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

Eva C. Serrano Balderas Institut de Recherche pour le Développement
UMR ESPACE DEV. Montpellier, France. Email: eva.serrano@ird.fr
Laure Berti-Equille Institut de Recherche pour le Développement
UMR ESPACE DEV. Montpellier, France. Email: laure.berti@ird.fr
Ma. Aurora Armienta Hernández Geophysics Institute
National Autonomous University of Mexico. Mexico city, Mexico. Email: victoria@geofisica.unam.mx
Corinne Grac LIVE, Laboratoire Image, Ville Environnement
Université de Strasbourg/ENGEES. Email: corinne.grac@engees.unistra.fr

***Abstract*—In many biological studies, statistical and data mining methods are extensively used to analyze the data and discover actionable knowledge. But, bad data quality causing incorrect analysis results and wrong interpretations may induce misleading conclusions and inadequate decisions. To ensure the validity of the results, avoid bias and data misuse, it is necessary to control not only the whole analytical pipeline, but most importantly the quality of the data with appropriate data preprocessing choices. Since various preprocessing techniques and alternative strategies may lead to dramatically different outputs, it is crucial to rely on a principled and rigorous method to select the optimal set of data preprocessing steps that depends both on the input data distributional characteristics and on the inherent characteristics of the targeted statistical or data mining methods. In this paper, we propose a method that selects, given a dataset, the optimal set of preprocessing tasks to apply to the data such that the overall data preprocessing output maximizes the quality of the analytical results for various techniques of clustering, regression, and classification. We present some promising results that validate our approach on biomonitoring data preparation.**

Index Terms— Biological Data Preprocessing, Data cleaning, Biomonitoring Data.

I. Introduction

Through the use of data mining (DM) and statistical methods, data scientists and researchers can discover relevant patterns from the data and gain crucial and actionable knowledge. However, it is necessary to adapt the volume, format and distributional characteristics of the input data to better suit the underlying assumptions and constraints of the DM and statistical methods to be applied. As data analysis input, it is essential to ensure quality data because erroneous data limit the performance of statistical methods induce misleading analytics and data mining results [4], [21] and finally, lead to faulty conclusions, costly decisions, and dramatic consequences.Data evolving over time can be big, noisy, unreliable, highly imbalanced, and heterogeneous. Biological data – both human- or environment-centered– are not an exception to this observation as shown in various studies in biomedical and environmental domains [2], [3], [14]. In environmental sciences, water pollution problems have boosted environmental research activities going from general to more detailed

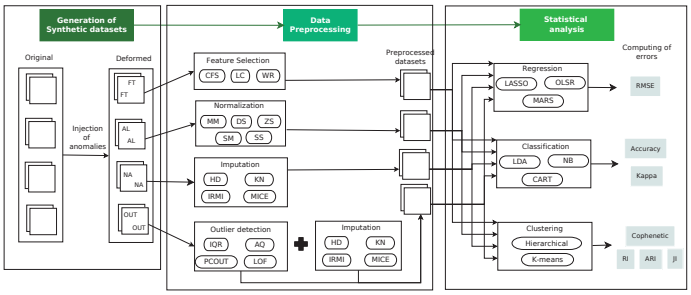

Fig. 1: Principled Approach for Data Preprocessing

measurements such as the use of biological aquatic and marine indicators. Biomonitoring metrics based on macroinvertebrates are commonly used as indicators of pollution on aquatic ecosystems. They are low-cost and easy-to-implement tools that provide valuable information about the ecological state of aquatic ecosystems [5]. Notwithstanding, data from biological survey of aquatic ecosystems may have a wide range of data anomalies such as missing values, inconsistencies, outlying data, duplicates, etc. Analysis results from data preprocessed inappropriately may lead to misleading conclusions and inadequate decision-making that can affect the survival of some species and even endanger some environmental ecosystems. To mitigate the impacts of data anomalies and thus provide quality analytical results, a first necessary step is to preprocess data appropriately [13].

Data preprocessing has been acknowledged as a primary sequence of tasks to correct the negative effects and bias that data anomalies may produce on analytical results.Common data preprocessing tasks mainly include: selection of features, data normalization, and treatment of missing values and outliers. Although there is a vast number of data preprocessing methods available to accomplish these tasks, the selection of an optimal strategy remains *ad hoc* difficult. To our knowledge, no rigorous methodological framework exists yet to guide, in a principled way, the orchestration of data preprocessing tasks and the selection of the most adequate methods. This topic is still an interesting, open research area. The focus of our work

is precisely to study the impact on the analysis results of the preprocessing methods used for selecting features, normalizing data, and dealing with missing values and outliers since these methods are the most frequently used in the analysis of biomonitoring data, and even more generally, beyond this particular application domain.

II. Overview of Our Approach for Principled Data Preprocessing

Our overall approach can be summarized in Figure 1. Given a dataset as input and a targeted analysis methods (e.g., clustering, regression, or classification), a clean dataset sample is first generated with similar distributional characteristics and the targeted statistical analysis is performed. Then, anomalies are injected with the same proportion and distribution as in the original input dataset; various data preprocessing techniques are applied including: feature selection, normalization, imputation of missing values, and outlier processing and the same targeted statistical analysis is performed again over the preprocessed data. Finally, the two sets of results are compared based on various evaluation parameters and the data preprocessing techniques that return the least difference in accuracy with respect to the original clean dataset sample is selected.The full pipeline is automated and was performed using different packages from the R environment for statistical computing;

A. Learning from Semi-Synthetic Data Samples

Semi-synthetic datasets are generated in order to learn the latent characteristics that will lead the choice of the preprocessing method that can optimize the quality of the final result for a given analysis method.

In our use case, the synthetic dataset generation is designed to be the most similar to our biological aquatic data as in water quality surveys in terms of distribution, correlation, etc. In our approach, we created clean datasets that we named "original"; then, we polluted each original dataset by injecting anomalies (i.e., missing values and outliers) or by voluntarily deforming the original characteristics of the datasets. This procedure was replicated 10 times to obtain different polluted datasets for a better control of the randomization of our experiments. All generated semi-synthetic datasets were flagged for better management on subsequent treatment. Statistical analysis were then conducted both on the clean and the polluted versions of the datasets and the results were compared. Typically, the most adequate preprocessing strategy would lead to the most similar analytical results that could be obtained from the original dataset.

Four synthetic datasets have been used to assess the impact of feature selection, normalization and imputation procedures on regression, classification, and clustering analysis. Each dataset follows a normal distribution with varying numbers of observations ($n = 21, 600, 4000,$ and 20000) and numerical non-correlated variables ($p = 8, 30, 53, 98$), plus one categorical variable uniformly distributed into five categories.

Missing data were introduced randomly (based on MCAR) in the numerical variable domains, for each dataset either with varying the rate from 5%, 10%, 15%, 20%, 25% to 30% of missing values or in the same proportions as in the original dataset. For each missing value rate, we replicated 10 times the generation of the datasets.

Similarly, outlying values were introduced randomly in each of the polluted version of the datasets with the following rates (replicated 10 times): 1.5%, 2.5%, 5%, 10% and 15% or in the same proportions of the original dataset.

B. Data preprocessing methods

The focus of our study was on four types of preprocessing techniques including: (1) feature selection for data reduction, (2) data normalization, (3) imputation methods to handle missing, and (4) outlier detection and replacement. Concerning feature selection, three methods were tested to select an optimal data subset, namely: correlation-based feature selection (CFS; [15]), linear-based correlation (LC; [9]), and wrapper subset evaluator (WR; [16]). One variable was randomly chosen as independent variable. Then, using each feature selection method, we identified the best data subset with respect to the pre-selected independent variable; Concerning normalization, three methods were applied to numerical data: min-max (MM), Z-score (ZS) and decimal scale normalization (DS); Four imputation methods were used to impute missing values from the following categories: two distance-based imputation methods: *Hot-deck* (HD [20]) and *K-NN* ([1]), and two model-based imputation methods: Multiple Imputation by Chained Equations (*MICE* [7]) and Iterative Robust Model-based Imputation (*IRMI* [19]); Outliers were detected using four methods including: an statistic-based approach (Inter Quartile Range, IQR), two multivariate outlier detection approaches (Adjusted-Quantile [11]), an algorithm using Principal Components decomposition (PCOUT [12]), and a density-based approach (Local Outlier Factor, LOF [6]). Finally, outliers were replaced using one of the previous imputation methods *Hot-Deck*, *k-NN*, *MICE* or *IRMI*.

C. Statistical Analysis

Our goal is to study how the way data are preprocessed can affect the results of statistical analysis. To assess the impact of data preprocessing on statistical analysis, we studied:

- Three regression methods: LASSO (Least Absolute Shrinkage and Selection Operator), OLSR (Ordinary Least Squares Regression), and MARS (Multivariate Adaptive Regression Splines);
- Three classification methods: LDA (Linear Discriminant Analysis), NB (Nave Bayes), and CART (Classification and Regression Trees); and
- Two clustering methods: HCA (Hierarchical Clustering) and K-means.

D. Evaluation Parameters

Next, in order to estimate preprocessing bias, we have compared the statistical results obtained from each preprocessed dataset variants by computing different parameters such as

statistical errors as follows. For regression methods, the observations were split into a training and test set where 66% of data was used for training and 34% for testing. The regression model was fit to the training set, and the fitted model was used to predict the responses for the observations in the testing set. The resulting validation was assessed using differences in RMSE (Root Mean Square deviation). We estimated the preprocessing error rate by comparing the RMSE value of the original non-polluted dataset and the preprocessed dataset.

Similarly, for classification methods, the observations were randomly split into a training and test sets with data amounts of 66% and 34% respectively. Classification models were fit to the training set, and the fitted model was used to predict the responses for the observations in the testing set. Resulting validation of the previous step was assessed through computation of accuracy, and Cohen's Kappa coefficient.

Clustering analysis was performed using K-means (KM) and Hierarchical clustering (HC) methods. To perform K-means clustering, we first specified a number of clusters K using the Elbow method. 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 K-means algorithm with the previously specified number of clusters. Hierarchical clustering was computed using the Ward's agglomeration method. Clustering results of K-means on preprocessed data were compared against the clustering results of the original non-deformed datasets using the Adjusted Rand Index (AR) and Jaccard Index (JI).

III. Results

A. Results for regression analysis

Results obtained on the synthetic datasets processed by feature selection showed that in general for semi-synthetic biomonitoring data, Linear correlation-based feature selection has 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.

With respect to datasets processed by imputation of missing values, we observed that in general, for a high number of missing values (25% and 30%), *Hot-deck* and *MICE* imputation methods give the lowest error values. 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%.

For outlier preprocessing, we observed that for a small number of outlying value (1.5% and 2.5%) and small dataset (e.g., $n < 100$ observations), the combined methods PCOUT-Hot-Deck and PCOUT-IRMI give the lowest preprocessing bias. While for large number of outliers (5%, 10%, or 15%) and large datasets (e.g., $n > 100$ observations), PCOUT and LOF outlier detection methods combined with imputation methods *MICE* and *IRMI* give the lowest preprocessing bias. We noticed that, in general, for large number of outliers (10% and 15%) preprocessing biases were significantly higher. For the specific characteristics of our synthetic datasets, we observed that, in general, multivariate methods give the best results, particularly on large datasets (i.e., $n > 400, p > 53$) with large number of missing values and outliers. We also observed that, simple methods (e.g., K-NN imputation method or Inter Quartile Range) provide the best results on our small datasets with small numbers of data anomalies. Tables I and II show respectively the preprocessing RMSE after imputation of missing values. The imputation method with the lowest preprocessing error values was considered as the method that impacted the least the statistical analysis results.

TABLE I: Data preprocessing study results on regression.

Missing rate (%)	LASSO		OLSR		MARS	
	Method	RMSE	Method	RMSE	Method	RMSE
5	HD	21.235	HD	42.840	HD	0
5	IR	1.601	IR	0.358	IR	0
5	KN	16.384	KN	0.274	KN	2.186
5	MI	1.052	MI	1.380	MI	0
10	HD	92.128	HD	0.460	HD	3.501
10	IR	3.798	IR	0.503	IR	1.098
10	KN	25.354	KN	0.617	KN	0.290
10	MI	82.562	MI	0.794	MI	8.417

B. Results for classification

Accuracy results on our synthetic datasets showed that the filtering method Correlation-based feature selection can give the most accurate results on all datasets for the three classifiers. Concerning Kappa, the best results for LDA were observed for the Correlation-based feature selector, whereas 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. CART is known to be highly non-robust, this explains its behavior. 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).

Concerning normalization, we observed that decimal-scale normalization provides the best results for dataset with $n < 100$ on the LDA and NB classifiers, while for datasets with $600 < n < 4000$, the min-max and z-score normalizations stand up as the best for LDA and NB. Except for the datasets $n < 4000$, the z-score normalization gives the best results when combined with CART classifier.

From Kappa results, we observed that, for small datasets decimal-scale and z-score give the best results. While for larger datasets (e.g., $n > 1000$), there is no clear winner. Our results suggest that the selection of a normalization method may provide dramatically different classification results. We assume that the difference with our results is due to the characteristics of our datasets (e.g., distribution, size) and on the differences of the learning style of the three classifiers.

For missing values preprocessing results show that, in general, for small datasets ($n < 21$), imputation methods *Hot-deck* and *k-NN* give low preprocessing error rates at small amounts of missing values (5%, 10%, and 15%). For large datasets ($n > 1000$) imputation methods IRMI and MICE show the lowest preprocessing error values in both, accuracy and Kappa, for the six amounts of missing values. Concerning the preprocessing error results for Kappa, we observed that small datasets ($n < 21$) and small amounts of missing values

(5%, 10%, and 15%) the imputation method *Hot-deck* have the lowest error values on the three classification methods for the four 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.

TABLE II: Data preprocessing study results on classification.

Missing	CART		LDA		NB	
rate (%)	Method	RMSE	Method	RMSE	Method	RMSE
5	HD	0	HD	0.200	HD	0
5	IR	0.200	IR	0.200	IR	0.200
5	KN	0	KN	0	KN	0.200
5	MI	0	MI	0.400	MI	0.200
10	HD	0.400	HD	0.200	HD	0.200
10	IR	0.200	IR	0	IR	0.200
10	KN	0.400	KN	0.200	KN	0.400
10	MI	0	MI	0	MI	0.200

C. Results for clustering

With respect to feature selection preprocessing applied before clustering, we observed that the Linear Correlation-based feature selection show the best result on both clustering methods. Concerning missing values preprocessing, 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 ($n < 21$) 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. We observed that the multivariate imputation methods MICE and IRMI stand up as the best since our datasets are multivariate.

For outlying data preprocessing, no universally best method to detect and impute outliers was observed. The impact of detection-imputation of outliers varies for the two clustering methods and for the synthetic datasets.

Our comprehensive study corroborates previous works [10] [18] such that the data characteristics (i.e., distribution, skewness, kurtosis, etc.) along with the assumptions of the statistical method at hand play a critical role in the selection of the adequate preprocessing methods. For instance, in order to make a prediction, – CART method may provide some improvements over - LDA method because CART is a non-parametric approach (i.e., no assumptions are made about data distribution) while - LDA classification method assumes that observations are drawn from a multivariate normal distribution.

IV. Case study: Biological indicators of water pollution

By applying our approach we aim at responding to a concrete applicative need in the context of water quality assessment. Precisely, our goal is to identify the most appropriate biological indicators to assess water quality of rivers. To do so, we have collected data that describe the physico-chemical, chemical and biological characteristics of four Mexican rivers (Tula, Humaya, Tamazula and Culiacan). The data include: 20 parameters (macro-pollutants) that are compounds naturally present in the rivers, necessary for the aquatic ecosystem; micro-pollutants that are compounds that do not occur naturally in the rivers (e.g., pesticides, pharmaceutical products), and biological data. Biological data are descriptions of the biological organisms (flora and fauna) living in rivers. We have selected macroinvertebrates in order to compute biological indices to assess the quality of the aquatic ecosystems of the Mexican rivers. Macroinvertebrates have been inventoried and the obtained list of taxa has been used to compute different indices that provide information about the diversity, fauna richness, and quality characteristics of the aquatic ecosystem. A total of 35 indices were computed as proposed in [17]. Our real-world dataset has a total of 78 numerical variables that include 20 macro-pollutants, 23 micro-pollutants and 35 biological indices.

We have performed a set of preprocessing tasks and statistical analysis to the data including: z-score normalization, imputation of missing values, and outlier processing. Our datasets contained 9.70% of missing data which were processed using *MICE* imputation method. Detection of outliers was performed using IQR method. A total amount of 7.61% of outlying values was detected. We decided to handle outliers by imputing them using k-NN imputation method.

To distinguish relationships among the different variables, we have partitioned the original dataset into three datasets named *macro*, *micro* and *metals*. A correlation analysis and a PCA analysis were performed on preprocessed and non-preprocessed datasets. The best number of clusters for the preprocessed dataset was 9 while for the non-preprocessed dataset was 2. When comparing PCA results, we observed a better distribution on the preprocessed dataset. Indeed, within the analysis of the preprocessed dataset, we can differentiate 3 main clusters that were classified as: `very polluted` (based on the amount of pollutants in sampling site), `moderately polluted`, and `clean`. While for the non-preprocessed dataset, the observations were only differentiated as `very polluted` and `clean`. Not surprisingly, the analysis of environmental data indicates that sites near to anthropogenic activities (i.e., agriculture, urbanity, industries) present higher amount of pollutants and poor aquatic biodiversity.

From the correlation matrices, we observed that out of the 35 biological indices, only 13 are positively correlated to macro-pollutants, 8 to micro-pollutants, and 19 to metals. From the PCA analysis, we obtained a visualization of the correlation between the different variables. Figure 2 shows the first two principal components for the macro-pollutants and micro-pollutants. We observed that some biological indices show a negative correlation with these pollutants. This observation indicates that at high concentrations of pollutants, low values of biological indices is observed. Such an observation is consistent with the biological indices: certain biological indices (i.e., BMWP, EBI, Shannon or Simpson's indices) show low values when poor aquatic biodiversity is observed and we can conclude that we have a polluted aquatic system. The Family Biotic Index (FBI) show a similar behavior compared to the other metrics. Actually, FBI has high values

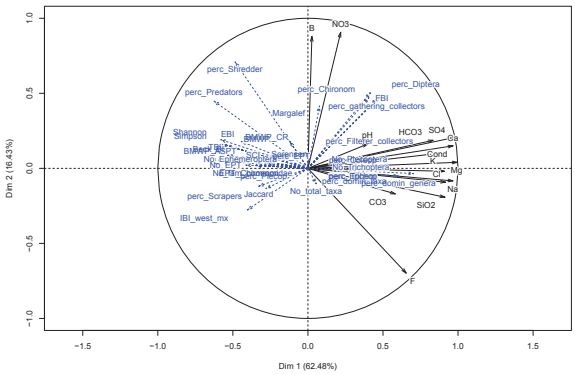

Fig. 2: PCA biplot for macro-pollutants (biomonitoring metrics in blue).

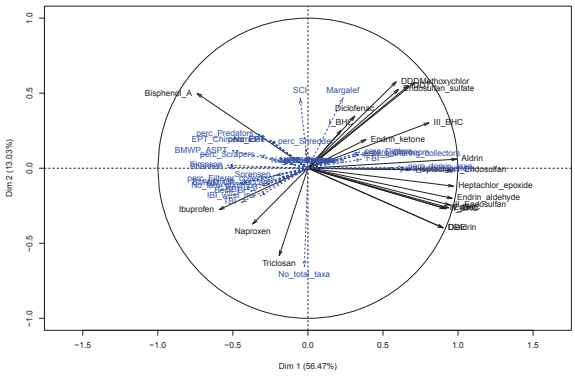

Fig. 3: PCA biplot for micro-pollutants (biomonitoring metrics in blue).

at high concentration levels of pollutants. As we can observe in Figures 2 and 3, this metric is positively correlated to the macro- and micro-pollutants. For the case of metals, our results do not provide enough information describing the relationship between the metals and the biomonitoring metrics. Interestingly, depending on the data preprocessing strategy, we may not have made the same conclusions and this demonstrates the need for a principled approach for data preprocessing.

V. Conclusions and Future Work

Our approach allowed us to identify the optimal data preprocessing strategies for each statistical method applied to our datasets. Although there is no universally adequate data preprocessing procedure (as it depends on many characteristics of the dataset in terms of size, distribution, Skewness, Kurtosis, correlation, etc.), we were able to learn hidden features for selecting the optimal data preprocessing for a specific dataset and analytical task. Our experimentations covered the most frequent data anomalies using the most general case (datasets with multivariate normal distribution). However, other data anomalies may appear on data (e.g., censored data, uncertainties) or may co-occur on datasets. Further study covering a wide spectrum of data anomalies, preprocessing procedures and types of data is part of our future agenda. Our experimental results show that an instantiation of the proposed framework can be effectively used to mine the hidden and valuable knowledge from biomonitoring data and it is extensible to address a wider spectrum of datasets in various application domains and the whole range of data preprocessing and data mining techniques.

References

[1] Batista, G., Monard, M.C., et. al. (2002). A Study of K-Nearest Neighbour as an Imputation Method, In HIS, vol. 87, p. 48.

[2] Berrahou L. , Lalande N., Serrano E., G. Molla G., Berti-Equille L., Bimonte 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, Comp. & Geosciences, 85, p. 126–135.

[3] Berti-Equille L., and Moussouni F. (2005). Quality-Aware Integration and Warehousing of Genomic Data. *In Proc. of the MIT Int. Conf. on Information Quality*.

[4] Berti-Equille, L. (2007). Measuring and Modelling Data Quality for Quality-Awareness in Data Mining, *Quality Measures in Data Mining*, Studies in Computational Intelligence, 43,p. 101–126, Springer.

[5] Bonada, N., Prat, N., Resh, V.H. and Statzner, B. (2006). Developments in Aquatic Insect Biomonitoring: A Comparative Analysis of Recent Approaches, Annu. Rev. Entomol., Vol. 51, p. 495–523.

[6] Breunig M M., Kriegel H.-P., Ng R.T., and Sander J. (2000). LOF: Identifying Density-Based Local Outliers. ACM Sigmod Record, vol. 29(2), p.93–104.

[7] Buuren, S., and Groothuis-Oudshoorn, K. (2011). MICE: Multivariate Imputation by Chained Equations in R. J. of Statistical Software, 45(3).

[8] Dasu T. (2013). Data Glitches: Monsters in Your Data. Handbook of Data Quality: Research and Practice, Springer, p.163–178.

[9] Eid, H.F., Hassanien, A.E., Kim, T., and Banerjee, S. (2013). Linear Correlation-Based Feature Selection for Network Intrusion Detection Model. In Advances in Security of Information and Communication Networks, p. 240–248.

[10] 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.

[11] Filzmoser P., Garrett R.G., Reimann D. (2005). Multivariate Outlier Detection in Exploration Geochemistry. Comp. & Geosciences, vol. 31, p. 579–587.

[12] Filzmoser P., Maronna R. and Werner M. (2008). Outlier Identification in High Dimensions. Computational Statistics and Data Analysis, vol. 52, p. 1694–1711.

[13] Gibert K., Izquierdo J., Holmes G., Athanasiadis I., Comas J., and Sánchez-Marré M. (2008). On the Role of Pre and Post-processing in Environmental Data Mining, In Proc. of the 4th Int. Congress on Environmental Modelling and Software.

[14] Guérin E., Marquet G., Burgun A., Loréal O., Berti-Equille L., Leser U., and Moussouni F. (2005). Integrating and Warehousing Liver Gene Expression Data and Related Biomedical Resources in GEDAW. In Proc. of Data Integration in the Life Sciences Workshop (DILS), p. 158–174.

[15] Hall M.A. (1999). Correlation-Based Feature Selection for Machine Learning. PhD thesis, Univ. of Waikato.

[16] Kohavi R., and John G. H. (1997). Wrappers for Feature Subset Selection. Artificial Intelligence, vol. 97, p. 273–324.

[17] Serrano Balderas E.C., Grac C., Berti-Equille L., and Armienta Hernandez M.A. (2016). Potential Application of Macroinvertebrates Indices in Bioassessment of Mexican Streams. Ecological Indicators, vol. 61, p. 558–567.

[18] Shreve J., Schneider, H., & Soysal, O. (2011). A Methodology for Comparing Classification Methods through the Assessment of Model Stability and Validity in Variable Selection. Decision Support Systems, 52(1), 247-257.

[19] Templ, M., Kowarik, A., and Filzmoser, P. (2011). Iterative Stepwise Regression Imputation Using Standard and Robust Methods. Computational Statistics & Data Analysis, vol. 55, p. 2793–2806.

[20] Troyanskaya O., Cantor M., Sherlock G., Brown P., Hastie T., Tibshirani R., Botstein D., and Altma R.B. (2001). Missing Value Estimation Methods for DNA Microarrays. Bioinformatics, vol. 17(6), p. 520–525.

[21] Wahlin, K., and Grimvall, A. (2008). Uncertainty in Water Quality Data and its Implications for Trend Detection: lessons from Swedish environmental data. Environmental Science and Policy, vol. 2, p. 115–124.