite, le modèle ajusté a été utilisé pour prédire les réponses pour les observations dans le jeu de données test ;
3. Le résultat de la validation a été évalué en calculant le RMSE ;
4. Enfin, l'erreur de pré-traitement a été calculé en comparant la valeur de RMSE du jeu de données non-deformé avec celle du jeu de données pré-traitée comme suit:

$$Error_{RMSE_{processing}} = \frac{RMSE_{Preprocessed} - RMSE_{original}}{RMSE_{original}} * 100\% \quad (7.1)$$

Les trois méthodes de classification qui ont été utilisées sont :

- Linear classification : Linear discriminant analysis (LDA) ;
- Non-Linear classifcation : Naïve Bayes (NB) ;
- Non-Linear classifcation with Regression Trees : Classification and Regression Trees (CART)

L'analyse de classification sur chaque jeu de données pré-traité a été exécutée dans quatre étapes :

1. Les données ont été divisées en deux sous-ensemble de données où 66% des données ont été utilisées pour training et 34% pour test ;
2. Le modèle de classification a été construit sur le sous-ensemble training, le modèle obtenue a par la suite été utilisé pour prédire les réponses pour les observations dans le sous-ensemble test ;
3. Le résultat de la classification a été évalué en calculant l'exactitude et le coefficient de Cohen's Kappa ;
4. Finalement, l'erreur de pré-traitement a été calculé en utilisant les formules suivantes :

$$Absolute\ Error_{Accuracy_{processing}} = |Accuracy_{Preprocessed} - Accuracy_{original}| \quad (7.2)$$

$$Absolute\ Error_{Kappa_{processing}} = |Kappa_{Preprocessed} - Kappa_{original}| \quad (7.3)$$

L'analyse de groupement a été implémenté en utilisant les méthodes K-means et regroupement hiérarchique sur les données pré-traitées. La méthode K-means a été implémenté en spécifiant tout d'abord le nombre de clusters K à l'aide de la méthode de coude. La méthode de coude à d'abord été appliquée à l'ensemble de données originaux non déformées et le nombre de clusters trouvé a été utilisé sur les données pré-traitées. La méthode de regroupement hiérarchique a été exécutée à l'aide de méthode d'agglomération de Ward.

Les résultats de regroupement de la méthode K-means de données pré-traitées ont été comparés contre les résultats de regroupement de données originaux non-déformées à l'aide des indices Rand, Adjusted Rand et Jaccard (Meila, 2007).

Concernant les analyses par regroupement hiérarchique les résultats des données pré-traitées ont été comparées contre ceux de données non-déformées à l'aide du coefficient Cophenetic de corrélation (Sokal and Rohlf, 1962).

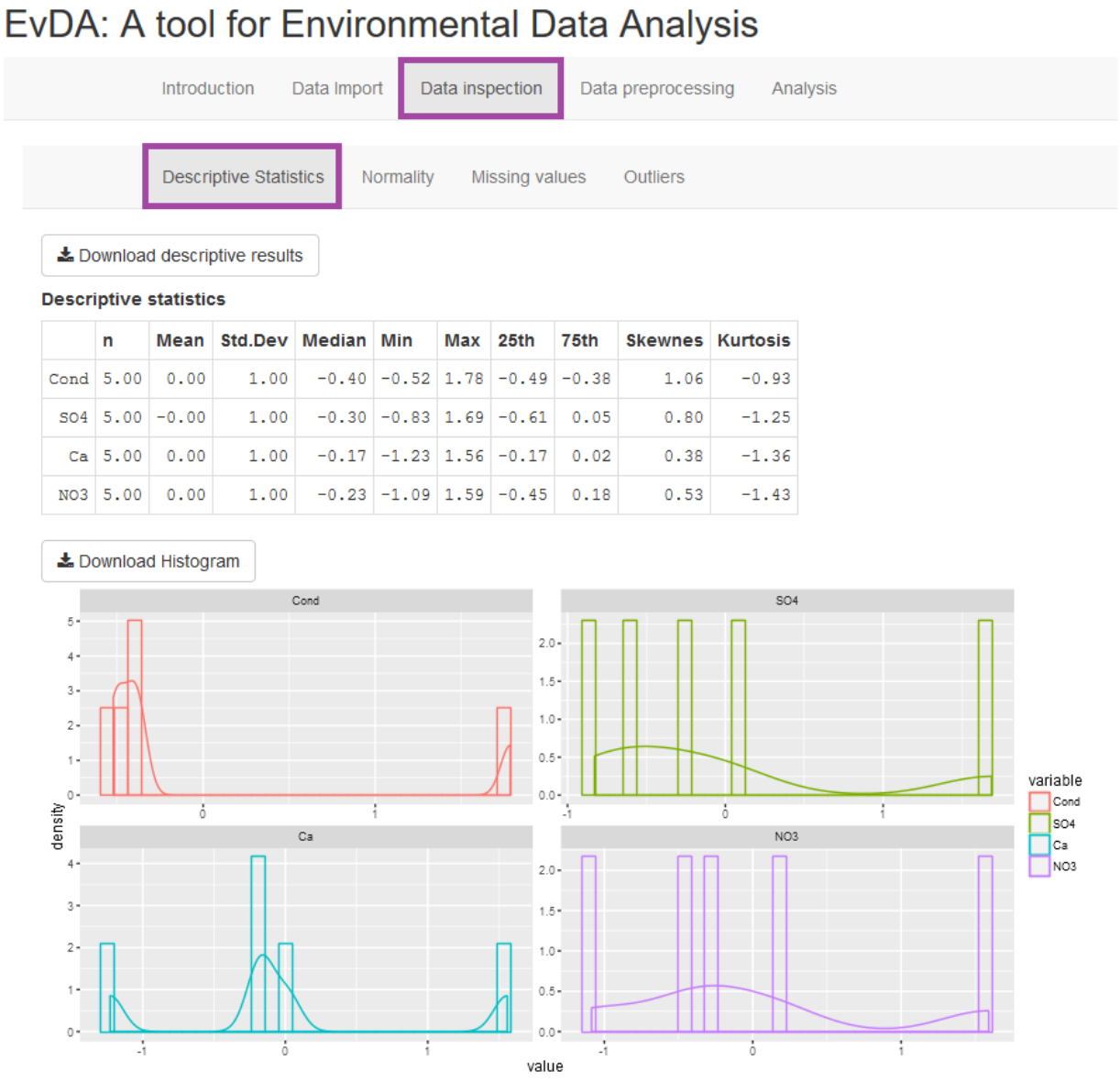
L'approche méthodologique pour évaluer l'impact des procédures de pré-traitement de données sur les résultats des analyses statistiques nous a permis d'identifier les meilleurs méthodes de pré-traiter de données pour chaque une de méthode d'analyse statistique étudiée. Nous avons découvert qu'il n'y a pas un procédure de prétraitement de données universel puisqu'il dépend des caractéristiques de données et termes de taille, distribution, Skewness, Kurtosis, etc. Les résultats de cette étude ont été utilisés pour établir un ensemble de règles permettant de combiner de façon optimale les procédures de pré-traitement et d'analyse de données. Cette règles par la suite ont été utilisés sur la construction d'un environnement analytique intégré sour la forme d'une application développée en R pour l'analyse statistique des données environnementales en général et l'analyse de la qualité de l'eau en particulier. L'étude réalisée ici a utilisé des données multivariée avec une distribution normale et sur quatre procédures pour préparer les données (normalisation et sélection des variables) et traiter des anomalies (valeurs manquantes et valeurs aberrantes). En conséquence, l'ensemble de règles qui ont été obtenus n'est pas exhaustive, mais il pourrait être utilisé comme base et encore plus important, il peut être complété par l'étude d'autres type des données, des anomalies de données et procédures de pré-traitement de données.

7.5 Développement

Un prototype nommée EvDa est présenté ici, il vise à inspecter, pré-traiter et analyser les données environnementales d'un manière plus facile. L'utilisateur peut télécharger ses propres données. Les procédures de pré-traitement peuvent être exécutées séquentiellement et l'analyse de données telles que : régression, classification, groupement ou ACP produisent des résultats utiles pour les experts du domaine à l'aide d'une interface graphique facile à utiliser.

EvDA se compose de quatre étapes principales : téléchargement, inspection, prétraitement et analyse des données. L'utilisateur peut cliquer sur les différents onglets pour se déplacer entre chaque étape. Les différents paramètres sont insérés par l'utilisation de widgets tels que des menus déroulants, case, etc. A chaque étape, l'utilisateur peut définir les paramètres et les résultats sont calculés et montrés de manière automatique (c.f. Figure 7·4).

- *Data input* : L'utilisateur peut télécharger des fichiers en format .csv, .txt ou .xls. Dans certain cas l'utilisateur peut être intéressé en étudier un sous-ensemble de données, dans ce cas l'utilisateur peut sélectionner les colonnes et lignes et utiliser le sous-ensemble de données comme nouveaux jeu de données. EvDA utilise uniquement des données numériques et catégorielles.
- *Inspection de données*: L'inspection de données est implémenté en utilisant des analyses statistiques descriptives, des tests de normalité, des analyses de valeurs manquantes et détection des valeurs aberrants. L'utilisateur peut sélectionner parmi les différents méthodes de normalisation et détection des valeurs aberrants.
- *Pré-traitement de données*: Différents procédures de pré-traitement de données sont proposés y compris: sélection des variables pertinents, normalisation, imputation des valeurs manquantes, et traitement des valeurs aberrants. L'utilisateur peut choisir parmi les différentes méthodes qui sont disponibles pour chaque tâche de prétraitement. Ils peuvent être effectués dans le jeu de données d'entrée original ou dans un jeu de données qui a résulté d'une tâche précédente de pré-traitement.
- Analyse de données: EvDA inclus des analyses par régression, classification, groupement et ACP.
  Pour une anlayse de régression, les variables indépendantes et dépendantes peu-



	n	Mean	Std.Dev	Median	Min	Max	25th	75th	Skewnes	Kurtosis
Cond	5.00	0.00	1.00	-0.40	-0.52	1.78	-0.49	-0.38	1.06	-0.93
SO4	5.00	-0.00	1.00	-0.30	-0.83	1.69	-0.61	0.05	0.80	-1.25
Ca	5.00	0.00	1.00	-0.17	-1.23	1.56	-0.17	0.02	0.38	-1.36
NO3	5.00	0.00	1.00	-0.23	-1.09	1.59	-0.45	0.18	0.53	-1.43

Figure 7.4 – Capture d'écran de l'application EvDA. Affichage de l'onglet inspection de données.

vent être spécifiés par l'utilisateur. Les modèles sont construits en créant des sous-ensemble de données pour training et test. La précision du modèle est ensuite calculée et les résultats peuvent être sauvegardés.
De manière similaire l'analyse par classification est effectuée en utilisant un sous-ensemble de données. La prédiction peut être faite en utilisant le modèle qui en résulte et l'utilisateur peut calculer l'exactitude et la valeur de Kappa.
L'analyse de regroupement par K-means est exécuté en spécifiant le nombre de clusters k. Les résultats sont affiches dans un graphique où l'utilisateur peut sélectionner le nombre de clusters k. Les résultats de la méthode de groupement hiérarchique sont fournis de manière graphique où les observations sont représentés de manière arborescente dans un dendrogramme.
La matrice de corrélation est implémenté en utilisant tous les variables, les résultats sont composés d'une liste de variables le plus corrélés et d'une représentation graphique de la matrice de corrélation.
L'ACP est construit en utilisant uniquement des données numériques. Les résultats de l'ACP sont représentés graphiquement pour les variables et les observations, où les deux premières composantes principales (Axe 1 et Axe 2) peuvent être visualisés. Les résultats numériques de l'ACP et les graphiques peuvent être sauvegardé.

Les approches méthodologiques développés en chimie de l'environnement, hydrobiologie et science des données ont par la suite été utilisées pour évaluer la pollution de l'eau dans quatre rivières mexicaines (Tula, Tamazula, Humaya et Culiacan). Les résultats des analyses sur les données du Mexique montrent que les activités d'agriculture, les déchets urbains et industrielles polluent l'eau des rivières étudiés. Les résultats observés indiquent que des mesures de protection et remédiation sont nécessaires pour améliorer et protéger la qualité des écosystèmes aquatiques dans les rivières mexicaines Tula, Tamazula, Humaya et Culiacan.

7.6 Conclusions

Les contributions multidisciplinaire de la thèse sont : (1) en chimie de l'eau : une procédure méthodologique permettant de déterminer les quantités de pesticides organochlorés dans des échantillons d'eau collectés sur le terrain en utilisant les techniques SPE-GC-ECD (Solid Phase Extraction – Gas Chromatography – Electron Capture Detection) ; (2) en hydrobiologie : une procédure méthodologique pour évaluer la qualité de l'eau dans quatre rivière Mexicaines en utilisant des indicateurs biologiques basés sur des macroinvertébrés ; (3) en science des données : une méthode pour évaluer et guider le choix des procédures de pré-traitement des données produites lors des deux précédentes étapes ainsi que leur analyse ; en fin (4) le développement d'un environnement analytique intégré sous la forme d'une application développée en R pour l'analyse statistique des données environnementales en général en l'analyse de la qualité de l'eau en particulier. Enfin, nous avons appliqué nos propositions sur le cas spécifique de l'évaluation de la qualité de l'eau des rivières Mexicaines Tula, Tamazula, Humaya et Culiacan dans le cadre de cette thèse qui a été menée en partie au Mexique et en France.

APPENDIX A

Biotic indices

Trend Biotic Index (Metcalfe, 1989) The organisms used for the computation of index are identified at family, genus or species level depending on the type of organism. The index values go from 0 to 10, with 10 representing clean streams, this value decreases with increasing pollution. In Table A.1, are represented the key groups and the Biotic index values necessary to estimate the Trent Biotic Index of a given sample. It is a two-entrance table: vertical entrance corresponds to the value of richness found and the horizontal entrance to the less tolerant organism.

Table A.1 – Trent Biotic Index (TBI). Key groups for the estimation of the Trent River Board Biotic Index and Biotic Index values related to the total number of groups present in a sample (Table taken from Metcalfe, 1989).

A group consists of :			Common name				
Each family of Trichoptera larvae			Caddis flies				
Each family of Coleoptera larvae and adults			Beetles				
Each family of Diptera (except blood worms)			Trues flies				
Each family of Annelida Oligochaeta			Worms				
Each genus of Plecoptera nymphs			Stoneflies				
Each genus of Ephemeroptera nymphs			May-flies				
Each species of Annelida Hirundinea			Leeches				
Each species of Mollusca			Sanils, limpets, etc.				
Each species of Crustacea			Shrimps, water hoglice				
Each species of Megaloptera larvae			Alder flies				
Chironomus thumni			Blood worms				
Trent River Board Biotic Index							
			Total number of groups present				
			0-1	2 – 5	6 – 10	11 – 15	16
			Biotic Index				
Clean	Plecoptera Nymphs Present	More than one species	-	VII	VIII	VIII	IX
		One species only	-	VI	VII	VIII	IX
Organisms in order of tendency to disappear as degree of pollution increase	Ephemeroptera nymphs present (excluding Baetis)	More than one species	-	VI	VII	VIII	IX
		One species only	-	V	VI	VII	VIII
	Trichoptera larvae or Baetis present	More than one species	-	V	VI	VII	VIII
		One species only	IV	IV	V	VI	VII
	Gammarus present	All above species absent	III	IV	V	VI	VII
	Asellus present	All above species absent	III	IV	V	VI	VII
	Tubificid worms and/or red Chironomid larvae present	All above species absent	I	II	III	IV	-
Polluted	All above species absent	Some organisms such as Eristalis tenax not requiring dissolved oxygen may be present	-	I	II	-	-

Extended Biotic Index (EBI) (Ghetti, 1997)

EBI values are determined considering the abundance of specific macroinvertebrate taxa which is ordered according to its tolerance to stress factors. EBI values go from 0 to 14. They are calculated using a two-entrance table (c.f. Table A.2): a vertical entrance – corresponding to the value of Richness found; and a horizontal entrance – corresponding to the less tolerant Systematic Units (SU) to stress factors. To determine the ecological quality of aquatic ecosystems, EBI values are calculated and converted into "quality classes". The organisms used to compute the Extended Biotic Index (EBI), are identified at family or genus level.

Table A.2 – Table to calculate Extended Biotic Index (EBI) values and conversion table to transform EBI values into Quality Classes (Ghetti, 1997). SU: Number of Systematic Units observed of the taxonomic group.

Faunal Groups which determine their presence to the horizontal inlet in table (horizontal input)		Total Number of Systematic Units constituting the community (vertical input)								
		0 – 1	2 – 5	6 – 10	11 – 15	16 – 20	21 – 25	26 – 30	31 – 35	36 - . . .
Plecoptera present (Leuctra)	More than one SU	-	-	8	9	10	11	12	13	14
	Only one SU	-	-	7	8	9	10	11	12	13
Ephemeroptera present (excluding Baetidae and Caenidae)	More than one SU	-	-	7	8	9	10	11	12	-
	Only one SU	-	-	6	7	8	9	10	11	-
Trichoptera present (Baetidae and Caenidae included)	More than one SU	-	5	6	7	8	9	10	11	-
	Only one SU	-	4	5	6	7	8	9	10	-
Gammarus	All the SU above absent	-	4	5	6	7	8	9	10	-
Asellus	All the SU above absent	-	3	4	5	6	7	8	9	-
Oligochaeta and/or Chironomid	All the SU above absent	1	2	3	4	5	-	-	-	-
All above organism	All the SU above absent	0	1	2	3	-	-	-	-	-

Conversion table of values of EBI in Quality Classes			
Quality class	EBI value	Quality judgment	Color relating to the quality class
Class I	10 – 11 – 12 - . . .	Environment is not altered in a sensitive manner	Blue
Class II	8 – 9	Environment with moderate symptoms of alteration	Green
Class III	6 – 7	Environment altered	Yellow
Class IV	4 – 5	Environment very altered	Orange
Class V	0- 1 – 2 – 3	Environment highly degraded	Red

Beck Biotic Index (Beck, 1955)

To calculate this index, macroinvertebrates are counted and organized according to the three classifications proposed by Beck (c.f. Table A.3). The three classes categorize macroinvertebrates according to their organic pollution tolerances. They are defined as follows: Class I Organisms (Sensitive or Intolerant): organisms that exhibit rapid response to aquatic environmental degradations and are reduced in number. Class II Organisms (Facultative): organisms that are capable to survive under polysaprobic conditions. Class III Organisms (Tolerant): organisms that have high resistance to adverse conditions within the aquatic environment. Once the organisms have been counted and classed the Beck's Biotic Index is calculated using the following expression:

$$BI = 2n1 + n2(7) \tag{A.1}$$

Where BI is the Beck's Biotic Index n1 is the number of Class I organisms identified n2 is the number of Class II organisms identified Index values range from 0 to about 40 and, at lower values of the index, the organic pollution is greater.

Table A.3 – Benthic macroinvertebrates clased according to Beck's biotic Index (BBI) Classes (Beck, 1955).

Invertebrate form	Class	Invertebrate form	Class
Caddisflies : Trichoptera		Crayfish : Crustacea	
Hydropsychidae	1	Astacidae	2
Hydroptilidae	1	Flatworms : Turbellaria	
Limnephilidae	1	Planaridae	2
Leptoceridae	1	Crane Flies : Diptera	
Helicopsychiade	1	Tipulidae	2
Psychomyiidae	1	Gill Snails : Mollusca	
Goeridae	1	Pleuroceridae	2
Stoneflies : Plecoptera		Horse Flies : Diptera	
Perlidae	1	Tabanidae	2
Perlodidae	1	Isopods : Crustacea	
Mayflies : Ephemeroptera		Asellidae (Aquatic Sowbugs)	2
Baetidae	1	Blackflies : Diptera	
Heptageniidae	1	Simuliidae	2
Ephemeridae	1	Air Breathing Snails : Mollusca	
Helligrammites : Megaloptera		Physidae	3
Corydalidae	1	Ancylidae (Limpets)	3
Freshwater Naiads (Clams) : Bivalvia		Aquatic Earthworms : Annelida	
Unionidae	1	Oligochaeta	3
Beetles : Coleoptera		Midges : Diptera	
Elmidae (Riffle Beetle)	1	Chironomidae	3
Psephenidae (Water Penny)	1	Leeches : Annelida	
Damselflies :Odonata zygoptera		Hirundinea	3
Coenagrionidae	2	Moth flies : Diptera	
Agrionidae	2	Psychodidae	3
Dragonflies : Odonata Anisoptera			
Aeschnidae	2		
Gomphidae	2		
Libellulidae	2		

Family Biotic Index (Hilsenhoff, 1988)

It includes other macroinvertebrates than arthropods and, uses family-level tolerance values (Hilsenhoff, 1988, Plafkin et al., 1989). Organisms are identified at family level. Tolerance values are assigned from 0 for organisms very intolerant to organic pollution to 10 for organisms very tolerant to organic pollution (c.f. Table A.4). FBI is computed using the following expression:

$$FBI = \sum x_i t_i / n \tag{A.2}$$

where xi is the number of individuals in the i_{th} taxon t_i is the tolerance value of the i_{th} taxon n is the total number of organisms in the sample.

Table A.4 – Evaluation of water quality using the Family Biotic Index (FBI) and the tolerance values for families of stream arthropods (Hilsenhoff, 1988).

Plecoptera		Trichoptera		Amphipoda	
Capniidae	1	Brachycentridae	1	Gammaridae	4
Chloroperlidae	1	Glossosomatidae	0	Talitridae	8
Leuctridae	0	Helicopsychidae	3		
Nemouridae	2	Hydropsychidae	4	Isopoda	
Perlidae	1	Hydroptilidae	4	Asellidae	8
Perlodidae	2	Lepidostomatidae	1		
Pteronarcyidae	0	Leptoceridae	4	Megaloptera	
Taeniopterygidae	2	Limnephilidae	4	Corydalidae	0
		Molannidae	6	Sialidae	4
Ephemeroptera		Odontoceridae	0		
Baetidae	4	Philpotamidae	3	Lepidoptera	
Baetiscidae	3	Phryganeidae	4	Pyralidae	5
Caenidae	7	Polycentropodidae	6		
Ephemerellidae	1	Psychomyiidae	2	Coleoptera	
Ephemeridae	4	Rhyacophilidae	0	Dryopidae	5
Heptageniidae	4	Sericostomatidae	3	Elmidae	4
Leptophlebiidae	2			Psephenidae	4
Metretopodidae	2	Diptera			
Oligoneuriidae	2	Athericidae	2		
Polymitarcyidae	2	Blephariceridae	0		
Potomanthidae	4	Ceratopogonidae	6		
Siphlonuridae	7	Blood-red Chironomidae (Chironomini)	8		
Tricorythidae	4	Other Chironomidae (including pink)	6		
		Dolichopodidae	4		
Odonata		Empididae	6		
Aeshnidae	3	Ephydridae	6		
Calopterygidae	5	Muscidae	6		
Coenagrionidae	9	Psychodidae	10		
Cordulegastridae	3	Simuliidae	6		
Corduliidae	5	Syrphidae	10		
Gomphidae	1	Tabanidae	6		
Lestidae	9	Tipulidae	3		
Libellulidae	9				
Macromiidae	3				

Family Biotic Index	Water Quality	Degree of Organic Pollution
0,00 – 3,75	Excellent	Organic pollution unlikely
3,76 – 4,25	Very good	Possible slightly organic pollution
4,26 – 5,00	Good	Some organic pollution probable
5,01 – 5,75	Fair	Fairly substantial pollution likely
5,76 -6,50	Fairly poor	Substantial pollution likely
6,51 – 7,25	Poor	Very substantial pollution likely
7,26 – 10,00	Very poor	Severe organic pollution likely

Biological Monitoring Working Party and Average Score per Taxon (National Water Council (1981), Armitage et al. (1983))

To calculate the BMWP, all the individual scores of all families present in a sample are summed. High BMWP values are characteristic of clean sites, while low BMWP values are typical of polluted sites (c.f. Table A.5).

Table A.5 – Biomonitoring Working Party Score System (BMWP) (National Water Council; 1981)

Families	Score
Siphlonuridae Heptageniidae Leptophlebiidae Ephemerellidae	10
Potamanthidae Ephemeridae	
Taeniopterygidae Leuctridae Capniidae Perlodidae Perlidae	
Chloroperlidae	
Aphelocheiridae	
Phryganeidae Molannidae Beraeidae Odontoceridae	
Leptoceridae Goeridae Lepidostomatidae Brachycentridae	
Sericostomatidae	
Astacidae	8
Lestidae Agriidae Gomphidae Cordulegasteridae Aeshnidae	
Corduliidae Libellulidae	
Psychomyiidae Philopotamidae	
Caenidae	7
Nemouridae	
Rhyacophilidae Polycentropodidae Limnephilidae	
Neritidae Viviparidae Ancylidae	6
Hydroptilidae	
Unionidae	
Corophiidae Gammaridae	
Platycnemididae Coenagriidae	
Mesoveliidae Hydrometridae Gerridae Nepidae Naucoridae	5
Notonectidae Pleidae Corixidae	
Haliplidae Hygrobiidae Dytiscidae Gyrinidae	
Hydrophilidae Crambidae Helodidae Dryopidae Elimidae	
Chrysomelidae Curculionidae	
Hydropsychidae	
Tipulidae Simuliidae	
Planariidae Dendrocoelidae	
Baetidae	4
Sialidae	
Piscicolidae	
Valvatidae Hydrobiidae Lymnaeidae Physidae Planorbidae	3
Sphaeriidae	
Glossiphoniidae Hirudidae Eropobdellidae	
Asellidae	
Chironomidae	2
Oligochaeta (whole class)	1

Average Score Per Taxon (ASPT) (Armitage et al., 1983). ASPT is calculated by dividing the BMWP score by the total number of contributing taxa. High ASPT values characterize clean sites with relatively large numbers of high scoring taxa and low ASPT values are distinctive of polluted sites that do not support many high scoring taxa (c.f. Table A.6).

Table A.6 – Biological Monitoring Working Party (BMWP) and Average Score Per Taxon (ASPT) scores and their related quality index (Armitage et al., 1983; Friedrich et al., 1996; National Water Council, 1981).

BMWP score	ASPT score	Category	Interpretation
0 – 10	3,6 or less	Very poor	Heavily polluted
11 – 40	3,61 – 4,2	Poor	Polluted or impacted
41 – 70	4,21 – 4,80	Moderate	Moderately impacted
71 – 100	4,81 – 5,4	Good	Clean but slightly impacted
>100	Over 5,4	Very good	Unpolluted, unimpacted

Functional Feeding Groups measures Six FFG measures are described in here, including filtering collectors, scrapers, shredders, predators and gathering collectors. The Functional feeding group approaches are based on the behavioural and feeding mechanisms by which macroinvertebrates obtain their food resources (c.f. Table A.7) (Merrit et al., 2008).

Table A.7 – Functional Feeding Groups (FFG): Categorization and food resources. Coarse Particulate Organic Matter (CPOM); Fine Particulate Organic Matter (FPOM) (Merrit et al., 2008).

Functional groups	Particle size feeding mechanisms	Dominant food ressources	Particle size range of food (mm)
Shredders	Chew conditioned letter or live vascular plant tissue, or gouge wood	CPOM-decomposing (or living hydrophyte) vasculat plants	>1,0
Filtering collectors	Suspension feeders : filter particles form the water column	FPOM-decomposing detrital particles ; algae, bacteria, and feces	0,01 – 1,0
Gathering collectors	Deposit feeders : ingest sediment or gather loose particles in depositional areas	FPOM-decomposing detrital particles ; algae, bacteria, and feces	0,05 – 1,0
Scrapers	Graze rock and wood surfaces or stems of rooted aquatic plants	Periphyton-attached, non-filamentous algae and associated detritus, microflora and fauna, and feces	0,01 – 1,0
Predators	Catpure and engulf prey or tissue, ingest body fluids	Prey-living animal	>0,5

Macroinvertebrate-based Index of Biotic Integrity (IBI designed for west-central Mexico streams) (Weigel et al., 2002).

The Macroinvertebrate-based Index of Biotic Integrity (IBI) designed for west-central Mexican streams was computed. It comprises eight metrics (catch per unit effort, generc richness, % EPT genera, % Chironomidae individuals, Hilsenhoff Biotic Index, % depositional individuals, % predator individuals and % gatherer genera). Macroinvertebrate IBI score was calculated by summing component metric scores (c.f. Table A.8 and Table A.9). The final IBI score is then used to qualify the stream. IBI classifies streams into five classes (very poor, poor, fair, good and very good). Low IBI scores correspond to streams with very poor quality.

Table A.8 – Scoring criteria of the Macroinvertebrate-based Index of Biotic Integrity (IBI designed for west-central Mexico streams) (Weigel et al., 2002).

Metric	Poor(0)	Fair(5)	Good(10)
Catch per unit effort (CPUE)	≤50	-	>50
Generic richness (GR)			
Basin area ≤400Km²	≤13	14 - 22	≥23
Basin area > 400 Km²	≤11	-	≥12
% EPT			
Basin area ≤400 Km²	<32	32 - 38	> 38
Basin area > 400 Km²	<35	35 - 55	> 55
% Chironomidae individuals (% Midge)	> 25	5 - 25	< 5
Hilsenhoff biotic Index (HBI)	> 5	4.25 - 5	< 4.25
% depositional individuals (%Depo)	> 75	55 - 75	< 55
% predator individuals (%Pred)	< 4	4 - 14	> 14
%gatherer genera (%Gath)	> 48	44 - 48	< 44

Table A.9 – Macroinvertebrates-based Index of Biotic Integrity (IBI) quality values and their related biological responses to environmental conditions (Weigel et al., 2002).

Value and qualitative rating	Biological response to environmental condition
75 - 80 Very good	Comparable to the minimum influence system in the region. Macroinvertebrates are abundant. GR and %EPT are near the maximum for the size of stream. Chironomids are absent or %Midge is very low. Low HBI scores and % Depo indicate no organic pollution and sedimentation are undetectable. High % Pred and low % Gath indicate a predominance of specialized feeders. Non-point-source pollution is minimum or distant from the site.
60 - 70 Good	Generally low levels of non-point-source pollution have influences some aspects of the macroinvertebrate assemblage. GR and % EPT are usually high but may not be at the maximum for the size of the stream. Often % Midge, % Depo and % Gath increase. Some organic pollution may be reflected in the HBI scores.
50 -55 Fair	Point-source pollution may be present distant from the site or in low quantities. Moderate to severe non-point-source pollution or diversion for irrigation are typically the major stressors. GR and % EPT are moderate to low for the size of stream. Odonates and other depositional taxa become more prevalent. HBI scores and % Midge typically run moderate to high.
25 - 45 Poor	Point-source pollution is generally present but is intermittent or not immediately at the site. Non-point-source pollution can be severe. Abundance is typically low but, if it is high, is due to a high % Midge. HBI scores and % Depo suggest very tolerant assemblages.
0 - 20 Very poor	System is severely degraded by point-source pollution. Zero values can result from all water being diverted for irrigation or collecting < 100 individuals in 4 CPUEs. Chironomids are typically the only macroinvertebrate fauna present and are usually abundant. If other taxa are present, their abundance is low.

APPENDIX B

Synthetic and semi-synthetic datasets

B.1 Datasets for assessment of feature selection

Table B.1 – N21 synthetic dataset ($n = 21$, $p = 8$) used to assess feature selection methods. The μ and σ^2 values of all variables are shown. Highly correlated variables are denoted as Y_k

	Variables							
	X1	X2	X3	X4	X5	Y1	Y2	Y3
μ	321.72	181.80	518.91	734.27	714.80	775.90	619.09	753.99
σ^2	1.11	1.12	0.87	1.43	1.20	0.84	0.86	0.74

Table B.2 – N600 synthetic dataset ($n = 600$, $p = 30$) used to assess feature selection methods. The μ and σ^2 values of all variables are shown. Highly correlated variables are denoted as Y_k

	Variables									
	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
μ	511.01	32.00	69.99	238.07	159.08	125.98	350.98	578.01	281.04	445.00
σ^2	0.95	1.06	1.06	1.02	1.01	1.03	0.91	1.05	0.96	0.97
	X11	X12	X13	X14	X15	X16	X17	X18	X19	X20
μ	243.03	472.04	57.99	127.05	9.01	407.00	78.06	27.01	581.05	260.94
σ^2	0.97	1.06	1.06	1.07	0.96	1.18	0.97	0.98	0.93	1.11
	X21	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8	Y9
μ	27.01	488.98	53.99	203.95	45.99	72.00	65.03	239.95	196.99	92.02
σ^2	0.98	0.94	0.96	1.06	1.04	0.99	1.05	0.98	1.01	1.01

Table B.3 – N4000 synthetic dataset ($n = 4000$, $p = 53$) used to assess feature selection methods. The μ and σ^2 values of all variables are shown. Highly correlated variables are denoted as Y_k.

	Variables									
	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
μ	769.98	155.98	663.00	19.01	720.99	684.00	467.02	599.98	477.00	326.01
σ^2	0.98	0.99	1.01	0.97	0.99	1.02	1.00	0.99	0.99	1.00
	X11	X12	X13	X14	X15	X16	X17	X18	X19	X20

μ	255.02	816.01	164.99	406.02	422.99	702.99	473.99	299.01	810.99	551.97
σ^2	1.05	1.00	0.99	1.04	1.02	0.99	0.99	1.00	0.95	1.00
	X21	X22	X23	X24	X25	X26	X27	X28	X29	X30
μ	865.99	891.00	69.99	508.01	650.99	512.98	668.99	611.99	363.00	361.99
σ^2	0.98	0.99	1.02	1.01	1.03	0.99	0.98	1.01	0.98	0.99
	X31	X32	X33	X34	X35	X36	X37	Y1	Y2	Y3
μ	294.98	593.97	599.01	427.99	604.98	745.99	726.00	665.00	28.00	248.99
σ^2	1.01	1.02	1.00	1.01	1.01	1.01	1.00	0.99	0.96	1.00
	Y4	Y5	Y6	Y7	Y8	Y9	Y10	Y11	Y12	Y13
μ	352.01	166.00	385.00	835.00	695.99	420.02	656.01	145.00	504.01	334.00
σ^2	0.97	0.99	1.01	0.99	0.99	0.99	1.00	1.01	0.97	0.98
	Y14	Y15	Y16							
μ	512.00	153.01	222.00							
σ^2	0.98	1.00	1.02							

Table B.4 – N20000 synthetic dataset ($n = 20000$, $p = 98$) used to assess feature selection methods. The μ and σ^2 values of all variables are shown. Highly correlated variables are denoted as Y_k.

	Variables									
	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
μ	23.00	142.00	746.01	351.00	459.99	270.00	55.99	132.00	804.99	783.01
σ^2	0.99	0.99	0.99	1.00	1.00	1.00	1.00	0.99	1.01	1.00
	X11	X12	X13	X14	X15	X16	X17	X18	X19	X20
μ	843.01	514.00	77.00	220.02	706.00	688.99	170.00	48.00	645.01	620.99
σ^2	1.00	1.02	1.00	0.99	1.01	1.02	0.98	1.01	1.00	0.99
	X21	X22	X23	X24	X25	X26	X27	X28	X29	X30
μ	735.00	562.99	692.01	158.01	260.00	528.00	653.00	663.99	117.01	268.00
σ^2	1.00	1.01	1.00	1.01	1.01	0.99	1.00	0.99	1.01	1.00
	X31	X32	X33	X34	X35	X36	X37	X38	X39	X40
μ	388.00	257.01	688.00	362.00	234.00	35.00	158.00	757.01	303.99	189.99
σ^2	1.01	1.00	1.00	0.99	1.01	1.00	1.02	0.99	1.00	1.00
	X41	X42	X43	X44	X45	X46	X47	X48	X49	X50
μ	687.99	495.01	627.00	664.00	514.99	103.01	545.01	384.01	635.01	701.00
σ^2	1.02	1.00	0.99	1.00	1.01	0.99	0.99	1.00	1.00	1.00
	X51	X52	X53	X54	X55	X56	X57	X58	X59	X60
μ	368.00	229.00	611.00	500.00	304.00	260.00	729.01	360.00	241.01	397.01
σ^2	1.00	1.00	0.99	1.01	1.01	0.99	1.00	0.99	0.99	0.99
	X61	X62	X63	X64	X65	X66	X67	X68	X69	Y1
μ	190.99	825.00	606.01	461.00	603.00	510.00	632.00	144.01	793.02	783.00
σ^2	1.01	1.01	0.99	1.00	1.00	1.01	1.00	0.99	1.00	1.00
	Y2	Y3	Y4	Y5	Y6	Y7	Y8	Y9	Y10	Y11
μ	772.00	506.99	451.99	451.00	721.00	478.99	411.00	334.00	638.99	109.00
σ^2	1.00	0.99	1.01	1.00	0.99	1.00	1.00	1.00	1.00	1.00
	Y12	Y13	Y14	Y15	Y16	Y17	Y18	Y19	Y20	Y21
μ	691.00	219.00	664.00	253.99	172.99	361.99	714.99	562.00	131.00	259.00
σ^2	0.99	0.99	0.99	1.00	1.00	1.00	0.99	0.99	1.00	0.99
	Y22	Y23	Y24	Y25	Y26	Y27	Y28	Y29		
μ	0.99	273.00	507.00	58.00	823.00	11.99	319.00	588.99		
σ^2	0.99	0.99	1.00	1.00	1.00	0.99	1.00	0.99		

B.2 Datasets for assessment of normalization

Table B.5 – N21 synthetic dataset ($n = 21$, $p = 8$) used to assess normalization methods. The μ and σ^2 values of all variables are shown.

	Variables							
	X1	X2	X3	X4	X5	X6	X7	X8
μ	213.98	255.75	761.98	562.19	35.16	411.48	348.67	393.36
σ^2	41343.79	62385.58	679029.21	129683.04	1010.23	221852.32	175850.52	122081.17

Table B.6 – N600 synthetic dataset ($n = 600$, $p = 30$) used to assess normalization methods. The μ and σ^2 values of all variables are shown.

	Variables									
	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
μ	171.95	193.49	72.89	107.32	633.29	310.19	244.98	558.62	443.10	156.39
σ^2	29166.08	40522.14	5041.46	12452.19	443577.50	82307.87	53645.94	333400.03	395958.17	20831.51
	X11	X12	X13	X14	X15	X16	X17	X18	X19	X20
μ	230.44	439.67	348.19	543.69	103.32	492.18	10.71	307.91	376.14	597.09
σ^2	48509.81	196197.88	126782.85	314880.02	10408.54	261451.92	120.93	104061.63	186155.22	336488.02
	X21	X22	X23	X24	X25	X26	X27	X28	X29	X30
μ	149.90	516.38	462.82	592.14	28.44	217.84	265.07	78.54	100.16	217.36
σ^2	20573.98	243214.29	198983.84	351947.00	935.20	47683.27	63771.27	7279.55	10179.21	47442.12

Table B.7 – N4000 synthetic dataset ($n = 4000$, $p = 53$) used to assess normalization methods. The μ and σ^2 values of all variables are shown.

	Variables									
	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
μ	259.17	227.73	838.73	67.93	165.33	323.62	741.48	531.30	620.14	789.61
σ^2	66203.82	51371.52	700756.56	4462.29	28778.14	99923.59	529437.41	271397.61	392837.24	605888.40
	X11	X12	X13	X14	X15	X16	X17	X18	X19	X20
μ	61.46	425.39	210.42	307.06	592.82	398.29	458.01	418.55	869.55	323.35
σ^2	3784.41	186875.75	45683.98	96196.95	367170.28	165341.61	207590.83	177688.14	734021.93	100877.20
	X21	X22	X23	X24	X25	X26	X27	X28	X29	X30
μ	880.13	0.00	328.66	887.06	729.59	604.98	415.10	26.18	654.61	793.93
σ^2	772333.28	0.00	110325.47	806008.34	624076.28	373426.31	175900.16	677.49	421312.08	618914.70
	X31	X32	X33	X34	X35	X36	X37	X38	X39	X40
μ	359.02	233.11	531.26	400.37	422.71	887.88	119.78	54.92	415.37	537.44
σ^2	134574.58	54284.41	294972.85	154544.61	117652.38	740684.41	14064.93	3131.38	163346.27	270536.16
	X41	X42	X43	X44	X45	X46	X47	X48	X49	X50
μ	300.09	775.59	390.67	258.02	36.27	231.47	474.80	618.16	212.09	576.81
σ^2	89469.43	612330.10	144625.56	68051.31	1286.68	54297.00	218089.49	379085.74	43515.72	329587.24
	X51	X52	X53							
μ	333.99	181.94	318.28							
σ^2	107663.96	31545.49	98725.43							

Table B.8 – N20000 synthetic dataset ($n = 20000$, $p = 98$) used to assess normalization methods. The μ and σ^2 values of all variables are shown.

	Variables									
	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10

μ	256.25	225.65	839.71	67.18	168.80	319.50	710.35	532.69	632.25	779.89
σ^2	66029.69	50712.29	683255.96	4518.53	27972.67	102515.49	500528.68	283402.55	407322.39	602853.08
	X11	X12	X13	X14	X15	X16	X17	X18	X19	X20
μ	60.64	436.79	206.82	307.70	587.59	383.29	457.25	407.74	873.67	327.31
σ^2	3709.63	186963.00	43167.64	95563.30	334101.60	147077.22	203124.77	163764.91	760757.50	107135.86
	X21	X22	X23	X24	X25	X26	X27	X28	X29	X30
μ	866.89	0.00	328.75	882.44	708.47	603.77	414.02	26.90	650.98	774.83
σ^2	731462.42	0.00	111380.71	780798.88	497773.79	364936.19	169903.44	716.31	419447.40	604890.00
	X31	X32	X33	X34	X35	X36	X37	X38	X39	X40
μ	357.79	230.36	541.34	392.98	426.97	891.83	118.02	56.37	420.54	522.53
σ^2	129495.90	51850.43	286181.76	155292.08	179373.43	801893.11	13926.27	3189.79	177238.05	269304.61
	X41	X42	X43	X44	X45	X46	X47	X48	X49	X50
μ	299.35	752.47	390.65	258.32	35.87	232.53	487.54	625.89	209.88	578.61
σ^2	89715.82	562883.22	149652.82	65563.29	1306.82	53869.40	237970.89	392352.66	42475.87	340018.65
	X51	X52	X53	X54	X55	X56	X57	X58	X59	X60
μ	331.93	178.58	311.62	257.89	224.85	829.31	67.04	169.27	323.43	715.00
σ^2	111161.35	32528.65	96610.98	66740.47	51360.61	701679.56	4431.92	28968.62	104694.17	504497.08
	X61	X62	X63	X64	X65	X66	X67	X68	X69	X70
μ	533.02	628.41	798.44	61.72	438.50	208.81	306.48	594.43	382.46	461.95
σ^2	286302.17	397655.36	634633.38	3724.25	194280.28	44181.82	96660.32	352346.07	146855.25	213417.76
	X71	X72	X73	X74	X75	X76	X77	X78	X79	X80
μ	405.62	888.18	327.85	866.30	0.00	325.81	889.64	712.29	597.94	414.89
σ^2	160561.01	789606.80	106419.25	747883.97	0.00	104118.77	80371.38	515921.77	356863.20	75971.83
	X81	X82	X83	X84	X85	X86	X87	X88	X89	X90
μ	27.23	640.82	790.15	355.04	229.49	534.95	393.80	429.26	878.11	119.23
σ^2	738.39	410935.81	626179.61	124148.05	51529.17	285859.45	154502.42	188158.64	783723.16	13926.79
	X91	X92	X93	X94	X95	X96	X97	X98		
μ	55.79	415.47	531.17	298.78	752.11	396.72	257.39	35.75		
σ^2	3055.07	172393.03	278408.02	89350.34	574736.61	157206.83	66238.50	1306.13		

B.3 Datasets for assessment of imputation of missing values

Table B.9 – N21 synthetic dataset ($n = 21$, $p = 8$) used to assess imputation methods. The μ and σ^2 values of all variables are shown.

	Variables							
	X1	X2	X3	X4	X5	X6	X7	X8
μ	191.25	156.16	536.92	544.79	693.06	766.90	744.81	269.08
σ^2	1.18	1.29	1.19	0.83	0.80	0.55	0.79	0.78

Table B.10 – N600 synthetic dataset ($n = 600$, $p = 30$) used to assess imputation methods. The μ and σ^2 values of all variables are shown.

	Variables									
	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
μ	216.99	512.02	244.01	333.99	172.01	241.02	211.04	310.99	158.94	402.06
σ^2	0.92	1.00	0.99	1.03	1.01	1.01	0.96	1.05	0.96	0.91
	X11	X12	X13	X14	X15	X16	X17	X18	X19	X20
μ	59.95	400.97	406.00	206.03	116.01	230.02	499.05	488.02	24.05	283.99
σ^2	0.91	1.06	0.97	0.94	1.14	1.06	1.03	0.87	0.95	0.98
	X21	X22	X23	X24	X25	X26	X27	X28	X29	X30
μ	564.03	179.04	517.02	164.00	541.04	344.94	33.01	245.98	71.98	81.95
σ^2	1.13	0.93	1.11	1.08	0.99	0.90	0.92	1.03	1.06	0.99

Table B.11 – N4000 synthetic dataset ($n = 4000$, $p = 53$) used to assess imputation methods. The μ and σ^2 values of all variables are shown.

	Variables									
	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
μ	97.02	513.00	517.00	487.99	362.98	893.01	355.99	688.98	661.01	817.00
σ^2	1.01	1.02	1.00	0.97	1.01	1.01	1.01	1.02	0.98	0.98
	X11	X12	X13	X14	X15	X16	X17	X18	X19	X20
μ	67.02	480.96	363.01	897.00	385.99	480.01	888.02	789.00	738.99	39.99
σ^2	0.99	1.03	0.96	0.95	0.99	1.01	0.98	1.04	0.99	1.01
	X21	X22	X23	X24	X25	X26	X27	X28	X29	X30
μ	670.98	655.01	769.99	384.00	396.00	765.01	10.01	118.02	243.99	608.96
σ^2	1.05	0.99	1.02	1.03	1.02	1.01	0.98	0.98	1.01	0.99
	X31	X32	X33	X34	X35	X36	X37	X38	X39	X40
μ742.00	199.99	246.01	562.00	188.99	576.96	86.96	424.01	640.00	422.99	
σ^2	1.00	1.00	0.99	1.03	1.04	1.01	0.99	0.99	1.02	0.97
	X41	X42	X43	X44	X45	X46	X47	X48	X49	X50
μ	260.99	527.04	659.01	237.00	61.03	165.99	227.00	102.00	596.99	214.01
σ^2	0.99	0.97	1.00	1.04	0.96	1.01	1.01	0.99	1.01	1.02
	X51	X52	X53							
μ	483.01	94.99	496.01							
σ^2	1.00	1.02	1.03							

Table B.12 – N20000 synthetic dataset ($n = 20000$, $p = 98$) used to assess imputation methods. The μ and σ^2 values of all variables are shown.

	Variables									
	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
μ	204.00	311.99	338.00	154.00	272.99	513.98	391.99	417.00	578.00	386.00
σ^2	0.99	1.00	1.00	1.00	1.01	0.99	1.00	0.99	1.00	0.99
	X11	X12	X13	X14	X15	X16	X17	X18	X19	X20
μ	560.01	771.01	172.00	116.00	587.99	11.99	265.99	827.00	544.99	300.00
σ^2	1.00	1.01	0.99	1.00	0.99	1.01	1.00	1.01	0.99	1.00
	X21	X22	X23	X24	X25	X26	X27	X28	X29	X30
μ	29.00	728.99	471.00	355.00	446.00	422.99	665.00	502.00	797.00	623.02
σ^2	0.99	1.02	1.00	1.01	0.99	0.99	1.00	0.99	1.00	1.01
	X31	X32	X33	X34	X35	X36	X37	X38	X39	X40
μ	789.01	620.00	7.99	386.99	137.00	246.01	821.00	370.00	204.00	786.00
σ^2	1.00	1.00	1.00	1.01	1.02	1.00	1.00	1.00	1.00	0.99
	X41	X42	X43	X44	X45	X46	X47	X48	X49	X50
μ	817.99	254.00	682.01	167.01	56.01	462.99	39.00	574.01	532.99	394.00
σ^2	1.00	1.01	1.02	0.99	1.00	1.00	0.99	0.99	0.99	1.02
	X51	X52	X53	X54	X55	X56	X57	X58	X59	X60
μ	454.00	440.00	704.00	615.00	794.00	65.00	168.00	335.01	845.00	252.01
σ^2	1.01	1.00	1.00	0.99	1.01	1.01	1.00	1.01	1.01	1.00
	X61	X62	X63	X64	X65	X66	X67	X68	X69	X70
μ	319.99	4.01	285.00	19.99	839.00	571.99	373.00	27.00	538.99	27.00
σ^2	1.01	0.98	1.00	1.01	1.03	0.99	1.01	0.98	1.00	1.00
	X71	X72	X73	X74	X75	X76	X77	X78	X79	X80
μ	366.00	740.01	4.00	834.99	631.01	442.01	367.00	716.00	279.00	730.99
σ^2	0.99	1.00	1.02	1.01	0.99	1.01	1.01	1.00	0.99	1.01
	X81	X82	X83	X84	X85	X86	X87	X88	X89	X90
μ	610.00	280.00	593.98	282.99	510.00	714.00	304.99	738.00	500.99	519.99
σ^2	1.01	1.00	1.00	1.03	0.98	1.01	1.00	1.00	1.01	0.99
	X91	X92	X93	X94	X95	X96	X97	X98		
μ	29.00	522.99	8.01	737.00	610.00	201.01	558.01	180.99		
σ^2	1.00	1.01	1.00	0.99	1.01	1.01	1.01	1.00		

B.4 Datasets for assessment of outliers processing

Table B.13 – N21 synthetic dataset ($n = 21$, $p = 8$) used to assess outliers processing methods. The μ and σ^2 values of all variables are shown.

	Variables							
	X1	X2	X3	X4	X5	X6	X7	X8
μ	190.88	155.89	536.80	544.77	693.12	766.84	744.70	268.90
σ^2	4.43	5.43	3.05	0.76	1.10	1.43	1.52	3.31

Table B.14 – N600 synthetic dataset ($n = 600$, $p = 30$) used to assess outliers processing methods. The μ and σ^2 values of all variables are shown.

	Variables									
	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
μ	216.79	511.82	243.71	333.76	171.73	240.75	210.81	310.80	158.73	401.81
σ^2	2.74	3.20	3.75	3.35	3.95	3.36	3.30	2.96	2.86	3.45
	X11	X12	X13	X14	X15	X16	X17	X18	X19	X20
μ	59.69	400.70	405.78	205.84	115.79	229.78	498.84	487.79	23.84	283.77
σ^2	3.07	3.91	3.03	2.83	3.68	3.55	3.19	2.92	2.98	3.20
	X21	X22	X23	X24	X25	X26	X27	X28	X29	X30
μ	563.79	178.79	516.76	163.80	540.84	344.65	32.78	245.72	71.75	81.70
σ^2	3.74	3.51	3.72	3.29	3.23	3.56	3.08	3.75	3.45	3.34

Table B.15 – N4000 synthetic dataset ($n = 4000$, $p = 53$) used to assess outliers processing methods. The μ and σ^2 values of all variables are shown.

	Variables									
	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
μ	96.81	512.78	516.77	487.76	362.76	892.80	355.78	688.74	660.76	816.75
σ^2	3.04	3.21	3.31	3.21	3.24	3.07	3.09	3.40	3.29	3.48
	X11	X12	X13	X14	X15	X16	X17	X18	X19	X20
μ	66.78	480.77	362.78	896.77	385.73	479.81	887.78	788.77	738.74	39.80
σ^2	3.29	3.11	3.28	3.17	3.36	3.11	3.31	3.34	3.37	3.09
	X21	X22	X23	X24	X25	X26	X27	X28	X29	X30
μ	670.76	654.75	769.78	383.76	395.80	764.82	9.76	117.78	243.80	608.72
σ^2	3.42	3.55	3.12	3.53	3.19	2.94	3.54	3.30	3.10	3.24
	X31	X32	X33	X34	X35	X36	X37	X38	X39	X40
μ	741.76	199.76	245.79	561.79	188.78	576.75	86.75	423.77	639.80	422.74
σ^2	3.39	3.39	3.04	3.20	3.17	3.19	3.10	3.37	3.05	3.39
	X41	X42	X43	X44	X45	X46	X47	X48	X49	X50
μ	260.75	526.78	658.79	236.77	60.77	165.77	226.77	101.76	596.79	213.80
σ^2	3.23	3.48	3.33	3.36	3.36	3.23	3.21	3.25	2.96	3.25
	X51	X52	X53							
μ	482.74	94.79	495.78							
σ^2	3.57	3.06	3.35							

B.5 Semi-synthetic datasets

Table B.16 – FQ16 semi-synthetic dataset ($n = 16$, $p = 13$). The μ and σ^2 values of all variables are shown.

	Variables									
	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
μ	9.87	0.72	89.95	0.17	62.00	32.01	54.24	13.64	169.27	3.98
σ^2	0.64	0.15	61.62	0.05	2418.85	3788.79	9718.26	44.65	12204.02	7.01
	X11	X12	X13							
μ		541.63	21.14	0.16						
σ^2	320283.22	352.27	0.01							

Table B.17 – FQ1000 semi-synthetic dataset ($n = 1504$, $p = 26$). The μ and σ^2 values of all variables are shown.

	Variables									
	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
μ	9.78	24.40	3.38	65.36	171.03	2.31	0.89	0.35	0.29	0.97
σ^2	1.21	2496.01	4.58	2619.00	11239.40	0.97	0.68	0.16	0.56	1.18
	X11	X12	X13	X14	X15	X16	X17	X18	X19	X20
μ	11.65	10.68	17.82	11.42	3.38	44.58	16.89	4.16	0.19	514.71
σ^2	146.14	64.15	394.60	1.90	2.35	10935.99	73.99	15.83	0.04	205676.86
μ	88.27	0.13	20.93	7.78	13.45	53.45				
σ^2	101.23	0.04	271.96	0.09	38.44	6617.83				

Table B.18 – FQ7000 semi-synthetic dataset ($n = 7520$, $p = 26$). The μ and σ^2 values of all variables are shown.

	Variables									
	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
μ	9.78	24.40	3.38	65.36	171.03	2.31	0.89	0.35	0.29	0.97
σ^2	1.21	2494.69	4.58	2617.60	11233.42	0.97	0.67	0.16	0.56	1.18
	X11	X12	X13	X14	X15	X16	X17	X18	X19	X20
μ	11.65	10.68	17.82	11.42	3.38	44.58	16.89	4.16	0.19	514.71
σ^2	146.06	64.12	394.39	1.90	2.34	10930.17	73.95	15.82	0.04	205567.45
	X21	X22	X23	X24	X25	X26				
μ	88.27	0.13	20.93	7.78	13.45	53.45				
σ^2	101.18	0.04	271.82	0.09	38.42	6614.31				

Bibliography

Abayomi, K., Gelman, A., and Levy, M. (2008). Diagnostics for multivariate imputations. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 57(3):273–291.

Abedjan, Z., Golab, L., and Naumann, F. (2016). Data profiling. In *32nd IEEE International Conference on Data Engineering, ICDE 2016, Helsinki, Finland, May 16-20, 2016*, pages 1432–1435.

Alameddine, I., Kenney, M., Gosnell, R. J., and Reckhow, K. (2010). Robust multivariate outlier detection methods for environmental data. *Journal of Environmental Engineering*, 136(11):1299–1304.

Alba-Tercedor, J. and Sánchez-Ortega, A. (1988). Un método rápido y simple para evaluar la calidad biológica de las aguas corriente basado en el de hellawell (1978). *Limnetica*, 4:51–56.

Alcántara-Concepción, V., Cram, S., Gibson, R., Ponce de León, C., and Mazari-Hiriart, M. (2013). Method development and validation for the simultaneous determination of organochlorine and organophosphorus pesticides in a complex sediment matrix. *Journal of AOAC International*, 96(4):854–863.

Alferes, J., Tik, S., Copp, J., and Vanrolleghem, P. A. (2013). Advanced monitoring of water systems using in situ measurement stations: Data validation and fault detection. *Water Science and Technology*, 68(5):1022–1030.

Alferes, J. and Vanrolleghem, P. (2014). Automated data quality assessment: Dealing with faulty on-line water quality sensors. In *Ames, D.P., Quinn, N.W.T., Rizzoli, A.E. (Eds.)*. Proceedings of the 7th International Congress on Environmental Modelling and Software, June 15-19, San Diego, California, USA. ISBN: 978-88-9035-744-2.

Altintas, I., Berkley, C., Jaeger, E., Jones, M., Ludascher, B., and Mock, S. (2004). Kepler: an extensible system for design and execution of scientific workflows. In *Scientific and Statistical Database Management, 2004. Proceedings. 16th International Conference on*, pages 423–424. IEEE.

Amaral Pereira, S., Trindade Trindade, C., Faria Albertoni, E., and Palma-Silva, C. (2012). Aquatic macrophytes as indicators of water quality in subtropical shallow lakes. *Southern Brazil Acta Limnologica Brasiliensia*, 24(1):52–63.

Anokhin, P., , and Motro, A. (2001). Data integration: Inconsistency detection and resolution based on source properties. In *Proceedings of FMII-01, International Workshop on Foundations of Models for Information Integration*.

APHA (1998). *Standard Methods for the Examination of Water and Wastewater*. 20th Edition. American Public Health Association: Washington, DC, USA.

APHA/AWWA/WPCF (2005). *Métodos normalizados para el análisis de aguas potables y residuales*. Ediciones Díaz de Santos, S.A.

Arlanturk, S., Siadat, M.-R., Ogunyemi, T., Killinger, K., and Diokno, A. (2016). Analysis of incomplete and inconsistent clinical survey data. *Knowl. Inf. Syst.*, 46:731–750.

Armienta, M., Zamora, V., and Juárez, S. (1987). *Manual para el análisis química de aguas naturales en el campo y laboratorio*. Comunicaciones técnicas, Serie Docencia y Divulgación, No.4, Instituto de Geofísica, UNAM, Laboratorio de Química Analítica.

Armitage, P., Moss, D., Wright, J., and Furse, M. (1983). The performance of the new biological water quality score system based on macroinvertebrates over a wide range of unpolluted runnig-water sites. *Water Research*, 17(3):333–347.

Assessment, M. E. (2005). *Ecosystems and Human Well-being: Synthesis*. Millennium Ecosystem Assessment. World Resources Institute. Island Press, Washington, DC.

Bady, P., Dolédec, S., Fesl, C., Gayraud, S., Bacchi, M., and Schöll, F. (2005). Use of invertebrate traits for the biomonitoring of european large rivers: the effects of sampling effort on genus richness and functional diversity. *Freshwater Biology*, 50:159–173.

Barba-Álvarez, R., De la Lanza-Espino, G., Contreras-Ramos, A., and González-Mora, I. (2013). Insectos acuáticos indicadores de calidad del agua en méxico: casos de estudio, ríos copalita, zimatán y coyula, oaxaca. *Revista Mexicana de Biodiversidad*, 84:381–383.

Barbato, G., Barini, E., Genta, G., and Levi, R. (2011). Features and performance of some outlier detection methods. *Journal of Applied Statistics*, 38(10):2133–2149.

Barbour, M., Gerritsen, J., Snyder, B., and Stribling, J. (1999). *Rapid Bioassessment Protocols for Use in Wadeable Streams and Rivers: Periphyton, Benthic Macroinvertebrates and Fish*. second ed. U.S. Environmental Protection Agency, Office of Water, Washington, DC.

Barnett, V. and Lewis, T. (1994). *Outliers in Statistical Data*. John Wiley & Sons, Chichester UK, 3rd edition.

Bartram, J. and Ballance, R. (1996). *Water Quality Monitoring: A Practical Guide to the Design and Implementation of Freshwater Quality Studies and Monitoring Programmes*. London: E & FN Spon.

Batista, G. E., Monard, M. C., et al. (2002). A study of k-nearest neighbour as an imputation method. *HIS*, 87(251-260):48.

Beck, W. (1955). Suggested method for reporting biotic data. *Sewage and Ind. Wastes*, 27:1193–1197.

Berrahou, L., Lalande, N., Serrano, E., Molla, G., Berti-Équille, L., Bimonte, S., Bringay, S., Cernesson, F., Grac, C., Ienco, D., et al. (2015). A quality-aware spatial data warehouse for querying hydroecological data. *Computers & Geosciences*, 85:126–135.

Berti-Équille, L. (2007a). Measuring and modelling data quality for quality-awareness in data mining. *Studies in Computational Intelligence (SCI)*, 43:101–126.

Berti-Équille, L. (2007b). *Quality awareness for managing and mining data*. PhD thesis, Université de Rennes 1.

Berti-Équille, L. and Dasu, T. (2009). Data quality mining: New research directions. In *Tutorial presented at IEEE International Conference on Data Mining (ICDM), Miami, Florida, USA*, volume 7.

Berti-Equille, L., Loh, J. M., and Dasu, T. (2015). A masking index for quantifying hidden glitches. *Knowledge and Information Systems*, 44(2):253–277.

Bervoets, L., Bruylants, B., Marquet, P., Vandelannoote, A., and Verheyen, R. (1989). A proposal for modification of the belgian biotic index method. *Hydrobiologia*, 179:223–228.

Bolón-Canedo, V., Sánchez-Maroño, N., and Alonso-Betanzos, A. (2013). A review of feature selection methods on synthetic data. *Knowledge and information systems*, 34(3):483–519.

Bonada, N., Prat, N., Resh, V., and Statzner, B. (2006). Developments in aquatic insect biomonitoring: A comparative analysis of recent approaches. *Annu. Rev. Entomol.*, 51:495–523.

Borgoni, R. and Berrington, A. (2013). Evaluating a sequential tree-based procedure for multivariate imputation of complex missing data structures. *Quality & Quantity*, 47(4):1991–2008.

Bray, J. and Curtis, J. (1957). An ordination of the upland forest communities of southern wisconsin. *Ecological Monographs*, 27:325–349.

Bremner, J., Rogers, S., and Frid, C. (2006). Methods for describing ecological functioning of marine benthic assemblages using biological traits analysis (bta). *Ecological Indicators*, 6:609–622.

Breunig, M., Kriegel, H.-P., Ng, R., and Sander, J. (2000). Lof: identifying density-based local outliers. In *ACM sigmod record*, volume 29(2), pages 93–104. ACM.

Brower, J., Zar, J., and C.N.von Ende (1997). *Field and laboratory methods for general ecology.* 4th Ed. WCB/MacGraw-Hill.

Bruzzoniti, M., Sarzanini, C., Costantino, G., and Fungi, M. (2006). Determination of herbicides by solid phase extraction gas chromatography-mass spectrometry in drinking waters. *Analytica Chimica Acta*, 578:241–249.

Buuren, S. and Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, 45(3).

Byers, S. and Raftery, A. (1998). Nearest neighbour clutter removal for estimating features in spatial point processes. *Journal of the American Statistical Association*, 93(442):577–584.

Cairns, J., Albaugh, D., Busey, F., and Chanay, M. (1968). The sequential comparison index – a simplified method for non-biologists to estimate relative differences in biological diversity in stream pollution studies. *Journal of the Water Pollution Control Federation*, 40:1607–1613.

Cairns, J. and Pratt, J. (1993). *A History of Biological Monitoring Using Benthic Macroinvertebrates. in Freshwater Biomonitoring and Benthic Macroinvertebrates.* Chapman & Hall, N.Y.

Cao, H., Si, G., Zhang, Y., and Jia, L. (2010). Enhancing effectiveness of density-based outlier mining scheme with density-similarity-neighbor-based outlier factor. *Expert Systems with Applications*, 37(12):8090–8101.

Charvet, S., Statzner, B., Usseglio-Polatera, P., and Bernard, D. (2000). Traits of benthic macroinvertebrates in semi-natural french streams: an initial application to biomonitoring in europe. *Freshwater Biology*, 43:277–296.

Chatterjee, A. and Segev, A. (1991). Data manipulation in heterogeneous databases. *ACM SIGMOD Record*, 20(4):64–68.

Chen, S., Wang, W., and van Zuylen, H. (2010). A comparison of outlier detection algorithms for ITS data. *Expert Systems with Applications*, 37:1169–1178.

Chessman, B. (1995). Rapid assessment of rivers using macroinvertebrates: A procedure based on habitat-specific sampling, family level identification and a biotic index. *Australian Journal of Ecology*, 20:122–129.

Clarke, J. U. (1998). Evaluation of censored data methods to allow statistical comparisons among very small samples with below detection limit observations. *Environmental Science & Technology*, 32(1):177–183.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46.

Cortes, C., Jackel, L. D., and Chiang, W.-P. (1995). Limits on learning machine accuracy imposed by data quality. In *KDD*, pages 57–62. AAAI Press.

Cottrell, G., Cot, M., and Mary, J.-Y. (2009). L'imputation multiple des données manquantes aléatoirement: concepts généraux et présentation d'une méthode monte-carlo. *Revue d'Épidémiologie et de Santé Publique*, 57(5):361–372.

Couceiro, S., Hamada, N., Forsberg, B., Pimentel, T., and Luz, S. (2012). A macroinvertebrate multimetric index to evaluate the biological condition of streams in the central amazon region of brazil. *Ecological Indicators*, 18:118–125.

Courtemanch, D. and Davies, S. (1987). A coefficient of community loss to assess detrimental change in aquatic communities. *Water Research*, 21:217–222.

Cousineau, D. and Chartier, S. (2015). Outliers detection and treatment: a review. *International Journal of Psychological Research*, 3(1):58–67.

Coussement, K., Van den Bossche, F., and De Bock, K. (2014). Data accuracy's impact on segmentation performance: Benchmarking rfm analysis, logistic regression, and decision trees. *Journal of Business Research*, 67:2751–2758.

Couvreur, C. (1997). The em algorithm: A guided tour. In *Computer Intensive Methods in Control and Signal Processing*, pages 209–222. Springer.

Crone, S. F., Guajardo, J., and Weber, R. (2006a). The impact of preprocessing on support vector regression and neural networks in time series prediction. In *DMIN*, pages 37–44.

Crone, S. F., Lessmann, S., and Stahlbock, R. (2006b). The impact of preprocessing on data mining: An evaluation of classifier sensitivity in direct marketing. *European Journal of Operational Research*, 173(3):781–800.

Damanik-Ambarita, Minar Everaert, G., Forio, M., Nguyen, T., Lock, K., Musonge, P., Suhareva, N., Dominguez-Granda, L., Bennetsen, E., Boets, P., and Goethals, P. (2016a). Generalized linear models to identify key hydromorphological and chemical variables determining the occurrence of macroinvertebrates in the guayas river basin (ecuador). *Water*, 8:297.

Damanik-Ambarita, M. N., Lock, K., Boets, P., Everaert, G., Nguyen, T. H. T., Forio, M. A. E., Musonge, P. L. S., Suhareva, N., Bennetsen, E., Landuyt, D., et al. (2016b). Ecological water quality analysis of the guayas river basin (ecuador) based on macroinvertebrates indices. *Limnologica-Ecology and Management of Inland Waters*, 57:27–59.

Dasu, T. (2013). *Data Glitches: Monsters in Your Data*, pages 163–178. Springer Berlin Heidelberg, Berlin, Heidelberg.

Dasu, T. and Loh, J. M. (2012). Statistical distortion: Consequences of data cleaning. *Proceedings of the VLDB Endowment*, 5(11):1674–1683.

De Pauwn, N. and Vanhooren, G. (1993). Method for biological quality assessment of watercourses in belgium. *Hydrobiologia*, 100:153–168.

De Veaux, R. and Hand, D. (2005). How to lie with bad data. *Statistical Science*, 20(3):231–238.

De Zwart, D. and Trivedi, R. (1994). *Manual on Integrated Water Quality Evaluation.* Report 802023003, National Institute of Public Health and Environmental Protection (RIVM), Bilthoven, The Netherlands.

Deng, F. and Rafiei, D. (2006). Approximately detecting duplicates for streaming data using stable bloom filters. In *Proceedings of the 2006 ACM SIGMOD international conference on Management of data*, pages 25–36. ACM.

DFID (1998). *Impact of wastewater reuse on ground-water in the Mezquital Valley, Hidalgo State, Mexico.* Department For International Development (DFID) Comisión Nacional del Agua, British Geological Survey, London School of Hygiene and Tropical Medicine, University of Birmingham. British Geological Survey Technical Report WC/98/42.

Diaz, T. G., Cabanillas, A. G., Soto, M. L., and Ortiz, J. (2008). Determination of fenthion and fenthion-sulfoxide, in olive oil and in river water, by square-wave adsorptive-stripping voltammetry. *Talanta*, 76(4):809–814.

Dolédec, S., Phillips, N., Scarsbrook, M., Riley, R., and Townsend, C. (2006). Comparison of structural and functional approaches to determining land use effects on grassland stream invertebrate communities. *Journal of the North American Benthological Society*, 25(1):44–60.

Dolédec, S. and Statzner, B. (2010). Responses of freshwater biota to human disturbances: contribution of j-nabs to developments in ecological integrity assessments. *J.N. Am. Benthol. Soc.*, 29(1):286–311.

Donders, A. R. T., van der Heijden, G. J., Stijnen, T., and Moons, K. G. (2006). Review: a gentle introduction to imputation of missing values. *Journal of clinical epidemiology*, 59(10):1087–1091.

Dongilli, P., Fillottrani, P., Franconi, E., and Tessaris, S. (2005). A multi-agent system for querying heterogeneous data sources with ontologies. In *SEBD*, pages 75–86.

Dudgeon, D., Arthington, A., Gessner, M., Kawabata, Z.-I., Knowler, D., Lévêque, C., Naiman, R., Prieur-Richard, A.-H., Soto, D., Stiassny, M., and Sullivan, C. (2006). Freshwater biodiversity: importance, threats, status and conservation challenges. *Biol. Rev.*, 81:163–182.

Eid, H. F., Hassanien, A. E., Kim, T.-h., and Banerjee, S. (2013). Linear correlation-based feature selection for network intrusion detection model. In *Advances in Security of Information and Communication Networks*, pages 240–248. Springer.

Elmagarmid, A., Ipeirotis, P., and Verykios, V. (2007). Duplicate record detection: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 19(1):1–16.

Engels, J. M. and Diehr, P. (2003). Imputation of missing longitudinal data: a comparison of methods. *Journal of clinical epidemiology*, 56(10):968–976.

EPA (2006). *Data Quality Assessment: A Reviewer's Guide. EPA QA/G-9S.* EPA, USA. Office of Environmental Information Washington, DC 20460.

Eppler, M. and Helfert, M. (2004). A classification and analysis of data quality costs. In *International Conference on Information Quality*, pages 311–325.

Eskelson, B. N., Temesgen, H., Lemay, V., Barrett, T. M., Crookston, N. L., and Hudak, A. T. (2009). The roles of nearest neighbor methods in imputing missing data in forest inventory and monitoring databases. *Scandinavian Journal of Forest Research*, 24(3):235–246.

Fahringer, T., Prodan, R., Duan, R., Hofer, J., Nadeem, F., Nerieri, F., Podlipnig, S., Qin, J., Siddiqui, M., Truong, H.-L., et al. (2007). Askalon: A development and grid computing environment for scientific workflows. In *Workflows for e-Science*, pages 450–471. Springer.

Famili, A., Shen, W.-M., Weber, R., and Simoudis, E. (1997). Data preprocessing and intelligent data analysis. *Intelligent Data Analysis*, 1:3–23.

Farhangfar, A., Kurgan, L., and Dy, J. (2008). Impact of imputation of missing values on classification error for discrete data. *Pattern Recognition*, 41(12):3692–3705.

Feio, M., Reynolds, T., Ferreira, V., and Graça, M. (2007). A predictive model for freshwater bioassessment (mondego river, portugal). *Hydrobiologia*, 589(1):55–68.

Filzmoser, P. (2004). *A multivariate outlier detection method*. na.

Filzmoser, P. (2005). Identification of multivariate outliers: A performance study. *Austrian Journal of Statistics*, 34(2):127–138.

Filzmoser, P., Garrett, R., and Reimann, C. (2005). Multivariate outlier detection in exploration geochemistry. *Computers & Geosciences*, 31(5):579–587.

Filzmoser, P., Maronna, R., and Werner, M. (2008). Outlier identification in high dimensions. *Computational Statistics and Data Analysis*, 52:1694–1711.

García, S., Luengo, J., and Herrera, F. (2015). *Data preprocessing in data mining*. Springer.

García-de la Parra, L., Cervantes-Mojica, L., González-Valdivia, C., Martínez-Cordero, F., Aguilar-Zárate, G., Bastidas-Bastidas, P., and Betancourt-Lozano, M. (2012). Distribution of pesticides and pcbs in sediments of agricultural drains in the culiacan valley, sinaloa, mexico. *Arch. Environ. Contam. Toxicol.*, 63:323–336.

García de Llasera, M. and Bernal-González, M. (2000a). Presence of carbamates pesticides in environmental waters from the northwest of mexico: Determination by liquid chromatography. *Wat. Res.*, 35(8):1933–1940.

García de Llasera, M. and Bernal-González, M. (2000b). Presence of carbamates pesticides in environmental waters from the northwest of mexico: Determination by liquid chromatography. *Wat. Res.*, 35(8):1933–1940.

García-Laencina, P. J., Sancho-Gómez, J.-L., Figueiras-Vidal, A. R., and Verleysen, M. (2009). K nearest neighbours with mutual information for simultaneous classification and missing data imputation. *Neurocomputing*, 72(7):1483–1493.

Geissen, V., Mol, H., Klumpp, E., Umlauf, G., Nadal, M., van der Ploeg, M., E.A.T.M. van de Zee, S., and J. Ritsema, C. (2015a). Emerging pollutants in the environment: a challenge for water resource management. *International Soil and Water Conservation Research*, 3:57–65.

Geissen, V., Mol, H., Klumpp, E., Umlauf, G., Nadal, M., van der Ploeg, M., van de Zee, S. E., and Ritsema, C. J. (2015b). Emerging pollutants in the environment: a challenge for water resource management. *International soil and water conservation research*, 3(1):57–65.

Ghetti, P. (1997). *Manuale di Applicazione: Indice Biotico Esteso – I macroinvertebrati nel controllo della qualità degli ambienti di acque correnti*. Provincia Autonoma di Trento, Servizio Protenzione Ambiente: Trento, Italy.

Ghosh, D. and Vogt, A. (2012). Outliers: An evaluation of methodologies. In *Joint Statistical Meetings*, pages 3455–3460. American Statistical Association San Diego, CA.

Gibert, K., Izquierdo, J., Holmes, G., Athanasiadis, I., Comas, J., and Sànchez-Marrè, M. (2008). On the role of pre and post-processing in environmental data mining. In *Sànchez-Marrè, Miquel and Béjar, J. and Comas J. and Rizzoli A. and Guarisa G. (Eds.)*. Proceedings of the 4th International Congress on Environmental Modelling and Software.

Gibert, K., Sànchez-Marrè, M., and Izquierdo, J. (2016). A survey on pre-processing techniques: Relevant issues in the context of environmental data mining. *AI Communications*, 29(6):627–663.

Gibson, R., Becerril-Bravo, E., Silva-Castro, V., and Jiménez, B. (2007). Determination of acidic pharmaceuticals and potential endocrine disrupting compounds in wastewaters and spring waters by selective elution and analysis by gas chromatography-mass spectrometry. *Journal of Chromatography A.*, 1169:31–39.

Güler, C., Thyne, G., McCray, J., and Turner, A. (2002). Evaluation of graphical and multivariate statistical methods for classification of water chemistry data. *Hydrogeology Journal*, 10:455–474.

Gopalan, P. and Radhakrishnan, J. (2009). Finding duplicates in a data stream. In *Proceedings of the twentieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 402–411. Society for Industrial and Applied Mathematics.

Grimvall, A., Wackernagel, H., and Lajaunie, C. (2001). Normalisation of environmental quality data. *Sustainability in the Information Society*, pages 583–590.

Groombridge, B. and Jenkins, M. (2000). *Global Biodiversity. Earth's Living Resources in the 21st Century*. World Conservation Monitoring Centre, Cambridge, U.K.

Guardia Rubio, M., Ruiz Medina, A., Pascual Reguera, M., and Fernández de Córdova, M. (2007). Multiresidue analysis of three groups of pesticides in washing waters from olive processing by solid-phase extraction-gas chromatography with electron capture and thermionic specific detection. *Microchemical Journal*, 85:257–264.

Gutiérrez-Fonseca, P. and Lorion, C. (2014). Application of the bmwp-costa rica biotic index in aquatic biomonitoring: sensitivity to collection method and sampling intensity. *Int. J. Trop. Biol. Conserv.*, 62:275–289.

Haase, P., Murray-Bligh, J., Lohse, S., Pauls, S., Sundermann, A., Gunn, R., and Clarke, R. (2006). Assessing the impact of errors in sorting and identifying macroinvertebrate samples. In *The Ecological Status of European Rivers: Evaluation and Intercalibration of Assessment Methods*, pages 505–521. Springer.

Hall, M. A. (1999). *Correlation-based feature selection for machine learning*. PhD thesis, The University of Waikato.

Han, J. and Kamber, M. (2000). *Data Mining: Concepts and Techniques*. Morgan Kaufmann.

Harmancioglu, N., Fistikoglu, O., Ozkul, S., Singh, V., and Alpaslan, M. (1999). *Water Quality Monitoring Network Design*. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Harrell, F. (2015). *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*. Springer.

Harvey, D. (2000). *Modern analytical chemistry*, volume 381. McGraw-Hill New York.

Haug, A., Zachariassen, F., and van Liempd, D. (2011). The costs of poor data quality. *Journal of Industrial Engineering and Management*, 4(2):168–193.

Hawkins, D. M. (1980). *Identification of outliers*, volume 11. Springer.

Heckman, C. (2006). *Encyclopedia of South American Aquatic Insects: Odonata - Anisoptera*. Springer, Dordrecht, The Netherlands.

Heckman, C. (2008). *Encyclopedia of South American Aquatic Insects: Odonata - Zigoptera*. Springer, Olympia, WA, USA.

Heckman, C. (2011). *Encyclopedia of South American Aquatic Insects: Hemiptera - Heteroptera*. Springer, Olympia, WA, USA.

Heitjan, D. F. (1997). Annotation: what can be done about missing data? approaches to imputation. *American Journal of Public Health*, 87(4):548–550.

Hellawell, J. (1978). *Biological surveillance of rivers, a biological monitoring hand-book*. Water Research Centre, Medmenham and Stevenag, UK.

Hellerstein, J. M. (2008). Quantitative data cleaning for large databases. *United Nations Economic Commission for Europe (UNECE)*.

Helsel, D. and Hirsch, R. (2002). *Statistical methods in water resources*. Elsevier.

Helson, J. and Williams, D. (2013). Development of a macroinvertebrate multimetric index for the assessment of lowland streams in the neotropics. *Ecological Indicators*, 29:167–178.

Henne, L., Scheider, D., and Martínez, L. (2002). Rapid assessment of organic pollution in a west-central mexican river using a family-level biotic index. *Journal of Environmental Planning and Management*, 45:613–632.

Hilsenhoff, W. (1982). *Using a biotic index to evaluate water quality in streams*. Technical Bulletin No. 132. Wisconsin Department of Natural Resources. Madison, W.I.

Hilsenhoff, W. (1988). Rapid field assessment of organic pollution with a family-level biotic index. *Journal of the North American Benthological Society*, 7(1):65–68.

Ho, P., Silva, M., and Hogg, T. (2001). Multiple imputation and maximum likelihood principal component analysis of incomplete multivariate data from a study of the ageing of port. *Chemometrics and Intelligent Laboratory Systems*, 55(1):1–11.

Hodge, V. and Austin, J. (2004). A survey of outlier detection methodologies. *Artificial Intelligence Review*, 22:85–126.

Holomuzki, J. and Biggs, B. (2000). Taxon-specific responses to high-flow disturbance in streams: implications for population persistence. *J.N. Am. Benthol. Soc.*, 19(4):670–679.

Holvoet, K. M., Seuntjens, P., and Vanrolleghem, P. A. (2007). Monitoring and modeling pesticide fate in surface waters at the catchment scale. *Ecological modelling*, 209(1):53–64.

Hron, K., Templ, M., and Filzmoser, P. (2010). Imputation of missing values for compositional data using classical and robust methods. *Computational Statistics & Data Analysis*, 54(12):3095–3107.

Huang, H., Lin, J., Chen, C., and Fan, M. (2006). Review of outlier detection. *Journal of Research on Computer Application*, 8:8–13.

Huang, S., Zhu, F., Jiang, R., Zhou, S., Zhu, D., Liu, H., and Ouyang, G. (2015). Determination of eight pharmaceuticals in an aqueous sample using automated derivatization solid-phase microextraction combined with gas chromatography–mass spectrometry. *Talanta*, 136:198–203.

INEGI (1990). *Anuario Estadístico de los Estados Unidos Mexicanos 1988-1989*. Instituto Nacional de Estadística, Geografía e Informática, Aguascalientes, México.

INEGI (1995). *Estudio Hidrológico del Estado de Sinaloa*. Instituto Nacional de Estadística, Geografía e Informática - México: INEGI.

INEGI (2013). *Anuario estadístico y geográfico de Hidalgo*. Instituto Nacional de Estadística y Geografía - México: INEGI.

INEGI (2014). *Anuario estadístico y geográfico de Sinaloa*. Instituto Nacional de Estadística y Geografía - México: INEGI.

ISO (1979). *Assessment of the biological quality of rivers by a macroinvertebrate 'score'*. International Organization for Standardization (ISO) ISO/TC 147/SC5/WG 6N5.

Jaccard, P. (1901). Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 37:547–579.

Jaccard, P. (1908). Nouvelles recherches sur la distribution florale. *Bull. Soc. Vaud. Sci. Nat. XLIV*, 163:223–269.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning*, volume 6. Springer.

Jerez, J. M., Molina, I., García-Laencina, P. J., Alba, E., Ribelles, N., Martín, M., and Franco, L. (2010). Missing data imputation using statistical and machine learning methods in a real breast cancer problem. *Artificial intelligence in medicine*, 50(2):105–115.

Jáimez-Cuéllar, P., Vivas, S., Casa, J., Ortega, M., Prado, I., Prat, N., Rieradevall, M., Sáinz-Cantero, C., Sánchez-Ortega, A., Suárez, M., Toro, M., Vidal-Abarca, M., Zamora-Muñoz, C., and Alba-Tercedor, J. (2004). Protocolo guadalmed. *Limnetica*, 21:187–204.

Johnson, R., Hering, D., Furse, M., and Clarke, R. (2006). Detection of ecological change using multiple organism groups: metrics and uncertainty. *Hydrobiologia*, 566:115–137.

Kadengye, D. T., Cools, W., Ceulemans, E., and Van den Noortgate, W. (2012). Simple imputation methods versus direct likelihood analysis for missing item scores in multilevel educational data. *Behavior research methods*, 44(2):516–531.

Kane, D., Gordon, S., Munawar, M., Charlton, M., and Culver, D. (2009). The planktonic index of biotic integrity (pibi): An approach for assessing lake ecosystem health. *Ecological Indicators*, 9:1234–1247.

Karr, J. (1981). Assessment of biotic integrity using fish communities. *Fisheries*, 6(6):21–27.

Karr, J. (1999). Defining and measuring river health. *Freshwater Biology*, 41:221–234.

Kilbane, G. and Holomuzki, J. (2004). Spatial attibutes, scale, and species traits determine caddisfly distributional responses to flooding. *J.N. Am. Benthol. Soc*, 23(3):408–493.

Kim, W., Choi, B.-J., Hong, E.-K., Kim, S.-K., and Lee, D. (2003). A taxonomy of dirty data. *Data Mining and Knowledge Discovery*, 7:81–99.

Knorr, E. and Ng, R. (1998). Algorithms for mining distance-based outliers in large datasets. In *In Proceedings of the 24th VLDB conference*, pages 392–403. New York, USA.

Knorr, E., Ng, R., and Tucakov, V. (2000). Distance-based outlier: Algorithms and applications. *VLDBJ*, 8(3-4):237–253.

Kohavi, R. and John, G. H. (1997). Wrappers for feature subset selection. *Artificial intelligence*, 97(1-2):273–324.

Kolev, B., Valduriez, P., Bondiombouy, C., Jiménez-Peris, R., Pau, R., and Pereira, J. (2015). Cloudmdsql: Querying heterogeneous cloud data stores with a common language. *Distributed and Parallel Databases*, pages 1–41.

Kolkwitz, R. and Marsson, M. (1908). Ökologie der pflanzlichen saprobien. *Berichte der Deutschen Botanischen Gesellschaft*, 26A:505–519.

Kotsiantis, S., Kanellopoulos, D., and Pintelas, P. (2006). Data preprocessing for supervised leaning. *International Journal of Computer Science*, 1(1):111–117.

Kriegel, H.-P., Kröger, P., and Zimek, A. (2010). Outlier detection techniques. In *Tutorial at the 16th ACM international conference on knowledge discovery and data Mining (SIGKDD), Washington, DC*.

Lafont, M., Vivier, A., Nogueira, S., Namour, P., and Breil, P. (2006). Surface and hyporheic oligochaete assemblages in a french suburban stream. *Hydrobiologia*, 564:183–193.

Lamastra, L., Balderacchi, M., and Trevisan, M. (2016). Inclusion of emerging organic contaminants in groundwater monitoring plans. *MethodsX*, 3:459–476.

Lane, C. and Brown, M. (2007). Diatoms as indicators of isolated herbaceous wetland condition in florida, usa. *Ecological Indicators*, 7:521–540.

Li, L., Zheng, B., and Liu, L. (2010). Biomonitoring and bioindicators used for river ecosystems: Definitions, approaches and trends. *Procedia Environmental Sciences*, 2:1510–1524.

Linke, S., Norris, R., Faith, D., and Stockwell, D. (2005). Anna: A new prediction method for bioassessment programs. *Freshwater Biology.*, 50:147–158.

Little, R. and Rubin, D. (2002). *Statistical analysis with missing data.* (2nd ed.) New York: John Wiley & Sons.

Liu, H., Li, J., and Wong, L. (2002). A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns. *Genome informatics*, 13:51–60.

Liu, H. and Motoda, H. (2012). *Feature selection for knowledge discovery and data mining*, volume 454. Springer Science & Business Media.

Liu, J., Pacitti, E., Valduriez, P., and Mattoso, M. (2015). A survey of data-intensive scientific workflow management. *Journal of Grid Computing*, 13(4):457–493.

López-Hernández, M., Ramos-Espinosa, M., and Carranza-Fraser, J. (2007). Multimetric analyses for assessing pollution in the lerma river and chapala lake, mexico. *Hidrobiológica*, 17:17–30.

Ma, J., Xiao, R., Li, J., Zhao, X., Shi, B., and Li, S. (2009). Determination of organophosphorus pesticides in underground water by spe-gc-ms. *Journal of Chromatographic Science*, 47:110–115.

Mafla, M. (2005). *Guia para Evaluaciones Ecologicas Rapidas con Indicadores Biologicos en rios de tamaño mediano Talamanca – Costa Rica. Macroinvertebrados (BMWP – CR – Biological Monitoring Working Party) y Habitat (SVAP – Stream Visual Assessment Protocol).* Centro Agronómico tropical de Investigación y Ensenanza (CATIE), Turriala, Costa Rica.

Mansilha, C., Melo, A., Rebelo, H., Ferreira, I., Pinho, O., Domingues, V., Pinho, C., and Gameiro, P. (2010). Quantification of endocrine disruptors and pesticides in water by gas chromatography-tandem mass spectrometry. method validation using weighted linear regression schemes. *Journal of Chromatorgraphy A.*, 1217:6681–6691.

Margalef, R. (1951). Diversidad de especies en las commundades naturales. *PubInes Inst. Biol. apl. Barcelona*, 9:5–27.

Markert, B., Wappelhorst, O., Weckert, V., Herpin, U., Siewers, U., Friese, K., and Breulmann, G. (1999). The use of bioindicators for monitoring the heavy-metal status of the environment. *Radioanal. Nucl. Chem.*, 240(2):425–429.

Masese, F. O., Omukoto, J. O., and Nyakeya, K. (2013). Biomonitoring as a prerequisite for sustainable water resources: a review of current status, opportunities and challenges to scaling up in east africa. *Ecohydrology & Hydrobiology*, 13:173–191.

Mathuriau, C., Mercado Silva, N., Lyons, J., and Martínez Rivera, L. (2011). *Fish and Macroinvertebrates as Freshwater Ecosystem Bioindicators in Mexico: Current State and Perspectives in Water Resources in Mexico: Scarcity, Degradations, Stress, Conflicts, Management and Policy.* Hexagon Series on Human and Environmental Security and Peace ed., Heidelberg, Springer-Verlag.

Meilă, M. (2007). Comparing clusterings—an information based distance. *Journal of multivariate analysis*, 98(5):873–895.

Mercer, T. G., Frostick, L. E., and Walmsley, A. D. (2011). Recovering incomplete data using statistical multiple imputations (smi): A case study in environmental chemistry. *Talanta*, 85(5):2599–2604.

Merrit, R., Cummins, K., and Berg, M. (2008a). *An Introduction to the Aquatic Insects of Norht America.* Kendall/Hunt Publishing, Dubuque, Iowa fourth ed.

Merrit, R., Cummins, K., and Berg, M. (2008b). *An introduction to the Aquatic Insects of North America.* 4th edition. Kendall/Hunt Publishing, Dubuque, Iowa.

Metcalfe, J. (1989). Biological water quality assessment of running waters based on macroinvertebrate communities: History and present status in europe. *Environmental Pollution*, 60:101–139.

Metzeling, L., Chessman, B., Hardwick, R., and Wong, V. (2003). Rapid assessment of rivers using macroinvertebrates: the role of experience, and comparisons with quantitative methods. *Hydrobiologia*, 510(1-3):39–52.

Meyers, R. (2000). *Encyclopedia of Analytical Chemistry, Applications, Theory and Instrumentation*. Jhon Wiley & Sons.

Mondy, C., Villeneuve, B., Archaimbault, V., and Usseglio-Polatera, P. (2012). A new macroinvertebrate-based multimetric index (i2m2) to evaluate ecological quality of french wadeable streams fulfilling the wfd demands: A taxonomical and trait approach. *Ecological Indicators*, 18:452–467.

Moya, N., Hughes, R., Domínguez, E., Gibon, F.-M., Goitia, E., and Oberdorff, T. (2011). Macroinvertebrate-based smultimetric predictive models for evaluating the human impact on biotic condition of bolivian streams. *Ecological Indicators*, 11:840–847.

Murray, K., Thomas, S., and Bodour, A. (2010). Prioritizing research for trace pollutants and emerging contaminants in the freshwater environment. *Environmental Pollution*, 158:3462–3471.

Murtaugh, P. A. (2009). Performance of several variable-selection methods applied to real ecological data. *Ecology Letters*, 12(10):1061–1068.

Mustaffa, Z. and Yusof, Y. (2011). A comparison of normalization techiques in predicting dengue outbreak. In *International Conference on Business and Economics Research, IACSIT Press*, volume 1.

National Water Council (1981). *River Quality:*. the 1980 Survey and Future Outlook.

Nawi, N. M., Atomi, W. H., and Rehman, M. (2013). The effect of data pre-processing on optimized training of artificial neural networks. *Procedia Technology*, 11:32–39.

Nebot, C., Gibb, S., and Boyd, K. (2007). Quantification of human pharmaceuticals in water samples by high performance liquid chromatography-tandem mass spectrometry. *Analytica Chimica Acta*, 598:87–94.

Neubauer, F., Hoheisel, A., and Geiler, J. (2006). Workflow-based grid applications. *Future Generation Computer Systems*, 22(1):6–15.

Norris, R. and Hawkins, C. (2000). Monitoring river health. *Hydrobiologia*, 435:5–17.

Novelo-Gutiérrez, R. (1997a). *Clave para la determinación de familias y géneros de las náyades de Odonata de México. Parte II. Anisoptera.* Dugesiana.

Novelo-Gutiérrez, R. (1997b). *Clave para la separación de familias y géneros de las náyades de Odonata de México. Parte I. Zygoptera.* Dugesiana.

Nowell, L. H., Crawford, C. G., Gilliom, R. J., Nakagaki, N., Stone, W. W., Thelin, G. P., and Wolock, D. M. (2009). Regression models for explaining and predicting concentrations of organochlorine pesticides in fish from streams in the united states. *Environmental Toxicology and Chemistry*, 28(6):1346–1358.

Oba, S., Sato, M.-a., Takemasa, I., Monden, M., Matsubara, K.-i., and Ishii, S. (2003). A bayesian missing value estimation method for gene expression profile data. *Bioinformatics*, 19(16):2088–2096.

Olinsky, A., Chen, S., and Harlow, L. (2003). The comparative efficacy of imputation methods for missing data in structural equation modeling. *European Journal of Operational Research*, 151(1):53–79.

Palarea-Albaladejo, J. and Martín-Fernández, J. (2008). A modified em alr-algorithm for replacing rounded zeros in compositional data sets. *Computers & Geosciences*, 34(8):902–917.

Papadimitriou, S., Kitagawa, H., Gibbons, P., and Faloutsos, C. (2003). Loci: Fast outlier detection using the local correlation integral. In *Data Engineering, 2003. Proceedings. 19th International Conference on*, pages 315–326. IEEE.

Park, E., Turner, S., and Spiegelman, C. (2003). Empirical approaches to outlier detection in intelligent transportation systems data. *Transportation Research Record*, 1840:21–30.

Pasha, M. Z. (2013). A comparative study on outlier detection techniques. *International Journal of Computer Applications*, 66(24):23–27.

Pawliszyn, J. (2010). Theory of extraction. *Handbook of Sample Preparation*, pages 1–24.

Peña-Álvarez, A. and Castillo-Alanís, A. (2015). Identificación y cuantificación de contaminantes emergentes en aguas residuales por microextracció en fase sólida-cromatografía de gases-espectrometría de masas (mefs-cg-em). *Revista Expecializada en Ciencias Químico-Biológicas*, 18(1):29–42.

Pegram, G., Li, Y., Quesne, T. L., Speed, R., Jianqiang, L., and Fuxin, S. (2013). *River basin planning: Principles, procedures and approaches for strategic basin planning.* Paris, UNESCO.

Petrie, B., Barden, R., and Kasprzyk-Hordern, B. (2015). A review on emerging contaminants in wastewaters and the environment: Current knowledge, understudied areas and recommendations for future monitoring. *Water Research*, 72:3–27.

Pigott, T. (2001). A review of methods for missing data. *Educational Research and Evaluation*, 7(4):353–383.

Pillmann, W., Geiger, W., and Voigt, K. (2006). Survey of environmental informatics in europe. *Environmental Modelling & Software*, 21:1519–1527.

Pinto, R., Patricio, J., Baeta, A., Fath, B., Neto, J., and Marques, J. (2009). Review and evaluation of estuarine biotic indices to assess benthic condition. *Ecol. Indic.*, 9:1–25.

Plafkin, J., Barbour, M., Porter, K., Gross, S., and Hughes, R. (1989). *Rapid Bioassessment Protocols for use in streams and rivers: Benthic Macroinvertebrates and Fish.* U.S. Environmental Protection Agency, EPA 444/4-89/001, US EPA, Washington.

Poff, N., Olden, J., Vieira, N., Finn, D., Simmons, M., and Kondratieff, B. (2006). Functional trait niches of north american lotic insects: traits-based ecological applications in light of phylogenetic relationships. *J. N. Am. Benthol. Soc*, 25(4):730–755.

Poquet, J., Alba-Tercedor, J., Puntí, T., Sánchez-Montoya, M., Robles, S., Alvarez, M., Zamora-Munoz, C., Sáinz-Cantero, C., Vidal-Abarca, M., Suárez, M., and Toro, M. (2009). The mediterranean prediction and classification system (medpacs): an implementation of the rivpacs/ausrivas predictive approach for assessing mediterranean aquatic macroinvertebrate communities. *Hydrobiologia*, 623(1):153–171.

Praveena, S. M., Kwan, O. W., and Aris, A. Z. (2012). Effect of data pre-treatment procedures on principal component analysis: a case study for mangrove surface sediment datasets. *Environmental monitoring and assessment*, 184(11):6855–6868.

R Core Team (2016). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

Ramaswamy, S., Rastogi, R., and Shim, K. (2000). Efficient algorithms for mining outliers from large data set. In *In Procedings fo ACM SIDMOD international conference on management of data*, pages 1–20.

Ramezani, M. and Fatemizadeh, E. (2010). Comparison of supervised classification methods with various data preprocessing procedures for activation detection in fmri data. In *Computational Neuroscience*, pages 75–83. Springer.

Reimann, C., Filzmoser, P., and Garrett, R. (2005). Background and threshold: critical comparison of methods of determination. *Science of the Total Environment*, 346:1–16.

Ren, D., Wang, B., and Perrizo, W. (2004). Rdf: A density-based outlier detection method using vertical data representation. In *Data Mining, 2004. ICDM'04. Fourth IEEE International Conference on*, pages 503–506. IEEE.

Resh, V. (2007). Multinational, freshwater biomonitoring programs in the developing world. lessons learned from african and southeast asian river surveys. *Environ. Manage*, 39:737–748.

Resh, V. and Jackson, J. (1993). *Rapid assessment approaches to biomonitoring using benthic macroinvertebrates. Freshwater Biomonitoring and Benthic Macroinvertebrates.* Rosenberg D.M. & Resh V.H., ed. Chapman and Hall, New York.

Reynoldson, T., Bailey, R., Day, K., and Norris, R. (1995). Biological guidelines for freshwater sediment based on benthic assessment of sediment (the beast) using a multivariate approach for predicting biological state. *Aust. J. Ecol.*, 20:198–219.

Riani, M., Atkinson, A., and Cerioli, A. (2009). Finding an unknown number of multivariate outliers. *Journal of Royal Statistical Society B*, 71:447–466.

Rieger, L., Takács, I., Villez, K., Siegrist, H., Lessard, P., Vanrolleghem, P. A., and Comeau, Y. (2010). Data reconciliation for wastewater treatment plant simulation studies—planning for high-quality data and typical sources of errors. *Water environment research*, 82(5):426–433.

Ritchie, J. C., Zimba, P. V., and Everitt, J. H. (2003). Remote sensing techniques to assess water quality. *Photogrammetric Engineering & Remote Sensing*, 69(6):695–704.

Robles-Molina, J., Lara-Ortega, F., Gilvert-López, B., García-Reyes, J., and Molina-Díaz, A. (2014). Multi-residue method for the determination of over 400 priority and emerging pollutants in water and wastewater by solid-phase extraction and liquid chromatography-time-of-flight mass spectrometry. *Journal of Chromatography A.*, 1350:30–43.

Roldán Pérez, G. (2003). *Bioindicación de la calidad del agua en Colombia. Uso del método BMWP/Col.* Medellin, Editorial Universidad de Antioquia.

Rosas, I., Mazari, M., Saavedra, J., and Báez, P. (1984). Benthic organisms as indicators of water quality in lake patzcuaro, mexico. *Water, Air and Soil Pollution*, 25:401–414.

Rosenberg, D. and Resh, V. (1982). *The use of artificial substrates in the study of freshwater benthic macroinvertebrates.* In: Artificial Substrates, Cairns, J. (Ed) Ann Arbor Science: Ann Arbor, MI, USA.

Rosenberg, D. and Resh, V. (1993). *Freshwater Biomonitoring and Benthic Macroinvertebrates.* Chapman & Hall, N.Y.

Rouessac, F. and Rouessac, A. (2008). *Chemical Analysis: Modern Instrumentation Methods and Techniques*, volume Second edition. John Wiley & Sons, Ltd.

Rubin, D. (1976). Inference and missing data. *Biometria*, 63:581–592.

Saeys, Y., Inza, I., and Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19):2507–2517.

Salvatierra-Stamp, V., Ceballos-Magaña, S., Gonzalez, J., Ibarra-Galván, V., and Muñiz-Valencia, R. (2015a). Analytical method development for the determination of emerging contaminants in water using supercritical-fluid chromatography coupled with diode-array detection. *Anal. Bioanal. Chem.*, 407:4219–4226.

Salvatierra-Stamp, V., Ceballos-Magaña, S., González, J., Jurado, J., and Muñiz-Valencia, R. (2015b). Emerging contaminant determination in water samples by liquid chromatography using a monolithic column coupled with a photodiode array detector. *Anal. Bioanal. Chem.*, 407:4661–4670.

Sarawagi, S. and Bhamidipaty, A. (2002). Interactive deduplication using active learning. In *Proc. Eight ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD '02)*, pages 269–278.

Schafer, J. and Graham, J. (2002). Misisng data: Our view of the state of the art. *Psychological Methods*, 7(2):147–177.

Sebastian-Coleman, L. (2012). *Measuring Data Quality for Ongoing Improvement: A Data Quality Assessment Framework.* Newnes.

Serrano Balderas, E., Grac, C., Berti-Equille, L., and Armienta Hernandez, M. (2016). Potential application of macroinvertebrates indices in bioassessment of mexican streams. *Ecological Indicators*, 61:558–567.

Serrano Balderas, E. C. (2012). *Study of Rhin-Meuse streams: inventary, quality of data and characterisation of physicochemical status.* PhD thesis, Université de Strasbourg.

Serrano Balderas, E. C. (2013). *Data pretreatment assessment on statistical results: imputation and clustering methods.* PhD thesis, Université Claude Bernard Lyon 1.

Shannon, C. (1948). The mathematical theory of communication. *Bell System Tech*, 27:379-423:623–656.

Shen, H. and Zhang, Y. (2008). Improved approximate detection of duplicates fro data streams over sliding windows. *Journal of Computer Science and Technology*, 23(6):973–987.

Simpson, E. (1949). Measurement of diversity. *Nature*, 163:163–688.

Sivaramakrishnan, K., Hannaford, M., and Resh, V. (1996). Biological assessment of the kaveri river catchment, south india, using benthic macroinvertebrates: applicability of water quality monitoring approaches developed in other countries. *Int. J. Ecol. Environ. Sci.*, 32:113–132.

Skoog Douglas, A., West Donald, M., and Holler, F. J. (2014). Fundamentals of analytical chemistry. *Chem. Listy*, 108:694–696.

Smital, T., Luckenbach, T., Sauerborn, R., Hamdoun, A., Vega, R., and Epel, D. (2004). Emerging contaminants-pesticides, ppcps, microbial degradation products and natural substances as inhibitors of multixenobiotic defense in aquatic organisms. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 552:101–117.

Smith, B., Clifford, N. J., and Mant, J. (2014). The changing nature of river restoration. *Wiley Interdisciplinary Reviews: Water*, 1:249–261.

Smith, M., Kay, W., Edward, D., Papas, P., Richardson, K. S. J., Simpson, J., Pinder, A., Cale, D., Horwitz, P., Davis, J., Yung, F., Norris, R., and Halse, S. (1999). Ausrivas: using macroinvertebrates to assess ecological condition of river in western australia. *Freshwater Biology*, 41:269–282.

Sokal, R. R. and Rohlf, F. J. (1962). The comparison of dendrograms by objective methods. *Taxon*, pages 33–40.

Sponseller, R., Benfield, E., and Valett, H. (2001). Relationships between land use, spatial scale and stream macroinvertebrate communities. *Freshwater Biology*, 46:1409–1424.

Sørensen, T. (1948). A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on danish commons. *Kongelige Danske Videnskabernes Selskab*, 5(4):1–34.

Statzner, B., Bady, P., Dolédec, S., and Schöll, F. (2005). Invertebrate traits for the biomonitoring of large european rivers: an initial assessment of trait patterns in least impacted river raches. *Freshwater Biology*, 50:2136–2161.

Statzner, B., Bis, B., Dolédec, S., and Usseglio-Polatera, P. (2001a). Perspectives for biomonitoring at large spatial scales: an unified measure for the functional composition of invertebrate communities in european running waters. *Basic and Applied Ecology*, 2:73–85.

Statzner, B., Hildrew, A., and Resh, V. (2001b). Species traits and environmental constraints: entomological research and the history of ecological theory. *Annu. Rev. Entomol*, 46:291–316.

Sutha, K. and Tamilselvi, J. J. (2015). A review of feature selection algorithms for data mining techniques. *International Journal on Computer Science and Engineering*, 7(6):63.

Tachet, H., Richoux, P., Bournaud, M., and Usseglio-Polatera, P. (2010). *Invertébrés d'eau douce. Systématique, biologie, écologie*. CNRS Editions. Paris.

Talia, D. (2013). Workflow systems for science: concepts and tools. *ISRN Software Engineering*, 2013.

Talia, D., Trunfio, P., and Verta, O. (2005). Weka4ws: a wsrf-enabled weka toolkit for distributed data mining on grids. In *European Conference on Principles of Data Mining and Knowledge Discovery*, pages 309–320. Springer.

Taylor, I., Shields, M., Wang, I., and Rana, O. (2003). Triana applications within grid computing and peer to peer environments. *Journal of Grid Computing*, 1(2):199–217.

Templ, M., Alfons, A., Kowarik, A., and Prantner, B. (2011a). Vim: Visualization and imputation of missing values. *URL http://CRAN.R-project.org/package=VIM. R package version 3.0.0*.

Templ, M., Kowarik, A., and Filzmoser, P. (2011b). Iterative stepwise regression imputation using standard and robust methods. *Computational Statistics & Data Analysis*, 55(10):2793–2806.

Thorndike, R. L. (1953). Who belongs in the family? *Psychometrika*, 18(4):267–276.

Thorne, R. and Williams, W. (1997). The response of benthic macroinvertebrates to pollution in developing countries: a multimetric system of bioassessment. *Freshw. Biol.*, 37:671–686.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., and Altman, R. B. (2001). Missing value estimation methods for dna microarrays. *Bioinformatics*, 17(6):520–525.

UNEP (2004). *Analytical Methods for Environmental Water Quality*. United Nations Environment Programme Global Environment Monitoring System (GEMS)/Water Programme.

U.S.E.P.A (2013). *National Rivers and Streams Assessment 2008-2009*. A Collaborative Survey. EPA/841/D- 13/001, Washington, DC: Office of Wetlands, Oceans & Watersheds and Office of Research & Development.

Usseglio-Polatera, P. and Beisel, J.-N. (2002). Longitudinal changes in macroinvertebrate assemblages in the meuse river: anthropogenic effects versus natural change. *River Research and Applications*, 18:197–211.

Usseglio-Polatera, P., Bournaud, M., Ricoux, P., and Tachet, H. (2000). Biological and ecological traits of benthic freshwater macroinvertebrates: relationships and definition of groups with similar traits. *Freshwater Biology*, 43:175–205.

Van den Broeck, J., Cunningham, S. A., Eeckels, R., and Herbst, K. (2005). Data cleaning: detecting, diagnosing, and editing data abnormalities. *PLoS Med*, 2(10):966–970.

Vlek, H., Verdonschot, P., and Nijboer, R. (2004). Toward a multimetric index for the assessment of dutch streams using benthic macroinvertebrates. *Hydrobiologia*, 516:173–189.

von der Ohe, P., Dulio, V., Slobodnik, J., De Deckere, E., Kühne, R., Ebert, R.-U., Ginebreda, A., De Cooman, W., Schüürmann, G., and Brack, W. (2011). A new risk assessment approach for the prioritization of 500 classical and emerging organic microcontaminants as potential river basin specific pollutants under the european water framework directive. *Science of the Total Environment*, 409:2064–2077.

von Laszewski, G. and Hategan, M. (2005). Java cog kit karajan/gridant workflow guide. Technical report, Technical Report, Argonne National Laboratory, Argonne, IL, USA.

Von Laszewski, G., Hategan, M., and Kodeboyina, D. (2007). Java cog kit workflow. In *Workflows for e-Science*, pages 340–356. Springer.

Vörösmarty, C., McIntyre, P., Gessner, M., Dudgeon, D., Prusevich, A., Green, P., Glidden, S., Bunn, S., Sullivan, C., Liermann, C., and Davies, P. (2010). Global threats to human water security and river biodiversity. *Nature*, 467:555–561.

Wahlin, K. and Grimvall, A. (2008). Uncertianty in water quality data and its implications for trend detection: lessons from swedish environmental data. *Environmental Science and Policy*, 2:115–124.

Wallace, J., Grubaugh, J., and Whiles, M. (1996). Biotic indices and stream ecosystem processes: results from an experimental study. *Ecol.Appl.*, 39:140–151.

Walley, W. and S.Džeroski (1996). *Biological monitoring: a comparison between Bayesian, Neural and Machine Learning methods of water quality classification*. Springer US.

Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244.

Wei, J., Jiang, H., Zhou, K., Feng, D., and Wang, H. (2011). Detecting duplicates over sliding windows with ram-efficient detached counting bloom filter arrays. In *Networking, Architecture and Storage (NAS), 2011 6th IEEE International Conference on*, pages 382–391. IEEE.

Weigel, B., Henne, L., and Martínez-Rivera, L. (2002). Macroinvertebrate-based index of biotic integrity for protection of streams in west-central mexico. *Journal of the North American Benthological Society*, 21(4):686–700.

WHO (2016). *World Health Statistics 2016: Monitoring health for the SDGs, sustainable development goals*. WHO Library Cataloguing-in-Publication Data.

Wiederkehr, J. (2015). *Estimation des incertitudes associées aux indices macroinvertébrés et macrophytes pour l'évaluation de l'état écologique des cours d'eau.* PhD thesis, Université de Strasbourg.

Wiederkehr, J., Grac, C., Fontan, B., Labat, F., Le Ber, F., and Trémolières, M. (2016). Experimental study of the uncertainty of the intrasubstrate variability on two french index metrics based on macroinvertebrates. *Hydrobiologia*, pages 1–15.

Wilson, I. D., Adlard, E., Cooke, M., and C.F., P. (2000). *Encyclopedia of Separation Science*. Academic Press.

Wiseman, A. (2006). Analyzing environmental data. walter w. piegorsch and a. john bailer. john wiley & sons ltd (2005). 512 pp. isbn 0-470-84836-7. *Journal of Chemical Technology and Biotechnology*, 81(8):1447–1448.

WMO (2013). *Planning of water quality monitoring systems, Technical report No.1113*. WMO, Geneva, Switzerland.

Woodiwiss, F. (1964). The biological system of stream classification used by the trent river board. *Chem Ind.*, 14:443–447.

Wright, J., Moss, D., Armitage, P., and Furse, M. (1984). A preliminary classification of running-water sites in great britain based on macro-invertebrate species and the prediction of community type using environmental data. *Freshwater Biology*, 14:221–256.

Wu, J., Lu, J., Wilson, C., Lin, Y., and Lu, H. (2010). Effective liquid–liquid extraction method for analysis of pyrethroid and phenylpyrazole pesticides in emulsion-prone surface water samples. *Journal of chromatography A*, 1217(41):6327–6333.

Yu, J., Bisceglia, K. J., Bouwer, E. J., Roberts, A. L., and Coelhan, M. (2012). Determination of pharmaceuticals and antiseptics in water by solid-phase extraction and gas chromatography/mass spectrometry: analysis via pentafluorobenzylation and stable isotope dilution. *Analytical and bioanalytical chemistry*, 403(2):583–591.

Zhang, S., Jin, Z., Zhu, X., and Zhang, J. (2009). Missing data analysis: a kernel-based multi-imputation approach. In *Transactions on Computational Science III*, pages 122–142. Springer.

Zhao, Y., Raicu, I., and Foster, I. (2008). Scientific workflow systems for 21st century, new bottle or new wine? In *2008 IEEE Congress on Services-Part I*, pages 467–471. IEEE.

Zimek, A., Schubert, E., and Kriegel, H.-P. (2012). A survey on unsupervised outlier detection in high-dimensional numerical data. *Statistical Analysis and Data Mining*, 5(5):363–387.