LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

INSTITUT FÜR STATISTIK

M. Schmid, T. Hothorn, K. O. Maloney,
D. E. Weller & S. Potapov

Geoadditive Regression Modeling of Stream Biological Condition

Technical Report Number 086, 2010
Department of Statistics
University of Munich

http://www.stat.uni-muenchen.de

Geoadditive Regression Modeling of Stream Biological Condition

Matthias Schmid[1] Torsten Hothorn[2] Kelly O. Maloney[3]
Donald E. Weller[3] Sergej Potapov[1]

Abstract

Indices of biotic integrity (IBI) have become an established tool to quantify the condition of small non-tidal streams and their watersheds. To investigate the effects of watershed characteristics on stream biological condition, we present a new technique for regressing IBIs on watershed-specific explanatory variables. Since IBIs are typically evaluated on an ordinal scale, our method is based on the proportional odds model for ordinal outcomes. To avoid overfitting, we do not use classical maximum likelihood estimation but a component-wise functional gradient boosting approach. Because component-wise gradient boosting has an intrinsic mechanism for variable selection and model choice, determinants of biotic integrity can be identified. In addition, the method offers a relatively simple way to account for spatial correlation in ecological data. An analysis of the Maryland Biological Streams Survey shows that nonlinear effects of predictor variables on stream condition can be quantified while, in addition, accurate predictions of biological condition at unsurveyed locations are obtained.

Keywords: Proportional odds model; Gradient boosting; Geoadditive regression; Stream biological condition; Maryland Biological Streams Survey

[1] Institut für Medizininformatik, Biometrie und Epidemiologie, Friedrich-Alexander-Universität Erlangen-Nürnberg, Waldstraße 6, 91054 Erlangen, Germany,
E-Mail: matthias.schmid@imbe.med.uni-erlangen.de
[2] Institut für Statistik, Ludwig-Maximilians-Universität München, Ludwigstraße 33, 80539 München, Germany
[3] Smithsonian Environmental Research Center, 647 Contees Wharf Rd., P.O. Box 28, Edgewater, Maryland 21037-0028 USA

1 Introduction

In view of the growing impact of humans on their natural environment, conserving and managing small streams and their watersheds have become important. Policy makers and land managers must assess the ecological effects of their decisions on streams, but also have to investigate the impacts of stream degradation on human health and the quality of life. For these reasons, a detailed understanding of the relationship between anthropogenic stressors and stream ecosystems is essential (Cushing and Allan 2001, USEPA 2006, Maloney et al. 2009). To develop that understanding, ecologists and statisticians need to quantify how watershed characteristics affect stream biological condition. Because small streams are numerous, assessing the biological condition of all streams in a landscape would be logistically impractical and cost prohibitive. It is therefore necessary to develop predictive models for site-specific stream condition using data from a limited number of sample sites. Ideally, those models would both quantify the effects of anthropogenic stressors on stream condition *and* accurately predict biological condition at unsurveyed locations (USEPA 2006). In recent years, a wide range of statistical tools to characterize and to model the condition of small streams have been developed (see, e.g., Barker et al. 2006, Collier 2009, Maloney et al. 2009, or Cooper 2009 for recent studies in this field).

The responses of streams to anthropogenic stress are often examined using biological metrics that describe biological conditions from structural and functional measures of the biological community (Karr 1991, Barbour et al. 1999). However, single metrics only measure a single feature of the community (e.g., number of taxa or diversity) and may not capture the effects of multiple stressors. Therefore, stream assessments usually compile several single metrics that are selected a priori to relate stream impairment to anthropogenic stress and then combine those metrics into a single multimetric index of biotic integrity ("IBI", Karr et al. 1986, Schleiger 2000, Southerland et al. 2005). These IBI indicators can then be statistically related to watershed-specific predictor variables using modeling approaches such as linear or ordinal regression, principal component analysis, or tree-based methods such as CART and random forests.

In this paper, we address the problem of developing predictive models for indices of biotic integrity for fish (FIBI) and benthic macroinvertebrates (BIBI). Both indices have become widely established tools for characterizing stream biological condition (Karr et al. 1986, Barbour et al. 1999, Southerland et al. 2005). When modeling FIBI and BIBI indicators, the following key issues need to be addressed:

1. IBI indicators are typically evaluated on an ordinal scale (e.g., using categories ranging from "poor condition" to "very good condition"). Although it is possible to use linear regression methods to model ordinal indicators, ordinal regression models such as the proportional odds model are a more appropriate choice (McCullagh 1980, Agresti 2002, Bigler et al. 2005). However, if maximum likelihood estimation is used to fit the model, and a large number of predictor variables are considered, proportional odds

models tend to overfit the data. Usually, this leads to a decrease in prediction accuracy. On the other hand, heuristic strategies to control the number of predictors are often biased and imprecise (Rawlings et al. 1998).

2. Stream condition is affected by many factors that are often highly correlated. Moreover, spatial correlation is usually evident in ecological data (Peterson and Urquhart 2006, Gelfand 2007). A statistical model must be able to identify the most important factors and to account for spatial correlation in the data. In addition, prediction models should be able to represent nonlinear relationships that often exist between predictors and indicators of stream biological condition.

3. It is well-known that maximizing prediction accuracy does not necessarily go hand in hand with finding a statistical model that is easy to interpret. Common examples are statistical learning techniques such as bagging or random forests (Breiman 2001, Cutler et al. 2007), which yield "black-box" predictions that are typically accurate but lack interpretability. This is not desirable in situations where effects of predictor variables need to be quantified.

The aim of this paper is to develop a statistical method for modeling IBI indicators that simultaneously addresses all issues outlined above. Following Agresti (2002), we use the proportional odds model framework to accomodate the ordinal structure of IBI indicators (issue 1). To avoid overfitting the data, however, we do not use classical maximum likelihood estimation to obtain model estimates but a *component-wise gradient boosting* approach (for an overview of boosting methods, see Bühlmann and Hothorn 2007). Because component-wise gradient boosting has a built-in mechanism for variable selection and shrinkage of estimates, the method can be used to obtain regularized fits of many types of statistical models. Consequently, heuristic techniques for variable selection and model choice are not needed.

In recent years, various authors have shown that gradient boosting can be modified such that prediction accuracy is optimized while, in addition, a meaningful interpretation of the model estimates is possible (Friedman et al. 2000, Bühlmann and Yu 2003, Bühlmann and Hothorn 2007, Kneib et al. 2009). Regarding issue 3, this is exactly what one wants to achieve: The structure and the interpretability of the proportional odds model is preserved while, in contrast to maximum likelihood estimation, prediction accuracy is maximized by fitting the model in a regularized way. Most notably, by using penalized regression splines to model effects of covariates, nonlinear relationships and spatial information can be easily incorporated into the prediction model. This is important with respect to issue 2, cf. Kneib et al. (2008, 2009).

While component-wise gradient boosting has become an established tool to fit continuous and binary data, it has not been possible so far to use boosting methods for fitting proportional odds models. The reason for this is that the boosting algorithms considered by Bühlmann and Hothorn (2007) do not allow for the estimation of scale parameters. The proportional odds model, however,

involves the constrained estimation of an ordered set of threshold parameters that have to be estimated simultaneously with the other model parameters. To take this problem into account, we construct a new boosting algorithm that combines the methods considered by Bühlmann and Hothorn (2007) with an estimation approach suggested by Schmid et al. (2010). With the latter approach, it is possible to adapt boosting algorithms to model families depending on a set of scale parameters. As will be shown, the method by Schmid et al. (2010) can be re-formulated such that fitting a proportional odds model via boosting techniques is feasible.

We will analyze IBI data from the Maryland Biological Streams Survey (MBSS) to demonstrate that the new algorithm is an efficient modeling tool for the biological assessment of small streams and their watersheds. Boosting predictions of FIBI and BIBI indicators are similar to predictions obtained from other established statistical techniques (see, e.g., Maloney et al. 2009), but spatial covariate patterns are detected and model estimates can be interpreted in a more meaningful way. This is possible because the structure of the proportional odds model allows for inspection and visualization of marginal predictor effects. As a consequence, the model can be used both for extrapolating estimates of stream biological condition to unsurveyed sites and exploring the determinants of biotic integrity.

The rest of the paper is organized as follows: In Section 2, the new algorithm is presented, along with a number of technical details involved in choosing appropriate tuning parameters. The characteristics of the algorithm are demonstrated in Section 3. Here, the new method is benchmarked against other regression techniques, and an analysis of the MBSS data is carried out. A summary and discussion of the main findings of the paper is given in Section 4. Additional figures and technical details are presented in the Appendix of the paper.

2 Methods

Proportional odds model

Let $\mathbf{Y}$ be an IBI outcome with K ordered categories and denote the vector of predictor variables by $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_p)$. Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a set of independent realizations of $(\boldsymbol{X}, \boldsymbol{Y})$. Define $X := (X_1, \ldots, X_n)$ and $Y := (Y_1, \ldots, Y_n)$.

The proportional odds model is given by

$$\mathrm{P}(\mathbf{Y} \leq k) = \frac{1}{1 + \exp(f(\mathbf{X}) - \theta_k)} \ , \ k = 1, \ldots, K, \tag{1}$$

where $f = f(\mathbf{X})$ is a prediction function depending on the predictor variables and

$$-\infty < \theta_1 < \ldots < \theta_{K-1} < \theta_K = \infty \tag{2}$$

is a set of threshold values that has to be estimated simultaneously with f. In many applications, f is restricted to being a linear function of the covariates

(see Agresti 2002). To take nonlinear predictor effects into account, we will use a more flexible approach: f will be modeled as the sum of (possibly nonlinear) marginal prediction functions $f_1(\mathbf{X}_1), \ldots, f_p(\mathbf{X}_p)$, i.e.,

$$f(\mathbf{X}) \equiv f(\mathbf{X}_1, \ldots, \mathbf{X}_p) = \sum_{j=1}^{p} f_j(\mathbf{X}_j) \tag{3}$$

(cf. Kneib et al. 2009). With this approach, the model has essentially the same structure as a generalized additive regression model (Hastie and Tibshirani 1990).

From (1) we obtain

$$\begin{aligned}
\mathrm{P}(\mathbf{Y} = 1|\mathbf{X}) &= \frac{1}{1+\exp(f-\theta_1)} \;, \\
\mathrm{P}(\mathbf{Y} = 2|\mathbf{X}) &= \frac{1}{1+\exp(f-\theta_2)} - \frac{1}{1+\exp(f-\theta_1)} \;, \\
&\vdots \\
\mathrm{P}(\mathbf{Y} = K-1|\mathbf{X}) &= \frac{1}{1+\exp(f-\theta_{K-1})} - \frac{1}{1+\exp(f-\theta_{K-2})} \;, \\
\mathrm{P}(\mathbf{Y} = K|\mathbf{X}) &= 1 - \frac{1}{1+\exp(f-\theta_{K-1})} \;,
\end{aligned} \tag{4}$$

which allows for specifying the log-likelihood of the proportional odds model (see Appendix A). By definition, the probability of observing a large outcome category increases with the magnitude of the estimates of f_j, $j = 1, \ldots, p$. For two sites with covariate vectors X_1 and X_2, (4) implies that the log ratio of cumulative odds does not depend on the category k under consideration:

$$\log\left(\frac{\mathrm{P}(\mathbf{Y} \leq k|X_1)/\mathrm{P}(\mathbf{Y} > k|X_1)}{\mathrm{P}(\mathbf{Y} \leq k|X_2)/\mathrm{P}(\mathbf{Y} > k|X_2)}\right) = f(X_2) - f(X_1) \;. \tag{5}$$

Equation (5) is the well-known "proportional odds assumption" which leads to estimates that are interpretable in terms of cumulative odds ratios.

Component-wise gradient boosting

As stated in the introduction, overfitting of the data can be avoided if component-wise boosting is used to fit the proportional odds model. In the following, we will adapt the component-wise gradient boosting algorithm considered by Bühlmann and Hothorn (2007) to the proportional odds model specified above.

In the boosting framework, the aim is to estimate the "optimal" prediction function f^* and the "optimal" set of threshold values $\theta^* := (\theta_1^*, \ldots, \theta_{K-1}^*)$ defined by

$$(f^*, \theta^*) := \underset{f,\theta}{\operatorname{argmin}}\, \mathbb{E}_{\boldsymbol{Y},\boldsymbol{X}}\left[\rho(\boldsymbol{Y}, f(\boldsymbol{X}), \theta)\right] \;, \tag{6}$$

where the loss function ρ is assumed to be differentiable with respect to f. In case of the proportional odds model, it is a natural choice to set ρ equal to the negative log-likelihood derived from (4). The full log-likelihood function and its derivative are given in Appendix A.

Instead of minimizing the theoretical mean given in (6), we consider the empirical risk $\mathcal{R} := \sum_{i=1}^{n} \rho(Y_i, f(X_i), \theta)$ and use the following new boosting algorithm to minimize the $\mathcal{R}$ over f and θ:

1. Initialize the n-dimensional vector $\hat{f}^{[0]}$ and the threshold parameter estimates $\hat{\theta}_1^{[0]}, \ldots, \hat{\theta}_{K-1}^{[0]}$ with offset values.
2. For each of the predictor variables specify a *base-learner*, i.e., a regression estimator with one input variable and one output variable. Set $m = 0$.
3. Increase m by 1.
4. (a) Compute the negative gradient $-\frac{\partial \rho}{\partial f}$ and evaluate at $\hat{f}^{[m-1]}(X_i)$, $\hat{\theta}^{[m-1]} = \left(\hat{\theta}_1^{[m-1]}, \ldots, \hat{\theta}_{K-1}^{[m-1]}\right)$, $i = 1, \ldots, n$. This yields the negative gradient vector

$$\begin{aligned} U^{[m]} &= \left(U_i^{[m]}\right)_{i=1,\ldots,n} \\ &:= \left(-\frac{\partial}{\partial f}\rho\left(Y_i, \hat{f}^{[m-1]}(X_i), \hat{\theta}^{[m-1]}\right)\right)_{i=1,\ldots,n} . \end{aligned}$$

   (b) Fit the negative gradient vector $U^{[m]}$ to each of the p predictor variables separately by using the base-learners specified in step 2. This yields p vectors of predicted values, where each vector is an estimate of the negative gradient vector $U^{[m]}$.

   (c) Select the base-learner that fits $U^{[m]}$ best according to the R^2 goodness-of-fit criterion. Set $\hat{U}^{[m]}$ equal to the fitted values of the best model.

   (d) Update $\hat{f}^{[m]} \leftarrow \hat{f}^{[m-1]} + \nu\, \hat{U}^{[m]}$, where $0 < \nu \leq 1$ is a real-valued step length factor.
5. Plug $\hat{f}^{[m]}$ into the empirical risk function $\sum_{i=1}^{n} \rho(Y_i, f, \theta)$ and minimize the empirical risk over θ. Set $\hat{\theta}^{[m]}$ equal to the newly obtained estimate of θ^*.
6. Iterate Steps 3 to 5 until the stopping iteration m_{stop} is reached (the choice of m_{stop} will be discussed below).

In the following, we will refer to the boosting algorithm introduced above as "proportional odds boosting" (P/O boosting). Steps 1 to 4 of the P/O boosting algorithm correspond to the classical gradient boosting approach discussed in Bühlmann and Hothorn (2007). From step 4 it is seen that the algorithm descends the gradient of the empirical risk $\mathcal{R}$, which is the main feature of all

gradient boosting algorithms. In each iteration, an estimate of the true negative gradient of $\mathcal{R}$ is added to the current estimate of f^*. Consequently, $\mathcal{R}$ is minimized in a stagewise fashion, and a structural (regression) relationship between $\mathbf{Y}$ and $\mathbf{X}$ is established. Obviously, using P/O boosting corresponds to replacing classical Fisher scoring algorithms for maximum likelihood estimation of f^* (McCullagh 1980) by a gradient descent algorithm in function space. As seen from steps 4(c) and 4(d), the P/O boosting algorithm additionally carries out variable selection, as only one base-learner (i.e., one component of $\mathbf{X}$) is selected for updating $\hat{f}^{[m]}$ in each iteration. Due to the additive update, the final boosting estimate at iteration m_{stop} can be interpreted as an additive prediction function, as defined in (3). In step 5, the estimation approach of Schmid et al. (2010) is used to obtain updates of θ. Here, the empirical risk is minimized over θ, using the current estimate of f^* as offset value. Note that step 5 of the P/O algorithm involves the *constrained* estimation of an ordered set of parameters, which has not been considered by Schmid et al. (2010). As shown in Appendix B, however, the constrained estimation problem can be re-formulated as an unconstrained problem, so that the method by Schmid et al. (2010) can be applied.

Specification of base-learners

It is clear from step 4 of the P/O boosting algorithm that the specification of the base-learners is crucial for interpreting the model fit. Here it is important to keep in mind that, due to the additive update in step 4(d), the estimate of a marginal function f_j at iteration m_{stop} has the same structure as the base-learner used in each iteration. For example, f_j is linear in $\mathbf{X}_j$ if the base-learner used to model f_j in step 4(b) is a simple linear model (cf. Bühlmann and Hothorn 2007, p. 484). Similarly, f_j is a smooth function of $\mathbf{X}_j$ if the corresponding base-learner is smooth as well.

Concerning the choice of appropriate base-learners, we follow the approach used by Kneib et al. (2009): The marginal functions f_j corresponding to *continuous predictors* are either modeled as linear functions or as penalized regression splines ("P-splines", cf. Wood 2006, Schmid and Hothorn 2008, Kneib et al. 2009), where selection of the best modeling alternative (smooth vs. linear) is carried out automatically by the P/O boosting algorithm. To do this, we modify step 2 of the P/O algorithm as follows: For each covariate, we specify two *competing* base-learners, namely a linear base-learner and a smooth P-spline deviation from the linear base-leaner (cf. Kneib et al. 2009, p. 628). Consequently, due to the base-learner selection carried out in step 4(c), the marginal functions f_j depending on continuous predictors become either linear or smooth.

To account for spatial dependency between neighboring sample sites, we additionally include a smooth function quantifying marginal *spatial effects* into the model. This function depends on the coordinates of the site locations and is added to the other functions specified in (3), see Kneib et al. (2008, 2009). As a base-learner for the marginal spatial effect we use a P-spline tensor product surface depending on the UTM easting and northing coordinates of the site

locations. Thus, denoting the coordinates by $\mathbf{X}_{\mathrm{E}}$ and $\mathbf{X}_{\mathrm{N}}$, the spatial effect becomes a smooth marginal surface $f_{\mathrm{sp}}(\mathbf{X}_{\mathrm{E}}, \mathbf{X}_{\mathrm{N}})$ depending on the bivariate "predictor" variable $(\mathbf{X}_{\mathrm{E}}, \mathbf{X}_{\mathrm{N}})$. As noted by Kneib et al. (2008), $f_{\mathrm{sp}}(\mathbf{X}_{\mathrm{E}}, \mathbf{X}_{\mathrm{N}})$ can be interpreted as the realization of a spatially correlated stochastic process, emphasizing the fact that one needs to account for spatial correlation in the data.

Finally, we model *categorical predictors* using dummy coded binary variables as base-learners. As a consequence, the functions f_j correspond to linear category effects in these cases. Detailed descriptions of P-splines and P-spline tensor product surfaces have been given by Fahrmeir et al. (2004), Wood (2006) and Kneib et al. (2009).

Tuning parameters

In the literature, it has been argued that boosting algorithms should generally not be run until convergence. Otherwise, overfits resulting in a suboptimal out-of-sample prediction accuracy are likely (see Bühlmann and Hothorn 2007). As a consequence of this "early stopping" strategy, the stopping iteration m_{stop} becomes the main tuning parameter of the P/O algorithm. In the following, we will use five-fold cross-validation to determine the value of m_{stop}, i.e., m_{stop} is the iteration with lowest predictive empirical risk. Alternatively, information criteria such as AIC or BIC could be used to determine the stopping iteration m_{stop}. For example, in case of Gaussian regression, a corrected AIC criterion could be calculated in each boosting iteration, and the stopping iteration would be given by the iteration with smallest AIC (Bühlmann and Hothorn 2007, p. 495). In this paper, however, we will consider cross-validation instead of information criteria because the latter have been criticized as being systematically biased towards stopping boosting algorithms too late (see Hastie 2007). In contrast to the choice of the optimal stopping iteration, the choice of the step length factor ν has been shown to be of minor importance for the predictive performance of a boosting algorithm. The only requirement is that the value of ν is "small", such that a stagewise adaption of the prediction function is possible (see Schmid and Hothorn 2008). We will set $\nu = 0.1$.

Regularization

A major consequence of the early stopping strategy is that the estimates of f^* are shrunken towards zero. In fact, using a small step length ν ensures that marginal function estimates increase "slowly" in the course of the P/O boosting algorithm but stop increasing as soon as the optimal stopping iteration m_{stop} is reached. As stated above, the optimal stopping iteration is chosen such that out-of-sample prediction accuracy is optimized within the proportional odds framework. In other words, stopping the P/O boosting algorithm at the optimal iteration implies that *the amount of shrinkage is chosen such that the predictive power of the proportional odds model is maximal.* Shrinkage is a well-established strategy to regularize model estimates: Estimates tend to have a slightly increased bias but a decreased variance, thereby improving prediction accuracy. On the other hand, the choice of the base-learners specified above en-

sures that black-box predictions are avoided and marginal effect estimates are obtained. Although, in contrast to maximum likelihood estimation, estimates are shrunken towards zero, the main characteristics (and thus the interpretability) of the proportional odds model are preserved.

Prediction

For given estimates of f^* and θ^*, the predicted outcome category (denoted by k^*) is the category with highest posterior probability, i.e.,

$$k^* = \max_k \widehat{\mathrm{P}}(\mathbf{Y} = k|\mathbf{X}) \ , \tag{7}$$

which can be computed from (4). Thus, misclassification rates and weighted kappa indices for ordinal data (Fleiss and Cohen 1973) can be used to evaluate the predictive power of the P/O boosting fit.

Confidence intervals

Since boosting estimates are shrunken towards zero, computation of confidence intervals for marginal functions is infeasible. This problem can also be found with other shrinkage methods such as ridge regression or the Lasso (Tibshirani 1996). With the help of bootstrap methods, however, it is possible to approximate the distribution of the boosting estimates in a non-parametric fashion (see Section 3). Consequently, the bootstrapped estimates can be used to assess whether a function estimate is systematically different from zero.

As an alternative to approximating the distribution of effect estimates via bootstrap sampling, Bayesian methods could be used to fit the proportional odds model and to calculate posterior distributions of marginal predictor effects. This approach would require Bayesian methods for shrinkage (such as the Bayesian Lasso, Park and Casella 2008) and variable selection (O'Hara and Sillanpää 2009) to be adapted to geoadditive proportional odds models. While being potentially useful, the Bayesian approach is conceptually different from the proposed P/O boosting algorithm and will therefore not be considered in this paper.

3 Analysis of the Maryland Biological Streams Survey

Data sources and pre-processing

In this section, we apply the P/O boosting algorithm to develop a predictive model for fish (FIBI) and benthic macroinvertebrates (BIBI) indicators of biological condition. Our study is focused on the $23,408$ km^2 part of Maryland, USA, lying in the Chesapeake Bay watershed (Figure 1). This area includes six Level III ecoregions: Central Appalachians, Ridge and Valley, Blue Ridge, Northern Piedmont, Southeastern Plains, and Middle Atlantic Coastal Plains (see Omernik 1987). Climate types range from cold with hot summers in the

mountainous western area to temperate with hot summers toward the southeast; vegetation patterns range from northern hardwood forests in the highlands to oak, hickory, pine, and southern mixed forests of the Coastal Plains. The Appalachian, Ridge and Valley, and Blue Ridge ecoregions are underlain largely by folded and faulted sedimentary rocks. The Piedmont ecoregion is underlain by crystalline igneous and metamorphic rocks, and the Plains ecoregions are underlain by unconsolidated sediments.

Our analysis is based on the Maryland Biological Streams Survey (MBSS), which is an on-going monitoring program designed to describe water quality in 1st- to 4th-order non-tidal streams within the state of Maryland, USA (USEPA 1999). MBSS scientists used a probabilistic sampling design stratified by major watershed and stream order to sample approximately 2500 sites from 1994 to 2004 (cf. Southerland et al. 2005). An MBSS site is a $\sim$ 75 m stream segment where data were collected for stream physical and hydrological attributes (e.g., flow, width, depth, and embeddedness), streamwater chemistry, location (latitude and longitude), riparian conditions, and biological communities (i.e., benthic macroinvertebrates and fish). For a detailed description of the MBSS, see http://www.dnr.state.md.us/streams. We considered only the first record for sites that were sampled more than once. This resulted in a database containing measurements at $n = 1573$ stream sites (see Figure 1). There were 96 sites that had no fish collected, and these sites were not used to examine FIBI, leaving $n = 1477$ sites for the FIBI models. Land cover data was obtained from the 2001 US National Land Cover Database (Homer et al. 2004). Watershed predictors were calculated in ARCGIS using watershed boundaries and relevant environmental coverages.

Individual IBIs were developed by MBSS scientists separately for each subregion of the study area and included individual metrics specific to each subregion (Appendix C, see Southerland et al. 2005 for IBI developments and for a complete list of metrics in each IBI). Following Maloney et al. (2009), we used an ordinal scale to quantify BIBI and FIBI indicators (1 = "very poor site", 2 = "poor site", 3 = "fair site", 4 = "good site"). FIBI and BIBI indicators were regressed on site-specific predictor variables using the P/O boosting algorithm introduced in Section 2. Predictors included UTM easting and northing coordinates, watershed land use, dominant ecoregion (Omernik 1987), and the "distance to mainstem" measured from the MBSS sampling site to a mainstem tributary with > 500 km^2 in upslope drainage area. A value of 0 was assigned to sites that drained into the Chesapeake Bay before reaching a mainstem river. For a detailed description of the predictor variables, see Appendix D. Predictors with a highly right-skewed distribution were log transformed before statistical analysis.

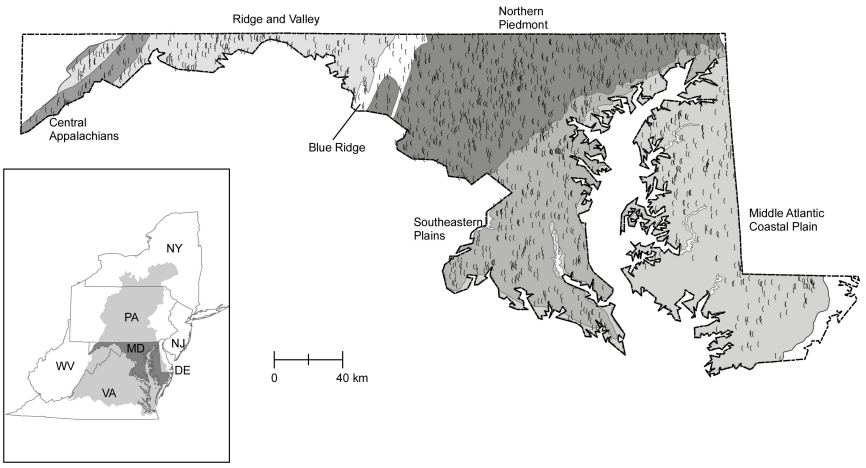

Figure 1: The Maryland portion of the Chesapeake Bay watershed, its major ecoregions, and stream assessment sites where data were collected. Inset shows the study area (dark gray) in relation to the Chesapeake Bay watershed (light gray).

Benchmark analysis

We first investigated the prediction accuracy of the P/O boosting algorithm, i.e., the ability of P/O boosting to predict the FIBI and BIBI values at unsurveyed sites. To do this, we carried out a benchmark experiment using 100 bootstrap samples drawn from the full data set. The bootstrap samples were used as training data sets, and the P/O boosting algorithm was applied to these samples. Five-fold cross-validation was carried out on the training data sets to determine the values of m_{stop}. In a next step, we applied the 100 prediction rules obtained from P/O boosting to the respective sets of out-of-bootstrap observations ("bootstrap cross-validation", see, e.g., Hothorn et al. 2005). The predictions and the true outcome values of the 100 out-of-bootstrap data sets were used to compute classification rates and weighted kappa values.

To benchmark the P/O boosting algorithm against other established techniques, we also considered the random forest method (Breiman 2001), a nonparametric classification technique based on ensembles of decision trees. Random forests have been shown to be one of the most accurate methods for predicting IBI indicators (Maloney et al. 2009). Because the random forest method neither requires the proportional odds assumption nor the additivity of the prediction function specified in (3), it is less restrictive than P/O boosting. As above, we estimated prediction accuracy using the same bootstrap samples to compute classification rates and weighted kappa values.

Table 1 shows that boosting and the random forest method had similar classification rates for both indices of biotic integrity. For the FIBI indica-

tor, mean classification rates were nearly equal (P/O boosting: 51.4%, random forest method: 51.2%) while in case of the BIBI indicator, classification rates were slightly lower for P/O boosting (45.4%) than for random forests (46.6%). Weighted kappa values obtained from P/O boosting were larger on average than the corresponding weighted kappa values obtained from random forests (Table 2). This result can be explained by the fact that P/O boosting explicitly accounts for the ordinal structure of the BIBI and FIBI indicators: Due to the structure of the proportional odds model, misclassification of observations into neighboring categories tends to be more likely than misclassification into categories that are "far away" from the true category. Note that *all* weighted kappa values obtained from P/O boosting were larger than 0.5, i.e., P/O boosting predicted significantly better than chance alone. Table 1 also suggests that classification rates of outcome categories are considerably higher if site-specific covariate information is used to obtain predictions than if the unconditional distribution of the FIBI and BIBI indicators is used. As seen from Table 1, classification rates are largest for sites with good biological condition. This result has previously been reported by Maloney et al. (2009).

Analysis of the full data set: FIBI indicator

After demonstrating that the prediction accuracy of P/O boosting is comparable to that of the random forest method, we analyzed the full data set to examine functional relationships between predictors and IBI indicators. We applied the P/O boosting algorithm to the full data set and visualized marginal function estimates using partial plots (Figures 2 to 5). Note that partial plots cannot be obtained from the random forest method because, in contrast to P/O boosting, random forests yield black-box estimates that are not easily interpreted. Although the random forest method can provide estimates of variable importance (Cutler et al. 2007), functional relationships between predictors and outcome variables cannot be quantified directly. We therefore did not use this method to analyze the full data set.

Let us first consider the FIBI indicator of biological condition. Figure 2 shows the most pronounced marginal effects obtained from P/O boosting. Light grey lines correspond to the function estimates obtained from the 100 bootstrap samples used in the benchmark experiment. Increases in watershed area and average watershed elevation have positive effects on the FIBI indicator, supporting previous findings of the importance of these factors on stream fishes (Angermeier and Schlosser 1989, Oberdorff and Hughes 1992, Matthews and Robison 1998, Joy and Death 2004). While the effect of watershed area is clearly nonlinear, a linear marginal prediction function was obtained for average watershed elevation. This demonstrates the ability of P/O boosting to incorporate both linear and smooth (nonlinear) predictor effects into the proportional odds model. Regarding the magnitude of its marginal function, watershed area is clearly the most important predictor for FIBI (Figure 2). For example, consider two stream sites with watershed areas $\mathrm{WA}_1 = 1$ km^2 and $\mathrm{WA}_2 = 10$ km^2. We obtain $\log(\mathrm{WA}_1) = 0$ and $\log(\mathrm{WA}_2) \approx 2.3$, which results in marginal function estimates $f_{\mathrm{WA}}(\mathrm{WA}_1) \approx -2$ and $f_{\mathrm{WA}}(\mathrm{WA}_1) \approx 1$ (see Figure 2). Assuming

Table 1: P/O boosting and random forest classification rates, as obtained from bootstrap cross-validation (uncond. = unconditional distribution of categories in the full data set).

FIBI classification rates, P/O boosting

	Mean	Min.	1st Qu.	Median	3rd Qu.	Max.	uncond.
all sites	0.514	0.461	0.500	0.514	0.528	0.563	
very poor sites	0.372	0.223	0.317	0.357	0.403	0.610	0.167
poor sites	0.221	0.056	0.183	0.219	0.259	0.361	0.172
fair sites	0.453	0.320	0.416	0.455	0.485	0.557	0.277
good sites	0.728	0.634	0.699	0.731	0.759	0.809	0.383

FIBI classification rates, random forests

	Mean	Min.	1st Qu.	Median	3rd Qu.	Max.	uncond.
all sites	0.512	0.466	0.500	0.513	0.526	0.561	
very poor sites	0.395	0.279	0.341	0.391	0.441	0.596	0.167
poor sites	0.278	0.152	0.241	0.278	0.309	0.405	0.172
fair sites	0.425	0.304	0.394	0.428	0.457	0.529	0.277
good sites	0.711	0.609	0.686	0.708	0.739	0.781	0.383

BIBI classification rates, P/O boosting

	Mean	Min.	1st Qu.	Median	3rd Qu.	Max.	uncond.
all sites	0.454	0.414	0.442	0.454	0.468	0.500	
very poor sites	0.435	0.285	0.399	0.435	0.473	0.568	0.181
poor sites	0.335	0.208	0.297	0.333	0.364	0.503	0.249
fair sites	0.483	0.329	0.453	0.482	0.516	0.590	0.295
good sites	0.546	0.440	0.508	0.552	0.578	0.653	0.275

BIBI classification rates, random forests

	Mean	Min.	1st Qu.	Median	3rd Qu.	Max.	uncond.
all sites	0.466	0.424	0.455	0.467	0.478	0.504	
very poor sites	0.474	0.350	0.443	0.480	0.513	0.579	0.181
poor sites	0.318	0.212	0.283	0.313	0.349	0.441	0.249
fair sites	0.436	0.279	0.408	0.431	0.466	0.519	0.295
good sites	0.629	0.532	0.601	0.632	0.655	0.736	0.275

Table 2: Weighted kappa values, as obtained from bootstrap cross-validation. Fleiss-Cohen weights were used to account for the ordinal structure of FIBI and BIBI indicators (cf. Fleiss and Cohen 1973).

FIBI weighted kappa values, P/O boosting

Mean	Min.	1st Qu.	Median	3rd Qu.	Max.
0.591	0.502	0.577	0.593	0.609	0.657

FIBI weighted kappa values, random forests

Mean	Min.	1st Qu.	Median	3rd Qu.	Max.
0.553	0.473	0.533	0.552	0.576	0.637

BIBI weighted kappa values, P/O boosting

Mean	Min.	1st Qu.	Median	3rd Qu.	Max.
0.586	0.527	0.568	0.583	0.603	0.643

BIBI weighted kappa values, random forests

Mean	Min.	1st Qu.	Median	3rd Qu.	Max.
0.580	0.501	0.565	0.580	0.596	0.634

constant values for the other predictor variables, it follows from equation (5) that the cumulative odds ratio of site 2 is approximately $\exp(1-(-2)) \approx 20$ times larger than the cumulative odds ratio of site 1. The strong positive effect of watershed area might be due to low natural richness of fishes in headwater streams (Angermeier and Schlosser 1989, Matthews and Robison 1998), which may affect individual metrics composing the IBI (Schleiger 2000).

Increases in the percentage of upstream watershed under impervious surface cover have a negative effect on FIBI scores, which is an often-reported pattern (e.g., Wang and Lyons 2003, Helms et al. 2009) that results from the numerous negative effects that impervious surfaces have on stream hydrologic and geomorphic factors (Paul and Meyer 2001, Walsh et al. 2005). The marginal function estimate for the distance from sampling location to the nearest main stem stream indicates that there is an inverted U-shaped relationship with the FIBI indicator, i.e., FIBI increases with low but increasing distance values but decreases again for large distance values. Sites with large distances from mainstems are likely headwaters having low FIBIs as discussed above. The lower FIBI scores for short distances to mainstem might reflect sites that directly drain into the Chesapeake Bay (which were given a distance value of 0). Marginal function estimates corresponding to other continuous predictors are relatively small in magnitude (relative to the estimates shown in Figure 2), indicating their minor importance for modeling FIBI. The function estimates corresponding to these predictor variables are shown in Appendix E.

Marginal effect estimates of the categorical covariate "percentage of bedrock that is calcareous in a watershed" were small and suggest that FIBI is lower when calcareous rock is present (percentage of calcareous bedrock $> 0\%$, see Table 3),

highlighting the importance of geology in structuring fish assemblages (Montgomery 1999, Joy and Death 2004). However, we caution over-interpretation of these findings because the range of the bootstrapped marginal effect estimates contains zero and because this covariate was analyzed at a coarse scale (presence/absence). The categorical effects of dominant ecoregions are also relatively small (Table 3), indicating that the dominant ecoregion is of minor importance for modeling FIBI.

A marginal spatial effect was still evident for the FIBI after accounting for all other covariates (Figure 3). Sites in the Blue Ridge region, the Ridge and Valley region, and the South-Eastern part of the Middle Atlantic Coastal plain tended to have lower FIBI scores than other sites. In contrast, the middle region of the Northern Piedmont region shows a very positive marginal effect on FIBI. It is important to remember that these effects are marginal and therefore not caused by variations in other predictors (such as the dominant ecoregion). They may reflect missing predictors, or, alternatively, problems with the calculation of FIBI itself. For example, the FIBI for Blue Ridge and Ridge and Valley ecoregions was stratified only into warmwater or coldwater streams (see Appendix C, Southerland et al. 2005). A more refined FIBI, possibly stratified by ecoregions or sub-ecoregions (Schleiger 2000), might reduce the marginal spatial patterns in the FIBI.

Analysis of the full data set: BIBI indicator

We next consider the BIBI indicator of biological condition. Figure 4 shows the most pronounced marginal effects obtained from P/O boosting. Obviously, increases in watershed area, distance from sampling location to the nearest main stem stream, and percentage of upstream watershed under tree cover have large positive effects on the BIBI indicator. In contrast, increases in the percentage of upstream watershed under impervious surface cover have a very strong negative effect on BIBI. Apart from a few sites with a small upstream population density (showing large variations in their effects on BIBI), the effect of population density on BIBI is also negative. The positive effect of the percentage of upstream watershed under tree cover and the negative effect of the percentage of upstream watershed under impervious surface cover on benthic macroinvertebrates support previous reports (Roy et al. 2003, Walsh et al. 2005, King et al. 2005, Maloney et al. 2009) and document the sensitivity of benthic macroinvertebrates to watershed conditions. The upstream population density was positively correlated with the percentage of upstream watershed under impervious surface cover ($r = 0.78$), so these covariates are likely to represent the same anthropogenic stressor (population pressure). The effect of distance to main stem demonstrates the importance of position within a stream network to the benthic community (Vannote et al. 1980). All functions shown in Figure 4 are nonlinear, demonstrating the ability of P/O boosting to account for nonlinear predictor effects. Marginal function estimates for other continuous predictors (Appendix F) are smaller in magnitude than those shown in Figure 4, indicating their minor importance for modeling BIBI. Marginal effect estimates of the categorical covariate "percentage of bedrock that is calcareous

Table 3: Effect of the categorical predictors "percentage of bedrock that is calcareous in a watershed" and "dominant ecoregion" on FIBI and BIBI indicators of stream condition. Values in columns 4 to 6 were obtained from applying P/O boosting to 100 bootstrap samples drawn from the full data.

FIBI indicator, P/O boosting

Predictor	Category	Full data set	Mean	Min.	Max.
% of calc. bedrock	= 0%	0			
	> 0%	-0.082	-0.144	-0.474	0.102
Ecoregion	Blue Ridge	0			
	Centr. Appalachian	-0.023	-0.188	-0.748	-0.001
	Mid. Atl. Coastal Plains	0.005	0.032	-0.004	0.141
	Northern Piedmont	-0.017	-0.082	-0.259	0.031
	Ridge & Valley	0.007	0.083	-0.017	0.337
	South Eastern Plains	0.004	0.019	-0.055	0.165

BIBI indicator, P/O boosting

Predictor		Full data set	Mean	Min.	Max.
% of calc. bedrock	= 0%	0			
	> 0%	-0.294	-0.308	-0.760	-0.013
Ecoregion	Blue Ridge	0			
	Centr. Appalachian	-0.605	-0.545	-1.017	-0.078
	Mid. Atl. Coastal Plains	-0.099	-0.086	-0.282	0.069
	Northern Piedmont	-0.238	-0.220	-0.459	0.031
	Ridge & Valley	0.511	0.450	0.041	0.878
	South Eastern Plains	0.266	0.254	0.031	0.574

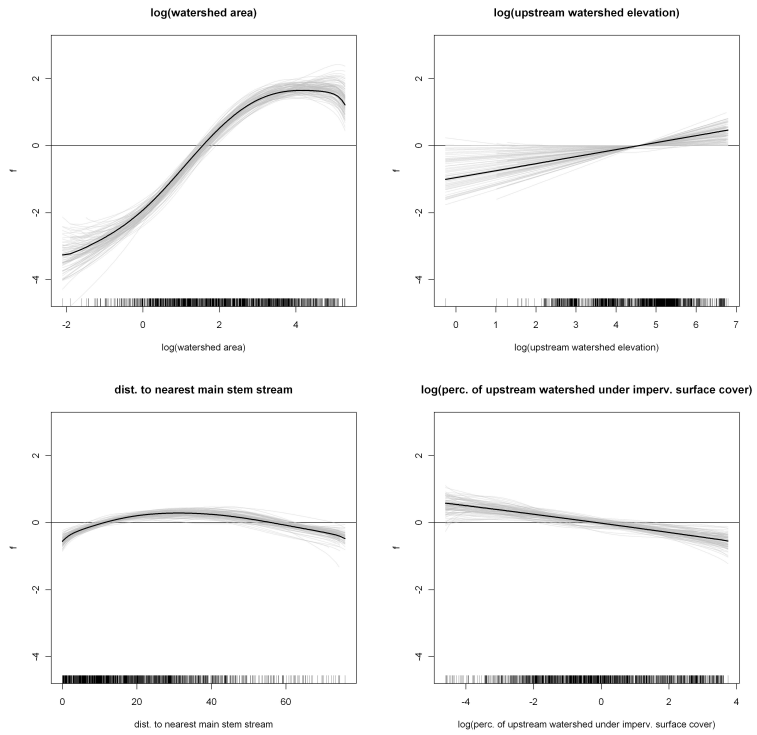

Figure 2: FIBI model - marginal function estimates obtained from applying P/O boosting to the full data set.

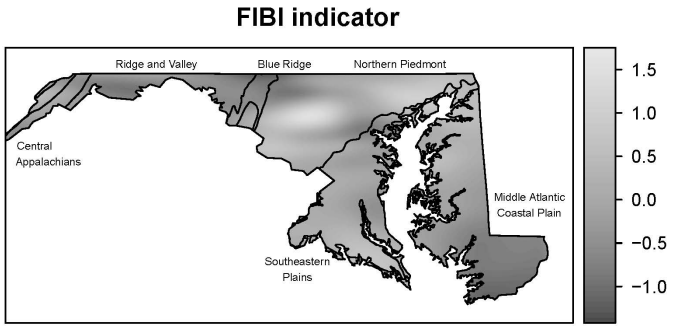

Figure 3: FIBI model - marginal spatial effect estimate obtained from applying P/O boosting to the full data set.

in a watershed" suggest that BIBI is lower when calcareous rock is present. This effect is much stronger than the effect obtained for the FIBI indicator (Table 3), reinforcing geology as an important structuring factor on local benthic macroinvertebrate assemblages (Montgomery 1999, Pyne et al. 2007). Again we caution over-interpretation of these results because of the coarse scale of this predictor. The categorical effects of three dominant ecoregions on BIBI are significantly different from zero (Table 3). The Central Appalachian ecoregion has the lowest marginal biotic integrity while the Ridge and Valley ecoregion has the largest positive effect on the BIBI indicator. Maloney et al. (2009) reported similar effects of ecoregions on BIBI.

The marginal spatial effect estimates show that sites in the Blue Ridge region and the eastern part of the Ridge and Valley region tended to have lower BIBI scores than other sites (Figure 5). These spatial effects may be due to missing predictors or coarse stratification during BIBI development. For example, the Ridge and Valley, Blue Ridge, and Central Appalachians ecoregions were lumped into a single "Combined Highlands" stratum during the BIBI construction (See Appendix C, Southerland et al. 2005).

4 Summary and conclusion

In recent years, much research has been undertaken to assess the degree of impairment in ecosystem structure and function due to anthropogenic disturbance in watersheds. As part of this research, biological assessments of stream condition have become an important tool to identify impairments and to develop appropriate management and conservation strategies. In this paper, we have considered indicators of biologic condition (Karr et al. 1986, Karr 1991, Bar-

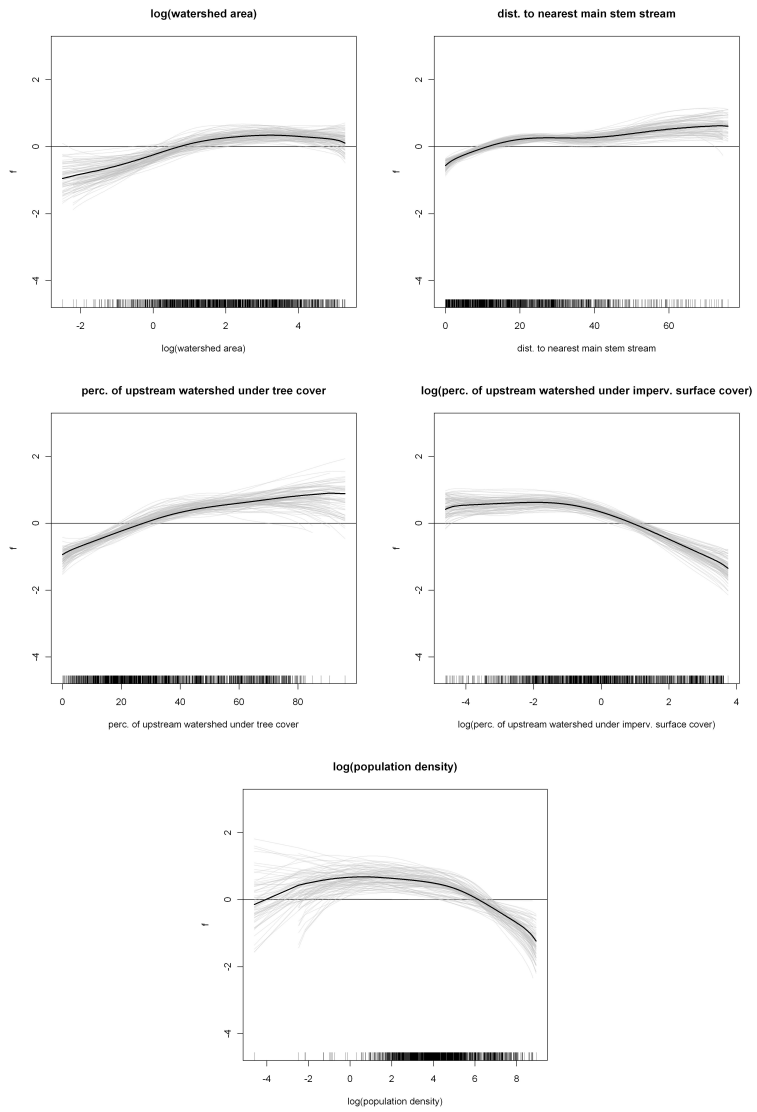

Figure 4: BIBI model - marginal function estimates obtained from applying P/O boosting to the full data set.

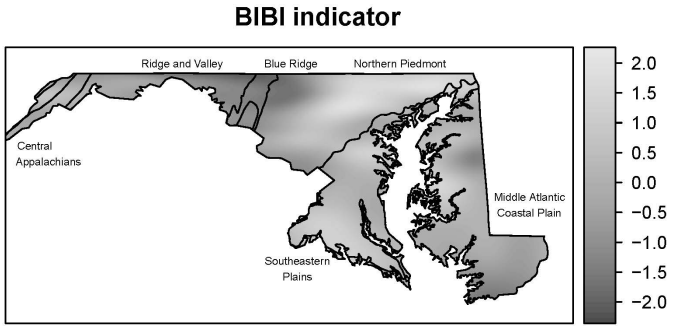

Figure 5: BIBI model - marginal spatial effect estimate obtained from applying P/O boosting to the full data set.

bour et al. 1999), which are indispensable tools for measuring and managing the health of streams and their watersheds.

We have developed a boosting algorithm for modeling indices of biotic integrity and applied our method to data collected from small non-tidal streams in the state of Maryland, USA. Because IBI indicators are often evaluated on an ordinal scale, the P/O boosting algorithm developed in this paper is based on the well-established proportional odds model introduced by McCullagh (1980). To obtain regularized model fits, we combined classical gradient boosting techniques with two recent advances: We used the modeling approach suggested by Kneib et al. (2009) to obtain prediction models accounting for nonlinear effects and spatial correlation, and we re-formulated the boosting method by Schmid et al. (2010) to obtain scale parameter estimates (which are necessary for adapting classical boosting techniques to the proportional odds model).

In summary, the boosting algorithm presented in this paper combines the following advantages:

1. P/O boosting allows for fully automatic variable selection and model choice. In particular, it does not require scientists to select predictor variables using heuristic approaches such as stepwise variable selection.

2. Although boosting estimates are typically different from classical maximum likelihood estimates, the P/O boosting algorithm preserves the structure of the proportional odds model. Therefore, boosting estimates are accessible for interpretation, which is a major advantage over estimation techniques that result in black-box estimates. On the other hand, of course, black-box predictions are expected to have a higher degree of flexibility than predictions that are linked to the pre-specified structure of the additive proportional odds model.

3. P/O boosting accounts for spatial effects. Although there are numerous methods to model spatial correlation in ecological data (cf. Bigler et al. 2005, Gelfand 2007), the interaction surfaces used in this paper have the advantage that marginal spatial effects can be visualized and information contained in unobserved (latent) predictor variables is quantified.

4. Spatial effects and nonlinear effects of predictor variables are estimated jointly based on penalized spline functions. Both the joint estimation and the additive structure of the prediction function facilitate a relatively simple interpretation of the results. For example, the use of P/O boosting led to clear interpretations of the relationships between watershed attributes and stream biological condition in the MBSS data, where both linear and nonlinear marginal predictor effects were present.

5. Due to the early stopping strategy, prediction accuracy of the P/O boosting fit is maximized in the proportional odds model framework. Of course, it is well known that there is no "uniformly best" prediction method for all types of data. Therefore, we do not claim that P/O boosting is generally superior to other methods for ordinal data. For the MBSS data, however, predictions obtained from P/O boosting turned out to be very similar to predictions obtained from the random forest method. This is remarkable because the random forest method is a completely non-parametric technique that is generally considered to be one of the most powerful statistical prediction methods (see Hastie et al. 2009, Chapter 15).

Although, for the sake of interpretation, we restricted ourselves to considering main-effects models in this paper, the gradient boosting framework can be extended to include interaction terms between predictor variables in the model formula. As demonstrated by Kneib et al. (2009), this can be accomplished by specifying additional sets of linear and smooth base-learners depending on the products of predictor variables.

When interpreting marginal function estimates, however, one should be aware of the fact that P/O boosting is based on two important assumptions: First, we assumed the proportional odds property to hold. Second, we assumed predictor effects to be additive. If these assumptions are not met, estimates might show a bias caused by model misspecification. Usually, this bias cannot be fully compensated by the high flexibility of the P/O boosting algorithm. Because the early stopping strategy results in regularized boosting estimates that are shrunken towards zero, it is generally difficult to derive tests on the appropriateness of model assumptions in the boosting framework. Therefore, assessing the robustness of boosting estimates against model misspecification constitutes an important issue of future research.

Instead of regularizing effect estimates via gradient boosting (in combination with early stopping), it would alternatively be possible to optimize out-of-sample prediction accuracy using penalized regression techniques. For example, the Lasso method (Tibshirani 1996), which is based on L_1-penalized likelihood estimation, would be a natural approach to incorporate shrinkage and variable selection into a proportional odds model. However, the original Lasso method has mainly been designed for regression models with a *linear* prediction function. Combining penalized estimation with sparse *nonlinear* additive modeling has only recently been accomplished (Meier et al. 2009). To date, there is no extension of the method developed by Meier et al. (2009) to geoadditive proportional odds models. On the other hand, gradient boosting and the Lasso are closely related, as both algorithms can be embedded into the LARS framework (Efron et al. 2004). Also, in case of Gaussian regression, there is evidence that the properties of gradient boosting are similar to those of the Lasso (Hastie et al. 2009, Chapter 16). These results suggest that the role of the boosting stopping iteration is similar to the role of the (inverse of the) shrinkage parameter used for the L_1 penalty of the Lasso method.

Finally, the boosting algorithm presented in this paper is not restricted to ecological applications but can be used to analyze very general types of ordinal data. Although developing a prediction method for stream biological condition was the primary goal of this paper, both the proportional odds model and the boosting framework are essentially independent of the application context in which they are used. It is therefore possible to apply the P/O boosting algorithm in many fields, for example in clinical or biomedical research.

Software

All computations were carried out with the R System for Statistical Computing (version 2.10.1, R Development Core Team 2009). The `gamboost` function of R package **mboost** (Hothorn et al. 2010) was used to calculate boosting estimates. Base-learners were made comparable by centering predictors at the beginning of the algorithm and by using the same degrees of freedom for each base-learner (see Kneib et al. 2009 for details). For example, in case of the FIBI model, the R code for specifying the model formula was given by

```
> library(mboost)
> FIBI.formula <- FIBI ~ bols(EASTING, intercept = FALSE) +
+                        bols(NORTHING, intercept = FALSE) +
+                        bspatial(EASTING, NORTHING, knots = 20, df = 1,
+                        differences = 1) +
+                        bols(DrainageDensity, intercept = FALSE) +
+                        bbs(DrainageDensity, center = TRUE, df = 1) +
                         ...
+                        bols(PerWet, intercept = FALSE) +
+                        bbs(PerWet, center = TRUE, df = 1) +
+                        bols(Ecoregion, intercept = FALSE, df = 1) +
+                        bols(INT, intercept = FALSE, df = 1)
```

where `FIBI` denotes the FIBI outcome, `EASTING` and `NORTHING` denote the UTM easting and northing coordinates of the site locations, respectively, and `DrainageDensity`, `PerWet` and `Ecoregion` are examples of predictor variables. The `bols()` and `bbs()` functions in R package **mboost** (using the `intercept = FALSE` and `center = TRUE` options) correspond to linear base-learners and smooth P-spline deviations from the linear base-learners, respectively (see Kneib et al. 2009 for details). Specifying the base-learners as shown above ensures that selection of the best modeling alternative (smooth vs. linear) is carried out automatically by the P/O boosting algorithm. Similarly, the `bspatial()` function in R package **mboost** (using the `center = TRUE` option) corresponds to a smooth P-spline tensor product deviation from a spatial linear surface. Note that specifying a `bbs()` base-learner for `Ecoregion` was not necessary because this predictor variable is categorical.

Using the model formula specified above, the proportional odds model was fitted with the help of the `PropOdds()` family in R package **mboost**. The corresponding R code was given by

```
> ctrl <- boost_control(mstop = 20000, nu = 0.1)
> FIBI.model <- gamboost(FIBI.formula, data = MBSS.training, family =
+                        PropOdds(), control = ctrl)
```

where `MBSS.training` is the name of the training data set containing the variables specified in `FIBI.formula` and where the step length ν and the initial number of boosting iterations were specified using the `boost_control()` function of R package **mboost**. Internal five-fold bootstrap cross-validation for

determining the optimal stopping iteration was carried out using the `cvrisk()` function of R package **mboost**:

```
> ntrain <- nrow(MBSS.training)
> bs5 <- rmultinom(5, ntrain, rep(1, ntrain) / ntrain)
> cvm <- cvrisk(FIBI.model, folds = bs5)
> st <- mstop(cvm)
```

After having determined the optimal stopping iteration (denoted by `st`), the "optimal" boosting fit at iteration `st` was calculated as follows:

```
> FIBI.optimal <- FIBI.model[st]
```

Afterwards, the `predict()` function of R package **mboost** was used to calculate predictions:

```
> pred <- predict(FIBI.optimal, newdata = MBSS.test, type = "response")
```

where `MBSS.test` denotes the test set of out-of-bootstrap observations (cf. Section 3). The `pred` object is a matrix containing the posterior class probabilities corresponding to the out-of-bootstrap observations. Using this object, the predicted outcome categories were calculated as described in Section 2. For a detailed description of the **mboost** package we refer to Hothorn et al. (2010).

Random forest analysis was carried out using the R package **randomForest** (Liaw and Wiener 2002, 2009). The random forest algorithm of R package `randomForest` has two main tuning parameters: (a) `ntree`, which is the number of trees used for the forest, and (b) `mtry`, which is the number of variables randomly selected at each node. To achieve sufficiently stable results, the number of trees was set to 2000 (see Cutler et al. 2007). The hyper-parameter `mtry` was tuned using additional internal 10-fold cross-validation.

Topographic surface plots were created using the R package **sp** (Pebesma and Bivand 2009). The `Kappa` function of R package **vcd** (Meyer et al. 2009) was used to compute weighted kappa indices.

Acknowledgements:

The work of Matthias Schmid and Torsten Hothorn was supported by Deutsche Forschungsgemeinschaft (DFG), grant HO 3242/1-3, and by the Interdisciplinary Center for Clinical Research (IZKF) at the University Hospital of the University of Erlangen-Nuremberg (Project J11). We thank the Maryland Department of Natural Resources for providing data from their Maryland Biological Stream Survey (MBSS). Funding for this project was provided by a Smithsonian Institution Post-Doctoral Fellowship awarded to KOM and by the US Environmental Protection Agency National Center for Environmental Research (NCER) Science to Achieve Results (STAR), grant #R831369. We thank the editor and two anonymous reviewers for helpful comments and suggestions.

Appendix

A Log-likelihood of the proportional odds model

The system of equations (4) implies that the log-likelihood of the proportional odds model is given by

$$\begin{aligned} l(f,\theta) &= -\,\mathrm{I}(\mathbf{Y}=1)\cdot\log(1+\exp(f-\theta_1)) \\ &\quad + \sum_{k=2}^{K-1}\mathrm{I}(\mathbf{Y}=k)\cdot\left[\log\left((1+\exp(f-\theta_k))^{-1}-(1+\exp(f-\theta_{k-1}))^{-1}\right)\right] \\ &\quad + \mathrm{I}(\mathbf{Y}=K)\cdot\log\left(1-(1+\exp(f-\theta_{K-1}))^{-1}\right)\ . \end{aligned}$$

Thus, the loss function used for the P/O boosting algorithm becomes $\rho = -l$. The negative derivative of ρ w.r.t. f is given by

$$\begin{aligned} -\frac{\partial\rho}{\partial f} &= \frac{\partial l}{\partial f} = -\,\mathrm{I}(\mathbf{Y}=1)\cdot(1+\exp(\theta_1-f))^{-1} \\ &+ \sum_{k=2}^{K-1}\mathrm{I}(\mathbf{Y}=k)\cdot\frac{1-\exp(2f-\theta_{k-1}-\theta_k)}{1+\exp(f-\theta_{k-1})+\exp(f-\theta_k)+\exp(2f-\theta_{k-1}-\theta_k)} \\ &+ \mathrm{I}(\mathbf{Y}=K)\cdot(1+\exp(f-\theta_{K-1}))^{-1}\ . \end{aligned}$$

B Unconstrained estimation of θ

In step 5 of the P/O boosting, we need to minimize the empirical risk $\mathcal{R}$ over θ subject to the constraint

$$-\infty < \theta_1 < \cdots < \theta_{K-1} < \theta_K = \infty\ .$$

This optimization problem can be re-formulated as follows: We introduce a vector $\delta = (\delta_1,\ldots,\delta_{K-1})'$ defined by

$$\begin{aligned} \theta_1 &= \delta_1\ , \\ \theta_2 &= \delta_1+\exp(\delta_2)\ , \\ \theta_3 &= \delta_1+\exp(\delta_2)+\exp(\delta_3)\ , \\ &\vdots \\ \theta_{K-1} &= \delta_1+\sum_{m=2}^{K-1}\exp(\delta_m)\ . \end{aligned}$$

Obviously, by using the system of equations presented above, the unconstrained minimization of $\mathcal{R}$ over δ is possible. Estimates of θ can subsequently be computed from the estimates of δ.

C Individual metrics used for IBI construction

Individual metrics used for construction of indices of biotic integrity (IBI) for fish and benthic macroinvertebrates. Modified from Southerland et al. (2005). Combined Highlands region comprises the Ridge and Valley, Blue Ridge, and Central Appalachians ecoregions. EPT = Ephemeroptera, Plecoptera, Trichoptera.

Fish Index of Biotic Integrity (FIBI)	Benthic Index of Biotic Integrity (BIBI)
Coastal Plain	Coastal Plain
Abundance per square meter	Number of Taxa
Number of Benthic species	Number of EPT Taxa
% Tolerant	Number of Ephemeroptera
% Generalist, Omnivores, Invertivores	% Intolerant Urban
% Round-bodied Suckers	% Ephemeroptera
% Abundance Dominant Taxa	Number of Scrapers
	% Climbers
Eastern Piedmont	
Abundance per square meter	Piedmont
Number of Benthic species	Number of Taxa
% Tolerant	Number of EPT Taxa
% Generalist, Omnivores, Invertivores	Number of Ephemeroptera
Biomass per square meter	% Intolerant Urban
% Lithophilic Spawners	% Chironomidae
	% Clingers
Warmwater Highlands	
Abundance per square meter	Combined Highlands
Number of Benthic species	Number of Taxa
% Tolerant	Number of EPT Taxa
% Generalist, Omnivores, Invertivores	Number of Ephemeroptera
% Insectivores	% Intolerant Urban
% Abundance of Dominant Taxa	% Tanytarsini
	% Scrapers
Coldwater Highlands	% Swimmers
Abundance per square meter	% Diptera
% Tolerant	
% Brook Trout	
% Sculpins	

D Predictor variables used for the analysis of the MBSS data

We included the following predictor variables in our analysis of the MBSS data:

- UTM easting and northing coordinates provided by MBSS (from Maryland State Plane Coordinate System). These predictors were used to take the spatial dependence structure of sample sites into account (predictor variables $\mathbf{X}_{\mathrm{E}}$ and $\mathbf{X}_{\mathrm{N}}$, see Section 2).
- Watershed Area, i.e., area of drainage upstream of sampling point (in km^2).
- Population density ($\#/\mathrm{km}^2$) of upstream watershed.
- Average upstream watershed elevation (in m).
- Average annual precipitation for upstream watershed elevation (in cm y^{-1}).
- Percentage of upstream watershed under tree cover.
- Percentage of upstream watershed under impervious surface cover.
- Percentage of upstream watershed under pasture cover.
- Percentage of upstream watershed under row crop cover.
- Percentage of upstream watershed under wetland cover.
- Percentage of upstream watershed under barren cover.
- Drainage density, defined as total stream length (in km) / watershed area (in km^2).
- Distance from sampling location to the nearest main stem stream (in km). Values of this predictor variable were set to zero for sites that drained directly into Chesapeake Bay.
- Average percentage of sand content in soil.
- Percentage of bedrock that is calcareous in a watershed.
- Dominant ecoregion (categorical predictor with six categories, see Section 3 and Omernik 1987).

A preliminary analysis of the data showed that the distributions of watershed area, population density, drainage density, upstream watershed elevation, and the percentages of upstream watershed under impervious surface, wetland, and barren cover were highly right-skewed. We therefore applied a log transformation to these predictor variables before fitting the proportional odds models. Since we observed a large number of zero percentages in calcareous bedrock, we transformed this predictor into a binary variable with categories "percentage of calcareous bedrock $= 0\%$" and "percentage of calcareous bedrock $> 0\%$".

E Marginal function estimates of predictor variables (FIBI model)

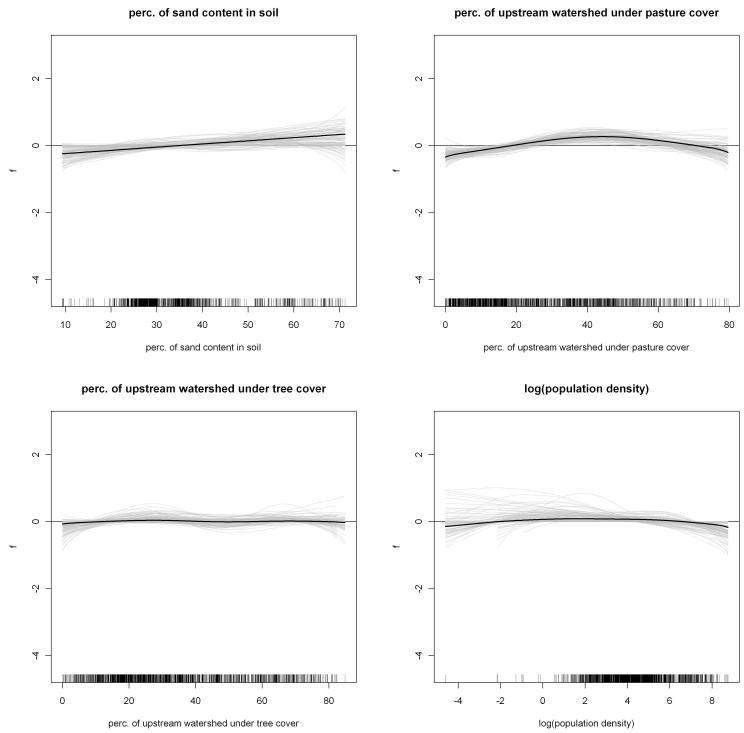

Figure 6: FIBI model - marginal function estimates corresponding to predictors "avg. percentage of sand content in soil", "percentage of upstream watershed under pasture cover", "percentage of upstream watershed under tree cover" and "population density of upstream watershed". These functions are relatively small in magnitude (compared to the functions shown in Figure 2 of the article) and are therefore not presented in Section 3 of the article.

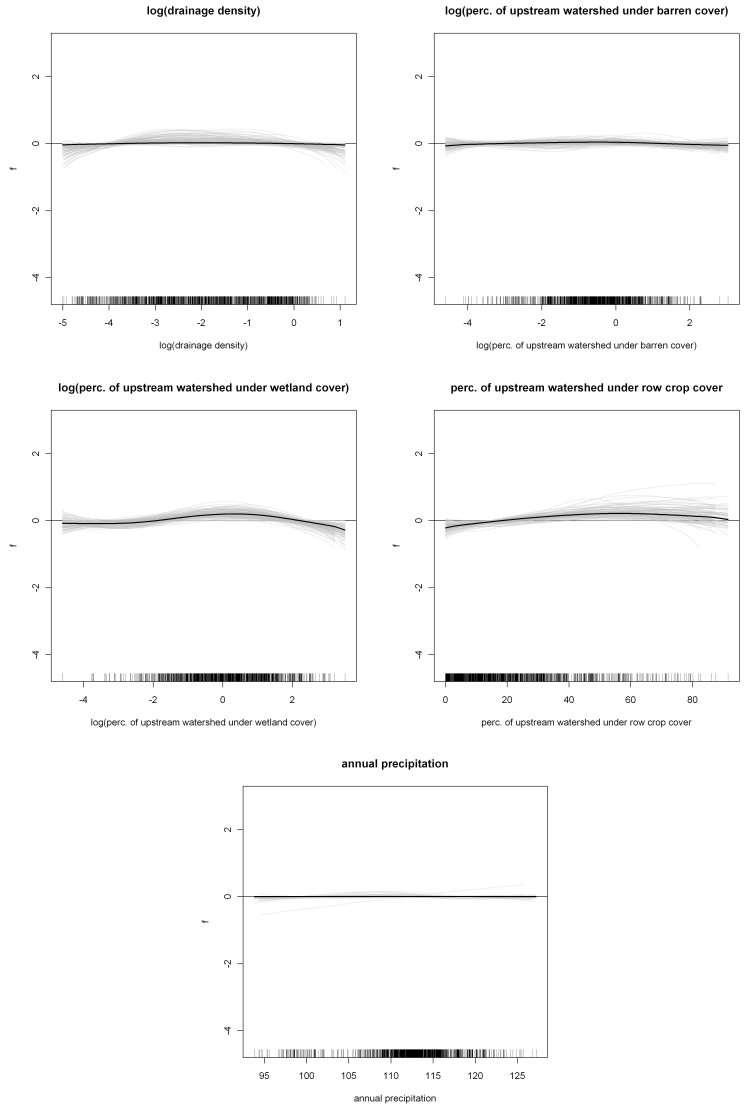

Figure 7: FIBI model - marginal function estimates corresponding to predictors "drainage density", "percentage of upstream watershed under barren cover", "percentage of upstream watershed under wetland cover", "percentage of upstream watershed under row crop cover" and "avg. annual precipitation for upstream watershed elevation". These functions are relatively small in magnitude (compared to the functions shown in Figure 2 of the article) and are therefore not presented in Section 3 of the article.

F Marginal function estimates of predictor variables (BIBI model)

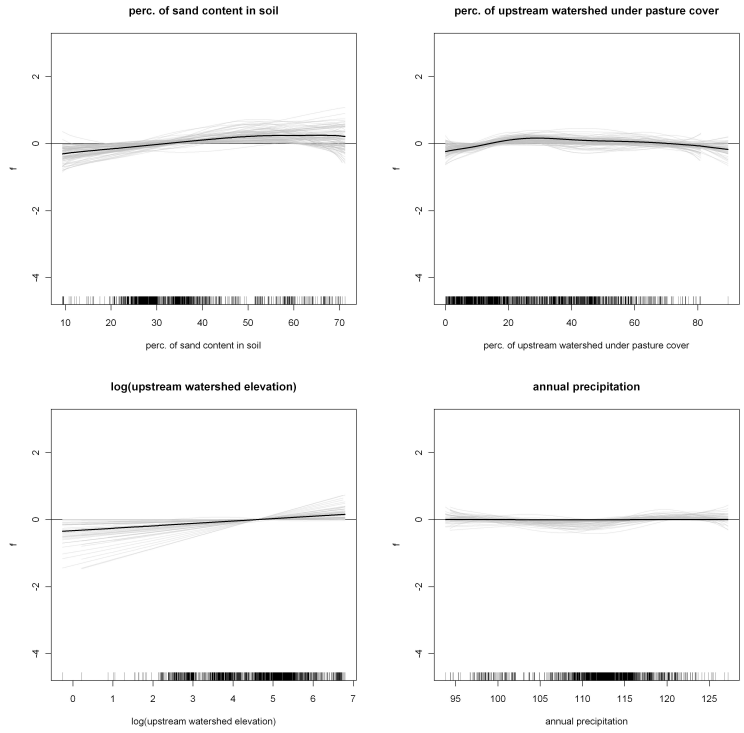

Figure 8: BIBI model - marginal function estimates corresponding to predictors "avg. percentage of sand content in soil", "percentage of upstream watershed under pasture cover", "avg. upstream watershed elevation" and "avg. annual precipitation for upstream watershed elevation". These functions are relatively small in magnitude (compared to the functions shown in Figure 4 of the article) and are therefore not presented in Section 3 of the article.

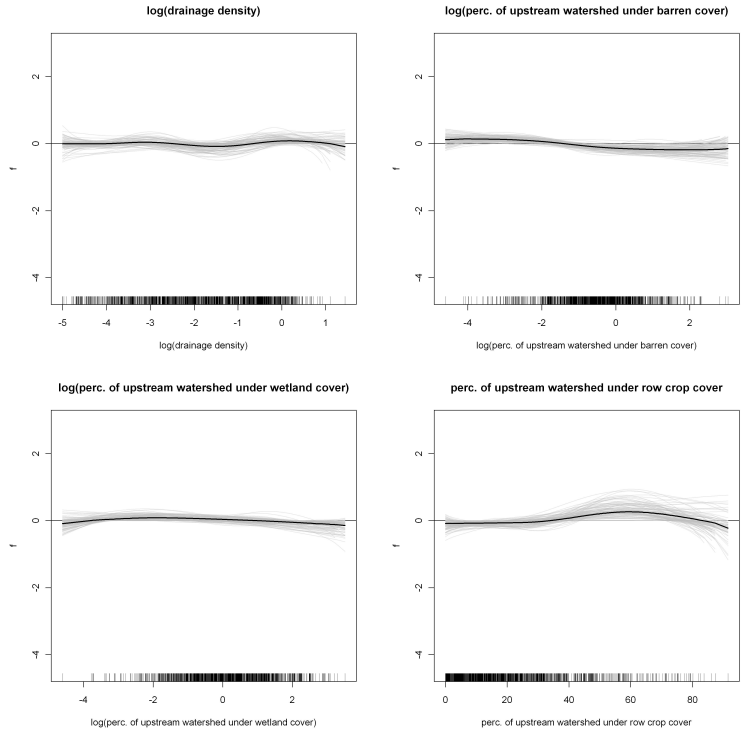

Figure 9: BIBI model - marginal function estimates corresponding to predictors "drainage density", "percentage of upstream watershed under barren cover", "percentage of upstream watershed under wetland cover" and "percentage of upstream watershed under row crop cover". These functions are relatively small in magnitude (compared to the functions shown in Figure 4 of the article) and are therefore not presented in Section 3 of the article.

References

Agresti, A.: Categorical Data Analysis. 2 edn. Wiley, New York (2002)

Angermeier, P.L., Schlosser, I.J.: Species-area relationship for stream fishes. Ecology **70**, 1450–1462 (1989)

Barbour, M.T., Gerritsen, J., Snyder, B.D., Stribling, J.B.: Rapid bioassessment protocols for use in streams and wadeable rivers: Periphyton, benthic macroinvertebrates and fish (2 ed.) (1999). Office of Water, US Environmental Protection Agency, Washington, DC

Barker, L.S., Felton, G.K., Russek-Cohen, E.: Use of Maryland Biological Stream Survey data to determine effects of agricultural riparian buffers on

measures of biological stream health. Environmental Monitoring and Assessment **117**, 1–19 (2006)

Bigler, C., Kulakowski, D., Veblen, T.T.: Multiple disturbance interactions and drought influence fire severity in Rocky Mountain subalpine forests. Ecology **86**, 3018–3029 (2005)

Breiman, L.: Random forests. Machine Learning **45**, 5–32 (2001)

Bühlmann, P., Hothorn, T.: Boosting algorithms: Regularization, prediction and model fitting (with discussion). Statistical Science **22**, 477–522 (2007)

Bühlmann, P., Yu, B.: Boosting with the L_2 loss: Regression and classification. Journal of the American Statistical Association **98**, 324–338 (2003)

Collier, K.J.: Linking multimetric and multivariate approaches to assess the ecological condition of streams. Environmental Monitoring and Assessment **157**, 113–124 (2009)

Cooper, C.: Assessing environmental impact on riparian benthic community vigor with unconditional estimates of quantile differences. Environmental and Ecological Statistics (2009). To appear

Cushing, C.E., Allan, J.D.: Streams: Their Ecology and Life. Academic Press, New York (2001)

Cutler, D.R., Edwards, T.C., Beard, K.H., Cutler, A., Hess, K.T.: Random forests for classification in ecology. Ecology **88**, 2783–2792 (2007)

Efron, B., Johnston, I., Hastie, T., Tibshirani, R.: Least angle regression. The Annals of Statistics **32**, 407–499 (2004)

Fahrmeir, L., Kneib, T., Lang, S.: Penalized structured additive regression: A Bayesian perspective. Statistica Sinica **14**, 731–761 (2004)

Fleiss, J.L., Cohen, J.: The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. Educational and Psychological Measurement **33**, 613–619 (1973)

Friedman, J.H., Hastie, T., Tibshirani, R.: Additive logistic regression: A statistical view of boosting (with discussion). The Annals of Statistics **28**, 337–407 (2000)

Gelfand, A.E.: Guest editorial: Spatial and spatio-temporal modeling in environmental and ecological statistics. Environmental and Ecological Statistics **14**, 191–192 (2007)

Hastie, T.: Discussion of "Boosting algorithms: Regularization, prediction and model fitting" by P. Bühlmann and T. Hothorn. Statistical Science **22**, 513–515 (2007)

Hastie, T., Tibshirani, R.: Generalized Additive Models. Chapman & Hall, London (1990)

Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2 edn. Springer, New York (2009)

Helms, B.S., Schoonover, J.E., Feminella, J.W.: Assessing influences of hydrology, physicochemistry, and habitat on stream fish assemblages across a changing landscape. Journal of the American Water Resources Association **45**, 157–169 (2009)

Homer, C., Huang, C.Q., Yang, L.M., Wylie, B., Coan, M.: Development of a 2001 National Land-Cover Database for the United States. Photogrammetric Engineering and Remote Sensing **70**, 829–840 (2004)

Hothorn, T., Bühlmann, P., Kneib, T., Schmid, M., Hofner, B.: mboost: Model-Based Boosting (2010). R package version 2.0-6. http://cran.r-project.org/web/packages/mboost/index.html

Hothorn, T., Leisch, F., Zeileis, A., Hornik, K.: The design and analysis of benchmark experiments. Journal of Computational and Graphical Statistics **14**(3), 675–699 (2005)

Joy, M.K., Death, R.G.: Predictive modelling and spatial mapping of freshwater fish and decapod assemblages using GIS and neural networks. Freshwater Biology **49**, 1036–1052 (2004)

Karr, J.R.: Biological integrity: A long-neglected aspect of water resource management. Ecological Applications **1**, 66–84 (1991)

Karr, J.R., Fausch, K.D., Angermeier, P.L., Yant, P.R., Schlosser, I.J.: Assessing Biological Integrity in Running Waters: A Method and its Rationale. 2 edn. Illinois Natural History Survey Special Publication 5, Champaign, Illinois (1986)

King, R.S., Baker, M.E., Whigham, D.F., Weller, D.E., Jordan, T.E., Kazyak, P.F., Hurd, M.K.: Spatial considerations for linking watershed land cover to ecological indicators in streams. Ecological Applications **15**, 137–153 (2005)

Kneib, T., Hothorn, T., Tutz, G.: Variable selection and model choice in geoadditive regression models. Biometrics **65**, 626–634 (2009)

Kneib, T., Müller, J., Hothorn, T.: Spatial smoothing techniques for the assessment of habitat suitability. Environmental and Ecological Statistics **15**, 343–364 (2008)

Liaw, A., Wiener, M.: Classification and regression by randomForest. R News **2**, 18–22 (2002)

Liaw, A., Wiener, M.: randomForest: Breiman and Cutler's random forests for classification and regression (2009). R package version 4.5-33. http://cran.r-project.org/web/packages/randomForest/index.html

Maloney, K.O., Weller, D.E., Russell, M.J., Hothorn, T.: Classifying the biological condition of small streams: An example using benthic macroinvertebrates. Journal of the North American Benthological Society **28**, 869–884 (2009)

Matthews, W.J., Robison, H.W.: Influence of drainage connectivity, drainage area and regional species richness on fishes of the Interior Highlands in Arkansas. American Midland Naturalist **139**, 1–19 (1998)

McCullagh, P.: Regression models for ordinal data (with discussion). Journal of the Royal Statistical Society, Series B **42**, 109–142 (1980)

Meier, L., van de Geer, S., Bühlmann, P.: High-dimensional additive modeling. The Annals of Statistics **37**, 3779–3821 (2009)

Meyer, D., Zeileis, A., Hornik, K.: vcd: Visualizing Categorical Data (2009). R package version 1.2-7. http://cran.r-project.org/web/packages/vcd/index.html

Montgomery, D.R.: Process domains and the river continuum. Journal of the American Water Resources Association **35**, 397–410 (1999)

Oberdorff, T., Hughes, R.M.: Modification of an index of biotic integrity based on fish assemblages to characterize rivers of the Seine Basin, France. Hydrobiologia **228**, 117–130 (1992)

O'Hara, R.B., Sillanpää, M.J.: A review of Bayesian variable selection methods: What, how and which. Bayesian Analysis **4**, 85–118 (2009)

Omernik, J.M.: Ecoregions of the conterminous United States. Annals of the Association of American Geographers **77**, 118–125 (1987)

Park, T., Casella, G.: The Bayesian Lasso. Journal of the American Statistical Association **103**, 681–686 (2008)

Paul, M.J., Meyer, J.L.: Streams in the urban landscape. Annual Review of Ecology and Systematics **32**, 333–365 (2001)

Pebesma, E.J., Bivand, R.: sp: Classes and Methods for Spatial Data (2009). R package version 0.9-47. http://cran.r-project.org/web/packages/sp/index.html

Peterson, E.E., Urquhart, N.S.: Predicting water quality impaired stream segments using landscape-scale data and a regional geostatistical model: A case study in Maryland. Environmental Monitoring and Assessment **121**, 615–638 (2006)

Pyne, M.I., Rader, R.B., Christensen, W.F.: Predicting local biological characteristics in streams: A comparison of landscape classifications. Freshwater Biology **52**, 1302–1321 (2007)

R Development Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2009). URL http://www.R-project.org

Rawlings, J.O., Pantula, S.G., Dickey, D.A.: Applied Regression Analysis: A Research Tool. 2 edn. Springer, New York (1998)

Roy, A.H., Rosemond, A.D., Paul, M.J., Leigh, D.S., Wallace, J.B.: Stream macroinvertebrate response to catchment urbanisation (Georgia, U.S.A.). Freshwater Biology **48**, 329–346 (2003)

Schleiger, S.L.: Use of an index of biotic integrity to detect effects of land uses on stream fish communities in west-central Georgia. Transactions of the American Fisheries Society **129**, 1118–1133 (2000)

Schmid, M., Hothorn, T.: Boosting additive models using component-wise P-splines. Computational Statistics & Data Analysis **53**, 298–311 (2008)

Schmid, M., Potapov, S., Pfahlberg, A., Hothorn, T.: Estimation and regularization techniques for regression models with multidimensional prediction functions. Statistics and Computing **20**, 139–150 (2010)

Southerland, M.T., Rogers, G.M., Kline, M.J., Morgan, R.P., Boward, D.M., Kazyak, P.F., Klauda, R.J., Stranko, S.A.: Maryland Biological Stream Survey 2000–2004, Volume XVI: New biological indicators to better assess the condition of Maryland streams (2005). DNR-12-0305-0100, Maryland Department of Natural Resources, Annapolis, Maryland

Tibshirani, R.: Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society, Series B **58**, 267–288 (1996)

USEPA: From the mountains to the sea: The state of Maryland's freshwater streams. (1999). EPA 903-R-99-023. Office of Research and Development, US Environmental Protection Agency, Washington, DC

USEPA: Wadeable streams assessment: A collaborative survey of the Nation's streams (2006). EPA 841-B-06-002. Office of Water, US Environmental Protection Agency, Washington, DC

Vannote, R.L., Minshall, G.W., Cummins, K.W., Sedell, J.R., Cushing, C.E.: The river continuum concept. Canadian Journal of Fisheries and Aquatic Sciences **37**, 130–137 (1980)

Walsh, C.J., Roy, A.H., Feminella, J.W., Cottingham, P.D., Groffman, P.M., Morgan, R.P.: The urban stream syndrome: Current knowledge and the search for a cure. Journal of the North American Benthological Society **24**, 706–723 (2005)

Wang, L., Lyons, J.: Fish and benthic macroinvertebrate assemblages as indicators of stream degradation in urbanizing watersheds. In: Simon, T.P. (ed.) Biological Response Signatures: Indicator Patterns Using Aquatic Communities, pp. 227–249. New York: CRC Press (2003)

Wood, S.: Generalized Additive Models: An Introduction with R. Chapman & Hall / CRC, Boca Raton (2006)