2013

An algorithmic and information-theoretic approach to multimetric index construction

Donald R. Schoolmaster Jr.
Five Rivers Services, schoolmasterd@usgs.gov

James B. Grace
U.S. Geological Survey, gracej@usgs.gov

E. William Schweiger
National Park Service

Glenn R. Guntenspergen
U.S. Geological Survey, Glenn_Guntenspergen@usgs.gov

Brian R. Mitchell
National Park Service

See next page for additional authors

Ecological Indicators 26 (2013) 14–23

Contents lists available at SciVerse ScienceDirect

Ecological Indicators

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

An algorithmic and information-theoretic approach to multimetric index construction

Donald R. Schoolmaster Jr.[a,*], James B. Grace[b], E. William Schweiger[c], Glenn R. Guntenspergen[d], Brian R. Mitchell[e], Kathryn M. Miller[f], Amanda M. Little[g]

[a] Five Rivers Services, LLC at U.S. Geological Survey, National Wetlands Research Center, 700 Cajundome Blvd., Lafayette, LA 70506, USA
[b] U.S. Geological Survey, National Wetland Research Center, 700 Cajundome Blvd., Lafayette, LA 70506, USA
[c] National Park Service, Rocky Mountain Network, 1201 Oakridge Drive, Fort Collins, CO 80525, USA
[d] U.S. Geological Survey, Patuxent Wildlife Research Center, Laurel, MD 20708, USA
[e] National Park Service, Northeast Temperate Network, Woodstock, VT 05091, USA
[f] National Park Service, Northeast Temperate Network, Acadia National Park, Bar Harbor, ME 04609, USA
[g] Biology Department, University of Wisconsin-Stout, Menomonie, WI 54751, USA

ARTICLE INFO

Article history:
Received 8 May 2012
Received in revised form 9 October 2012
Accepted 18 October 2012

Keywords:
Index of biological integrity
Information theory
Bioassessment
Disturbance

ABSTRACT

The use of multimetric indices (MMIs), such as the widely used index of biological integrity (IBI), to measure, track, summarize and infer the overall impact of human disturbance on biological communities has been steadily growing in recent years. Initially, MMIs were developed for aquatic communities using preselected biological metrics as indicators of system integrity. As interest in these bioassessment tools has grown, so have the types of biological systems to which they are applied. For many ecosystem types the appropriate biological metrics to use as measures of biological integrity are not known a priori. As a result, a variety of ad hoc protocols for selecting metrics empirically has developed. However, the assumptions made by proposed protocols have not be explicitly described or justified, causing many investigators to call for a clear, repeatable methodology for developing empirically derived metrics and indices that can be applied to any biological system. An issue of particular importance that has not been sufficiently addressed is the way that individual metrics combine to produce an MMI that is a sensitive composite indicator of human disturbance. In this paper, we present and demonstrate an algorithm for constructing MMIs given a set of candidate metrics and a measure of human disturbance. The algorithm uses each metric to inform a candidate MMI, and then uses information-theoretic principles to select MMIs that capture the information in the multidimensional system response from among possible MMIs. Such an approach can be used to create purely empirical (data-based) MMIs or can, optionally, be influenced by expert opinion or biological theory through the use of a weighting vector to create value-weighted MMIs. We demonstrate the algorithm with simulated data to demonstrate the predictive capacity of the final MMIs and with real data from wetlands from Acadia and Rocky Mountain National Parks. For the Acadia wetland data, the algorithm identified 4 metrics that combined to produce a −0.88 correlation with the human disturbance index. When compared to other methods, we find this algorithmic approach resulted in MMIs that were more predictive and comprise fewer metrics.

1. Introduction

Multimetric indices (MMIs) have become important tools for the assessment of the quality of biological resources. MMIs consist of a number of biological and/or ecological indicators (often called metrics because they are measures of the system) that are combined to act as a one-dimensional indicator of the biological or ecological condition of a system. The biological or ecological condition of systems can then, using MMIs, be graded based on how much they have been altered by human disturbance (Karr and Chu, 1997).

MMIs constitute an outgrowth of the bioindicator tradition (Karr, 1981). In this tradition, intensive studies are used to measure both biological responses and the degree of human disturbance in order to develop a bioindicator MMI that can then be used more extensively. This approach is typically used where sources of human disturbance occur at unknown and multiple spatial and temporal scales and the disturbance history of areas is undocumented or unknown. In such situations, the component metrics of MMIs are often easier to observe and measure than human

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

1470-160X/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.ecolind.2012.10.016

disturbance itself. As a result, biologists, ecologists and managers commonly rely on MMIs to estimate the degree to which the biological responses of a system have been affected by anthropogenic disturbance and thereby grade system condition.

Since the introduction of MMIs as tools for the bioassessment of water quality (Karr, 1981), the concept has been applied to a rapidly growing set of biological systems, including wetland plants (Mack, 2001; DeKeyser et al., 2003; Ferreira et al., 2005; Miller et al., 2006; Rocchio, 2006; Rothrock et al., 2007), 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). Given the rapid growth in the development and use of these tools, it is essential to establish clear and repeatable methods for the selection of metrics.

In some cases, where systems and their stressors are well-known, metrics are chosen to reflect expert knowledge of the state of a "healthy" system. This is the approach advocated by Karr (1981) in developing an Index of Biological Integrity (IBI) for fish assemblages. In these cases, the scientist or manager employs the index to assess deviations from the predetermined high-integrity state. This approach works well when the causal pathways between the elements of human disturbance and the elements of biological integrity (e.g. the metrics) are well established (i.e. the causal structure of Fig. 1a is known). Because they are based on knowledge of the causal processes, MMIs developed from well-supported theoretical understanding of the system have the significant benefit of being general, so that the same MMI can be used in many places, and the results meaningfully compared.

However, in many other cases, the causal networks linking human disturbance to the measured biological metrics are complex and the mediating factors are unknown or under debate. Managers are nonetheless charged with assessing the impact of human disturbance on systems, allocating resources to remediate the impacts and measuring the effects of remediation efforts. In this case, where theory is insufficient to determine what metrics best measure the impact of human disturbance, it is necessary to determine the combinations of factors that are most sensitive to human disturbance from the biological response data. An overview of the construction of empirically derived MMIs and their connection of causal processes is outlined in Fig. 1a–c. The drawback of empirically derived MMIs is that because they are phenomenological, as opposed to being based on causal knowledge, they are likely to be less general than those derived from theory. Until the causal network linking the elements for human disturbance to the metrics of the MMI is established, applying MMIs to different ecological contexts or spatial scales is largely a descriptive procedure (Ode et al., 2008).

The benefit of deriving a MMI empirically is that it is often possible to create a model that is highly predictive, i.e., is highly correlated with human disturbance in data outside of those used to construct the index. In addition, empirically derived MMIs can be used as a starting point for further analyses, via structural equation modeling (e.g. Riseng et al., 2006; Grace et al., 2012), to establish the network of causation in the system, uncover the mediating variables, and thus further the theoretical understanding of the system (Grace, 2006). We refer to the empirically derived MMIs as a BioIndicator Indices (BIIs) because they are not purely indices of biological/ecological integrity.

The data used to derive and calibrate MMIs usually consist of, for each of a number of sites, some measure of human disturbance and an associated table of candidate response metrics. The candidate metrics are usually chosen to span different levels of biological organization from individual condition (e.g., number of diseased individuals), to community level measures (e.g. species richness). The goal is create an index from the subset of candidate metrics that is sensitive to human disturbance (Fig. 1). Once this is achieved, deviation from a non-disturbed, or reference state (Stoddard and Larsen, 2006) can be computed and the quantitative relationship between the measure of human disturbance and the MMI can be used to identify those sites that deviate from the natural variation in the (hypothetical) non-disturbed state (Appendix A).

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ecolind.2012.10.016.

Currently, there are a number of published methodologies for constructing empirically derived MMIs from data (Ofenböck et al., 2004; Hering et al., 2006; Whittier et al., 2007; Stoddard et al., 2008). These methodologies present a number of useful steps for narrowing a set of candidate metrics based on particular benchmarks. For example, Ofenböck et al. (2004) suggests that metrics for which the value at most sites is zero be eliminated for the set of candidates. Stoddard et al. (2008) suggest, among other things, eliminating metrics which display a highly temporally variable relationship with human disturbance (which they refer to as signal to noise ratio). These steps are useful for quickly narrowing the candidate set of metrics to those that could be potentially beneficial to the final MMI. However, each of these methodologies ultimately

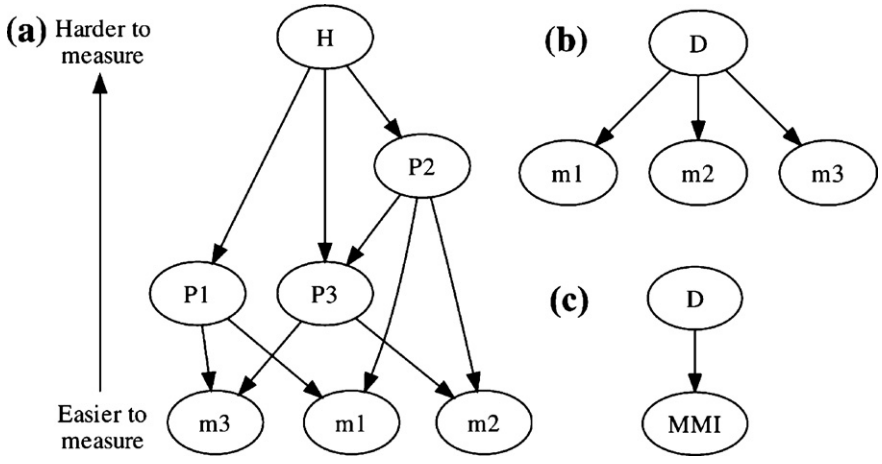

Fig. 1. Schematic of multimetric indices (MMI) construction process. (a) Causal network showing our assumption that human activities (H) affect physical and chemical properties in the environment (P1–P3), which in turn affect biological community metrics (m1–m3). In the process for developing an overall index of human disturbance (D), multiple measures of physical and chemical impacts (P1–P3) are combined. (b) The human disturbance index is then used to evaluate candidate metrics for use in constructing an MMI. Finally, (c) metrics are combined into a single MMI, effectively reducing a many dimensional system (i.e. graph a) into a two dimensional system c. Adapted from Schoolmaster et al. (2012).

recommends selecting or eliminating metrics from the set based on the magnitude and statistical significance of the bivariate relationship with human disturbance. It is our contention that, since human disturbance has a complex, multivariate affect on biological systems, it is the multivariate relationship with human disturbance that needs to be evaluated and incorporated into index construction. In practice, this means selecting sets of metrics based on how they respond to human disturbance as a group, as opposed to how each metric responds individually.

In a recent quantitative analysis of the general properties of MMIs, Schoolmaster et al. (2012) showed that the characteristics of metrics that will combine to produce the most sensitive MMI cannot be deduced from the bivariate relationships between the metrics and human disturbance. Instead, whether a candidate metric improved the sensitivity of an MMI (or not) depends on the strength of the disturbance signal in the candidate metric and its unique explanatory power within the set of candidate metrics. This suggests that the way that metrics combine to convey information about disturbance should be incorporated into empirical MMI construction.

In addition to the complexities associated with determining the identity of metrics to use for an MMI, there is also a decision to be made about how many metrics should be included. Some recent descriptions of MMI construction advocate classifying metrics into groups based on level of biological organization and then choosing at least one from each group (Karr and Chu, 1997; Hering et al., 2006) or at most one from each group (Whittier et al., 2007; Stoddard et al., 2008). No statistical criteria has been established to determine which of these choices is most effective, how many groups should be identified or, in the former case, how many metrics to select from each group. Schoolmaster et al. (2012) found that, under conditions expected to be common for most studies, the sensitivity of an MMI is a non-monotonic function of the number of metrics included; that is, there is some number of metrics that will maximize the sensitivity, but that number depends on the distribution of disturbance signal across metrics, and the degree of inter-correlation among metric errors.

In this article, we present an algorithmic approach to MMI construction. The goal of the algorithm is to produce a maximally sensitive MMI from a given set of candidate metrics and a measure of human disturbance in a fashion that is explicitly determined and repeatable. Where there are decisions to be made among alternative choices, such as whether to add another metric to the MMI, we employ an information theoretic criterion (Burnham and Anderson, 2002) to inform the process. The purpose of this is to make empirical MMI construction less subjective, more efficient and reproducible. In the treatment that follows, we first describe the steps necessary to perform this method and then demonstrate the approach first using simulation and finally with real data from wetland research projects at Acadia National Park and Rocky Mountain National Park, USA.

2. Methods

2.1. Data preparation before MMI assembly

The first step necessary to prepare data for analysis is to remove from the set of candidate metrics any that are temporally inconsistent (Stoddard et al., 2008), contain a large proportion of zeros (Ofenböck et al., 2004; Stoddard et al., 2008) or duplicate another metric. There are a number of situations that can result in duplicate metrics. For example, in a system without woody plants, metrics defined as percent of the community composed of forbs and the percent composed of dicots would contain the exact same values. Additionally, metrics defined by mutually exclusive categorizations may contain the exact same information. For example, the percent of a community composed of native species and the percent that is non-native. As a useful rule, duplicate metrics can be identified by looking for very high values of inter-metric correlation (e.g. >0.95 or <−0.95). Note that this step is intended to remove identical metrics and differs from the step of removing "redundant" metrics suggested by others (Hering et al., 2006; Whittier et al., 2007; Stoddard et al., 2008), and should not be judged using simple inter-metric correlation (Schoolmaster et al., 2012).

The next step after the initial culling step above is to adjust the metrics for environmental covariate effects. Environmental covariates can obscure or exaggerate the relationship between a metric and human disturbance and can also inflate the correlation of metric errors. Schoolmaster et al. (unpublished data) found that the most robust and efficient method of adjusting for environmental covariates is Whole Set Residualization (WSR). This method models each metric as a function of human disturbance and the environmental covariates simultaneously and then adjusts metrics based on estimates from the model where the parameters associated with the environmental covariates are set to zero. For example, if human disturbance and the value of a metric such as species diversity both decrease with elevation, we would estimate the parameters β of the model

$$m = \beta_0 + \beta_D D + \beta_E E + \varepsilon$$

where E is the measure of elevation and ε contains the residuals. The value of the adjusted metric is calculated as

$$m_{adj} = \beta_0 + \beta_D D + \varepsilon,$$

leaving the signal associated with a response to disturbance in the adjusted metric.

After metrics are adjusted, they must be rescaled (sometimes called scored) to unitless measures with similar ranges. This allows metrics, which naturally have different units, to be combined and allows for each to have similar weight on the final MMI score. Blocksom (2003) compared various methods for rescaling metrics and concluded that a continuous scaling method that recodes outliers performed the best for MMI construction. For example,

$$m_{scaled} = \frac{m - L}{U - L} \quad (1)$$

where m_{scaled} is the rescaled metric score, L and U are the 2.5 and 97.5 percentile values of m. Values of m that are less than L or greater than U are set to L or U respectively.

While we have been drawing distinct lines between the empirically based and theoretical-based approaches to MMI construction, it is possible to use theory, expert opinion or other criteria such as the temporal stability to influence the selection of metrics by this algorithm. To do this, one could represent each metric as an influence of disturbance plus some residual error; however, the residual error could be scaled by the inverse of a weighting factor that reflects a theoretical bias for or against the metric. This has the effect of increasing (or decreasing) the correlation of the adjusted metric with human disturbance, and its likelihood of being selected by the algorithm, while leaving the inter-correlation structure of metric errors unchanged. After the assembly algorithm (described below) is applied to choose metrics, the final MMI scores are calculated from metric scores to which the weightings have not been applied. We refer to indices assembled using a weight vector that favors conceptually preferred metrics (e.g. native species richness might be preferred over total species richness) as "value-weighted". This approach to value-weighting permits explicit and repeatable consideration of human valuation in metric weighting.

2.2. MMI assembly

The goal of empirical MMI assembly is to construct an MMI from candidate metrics that has the strongest (predictive) relationship with human disturbance (without being overfitted, e.g., unable to be extrapolated). This idea can be re-expressed symbolical as choosing the vector of parameters α of the expression

$$\mathrm{MMI} = \frac{1}{\sum_j^N \alpha_j} \sum_i^N \alpha_i m_i \quad (2)$$

that maximizes $P(D|\mathrm{MMI})$, where N is the number of candidate metrics and the parameters in α can take on values of either 0 or 1 (i.e. $\alpha \in \{0,1\}$). Note that upon choosing the values of α, Eq. (2) is the mean of the metric values m_i for which $\alpha_i = 1$.

One approach for finding the values of α that maximize $P(D|\mathrm{MMI})$ is to do a complete search of all possible combinations of the parameters α_i. Such a search requires 2^N calculations and is infeasible for $N > 20$. We propose a heuristic for this optimization problem that is much faster and not limited by the size of the set of candidate metrics. Our solution is a form of "greedy algorithm" (Cormen et al., 2009), which attempts to solve the problem by breaking it into a number of sub-problems and making a series of locally optimal choices with the goal of finding a globally optimal solution.

The strategy is to use each candidate metric to inform the selection process, resulting, initially, in as many candidate MMIs as there are candidate metrics. We then use various criteria to narrow the set of candidate MMIs to a much smaller set of those with the highest predictive capacity. A final MMI can be chosen from this smaller set or, alternatively, the unique information in each MMI in the smaller set can be combined using a model averaging approach (Burnham and Anderson, 2002).

What follows is a list of steps (i.e., a pseudo-code for an algorithm) that can be used to generate a sensitive MMI from a given set of metrics. These steps can be developed into computer code to quickly generate a set of candidate MMIs. This method assumes one wants an MMI with the strongest possible negative correlation with human disturbance and that metrics have already been adjusted for correlated environmental gradients and scaled. This method, as presented also assumes that metrics and the final MMI are linear functions of the measure of disturbance, and that disturbance is measured on a continuous scale. These assumptions are made for the sake of simplicity and the method can be easily generalized to account for non-linear relationship and discrete measures of human disturbance (e.g. reference/disturbed). The algorithm is as follows:

1. Reflect metrics that are positively correlated with disturbance D about their midpoint (i.e. $m_{ref} = \max(m) - m$) to ensure MMI is negatively correlated with D.
2. Select an initial metric to include in the MMI, m_1. This selection can be made arbitrarily since all metrics will eventually be used as an initial metric
3. Add m_1 to each of the rest of the metrics, m_j, site-by-site
4. For each m_j, find which combination $m_1 + m_j$ that has the strongest negative correlation with D. Select that one.
5. Record the correlation (or other statistic) of the relationship between the assembled index with D
6. Add the index to each of the remaining metrics m_j site-by-site
7. Find the combination of index + m_j has the strongest negative correlation with D. Select that one.
8. Record the correlation or other statistics of the relationship between the assembled index with D
9. Continue steps 6–8 until a stopping rule (discussed below) is satisfied, or all metrics have been used
10. Repeat steps 2–9 until all metrics have been used as initial metric, m_1

Notice that each metric in the candidate set is used as the initial metric to inform a candidate MMI, thus this algorithm outputs one MMI for each candidate metric. The problem of selecting the best MMI then becomes a model selection problem.

2.3. Decision rules to narrow set of candidate MMIs

As stated previously, the algorithmic process described above results in a number of (not necessarily unique) candidate MMIs equal to the number of candidate metrics. Each candidate MMI is assembled from a list of metrics (given the initial metric) ordered by ability of added metrics to *combine* with those above it to most strengthen (or least weaken) the predictive capacity between the MMI and D (the human disturbance measure). The first decision that must be made is how many metrics should be included in each candidate MMI. Analytical theory (Schoolmaster et al., 2012) suggests that in most cases MMIs can be expected to reach a maximum predictive capacity then decline as metrics are added (Fig. 2). One potential method is to include the metrics that result in the strongest correlation (highest R^2) of the MMI and D. However, as the maximum is approached, each additional metric tends to improve the MMI less and less (Fig. 2). Given that metric data may be hard or expensive to obtain, it is desirable to exclude metrics that do not contribute "significantly" to the correlation.

As a means of deciding what constitutes a significant contributor, we suggest the use of the likelihood ratio test. In this context, the likelihood ratio test is calculated as

$$\Lambda = -2(L(D|\theta, \mathrm{MMI}_n) - L(D|\theta, \mathrm{MMI}_{n+1})),$$

where L is the log likelihood of D, given θ. In this function, θ is a vector of the parameters α and the parameters of linear model $D = \beta_0 + \beta_1 \mathrm{MMI}_n$ that maximizes L. MMI_k $k \in \{n, n+1\}$ is the MMI comprised of the first k metrics of the list for the candidate MMI. The value Λ is approximately chi-square distributed with one degree of freedom (i.e., the difference in the number of metrics in each

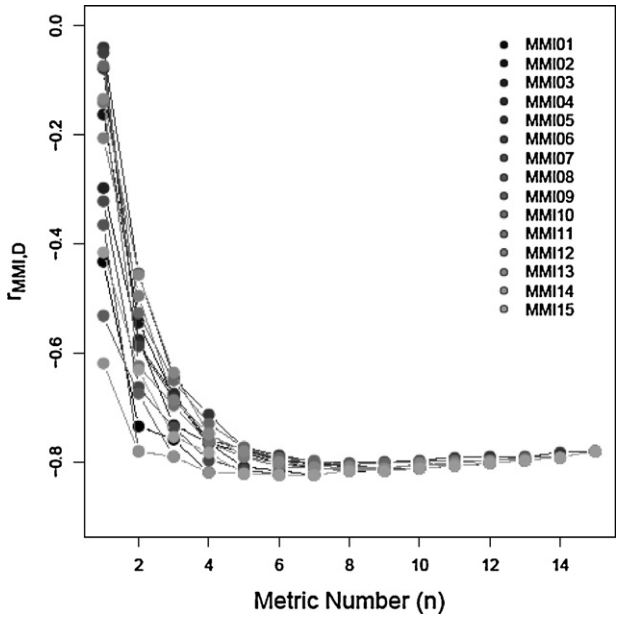

Fig. 2. Correlation between the MMI and human disturbance at each step of the MMI selection process using simulated data. For each MMI, the correlation reaches a minimum at some intermediate number of metrics, but they differ in the number and identity of metrics that result is the strongest correlation.

MMI01	MMI02	MMI03	MMI04	MMI05	MMI06	MMI07	MMI08	MMI09	MMI10	MMI11	MMI12	MMI13	MMI14	MMI15
m01	m02	m03	m04	m05	m06	m07	m08	m09	m10	m11	m12	m13	m14	m15
m14	m08	m14	m14	m14	m14	m14	m14	m08	m14	m14	m14	m01	m08	m14
m09	m14	m08	m09	m08	m09	m08	m09	m14	m01	m01	m08	m14	m09	m08
m08	m01	m09	m08	m09	m08	m09	m01	m01	m09	m09	m01	m08	m01	m09
↑	m09	m01	m01	m01	m15	m01	↑	↑	m08	m08	m09	m09	↑	m01
	m07		m15	m15	m01	m15			m15	m15	m07	m07		
				m07							m15	m15		

Fig. 3. Snapshot of resulting candidate MMIs after metric elimination criterion (based on likelihood ratio) was applied. Arrows indicate duplicate MMIs.

MMI). Thus, we can define a rule to stop adding metrics to the MMI where $\Lambda < 3.84$, which is the value of the chi-squared distribution that corresponds with $p = 0.05$.

As an example, to decide if the $n+1$th metric should be retained in the candidate MMI, we suggest the following: (1) Calculate the MMI scores using the first n metrics on the candidate MMI list (MMI_n) and the first $n+1$ (MMI_{n+1}), for each of these calculate $L(D|\theta, \text{MMI}_k)$. (2) Calculate Λ. (3) If $\Lambda > 3.84$ then retain the $n+1$th metric in the MMI and test the next. If $\Lambda < 3.84$ then the $n+1$th metric is not retained and no additional metrics are tested (example of output results is shown in Fig. S1). There are other choices for stopping rules, such as AIC, but we chose the likelihood ratio test for this application because it is conservative at the degrees of freedom used here.

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ecolind.2012.10.016.

The next task is to choose from among the candidate MMIs. Since it is likely that some of the candidate MMIs will duplicate others, the set of candidate MMIs must be narrowed by eliminating any duplicates (Fig. 3).

Given the narrowed set of candidate MMIs we choose the "best" by calculating the Akaike Information Criterion (AIC) of each MMI. AIC is a measure of the goodness of fit of a model relative to those in a set of candidate models (Burnham and Anderson, 2002). It is an unbiased estimate of a model's expected predictive capacity. AIC is calculated as $\text{AIC}_j = 2L_j + 2k_j$, where L_j, is the log-likelihood of the jth model $D = \beta_0 + \beta_1 \text{MMI}_j$ and k_j is the dimensionality of the set of predictors ($2 + \sum_i^N \alpha_i$ or number of metrics in $\text{MMI}_j + 2$) in the jth model. In cases where the sample size (number of sites) is not much larger than k^2, Burnham and Anderson (2002) suggest using a measure of AIC that corrects for small sample bias, called AICc (Hurvich and Tsai, 1989),

$$\text{AICc} = -2L + 2k + \frac{2k(k+1)}{n-k-1},$$

where n is the sample size (number of sites). AIC is an estimate of amount of information "lost" relative to a true model (Burnham and Anderson, 2002). Therefore, smaller values of AIC (or AICc) indicate better models.

After the AIC of each MMI has been calculated, each is judged based on the difference between its AIC and that of the MMI with the lowest AIC in the set, $\Delta_j = \text{AIC}_j - \min(\text{AIC})_j$. MMIs with values of $\Delta < 2$ are judged to have substantial support, and should be considered as viable alternatives to the model with the lowest AIC (Burnham and Anderson, 2002). Those with $\Delta > 2$ can be eliminated from the set of candidate MMIs. The threshold of $\Delta = 2$ is not a hard rule. A researcher is free to set the threshold to any number they are comfortable with.

Additionally, MMIs that contain all the metrics of the MMI with the lowest AIC plus one more and that have a value of Δ near the penalty for the additional parameter (i.e. 2 for AIC, larger for AICc), could be reasonably dropped from the candidate list since the additional variable does not increase the log-likelihood of the model. This situation is referred to as the "pretending variable problem" by Anderson (2008), though "pretending variables" are unlikely to occur in this application. In traditional model selection problems, where each variable is accompanied by a continuous real-valued parameter that must be estimated, "pretending variables" occur when the value of the parameter for the additional variable is close to zero. In this application, each metric is given equal weight. As a result, the possibility of "pretending variables" is greatly decreased.

2.4. Creating a model-weighted MMI

If all MMIs except the one with the lowest Δ_j have $\Delta_j > 2$, then the candidate MMI with the lowest Δ_j is selected and no further analysis is necessary. Alternatively, any one MMI can be chosen from the set of models with $\Delta_j < 2$ without much loss of predictive capacity. However, if multiple candidate MMIs have $\Delta_j < 2$, then there is uncertainty which MMI will best predict D. This uncertainty can be addressed by model averaging. Model averaging uses a measure of the probability that each model in a set is the best to create weighted averages over the models' parameter values and model predictions. Akaike weights are calculated as

$$w_j = \frac{\exp(-\Delta_j/2)}{\sum_z^Z \exp(-\Delta_z/2)},$$

where Z is the number of models with $\Delta_j < 2$, and z in the index of summation. Model weighted MMI values are calculated as $\text{MMI}_{wt} = M\bar{w}$, where M is at matrix of MMI values made by combining the model values of the remaining MMIs column-wise and $\bar{w}$ is the vector of model weights.

The Akaike weights can also be used to measure the influence of each metric on the final weighted MMI. These metric weights can be calculated as

$$\text{met}_{wt} = C\bar{w}$$

where C is a matrix with a row for each metric, a column for each remaining MMI, and filled with 1s and 0s to indicate the presence or absence each metric in each MMI. The values in met_{wt} range from 0 to 1. If the model weights are each scaled by the number of metrics in the respective MMI, n_j, $w_j^* = w_j/n_j$ then the model weighted MMI scores can be computed directly as

$$\text{MMI}_{wt} = SC\tilde{w}^*$$

where S is the matrix of adjusted, scaled and reflected (where necessary) metric scores.

3. Examples

In this section we demonstrate the use of the empirical MMI construction algorithm with two data sets, one simulated and one from a wetland study at Acadia National Park. Using simulated data allows us to test the predictive capacity of the MMI on other simulated data sets of similar form, while the Acadia MMI demonstrates the method's efficacy in the face of the complexities of real data.

3.1. Example 1: simulation data

To demonstrate our MMI construction approach, we simulated 15 metrics as

$$m_i = \beta_i D + \varepsilon_i,\ \varepsilon \sim N(0, \Sigma),$$

where m_i, ε_i and D are vectors with one value per each of 100 sites and ε is multivariate normally distributed with mean of zero and variance/covariance equal to Σ. Values of the measures of human disturbance D were generated as uniform distributed random variables between 0 and 10. Values of β were drawn from a standard normal distribution ($\mu = 0$ and $\sigma = 1$). The matrix of metric errors ε was simulated as follows. A matrix $A = aa^T$ was created, where a is a column vector of random variables selected from a beta distribution with parameters $\alpha = 2$ and $b = 3$. This resulted in a symmetric, positive definite matrix. The diagonal of A was then replaced with 1s. A 100×15 matrix, E of uncorrelated errors was created by sampling a normal distribution with $\mu = 0$ and $\sigma = 10$. Finally, ε was calculated as $\varepsilon = EL$, where L is the result of the Cholesky decomposition of A. The candidate metrics were scaled using the method described in Eq. (1).

The simulated set of metrics was sent to a function built on the R platform (2008) to carry out the algorithmic assembly described above. The result was an initial set of 15 candidate MMIs containing 15 metrics (Fig. S1). Each candidate MMI has an associated list of the order in which metrics were added to it to produce an MMI with the best possible correlation with D. The ordered list of metrics was used to calculate the negative log-likelihood of the linear relationship between the implied MMI and D. Likelihood ratios were calculated from the list. We used the list of likelihood ratios to find the first metric in each list that failed to result in a likelihood ratio greater than 3.84. This metric, and all those below it were discarded from the candidate MMI (indicated by gray values in Fig. S1). For our simulated data, this resulted in candidate MMIs consisting of between two and nine metrics (Fig. 3)

Given the shortened set of candidate MMIs (Fig. 3), we discarded all but one in any subset that were duplicates of one another (it does not matter which you keep). For our simulated data, MMI01 was duplicated by MMI08, MMI09 and MMI14 (all had the same metrics, but in different orders). We then eliminated MMI08, MMI09 and MMI14 from the set of candidate MMIs. For each remaining candidate MMI, we calculated AICc and Δ (Fig. 4). Of the twelve remaining candidate MMIs six had $\Delta > 10$ and only three had $\Delta < 2$. We excluded all with $\Delta > 2$ and calculated AIC weights for the remaining MMIs (Fig. 5). Since MMI15 contains all the metrics in the best model plus one additional metric (m15), we examined the possibility that m15 was acting as a "pretending variable". Since we are using AICc the penalty for MMI15's additional parameter is 2.21. Since its value of Δ (1.18) is much smaller than the penalty, we are assured that m15 is adding information to the MMI and we retain MMI15 in the final set of MMIs. Each of the three remaining MMIs were strongly correlated with D, but each made different estimates for the MMI score of a given site (Fig. 6a). A model-weighted MMI score was calculated for each site using the scores from each of the MMIs and the AIC weights. The weighted MMI showed a stronger correlation with D than did any of the component MMIs (Fig. 6b).

MMI01	MMI07	MMI15
m01	m07	m15
m14	m14	m14
m09	m08	m08
m08	m09	m09
	m01	m01
	m15	
ΔAIC 0	1.97	1.18
AIC_{wt} 0.519	0.194	0.287

Fig. 5. Final snapshot table of candidate MMIs that pass the criterion $\Delta AIC < 2$.

3.1.1. Testing the predictive capacity of simulated MMI: methods

One potential drawback of algorithmic procedures such as the one proposed here is overfitting. Overfitting is the process of using additional degrees of freedom to fit the random variation in the data in hand. It has the effect of decreasing the predictive performance of the model when applied to new data. The information-theoretic tools we employ, such as decision rules based on AIC are designed to reduce the risk of overfitting. We tested the predictive performance of the MMIs generated by the algorithmic process against new data simulated sets of similar structure and assess their ability using the correlation between MMI scores and D for the new data.

To measure the variability in predictive performance, we recorded the correlation between MMI scores and D over 1000 simulated data sets. These new data sets were simulated using the same β and L used to generate the data used to create the MMI. However for each new data set new values of D and E where generated.

3.1.2. Testing predictive capacity of simulated MMI: results

Simulating new data sets (test sets) with the same structure as the one used to derive the MMI (calibration set) resulted in significant variation in the component metrics and the resulting correlation with D (Fig. 7). In some cases, the average of the correlation of the metrics with D in the test data was much different than the values in the calibration set used to derive the MMI. For

MMI01	MMI02	MMI03	MMI04	MMI05	MMI06	MMI07	MMI10	MMI11	MMI12	MMI13	MMI15
m01	m02	m03	m04	m05	m06	m07	m10	m11	m12	m13	m15
m14	m08	m14	m14	m14	m14	m14	m14	m14	m14	m01	m14
m09	m14	m08	m09	m08	m09	m08	m01	m01	m08	m14	m08
m08	m01	m09	m08	m09	m08	m09	m09	m09	m01	m08	m09
	m09	m01	m01	m01	m15	m01	m08	m08	m09	m09	m01
	m07		m15	m15	m01	m15	m15	m15	m07	m07	
				m07					m15	m15	
AIC 397.87	409.53	404.56	412.05	413.49	413.11	399.84	414.16	407.36	407.34	409.71	399.05
ΔAIC 0	11.66	6.69	14.18	15.62	15.25	1.97	16.29	9.49	9.48	11.84	1.18

Fig. 4. Snapshot table of candidate MMIs and resulting values of AIC and ΔAIC after duplicate MMIs were removed.

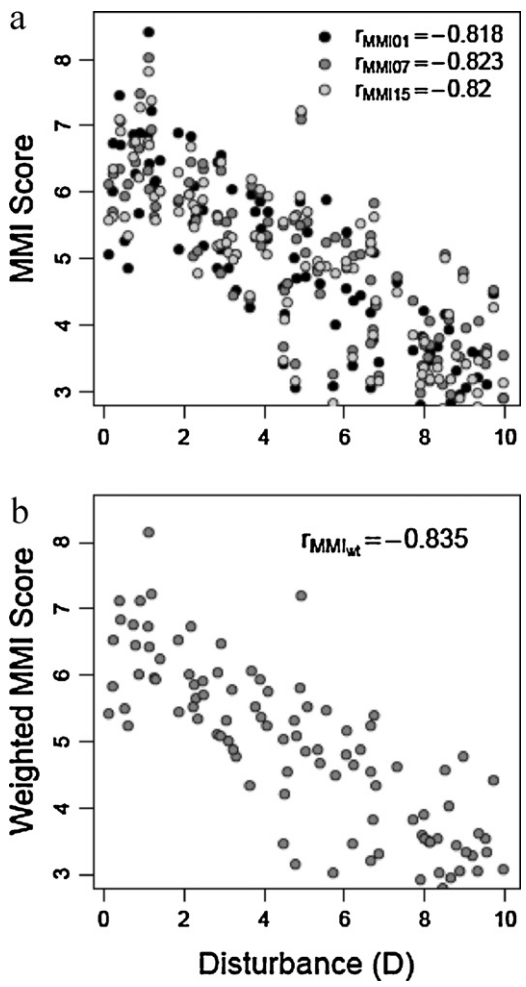

Fig. 6. (a) MMI scores as a function of human disturbance for each final candidate MMI and (b) model-weighted MMI, from simulated data.

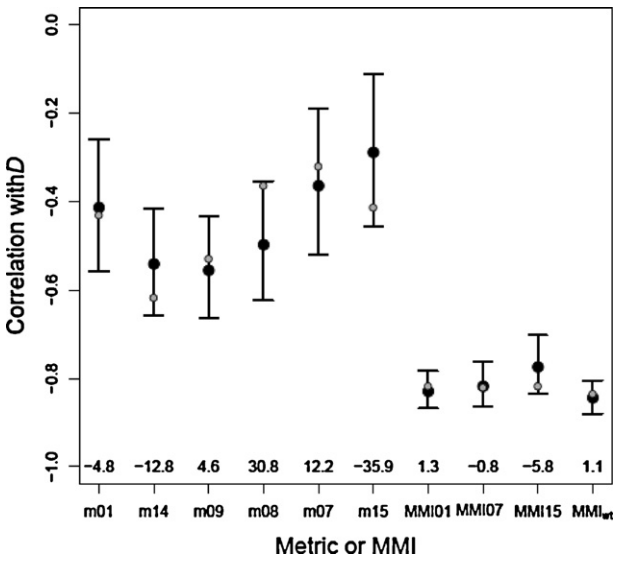

Fig. 7. Mean, 5th and 95th percentiles of distributions of correlations between selected metrics, candidate MMIs and model weighted MMI for 1000 simulated "test" data sets. Gray points indicate the observed values of the "calibration" data used to construct the MMIs. The numbers along the *x*-axis indicate the difference in correlation between the "calibration" data and the average of the "test" data (100×). Notice that the variation in correlation of the MMIs with human disturbance is much less than any of the component metrics (the major benefit of creating an index), and that variation in the model weighted MMI is less that that of the individual MMIs. If the MMI algorithm resulted in overfit models, the gray points would lie outside of the distribution of "test" sets.

example, the relationships between m14 and m15 with *D* in the calibration set (shown as gray points in Fig. 7) were much stronger than the average over the test data sets. This phenomenon is the type that can compromise the predictive performance of single metrics and that creating an index is supposed to reduce. Indeed, in the example, the benefits of indexing resulted in a correlation between each MMI and *D* that was much less variable than in the component metrics (Fig. 7). The variation in the MMIwt, which also compensates for model selection uncertainty, was even less variable. The predictive performance of each MMI was very good, although better for MMI01 and MMI07 than MMI15 due to the latter's dependence on m15. The percent difference between the mean correlations across test sets and those in the calibration set ranged from −5.8% for MMI15 to 1.3% for MMI01. MMIwt performed very well in the simulations. Among all the MMIs, the MMIwt had the strongest correlation with *D*, the smallest variation in correlation with *D* over test data sets, and the mean correlation between it and *D* in the test data sets was even stronger than in the calibration set, indicating no reduction in performance due to overfitting.

3.2. Example 2: Acadia National Park wetland data

Acadia National Park is located on Mount Desert Island on the coast of Maine (USA). Mount Desert Island is a 24,000 ha granite bedrock island and includes the highest mountain on the Atlantic coast of the U.S. The park occupies roughly 2/3 of the island, with private lands around its perimeter. Acadia is considered a high-integrity ecosystem (Tierney et al., 2009), though its unbuffered soils may result in some habitats being vulnerable to acidification and other changes in water chemistry. As a consequence of its mountainous topography, wetlands on the island are in relatively small catchments. The soils are shallow in the uplands while the wetlands are often peat-forming, receiving their water largely from acidic and low-nutrient inputs from rain and surface runoff (Kahl et al., 2000).

Little et al. (2010) have recently quantified human activities and measured biological characteristics of wetland plant communities on Mount Desert Island with the purpose of gauging their condition and vulnerability to human impacts. In these studies, 37 non-forested wetlands were examined as part of a bioassessment. In addition to measuring characteristics of the biological community, the investigators measured hydrological characteristics, water quality in the wetlands and potential natural environmental gradients. We used these data to quantify the degree and type of human disturbance in the vicinity of wetland catchments using a modification of the ORAM rating system developed for wetlands by the Ohio EPA (Mack, 2001). We then constructed a human disturbance index (HDI) for Acadia wetlands through modification of the ORAM assessment.

We applied the MMI construction methods detailed above by first creating forty-three candidate metrics of the wetland vegetation sampled at Acadia. Metrics were of three types, those that indicate (1) vegetation cover and (2) richness of plant groups at various taxonomic and functional levels (i.e. number of ferns species, cover of perennial plants) as well as (3) composite measures of plant community composition and structure such as average wetland indicator score, and height of dominant vegetation (Table 1). All metrics with fewer than 25% (i.e., 9) non-zero entries were eliminated from the set of candidate metrics. This resulted in the rejection of 16 metrics.

Next, we identified two environmental factors that could potentially confound measurement of the relationship between the metrics and human disturbance, distance from the ocean, and area

Table 1
Final best set of candidate MMIs selected by algorithm from unweighted metrics for Acadia National Park wetland vegetation data.

Statistic	MMI1	MMI2
	Forb richness	Log(*Typha* spp. cover)
	Forb cover	Perennial richness
	Log(*Typha* spp. cover)	
AIC wt.	0.728	0.272
Cor w/HDI	−0.872	−0.855

of the wetland. As described above, the metrics were adjusted for these factors (where necessary) and scaled using the WSR method and Blocksom's (2003) CAUL method respectively. Finally, we transformed variables as necessary to meet the assumptions of linear models.

We applied the index assembly algorithm described above to the set of remaining candidate metrics. We applied the algorithm twice; once on the unweighted metric values and a second time to demonstrate the effect of value-weighting metrics. In the value-weighted case, candidate metrics measuring the contribution of native species (i.e. native cover and native richness) were given twice the weight (weight = 1) of the remaining metrics (weight = 0.5). This was done to simulate a case where a resource manager who will use the MMI values the contribution of native species above other metrics.

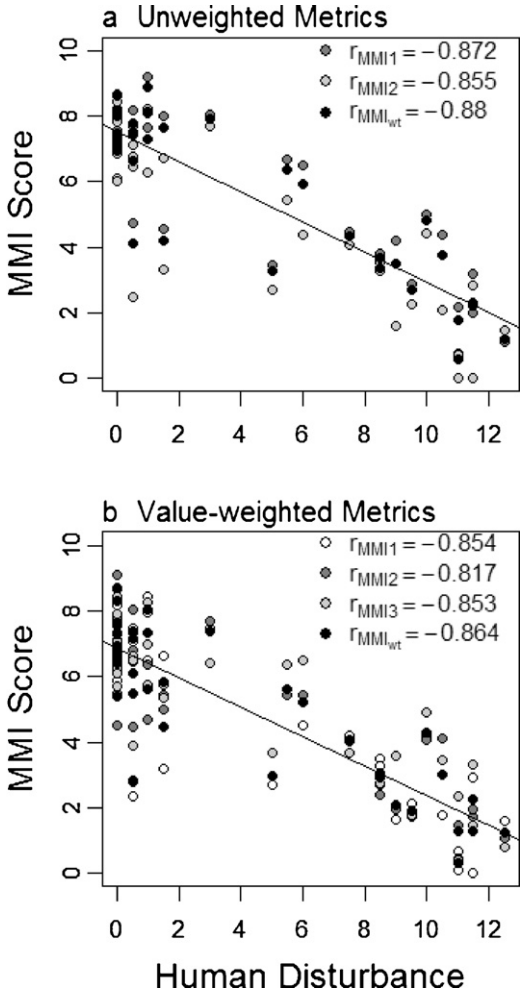

Fig. 8. Acadia National Park wetland MMI scores as a function of human disturbance for each final candidate MMI and model-weighted MMI for MMIs derived from (a) unweighted metric and (b) value-weighted metrics.

When applied to the unweighted collection of candidate metrics, the MMI assembly algorithm identified two MMIs with ΔAIC < 2 consisting of 4 unique metrics (Table 1). Both MMIs exhibit a strong correlation ($r = -0.872$, $r = -0.855$) with the human disturbance measure, as does the model-weighted MMI ($r = -0.880$) (Fig. 8a). We refer to the model-weighted MMI as the Acadia BioIndicator Index (BII). The BII is highly correlated, in this case, with both the biological metrics that comprise it (as expected), as well as other metrics that are of potential importance from conservation and management perspectives (Table 2). For example, the BII is highly positively correlated with the richness of native species ($r = 0.784$) and cover of *Sphagnum* ($r = 0.797$), while highly negatively correlated with average wetland indicator score ($r = -0.736$) and height of dominant vegetation ($r = -0.714$); suggesting that human disturbance has possibly caused increased water levels and productivity, the increases in *Typha* spp. cover and a reduction of species rich, *Sphagnum*-associated plant communities.

When applied to the value-weighted set of candidate metrics, the algorithm identified three candidate MMIs with ΔAIC < 2 consisting of 4 metrics (Table 3). The set of metrics comprising the final MMIs include both metrics that we biased toward inclusion by weighting. Each of the three MMIs are strongly correlated with the human disturbance measure ($r = -0.854$, $r = -0.817$, $r = 0.853$), as is the value-weighted, model-weighted MMI ($r = -0.864$) (Fig. 8b), which we will refer to as the value-weighted BII. Although the value-weighted BII is slightly less sensitive to human disturbance (as evidence by the difference in correlations) than the unweighted BII, it also is highly correlated with metrics that support the same interpretation of as the unweighted BII (Table 2). In Appendix B we compare these results with those produced by some existing approaches for both the Acadia wetland study and for riparian wetlands in Rocky Mountain National Park.

4. Discussion

The goal of the method of MMI assembly suggested here is to produce the most robust and sensitive MMI possible from the given data in a transparent and repeatable fashion. This method differs from many other approaches (Karr and Chu, 1997; Stoddard et al., 2008; Whittier et al., 2007) in several ways. At least part of the difference is attributable to a difference in goals and priorities. Methodologies that select metrics from conceptual categories introduce an information selection process that fails to allow the data to speak for themselves (specifically, by ignoring the multivariate character of the biological response). Such approaches often result in a product that exhibits some but not a full measure of the statistical benefits of indexing. What is accomplished is the creation of an index that partly reflects theoretical understanding and partly reflects the signals in the data (i.e. Fairwether, 1999).

Algorithmic index assembly is designed to maximize the statistical benefits of index construction to create a sensitive, robust indicator of human disturbance. The index that results from the algorithmic process may or may not comprise individual metrics that are of highest interest in terms of management, but will likely be correlated with such metrics. The utility of an index that is maximally sensitive to human disturbance is similar to the utility of having canaries in coal mines; they act as a sensitive indicator of other biological and ecological impacts that may be of greater interest than the metrics in the index. Thus, algorithm-based MMIs are more purely bioindicators.

In cases where there are a few candidate metrics that are much more strongly related to the measure of human disturbance than any of the others, the algorithmic method of index assembly is unlikely to result in much more sensitive index than that constructed by less computationally intense methods. This is because, if there are a few very strong candidate metrics, there will be very

Table 2
Correlations between each candidate metric and the human disturbance index (HDI) the model-weighted MMI derived from unweighted metrics (a BioIndicator Index, BII) and the model-weighted MMI derived from value-weighted metrics (val-wt. BII). Check marks indicate that the metric was included in the index.

Metric	Cor w/HDI	Cor w/BII	BII	Cor w/val-wt. BII	Val-wt. BII
Cover metrics					
Total	0.23	−0.17		−0.23	
Tree	−0.35	0.37		0.36	
Sphagnum	−0.66	0.79		0.78	
Annuals	−0.21	0.19		0.17	
Perennials	0.47	−0.43		−0.49	
Typha spp.	0.74	−0.81	√	−0.80	√
Dicots	0.55	−0.57		−0.48	√
Monocots	−0.23	0.23		0.07	
Ferns	0.36	−0.39		−0.37	
Forbs	0.62	−0.70	√	−0.66	
Shrubs	−0.04	0.14		0.19	
Introduced spp.	−0.01	0.08		0.14	
Native spp.	0.40	−0.41		−0.55	√
Richness metrics					
Total	−0.54	0.67		0.71	
Annuals	−0.31	0.48		0.45	
Perennials	−0.63	0.78	√	0.78	
Dicots	−0.47	0.62		0.63	
Monocots	−0.66	0.81		0.79	
Ferns	−0.35	0.29		0.26	
Forbs	−0.55	0.66	√	0.62	
Shrubs	−0.45	0.62		0.67	
Trees	−0.51	0.68		0.73	
Introduced spp.	−0.12	0.26		0.22	
Native spp.	−0.64	0.79		0.78	√
Comp. and struct. metrics					
Avg. Wetland Indicator Score	0.60	−0.75		−0.77	
Height of Dom. Vegetation	0.60	−0.71		−0.68	

little uncertainty about which candidate index is the best, as long as it includes the strongest ones. However, in cases where there are many candidate metrics which display similar strength in the relationship to human disturbance and complex relationships with one another, which is common for some types of systems, the algorithmic approach is likely to result in an index with a much more sensitive index than any produced by less computational approaches. This point is illustrated by a comparison of the indices developed for Acadia National Park wetland and the riparian wetlands of Rocky Mountain National Park in Appendix A. For both data sets, it is shown that the MMI constructed using algorithmic methods had a substantially stronger relationship to human disturbance than did the MMI developed using the alternative method.

The algorithmic approach to MMI assembly we advocate shares features with algorithmic statistical procedures such as stepwise regression in that metrics are added one at a time to avoid computational overload. Stepwise regression has been criticized based on grounds that it results in overfitting, leading to poor predictive capacity of the model when confronted with new data (e.g. Babyak, 2004; Whittingham et al., 2006). However, when confronted with new data in simulations the algorithm described here did not result in overfitting (Fig. 7). The correlation between the simulated MMI and HDI for new data sets was very similar to the correlation observed for the "calibration" data. The method described here avoids overfitting by using conservative model selection criteria (e.g., Likelihood ratio test) and information theoretic principals (e.g., AIC) designed to maximize predictive capacity and thus minimize overfitting.

As presented, our approach to MMI assembly selects or rejects metrics based on the linear fits of the models and assumes the residual errors are normally distributed. While this is a limitation of this method, we expect these assumptions will not severely affect the method's ability to produce robust, predictive MMIs. There are two reasons for this confidence. The first is that, while the assumption of linearity is often used to generate approximate statistical model, in this context it functions as a constraint. Since it is a criterion for the inclusion of a metric, the algorithm will be biased toward, but not guaranteed to, select metrics that are linearly related (or can be transformed to be linear) to the measure of human disturbance and thus bias the final MMI toward having a linear relationship human disturbance. Secondly, the assumption of normally distributed residual errors is less problematic in this application than in other statistical contexts. Since the process of constructing an index involves a summation of random variables (i.e., the metrics), the central limit theorem suggests that the distribution will approach a (approximately) normal distribution as metrics are included. This is aided by the tendency for the selection algorithm to select metrics with low inter-correlation of metric errors, which is a requirement of the theorem. Finally, if it is known that many of the metrics one may wish to include in the set of candidate metrics exhibit a non-linear function with human disturbance, a non-linear measure of association with human disturbance, such

Table 3
Final set of "best" MMIs selected by algorithm from value-weighted metrics for Acadia National Park wetland vegetation data.

Statistic	MMI1	MMI2	MMI3
	Log(*Typha* spp. cover) Rich. of native spp.	Cover of native spp. Rich. of native spp. Log(*Typha* spp. cover)	Dicot cover Log(*Typha* spp. cover) Rich. of native spp. Cover of native spp.
AIC wt.	0.422	0.408	0.17
Cor w/HDI	−0.854	−0.817	−0.853

as Efron's R^2 (Efron, 1978) from a LOESS regression, could replace the simple correlation in the algorithm.

5. Conclusions

As the use and profile of MMIs for bioassessment expands, it is important that they are constructed in such a way that they are as sensitive as possible to human disturbance, robust, and that the decisions made during their assembly are transparent and repeatable. The algorithmic approach to MMI assembly presented here meets those goals. In addition, the ability to incorporate information from expert opinion or biological theory in an explicit, repeatable way makes this approach sufficiently flexible for use across many purposes and many biological systems. Finally, while the method currently relies of relationships that are approximately linear-Gaussian, we believe that those details can be generalized to account of functional relationships of arbitrary complexity.

Acknowledgments

We thank Acadia National Park and especially David Manski for support in data collection. We also thank Ed Hall and two anonymous reviewers for constructive comments on an earlier draft of the manuscript. This work was supported, in part, by funding from the USGS Status and Trends, Ecosystems, and Global Change Programs. The use of trade names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

References

Anderson, D.R., 2008. Model Based Inference in the Life Sciences: A Primer on Evidence. Springer-Verlag New York Inc., New York.

Babyak, M.A., 2004. What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. Psychosom. Med. 66, 411–421.

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.

Burnham, K.P., Anderson, D.R., 2002. Model Selection and Inference: A Practical Information-Theoretic Approach. Springer-Verlag New York Inc., New York.

Cormen, T.H., Leiserson, C.E., Rivest, R.L., Stein, C., 2009. Introduction to Algorithms, 3rd ed. MIT Press and McGraw-Hill.

DeKeyser, E.S., Kirby, D.R., Ell, M.J., 2003. An index of plant community integrity: development of the methodology for assessing prairie wetland plant communities. Ecol. Indic. 3, 119–133.

Efron, B., 1978. Regression and ANOVA with zero-one data: measures of residual variation. J. Am. Stat. Assoc. 73, 113–121.

Ferreira, M.T., Rodriguez, P.M., Aguiar, F.C., Albuquerque, A., 2005. Assessing biotic integrity in Iberian rivers: development of a multimetric plant index. Ecol. Indic. 5, 137–149.

Fairwether, P.G., 1999. State of environment indicators of 'river health': exploring the metaphor. Freshwater Biol. 41, 211–220.

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

Grace, J.B., Schoolmaster Jr., D.R., Guntenspergen, G.R., Little, A.M., Mitchell, B.R., Miller, K.M., Schweiger, E.W., 2012. Guidelines for a graph-theoretic implementation of structural equation modeling. Ecosphere 3 (8), article 73, 44 pp.

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.

Hurvich, C.M., Tsai, C., 1989. Regression and time series model selection in small samples. Biometrika 76, 297–307.

Kahl, S., Manski, D., Flora, M., Houtman, N., 2000. Water Resources Management Plan. Acadia National Park, United States Department of Interior, National Park Service, Bar Harbor.

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

Karr, J.R., Chu, E.W., 1997. Biological Monitoring and Assessment: Using Multimetric Indexes Effectively. EPA 235-R97-001. University of Washington, Seattle.

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

Little, A.M., Guntenspergen, G.R., Allen, T.F.H., 2010. Conceptual hierarchical modeling to describe wetland plant community organization. Wetlands 30, 55–65.

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, Volume 1. Wetland Ecology Group, Division of Surface Water, Ohio Environmental Protection Agency, Columbus, OH.

Miller, S.J., Wardrop, D.H., Mahaney, W.M., Brooks, R.P., 2006. A plant-based index of biological integrity (IBI) for headwater wetlands in central Pennsylvania. Ecol. Indic. 6, 290–312.

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.

Ode, P.R., Hawkins, C.P., Mazor, R.D., 2008. Comparability of biological assessments derived from predictive models and multimetric indices of increasing geographic scope. J. N. Am. Benthol. Soc. 27, 967–985.

Ofenböck, T., Moog, O., Gerritsen, J., Barbour, M., 2004. A stressor specific multimetric approach for monitoring running waters in Austria using benthic macro-invertebrates. Hydrobiologia 516, 251–268.

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.

R Development Core Team, 2008. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0 http://www.R-project.org

Riseng, C.M., Wiley, M.J., Stevenson, R.J., Zorn, T.G., Seelbach, P.W., 2006. Comparison of Course versus fine scale sampling on statistical modeling of landscape effects and assessment of fish assemblages of the Muskegon River, Michigan. Am. Fish. Soc. Symp. 48, 555–575.

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.

Rothrock, P.E., Simon, T.P., Stewart, P.M., 2007. Development, calibration, and validation of a littoral zone plant index of biotic integrity (PIBI) for lacustrine wetlands. Ecol. Indic. 8, 79–88.

Schoolmaster Jr., D.R., Grace, J.B., Schweiger, E.W., 2012. A general theory of multimetric indices and their properties. Methods Ecol. Evol. 3, 773–781.

Stoddard, J.L., Larsen, D.P., 2006. Setting expectations for the ecological condition of streams – the concept of reference condition. Ecol. Appl. 16, 1267–1276.

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.

Tierney, G.L., Faber-Langendoen, D., Mitchell, B.R., Shriver, W.G., Gibbs, J.P., 2009. Monitoring and evaluating the ecological integrity of forest ecosystems. Front. Ecol. Environ. 7, 308–316.

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.

Whittingham, M.J., Stephens, P.A., Bradbury, R.B., Freckleton, R.P., 2006. Why do we still use stepwise modeling in ecology and behaviour. J. Anim. Ecol. 75, 1182–1189.