Water quality assessment with emphasis in parameter optimisation using pattern recognition methods and genetic algorithm

Gonzalo Sotomayor a,*, Henrietta Hampel b, a, Raúl F. Vázquez c, a

a Laboratorio de Ecología Acuática, Departamento de Recursos Hídricos y Ciencias Ambientales, Universidad de Cuenca, Av. 12 de Abril S/N, Cuenca, Ecuador
b Facultad de Ciencias Químicas, Universidad de Cuenca, Av. 12 de Abril S/N, Cuenca, Ecuador
c Facultad de Ingeniería, Universidad de Cuenca, Av. 12 de Abril S/N, Cuenca, Ecuador

A B S T R A C T

A non-supervised (k-means) and a supervised (k-Nearest Neighbour in combination with genetic algorithm optimisation, k-NN/GA) pattern recognition algorithms were applied for evaluating and interpreting a large complex matrix of water quality (WQ) data collected during five years (2008, 2010–2013) in the Paute river basin (southern Ecuador). 21 physical, chemical and microbiological parameters collected at 80 different WQ sampling stations were examined. At first, the k-means algorithm was carried out to identify classes of sampling stations regarding their associated WQ status by considering three internal validation indexes, i.e., Silhouette coefficient, Davies-Bouldin and Calinski-Harabasz. As a result, two WQ classes were identified, representing low (C1) and high (C2) pollution. The k-NN/GA algorithm was applied on the available data to construct a classification model with the two WQ classes, previously defined by the k-means algorithm, as the dependent variables and the 21 physical, chemical and microbiological parameters being the independent ones. This algorithm led to a significant reduction of the multidimensional space of independent variables to only nine, which are likely to explain most of the structure of the two identified WQ classes. These parameters are, namely, electric conductivity, faecal coliforms, dissolved oxygen, chlorides, total hardness, nitrate, total alkalinity, biochemical oxygen demand and turbidity. Further, the land use cover of the study basin revealed a very good agreement with the WQ spatial distribution suggested by the k-means algorithm, confirming the credibility of the main results of the used WQ data mining approach.

© 2017 Elsevier Ltd. All rights reserved.

1. Introduction

Present-day changes to fluvial systems include a great variety of direct and indirect anthropogenic activities, which in many cases result in a drastic deterioration of the water quality (Barnett et al., 2008; Harper et al., 2008). A variety of contaminants, in addition to a multitude of imprudent water management practices and destructive land uses, are currently threatening aquatic systems on a world-wide scale (Sanderson et al., 2002). Moreover, water of good quality is a crucial component for sustainable socio-economic development in any region of the world (Bartram and Ballance, 1996). In response to this, and owing to spatial-temporal variations in water chemistry, a monitoring program that provides a representative and reliable estimation of the water quality (WQ) is necessary (Simeonov et al., 2003). The large number of samples that need to be considered in appropriate WQ assessments and the number of constituents that must be considered per sample gives rise to data sets of enormous size and complexity.

Furthermore, the relationships investigated in these data sets usually cannot be expressed in quantitative terms and they are better expressed in terms of similarity or dissimilarity among groups of multivariate data (Lavine and Rayens, 2009). Herein, the application of different supervised and non-supervised pattern recognition techniques has the potential of facilitating the interpretation of complex data matrices and, as such, of understanding the WQ status of the studied systems without losing important information; moreover, they constitute valuable tools for reliable management of water resources (Shrestha and Kazama, 2007; Juahir et al., 2010; Bücker et al., 2010).

In Ecuador, governmental efforts have been carried out in the past to establish various WQ monitoring programs in some
particular locations of the country and, more recently, to a national level (SENAAGUA, 2016). However, the resulting observed data are not really public available, and as such the data can hardly be used for research or management purposes. Nevertheless, the Ecuadorian National Secretary of Water (SENAAGUA) - Santiago River Hydrographic Demarcation (DHS) generated, and made available to the current study, an extensive database derived from 80 monitoring stations located in the Paute river basin, which is one of the most important hydrographic systems of southern Ecuador.

As such, in the present study, a large and complex database (including data on 14427 observations), collected throughout a 5-year monitoring program (year 2008 and period 2010–2013), was used in the context of supervised and non-supervised pattern recognition techniques that were applied with the aim of handling the complexity of the database and mining relevant information about: (i) the WQ similarities or dissimilarities between the 80 monitoring stations based on the measured river WQ parameters; (ii) the most significant WQ related parameters that explain the WQ classes identified through the previous objective; and (iii) the correspondence among the identified WQ classes, the most significant WQ related parameters and the land cover of the study river basin.

2. Study area and methods

2.1. Study area and water quality (WQ) monitoring sites

The study area, the Paute river basin, is located in the south of Ecuador (Fig. 1). It has a surface of about 6442 km² and its main reach length is approximately 120.4 Km. This is one of the most important hydrographic systems of southern Ecuador (Da Ros, 1995) and discharges into the Upano river, which belongs to the Amazon river system. Two important cities are located inside the river basin, respectively Cuenca and Azogues with approximately 500000 and 33850 inhabitants. The main pollutant loads, from both point and non-point sources, include domestic wastewaters, agricultural runoff, animal husbandry and industrial effluents (Da Ros, 1995).

The geological features of the basin are very complex, primarily due to the subduction of the Nazca plate towards the South American plate, which causes that this basin is still in the process of rising, sharpening its slope, resulting in large amounts of suspended sediments that are transported by the river to the Amazon basin (Astudillo et al., 2010). The elevation ranges approximately between 500 and 4250 m above the sea level (a.s.l.), with the majority (61.3%) of the basin in the elevation range 2550–3575 m, 4.3% in the range 500–1525 m, 13.7% in the elevation band 1525–2550 m, and 20.7% of the basin is situated above 3575 m. In average terms, slopes vary between 25% and 50%; in the upper part of the basin (west) a mountainous relief is dominant whilst a gentler relief is representative in the central and lower parts (east).

Multi-year average temperature varies between 4.4 °C and 18.6 °C. The lower temperatures correspond to the western Andes range with an average of 6 °C (Páramo), while the warmest areas are situated in the valleys and subtropical zones (Amazonia regions), with an average temperature fluctuating between 22 °C and 26 °C. Due to the wide elevation range, rainfall oscillates in intensity and duration, with maximum annual averages between 2500 mm and 3000 mm at higher elevations and minimum annual averages between 600 mm and 800 mm in the valleys.

The sampling design was planned to cover a wide range of parameters (21 in total) at key monitoring sites (80 in total: Fig. 1), aiming at representing as accurately as possible the WQ distribution in the study basin (SENAAGUA, 2016).

2.2. Sampled WQ parameters

The studied 21 WQ parameters include: aluminium (Al), ammonia (N-NH₄), 5-day biochemical oxygen demand (BOD₅), chloride (Cl), dissolved oxygen (DO), electric conductivity (EC), faecal coliforms (FC), fluoride (F), iron (Fe), nickel (Ni), nitrate (N-NO₃), pH, phosphates (P-Po₄), potassium (K), sodium (Na), total alkalinity (TALK), total hardness (TH), total phosphorus (P-tot), total solids (TS), turbidity (TU) and water temperature (WT). On average, the monitoring stations were visited five times per year, except in 2008, when they were only sampled three times. Some stations were sampled more often, because they were located either in highly polluted sites or, on the contrary, in unaltered environmental (i.e., reference) locations. As a result of the applied monitoring protocol, a large and complex WQ database was established for the nrep = 80 monitoring stations, upon nrep = 687 sampling replicates and the surveying of the nobs = 21 WQ parameters per replicate and monitoring station, resulting in a total of nobs = nrep X nrep = 14427 observations, that are represented by xij, with i = 1, 2, ..., nobs and j = 1, 2, ..., nrep (Fig. 2). Unfortunately, despite the fact that this is a very interesting and rich data set, there was never a scientifically driven monitoring plan to collect it but, rather, its structure responds to political/personal initiatives that varied throughout time. Table 1 lists the basic statistics for each of the studied WQ parameters.

2.3. Processing land coverage data for assessing the spatial congruency of WQ classification

The correspondence of the spatial distribution of the WQ classes was assessed by comparing it with the land cover distribution of the Paute river basin, using as well auxiliary topographical...
information. Land cover raster data was available for the year 2001, covering the whole extent of the study catchment. This is the most recent dataset that is available publicly. Geographical Information Systems (GIS) algorithms were applied on the original land cover data so that it was reclassified to a more convenient form, which enabled a direct association between land cover and the spatial distribution of the WQ classes.

The considered land cover classes were: (i) altered vegetation; (ii) woody native vegetation; (iii) without cover/urbanised; and (iv) Páramo (unaltered). Additionally, a digital elevation (raster) model of the whole catchment was available with a resolution of 50 × 50 m². The referred congruency assessment was based on the visual inspection of the geographical distribution of the WQ classes, the land cover and the topography. ArcGIS® was used for all GIS analyses, including the latter visual inspection.

2.4. Data processing and multivariate statistical assessment

Fig. 2 depicts the multivariate statistical protocol applied in this study. All the statistical analyses involved in it were implemented with MATLAB® (Hanselman and Littlefield, 2012) version 2014. As a first step, the Shapiro-Wilk test (Shapiro and Wilk, 1965) was used to evaluate whether the distributions of the WQ parameters were normal. This test showed with a 95% confidence level that none of the WQ parameters are normally distributed (p probability value = 0.0001 lower than the significance level α = 0.05), exhibiting in all cases a positive skew (Table 1).

Therefore, it was decided to continue with the statistical assessment by using a central tendency value (i.e. aggregated value) of the distribution of every WQ parameter, for a given monitoring station, instead of using the whole set of available observations. Further, this aggregated parameter was standardised through range scaling so that a more normal associated distribution could be obtained; this kind of standardisation procedure is recommended in a cluster analysis, particularly for the k-means method (Steinley, 2004).

Herein, the applied aggregation procedure reduced the dimension of the analysis to \( n_{med} = n_x \times n_{med} = 1680 \) median observations \( \left(x_{im}\right) \) with \( i = 1, 2, \ldots, n_x \) and \( m = 1, 2, \ldots, n_{med} \). Given the significant skewness of the WQ parameters distributions, the use of median

Table 1

<table>
<thead>
<tr>
<th>WQ Parameter</th>
<th>Mean</th>
<th>Median</th>
<th>STD</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>TALK (mg L⁻¹)</td>
<td>0.62</td>
<td>0.06</td>
<td>1.50</td>
<td>0.00–23.00</td>
</tr>
<tr>
<td>Al (mg L⁻¹)</td>
<td>0.06</td>
<td>0.00</td>
<td>0.22</td>
<td>0.00–1.59</td>
</tr>
<tr>
<td>NH₄ (mg L⁻¹)</td>
<td>0.50</td>
<td>0.00</td>
<td>1.25</td>
<td>0.00–11.35</td>
</tr>
<tr>
<td>BOD₅ (mg L⁻¹)</td>
<td>7.2</td>
<td>2.3</td>
<td>11.9</td>
<td>0.0–141.0</td>
</tr>
<tr>
<td>CL (mg L⁻¹)</td>
<td>4.26</td>
<td>0.82</td>
<td>16.63</td>
<td>0.00–263.06</td>
</tr>
<tr>
<td>DO (mg L⁻¹)</td>
<td>6.77</td>
<td>6.79</td>
<td>0.80</td>
<td>3.30–9.75</td>
</tr>
<tr>
<td>EC (μS cm⁻¹)</td>
<td>130.3</td>
<td>76.4</td>
<td>178.0</td>
<td>0.16–1810.00</td>
</tr>
<tr>
<td>FC (bacteria 100⁻¹ ml⁻¹)</td>
<td>5675.2</td>
<td>1700.0</td>
<td>6668.5</td>
<td>0.0–16000.0</td>
</tr>
<tr>
<td>FL (mg L⁻¹)</td>
<td>1.03</td>
<td>0.40</td>
<td>4.51</td>
<td>0.00–54.6</td>
</tr>
<tr>
<td>TH (mg L⁻¹)</td>
<td>47.8</td>
<td>36.8</td>
<td>55.8</td>
<td>0.00–657.0</td>
</tr>
<tr>
<td>Fe (mg L⁻¹)</td>
<td>0.63</td>
<td>0.00</td>
<td>2.28</td>
<td>0.00–31.00</td>
</tr>
<tr>
<td>Ni (mg L⁻¹)</td>
<td>0.01</td>
<td>0.00</td>
<td>0.09</td>
<td>0.00–1.51</td>
</tr>
<tr>
<td>N-N0₃ (mg L⁻¹)</td>
<td>0.57</td>
<td>0.06</td>
<td>0.06</td>
<td>0.00–23.47</td>
</tr>
<tr>
<td>pH</td>
<td>7.53</td>
<td>7.58</td>
<td>0.68</td>
<td>4.17–9.43</td>
</tr>
<tr>
<td>P-P0₄ (mg L⁻¹)</td>
<td>1.09</td>
<td>0.41</td>
<td>1.63</td>
<td>0.00–7.72</td>
</tr>
<tr>
<td>P-tot (mg L⁻¹)</td>
<td>0.58</td>
<td>0.19</td>
<td>0.92</td>
<td>0.00–4.76</td>
</tr>
<tr>
<td>K (mg L⁻¹)</td>
<td>2.64</td>
<td>0.47</td>
<td>13.06</td>
<td>0.00–227.06</td>
</tr>
<tr>
<td>Na (mg L⁻¹)</td>
<td>6.57</td>
<td>3.48</td>
<td>12.38</td>
<td>0.00–112.89</td>
</tr>
<tr>
<td>TS (mg L⁻¹)</td>
<td>12.47</td>
<td>0.01</td>
<td>76.36</td>
<td>0.00–1160.00</td>
</tr>
<tr>
<td>TU (NTU)</td>
<td>23.3</td>
<td>2.9</td>
<td>80.9</td>
<td>0.00–1190.5</td>
</tr>
<tr>
<td>WT (°C)</td>
<td>15.2</td>
<td>14.6</td>
<td>3.5</td>
<td>8.1–23.8</td>
</tr>
</tbody>
</table>

Legend: STD = standard deviation; TALK = total alkalinity; Al = Aluminium; NH₄ = Ammonia; BOD₅ = 5-day biochemical oxygen demand; CL = chlorides; DO = dissolved oxygen; EC = electric conductivity; FC = faecal coliforms; Fe = iron; Ni = nickel; N-N0₃ = nitrate; P-P0₄ = phosphates; K = potassium; Na = sodium; TH = total hardness; P-tot = total phosphorus; TS = total solids; TU = turbidity; WT = water temperature.

Fig. 2. Flowchart of the modelling protocol that was implemented in the current study and that combines a non-supervised (k-means) and a supervised (k-NN/GA) pattern recognition methods.
rather than of mean values was preferred since the median is expected to be a better central tendency measure (Anderson and Finn, 1996; Helsel and Hirsch, 2002). Standardisation of \( x_{i,m} \) was achieved by means of:

\[
Z_{i,m} = \frac{X_{i,m} - L_m}{U_m - L_m} \quad \text{for } i = 1, 2, \ldots, n_v \text{ and } m = 1, 2, \ldots, n_{sp}
\]

where \( L_m \) and \( U_m \) are the minimum and maximum limits of the parameter range, so that \( z_{i,m} \) varies between 0 and 1 (Frank and Todeschini, 1994)

Then, the aggregated dataset of \( n_{med} = 1680 \) \( z_{i,m} \) values was used in the context of the k-means algorithm (MacQueen, 1967), a non-hierarchical cluster analysis (CA) and a non-supervised pattern recognition method, were carried out with the intention of defining \( k \) groups of monitoring stations with common WQ characteristics. As a prior process, validation indices (Wang et al., 2009) were calculated to determine the number of WQ clusters, \( k \), for the application of the k-means method.

Once the \( k \) WQ groups of monitoring stations were defined through the k-means process (Fig. 2), for a given monitoring station, the respective identifier (or “label”) of the corresponding WQ (k-means) group (or class) was assigned to the original observations of the station. After following this procedure for all of the sampling stations, on the resulting new data matrix of \( n_{obs} \) observations \( \{x_{i,j}\} \), the k-Nearest Neighbour (k-NN), a supervised pattern recognition method, in combination with a genetic algorithm (GA) for optimisation, was applied.

It is worth noticing that the WQ related parameters frequently are of different nature or have different units. Many pattern recognition techniques, like the ones used in this study, are very sensitive to these issues (Todeschini et al., 2015). Therefore, range scaling (Eq. (1)) was used to eliminate this dependence on the referred issues, regardless of whether the applied recognition techniques were parametric or non-parametric.

### 2.4.1. The non-supervised k-means agglomerative method

The goal of CA is to determine the intrinsic grouping in a set of unlabelled (i.e. unclassified) data, upon the similar characteristics that their members possess (Frank and Todeschini, 1994). In this context non-hierarchical clustering techniques, such as k-means, have been widely used in different applications, among them the investigation of the hydro-chemical processes in-stream water quality (Güler et al., 2002; Caccia and Boyer, 2005). Cluster membership is determined by calculating the centroid for each group and assigning each object to the group with the closest centroid, which minimises the overall within-cluster dispersion by iterative reallocation of cluster members (Hartigan and Wong, 1979). In this study, this method was applied using the Euclidean distance (expressed in the units of measure of \( v \)) as the measure of similarity between objects (Chen et al., 2002):

\[
D_{ij} = \sqrt{\sum_{v=1}^{n_v} (z_{i,v} - z_{j,v})^2}
\]

where \( z_{i,v} \) and \( z_{j,v} \) are the values of the normalised parameter \( v \) for object \( i \) and \( j \), respectively, where \( i \) is different from \( j \) and both are \( \leq n_{sp} \).

Further, the evaluation of clustering solutions is fundamental in CA. Validity indices, both external and internal, are used for this purpose (Wang et al., 2009). An external index measures the agreement between a priori known clustering structure and the result from a current clustering procedure (Dudoit and Fridlyand, 2002), whilst an internal index measures the appropriateness (or goodness) of a clustering partition without external information, using quantities and features inherent in the data (Thalamuthu et al., 2006). Internal validity indices were applied in the current analysis, namely, the Silhouette Coefficient (SC), the Davies-Bouldin (DB) index and the Calinski-Harabasz (CH) index. In this context, the Cluster Validity Analysis Platform (CVAP) was used for this purpose (Wang et al., 2009). These were adopted herein since they are widely used for \( k \) estimation and clustering quality evaluation (Wang et al., 2009).

SC (Kaufman and Rousseeuw, 1990) is a dimensionless measure that evaluates the quality of compactness and separation of clusters; with an upper bound equal to 1, the optimum \( k \) value corresponds to its largest average (Chen et al., 2002). DB (Davies and Bouldin, 1979) is a function of the ratio of the sum of within-cluster scatter to between-cluster separation; as such, the objective is to minimise it, which implies minimising the within-cluster scatter and maximising the between-cluster separation (Ray and Turi, 1999). CH (Calinski and Harabasz, 1974), or variance ratio criterion, is a measure of the between-cluster isolation and the within-cluster coherence; the objective is to maximise it (Kryszczuk and Hurley, 2010).

Each of the \( n_{sp} = 80 \) monitoring stations was assigned to either of the identified k WQ-clusters, as the result of the application of the k-means methodology. For a given WQ monitoring station, the identifier (ID) of the associated WQ-cluster was designated to the class the newly assigned class matches the previously assigned class of the sample point (de ned previously in this study through the k-means) of most of the NN-in a sample point, the sample point is assigned that class; when the newly assigned class matches the previously assigned class of the sample point (defined previously in this study through the k-means) of most of the NN in a sample point, the sample point is assigned that class; when the newly assigned class matches the previously assigned class of the sample point (defined by the k-means method), the k-NN algorithm is considered successful (Lavine and Rayens, 2009). Thus, the k-NN classification has two steps, namely, (i) the definition of the neighbouring points (NN-points) of the sample point; and (ii) the subsequent determination of the class of the sample point, using the information provided by those NN-points.
The k-NN algorithm performance was measured through the non-error rate (NER), which is defined as the percentage of monitoring stations correctly assigned by the k-NN algorithm to the classes determined previously (Fig. 2) by the k-means method (classification accuracy criteria; Hand, 2012). In this study, NER was estimated using a testing data set (i.e., 40% of total data). Sample points were classified according to the WQ parameters (physico-chemical and microbiological) that are known to be indirectly related to WQ, usually through some undetermined mathematical relationships. A training data set (i.e., 60% of the total data) for which the property of interest (WQ) and the indirectly related descriptors (in the current case the physicochemical and microbiological parameters measured at every WQ monitoring station) are known was used for developing a classification rule that was in turn utilised to predict the WQ of data samples that were not part of the training set (Lavine and Rayens, 2009).

Besides the fact that, in general, the more the number of measured parameters the more a classification algorithm gets confused, for the particular case of the k-NN algorithm, there are three main limitations, namely (Frank and Todeschini, 1994; Suguna and Thanushkodi, 2010): (i) calculation complexity due to the fact that all of the similarities between the training sample points need to be computed for each chromosome; (ii) the performance of the classification algorithm is over-dependent on the training set; and (iii) there is no weight difference between the training sample points, despite the implicit differences in terms of available data for every one of those points.

Thus, the combination of the k-NN algorithm with a GA was applied to overcome these limitations by: (i) generating a profile about the internal data structure of every class; and (ii) getting the contribution of individual related parameters (measured at every WQ monitoring station) to the classification of the main property of interest (in the current case, WQ). In this context, GAs demonstrated to be robust searching techniques that in most cases outperform traditional optimisation methods in water resources applications (Mulligan and Brown, 1998; Ng and Perera, 2003; Liu et al., 2007), through the emulation of evolution by natural selection and genetic inheritance, so that a population of competing solutions evolve over time to converge to a single optimal one (Holland, 1975).

Thus, every model parameter (every WQ related parameter, in this study) is a gene, while a complete set of genes (21 parameters) is a chromosome. Every gene adopts the value of the respective WQ related parameter encoded as a variable-length binary number. The length of this binary number depends on the magnitude of the WQ parameter value. For example, if a gene represents pH (4.17 < pH < 9.43) the length of the respective binary number will be shorter than the corresponding length when a given gene represents, say, faecal coliforms (varying between 0 and 16000), owing to the magnitude difference of the values adopted by pH and faecal coliforms. The advantage of using this binary system is that very small value variations of a given WQ parameter can be taken into account in the analysis.

The optimisation process is carried out by using three biological operators (Sivanandam and Deepa, 2008; Liu et al., 2007), namely: (i) selection; (ii) cross-over (reproduction); and (iii) mutation. Selection makes sure that only the best chromosomes (solutions) could cross-over or mutate. During successive iterations (generations), initial chromosomes evolves into stronger ones by reproduction among members of the previous generation. Each GA run consists of several generations with constant population size n chromosomes. The strength of each chromosome is evaluated by means of a fitness function. In the current study, a pattern recognition function through a classification model, NER, was used to characterise the performance of the k-NN/GA method regarding the prior classification produced by the unsupervised k-means method. So, the highest NER values were obtained for a cross-over probability of 60%, a mutation probability of 3% and a constant population of 21 chromosomes that were evolved over 100 generations. Thus, subsequent generations are formed by combining (stronger) chromosomes, with associated higher fitness values, from the previous (or parent) population. Further, a random change of the value of a gene to form a new chromosome is achieved through mutation, on a bit-by-bit basis (from 0 to 1 or in reverse order).

The combination of the k-NN and GA that was used in the current study follows to a great extent the approach presented by Suguna and Thanushkodi (2010); alternatives are for instance Chang and Lippmann (1990) and Raymer et al. (2000). Hereafter, the approach involved the following steps: (i) data processing in general, including the encoding of the genes that form the chromosomes; (ii) selecting the distance (or similarity) measure among the test stations whose category are being predicted and the respective neighbouring training stations; (iii) choosing the k-number of samples (or objects) from the training set to generate the initial population on which genetic operators are applied during optimisation; (iv) calculating the distance measures among testing and training samples through a goodness of fit (or similarity) index; (v) choosing the chromosome with the highest goodness of fit value and storing it as the global maximum; and (vi) refining the global maximum. The latter step is performed through an iterative procedure that is based on the application of genetic operators, namely, reproduction, crossing-over and mutation to get a new population upon which a newer evaluation of the goodness of fit measure takes place to define a local maximum; only if the newer goodness of fit measure (local maximum) is higher than the one associated to the older global maximum, the chromosome that corresponds to the newer local maximum is the newer global maximum. The iterative process is over once the total number of generations (100 in this study) are taken into account. The chromosome that corresponds to the global maximum has the optimum k-neighbours and the respective classification results.

Fig. 3. Spatial distribution in the Paute river basin of (i) the soil cover (year 2001); (ii) the 18 sub-basins; and (iii) the water quality (WQ) monitoring stations, classified according to the k-NN/GA algorithm as having associated low (C1) or high (C2) pollution. Sub-basins are: 1 – Sidcay, 2 – Collay, 3 – Cuenca, 4 – Jadan, 5 – Paute, 6 – Machangara, 7 – Magdalena, 8 – Mazar, 9 – Juval, 10 – Pindililig, 11 – Pulpito, 12 – Sta. Bárbara, 13 – Burgay, 14 – Tarqui, 15 – Tomebamba, 16 – Yanuncay, 17 – Paute bajo, and 18 – Negro.
3. Results

3.1. Land cover processing

Based on the adopted classes, the land cover percentages within the study basin are respectively altered vegetation (37.12%), woody native vegetation (34.38%), without cover/urbanised (3.54%), and Páramo (unaltered vegetation) (24.96%). Fig. 3 illustrates, in addition to the distribution of land cover, the WQ class, C1 or C2, of each monitoring station and the bounds of the 18 sub-basins in which the Paute river basin was sub-divided. The altered vegetation is present particularly in the mid corridor of the basin; both, the higher western and lower eastern sides of the basin still have a significant presence of woody native vegetation and Páramo. Woody native vegetation is present for instance in the sub-basins Negro (77.8%), Juval (86.4%), Pulpito (77.3%) and Machángara (76.4%). Páramo (unaltered) is present in the sub-basins Yanuncay (60.1%), Pulpito (59.5%), Machángara (56.7%) and Tomebamba (48.3%). Further, the Paute basin includes important extents of protected areas (Fig. 1). The most renowned are the Cajás National Park (PNP, located in the western, higher, extreme of the basin; Mosquera et al., 2017), which is a Ramsar-Convention (RAMSAR) wetland site, and the Sangay National Park (PNS, located at the north-east extreme); both recognised by the United Nations Educational, Scientific and Cultural Organization (UNESCO) as World Heritage Sites. Nearly 41% of the Paute basin area is subjected to management and conservation programs.

3.2. k-means algorithm: spatial similarity and monitoring stations grouping

The maximum SC value ($SC_{\text{max}}$), the minimum DB value ($DB_{\text{min}}$), and the maximum CH value ($CH_{\text{max}}$) were all obtained for $k = 2$ (Table 2), that is, implying that there are two statistically significant clusters. As for the results of the k-means method, these two groups of sites in the Paute river basin coincide with prior WQ studies in the study region (Da Ros, 1995; Pauta-Calle and Chang-Gómez, 2014); cluster 1 (C1) corresponds to relatively less polluted sites (67 monitoring stations) while cluster 2 (C2) matches highly polluted sites (13 monitoring stations).

Table 2

<table>
<thead>
<tr>
<th>Internal Indexes</th>
<th>Inspected number of clusters k</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2</td>
</tr>
<tr>
<td>SC</td>
<td>0.60</td>
</tr>
<tr>
<td>DB</td>
<td>1.31</td>
</tr>
<tr>
<td>CH</td>
<td>28.22</td>
</tr>
</tbody>
</table>

Legend: SC – Silhouette coefficient; DB – Davies-Bouldin; CH – Caliński-Harabasz.

3.3. k-NN/GA: pattern recognition performance and significant WQ parameters

In this analysis, for a given monitoring station, the respective (k-means) WQ class was the (dependent) grouping variable, whilst the corresponding measured parameters constituted the independent variables. A good classification accuracy was obtained with the k-NN/GA algorithm, since the average NER for 100 runs (Fig. 4a) was about 0.87 ($\pm$STD = $\pm$0.006), implying that about 87% of the WQ monitoring stations were correctly assigned with respect to the previously, k-means derived WQ classes.

With the objective of getting the most important independent variables (i.e. measured parameters) that explain the structure of the two WQ classes, the final stepwise selection with the best NER indicated that nine parameters are the optimum to interpret the two WQ classes (Fig. 4b), namely, EC, FC, DO, TH, CL, N–NO3, BOD5, TALK and TU, that account for most of the expected spatial variations of the WQ in the Paute river basin (Fig. 4c). Further, the frequency of the k-NN/GA variable selection throughout the classification process (Fig. 4c), indicates that the EC parameter is the most significant by the classification model, whilst the remaining 8 WQ parameters seem to have a similar (lower) significance.

In the following, the sub-index C1 indicates that a given parameter or characteristic is associated to the WQ class C1; whilst the sub-index C2 refers to WQ class C2. Fig. 5 shows the Box plots of the 9 significant WQ parameters as a function of the two WQ classes C1 (less polluted) and C2 (highly polluted). In general, the values adopted by EC (Fig. 5a), FC (Fig. 5b), TH (Fig. 5d), CL (Fig. 5e), N–NO3 (Fig. 5f) and TU (Fig. 5i) are higher for the sampling points belonging to C2 than for the ones belonging to C1 (average values are $EC_{C1} = 87.7 \mu S \text{ cm}^{-1}$, $EC_{C2} = 280.3 \mu S \text{ cm}^{-1}$; $FC_{C1} = 3936.7$ bacteria $100^{-1}$ $\text{ml}^{-1}$, $FC_{C2} = 117943.7$ bacteria $100^{-1}$ $\text{ml}^{-1}$; $TH_{C1} = 37.3 \text{ mg L}^{-1}$, $TH_{C2} = 84.5 \text{ mg L}^{-1}$; $CL_{C1} = 1.7 \text{ mg L}^{-1}$, $CL_{C2} = 13.2 \text{ mg L}^{-1}$; $N–NO3_{C1} = 0.3 \text{ mg L}^{-1}$, $N–NO3_{C2} = 1.5 \text{ mg L}^{-1}$; and $TU_{C1} = 14.7 \text{ mg L}^{-1}$, $TU_{C2} = 53.5 \text{ mg L}^{-1}$). For the DO (Fig. 5c) the inverse condition was observed, that is, lower waters for C2 (average $6.32 \text{ mg L}^{-1}$) than for the waters in C1 (average $6.90 \text{ mg L}^{-1}$). For TALK (Fig. 5g) and BOD5 (Fig. 5h) no considerable differences between waters belonging to C1 and the ones belonging to C2 were observed (average TALK$_{C1}$ = 0.62 mg L$^{-1}$, TALK$_{C2}$ = 0.61 mg L$^{-1}$ and BOD5$_{C1}$ = 6.43 mg L$^{-1}$, BOD5$_{C2}$ = 9.92 mg L$^{-1}$); however, the dispersion is slightly lower for the peak values of waters in C2 than for the waters in C1, similarly to what is observed for the DO. It is important to observe that most of the peak values (Fig. 5) were observed at the Burgay sub-basin (Fig. 3); thus, they do not seem to be outliers, but they are simply characterising the extreme WQ conditions taking place within such sub-basin.

4. Discussion

Because the pre-selection of the number of clusters (k) at the start of the k-means method strongly influences its outcome (Güler et al., 2002), the reliability of the two clusters was satisfactorily tested through the internal validation indices SC, DB and CH. In the current case a SC value of 0.6 corresponds to a reasonable structure has been found” category (Kaufman and Rousseeuw, 1990). DB and CH are not valued with a categorisation likewise SC; however, the fact that both suggest two WQ classes emphasises the similarity with the results of the k-means algorithm that partitioned the monitoring stations into two homogeneous groups.

Further, the current application of the k-means algorithm was based on a non-conventional approximation. Namely, due to the significant skewness present in the original WQ parameters, a
simplified data matrix of the median values of the WQ parameters, that were previously standardised, was used (Ouyang, 2005). Standardisation of the WQ parameters was achieved by considering their ranges of variation rather than by using the z-scale procedure, as conventionally done in most WQ studies (Shrestha and Kazama, 2007; Kannel et al., 2007; Singh et al., 2004). In this context, Steinley (2004) concludes that standardisation of the studied parameters, by considering their range of variation, was the most effective approach while using the k-means method.

In classification models, parametric methods, such as the linear discriminant analysis, widely used in surface water resources (Hajigholizadeh and Melesse, 2017; Gholizadeh et al., 2016; Kovács et al., 2014), require a priori knowledge of the probability density functions of the classes. However, in most real-world applications these statistical properties are very hard to be known in advance (Lavine and Rayens, 2009) and/or classes substantially depart from normality (Hirsch and Alexander, 1991). Consequently, parametric methods are no longer adequate (Huberty and Olejnik, 2006) and non-parametric methods, such as the k-NN algorithm, are a more appropriate alternative (Hirsch et al., 1982; Towler et al., 2009), despite of which the k-NN approach is normally not chosen as a primary classification method in surface water studies.

The k-NN/GA method gave relevant results in selecting appropriate explanatory variables representative of the WQ status in the study basin. Hence, the method yielded an important data reduction as it identified only nine (EC, FC, DO, TH, CL, N-NO3, BOD5, TALK and TU) out of the 21 WQ measured parameters, that are explaining about 86% of the cluster assignments in supervised pattern recognition, reflecting most of the WQ spatial variation in the study basin.

Hence, the average EC2 (280.3 μS cm⁻¹) was almost three times higher than the respective value for EC1 (87.7 μS cm⁻¹). Because EC increases nearly linearly with increasing ion concentration, it implies that inorganic inputs are high in waters belonging to C2. Most of C2 associated monitoring stations belong to the Burgay sub-basin (Fig. 3), where wastewater effluents often contain high amounts of dissolved salts from inorganic pollution sources such as domestic sewage, municipal storm water drainage and industrial effluent discharges (Da Ros, 1995; Pauta-Calle and Chang-Gómez, 2014). Further, other variables related to inorganic pollution sources such as TH, CL, TU, N-NO3 and TALK have similar evolution to the one of EC. In this regard, Olajire and Imeokparia (2001), Daniel et al. (2002) and Begum et al. (2009) observed in similar studies that in water with high EC values nitrate and chloride ions and calcium or magnesium salts are predominant.

The high TH values associated to C2 might be linked to the use of inorganic fertilisers (WHO, 1998). Hereafter, high concentrations of the nitrate ions (inorganic nutrient) associated to C2 might be attributed to the runoff of fertiliser residues from agricultural lands, mainly urea (CH4N2O). Further, FC, DO, and BOD5 normally reflect contamination by microbial communities and human activities (Liou et al., 2004). Hereafter, the trends observed for these variables suggest an important organic pollution from anthropogenic source associated to the WQ monitoring sites belonging to C2, resulting in high oxygen-demand consumption, characterised by average BOD5 values higher in C2 (9.92 mg L⁻¹) than in C1 (6.43 mg L⁻¹).

A similar trend was found in other studies such as Vega et al. (1998), Singh et al. (2004) and Kannel et al. (2007), which reported that high levels of dissolved organic matter (DOM, i.e., carbohydrates, proteins, lipids, etc.) consume large amounts of oxygen, which depletes the amount of available DO. The latter is in line with the results of the current study. However, on the contrary to what is addressed in those studies, in our study basin, the DO did not drop to a minimum level and, as such did not induce anaerobic fermentation leading to development of ammonia and organic acids. This is emphasised by the evolution of the pH distributions associated to the C1 and C2 monitoring sites, which shows that the acidity of the C2-associated waters is not systematically higher than
the one recorded in the C1-associated waters, except for very few pH data near the lower, i.e. acidic, extreme of the pH_{C2} range of variation (pH_{C1} range = 5.3–9.43, pH_{C2} range = 4.17–8.95).

Several authors emphasised the importance of finding land use/cover attributes that explain the spatial variation of WQ (Bücker et al., 2010; Bu et al., 2014; Cunha et al., 2016). Examination of the spatial distribution of both, the C1 and C2 stations, as well as the land cover, suggest that most of the WQC2 monitoring stations (Fig. 3) are not directly exposed to, nor have a significant influence from, native forest and unaltered vegetation cover, which is particularly the case in the Burgay sub-basin, where most of the WQC2 monitoring stations are located and where 55.3% of the surface has altered native forest and vegetation (resulting from the introduction of exotic species such as eucalyptus, or replacement of native vegetation by pastures). Further, besides the Azogues city, which does not possess a water treatment system prior to final in-river disposal, this sub-basin has an important number of small villages that produces a significant contribution of untreated sewage (Da Ros, 1995). This problem is even accentuated by the industrial production of cement (Da Ros, 1995; Pauta-Calle and Chang-Gómez, 2014). Similarly, the neighbouring Magdalena sub-basin has a very high proportion (70.7%) of its surface covered by altered vegetation and a limited native forests cover (24.2%), which explains the presence of a WQC1 station located in the native forest area and a WQC2 station by the outlet of the sub-basin. Both, the Burgay and the Magdalena sub-basins, do not have adequate land management programs.

In the other sub-basins, with the presence of WQC2 stations (Pauite, Pindilig and Sta. Barbara), the proportion of surface with altered vegetation cover is significant, even though the majority of the WQ sampling stations are classified as C1. All WQ sampling stations of the Paute sub-basin (Fig. 3), except one, are classified as C1 and are located in areas where no significant human settlements are present; only the WQ station located at the outlet of the sub-basin is classified as C2, which can be explained not only by the direct influence of human settlements located immediately

---

**Fig. 5.** Box plots of the WQ significant parameters as a function of the two WQ classes C1 (less polluted) and C2 (highly polluted): (a) electric conductivity (EC); (b) faecal coliforms (FC); (c) dissolved oxygen (DO); (d) total hardness (TH); (e) chlorides (CL); (f) nitrates (N\textsubscript{NO3}); (g) total alkalinity (TALK); (h) 5-day biochemical oxygen demand (BOD\textsubscript{5}); and (i) turbidity (TU).
upstream along the river course, but also by the direct contribution from the Magdalena sub-basin into the lower part of the Paute sub-basin (Figs. 2 and 3).

For the Pindilig and Santa Bárbara sub-basins that contain WQ sub-stations, the location of the WQC1 stations is congruent with its location near a native forest and unaltered vegetation. Further, a significant proportion of their surface (50.2%) for the Santa Bárbara sub-basin and 32.6% for the Pindilig sub-basin) is covered by altered vegetation and there is only a very scarce presence of human settlements. In this regard, once again, the land cover distribution suggests point source pollution from focalised human settlements, which is accentuated in the Santa Bárbara sub-basin by the fact that the average FCC2 (13573.1 bacteria 100 ml−1) doubles the respective average FCC1 (6272.4 bacteria 100 ml−1). For the Pindilig sub-basin, this is confirmed by the average FCC2 value (12621.4 bacteria 100 ml−1) that nearly doubles the FCC1 value (6909.8 bacteria 100 ml−1), as well as, by the fact that the average (BOD3,C2 value (22.6 mg L−1) triplicates the respective (BOD3,C1 value (6.1 mg L−1).

Hereafter, point source pollution locations seem to have a low influence on the downstream WQ, as the presence of human settlements nearly vanishes, there exists an increment of discharge and the presence of native and altered vegetation prevails. Particularly, the presence of native vegetation has the potential of determining the existence of riparian ecosystems in several spots along the river courses forming buffer systems for the enhancement of riparian ecosystems and, finally, of the downstream WQ. Formation of these riparian buffers should be encouraged in the study basin, for instance for the removal of nitrates (Lowrance et al., 1997; Sweeney and Newbold, 2014; Connolly et al., 2015), that in WQC2 sites reaches an average value of 1.51 mg L−1, which is about 5 times the average in the WQC1 sites (0.31 mg L−1).

On the other hand, the sampling sites located by Cuenca city (Tomebamba and Yanuncay sub-basins) are classified as belonging to the C1 group, despite of being located at the most urbanised area of the Paute river basin. This is logic and explained by the fact that in the city of Cuenca an acceptable treatment of most of the wastewater takes place, with favourable consequences for the WQ of the Paute river. In the remaining sub-basins of the Paute river basin the presence of native forest and unaltered vegetation (such as páramo) is significant and the level of anthropisation is low, which emphasises the congruence of the current WQ classification results.

5. Conclusions

The study illustrated the usefulness of the utilised non-supervised (k-means) and supervised (k-NN/CA) pattern recognition algorithms for the analysis and interpretation of complex data sets, in the context of river WQ assessment, the identification of pollution sources/factors and the understanding of spatial variations in river WQ, offering the potential of contributing positively to effective river WQ management in the study basin, such as, reducing the number of parameters to be monitored in the future, which shall result in a significant reduction of the monitoring costs, without surrendering on accuracy. Hence, the study revealed as well the worth of combining data mining techniques and common GIS tools to cross-check on congruency of results.

Acknowledgements

This study was conducted in the context of the Master’s thesis (National University of La Plata, Argentina) carried out by the first author and co-directed by the second author. Additional work took place within the frame of the fellowship awarded to the second author by the Ecuadorian Secretary for Higher Education, Science, Technology and Innovation (SENESCYT) through its “PROMETEO Viejos Sabios” Program. The authors would like to express their gratitude to SENAGUA - DHS for making the raw data accessible to the current study. Further, the authors are thankful to Jan Feyen for going through the first draft of this manuscript and to the anonymous reviewers for their important remarks.

References
