Application of classification trees to determine biological and chemical indicators for river assessment: case study in the Chaguana watershed (Ecuador)

L. Dominguez-Granda, K. Lock and P. L. M. Goethals

ABSTRACT

Benthic macroinvertebrates were sampled in the Chaguana river basin in SW Ecuador in March (wet season) and September (dry season) of 2005 and 2006. Aquatic insects dominated the macrobenthos, with Trichoptera, Diptera, Ephemeroptera, Hemiptera and Odonata being the orders with the highest diversity and Ephemeroptera and Diptera being most abundant. No systematic differences in richness and abundance were observed between dry and wet seasons, which is in agreement with the literature. It is concluded that, in the neotropics, macroinvertebrates can probably be sampled for water quality assessments during the whole year: however, sampling soon after spates should be avoided. Using multivariate analysis, stations could be clustered into three groups based on their macroinvertebrate community composition: sites with low, intermediate and high human impact. Classification trees indicated that stations with low human impact had low conductivities, while stations with high conductivities were characterised as highly impacted if the dissolved oxygen concentration was low and intermediately impacted if the dissolved oxygen concentration was high. Classification trees also indicated that Leptophlebiidae (Ephemeroptera) were characteristic for sites with low impact; in sites with intermediate impact, this family was absent but Hydropsychidae (Trichoptera) were present; when both families were absent, impact was high.

Key words | aquatic insects, community analysis, river basin, water quality assessment

L. Dominguez-Granda
Escuela Superior Politécnica del Litoral,
Instituto de Ciencias Químicas y Ambientales,
Campus Gustavo Galindo, Km. 30.5 Vía Perimetral,
PO Box 09-01-5863,
Guayaquil,
Ecuador

K. Lock (corresponding author)
P. L. M. Goethals
Ghent University,
Laboratory of Environmental Toxicology
and Aquatic Ecology,
J. Plateaustraat 22,
B-9000 Gent,
Belgium
E-mail: *Koen.Lock@UGent.be*

INTRODUCTION

Seasonality is known to play an important role in structuring the macroinvertebrate community in temperate watercourses and can influence the outcome of bioassessment methods (Linke *et al.* 1999; Clarke *et al.* 2002). Although several studies have already been performed in order to increase the knowledge of seasonal patterns of benthic macroinvertebrates in the neotropics (Flecker & Feifarek 1994; Jacobsen 1998; Jacobsen & Encalada 1998; Melo & Froehlich 2001; Carrera Burneo & Gunkel 2003, Buss *et al.* 2004), further studies exploring the influence of seasonal variation on existing and forthcoming bioassessment methods are required.

From a spatial viewpoint, latitude and altitude can both influence aquatic invertebrate communities in the tropics. The majority of the reviewed assessment methods for neotropical areas are generally based on a biotic approach, in which the confidence relies on its application within a delimitated geographical area. Therefore, regions should be defined in which the method could be confidently applied to similar aquatic macroinvertebrate communities. These geographical areas, called ecoregions, provide the basis for the implementation of biological methods with score systems related to reference conditions (Gerritsen *et al.* 2000).

Artificial intelligence has played a crucial role in modelling the distribution of organisms as a function of the abiotic environment, often called habitat suitability modelling. A range of modelling techniques have been applied for the

doi: 10.2166/hydro.2010.082

assessment of running waters based on the distribution of macroinvertebrates. Artificial neural networks (Dedecker *et al.* 2007), fuzzy logic (Adriaenssens *et al.* 2004a; Mouton *et al.* 2009), classification trees (Dzeroski *et al.* 2000; Dakou *et al.* 2007), Bayesian belief networks (Adriaenssens *et al.* 2004b) and support vector machines (Hoang *et al.* 2010) have proven to have a high potential in macroinvertebrate habitat suitability modelling. Other promising techniques are, for example, evolutionary polynomial regression (Elshorbagy & El-Baroudy 2009; Giustolisi & Savic 2009; Savic *et al.* 2009) and cellular automata (Chen *et al.* 2009). However, in the present study, classification trees were used because this technique is applicable for relatively small datasets and because their transparency makes them suitable as a practical tool for decision support by water quality managers.

While Walley & Dzeroski (1995) already applied data mining techniques for biological river quality assessment, Dzeroski *et al.* (1997) were among the first to describe applications of classification trees in river community analysis. These included the biological classification of British rivers based on biological data, the analysis of the influence of physical and chemical parameters on selected bioindicator organisms in Slovenian rivers and the biological classification of Slovenian rivers based on physical and chemical parameters as well as bioindicator data. Blockeel *et al.* (1999) made simultaneous predictions of multiple physical–chemical properties of the water from its biological properties using a single decision tree and also predictions of past physical–chemical properties of the river water from its current biological properties. Dzeroski *et al.* (2000) predicted physical–chemical variables on the basis of biological data; taxa that occurred in many trees were considered as useful indicator taxa. Dakou *et al.* (2007) induced decision trees to predict the habitat suitability of six macroinvertebrate taxa in the river Axios (Northern Greece). predicted the presence of several macroinvertebrate taxa in Flanders based on a selection of environmental variables. Turak & Koop (2008) defined river types for abiotic features, riffle zones, fish assemblages and macroinvertebrate assemblages from river edges on the basis of reference sites in New South Wales (Australia).

Thus far, most papers dealing with classification trees were mainly focused on the optimisation of model performance, while in the present paper, classification trees were used as a practical tool: indicator taxa and key environmental parameters were derived that can be used for water quality assessment. In our case study in the neotropical Chaguana basin, the water quality was predicted on the basis of the occurring macroinvertebrates and the measured environmental parameters, taking spatial and temporal variability into account.

METHODS

Study area

The Chaguana river basin is located in El Oro Province in SW Ecuador (Figure 1). The basin drains an area of approximately 32,000 ha, flowing from the occidental slope of the W Andes to a larger watershed called the Pagua river basin. The Chaguana river basin drains a mountainous area of difficult access, with headwaters located at about 2900 m above sea level. The Chaguana basin, which is inhabited by about 7600 inhabitants, is predominantly an agricultural basin, with agricultural activities covering 55% of the watershed surface, mostly for intensive banana farming in the lowlands (Matamoros *et al.* 2005). Near-natural conditions (humid forests and brushes, mangroves and uncultivated land) cover 37% of the land, while activities performed in the basin are shrimp farming, human settlements and, in recent years, gold mining (Matamoros *et al.* 2005).

Three gauging stations in the basin registered monthly water flow measurements from 1979 to 1982. Based on the registered data, Matamoros *et al.* (2005) reported median values per month at the Chaguana village (which corresponds to our sampling site 28) between 0.4 and 4.4 m^3/s, with April and November being the months with the highest flows. During four sampling campaigns in 2001 and 2002, Matamoros *et al.* (2005) measured biological oxygen demand (BOD) values that were mostly below 4 mg/L, suggesting that organic matter discharges into the river were not the main concern for the ecosystem at the moment of sampling. However, a wide array of pesticides are applied on banana crops for pest control, of which propiconazole, imazalil, thiabendazole, glyphosate and tridemorph are the most frequently used (Matamoros *et al.* 2005). However, because their residence time in the basin is estimated to be only approximately 24 h (Matamoros *et al.* 2005), it is difficult to measure

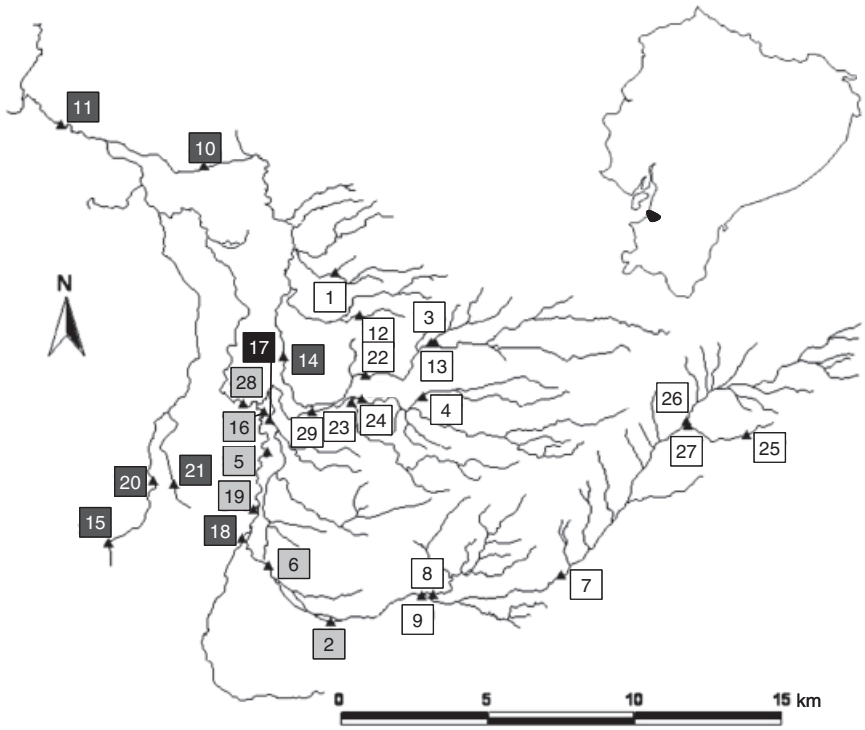

Figure 1 | Location of the sampling stations in the Chaguana river basin, with indication of the three identified clusters (black: high impact; grey: intermediate impact: white: low impact).

peak concentrations and to assess their effect on macroinvertebrates. Despite pesticides probably strongly affecting macroinvertebrates in the Chaguana watershed, they were not measured during the present study because pesticide concentrations at one point in time are not representative of their potential effect. More detailed studies are required to assess the impact of pesticides in the Chaguana watershed. In addition to pesticides, there are also effluents from banana packing plants, discharges of canals draining the crops and wastewater discharges. Construction of artificial embankments to obtain a water level suitable for irrigation, removal of river bed sediments, either for deeper channels suitable for agricultural purposes or for the extraction of materials for construction, clearance of riparian vegetation and artificial embankment for flood alleviation are all factors that affect the river structure in the Chaguana river basin.

Sampling

Within the Chaguana river basin, 29 sampling sites, distributed among different land use categories, were selected to be surveyed in March (wet season) and September (dry season) of 2005 and 2006. However, in 2006, access to some stations was denied as a result of land use conflicts. In total, 104 samples were taken: 29 during the dry season of 2005, 29 during the wet season of 2005, 24 during the dry season of 2006 and 22 during the wet season of 2006. Representative measurements of width and depth as well as current velocity were taken in each station. Dissolved oxygen, oxygen saturation and pH were measured with an OAKTON®35632, while conductivity and water temperature were measured with a YSI®30. Macroinvertebrate samples were always collected by the same operator by means of a standard hand net consisting of a metal frame holding a conical net (20×30 cm, 300 μm mesh size). Sampling duration was 3 min active sampling in 2005; in 2006 sampling duration was increased to 10 min active sampling. Organisms were collected from the different habitats present at the sampling site. Riffle habitats were sampled by holding the net downstream while the operator disturbed the substratum by kicking directly in front of the net opening. Stream edge habitats were sampled by vigorously sweeping along the stream margins disturbing bottom and bank substrata. The objective of the sampling was to collect the most representative diversity of macroinvertebrates at the

site examined. After separation, macroinvertebrates were identified under a stereomicroscope. The taxonomical knowledge of stream fauna in Ecuadorian streams is still scarce; therefore aquatic insects were identified at family level with the available literature containing identification keys and descriptions of the riverine fauna of the region (Roldán 1988; Domínguez *et al.* 1994; Fernández & Dominguez 2001). Non-insects were mostly identified at higher taxonomic levels.

Statistics

Because the number of sampling sites as well as the sampling effort differed between 2005 and 2006, both years were analysed separately. The *t*-test for dependent samples was used to identify differences between the wet and dry seasons. The stations were classified into clusters according to the composition of the macrobenthic community, using the classification program TWINSPAN (Two-Way INdicator SPecies ANalysis) (Hill 1979). TWINSPAN also yields indicator species characterising the various assemblages. In order to take the abundance of the macroinvertebrates into account, the five cutlevels used in this analysis were 1, 2, 4, 8 and 16 in 2005 and 1, 3, 10, 33 and 100 in 2006, when the sampling time was increased. In this way, abundant taxa will occur at a higher cutlevel, which increases their weight in the analysis. To check the stability of the TWINSPAN results, the Canonical Correspondence Analysis (CCA) option from the program package CANOCO (Ter Braak & Smilauer 2002) was applied. Prior to CCA analysis, all data were log transformed, except pH, which is already on a log scale. Since dissolved oxygen concentration and oxygen saturation values were strongly correlated, only the former was retained for analysis. Based on a training set of 104 samples, classification trees were built using Weka and applying the J48 algorithm and a 10-fold cross-validation (Witten & Frank 2005). To evaluate the performance of the classification trees, the percentage of correctly classified instances (% CCI) and kappa statistics (K) were used (Witten & Frank 2005). CCI is calculated as the percentage of the true predictions: however, K is recommended to be used as a more reliable performance measure because it is a derived statistic that measures the proportion of all possible outcomes that are predicted correctly by a model after accounting for chance predictions (Cohen 1960). When predicting two states based on field data (i.e. presence/absence), models with CCI higher than 70% and K higher than 0.4 are considered reliable (D'heygere *et al.* 2006; Dakou *et al.* 2007), while Landis & Koch (1977) indicated the degree of agreement that exists when Cohen's kappa is found to be in various ranges such as <0 (poor); 0–0.2 (slight); 0.2–0.4 (fair); 0.4–0.6 (moderate); 0.6–0.8 (substantial) and 0.8–1 (almost perfect). However, predicting more than two states will, of course, be more difficult.

RESULTS

In 2005, significantly higher values were observed during the wet season than during the dry season for water temperature ($P<0.001$), water depth ($P<0.001$) and current velocity ($P=0.0040$). Also in 2006, significantly higher values were measured during the wet season than during the dry season for water temperature ($P<0.001$), water depth ($P=0.0013$) and current velocity ($P<0.001$). In addition, the stream width was significantly higher in the wet season of 2006 than in the dry season of 2006 ($P=0.0021$), while no significant difference could be observed between seasons in 2005 ($P=0.13$). The dissolved oxygen concentration was significantly higher during the wet season of 2005 than during the dry season of 2005 ($P<0.001$), which was also the case in 2006 ($P<0.001$). However, no significant difference in oxygen saturation was observed between seasons in 2005 ($P=0.93$) and in 2006 ($P=0.53$). Conductivity ($P<0.001$) and pH ($P<0.001$) were significantly higher during the dry season of 2005 than during the wet season of 2005, but no significant differences in conductivity ($P=0.86$) and pH ($P=0.20$) were observed between seasons in 2006.

The macroinvertebrate community was dominated by aquatic insects: insects constituted 89% of the 28,469 collected individuals. Trichoptera (9 families), Diptera and Ephemeroptera (both 8 families), Hemiptera and Odonata (both 7 families) were the orders with the highest diversity. The most abundant taxa were the Ephemeroptera families Baetidae and Leptohyphidae (24% and 18% of the collected animals, respectively), the Diptera family Chironomidae (11% of all organisms) and the Trichoptera family Hydropsychidae (9% of all organisms). During the wet season of 2005 and 2006, Ephemeroptera was the dominant taxon in

most of the surveyed stations, while sites 11, 15, 20 and 21 were mainly dominated by non-insects and Diptera were dominant in station 6 in 2005 and in stations 18 and 19 in 2006 (Figure 2). During the dry season, Ephemeroptera were the dominant taxon in fewer stations and in most stations no clear dominance of any given taxon was observed: however,

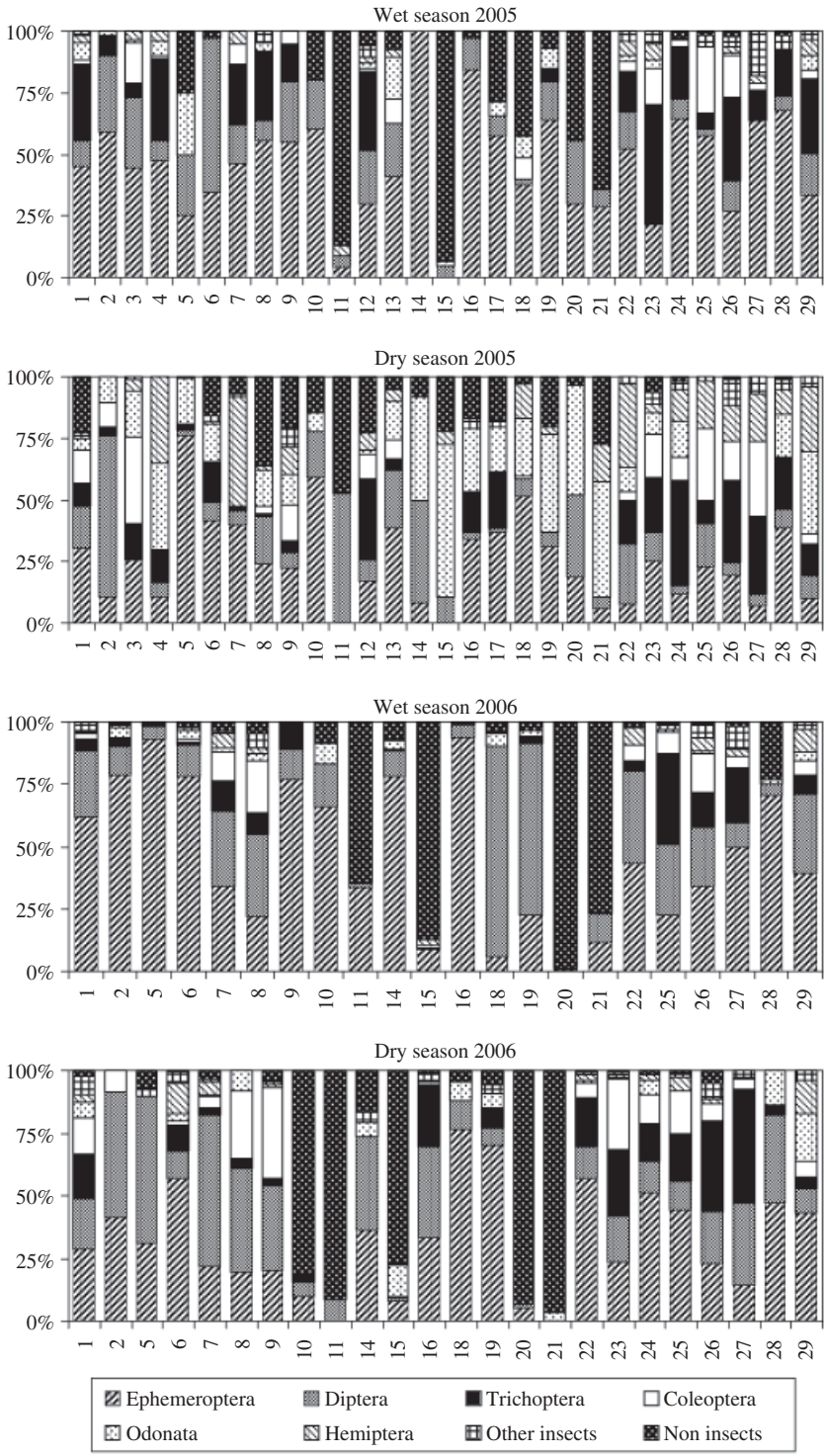

Figure 2 | Relative abundances of the main taxa encountered in the sampling stations during the four sampling campaigns.

in 2005 Diptera dominated in stations 2 and 11, while in 2006 Diptera dominated in stations 2, 5 and 7 and non-insects in stations 10, 11, 15, 20 and 21 (Figure 2).

Odonata ($P<0.001$) and Hemiptera ($P=0.014$) were significantly more abundant during the dry season than during the wet season in 2005, while Ephemeroptera ($P=0.015$) and Diptera ($P=0.079$) were significantly more abundant during the wet season than during the dry season in 2006. The number of taxa was significantly higher during the dry season than during the wet season in 2005 ($P<0.001$): however, slightly higher values were observed in the wet season than in the dry season in 2006 ($P=0.10$). The total abundance was slightly higher during the dry season than during the wet season in 2005 ($P=0.56$): however, significantly higher values were observed during the wet season than during the dry season in 2006 ($P=0.017$). No systematic differences in richness and abundance were therefore observed between dry and wet seasons of 2005 and 2006.

The results of the TWINSPAN for both 2005 and 2006 indicated that, based on the macroinvertebrate composition, the stations with a poor water quality are separated from those with a good water quality in the first division. Stations with a good water quality were characterised by a higher taxa richness and the presence of more pollution-sensitive taxa, whereas stations with a poor water quality were characterised by a lower taxa richness and the presence of more pollution-tolerant taxa. In 2005, indicator taxa for the sites with a good water quality are Hydropsychyidae, Leptophlebiidae and Naucoridae and an indicator taxon for sites with a poor water quality is Pleuroceridae. In 2006, Hydropsychidae and Elmidae are indicator taxa for sites with a good water quality. On the basis of the TWINSPAN results, the stations are separated into three groups: (1) a group containing stations with a high impact, which had a poor water quality during all four sampling campaigns, (2) a group containing stations with a low impact, which had a good water quality during all sampling campaigns, and (3) a group containing stations with an intermediate impact that are sometimes classified as stations with a poor water quality and sometimes as stations with a good water quality, depending on the season and the sampling year. The three identified clusters are indicated on the map of the Chaguana river basin (Figure 1).

The Canonical Correspondence Analysis (CCA) seems to justify this division into three clusters which are separated along the first axis (eigenvalue 0.32 in 2005 and 0.36 in 2006) (Figure 3). The second axis explained less of the variance (eigenvalue 0.18 in 2005 and 0.16 in 2006). For both years, stations with a high human impact are characterised by high conductivities and low oxygen concentrations, while stations with a low human impact are characterised by low conductivities and high oxygen concentrations.

A classification tree was developed that predicted the water quality on the basis of the occurring macroinvertebrates (CCI = 63%, $K=0.41$) (Figure 4). Sites with a low impact were characterised by the presence of the mayfly family Leptophebiidae; sites where Leptophlebiidae were absent but the caddisfly family Hydropsychidae was present had an intermediate impact; in sites with a high impact, both families were absent. The confusion matrix indicated that poor and good water qualities were better predicted than moderate water quality and that only on a few occasions was a good water quality predicted when the observed water quality was bad and vice versa (Table 1). Based on this classification tree, the water quality of the samples of 2005 was somewhat better predicted than for those of 2006: the water quality was predicted accurately in 76% and 62% of the samples, respectively (Table 1). An unpruned classification resulted in a more complicated tree with 24 leaves: however, such a tree did not perform better (CCI = 61%, $K=0.38$).

Another classification tree was developed for predicting the water quality based on the environmental parameters (CCI = 71%, $K=0.52$) (Figure 5). Sites with a low impact had a low conductivity, while sites with a high conductivity were highly impacted when the oxygen concentration was low and intermediately impacted when the oxygen concentration was high. The confusion matrix indicated that this tree tended to overpredict the water quality: only on two occasions was the water quality underpredicted, while the water quality was overpredicted on 23 occasions (Table 1). The developed classification tree had a similar performance for the samples of 2005 and 2006: the water quality was predicted accurately in 74% and 77% of the samples, respectively (Table 1). An unpruned classification tree resulted in a more complicated tree with 10 leaves: however, such a tree performed only marginally better (CCI = 73%, $K=0.58$). When a classification tree was developed based on the occurring

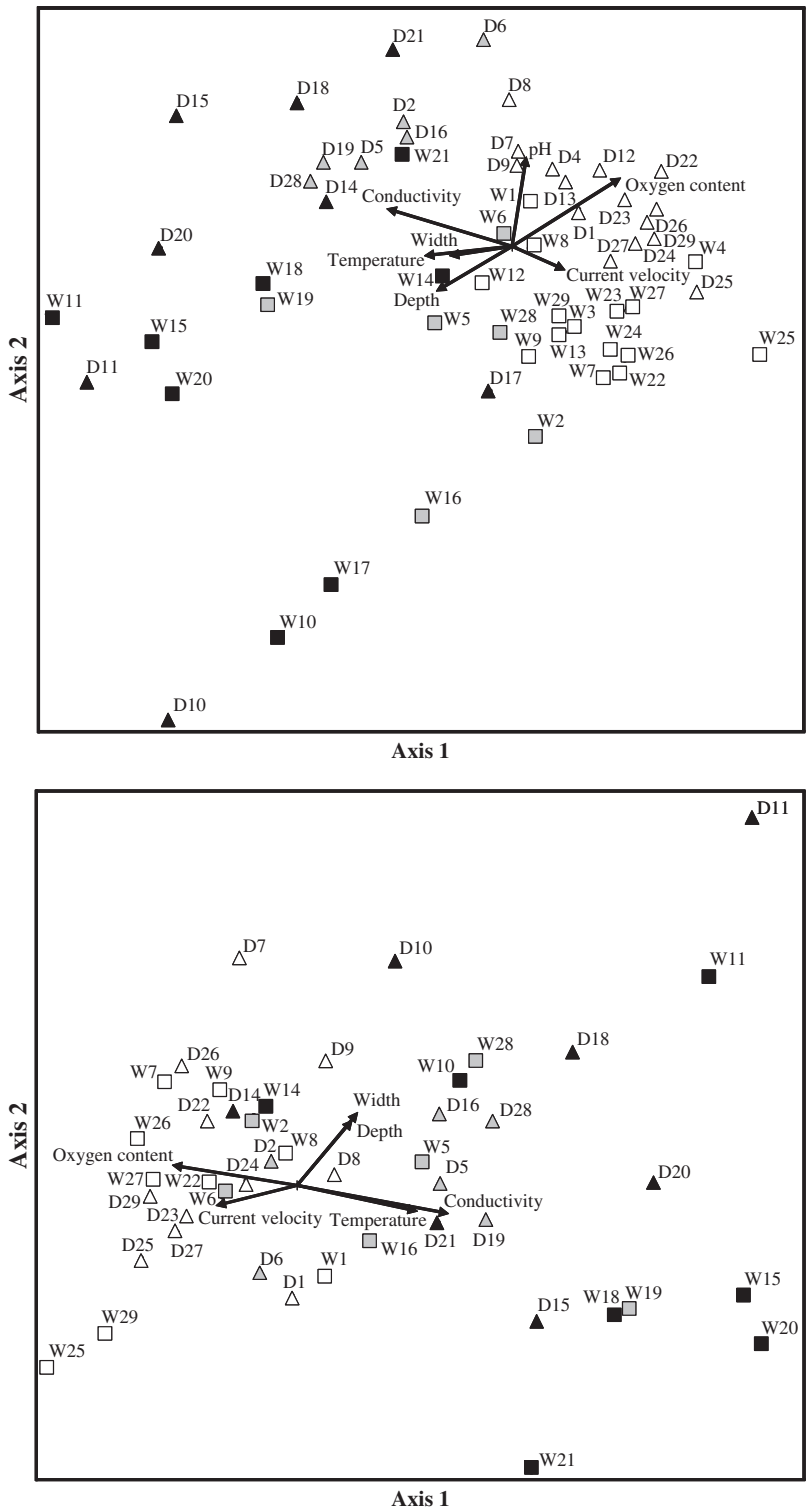

Figure 3 | Bi-plot of the sample scores and the environmental variables with indication of the three identified clusters (black: high impact; grey: intermediate impact: white: low impact) and the sampling season (dry season: triangles; wet season: squares) for 2005 and 2006.

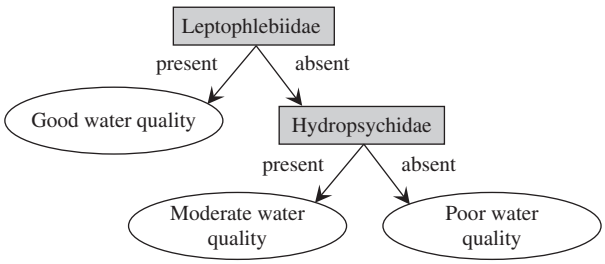

Figure 4 | Classification tree which predicted the water quality based on the occurrence of macroinvertebrate taxa.

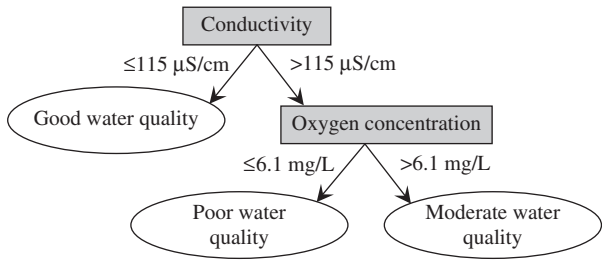

Figure 5 | Classification tree which predicted the water quality based on the environmental variables.

macroinvertebrates and the environmental parameters together, only environmental parameters were retained in the classification tree.

DISCUSSION AND CONCLUSIONS

The influence of altitude on the invertebrate community composition of running waters in the neotropics has recently been studied by several authors (e.g. Carrera Burneo & Gunkel 2003; Sites *et al.* 2003; Jacobsen 2005). Relationships are frequently attributed to natural changes based on the river continuum concept (Vannote *et al.* 1980): however, the influence of human activities could be obscuring these natural patterns. An inverse relationship is frequently observed between the altitude and the intensity of human activities, such as agriculture, human settlements and cattle grazing, and its related environmental degradation (e.g. organic, chemical and morphological) (Lang & Reymond 1993). Also in the Chaguana basin, major human activities are carried out in the lowlands, mainly banana farming, which is the dominant crop cultivated in the basin (Matamoros *et al.* 2005). The results of the present study confirmed that the higher reach of the basin is only slightly affected by human settlements, leaving streams and its macroinvertebrates in that area in a nearly pristine state (Figure 1). However, based on the macroinvertebrate communities, the water quality in the downstream sites was evaluated as poor due to the high human impact. As the highest sampling point in the present study was only 1600 m above sea level, altitude was of lesser importance as a factor limiting oxygen availability by reduced atmospheric pressure.

In lotic ecosystems, the physical environment exerts substantial control over the population abundances and hence the community composition of insects. Some of the physical factors that are of particular importance to aquatic insects include dissolved oxygen concentrations, water temperature and hydrodynamics, and all these factors are characterised by seasonal fluctuations. In contrast to temperate areas, tropical regions are not characterised by strong thermal changes over time. Most tropical regions experience, rather, some degree of seasonality related to rainfall. Also in the Chaguana basin, the annual rainfall is highly seasonal, with major peaks during the months of April and November and significant differences between seasons were detected for parameters such as water temperature, dissolved oxygen and current velocity.

In a coastal Atlantic Forest river in SE Brazil, no differences were found between macroinvertebrate communities at

Table 1 | Confusion matrix of the developed trees predicting the water quality based on the occurrence of macroinvertebrate taxa and based on the environmental variables. The values for 2005 and 2006 are given separately

	Predicted water quality (based on macroinvertebrate taxa)			Predicted water quality (based on environmental variables)		
	Poor	Moderate	Good	Poor	Moderate	Good
Observed water quality poor	15/10	1/2	0/2	8/8	3/2	5/4
Observed water quality moderate	6/2	4/4	2/6	1/1	5/8	6/3
Observed water quality good	1/3	4/2	25/15	0/0	0/0	30/20

the end of the rainy season, the dry season and the beginning of the next rainy season: 95–100 taxa were observed during each sampling campaign (Baptista *et al.* 2001). Slightly more taxa were registered for the wet (38 taxa) than for the dry period (32 taxa) in two streams of the São Francisco river basin in SE Brazil (Tupinambás *et al.* 2007). Qualitatively, communities were also not influenced by seasonal changes in SE Brazil, but abundance was negatively influenced by rainfall: 73 taxa were observed during the dry season, while 66 taxa were observed during the wet season and at the end of the wet season, 51 taxa were present during the three sampling campaigns (Buss *et al.* 2004). Melo & Froehlich (2001) found no difference in macroinvertebrate richness and abundance between wet and dry seasons within the Carmo river catchment in Brazil. In two streams in S Brazil, only small differences in the number of families were observed between seasons (Bueno *et al.* 2003). Also in four highland streams in S Brazil, no systematic seasonal differences were found in the number of taxa and other diversity indices (Buckup *et al.* 2007). Invertebrate drift in two lowland streams in Costa Rica revealed that all major taxa were found every month; there was a trend for higher densities during the dry season, but these trends were not significant (Ramirez & Pringle 2001). In eight Ecuadorian highland streams, the number of individuals and species were significantly higher in the dry season than in the wet season; in all streams the composition of the fauna differed markedly between the two seasons, but no consistent changes in community structure were found for the eight streams (Jacobsen & Encalada 1998). In these Ecuadorian highland streams, more taxa were collected in the dry season at all upstream sites, while at the polluted downstream sites, more taxa were collected during the wet season (Jacobsen 1998). In the Andean stream Rio Itambi (Otavalo, Ecuador), 40 taxa were observed during the wet season, while 32 taxa were observed during the dry season (Carrera Burneo & Gunkel 2003). In the present study, no systematic differences in richness and abundance were observed between the dry and wet seasons in the Chaguana catchment. Overall, it can thus be concluded that only minor differences in species richness were observed between seasons.

Not seasons, but spates, are thought to be the major factor regulating benthic densities in the neotropics. Turcotte & Harper (1982) found that benthos densities in a small headwater stream in the Amazon drainage in Ecuador followed a bimodal pattern of rapid decreases and intervening gradual recoveries after spates. Also, in the Venezuelan Andes, major declines in invertebrate numbers were generally associated with spates (Flecker & Feifarek 1994). Monthly invertebrate densities were inversely related to rainfall at each site and a highly significant positive relationship was observed between invertebrate density and the number of days elapsed since the previous major rainstorm event (Flecker & Feifarek 1994). Also Buss *et al.* (2004) found that, during the wet season, the high precipitation observed before the sampling period led to a decrease in richness at upstream sites and an increase in richness at downstream sites. It can therefore be concluded that spates affect macroinvertebrate richness: due to washout, less species are found upstream, while more species can be observed downstream. The influence of spates on macroinvertebrates should be further investigated by coupling hydrological models with water quality models (van Griensven *et al.* 2006). However, in the framework of water quality assessments, sampling macroinvertebrates shortly after spates should be avoided.

In the present study, classification trees were successfully induced to assess the water quality based on the occurrence of macroinvertebrate indicator taxa as well as on the physical–chemical key variables. Dzeroski *et al.* (1997) indicated that the reduction of the dimensionality of a tree contributed to an easier interpretation of the revealed trends in the data, focusing attention on the important variables. Dakou *et al.* (2007) found that optimal pruning is an important mechanism as it improves the transparency of the induced trees by reducing their size and enhances the classification accuracy by eliminating errors that are present due to noise in the data. In the present study, tree size was reduced by using a small confidence factor and using a large minimum group size for division. In this way, small trees were obtained which were easy to interpret and which can be used for decision support by water quality managers. Unpruned trees resulted in more complex trees, which did not perform better when based on the occurrence of macroinvertebrates and only marginally better when based on environmental variables. In addition, more complex trees are not only more difficult to interpret, their generalisation capacity is usually also lower due to overfitting to the dataset used.

In the present study, sampling stations were clustered into three groups using multivariate analysis (TWINSPAN): stations

with a low, an intermediate and a high human impact. No systematic differences in richness and abundance were observed between the dry and wet seasons of 2005 and 2006. By applying classification trees, it was possible to predict the water quality based on the occurring macroinvertebrates. Classification trees based on the environmental variables performed even better and resulted in more stable results when samples from 2005 and 2006 were compared. In addition, the transparency of classification trees, especially pruned ones, makes them easy to interpret. It can therefore be concluded that classification trees are a useful practical tool for decision support by water quality managers.

ACKNOWLEDGEMENTS

The first author is grateful for the financial support of the VLIR-ESPOL IUC programme in Ecuador and SENACYT. In particular, we would like to thank Magda Vincx, coordinator of the VLIR-ESPOL IUC programme, Nancy Fockedey, pioneer of this programme in Ecuador, and Pilar Cornejo, director of this program in Ecuador. We would also like to thank Galo, Erick, Christian L., Christian R., Christian V., Felix and Santiago for help during the field work. KL is currently supported by a post-doctoral fellowship from the Fund for Scientific Research (FWO-Vlaanderen, Belgium).

REFERENCES

Adriaenssens, V., De Baets, B., Goethals, P. L. M. & De Pauw, N. 2004a Fuzzy rule-based models for decision support in ecosystem management. *Sci. Total Environ.* **319**, 1–12.

Adriaenssens, V., Goethals, P. L. M., Charles, J. & De Pauw, N. 2004b Application of Bayesian Belief Networks for the prediction of macroinvertebrate taxa in rivers. *Ann. Limnol. Int. J. Limnol.* **40**, 181–191.

Baptista, D. F., Dorvillé, L. F. M., Buss, D. F. & Nessimian, J. L. 2001 Spatial and temporal organization of aquatic insect assemblages in the longitudinal gradient of a tropical river. *Braz. J. Biol.* **61**, 295–304.

Blockeel, H., Dzeroski, S. & Grbovic, J. 1999 Simultaneous prediction of multiple chemical parameters of river water quality with TILDE. *Lecture Notes Artif. Intell.* **1704**, 32–40.

Buckup, L., Bueno, A. A. P., Bond-Buckup, G., Casagrande, M. & Majolo, F. 2007 The benthic macroinvertebrate fauna of highland streams in S Brazil: composition, diversity and structure. *Rev. Bras. Zool.* **24**, 294–301.

Bueno, A. A. P., Bond-Buckup, G. & Ferreira, B. D. P. 2003 Community structure of benthic invertebrates in two watercourses in Rio Grande do Sul State, southern Brazil. *Rev. Bras. Zool.* **20**, 115–125.

Buss, D. F., Baptista, D. F., Nessimian, J. L. & Egler, M. 2004 Substrate specificity, environmental degradation and disturbance structuring macroinvertebrate assemblages in neotropical regions. *Hydrobiologia* **518**, 179–188.

Carrera Burneo, P. & Gunkel, G. 2003 Ecology of a high Andean stream, Rio Itambi, Otavalo, Ecuador. *Limnologica* **33**, 29–43.

Chen, Q. W., Ye, F. & Li, W. F. 2009 Cellular-automata-based ecological and ecohydraulics modelling. *J. Hydroinf.* **11**, 252–265.

Clarke, R. T., Furse, M. T., Gunn, R. J. M., Winder, J. M. & Wright, J. F. 2002 Sampling variation in macroinvertebrate data and implications for river quality indices. *Freshwat. Biol.* **47**, 1735–1751.

Cohen, J. 1960 A coefficient of agreement for nominal scales. *Educ. Physiol. Meas.* **20**, 37–46.

Dakou, E., D'heygere, T., Dedecker, A. P., Goethals, P. L. M., Lazaridou-Dimitriadou, M. & De Pauw, N. 2007 Decision tree models for prediction of macroinvertebrate taxa in the river Axios (Northern Greece). *Aquat. Ecol.* **41**, 399–411.

Dedecker, A., Van Melckebeke, K., Goethals, P. L. M. & De Pauw, N. 2007 Development of migration models for macroinvertebrates in the Zwalm river basin (Flanders, Belgium) as tools for restoration management. *Ecol. Modell.* **203**, 72–86.

D'heygere, T., Goethals, P. L. M. & De Pauw, N. 2006 Genetic algorithms for optimisation of predictive ecosystems models based on decision trees and neural networks. *Ecol. Modell.* **195**, 20–29.

Domínguez, E., Hubbard, M. D. & Pescador, M. L. 1994 Los Ephemeroptera en Argentina. *Fauna de Agua Dulce de la Republica Argentina* **33**, 1–142.

Dzeroski, S., Demsar, D. & Grbovic, J. 2000 Predicting chemical parameters of river water quality from bioindicator data. *Appl. Intell.* **13**, 7–17.

Dzeroski, S., Grbovic, J. & Walley, W. J. 1997 Machine learning applications in biological classification of river water quality. In: *Machine Learning and Data Mining: Methods and Applications* (Michalski, R. S., Bratko, I. & Kubat, M. (Eds.)). John Wiley & Sons, New York, pp 429–448.

Elshorbagy, A. & El-Baroudy, I. 2009 Investigating the capabilities of evolutionary data-driven techniques using the challenging estimation of soil moisture content. *J. Hydroinf.* **11**, 237–251.

Fernández, H. R. & Dominguez, E. 2001 *Guía para la determinación de los artrópodos bentónicos sudamericanos.* Universidad Nacional de Tucumán, Tucumán.

Flecker, A. S. & Feifarek, B. 1994 Disturbance and the temporal variability of invertebrate assemblages in two Andean streams. *Freshwat. Biol.* **31**, 131–142.

Gerritsen, J., Barbour, M. T. & King, K. 2000 Apples, oranges, and ecoregions: on determining pattern in aquatic assemblages. *J. North Am. Benthol. Soc.* **19**, 487–496.

Giustolisi, O. & Savic, D. A. 2009. Advances in data-driven analyses and modelling using EPR-MOGA. *J. Hydroinf.* **11**, 225–236.

Hill, M. O. 1979 *A FORTRAN Program for Arranging Multivariate Data in an Ordered Two-way Table by Classification of the Individuals and Attributes*. Cornell University, Ithaca, NY.

Hoang, T. H., Lock, K., Mouton, A. & Goethals, P. L. M. 2010 Application of decision trees and support vector machines to model the presence of macroinvertebrates in rivers in Vietnam. *Ecol. Inform.* **5**2, 140–146.

Jacobsen, D. 1998 The effect of organic pollution on the macroinvertebrate fauna of Ecuadorian highland streams. *Arch. Hydrobiol.* **143**, 179–195.

Jacobsen, D. 2005 Temporally variable macroinvertebrate-stone relationships in streams. *Hydrobiologia* **544**, 201–214.

Jacobsen, D. & Encalada, A. 1998 The macroinvertebrate fauna of Ecuadorian highland streams in the wet and dry season. *Arch. Hydrobiol.* **142**, 53–70.

Landis, J. R. & Koch, G. G. 1977 The measurement of observer agreement for categorical data. *Biometrics* **33**, 159–174.

Lang, C. & Reymond, O. 1993 Empirical relationships between diversity of invertebrate communities and altitude in rivers: application to biomonitoring. *Aquat. Sci.* **55**, 188–196.

Linke, S., Bailey, R. C. & Schwindt, J. 1999 Temporal variability of stream bioassessments using benthic macroinvertebrates. *Freshwat. Biol.* **42**, 575–584.

Matamoros, D. E., van Griensven, A., van Biesen, L. & Vanrolleghem, P. A. 2005 Development of a geographical information system for pesticide assessment on an Ecuadorian watershed. *Wat. Sci. Technol.* **52**, 259–265.

Melo, A. S. & Froehlich, C. G. 2001 Macroinvertebrates in neotropical streams: richness patterns along a catchment and assemblage structure between 2 seasons. *J. North Am. Benthol. Soc.* **20**, 1–16.

Mouton, A. M., De Baets, B. & Goethals, P. L. M. 2009 Knowledge-based versus data-driven fuzzy habitat suitability models for river management. *Environ. Modell. Software.* **24**, 982–993.

Ramirez, A. & Pringle, C. M. 2001 Spatial and temporal patterns of invertebrate drift in streams draining a Neotropical landscape. *Freshwat. Biol.* **46**, 47–62.

Roldán, G. 1988 *Guía para el estudio de los macroinvertebrados acuáticos del Departamento de Antioquia*. Fondo FEN Colombia, Conciencias-Universidad de Antioquia, Santafé de Bogota.

Savic, D. A., Giustolisi, O. & Laucelli, D. 2009 Asset deterioration analysis using multi-utility data and multi-objective data. *J. Hydroinf.* **11**, 211–224.

Sites, R. W., Willig, M. R. & Linit, M. J. 2003 Macroecology of aquatic insects: a quantitative analysis of taxonomic richness and composition in the Andes mountains of Northern Ecuador. *Biotropica* **35**, 226–239.

Ter Braak, C. J. F. & Smilauer, P. 2002 *CANOCO Reference Manual and CanoDraw for Windows User's Guide: Software for Canonical Community Ordination (version 4.5)*. Microcomputer Power, Ithaca, NY.

Tupinambás, T. H., Callisto, M. & Santos, G. B. 2007 Benthic macroinvertebrate assemblages structure in two headwater streams, south-eastern Brazil. *Rev. Bras. Zool.* **24**, 887–897.

Turak, E. & Koop, K. 2008 Multi-attribute ecological river typology for assessing ecological condition and conservation planning. *Hydrobiologia* **603**, 83–104.

Turcotte, P. & Harper, P. P. 1982 The macro-invertebrate fauna of a small Andean stream. *Freshwat. Biol.* **12**, 411–419.

van Griensven, A., Breuer, L., Di Luzio, M., Vandenberghe, V., Goethals, P., Meixner, T., Arnold, J. & Srinivasan, R. 2006 Environmental and ecological hydroinformatics to support the implementation of the European Water Framework Directive for river basin management. *J. Hydroinf.* **8**, 239–252.

Vannote, R. L., Minshall, G. W., Cummins, K. W., Sedell, J. R. & Cushing, C. E. 1980 The river continuum concept. *Can. J. Fish. Aquat. Sci.* **37**, 130–137.

Walley, W. J. & Dzeroski, S. 1995 Biological monitoring: a comparison between Bayesian, neural and machine learning methods of water quality classification. In: *Environmental Software Systems* (Denzer, R., Shimak, G. & Russell, D. (Eds.)). Chapman & Hall, London, pp 229–240.

Witten, I. H. & Frank, E. 2005 *Data Mining: Practical Machine Learning Tools and Techniques* 2nd edn. Morgan Kaufmann, San Francisco.

First received 29 September 2009; accepted in revised form 22 February 2010. Available online 1 October 2010