Techniques to Improve Ecological Interpretability of Black-Box Machine Learning Models

Case Study on Biological Health of Streams in the United States with Gradient Boosted Trees

Thomas WELCHOWSKI[✉], Kelly O. MALONEY[✉], Richard MITCHELL, and Matthias SCHMID[✉]

Statistical modeling of ecological data is often faced with a large number of variables as well as possible nonlinear relationships and higher-order interaction effects. *Gradient boosted trees* (GBT) have been successful in addressing these issues and have shown a good predictive performance in modeling nonlinear relationships, in particular in classification settings with a categorical response variable. They also tend to be robust against outliers. However, their black-box nature makes it difficult to interpret these models. We introduce several recently developed statistical tools to the environmental research community in order to advance interpretation of these black-box models. To analyze the properties of the tools, we applied gradient boosted trees to investigate biological health of streams within the contiguous USA, as measured by a benthic macroinvertebrate biotic index. Based on these data and a simulation study, we demonstrate the advantages and limitations of *partial dependence plots* (PDP), *individual conditional expectation* (ICE) curves and *accumulated local effects* (ALE) in their ability to identify covariate–response relationships. Additionally, interaction effects were quantified according to interaction strength (IAS) and Friedman’s H^2 statistic. Interpretable machine learning techniques are useful tools to open the black-box of gradient boosted trees in the environmental sciences. This finding is supported by our case study on the effect of impervious surface on the benthic condition, which agrees with previous results in the literature. Overall, the most important variables were ecoregion, bed stability, watershed area, riparian vegetation and catchment slope. These variables were also present in most identified interaction effects. In conclusion, graphical tools (PDP, ICE, ALE) enable visualization and easier interpretation of GBT but should be supported by analytical statistical measures. Future methodological research is needed to investigate the properties of interaction tests. Supplementary materials accompanying this paper appear on-line.

T. Welchowski (✉) · M. Schmid, Department of Medical Biometry, Informatics and Epidemiology, Medical Faculty, University of Bonn, Venusberg-Campus 1, 53127 Bonn, Germany
(E-mail: welchow@imbie.uni-bonn.de). K. Maloney, U.S. Geological Survey (USGS) Eastern Ecological Science Center at the Leetown Research Laboratory, 11649, Leetown Rd, 25430 Kearneysville, WV, USA. R. Mitchell, U.S. Environmental Protection Agency Office of Water Washington, Washington, DC 20460, USA.

© 2021 The Author(s)

Journal of Agricultural, Biological, and Environmental Statistics, Volume 27, Number 1, Pages 175–197
<https://doi.org/10.1007/s13253-021-00479-7>

Key Words: Boosting; Interpretable machine learning; Interaction terms; Macroinvertebrates; Stream health.

1. BACKGROUND

Ecological data are often high-dimensional and exhibit complex interactions among variables. Higher interactions may become a problem for supervised statistical modeling techniques that investigate the effects of additive combinations of a set of covariates on a response variable of interest. Important examples are generalized linear models (GLMs) and generalized additive models (GAMs) which are difficult to apply in scenarios where a data-driven detection of interaction effects is desirable (De'ath and Fabricius 2000). Here, we focus on classification procedures, which are special types of supervised modeling techniques for categorical responses. Typical applications in ecology include vegetation mapping by remote sensing (Steele 2000) and species distribution modeling (Guisan and Thuiller 2005). Among classification procedures, gradient boosted trees (GBT, Friedman 2001) have been widely used in environmental research because they often show a high prediction accuracy (e.g., De'ath 2007, Elith et al. 2008). Of note, GBT do not assume an additive relationship between the covariates and the response, and they are able to account for higher-order interactions among covariates. Furthermore, model fitting tends to be robust toward outliers in the covariates (Breiman et al. 1984), and the overall “strengths” of the covariate effects can be assessed using variable importance measures (Fisher et al. 2016). By definition, these measures estimate how much a covariate influences the prediction error compared to a situation where the covariate has no effect under a given model.

A drawback of boosted trees is that they yield so-called black-box models, which are usually not accessible to a simple interpretation of the direction and/or the shape of covariate effects. This is a consequence of the definition and the construction of GBT, which is obtained by the (non-interpretable) aggregation of classification trees. The lack of interpretability of covariate effects is a major issue in environmental applications where an objective is often to characterize and describe such effects. Note that estimating only variable importance is not sufficient for this purpose: Although variable importance measures quantify the impact on prediction, they provide no information on the characteristics of the underlying covariate–response relationships. The term *black-box* is typically used to describe a property of a statistical model that has been trained to optimize predictive performance but is complex to understand and does not provide clear interpretations itself without use of other tools. In contrast, “white box” models like linear models with few covariates or single decision trees with limited depth can be interpreted with less effort. These statistical models provide exact explanations, but are often less suited to capture the full complexity of the underlying biological phenomenon of interest.

In view of the aforementioned issues, statisticians have recently developed tools to improve the interpretability of black-box analyses. These tools can be of great value for researchers interested in gaining insight in the nature of covariate–response relationships while taking advantage of the flexibility of black-box methods. Based on a case study with

data of the National Rivers and Streams Assessment (NRSA) Survey (EPA 2016a), we analyze these tools and demonstrate how they can be utilized in environmental applications. Complementary content related to the case study is presented in the supplemental material.

2. METHODS AND RESULTS

2.1. US NATIONAL RIVERS AND STREAMS ASSESSMENT SURVEY

Our main objective with this case study was to examine interpretable machine learning techniques of gradient boosted trees and show properties as well as limitations of these in practice. For this, we used data taken as part of the US Environmental Protection Agency's (EPA) 2008–2009 US National Rivers and Streams Assessment (NRSA) Survey (EPA 2016a), which sampled streams across the conterminous USA to assess physiochemical and biological condition based on a generalized random tessellation stratified design (Stevens and Olsen 2004). The data acquisition followed the protocols of field methods described in EPA (2007). Analytical methods are given by EPA (2016b), and the data source is publicly available (EPA 2020). Especially in this ecological context, complex interactions between covariates exist, but it is less clear among which covariates and how these interactions are related. Altogether data were available for 994 streams. For our response, we used the benthic multimetric index (MMI) condition class, which evaluates stream condition based on the benthic macroinvertebrate community (Stribling and Dressing 2015). Benthic macroinvertebrates are often used to assess stream biological condition because they occur in a wide variety of habitats, are highly diverse and thus capture a range of sensitivity to disturbance, are relatively sedentary and have relatively long life cycles (Rosenberg and Resh 1993). They have been used worldwide to assess stream health (Barbour et al. 1999) with an aquatic biomonitoring system (Poquet et al. 2009). However, different stream regions require adaptations of the sampling and evaluation of benthic MMI condition classes. To reduce heterogeneity of the measurements in different stream regions due to sampling or evaluation adaptations the EPA categorized the response into three classes (poor, fair, good, with relative frequencies 45%, 23% and 32%, respectively, in our data) (Barbour et al. 1999).

Local benthic macroinvertebrate communities are governed by cumulative upstream and local factors (Allan 2004; Richards et al. 1997; Richards and Host 1994). Cumulative upstream factors (= covariates) in our model included natural landscape conditions (e.g., elevation, slope, sandy soils, maximum temperature and contributing area) and land use/land cover measures (areal coverage impervious surface, agriculture, wetlands, shrubs and grasslands, and tree canopy) (EPA 2007, 2016a). We also included average upstream nitrate and sulfate ions wet deposition as covariates to account for possible outside basin anthropogenic sources, and a generalized overall human disturbance index to capture any anthropogenic stressors not included in the above covariates. Local factors in our model included indices of bed stability, riparian vegetation condition and fish habitat cover. Lastly, we included the Omernick Level III ecoregion (Omernick 1987) to provide additional information missing in the previous datasets. A descriptive summary of all variables is provided in Table 1. The R Language and Environment for Statistical Computing, version 3.5.3 (R Core Team 2017), including the packages *gbm*, version 2.1.5 (Ridgeway 2017), *ICEbox*,

Table 1. Variables descriptive summary

Identifier	Units	ProportionMin in %	25 % Quan- tile	Median	75 % Quan- tile	Max	
Categorical variables							
Ecoregion	Coastal plains	13.93	—	—	—	—	
"	Northern Appalachians	10.81	—	—	—	—	
"	Northern plains	9.31	—	—	—	—	
"	Southern Appalachians	17.06	—	—	—	—	
"	Southern plains	8.86	—	—	—	—	
"	Temperate plains	9.87	—	—	—	—	
"	Upper Midwest	8.42	—	—	—	—	
"	Western Mountains	11.48	—	—	—	—	
"	Xeric West	10.26	—	—	—	—	
Benthic MMI condition	Poor	45.21	—	—	—	—	
"	Fair	23.19	—	—	—	—	
"	Good	31.60	—	—	—	—	
Continuous covariates							
Bed stability	Numeric	—	−4.42	−1.65	−0.83	0.00	3.46
Riparian vegetation condition	Numeric	—	−2.00	−0.45	−0.16	0.01	0.41
Fish cover	Numeric	—	−2.00	−0.81	−0.53	−0.31	0.36
Elevation	Meters	—	5.49	256.75	433.89	1170.79	3519.15
NO ₃ Deposition	Numeric	—	0.56	3.32	6.63	8.54	12.35
SO ₄ Deposition	Numeric	—	0.36	2.45	6.92	10.88	17.38
Tree canopy	Percent	—	0.00	9.17	32.00	58.30	96.46
Impervious surface	Percent	—	0.00	0.22	0.46	1.10	51.43
Sandy soils	Percent	—	2.02	29.86	39.42	51.37	95.74
Catchment slope	Percent	—	0.09	2.10	4.33	9.18	29.48
Agriculture	Percent	—	0.00	1.24	11.51	37.78	98.41
Shrub/Grass	Percent	—	0.00	2.24	10.75	43.95	99.90
Wetlands	Percent	—	0.00	0.14	0.95	3.43	95.58
Max temperature	Degrees in Celsius	—	5.62	13.99	16.94	21.33	31.08
Human Disturbance Index	Numeric	—	0.00	0.27	0.83	1.50	6.61
Watershed area	Square km	—	0.02	22.36	546.70	4943.82	3195990.00

Dashes represent non-applicable fields

version 1.1.2 (Goldstein et al. 2014), *corrplot* (Wei and Simko 2017), version 0.84 and *iml*, version 0.9.0 (Molnar et al. 2018) was used to conduct the data analysis. *R*-scripts to reproduce the case study and the simulations are available as supplemental material.

2.2. GRADIENT BOOSTED TREES, PREDICTIVE PERFORMANCE AND VARIABLE IMPORTANCE

The GBT model is an aggregation of multiple decision trees (Friedman 2001), which is used here to predict benthic MMI condition (multi-class response with three levels). Let $\mathbf{y} \in \{1, \dots, Q\}^n$ be the vector of observed response levels with $i = 1, \dots, n$ observations ($Q = 3$; $y_i = 1 \Rightarrow$ poor; $y_i = 2 \Rightarrow$ fair; $y_i = 3 \Rightarrow$ good) and the matrix of covariates

$\mathbf{X} \in \mathbb{R}^{n \times p}$. The vector $\mathbf{X}_{:,j} \in \mathbb{R}^n$ denotes all observations of the j -th covariate in $\mathbf{X}$ and $\mathbf{X}_{i,:} \in \mathbb{R}^p$ is defined as the i -th observation of the dataset. Further define $\mathbf{X}_{,-j} \in \mathbb{R}^{n \times (p-1)}$ to be the matrix without the j -th covariate. Following [Friedman \(2001\)](#), the models considered in this paper are based on the optimization of the multinomial log-likelihood function

$$\sum_{i=1}^n \sum_{q=1}^Q I_{y_i}(q) \log(f_{i,q}(\mathbf{X}_i)) , \quad (1)$$

which yields covariate-specific $\mathbf{X}_i$, predicted probabilities $f_{i,q}(\cdot)$ of the i -th sample for each of the levels of the categories $q = \{1, \dots, Q\}$. The indicator function $I_A(a)$ equals one if and only if the argument a equals the index A , which means in our study that only the prediction probabilities $f_{i,q}(\cdot)$ are evaluated if it matches the observed class y_i . With this specification, the decision trees are successively fitted to the gradient of the multinomial log-likelihood function, yielding iterative updates of the class probability estimates. In each iteration, the tree splits the covariates a predefined number of times, searching in each split for the best covariate $\mathbf{X}_{:,j}$ and the best binary threshold within the selected covariate, in order to produce sets of observations that are as heterogeneous as possible with respect to the gradient. The maximum number of splits from the root node to any of the terminal nodes is called the *depth* of the decision tree. A tree may split the same variable many times, and therefore, it is not possible to identify interaction effects only from the depth of the trees. Details of the algorithm are given in [Friedman \(2001\)](#); [Ridgeway \(2017\)](#).

Note that the multinomial model above does not account for the ordinal structure of the response levels (poor, fair, good). An alternative strategy for analysis would be to fit an ordinal regression model, e.g., along the lines of the approach by [Schmid et al. \(2011\)](#). When comparing this approach to the multinomial model used here, it should be noted that the multinomial model is far more flexible, as it involves less restrictions for the covariate effects due to ignoring the monotonicity constraints. Also, it would be difficult to perform a simplified analysis by fitting a binary classification model with a collapsed number of categories (e.g., by collapsing the poor and fair levels), as these categories are well established and have a distinct meaning to ecological researchers. Therefore, and since the aim of this work is not to benchmark different models with regard to their predictive performance but to explore and illustrate methods for interpretable machine learning, we applied a GBT without monotonicity constraints.

For our case study, we tuned the GBT model with respect to two hyperparameters: (1) the number of trees (or iterations) and (2) the depth of the trees. Note that higher depths are necessary for incorporating higher-order interaction effects. For tuning, we applied 10-fold stratified cross-validation (see, e.g., Sect. 7.10 in [Hastie et al. 2009](#)) on the complete dataset using the predictive multinomial log-likelihood as performance measure. Stratification ensures that the distribution of the response is approximately equal in the training and test datasets. Following [Ridgeway \(2017\)](#), random subsamples containing 50% of the respective training samples were used for model fitting in each iteration. The optimal gradient boosting model used 2990 trees with a maximum depth of 17.

Prediction accuracy was evaluated using a 10×10 -fold nested stratified cross-validation approach on the complete dataset. In this process, the model was tuned on each of the 10×10

Table 2. Aggregated confusion matrix obtained from 10×10 nested stratified cross-validation

Predicted/observed	Poor	Fair	Good
Poor	613	194	131
Fair	97	76	100
Good	101	146	336

training sets using 10-fold stratified cross-validation as described above. After refitting the tuned models to the respective 10×10 training sets, prediction accuracy was evaluated on the 10×10 test datasets. Note that a high maximum depth value is an indicator of either strong interaction effects or nonlinear effects. Overall the classification accuracy of the tuned GBT approach was 57.13%, which is much higher than obtained from random guessing (33%). The sum of the outer fold confusion matrices is given in Table 2. Poor condition was most accurately predicted (sensitivity of “poor” condition 75.59 %), followed by good condition (sensitivity of “good” condition 59.26 %) and then fair condition (sensitivity of “fair” condition 18.27 %, Table 2).

The overall contribution of each variable to prediction models is often evaluated using variable importance measures. A model independent type of variable importance, which is also considered in this paper, is based on a covariate-wise permutation of the data values (Fisher et al. 2016). Note that in this article the permutation variable importance is estimated from the training data, as our goal was not to evaluate predictive characteristics but to investigate the properties of the final trained model. This is in line with the work by Breiman (2001); Friedman (2001) on random forests and GBT that also evaluated variable importance only on the training data. For a discussion on the use of test data in the assessment of variable importance, see Molnar (2019). To compute the importance value of a covariate, the predictive performance of a fitted model is first evaluated on the complete (training) dataset. Afterward, given the fitted model, the association between the response and the covariate is removed by randomly permuting the values of the covariate of interest. The permutation variable importance is then defined by the average difference between the model’s performance on the unpermuted data and the respective performance on the permuted data. It is computed separately for each covariate. In general, variable importance evaluates the influence of the covariates relative to each other (with respect to a given performance measure), but neither measures how each variable affects the predictions in the fitted model nor assesses interaction effects between covariates. Note that we did not use variable importance based on the estimation process of decision trees that measures how well a split improves the prediction, because it was shown to be biased by the scaling of covariates (e.g., higher number of unique values) (Strobl et al. 2007).

The most important variable in the GBT model was ecoregion (Fig. 1). Log relative bed stability and watershed area contributed most to the improvement of the model within the set of continuous covariates.

Figure 1. Permuted variable importance for the top 15 most influential covariates use in the GBT model, organized in decreasing order. Descriptions for covariates provided in Supplemental Table 1.

2.3. MARGINAL EFFECTS OF COVARIATES ON THE RESPONSE

2.3.1. Partial Dependence Plots and Individual Conditional Expectation Plots

Partial dependence plots (PDP) were suggested by [Friedman \(2001\)](#) to explore black-box models graphically. The vector of random variables referring to the data-generating process of the covariates is denoted by $\mathbf{x} \in \mathbb{R}^p$. The matrix $\mathbf{Z} \in \mathbb{R}^{n_2 \times p}$ consists of n_2 pre-specified grid values of the covariates of interest. PDPs estimate the expectation of the distribution function of the random variable $\mathbf{x}_{(j, \dots, k)}$ conditional on the remaining variables $\mathbf{x}_{-(j, \dots, k)}$, defined as

$$f_{PDP}(\mathbf{x}_{(j, \dots, k)}) = \mathbb{E}_{\mathbf{x}_{-(j, \dots, k)}} (f(\mathbf{x}_{(j, \dots, k)})) = \int f(\mathbf{x}_{(j, \dots, k)} | \mathbf{x}_{-(j, \dots, k)}) p(\mathbf{x}_{-(j, \dots, k)}) d\mathbf{x}_{-(j, \dots, k)}, \quad (2)$$

and estimated via

$$\hat{f}_{PDP}(\mathbf{Z}_{s, (j, \dots, k)}) = \frac{1}{n} \sum_{i=1}^n f(\mathbf{Z}_{s, (j, \dots, k)} | \mathbf{X}_{i, -(j, \dots, k)}), \quad (3)$$

where $f(\cdot)$ denotes the prediction function of the fitted black-box model. In case of GBT with categorical responses, $f(\cdot)$ is measured on the probability scale. Consider the simplest case if the set of variables consists of just one covariate j . Then, the j -th variable of interest is varied according to pre-specified values (usually all observations). For each specified value, the prediction function is evaluated on the dataset with the j -th covariate replaced by

Figure 2. Centered marginal effect estimates of log relative bed stability on benthic MMI condition. Each line represents one observation in the dataset. Thick marked yellow lines correspond to the PDP curve (pointwise average over all curves). On the y-axis, f_1 denotes the prediction of response class “Poor”, f_2 represents prediction of response class “Fair”, and f_3 equals prediction of response class “Good” as defined in Sect. 2.2. ICE curves are colored by ecoregion, the most important covariate (Fig. 1).

the first pre-specified value. This process is then repeated for all grid values separately. If two variables $\{j, k\}$ are chosen a grid is constructed of all possible values of those variables. The effect of all other variables $\mathbf{x}_{-(j,k)}$ is averaged out. Thus, the resulting PDP shows the average relation of the chosen variable to the response. In the case of many complex interactions, as usually encountered in environmental research, the average may not appropriately reflect the relationship, because positive and negative interaction effects may cancel and the variability across different observations is not captured. In Goldstein et al. (2014), the PDP framework was extended to consider all observations as disaggregated *individual conditional expectation* (ICE) curves rather than the average value. This means each observation i is represented by one line in the ICE plot given other covariates $\mathbf{X}_{i,-(j,\dots,k)}$ evaluated over the grid values of the set of variables $\{j, \dots, k\}$. In contrast to the average PDP curve, ICE visualizes the variability of predictions across different covariates instead of a single covariate of interest. For further details of ICE curves, we refer to Goldstein et al. (2014).

To illustrate the utility of these methods, we focus mostly on the covariates *log relative bed stability* (being the most important continuous variable in the GBM according to Figs. 1) and *impervious surface* (given the importance of this predictor in stream ecology Walsh

Figure 3. Centered marginal effect estimates of impervious surface on benthic MMI condition. Each line represents one observation in the dataset. Thick marked yellow lines correspond to the PDP curve (pointwise average over all curves). ICE curves are colored by ecoregion, the most important covariate (Fig. 1) (Color figure online).

et al. 2005). The ICE plot of the complete class distribution for benthic MMI condition with respect to bed stability is shown in Fig. 2. For log relative bed stability, PDPs suggest that higher bed stability does on average lower the probability of a poor response and there is on average a peak for the good condition between -1 and 1 (Fig. 3). However, the ICE curves show much variation around these values, revealing high uncertainty around the average PDP line. When the ICE plots between log bed stability and benthic condition were examined by ecoregion, different patterns emerged. For example, in the Northern Appalachians, the probability of a poor condition decreased with increase in log bed stability, whereas most ICE lines for fair and good conditions increased (Fig. 2). In the Southern plains, the probability of a poor condition increased, fair condition decreased with increase in bed stability, and the probability of a good condition was less clear as many ICE lines fall above and below the PDP curve.

For impervious surface, the PDPs suggest that the probability of a poor response increases with impervious cover, whereas those for fair and good decrease; for good probability, a break can be seen at 1–3% (Fig. 3) which is slightly below currently believed thresholds (Paul and Meyer 2001; Schueler et al. 2009) but is within other findings that use more detailed analyses (Baker and King 2010; Maloney et al. 2012). Unlike log bed stability, ICE lines showed no large shifts in the probability of a good or fair condition (all decreased) or poor condition (all increased, Fig. 3). In the left side panel of Fig. 3, most observations in the northern Appalachians regions show increasing poor probability. Conversely, most Southern

plains have a higher probability of good response than Northern Appalachians (right side panel). Most ICE lines in Coastal Plains show decreasing probability of a fair response (middle panel). However, again here the ICE curves reveal high variations among sites that cannot be explained by one additional variable, ranging from relatively flat curves to steep curves for a good condition score. Lee et al. (2012). Irrespective of this variation, ICE plot support higher impervious surfaces increase the probability of a poor benthic MMI condition class and decreases the probabilities of good benthic conditions supporting the findings of Lee et al. (2012). In our case study, the PDP curves also show that for increasing levels of the human disturbance index (see, e.g., Kaufmann et al. 1999), the probability of poor benthic MMI condition increases (Fig. S1 in the Supplements) and that the probability of a good condition stays up to a value of one and then drops rapidly. Furthermore, it is well known that human disturbance can have a negative impact on aquatic ecosystems (Shen et al. 2017). Increasing riparian vegetation along rivers, shown in Fig. S2 in the Supplements, yields on average higher probabilities for good benthic conditions up to a maximum of 0.2. This agrees with other studies (Anbumozhia et al. 2005), which found that riparian vegetation buffer zones reduce impairment of water quality.

In all previous ICE plots (Figs. 2, 3 and Figs. S1, S2 in the Supplements) there is a lot of variability in the ICE curves. These substantial changes in the directions of the effects on the response level are not explained by the “marginal” variables only. If a variable would determine the functional form of the ICE curves independent of other predictor variables, the curves should mostly overlap with the PDP curve. When other covariates have only additive effects all curves in the ICE plots should be parallel to each other. These findings are in line with Goldstein et al. (2014), who proposed ICE curves to fix disadvantages of PDPs. From the ICE plots it is however not clear, which variables interact with each other, highlighting that these curves are not suited to identify interaction effects with more than two variables. In some situations, a solution to this problem could be the coloring of ICE plots by the levels of third variables. For example, in the middle panel of Fig. 6 there is a cluster of ICE curves below the PDP curve with high values of maximal annual temperature affecting the impact of elevation on the fair benthic multimetric condition. However, coloring by third variables may not help in situations where there are multiple interaction effects present (for comparison see left and right panels of Fig. 6) which will be discussed in Sect. 2.4.2. Note that in Supplement Section 1.3, we additionally provide alternative figures depicting probability integral transformed x-axes in the range [0, 1]. In these figures, the values on the x-axis are more evenly spaced; however, they are not as easy to interpret as on the original scale.

2.3.2. Accumulated Local Effect Plots

Accumulated local effects (ALE) plots (Apley 2016) have the same purpose as PDPs, which is to estimate the marginal effect one or two features have on the predicted response of a machine learning model. ALE plots are formally defined by

$$f_{ALE}(\mathbf{x}_j) = \int_{z_{0,j}}^{\mathbf{x}_j} E_{\mathbf{x}_{-j}} \left(\frac{\partial f(\mathbf{x}_j)}{\partial \mathbf{x}_j} \middle| \mathbf{x}_j = \mathbf{z}_j \right) d\mathbf{z}_j, \quad (4)$$

where $z_{0,j}$ denotes a minimum value of the covariate of interest. In our case study, the derivative of the function $\partial f(\mathbf{x}_j)/\partial \mathbf{x}_j$ can be estimated because the predicted probabilities $f_{i,q}(\mathbf{x})$ of a specific benthic MMI condition q are continuous. The expected value is taken with respect to the random variables $\mathbf{x}_{-j}$ and evaluated at the specific covariate $\mathbf{x}_j = \mathbf{z}_j$. ALE plots are estimated from data via the following equations, which are explained in more detail below:

$$\hat{f}_{ALE}(\mathbf{X}_{i,j}; i) = \sum_{l=1}^{T_{i,j}} \frac{1}{n_j(l)} \sum_{s: \mathbf{X}_{s,j} \in N_j(l)} (f(\mathbf{Z}_{l,j}, \mathbf{X}_{s,-j}) - f(\mathbf{Z}_{l-1,j}, \mathbf{X}_{s,-j})) , \quad (5)$$

$$T_{i,j} = l : \mathbf{X}_{i,j} \in N_j(l) , \quad (6)$$

$$N_j(l) = \{(\mathbf{Z}_{l-1,j}, \mathbf{Z}_{l,j})\} , \quad l = 1, \dots, L , \quad (7)$$

$$n_j(l) = \sum_{i=1}^n I(\mathbf{X}_{i,j} \in N_j(l)) ; \quad \sum_{l=1}^L n_j(l) = n . \quad (8)$$

By definition, ALE plots estimate the cumulative average local changes of the predictions for a covariate of interest (Equation (4)). The conditional derivative $\partial f(\mathbf{x}_j)/\partial \mathbf{x}_j | \mathbf{x}_j = \mathbf{z}_j$ is approximated by finite differences on a pre-specified grid matrix $\mathbf{Z}$ of intervals for each covariate j (Eq. (5)). In each interval $N_j : l = 1, \dots, L$, the differences of the prediction model $f(\mathbf{Z}_{l,j}, \mathbf{X}_{s,-j}) - f(\mathbf{Z}_{l-1,j}, \mathbf{X}_{s,-j})$ are averaged over the set of $n_j(l)$ available observed values $\mathbf{X}_{s,-j}$ out of total n observations. The set of L intervals was constructed using empirical quantiles $\hat{\psi}$ of one specific covariate (equally distributed between the minimal observed value $\hat{\psi}_0$ and the maximal observed value $\hat{\psi}_1$ into intervals $N_j(1) = [\hat{\psi}_0, \hat{\psi}_{1/(L+1)}]$, $N_j(2) = (\hat{\psi}_{1/(L+1)}, \hat{\psi}_{2/(L+1)}]$, $\dots$, $N_j(L) = (\hat{\psi}_{L/(L+1)}, \hat{\psi}_1]$.

The average local changes are accumulated from the minimal value $z_{0,j}$ up to a given value of the covariate of interest (integral in Eq. (4)). ALE curves are centered by the average value of $f_{ALE}(\mathbf{x}_j)$. Note that although the index i is not visible in Eq. (5), all observations of $\mathbf{X}$ are used implicitly in the sum $\sum_{s: \mathbf{X}_{s,j} \in N_j(l)}$ when the complete ALE curve is estimated up to the maximal observed value $\hat{\psi}_1$. It is not meaningful to compute ALE curves for a single observation, because to approximate the expectation of the conditional derivative $E_{\mathbf{x}_{-j}}(\partial f(\mathbf{x}_j)/\partial \mathbf{x}_j | \mathbf{x}_j = \mathbf{z}_j)$ by the empirical mean it is necessary to use more than one observation.

ALE plots differ from PDPs in the following ways: First, the observations are averaged conditional on observed combinations of data values and not over the complete range of (possibly unobserved) values. This conditional averaging is especially relevant when the covariates are correlated. Second the changes in the predictions are evaluated within a local interval around the pre-specified value of $\mathbf{Z}_{l,j}$ instead of the predictions $f(\mathbf{X}_{i,j})$. The rationale behind this evaluation is that without conditioning on other covariates the estimated main effect of the covariate of interest is biased when covariates are correlated (Greene 2018). The calculation of differences within local intervals keeps the values of other covariates at comparable values and can thus separate the main effect of the covariate of interest. The third difference between ALE plots and PDPs is the accumulation of the estimated average derivatives up to the given value $\mathbf{X}_{i,j}$. Furthermore in the case of two-dimensional ALE

Figure 4. Accumulated local effect (ALE) plot of the effect of log relative bed stability on the good benthic MMI condition. The y-axis displays the marginal effect of log relative bed stability on the prediction probability for class good. Each curve was computed from a nonparametric bootstrap sample using a pre-specified number of intervals. The variability of the ALE curve increases with the number of intervals; however, the overall trend does not change much.

effects, the main effects of the corresponding covariate are subtracted (Apley 2016), which means that these show the possible interaction effects between two covariates. Disadvantages of ALE plots include the need for a prior specification of the number of intervals and the lack of an extension to individual observations in order to display variability of curves (analogous to the extension of PDPs to ICE plots). If the number of intervals is large, ALE curves can be estimated on a finer grid and thus provide more information on the relationships between covariates and response. However, the uncertainty in the estimation increases with the number of intervals, as there are less observations available for estimation within each interval (Molnar 2019). Consider, for example, the ALE plot of the effect of log relative bed stability on the good benthic condition in Fig. 4. In this figure, only one level of the response is considered in isolation because one wants to focus on the behavior of ALE plots with respect to the number of intervals. The final GBT model was held fixed, and each curve represents one out of 100 nonparametric bootstrap samples used to investigate the variability of the ALE curve estimate. Especially in the lower and upper regions of log relative bed stability (where there are fewer observations), the variance of the ALE curve

Figure 5. Centered marginal effects of impervious surface on benthic MMI condition. The plots present a comparison between PDP curves (black lines) and ALE curves (dashed red lines) based on the gradient boosting model fitted to the complete dataset (Color figure online).

increased with the number of intervals. The PDP curve in Fig. 2 shows a similarly shaped parabolic relationship between log relative bed stability and probability of a good response as displayed in the upper left plot of Fig. 4.

We evaluated ALE plots in Fig. 5 on impervious surface to see how those compare to the PDP in Fig. 3 with a grid size of 100. The trends of ALE curves are comparable across the responses poor, fair and good to the PDP curves. Figure S13 in the Supplements shows that impervious surface is positively correlated to NO_3 and SO_4 deposition and negatively to elevation and grasslands. PDP do not properly take these correlations into account, and therefore, the averaging may include unrealistic cases (analogous to the theoretical exam-

ple in Fig. S15 in the Supplements). In general, this assumption should be checked when there are high correlations between variables. In Fig. 5, the shapes of the PDP curves look quite similar to the ALE curves, indicating that averaging over the sample distribution was robust to correlations of impervious surface with several other variables. In Fig. 4, all ALE curves showed comparable trends just like the PDP curve in the rightmost part of Fig. 2. The additional features of ALE plots (derivative, conditional distribution) did not influence the resulting marginal curves. ICE plots have the additional advantage of displaying heterogeneity between observations, so we suggest using these for interpretation instead of ALE for the marginal effect of impervious surface and bed stability in this case study.

2.4. INTERACTION EFFECTS IN BLACK-BOX MODELS

2.4.1. Global Interaction Strength of Black-Box Models

There have been some attempts in environmental research to identify interaction effects, e.g., with a linear regression approach (Elith et al. 2008; De'ath 2007; Fraker and Peacor 2008). In this framework, the first step is to estimate partial dependence function from the GBT model given two covariates $\hat{f}_{PDP}(\mathbf{Z}_{s,(j,k)})$ and separately for each covariate $\hat{f}_{PDP}(\mathbf{Z}_{s,j})$, $\hat{f}_{PDP}(\mathbf{Z}_{s,k})$. Then, the relationship between the partial dependence values $\hat{f}_{PDP}(\mathbf{Z}_{s,(j,k)})$ as response and the partial dependence values of the two covariates $\hat{f}_{PDP}(\mathbf{Z}_{s,j})$, $\hat{f}_{PDP}(\mathbf{Z}_{s,k})$ is fitted by a linear model. The authors argue that when the residual variance of the model is high, this indicates an interaction effect. However, this approach has some disadvantages. First, it does not take the nonlinear functions of the covariates into account. For example, consider the quadratic equation $y = x_1^2 + x_2^2$. A linear model with untransformed covariates is not able to fit this surface and will have high residual variance. Secondly, if the null hypothesis of no interaction is true, the overall F test statistic (null hypothesis that all covariate coefficients equal zero (Weisberg 2014)) must not be zero due to sample fluctuations and multi-collinearity of covariates. Furthermore, the residual variance will increase with the scale of the response variable y .

In addition to graphical analysis (PDP, ICE and ALE plots), we investigated the use of analytic statistics to measure interactions. An overall measure of global interaction strength (IAS) can be defined by the expectation of a pre-specified loss function of the predictions and an additive model with only main effects (Molnar et al. 2019). Here, we consider the squared error loss function and generalize IAS to include multiple responses with categories $q = 1, \dots, Q$ (yielding predictions $\tilde{f}_1, \dots, \tilde{f}_Q$ as random variables):

$$\text{IAS} = \frac{\mathbb{E}_{\mathbf{x}} \left(\sum_{q=1}^Q (\tilde{f}_q - \tilde{f}_{1,q})^2 \right)}{\mathbb{E}_{\mathbf{x}} \left(\sum_{q=1}^Q (\tilde{f}_q - \tilde{f}_{0,q})^2 \right)}, \quad (9)$$

$$\tilde{f}_{0,q} = \mathbb{E}_{\mathbf{x}} \left(\tilde{f}_q \right), \quad (10)$$

$$\tilde{f}_{1,q} = \tilde{f}_{0,q} + \sum_{j=1}^J f_{ALE}(\mathbf{x}_j). \quad (11)$$

The additive approximation $\tilde{f}_{1,q}$ of the black-box model $\tilde{f}_q$ is derived by summing up the expected prediction $\tilde{f}_{0,q}$ and the ALE functions (= “main effects”) of all covariates. For example, if the prediction model consists of only main effects $\tilde{f}_q = \tilde{f}_{0,q} + \sum_{j=1}^J \tilde{f}_{ALE,q}(\mathbf{x}_j)$, then the nominator in Eq. 9 equals zero and therefore $IAS = 0$. Conversely, if a prediction model contains only interaction effects, i.e., $\tilde{f}_{0,q} = 0$ and $\sum_{j=1}^J \tilde{f}_{ALE,q}(\mathbf{x}_j) = 0$ then the nominator and denominator in Eq. 9 are equal and $IAS = 1$. Analogous to the R^2 measure in linear regression, values of IAS greater than zero indicate how much the interaction effects contribute to the variance of the predictions. In practice, IAS is estimated by replacing the theoretical ALE functions in Eq. 11 by the estimate defined in Eq. 5.

To characterize the behavior of IAS , we consider an example using the species-discharge relationship for Cahaba, Black Warrior and Tallapoosa river fishes in the southeastern part of the USA (Mcgarvey and Ward 2008). The authors showed that in their study the logarithmic species richness and the logarithmic discharge had a correlation of 83%. Another study showed a correlation of 73% between dissolved oxygen levels and flow discharge (Dou et al. 2018). Benthic macroinvertebrate richness also depends on dissolved oxygen levels and is additionally sensitive to pH (Courtney and Clements 1998). Contemplate the linear model of logarithmic benthic richness (y) with two main effects β_1 of logarithmic dissolved oxygen (x_1) and β_2 of logarithmic pH level (x_2) and an interaction effect $\beta_{1,2}$. This model would be defined by $E(y|\mathbf{x}) = \beta_1 x_1 + \beta_2 x_2 + \beta_{1,2} x_1 x_2$. From an ecological point of view, it is reasonable to assume a positive effect of logarithmic dissolved oxygen ($\beta_1 > 0$) and a negative effect of pH ($\beta_2 < 0$). Further assume that the fitted model has sufficient data to approximate the population model well. Since $f_{1,q} = \beta_1 x_1 + \beta_2 x_2$ it follows that $IAS = E_{\mathbf{x}}((\beta_{1,2} x_1 x_2)^2) / E_{\mathbf{x}}((\beta_1 x_1 + \beta_2 x_2 + \beta_{1,2} x_1 x_2)^2)$. Assuming (for a moment) independence of dissolved oxygen and pH (i.e., $x_1 \perp x_2$) and standard normal distributions on the logarithmic scale (i.e., $x_1, x_2 \sim N(0, 1)$). In this case, the theoretical value of IAS is $\beta_{1,2}^2 / (\beta_1^2 + \beta_2^2 + \beta_{1,2}^2)$. For example, if all coefficients are equal, i.e., $\beta_1 = \beta_2 = \beta_{1,2}$, then $IAS = 1/3$. If the interaction effect is $\beta_{1,2} = \sqrt{2}\beta_1 = \sqrt{2}\beta_2$, then IAS would take a medium value of 0.5. If the interaction effect is four times larger than both main effects, i.e., $\beta_{1,2} = 4\beta_1 = 4\beta_2$, then an IAS of 8/9 would be the result. Thus, IAS measures how strong interaction effects between dissolved oxygen and flow discharge affect the model predictions. However in the case of more than two covariates, IAS cannot differentiate between which covariates the interaction occurs within the prediction model.

In our case study, the GBT prediction model for benthic MMI condition yielded a global interaction strength of $IAS = 0.5554$, which means that more than 50% of the total variance of the predictions can be attributed to interaction effects. If predictions for each class are analyzed separately, the poor, fair and good conditions yielded $IAS = 0.5032$, $IAS = 0.6964$, and $IAS = 0.5681$, respectively. Obviously, the interaction effects were stronger in the fair category than in the other conditions.

2.4.2. Local Interaction Strength of Black-Box Models for Specific Covariates

In contrast to IAS , Friedmans H^2 statistic (Friedman and Popescu 2008) assesses the local interaction strength for a specific set of covariates as given in Eq. 12. In this section,

we assume that all the PDP functions are centered, which means that their expectation with respect to the covariates of interest is zero. Under these conditions, a central concept is the excess variance of the PDP function of two covariates $E_{\mathbf{x}_{-(j,k)}}(f(\mathbf{x}_{(j,k)}))$ and their corresponding PDP functions of each covariate $E_{\mathbf{x}_{-j}}(f(\mathbf{x}_j))$, $E_{\mathbf{x}_{-k}}(f(\mathbf{x}_k))$. This excess variance is then divided by the total variance of the two covariate PDP function that gives the fraction of unexplained variance by the PDP functions with one covariate. When two covariates, j and k , interact with each other, changes in the prediction function according to variable j depend on the value of variable k . In practice,

$$H_{j,k}^2 = \frac{\text{Var}_{\mathbf{x}_{(j,k)}}(E_{\mathbf{x}_{-(j,k)}}(f(\mathbf{x}_{(j,k)})) - E_{\mathbf{x}_{-j}}(f(\mathbf{x}_j)) - E_{\mathbf{x}_{-k}}(f(\mathbf{x}_k)))}{\text{Var}_{\mathbf{x}_{(j,k)}}(E_{\mathbf{x}_{-(j,k)}}(f(\mathbf{x}_{(j,k)})))} \quad (12)$$

is estimated by replacing each PDP function by the estimator in (3), which yields

$$\hat{H}_{j,k}^2 = \frac{\sum_{i=1}^n \left(\hat{f}_{PDP}(\mathbf{X}_{i,(j,k)}) - \hat{f}_{PDP}(\mathbf{X}_{i,j}) - \hat{f}_{PDP}(\mathbf{X}_{i,k}) \right)^2}{\sum_{i=1}^n \hat{f}_{PDP}^2(\mathbf{X}_{i,(j,k)})}. \quad (13)$$

To have meaningful reference values for the observed local interaction strength $\hat{H}_{j,k}^2$, we follow [Friedman and Popescu \(2008\)](#) and use parametric bootstrapping to test whether an interaction effect of covariates j , k is present. The null hypothesis H_0 is that the fraction of unexplained variance equals zero $H_{j,k}^2 = 0$ and the alternative hypothesis H_1 is that this is not the case $H_{j,k}^2 \neq 0$. First, a model with no interactions between covariates is fitted. For example in the context of a GBT model, this is equivalent to setting the depths of all decision trees to one, because then each tree has only two terminal nodes. This model is tuned by the predictive multinomial log-likelihood (see Eq. 1) based on stratified cross-validation of the number of boosting iterations. In the next step, artificial datasets are created by randomly generating categorical responses with multinomial distribution given the covariates of the original dataset according to the fitted GBT model without interactions. For each generated artificial dataset, an unrestricted GBT is fitted. Based on this model, the local interaction strength statistic $\hat{H}_{j,k}^2$ is evaluated. These bootstrapped values are used to estimate the reference null distribution H_0 of $\hat{H}_{j,k}^2$, which will be then compared with the statistic of the original dataset. The bootstrap p values are computed by $\hat{p} = \sum_{l=1}^m \left(H_{j,k,l}^2 \geq \hat{H}_{j,k}^2 + 1 \right) / (m + 1)$ where $H_{j,k,l}^2$ denotes the l -th bootstrap replication of $\hat{H}_{j,k}^2$ ([North et al. 2002](#)) and corrected for multiple testing with the Benjamini–Yekutieli procedure ([Benjamini and Yekutieli 2001](#)) controlling false discovery rate to 0.05. The result is considered significant if a (pre-specified) significance threshold α is above the p value $\hat{p}$. The advantage of this approach is that possible nonlinear covariate effects and their interactions can be detected, because this bootstrap approach can be used with any black-box prediction model. Furthermore $\hat{H}_{j,k}^2$ can be extended to check for higher-order interaction effects, e.g., with three or more covariates. It should also be noted that results could be further analyzed using a visual test for additivity as presented in [Goldstein et al. \(2014\)](#), but this method is not explored here.

Figure 6. ICE plots visualizing the three strongest significant two-way interaction effects (centered), as identified by parametric bootstrap test based on $\hat{H}_{j,k}^2$: NO₃ colored by SO₄ deposition (left panel), elevation colored by maximal temperature (middle panel), agriculture colored by catchment slope (right panel). Blue color corresponds to low and red color to high values. Here, PDP highlights an interaction effect between elevation and maximum temperature regarding occurrence of fair benthic MMI condition (Color figure online) .

To compare H^2 with IAS, consider the following extension of the linear regression example of Sect. 2.4.1: In addition to the assumptions of the previous example, assume a linear regression model that additionally includes the logarithmic waste concentration per m^3 as covariate (x_3). The model with all pairwise interactions between the three covariates is then given by $E(y|x) = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{1,2} x_1 x_2 + \beta_{1,3} x_1 x_3 + \beta_{2,3} x_2 x_3$. It can be shown that $H_{j,k}^2 = \beta_{j,k}^2 / (\beta_j^2 + \beta_k^2 + \beta_{j,k}^2)$. In comparison with $IAS = (\beta_{1,2}^2 + \beta_{1,3}^2 + \beta_{2,3}^2) / (\beta_1^2 + \beta_2^2 + \beta_3^2 + \beta_{1,2}^2 + \beta_{1,3}^2 + \beta_{2,3}^2)$, Friedmans H^2 statistic only uses the main and interaction effects of the specified subset j, k of covariates. For example,

Figure 7. Two-dimensional ALE plot showing interactions between catchment slope and percentage of agriculture on the fair benthic condition. Brighter regions correspond to larger changes in prediction, which tend to be more present for higher values of catchment slope and percentage of agriculture .

if $j = 1$ and $k = 2$, the squared interaction effect of logarithmic dissolved oxygen and logarithmic pH level of the response logarithmic benthic richness $\beta_{1,2}^2$ would be compared against all other squared effects $\beta_1^2 + \beta_2^2 + \beta_{1,2}^2$ that include these two covariates. IAS instead uses the sum of all pairwise squared interaction effects and compares that to the squared effects of all covariates included in the model.

In our case study on benthic MMI condition, parametric bootstrap tests of the H^2 statistic yielded 17 out of 136 possible significant two-way interactions that affected the probability of a fair MMI condition and 2 out of 680 possible three-way interactions that affected the good benthic MMI condition. The coloring in the ICE curves in Fig. 6 visualizes the strongest two-way interactions. The first covariates on the x-axis (NO_3 , SO_4 deposition and agriculture) were chosen to be more important according to variable importance. In the left plot, lower values of SO_4 deposition are more common above the PDP line than below. This suggests a negative interaction of SO_4 deposition on the marginal effect of the NO_3 deposition. In the right plot of Fig. 6, there is a lot of variation due to interaction effects. In fact there is no clear pattern regarding the coloring by agriculture percentage, which is an indicator that more covariates affect the shape of the marginal effect of catchment slope. All significant two-way and three-way interaction effects are given in the Supplements (Tables S3, S4, S5). Note that a strong interaction effect does not always translate to a high variable importance. For example, NO_3 and SO_4 deposition had the strongest significant interaction

effect, but were less important and did not contribute much to the predictive performance (see Fig. 1). Furthermore, a two-dimensional ALE plot of catchment slope and agriculture showed a strong interacting effect on the fair benthic MMI condition (Fig. 7).

In addition, a sensitivity analysis was conducted using a parametric approach with multinomial logistic regression. Note that with this method there is no prior knowledge available to specify the relevant interactions effects. We do not want to incorporate results from GBT into the parametric model, because this would not allow a fair comparison between the results. Also, it is not feasible to compute the full set of all possible models with two- and three-way interactions and their combinations with reasonable effort and precision. For this reason, we carried out forward/backward variable selection based on generalized AIC with penalty $\sqrt{\log(n)}$ (as a compromise between AIC and BIC, see [Stasinopoulos et al. 2017](#)). The starting model of the forward/backward procedure included the main effects of all covariates (Supplements Section 1). The upper bound in model specification would have main effects plus all two- and three-way interaction effects. To take the Hauck–Donner effect ([Yee 2020](#)) into account, likelihood ratio tests were used to investigate population inference. The estimates of the selected main effects are presented in Table S7 and the corresponding interaction effects in Table S8 in the Supplements. In comparison with the permutation variable importance of GBT in Fig. 1, the main effects with odds ratios higher than two did not include watershed or catchment slope, but included fish cover. The magnitude of the parametric interaction effects was low except for the interaction between human disturbance index and catchment slope $\text{good} = 0.9588$ as well as between log relative bed stability and fish cover $\text{fair} = 1.6885$, $\text{good} = 1.6145$. The strongest interaction of fair condition in the GBT model was NO_3 deposition and SO_4 deposition and these variables have no p values below 0.05 in the parametric model. No three-way interactions were selected in multinomial logistic regression.

3. SUMMARY AND DISCUSSION

In this study, interpretable machine learning techniques were explained and their strengths and weaknesses were demonstrated in a case study on stream condition, as measured by benthic MMI condition class. A gradient boosted tree model was tuned for predictive performance, and permutation variable importance was used to investigate which covariates contributed most to the performance. ICE, including PDPs, and ALE plots were used to estimate how one or two covariates affected the functional forms of the covariate–response relationships. An overall measure of interaction strength was estimated by IAS, and covariate combinations were tested for interactions using the H^2 statistic. An advantage of all of these methods is that they can be applied to any black-box prediction model. From an ecological point of view, most of the estimated average functional relationships (e.g., impervious surface on benthic MMI condition) were highly plausible. In the Supplements Section 2, we present additional simulation experiments to explore the relationship between ICE and ALE plots.

While the aforementioned methods in the last paragraph proved to be very useful to gain insight in the characteristics of the (black-box) GBT model, our case study also revealed

several shortcomings of these methods when applied to ecological data. For example, permutation variable importance yields a qualitative order of covariates; however, it is a point measure and the uncertainty of this estimator is usually not considered in practice. There are, however, some approaches to conduct statistical inference for variable importance measures (Van der Laan 2006; Janitza et al. 2016).

H^2 bootstrap tests of interaction effects also can be applied to higher-order interactions, but are computationally quite expensive to evaluate. Therefore, it is not possible to evaluate all possible sets of covariates. By definition, PDP curves do not indicate interaction effects. This may be fixed by considering ICE curves. However, ICE plots are not able to identify which interactions are present, and coloring by another covariate rarely produces straightforward graphical structures that are easy to interpret. Furthermore, ICE plots do not average over the conditional distribution of the covariate of interest, because these are based on the PDP definition. Therefore, ICE plots may include combinations of covariate values that are rarely observed. Nevertheless, it is possible that multiple covariates are correlated and that ALE and PDP curves have a big overlap. Overall further research is needed to investigate, in particular, the reliability of H^2 interaction tests and the relation between ALE and PDP curves in case of multivariate dependencies between covariates.

Another point of discussion is whether the expanded measure IAS would benefit from weighting by prevalences of the responses. On the one hand, with the addition of weighting IAS could focus more on the most frequent predictions. On the other hand, the benefit of weighting depends on whether the response population is unbalanced and on the specification of the prediction model. There are several statistical approaches to explicitly take unbalanced data into account, e.g., Chawla et al. (2002); Cieslak and Chawla (2008), and in this case, weighting may be not necessary. Furthermore, if the goal of the analysis is to compare the relative amount of interactions in several statistical models, then the predictions of each category should be taken into account equally, because interaction effects in one category affect the probabilities that other categories occur.

ACKNOWLEDGEMENTS

Financial support for from Deutsche Forschungsgemeinschaft is gratefully acknowledged. Support for was provided by the Fisheries Program of the USGS's Ecosystems Mission Area and Priority Ecosystems Science (Chesapeake Bay Study Unit) programs. Thanks to Paul Stackelberg (USGS) for comments on an early version on this manuscript. Use of trade, product or firm names does not imply endorsement by the US Government. The views expressed in this article are those of the authors and do not necessarily represent the views or policies of the US EPA. We have no conflict of interest to declare.

Funding Open Access funding enabled and organized by Projekt DEAL.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

[Received October 2020. Revised September 2021. Accepted September 2021. Published Online October 2021.]

REFERENCES

- Allan JD (2004) Landscapes and riverscapes: the influence of land use on stream ecosystems. *Annu Rev Ecol Evol Syst* 35(1):257–284. <https://doi.org/10.1146/annurev.ecolsys.35.120202.110122>
- Anbumozhia V, Radhakrishnan J, Yamajic E (2005) Impact of riparian buffer zones on water quality and associated management considerations. *Ecol Eng* 24(5):517–523. <https://doi.org/10.1016/j.ecoleng.2004.01.007>
- Apley DW (2016) Visualizing the effects of predictor variables in black box supervised learning models. Technical Report [arXiv:1612.08468](https://arxiv.org/abs/1612.08468)
- Baker ME, King RS (2010) A new method for detecting and interpreting biodiversity and ecological community thresholds. *Methods Ecol Evol* 1(1):25–37. <https://doi.org/10.1111/j.2041-210X.2009.00007.x>
- Barbour MT, Gerritson J, Snyder BD, et al (1999) Rapid bioassessment protocols for use in streams and Wadeable rivers: Periphyton, benthic macroinvertebrates and fish. second edition. Technical report, United States Environmental Protection Agency, Washington, DC, EPA/841/B-99/002
- Benjamini Y, Yekutieli D (2001) The control of the false discovery rate in multiple testing under dependency. *Ann Stat* 29:1165–1188
- Breiman L (2001) Random forests. *Mach Learn* 45:5–32
- Breiman L, Friedman J, Stone CJ et al (1984) *Classification and Regression Trees*. The Wadsworth and Brooks-Cole statistics-probability series. Taylor and Francis, Oxfordshire, ISBN 9780412048418. <https://doi.org/10.1201/9781315139470>
- Chawla NV, Bowyer KW, Hall LO et al (2002) Smote: synthetic minority over-sampling technique. *J Artif Intell Res* 16:321–357. <https://doi.org/10.1613/jair.953>
- Cieslak DA, Chawla NV (2008) Learning decision trees for unbalanced data. In: Daelemans W, Goethals B, Morik K (eds) *Machine learning and knowledge discovery in databases*. Springer, Berlin Heidelberg, pp 241–256
- Courtney LA, Clements WH (1998) Effects of acidic pH on benthic macroinvertebrate communities in stream microcosms. *Hydrobiologia* 379:135–145
- De'ath G (2007) Boosted trees for ecological modeling and prediction. *Ecology* 88(1):243–251. [https://doi.org/10.1890/0012-9658\(2007\)88\[243:BTFEMA\]2.0.CO;2](https://doi.org/10.1890/0012-9658(2007)88[243:BTFEMA]2.0.CO;2)
- De'ath G, Fabricius KE (2000) Classification and regression trees: a powerful yet simple technique for ecological data analysis. *Ecology* 81(11):3178–3192. [https://doi.org/10.1890/0012-9658\(2000\)081\[3178:CARTAP\]2.0.CO;2](https://doi.org/10.1890/0012-9658(2000)081[3178:CARTAP]2.0.CO;2)
- Dou B, Hosseini Y, Lee C, Rosenberg C, Wu N (2018) The relationship between stream discharge and dissolved oxygen levels at canyon creek, and implications towards salmon performance. *Open J Syst Exped*, 8
- Elith J, Leathwick JR, Hastie T (2008) A working guide to boosted regression trees. *J Anim Ecol* 77(4):802–813. <https://doi.org/10.1111/j.1365-2656.2008.01390.x>
- EPA. National rivers and streams assessment: Field operations manual. Technical report, United States Environmental Protection Agency, 2007. EPA-841-B-07-009
- EPA. National rivers and streams assessment 2008-2009: A collaborative survey. Technical report, United States Environmental Protection Agency, Washington, D.C., 2016a. EPA/841/R-16/007
- EPA. National rivers and streams assessment 2008-2009 technical report. Technical report, United States Environmental Protection Agency, 2016b. EPA/841/R-16/008
- EPA. Data from the national aquatic resource surveys. <https://www.epa.gov/national-aquatic-resource-surveys/data-national-aquatic-resource-surveys>, 2020. Accessed: 2020-09-29
- Fisher A, Rudin C, Dominici F (2016) All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. Technical Report [arXiv:1801.01489](https://arxiv.org/abs/1801.01489)
- Fraker ME, Peacor SD (2008) Statistical tests for biological interactions: a comparison of permutation tests and analysis of variance. *Acta Oecol* 33:66–72. <https://doi.org/10.1016/j.actao.2007.09.001>

- Friedman JH (2001) Greedy function approximation: a gradient boosting machine. *Ann Stat* 29(5):1189–1232
- Friedman JH, Popescu BE (2008) Predictive learning via rule ensembles. *Ann Appl Stat* 2(3):916–954. <https://doi.org/10.1214/07-AOAS148>
- Goldstein A, Kapelner A, Bleich J et al (2014) Peeking inside the black box visualizing statistical learning with plots of individual conditional expectation. *J Comput Graph Stat* 24(1):44–65. <https://doi.org/10.1080/10618600.2014.907095>
- Greene WH (2018) *Econometric analysis*, 8th edn. Harlow, Pearson
- Guisan A, Thuiller W (2005) Predicting species distribution: offering more than simple habitat models. *Ecol Lett* 8(9):993–1009. <https://doi.org/10.1111/j.1461-0248.2005.00792.x>
- Hastie T, Tibshirani R, Friedman JH (2009) *The elements of statistical learning: data mining, inference, and prediction, 2nd Edition, corrected 12th printing*. Springer series in statistics. Springer, Berlin <https://doi.org/10.1007/978-0-387-84858-7>
- Janitza S, Celik E, Boulesteix AL (2016) A computationally fast variable importance test for random forests for high-dimensional data. *Adv Data Anal Classif* 12(4):885–915. <https://doi.org/10.1007/s11634-016-0276-4>
- Kaufmann PR, Levine P, Robison EG et al (1999) Quantifying physical habitat in Wadeable streams. Technical report, U.S. United States Environmental Protection Agency, Washington, DC
- Lee B-Y, Park S-J, Paule MC et al (2012) Effects of impervious cover on the surface water quality and aquatic ecosystem of the kyeongan stream in south korea. *Water Environ Res* 84(8):635–645. <https://doi.org/10.2175/106143012X13373550426878>
- Maloney KO, Schmid M, Weller DE (2012) Applying additive modelling and gradient boosting to assess the effects of watershed and reach characteristics on riverine assemblages. *Methods Ecol Evol* 3(1):116–128. <https://doi.org/10.1111/j.2041-210X.2011.00124.x>
- Mcgarvey DJ, Ward MG (2008) Scale dependence in the species-discharge relationship for fishes of the southeastern USA. *Freshw Biol* 53:2206–2219
- Molnar C (2019) *Interpretable Machine Learning*. Leanpub, <https://christophm.github.io/interpretable-ml-book/>
- Molnar C, Bischl B, Casalicchio G (2018) IML: an R package for interpretable machine learning. *JOSS* 3(26):786. <https://doi.org/10.21105/joss.00786>
- Molnar C, Casalicchio G, Bischl B (2019) Quantifying interpretability of arbitrary machine learning models through functional decomposition. Technical Report [arXiv:1904.03867](https://arxiv.org/abs/1904.03867)
- North BV, Curtis D, Sham PC (2002) A note on the calculation of empirical p values from monte carlo procedures. *Am J Hum Genet* 71(2):439–441. <https://doi.org/10.1086/346173>
- Omerik JM (1987) Ecoregions of the conterminous united states. *Ann Assoc Am Geogr* 77(1):118–125. <https://doi.org/10.1111/j.1467-8306.1987.tb00149.x>
- Paul MJ, Meyer JL (2001) Streams in the urban landscape. *Annu Rev Ecol Syst* 32(1):333–365. <https://doi.org/10.1146/annurev.ecolsys.32.081501.114040>
- Poquet JM, Alba-Tercedor J, Punti T et al (2009) The mediterranean prediction and classification system (med-pacs): An implementation of the rivpacs/ausrivas predictive approach for assessing mediterranean aquatic macroinvertebrate communities. *Hydrobiologia* 623:153–171. <https://doi.org/10.1007/s10750-008-9655-y>
- R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017. URL <https://www.R-project.org/>
- Richards C, Host G (1994) Examining land use influences on stream habitats and macroinvertebrates: A GIS approach. *J Am Water Resour Assoc* 30(4):729–738. <https://doi.org/10.1111/j.1752-1688.1994.tb03325.x>
- Richards C, Haro R, Johnson L et al (1997) Catchment and reach-scale properties as indicators of macroinvertebrate species traits. *Freshw Biol* 37(1):219–230. <https://doi.org/10.1046/j.1365-2427.1997.d01-540.x>
- Ridgeway G (2017) *gbm: Generalized Boosted Regression Models*, URL <https://CRAN.R-project.org/package=gbm>. R package version 2.1.3
- Rosenberg DM, Resh VH (eds) (1993) *Freshwater Biomonitoring and Benthic Macroinvertebrates*. Chapman/Hall, New York

- Schmid M, Hothorn T, Maloney KO, Weller DE, Potapov S (2011) Geoadditive regression modeling of stream biological condition. *Environ Ecol Stat* 18(4):709–733
- Schueler TR, Fraley-McNeal L, Cappiella K (2009) Is impervious cover still important? review of recent research. *J Hydrol Eng* 14(4):309–315. [https://doi.org/10.1061/\(ASCE\)1084-0699\(2009\)14:4\(309\)](https://doi.org/10.1061/(ASCE)1084-0699(2009)14:4(309))
- Shen Y, Cao H, Tang M et al (2017) The human threat to river ecosystems at the watershed scale: An ecological security assessment of the songhua river basin, northeast china. *Water* 9(3):14. <https://doi.org/10.3390/w9030219>
- Stasinopoulos Mikis D, Rigby RA, Heller GZ, Voudouris V, Bastiani FD (2017) Flexible regression and smoothing: using GAMLSS in R. Chapman and Hall/CRC, Boca Raton
- Steele BM (2000) Combining multiple classifiers: an application using spatial and remotely sensed information for land cover type mapping. *Remote Sens Environ* 74(3):545–556. [https://doi.org/10.1016/S0034-4257\(00\)00145-0](https://doi.org/10.1016/S0034-4257(00)00145-0)
- Stevens DL Jr, Olsen AR (2004) Spatially balanced sampling of natural resources. *J Am Stat Assoc* 99(465):262–278. <https://doi.org/10.1198/016214504000000250>
- Stribling JB, Dressing SA (2015) Applying benthic macroinvertebrate multimetric indexes to stream condition assessments. Technical report, United States United States Environmental Protection Agency (EPA)
- Strobl C, Boulesteix A-L, Zeileis A et al (2007) Bias in random forest variable importance measures. *BMC Bioinf* 8(1):1471–2105
- Van der Laan MJ (2006) Statistical inference for variable importance. *Int J Biostat*. <https://doi.org/10.2202/1557-4679.1008>
- Walsh CJ, Roy AH, Feminella JW et al (2005) The urban stream syndrome: current knowledge and the search for a cure. *J N Am Benthol Soc* 24(3):706–723. <https://doi.org/10.1899/04-028.1>
- Wei T, Simko V (2017) *R package “corrplot”: Visualization of a Correlation Matrix*, URL <https://github.com/taiyun/corrplot>. (Version 0.84)
- Weisberg S (2014) *Applied linear regression*. 4th edn. Wiley, Hoboken NJ, <http://z.umn.edu/alr4ed>
- Yee TW (2020) On the hauck-donner effect in wald tests: Detection, tipping points, and parameter space characterization