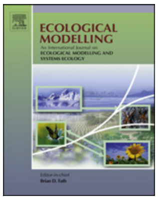
Ecological Modelling 220 (2009) 2393–2406

Contents lists available at ScienceDirect

Ecological Modelling

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



Instream and offstream environmental conditions and stream biotic integrity Importance of scale and site similarities for learning and prediction

David Bedoya[a,*,1], Vladimir Novotny[a,2], Elias S. Manolakos[b,3]

[a] *Civil & Environmental Engineering Department, Northeastern University, 400 Snell Engineering Center, 360 Huntington Avenue, Boston, MA 02115, USA*
[b] *Department of Informatics & Telecommunications, National and Kapodistrian University of Athens, Panepistimiopolis, Ilissia, Athens 15784, Greece*

ARTICLE INFO

Article history:
Received 11 June 2008
Received in revised form 3 May 2009
Accepted 3 June 2009
Available online 17 July 2009

Keywords:
Index of Biotic Integrity (IBI)
k-Nearest neighbors
Self-organizing maps (SOM)
Environmental stressors
Multiple regression models

ABSTRACT

Two different methods to predict biotic integrity were tested and compared in the present paper. The first one tries to predict the fish indices of biotic integrity (IBI) at the state or regional scale based on the most similar observations to a specific target site of interest using the simple to implement *k-nearest neighbors* (or kNN) method. Two different distance functions were considered to find the *k*-nearest neighbors: the Euclidean and the Mahalanobis. The second method was applied on the same datasets and consisted of a step-wise multiple regression. The two modeling approaches yielded similar results but kNN proved to be more time-efficient and very fast computationally for the given dataset sizes, which allowed applying a leave-one-out cross validation.

In an attempt to reveal the importance of scale in the prediction of IBI, regression models were constructed at the state (or regional) scale and at the more refined cluster of sampling sites scale. Clusters of sites were extracted using Kohonen's self-organizing maps (SOM) followed by *k*-means clustering of the SOM neurons. Cluster-level regression models, constructed after site patterning, performed better in IBI prediction than global regression models constructed without any previous site patterning. The importance of identifying groups of sites with similar environmental characteristics affecting the IBI was revealed. The combined use of site patterning and regression modeling for IBI prediction also helped identifying important variables acting at the local scale which remain latent at the global scale.

1. Introduction

The main objective of the U.S. Clean Water Act of 1972, is *to restore, and maintain the chemical, physical, and biological integrity of the Nation's waters.* Integrity is defined as the ability to support and maintain a balanced integrated, adaptive community of organisms comparable to that of a natural biota of the region (Karr and Dudley, 1981; Karr et al., 1986). This objective highlights the multidimensionality of the aquatic system; physical, chemical and biological. For many years, most of the emphasis was given to the chemical quality of receiving waters. Great progress in the control of point sources was achieved between 1980s and the end of the last century by using the *National Pollution Discharge Elimination System* (NPDES) permits. Even though the NPDES applies only to some diffuse sources, its application is far more challenging because these are driven by difficult-to-measure parameters such as land use, atmospheric conditions, and sub terrain sources and are affected by meteorological events. Urban or feedlot runoff events are just two examples (Novotny and Olem, 1994).

Even though the dynamism and effects of the highly intertwined and mutually dependant stressors on the biological community (and therefore, on stream health) have been widely studied and proved, the regulatory system has failed to keep pace with this reality and continues to neglect it in fresh water systems. Oversimplification, lack of multidisciplinary work, and patchiness of the U.S. regulatory system have been major impediments in developing a new framework that incorporates this whole spectrum of the problem (Karr, 1991). This becomes more evident when the sources of impairment originate from diffuse/non-point sources. It is very challenging to identify a steady state between land use changes, non-point source pollution, and biotic integrity. Land use changes can result in drastic changes in the water quality (i.e. fertilizers and pesticides transported by runoff), hydrologic regime (i.e. increase of peak flows in paved areas), or increase in siltation

---

[*] Corresponding author. Tel.: +1 617 2187147; fax: +1 617 3734419.
E-mail address: dbedoya@coe.neu.edu (D. Bedoya).

[1] Present address: Office 431, Civil & Environmental Engineering Department, Northeastern University, 400 Snell Engineering Center, 360 Huntington Avenue, Boston, MA 02115, USA.

[2] Present address: Office 429, Civil & Environmental Engineering Department, Northeastern University, 400 Snell Engineering Center, 360 Huntington Avenue, Boston, MA 02115, USA.

[3] Present address: Office B23, Department of Informatics & Telecommunications, National and Kapodistrian University of Athens, Panepistimiopolis, Ilissia, Athens 15784, Greece.

0304-3800/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.ecolmodel.2009.06.017

due to denudation processes that subsequently impair stream habitat. It is highly unlikely that a sole stressor (e.g. a point source) would be responsible for the whole degradation of a water body. Therefore, many other sources of pollution remain unregulated.

Chemical water quality plays a key role in the overall integrity of a stream. However, this only represents one out of its five main integrity components in fresh water systems. Habitat structure and quality, flow regime, energy sources, and biotic interactions are the other four (Karr et al., 1986). System alterations which produce a deviation from the reference state in any of these constituents will have an impact in the biological community due to the propagation of stresses through the stressor hierarchy until the system's endpoint (the biological community) is reached (Karr et al., 1986; Novotny, 2003). The state of a system is usually assessed with environmental indicators, which can be categorized as stressors, exposure and response indicators (Yoder and Rankin, 1998; Yoder et al., 2000). Stressors may include point and non-point loading, land use changes, stream modifications, and other large scale influences that generally result from anthropogenic activities. Exposure indicators include chemical concentrations, whole effluent toxicity, tissue residue, sediment contamination, habitat degradation and other parameters that result in a risk for the biota. A biotic assessment endpoint or response indicator is a direct measure of ecological integrity of a water body because the biota is the point of convergence for all the stresses throughout the ecosystem (Yoder and Rankin, 1998; Yoder et al., 2000; Novotny et al., 2005).

Biological indices are nowadays regarded as valid overall stream health response indicators. However, in the past, most of the biological indices were based on the macroinvertebrate or microscopic community and designed to identify impacts from specific pollutants, especially from organic matter. Because stream degradation may originate from many different sources of pollution, these type of indices were hardly criticized as valid indicators of overall stream health (Doudoroff and Warren, 1957). Valid environmental and biodiversity indicators should be sensitive enough to track changes from reference conditions, applicable in large geographic areas, capable of providing a continuous assessment over a wide range of stress, and differentiate between natural cycles or trends and anthropogenic stress (Ott, 1978). These indices need to be calibrated within homogeneous environmental and geographic units (e.g. ecoregions) because reference species composition, distribution, and richness (i.e. biological integrity) differ between regions. Therefore, there is a need to define what truly ecological health is within each ecological context in order to have truly reliable indicators of streams's overall health (Karr, 1991).

In work by Doudoroff and Warren (1957) fish populations were already targeted as potential indicators of overall ecological conditions in fresh water bodies. Nowadays, one of the most widely used and accepted indices of stream's integrity in the United States is the fish Index of Biological Integrity (IBI) developed by Karr et al. (1986). Karr's IBI is currently accepted as a sensitive index to human impacts and applied successfully to aquatic communities (Richards et al., 1996; Roth et al., 1996; Dyer et al., 1998; Lammert and Allan, 1999; Dyer et al., 2000; Wang et al., 2001; Karr and Yoder, 2004; Yuan and Norton, 2004; Yoder et al., 2005; Manolakos et al., 2007). Many public agencies have adopted it as a framework for their own calibrated version at the state or regional scales (Ohio EPA, 1987; Bode, 1988; Roth et al., 1998; Niemela and Feist, 2000; Lyons et al., 2001; Lyons, 2006). One of the main advantages of biological approaches is that they have environmental memory, i.e. they can reflect impacts that occurred in the past or might reflect impacts which would go unnoticed in standard, periodic water

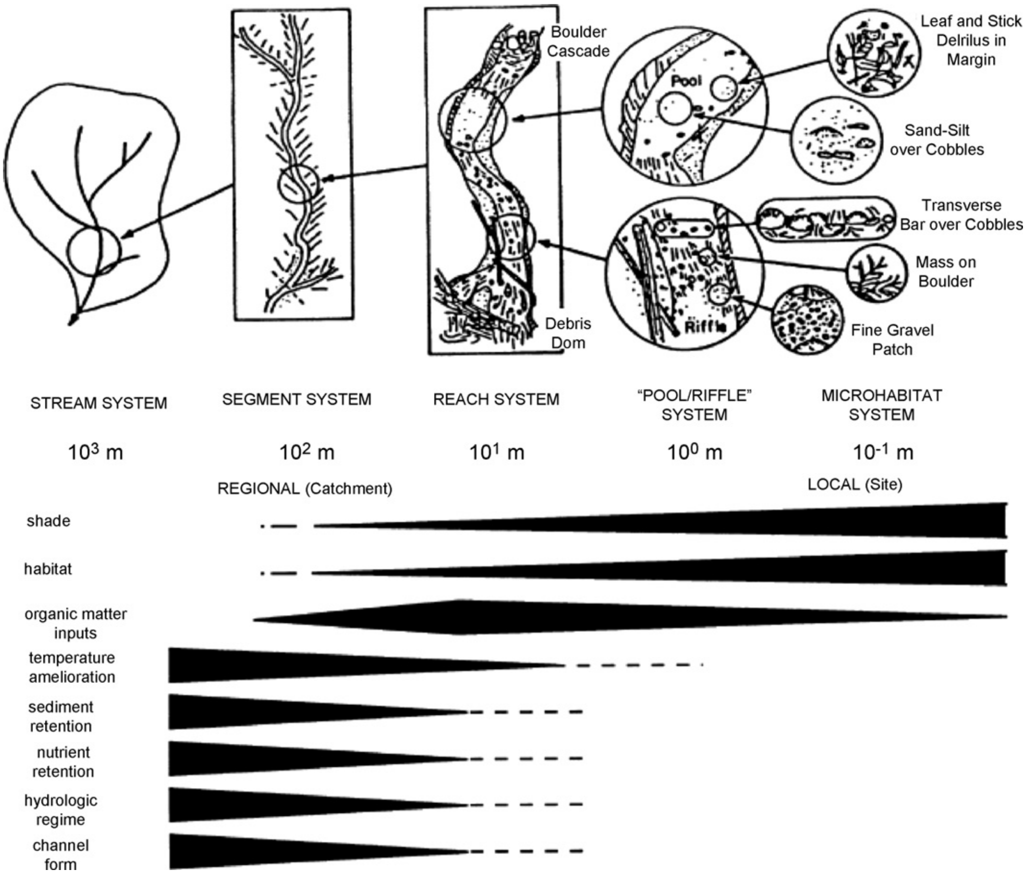


Fig. 1. Landscape influences structure and function across spatial scale. The thickness of the black stripes is proportional to the importance at each scale (adapted from Allan et al. (1997)).