International Environmental Modelling and Software Society (iEMSs)
8th International Congress on Environmental Modelling and Software
Toulouse, France, Sabine Sauvage, José-Miguel Sánchez-Pérez, Andrea Rizzoli (Eds.)
http://www.iemss.org/society/index.php/iemss-2016-proceedings

Water Quality Data Analytics

Eva Carmina Serrano Balderas[1], Laure Berti-Equille[1,2], Ma. Aurora Armienta Hernandez[3], Jean-Christophe Desconnets[1]

[1]IRD, Institut de Recherche pour le Développement, UMR ESPACE DEV, Maison de la Télédétection, 500 Avenue Jean-François Breton, F-34093 Montpellier cedex 5, France (eva.serrano, laure.berti, jean-christophe.desconnets@ird.fr)

[2] QCRI, Qatar Computing Research Institute, HBKU, P.O. Box 5825, Doha, Qatar

[3] Geophysics Institute, Department of Analytical Chemistry, National Autonomous University of Mexico, 04510, Coyoacan, Mexico city, Mexico (victoria@geofisica.unam.mx)

Abstract: Water quality monitoring is a regular practice to assess the presence of pollutants in the water. The importance of monitoring is justified by the need to know the current state of aquatic ecosystems to design appropriate conservative and protective actions (Serrano Balderas *et al.*, 2015). Data from water quality monitoring may be prone to have various problems (i.e., incomplete, inconsistent, inaccurate, or outlying data) that may result in misleading analysis interpretation (Berrahou *et al.*, 2015). Incomplete data for instance, can be replaced by imputed values so that the statistical methods commonly used to describe patterns on water quality assessment (such as PCA, Hierarchical Classification, Kohonen-SOM) can be achieved. But imputation of missing values may impact statistical results. In this study, our goal is to assess the impact of imputation methods, and more generally of pre-processing, on the results of various statistical analyses. To this purpose, we studied five imputation methods (Mean, Hot-Deck, Sequential Imputation, Multiple Imputation and Iterative Stepwise Regression Imputation) on four statistical methods (Correlation, PCA, Kohonen-SOM and Hierarchical Classification) and developed a fully integrated analytics environment in R for statistical analysis of environmental data in general and for water quality data analytics in particular. The results obtained indicated that the imputation methods IRMI and MI generally improve the accuracy of the tested statistical methods when compared to methods without imputation. Our findings demonstrated that reliable results could be obtained when robust imputation methods are used to pre-process incomplete data.

REFERENCES

Serrano Balderas E.C., Grac C., Berti-Equille L., Armienta Hernandez M.A. (2015) Potential Application of Biological Indices Based on Macroinvertebrates on Mexican Streams. *Ecological Indicators*, 61:558-567.

Berrahou L., Lalande N., Serrano E., 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, *Computers & Geosciences*, 85: 126-135.