Original Research / Articulo original - Tipo 1

Water quality warnings based on cluster analysis in Colombian river basins

Edwin Ferney Castillo / efcastillo@unicauca.edu.co

Wilmer Fernando Gonzales / wfgtulcan@unicauca.edu.co

Iván Darío López / navis@ unicauca.edu.co

Apolinar Figueroa, Ph.D. / apolinar@unicauca.edu.co

David Camilo Corrales / dcorrales@unicauca.edu.co

Miller Guzmán Hoyos / mguzman@unicauca.edu.co

Juan Carlos Corrales, Ph.D. / jcorral@unicauca.edu.co

Universidad del Cauca, Popayán-Colombia

ABSTRACT Fresh water is considered one of the most important renewable natural resources in the world. Among all the countries, Colombia is one of the places with the highest water supply, and has five watersheds: the Caribbean, Orinoco, Amazon, Pacific and Catatumbo. It is therefore vital to study and evaluate the water quality of the rivers and/or lotic systems. In recent studies, some scientists made use of biological indices to calculate water quality, while others detected water quality through machine learning techniques. However, these studies do not allow users to easily interpret the results. These investigations motivated us to propose a dataset for generating water quality alerts in Piedras river basin based on the analysis of the K-Means clustering algorithm and C.4.5 classification technique.

KEYWORDS Clustering; water quality data; aquatic macro-invertebrates; taxon; C.4.5 decision tree.

Alertas de calidad del agua basadas en análisis de agrupamiento en las cuencas de los ríos colombianos

RESUMEN El agua dulce es considerada uno de los recursos naturales renovables más importantes, Colombia se ubica entre los países con mayor oferta hídrica del mundo con cinco vertientes: Caribe, Orinoco, Amazonas, Pacifico y Catatumbo. En este sentido es de vital importancia estudiar y evaluar la calidad del agua de sus ríos y/o sistemas lóticos. Hoy por hoy, algunos científicos hacen uso de índices biológicos para calcular la calidad del agua, mientras que otros detectan la calidad del agua por medio de técnicas de aprendizaje automático, sin embargo los trabajos encontrados hasta el momento no permiten al usuario interpretar fácilmente los resultados. Estas investigaciones motivaron a proponer un conjunto de datos para la generación de alertas de la calidad del agua en la cuenca Rio Piedras basado en el análisis del algoritmo de agrupamiento K-Means y la técnica de clasificación C.4.5.

PALABRAS CLAVE Agrupamiento; datos de calidad del agua; macro-invertebrados acuáticos; taxón; árbol de decisión C4.5.

Alertas de qualidade da água com base na Análise de Agrupamento nas bacias dos rios colombianos

RESUMO A água doce é considerada um dos recursos naturais renováveis mais importantes, a Colômbia é um dos países com a maior oferta hídrica do mundo, com cinco vertentes: Caribe, Orinoco, Amazonas, Pacífico e Catatumbo. Neste sentido, é de vital importância estudar e avaliar a qualidade da água dos seus rios e / ou sistemas lóticos. Atualmente, alguns cientistas fazem uso de índices biológicos para calcular a qualidade da água, enquanto outros detectam a qualidade da água através de técnicas de aprendizado de máquina, no entanto os trabalhos encontrados até a data ainda não permitem que os usuários possam facilmente interpretar os resultados. Essas investigações levaram a propor um conjunto de dados para a geração de alertas de qualidade da água na bacia do rio Piedras, com base na análise do algoritmo de agrupamento K-Emanes e na técnica de classificação C.4.5

PALAVRAS-CHAVE Agrupamento; dados de qualidade da água; macroinvertebrados aquáticos; taxon; C4.5 árvore de decisão.

Received / Recepción: Abril 22, 2015 - Accepted / Aceptación: Junio 17, 2014 *doi: 10.18046/syt.v13i33.2077*

I. Introduction

Fresh water is considered one of the most important renewable natural resources. Among all the countries, Colombia has the largest water supply in the world (Viceministerio de Ambiente, 2010) with five watersheds: the Caribbean, Orinoco, Amazon, Pacific and Catatumbo. It is therefore vital to study and evaluate the water quality of rivers and/or lotic systems. However, determining the environmental status of these systems becomes a particularly complex task when the streams' reference conditions are unknown and also when they have been exposed for a long period of time to anthropogenic perturbations (Arango, Álvarez, Arango, Torres, & Monsalve, 2008).

In this type of ecosystem the macro-invertebrate community is highly diverse. Due to their tolerance levels to different environmental changes (Alba-Tercedor, 1996), they have potential to be used in lotic system monitoring (Pino et al., 2003), supplemented by analysis of physicochemical variables (pH, dissolved oxygen, etc.) and environmental variables (temperature, humidity, rainfall, sunlight, etc.). Scientists currently make use of biological indices for calculating water quality (Rico, Paredes, & Fernandez, 2009; Park, Chon, Kwak, & Lek, 2004). In Colombia, commonly used bio-indicators in river systems are the Biological Monitoring Working Party (BMWP) and Average Score per Taxon (ASPT), which are adapted for each region of the country, due to the diversity of climates and reliefs (Pérez, 2003).

Some studies detect water quality through Support Vector Machines (SVM) (Singh & Gupta, 2012; Bae & Park, 2014; Liu et al., 2012) and Artificial Neural Networks (ANN) (Bucak & Karlik, 2011), with the objective of effectively monitoring and controlling water quality. However, these studies do not define classes that allow users to easily interpret the results obtained by the classifiers. We therefore propose a data set to generate alerts for water quality in Piedras river basin based on the analysis of clustering algorithms. This paper is organized as follows: Section II presents the process of understanding the data and the techniques used to build it. Section III describes the cluster validation methods and experimental results. And Section IV summarizes the study and provides conclusions.

II. Materials and methods

This section describes the process of data collection, and techniques used to build it, through clustering and classification algorithms.

I. Introducción

El agua dulce es considerada uno de los recursos naturales renovables más importantes, Colombia se ubica entre los países con mayor oferta hídrica del mundo (Viceministerio de Ambiente, 2010) con cinco vertientes: Caribe, Orinoco, Amazonas, Pacifico y Catatumbo. En este sentido es de vital importancia estudiar y evaluar la calidad del agua de los ríos y/o sistemas lóticos. Sin embargo, determinar el estatus ambiental de estos sistemas se convierte en una tarea compleja cuando las condiciones de referencia de los ríos son desconocidas y también cuando han sido expuestas por un largo periodo de tiempo a perturbaciones antropogénicas (Arango, Álvarez, Arango, Torres, & Monsalve, 2008).

En este tipo de ecosistemas, la comunidad de macroinvertebrados es altamente diverso. Debido a sus niveles de tolerancia para diferentes cambios ambientales (Alba-Tercedor, 1996), tienen el potencial para ser usados en el monitoreo de sistemas lóticos (Pino et al., 2003) complementados por análisis de variables fisicoquímicas (pH, oxígeno disuelto, etc.) y variables ambientales (temperatura, humedad, precipitaciones, luz solar, etc.). Actualmente los científicos hacen uso de los índices biológicos para medir la calidad del agua (Rico, Paredes, & Fernandez, 2009; Park, Chon, Kwak, & Lek, 2004). En Colombia, es común el uso de bioindicadores en sistemas fluviales: Biological Monitoring Working Party [BMWP] y Average Score per Taxon [ASPT], los cuales son adaptados para cada región del país, debido a la diversidad de climas y relieves (Pérez, 2003).

Algunas investigaciones detectan la calidad del agua a través de *Support Vector Machines* [SVM] (Singh & Gupta, 2012; Bae & Park, 2014; Liu et al., 2012) y *Artificial Neural Networks* [ANN] (Bucak & Karlik, 2011), con lo que logran monitorear y controlar su calidad, efectivamente. Sin embargo, estos estudios no definen clases que permitan a los usuarios interpretar fácilmente los resultados obtenidos por los clasificadores. En cierto modo, se ha propuesto un conjunto de datos para generar alertas para la calidad del agua en la cuenca del río Piedras, basado en los análisis de algoritmos de agrupamiento.

Este artículo está organizado de la siguiente manera: La sección II presenta el proceso de comprensión de datos y técnicas usadas para desarrollarlo; la sección III describe los grupos de métodos de validación y resultados experimentales; y la sección IV resume el estudio y ofrece las conclusiones.

II. Materiales y métodos

Esta sección describe el proceso de recolección de datos y las técnicas usadas para desarrollarlo, a través de algoritmos de agrupamiento y clasificación.

A. Descripción del área de estudio

Los datos usados en este estudio fueron recolectados de la cuenca del río Piedras, localizado en el departamento del Cauca, Colombia (Fuente: 76° 31' 10" oeste de Greenwich y 2° 21' 45" de latitud norte. Desembocadura: 76° 23' 45" de longitud oeste y 2° 25' 40" de latitud norte del rio Cauca), por el Grupo de Estudios Ambientales en la Universidad del Cauca, entre

el 2011 y 2013, implementando la metodología aplicada por Pérez (2003).

Las muestras capturadas contenían variables biológicas (macroinvertebrados) y fisicoquímicas de tres puntos de vertientes: *Puente Alto, Puente Carretera y la bocatoma El Diviso* (**Figura 1**), en diferentes periodos de precipitaciones: altas (Octubre-Noviembre), medias (Junio-Julio) y bajas (Agosto-Septiembre). Por lo tanto, se han capturado diez indicadores fisicoquímicos, cinco biológicos y tres propiedades que describen el periodo de precipitación. Estos indicadores se definen a continuación.

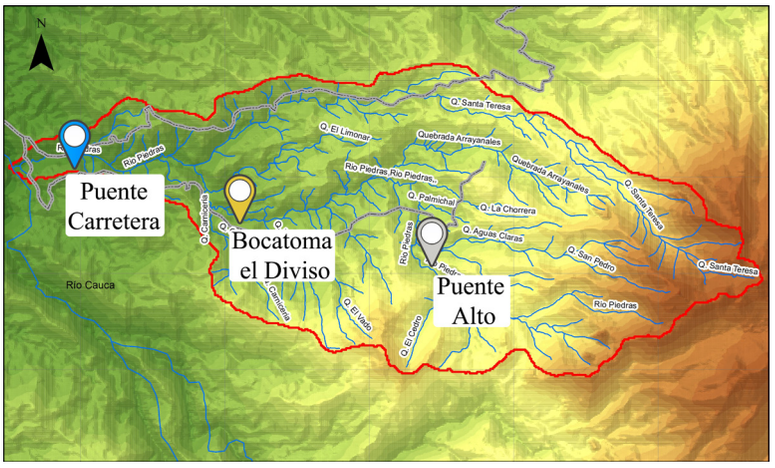

Figure 1. Location sampling points / Figura 1. Localización de los puntos de muestreo

Indicadores fisicoquímicos

Las variables fisicoquímicas son descritas por Pérez (2003), y se presentan de la siguiente manera:

- Temperatura (T): actúa en procesos de absorción de oxígeno, actividad biológica, precipitación de compuestos, formación de depósitos y modificación de la solubilidad de las sustancias; unidad de medida: grados celsius (°C).
- Conductividad (C): se usa como un coeficiente de concentración del soluto (entre sólidos) disuelto en agua; unidad de medida: µs/cm.
- Total de Sólidos Disueltos [TDS]: Mide sustancias orgánicas e inorgánicas en forma molecular o agua ionizada microgranular; unidad de Medida: mg/L.
- Oxígeno Disuelto [DO]: cantidad de oxígeno disuelto en el agua, indicador de la medida de la contaminación del agua. Un alto nivel de oxígeno disuelto indica mejor calidad del agua; unidad de medida: mg/L.
- PH: mide la concentración de iones de hidrógeno en el agua; las aguas naturales (descontaminadas) muestran un rango de PH de 5-9.
- Amoníaco [Am]: se forma durante la biodegradación de compuestos de nitrógeno orgánico, un alto nivel causa daños a ríos o estanques; unidad de medida: mg/L.
- Nitratos [Nitra]: son nutrientes requeridos por plantas y animales acuáticos para la creación de proteínas; la descomposición de plantas y animales muertos y el excremento de nitratos de animales vivos son descargados en ecosistemas acuáticos; unidad de medida: mg/L.

A. Study area description

The data used in this study was collected from Piedras river basin, located in Cauca, Colombia (Source: 76° 31' 10" west of Greenwich and 2° 21' 45" north, Outfall: 76° 23' 45" west and 2° 25' 40" north of Cauca river) by the Environmental Studies Group at the Universidad del Cauca, between 2011 and 2013, implementing the methodology followed by Pérez (2003).

The captured samples contained biological (macro-invertebrate) and physicochemical variables from three watershed points: *Puente Alto, Puente Carretera and El Diviso intake* (**Figure 1**), at different precipitation periods: high (October–November), medium (June–July) and low (August–September).

Thus, 10 physicochemical and, 5 biological indicators and 3 attributes were captured, which describe the precipitation period. These indicators are defined below:

Physicochemical indicators

The physicochemical variables are described by the authors (Pérez, 2003), and are as follows:

- Temperature (T): acts on oxygen absorption processes, biological activity, precipitation of compounds, deposit formation and modification of the solubility of substances. Measuring unit: degrees Celsius (° C).
- Conductivity (C): used as the solute concentration ratio (amount of solids) dissolved in water. Measuring unit: µs/cm.
- Total Dissolved Solids (TDS): measuring organic and inorganic substances, in molecular form, or micro-granular ionized water. Measuring unit: mg/L.
- Dissolved oxygen (DO): amount of oxygen dissolved in the water. It is an indicator that measures water pollution. A high level of dissolved oxygen indicates better water quality. Measuring unit: mg/L.
- PH: measures the concentration of hydrogen ions in the water. Natural waters (uncontaminated) exhibit a pH range 5–9.
- Ammonia (Am): formed during the biodegradation of organic nitrogen compounds. A high level causes damage to rivers or ponds. Measuring unit: mg/L.
- Nitrates (Nitra): nutrient required for aquatic plants and animals for creating proteins. The decomposi-

tion of dead plants and animals and excrement of live animal nitrates is discharged in aquatic ecosystems. Measuring unit: mg/L.

- Nitrite (Nitri): naturally transformed from nitrates, and its presence in water indicates fecal contamination. Measuring unit: mg/L.
- Phosphates (F): are essential nutrients for aquatic organisms in both natural waters and sewage; they are necessary for reproduction and synthesis of new cell tissue. Measuring unit: mg/L.
- Turbidity (Tu): lack of transparency in the water due to insoluble materials in suspension or colloids (clay, silt, dirt, etc.); the more material is in the water (high turbidity), the lower the concentration of oxygen in the same. Measuring unit: NTU.

Biological indicators

Biological samples were collected at the sampling points and they were classified by taxonomic keys, including: class, order, family, taxon and number of individuals (Pérez, 2003). A brief description of the collected samples is presented below:

- Acari: living in clean and highly oxygenated inland (freshwater), lotic (flowing) and lentic (stagnant) waters.
- Pelecypoda: belong to highly oxygenated marine and inland aquatic ecosystems, are highly sensitive to pollution and are thus considered excellent biomarkers to determine water quality.
- Plecoptera: living in clean, rough, cold, highly oxygenated lotic inland waters; used as biomarkers to determine water quality.
- Lepidoptera: living in both lentic and lotic waters, on stony bottoms, and highly oxygenated submerged vegetation. The species are intolerant to eutrophication (chemical pollution of water).
- Coleoptera: inhabit clean, shallow lotic and lentic inland water with high concentrations of oxygen, average temperatures and low speed.
- Diptera: living in terrestrial niches in deep and shallow lotic and lentic inland waters. This order includes parasites, predators and degraders, and has by virtue of this become of major health significance (poor water quality). Species are tolerant to different degrees of contamination.
- Ephemeroptera: live in clear, clean lotic inland wa-

- Nitrito [Nitri]: transformación natural de los nitratos, su presencia en agua indica características de contaminación fecal; unidad de medida: mg/L.
- Fosfatos [F]: nutrientes esenciales para organismos acuáticos en aguas naturales y de alcantarillado; son necesarios para la reproducción y síntesis de nuevos tejidos celulares; unidad de medida: mg/L.
- Turbidez [Tu]: falta de transparencia en el agua debido a materiales insoluble en suspensión o coloides (arcilla, sedimento, suciedad, etc). Si la mayoría del material está en el agua (alta turbidez), hay menor concentración de oxígeno en ella; unidad de medida: NTU.

Indicadores biológicos

Las muestras biológicas fueron recolectadas en los mismo puntos de muestreo y clasificadas por claves taxonómicas, incluyendo: clase, orden, familia, taxón y número de individuos (Pérez, 2003). A continuación se presenta una breve descripción de las muestras recolectadas:

- Acari: viven en aguas continentales (de agua dulce), lóticas (aguas corrientes) y lénticas (aguas estancadas), altamente oxigenadas.
- Pelecypoda: pertenecen a ecosistemas acuáticos marinos y continentales altamente oxigenados, son altamente sensibles a la contaminación, por lo que son considerados excelentes indicadores biológicos para determinar la calidad del agua.
- Plecoptera: viven en aguas continentales limpias, ásperas, frías y lóticas altamente oxigenadas, se usan como indicadores biológicos para determinar la calidad del agua.
- Lepidoptera: viven en aguas lénticas y lóticas, en fondos pedregosos y en vegetación sumergida altamente oxigenada, las especies son intolerantes a la eutrofización (contaminación química del agua).
- Coleoptera: habitan en aguas lóticas y lénticas limpias poco profundas (superficiales) con altas concentraciones de oxígeno, a temperaturas medias y de baja velocidad.
- Diptera: viven en nichos terrestres, como en aguas continentales profundas y superficiales, lóticas y lénticas. Este orden incluye parásitos, depredadores y degradadores, en virtud de lo cual, se han convertido en un importante riesgo para la salud (baja calidad del agua); son especies tolerantes a diferentes grados de contaminación.
- Ephemeroptera: viven en aguas continentales lóticas, claras y limpias, bien oxigenadas y con bajo contenido de residuos; son usados como indicadores biológicos de la calidad del agua.
- Hemiptera: viven en aguas continentales lóticas y lénticas de baja velocidad; algunas especies (neuston) soportan un cierto grado de salinidad y temperaturas altas; se usan como un marcador biológico en aguas superficiales.
- Odonata: viven en aguas continentales lóticas y lénticas de baja velocidad, superficiales y rodeadas de abundan-

te vegetación acuática sumergida o emergente; algunas especies pueden soportar un cierto grado de contaminación.

- Trichoptera: la mayoría de especies lóticas viven en aguas continentales (bajo rocas, troncos y material vegetal) y unas pocas en aguas lénticas, limpias y oxigenadas.
- Tricladida: viven en aguas poco profundas, tanto lénticas, como lóticas; la mayoría vive en aguas bien oxigenadas, pero algunas especies pueden soportar altos niveles de contaminación orgánica.
- Isopoda: comunes en los hábitats marinos, aunque algunas especies son de agua dulce y terreno; un gran número de especies de este orden, indica enriquecimiento orgánico.
- Glossiphoniformes: ectoparásitos de peces de aguas continentales, este orden tiene una alta tolerancia a la contaminación del agua.
- Haplotaxida: la mayoría de organismos de este orden viven en aguas eutróficas lénticas y lóticas, de fondo fangoso y con muchos residuos; son especies altamente tolerantes a la contaminación orgánica.

Periodo de precipitación

Se describe el periodo de precipitación en el cual fueron tomadas las muestras. Se presentan por año, mes y código del punto de muestreo. Estos indicadores se describen en la **Tabla 1**.

ters with well oxygenated low organic content of waste; used as bio-indicators of water quality.

- Hemiptera: living in lotic and lentic low speed inland waters. Some species (Neuston) withstand some degree of salinity and high temperatures; used as a biomarker in surface waters.
- Odonata: living in shallow, low speed lotic and lentic inland waters, surrounded by abundant submerged or emergent aquatic vegetation. Some species can withstand a certain degree of contamination.
- Trichoptera: most lotic species live in inland waters (under rocks, logs and plant material) and a few live in lentic, clean and oxygenated water.
- Tricladida: live in both lentic and lotic shallow water. Most live in well oxygenated waters, but some species can withstand high levels of organic contamination.
- Isopoda: common in marine habitats, but some are freshwater species and many are land-based. Large numbers of species of this order indicate organic enrichment.
- Glossiphoniformes: ectoparasites of fish in inland waters; this order has a high tolerance to water contamination.
- Haplotaxida: most living organisms of this order live in eutrophic lentic and lotic waters, with a muddy bottom and plenty of waste. The species are highly tolerant of organic contamination.

Precipitation period

This describes the precipitation period in which the samples were taken. They are presented by: year, month and sampling point code. These indicators are described in **Table 1**.

Table 1. Piedras river database attributes / Tabla 1. Base de datos de las propiedades del río Piedras

Category / *Categoría*	**Attribute** / *Propiedades*	**Measuring Unit** / *Unidad de medida*	**Range** / *Rango*
Physicochemical Indicators / *Indicadores Fisicoquímicos*	Temperature / *Temperatura*	°C	13.0 - 17.8
	Conductivity / *Conductividad*	µs/cm	35.2 - 89.0
	Total Dissolved Solids / *Total de Sólidos Disueltos*	mg/L	16.5-42.1
	Dissolved Oxygen / *Oxígeno disuelto*	mg/L	7.17-8.23
	pH	mg/L	6.62-8.17
	Ammonia / *Amoníacos*	mg/L	0.01-0.04
	Nitrates / *Nitratos*	mg/L	0.01-0.09
	Nitrite / *Nitrito*	mg/L	0.01-0.06
	Phosphate / *Fosfato*	mg/L	0.08-0.24
	Turbidity / *Turbidez*	mg/L	1.0-9.8
Biological Indicators / *Indicadores Biológicos*	Class / *Clase*	-	-
	Order / *Orden*	-	-
	Family / *Familia*	-	-
	Taxon / *Taxón*	-	-
	Number of Individuals / *Número de Individuos*	-	-
Precipitation Periods / *Periodos de Precipitación*	Month / *Mes*	-	-
	Year / *Año*	-	-
	Sampling Point Code / *Código de Punto de Muestreo*	-	-

B. Background

Clustering algorithms are among the unsupervised learning methods, which divide a dataset into a number of groups, so that the elements (observations) of a same group are homoge-

neous (similar) (Gurrutxaga, Muguerza, Arbelaitz, Pérez, & Martín, 2011; Arbelaitz, Gurrutxaga, Muguerza, Pérez, & Perona, 2013). Several authors (Lin & Chen, 2005; Gan, Ma & Wu, 2007) have classified clustering algorithms into three types, which are presented below.

Partitional algorithms

Partitional algorithms assume in advance the number of non-overlapping groups that the dataset should be divided into, determining a split into groups or classes that ensures that an observation belongs to exactly one group (Gan et al., 2007; Arbelaitz et al., 2013). The most commonly used search algorithms are: K-Means and K-Medoids. K-Means defines k centroids (one for each group) and then takes each element of the data set in order to locate the nearest centroid. It recalculates the centroid of each group and redistributes each element of the data set using the same criteria (Velmurugan & Santhanam, 2010). The K-Means algorithm takes the average value of the objects in a group as a reference point (seed), indicating that the centroids do not need to match the values of the objects in the cluster to which they belong. The K-Medoids algorithm is a variant of the K-Means algorithm. In contrast to the K-means algorithm, K-Medoids choose points belonging to the dataset as centers or reference points (seed).

Hierarchical algorithms

These algorithms establish a hierarchical structure as a result of grouping (Theodoridis, Pikrakis, Koutroumbas, & Cavouras, 2010). Such techniques decompose the dataset into levels or stages, such that on each level (agglomerative) or bind divide (partition) the previous level groups (Pang-Ning, Steinbach, & Kumar, 2006; Theodoridis et al., 2010). The agglomerative analysis method starts with as many groups as there are observations in the dataset. From here, new groups are formed until the end of the process, when all the cases treated are contained in a single group, while the dividing technique is the reverse process, obtaining at the end of the process as many groups as cases have been treated.

It should be noted that the hierarchical algorithms used in the environmental sector are actually of the agglomerative type (Moreno, 2000) and they are much more understandable (easy to interpret) and effective (less complex) than divisive algorithms, since in the first group, merging the groups corresponds to a high degree of similarity, thus making them more comprehensible,

B. Antecedentes

Los algoritmos de agrupamiento son parte de los métodos de aprendizaje sin supervisión, que dividen los conjuntos datos en un número de grupos, de tal manera que los elementos (observaciones) de un mismo grupo son homogéneos (Gurrutxaga, Muguerza, Arbelaitz, Pérez, & Martín, 2011; Arbelaitz, Gurrutxaga, Muguerza, Pérez, & Perona, 2013). En este sentido, Lin y Chen (2005) y Gan, Ma y Wu (2007) clasificaron los algoritmos de agrupamiento en los siguientes tres tipos: particionales, jerárquicos y basado en la densidad.

Algotirmos particionales

Asumen de antemano el número de grupos no sobrepuestos en que se debería dividir el conjunto de datos, llegan a una división de grupos o clases que asegura que una observación pertenece a un grupo en particular (Gan et al., 2007; Arbelaitz et al., 2013). Los algoritmos de búsqueda más comúnmente usados son: K-Means y K-Medoids. K-Means define K centroides (uno por cada grupo) y luego toma cada elemento del conjunto de datos –con el fin de localizar el centroide más cercano–, recalcula el centroide de cada grupo y redistribuye cada elemento del conjunto de datos, usándolo (Velmurugan & Santhanam, 2010). El algoritmo K-Means toma el valor promedio de los objetos en un grupo como punto de referencia (semilla), indicando que los centroides no necesitan coincidir con los valores de los objetos en el grupo al cual pertenecen. El algoritmo K-Medoids es una variante del algoritmo K-Means. A diferencia de él, escoge puntos pertenecientes al conjunto de datos como centros o puntos de referencia (semilla).

Algoritmos jerárquicos

Establecen una estructura jerárquica como resultado de una agrupación (Theodoridis, Pikrakis, Koutroumbas, & Cavouras, 2010). Descomponen el conjunto de datos en niveles o etapas, de tal manera que en cada nivel (aglomerativo) o unión dividen (particionan) los grupos de los niveles anteriores (Pang-Ning, Steinbach, & Kumar, 2006; Theodoridis et al., 2010). El método del análisis aglomerativo empieza con tantos grupos como observaciones hayan en el conjunto de datos; desde ahí se forman nuevos grupos hasta que, al final del proceso, todos los casos tratados están en un solo grupo; mientras, la técnica de separación es el proceso inverso al previamente obtenido, al final del proceso, tanto grupos, como casos, han sido tratados.

Se debe notar que los algoritmos jerárquicos usados en el sector ambiental son de tipo aglomerativo (Moreno, 2000) y mucho más entendibles (de fácil interpretación) y efectivos (menos complejos) que los algoritmos divisivos; desde el primer grupo, se fusionan los grupos similares, de forma tal que se tornan más comprensivos, mientras que la última división de grupo apunta a minimizar, la variación general del grupo, lo que conlleva a un entendimiento mucho más complejo de los grupos (Sasirekha & Baby, 2013; de Mantaras, & Saitia, 2004). Por lo tanto, el resultado de este tipo de algoritmos debería ser examinado exhaustivamente para asegurar que tiene sentido; por esta razón este trabajo se enfoca en los análisis de los métodos aglomerativos. Las estrategias de búsqueda más repre-

sentativas son: la mínima distancia (Enlace Simple), la máxima distancia (Enlace Completo) y la distancia promedio (Enlace Promedio) (Madhulatha, 2012; Pang-Ning et al., 2006). Un paso clave en el proceso de agrupamiento jerárquico es seleccionar la medida que describe la distancia entre observaciones; Madhulatha (2012) argumenta que la más común, de mayor uso y con resultados confiables en la mayoría de los casos es la medida de distancia euclidiana, la misma que se usa en la presente investigación para medir la similitud entre los grupos.

Algoritmos basados en la densidad

Esta clase de algoritmos divide el conjunto de datos en grupos teniendo en cuenta la distribución de la densidad de los elementos, permitiendo que los grupos sean formados con una alta densidad de puntos en su interior. Entre los algoritmos usados en esta área, Density-Based Spatial Clustering of Applications with Noise [DBSCAN] agrupa las observaciones de un conjunto de datos en regiones de alta y baja densidad. Inicialmente el algoritmo define un conjunto de objetos centrales (objetos que son, en su vecindario, una cantidad mayor o igual que unos puntos límite específicos), objetos de borde o límite (objetos que son, en su vecindario, una cantidad menor que un punto límite específico, pero que se encuentran en la vecindad de un objeto central) y un objeto atípico o ruido (objetos que no entran en las categorías anteriores). Una vez los objetos estén definidos, DBSCAN selecciona un elemento arbitrario p; si p es un objeto central, un grupo en el cual estén localizadas todas las observaciones de densidad alcanzable p, está listo. Si p no es un objeto central se escoge otro elemento del conjunto de datos. El proceso continúa hasta que todos los elementos hayan sido asignados a un objeto. Aquellos que se encuentran por fuera de los grupos, forman objetos llamados valores atípicos o puntos de ruido, y los elementos que no son atípicos o valores fundamentales reciben el nombre de puntos de borde (Gan et al., 2007; Pang-Ning et al., 2006). Es notable que la técnica DBSCAN puede manejar grupos de diferentes formas y tamaños, pero presenta limitaciones cuando el conjunto de datos de grupos de dimensión unidos se encuentra superpuesto y en la presencia de ruido (González, 2010), por lo tanto, en esta investigación no se tuvo en cuenta esta técnica debido al conjunto de datos con el que se contaba.

III. Resultados y discusión

En esta sección se presentan las evaluaciones de algoritmos de agrupamiento ya discutidas, aplicadas al conjunto de datos que se describió en la sección II.

A. Resultados experimentales

Se llevó a cabo una evaluación experimental para determinar el comportamiento de los métodos de agrupación jerárquica y particional aplicados en diferentes conjuntos de datos (Tabla 2), usando solo índices de grupos de validación (IVG) debido a la ausencia de la partición correcta. Los IVG se basan principalmente en el concepto de cohesión del conjunto y el espacio entre ellos (Pang-Ning et al., 2006). Las medidas de cohesión determinan el grado de relación u homogeneidad de objetos en un grupo (pertenencia), mientras que la segregación determina

while in the latter group the division aims to minimize the overall variance of the group, leading to a much more complex understanding of the groups (Sasirekha & Baby, 2013; de Mantaras, & Saitia, 2004). Therefore the results of this type of algorithm should be examined thoroughly to ensure that they make sense, and for this reason, this work focuses on the analysis of the agglomerative methods. The most representative search strategies are: the minimum distance (single linkage), the maximum distance (complete linkage) and the average distance (average linkage) (Madhulatha, 2012; Pang-Ning et al., 2006). A key step in the process of hierarchical clustering is to select the measurement that describes the distance between observations; in Madhulatha (2012), the author argues that the most common and most often used measurement, and one which gives reliable results in most cases, is the Euclidean distance measure. In this research, we use the Euclidean distance to measure the similarity between clusters.

Density-based algorithms

This class of algorithms divides the dataset into groups, taking into account the density distribution of the elements, so that the groups are formed having a high density of points inside. Among the algorithms used in this area is DBSCAN (Density-Based Spatial Clustering of Applications with Noise), which groups the observations of a dataset into regions of high and low density. Initially, the algorithm defines a set of central objects (objects that have in their neighborhood an amount greater than or equal to a specific threshold point), edge or boundary objects (objects that have in their neighborhood a number of points less than a specified threshold, but which are in the vicinity of a central object) and atypical objects or noise (objects that do not fall into the above categories). Once the objects are defined, DBSCAN selects a p arbitrary element; if p is a central object, a group in which are located all observations density-reachable p is built. If p is not a central object, another element of the dataset is visited. The process continues until all elements have been assigned to an object. Those outside the groups that have formed objects are called outliers or noise points, while items that are not atypical or core values are called edge points (Gan et al., 2007; Pang-Ning et al., 2006). It is noteworthy that the DBSCAN technique can handle groups of different shapes and sizes, but has limitations when the dataset has tied dimensions, when groups are overlapping, and in the presence

of noise (González, 2010). Therefore, in this research this technique is not taken into account because of the datasets involved.

III. Results and discussion

This section presents the clustering assessments algorithms discussed above, applied to the dataset described in section II.

A. Experimental results

An experimental evaluation was conducted to determine the performance of the methods of hierarchical and partitional clustering applied on different data sets (Table 2), using only the Indexes Validation Group (IVG) due to the absence of the correct partition. The IVG is mostly based on the concept of cohesion of the assembly and the spacing between them (Pang-Ning et al., 2006). Cohesion measures determine the degree of relationship or homogeneity of objects in a group (membership), while the segregation determines the degree of separation between the groups (non-membership).

For the experiments, four datasets derived from the Piedras river database were used, which contain information on the physicochemical and biological samples collected. It should be mentioned that such referrals in the dataset were made taking into account the reduction technique attributes of Principal Component Analysis (PCA), which aims to select the most appropriate subset of features of the original dataset, while discarding redundant attributes (correlated), and any that are inconsistent, irrelevant, etc. (useless attributes). These attributes are described in **Table 2.**

In this way, the datasets in question were used with the techniques provided above to give the results displayed further on.

Performance analysis of partitional algorithms

To evaluate the performance of the partitional algorithms K-Means and K-Medoids, we must first find an optimal set of groups that fit in the best possible way the natural group of input data for the range tested K=3 to K=20, and the result of each iteration with the IVG des-

el grado de separación entre los grupos (no pertenecientes).

Para los experimentos, se usaron cuatro conjuntos de datos derivados del conjunto de datos del río Piedras, los cuales contenían información en las muestras fisicoquímicas y biológicas que fueron recolectadas. Estas referencias en el conjunto de datos fueron hechas tomando en cuenta los atributos de la técnica de reducción Principal Component Analysis (PCA), la cual permite seleccionar el subconjunto más apropiado de características provenientes del conjunto de datos original, descartando atributos redundantes (correlacionadas), inconsistentes, irrelevantes, etc. (atributos inútiles). Estos atributos se describen en la **Tabla 2.**

Table 2. Attributes of the respective datasets / Tabla 2. Atributos de los respectivos conjuntos de datos

Dataset / *Conjunto de Datos*	**Number of Attributes** / *Número de Atributos*	**Attributes** / *Atributos*
DS0	18	T, C, TSD, OD, pH, Am, Nitri, Nitra, F, Tu, Cl, Or, Fam, Tax, NI, Year, Month, CM
DS1	5	Year, Month, CM, Tax, NI
DS2	8	Year, Month, CM, Cl, Or, Fam, Tax, NI
DS3	15	Year, Month, CM, Tax, NI, T, C, TSD, OD, pH, Am, Nitri, Nitra, F, Tu

De esta manera, los conjuntos de datos en cuestión fueron usados con las técnicas proporcionadas anteriormente para reproducir los resultados

Análisis del rendimiento de algoritmos particionales

Para evaluar el rendimiento de los algoritmos particionales K-Means y K-Medoids, en primer lugar se encontró un conjunto óptimo de grupos que encajaran de la mejor manera posible en el grupo natural de datos de entrada para ese rango evaluado de K=3 a K=20, y el resultado de cada iteración con el IVC descrito en la sección previa. De la misma manera, este procedimiento se llevó a cabo para 4 conjuntos de datos descritos en la **Tabla 2**. La **Figura 2** permite describir el comportamiento de los resultados del algoritmo K-Means (**Figura 2a**) y K-Medoids (**Figura 2b**) respectivamente en el conjunto de datos DS0, en donde el color azul simboliza el índice DB (basado en la similitud media entre dos grupos que denotan el valor mínimo y la mejor partición), y referencias rojas para el índice Silueta (combina tanto cohesión como separación, entre más grande sea la Silueta, es más compacta y de grupos separados, es decir, su valor máximo denota la mejor partición).

Como se menciono, un valor máximo de la figura del índice Silueta y un valor mínimo del índice DB representan la mejor partición, y por lo tanto, el valor óptimo de K. Del mismo modo, en la **Figura 2** el número óptimo de grupos para el método K-Means se encuentra entre 5 y 6 grupos, mientras que para K-Medoids corresponde a 7 grupos. En este caso, las medidas para definir el número apropiado de grupos son mejores para K-Medoids en comparación con K-Means, desde que los valores del índice DB coinciden con el índice Silueta y K=7.

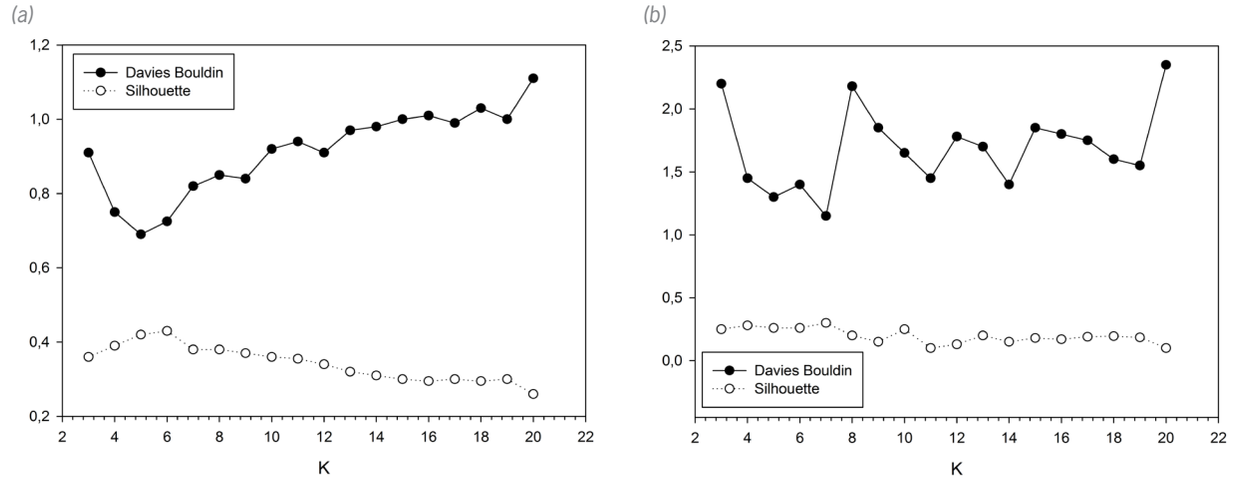

Figure 2. K-Means vs. K-Medoids behavior in DS0 with respect to Silhouette and DB. (a) K-Means (b) K-Medoids / Figura 2. El comportamiento de K-Means vs. K-Medoids en DS0 respecto a los índices Silueta y DB. (a) K-Means (b) K-Medoids

De igual manera se analiza el comportamiento de los algoritmos aplicados en cuestión al conjunto de datos del DS1 (**Figura 3**). En la **Figura 2** (a) y (b), se puede observar que los resultados están invertidos en comparación con el caso previo, es decir, K-Means adquiere mayor precisión para el número apropiado de grupos (K=3), mientras K-Medoids ofrece dos posibilidades para este valor (K=3 o K=4).

La **Figura 4** muestra los resultados obtenidos cuando el conjunto de datos DS2 es evaluado en K-Means y K-Medoids.

Means obtenido con la mayor precisión percibida en número hace referencia a una agrupación óptima (K=3) versus el algoritmo K-Medoids. Estos resultados indican que los conjuntos de datos DS1 y DS2 son muy similares.

La **Figura 5** muestra los resultados del conjunto de datos DS3, donde se obtienen resultados similares a DS0. A partir de aquí se puede asumir que existe información redundante entre los conjuntos de datos DS0-DS3 y DS1-DS2. Por tanto, se puede decir que DS3 representa la misma información que DS0, a pesar de que el primero posee menos atributos. Lo mis-

cribed in the previous section is evaluated. In the same way, note that this procedure was performed for the four datasets described in **Table 2**. **Figure 2** describes the behavior of the results of the K-Means algorithm (**Figure 2a**) and K-Medoids (**Figure 2b**) respectively on the dataset DS0, where the blue color symbolizes the DB index (based on the similarity measure between two groups denoting the minimum value and the best partition) and red refers to the Silhouette index (combining both cohesion and separation, how greater the Silhouette, more compact and separate the groups, i.e., its maximum value denotes the best partition).

As mentioned above, a maximum value of the Silhouette index figure and a minimum value of the DB index represent the best partition and thus the optimum value of K. Therefore from **Figure 2** the optimal number of groups for the K-Means method is between 5

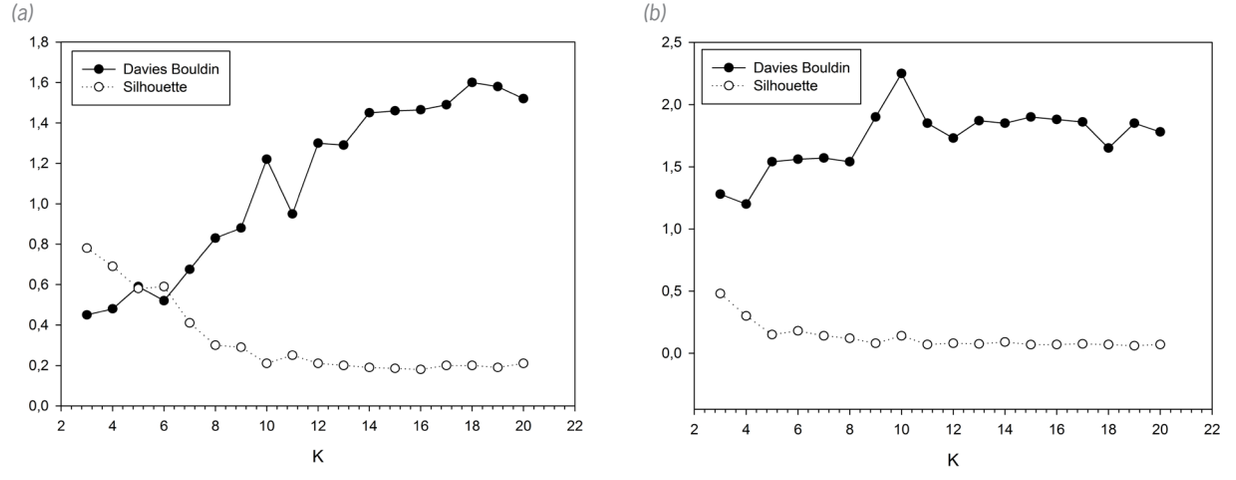

Figure 3. K-Means vs. K-Medoids behavior in DS1 with respect to Silhouette and DB. (a) K-Means (b) K-Medoids / Figura 3. El comportamiento K-Means vs. K-Medoids en DS1 respecto a los índices Silueta y DB. a) K-Means (b) K-Medoids

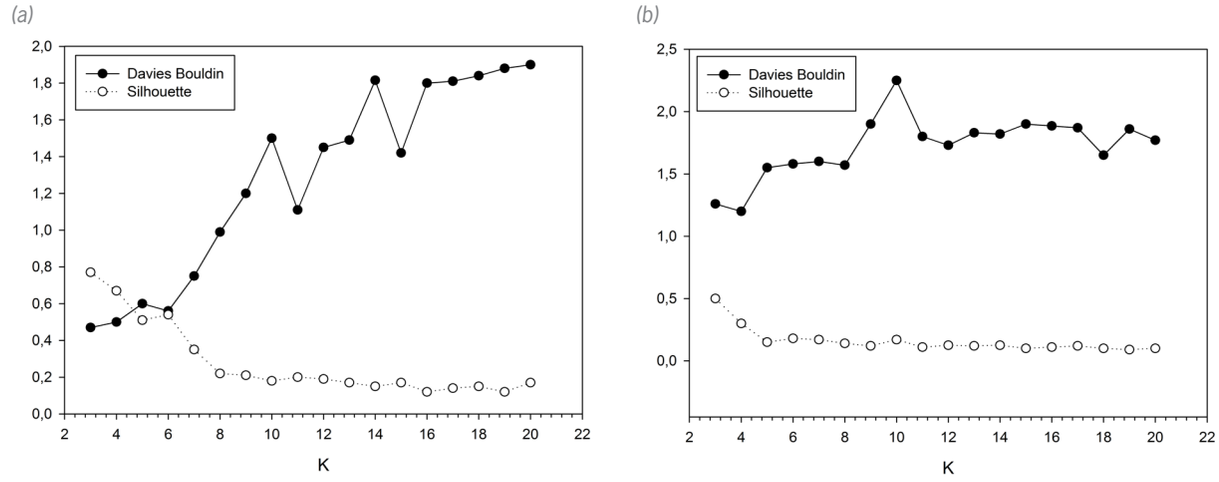

Figure 4. K-Means vs. K-Medoids behavior in DS2 with respect to Silhouette and DB. (a) K-Means (b) K-Medoids / Figura 4. El comportamiento de K-Means vs. K-Medoids en DS2 respecto a los índices Silueta y DB. (a) K-Means (b) K-Medoids

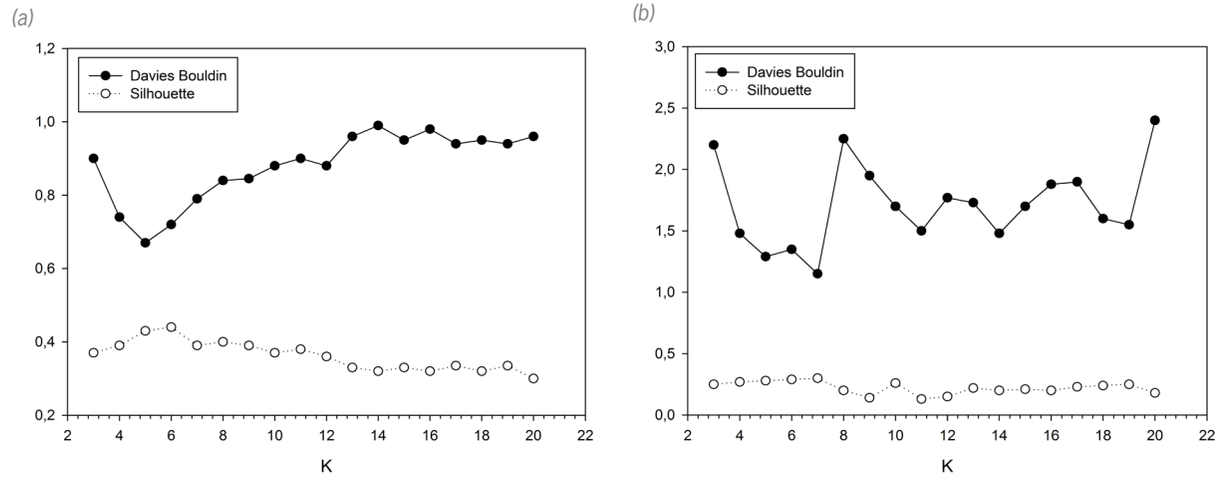

Figure 5. K-Means vs. K-Medoids behavior in DS3 with respect to Silhouette and DB. (a) K-Means (b) K-Medoids / Figura 5. El comportamiento K-Means vs. K-Medoids en DS1 respecto a los índices Silueta y DB. (a) K-Means (b) K-Medoids

and 6 groups, while for the K-Medoids it is 7 groups. In this case, measures to define the appropriate number of groups are better for K-Medoids than for K-Means, since the values of DB agree with Silhouette and K = 7.

Similarly, the algorithms' behavior in question applied to the dataset of DS1 (**Figure 3**) is analyzed. In **Figure 2** (a) and (b), note that the results are reversed compared to the previous case; that is, K-Means acquires accurately the appropriate number of groups (K=3), while K-Medoids delivers two possibilities for this value (K=3 or K=4).

Now, **Figure 4** shows the results obtained when the dataset DS2 is evaluated on K-Means and K-Medoids.

Here the same behavior of one data record (DS1) is perceived, wherein the K-Means partitional method obtained accurately perceived in number refers to the opmo ocurre con DS1 vs. DS2. Este comportamiento indica que las bases de datos DS0 y DS2 podrían estar representadas por DS1 y DS3, respectivamente.

A continuación, se resumen los mejores valores de los índices Silueta y DB (mínimo y máximo respectivamente) obtenidos por los algoritmos K-Means y K-Medoids en cada una de las bases de datos. En este sentido, los mejores resultados se obtienen cuando los algoritmos K-Means y K-Medoids procesan el conjunto de datos DS1 porque obtienen el valor mínimo del índice DB (0.458 y 1.101, respectivamente) y los más altos resultados del índice Silueta (0.773 y 0.526, respectivamente). A partir de estos resultados, K-Means alcanza la mejor calidad para el proceso de agrupamiento de DS1 con K=3 para ambos casos. De modo significativo, K-Means, de manera diferente a K-Medoids, presenta los mejores valores para la validación de los índices DB y Silueta, mínimos y máximos respectivamente.

Análisis de rendimiento de algoritmos jerárquicos

En este punto se evaluaron los algoritmos: enlace simple, completo y promedio. El análisis de rendimiento se desarrolló

para cada conjunto de datos definidos en la Tabla 2, basados en los índices CCC y Silueta con medidas del grado de distorsión, y el número óptimo de agrupaciones de manera respectiva.

La **Figura 6** muestran el comportamiento de estos algoritmos para diferentes números de agrupamientos (3-20). Como se puede notar claramente, todos los algoritmos presentan los mejores valores en Silueta cuando se agrupan los datos en K=3, obteniendo el mismo número óptimo de agrupaciones como de métodos particionales.

El Enlace Simple presenta los mejores resultados de Silueta comparado con los Enlaces Completo y Promedio (0.8, 0.88, 0.88, 0.8 respectivamente) en los cuatro conjuntos de datos analizados. Al analizar estos resultados en detalle, se puede observar que el mismo comportamiento de los conjuntos de datos se obtiene a partir del análisis de los métodos particionales, corroborando así la existencia de datos redundantes.

Por otra parte, dado que los algoritmos jerárquicos presentan una estructura estratificada como resultado del grupo (dendrograma), la mejor manera de obtener una evaluación precisa es utilizando el índice CCC. La **Tabla 3** muestra el grado de distorsión de las relaciones entre las observaciones.

Se observó que el algoritmo Enlace Promedio obtuvo los mejores valores para el índice CCC (0,631, 0,681, 0,69 y 0,713, respectivamente), los cuatro conjuntos de datos analizados. Esto permite concluir, que el Enlace Promedio es la estrategia que genera menor distorsión en las relaciones y por tanto, se obtienen observaciones relacionadas de manera más apropiada.

Según este enfoque, los mejores conjuntos de datos agrupados fueron DS1 y DS2, los cuales obtuvieron los mejores valores de CCC (0,681 y 0,69, respectivamente), teniendo presente el mayor valor de Silueta alcanzado al agrupar el conjunto de datos DS1. Por esta razón DS1 es considerado el mejor conjunto de datos, lo cual permite corroborar los resultados obtenidos por el algoritmo particional.

B. Entendiendo los datos

En la sección anterior se analizaron algoritmos de agrupamiento particional y jerárquicos, donde se obtuvieron los mejores resultados para K-Means (K = 3) y Enlace Promedio, respectivamente, así como DS1 fue el conjunto de datos que demostró los mejores resultados. Esta sección ofrece un análisis de los grupos obtenidos a través de la ejecución del algoritmo K-Means. El análisis de Enlace Promedio no es tenido en cuenda dado que los algoritmos de agrupamiento jerárquico se corrompen cuando el conjunto de datos es de altas dimensiones debido a sus altos costos y su complejidad no lineal en el tiempo (Entregan resultados descriptivos y poco confiables). Por lo tanto, la literatura sugiere que este tipo de técnicas son muy eficaces para conjuntos de datos de bajas dimensiones (Quiroz, Pla, Badia, & Chover, 2007; Saraçli, Doğan, & Doğan, 2013).

El árbol de decisión C.4.5 fue utilizado para interpretar la composición de los grupos generados por K-means, teniendo en cuenta que este algoritmo es uno de los más utilizados

timal clustering (K=3) versus the K-Medoids algorithm. These results indicate that the DS1 and DS2 datasets are very similar.

However, **Figure 5** shows the results of the DS3 dataset, where similar results are obtained to those for DS0. From this it is possible to assume that there is redundant information between datasets DS0-DS3 and DS1-DS2. In other words, we can say that DS3 represents the same information as DS0, even though the former has fewer attributes. The same goes for DS1 vs. DS2. This behavior indicates that the datasets DS0 and DS2 can be represented by DS1 and DS3 respectively.

Below are summarized the best values of the Silhouette and DB indices (minimum and maximum respectively) obtained by the K-Means algorithm and K-Medoids on each of the datasets. The best results are obtained when the K-Means algorithm and K-Medoids process data set DS1, because they obtain the minimum values of DB (0.458 and 1.101 respectively) and the highest values of Silhouette (0.773 and 0.526 respectively). From these results, K-Means reaches the best quality for the DS1 clustering process with K=3, while K-Medoids for DS1 reaches K=3 or K=4. Given these criteria, it can be concluded that the optimal number of groups is K=3 for both cases. Significantly, K-Means, unlike K-Medoids, presents the best values for validation of the DB and Silhouette indexes, the minimum and maximum respectively.

Performance analysis of hierarchical algorithms

In this item are evaluated single, complete and average linkage algorithms. The performance analysis was performed for each dataset defined in **Table 2**, based on the CCC and Silhouette indices which measure the degree of distortion, and the respective optimal number of clusters.

Figure 6 show the behavior of these algorithms for different numbers of clusters (3–20). As clearly noted, all the algorithms have the best values in Silhouette when grouping the data set K=3, obtaining the same optimum number of clusters as the partitional methods.

Single Linkage presents the best results for Silhouette compared with Complete and Average Linkage (0.8, 0.88, 0.88, 0.8 respectively) on the four datasets analyzed. If these results are analyzed in detail, it is noted that the behavior of the datasets is the same as is obtained from

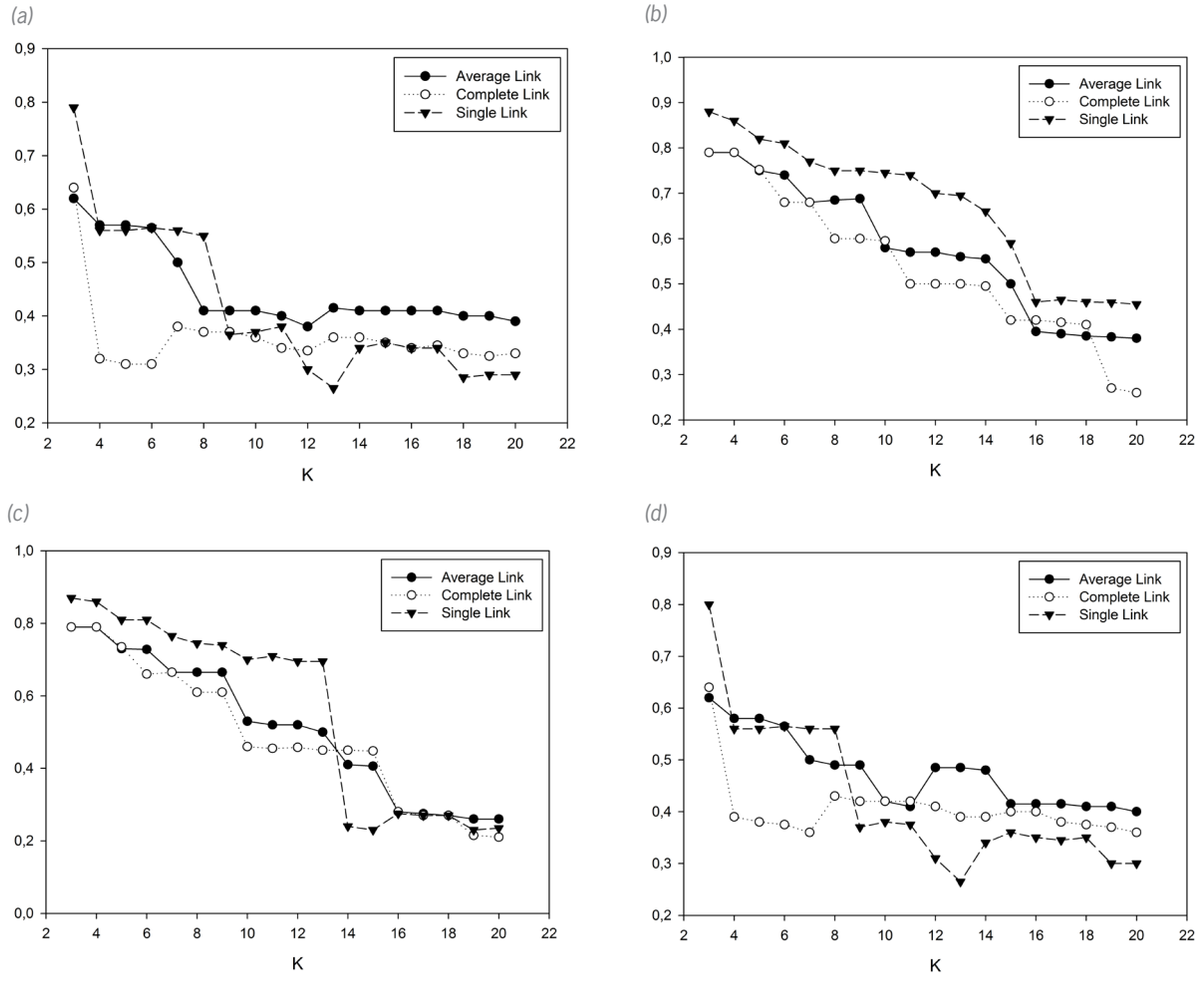

Figure 6. Behavior of hierarchical agglomerative algorithms. (a) DS0- Silhouette (b) DS1- Silhouette (c) DS2- Silhouette (d) DS3- Silhouette / Figura 6. Comportamiento de algoritmos aglomerativos jerárquicos. a) DS0- Silhouette (b) DS1- Silhouette (c) DS2- Silhouette (d) DS3- Silhouette

the analysis of partitional methods, thus corroborating the occurrence of data redundancy.

Moreover, taking into account that the hierarchical algorithms establish a tiered structure as a result of the group (dendrogram), the best way to get an accurate assessment is by using the CCC index. **Table 3** lists the degree of distortion of relations between observations.

It is observed that the Average Linkage algorithm obtained the best values for the CCC index (0.631, 0.681, 0.69 and 0.713 respectively) for the four data sets analyzed. Thus, Average Linkage is the strategy that generated the least distortion in relationships, or in other words, more appropriately related observations.

Considering this approach, it appears that the best datasets grouped were DS1 and DS2, which obtained the best values of CCC (0.681 and 0.69, respectively). However, considering that the greatest value of Silhouette was achieved when the dataset DS1 was grouped, for

Table 3. CCC index values / Tabla 3. Valores del índice CCC

	Hierarchical Algorithms / *Algoritmos Jerárquicos*		
	Single Linkage / *Enlace Simple*	Average Linkage / *Enlace Promedio*	Complete Linkage / *Enlace Completo*
Dataset	CCC	CCC	CCC
DS0	0.610	0.631	0.471
DS1	0.673	0.681	0.666
DS2	0.606	0.690	0.572
DS3	0.681	0.613	0.534

para este tipo de tareas (Corrales, Corrales, & Figueroa-Casas, 2015; Pérez, 2003). La **Figura 7** muestra el árbol de decisión resultante.

El conjunto de datos DS1 consiste en un total de 5590 individuos macroinvertebrados acuáticos, pertenecientes a 63 taxones, además de propiedades como año, mes, código de estación y número de individuos, como lo indica la **Tabla 2**. Ahora, usando la técnica de agrupación K-Means se dividió el conjunto de datos en tres grupos donde cada uno es interpretado a través del árbol de decisión C.4.5, ya descrito. De este modo,

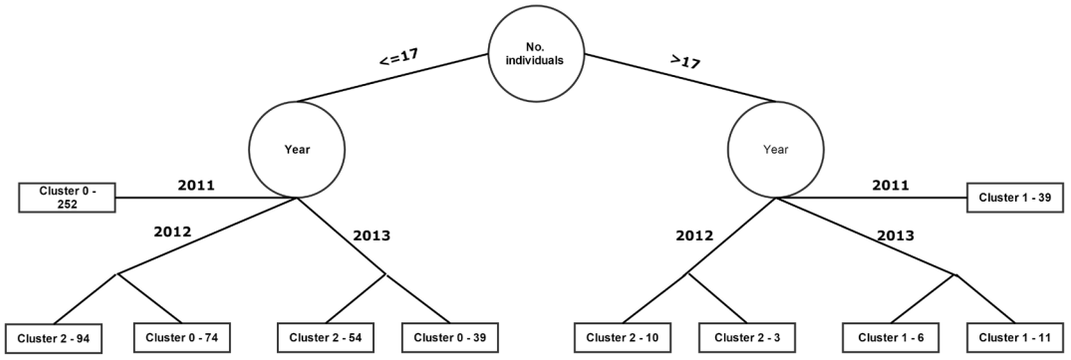

Figure 7. Decision tree / Figura 7. Árbol de decisión

la **Tabla 4** muestra la distribución de casos obtenidos por el árbol de decisión C.4.5 para individuos macro-invertebrados acuáticos que se encuentran en la cuenca del río Piedras. Asi mismo, se clasifica el porcentaje de individuos en cada grupo, teniendo en cuenta la metodología (Pérez, 2003), donde el autor ha etiquetado cada macro-invertebrado con un número que indica el grado de sensibilidad a los contaminantes. Estos números varían de forma gradual y sucesiva en un rango de 1 a 10, donde el número 1 indica la menor sensibilidad (contaminantes aceptados) y el número 10 la mas alta (acepta cualquier tipo de contaminantes). Además, considera la calidad del agua de la abundancia taxonómica mediante el cálculo de los índices biológicos BMWP y ASPT, y también el color clasificado de acuerdo a su calidad.

Por lo tanto, esta metodología es tomada para interpretar dichas agrupaciones, en la cual cada grupo tiene tres niveles de alerta para la calidad del agua (representado por los taxones de individuos encontrados) distinguidos por el color azul, verde y amarillo de acuerdo a su sensibilidad a los contaminantes.

Como lo muestra la **Tabla 4**, las agrupaciones 1 y 3 tienen la mayor diversidad de macroinvertebrados, alcanzando una representación del 40,56% y el 44,4% de los taxones recolectados, respectivamente. Donde el porcentaje de taxones, indicador de alta y buena calidad del agua, es mucho mayor en comparación con los indicadores de taxones de la calidad del

Table 4. Diversity percentage of individuals in the clusters (Dataset DS1) / Table 4. Porcentaje diversidad de los individuos en las agrupaciones (Conjunto de datos DS1)

Color / *Color*	**Represents** / *Representación*	**Cluster 1** / *Agrupación 1*	**Cluster 2** / *Agrupación 2*	**Cluster 3** / *Agrupación 3*
Blue / *Azul*	High biological water quality (very clean water) / *Alta calidad biológica del agua (agua muy limpia)*	23.63	2.2	17
Green / *Verde*	Good biological water quality (slightly polluted water) / *Buena calidad biológica del agua (agua ligeramente contaminada)*	12.93	12.68	25
Yellow / *Amarillo*	Water quality doubtful or regular / *Dudosa o regular calidad del agua*	4	0	2.4
	Total	40.56	14.88	44.4

that reason DS1 is considered the best dataset, corroborating the results obtained by the partitional algorithm.

B. Understanding the data

In the previous section we analyzed partitional and hierarchical clustering algorithms, where the best results for the K-Means (K=3) and Average Linkage were obtained respectively, just as DS1 was the dataset that showed the best results. This section provides an analysis of the groups obtained from the K-Means algorithm performance, but Average Linkage analysis is omitted because when the dataset is high-dimensional, hierarchical clustering algorithms decompose (with some descriptive results and unreliable delivery) due to their nonlinear time complexity and high cost, so the literature shows that this kind of technique is quite effective for low-dimensional datasets (Quiroz, Pla, Badia, & Chover, 2007; Saraçli, Doğan, & Doğan, 2013).

However, to interpret the composition of groups generated by K-Means, a C.4.5 decision tree was used, considering that this algorithm is one of the most used for such tasks (Corrales, Corrales, & Figueroa-Casas, 2015; Pérez, 2003). The resulting decision tree is shown in **Figure 7**.

The DS1 dataset consists of a total of 5590 individuals of aquatic macro-invertebrates, belonging to 63 taxa, plus attributes: year, month, station code and number of individuals, as indicated in Table 2. Now, by dividing the dataset into three groups by the K-Means clustering technique, each is interpreted through the C.4.5 decision tree, as indicated above. **Table 4** shows the

distribution of instances obtained by the C.4.5 decision tree for individual aquatic macro-invertebrates found in the Piedras river basin. In turn, the percentage of individuals in each group was categorized, taking into account the methodology (Pérez, 2003), where the author has labeled each macro-invertebrate with a number indicating the degree of sensitivity to pollutants. These numbers range from 1 to 10, where 1 indicates the least sensitive (accepts contaminants), and so on, gradually, until the number 10 (accepts no kind of contaminants) pointing to the most sensitivity. Additionally, water quality is considered from the taxonomic abundance by calculation of the biological indices BMWP and ASPT, and also the color is categorized according to their quality.

Thus, to interpret the clusters, this methodology is taken as a starting point, where each group has three alert levels for water quality (represented by individual taxa found), discriminated as blue, green and yellow according to their sensitivity to contaminants.

As shown in **Table 4**, clusters 1 and 3 have the greatest diversity of macro-invertebrates, reaching a representation of 40.56% and 44.4% of the collected taxa respectively. The taxa percentages indicating high and good water quality are much higher than the taxa indicators of dubious water quality, indicating good condition alerts on water quality in the three clusters.

The results obtained by the classifier are discussed in more detail, and are organized so that these can visualize the behavior of the taxonomic community at different times and sampling points (Puente Alto, Puente Carretera and El Diviso intake). **Figure 8** explains this behavior.

In general, it is clearly seen that indicators of high and good water quality taxa can be compared with taxa representing water of questionable quality in the three sampling points, thereby expressing quality alerts for relatively good water in the three sites.

Similarly, analyzing this figure more thoroughly, it displays in the first instance for the period 2011 that cluster number one (C1) is a good representative of water quality alerts due to the dominance of individuals belonging to this category (blue), followed by two (C2) and three (C3), where the warnings of good water quality (green) in this study are not considered alerts of wa-

agua dudosa, lo que indica las buenas condiciones de alerta en la calidad del agua en las tres agrupaciones.

Los resultados obtenidos por el clasificador se analizan con más detalle, y se organizan de manera que éstos puedan develar el comportamiento de la comunidad taxonómica en diferentes tiempos y puntos de muestreo (Puente Alto, Puente Carretera y la ingesta El Diviso). La **Figura 8** explica dicho comportamiento.

En general, los indicadores de alta y de buena calidad de agua de taxones resaltan claramente comparados con taxones que representan agua de calidad cuestionable en los tres puntos de muestreo, mostrando de este modo alertas de calidad de agua relativamente buena en los tres sitios.

Del mismo modo, la **figura 8** es estudiada, la cual, muestra en primera instancia al Cluster o agrupación numero 1 (C1) como un fiel representante de alertas de calidad de agua debido al dominio sobre los individuos que pertenecen a su categoría (Azul), seguido del Cluster numero dos (C2) y el numero tres (C3) que representan alertas de buena calidad de agua (Verde). Otras categorías no hacen parte de la presente investigación debido a la falta de diversidad de taxones de este tipo en la base de datos y a la calidad dudosa del agua.

Además, para 2012 se aprecia el crecimiento en el indicador de especies en aguas de buena y dudosa calidad así mismo como una pequeña reducción en las especies representativas. Esto ultimo, genera alertas de incrementos de agentes contaminantes (comparados con 2011) en los tres puntos de muestreo, especialmente en las tomas de agua de Diviso y Puente Carretera en las cuales este fenómeno es mas evidente. El mismo comportamiento se presenta para los mismos grupos en el año anterior, donde C1 representa alertas de alta calidad de agua y tanto C2 como C3 representan a su vez alertas de buena calidad de agua.

De la misma manera, se detectó que la abundancia taxonómica para 2013 sigue disminuyendo, la cual se obtiene a partir del numero de especies indicadoras de agua de alta calidad, así mismo como la calidad del agua en general de las áreas de muestreo con la excepción del punto de Puente Alto, el cual parece seguir un constante comportamiento desde que la co-

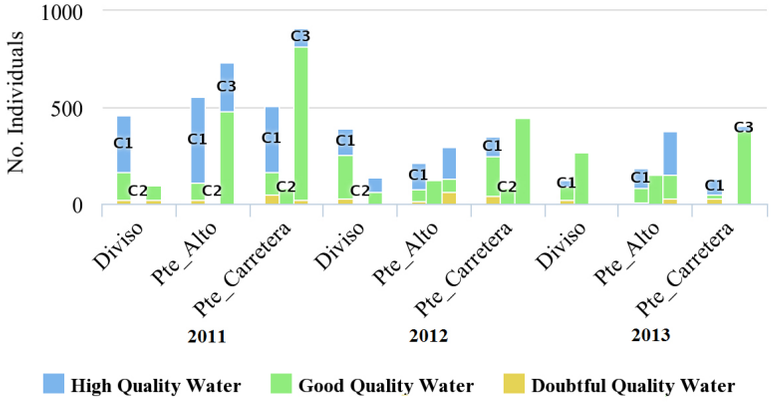

Figure 8. Taxonomic abundance - Piedras river 2011-2013 / Figura 8. Abundancia taxonómica 2011-2013

munidad pendiente de las especies indicadoras de alta y buena calidad reportan la misma proporción durante los tres años estudiados. Los grupos siguen el mismo patrón para el mismo período, sin embargo, los grupos y alertas C1, C2 y C3 solo representan buena calidad de agua.

Con base en los resultados obtenidos es posible asumir que el punto de muestreo de Puente Alto es el sistema que ha sufrido un menor grado de alteraciones por las actividades humanas en comparación con las áreas de muestreo de Puente Carretera y de la toma de agua de Diviso, teniendo en cuenta que posee mayor riqueza de individuos que representan alta calidad en todos los períodos de muestreo.

Se evidencia muestra una diversidad de 43% en los indicadores macroinvertebrados acuáticos de calidad de agua en los tres puntos de muestreo para 2011, sin embargo, se redujeron hasta 30% y 27% en los años 2012 y 2013, respectivamente. Por otro lado, la población total de taxones resistentes a la contaminación creció con el paso del tiempo de 53% en 2011 a 62% en 2012 y 67% en 2013, lo que indica una notable reducción en la calidad del agua.

IV. Conclusiones

El análisis manual de la calidad del agua a través de métodos tradicionales es muy engorroso, costoso y consume mucho tiempo cuando el conjunto de datos es demasiado amplio. Por esta razón, este proceso requiere herramientas especializadas que sean apropiadas para un análisis preciso y efectivo de la información, del mismo modo que las técnicas de aprendizaje automático utilizan el conocimiento existente para obtener las mismas conclusiones, mediante el uso de diferentes y menos complicadas formas de aprendizaje.

El objetivo de este estudio se basó en la generación de alertas mediante análisis grupales de calidad de agua en la cuenca del rio Piedras. Diferentes tipos de métodos de validación fueron revisados y abordados con el presente propósito. Después de analizar los resultados de todos los experimentos, se ha llegado a la conclusión de que el agua de las cuencas es de buena calidad en los tres puntos de muestreo analizados, a pesar de que se reduce con el paso del tiempo. La pérdida de la calidad del agua se debe al aumento de aproximadamente 14% en el número de organismos resistentes, el cual, se debe a los diferentes grados de contaminación y la reducción de 16% en el número de individuos indicadores de calidad de agua.

El punto de muestreo de Puente Alto es el sistema que ha sido alterado en un menor grado por las actividades humanas en comparación con Puente Carretera y la toma de agua de Diviso. Este punto tiene la mayor cantidad de individuos que representan alta calidad en todos los períodos de muestreo.

Por otro lado, el uso de la metodología con el algoritmo K-means y el árbol de decisión C.4.5 pueden generar alertas de calidad del agua de fácil interpretación para todos los usuarios. Hay un gran interés en el grupo para poder llevar a cabo este tipo de análisis con otras cuencas donde la abundancia taxonómica de organismos indicadores de baja calidad es significativa.

ter of dubious quality groups, due to the lack of diverse taxa of this type in the database.

Moreover, for 2012 the growth of indicator species of good water and dubious quality can be observed, with a small reduction of the representatives species of water quality, generating alerts for increasing pollutants (compared to 2011) in the three sampling points, especially in the Diviso intake and Puente Carretera, where this phenomenon is most clearly noted. The behavior of groups is the same as in the previous year, where C1 represents alerts of high quality water, while the C2 and C3 alerts still represent good water quality.

In the same way, from the taxonomic abundance for 2013 it is obtained that the number of indicator species of high quality continues to decline and with this the water quality of the sampling points, with the exception of the sampling point analyzed in Puente Alto, which follows an approximately constant behavior, since at this point the community of indicator species of high quality and good quality shows the same proportions during the three years analyzed. As for the groups, these follow the same pattern as in previous years, with the difference that the C1, C2 and C3 groups and alerts are representing only good water quality.

Based on the results found it can be assumed that the sampling point at Puente Alto is the system that has been altered the least by human activities compared with the sampling points at Puente Carretera and the Diviso intake, since this point has the greatest wealth of individuals representing high quality at all sampling periods.

The diversity of aquatic macro-invertebrate indicators of water quality in the three sampling points for 2011 was 43%. However, in 2012 and 2013 this community of macro-invertebrates was reduced to 30% and 27% respectively, by contrast with the population of contamination-resistant taxa, which increased with the passing of time. That is, the community rose from 53% of the total population in 2011 to 62% and 67% in 2012 and 2013 respectively, thus indicating that the waters are declining in quality over time.

IV. Conclusions

Manual analysis of water quality through traditional methods is very cumbersome, expensive and time-consuming when the data set is too large. For this reason, this process requires specialized tools that are appropriate for the accurate and effective analysis of informa-

tion, as well as machine-learning techniques that utilize existing knowledge to arrive at the same conclusions, by using different and less complicated methods.

The objective of this study is based on the generation of alerts by cluster analysis for water quality in the Piedras river basin. For this purpose, different types of validation methods were reviewed and addressed. After analyzing the results of all the experiments, we concluded that the basins' water is of good quality at the three analyzed sampling points. However, these basins are declining in water quality with the passage of time. The reduction in water quality is due to the increase by approximately 14% of the number of resistant organisms. This increase is caused by the different degrees of pollution and the reduction by 16% of individuals that are indicators of water quality.

The sampling point at Puente Alto is the system that has been least altered by human activities compared to Puente Carretera and the Diviso intake. This point has the greatest wealth of individuals representing high quality in all sampling periods.

On the other hand, the use of the methodology with the K-Means algorithm and C.4.5 decision tree can generate water quality alerts that are easy for all users to interpret. There is a strong interest in the group to be able to perform this type of analysis in other basins where the taxonomic abundance of poor quality indicator organisms is significant.

As a general recommendation, we suggest constant control and monitoring of activities around Piedras River. Although the water is of good quality at the basin, the reduction of water quality indicator organisms is an alarm, warning of the need to promote continuous monitoring, with the overall objective of preserving the sources of water supply for Popayan city in the State of Cauca.

Acknowledgments

The authors of this article would like to thank the Universidad del Cauca, the Environmental Studies Group (GEA), the Telematics Engineering Group (GIT), Colciencias Doctoral Scholarship (awarded to MSc. David Camilo Corrales) and the program AgroCloud of RICCLISA project, for technical-scientific support. ST

Como recomendación general, se aconseja el control y seguimiento constante de las actividades en torno al Río Piedras. Aunque el agua es de buena calidad en la cuenca, la reducción de organismos indicadores de calidad de agua es una alarma para promover un monitoreo continuo, todo con el objetivo de preservar las fuentes de abastecimiento de agua para la ciudad de Popayán, en el Estado del Cauca.

Agradecimientos

Los autores de este artículo quisieran dar las gracias a la Universidad del Cauca, el Grupo de Estudios Ambientales (GEA), el Grupo de Ingeniería Telemática (GIT), Beca Colciencias Doctorado (que fue otorgado al MSc. David Camilo Corrales) y el programa AgroCloud del proyecto RICCLISA, por el apoyo técnico-científico. ST

References / Referencias

Alba-Tercedor, J. (1996). Macroinvertebrados acuáticos y calidad de las aguas de los ríos. In*IV Simposio del agua en Andalucía (SIAGA). Almería*(vol. 2, pp. 203-213).

Arango, M. C., Álvarez, L. F., Arango, G. A., Torres, O. E., & Monsalve, A. D. J. (2008). Calidad del agua de las quebradas La Cristalina y La Risaralda, San Luis, Antioquia.*Revista EIA*, *9*, 121-141.

Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Pérez, J. M., & Perona, I. (2013). An extensive comparative study of cluster validity indices.*Pattern Recognition*,*46*(1), 243-256.

Bae, M. J., & Park, Y. S. (2014). Biological early warning system based on the responses of aquatic organisms to disturbances: a review.*Science of the Total Environment*,*466*, 635-649.

Bucak, I. O., & Karlik, B. (2011). Detection of drinking water quality using CMAC based artificial neural Networks.*Ekoloji*,*20*(78), 75-81.

Corrales, D. C., Corrales, J. C., & Figueroa-Casas, A. (2015). Towards detecting crop diseases and pest by supervised learning. *Ingeniería y Universidad*, *9*(1) 207-228. doi:10.11144/Javeriana.iyu19-1.tdcd

Gan, G., Ma, C., & Wu, J. (2007).*Data clustering: theory, algorithms, and applications*(vol. 20). Philadelphia, PA: Society for Industrial and Applied Mathematics.

González, D. P. (2010). *Algoritmos de agrupamiento basados en densidad y validación de clusters* (Thesis), Universitat Jaume I: Castellón, España.

Gurrutxaga, I., Muguerza, J., Arbelaitz, O., Pérez, J. M., & Martín, J. I. (2011). Towards a standard methodology to evaluate internal cluster validity indices. *Pattern Recognition Letters*,*32*(3), 505-515.

Lin, C. R., & Chen, M. S. (2005). Combining partitional and hierarchical algorithms for robust and efficient data clustering with cohesion self-merging. *Knowledge and Data Engineering, IEEE Transactions on*,*17*(2), 145-159.

Liu, S., Tai, H., Ding, Q., Li, D., Xu, L., & Wei, Y. (2013). A hybrid approach of support vector regression with genetic algorithm optimization for aquaculture water quality prediction.*Mathematical and Computer Modelling*,*58*(3), 458-465.

Madhulatha, T. S. (2012). An overview on clustering methods. *IOSR Journal of Engineering, 2*(4), 719-725.

de-Mantaras, R. L., & Saitia, L. (2004). Comparing conceptual, divisive and agglomerative clustering for learning taxonomies from text. In*ECAI 2004: 16th European Conference on Artificial Intelligence, August 22-27, 2004, Valencia, Spain: Including Prestigious Applications of Intelligent Systems (PAIS 2004): Proceedings*(vol. 110, p. 435). IOS Press. Addison.

Moreno, A. H. (2000). *La clasificación numérica y su aplicación en la ecología*. Santo Domingo, República Dominicana: Instituto Tecnológico de Santo Domingo.

Pang-Ning, T., Steinbach, M., & Kumar, V. (2006). *Introduction to data mining*. Boston, MA: Addison Wesley.

Park, Y. S., Chon, T. S., Kwak, I. S., & Lek, S. (2004). Hierarchical community classification and assessment of aquatic ecosystems using artificial neural networks.*Science of the Total Environment*,*327*(1), 105-122.

Pérez, G. R. (2003). *Bioindicación de la calidad del agua en colombia: Propuesta para el uso del método BMWP Col*. Medellín, Colombia Universidad de Antioquia.

Pino, W., García, D. M., Mosquera, M. L., Caicedo, K. P., Palacios, J. A., Castro, A. A., & Guerrero, J. E. (2011). Diversidad de macroinvertebrados y evaluación de la calidad del agua de la quebrada La Bendición, municipio de Quibdó (Chocó-Colombia).*Acta Biológica Colombiana*,*8*(2), 23-30.

Quiroz, R., Pla, F., Badia, J. M., Chover, M. (Eds.). (2007).*Métodos informáticos avanzados*. Castellón, España: Universitat Jaume I.

Rico, C., Paredes, M., & Fernandez, N. (2009). Modelación de la estructura jerárquica de macroinvertebrados bentónicos a través de redes neuronales artificiales.*Acta Biológica Colombiana*,*14*(3), 71-96.

Saraçli, S., Doğan, N., & Doğan, İ. (2013). Comparison of hierarchical cluster analysis methods by cophenetic correlation.*Journal of Inequalities and Applications*,*2013*(1), 1-8.

Sasirekha, K., & Baby, P. (2013). Agglomerative hierarchical clustering algorithm–A Review. *International Journal of Scientific and Research Publications, 3*(3) [on-line]. Retrieved from http://www.ijsrp.org/research-paper-0313/ijsrp-p1515.pdf

Singh, K. P., & Gupta, S. (2012). Artificial intelligence based modeling for predicting the disinfection by-products in water.*Chemometrics and Intelligent Laboratory Systems*,*114*, 122-131.

Theodoridis, S., Pikrakis, A., Koutroumbas, K., & Cavouras, D. (2010). *Introduction to Pattern Recognition: A Matlab Approach*. Punta Gorda, FL: Academic Press.

Velmurugan, T., & Santhanam, T. (2010). Computational complexity between K-means and K-medoids clustering algorithms for normal and uniform distributions of data points. *Journal of Computer Science, 6*(3), 363-368.

Viceministerio de Ambiente. (2010). *Política nacional para la gestión integral del recurso hídrico*. Bogotá Colombia: Ministerio de Ambiente, Vivienda y Desarrollo Territorial.

CURRICULUM VITAE

Edwin Ferney Castillo Currently an undergraduate student of the last semester in Electronics and Telecommunications Engineering at Universidad del Cauca, Colombia. His research interests focus on machine learning, data analysis and the area of Telecommunications and Telematics. / Estudiante de pregrado de último semestre en Electrónica e Ingeniería de Telecomunicaciones en la Universidad del Cauca, Colombia. Sus intereses de investigación se centran en el aprendizaje automático, el análisis de datos, las telecomunicaciones y la telemática.

Wilmer Fernando Gonzales Currently an undergraduate student of the last semester in Electronics and Telecommunications Engineering at Universidad del Cauca, Colombia. His research interests focus on machine learning, data analysis and the area of Telematics. / Estudiante de pregrado del último semestre de Electrónica e Ingeniería de Telecomunicaciones en la Universidad del Cauca, Colombia. Sus intereses de investigación se centran en el aprendizaje automático, el análisis de datos y la telemática.

David Camilo Corrales Received degrees in Informatics Engineering and Master in Telematics Engineering at Universidad del Cauca, Colombia, in 2011 and 2014 respectively. Currently he is a Ph.D. student in Telematics Engineering at the Universidad del Cauca and Science and Informatics Technologies at Universidad Carlos III de Madrid. His research interests focus on data mining, machine learning and data analysis. / Recibió los grados en Ingeniería Informática y Máster en Ingeniería Telemática de la Universidad del Cauca, Colombia, en 2011 y 2014 respectivamente. Actualmente es estudiante de doctorado en Ingeniería Telemática de la Universidad del Cauca y Ciencia y Tecnologías Informáticas en la Universidad Carlos III de Madrid. Sus intereses de investigación se centran en la minería de datos, el aprendizaje automático y el análisis de datos.

Iván Darío López Received the Engineering degree in Information Systems from Universidad del Cauca, Colombia, in 2011, and is an MSc. student in Telematics Engineering in the same institute. His current research interests are applications of computational intelligence techniques to modeling and data mining problems. / Recibió el título de Ingeniero en Sistemas de Información de la Universidad del Cauca, Colombia, en 2011, donde actualmente estudia la Maestría en Ingeniería Telemática. Sus intereses de investigación se centran en las técnicas de inteligencia computacional para problemas de modelado y minería de datos.

Miller Guzmán Hoyos Biologist at Universidad del Cauca, Colombia, and currently an MSc. student in Continental Hydrobiological Resources in the same institute. Currently also a researcher at the hydro-biological component of the Group for Environmental Studies at the Universidad del Cauca. His research interests focus on water quality based on benthic macro-invertebrates and water physical and chemical characteristics. / Biólogo de la Universidad del Cauca, Colombia, institución donde actualmente estudia la Maestría en Recursos Hidrobiológicos Continentales. Es investigador, en el componente hidrobiológico, del Grupo de Estudios Ambientales de la Universidad del Cauca. Sus intereses se centran en la investigación de la calidad del agua basada en macroinvertebrados bentónicos y las características químicas y físicas del agua.

Apolinar Figueroa Received a degree in biology from Universidad del Cauca, Colombia, in 1982, a master's degree in Ecology from Universidad de Barcelona, Spain, in 1986, and a Ph.D. in Biological Sciences from Universidad de Valencia, Spain, in 1999. Presently, he is full Professor and leads the Environmental Studies Group at Universidad del Cauca. His research interests focus on environmental impact assessment and biodiversity management. / Recibió el grado en biología de la Universidad del Cauca, Colombia, en 1982, una maestría en Ecología de la Universidad de Barcelona, España, en 1986, y el doctorado en Ciencias Biológicas de la Universidad de Valencia, España, en 1999. Es profesor titular y líder del Grupo de Estudios Ambientales de la Universidad del Cauca. Sus intereses de investigación se centran en la evaluación del impacto ambiental y la gestión de la biodiversidad.

Juan Carlos Corrales Engineer (1999) and Master in Telematics Engineering (2004) from the Universidad del Cauca, Colombia, and Ph.D. in sciences, specialty Computer Science, from the University of Versailles Saint-Quentin-en-Yvelines, France (2008). Presently, he is full time Professor and leads the Telematics Engineering Group at the Universidad del Cauca. His research interests focus on service composition and data analysis. / Ingeniero Telemático y Máster en Ingeniería Telemática de la Universidad del Cauca, Colombia (1999 y 2004 respectivamente); Doctor en Ciencias –énfasis en Ciencias de la Computación– de la Universidad de Versailles Saint-Quentin-en-Yve-lines, Francia (2008). Es profesor titular y líder del Grupo de Ingeniería Telemática de la Universidad del Cauca. Sus intereses de investigación se centran en la composición de servicios y el análisis de datos.