Principled Data Preprocessing : Application to Biological Aquatic Indicators of Water Pollution

Eva C. Serrano Balderas, Laure Berti-Equille, Ma. Aurora Armienta Hernandez, Corinne Grac

IRD, Institut de Recherche pour le Développement, Espace DEV
Institute of Geophysics, National Autonomous University of Mexico, UNAM
ENGEES, Ecole Nationale de Génie de l'Eau et de l'Environnement de Strasbourg

Project presentation

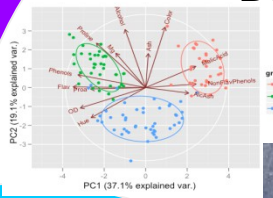
Multidisciplinary approach
Bi-national between Mexico and France

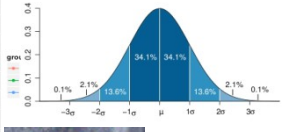
Statistics & Computer Science

IRD, Montpellier
Dr. Laure Berti-Equille

Stay of 12 months

Data preprocessing
Data quality control
Data Mining





Environmental Chemistry

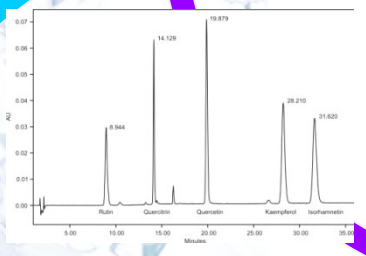
Hydrobiology

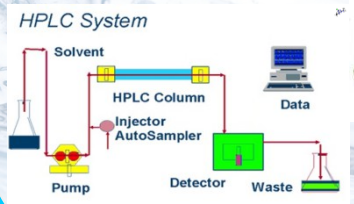
UNAM, Mexico
Dr. Ma.Aurora Armienta Hernández

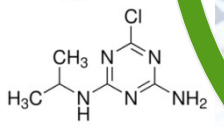
Stay of 18 months

LIVE, Strasbourg
Corinne Grac

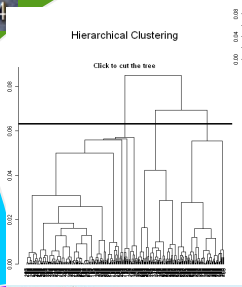
Stay of 6 months

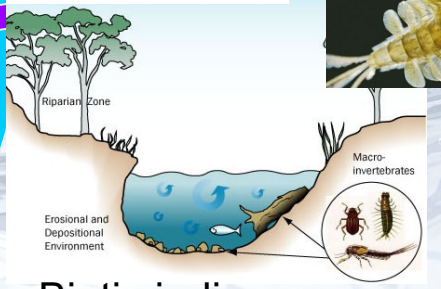






Chemical &
Physicochemical analysis





Biotic indices
Macroinvertebrates

Introduction and Motivation

- In the knowledge discovery process, data preparation is a crucial step
- An appropriate data preparation will assure the accuracy and reliability of analysis results
- Imprecision and uncertainty can lead to misinterpretations and wrong conclusions
- Standardized protocols for data preprocessing have not been established
- It is necessary to establish data preparation and data quality validation procedures in order to ensure reliable results

Application to environmental data

Hydro-biological data

Chemical and Physicochemical data

Assess the impact of emerging pollutants on Mexican rivers

General objectives

- To define the reliability of the results obtained during environmental evaluations
- To specify the data treatment procedures necessary to obtain accurate results
- To evaluate the pollution of Mexican streams by heavy metals, metalloids, pharmaceutical products, pesticides and halogenated organic compounds
- To determine the ecological state of rivers by using macro invertebrates

Specific objectives

Environmental Chemistry

- To evaluate the impact of the industrial, agricultural and urban activities on the water quality of the rivers Tula, Taxco, Toliman, Culiacan and Humaya
- To analyze the content of the ions Ca^{+}, Mg^{+}, Na^{+}, F^{-}, Cl^{-}, K^{+}, of nitrates and heavy metals (Pb, Zn, Cu, Mn, Cd, Fe and As) on the rivers
- To adapt the Solid Phase Extraction (SPE) and the High-Performance Liquid Chromatography coupled to an Ultra-Violet detector (HPLC) for the analysis of pesticides and hormones in liquid samples

Specific objectives

Hydrobiology

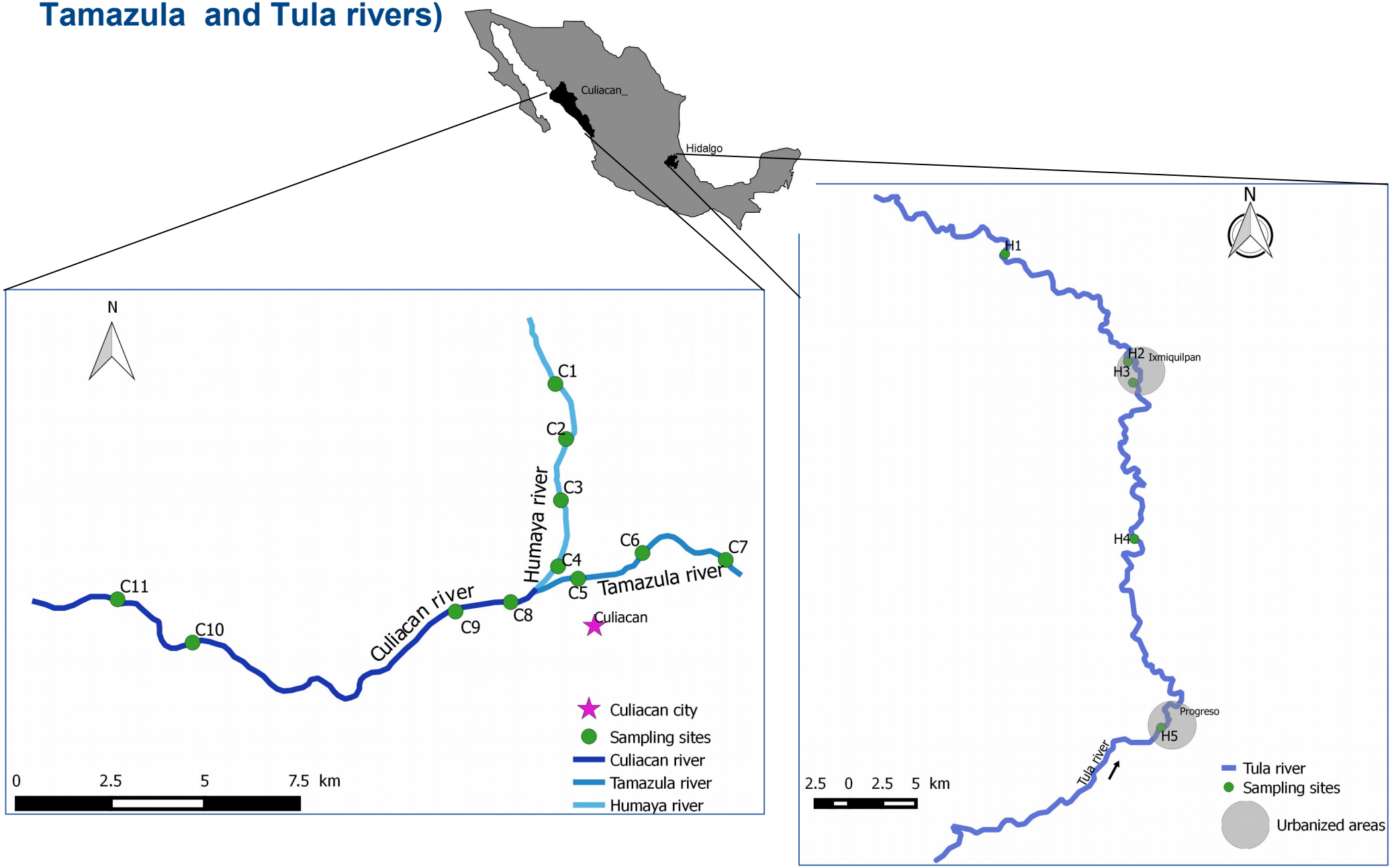
- To define the correlation between macro invertebrates and the ecological state of the rivers
- To define the correlation between macro invertebrates and the presence of emergent pollutants on the rivers
- To adapt the existing and commonly-used sampling and analytical methods for the identification of macro invertebrates in Mexican streams
- To identify the macro invertebrates-based biotic metrics most suitable for the assessment of the ecological status of Mexican streams
- To define the utility of macro invertebrates as complementary tool in the evaluation of the pollution on Mexican rivers

Specific objectives

Statistics and Computer Science

- To define the chain of procedures necessary to clean, prepare and analyze data with control over the quality of data
- To examine the utility to data quality evaluation on the interpretation of environmental analysis results
- To generalize the approach of control and improvement of the quality of data to other data collection procedures applied to the context of the monitoring of pollution of the environment in Mexico and France

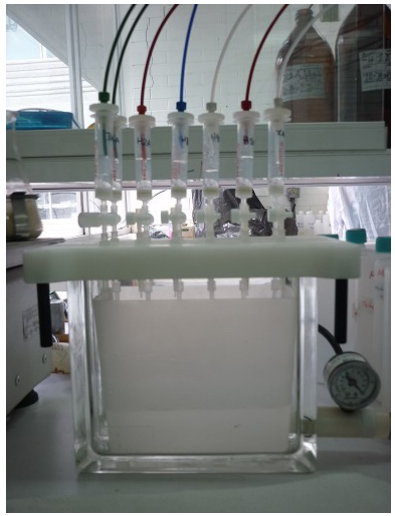
Achievements 2014-2015: Data acquisition (Culiacan, Humaya and Tamazula and Tula rivers)



Achievements 2014-2015: Data acquisition

Sampling of liquid and biological samples

Analysis of samples



Solid Phase Extraction (SPE)

Coenagrionidae

Velidae

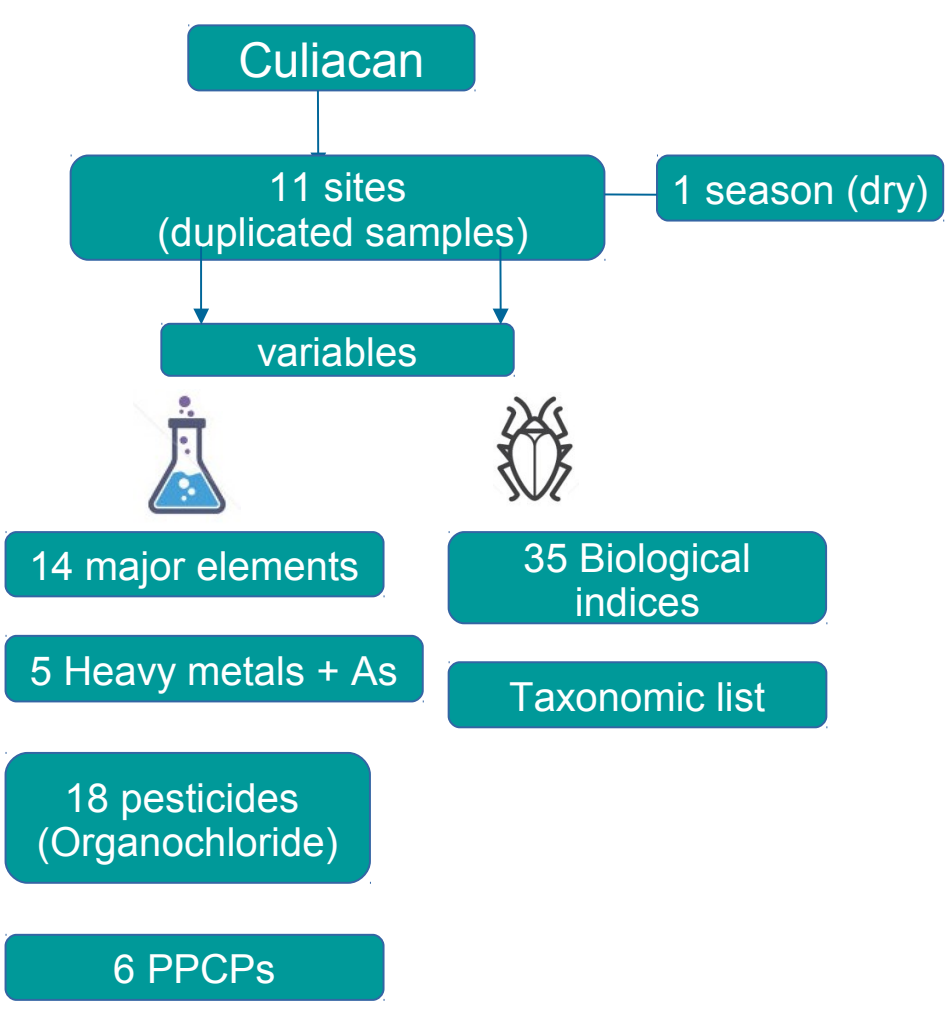
Corixidae

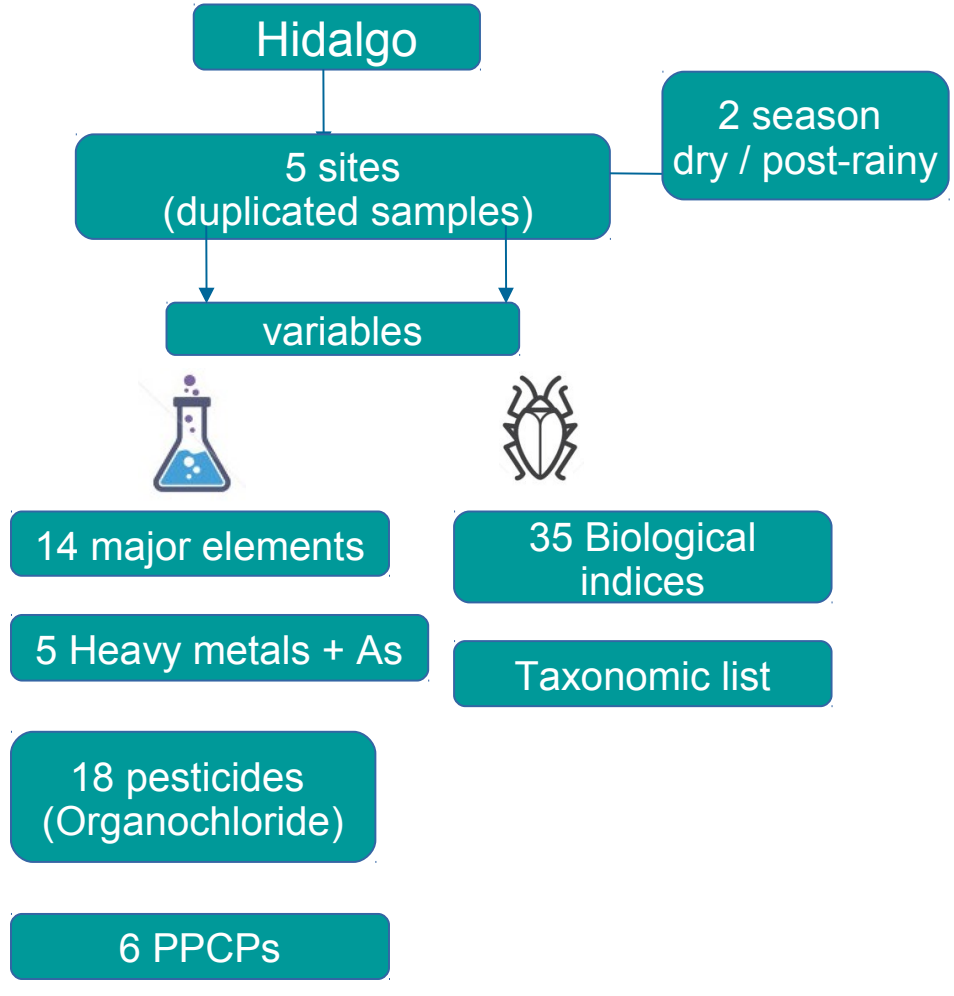
Physidae

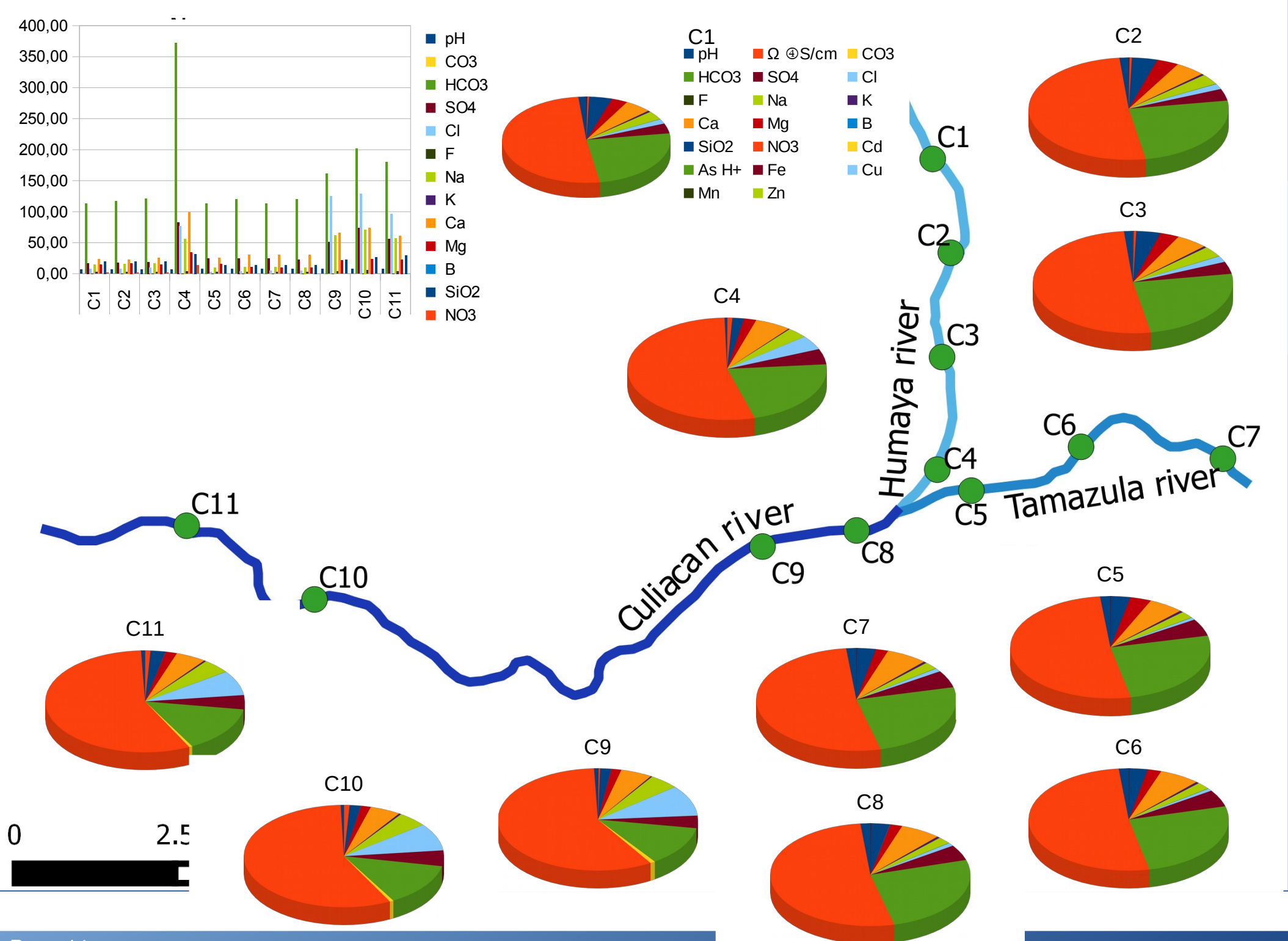
Pyralidae

Achievements 2014-2015: Data acquisition

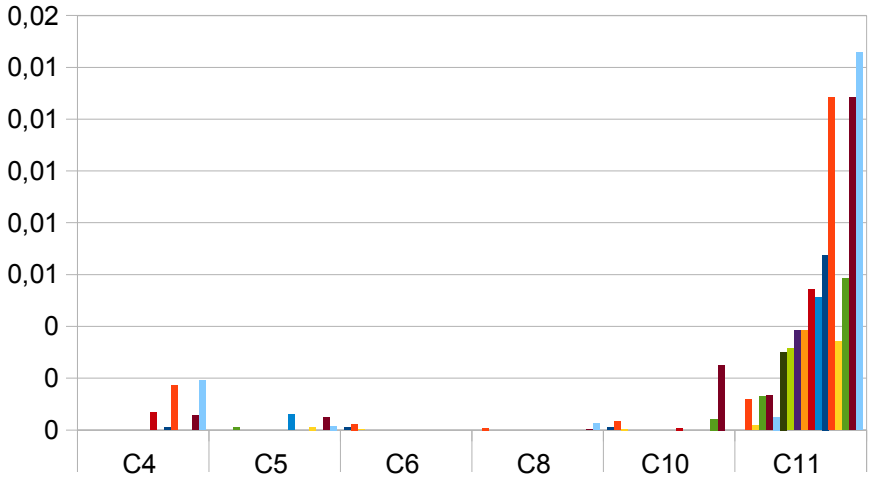
Description of data

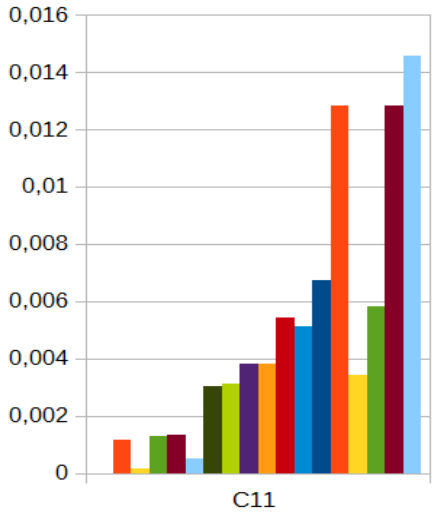


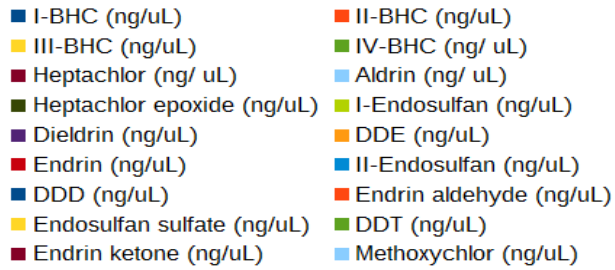


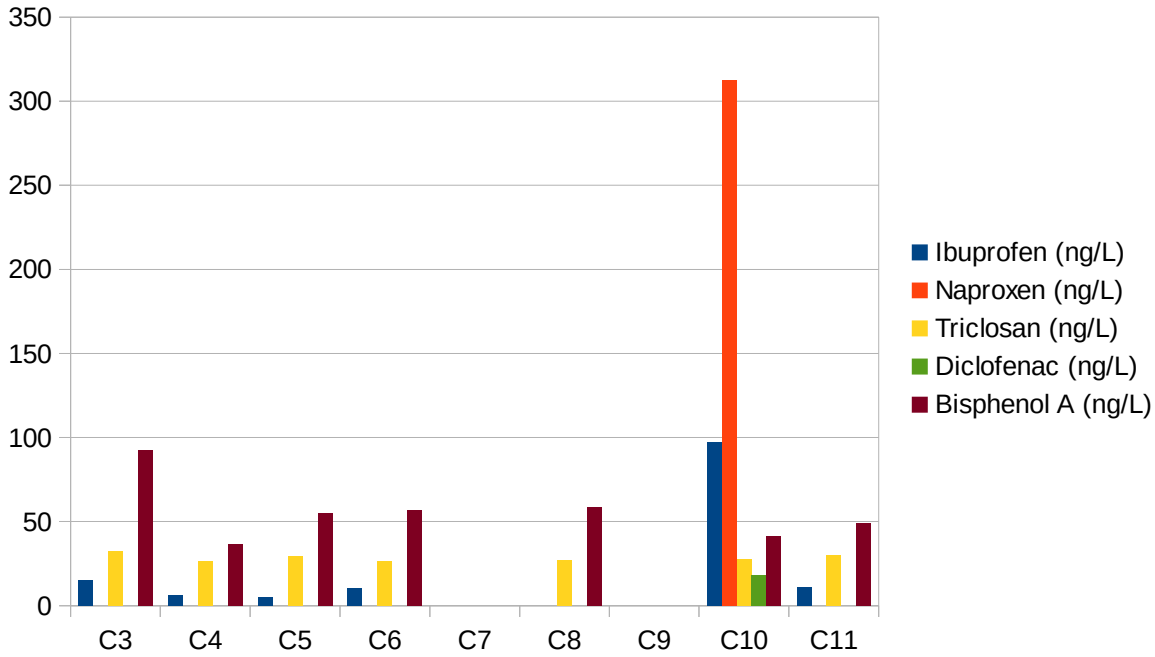

400,00
350,00
300,00
250,00
200,00
150,00
100,00
50,00
0,00
C1
C2
C3
C4
C5
C6
C7
C8
C9
C10
C11
pH
CO3
HCO3
SO4
Cl
F
Na
K
Ca
Mg
B
SiO2
NO3
C1
pH
Ω ④S/cm
CO3
HCO3
SO4
Cl
F
Na
K
Ca
Mg
B
SiO2
NO3
Cd
As H+
Fe
Cu
Mn
Zn
C2
C3
C4
C1
C2
C3
Humaya river
C4
C6
C7
C11
C10
Culiacan river
C9
C8
C5
Tamazula river
C5
C11
C7
C9
C6
C10
C8
0
2.5

Achievements 2014-2015: Preliminary analysis (Culiacan)



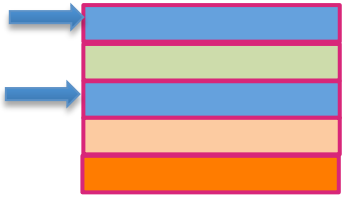




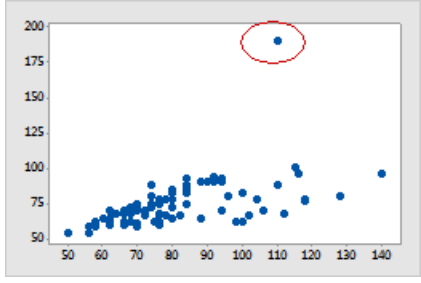


Some facts on data from biomonitoring surveys

- **Prone to anomalies :**
  - Duplicates
  - Inconsistencies
  - Outliers
  - Incomplete and missing data



pH = 24



C
3.2
6.3
NA

Bad data quality

➡ **Inaccurate and biased test results**

➡ **Wrong diagnosis and wrong interpretations**
(Wahlin and Grimvall, 2008)

Quality data is needed

Data processing to mitigate the impacts of data anomalies

1. **Preprocessing to correct data and improve data quality**
2. **Preparation of corrected data before data analysis**

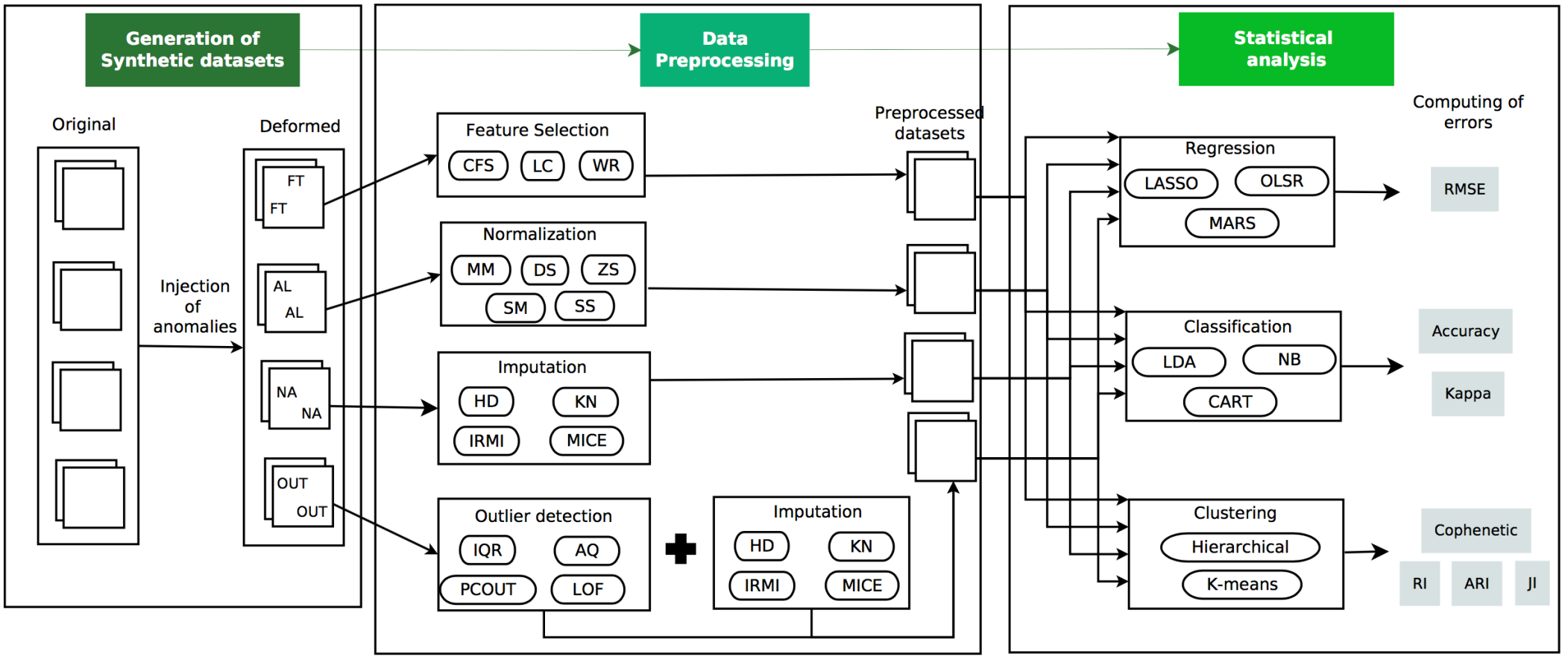
Limitations

- Few studies on the appropriate selection of procedures to preprocess and prepare data
- Little work on data from biomonitoring surveys

Objectives

- Propose a methodological framework to guide on the orchestration of data preprocessing tasks and the selection of the most adequate methods
- Provide a comparative study to make a better selection of preprocessing procedures
- Quantify the bias introduced by a preprocessing strategy

Principled Approach for Data Preprocessing

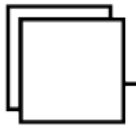


Principled Approach for Data Preprocessing

Step 1. Generation of synthetic and semi-synthetic data

1 Control of anomalies ⇩ Calculation of bias

Original

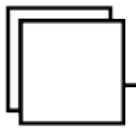


- N = 21, V = 8
- N = 600, V = 30
- N = 4000, V = 53
- N = 20000, V = 98

- 4 numerical datasets
- Simulation of biomonitoring data
- Multivariate data
- Normal distribution

- Generation of semi-synthetic data

Original



- N = 16, V = 13
- N = 1504, V = 26
- N = 7520, V = 26

- Generation from real dataset (Rhin-Meuse rivers, France)
- Elimination of anomalies
- 3 numerical datasets

Principled Approach for Data Preprocessing

- **Step 2. Data deformation**

2 Control of anomalies

- Creation of data distribution
  - Non-normalized
  - No relevant variables

→ Weibull Distribution
→ Strong correlation

- Random Injection
  - Missing data
  - Outlying data

Different amount of injection

5%	10%	15%	20%	25%	30%

0,1%	0,5%	5%	10%	15%

→ Procedure replicated X 10 times

Step 3. Data preprocessing (1/3)

Preparation

3.1 Normalization

- Min-Max
- Zero-mean
- Decimal scale

More frequently used methods

Different techniques

3.2 Feature selection

- Correlation-based : Ranks based on correlations
- Linear correlation : Ranks based on the linear correlations according to a certain threshold
- Wrapper subset evaluator : Evaluation of different data subsets

Principled Approach for Data Preprocessing

Step 3. Data preprocessing (2/3)

Correction

Imputation of missing values

- Hot-Deck: replacement by an observed value
- k-NN : replacement by the nearest value
- Multiple Imputation
  - MICE (*Multiple Imputation by Chained Equations*)
  - IRMI (*Iterative Step-wise Regression Multiple Imputation*)

Principled Approach for Data Preprocessing

Step 3. Data preprocessing (3/3)

Correction

3.4 Outlier detection + imputation

- Inter Quartile Range (IQR)
- Adjusted-Quantile
- Principal Component Decomposition (PCOUT)
- Local Outlier Factor (LOF)

- Hot-Deck
- k-NN
- MICE
- IRMI

Different techniques
- univariate detection (IQR)
- multivariate detection

→ Outliers were treated as missing values

Principled Approach for Data Preprocessing

Step 4. Comparative study and calculation of bias

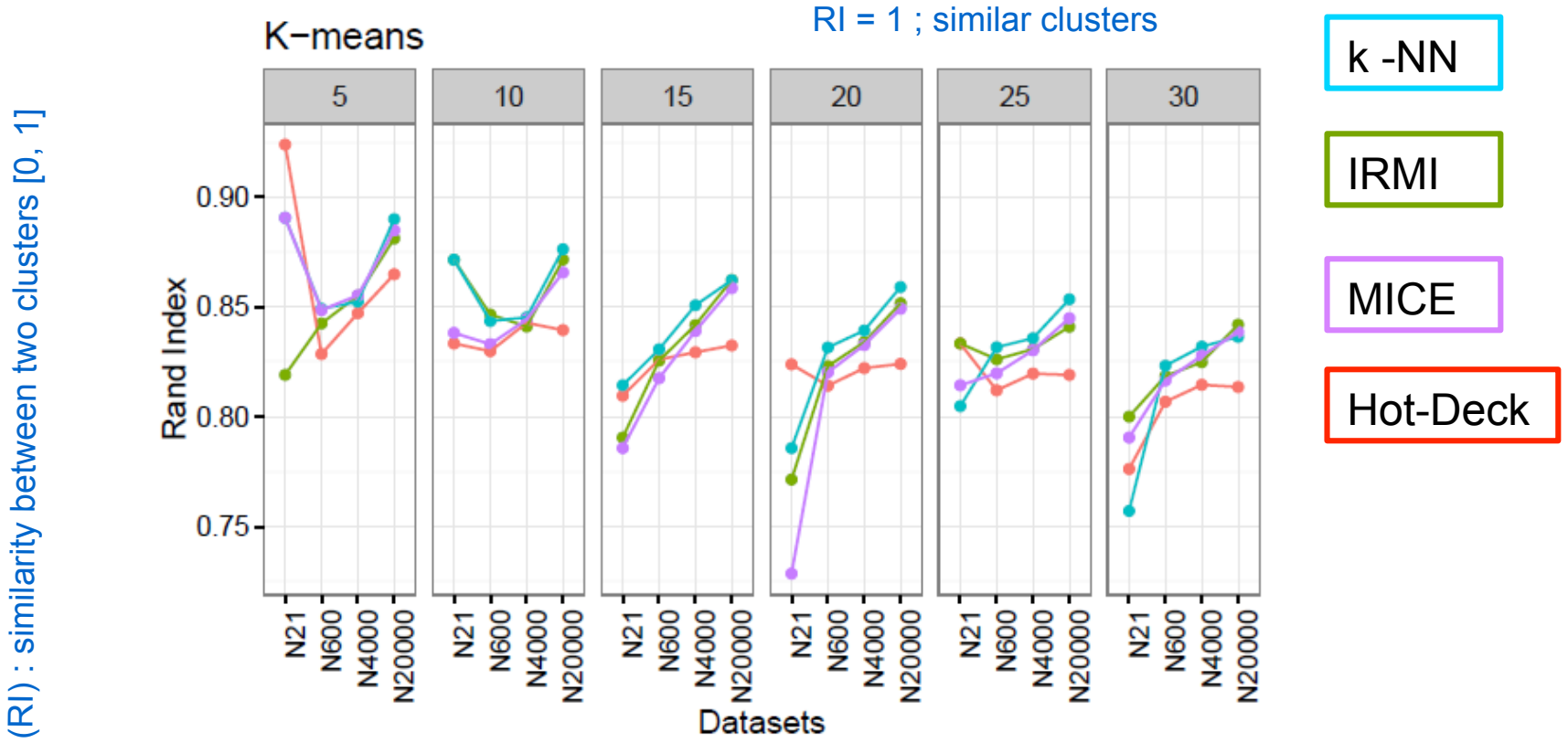
- Regression — RMSE
- Classification — Precision, Kappa
- Clustering — Rand Index, Adjusted Rand Index, Jaccard Index

Original data vs Pre-processed data

Focus on a sample of results

- K-means Clustering
- Anomaly : Missing values

1) Imputation
2) Rank of the pre-treatment



Classification of the methods is specific to the characteristics of the data and the type of anomalies

CASE STUDY

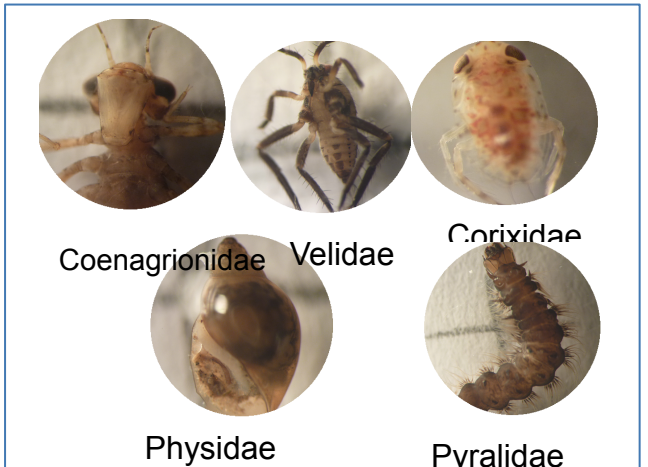
Biological Aquatic Indicators of Water Pollution

Water quality assessment using:

- **43** Physical-chemical variables

- pH
- Metals
- Nitrates
- Phosphates
- Pesticides, etc.

- **35** Biological indicators based on the presence/ absence of aquatic macroinvertebrates



- Measures of richness and enumerations
- Diversity and similarity indices
- Biotic indices
- Functional feeding groups measures
- Multimetric approach

CASE STUDY

Biological Aquatic Indicators of Water Pollution

Data preprocessing

Missing values : 9.7 % → MICE

Outliers (IQR) : 7.61 % → k-NN

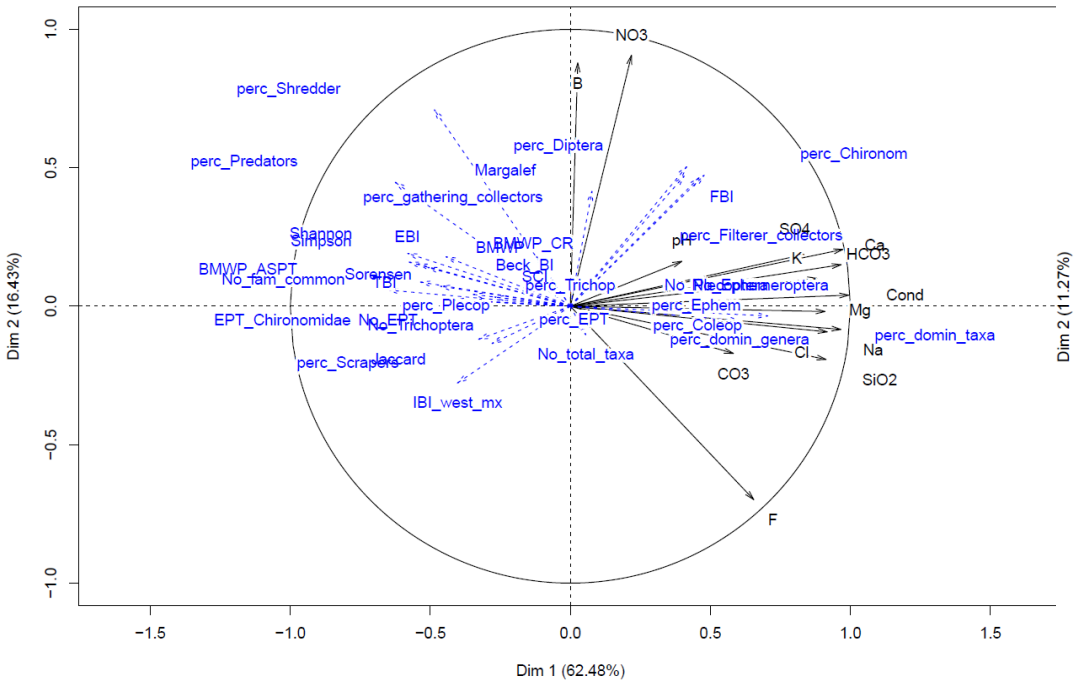
Non-normalized data → Z-score

- Selection of preprocessing strategies :

  → According to the results of the comparative study of preprocessing procedures

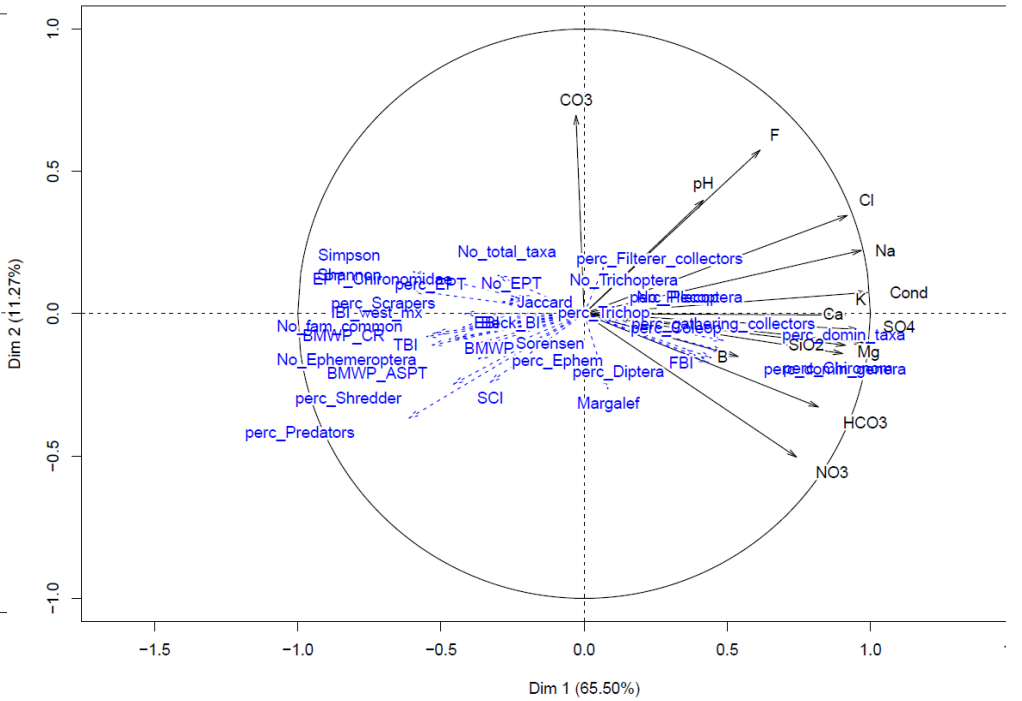
Relevance of the proposed approach (1/2)

Preprocessed data



Physical-chimical degradation →

Non preprocessed data



Physical-chimical degradation →

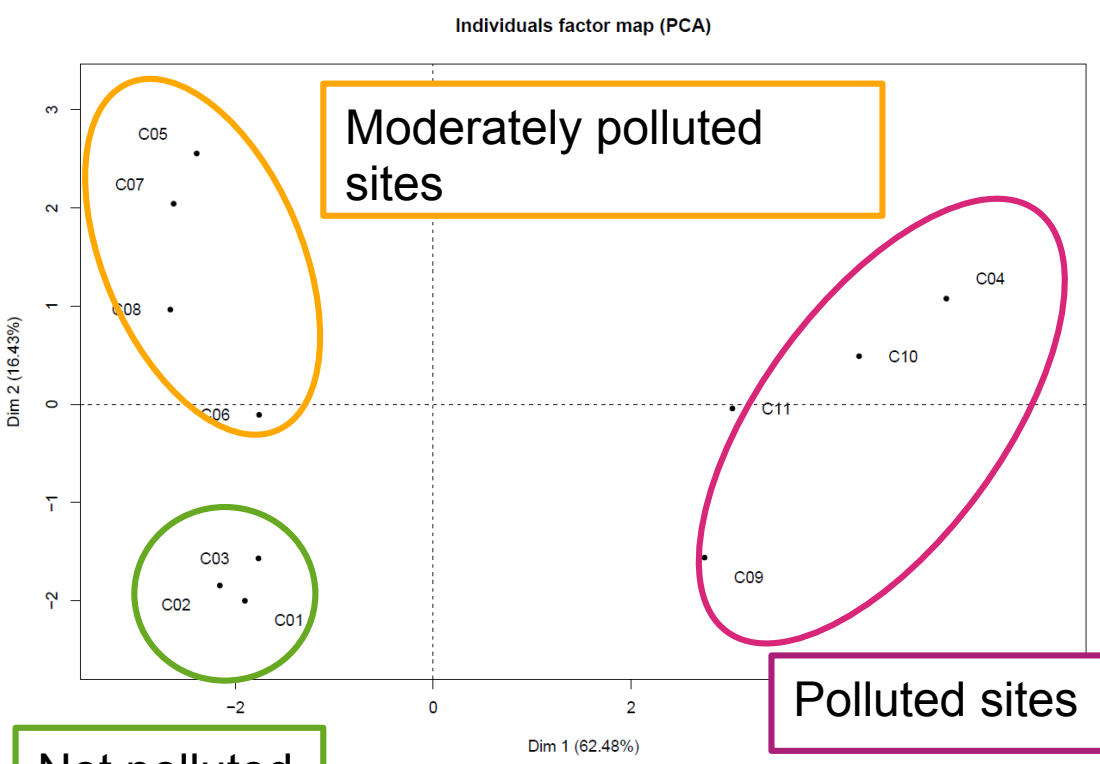
Results with similar trend

→ But different distribution of variables

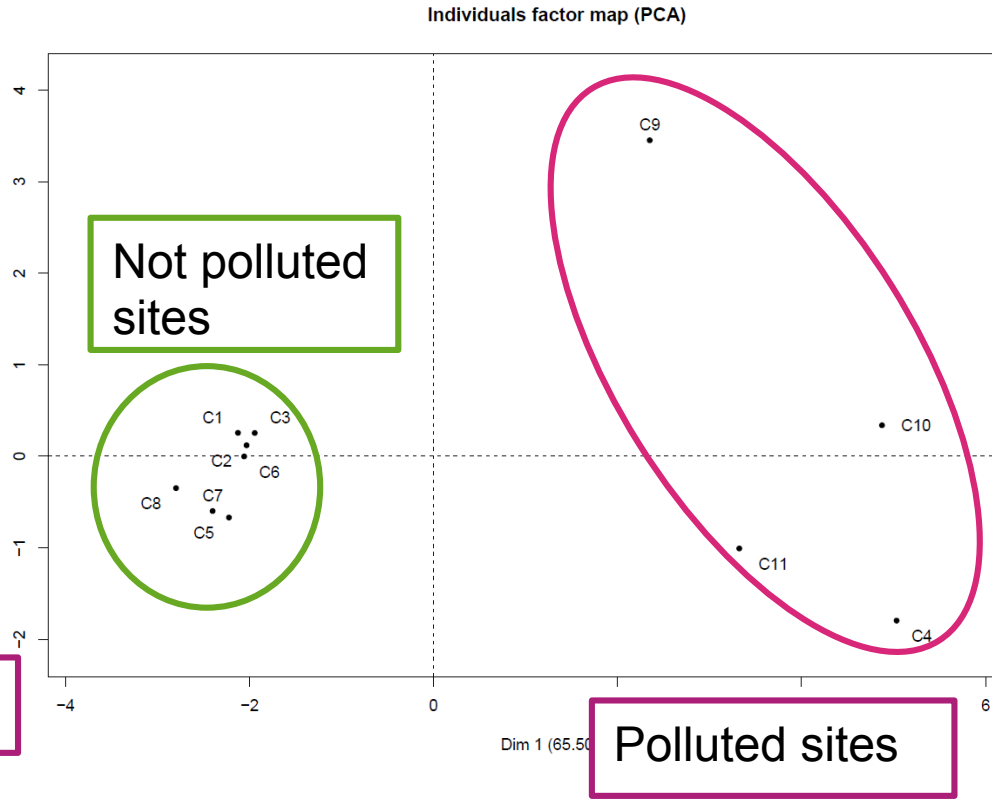
Relevance of the proposed approach (2/2)

Preprocessed data

➔ Better classification of sites



Non preprocessed data



Conclusions

- There is not a single universally adequate data preprocessing procedure (as it is dataset-dependent)
- Effectiveness of preprocessing decreases with the percentage of anomalies
- An instantiation of the proposed framework can be effectively used to mine the hidden knowledge from biomonitoring data
- Next : Leveraging machine learning to find the optimal data preprocessing strategy for any given dataset

Thank you for your attention

Publications

Revues internationales

- L. Berrahou, N. Lalande**, E. Serrano**, G. Molla, L. Berti-Equille, S. Bimonte, S. Bringay, F. Cernesson, C. Grac, D. Ienco, F. Le Ber, M. Teisseire, 2015. A quality-aware spatial data warehouse for querying hydroecological data, Computers & Geosciences, 85 : 126-135.

- **E. C. Serrano Balderas**, C. Grac, L. Berti-Equille, M.A. Armienta Hernandez. Potential application of biological indices based on macroinvertebrates on Mexican streams. Ecological Indicators, 61 : 558-567.

Conférences avec comité de lecture

- **E. C. Serrano Balderas**, L. Berti-Equille, M.A. Armienta Hernandez, J-C. Desconnets, « Water Quality Data Analytics» iEMSs (International Environmental Modelling and Software Society, Juin 2016. Toulouse, France.

- **E. C. Serrano Balderas**, L. Berti-Equille, M.A. Armienta Hernandez, C. Grac, « Impacts of data Quality on Environmental Analysis : Application to Mexican Rivers Pollution » WomENcourage (ACM-W womENcourage Celebration of Women in Computing), Mars 2014. Manchester, UK.

- **E. C. Serrano Balderas**, C. Grac, L. Berti-Equille, « Data processing for controlling data quality on surface water quality assessment » Systèmes d'Information pour l'environnement. INFORSID Mai 2014. Lyon, France.