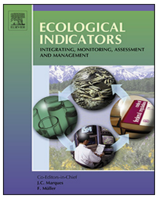
Ecological Indicators 111 (2020) 106037

Contents lists available at ScienceDirect

Ecological Indicators

journal homepage: www.elsevier.com/locate/ecolind



Multivariate-statistics based selection of a benthic macroinvertebrate index for assessing water quality in the Paute River basin (Ecuador)

Gonzalo Sotomayor[a,d,*], Henrietta Hampel[a,b], Raúl F. Vázquez[a,c], Peter L.M. Goethals[d]

[a] *Laboratorio de Ecología Acuática (LEA), 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*
[d] *Laboratory of Environmental Toxicology and Aquatic Ecology, Department of Applied Ecology and Environmental Biology, Ghent University, Coupure Links 653, 9000 Ghent, Belgium*

ARTICLE INFO

Keywords:
SIMCA
Pattern recognition
Biotic index

ABSTRACT

Multivariate statistics -Soft Independent Modelling of Class Analogies (SIMCA), Principal Components Analysis (PCA), Multiple Regression (MR)- were used to search for key biotic and water quality (WQ) variables within a dataset/matrix collected over a five-year period in the Paute River Basin (Ecuador). Benthic macroinvertebrates and 27 descriptive physical, chemical, microbiological, hydrological and geomorphological variables were collected from 64 monitoring sites across the basin. Nine macroinvertebrate biotic indices were calculated. The SIMCA method was applied to find the most accurate biotic index that best discriminated among less polluted (C1), moderately polluted (C2) and highly polluted (C3) sites. A cross-validation scheme was applied to evaluate the performance of the modelling process. Within the PCA that was further refined using a Kruskal-Wallis test, the key WQ variables that mostly contributed to the macroinvertebrate-based WQ classification were identified. The results showed that the Elmidae-Plecoptera-Trichoptera (ElmPT) index was the most accurate biotic classifier. Riparian vegetation and streambed heterogeneity were the best predictors of the C1 class, while the concentration of faecal coliforms, pH, temperature and dissolved oxygen, best predicted the C3 class. The reduction of the field monitored parameters could help designing more cost-effective but equally accurate future WQ monitoring schemes in the basin.

1. Introduction

Surface water quality (WQ) is affected by many interacting processes (Lischeid and Bittersohl, 2008). Commonly, these are a combination of natural and anthropogenic factors whose relative influence changes in both time and space (Baker, 2005; Barnett et al., 2008; Harper et al., 2008). Good quality water is a crucial component for sustainable socio-economic development (Bartram and Ballance, 1996). Consequently, monitoring programmes that provide spatiotemporal representations and reliable WQ estimations are necessary (Simeonov et al., 2003).

In this context, the use of variables measuring different physical, chemical and biological properties of surface water bodies is a useful way to assess their health. From an ecohydrological perspective, it is necessary to understand the links between physicochemical stressors (that cause pollution/contamination) and biological receptors (such as benthic macroinvertebrates) to support policies for the sustainable management of natural resources (Loinaz, 2012). To this end, implementation of control, as well as protection policies, should be based on indices of proven reliability, that is, on indices that correctly detect the health status of the assessed environments (Dos Santos et al., 2011), contributing decisively to define which and how much improvement is needed in a given ecosystem (Feio et al., 2014).

Worldwide experience has demonstrated that the most useful biological assessment methods for freshwater monitoring are based on benthic macroinvertebrates. Thereby, many indices have been developed using them for evaluating the ecological status of lotic systems (Herman and Nejadhashemi, 2015).

Biomonitoring with the use of these indices aims at solving the problem of determining if a given stream should be considered degraded or not, as a result of any anthropic impact. This problem corresponds to a classification process using only two classes, namely, “degraded” or “not degraded” (Dos Santos et al., 2011). Hence, multivariate-statistics methods are increasingly in use for tackling

[*] Corresponding author.
E-mail addresses: gonpa83@yahoo.com, gonzalo.sotomayor@ugent.be (G. Sotomayor), hennihampel@gmail.com (H. Hampel), raulfvazquezz@yahoo.co.uk (R.F. Vázquez), peter.goethals@ugent.be (P.L.M. Goethals).

https://doi.org/10.1016/j.ecolind.2019.106037
Received 15 July 2019; Received in revised form 9 December 2019; Accepted 21 December 2019
1470-160X/ © 2019 Elsevier Ltd. All rights reserved.

classification problems in WQ assessments based on large data sets (Einax et al., 1997; Kannel et al., 2007). In this regard, supervised pattern recognition, a set of multivariate-statistics techniques, has been developed to solve class membership problems (Brereton, 2007; Lavine and Rayens, 2009; Sotomayor et al., 2018).

Various WQ monitoring efforts by local and national Ecuadorian governments have been carried out in the past at some particular locations of the country as well as nationwide (SENAGUA, 2016). However, the resulting data cannot be used for research or management purposes as it is not publicly available. As an exception, the Ecuadorian National Secretary of Water (SENAGUA) – Santiago River Hydrographic Demarcation (DHS), provided the current study with an extensive database developed from the records of 64 monitoring sites located in the Paute River Basin. The database includes 8127 observations, within a 5-year monitoring program (2008, 2010–2013), containing information about physical, chemical, hydrological, geomorphological and microbiological WQ descriptive variables, as well as, about benthic macroinvertebrates.

The purpose of this study was (i) to identify and select a benthic macroinvertebrate index that best reflects the status of the aquatic ecosystem in the Paute River Basin; and (ii) to identify relationships between physicochemical, hydromorphological and microbiological descriptors-variables and the benthic macroinvertebrates. Multivariate pattern recognition methods, mainly the Soft Independent Modelling of Class Analogy (SIMCA) method (Wold, 1976; Wold and Sjöström, 1977) were used. To the best of our knowledge, this is the first ecohydrological application of the SIMCA method worldwide.

2. Materials and methods

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

The Paute River Basin (PRB) is located in the south of Ecuador (Fig. 1) and covers an area of 6442 km^2. The length of the river's main stem is approximately 120.4 km. The PRB is one of the most important hydrographic systems of Ecuador due to the exploitation of its high hydroelectric potential, which currently supports more than 40% of the energy demand of the country (CONELEC, 2011). The annual average discharge of the river is about 136 m^3 s^{-1}; its temporal variation is strongly influenced by the presence of hydroelectricity generation subsystems that are conforming the Integrated Paute System (CONELEC, 2009). The Paute River provides water resources to agricultural, rural, urban and industrial areas of an important part of the southern Andes of Ecuador (Da Ros, 1995) and runs to the Upano River, which belongs to the Amazon River system. Two major cities, namely Cuenca and Azogues are located in the PRB with approximately 500,000 and 33,850 inhabitants, respectively (2010 census). Pollution in the PRB comes from both point and non-point sources, including 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 processes leading to ground up-lift, slope sharpening and resulting in large amounts of suspended sediments that are transported by the river (Astudillo et al., 2010).

Altitudes range between 500 m and 4250 m above sea level (a.s.l.), with the majority (61.3%) being in the 2550–3575 m range, 4.3% in the 500–1525 m range, 13.7% in the 1525–2550 m range, and 20.7% of the basin is situated above 3575 m a.s.l. On average, slopes vary between 25% and 50%; in the upper part of the basin a mountainous relief is dominant whilst a gentler relief is representative of the central and lower parts.

Average air 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 region), with an average 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.

2.2. Sampling WQ descriptive variables

Twenty-seven WQ descriptive variables were sampled in 64 monitoring sites (Fig. 1), representing the WQ distribution in the study basin (SENAGUA, 2016). These variables were: aluminium (Al), ammonium-nitrogen (N-NH$_4$), cadmium (Cd), copper (Cu), chloride (CL), fluoride (FL), iron (Fe), nickel (Ni), nitrate-nitrogen (N-NO$_3$), lead (Pb), pH, potassium (K), sodium (Na), total alkalinity (TALK), total hardness (TH), total phosphorus (P-tot), dissolved oxygen (DO), 5-day biochemical oxygen demand (BOD$_5$), faecal coliforms (FC), river slope (Slp), Shreve river order (Shreve), elevation (Elev.), electric conductivity (EC), total solids (TS), turbidity (TU), water temperature (WT) and the fluvial habitat index of the Environmental Protection Agency (FHI-EPA) (Barbour et al; 1999). The FHI-EPA is focused on the visual assessment of streambed and riparian habitat, the alteration of which is considered one of the major stressors of aquatic systems (Barbour et al; 1999).

On average, the monitoring sites were visited five times per year.

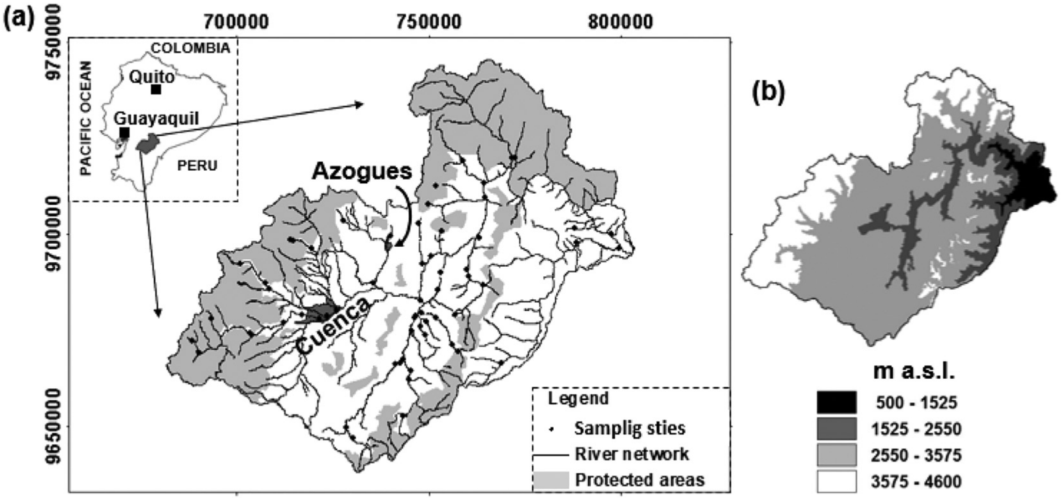


Fig. 1. Study area: (a) The Paute River basin in the continental Ecuador, its two largest cities (Quito and Guayaquil) and the location of the 64 water sampling sites. (b) A Digital Elevation Model of the basin.

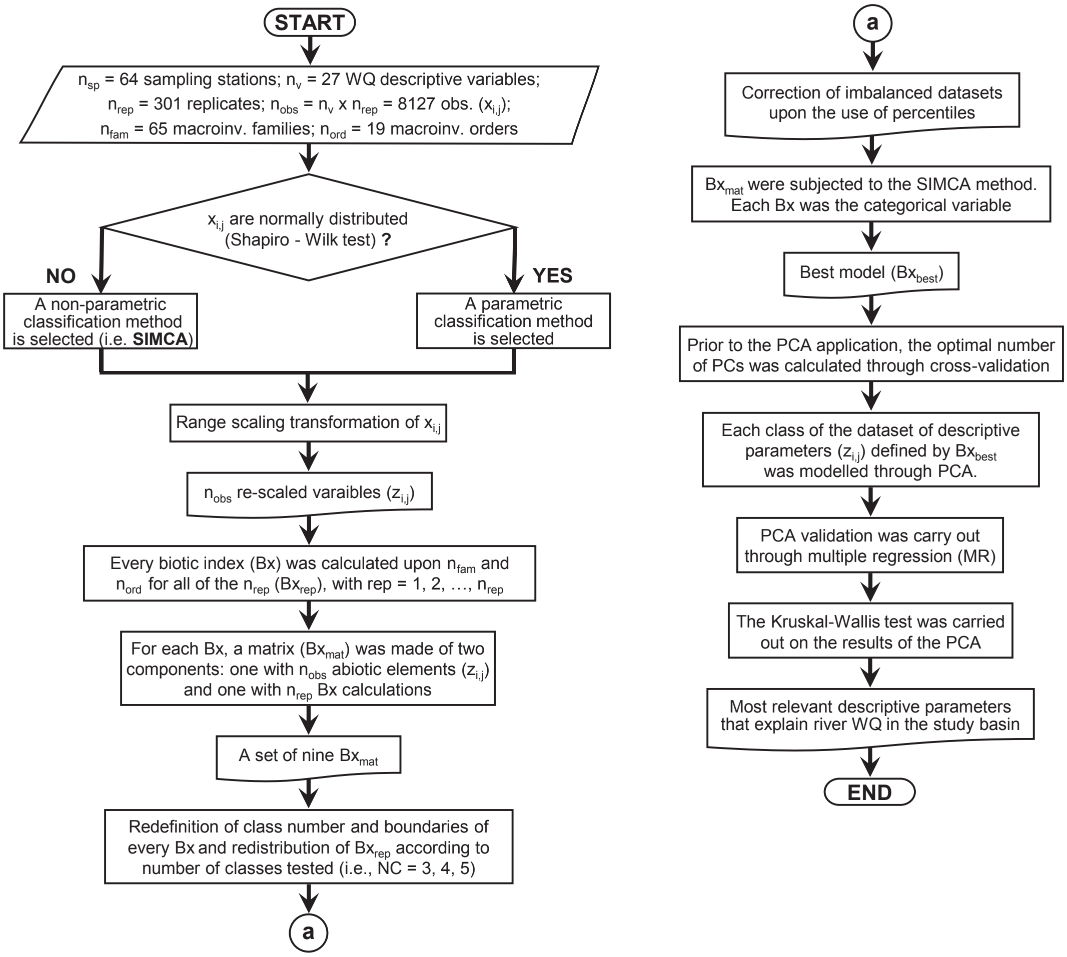

START
n_{sp} = 64 sampling stations; n_v = 27 WQ descriptive variables;
n_{rep} = 301 replicates; n_{obs} = n_v x n_{rep} = 8127 obs. ($x_{i,j}$);
n_{fam} = 65 macroinv. families; n_{ord} = 19 macroinv. orders
$x_{i,j}$ are normally distributed (Shapiro - Wilk test) ?
NO
YES
A non-parametric classification method is selected (i.e. SIMCA)
A parametric classification method is selected
Range scaling transformation of $x_{i,j}$
n_{obs} re-scaled varaibles ($z_{i,j}$)
Every biotic index (Bx) was calculated upon n_{fam} and n_{ord} for all of the n_{rep} (Bx_{rep}), with rep = 1, 2, …, n_{rep}
For each Bx, a matrix (Bx_{mat}) was made of two components: one with n_{obs} abiotic elements ($z_{i,j}$) and one with n_{rep} Bx calculations
A set of nine Bx_{mat}
Redefinition of class number and boundaries of every Bx and redistribution of Bx_{rep} according to number of classes tested (i.e., NC = 3, 4, 5)
a
a
Correction of imbalanced datasets upon the use of percentiles
Bx_{mat} were subjected to the SIMCA method. Each Bx was the categorical variable
Best model (Bx_{best})
Prior to the PCA application, the optimal number of PCs was calculated through cross-validation
Each class of the dataset of descriptive parameters ($z_{i,j}$) defined by Bx_{best} was modelled through PCA.
PCA validation was carry out through multiple regression (MR)
The Kruskal-Wallis test was carried out on the results of the PCA
Most relevant descriptive parameters that explain river WQ in the study basin
END

Fig. 2. Flowchart of the modelling protocol that was implemented in the current study.

Some were sampled more frequently, because they were located either at highly polluted sites or, on the contrary, at unaltered environmental (i.e., reference) locations. As a result, a WQ database was developed for the $n_{sp} = 64$ monitoring sites with $n_{rep} = 301$ sampling replicates of $n_v = 27$ WQ descriptive variables, resulting in a total of $n_{obs} = n_{rep} \times n_v = 8127$ observations, represented as $x_{i,j}$, where $i = 1, 2, …., n_v$ and $j = 1, 2, …, n_{rep}$ (Fig. 2). Table 1 lists the main statistics for each of the studied WQ descriptive variables.

2.3. Sampling and analysis of benthic macroinvertebrates

Benthic macroinvertebrates were sampled at each of the 64 monitoring sites (Fig. 1). At each monitoring site, a 20-m long reach was selected. Three transects, evenly spaced across the transect (located at a 0 m, 10 m and 20 m distance, respectively) were delineated (Von Ellenrieder, 2007). A macroinvertebrate sample was collected along each transect. Sampling was carried out for three minutes, encompassing all existing microhabitats characterised by different depths, substrates and water velocities. Samples were collected by following the standardised ‘kick-sampling’ process, using a 25 × 25 cm^2 nylon handnet (mesh opening size: 0.5 mm) placed vertically on the stream bottom, in front of which (upstream) the substratum was vigorously stirred by kicking it (Jacobsen et al., 1997). The three samples from each transect were pooled together, and sampling continued by visually inspecting (for about 20 min) the substrate and aquatic vegetation for checking that tightly clinging taxa (e.g. Blephariceridae) that may have not been dislodged by kick-sampling (Roldán, 2003) have also been collected.

Macroinvertebrates samples were preserved in 70% ethanol solution and mostly identified to the family level with the use of a stereomicroscope. $n_{fam} = 65$ families were identified and grouped into $n_{ord} = 19$ taxonomic groups (in its great majority orders).

2.4. Assessing biotic indices

Nine biotic indices (Bx) were calculated using n_{fam} and n_{ord}. Each Bx was calculated for each of the n_{rep} replicates and a matrix (Bx_{mat}) was produced, which had an abiotic and a biotic component. The abiotic component is made of all of the n_{obs} abiotic observations, whilst the biotic component is made of all of the n_{rep} definitions of the particular Bx. Thus, the resulting matrices differ only in the biotic component since the abiotic component is the same in all of the matrices. The biotic indices were considered as biological response (dependent) variables in the classification models. The indices that were calculated and included were (1) the Biological Monitoring Working Party (BMWP) (Armitage et al., 1983), calibrated for Colombia (BMWP_Col) (Roldán, 2003); (2) the Andean Biotic Index (ABI), similar to the BMWP but adapted for the northern and central Andean streams above 2000 m a.s.l. (Ríos-Touma et al., 2014); (3) a combined ABI_BMWP_Col index as proposed by Sotomayor (2016), namely, (i) ABI for streams located above 2000 m a.s.l.; and (ii) BMWP_Col for lower elevations (< 2000 m a.s.l.); (4) the Ephemeroptera-Plecoptera-Trichoptera (EPT) index, expressed as the number of the EPT individuals divided by the total macroinvertebrate abundance (Lenat, 1988; Carrera and Fierro, 2001); (5) the Elmidae-Plecoptera-Trichoptera (ElmPT) index (Von Ellenrieder, 2007), calculated similarly to the EPT index; (6) family richness; (7) the Average Score Per Taxon (ASPT) (Walley and Hawkes, 1996). In this study, three variations of this ASPT index were calculated, namely, (i) the

Table 1
Main statistics associated to the water quality (WQ) descriptive variables that were monitored throughout years 2008 and 2010–2013 in the Paute River basin, Ecuador (Fig. 1). Legend: Al = aluminium; BOD_5 = 5-day biochemical oxygen demand; Cd = cadmium; CL = chlorides; Cu = cooper; DO = dissolved oxygen; EC = electric conductivity; Elev = elevation, FC = faecal coliforms; Fe = iron; FHI–EPA = fluvial habitat index-Environmental Protection Agency (Barbour et al., 1999); FL = fluorides; K = potassium; N-NH_4 = ammonium-nitrogen; Na = sodium; Ni = nickel; N-NO_3 = nitrate-nitrogen; Pb = lead; P-tot = total phosphorus; Shreve = river order calculated with the Shreve method; Slp = slope; TALK = total alkalinity; TH = total hardness; TS = total solids; TU = turbidity; WT = water temperature; STD = standard deviation; and m a.s.l. = meters above sea level.

WQ Parameter	Mean	Median	STD	Range
Al (mg L^{-1})	0.06	0.00	0.23	0.00–1.59
BOD_5 (mg L^{-1})	10.30	9.98	8.08	0.00–55.82
Cd (mg L^{-1})	0.01	0.00	0.05	0.00–0.61
CL (mg L^{-1})	5.26	0.73	27.32	0.00–363.86
Cu (mg L^{-1})	0.02	0.00	0.12	0.00–1.32
DO (mg L^{-1})	6.83	6.89	0.75	4.12–9.75
EC (μS cm^{-1})	122.43	70.00	197.36	2.96–1810.00
Elev (m a.s.l.)	2419.07	2420.00	730.01	480.00–3780.00
FC (bacteria 100^{-1} ml^{-1})	5038.32	1600.00	6451.66	1.00–16000.00
Fe (mg L^{-1})	0.21	0.00	0.49	0.00–3.73
FHI – EPA	129.49	128.00	28.54	71.00–184.00
FL (mg L^{-1})	1.56	0.42	6.83	0.00–67.89
K (mg L^{-1})	1.63	0.37	5.29	0.00–69.86
N-NH_4 (mg L^{-1})	0.78	0.00	1.57	0.00–15.00
Na (mg L^{-1})	5.28	3.37	9.15	0.00–112.89
Ni (mg L^{-1})	0.02	0.00	0.14	0.00–1.51
N-NO_3 (mg L^{-1})	0.56	0.13	2.05	0.00–20.79
Pb (mg L^{-1})	0.02	0.00	0.07	0.00–0.85
pH	7.52	7.56	0.65	5.30–9.43
P-tot (mg L^{-1})	0.38	0.16	0.50	0.00–2.09
Shreve	334.15	51.00	1048.84	1.00–5760.00
Slp (%)	25.82	10.31	35.05	0.00–142.30
TALK (mg L^{-1})	0.74	0.08	1.35	0.00–7.80
TH (mg L^{-1})	33.34	23.60	36.50	0.00–263.56
TS (mg L^{-1})	2.00	0.01	10.42	0.00–116.00
TU (NTU)	19.48	0.92	88.16	0.00–1136.81
WT (°C)	14.57	14.00	3.27	8.70–23.50

calculation focused on the BMWP index validated for Colombia ASPT_BMWP_Col; (ii) the ASPT_ABI index focused on the ABI index; and (iii) the ASTP_ABI_BMWP_Col index focused in turn on the ABI_BMWP_Col index.

2.5. Data pre-processing

Fig. 2 depicts the flowchart of the multivariate-statistics based protocol applied on this study. The Shapiro-Wilk test (Shapiro and Wilk, 1965; Yap and Sim, 2011) was applied to evaluate the normality of the distribution associated with every WQ descriptive variable. This test showed that none of the WQ descriptive variables (with exception of pH) were normally distributed (Table 1). Thus, for the implemented supervised pattern recognition process, a non-parametric method, SIMCA, was selected. Prior to the SIMCA application, all WQ variables were re-scaled in a 0–1 interval as follows:

$$z_{i,\,j} = \frac{x_{i,j} - L_j}{U_j - L_j} \quad \text{for i= 1, 2, ..., } n_v \text{ and j= 1, 2, ..., } n_{rep} \tag{1}$$

where L_j and U_j are respectively the lower and upper limits of the variable range, so that all of the $z_{i,j}$ ranges between 0 and 1 (Frank and Todeschini, 1994), eliminating as such the discrepancy of the extent of ranges of variation of the original variables.

2.6. Classification process

Classification models or supervised pattern recognition methods are statistical tools aimed at finding out a model capable of assigning every sampling site to its class using some independent variables (descriptors). The class variable is the dependent categorical variable (Frank and Todeschini, 1994). In the current study, the 27 WQ descriptive variables were the independent ones and each Bx was the dependent variable, introduced in the SIMCA model using not the actual Bx values but the class number corresponding to each value.

2.6.1. The soft independent modelling of class analogy (SIMCA) method

SIMCA carries out "soft modelling", i.e., this method can identify sampling sites as belonging to multiple classes and not necessarily producing a classification of sampling sites into non-overlapping classes (Pomerantsev, 2014). To build up the classification models, the first step is to apply a Principal Component Analysis (PCA) for independently modelling each predefined class, which could be described by a different number of principal components -PCs- (Brereton, 2007). In this way, SIMCA defines G class models. In a final step, a new sampling site is projected on each PCA subspace (for every class) and the respective distances between the projections and the classes are evaluated to assess on the adequacy of class assignment. Distances are calculated on the basis of normalised Q residual variances (the variation in the data not explained by the PC models) and normalised Hotelling T2 values (Hotelling, 1936), the multivariate extension of the Student's *t*-test (Balabin et al., 2010; Ballabio and Consonni, 2013). Further details about the SIMCA method can be found for instance in Massart et al., (1988), Brereton (2007) and Pomerantsev (2014).

SIMCA was used in this study to assess the appropriateness of each of the nine Bx applied to evaluate the surface WQ via benthic macroinvertebrates. Thus, the method was used for verifying whether the allocation of WQ classes to the sampling sites, according to the WQ descriptive variables of the sites (pattern recognition), matched the classes determined previously by the nine Bx. Thus, the best Bx was defined in the framework of the best SIMCA classification model adjustment.

2.6.2. Principal components analysis (PCA) in SIMCA

PCA is an intrinsic part of the SIMCA algorithm. In this method, the original data matrix X with n_v variables and n_{rep} replicates is reduced to the parts A (factor loadings) and U (factor scores). The loadings indicate how much an original variable is 'loaded into' a PC, that is, how relevant the variable is. Scores are the coordinates of one sampling site in the new coordinate system (Zupan, 1990). These factors are new non-correlated synthetic variables and explain the total variance of all the original variables. In the SIMCA method the PCA gives the necessary information about which variables are useful for discriminating classes.

Further, an important prior task of a PCA is selecting an optimal number of PCs, for which, a cross-validation process through the Venetian blinds method (Ballabio and Consonni, 2013; Ballabio, 2015) was used. This method is based on the use of segments of consecutive sampling sites, where a test segment is determined by selecting every s-th sampling site in the data set, starting at sampling site numbered 1 through s, with s being the number of data splits (Mevik and Wehrens, 2015). The data (n_{rep}) was split into 5 groups for cross-validation implying that for every group the PCA used 80% of the data for model fitting, leaving out 20% of the data for model validation.

2.6.3. Evaluating the performance of classification models

For evaluating whether a given SIMCA classification model (out of the nine ones) correctly allocated WQ classes to sampling sites, the three following classification measures were used: accuracy (Acc), the F measure (F) and the non-error rate (NER). The variation range of these three classification measures is between 0 and 1, with 1 being their optimal value. Acc is the most used empirical measure to assess a supervised pattern recognition problem (Hand, 2012); it is the ratio of correctly assigned sampling sites:

$$\mathrm{Acc} = \frac{\sum_{g=1}^{G} n_{gg}}{n_{rep}}, \quad (2)$$

where G is the number of classes; n_{gg} is the number of sampling sites of the g-th class that are correctly classified. Not assigned sampling sites are not considered in the Acc calculation. NER is the average of all of the class sensitivity values (Ballabio and Consonni, 2013), with the sensitivity of the g-th class (Sn_g) being the model ability to correctly recognise sampling sites belonging to the g-th class. It is defined as the ratio between n_{gg} and the total number of sampling sites belonging to the g-th class (n_g). Not assigned sampling sites are not considered in the calculation of Sn_g.

F is a classification measure (Ballabio et al., 2018) that considers the NER and the mean precision (Pr_{mean}), with the precision of the g-th class (Pr_g) representing the capability of a classification model to avoid assigning sampling sites of other classes into the g-th class. Pr_{mean} is defined as the ratio between n_{gg} and the total number of sampling sites assigned to that class (n'_g):

$$F = 2\,\frac{(NER)Pr_{mean}}{NER + Pr_{mean}} \quad (3)$$

The assessment of these classification performance measures was carried out in the cross-validation phase by using also in this case the Venetian blinds cross-validation method, with 5 splits. This implies that 20% of the observed data was used during this validation phase to judge on the correct class assignment process (i.e., defining n_{gg}, sn_g, etc.).

2.6.4. Number and limits of biotic indices classes

The standard methodology to establish the number of biotic classes of benthic macroinvertebrates and their boundaries is based on the Reference Condition Approach (RCA), where the biological integrity of a site is defined by the “distance” (or alteration gradient) between current and reference conditions. The alteration gradient is often divided into, say, five classes, reflecting qualitative levels of biological integrity (excellent, good, moderate, poor and bad conditions), which is a classification arrangement particularly suitable for WQ studies. Nevertheless, within the framework of the SIMCA classification models implemented in this study, some particular aspects were taken into account, namely:

- The “standard” version of the ElmPT and FamR indices has no specific number of classes. However, the standard version of the ABI, BMWP_Col, ASPT_BMWP_Col, ASPT_AB and ASPT_ABI_BMWP_Col indices has five classes (Armitage et al., 1983; Ríos-Touma et al., 2014), whilst the “standard” version of the EPT index has four classes (Carrera and Fierro, 2001). There is then a great variability in terms of the number of biotic classes as a function of the biotic indexes. However, in the framework of the current application of the SIMCA method, for an appropriate comparison of the values of the nine Bx, the number of classes must be equal for all of these Bx. Thus, there is the need of adopting the same number of classes for all of the nine studied Bx.
- A dataset is imbalanced if the classification categories are not similarly represented, i.e., the sampling site distributions among classes are skewed (Martina et al., 2017). As a result, classification algorithms are affected in their accuracy performances causing misclassification of sampling sites belonging to the under-represented classes (Chawla, 2009). For avoiding this problem, it was decided to discretise the different Bx using a set of threshold values (Forman, 2003; Blagus and Lusa, 2010) defined upon percentiles throughout the distribution of the Bx_{rep}, for determining the limits (i.e., bounds) of the classes or categories.

In this context, three analyses were carried out varying the number of classes (NC) of the different Bx and, as such, the respective threshold values, namely, NC = 5 (80-th, 60-th, 40-th and 20-th percentiles), NC = 4 (75-th, 50-th and 25-th percentiles) and NC = 3 (66.7-th and 33.3-th percentiles). To the best of the knowledge of the authors, no study has been conducted for WQ assessments by using benthic macroinvertebrates biotic indices that consider only two biological classes. Thus, the current research did not consider NC = 2 as being a feasible number of classes. Using percentiles does not only enable the inspection of the performance of the different Bx as a function of the NC but also deals with the problem of imbalanced datasets.

2.6.5. Choosing the best biotic index through the SIMCA method

The best biotic index, was defined upon the calculation of the above depicted classification measures Acc, NER and F for each of the nine studied Bx. After comparing the performance of the nine different SIMCA classification models, the model with the highest values of the associated classification measures was referred to as the best one. Correspondingly, the respective Bx, the dependent variable of the best classification model, was chosen as the best one.

2.6.6. Assessing the most significant WQ descriptive variables

Upon the selection of the best classification model, simultaneously, the PCA component of the SIMCA process identifies the WQ descriptive variables that explain each one of the biotic classes.

In the context of the PCA, with the aim of determining the number of interpretable (i.e., non-trivial) ordination axes, the “cutoff rule” criterion was applied, which regards loadings as being significant when $|\text{loadings}| > A_{th}$, with A_{th} being a subjectively pre-established threshold value (Peres-Neto et al., 2003) that was herein fixed as 0.25 (Chatfield and Collins, 1980).

Moreover, for improving the identification, previously achieved by the PCA, of the informative WQ variables that describe suitably all of the biotic classes, a further analysis, the non-parametric Kruskal-Wallis test (K-W) (Kruskal and Wallis, 1952), was carried out. Herein, the set of n_{rep} data of each informative WQ variable (i.e., with $|\text{loadings}| > 0.25$) was divided into biotic classes, after which the K-W test was carried out to assess whether there is a significant difference among these classes.

2.6.7. Testing the reliability of the PCA

The reliability of the PCA must be tested (Ouyang, 2005), for instance, by comparing the outcome of multiple regression (MR) analyses performed with and without the WQ descriptive variables that the PCA identified as being informative (i.e., the WQ descriptive variables with $|\text{loadings}| > 0.25$).

Three cases were developed for comparison through MR analyses, using as the goodness of fit measure the adjusted R^2 statistic (R^2_{adj}), which enables comparing the performance of models involving different numbers of independent variables (Rawlings et al., 1998). In the MR analyses, the dependent variable was always the best biotic index and the independent variables were either: (i) all of the WQ descriptive variables (MR-model 1); (ii) WQ descriptive variables with associated $|\text{loadings}| > 0.25$ (MR-model 2); and (iii) WQ descriptive variables with associated $|\text{loadings}| \leq 0.25$ (MR-model 3). If the PCA was successful identifying the informative variables that explain most of the WQ variability, the R^2_{adj} values should not vary significantly with regard to MR-models 1 and 2, whilst for MR-model 3 its value is expected to be markedly lower than for the other two MR-models.

All the statistical analyses were implemented with MATLAB® (Hanselman and Littlefield, 2012) version 2014, using the Classification toolbox version 5.0 (Ballabio and Consonni, 2013), as well as specific-purpose subroutines, developed particularly for this study.

2.6.8. Congruency of the resulting WQ classification

The correspondence of the spatial distribution of the biotic WQ classes was assessed by its comparison with the land cover (LC) distribution of the Paute River basin, using as well auxiliary topographical information. LC raster data (year 2013) was available for the whole extent of the basin. This is the most recent dataset that is available publicly by the Ministry of Environment of Ecuador (MAE, 2013). Geographic Information Systems (GIS) algorithms were applied on the original LC data so that it was reclassified to a more convenient form, which enabled a direct association between LC and the spatial distribution of the WQ classes. The considered LC classes were: (i) altered vegetation; (ii) woody native vegetation; (iii) without cover/urbanised; and (iv) páramo (unaltered). Additionally, a raster digital elevation model (DEM) of the whole catchment was available with a resolution of 50 × 50 m^2. The referred congruency assessment was based on the visual inspection of the geographical distribution of the WQ classes, the LC, and the DEM. ArcGis® and TerrSet® were used for all of the GIS analyses.

3. Results

3.1. Sampling and analysing benthic macroinvertebrates

53,452 macroinvertebrate individuals belonging to 65 families were collected and grouped into 19 taxonomic groups. Ephemeroptera were the most dominant, accounting for 63.7% of all individuals, followed by Diptera (11.2%), Trichoptera (8.2%), Coleoptera (6.3%) and Oligochaeta (5.2%). The rest of the groups accounted for only 5% of the total abundance.

3.2. Number and limits of biotic indices classes

The results showed that a five-class Bx scheme (standard methodology) for defining the number of biotic classes of benthic macroinvertebrates produces an imbalanced distribution with reduced sampling site concentrations at the extreme classes (C1 and C5) (Fig. 3). In contrast, the use of percentiles produces uniform sampling site distributions throughout the different classes, thus resolving the problem of imbalanced data.

The results suggest that the distributions of the Acc, NER and F classification measures are significantly different from each other as a function of the NC (Fig. 4). Further the results depict that the maximum performance was obtained for NC = 3 (Fig. 4). Thus, it was decided to use this number of classes for all of the nine studied Bx. All Bx were accordingly discretized using the 33.3-th and 66.7-th percentiles. Each of the three classes was assigned a WQ attribute; C1: less polluted, C2: moderately polluted, C3: highly polluted.

3.3. Choosing the best biotic index through the SIMCA method

Best performance was obtained for the ElmPT biotic index for which the maximum values of the three referred classification measures were observed (Fig. 5). The ElmPT index varied between 0.0% and 87.0% with an average value of 20.2% and a standard deviation of ± 19.3%. The class limits for NC = 3 were 7.6% (33.3-th percentile) and 24.6% (66.7-th percentile).

3.4. Assessing the most significant WQ descriptive variables

The optimal number of PCs chosen a priori in the framework of the SIMCA method were 4, 9 and 3 (Table 2), respectively, for the biotic classes C1, C2 and C3 (ElmPT index). As a result, for the three PCA models, > 80% of the explained variance is present in its first three components.

Fifteen WQ descriptive variables (Table 3) were regarded as explaining most of the surface WQ variability, represented in the three biotic classes C1, C2 and C3, that is, Al, DO, EC, Elev, FC, Fe, FHI-EPA, FL, N-NH_4, P, pH, Shreve, Slp, TALK and WT. Furthermore, considering the results of the K-W test, only seven of the fifteen WQ descriptive variables were found significant ($p < 0.05$), namely, DO, EC, FC, FHI-EPA, pH, Shreve and Slp (the p-values were $p_{Al} = 0.91$, $p_{DO} \leq 0.0001$, $p_{EC} = 0.0001$, $p_{Elev} = 0.51$, $p_{FC} \leq 0.0001$, $p_{Fe} = 0.46$, $p_{FHI-EPA} \leq 0.0001$, $p_{FL} = 0.28$, $p_{N\text{-}NH_4} = 0.4033$, $p_P = 0.50$, $p_{pH} = 0.0055$, $p_{Shreve} \leq 0.0001$, $p_{slope} = 0.0004$, $p_{TALK} = 0.29$ and $p_{WT} = 0.0553$).

Fisher's least significant difference (LSD) test was used to calculate intervals around the means of these seven most significant WQ descriptive variables, as a function of the three biotic classes C1, C2 and C3 (Fig. 6) for identifying the statistical populations whose averages are statistically different from each other (Dodge, 2008). The values adopted by pH (Fig. 6a), FC (Fig. 6b), EC (Fig. 6c) and Shreve (Fig. 6d) are larger for the sampling points belonging to C3 than for the ones belonging to C1 and C2 (averages of FC expressed in bacteria 100^{-1}

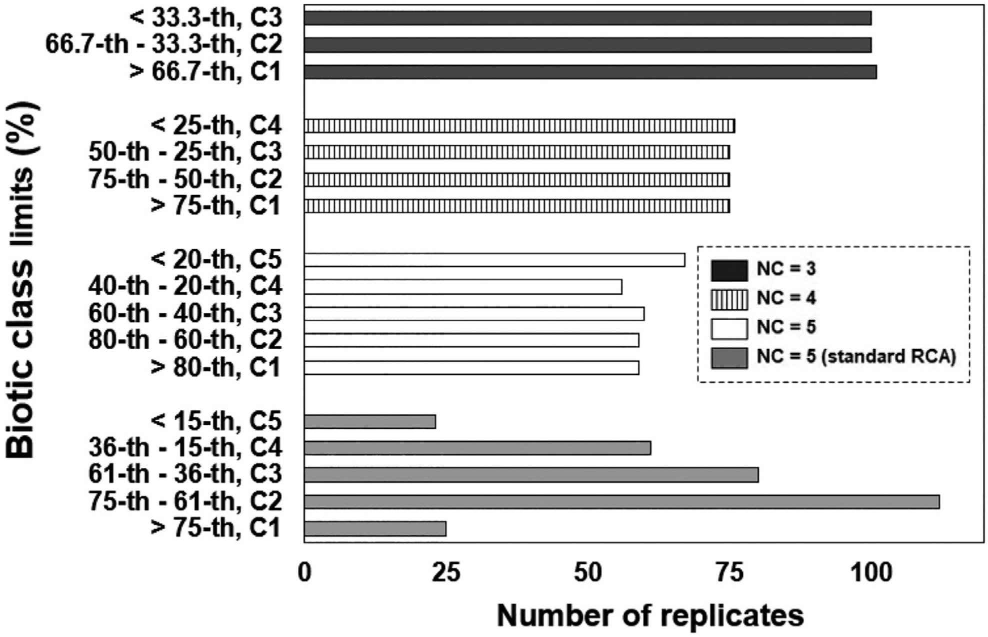


Fig. 3. Distribution of replicates as a function of the number of biotic classes (NC), for the BMWP_Col index (RCA: Reference Condition Approach).

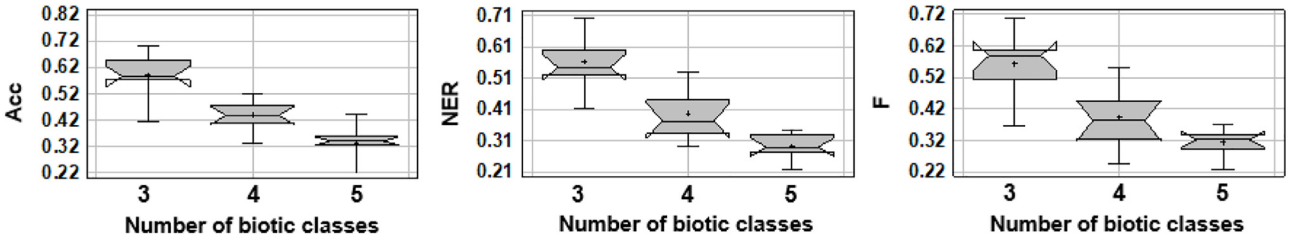


Fig. 4. Notched box plots of Acc, NER and F classification measures as a function of the number of biotic classes (NC), considered for every one of the nine inspected biotic indices.

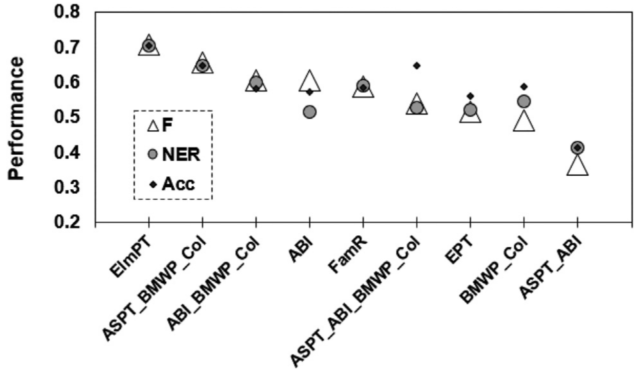


Fig. 5. Performance of each biotic index (Bx) based on F, NER and Acc. Classification was considered correct when the Bx class of a sample matched that of the water quality classification. For Bx abbreviations see section 2.4.

Table 2
Eigenvalues, explained variance and cumulative variance for the three PCA model that correspond to the three biotic classes C1: less polluted, C2: moderately polluted, C3: highly polluted.

Class	PC	Eigenvalue	Variance (%)	Cumulative variance (%)
C1	1	1.7	71.6	71.6
	2	0.2	7.4	79.0
	3	0.1	4.4	83.4
	4	0.1	3.4	86.8
C2	1	1.6	71.3	71.3
	2	0.2	7.7	79.0
	3	0.1	3.9	82.9
	4	0.1	3.2	86.1
	5	0.1	2.7	88.8
	6	0.1	2.2	91.0
	7	0.0	1.8	92.9
	8	0.0	1.4	94.3
	9	0.0	1.2	95.5
C3	1	1.6	69.8	69.8
	2	0.2	8.1	77.8
	3	0.1	4.9	82.8

ml^{-1} are 3378.5 for C1, 5079.9 for C2 and 6740.3 for C3; averages of pH are 7.3 for C1, 7.5 for C2 and 7.7 for C3; averages of EC expressed in $\mu S\ cm^{-1}$ are 86.1 for C1, 93.8 for C2 and 185.0 for C3; and averages of Shreve are 221.7 for C1, 233.3 for C2 and 551.0 for C3). For FHI-EPA, DO and slope (Fig. 6e, 6f and 6 g respectively) the inverse trend was observed (averages of FHI-EPA are 145.7 for C1, 132.5 for C2, 111.3 for C3; averages of DO expressed in mg L^{-1} are 7.0 for C1, 6.8 for C2, 6.6 for C3 and of river slope expressed in % are 32.8 for C1, 29.2 for C2 and 14.5 for C3).

3.5. Testing the reliability of the PCA

The R^2_{adj} value for MR-model 1 (27.0%) was not affected that much when the uninformative WQ descriptive variables (|loadings| ≤ 0.25) were removed from it to produce MR-model 2; on the contrary, its value (28.1%) increased slightly. For the MR-model 3 the R^2_{adj} value (3.1%) was very poor due to the uninformative WQ descriptive variables included in the model. Thus, the PCA was regarded as being reliable.

3.6. Congruency of the resulting WQ classification

The distribution of each biotic class (C1, C2, C3) across the basin, in relation to the relevant land cover is shown in Fig. 7a. Intervened areas cover 37.1% of the basin, woody native vegetation covers 34.4%, without cover/urbanised 3.5% and Páramo 25.0% (Fig. 7b). Sub-basins with increased physicochemical degradation had significantly higher proportion of class C3 sites. Most C3 sites are located around flat areas with increased urbanization and anthropogenic activity (Fig. 7). Thus, flatter values of the Slp. variable suggest potential anthropogenic causes for C3 conditions. Thus, the congruency of the results seems appropriate.

4. Discussion

The results of the study showed that maximum performance is obtained using a 3-class system (Fig. 4), which is in agreement with previous research (Theodoropoulos et al., 2018). As the number of classes increased, F, NER and Acc values decreased. This is probably a multiclass classification issue; the classifier's effort to carry out pattern recognition, and distinguish and correctly classify when NC > 2, is higher (Silva-Palacios et al., 2017), and this effort potentially results to biased determination. However, this study does not necessarily suggest that NC = 3 should be strictly followed elsewhere instead of NC = 5, but a 3-class system had maximum performance in the current study.

Within the selected NC = 3 system, ElmPT had the highest classification performance (> 0.7) (Fig. 5). Previous studies use Plecoptera and Trichoptera as biological indicators for WQ monitoring, mainly as part of the EPT index, but Elmidae have not always been included. Their potential as bio-indicators has been studied in Europe (García-Criado and Fernández-Alaez, 1995; 2001) and USA (Ode et al., 2005; Muenz et al., 2006). Similar findings were observed from studies in South America. Von Ellenrieder (2007) worked in a mountain rainforest of the Andes (Bolivia and Argentine) and found that ElmPT index was best correlated with the local disturbance gradient. Dos Santos et al., (2011) compared the diagnostic capabilities of benthic macroinvertebrates metrics through a classification method. The IBY-4 index that includes Plecoptera, Trichoptera and Elmidae (besides Megaloptera) achieved the best performance in the Yungas Mountains. Elmidae exhibited a decreasing trend from C1 to C3 with respect to presence and total abundance, i.e., 77.5% (C1), 66.3% (C2) and 32% (C3) and 2005 ind. (C1), 807 ind. (C2) and 180 ind. (C3).

Several studies have shown that Elmidae are sensitive to stream degradation (Brown, 1987; Elliott, 2008; Miserendino et al., 2000; Miserendino and Archangelsky, 2006), particularly when caused indirectly through the reduction of the riparian vegetation. Braun et al. (2018) suggested the maintenance of a buffer riparian vegetation to conserve environmental integrity of Brazilian streams. Nessimian et al. (2008) found that elmids were positively correlated with the Habitat

Table 3
Loading values for the principal components of the three PCA models (C1 (less polluted), C2 (moderate polluted) and C3 (highly polluted). Bold values indicate strong influence of the WQ descriptive variables (i.e., |loadings| > 0.25). Aluminium (Al), 5-day biochemical oxygen demand (BOD5), cadmium (Cd), chloride (CL), copper (Cu), dissolved oxygen (DO), electric conductivity (EC), elevation (Elev), faecal coliforms (FC), iron (Fe), fluvial habitat index of the Environmental Protection Agency (FHI-EPA), fluoride (FL), potassium (K), sodium (Na), nickel (Ni), ammonium-nitrogen (N-NH4), nitrate-nitrogen (N-NO3), total phosphorus (P-tot), lead (Pb), pH, Shreve river order (Shreve), river slope (Slp), total alkalinity (TALK), total hardness (TH), total solids (TS), turbidity (TU) and water temperature (WT).

Parameter	Class 1				Class 2									Class 3		
	PC1	PC2	PC3	PC4	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC1	PC2	PC3
Al	−0.04	0.08	−0.08	−0.15	−0.03	−0.06	0.01	−0.07	−0.08	−0.24	−0.07	**0.56**	**0.30**	−0.03	−0.04	−0.02
BOD_5	−0.15	0.00	−0.02	0.14	−0.14	0.04	−0.08	−0.03	−0.19	0.02	−0.07	−0.14	0.15	−0.15	0.01	0.09
Cd	−0.02	0.03	0.00	−0.03	−0.02	0.04	−0.02	0.01	−0.15	−0.14	−0.11	0.02	−0.20	−0.01	0.01	0.00
CL	0.00	0.01	−0.01	−0.01	−0.01	0.03	0.01	−0.03	0.01	0.02	0.11	−0.04	0.08	−0.02	−0.04	−0.02
Cu	−0.02	0.11	−0.05	−0.04	0.00	0.01	0.00	0.00	−0.01	−0.02	−0.02	0.01	−0.03	−0.01	0.03	0.01
DO	**−0.40**	0.15	0.03	−0.09	**−0.38**	0.14	−0.09	−0.05	−0.18	−0.08	**0.43**	**0.44**	**−0.40**	**−0.35**	**0.25**	−0.05
EC	−0.04	0.03	−0.04	−0.03	−0.04	0.04	−0.03	−0.13	−0.13	−0.18	−0.11	**−0.25**	0.09	−0.09	−0.04	−0.07
Elev	**−0.44**	**−0.42**	**0.42**	**−0.30**	**−0.49**	0.13	**0.49**	−0.11	**0.33**	**−0.30**	0.05	−0.06	0.17	**−0.47**	0.06	**0.55**
FC	−0.17	**−0.56**	**−0.73**	**−0.27**	**−0.28**	**−0.91**	−0.09	**−0.26**	−0.05	0.00	−0.04	−0.05	−0.08	**−0.38**	**−0.82**	−0.15
Fe	−0.04	0.07	0.00	−0.20	−0.06	−0.01	0.09	−0.04	−0.03	0.20	**−0.38**	**0.44**	**0.44**	−0.03	0.03	0.01
FHI – EPA	**−0.52**	−0.10	0.20	**0.38**	**−0.44**	−0.01	0.02	**0.57**	0.17	0.13	**−0.51**	−0.05	**−0.28**	**−0.27**	0.20	−0.06
FL	−0.01	0.02	−0.02	0.01	−0.02	0.07	−0.08	−0.16	−0.09	−0.24	−0.14	**−0.28**	0.15	−0.02	0.04	−0.08
K	−0.01	0.01	0.00	−0.01	−0.02	0.02	0.00	−0.03	−0.03	−0.04	−0.02	−0.01	−0.03	−0.03	0.04	0.01
Na	−0.03	0.07	−0.07	−0.12	−0.05	−0.03	−0.07	0.01	−0.02	**−0.27**	0.07	−0.03	0.05	−0.05	−0.03	−0.01
Ni	−0.03	0.04	−0.03	−0.03	−0.02	0.01	−0.01	−0.01	−0.03	−0.05	0.00	−0.02	0.00	−0.06	−0.03	−0.08
$N\text{-}NH_4$	0.00	−0.01	0.03	−0.04	0.00	0.01	0.01	−0.01	−0.02	0.03	−0.02	−0.01	0.06	−0.02	0.06	0.09
$N\text{-}NO_3$	−0.02	0.02	0.02	0.01	−0.01	0.01	−0.01	0.00	−0.02	−0.03	−0.01	−0.01	−0.01	−0.04	−0.08	−0.10
P-tot	−0.16	−0.13	−0.23	**0.68**	−0.15	−0.11	**−0.32**	**0.58**	−0.07	**−0.31**	**0.33**	−0.09	**0.40**	−0.15	−0.16	−0.05
Pb	−0.01	0.04	−0.02	−0.01	−0.01	0.00	−0.03	0.03	−0.03	−0.07	0.00	0.00	−0.04	−0.03	−0.04	0.00
pH	**−0.39**	0.12	0.07	−0.21	**−0.42**	0.15	0.18	−0.22	−0.12	**0.34**	0.24	−0.21	0.11	**−0.47**	0.21	0.11
Shreve	−0.02	0.10	−0.07	0.00	−0.03	0.04	−0.13	0.07	**−0.38**	0.11	−0.06	0.01	−0.18	−0.07	0.24	**−0.59**
Slp	−0.18	**0.39**	−0.11	−0.13	−0.17	0.19	**−0.69**	**−0.32**	**0.54**	−0.01	−0.11	0.04	−0.02	−0.08	0.04	0.09
TALK	−0.07	0.19	−0.05	−0.17	−0.08	0.15	0.00	−0.15	**−0.30**	**−0.47**	**−0.32**	0.07	−0.23	−0.08	0.08	0.04
TH	−0.09	0.02	−0.07	0.16	−0.08	−0.01	−0.09	0.08	0.03	−0.04	0.12	−0.03	0.06	−0.14	−0.12	−0.23
TS	−0.01	0.04	−0.01	−0.03	−0.01	0.05	−0.01	−0.10	−0.14	−0.21	−0.16	−0.23	0.05	−0.02	0.04	0.04
TU	−0.01	−0.02	−0.02	0.00	−0.01	−0.01	0.01	0.02	0.00	0.03	−0.05	0.05	0.07	−0.03	−0.06	−0.07
WT	**−0.31**	**0.46**	**−0.38**	0.02	**−0.28**	0.15	**−0.28**	−0.08	**−0.41**	**0.31**	−0.09	−0.05	**0.25**	**−0.34**	0.20	**−0.44**

Integrity Index (HII), which encapsulates information on land use, riparian zone, stream-bed characteristics and stream-channel morphology, and this is in accordance with the results of this study that showed decreasing FHI-EPA values as degradation increased (C1: 145.7, C2: 132.5, C3: 111.3).

Plecoptera and Trichoptera exhibited trends similar to the one shown by Elmidae with respect to presence and total abundance. Trichoptera proved to be the second most diverse group, with 11 families. It is considered to be indicative of good WQ status (De Moor and Ivanov, 2008). However, this group is very diverse and this could result in lower performance in terms of presence, compared to Plecoptera and Elmidae, in adequately reflecting the WQ status (Dohet, 2002). Plecoptera are generally considered an appropriate bioindicator of good ecological status (Fochetti and Tierno De Figueroa, 2008), which is congruent with the current results.

All of the above suggest that elmids are a critical bio-indicator taxon and, in this study, their use in a biotic index, such as the ElmPT, resulted in increased performance over the EPT index (Fig. 5). EPT showed poor performance in assessing the WQ status in the basin, as indicated by the presence and abundance of Ephemeroptera (i.e., C1: 92.3%, C2: 96.6% and C3: 87.6% and C1: 4548 individuals, C2: 7586 individuals and C3: 20,025 individuals). Similar conclusions were drawn by Dos Santos et al. (2011), Nessimian et al. (2008) and Von Ellenrieder (2007). The lower EPT performance is probably caused by the presence of tolerant taxa of the genus *Baetodes* (family Baetidae). *Baetodes* individuals occurred in both clean and heavily disturbed sub-basins, such as Burgay, due to physiological adaptations (Baptista et al., 2006; Buss and Salles, 2007) that enable them tolerate eco-hydrological disturbances. Thus, the use of Ephemeroptera as a biotic index is not recommended (Btista et al., 2007).

In combination with other advantages of the ElmPT index (simple calculation and easy identification of Elmidae, large body size of Elmidae, Plecoptera and Trichoptera groups), it can be inferred that the ElmPT index is a powerful tool for biomonitoring for non-taxonomists.

Increased concentrations of faecal coliforms were found in C2 and C3 sites compared to C1, suggesting that C2 and C3 sites had higher organic inputs.

Most of the monitoring sites located within the Burgay and the Magdalena sub-basins (Fig. 7a) present wastewater effluents that often contain high amounts of organic pollution from sources such as domestic sewage and municipal stormwater drainage (Da Ros, 1995; Pauta-Calle and Chang-Gómez, 2014; Sotomayor et al., 2018); as such, they belong to WQ classes C3 and C2. Although faecal coliforms consume large amounts of oxygen, reducing the dissolved oxygen available for aquatic invertebrates (Varnosfaderany et al., 2010), the DO values recorded at the C3 sites were not systematically lower than those of the C1 and C2 sites. However, the lowest DO values show a decreasing trend from C1 to C3 sites, i.e., [5.4, 9.2] mg L^{-1} for C1, [4.4, 9.8] mg L^{-1} for C2 and [4.1, 8.5] mg L^{-1} for C3.

FHI-EPA was higher in C1 sites than C2 and C3 (Fig. 6e), in accordance with previous research, suggesting that habitat and biological diversity are closely linked (Barbour et al., 1999). Loss of riparian ecosystems turns into loss of instream quality because, due to resulting changes in hydrology, erosion and suspended solids increase, which may even be accentuated by an increment of concentrations of pollutants because of the lack of natural riparian ecosystems filtering (Tate et al., 2004; Sovell et al., 2000). Correspondingly, FC, $N\text{-}NH_4$, FL, TS, K and CL were higher at C3 monitoring sites where poor riparian ecosystems exist. Nevertheless, the contribution of these variables to the WQ classification, except for FC, was not significant. Furthermore, they are often correlated with EC (Singh et al., 2004). On average, EC was higher in C3 sites compared to C1 and C2 ones (Fig. 6c), in agreement with the results of Azrina et al. (2006), which also match the trend observed for $N\text{-}NH_4$, FL, TS, K and CL.

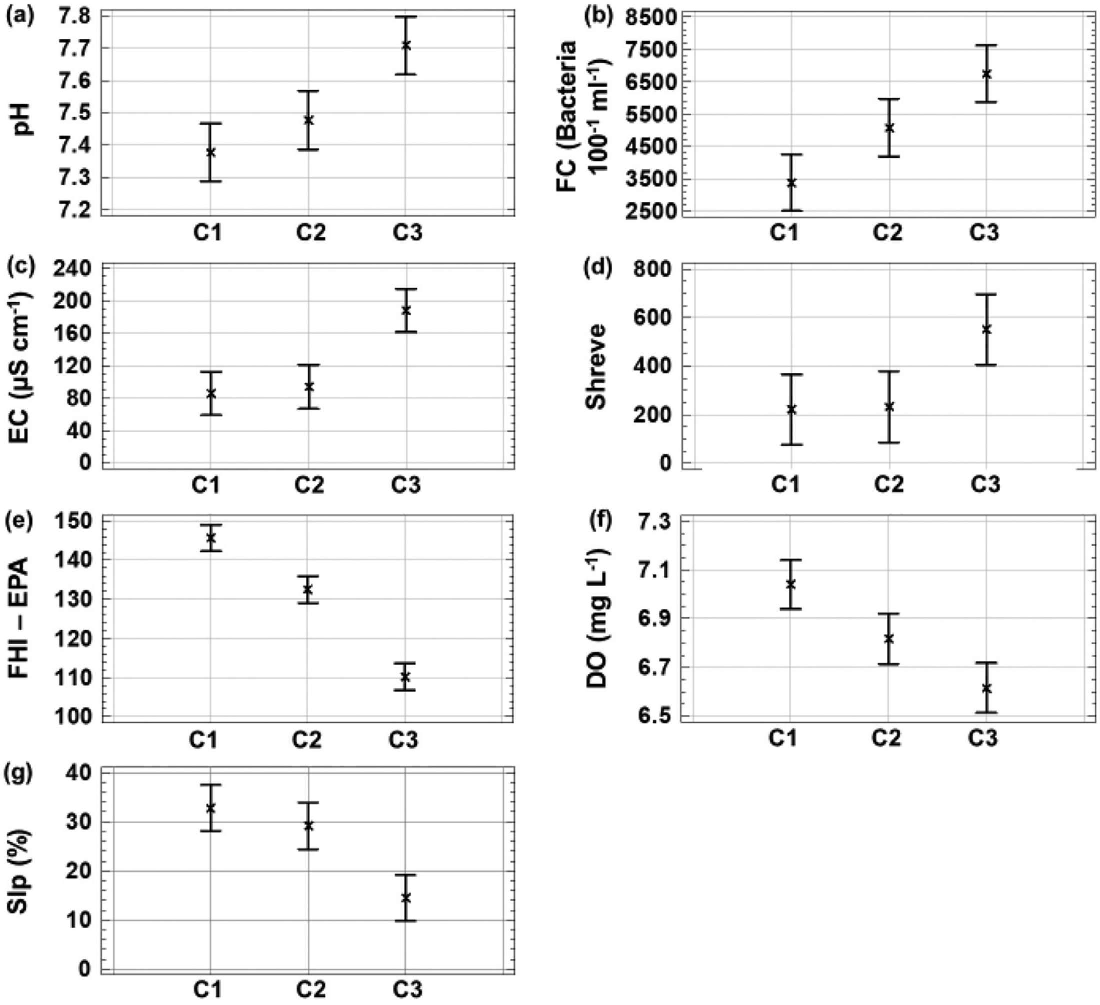


Fig. 6. Means and Fisher's-based intervals of the significant WQ descriptive variables for each biotic class (C1: less polluted, C2: moderately polluted, C3: highly polluted). FHI-EPA: Fluvial habitat integrity, DO: Dissolved oxygen, FC: Faecal coliforms, EC: Electric conductivity. Mean values are depicted with an × symbol.

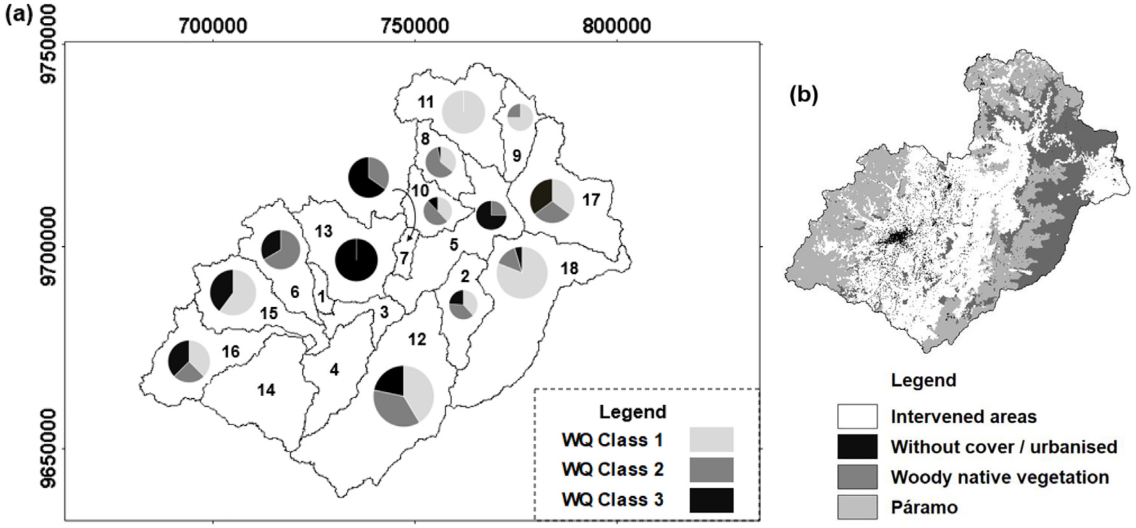


Fig. 7. (a) Water quality (WQ) classes at each sub-basin. 1 = Sidcay, 2 = Collay, 3 = Cuenca, 4 = Jadán, 5 = Paute, 6 = Machángara, 7 = Magdalena, 8 = Mazar, 9 = Juval, 10 = Pindilig, 11 = Pulpito, 12 = Santa Bárbara, 13 = Burgay, 14 = Tarqui, 15 = Tomebamba, 16 = Yanuncay, 17 = Paute bajo and 18 = Negro; (b) Land use/cover in the Paute River Basin (data from 2013).

Relationships between benthic macroinvertebrates and stream order were evidenced in this study and have been previously documented (Harrel and Dorris, 1968). Based on the PCA, stream order significantly influenced WQ classification in the basin (Fig. 6d). Generally, higher-order, lowland streams receive most of the groundwater flow, concentrating pollution and thus reducing WQ. In this study, however, most C3 sites are located in higher elevations in the Burgay and Magdalena sub-basins. Consequently, these sites have associated low stream order values, which would produce for the entire basin a lower average Shreve value associated to C3. This average Shreve value for the entire basin and associated to C3 turned, nevertheless, higher only because there exists one C3 monitoring station, located within the Paute sub-basin (Fig. 7a), with an associated larger order (i.e., 5760), which replicates six times in the database.

Slope was another variable significantly contributing to WQ classification, showing a decreasing trend from C1 to C3 sites (Fig. 6g). The influence of slope on macroinvertebrates has been previously reported by Roy et al. (2003). Despite the fact that most of the monitoring sites belonging to class C3 are located at higher elevation zones, the analysis of the DEM reveals that the slopes of these zones are generally flatter. Anthropogenic activities from human settlements tend to occur at flatter slopes. These suggest that the WQ associated with C3 monitoring sites, located at higher elevations, respond to anthropogenic perturbations, which is emphasised by the land use/coverage information.

Degraded areas are located in the mid-route of the Paute River at mid elevations. Western and eastern areas of higher and lower elevations, respectively, 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 River Basin is characterized by extended protected areas (Fig. 1a). However, the study showed that a large part of the basin has been altered due to anthropogenic activities and, consequently, its surface WQ has been degraded (Fig. 7a).

4.1. Future research

This study is the first assessment of the ecological requirements of elmids in the Paute River Basin. Considering the importance of this taxon for accurate biotic classifications, reported in this study, a deep knowledge of their ecological and taxonomic aspects is required. Ecological monitoring tools based on elmids distribution models in the basin could be a very effective management tool.

The results regarding the EPT index suggest that future research should focus on the taxonomical resolution level of Ephemeroptera and the distortion that the genus *Baetodes* may cause in the EPT index. Further, research associated with the allocation of correct scores to taxa is needed for the Paute River basin.

5. Conclusions

This study highlighted the effectiveness of the SIMCA supervised pattern recognition algorithm for the analysis and interpretation of complex datasets, which in this case resulted in identifying the ElmPT biotic index as the most appropriate for assessing the surface WQ to be monitored in the future in the basin. This would result in cost-effective but equally accurate monitoring schemes. The combination of multivariate statistics with GIS tools is also useful for assessing the congruency between WQ-biotic classification and field or land-use based information.

CRediT authorship contribution statement

Gonzalo Sotomayor: Conceptualization, Methodology, Software, Data curation, Formal analysis, Writing - original draft. **Henrietta Hampel:** Methodology, Writing - original draft, Writing - review & editing. **Raúl F. Vázquez:** Methodology, Writing - original draft, Writing - review & editing. **Peter L.M. Goethals:** Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors would like to express their gratitude to SENAGUA – DHS for making the row data accessible for the development of the current study. The preparation of the current manuscript took place in the framework of the development of the ERASMUS + project "Water Management and Climate Change in the Focus of International Master Courses (WATERMAS)" financed by the European "Education, Audiovisual and Culture Executive Agency (EACA)". This publication reflects only the authors' views; thereby, neither the European Union nor EACA is liable for any use that may be made of the information contained herein.

References

Armitage, P.D., Moss, D., Wright, J.F., Furse, M.T., 1983. The performance of a new biological water quality score system based on macroinvertebrates over a wide range of unpolluted running-water sites. Water Res. 17, 333–347. https://doi.org/10.1016/0043-1354(83)90188-4.

Astudillo, S., Astudillo, P., Cisneros, P., Coello, C., García, J., González, C., Pacheco, E., Rengel, A., Stoop, B., Van Noten, S., Wijffels, A., Zúñiga, A., 2010. Atlas de la cuenca del río Paute, PROMAS. ed. Consejo de gestión de aguas de la cuenca del Paute, Cuenca – Ecuador.r.

Azrina, M.Z., Yap, C.K., Rahim Ismail, A., Ismail, A., Tan, S.G., 2006. Anthropogenic impacts on the distribution and biodiversity of benthic macroinvertebrates and water quality of the Langat River, Peninsular Malaysia. Ecotoxicol. Environ. Saf. 64 (3), 337–347. https://doi.org/10.1016/j.ecoenv.2005.04.003.

Baker, A., 2005. Land use and water quality. In: Anderson, M.G. (Ed.), Encyclopedia of Hydrological Sciences. John Wiley & Sons, Inc., pp. 6. https://doi.org/10.1002/0470848944.

Balabin, R.M., Safieva, R.Z., Lomakina, E.I., 2010. Gasoline classification using near infrared (NIR) spectroscopy data: Comparison of multivariate techniques. Anal. Chim. Acta 671, 27–35. https://doi.org/10.1016/j.aca.2010.05.013.

Ballabio, D., 2015. A MATLAB toolbox for Principal Component Analysis and unsupervised exploration of data structure. Chemom. Intell. Lab. Syst. 149, 1–9. https://doi.org/10.1016/j.chemolab.2015.10.003.

Ballabio, D., Consonni, V., 2013. Classification tools in chemistry. Part 1: linear models. PLS-DA. Anal. Methods 5 (16), 3790–3798.

Ballabio, D., Grisoni, F., Todeschini, R., 2018. Multivariate comparison of classification performance measures. Chemom. Intell. Lab. Syst. 174, 33–44. https://doi.org/10.1016/j.chemolab.2017.12.004.

Baptista, D.F., Buss, D.F., Dias, L.G., Nessimian, J.L., Da Silva, E.R., De Moraes Neto, A.H.A., de Carvalho, S.N., De Oliveira, M.A., Andrade, L.R., 2006. Functional feeding groups of Brazilian Ephemeroptera nymphs: ultrastructure of mouthparts. Ann. Limnol. – Int. J. Limnol. 42, 87–96. https://doi.org/10.1051/limn/2006013.

Btista, D.F., Buss, D.F., Egler, M., Giovanelli, A., Silveira, M.P., Nessimian, J.L., 2007. A multimetric index based on benthic macroinvertebrates for evaluation of Atlantic Forest streams at Rio de Janeiro State, Brazil. Hydrobiologia 575, 83–94. https://doi.org/10.1007/s10750-006-0286-x.

Barbour, M.T., Gerritsen, J., Snyder, B.D., Stribling, J.B., 1999. Rapid bioassessment protocols for use in streams and wadeable rivers: periphyton, benthic macroinvertebrates and fish, Second Edi. ed. EPA 841-B-99-002. US Environmental Protection Agency, Office of Water., Washington, D.C.C.

Barnett, T., Pierce, D., Hidalgo, H., Bonfils, C., Santer, B., Das, T., Bala, G., Wood, A., Nozawa, T., Mirin, A., Cayan, D., Dettinger, M., 2008. Human-induced changes in the hydrology of the western United States. Science (80-). 319, 1080–1083. https://doi.org/10.1126/science.1152538.

Bartram, J., Ballance, R., 1996. Water Quality Monitoring - A Practical Guide to the Design and Implementation of Freshwater Quality Studies and Monitoring Programmes, First Edit. ed. United Nations Environment Programme/World Health Organization. https://doi.org/10.1159/000170272.

Blagus, R., Lusa, L., 2010. Class prediction for high-dimensional class-imbalanced data. BMC Bioinformatics 11, 523. https://doi.org/10.1038/4462.

Braun, B.M., Pires, M.M., Stenert, C., Maltchik, L., Kotzian, C.B., 2018. Effects of riparian vegetation width and substrate type on riffle beetle community structure. Entomol. Sci. 21, 66–75. https://doi.org/10.1111/ens.12283.

Brereton, R.G., 2007. Applied Chemometrics for Scientists. John Wiley Sons, Ltd.

Brown, H., 1987. Biology of riffle beetles. Annu. Rev. Entomol. 32, 253–273. https://doi.org/10.1146/annurev.ento.32.1.253.

Buss, D.F., Salles, F.F., 2007. Using Baetidae species as biological indicators of environmental degradation in a Brazilian river basin. Environ. Monit. Assess. 130, 365–372. https://doi.org/10.1007/s10661-006-9403-6.

Carrera, C., Fierro, K., 2001. Manual de monitoreo: los macroinvertebrados acuáticos como indicadores de la calidad del agua. EcoCiencia, Quito – Ecuador.

Chatfield, C., Collins, A.J., 1980. Introduction to Multivariate Analysis, 1st ed. Chapman and Hall. https://doi.org/10.1007/978-1-4899-3184-9.

Chawla, N.V., 2009. Data Mining for Imbalanced Datasets: An Overview, in: Maimon, O., Rokach, L. (Eds.), Data Mining and Knowledge Discovery Handbook. Springer Science + Business Media, pp. 875–886. https://doi.org/10.1007/978-0-387-09823-4_45.

Consejo Nacional de Electricidad (CONELEC), 2011. Estadística del sector eléctrico ecuatoriano - Folleto resumen.

Consejo Nacional de Electricidad (CONELEC), 2009. Plan maestro de electrificación del Ecuador.

Consejo Nacional de Electricidad (CONELEC), 2011. Estadística del sector eléctrico ecuatoriano - Folleto resumen.

De Moor, F.C., Ivanov, V.D., 2008. Global diversity of caddisflies (Trichoptera: Insecta) in freshwater. Hydrobiologia 595, 393–407. https://doi.org/10.1007/s10750-007-9113-2.

Dodge, Y., 2008. The Concise Encyclopedia of Statistics. Springer Science + Business Media, LLC. https://doi.org/10.1117/12.2084301.

Dohet, A., 2002. Are caddisflies an ideal group for the biological assessment of water wuality in streams? Nov. Suppl. Entomol. (Proceedings 10th Int. Symp. Trichoptera) 507–520.

Dos Santos, D.A., Molineri, C., Reynaga, M.C., Basualdo, C., 2011. Which index is the best to assess stream health? Ecol. Indic. 11, 582–589. https://doi.org/10.1016/j.ecolind.2010.08.004.

Einax, J.W., Zwanziger, H.W., Geiss, S., 1997. Chemometrics in environmental analysis. Wiley-VCH Verlag GmbH. https://doi.org/10.1002/352760216X.

Elliott, J.M., 2008. The Ecology of Riffle Beetles (Coleoptera: Elmidae). Freshw. Rev. 1, 189–203. https://doi.org/10.1608/FRJ-1.2.4.

Feio, M.J., Ferreira, J., Buffagni, A., Erba, S., Dörflinger, G., Ferréol, M., Munné, A., Prat, N., Tziortzis, I., Urbanič, G., 2014. Comparability of ecological quality boundaries in the Mediterranean basin using freshwater benthic invertebrates. Statistical options and implications. Sci. Total Environ. 476–477, 777–784. https://doi.org/10.1016/j.scitotenv.2013.07.085.

Fochetti, R., Tierno De Figueroa, J.M., 2008. Global diversity of stoneflies (Plecoptera; Insecta) in freshwater. Hydrobiologia 595, 365–377. https://doi.org/10.1007/s10750-007-9031-3.

Forman, G., 2003. An Extensive Empirical Study of Feature Selection Metrics for Text Classification. J. Mach. Learn. Res. 3, 1289–1305. https://doi.org/10.1162/153244303322753670.

Frank, I.E., Todeschini, R., 1994. The Data Analysis Handbook. B.V., Elsevier Science. https://doi.org/10.1016/S0922-3487(08)70048-0.

Garcia-Criado, F., Fernandez-Alaez, M., 2001. Hydraenidae and Elmidae assemblages (Coleoptera) from a Spanish river basin: good indicators of coal mining pollution? Arch. Fur Hydrobiol. 150, 641–660. https://doi.org/10.1127/archiv-hydrobiol/150/2001/641.

Garcia Criado, F., Fernandez Alaez, M., 1995. Aquatic Coleóptera (Hydraenidae and Elmidae) as indicators of the chemical characteristics of water in the Orbigo River basin (N-W Spain). Ann. Limnol. 31, 185–199. https://doi.org/10.1051/limn/1995017.

Hand, D.J., 2012. Assessing the Performance of Classification Methods. Int. Stat. Rev. 80, 400–414. https://doi.org/10.1111/j.1751-5823.2012.00183.x.

Hanselman, D., Littlefield, B., 2012. Mastering MATLAB®. Pearson Education Limited.

Harper, D., Zalewski, M., Pacini, N., 2008. Ecohydrology: processes, models and case studies - An approach to the sustainable management of water resources. International, CAB. https://doi.org/10.1111/j.1365-2427.2010.02451.x.

Harrel, R.C., Dorris, T.C., 1968. Stream order, morphometry, physico-chemical conditions, and community structure of benthic macroinvertebrates in an intermittent stream system. Am. Midl. Nat. 80, 220–251.

Herman, M.R., Nejadhashemi, A.P., 2015. A review of macroinvertebrate- and fish-based stream health indices. Ecohydrol. Hydrobiol. 15, 53–67. https://doi.org/10.1016/j.ecohyd.2015.04.001.

Hotelling, H., 1936. Relations between two sets of variates. Biometrika 28, 321–377.

Jacobsen, D., Schultz, R., Encalada, A., 1997. Structure and diversity of stream invertebrate assemblages: the influence of temperature with altitude and latitude. Freshw. Biol. 38, 247–261. https://doi.org/10.1046/j.1365-2427.1997.00210.x.

Kannel, P.R., Lee, S., Kanel, S.R., Khan, S.P., 2007. Chemometric application in classification and assessment of monitoring locations of an urban river system. Anal. Chim. Acta 582, 390–399. https://doi.org/10.1016/j.aca.2006.09.006.

Kruskal, W.H., Wallis, W.A., 1952. Use of ranks in one-criterion variance analysis. J. Am. Stat. Assoc. 47, 583–621.

Lavine, B.K., Rayens, W.S., 2009. Classification: Basic Concepts, in: Brown, S.D., Tauler, R., Walczak, B. (Eds.), Comprehensive Chemometrics. Elsevier B.V., pp. 507–515.

Lenat, D.R., 1988. Macroinvertebrates water quality assessment of streams using a qualitative collection method for benthic macroinvertebrates. J. North Am. Benthol. Soc. 7, 222–233. https://doi.org/10.2307/1467422.

Lischeid, G., Bittersohl, J., 2008. Tracing biogeochemical processes in stream water and groundwater using non-linear statistics. J. Hydrol. 357, 11–28. https://doi.org/10.1016/j.jhydrol.2008.03.013.

Loinaz, M., 2012. Integrated ecohydrological modeling at the catchment scale. Technical University of Denmark.

Ministerio del Ambiente de la República del Ecuador (MAE), 2013. Sistema de clasificación de ecosistemas del Ecuador continental. Subsecretaría de Patrimonio Natural - Proyecto mapa de vegetación, Quito.

Martina, F., Beccuti, M., Balbo, G., Cordero, F., 2017. Peculiar genes selection: A new features selection method to improve classification performances in imbalanced data sets. PLoS One 12, 1–18. https://doi.org/10.1371/journal.pone.0177475.

Massart, D.L., Vandeginste, B.G.M., Deming, S.N., Michotte, Y., Kaufman, L., 1988. DATA HANDLING IN SCIENCE AND TECHNOLOGY - VOLUME 2 - Chemometrics: a textbook. Elsevier Science B.V.

Mevik, B., Wehrens, R., 2015. Introduction to the pls Package. Help Sect. "pls" Packag. RStudio Softw. 1–23.

Miserendino, M.L., Archangelsky, M., 2006. Aquatic coleoptera distribution and environmental relationships in a large Patagonian river. Int. Rev. Hydrobiol. 91, 423–437. https://doi.org/10.1002/iroh.200510854.

Miserendino, M.L., Pizzolon, L.A., 2000. Macroinvertebrates of a fluvial system in Patagonia: altitudinal zonation and functional structure. Arch. Fur Hydrobiol. 150, 55–83. https://doi.org/10.1127/archiv-hydrobiol/150/2000/55.

Muenz, T.K., Golladay, S.W., Vellidis, G., Smith, L.L., 2006. Stream buffer effectiveness in an agriculturally influenced area. Southwestern Georgia. J. Environ. Qual. 35, 1924. https://doi.org/10.2134/jeq2005.0456.

Nessimian, J.L., Venticinque, E.M., Zuanon, J., De Marco, P., Gordo, M., Fidelis, L., D'arc Batista, J., Juen, L., 2008. Land use, habitat integrity, and aquatic insect assemblages in Central Amazonian streams. Hydrobiologia 614, 117–131. https://doi.org/10.1007/s10750-008-9441-x.

Ode, P.R., Rehn, A.C., May, J.T., 2005. A quantitative tool for assessing the integrity of southern coastal California streams. Environ. Manage. 35, 493–504. https://doi.org/10.1007/s00267-004-0035-8.

Ouyang, Y., 2005. Evaluation of river water quality monitoring stations by principal component analysis. Water Res. 39, 2621–2635. https://doi.org/10.1016/j.watres.2005.04.024.

Pauta Calle, G., Chang Gómez, J., 2014. Indices de calidad del agua de fuentes superficiales y aspectos toxicológicos, evaluación del Río Burgay. MASKANA, I + D + ingeniería 165–176.

Peres-Neto, P.R., Jackson, D.A., Somers, K.M., 2003. Giving meaningful interpretation to ordination axes: Assessing loading significance in principal component analysis. Ecology 84, 2347–2363. https://doi.org/10.1890/00-0634.

Pomerantsev, A.L., 2014. Chemometrics in excel, 1st ed. John Wiley & Sons Inc.

Rawlings, J.O., Pantula, S.G., Dickey, D.A., 1998. Applied Regression Analysis: A Research Tool, Second. Springer-Verlag, New York Inc.

Ríos-Touma, B., Acosta, R., Prat, N., 2014. The Andean biotic index (ABI): Revised tolerance to pollution values for macroinvertebrate families and index performance evaluation. Rev. Biol. Trop. 62, 249–273. https://doi.org/10.15517/rbt.v62i0.15791.

Roldán, G., 2003. Bioindicación de la calidad del agua en colombia: Propuesta para el uso del método BMWP Col. Universidad de Antioquia.

Roy, A.H., Rosemond, A.D., Paul, M.J., Leigh, D.S., Wallace, J.B., 2003. Stream macroinvertebrate response to catchment. Freshw. Biol. 48, 329–346.

Secretaría Nacional del Agua (SENAGUA), 2016. Plan de monitoreo de calidad del agua de los sistemas de agua de abastecimiento público de Portoviejo, Manta, Chone, Pedernales, Jama, Bahia de Caráquez, San Vicente, Canoa, Calceta, Junín, Tosagua, Flavio Alfaro y Muisne.

Shapiro, S.S., Wilk, M.B., 1965. An Analysis of Variance Test for Normality (Complete Samples). Biometrika Trust 52, 591–611. https://doi.org/10.1093/biomet/52.3-4.591.

Silva-Palacios, D., Ferri, C., Ramírez-Quintana, M.J., 2017. Improving performance of multiclass classification by inducing class hierarchies. Procedia Comput. Sci. 108, 1692–1701. https://doi.org/10.1016/j.procs.2017.05.218.

Simeonov, V., Stratis, J.A., Samara, C., Zachariadis, G., Voutsa, D., Anthemidis, A., Sofoniou, M., Kouimtzis, T., 2003. Assessment of the surface water quality in Northern Greece. Water Res. 37, 4119–4124. https://doi.org/10.1016/S0043-1354(03)00398-1.

Singh, K.P., Malik, A., Mohan, D., Sinha, S., 2004. Multivariate statistical techniques for the evaluation of spatial and temporal variations in water quality of Gomti River (India) - a case study. Water Res 38, 3980–3992. https://doi.org/10.1016/j.watres.2004.06.011.

Sotomayor, G., 2016. Evaluación de la calidad de las aguas superficiales mediante técnicas de estadística multivariante: Un estudio de caso en la cuenca del Río Paute, al sur de. Ecuador. Universidad Nacional de La Plata.

Sotomayor, G., Hampel, H., Vázquez, R.F., 2018. Water quality assessment with emphasis in parameter optimisation using pattern recognition methods and genetic algorithm. Water Res. 130, 353–362. https://doi.org/10.1016/j.watres.2017.12.010.

Sovell, L.A., Vondracek, B., Frost, J.A., Mumford, K.G., 2000. Impacts of rotational grazing and riparian buffers on physicochemical and biological characteristics of Southeastern Minnesota, USA, streams. Environ. Manage. 26, 629–641. https://doi.org/10.1007/s002670010121.

Tate, K.W., Pereira, M.D.G.C., Atwill, E.R., 2004. Efficacy of vegetated buffer strips for retaining Cryptosporidium parvum. J. Environ. Qual. 33, 2243–2251. https://doi.org/10.2134/jeq2004.2243.

Theodoropoulos, C., Vourka, A., Skoulikidis, N., Rutschmann, P., Stamou, A., 2018. Evaluating the performance of habitat models for predicting the environmental flow requirements of benthic macroinvertebrates. Journal of Ecohydraulics. 3, 30–44. https://doi.org/10.1080/24705357.2018.1440360.

Varnosfaderany, M.N., Ebrahimi, E., Mirghaffary, N., Safyanian, A., 2010. Biological assessment of the Zayandeh Rud River, Iran, using benthic macroinvertebrates. Limnologica 40, 226–232. https://doi.org/10.1016/j.limno.2009.10.002.

Von Ellenrieder, N., 2007. Composition and structure of aquatic insect assemblages of

Yungas mountain cloud forest streams in NW Argentina. Rev. de la Soc. Entomológica Argentina 66, 57–76.

Walley, W.J., Hawkes, H.A., 1996. A computer-based reappraisal of the biological monitoring working party scores using data from the 1990 river quality survey of England and Wales. Water Res. 30, 2086–2094. https://doi.org/10.1016/0043-1354(96)00013-9.

Wold, S., 1976. Pattern recognition by means of disjoint principal components models. Pattern Recognit. 8, 127–139. https://doi.org/10.1016/0031-3203(76)90014-5.

Lavine, B.K., Rayens, W.S., 2009. Classification: Basic Concepts, in: Brown, S.D., Tauler, R., Walczak, B. (Eds.), Comprehensive Chemometrics. Elsevier B.V., pp. 507–515.

Yap, B.W., Sim, C.H., 2011. Comparisons of various types of normality tests. J. Stat. Comput. Simul. 81 (12), 2141–2155. https://doi.org/10.1080/00949655.2010.520163.

Zupan, J. (Ed.), 1990. PCs for chemists. Elsevier Science Publishers B.V., Amsterdam. https://doi.org/10.1016/0097-8485(91)80009-B.