Multivariate functional data analysis for outlier detection in environmental data

X. Rigueira[♭], J. Martinez[♮ ♢,1] M. Araujo[♭], M. Pazo-Rodriguez[♭], and J. Taboada[♭]

(♭) CINTECX, GESSMin group, Department of Natural Resources and Environmental Engineering, University of Vigo,
Campus Lagoas-Marcosende, Rúa Maxwell, S/N, 36310.

(♮) CINTECX, GESSMin group, Department of Applied Mathematics I, University of Vigo,
Campus Lagoas-Marcosende, Rúa Maxwell, S/N, 36310.

(♢) Galician Center for Mathematical Research and Technology,
Technological Research Institute, floor -1, Campus Vida, Rúa de Constantino Candeira S/N.
15782.

Abstract

Authorities at an international, national, and local level are approving stricter water quality regulations in a coordinated effort to preserve this important natural element. In order to make sure that the limits defined in the regulations are respected, the data from numerous measuring stations are analyzed with mathematical and computational tools. This kind of information is defined by several features: (i) various physical-chemical water quality parameters, (ii) time-series type of data, (iii) not usually normally distributed, and (iv) hidden dependencies between the different variables. The present research work proposes a multivariate approach to the study of water quality data and the detection of outliers. From a mathematical perspective, functional data analysis [1], which is based in the transformation of discrete points into functions over a continuum, is able to study the trend of the data, and does not depend on its distribution. Moreover, the implementation of the rank-based copula cross-periodogram (CCR-Periodogram) [2] allows the analysis of the information from a multivariate approach, while the outlier detection is computed with the halfspace depth [3]. This methodology was applied to a water quality database from the Ebro Basin in Spain. The variables studied are ammonium, conductivity, nitrates, dissolved oxygen, pH, temperature, and turbidity with samples taken every 15 minutes from January 1st, 2019 to December 31st, 2021. The methodology introduced was developed in Python 3.7 and in R 4.2 with aid of the library *mlmts*. The aforementioned method is able to compute an accurate detection of those outlying curves and unveils the hidden relations between variables in the database. Consequently, it makes easier the selection of specific corrective measures for polluting events and allows the tracking of the results, as well as the study of the synergies in the basin as a whole.

References

[1] Ramsay, J.O., Silverman, B.W., Functional Data Analysis (1ts ed.). New York City, Springer International Publishing, 2005.

[2] Lopez-Oriona, A., Vilar, J. A., Outlier detection for multivariate time series: A functional data approach *Knowledge-Based Systems*, Volume(233):107527, 2021.

[3] Dyckerhof, R., Mozharovskyi, P., Exact computation of the halfspace depth *Computational Statistics & Data Analysis*, Volume 98:19–30

[1] javmartinez@uvigo.es