Ecological Indicators 11 (2011) 582–589

Contents lists available at ScienceDirect

Ecological Indicators

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

Original article

Which index is the best to assess stream health?

D.A. Dos Santos*, C. Molineri, M.C. Reynaga, C. Basualdo

CONICET-Facultad de Ciencias Naturales e Instituto Miguel Lillo, Universidad Nacional de Tucumán, Miguel Lillo 205, San Miguel de Tucumán, C.P. 4000, Tucumán, Argentina

ARTICLE INFO

Article history:
Received 28 February 2010
Received in revised form 2 June 2010
Accepted 22 August 2010

Keywords:
ROC methodology
Biotic index
BMWP
EPT
Yungas
Benthic macroinvertebrates

ABSTRACT

Biotic indices are widely used in monitoring the health status of various ecosystems. The choice of the best index is generally done qualitatively depending on a variety of aspects including cost and time. ROC (Receiver Operating Characteristic) methodology constitutes a valuable tool to compare objectively the diagnostic capabilities of different tests in addition to obtain decision thresholds. In this manuscript, ROC methodology is described and implemented for the first time in the context of stream bioassessment through benthic macroinvertebrates. Cut-off values that distinguish impaired from healthy sites are suggested. A new index called IBY-4 is also developed. IBY-4 accounts for the occurrence of Megaloptera, Plecoptera, Trichoptera and Elmidae in a target site and may achieve the best general performance in the study region concerning to Andean Tropical streams.

1. Introduction

Freshwater ecosystems are one of the most endangered of the world. Moreover, the natural services they provide (mainly water) and the biodiversity they support are also threatened (UNESCO, 2009). Worldwide anthropic disturbances include channelization of stream bottom, dams, removal of riparian trees, wastewaters allocation, replacement of native forests by grasslands along the watershed and invasive species. Habitat transformation followed by biodiversity loss constitutes a consequence associated to them. Thus, the frequent supervision of the ecosystem integrity represents a priority task for water resource management. In this context, biotic indices based on aquatic macroinvertebrates have been developed as one type of diagnostic test of ecosystem integrity (for review see Bonada et al., 2006). The main premise underlying biotic indices development is that an assessment of stream integrity (and water quality) could be achieved by evaluating the community structure.

Worldwide experience has demonstrated that the most useful biological assessment methods for freshwater monitoring are based on benthic macroinvertebrates (Sivaramakrishnan, 2000). An extensive literature on this topic is available (e.g., Rosenberg and Resh, 1993; Chessman and McEvoy, 1998; Reynoldson et al., 2001; Resh, 2008). Alleged reasons are ubiquity, susceptibility to disturbances, large number of species that offers a spectrum of responses to perturbations, accessibility, inexpensive equipment for sampling, etc. (Resh, 1995). This monitoring strategy has been also incorporated to South America: e.g. Argentina (Domínguez and Fernández, 1998; Rodrigues Capítulo et al., 2001), Brasil (Junqueira and Campos, 1998), Chile (Figueroa et al., 2003, 2007), Colombia (Roldán, 1999) but with much effort devoted to adapt tolerance values or suggest the most suitable biotic index for each region (Prat et al., 2009). Fernández et al. (2002) and Von Ellenrieder (2007) represent the latest contributions studying relationships between macrobenthos and environmental variables associated to basin disturbance in the study area.

Implementation of control and protection policies should be based on indices of proven reliability. Such reliability refers to the ability of the index to detect the correct status about the health of the assessed environment and has been commonly evaluated qualitatively (e.g. Bonada et al., 2006). Nonetheless, studies that compare the performance of different biotic indices providing a statistical significance of their results are much rarer (Barbour et al., 1996; Murtaugh, 1996; Hale et al., 2004; Hale and Heltshe, 2008; Sánchez-Montoya et al., 2010).

The accuracy of a biotic index can be calculated by comparing the results of the test to the true health status of the ecosystem. True status has to be determined with reference standard procedures (chemical analyses, analysis of disturbance in the basin, etc.). To compare different biotic indices is necessary to know the following accuracy ratios: sensitivity (number of true positive predictions vs. number of actually positive cases) and specificity (number of true negative predictions vs. number of actually negative cases). The Receiver Operating Characteristic (ROC) curve is a plot of sensitivity (y coordinate) versus 1 − specificity (x coordinate). ROC curves are graphic tools especially suitable for evaluating diagnostic tests because they capture the trade-off between sen-

* Corresponding author.
E-mail addresses: dadossantos@csnat.unt.edu.ar, pseudalopex_79@yahoo.com (D.A. Dos Santos).

1470-160X/$ – see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ecolind.2010.08.004

Table 1
Two-by-two confusion matrix.

		Stream actual status	
		Perturbed (+)	Healthy (−)
Stream predicted status	Perturbed (+)	True positive (TP)	False positive (FP)
	Healthy (−)	False negative (FN)	True negative (TN)

sitivity and specificity over the range of test values (Lasko et al., 2005).

ROC curves have been applied to many disciplines, including medicine (e.g. Lusted, 1971), industrial quality control (Drury and Fox, 1975) and estuarine ecology (Hale et al., 2004; Hale and Heltshe, 2008). To our knowledge, this work represents the first contribution to the use of ROC methodology in the context of freshwater bioassessment through benthic macroinvertebrates.

The general aim of this article is to introduce basic concepts of the ROC methodology to an audience interested on freshwater biomonitoring and to emphasize its role in the appraisal of biological index performance. Specific objectives include the use of ROC curves (1) to compare the diagnostic capabilities of some widely used metrics in addition to a new index (IBY-4) applied on a large data set from Tropical Andes streams; (2) to identify thresholds of decision for those indices in order to be used in biomonitoring programs; and (3) to analyze the response of different indices to increasing levels of perturbation.

2. Materials and methods

2.1. ROC methodology

A stream is considered perturbed if it receives some anthropic impact directly on it (e.g., water chemistry or channel shifts) or on surrounding areas (e.g., riparian or watershed area denudation) to the extent of impairing the stream capability to hold a biodiversity otherwise different at the pristine condition. Basically, biomonitoring aims to determine if a given stream should be considered perturbed or not. This corresponds to a classification problem using only two classes. Formally, each instance (stream) is mapped to one element of the set {+, −} of positive (perturbed) and negative (non-perturbed) class labels (Fawcett, 2005). Classifiers are used to predict the membership of items to one of the two alternative classes. Biological metrics are classifiers that may surrogate expensive and time consuming procedures to assess the stream quality (Cullen, 1990). However, the outputs of these metrics are not single scores, they span over a range of values to which different thresholds may be applied to predict class membership. We are interested in achieving good predictions, i.e. the predicted class should agree with the actual class of stream perturbation.

Sensitivity and specificity. In dealing with predictive tasks, there are four possible outcomes: (1) *true positive*, when a perturbed stream is correctly classified; (2) *false positive*, when a healthy stream is considered an altered one; (3) *true negative*, when a preserved stream is assigned to the right class; (4) *false negative*, when a damaged stream is wrongly mapped to the non-perturbed class. The counts of correct yes-forecasts and false alarms can be arranged into a two-by-two confusion matrix (Table 1).

We will focus on two ratios, viz. the *True Positive Rate* (TPR) and the *False Positive Rate* (FPR). TPR denotes the proportion of perturbed streams correctly predicted: TPR = TP/(TP + FN); whereas FPR concerns to the proportion of negatives incorrectly classified: FPR = FP/(FP + TN). Sensitivity is equivalent to the TPR score, while specificity refers to 1 − FPR, that is the proportion of negatives correctly classified: 1 − FPR = TN/(FP + TN). Sensitivity and specificity are the basic measures of accuracy of a diagnostic test (Obuchowski, 2003); for our purposes, they describe the ability of a biological metric to correctly diagnose perturbation when perturbation is actually present and to correctly dismiss perturbation when it is truly absent.

ROC plot. Biological metrics yield a range of values rather than a dichotomous response. One strategy for obtaining binary predictions is to select a cut point and record the cases lying above and below that point. Nevertheless, the choice of a unique cut point is an arbitrary procedure that blurs the information contained in the data. As the cut point changes, specificity and sensitivity shifts (Obuchowski, 2003). A fruitful alternative is to explore the entire range of values, calculating for each possible cut point the respective sensitivity/specificity pair. The graphical display of all those pairs connected by segment lines, with sensitivity and 1 − specificity plotted on the *y* and *x* axes respectively, is known as the empirical ROC curve. Table 2 shows a workable example with the scores provided by a hypothetical metric *H* applied on 10 streams (5 perturbed and 5 non-perturbed) to illustrate how to construct the respective ROC curve (Fig. 1a). It should be considered that the true health status (gold standard) has to be fixed in a first stage of analysis and the diagnostic performance of the test has to be evaluated afterwards.

Observe that the lower the score of *H* metric, the higher the chance of predicting a positive result. A cut point at each value of *H* is established. Thus, for example, the predictions under the first criterion (i.e. perturbed if $H < 1$, otherwise non-perturbed) yield 0 for the sensitivity and 1 for the specificity, that is the point (0, 0) in the

Table 2
Hypothetical data illustrating ROC analysis. Streams actually perturbed are coded 1, otherwise they are coded 0. Values are given for an imaginary diagnostic metric called *H*. The random *H* metric is obtained via randomization of vector *H*. ROC analysis is performed below the table. For each decision threshold, 1 − specificity (1 − Spe) and sensitivity (Sen) values have been calculated. The performance of each metric can be evaluated through the respective ROC curves in Fig. 1.

	Item (status)	*H* metric Value	Random *H* Value
Raw data	Stream A (1)	1	10
	Stream B (1)	3	14
	Stream C (1)	5	1
	Stream D (1)	7	8
	Stream E (1)	10	5
	Stream F (0)	8	3
	Stream G (0)	9	12
	Stream H (0)	12	9
	Stream I (0)	14	16
	Stream J (0)	16	7
	Cut point	***H* metric (1 − Spe)/Sen**	**Random *H* (1 − Spe)/Sen**
ROC analysis	<1	0/0	0/0
	≤1	0/0.2	0/0.2
	≤3	0/0.4	0.2/0.2
	≤5	0/0.6	0.2/0.4
	≤7	0/0.8	0.4/0.4
	≤8	0.2/0.8	0.4/0.6
	≤9	0.4/0.8	0.6/0.6
	≤10	0.4/1	0.6/0.8
	≤12	0.6/1	0.8/0.8
	≤14	0.8/1	0.8/1
	≤16	1/1	1/1