A Computational Framework for Monitoring Black-Odorous Water from Remote Sensed Data and Data Mining Techniques

Yichun Xie[1,*], Ji Yang[2], Alicia Zhou[3], Chenghu Zhou[4], Liusheng Han[2], Yong Li[2]

[1] Institute for Geospatial Research and Education, Eastern Michigan University, Ypsilanti, Michigan 48197, USA

[2] Guangdong Key Laboratory of Geospatial Information Technology and Application, Guangzhou Institute of Geography, Guangzhou 510070, China

[3] Alicia Zhou, Department of Statistics, Boston University, azhou96@bu.edu

[4] Chenghu Zhou, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, zhouch@lreis.ac.cn

**** Corresponding author: yxie@emich.edu, +1 734-4877588.***

Abstract

China, through 40 years of the economic reforms, has witnessed fast economic growth, which has been accompanied with severe environmental challenges. One of those concerns is the surface water pollution, called black-odorous water, which has occurred in almost all big rivers as well as small streams. The Sate Council of China on April 16, 2015 unveiled its first Action Plan for Water Pollution Prevention and Control, aiming at cleaning up 70% black-odorous water in major rivers and cities by 2020. However, there are several technical challenges hindering this Action Plan: 1) What are the key indicators of black-odorous water? 2) Is the traditional direct measurement approach too time-consuming and costly for quick actions; 3) Can remote-sensing techniques be used to replace or enhance the traditional direct monitoring approach? 4) Are there effective computational algorithms for using the remote sensed data to assess the severity of black-odorous water and to identify key indicators of black-odorous water? From the perspectives of remote sensing, an effective computational framework should contain the following functions: 1) examining how each water quality indicator contributes to the severity of black-odorous water; 2) identifying which wavelength reflectances closely relate to the severity of black-odorous water; and 3) assessing which wavelength reflectances complement with what water quality indicators to reveal the most sensitive responses to the severity of black-odorous water. The paper intends to develop a computational data mining framework, which innovatively integrates the data mining techniques, such as hierarchical clustering, semisupervised discriminating, factorical analysis, spatial lasso selecting and linear regression, to answer these questions. A case study in Guangzhou City, China is carried out to demonstrate the applicability of this remote sensing based black-odorous water computational monitoring framework.

Keywords: Black-odorous water, Computational modeling, Data mining, Remote sensing, Spectroradiometer

1. Introduction

Black and odorous water occurrence is a severe environmental pollution especially in urban areas, which threatens the life and welfare of residents and severely damages the urban image and its ecological environment (Qian et al., 2012). We need to have an effective monitoring system to provide scientific data to identify the causes and sources of black-odorous water and eventually to eliminate the occurrence of black-odorous water (Hasan et al., 2015). However, the occurrence of black-odorous water is a disastrous event, which is dynamic in time, and hard to predict in location (Shen et al, 2017). Therefore, the traditional approach of direct measurement of water pollution indicators is not efficient to capture abrupt events of black-odorous water. Intuitively the remote sensed techniques, including Internet of Things, drone, airborne and spaceborne based remote sensing could provide a solution. Remote sensing techniques are quicker, less costly and more targeted for monitoring black-odorous water because of the widespread use and easy deployment of remote sensed devices.

However, at present, the remote sensing based black-odorous water monitoring is still at an experimental stage. Very few studies on the optical characteristics of urban black-odorous water and the inversion of key water quality indicators through remote sensing have been reported. In other words, whether the characteristics and severity of black-odorous water can be detected through analysing the water's spectral characteristics is still a challenge. Moreover, the hyperspectral imaging techniques have been developed and applied rapidly in environmental monitoring since a decade ago (Hall et al., 2009). Hyperspectral images contain thousands of narrow bands, providing better capacity to distinguish the undetectable subjects or characteristics that cannot be done with the multispectral images that usually include a few to a dozen wider bands. Unfortunately, hyperspectral images have presented some challenges to image analysis and processing. One straightforward challenge is how the correspondences between the narrow spectral bands of hyperspectral images and the chemical characteristics of black-odorous water can be established. As explained above, the hyperspectral images have hundreds of or thousands of bands. There are critical needs for developing new data mining techniques to supplement traditional statistical methods in order to detect the spectral correspondences of the hyperspectral images to the chemical characteristics of the polluted water.

2. The Computational Data Mining Framework for Monitoring Black-odorous Water

The process of establishing the spectral correspondences of the hyperspectral images to the chemical characteristics of the polluted water involves at least five steps: (1) classifying the severity level of black-odorous water; (2) determining which water quality parameters are meaningful indicators of the black-odorous water; (3) pre-processing the spectral bands of hyperspectral images in order to have better choices of spectral bands that can reveal spectral responses to the severity of black-odorous water; (4) detecting which hyperspectral bands correspond with the selected water quality indicators in order to find the most sensitive bands revealing the severity of black-odorous water; and (5) validating the outcomes in the previous steps.

The techniques of data mining and machine learning chosen to construct the black-odorous water computational data mining framework are depicted in Figure 1. The computational framework requires two sets of data: (1) the measurements of water quality parameters commonly collected for water quality monitoring; and (2) the hyperspectral data that were collected simultanenously with the water quality samples. Since the hyperspectral data can collect spectral bands from 350nm to 2500nm at 1nm spectral resolution, there is a huge amount of spectral data. Therefore, some types of band transformation or derivation

processes are usually adopted to reduce the number of bands or select adequate bands for analysis. These methods include simple aggregation, specific water quality indices such as the water adjusted vegetation index (Villa et al., 2014), and various modeling approaches (Jay et al., 2017).

The data mining starts with a classification of the black-odorouos water samples into several groups, which either reflects the severity of the black-odor based on the water quaility parameters, or the spatial adjacency and water quality disimilarity judged by both the spatial adjacency of the sampling sites and the characteristics of the water quality parameters. Since no prior knowledge exisits, an unsupervised hierachical clustering either based on a disimilarity (the divisive method) or a similarity measurement (the agglomerative method) (Rodriguez et al., 2019) can be applied to classify the water quality samples into several clusters or groups. Based on the outcome of the hierarchical data mining, a semisupervised learning (a linear discriminant analysis) can be followed to validate the performance and accuracy of the hierarchical clustering.

Next, we want to explore how water quality parameters are correlated to affect the severity or variability of water black-odor. A factorial data mining analysis will be conducted to find out whether a lower number of latent (i.e., the unobserved) variables exist to reveal which latent variable (s) are the most important for describing the severity and variability of black-odorous water (Lee et al., 2005).

Third, the spectral bands of hyperspectral data are either combined into adequate band widths, or/and pre-processed as unique indexes through applying some computations among certain bands, which are assumed to be closely related to water quality through prior knowledge or literature view.

Now, the severity of black-odorous water is classified, the water quality parameters are reduced to fewer numbers in the context of black-odour severity, and the spectral bands of hyperspectral data are also pre-processed into fewer ones. The selected hyperspectral bands are treated as the independent variables while the latent variables derived from the water quality parameters are handled as the dependent variables. However, there are still too many independent variables, which creates challenges for us to determine correct regression model specifications to examine whether there are causal relationships between the selected spectral bands and the latent dependent variables. Therefore, a data mining technique, Least Absolute Shrinkage and Selection Operator (LASSO), is adopted to select significantly fewer but highly correlated independent variables (Heinze and Dunkler, 2017; Heinze et al., 2018). After the mining through LASSO, a sub-set of spectral bands will be chosen to be highly correlated to each dependent latent variable, respectively. In other words, whether certain spectral bands are sensitive to a set of water quality parameters (which attribute to a latent dependent variable) is determined.

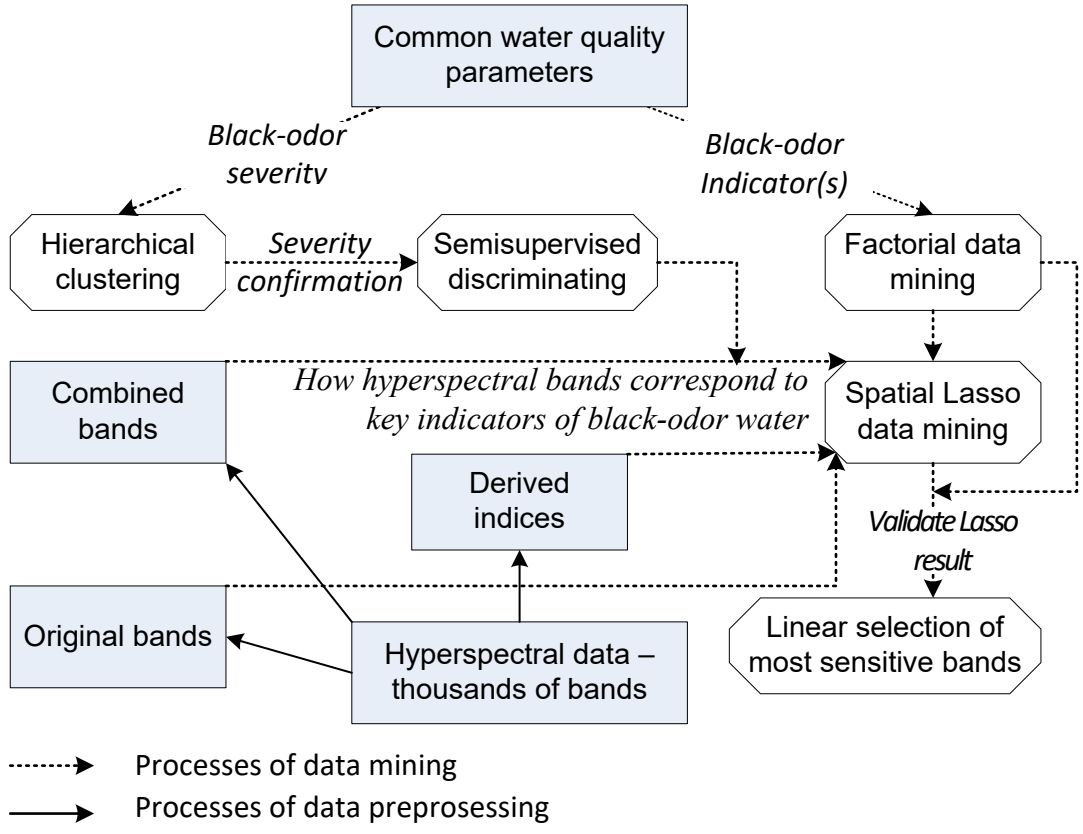

Figure 1 The flowchart of detecting key indicators of black-odor water through remote sensing and data mining

Once some causal relationship between a latent water quality indicator and a set of spectral bands is confirmed, an ordinary least-square linear regression model can be applied to quantify the explanation power of the variances of the latent water quality indicator by these selected spectral bands. In combination with the variance of the black-odour severity explained by each latent water quality indicator in the process of factorial data mining, the sensitivity of the spectral bands to the black-odour water severity can be analysed and estimated.

3. A Case Study in Guangzhou City, China

We surveyed current water quality indicators through literature review and standard operational manuals. We took 52 black-odorous water samples in the Chebei and Yonghe Rivers in the City of Guangzhou from September to October in 2017. We measured the indicators of dissolved oxygen (DO), ammonia nitrogen (NH3-N), total phosphorus (TP), total nitrogen (TN), chemical oxygen demand (COD), five-day biochemical oxygen demand (BOD5), and suspended solids (TSS). We also collected the full spectral data from 350nm to 2500nm at 1nm spectral resolution synchronously by using a spectroradiometer. Together the water quality indicators and the spectral reflectances composed a comprehensive data-set.

We are in the middle of pre-processing the hyperspectral data and expect to start the data mining procedures in the summer and to report the findings at The International Conference of Geo-Computation in September 2019.

References:

Hall, R.K., Watkins, R.L., Heggem, D.T. et al. (2019). *Environ Monit Assess* 159: 63. https://doi.org/10.1007/s10661-008-0613-y

Hasan, H. H., Jamil, N. R., and Aini, N. (2015). Water Quality Index and Sediment Loading Analysis in Pelus River, Perak, Malaysia. *Procedia Environmental Sciences*, 30: 133-138.

Heinze G, Wallisch C, and Dunkler D. (2018). Variable selection—a review and recommendations for the practicing statistician. Biometrical J., 60:431–49. https://doi.org/10.1002/bimj.201700067 PMID: 29292533

Heinze, G. and Dunkler, D. (2017). Five myths about variable selection. *Transplant International*, *30*: 6–10.

Jay, S.; Guillaume, M.; Minghelli, A.; Deville, Y.; Chami, M.; Lafrance, B.; Serfaty, V. (2017) Hyperspectral remote sensing of shallow waters: Considering environmental noise and bottom intra-class variability for modeling and inversion of water reflectance. *Remote Sens. Environ.* 200: 352–367.

Lee, Z.H., Peterson, R.L., Chien, C.F., and Xing, R. (2005). Factor Analysis in Data Mining. In John Wang, Encyclopedia of Data Warehousing and Mining, *IGI-Global*, pp1382.

Qian, C. P., Liu, Y. L., Wang, J. and Chen, Z. L. (2012). Empirical Study on the Performance Evaluation of the Black Color and Odor Urban Rivers Treatment Projects. *Advanced Materials Research* 518-523: 2104-2108.

Rodriguez MZ, Comin CH, Casanova D, Bruno OM, Amancio DR, Costa LdF, et al. (2019) Clustering algorithms: A comparative approach. *PLoS ONE* 14(1): e0210236. https://doi.org/10.1371/journal.pone.0210236.

Villa, P., Mousivand, A., & Bresciani, M. (2014). Aquatic vegetation indices assessment through radiative transfer modeling and linear mixture simulation. International *Journal of Applied Earth Observation and Geoinformation*, 30: 113–127.