Zhang, M.; Muñoz Mas, R.; Martinez-Capel, F.; Qu, X.; Zhang, H.; Peng, W.; Liu, X. (09-2). Determining the macroinvertebrate community indicators and relevant environmental predictors of the Hun-Tai River Basin (Northeast China): A study based on community patterning. The Science of The Total Environment. 634:749-759. https://doi.org/10.1016/j.scitotenv.2018.04.021

Determining the macroinvertebrate community indicators and relevant environmental predictors of the Hun-Tai River Basin (Northeast China): a study based on community patterning

Min Zhang[1,2], Rafael Muñoz-Mas[3], Francisco Martínez-Capel[3], Xiaodong Qu[1,2,*], Haiping Zhang[1,2], Wenqi Peng[1,2], Xiaobo Liu[1,2]

[1] State Key Laboratory of Simulation and Regulation of Water Cycle in River Basin, China Institute of Water Resources and Hydropower Research, Beijing 100038, China

[2] Department of Water Environment, China Institute of Water Resources and Hydropower Research, Beijing 100038, China

[3] Institut d'Investigació per a la Gestió Integrada de Zones Costaneres (IGIC), Universitat Politècnica de València, C/ Paranimf 1, Grau de Gandia, València 46730, Spain.

*Corresponding author: Xiaodong Qu, e-mail: davidqu0108@126.com

Abstract

It is essential to understand the patterning of biota and environmental influencing factors for proper rehabilitation and management at the river basin scale. The Hun-Tai River Basin was extensively sampled four times for macroinvertebrate community and environmental variables during one year. Self-Organizing Maps (SOMs) were used to reveal the aggregation patterns of the 355 samples. Three community types (*i.e.,* clusters) were found (at the family level) based on the community composition, which showed a clearly gradient by combining them with the representative environmental variables: minimally impacted source area, intermediately anthropogenic impacted sites, and highly anthropogenic impacted downstream area, respectively. This gradient was corroborated by the decreasing trends in density and diversity of macroinvertebrates. Distance from source, total phosphorus and water temperature were identified as the most important variables that distinguished the delineated communities. In addition, the sampling season, substrate type, pH and the percentage of grassland were also identified as relevant variables. These results demonstrated that macroinvertebrates communities are structured in a hierarchical manner where geographic and water quality prevail over temporal (season) and habitat (substrate type) features at the basin scale. In addition, it implied that the local-scale environment variables affected macroinvertebrates under the longitudinal gradient of the geographical and anthropogenic pressure. More than one families were identified as the indicator for each type of community. Abundance contributed significantly for distinguishing the indicators, while Baetidae with higher density indicated minimally and intermediately impacted area and lower density indicated highly impacted area. Therefore, we suggested the use of abundance data in community patterning and classification, especially in the identification of the indicator taxa.

Keywords

Macroinvertebrate patterning, indicators, Self-Organizing Map, decision tree, site classification

1 Introduction

Currently, human activities are greatly influencing the flow rate, water yield, sediment transport and nutrient releases in freshwater ecosystems at scales that far exceed those of natural phenomena (Habersack et al., 2014). Accordingly, water resources are currently over-exploited in many regions, which has resulted in 65% of rivers worldwide being under moderate-to-high threats in terms of human water security and biodiversity loss (Vörösmarty et al., 2010).

Biotic assemblages in freshwater ecosystems integrate these impacts throughout the drainage basins; thus, these assemblages can be considered as indicators of ecosystem health (Habersack et al., 2014). Consequently, the classification and delineation of the ecological statuses of rivers based on the biotic assemblages is an essential prerequisite for river ecosystem assessment, restoration and management (Heino et al., 2002; Marchant et al., 2000; Siddig et al., 2016; Tsai et al., 2017).

Macroinvertebrate assemblages have been widely used as indicators of ecosystem changes because macroinvertebrate communities encompass a diverse group with a wide range of life-history requirements (O'Brien et al., 2016). Macroinvertebrates vary spatially and temporally and integrate ecosystem changes as a result of their suite of feeding strategies and lifestyles and their different sensitivities to changes in physical habitat and water quality (Milošević et al., 2016; Ogbeibu and Oribhabor, 2002). According to a recent review on indicator species over the last 14 years, nearly 50% of the taxa used as indicators were animals, and 70% of these were invertebrates (Siddig et al., 2016). However, data on macroinvertebrate assemblages are highly complex and difficult to analyze because macroinvertebrate assemblages consist of numerous species that respond in complex manners to natural and anthropogenic pressures (Kim et al., 2013; Tsai et al., 2017). In this situation, supervised machine learning approaches, which make use of techniques from mathematical programming and statistics, have been used to scrutinize and model the environmental requirements

63 of relevant macroinvertebrate taxa, and these techniques include decision trees (C4.5 – D'heyere et
64 al., 2003) or multilayer perceptrons (Edia et al., 2010).

65 In addition, macroinvertebrate datasets include numerous taxa and a large number of samples, which
66 can also cause difficulties for community analysis and river regionalization (Kim et al., 2013). In
67 particular, ordination techniques and unsupervised machine learning approaches have been used to
68 explore patterns of occurrence and community shifts and their relationships with environmental
69 factors (Adriaenssens et al., 2007; Giraudel and Lek, 2001; Zhang et al., 2012a). Nevertheless, each
70 ordination technique may have important limitations and assumptions that are incompatible with the
71 over-dispersion and nonlinear nature of ecological data (Paliy and Shankar, 2016). Researchers have
72 advocated for the use of a type of unsupervised artificial neural network called Kohonen
73 Self-Organizing Maps (SOMs) (Kohonen, 1982), which have been demonstrated to be particularly
74 competent in analyses such as macroinvertebrate community delineations (Chon, 2011; Park et al.,
75 2007; Kim et al., 2013; Sroczyńska et al., 2017).

76 The freshwater ecosystems of China are a clear example of the abovementioned human-induced
77 impacts. For instance, more than 40% of the rivers in China are notably polluted, which has led to
78 poor drinking water quality for approximately 300 million rural residents (Liu and Yang, 2012). The
79 river ecosystems in the northeast have also degraded due to industrial and agricultural development;
80 thus, some river restoration work has been conducted in this area (Kong et al., 2013; Zhang et al.,
81 2011; Zhang et al., 2013). The Hun-Tai River Basin is a large river basin with a basin area of
82 2.73×10^4 km^2. It represents the overall status of the water in the Liaohe River Basin in northeast
83 China and is undergoing degradation. Many field surveys and studies using macroinvertebrates as
84 important indicators in river health assessments showed the ecosystem were not in good conditions
85 (Qu et al., 2016; Zhang et al., 2011; Zhang et al., 2013). However, most of these studies have mainly
86 focused on small rivers or tributaries, and cannot reflect the overall status of the whole watershed,
87 especially in such a large river basin. Up to now, little is known about the macroinvertebrate

community and related environmental variables in the entire basin. We hypothesized when data from a large river basin and temporal span is merged, the geographical features could override the local environmental variables (e.g. water quality) in structuring the macroinvertebrate community, because the geographical features (e.g. elevation, distance from the river source) usually determine the macroinvertebrate structures when communities are studied at broader scales (Dedieu et al., 2014; Gaston, 2000).

This study analyzed the macroinvertebrate assemblages present in the large Hun-Tai River Basin to elucidate the existence of different types of communities and determine the indicator families and main environmental predictors for their occurrence. SOMs were used to reveal the existence of these macroinvertebrate communities (*i.e.*, clusters) in different areas of the river basin. Then, a genetically optimized C5.0 algorithm (Quinlan, 1992) (*i.e.*, a type of decision tree) was used to reveal the most important set of environmental predictors and indicator taxa of each community in the Hun-Tai River Basin.

2 Materials and methods

2.1 Study area

The Hun-Tai River is located in Liaoning Province of Northeast China, and it has two main tributaries, the Taizi and the Hunhe Rivers (Fig. 1). The lengths of the Hunhe and Taizi Rivers are approximately 415 and 413 km, respectively. The climate in this area is typical continental monsoon, with the highest temperature (34.3 ºC) in the summer and lowest temperature (-25.2 ºC) in the winter. The precipitation follows the temperature pattern, with the annual average precipitation 778 mm, 63% of which occurs in summer (Bu et al. 2014).

Field surveys were carried out in May 2009 (spring), August 2009 (summer), October 2009 (autumn) and May 2010 (spring). These surveys encompassed the entire river basin. In total, 287 sites from May 2009 to May 2010 in the Hun-Tai River Basin, in which 68 of Taizi River Basin were sampled twice (The number of sampling sites in each river and the codes of each river are located can be found in Appendix A). Consequently, 355 samples were ultimately collected where the selected environmental variables were measured *in situ* or obtained from reference databases.

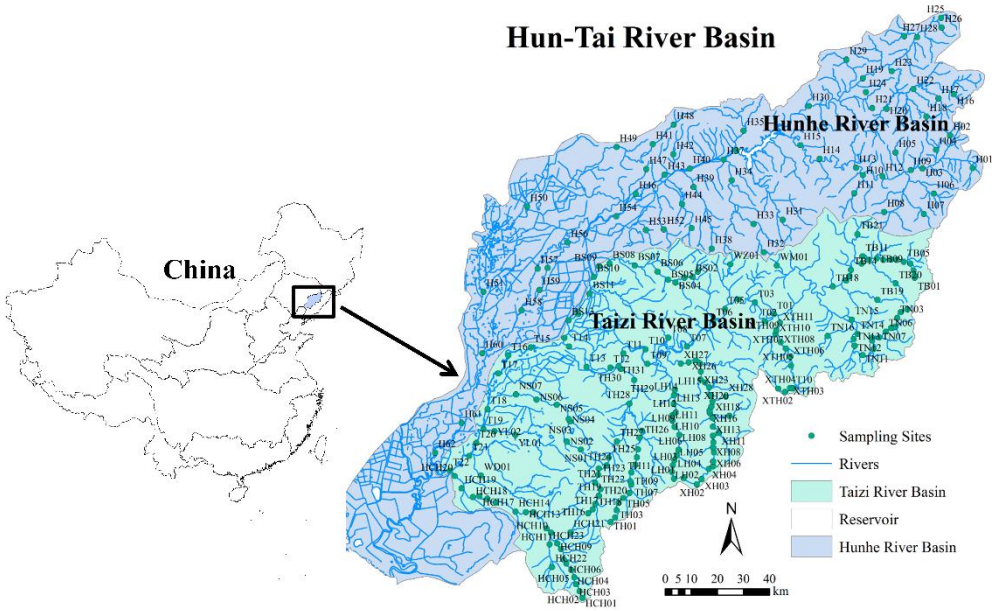

Fig. 1. Location of the study area and sampling sites within the Hun-Tai River Basin. Further details about the codes depicted in this map can be found in Appendix A.

2.2 Data collection

2.2.1 Environmental variables

All sampling sites were characterized using variables determining the geography, hydrology, climate, landuse, water quality and habitat, and there were 30 variables and 1 binary control variable (wadeable or non-wadeable) in total (Table 1). For each site, a handheld global positioning system

(GPS, Trimble Juno SA) was used to obtain the latitude, altitude and elevation above sea level (m a.s.l.). Distance from the source was extracted from the digital map of the river basin using ArcGIS 10.2. The river widths and water depths were directly measured with tape. The river width was measured by randomly selecting three transects per sampling station While the water depth was measured at the macroinvertebrate sampling sites. Air temperature, water temperature, pH, dissolved oxygen (DO), electrical conductivity (EC), and total dissolved solids (TDS) were measured by a multiparameter water quality probe (YSI-Pro Plus©, YSI Inc., USA).

One liter of stream water was collected and transported in a portable fridge. Chemical oxygen demand (COD), total nitrogen (TN), ammonia nitrogen (NH_3-N), nitrate-nitrogen (NO_3-N), and total phosphorus (TP) were measured in the laboratory according to the "Environmental quality standards for surface water" of China (GB3838-2002). COD and NO_3-N measurements were not collected in BeiShahe, NanShahe, TangHe and XiaoTangHe Rivers in spring 2010 (Table 1). Suspended solids (SS) were measured by filtration through pre-dried cellulose acetate membranes (0.45 μm) according to the Chinese standard (GB11901-89). The particle sizes of the substrate were measured by using a series of stainless steel mesh sizes according to the modified Wentworth classification of substrate and then expressed using a percentage. The substrate sizes were classified as boulder ($Ø > 256$ mm), large cobble ($128 > Ø \leq 256$ mm), cobble ($64 > Ø \leq 128$ mm), large pebble ($32 > Ø \leq 64$ mm), pebble ($16 > Ø \leq 32$ mm), large gravel ($8 > Ø \leq 16$ mm), gravel ($4 > Ø \leq 8$ mm) and small gravel, sand and silt ($Ø < 4$ mm) (Cummins, 1962). The substrates in the HaichengHe, LanHe and XiHe Rivers were not monitored in summer 2009. Finally, the proportion of each landuse type was extracted over 3 km upstream and 500 m wide buffer zones (Zhang et al., 2013) from a digital landuse map using ArcGIS 10.2. The landuse map was interpreted from landsat TM data of the year 2010. These variables and the sampling season were used to infer the most determinant environmental variables of the macroinvertebrate communities of the Hun-Tai River Basin. The basic statistics and the number of unavailable data for the whole variables' set are shown in Table 1.

150 Table 1. Summary and units of environmental variables collected in the Hun-Tai River Basin

Variable group	Variables	Min	Mean	Median	Max	Number of unavailable data
Geographic	Elevation (m a.s.l.)	4.53	205.38	188.00	663.00	0
	Distance from source (km)	0.12	65.55	34.52	380.68	0
Landuse	Grassland (%)	0.00	4.47	0.00	54.91	0
	Agricultural land (%)	0.00	48.59	48.11	97.04	0
	Residential land (%)	0.00	14.30	4.60	96.90	0
	Forest (%)	0.00	32.63	31.03	100.00	0
Hydrologic	Discharge (m^3/s)	0.00	3.40	1.04	94.76	0
	River width (m)	0.30	40.41	15.00	420.00	0
	Water depth (cm)	0.80	24.60	22.33	130.00	0
	Velocity (m/s)	0.00	0.40	0.38	1.14	0
Climatic	Air temperature (ºC)	5.20	21.22	21.80	36.80	0
Water quality	Water temperature (ºC)	6.80	17.42	18.00	27.80	0
	pH	6.01	8.39	8.40	10.12	0
	DO (mg/l)	0.00	8.93	9.63	15.63	0
	EC (μs/cm)	3.90	317.51	272.00	1431.00	0
	SS (mg/l)	1.00	67.08	19.50	1110.00	0
	TDS (mg/l)	14.95	235.15	192.00	995.00	0
	COD (mg/l)	0.00	22.96	16.00	151.00	62
	TN (mg/l)	0.47	5.19	3.49	22.60	0
	NH_3-N (mg/l)	0.00	1.15	0.38	20.60	0
	NO_3-N (mg/l)	0.01	2.16	1.60	15.90	70
	TP (mg/l)	0.00	0.19	0.07	3.03	0
Habitat	Boulder (%)	0.00	12.74	0.00	90.56	119
	Large cobble (%)	0.00	19.91	18.64	79.70	119
	Cobble (%)	0.00	15.65	16.42	58.38	119
	Large pebble (%)	0.00	11.77	11.26	77.78	119
	Pebble (%)	0.00	8.82	8.18	47.13	119
	Large gravel (%)	0.00	3.64	3.01	20.00	119
	Gravel (%)	0.00	4.34	2.85	29.05	119
	Small gravel, sand and silt (%)	0.00	23.12	4.84	100.00	119

151

152 **2.2.2 Macroinvertebrate sampling**

153 Macroinvertebrates were collected using a Surber net (30×30 cm, 500 μm mesh). At the sites that
154 could be waded, three replicates were obtained from two riffles and one shallow pool. For the sites
155 that could not be waded, a Surber net was used to collect three replicates in shallow water along the
156 riverside.

The samples were passed through a 500 μm mesh sieve, and organisms retained on the sieve were fixed and preserved in 10% formaldehyde. Most taxa were identified based on the available references (Brinkhurst, 1986; Merritt and Cummins, 1996; Morse et al., 1994; Wiggins, 1996). Finally, 90 families were identified and they were used to cluster the sampling sites (community delineation) and to obtain the community indicators. The density (individuals/m^2 – ind./m^2), richness (number of species) and the Shannon–Wiener diversity and Pielou evenness indices were calculated for each sample to characterize the macroinvertebrate communities of the Hun-Tai River Basin.

The wadeable *vs* non-wadeable nature of each site was used as a control variable during the process to determine the most relevant environmental variables. Ruling out this variable indicated us that the sampling protocol had no impact in the delineated communities (*i.e.*, clusters).

2.3 Statistical analysis

2.3.1 Self-Organizing Maps (SOMs)

Delineating macroinvertebrate communities with Self-Organizing Maps (SOMs) is done in two steps (see e.g., Edia et al., 2010; Kim et al., 2013; Park et al., 2007) (Fig. 2). First, the SOM implements an ordered dimensionality-reducing mapping of input variables (Kohonen, 1982). Therefore, SOM provided a projection of the matrix of n = 335 rows and p = 90 families onto a topological structure (i.e., a XY 2D map of unit neurons nodes) of smaller dimensionality (*i.e.*, $X << n$ and $Y << p$). The entire process is carried out preserving the original topology of the input data. In accordance, the neurons that are located near to each other in the SOM had similar associated input samples (i.e., macroinvertebrate communities). Then, these neuron nodes are clustered, usually employing the Ward's linkage method (Murtagh and Legendre, 2014), allowing exclusively the aggregation of contiguous neuron nodes. Finally, the samples (i.e., macroinvertebrate communities) assigned to the neuron nodes aggregated in a given cluster are grouped (i.e. clustered) together, and upon these

communities or clusters, further analyses were performed (see e.g., Edia et al., 2010; Kim et al., 2013; Park et al., 2007).

The training and visualization of the SOM was performed using the functionalities implemented within the *R* package *kohonen* (Wehrens and Buydens, 2007). SOMs were trained for all possible combinations from 2 to 20 neuron nodes for the *X* and *Y* axes and the quality of these alternative topologies was evaluated using the quantization error (QE), which evaluates the resolution of the map, and the topographic error (TE), which indicates the accuracy of the topology preservation of the map (Kim et al., 2013; Tsai et al., 2017). The learning rate (α) varied linearly from 0.05 and 0.01, and the neighborhood function was Gaussian. Conversely, the initial radius varied in accordance with the SOM dimensions (*i.e.*, *X* and *Y*) after the $argmax\{\left\lfloor\frac{X}{3}\right\rfloor, \left\lfloor\frac{Y}{3}\right\rfloor\}$. In terms of densities (ind./m^2), the macroinvertebrate data depicted large numerical differences. Therefore, data were transformed (log+1) prior to training the SOMs (Adriaenssens et al., 2007; Kim et al., 2013; Tsai et al., 2017). Finally, the input sample data were presented to each SOM 500 times and the selected distance measure was the Euclidean distance (See Appendix B for further details about the optimization of SOMs).

Once the optimal dimensions of the SOM were determined based on the QE and TE, the Ward's linkage method was applied to the SOM to cluster the unit neurons and, consequently, the macroinvertebrate samples (Chon, 2011; Park et al., 2007; Tsai et al., 2017). The function *NbClust* included in the homonymous *R* package (Charrad et al., 2014) was used to determine the optimal number of clusters. The latter function calculates 30 quality indices, from *ball* (Ball and Hall, 1965) to *gap* (Tibshirani et al., 2001), and the optimal number of clusters is determined using the majority rule. In this case, the optimal number of cluster between 2 and 15 was sought. Once the optimal number of clusters was determined, the distribution of the categorical variables (season) was

visualized with the SOM, and the family density, richness and the diversity indices of the samples assigned to each cluster were scrutinized with violin plots.

2.3.2 C5.0 algorithm

The most relevant indicator families and environmental predictors of the macroinvertebrate communities that were delineated by the SOM (*i.e.*, clusters) were identified with a genetically optimized C5.0 algorithm (Quinlan, 1992) (Fig. 2). The C5.0 algorithm was selected because it is a kind of decision tree that is able to handle missing or unavailable data. In addition, C5.0 is able to collapse the former tree-like structure into a compact list of IF-THEN rules. The missing data was down-weighted when the entropy gain is calculated. The proportion of missing data do not necessarily reduces the predictive capacity of a variable.

To prevent overfitting, a *wrapper* approach involving cross-validation and the Genetic Algorithm (GA) (Holland, 1992) implemented within the *R* package *rgenoud* (Mebane Jr & Sekhon, 2011) was used to find the optimal variable set and C5.0 hyperparameters (Muñoz-Mas et al., 2016). The parameters of the GA were selected to avoid premature convergence (Muñoz-Mas et al., 2016) whereas, compared to previous studies (D'heygere et al., 2003; Gobeyn et al., 2017), the population size and number of generations were set very large (i.e. to 1000), and the optimization halted after 250 generations without improvement (See Appendix B for further detail about the optimization with the GA). The optimization took place by maximizing the product of the individual sensitivities (Sn) for each class (*i.e.*, cluster), as described in Equation 4 (Caballero et al., 2010; Pérez-Ortiz et al., 2015). It was performed following a threefold cross-validation ($3 \times 3_{cross-validation}$) scheme (Muñoz-Mas et al., 2016), with every fold presenting a similar proportion of samples per community (*i.e.*, samples per cluster) and favoring the use of each variable in every of the nine decision trees (Equation 4).

$$Fitness = \frac{1}{9}\sum_{1}^{9}\left(\prod_{i}^{k} Sn_i \times \frac{\#\ variables\ used}{\#\ variables\ selected}\right) \quad \text{(Equation 4)}$$

Where k corresponds to the ultimate number of communities delineated by the SOM (*i.e.*, the ultimate number of different clusters). To avoid redundancy in the optimal variables' set (community indicators or environmental predictors), the number of variables was restricted by preventing correlated variable combinations (*i.e.*, r^2 > 66%) (Additional information about data correlation can be found in Appendix B).

To allow comparison with previous studies, other performance criteria were calculated, namely, accuracy or correctly classified instances, Cohen's kappa, sensitivity, specificity, and balanced accuracy (i.e. the number of correctly predicted cases weighted by the rarity of the community) (see Mouton et al., 2010 for additional details about performance criteria).

Once the optimal hyperparameters and the most relevant environmental variables or macroinvertebrate families were obtained, a single C5.0 decision tree was trained using the entire dataset (i.e. without cross-validation) (Fukuda et al., 2013; Muñoz-Mas et al., 2016). The resulting models were used to calculate the variable importance based on *usage*, which measures the percentage of training set samples that fall into all the terminal nodes after the split, and *splits*, which measures the percentage of splits associated with each variable (Fig. 2). Finally, the relationship between the environmental variables or macroinvertebrate families and the probability of occurrence of each community or cluster was scrutinized with partial dependence plots (Friedman, 2001) to accommodate the tree-like or rule-based structure of the optimal C5.0s (Fig. 2). This was done adapting the code implemented in the *randomForests* package (Liaw and Wiener, 2002).

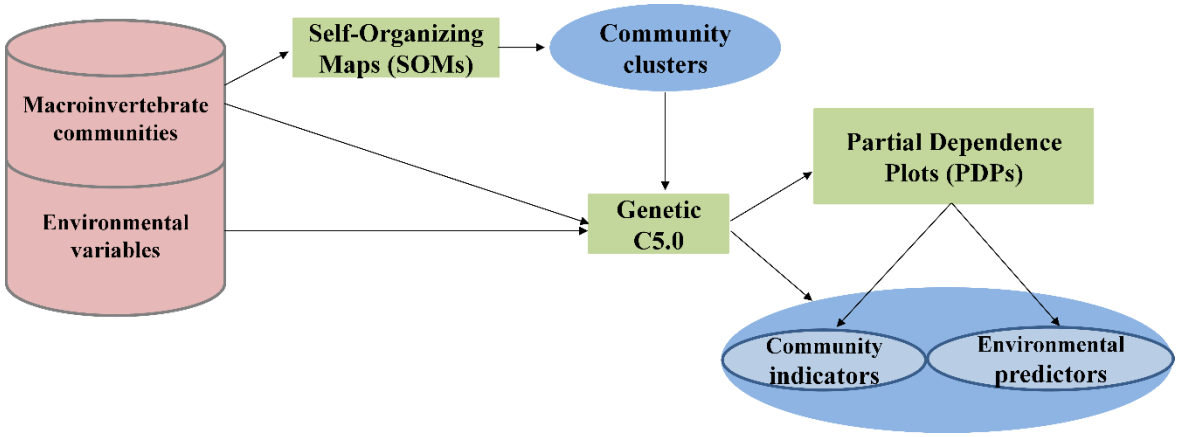

Fig. 2. Flowchart depicting the process followed to delineate the macroinvertebrate communities and to identify the family indicators and most relevant environmental drivers.

3 Results

3.1 Macroinvertebrate patterning with Self-Organizing Maps (SOMs)

The SOMs that simultaneously minimized the quantization and topographic errors (4.24 and 0.31, respectively) had a lattice of 17 x 19 neurons. This SOM presented 32.8% of empty neurons, and Ward's approach distinguished three clusters (hereafter clusters I, II and III) (Fig. 3).

Cluster I encompassed the largest number of samples (n = 178) and mainly included samples from both spring and summer. Within this cluster, there were many samples in some neurons from the LanHe River (LH), HunHe River (H) and Taizi River mainstream (T), and these neurons are located in the upper part of the SOM. Some downstream sites in the XiHe (XH), HaiChengHe (HCH) and TangHe (TH) Rivers were also located in this cluster.

Cluster II mainly contained samples from all seasons (n = 51) from sites in the Taizi Nan River (TN) and Taizi Bei River (TB), and it included most sites in the XiaoTangHe (XTH) river and some upstream sites of LanHe, XiHe Rivers.

Cluster III included many sites in the autumn and spring and some sites in the summer (n = 106). However, within this cluster, the spring sites exhibited higher similarity between each other, which

was determined based on the information shown by the neurons in the upper part of the SOM. The sites were relatively dissimilar in the summer and autumn and were further from the sites in the spring. Spatially, the sites in cluster III were mainly composed of sites in the XH, TH, XTH and HCH Rivers (especially the downstream sites), which were located in the lower reach of the Taizi River mainstream.

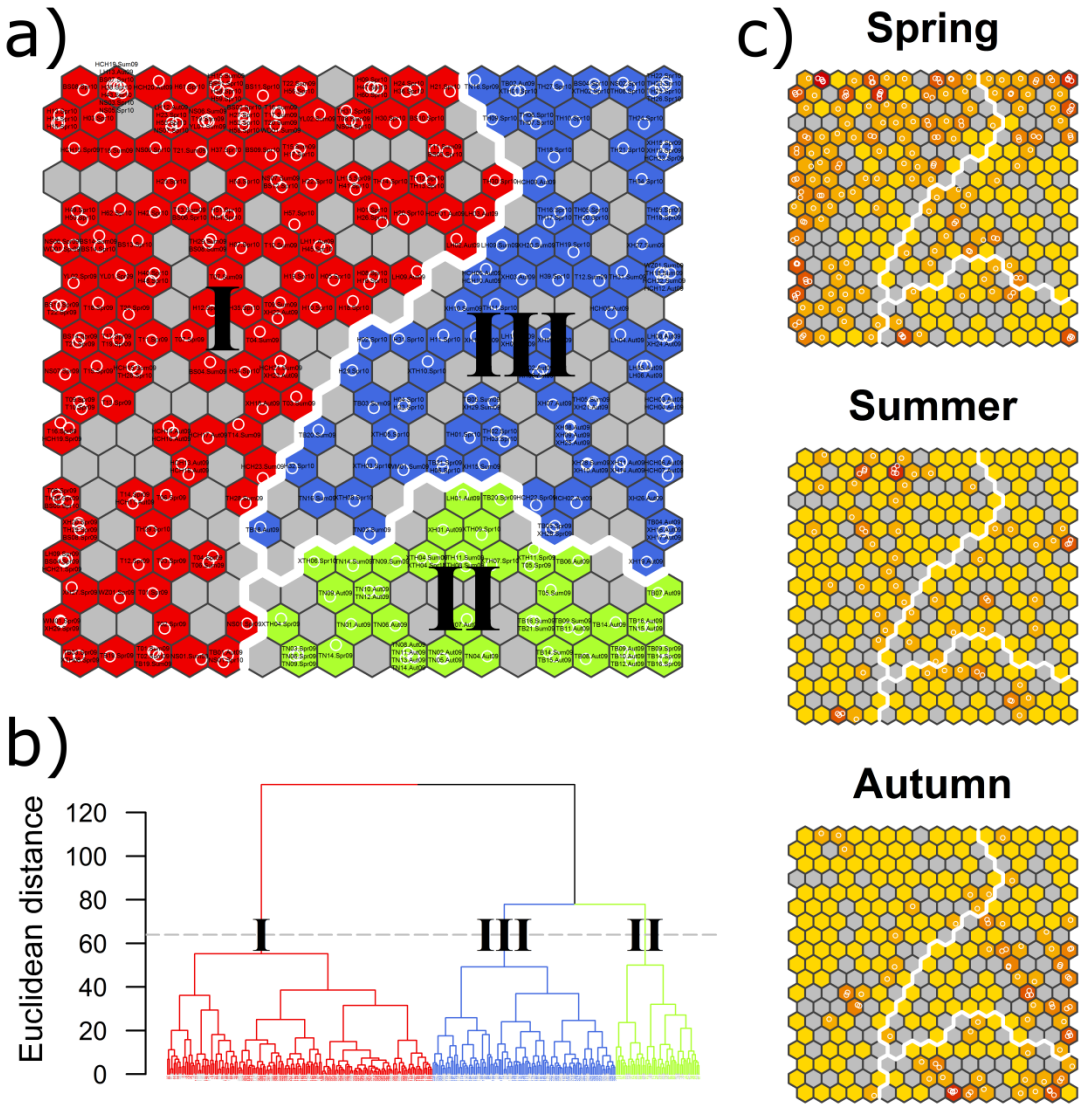

Fig. 3. a) Classification of the samples collected in Self-Organizing Map (SOM), b) dendrogram clustering of the SOM and c) mapping of the samples collected in each season (spring, summer and autumn). The darker the cell the larger the number of samples mapped in the corresponding neuron node.

3.2 Relevant Macroinvertebrate community characteristics

In the three clusters, cluster I had the lowest macroinvertebrate density and family richness whereas cluster II had the highest (Fig. 4). Shannon-Weiner diversity and Pielou evenness showed similar patterns. The most abundant families were Chironomidae, Baetidae, Hydropsychidae, Ephemerellidae and Tipulidae, while the least abundant ones were Peltoperlidae, Neoschoenobia, Corixidae, Hemiptera sp., and Cordullidae. The most abundant families reproduced the general patterns on abundance (Fig. 4). Therefore, cluster II encompassed the samples with higher densities and cluster I those with the lower.

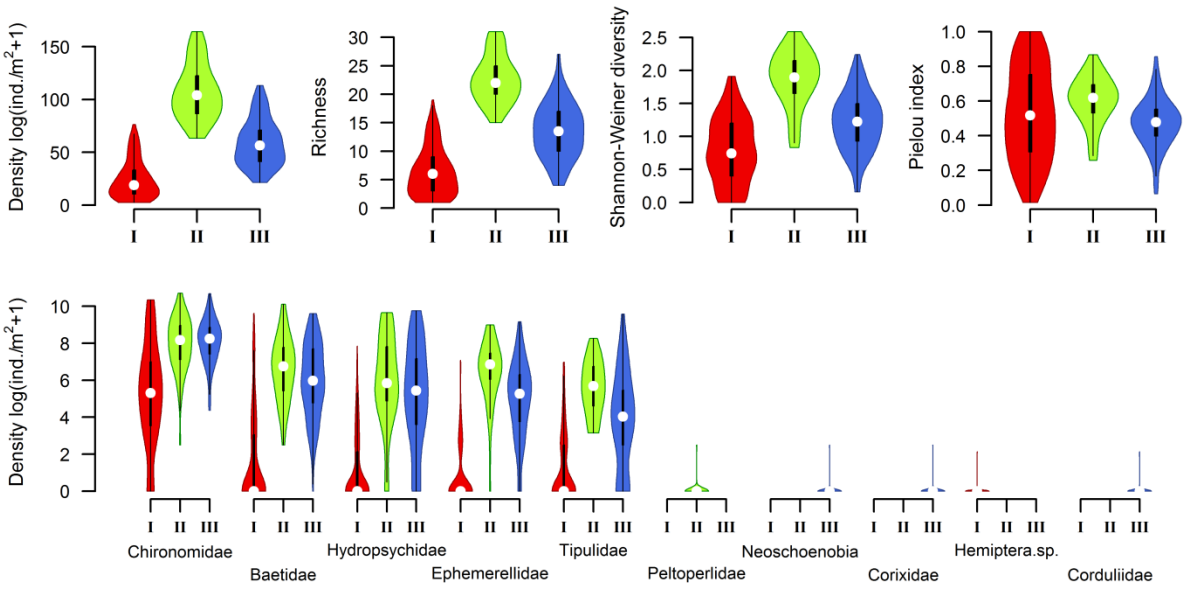

Fig. 4. Total density (log (ind./m^2+1)) and richness (families) and diversity indices (Shannon-Wiener diversity and Pielou evenness index) of the samples included in each cluster (upper sequence). The density of the most and least abundant families per cluster is depicted in the lower sequence.

3.3 Identification of the indicators and environmental predictors

The genetically optimized C5.0 model to infer the most relevant macroinvertebrate families outperformed the model to discover the main environmental predictors. However, both models

299 achieved high values for each performance criteria. In both cases, cluster II presented higher
300 performance criteria, and cluster III presented lower performance (Table 2).

301
302 Table 2. Summary of the accuracy or correctly classified instances Cohen's kappa, sensitivity,
303 specificity and balanced accuracy calculated during the $3 \times 3_{cross-validation}$ (nine models).

Classification based on	Cluster	Accuracy/CCI	Cohen's kappa	Sensitivity	Specificity	Balanced accuracy
Environmental variables	I	0.77±0.08	0.61±0.12	0.82±0.04	0.85±0.11	0.83±0.06
	II			0.88±0.19	0.94±0.03	0.91±0.09
	III			0.70±0.16	0.85±0.05	0.78±0.10
Indicator families	I	0.80±0.08	0.65±0.14	0.81±0.07	0.88±0.1	0.85±0.07
	II			0.88±0.16	0.93±0.03	0.91±0.08
	III			0.74±0.12	0.87±0.06	0.81±0.09

304

305 **3.3.1 Environmental predictors of the macroinvertebrate communities**

306 Based on the metrics *splits* and *usage*, the most relevant environmental predictors for the three
307 macroinvertebrate communities (*i.e.,* clusters) were distance from source followed by total
308 phosphorous (Fig. 5). The lesser important factors were pH and grassland whereas water
309 temperature, season, boulder and gravel occupied intermediate positions. Nevertheless, there were
310 differences in the ranges of the variables that characterized each cluster.

311 Cluster I, which principally encompassed samples that were collected in the spring and summer, was
312 characterized by high values of total phosphorous, long distance from source and intermediate water
313 temperature. The samples delineated in this community were collected in areas with low percentage
314 of boulder, high percentage of gravel and lower pH whereas the percentage of grassland had almost
315 no influence in its probability of presence.

316 Cluster II included samples collected in summer and especially in autumn. This community was
317 characteristic of river segments with low total phosphorous, distance from source and water
318 temperature. It occurred in segments with coarse substrate, high percentage of boulder and low of
319 gravel and higher pH. Finally, low percentage of grassland had a positive effect on its presence.

Cluster III included samples collected in spring and autumn and the partial dependence plots for the most relevant variables resembled those for cluster II. Therefore, this community was characteristic of river segments with low total phosphorous and distance from source. However, its presence was favored by relatively higher water temperatures. It occurred in intermediate substrate granulometries (*i.e.*, low percentage of boulder and relatively low percentage of gravel) and higher pH whereas the percentage of grassland had a positive effect on its presence.

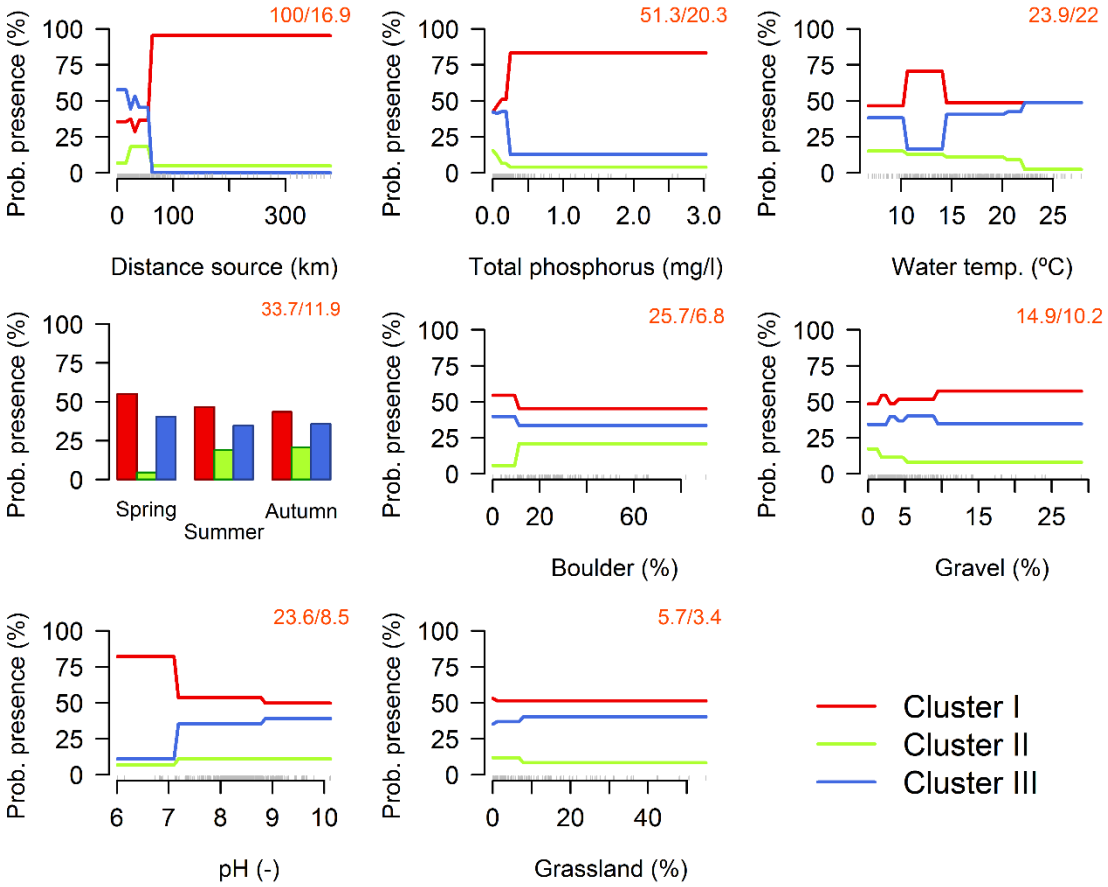

Fig. 5. Partial dependence plots relating the probabilities of the presence of each cluster and the selected environmental variables. The variable importance based on *usage*/*splits* is tagged above each panel. The tick marks near the x-axis depict the collected data.

3.3.2 Relevant indicators of the macroinvertebrate communities

The genetically optimized C5.0 algorithm found eight representative indicator families for this study area, namely, Baetidae, Stenopsychidae, Tipulidae, Dytiscidae, Tabanidae, Perlodidae Psychomyiidae, and Athericidae, (Fig. 6.). Among these families, Baetidae was the most characteristic family of the entire river basin. Cluster I was characterized by low densities of Baetidae, Stenopsychidae, Tipulidae, and Dytiscidae. For cluster II, higher densities of Stenopsychidae, Perlodidae and Athericidae were the most prominent mark. Cluster III was characterized by high densities of Baetidae, Tipulidae and Psychomyiidae, and low densities of Stenopsychidae, Perlodidae and Athericidae.

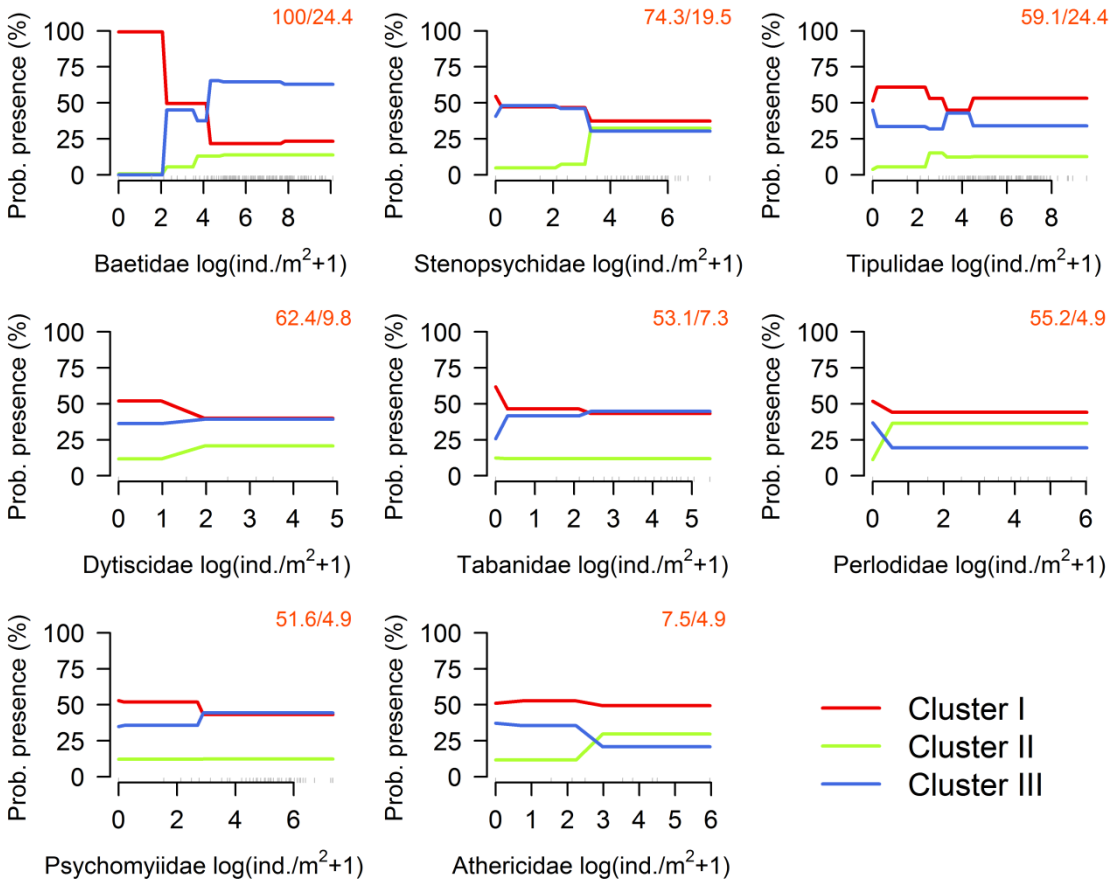

Fig. 6. Partial dependence plots relating the probabilities of the presence of each cluster and the selected community indicator families. The variable importance based on *usage*/*splits* is tagged above each panel. The tick marks near the x-axis depict the collected data.

4 Discussion

4.1 Community patterning and the environmental predictors

All study sites were aggregated into three different clusters. This number of clusters is significantly inferior when compared to previous studies (Kim et al., 2013; Park et al., 2007) but in line with the spatial and temporal extension of the study (they do not encompass multiple river basins and long periods). Given that our working dataset was fivefold larger than the datasets in previous studies, the resulting number of clusters may highlight the robustness of the analysis.

Macroinvertebrate community patterns have been usually linked to anthropogenic stress gradients (Álvarez-Cabria et al., 2011). In this study, the delineated communities (*i.e.*, clusters) portrayed a very clear spatial gradient that can be observed by comparing the sites in the three clusters. By combining the representative environmental variables, the three clusters could be defined as: minimally impacted source area, intermediately anthropogenic impacted area, and highly anthropogenic impacted downstream area, respectively. This classification was in concordance with some previous assessments in this area (Li et al., 2013; Qu et al., 2016; Zhang et al., 2013). The sites sampled in the TaiziNan and TaiziBei rivers, which were mostly encompassed in cluster II, are located in a protected area, and, in these studies, they exhibited good ecological conditions. Most of the sites in Hunhe River and the area in the tributaries and downstream of Taizi River, which were encompassed in cluster I, were determined to be in a poor ecological status (Li et al., 2013), while there were also some studies showing very low macroinvertebrate diversities in this area (Kong et al., 2013).

This gradient in community structure likely reflected the longitudinal changes in natural (gradient, temperature) and anthropogenic (water quality always deteriorated from source to downstream areas) factors that influenced the study river basin, as previous studies have shown (Álvarez-Cabria et al.,

2011; Traversetti and Scalici, 2014). The macroinvertebrate density and diversity both showed a decreasing trend with the pollution increase. This result is typical because reductions in macroinvertebrate density and diversity have been observed in many studies as a response of the benthic communities to pollution and habitat alterations (Boehme et al., 2016; Ogbeibu and Oribhabor, 2002), which is in agreement with the predicted effects first of distance from source, total phosphorous and water temperature and then on habitat quality (*i.e.*, substrate) (D'heyere et al., 2003).

In our study, distance from source and total phosphorous had arguably almost equal importance to the community classification, and to a lesser extent, water temperature, although the former one was a little more important. In the original hypothesis, we expected the geographical variables to override the local environment (e.g. water quality), because over larger scales, geographical gradient and variability appear to have stronger influence (Allan, 2004; Mykrä et al., 2007), and spatial structuring may mask the effect of the local environment on the macroinvertebrate community structures (Tonkin et al., 2017). Our results support the hypothesis to some degree. The difference is the local environment variables (e.g. water quality) were also a determinant factor for classifying the macroinvertebrate communities. Interestingly, the landuse type (a larger scale variable) just showed, apparently, minor influence to the community structure. In general terms, local habitat and biological diversity of streams and rivers are strongly influenced by landuse type within the surrounding valley at multiple scales (Allan, 2004). The percentage of the landuse type correlated well with distance from source (See Appendix B). Human settlements, and hence anthropogenic impacts (*e.g.*, diffuse pollution, landuse changes), are negatively correlated with elevation (Kummu et al., 2016), which is a general pattern also observed in the Taizi River Basin. Therefore, it is plausible to consider distance from source a proxy of these large scale processes. Then, we could conclude that the influence of local scale variables on macroinvertebrate depended on larger-scale longitudinal

394 gradient under anthropogenic pressure, which is similar with what other studies have shown
395 (Manfrin et al., 2016).

396 In addition, within each cluster, the distribution patterns of some sites showed high relevance with
397 the seasonality. The sites subjected to seasonality had higher similarities and were grouped into
398 specific neurons (Fig. 3a). In the relatively clean areas (cluster II and III), sites in different seasons
399 were grouped into different neurons. This seasonal pattern overlapped with the spatial zonation
400 pattern, which was consistent with Sroczyńska et al. (2017) for a temporary Mediterranean river.
401 Conversely, in cluster I which occupied the upper part of the SOM, the sites in the spring and
402 summer and even some sites in the autumn were concentrated in nearby neurons, indicating the high
403 similarity of the community structure at different sites and in different seasons. Kim et al. (2013)
404 showed that the variability of the macroinvertebrate density (mainly tolerant species) was very small
405 in different seasons at severely polluted sites. This result may indicate that seasonality played an
406 important role in the community patterning or sites classification in clean areas with little or
407 intermediate anthropogenic influence, while this seasonality effect was minimized by other factors in
408 severely anthropogenic influenced areas, as the case in our study area.

409 Substrate is very important to macroinvertebrates (Connolly and Pearson, 2007; Sroczyńska et al.,
410 2017). However, in this case it showed intermediate influence for the community classification. This
411 might be caused by the taxonomic level employed to delineate the communities. Thus, low-level or
412 trait-based analyses may have rendered different variable rankings (Soininen et al., 2016; Wu et al.,
413 2014). Communities are product of the environmental variables that act at multiple spatial scales
414 (Boyero and Bailey, 2001; Liu et al., 2016). Therefore, the links between the environmental variables
415 and species may be masked at this level because different species adapt to changes in the
416 environment differently by a multitude of strategies (Lamouroux et al., 2004; Resh et al., 1994).
417 However, community incorporates all of the species information and reflects all environmental
418 changes at multiple scales (Heino et al., 2002; Marchant et al., 2000; Zhang et al., 2012b), and could

provide a comprehensive reflection of the ecosystem to indicate more relevant ecoregions for river management and restoration plans.

4.2 Family indicators and the relationships with the environmental variables

As mentioned above, family level was used to identify the indicators. It was considered satisfactory to characterize the ecological status because investigations at the species level at large spatial scales are particularly costly (Heino et al., 2002; Rosenberg and Resh, 1993; Sroczyńska et al., 2017). In this regard, some researcher compared the taxonomic resolution's influence to macroinvertebrate community patterning and found that only little information (<6%) was lost using family level, as opposed to species level. And they concluded that family level abundance was a better resolution for patterning the macroinvertebrate community (Marshall et al., 2006). This also in agreement with the standard taxonomic level employed in a number classic and new rapid bioassessment protocols (*e.g.*, Kaaya et al., 2015).

Four out of the eight families that were considered as the most characteristic families (Fig. 9) in the Hun-Tai River Basin belong to the orders Ephemeroptera, Plecoptera or Trichoptera (EPT), which are widely accepted as sensitive indicators of habitat conditions (Boehme et al., 2016). Other families had also lower tolerance except Tabanidae, which is considered a family slightly tolerant to pollution (Mandaville, 2002; Qin et al., 2014).

Baetidae was the most characteristic family in our study area. A higher density of Baetidae accompanied its higher probability of presence of the minimally impacted source area (cluster II) and the intermediate impacted area (cluster III), which is in agree with other studies that performed variable selection with GA which also found Baetidae intended to live in the conditions characterized by no pollution (Gobeyn et al., 2017). In addition, this is a phenomenon observed in numerous streams (Kasangaki et al., 2006; Törnblom et al., 2011).

443 The evenness in the three clusters were all very high, implying that there might be more than one
444 representative families as indicators. This has been proven by our indicator analysis results (Fig. 6).
445 The combination of three families with low tolerance values (Mandaville, 2002; Qin et al., 2014)
446 appearing at the same time with higher density was regarded as the indicators of the clean river
447 source areas (cluster II), while lower density of four sensitive families were identified as the
448 indicators of the most anthropogenic impacted area (cluster I).

449 Usually, the presence/absence of taxa is used to delineate the community patterning or for
450 identification of the indicator species (Paini et al., 2010; Tonkin et al., 2017). Some researchers also
451 demonstrated that species assemblage patterns were adequately reproduced at the resolution of
452 family using presence/absence data (Wright et al., 1989; Rutt et al., 1990). However, our indicator
453 analysis results showed that density (or abundance) could be better in community patterning For
454 instance, Baetidae appeared in many sites of both cluster I and III, but lower abundance indicated
455 cluster I while higher abundance indicated cluster III. Additionally, some species with broad niche
456 could live under different environmental condition, e.g. *Limnodrilus hoffmeisteri*, which could live in
457 many environment (Kim et al., 2013; Song et al., 2006; Zhang et al., 2012a), but only the very high
458 abundance could indicate an organic pollution (Chapman et al., 1982). In accordance, we concluded
459 that abundance could give more biological information than presence/absence data, which is in
460 agreement with other studies that highlighted the benefits of abundance data over presence/absence
461 data (Fukuda et al., 2011; Marshall et al., 2006). Therefore, we suggest the use of abundance rather
462 than presence/absence data in patterning the macroinvertebrate assemblages, especially to identify
463 the most indicative taxa.

464

5 Conclusions

In this study, the combination of SOMs and decision trees was demonstrated to be a proficient tool for community patterning and identification of the most relevant environmental predictors and indicator taxa (*i.e.*, families). A gradient of three types of sites (communities) were distinguished: minimally impacted source area, intermediately anthropogenic impacted area, and highly anthropogenic impacted downstream area. Distance from source and total phosphorus were considered to be the most important environmental factors determining the presence of each community (*i.e.*, cluster), which indicated that local environmental factors affect macroinvertebrate community composition under the geographical gradient. At the same time, water temperature, season, substrate, pH and the percentage of grassland were also identified as distinguishing factor. These results support the hypothesis that the importance of environmental predictors are spatial-scale dependent because macroinvertebrates' presence was primarily regulated by processes operating at larger spatial scales (*i.e.*, summarized in the variable distance from source), while they responded, only subsequently, to the water quality and habitat structure.

Eight families were identified as the indicators of the community in the study area with families of the order Ephemeroptera, Plecoptera or Trichoptera (EPT) leading the ranking. Particularly, the abundance of Baetidae, a sensitive family to pollution, was the most relevant indicator for the three delineated communities. Baetidae was particularly abundant in less impacted sites, although combinations of more than one family were identified as indicators for each community (*i.e.*, cluster) in such an ecosystem with high species evenness. Abundance contributed a lot for distinguishing the indicators in different areas, so that the indicators identified in our study were the families with their density data. Therefore, we suggested to use the abundance rather than the presence/absence data in community patterning or specific zonation of an ecosystem, especially in the identification of the indicators. These results should help to improve Hun-Tai River management plans and to refine future bio-monitoring programs in the region and the similar bioclimatic river regions.

6 Acknowledgments

This work was supported by the National Natural Science Foundation of China (51779275, 41501204, 51479219) and the IWHR Research & Development Support Program (WE0145B532017).

7 References

Adriaenssens V., Verdonschot P.F.M., Goethals P.L.M., Pauw N.D., 2007.Application of clustering techniques for the characterization of macroinvertebrate communities to support river restoration management. Aquat. Ecol. 41, 387-398.

Allan, J. D., 2004. Landscapes and riverscapes: the influence of land use on stream ecosystems. Annu. Rev. Ecol. Evol. Syst. 35, 257–284.

Álvarez-Cabria M., Barquín J., Juanes J.A., 2011. Macroinvertebrate community dynamics in a temperate European Atlantic river. Do they conform to general ecological theory? Hydrobiologia 658, 277-291.

Ball G.H., Hall D.J., 1965. ISODATA, a novel method of data analysis and pattern classification. Stanford research inst Menlo Park CA.

Boehme E.A., Zipper C.E., Schoenholtz S.H., Soucek D.J., Timpano A.J., 2016.Temporal dynamics of benthic macroinvertebrate communities and their response to elevated specific conductance in Appalachian coalfield headwater streams. Ecol. Indic. 64, 171-180.

Boyero L., Bailey R.C., 2001.Organization of macroinvertebrate communities at a hierarchy of spatial scales in a tropical stream. Hydrobiologia 464, 219-225.

Brinkhurst R.O., 1986.Guide to the freshwater aquatic microdrile oligochaetes of North America. Ottawa, Canada, Dept. of Fisheries and Oceans.

Brown B.L., Swan C.M., 2010.Dendritic network structure constrains metacommunity properties in riverine ecosystems. J. Anim. Ecol. 79, 571-80.

Bu H., Meng W., Zhang Y., 2014. Spatial and seasonal characteristics of river water chemistry in the Taizi River in Northeast China. Environ. Monit. Assess. 186, 3619-3632.

Caballero J.C.F., Martínez F.J., Hervás C., Gutiérrez P.A., 2010.Sensitivity versus accuracy in multiclass problems using memetic pareto evolutionary neural networks. IEEE Trans. Neural Networks 21, 750-770.

Chapman, P.M., Farrell, M.A. and Brinkhurst, R.O., 1982. Effects of species interactions on the survival and respiration of Limnodrilus hoffmeisteri and Tubifex tubifex (Oligochaeta Tubificidae) exposed to various pollutants and environmental factors. Water Res. 16, 1405-1408.

Charrad M., Ghazzali N., Boiteau V., Niknafs A., 2014.An R Package for Determining the Relevant Number of Clusters in a Data Set. J. Stat. Softw. 61, 1-36.

Chon T. S., 2011.Self-Organizing Maps applied to ecological sciences. Ecol. Inform. 6, 50-61.

526 Connolly N.M., Pearson R.G., 2007.The effect of fine sedimentation on tropical stream macroinvertebrate
527 assemblages: a comparison using flow-through artificial stream channels and recirculating
528 mesocosms. Hydrobiologia 592, 423-438.

529 Cummins K.W., 1962.An evaluation of some techniques for the collection and analysis of benthic samples
530 with special emphasis on lotic waters. Am. Midl. Nat. 67, 477-504.

531 Dedieu N., Vigouroux R., Cerdan P., Céréghino R., 2014.Invertebrate communities delineate
532 hydro-ecoregions and respond to anthropogenic disturbance in East-Amazonian streams.
533 Hydrobiologia 742, 95-105.

534 D'heygere, T., Goethals P. L. M., De Pauw N., 2003. Use of genetic algorithms to select input variables in
535 decision tree models for the prediction of benthic macroinvertebrates. Eco. Model. 160: 291–300.

536 Edia E.O., Gevrey M., Ouattara A., Brosse S., Gourène G., Lek S., 2010.Patterning and predicting aquatic
537 insect richness in four West-African coastal rivers using artificial neural networks. Knowl. Manag.
538 Aquat. Ecosyst. 63, 78-83.

539 Friedman J.H., 2001.Greedy function approximation: a gradient boosting machine. Ann. Stat., 1189-1232.

540 Fukuda S., De Baets B., Waegeman W., Verwaeren J., Mouton A.M., 2013.Habitat prediction and knowledge
541 extraction for spawning European grayling (Thymallus thymallus L.) using a broad range of species
542 distribution models. Environ. Model. Software 47, 1-6.

543 Fukuda, S., Mouton, A.M., and De Baets, B., 2011. Abundance versus presence/absence data for modelling
544 fish habitat preference with a genetic Takagi-Sugeno fuzzy system. Environ. Monit. Assess. 184,
545 6159–6171.

546 Gaston K.J., 2000.Global patterns in biodiversity. Nature 405, 220-227.

547 Giraudel J.L., Lek S., 2001.A comparison of self-organizing map algorithm and some conventional statistical
548 methods for ecological community ordination. Ecol. Model. 146, 329-339.

549 Gobeyn, S., Volk M., Dominguez-Granda L., Goethals P. L. M., 2017. Input variable selection with a simple
550 genetic algorithm for conceptual species distribution models: A case study of river pollution in
551 Ecuador. Environ. Model. Software 92: 269–316.

552 Habersack H., Haspe D., Kondolf M., 2014.Large Rivers in the Anthropocene- Insights and tools for
553 understanding climatic, land use, and reservoir influences. Water Resour. Res. 50, 3641-3646.

554 Heino J., Muotka T., Paavola R., Hämäläinen H., Koskenniemi E., 2002.Correspondence between regional
555 delineations and spatial patterns in macroinvertebrate assemblages of boreal headwater streams. J. N.
556 Am. Benthol. Soc. 21, 397-413.

557 Kaaya, L.T., Day, J.A., and Dallas, H.F. 2015. Tanzania River Scoring System (TARISS): a
558 macroinvertebrate-based biotic index for rapid bioassessment of rivers. African J. Aquat. Sci. 40,
559 109–117.

560 Kasangaki, A., Babaasa, D., Efitre, J., McNeilage, A., and Bitariho, R. (2006) Links Between Anthropogenic
561 Perturbations and Benthic Macroinvertebrate Assemblages in Afromontane Forest Streams in Uganda.
562 Hydrobiologia 563, 231–245.

563 Kim D.-H., Cho W.-S., Chon T.-S., 2013.Self-organizing map and species abundance distribution of stream
564 benthic macroinvertebrates in revealing community patterns in different seasons. Ecol. Inform. 17,
565 14-29.

566 Kohonen T., 1982.Self-organized formation of topologically correct feature maps. Biol. Cybern. 43, 59-69.

567 Kong W., Meng W., Zhang Y., Gippel C., Qu X., 2013.A freshwater ecoregion delineation approach based on
568 freshwater macroinvertebrate community features and spatial environmental data in Taizi River
569 Basin, northeastern China. Ecol. Res. 28, 581-592.

Kummu, M., de Moel, H., Salvucci, G., Viviroli, D., Ward, P.J., and Varis, O. (2016) Over the hills and further away from coast: global geospatial patterns of human and environment over the 20th–21st centuries. Environ. Res. Lett. 11 (3), 34010.

Lamouroux N., Dolédec S., Gayraud S., 2004.Biological traits of stream macroinvertebrate communities: effects of microhabitat, reach, and basin filters. J. N. Am. Benthol. Soc. 23, 449-466.

Li Y., Xu Z., Yang X., 2013.Health assessment by using a benthic-index of biotic integrity in the Huntai River basin. J Beijing Normal Univ (Nat Sci) 49, 297-303.

Liaw A., Wiener M., 2002.Classification and regression by randomForest. R news 2, 18-22.

Liu J., Yang W., 2012.Water sustainability for China and beyond. Science 337, 649-650.

Liu S., Xie G., Wang L., Cottenie K., Liu D., Wang B., 2016. Different roles of environmental variables and spatial factors in structuring stream benthic diatom and macroinvertebrate in Yangtze River Delta, China. Ecological Indicators 61, 602-611.Mandaville S.M., 2002.Benthic macroinvertebrates in freshwaters_ Taxa tolerance values, metrics, and protocols, Soil & Water Conservation Society of Metro Halifax.

Manfrin, A., Traversetti, L., Pilotto, F., Larsen, S. and Scalici, M., 2015. Effect of spatial scale on macroinvertebrate assemblages along a Mediterranean river. Hydrobiologia, 765(1): 185-196.

Marchant R., Wells F., Newall P., 2000.Assessment of an ecoregion approach for classifying macroinvertebrate assemblages from streams in Victoria, Australia. J. N. Am. Benthol. Soc. 19, 497-500.

Marshall, J.C., Steward, A.L. and Harch, B.D., 2006. Taxonomic Resolution and Quantification of Freshwater Macroinvertebrate Samples from an Australian Dryland River: The Benefits and Costs of Using Species Abundance Data. Hydrobiologia 572, 171-194.

Mebane Jr W.R., Sekhon J.S., 2011.Genetic optimization using derivatives: the rgenoud package for R. J. Stat. Softw. 42, 1-26.

Merritt R.W., Cummins K.W., 1996.An introduction to the aquatic insects of North America, Kendall Hunt.

Milošević D., Čerba D., Szekeres J., Csányi B., Tubić B., Simić V., Paunović M., 2016.Artificial neural networks as an indicator search engine: The visualization of natural and man-caused taxa variability. Ecol. Indic. 61, 777-789.

Morse J.C., Yang L., Tian L., 1994.Aquatic insects of China useful for monitoring water quality, Hohai University Press.

Mouton, A.M., De Baets, B., Goethals, P.L.M., 2010. Ecological relevance of performance criteria for species distribution models. Ecol. Model. 221 (16), 1995–2002.

Muñoz-Mas R., Fukuda S., Vezza P., Martínez-Capel F., 2016.Comparing four methods for decision-tree induction: A case study on the invasive Iberian gudgeon (Gobio lozanoi; Doadrio and Madeira, 2004). Ecol. Inform. 34, 22-34.

Murtagh F., Legendre P., 2014.Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion? J. Classif. 31, 274-295.

Mykrä, H., Heino, J. and Muotka, T., 2007. Scale-related patterns in the spatial and environmental components of stream macroinvertebrate assemblage variation. Global Ecol. Biogeogr. 16, 149–159.

O'Brien A., Townsend K., Hale R., Sharley D., Pettigrove V., 2016.How is ecosystem health defined and measured? A critical review of freshwater and estuarine studies. Ecol. Indic. 69, 722-729.

Ogbeibu A.E., Oribhabor B.J., 2002.Ecological impact of river impoundment using benthic macro-invertebrates as indicators. Water Res. 36, 2427–2436.

Pérez-Ortiz M., Gutiérrez P.A., Hervás-Martínez C., Yao X., 2015.Graph-based approaches for over-sampling in the context of ordinal regression. IEEE Tran. Knowledge and Data Engineering 27, 1233-1245.

615 Paini, D.R., Worner, S.P., Cook, D.C., De Barro, P.J., Thomas, M.B., 2010. Using a self-organizing map to
616 predict invasive species: sensitivity to data errors and a comparison with expert opinion. J. Appl. Ecol.
617 47, 290–298.

618 Paliy O., Shankar V., 2016.Application of multivariate statistical techniques in microbial ecology. Mol. Ecol.
619 25, 1032-1057.

620 Park Y.S., Céréghino R., Compin A., Lek S., 2003.Applications of artificial neural networks for patterning
621 and predicting aquatic insect species richness in running waters. Ecol. Model. 160, 265-280.

622 Park Y.-S., Song M.-Y., Park Y.-C., Oh K.-H., Cho E., Chon T.-S., 2007.Community patterns of benthic
623 macroinvertebrates collected on the national scale in Korea. Ecol. Model. 203, 26-33.

624 Qin C.Y., Zhou J., Cao Y., Zhang Y., Hughes R.M., Wang B.X., 2014.Quantitative tolerance values for
625 common stream benthic macroinvertebrates in the Yangtze River Delta, Eastern China. Environ.
626 Monit. Assess. 186, 5883-5895.

627 Qu X., Zhang H., Zhang M., Liu M., Yu Y., Xie Y., Peng W., 2016.Application of multiple biological indices
628 for river health assessment in northeastern China. Annales de Limnologie - International Journal of
629 Limnology 52, 75-89.

630 Quinlan, J.R., 1992. C4.5: programs for machine learning. Morgan Kaufmann Publishers, Inc., San Mateo,
631 CA (USA).Resh V.H., Hildrew A.G., Statzner B., Townsend C.R., 1994.Theoretical habitat templets,
632 species traits, and species richness: a synthesis of long-term ecological research on the Upper Rhone
633 River in the context of concurrently developed ecological theory. Freshwat. Biol. 31, 539-554.

634 Rosenberg D.M., Resh V.H., 1993.Freshwater biomonitoring and benthic macroinvertebrates. New York,
635 Chapman and Hall.

636 Rutt, G.P., Weatherley, N.S. and Ormerod, S.J., 1990. Relationships between the physicochemistry and
637 macroinvertebrates of British upland streams: the development of modelling and indicator systems for
638 predicting fauna and detecting acidity. Freshwat. Biol. 24(3), 463-480.

639 Siddig A.A.H., Ellison A.M., Ochs A., Villar-Leeman C., Lau M.K., 2016.How do ecologists select and use
640 indicator species to monitor ecological change? Insights from 14 years of publication in Ecological
641 Indicators. Ecol. Indic. 60, 223-230.

642 Song, M.-Y., Park, Y.-S., Kwak, I.-S., Woo, H. and Chon, T.-S., 2006. Characterization of benthic
643 macroinvertebrate communities in a restored stream by using self-organizing map. Ecol. Inform. 1(3):
644 295-305.

645 Soininen, J., Jamoneau, A., Rosebery, J. and Passy, S.I., 2016. Global patterns and drivers of species and trait
646 composition in diatoms. Global Ecol. Biogeogr. 25, 940-950.

647 Sroczyńska K., Claro M., Kruk A., Wojtal-Frankiewicz A., Range P., Chícharo L., 2017.Indicator
648 macroinvertebrate species in a temporary Mediterranean river: Recognition of patterns in binary
649 assemblage data with a Kohonen artificial neural network. Ecol. Indic. 73, 319-330.

650 Traversetti, L. and Scalici, M., 2014. Assessing the influence of source distance and hydroecoregion on the
651 invertebrate assemblage similarity in central Italy streams. Knowl. Manag. Aquat. Ecosyst. 414, 02.

652 Tibshirani R., Walther G., Hastie T., 2001. Estimating the number of clusters in a data set via the gap statistic.
653 J. Roy. Stat. Soc. Ser. B. (Stat. Method.) 63, 411-423.

654 Tonkin J.D., Tachamo Shah R.D., Shah D.N., Hoppeler F., Jähnig S.C., Pauls S.U., 2017.Metacommunity
655 structuring in Himalayan streams over large elevational gradients: the role of dispersal routes and
656 niche characteristics. J. Biogeogr. 44, 62-74.

657 Törnblom, J., Roberge, J.-M., and Angelstam, P., 2011. Rapid assessment of headwater stream
658 macroinvertebrate diversity: an evaluation of surrogates across a land-use gradient. Fundam. Appl.
659 Limnol. 178, 287–300.

Tsai W.P., Huang S.P., Cheng S.T., Shao K.T., Chang F.J., 2017.A data-mining framework for exploring the multi-relation between fish species and water quality through self-organizing map. Sci. Total Environ. 579, 474-483.Vörösmarty C.J., McIntyre P.B., Gessner M.O., Dudgeon D., Prusevich A., Green P.,Glidden S., Bunn S. E., Sullivan C.A., Reidy Liermann C.& Davies P.M., 2010. Global threats to human water security and river biodiversity. Nature467, 555–561.

Wehrens R., Buydens L.M., 2007.Self-and super-organizing maps in R: the Kohonen package. J. Stat. Softw. 21, 1-19.

Wright, J.F., Armitage, P.D., Furse, M.T. and Moss, D., 1989. Prediction of invertebrate communities using stream measurements. River. Res. App. 4(2), 147-155.

Wiggins G.B., 1996.Trichoptera families. An introduction to the aquatic insects of North America, 309-349.

Wu, N., Qu, Y., Guse, B., Makarevičiūtė, K., To, S., Riis, T. and Fohrer, N., 2018. Hydrological and environmental variables outperform spatial factors in structuring species, trait composition, and beta diversity of pelagic algae. Ecol. Evol. 8, 2947-2961.

Zhang M., Cai Q., Xu Y., Kong L., Tan L., Wang L., 2012a.Spatial Distribution of Macroinvertebrate Community along a Longitudinal Gradient in a Eutrophic Reservoir- Bay during Different Impoundment Stages, China. Int. Rev. Hydrobiol. 97, 169-183.

Zhang M., Cai Q., Xu Y., Wang L., Kong L., 2012b.Zonation and its influencing factors of a large subtropical reservoir (Danjiangkou Reservoir in central Chian), based on macroinvertebrates. Fresenius Environ. Bull. 21, 2095-2104.

Zhang Y., Guan D., Jin C., Wang A., Wu J., Yuan F., 2011.Analysis of impacts of climate variability and human activity on streamflow for a river basin in northeast China. J. Hyd. 410, 239-247.

Zhang Y., Zhao R., Kong W., Geng S., Bentsen C.N., Qu X., 2013.Relationships between macroinvertebrate communities and land use types within different riparian widths in three headwater streams of Taizi River, China. J. Freshwat. Ecol. 28, 307-328.

Supplementary Appendix A

Click here to download Supplementary material for on-line publication only: Appendix A.docx

Supplementary Appendix B

Click here to download Supplementary material for on-line publication only: Appendix B.docx