International Journal of Agricultural and Environmental Information Systems
Volume 9 • Issue 2 • April-June 2018

Statistical and Data Mining Techniques for Understanding Water Quality Profiles in a Mining-Affected River Basin

Jose Simmonds, Universidad Carlos III de Madrid, Leganés, Spain

Juan A. Gómez, Universidad de Panamá, Panama City, Panamá

Agapito Ledezma, Universidad Carlos III de Madrid, Leganés, Spain

ABSTRACT

This article contains a multivariate analysis (MV), data mining (DM) techniques and water quality index (WQI) metrics which were applied to a water quality dataset from three water quality monitoring stations in the Petaquilla River Basin, Panama, to understand the environmental stress on the river and to assess the feasibility for drinking. Principal Components and Factor Analysis (PCA/FA), indicated that the factors which changed the quality of the water for the two seasons differed. During the low flow season, water quality showed to be influenced by turbidity (NTU) and total suspended solids (TSS). For the high flow season, main changes on water quality were characterized by an inverse relation of NTU and TSS with electrical conductivity (EC) and chlorides (Cl), followed by sources of agricultural pollution. To complement the MV analysis, DM techniques like cluster analysis (CA) and classification (CLA) was applied and to assess the quality of the water for drinking, a WQI.

KEYWORDS

Classification, Cluster Analysis, Decision Tree, Multivariate, Water Quality Index

1. INTRODUCTION

Minera Panamá S.A. (MPSA), wholly owned by Minera Panamá S.A-First Quantum Minerals Ltd (MPSA-FQML), is investigating the feasibility of developing the MPSA Project Mina de Cobre Panamá (the Project). The proposed Project would mine and process copper sulfide ore in the Petaquilla Concession, Panamá. This concession covers an area of 130 square kilometers (km2) and is located in the District of Donoso, Colón Province, in north-central Panamá. The concession contains at least three spatially distinct copper ore bodies (Colina, Botija and Valle Grande) and three conventional open pit mines are currently planned to exploit these ore bodies (EIAs, 2010).

The copper sulfide ore will be mined using conventional open pit mining and will be processed using crushing, milling, flotation recovery and concentrate dewatering. The proposed design ore feed to the processing plant is 150,000 tons per day (t/d). It is expected that this will be expanded to 225,000 t/d at year ten by the addition of a third processing line. The Project will export materials through a port site to be constructed on the Caribbean coast at Punta Rincón and linked to the main Project site by a road, a power line corridor, and buried pipelines for transfer of products and other materials. As the nation of Panama develops, increasing industrialization and urbanization has led to a wide-scale contamination of many surface water resources from industrial effluents, domestic sewage discharges, and excessive use of fertilizers, pesticides and the emerging mining activities. Then, it

DOI: 10.4018/IJAEIS.2018040101

may be inferred that the increased anthropogenic pressures and natural processes are accounting for degradation in surface water and groundwater quality (Carpenter et al., 1998). Hence, given these pressures experienced on the water resources in the area, the main objectives of conservation must be in the control and minimization of pollution occurrences and problems facing these pollutants and to provide water of an adequate quality that can serve different purposes, such as drinking water, irrigation water (Dinar et al., 1995). Then, the monitoring of water quality for any water body must be one of the highest priorities for their protection policy (Lewis, 2000).

Multivariate statistical methods such as factor analysis and principal components have been used successfully in hydrochemistry for many years. Nowadays, with the emerging technique offered by data- mining, the water quality of a given river state can reveal features otherwise not seen by conventional methods. Multivariate techniques allow us to discover the information hidden in the data set about the possible environmental influences on water quality (Spanos et al., 2003). Today, data mining is popular among researchers of water quality investigations, for example in regard to chlorophyll levels, Lu & Huang (2009) proposed Decision-making tree to forecast levels for the next day. Also, Fu-Cheng & Xue-Zhao (2013), suggested the use of fuzzy c-means clustering method to classify and assess rural surface water quality built on monitoring data from 33 water quality stations in 23 rural rivers and 4 reservoirs in Lianyungang city (China). Multivariate methods have several shortcomings such as the presence of mathematical calculations, equal treatment and process to the old and new data, problems with prediction and classification task due to multivariate overlapping of the parameters. Notwithstanding, data mining and machine learning techniques have shown to achieve great success in many disciplines (Mjolsness & DeCoste, 2001). Nevertheless, it is a well-known fact that data mining algorithms work best on large data sets, yet there are several studies which encourages its application on small databases (Jiang et al., 2009; Andonie, 2010; Natek & Zwilling, 2014).

In this study, we evaluated the possibility that a smaller group of water quality parameters could provide sufficient information for assessing water quality. For this reason, Factor analysis and data mining methods were applied to water quality data obtained from the surface waters of three (3) water quality monitoring stations at the Petaquilla River Basin during two hydrological seasons (high and low flows).

The paper is organized as follows. Section 2, describes the problem of the overall study location and provides information on the study site. Section 3 illustrates the resources and methodology that has been used to tackle the problem. The experimental setup and analysis results of the study are presented in Section 4. Finally, conclusions and future work are discussed in Section 5.

2. THE PROBLEM

2.1. Data Source and Study Area

The Petaquilla basin is the westernmost drainage basin at the mine site. Surface runoff in this basin reports to the Petaquilla River, where it subsequently flows northwest and discharges directly to the Caribbean Sea. Two open pits (Colina and Valle Grande), as well as the southwest waste rock storage facility and associated sedimentation ponds, will be developed in the southeast region of the Petaquilla basin. The community of Nueva Lucha and the Faldalito sector are also located in this basin. The three-surface water baseline sampling stations established in the Petaquilla River basin, specifically in the Petaquilla River, are described in Table 1 and shown in Figure 1.

3. RESOURCES AND METHODS

The Isthmus of Panama has basically two seasons: the dry low flow season (January to April) and the high flow season (May to December). The climate in the region of the Petaquilla River Basin is typically governed by these two seasons. Therefore, the hydrological conditions during the low and