Desalination and Water Treatment
www.deswater.com
doi: 10.5004/dwt.2020.25481

182 (2020) 127–134
April

Characterizing water quality and quantity profiles with poor quality data in a machine learning algorithm

Zhonghyun Kim[a], Heewon Jeong[b], Sora Shin[b], Jinho Jung[a], Joon Ha Kim[b], Seo Jin Ki[c,*]

[a]*Division of Environmental Science and Ecological Engineering, Korea University, Seoul 02841, Republic of Korea*
[b]*School of Earth Sciences and Environmental Engineering (SESE), Gwangju Institute of Science and Technology (GIST), Gwangju 61005, Republic of Korea*
[c]*Department of Environmental Engineering, Gyeongnam National University of Science and Technology, Jinju 52725, Republic of Korea, email: seojinki@gntech.ac.kr*

Received 10 September 2019; Accepted 3 January 2020

ABSTRACT

Statistical analyses are often subject to misinterpretation due to poor data quality which is inaccurate, incomplete, or unavailable. This study describes how incomplete data diminishes the screening accuracy of water pollution hotspots using a self-organizing map (SOM), a popular algorithm in reducing the dimension of complex data in a nonlinear fashion. A full data set consisting of 12 water quality and quantity parameters monitored monthly over 3 years at the Yeongsan River in Korea was provided to SOM as a reference input. For purposes of comparison, SOM was further allowed to accept three incomplete data sets in terms of variable availability as well as data loss for single and multiple parameters and different pollution levels. We found that data loss of either single or multiple parameters exceeding 15% of the entire data set led to significant changes in spatial and temporal patterns of the original data. However, the variables intentionally unavailable in the given data set affected the screening performance of water pollution hotspots in SOM, to a less obvious extent, as long as the percentage of missing data fell below 10%. The same applied to data loss with three pollution levels, from high through moderate to low concentrations of one important variable. Therefore, we recommend the use of multiple approaches that couple dimensionality reduction algorithms with reasonable imputation methods for the data set with a high percentage (e.g. above 15%) of missing values.

Keywords: Non-linear data analysis; Dimensionality reduction; Water quality data; Pollution hotspots; Incomplete data; Self-organizing map

* Corresponding author.