Ecological Indicators 29 (2013) 411–419

Contents lists available at SciVerse ScienceDirect

Ecological Indicators

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

A causal examination of the effects of confounding factors on multimetric indices

Donald R. Schoolmaster Jr.[a,*], James B. Grace[b], E. William Schweiger[c], Brian R. Mitchell[d], Glenn R. Guntenspergen[e]

[a] Five Rivers Services, LLC at U.S. Geological Survey, National Wetlands Research Center, United States
[b] U.S. Geological Survey, National Wetland Research Center 700 Cajundome Blvd., Lafayette, LA 70506, United States
[c] National Park Service, Rocky Mountain Network, 1201 Oakridge Drive, Fort Collins, CO 80525, United States
[d] National Park Service, Northeast Temperate Network 54 Elm Street, Woodstock, VT 05091, United States
[e] U.S. Geological Survey, Patuxent National Wildlife Research Center 12100 Beech Forest Road, Laurel, MD 20707, United States

ARTICLE INFO

Article history:
Received 17 November 2011
Received in revised form
11 December 2012
Accepted 18 January 2013

Keywords:
Multimetric index
Metric adjustment
Causal networks
Biological integrity
Bioassessment
Human disturbance
Environmental covariates

ABSTRACT

The development of multimetric indices (MMIs) as a means of providing integrative measures of ecosystem condition is becoming widespread. An increasingly recognized problem for the interpretability of MMIs is controlling for the potentially confounding influences of environmental covariates. Most common approaches to handling covariates are based on simple notions of statistical control, leaving the causal implications of covariates and their adjustment unstated. In this paper, we use graphical models to examine some of the potential impacts of environmental covariates on the observed signals between human disturbance and potential response metrics. Using simulations based on various causal networks, we show how environmental covariates can both obscure and exaggerate the effects of human disturbance on individual metrics. We then examine from a causal interpretation standpoint the common practice of adjusting ecological metrics for environmental influences using only the set of sites deemed to be in reference condition. We present and examine the performance of an alternative approach to metric adjustment that uses the whole set of sites and models both environmental and human disturbance effects simultaneously. The findings from our analyses indicate that failing to model and adjust metrics can result in a systematic bias towards those metrics in which environmental covariates function to artificially strengthen the metric–disturbance relationship resulting in MMIs that do not accurately measure impacts of human disturbance. We also find that a "whole-set modeling approach" requires fewer assumptions and is more efficient with the given information than the more commonly applied "reference-set" approach.

1. Introduction

The enterprise of bioassessment (evaluation of the condition of an ecosystem using biological surveys – Barbour et al., 1999) has had a long history and is increasingly relied upon to guide the management of natural resources. One tool of bioassessment that is increasingly being used is the multimetric index (MMI, Hering et al., 2006). MMIs use biological or ecological measurements, often compiled into metrics, to quantify and serve as a surrogate for the degree to which human disturbance has influence on biological communities. While originally applied to streams under the name Indices of Biological Integrity (IBI) (Karr, 1981, 1991), MMIs have now been developed for a number of different systems, including wetland plants (Mack, 2001; Rocchio, 2006), terrestrial invertebrates (Kimberling et al., 2001) and lakes (O'Connor et al., 2000) and have been applied at a range of spatial scales from local (Wallace et al., 1996) to continental (Pont et al., 2006). Indeed the concept represented by MMIs has been suggested to represent an important integrative concept in ecology (Ford and Ishii, 2000). To be useful to resource managers, an index must meet at least three criteria; (1) it must be sensitive to human disturbance (2) it should measure variation in metrics and disturbance at a scale that is useful for management and (3) it should include interpretable metrics. Individual metrics are typically combined into a "multimetric" index (MMI), which provides an overall score of integrity for a system (see Kurtz et al., 2001 and Andreasen et al., 2001 for discussion of criteria for MMIs)

The development of a MMI requires a number of decisions. These decisions relate to, for example, the criteria for selecting metrics (Karr and Chu, 1997; Barbour et al., 1999; Stoddard et al., 2008) and the scaling of metrics (Blocksom, 2003). One particular aspect of MMI development that has been receiving increasing

* Corresponding author. Tel.: +1 337 266 8653.
E-mail address: schoolmasterd@usgs.gov (D.R. Schoolmaster Jr.).

1470-160X/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.ecolind.2013.01.015

attention is the potential impacts of environmental covariates on the interpretability of individual biological or ecological metrics (Wiley et al., 2003; Baker et al., 2005; Cao et al., 2007; Whittier et al., 2007; Stoddard et al., 2008; Hawkins et al., 2010). In this context, environmental covariates refer to natural gradients such as soil texture, elevation, aspect, etc., that may affect the degree of human disturbance at a site and/or aspects of the biological community.

Environmental covariates can interfere with effective metric selection in two ways: (1) Certain patterns of causal connections between environmental covariates and ecological metrics can obscure the true effect of disturbance, resulting in the non-selection of ecosystem components that are in fact strongly affected by human activities, (2) some patterns of causal connections among covariates and metrics can exaggerate the true effects of human disturbance and result in selection of metrics that are not informative measures of system response. Explaining how each of these may occur is a goal of this work and is described in later sections. But, we should expect that in most environments, complex environmental influence on bioassessment measures will exist and both obscuring and exaggerating influences may be occurring simultaneously, with unknown net effects.

Currently, there are at least three strategies employed to mitigate the effects of environmental covariates on MMIs. The apparently oldest and most frequently used approach is to attempt to avoid problems by developing separate indices for (presumably) different homogeneous sub-regions of the study area (Barbour et al., 1999). We do not address this approach here, except to note that it may be impractical at smaller spatial scales where sample sizes are not large, it cannot address interactions among environmental covariates and it results in multiple MMIs where one, more general MMI would be preferable. Another prominent approach is to use a "minimally impacted" reference set of sites to develop models of the effects of environmental covariates on metrics and then use these models to remove such effects in the entire (reference and non-reference) set of sites (Cao et al., 2007; Whittier et al., 2007; Stoddard et al., 2008; Hawkins et al., 2010). A third, less-commonly applied approach uses the set of all sites to simultaneously model the effects of disturbance and environmental covariates on metrics and then adjusts metrics based on model parameters (Wiley et al., 2003; Baker et al., 2005).

Controlling for confounding effects so that focal relationships of interest (e.g., the effects of human disturbance on biotic conditions) can be interpreted causally is a long-standing dilemma in statistics (e.g., Pedhazur, 1997; Cohen et al., 2003; Pearl, 2009). Increasingly, it is recognized that a structural equation approach is needed (Grace, 2006) and that such an approach should be informed by a graphical modeling perspective that compensates for the absence of a causal language in probability theory (Pearl, 2010). Absent a graphical specification of causal assumptions, statistical adjustment oversimplifies the interpretive implications and fails to guide the scientist as to the various options available for modeling their data. These ideas are expanded upon and an example of the use of causal networks for metric adjustment is described in Appendix A.

Our goals in this paper are (1) to demonstrate the negative impact that the effects of environmental covariates can have on producing effective and interpretable MMIs and (2) to evaluate the efficacy of different methods of metric adjustment. We believe it is important for the conceptual development of this large and complex topic that underlying assumptions be conveyed as clearly as possible and we use graphical models of causal networks (Grace, 2006; Pearl, 2009) to describe various scenarios so that we can better interpret the effects of various relationships among covariates, human disturbance and biological metrics (Fig. 1). We end our treatment with a suggested set of steps for modeling relationships and adjusting metrics to aid in their selection for MMIs.

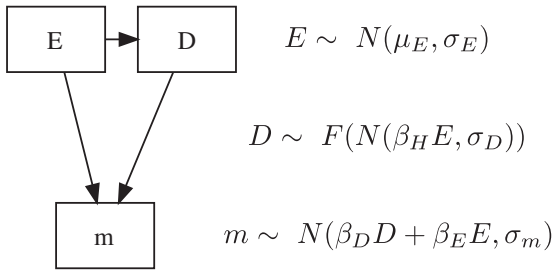

Fig. 1. Causal network showing situation where an environmental covariate (E) influences both patterns of human disturbance (D) and metric expressions (m). The network on the right side of the figure implies the equations given for the statistical relationships on the left. See Appendix A for more details.

2. Effects of environmental covariates

2.1. Scenario methods

In order to examine the potential effects of environmental covariates on the MMI construction process and to test methods of metric adjustment, it is necessary to construct scenarios where the effects of all factors are known. For this reason, we have simulated a series of data sets based on a variety of causal situations commonly encountered in real data. Each simulated data set includes a measure of human disturbance (D) and one or more biological metrics (m). For each scenario, we embed the essential relationships in a causal network that includes one to many environmental covariates (E) that may be associated with the metric and/or the disturbance measure as well as with one another. For the purpose of this paper, we consider environmental covariates to be factors that are exogenous with respect to D and m (i.e., the covariates have no arrows pointing to them from other variables in the model). In other words, we will consider systems described by graphical models in which E may influence D and m, but not vice versa (Note that in cases where D influences E, adjusting for E will remove part of the effect of D; thus, this situation is one where adjustment is not appropriate.). This is the typical assumption for metric adjustment (e.g. Stoddard et al., 2008). We examine (1) linear networks of increasing complexity, (2) networks that include multiplicative effects and (3) networks that include non-linear relationships.

For simulations, exogenous variables in networks were instantiated by drawing normally distributed pseudo-random numbers, using the "Mersenne Twister" algorithm of Matsumoto and Nishimura (1998). Values for endogenous variables (i.e., those affected by other variables in the model and thus having arrows pointing to them in the graphical models) were calculated by applying the network-implied equations, plus normally distributed random error (Fig. 1). Unless otherwise stated, an arrow in a network diagram indicates a linear effect of the variable at the tail on the variable at the head. Specific equations used for simulation are given in Appendix B.

The disturbance variable (D) was transformed to have a uniform distribution of values between 0 and 9. The transformation function used was $F(D) = \min(I > \mathrm{rank}(D_i) \times 10/n) - 1$, where I is the set of integers and n is the number of simulated sample sites. The function $F(x) = \min(I > x)$ is implemented on many programming platforms as the "ceiling" function. We chose this form of D to provide even coverage across potential values.

In this section, we demonstrate the variety of effects that different scenarios can have on the observed correlation between disturbance and the metric. We chose correlations as the parameter of interest because it is the summary statistic most often used in the MMI construction process to detect the association between a metric and disturbance. We quantify the effects of environmental covariates as the difference in strength of correlation between D

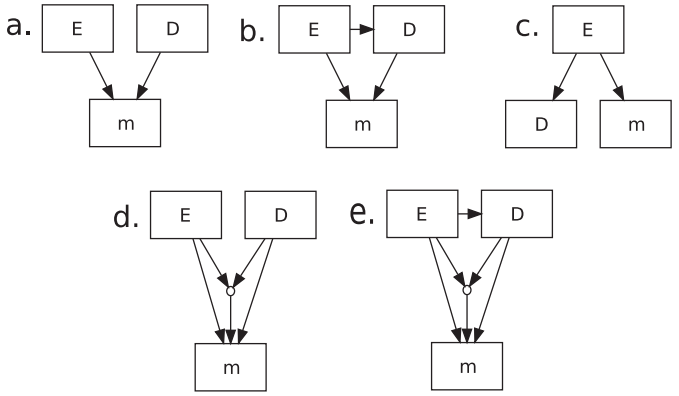

Fig. 2. Causal networks representing different scenarios for environmental covariates. Network (a) represents Scenarios I and III. Network (b) and (c) are different cases of Scenario II. Networks (d) and (e) have both direct individual effects and an interactive effect and are discussed in Scenario IV.

and m with and without the effects of the network of covariates E. That difference is expressed as $\Delta\text{cor} = \text{cor}(m_{E=0},D) - \text{cor}(m,D)$, where $m_{E=0}$ refers to a metric unaffected by covariates. For this analysis, we assume that biological metrics are adversely affected by disturbance; thus, we expect $\text{cor}(m_{E=0},D) < 0$. For real data, raw metrics may be positively or negatively related to D and in such cases, those that are positively related to D are often "reflected" (reversed) before being included in a MMI. Values of $\Delta\text{cor} < 0$ indicate that the network of environmental covariates obscures the true strength of effect of D on m.

For the simulation method described above, the possible outcomes for measured association between D and m can be described analytically by recognizing that a correlation can be understood as the magnitude of the bivariate relationship standardized by the total variation in the system. We use this insight to develop analytical models of the potential effects of environmental covariates on the measured association between D and m. Those analyses, which follow the scenarios described below, can be found in Appendix B. Since there are many ways that environmental covariates may affect the relationship between human disturbance and biological metrics, we proceed in this paper by presenting a number of instructive hypothetical scenarios to examine the types of effects environmental variables can have on the observed correlation between D and m. In these scenarios, we assume only a single covariate is involved for simplicity.

Scenario I: Only the metric is affected by environmental covariate (Fig. 2a). For example, small scale variation in soil properties that affect plant growth, but do not affect the probability or degree of human disturbance.

Scenario II: Metric and disturbance are both influenced by environmental covariate (Fig. 2b). For example, elevation can affect both accessibility to humans and be strongly correlated with factors that affect plant and animal communities.

Scenario IIa: A special case of Scenario II is spurious correlation (Fig. 2c), in which the covariate affects both the metric score and Human disturbance, but human disturbance does not directly affect the metric. Spurious correlation can result in selecting metrics for a MMI that are not responsive to disturbance.

Scenario III: Non-linear environmental covariates. Many metrics, especially community level metrics are known to vary non-linearly along environmental gradients. For example, plant species richness is often found to be a unimodal function of productivity (Grace, 1999; Gough et al., 2000; Mittelbach et al., 2001).

Scenario IV: Interactive networks. It is possible for environmental covariates and disturbance to interact in such a fashion as to have a multiplicative effect on a metric (Fig. 2d,e). This can happen if the effect of an environmental covariate on a metric is a function of the degree of disturbance. For example, plant productivity can be affected by water availability and also by disturbance from cattle grazing. For physiological reasons, the efficiency with which plants produce biomass at a given water level depends on the level of grazing damage sustained by the plants. In symbols,

$$m = \beta_E(D)E + \beta_D D \text{ and } \beta_E(D) = \beta_{E_0} + \beta_{E\times D}D,$$
$$\therefore m = \beta_{E_0}E + \beta_{E\times D}DE + \beta_D D$$

where $\beta_D < 0$, $\beta_{E\times D} < 0$ and $\beta_{E_0} > 0$.

2.2. Scenario simulations

2.2.1. Scenario I

The observed relationship in simulations where there were no environmental covariates affecting either metrics or disturbance exhibited a correlation of ~−0.65 (based on a sample size of 200). In subsequent comparisons, we will refer to the standardized unobscured effect of D on m as the "true" effect for simplicity, where the true effect is measured as the partial correlation of m and D. For Scenario I (Fig. 2a), simulation results (Fig. 3) confirm the analytical expectation (Appendix B) that the influence of an independent environmental covariate on a metric will be to decrease the strength of the observed correlation between disturbance and metric (holding constant the true effect of D on m, which is the case for all of the simulations). Fig. 3a shows that the correlation between D and m in the absence of the environmental covariate is stronger than when the effect of the environmental covariate included on the metric (Fig. 3b). This result will be quite general because any additional cause of variation in m independent of the effect of disturbance (such as an E uncorrelated with D) will elevate the unexplained error variance for m, decreasing the observed strength of association.

2.2.2. Scenarios II and III

For the case where there is an environmental covariate (perhaps a topographic gradient) that influences both patterns of human development and native ecosystem characteristics (a situation represented in Fig. 2b), more complex influences on observed D–m relationships are possible. Simulation results confirm the analytical expectations described in Appendix B. If the indirect effect of E on m via D is of the same sign as the direct effect of E on m, we can expect a correlation between D and m that is stronger than the true effect (compare Fig. 4b to Fig. 4a, for an example with real data see Appendix C: Example 2). In the case where the indirect effect of E on m via D is of opposing sign to the direct effect of E on m, the observed correlation between D and m can be substantially weaker than the true effect (compare Fig. 4c to Fig. 4a) or even of opposite sign (compare Fig. 4d to Fig. 4a, for an example with real data see Appendix C: Example 1). Because indirect effects of D on m can be quite strong, it is possible to observe significant correlations between D and m even when the true effect is zero. Such relationships are often referred to as spurious (see Fig. 5 for an example). Such results can be generated regardless of whether effects in the model are linear or non-linear.

2.2.3. Scenario IV

As in Scenarios I and II, for the interactive case (Scenario IV), environmental covariates can have a wide range of influences on the bivariate relationship between D and m (Fig. 6). In general, the expectations are the same as in Scenarios I and II; however, the results can include curvilinearities in responses that bring an additional complexity to the metric adjustment enterprise.

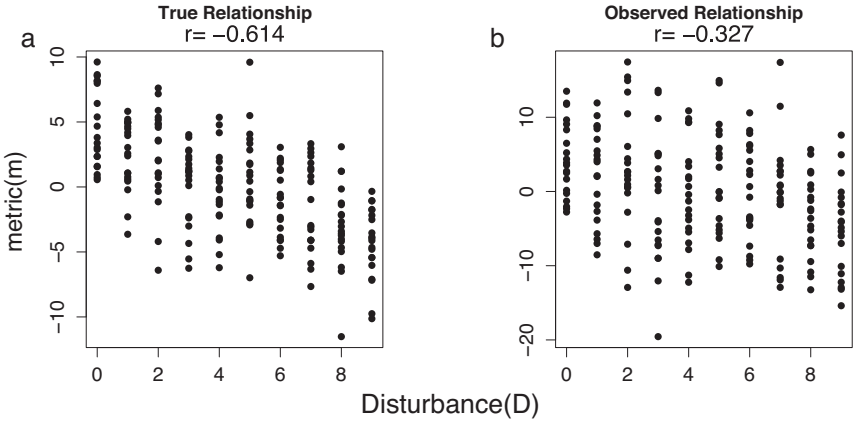

Fig. 3. Simulation of Scenario I (based on 200 simulations). (a) The true relationship between D and m. (b) The observed correlation between D and m is weakened by the environmental effect on the metric (β_E).

3. Quantitative assessment of metric adjustment procedures

3.1. Quantitative assessment methods

As stated previously a MMI should satisfy a number of criteria; two of which being, that it is sensitive to the effects of human disturbance and that it be comprised of interpretable metrics. These require that the confounding influences of environmental covariates are removed from the metrics. Recently, two methods have been used to quantitatively determine and remove the effect of environmental covariates from candidate metrics; we refer to them as "Reference-set residualization" (RSR) (Whittier et al., 2007; Stoddard et al., 2008), and "Whole-set residualization" (WSR) (called "regional normalization" by Wiley et al., 2003).

Both methods are similar in that they consist of two steps, estimation and residualization. Each models the metric as a function of environmental covariates and then uses the model to adjust the observed metric values. However, they differ in how they deal with human disturbance. The RSR methodology models only the subset of metric values that come from "reference" sites, which are presumably free from the effects of human disturbance. This

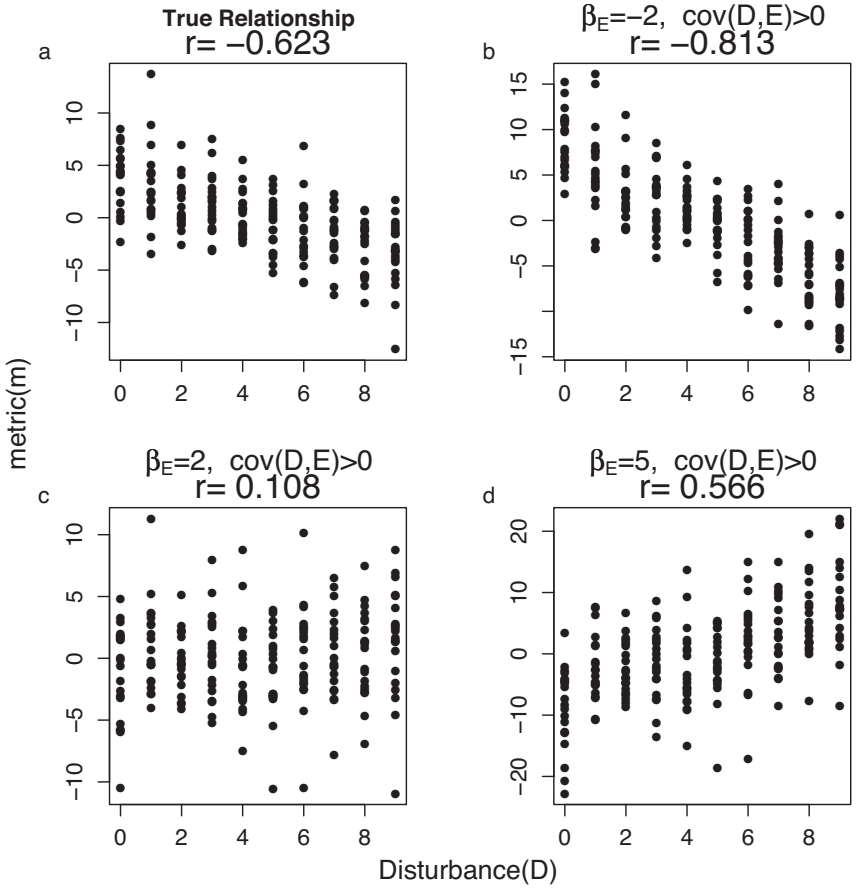

Fig. 4. Simulation of Scenario II (based on 200 simulations). a) shows the true relationship between D and m. (b) The observed correlation is inflated if the sign of β_E $cov(D, E)$ is the same the sign of β_D. Otherwise, the observed correlation can be weakened (c) or even appear to be strong in the opposite direction (d).

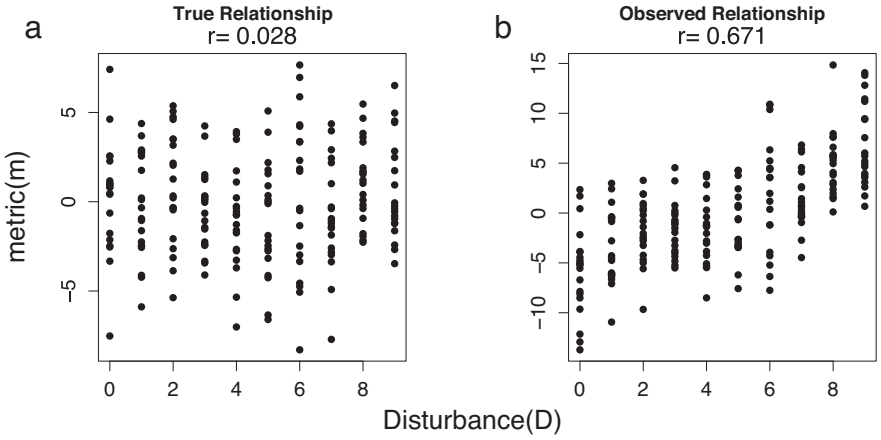

Fig. 5. (a) True and (b) observed spurious relationship between disturbance and a metric when they are only causally related through an environmental covariate.

model is then used to adjust metrics from all sites. The rationale behind this method is that using data only from the reference sites allows disturbance-free estimates of the effect of the environmental covariates. The WSR methodology models the metrics from the whole set of sites and includes measures of disturbance in the model. Metrics are adjusted as the predicted metric values in the absence of environmental covariates ($E = 0$). Thus, the new metric values contain only the portion of the variance attributable to variation in disturbance plus the variance unexplained by the model.

In this section, we examine the ability of both RSR and WSR to recover the actual relationship between the metric and disturbance in each of the scenarios above. To do this, we create a data set corresponding to a causal network, apply the RSR or WSR method and measure the percent error as the actual correlation versus that observed after the metric adjustment procedure, $[\mathrm{cov}(m_{E=0},D) - \mathrm{cov}(m_{\mathrm{adj}},D)]/\mathrm{cov}(m_{E=0},D)\times 100$. Positive values of this metric indicate that the adjustment method has resulted in an artificially low correlation; negative values indicate that metric adjustment resulted in an artificially high correlation.

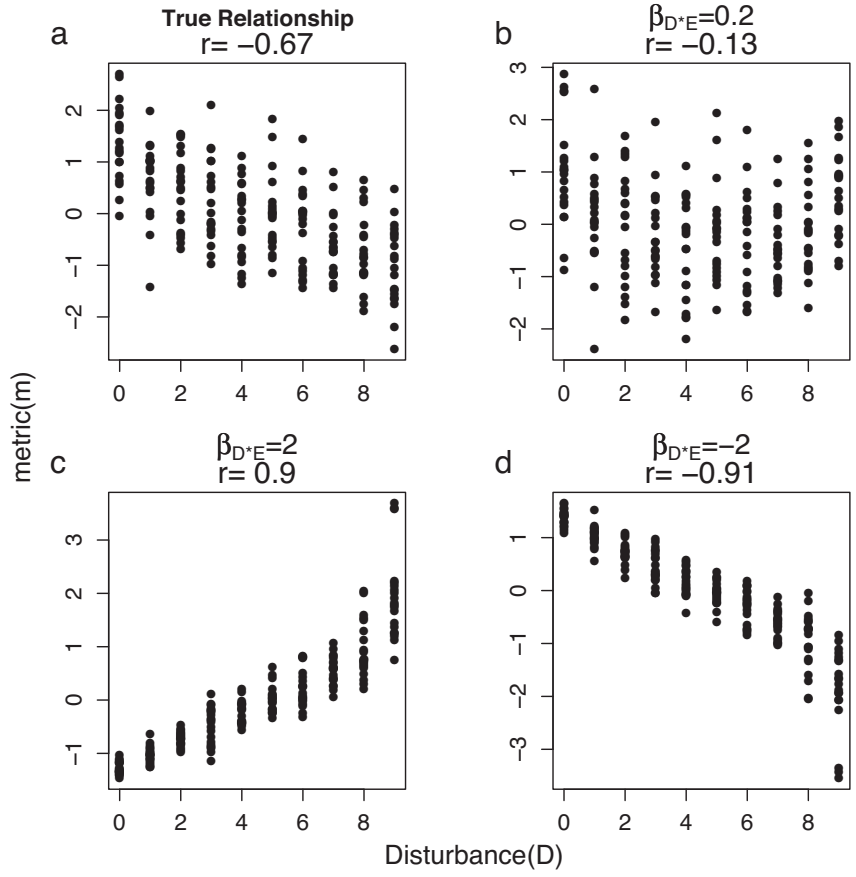

Fig. 6. True relationship (a) and the variety of outcomes (b)–(d) that may result for different values of the interactive effect of an environmental covariate and disturbance on the metric.

For each causal network, we repeated the process 1000 times at sample sizes ranging from 50 to 3200 to examine accuracy (percent error), precision (variation in percent error), and efficiency (error range/sample size) of each method. Because our goal is to examine the effectiveness of metric adjustment methods, we base the estimated coefficients used for adjusting on the true model. We did not add a model selection step in this case, which would be necessary if the true model were unknown. For the simulations, sites with a value of $D < 3$ (D ranged from 0 to 9) were defined as members of the reference set.

3.2. Scenario I

For situations structured as in scenario I (Fig. 2a), in which the environmental covariate affects only the metric, both methods of metric adjustment result in average errors of less than 1%. However, at all sample sizes, the accuracy of metrics adjusted by Whole-set residualization is about twice that made by Reference-set residualization. In addition, the precision of WSR adjusted metrics is much greater than that of RSR, especially at small sample sizes. The larger variation of the RSR method results directly from the smaller sample size used in the models to estimate the effect of the covariate. For example, for the results shown in Fig. 7a, the standard deviation of the estimated coefficient (β_E) was over twice as large as that of the WSR method.

3.3. Scenario II

When the environmental covariate affects both the metric and the disturbance measure (Fig. 2b), RSR results in some bias, even at very larger sample sizes (Fig. 7b). The average percent error is positive, indicating that on average the RSR tends to over-adjust, thus discarding part of the disturbance signal in the metric. As a function of sample size, the average percent error fell from near 19% at the smallest sample size to 6.6% at the largest (Fig. 7b). In fact, at the largest two sample sizes, the 95% confidence intervals of the simulations do not include zero, suggesting that RSR is asymptotically biased (i.e., it will never converge to the correct answer even with infinite sample size). The RSR method also tends to make large errors, especially at small sample size. At the sample size of 50, RSR errors of over 100% fall within the 95% confidence interval indicating that this method may result in adjusted metrics whose sign of correlation with D is opposite of the true relationship.

WRS resulted in average percent error of just over 1% at the smallest sample size, but well under 1% for all other sample sizes. As with RSR, the size of the errors that WRS tended to make was larger for networks of this structure than Scenario I, although these tend to decrease quickly as sample size is increased. The increased variation in the measurement is caused by increased variation in the metric m which is caused by the covariance between D and E.

3.4. Scenario III

The relative abilities of RSR and WSR to adjust metrics in the case of curvilinear environmental covariates and $\mathrm{cov}(D,E) = 0$ is similar to the linear case; RSR making larger errors on average and generating greater variation in the distribution of errors (Fig. 7c). Again, the larger errors come from the smaller sample size used by the RSR estimating models.

Both the accuracy and precision of the RSR method are greatly reduced if $\mathrm{cov}(D,E) \neq 0$ often resulting in metrics with error over 100% for samples sizes under 400 (Fig. 8). This happens because the covariance between D and E causes reference sites to only sample a portion of the environmental covariate, thus making it difficult for the model to make accurate estimates of the non-linearity (Fig. 8a). Thus, RSR uses a smaller sample size and a biased sample to estimate the non-linear effect of the environmental gradient. WSR performs as well in this case as it did in the simpler linear case (Scenario II).

3.5. Scenario IV

In the case where D and E interact to determine m, RSR fails to make accurate adjustment regardless of sample size (Fig. 7d). As in the non-linear case, this happens because the relationship between E and m in the reference set is not representative of the relationship in the whole set (Fig. 8b). If the interaction term ($\beta_{D\times E}$) is the opposite sign of the main-effect (β_E), this effect is even worse, producing average percent errors well over 100 at all sample sizes.

The WRS produces accurate adjustments to metrics with interactive effects, although the WRS adjusted metrics are less precise than comparable non-interactive networks. This reflects the increased variation in the metric relative to non-interactive scenarios and the increased difficulty in obtaining accurate estimates from interactive models.

4. Effect of environmental covariates on MMI sensitivity to human disturbance

One reason for combining metrics into MMIs is to gain a more robust characterization of ecosystem responses than could be achieved by any of the individual metrics alone. Although one hopes to assemble metrics that are both sensitive to disturbance and interpretable, these goals are not the same. We have shown how environmental covariates can interfere with both the interpretability and sensitivity of individual metrics. In this section, we use simulations to examine whether sensitivity to disturbance can be recovered through the process of creating an index from multiple unadjusted metrics.

We generated environmental and disturbance variables as described above. In addition, for each of the five metrics, 15 candidate metrics were generated as $m = \beta_D D + \beta_E E + \varepsilon$ where $\varepsilon \sim N(0, 4)$. We generated five sets of metrics corresponding to the different scenarios described above. The first metric type, m_1, which represents the true relationship, was generated from metric scores that range from 0 to $-\infty$ and are influenced only by human disturbance; thus, with $-\beta_D \sim \Gamma(1,2)$, $\beta_E = 0$, where $x \sim \Gamma(a,b)$ indicates that x is gamma-distributed with shape parameter a and scale parameter $1/b$. The next, m_2 corresponds to Scenario I, where metric scores are influenced by both disturbance and an independent environmental covariate; thus, $-\beta_D \sim \Gamma(1,2)$, $\beta_E = \Gamma(1,2)$, $\mathrm{cov}(D,E) = 0$. We also generated three metrics related to possibilities of Scenario II where disturbance and the environmental covariate are correlated either negatively or positively: $m_3(-\beta_D \sim \Gamma(1,2),\ \beta_E = \Gamma(1,2),\ \mathrm{cov}(D,E) < 0)$, $m_4(-\beta_D \sim \Gamma(1,2),\ \beta_E = \Gamma(1,2),\ \mathrm{cov}(D,E) > 0)$ and $m_5(\beta_D = 0,\ \beta_E = \Gamma(1,2),\ \mathrm{cov}(D,E) < 0)$.

Environmental covariates generally strengthen the observed correlation between D and m in metrics of the m_3 type because of the positive correlation between disturbance and covariate. For metrics of the m_4 type, where disturbance and covariate are negatively correlated, covariates generally weaken the observed correlation between D and m. Metrics of type m_5 represent a spurious relationship between D and m due to mutual dependence on the environmental covariate.

The candidate metrics were scaled to unitless quantities using the Blocksom CAUL method (Blocksom, 2003), which scales metrics to values between 0 and 10. Metrics were then screened for sensitivity to disturbance. Scaled metrics exhibiting significant correlation with D at $\alpha = 0.05$ were accepted for inclusion in the index. Selected metrics exhibiting a positive relationship with D were reflected as $m' = 10 - m$ to ensure an index with a negative

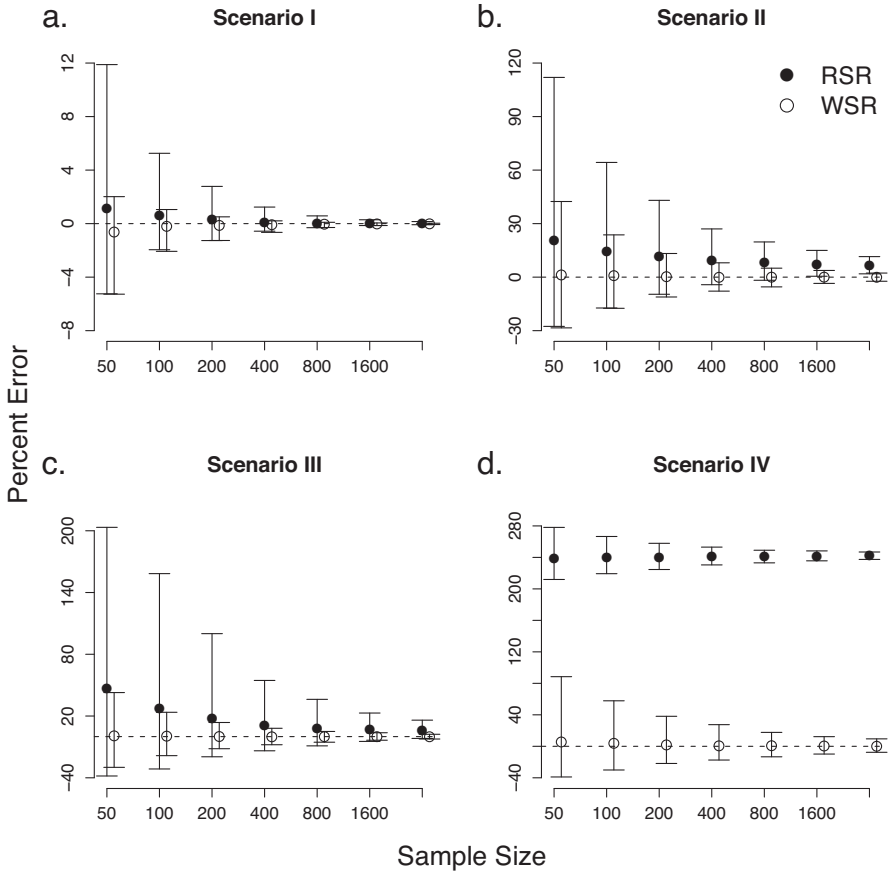

Fig. 7. Average, 5th and 95th percentiles of percent error resulting from adjusting metrics using Reference-site residualization methodology (filled circles) and Whole-set residualization (open circles).

relationship with disturbance. The MMI was created by calculating the mean of the (up to) 10 metrics with the highest observed strength of correlation metrics. The sensitivity of the index was measured as its correlation with D. This process was repeated 1000 times to allow us to estimate the variability of the result.

The average sensitivity of the simulated indexes varied in ways that could be predicted from the ways that environmental covariates affected the component metrics. Fig. 9 shows the mean, 5th and 95th percentiles of the sensitivity of each metric type. These MMIs were composed of metrics generated by Scenarios I and II. MMI m_1 consists of metrics in which there were no environmental covariates. MMI m_2 consists of metrics simulated by Scenario I. MMIs m_3–m_5 were generated from Scenario II in cases where environmental covariates strengthen observed correlation between D and m (m_3), weakened it (m_4) or resulted in purely spurious relationship (m_5). Numbers at top show values of correlation of individual metrics and D averaged over all simulations and suggest that MMIs with relatively high sensitivities are possible with unadjusted

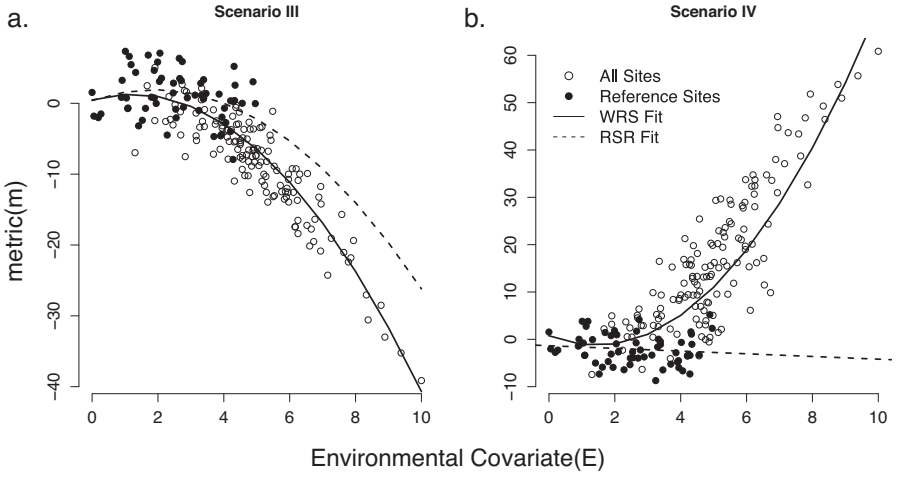

Fig. 8. Fits of models used by RSR (dashed) and WRS (models) to adjust metrics in non-linear (a) and interactive (b) scenarios.

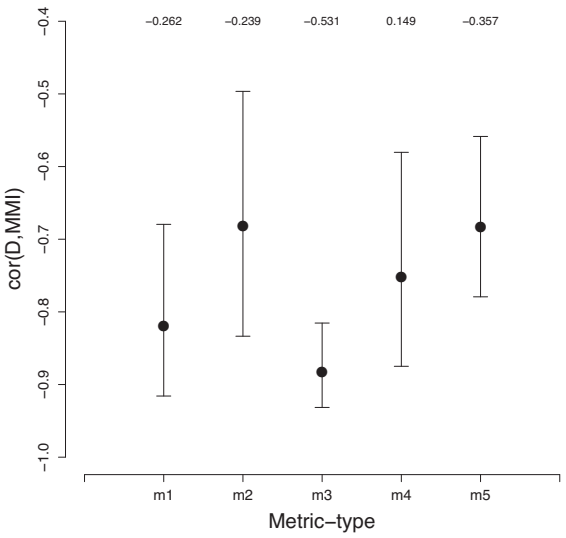

Fig. 9. Average, 5th and 95th percentile correlations between MMI scores simulated under Scenarios I and II and disturbance. Metric descriptions: m_1 = no environmental covariates; m_2 = Scenario I; m_3 = Scenario II, covariate inflates correlation; m_4 = Scenario II, covariate weakens correlation; m_5 = Scenario II, purely spurious relationship. Numbers at top show values of correlation of individual metrics with D averaged over all simulations.

metrics even if the average correlation between candidate metrics and D is low.

The difference in sensitivity between m_1 (not obscured by covariates) and the others at each realization of the simulation indicates the effect not adjusting metrics would have on the sensitivity of the MMI. Failing to adjust metrics of the m_2-type (Scenario I) for environmental covariates led to a MMI, on average, 13% less sensitive than an index constructed with correctly adjusted metrics and resulted in an index that was just as or more sensitive than the unobscured index 0.6% of the time. Simulated metrics of the m_3-type, in which covariates inflate the correlation between D and m, also increased average sensitivity of the MMI (by 7.1%) to lead to a more sensitive MMI than those constructed from correctly adjusted metrics 99.3% of the time. Metrics of type m_4, which are similar to m_3 but with a change in sign of one path, result in MMIs that are 8.1% less sensitive on average, but would result in an index as strong or stronger than one made with correctly adjusted metric 7.8% of the time. Finally, indexes made with metrics that had only spurious correlation with D, were 17.5% weaker on average and could be expected to produce a MMI as sensitive as one constructed with adjusted metrics 4.5% of the time.

5. Discussion

Our analyses show that the network of environmental covariates can affect the observed relationship between a biological or ecological metric and human disturbance. Environmental covariates may strengthen or weaken observed relationships depending on the structure of the network and the functional form of the relationships. This poses serious problems for effective metric selection and interpretability of MMIs. Because most MMI construction procedures (Karr and Chu, 1997; Barbour et al., 1999; Stoddard et al., 2008) involve selection of metrics that show the strongest relationship with the measure of human disturbance, failing to model and adjust candidate metrics will result in a systematic bias towards those metrics that are products of causal networks that artificially strengthen the metric–disturbance relationship and are most strongly affected by environmental covariates.

As evidence of this potential problem, we were able to simulate a MMI with what would be regarded as satisfactorily correlated with human disturbance, from metrics that have only purely spurious relationships to disturbance (m_5-type MMI). While the metrics in those simulated MMIs each had some ability to predict local disturbance, there was no direct causative relationship between disturbance and the metrics (these metrics fail to satisfy Pearl's back-door criterion for causal relations (Pearl, 2009)). As a result, the MMI (and its component metrics) would be insensitive to any management action taken to reduce human disturbance. As levels of human disturbance were reduced, one would not find that metric scores improved, but that they no longer were predictive of human disturbance.

Others have recognized that environmental covariates could interfere with MMI performance and have suggested methods for adjusting the metrics (e.g., Stoddard et al., 2008). Of the methods we tested, we find that metrics can be adjusted for known covariates most effectively with a "Whole-set" adjustment method that uses all available data to model metrics as a function of the known gradients and disturbance. This method not only produced more accurate, precise and efficient adjustments, but it also eliminates the need for classification of the disturbance state of sites into "reference" and "impacted" sites (another source of potential error). Such an approach does, however, require the ability to estimate human disturbance scores for individual sites (which may not be compatible with certain large-scale surveys). These models of metrics can be used to make predictions of the disturbance-free range of variation of the metric (Dodd and Oakes, 2004; Kilgour and Stanfield, 2006). While this approach has been criticized for extrapolating beyond the data, it makes predictions based on the largest set of data available and allows the assumptions going into the designation of "reference sites" to be identified. In fact, the WSR methodology could be characterized as using the set of all available data to extrapolate one point on the disturbance gradient. Considering it as such is helpful for understanding why it is more effective than "Reference-set" adjustment methods.

Where "Reference set" residualization fails, one reason it does so is because it takes the opposite approach of WSR to extrapolation; it uses a subset of data from one point on a disturbance gradient to extrapolate to the rest of the gradient. This leads to two kinds of errors, those associated with producing accurate model estimates from the reduced sample size, and those made because reference samples systematically fail to sample the variation in the environmental covariates. An example of this latter effect comes from the multiplicative example described in Scenario IV. Where disturbance and the environmental covariate interact to determine the metric, the relationship between the metric and the covariate in the reference set will not accurately represent the relationship elsewhere along the disturbance gradient (Fig. 8), resulting in highly biased adjustments. This does not only happen in interactive cases, but wherever there is covariance between environmental covariate and disturbance. The negative effects of this phenomenon tend to be small when all relationships are linear, but can be very large if any are non-linear.

6. Suggestions for constructing MMIs

Our analyses and simulations suggest that in order to produce an interpretable MMI, one should model the metrics for known environmental covariates. This approach requires that before data are collected, one considers what the major environmental covariates are and how to collect data on them. It is also helpful for one to consider potential causal networks for the system. Using graphical models, such as Fig. 2 (and Appendix A), to represent the hypothetical causal structure of the system will help determine how data

should be collected and which factors may be usefully modeled as environmental covariates.

Having a causal network hypothesis is important to determine which variables may usefully be included in the analysis as environmental covariates. As stated earlier, we consider only exogenous variables that may affect either disturbance or metrics, but not the other way around. An environmental measure that is affected by disturbance and in turn affects the metric does not function merely as a covariate, but as a causal mediator (Judd and Kenny, 1981; Grace, 2006) or mechanism through which disturbance affects the metric. Adjusting for variables that act as mediators would result in discarding part of the true relationship between D and m, a serious error of a different sort. This provides yet another reason why the development of a causal network for a system can guide the MMI development process.

Failing to adjust for environmental covariates can lead to biased metrics and MMIs. However, these effects can be mitigated by modeling and adjusting metrics. Both reference-set and whole-set residualization can be effectively used when the causal relationships among the metrics, human disturbance and the environmental covariates are simple. But, where the relationships are complex, only Whole-Set Residualization results in robust, precise and efficient adjustment of metrics.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ecolind.2013.01.015.

References

Andreasen, J.K., O'Neill, R.V., Noss, R., Slosser, N.C., 2001. Considerations for the development of a terrestrial index of ecological integrity. Ecol. Indic. 1, 21–35.

Baker, E.A., Wehrly, K.E., Seelbach, P.W., Wang, L., Wiley, M.J., Simon, T., 2005. A multimetric assessment of stream condition in the northern lakes and forests ecoregion using spatially explicit statistical modeling and regional normalization. Trans. Am. Fish. Soc. 134, 697–710.

Blocksom, K.A., 2003. A performance comparison of metric scoring methods for a multimetric index for mid-Atlantic highlands streams. Environ. Manage. 31, 670–682.

Barbour, M.T., Gerritsen, J., Snyder, B.D., Stribling, J.B., 1999. Rapid Bioassessment Protocols for Use in Streams and Rivers: Periphyton, Benthic Macroinvertebrates and Fish, EPA 841-B-99-002. United States Environmental Protection Agency, Office of Water, Washington, DC, USA.

Cao, Y., Hawkins, C.P., Olson, J., 2007. Modeling natural environmental gradients improves the accuracy and precision of diatom-based indicators. J. N. Am. Benthol. Soc. 26, 566–585.

Cohen, J., Cohen, P., West, S.G., Aiken, L.S., 2003. Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences, Third edition. Routledge Publications, New York.

Dodd, W.K., Oakes, R.M., 2004. A technique for establishing reference nutrient concentrations across watershed affected by humans. Limnol. Oceanogr. Methods 2, 333–341.

Ford, E.D., Ishii, H., 2000. The method of synthesis in ecology. Oikos 93, 153–160.

Gough, L., Osenberg, C.W., Gross, K.L., Collins, S.L., 2000. Fertilization effects on species density and primary productivity in herbaceous plant communities. Oikos 89, 428–439.

Grace, J.B., 1999. The factors controlling species density in herbaceous plant communities: an assessment. Perspect. Plant Ecol. 2, 1–28.

Grace, J.B., 2006. Structural Equation Modeling and Natural Systems. Cambridge University Press, New York.

Hawkins, C.P., Yong, C., Roper, B., 2010. Method of predicting reference condition biota affects the performance and interpretation of ecological indices. Freshwater Biol. 55, 1066–1085.

Hering, D., Feld, C.K., Moog, O., Ofenbock, T., 2006. Cook book for the development of a Multimetric Index for biological condition of aquatic ecosystems: experiences from the European AQEM and STAR project and related initiatives. In: Furse, M.T., Hering, D., Brabec, K., Buffagni, A., Sandin, L., Verdonschot, P.F.M. (Eds.), The Ecological Status of European Rivers: Evaluation and Intercalibration of Assessment Methods. Hydrobiologia 566, 311–324.

Judd, C.M., Kenny, D.A., 1981. Process analysis: estimating mediation in treatment evaluations. Eval. Rev. 5, 602–619.

Karr, J.R., 1981. Assessment of biotic integrity using fish communities. Fisheries 6, 21–27.

Karr, J.R., 1991. Biological integrity: a long-neglected aspect of water resource management. Ecol. Appl. 1, 66–84.

Karr, J.R., Chu, E.W., 1997. Biological Monitoring and assessment: Using Multimetric Indexes Effectively, EPA 235-R97-001. United States Environmental Protection Agency, Office of Water, Washington, DC, USA.

Kilgour, B.W., Stanfield, L.W., 2006. Hindcasting reference conditions in streams. In: Hughes, R.M., Wang, L., Seelbach, P.W. (Eds.), Influences of Landscapes on Stream Habitats and Ecological Assemblages. American Fisheries Society, Bethesda, Maryland, pp. 623–639.

Kimberling, D.N., Karr, J.R., Fore, L.S., 2001. Measuring human disturbance using terrestrial invertebrates in the shrub-steppe of eastern Washingtion(USA). Ecol. Indic. 1, 63–81.

Kurtz, J.C., Jackson, L.E., Fisher, W.S., 2001. Strategies for evaluating indicators based on guidelines from the Environmental Protection Agency's Office of Research and Development. Ecol. Indic. 1, 49–60.

Mack, J.J., 2001. Vegetation Indices of Biotic Integrity (VIBI) for Wetlands: ecoregional, hydrogeomorphic, and plant community comparison with preliminary wetland aquatic life use designations. Final Report to U.S. EPA Grant No. CD985875, vol. 1. Wetland Ecology Group, Division of Surface Water, Ohio Environmental Protection Agency, Columbus, Ohio.

Matsumoto, M., Nishimura, T., 1998. Mersenne twister: a 623-dimensionally equidistributed uniform pseudorandom number generator. ACM Trans. Model. Comput. Simul. 8, 3–30.

Mittelbach, G.G., Steiner, C.F., Scheiner, S.M., Gross, K.L., Reynolds, H.L., Waide, R.B., Willig, M.R., Dodson, S.I., Gough, L., 2001. What is the observed relationship between species richness and productivity? Ecology 82, 2381–2396.

O'Connor, R.J., Walls, T.E., Hughes, R.M., 2000. Using multiple taxonomic groups to index the ecological condition of lakes. Environ. Monit. Assess. 61, 207–229.

Pearl, J., 2009. Causality. Cambridge University Press, New York.

Pearl, J., 2010. An introduction to causal inference. Int. J. Biostat. 6, 1–59.

Pedhazur, E.J., 1997. Multiple Regression in Behavioral Research, Third edition. Wadsworth Publ., Belmont, California.

Pont, D., Hugueny, G., Beier, U., Goffaux, D., Melcher, A., Noble, R., Rodgers, C., Roset, N., Schmutz, S., 2006. Assessing river biotic condition at a continental scale: a European approach using functional metrics and fish assemblages. J. Appl. Ecol. 43, 70–80.

Rocchio, J., 2006. Vegetation index of biotic integrity for Southern Rocky Mountain fens, wet meadows, and riparian shrublands: phase 1 final report. Unpublished report prepared for the Colorado Department of Natural Resources and US EPA Region 8. Colorado Natural Heritage Program, Colorado State University, Fort Collins, Colorado.

Stoddard, J.L., Herlihy, A.T., Peck, D.V., Hughes, R.M., Whittier, T.R., Tarquinio, E., 2008. A process for creating multimetric indices for larger-scale aquatic surveys. J. N. Am. Benthol. Soc. 27, 878–891.

Wallace, B.J., Grubaugh, J.W., Whiles, M.R., 1996. Biotic indices and stream ecosystem processes: results from and experimental study. Ecol. Appl. 6, 140–151.

Whittier, T.R., Hughes, R.M., Stoddard, J.L., Lomincky, G.A., Peck, D.V., Herlihy, A.T., 2007. A structured approach for developing indices of biotic integrity: three examples from streams and rivers in the Western USA. Trans. Am. Fish. Soc. 136, 718–735.

Wiley, M.J., Seelbach, P.W., Wehrly, K., Smith, J.S., 2003. Regional ecological normalization using linear models: a meta-method for scaling stream assessment indicators Chapter 12. In: Simon, T.P. (Ed.), Biological Response Signatures: Indicator Patterns Using Aquatic Communities. CRC Press, Boca Raton, Florida.