Experimental and environmental factors affect spurious detection of ecological thresholds

JONATHAN P. DAILY,1 NATHANIEL P. HITT, DAVID R. SMITH, AND CRAIG D. SNYDER

U.S. Geological Survey, Leetown Science Center, Aquatic Ecology Branch, 11649 Leetown Road, Kearneysville, West Virginia USA 25430

Abstract. Threshold detection methods are increasingly popular for assessing nonlinear responses to environmental change, but their statistical performance remains poorly understood. We simulated linear change in stream benthic macroinvertebrate communities and evaluated the performance of commonly used threshold detection methods based on model fitting (piecewise quantile regression [PQR]), data partitioning (nonparametric change point analysis [NCPA]), and a hybrid approach (significant zero crossings [SiZer]). We demonstrated that false detection of ecological thresholds (type I errors) and inferences on threshold locations are influenced by sample size, rate of linear change, and frequency of observations across the environmental gradient (i.e., sample–environment distribution, SED). However, the relative importance of these factors varied among statistical methods and between inference types. False detection rates were influenced primarily by user-selected parameters for PQR ($s$) and SiZer (bandwidth) and secondarily by sample size (for PQR) and SED (for SiZer). In contrast, the location of reported thresholds was influenced primarily by SED. Bootstrapped confidence intervals for NCPA threshold locations revealed strong correspondence to SED. We conclude that the choice of statistical methods for threshold detection should be matched to experimental and environmental constraints to minimize false detection rates and avoid spurious inferences regarding threshold location.

Key words: ecological thresholds; nonparametric change point analysis; piecewise quantile regression; SiZer; stream benthic macroinvertebrates; type I error.

INTRODUCTION

Ecological communities commonly exhibit abrupt, nonlinear responses to environmental change due to rapid transitions among stable states (May 1977). Environmental conditions associated with rapid community change have been described as "ecological thresholds" (Groffman et al. 2006) and have been attributed to many environmental factors, including eutrophication in lakes and rivers (Scheffer 1990), climate change in arid and marine ecosystems (Hare and Mantua 2000), biotic interactions in rocky intertidal zones (Paine 1966), watershed urbanization in streams (Wang et al. 2007, Utz et al. 2009), and grazing and soil loss in grasslands (van de Koppel et al. 1997). Additionally, ecological thresholds may exist as lag times for community recovery from anthropogenic disturbances (Harding et al. 1998, Clements et al. 2010). Interest in ecological thresholds is growing rapidly, as reflected by the increasing number of research papers on this topic over the last 15 years (Dodds et al. 2010).

Threshold detection studies rely on various statistical approaches. Model-fitting approaches quantify thresholds as model parameters. Such methods include piecewise least-squares regression (Toms and Lesperance 2003), piecewise autoregressive models (Tong 1983), piecewise quantile regression (PQR; Chaudhuri and Marron 1999, Clements et al. 2010), and various polynomial spline models (Anderson 2008). In contrast, data-partitioning methods typically identify threshold locations by minimizing residual variance through recursive partitioning (Breiman et al. 1984). Nonparametric change point analysis (NCPA) is commonly used to identify potential thresholds through data partitioning (King and Richardson 2003, Qian et al. 2003). For example, Evans-White et al. (2009) applied NCPA for analysis of macroinvertebrate community-level responses to water-quality gradients. A hybrid approach incorporating model-fitting and data-partitioning was developed by Chaudhuri and Marron (1999) based on the second derivative of polynomial regressions (significant zero crossings, SiZer) and this method has been recently applied for ecological threshold detection in stream benthic macroinvertebrate communities (Sonderegger et al. 1999, Clements et al. 2010). Each of these methods estimate a change point that ecologists often interpret as an ecological threshold (Evans-White et al. 2009) and we use terms "change point" and "threshold" interchangeably throughout this paper.
Despite conceptual and mathematical differences in threshold detection methods, the relative performance of methods remains largely unknown (Andersen et al. 2009). Brenden et al. (2008) demonstrated significant differences in the sensitivity of model-fitting and data-partitioning methods to detect a modeled ecological threshold (i.e., type II errors). Similarly, Dodds et al. (2010) reported a three-fold difference in ecological threshold estimates among methods using a common data set. However, prior studies have assumed the existence of thresholds and therefore the problem of spurious detection has not been addressed. Furthermore, prior threshold simulation studies assumed that sample observations were uniformly distributed across the environmental gradient but did not evaluate the implications of this key assumption. False detection of ecological thresholds could lead to a misallocation of resources and may undermine the practical application of threshold detection tools in management and conservation. In this paper, we evaluate how experimental and environmental factors affect the rate and location of spurious ecological thresholds and how these effects vary among commonly used statistical methods.

**METHODS**

We used stochastic simulations to evaluate spurious detection of ecological thresholds in stream benthic macroinvertebrate responses. First, we generated species-abundance data based on a regional data set from the mid-Atlantic highlands, USA. Second, we modeled variation in sample size, the frequency of sample observations along an environmental gradient (i.e., sample-environment distribution, SED), and linear rates of change in taxonomic abundances. Third, we used bootstrapping methods to evaluate the relative importance of experimental and environmental factors for spurious threshold detection rates in PQR and SiZer and spurious change point locations in PQR, SiZer, and NCPA.

For model-fitting methods (PQR and SiZer), any significant threshold was interpreted as a spurious detection because the simulated communities exhibited linear change. In contrast, NCPA will identify change points based on deviance reduction in recursive partitioning (Qian et al. 2003) whether or not the underlying change is nonlinear. We therefore did not evaluate spurious detection rates for NCPA because we modeled abundance with changing mean and variance across the environmental quality gradient. However, NCPA threshold locations would be randomly distributed across a linear change gradient if the model provided unbiased threshold estimates. We therefore evaluated the confidence intervals of NCPA threshold locations based on bootstrapped 5th, 50th, and 95th percentiles (Baker and King 2010) as an indicator of spurious threshold detection.

We generated ranges for simulation model parameters from the USEPA’s Environmental Monitoring and Assessment Program (EMAP) stream benthic macroinvertebrate data set in the mid-Atlantic highlands, USA (Herlihy et al. 2000). This data set encompasses a large geographic extent (>200,000 km²) and has been used for ecological dose–response modeling (Yuan and Norton 2003). We assembled simulated taxonomic abundances with random draws from the negative binomial distribution (Hall et al. 1997, Kunin et al. 2000, Royle and Nichols 2003) and simulated change by parameterizing mean abundances as a linear function of a hypothetical environmental quality gradient. Responses of EMAP taxa to an empirical land-use intensity gradient ranged from extirpation to roughly doubling in abundance (C. D. Snyder, N. P. Hitt, D. S. Smith, and J. P. Daily, unpublished manuscript), so we modeled rates of linear change in the negative binomial mean as either 0, ±0.25, ±0.75, or +1.50. Extirpations were not permitted to avoid inducing nonlinearity. To reduce the number of zero-abundance counts associated with negative binomial draws at low mean abundances, we chose an initial mean abundance of 50 for all taxa.

We modeled six SEDs to represent the number of sample observations along gradients of decreasing environmental quality (range 0–100): uniform, exponentially decreasing, exponentially increasing, normal, wedge-shaped, and bimodal distributions. The uniform distribution provided a null model wherein sample observations were equally abundant across all levels of environmental quality. Exponential SEDs represented cases where the number of sample observations increased or decreased dramatically across an environmental gradient of interest. Such SEDs are commonplace in the ecological literature, as observed in studies of vegetative biomass (Austin 2007), stream benthic macroinvertebrates (Wang et al. 2007, Bryce et al. 2010), freshwater fishes (Freund and Petty 2007), forest birds (Melles et al. 2003), and mammals (Badgley and Fox 2000). A normal distribution provided the majority of sample observations at intermediate environmental quality levels, and has been observed in various floral and faunal communities (Gentry 1988, Anderson 2008, Evans-White et al. 2009). For comparison, the wedge-shaped distribution represented a distinct breakpoint in the number of sample sites across the environmental gradient and the bimodal distribution represented concentrations of sites at the extremes of the environmental gradient.

We evaluated sample sizes of 30, 45, 60, 75, 90, 120, 180, 240, 360, 480, and 600 sites to represent a reasonable range of sampling efforts. For instance, Bryce et al. (2010) used a 557-site database while King and Richardson (2003) used 36 samples for each date. We generated 1000 simulations for each combination of sample size, SED, and magnitude of linear change (Cao and Hawkins 2005). All analyses were conducted in R (version 2.12.0; R Development Core Team 2010) and code is presented in the Supplement.
We fit PQR models to the simulated linear changes in abundances and used a Bayesian information criterion (BIC) for model selection (Schwarz 1978). We assessed PQR threshold locations as the one-degree spline model with the lowest BIC scores for a given regression quantile (τ). We calculated thresholds for τ values of 0.5 and 0.9 and for an optimized τ value obtained by minimizing BIC for a linear quantile model. Quantile regressions were performed using the R package quantreg (Koenker 2010).

NCPA was implemented using a recursive partitioning algorithm from the R package rpart (Therneau and Atkinson 2010) and the significance test of Qian et al. (2003). We specified a maximum tree length of 1 and reported a change point only in cases where a split was reported with a $\chi^2 P < 0.05$ (Qian et al. 2003). We calculated 90% confidence intervals using bootstrapping (King and Richardson 2003, Evans-White et al. 2009). While Buhlmann and Yu (2002) found subsample bootstrapping to be more accurate in many cases, we used standard bootstrap aggregation to reflect the more common practice.

We implemented the SiZer method using the R package SiZer (Sonderegger 2010) and evaluated a broad range of smoothing bandwidths ($h = 1, 5, 50, 100, and 200$) over 100 environmental condition bins. The second derivative indicates changes in slope magnitude, so any significant crossing of zero was interpreted as a threshold (Chaudhuri and Marron 1999). If multiple thresholds were found for a single taxon, we included all thresholds when analyzing location.

We used random forests (Breiman 2001) to evaluate the relative importance of sample size, SED, and rate of linear change on spurious detection rates and threshold locations. To evaluate threshold detection rates, we used 1000 two-factor tree random forests with sample size, SED, and magnitude of change as predictors and the proportion of model runs indicating thresholds as the response variable (Breiman 2001). For PQR and SiZer models, we included τ and h as additional predictor variables, respectively. To evaluate spurious threshold locations, we repeated random forest analysis using threshold location as the response variable. Factors were ranked based on increased node purity, a measure of variable importance, and calculated in R package randomForest (Liaw and Wiener 2002).

**RESULTS**

Explanatory factors accounted for 97.7% and 83.5% of the total variance in threshold detection for PQR and SiZer, respectively (Fig. 1). Error rates were influenced primarily by user-selected parameters: relative node purity for τ was 3.5-times that of the second-most important variable in PQR ($n$) and 1.7-times that of the second-most important variable in SiZer (SED; Fig. 1). Explanatory factors for threshold locations accounted for 46.4% of the total variance in PQR, 7% in SiZer, and 54.8% in NCPA (Fig. 2B). In all models, SED was the most influential factor for threshold locations, with relative node purity values accounting for over 67% of the explained variance in SiZer and over 90% of the explained variance in PQR and NCPA (Fig. 1). Error rates for each scenario are presented in the Appendix.
User-selected parameters exhibited different asymptotic error rates along sample size gradients (Fig. 2). PQR models using median values ($\tau = 0.5$) showed error rates declining from 33% to 10% with increasing sample size. However, at $\tau = 0.9$, increasing sample sizes yielded decreases in PQR error rates from 69% to 38%, revealing an asymptotic error rate approximately 30% greater than regression on median values (Fig. 2A). At low sample sizes, linear optimization for $\tau$ yielded marginally higher error rates than regressions at $\tau = 0.5$; but these differences were largely within 1 standard deviation at sample sizes greater than 100 (Fig. 2A). In all PQR models, increasing sample size from 30 to 200 improved error rates more than increasing sample size from 200 to 600 (Fig. 2A).

SiZer error rates also decreased with sample size and revealed different asymptotic rates among user-selected parameters (Fig. 2B). Error rates were smallest when $h = 50$ and were highest when $h = 200$, averaging 9% and 21% across sample sizes, respectively (Fig. 2B). At high sample sizes, asymptotic error rates typically increased with bandwidth size from approximately 10% at $h = 50$ to 20% at $h = 200$ (Fig. 2B). Bandwidths of $h = 5$ exhibited high model-fitting failure rates (ranging from 10% to 60%) and failed to fit models most often at the smallest sample sizes (Fig. 2B). Results for $h = 1$ were discarded because model-fitting high failure rates exceeded 70%. Overall spurious threshold error rates were lower for SiZer than for other methods.

Predicted threshold locations exhibited substantial variation among statistical methods and SEDs (Fig. 3). At bandwidths >5, SiZer was more invariant to effects of SED than PQR or NCPA; predicted threshold locations for PQR and NCPA more closely tracked the SEDs for uniform, exponential, wedge-shaped, and binomial distributions than did SiZer (Fig. 3). All models except SiZer at $h = 5$ were largely concordant with the normal SED (Fig. 3). NCPA threshold location confidence intervals also closely tracked SED in all cases but the uniform SED (Fig. 3). The 5th and 95th percentiles captured lower and upper threshold locations for the uniform distribution, respectively (Fig. 3).

**DISCUSSION**

Several methods for ecological threshold detection have recently been developed (Dodds et al. 2010), and valid inference from these methods requires an understanding of their statistical properties and performance. The first inference from a threshold analysis involves the determination of whether or not a significant change point can be detected; the next step is to estimate the location of the change point. Our study provides one of the first analyses of sequential levels of inference for spurious thresholds (type I errors; Duggins 2010) and is the first to consider the effect of SEDs. We found that user-selected model parameters, sample size, and SED have important influences on both the rate of threshold detection and the location of reported thresholds. Some
of these effects may be minimized through experimental design considerations (i.e., sample size and model parameters), but other factors are intrinsic to study areas or population dynamics (i.e., SED and rate of change) and largely beyond investigators’ control.

Spurious detection rates were generally lower for SiZer than PQR but both approaches were influenced by user-selected parameters ($\tau$ and $h$). Cade and Noon (2003) recommended quantile regression in part for the ability to indicate different bivariate relationships in ecological data through the choice of different $\tau$ values. Based on our results, median or optimized $\tau$ values are expected to minimize spurious detection of ecological thresholds in quantile regression analysis. Our results also support Sonderegger et al.’s (2009) assertion that different applications will have different bandwidth optima (e.g., count data or pH levels). Bandwidths are expected to vary as ranges of the environmental gradient change, but our results indicate that an optimal relative bandwidth may exist for a given range of environmental data. Considering that there are no standards for either $\tau$ or $h$ selection, new research into parameter selection could greatly benefit threshold analysis.

Threshold locations were most sensitive to SED in the data-partitioning model (NCPA). The availability of sample observations across an environmental gradient should therefore be carefully considered in applications of NCPA. King and Richardson (2003) recognized that interpretation of NCPA requires an understanding of the uncertainty associated with change point locations and recommended quantifying this uncertainty using bootstrapping methods as described by Qian et al. (2003). However, Buhlmann and Yu (2002) asserted that bootstrap methods serve as smoothing operators on discrete thresholds that result from data partitioning. Our results support this assertion and further showed that NCPA bootstrapped confidence intervals were sensitive to SEDs, demonstrating the practical importance of SED for inferences on threshold location. We therefore suggest that caution is warranted when interpreting NCPA threshold locations from non-uniform SEDs.

We suggest that the choice of threshold detection method may be optimized based on experimental and environmental constraints in ecological studies. For instance, studies limited to small sample sizes (i.e., <50 observations) would minimize spurious threshold detection by using SiZer instead of PQR methods. However, we caution that bandwidth selection in SiZer should be carefully considered. Accounting for SED presents a challenge for ecological researchers because field-based studies often seek to characterize environmental conditions in the proportion to their availability (e.g., Herlihy et al. 2000). In contrast, paired laboratory-field analysis of thresholds (e.g., King and Richardson 2003) may be useful to account for SED by designing treatments to

![Figure 3. Threshold location density plots. Nonparametric change point analysis (NCPA) indicates the density of change points (CP) and bootstrapped 5th, 50th, and 95th percentiles that define the 90% confidence interval (CI) (dark gray, black, light gray, respectively). Significant zero crossing (SiZer) threshold locations indicate bandwidth size. Piecewise quantile regression (PQR) threshold locations indicate $\tau$ values (0.5 and 0.9) and optimized $\tau$ (OPT). Sample–environment distributions are represented with black lines.](image-url)
achieve equivalent representation of sample observations among dose levels.

Our results provide the first analysis of SED on spurious threshold detection and we demonstrated important effects of SED in statistical methods using conceptually distinct model-fitting and data-partitioning approaches. However, additional analysis of type II error rates and accuracy are necessary to understand the full trade-offs for statistical detection of ecological thresholds. We agree with Buhl-Mortensen (1996) that the consequences of type II errors may be more harmful than type I errors for conservation and management due to irreversible no-action decisions. Our analysis simply provides a cautionary tale to assist researchers in avoiding one type of statistical error and thus to optimize the detection of ecological thresholds.

Acknowledgments

Funding for this work was provided by USGS National Park Monitoring Project, Status and Trends of Biological Resources (S&T) Program. We thank P. Geissler for assistance with the Monitoring Project, Status and Trends of Biological Resources. We thank B. Cade, D. Sonderegger, and R. King for helpful comments on the manuscript. We thank P. Geissler for assistance with the Monitoring Project, Status and Trends of Biological Resources.

LITERATURE CITED


SUPPLEMENTAL MATERIAL

Appendix
Threshold fitting rates for each method by sample size, slope of change, and environmental distribution (Ecological Archives E093-003-A1).

Supplement
R scripting language code used for data generation in simulations (Ecological Archives E093-003-S1).