Benchmark of outlier detection methods for spectral data

Mathieu Lepot[1], Alma Mašić[2], Jean-Baptiste Aubin[3] and François H.L.R. Clemens[1,4]

[1]Department of Water Management, Faculty of Civil Engineering and Geosciences, Technical University of Delft, Stevinweg 1 (Building 23), 2628CN Delft, The Netherlands (Email: *m.j.lepot@tudelft.nl; f.h.l.r.clemens@tudelft.nl*)
[2]EAWAG, Department Process Engineering, Überlandstrasse 133, CH-8600 Dübendorf, Switzerland (Email: *alma.masic@eawag.ch*)
[3]Univesity of Lyon, INSA Lyon, DEEP, 34 avenue de arts, F-69621 Villeurbanne Cedex, France (Email: *jean-baptiste.aubn@insa-lyon.fr*)
[4]Deltares, Rotterdamweg 185, 2629HD, Delft, The Netherlands (Email:*Francois.ClemensMeyer@deltares.nl*).

Abstract
UV/Vis spectrophotometers have been used to monitor water quality since the early 2000s. Calibration of these devices requires sampling campaigns to elaborate relations between recorded spectra and measured concentrations. Recent sensor improvements allow recordings of a spectrum in as little as 15 seconds, making it possible to record several spectra for the same sample. Spectrum repetitions provide new opportunities to detect outliers – a task that is difficult in non-repetitive spectra recordings. A well-executed outlier detection can e.g. result in a more accurate calibration of the spectrophotometer or an improved construction of a regression model. In this work, two methods are presented and tested to detect outliers in repetitions of spectral data: one based on data depth theory (DDT) and one on principal component analysis (PCA). Results show that the two methods are generally consistent in identifying outliers, with only small differences between the methods.

Keywords
DDT, detection, outlier, PCA, spectra, UV/Vis

INTRODUCTION

Since the early 2000s, UV/Vis spectrophotometers have been used to monitor water quality (Rieger *et al.*, 2004). The accuracy and robustness of the concentration estimation through the recorded UV/Vis spectra require a local calibration (Langergraber *et al.*, 2003). The calibration allows taking into account the local characteristics of the water, the so called background matrix. To this end, water samples are collected, spectra are recorded with the spectral device, and compound concentrations are measured with laboratory analyses. Single spectral recordings per sample pose a difficult task of detecting any outliers that can potentially lead to a suboptimal calibration of the spectrophotometer. Recent technological advances have shortened the spectral recording time down to as little as 15 seconds. This short recording time makes it possible to quickly collect several spectra for each sample. Such repetitions of spectral recordings provide new opportunities to detect outliers. The calibration of a device could therefore be divided into several steps, one of which would consist of outlier detection, thereby improving the accuracy of the entire calibration. In this paper, we present two methods that can be used as a preliminary step for outlier detection in spectral calibration.

MATERIALS AND METHODS

Data sets

Dry weather samples in a WWTP. 94 1L inlet water samples were collected at the Fontaines-sur-Saône WWTP (30 000 inhabitants, combined sewer). Two kinds of data have been recorded: *i)* up to 25 spectra for each sample, and *ii)* laboratory analysis concentrations of TSS, total and dissolved COD. The spectrophotometer used in this study was a spectro::lyser (s::can, Austria),

with an optical path length of 2 mm, recording in the UV/Vis spectrum (200-750 nm) with a wavelength step of 2.5 nm.

Source-separated nitrified urine samples. 30 3L samples from a nitrification reactor treating source-separated urine were collected during 10 weeks, with additions of nitrite and nitrate via stock solutions. Each sample was subjected to combinations of pre-treatments [(un)-filtered/(un)-diluted], resulting in four sample groups. The spectral device used was a spectro::lyser, with a path length of 0.5 mm, recording in the UV spectrum (220-399 nm) with a resolution of 1 nm and recorded 5 spectra per sample.

Methods

DDT. Initially developed by Lepot (2012), the three steps method has been improved: 1) remove outliers according to the relative positions of spectra or Euclidean distances between spectra, 2) identify the most representative spectrum (as the "most in the middle") and 3) study of uncertainty associated to each wavelength.

PCA. This method relies on principal component analysis and focuses on the scores of the first principal component (Mašić *et al.*, 2015). The user determines the outliers based on a comparison of the scores for the spectral repetitions for each sample.

RESULTS AND DISCUSSION

The DDT method with the first step based on relative position is consistent with other methods for outlier detection in the WWTP samples. On the other hand, the method is unable to identify outliers in some of the urine sample groups due to strong saturation effects in parts of the spectra. PCA and DDT give consistent results for filtered urine samples, not for unfiltered. These small differences in the method behaviour still need to be investigated.

CONCLUSIONS AND PERSPECTIVES

This study deals with a new topic and can be useful for researchers who record repetitive spectra. Two independent methods provide often-consistent detected outliers, but some differences still need to be studied. Further work should *i)* focus on links between method behaviours and water background matrices and *ii)* be performed on additional data sets.

ACKNOWLEDGEMENTS

Marie Curie ITN QUICS project; EU's 7th Framework Programme for research, technological development and demonstration (grant 607000); MAC-Nut Eawag Discretionary Funds (5221. 00492.007.10); Kris Villez (Eawag, Switzerland), Ana Santos (U. Nova de Lisboa, Portugal).

REFERENCES

Langergraber G., Fleischmann N., and Hofstädter F. (2003). A multivariate calibration procedure for UV/VIS spectrometric quantification of organic matter and nitrate in wastewater. Water Science and Technology, 47(2), 63-71.

Lepot M. (2012). Mesurage en continu des flux polluants en MES et DCO en réseau d'assainissement (Continuous monitoring of pollutant fluxes - TSS and COD – in sewers), PhD thesis, INSA Lyon, 257p.

Mašić, A., Santos, A.T.L., Etter, B., Udert, K.M., Villez, K. (2015). Estimation of nitrite in source-separated nitrified urine with UV spectrophotometry. Water Research, 85, 244-254.

Rieger L., Langergraber G., Thomann M., Fleischmann N. and Siegrist H. (2004). Spectral in-situ analysis of NO2, NO3, COD, DOC and TSS in the effluent of a WWTP. Water Science and Technology, 50(11), 141-152.