Ecological Modelling 345 (2017) 21–29

Contents lists available at ScienceDirect

Ecological Modelling

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

A Bayesian network assessment of macroinvertebrate responses to nutrients and other factors in streams of the Eastern Corn Belt Plains, Ohio, USA

Douglas B. McLaughlin[a,*], Kenneth H. Reckhow[b]

[a] National Council for Air and Stream Improvement, P.O. Box 40, Mattawan, MI, 49071, USA
[b] Professor Emeritus, Duke University, Box 90328, Durham, NC, 27708, USA

ARTICLE INFO

Article history:
Received 7 April 2015
Received in revised form
23 November 2016
Accepted 7 December 2016
Available online 14 December 2016

Keywords:
Bayesian network models
Nutrients
Water quality
Macroinvertebrates
Habitat quality
Cause and effect

ABSTRACT

Over the past several years, the United States Environmental Protection Agency has urged states to adopt numeric nutrient criteria to protect water quality. In a number of states, new numeric nutrient criteria have incorporated both a nutrient (nitrogen and/or phosphorus) criterion and a biological endpoint (e.g., chlorophyll *a* in lakes or benthic macroinvertebrates in streams). While the causal relationship between nutrient levels and chlorophyll in lakes is well-established, quantifying causal relationships between nutrients, chlorophyll a, and benthic macroinvertebrate communities in rivers and streams has been more elusive. This is especially true in highly agricultural ecoregions of the upper midwest United States where a number of confounding factors may be present. Predictive relationships derived from field-collected data can provide important support for setting numeric criteria and identifying management alternatives that can achieve and sustain water quality goals. In this paper, we examine the empirical basis for a causal relationship between nutrients, chlorophyll, and benthic macroinvertebrates in the context of other potential macroinvertebrate stressors in a highly agricultural region. We developed a Bayesian network (BN) model for the number of Ephemeroptera, Plecoptera, and Trichoptera taxa (EPT) present in streams of the Eastern Corn Belt Plains Ecoregion of Ohio using Ohio Environmental Protection Agency data collected over roughly a decade (2005–2013). For the data evaluated in this study, useful relationships between nutrients, chlorophyll a, and EPT were not found. An alternative BN model including total Kjeldahl nitrogen with habitat quality and dissolved oxygen appeared to reflect stronger causal influences on this indicator of stream macroinvertebrate quality. However, the predictive power of the BN was relatively low, suggesting that other factors not accounted for in the model may contribute significantly to EPT taxa abundance in these streams.

1. Introduction

The National Research Council review of the TMDL program (NRC, 2001) recommended that water quality criteria be positioned as closely as possible to the biological (or human health) response in the causal chain. For nutrient criteria, this means that a measure of algal density (e.g., chlorophyll *a*), submerged aquatic vegetation, and/or macroinvertebrate indices should serve to augment or replace phosphorus and nitrogen criteria. As two relatively recent USEPA Science Advisory Board reviews have made clear, the presence of an underlying cause and effect in the stressor-response relationship is critical to the effectiveness of such water quality criteria (USEPA, 2010, 2011). Yet there is little guidance on how to develop sufficient evidence to support cause-effect conclusions. There has been increased emphasis in recent years on the use of field-collected data to support the development of numeric criteria or benchmark values that can help guide water quality management decisions and mitigate ecosystem risks (Cormier and Suter, 2013; Suter and Cormier, 2008). Furthermore, there is considerable interest in quantifying the impacts of increasing nutrient concentrations on the impairment of beneficial uses in streams and rivers in order to establish numeric nutrient criteria that are meaningful in the context of other lotic ecosystem stressors (Hieskary and Bouchard, 2015; Qian and Miltner, 2015; Miltner, 2010; Reckhow et al., 2005; Dodds, 2007). However, there remains a considerable need for more information on the use of field-based methods to support causal inference in complex aquatic ecosystems where

* Corresponding author.
E-mail addresses: dmclaughlin@ncasi.org (D.B. McLaughlin), Reckhow@duke.edu (K.H. Reckhow).

http://dx.doi.org/10.1016/j.ecolmodel.2016.12.004
0304-3800/

biological conditions of interest may be influenced by many environmental factors. Without such information, there is an increased likelihood that water quality criteria will be proposed or established in regulations without sufficient basis for establishing cause and effect, possibly resulting in over-/under- protective, ineffective criteria.

One of the approaches that has been used increasingly to model environmental ecosystems involves Bayesian Network (BN) models (Aguilera et al., 2011; Barton et al., 2012). BN models (Pearl, 2009) form the statistical and probabilistic basis for many recent studies addressing causal inference for multivariate observational datasets (Reckhow, 2010). Numerous applications of BN models to address a range of water resources questions are now present in the literature, including stream ecosystem risks (Van Looy et al., 2015; Death et al., 2015; Adriaenssens et al., 2004), fisheries management (Rahikainen et al., 2014), river restoration (Borsuk et al., 2012), groundwater management (Farmani et al., 2012; Henriksen et al., 2012), forest management (Ayre and Landis, 2012), and applications involving geographic information systems (Johnson et al., 2011). Several papers provide reviews of and information on best practices for the use of BN models, including Marcot et al. (2006), Marcot (2012) and Uusitalo (2007). Chen and Pollino (2012) review several important aspects of environmental BN model applications including conceptual model development, discretization of continuous variables, parsimony in model construction, and model evaluation. For BNs to have a sound statistical basis for causal inference, the model should be considered unbiased (Weisberg, 2010); this means that the model structure should mimic theoretical understanding. Accordingly, an important element of our work is the exploration of alternative model structures that are equally plausible and provide similar predictive performance, thus representing possible alternative cause-effect mechanisms. The better the true mechanisms are understood, the more likely that effective management techniques may be identified and implemented to achieve water quality goals.

The objectives of this paper are to describe and implement a Bayesian Network (BN) modeling approach using field-collected data for evaluating regionally-relevant causal relationships between nutrients and stream macroinvertebrate communities in light of other water quality stressors. The approach is applied to an index of macroinvertebrate community health, the number of Ephemeroptera, Plecoptera, and Trichoptera taxa (EPT), measured in streams in an agricultural ecoregion, the Eastern Corn Belt Plains (ECBP), in Ohio, USA. Use of the Ohio data provides an opportunity to build knowledge regarding the ability of BN models constructed using field data from water quality sampling programs implemented by state environmental agencies to make useful predictions of the impacts of various water quality management scenarios.

2. Methods

2.1. Data set

Data on macroinvertebrate communities and stream water quality were provided by the Ohio Environmental Protection Agency (OEPA). Data collected from within the ECBP ecoregion, Level III Ecoregion 55, were selected for BN modeling after exploratory data analysis indicated the presence of several relatively strong relationships among environmental variables and EPT. This ecoregion covers a large portion of the Western half of Ohio, and is characterized by extensive corn, soybean, and livestock production. The data set consists of variables measured during 128 visits to 122 stations (some stations were visited more than once) between 2005 and 2013. Data collected prior to 2008 were from OEPA "developmental" studies used to capture a nutrient gradient in certain rivers and minimize the effect of confounding variables (R. Miltner, personal communication). These data comprised roughly 20% of the values used. The remaining values were "basin survey data" obtained during routine monitoring where confounding variables may have a greater effect on observed relationships (R. Miltner, personal communication). For the purposes of this project, we assumed both types of data could be combined without adding significant bias to the modeling results. Analysis of bivariate scatter plots of data records with no missing values supported this assumption although patterns were driven by the larger basin survey data set.

2.2. Treatment of missing data and outliers

The Ohio ECBP data set contained a number of missing observations for EPT and dissolved oxygen (DO). Data sets with missing data are not uncommon. The standard practice of deleting those cases (e.g., data rows) is the simplest strategy in preparing data for analysis, but this may result in a smaller than desirable data set, leading to weak statistical inferences. An alternative approach is data imputation, which was applied here using Amelia II (Honaker and King, 2010; Honaker et al., 2010) for multiple imputation to allow a comparison of BN models developed with and without records containing imputed data. It is reasonable to conceptualize the estimation of missing data as being based on a multiple regression of each variable (with missing data) on all other variables. It is recommended that additional correlative variables (beyond those to be used in the statistical analysis of interest) be included in the multiple imputation, even if they are not being used in the actual statistical analysis. Amelia II is most reliable for data sets that are multivariate normal, although the authors claim that the program is relatively robust to non-normality. For water quality data, this generally means that variables should be log-transformed before analysis, which we did. Two records were removed from the data set, one from 2010 in which the pheophyton concentration is greater than the chlorophyll a value) and one from 2013 which is a low statistical outlier with respect to chlorophyll concentration.

2.3. Conceptual model development

It is common modeling practice to develop a flow diagram consisting of boxes and arrows to describe relationships (arrows) among key variables (boxes) in an aquatic ecosystem. This graphical "conceptual model" can be used as a device to guide model development and explanation. In ecology, such models commonly display a conceptual understanding of the flow of materials or energy in an ecosystem. For a BN, however, the "flows" indicated by the arrows in a graphical model do not represent material flow or energy flow; rather, they represent conditional probability. Thus, while the BN modeling approach used here begins with a graphical model, in this case the presence (absence) of an arrow connecting two boxes specifies conditional dependence (independence). Bayesian networks can be as simple or as complex as scientific needs, knowledge, and data allow. The relationships may reflect direct causal dependencies based on process understanding or on a statistical, aggregate summary of more complex associations. In either case, the relationships are characterized by conditional probability distributions that reflect the aggregate response of each variable to changes in its up-arrow (or parent) predecessor, together with the uncertainty in that response. Conditional probability relationships may be based on observational/experimental data or on expert scientific judgment. Model structure and parameter (conditional probability) learning may be undertaken using machine learning algorithms (see Cowell et al., 1999; Jensen and Nielsen, 2007; Kjaerulff and Madsen, 2008; Korb and Nicholson, 2004; Pearl, 1988 for descriptions). Structure and parameter learning are particularly useful when the relationships among variables are uncertain.