Boosted Beta Regression

Matthias Schmid[1]*, Florian Wickler[2], Kelly O. Maloney[3], Richard Mitchell[4], Nora Fenske[2], Andreas Mayr[1]

[1] Department of Medical Informatics, Biometry and Epidemiology, Friedrich-Alexander University Erlangen-Nuremberg, Erlangen, Germany, [2] Department of Statistics, University of Munich, Munich, Germany, [3] USGS - Leetown Science Center, Wellsboro, Pennsylvania, United States of America, [4] USEPA Office of Wetlands, Oceans, and Watersheds, Washington, DC, United States of America

Abstract

Regression analysis with a bounded outcome is a common problem in applied statistics. Typical examples include regression models for percentage outcomes and the analysis of ratings that are measured on a bounded scale. In this paper, we consider beta regression, which is a generalization of logit models to situations where the response is continuous on the interval (0,1). Consequently, beta regression is a convenient tool for analyzing percentage responses. The classical approach to fit a beta regression model is to use maximum likelihood estimation with subsequent AIC-based variable selection. As an alternative to this established - yet unstable - approach, we propose a new estimation technique called boosted beta regression. With boosted beta regression estimation and variable selection can be carried out simultaneously in a highly efficient way. Additionally, both the mean and the variance of a percentage response can be modeled using flexible nonlinear covariate effects. As a consequence, the new method accounts for common problems such as overdispersion and non-binomial variance structures.

Citation: Schmid M, Wickler F, Maloney KO, Mitchell R, Fenske N, et al. (2013) Boosted Beta Regression. PLoS ONE 8(4): e61623. doi:10.1371/journal.pone.0061623

Editor: Mauro Gasparini, Politecnico di Torino, Italy

Received February 7, 2013; **Accepted** March 11, 2013; **Published** April 23, 2013

Funding: The work of Matthias Schmid and Andreas Mayr was supported by Deutsche Forschungsgemeinschaft (DFG) (www.dfg.de), grant SCHM 2966/1-1. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: matthias.schmid@imbe.med.uni-erlangen.de

Introduction

The analysis of percentage data is a common issue in quantitative research. Percentage data arise in many scientific fields, for example in ecology [1–4], in econometrics [5,6], and in medical research [7,8]. A recent survey conducted by Warton & Hui [2] even found that nearly one third of papers published in *Ecology* in 2008/09 dealt with the analysis of percentage data.

From a statistical perspective, the analysis of percentage data is a challenging problem. This problem primarily concerns the development of *regression models* for percentage outcomes, which may be biased and inefficient if the specific nature of percentage outcomes is not taken into account. Although it would be convenient to use percentage responses as outcome variables in ordinary least squares (OLS) regression, this approach is problematic because OLS regression does not account for the fact that percentages are bounded by the interval c [9]. Hence, in order to avoid biased estimators and hypothesis tests, regression techniques that are tailored to the analysis of percentage outcomes are needed.

In the literature, various alternative methods to model percentage data have been proposed. A well-known strategy is to transform the percentage outcome Y and to carry out OLS regression using the transformed response values. Typical examples of transformations include the arcsine square root transformation (to stabilize the variance of Y, see [2]) and the logit transformation $\log(Y/(1-Y))$ (to map the interval (0,1) to the real line). While transformations followed by OLS regression are popular among analysts, their use is currently being challenged [2]. This is because (a) the assumptions of OLS regression are often not met despite the transformation of the data, and because (b) a reasonable interpretation of estimation results is only possible on the transformed scale but not on the original percentage scale [10].

An alternative approach for the analysis of percentage outcomes is to use regression models that are based on the binomial distribution [2,11]. These models are suitable if the outcome is of the form "x out of N" (with $Y := x/N$). Binomial models are, however, inapplicable in situations where the raw numbers x and N are not available. In addition, there are numerous applications where percentage outcomes are non-binomial (e.g., in ecological research, where fractions of communities and ecosystem measures are often of interest).

To overcome the aforementioned problems and limitations, we consider *beta regression* [6], which is an alternative to variable transformations and binomial models. In beta regression, the response variable is assumed to follow a beta distribution on the interval (0,1). Because the beta distribution has a highly flexible shape, it is suitable to represent arbitrary outcome variables measured on the percentage scale [10]. Consequently, beta regression is appropriate for analyzing both binomial and non-binomial data. Moreover, the results of a beta regression model have essentially the same interpretation as logistic regression. Estimates of the model parameters can conveniently be obtained using maximum likelihood estimation [6].

In the statistical literature, beta regression has been established as a powerful technique to model percentages and proportions [10]. Also, the method has been used in a variety of research fields [3,8,12]. There are applications, however, where classical beta regression methodology still has a number of limitations:

1. Scientific databases often involve large numbers of potential predictor variables that could be included in a regression model.

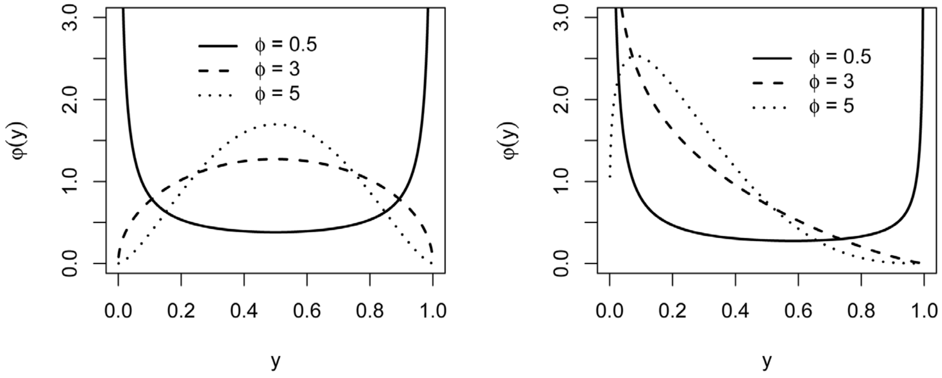

Figure 1. Probability density functions for beta distributions. Probability density functions for beta distributions with $\mu = 0.5$ (left) and $\mu = 0.25$ (right).
doi:10.1371/journal.pone.0061623.g001

Consequently, if maximum likelihood estimation is used to fit a beta regression model, the model may become too complex and may thus overfit the data. This usually leads to a large variance and to a high uncertainty about the predictor-response relationships. As a consequence, techniques for *variable selection* in beta regression models are needed.

2. Statistical models often suffer from *multicollinearity* problems, meaning that predictor variables are highly correlated. Also, observations of the response variable may be affected by *spatial correlation*, which is, for example, a common problem in ecology [13,14]. To date, these issues have not been incorporated into beta regression methodology.

3. In many applications, predictor-response relationships are *nonlinear* in nature [15,16]. This means that the linear predictor $\mathbf{X}^T\beta$ of the classical beta regression model needs to be replaced by a more flexible function that allows for an appropriate quantification of nonlinear predictor effects. Although Simas et al. [17] have recently suggested an approach to incorporate nonlinear effects into beta regression models, this approach requires the functional form of the predictor-response relationships (e.g., quadratic or exponential) to be specified in advance. In cases where the functional forms of predictor effects are unknown, a more flexible approach based on smooth nonlinear effects is desirable.

4. Percentage outcomes that are based on the binomial model $Y = x/N$ are often *overdispersed*, meaning that they show a larger variability than expected by the binomial distribution. Classical beta regression models conveniently account for overdispersion by including a precision parameter ϕ to adjust the conditional variance of the percentage outcome (see the next section for

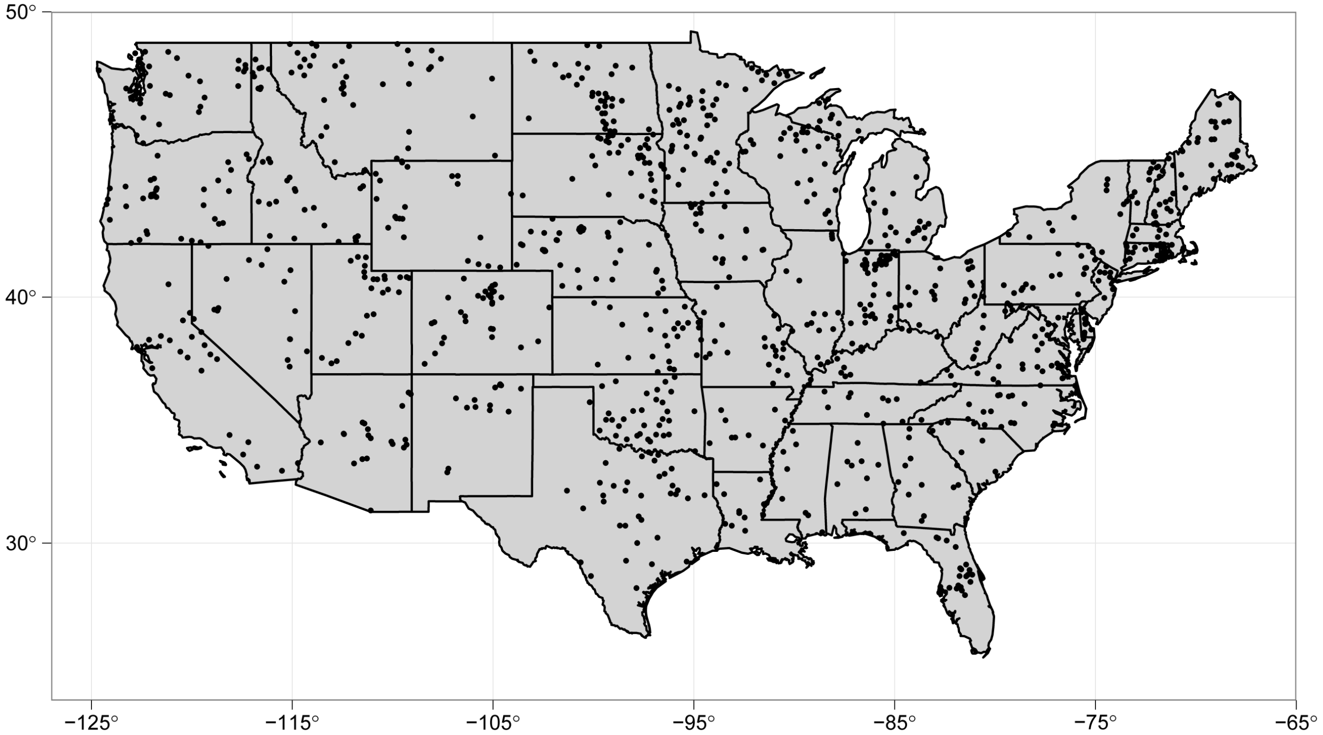

Figure 2. Distribution of lakes. Distribution of lakes that were sampled for the 2007 U.S. National Lakes Assessment.
doi:10.1371/journal.pone.0061623.g002

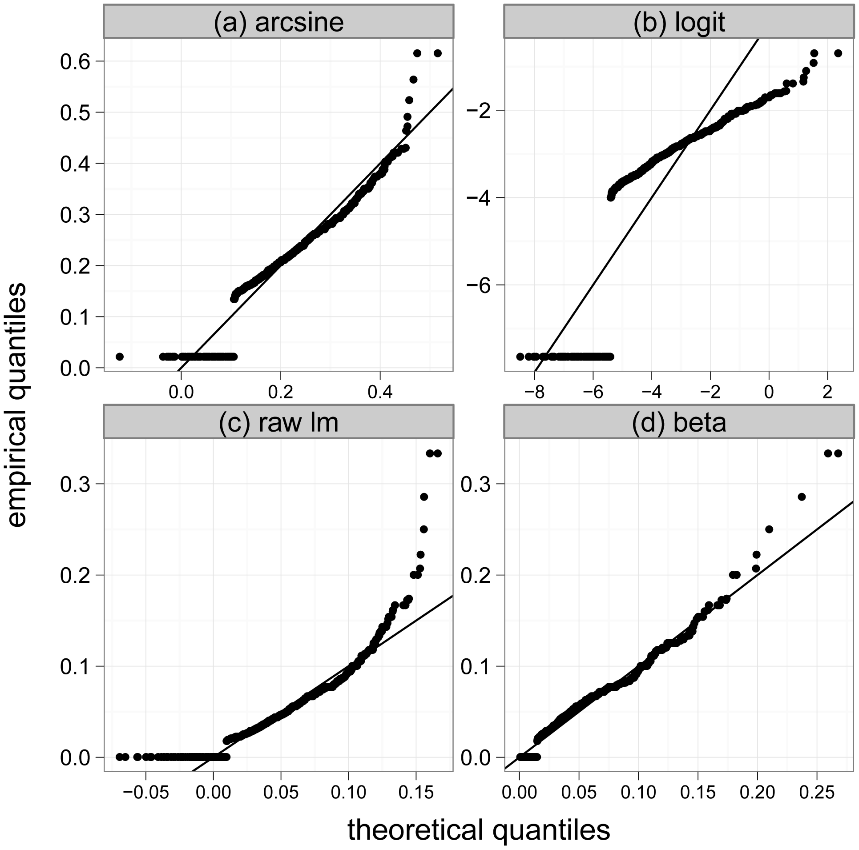

Figure 3. Normal quantile-quantile plots. Normal quantile-quantile plots of arcsine-square-root-transformed ("arcsine"), logit-transformed ("logit") and untransformed ("raw lm") EPHEptax values (panels (a) - (c)). Panel (d) shows a beta quantile-quantile plot using the untransformed EPHEptax values. It is seen that EPHEptax is best approximated by a beta distributed random variable.
doi:10.1371/journal.pone.0061623.g003

details). On the other hand, it is often observed that overdispersion depends on the values of one or more predictor variables [17]. In the context of a beta regression model, this implies that ϕ is not constant but needs to be regressed to the predictor variables. This issue makes variable selection even more complicated because analysts need to identify the predictor variables that affect ϕ.

The aim of this paper is to extend the classical framework and to develop a statistical methodology for beta regression that addresses the aforementioned issues. To this purpose, we develop an estimation technique called *boosted beta regression*. Following the approach by Ferrari & Cribari-Neto [6], we use the beta regression framework to account for the fact that responses are bounded by $(0,1)$. To avoid overfitting the data, however, we do not use classical maximum likelihood estimation but focus on a recently developed algorithm called gamboostLSS [18]. GamboostLSS is a **boost**ing method to fit **g**eneralized **a**dditive **m**odels for **L**ocation, **S**cale and **S**hape (GAMLSS, [19]). The GAMLSS class includes beta regression as a special case and constitutes a flexible class of regression models that allow for modeling multiple parameters of the response distribution (not only the conditional mean of Y as in classical regression). The gamboostLSS technique has a built-in mechanism for variable selection, so that the method can be conveniently used to address the selection of predictor variables in beta regression models. Specifically, this approach avoids the use of heuristic variable selection techniques that are often biased and unstable [20,21]. Because gamboostLSS is based on the gradient boosting framework [22,23], boosted beta regression additionally results in a prediction-optimized model that is suitable for estimating future or unsurveyed response values. Still, the method preserves the structure of the classical beta regression model and thus provides a meaningful interpretation of predictor-response relationships. Furthermore, by using spline modeling, boosted beta regression allows for incorporating nonlinear predictor-response relationships and spatial information even if the functional forms of the relationships are unknown (cf. [15,24]).

To illustrate our method, we use data collected during the 2007 U.S.A. National Lakes Assessment (NLA) Survey [25]. The 2007 U.S.A. NLA is an example of ecological research that often involves the analysis of percentages: the assessment of aquatic biological health. In these studies, percentages of the biological community, often those deemed intolerant or tolerant to stressors, are used as indicators of stream or lake biological condition [26] and are often related to predictor variables such as water chemistry (temperature, dissolved oxygen, pH) and geographical information (site elevation, size of basin area, ecoregion). As response variable for our comparative analysis of modeling approaches we focus on the percentage of benthic macroinvertebrate taxa collected that are in the order Ephemeroptera (mayflies, here denoted as EPHEptax). Ephemeroptera are taxa sensitive to anthropogenic disturbance and are therefore often used to evaluate stream health [27,28]. As will be demonstrated in the results section of the paper, analyzing EPHEptax suggests that beta regression outperforms other approaches in terms of both model fit and prediction accuracy. Hence, by applying boosted beta regression to the 2007 NLA data, this paper builds directly on

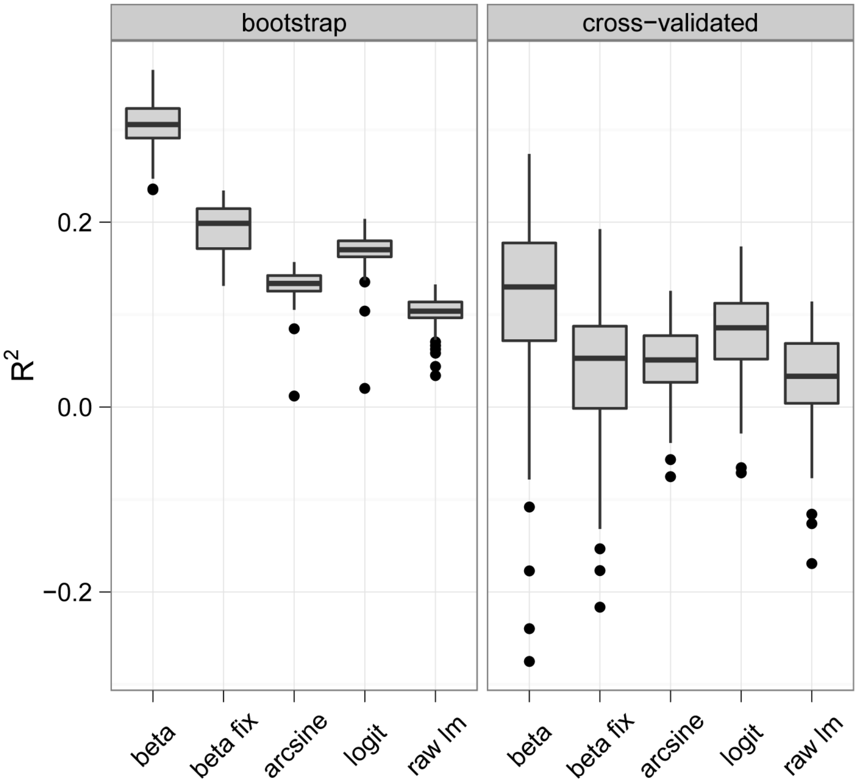

Figure 4. Boxplots of R^2 values. Analysis of the NLA Data. The figure contains boxplots of R^2 values obtained from the 100 bootstrap samples (left panel) and from the 100 sets of out-of-bootstrap observations (right panel).
doi:10.1371/journal.pone.0061623.g004

the modeling approaches of Warton & Hui [2], who argued that the arcsine square root transformation should no longer be applied to analyze percentage outcomes in ecology.

The rest of the paper is organized as follows: In the next section, boosted beta regression is presented in detail, along with a description of the classical beta regression and gamboostLSS approaches. Additionally, we briefly review the arcsine square root and logit transformation approaches and discuss their limitations when used for modeling percentage outcomes. The characteristics of boosted beta regression are demonstrated in the results section of the paper, where the new method is benchmarked against a number of alternative regression models. Using the NLA data, we further show how to apply the new method to derive an easy-to-interpret regression model for the EPHEptax response. A summary and discussion of the main findings of the paper is given in the final section of the paper. Technical details on boosted beta regression are presented in the Supporting Information of the paper.

Methods

Transformation Models for Percentage Outcomes

In this subsection, we briefly review transformation models for percentage outcomes. This model class comprises both classical OLS regression and OLS regression with arcsine-square-root-transformed response. Transformation models are based on the model equation

$$h(Y) = \mathbf{X}^T\beta + \epsilon, \qquad (1)$$

where $Y \in [0,1]$ denotes the percentage outcome, $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_p)$ is a vector of predictor variables, β is an unknown vector of coefficients, ϵ is a normally distributed noise variable with zero mean and constant variance, and h is the transformation function. Typical examples of h include the identity function $h(Y) = Y$ (leading to the classical OLS regression model), the arcsine square root transformation, and the logit transformation $h(Y) = \log(Y/(1-Y))$. Estimates of β are obtained by applying OLS regression to the transformed data.

As noted in [9], two problems arise if classical OLS regression is used to fit model (1): First, because percentages are bounded by the interval $[0,1]$ while the predictor $\mathbf{X}^T\beta$ is not, the expectation of Y conditional on $\mathbf{X}$ must be nonlinear. This is contradictory to the classical OLS assumption $\mathbb{E}(Y|\mathbf{X}) = \mathbf{X}^T\beta$ with $\mathbb{E}(Y|\mathbf{X})$ being linear in $\mathbf{X}$. Second, the variance of a percentage response is not constant but will approach zero near the boundary points 0 and 1 ("heteroscedasticity"). This is contradictory to the homoscedasticity assumption made in classical OLS regression (where $\text{Var}(Y|\mathbf{X})$ is assumed to be constant for all $\mathbf{X}$). Violations of the linearity and homocedasticity assumptions result in biased OLS estimates and hypothesis tests.

To overcome the problems with classical OLS regression, it is a common strategy to transform Y using the arcsine square root function and to carry out OLS regression using the transformed data. Applying this strategy can be justified theoretically by the fact that the arcsine square root transformation leads to asymptotic homoscedasticity in situations where Y is binomial [2]. It has been argued, however, that the approximation is often poor, especially near the boundary values 0 and 1. Also, there is no specific reason for applying the arcsine square root transformation in situations where the response is non-binomial. In the latter cases, it has been

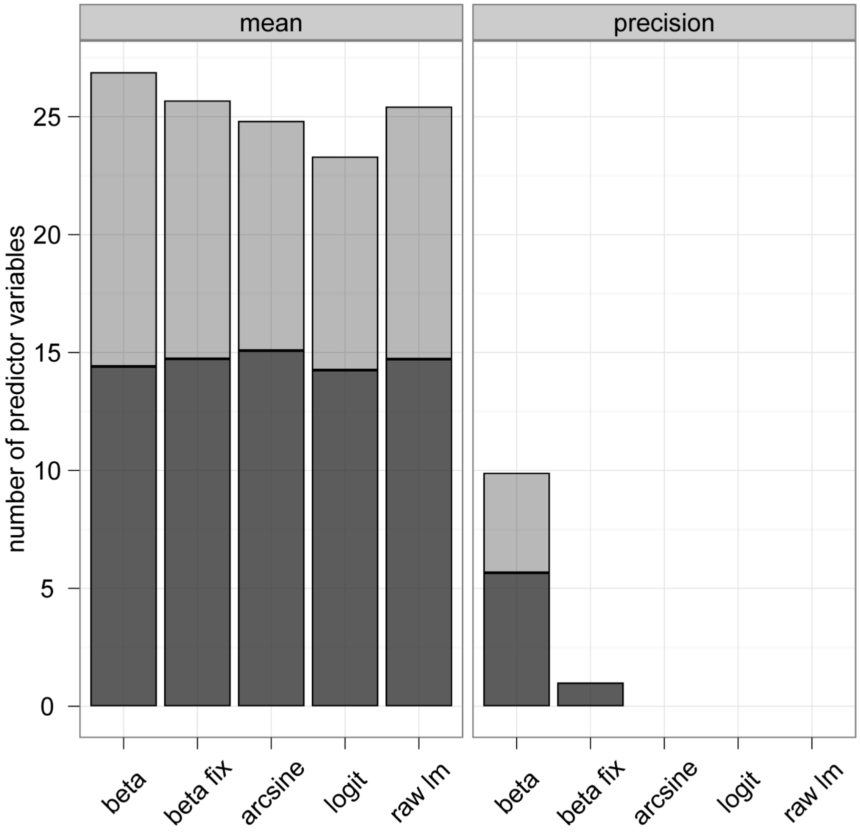

Figure 5. Number of selected predictor variables. Analysis of the NLA Data. The two panels contain the number of selected predictor variables (averaged over 100 bootstrap samples) for various modeling approaches. Dark grey bars represent linear effects, light grey bars represent non-linear effects. In case of beta regression with fixed precision parameter ("beta fix"), the precision model contains only one predictor (namely, the intercept). doi:10.1371/journal.pone.0061623.g005

suggested to fit an OLS regression model with logit-transformed response variable [2].

Regardless of the choice of the transformation function, a major problem of transformation models remains: Unless the identity transformation $h(Y)=Y$ is used, transformation models cannot be interpreted in terms of the conditional mean $\mathbb{E}(Y|\mathbf{X})=\mathbf{X}^T\beta$ (as we would expect from any unbiased regression model). Instead, interpretation is only possible in terms of the *transformed mean* $\mathbb{E}(h(Y)|\mathbf{X})=\mathbf{X}^T\beta$. This makes an appropriate quantification of predictor-response relationships difficult [10].

Beta Regression for Percentage Outcomes

To overcome the problems and limitations discussed in the previous subsection, Ferrari & Cribari-Neto [6] introduced beta regression for proportions and percentage outcomes. In this subsection, we outline the main characteristics of the classical beta regression method.

In the following, we assume that $Y\in(0,1)$ follows a beta distribution with density

$$\varphi(y,\mu,\phi)=\frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)}\;y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}\;,\,y\in(0,1)\;,\quad(2)$$

where $\mu\in(0,1)$ is the mean of Y and $\phi>0$ is a precision parameter. The variance of Y is given by $\mu(1-\mu)/(1+\phi)$ (see [6]). Hence the variance of a beta-distributed random variable is a scaled version of the binomial variance $\mu(1-\mu)$. The precision parameter ϕ generally allows for a wide range of shapes for the density (2) (see Figure 1). Note that (2) assumes y to be strictly larger than 0 and strictly smaller than 1. In applications where Y may assume the boundary values 0 and 1, it is common practice to replace y by $(y\cdot(n-1)+0.5)/n$, where n is the sample size [10,29].

To relate the conditional mean $\mu_x:=\mathbb{E}(Y|\mathbf{X})$ to the predictor variables, the classical beta regression model assumes a predictor-response relationship given by

$$g(\mu_x)=\mathbf{X}^T\beta\;,\qquad(3)$$

where g is an invertible link function. Estimation of β is accomplished using maximum likelihood (ML) estimation, which is consistent and asymptotically efficient for β [6]. Although, in principle, many types of link functions are possible, we focus on the logit transformation $g(\mu_x)=\log(\mu_x/(1-\mu_x))$ in this paper. Apart from being a suitable link function for proportions [30], the logit link has the nice property that it is interpretable in terms of the odds ratio. Consider, for example, the EPHEptax response discussed in the introduction and suppose that the i-th predictor variable $\mathbf{X}_i$ of a beta regression model for EPHEptax is increased by one unit. Then the odds of the proportion of the assemblage richness as Ephemeroptera increases by the factor $\exp(\beta_i)$, where β_i is the regression coefficient for $\mathbf{X}_i$ [6]. This interpretation is exactly the same as the classical interpretation of a logistic regression model.

In contrast to logistic regression, however, the conditional variance $\sigma_x^2:=Var(Y|\mathbf{X})$ of a beta regression model is not restricted to $\mu_x(1-\mu_x)$ but is of the more flexible form

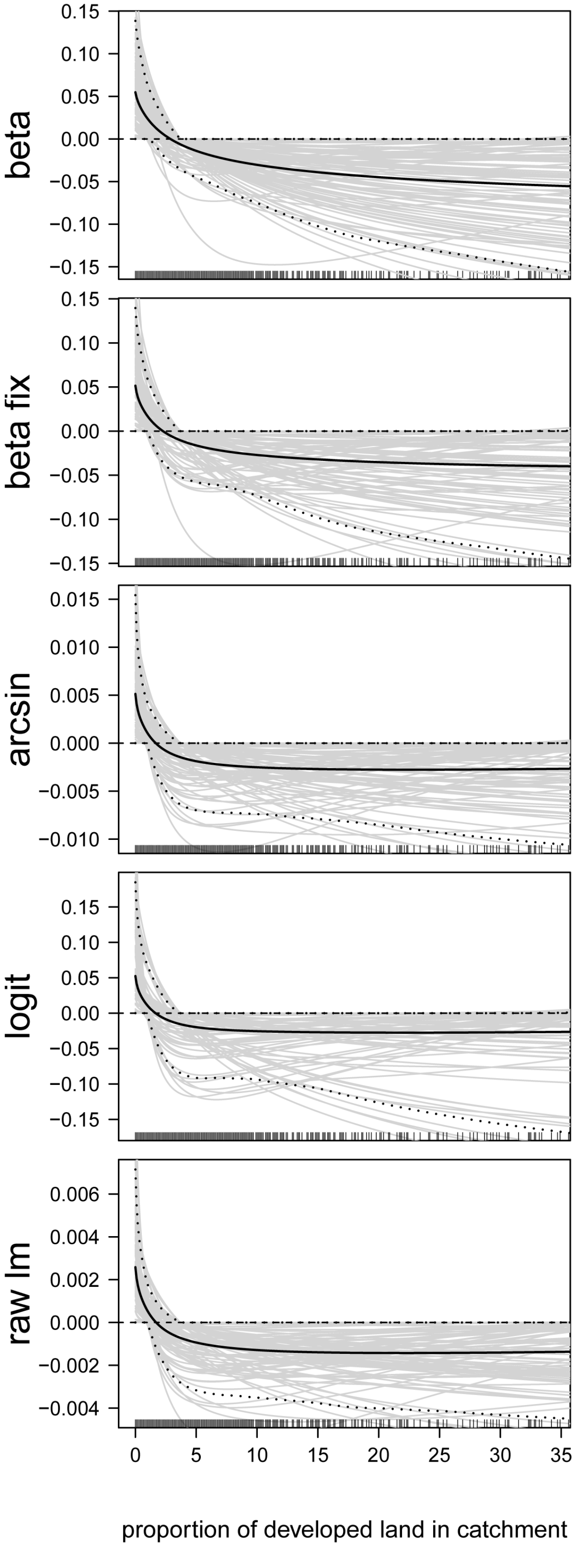

Figure 6. Function estimates for the proportion of developed land in catchment. Analysis of the NLA Data. The five panels contain the function estimates for the proportion of developed land in catchment (computed from 100 bootstrap samples). In case of beta regression, estimates present the effects of the proportion of developed land in catchment on the mean parameter μ. Black lines correspond to the mean and the 0.05 and 0.95 quantiles of the function estimates. For reasons of interpretability, the range of the x-axes was restricted to the lower 95% of the sample values.
doi:10.1371/journal.pone.0061623.g006

$\mu_x(1-\mu_x)/(1+\phi)$. Consequently, the model allows for variances that are larger than those expected by the binomial model ("overdispersion"). Also note that, in contrast to logistic regression, beta regression does not depend on the binomial counts x and N (as defined in the introduction). The method is therefore suitable for both binomial and non-binomial percentage outcomes.

In cases where overdispersion depends on the values of the predictor variables, it is further possible to extend the beta regression model by regressing the precision parameter $\phi \equiv \phi(\mathbf{X})$ to the predictor variables $\mathbf{X}$. This is accomplished by assuming the relationship

$$\tilde{g}(\phi(\mathbf{X})) = \mathbf{X}^T\gamma \qquad (4)$$

with link function $\tilde{g}$ and parameter vector γ ("variable dispersion beta regression", [17]). A common choice for $\tilde{g}$ is the log link $\tilde{g}(\phi(\mathbf{X})) = \log(\phi(\mathbf{X}))$, which will also be considered in this paper. Estimates of γ are again obtained by using maximum likelihood estimation [17]. Efficient implementations of this "classical" beta regression method are provided by the R add-on packages **betareg** [10] and **gamlss** [31].

GamboostLSS

Following its introduction by Ferrari & Cribari-Neto [6], beta regression has been used to model percentage outcomes in various fields of research. However, the classical version of the method still has several shortcomings. For example, scientific databases often contain a large number of possible predictor variables (relative to the sample size). It is well known that classical maximum likelihood estimators suffer from large variances in this case. This problem leads to overfitting and therefore to a decreased prediction accuracy of the classical beta regression model. To avoid overfitting (and also to improve the interpretability of the model), it is desirable to carry out variable selection, i.e., to include only the most "important" predictors in the model. Although there exist many "classical" techniques for variable selection (e.g., stepwise variable selection based on information criteria or hypothesis tests), these methods are known to be unreliable and require the model to be fitted multiple times [21].

To address the issue of variable selection in beta regression models, we propose a new fitting method called *boosted beta regression*. Boosted beta regression is based on the gamboostLSS algorithm, which has been introduced in [18] as a boosting method for generalized additive models for location, scale and shape (GAMLSS, [19]). Because beta regression is a special case of GAMLSS, the theory presented in [18] applies: Similar to ML estimation, gamboostLSS uses the log-likelihood function of Y as optimization criterion for deriving a regression model. In contrast to the original beta regression method proposed in [6], however, gamboostLSS is not based on (quasi-)Newton algorithms but on the gradient boosting framework [23,32] (hence the name "boosted" beta regression). Broadly speaking, gamboostLSS uses gradient descent techniques to optimize arbitrary differentiable

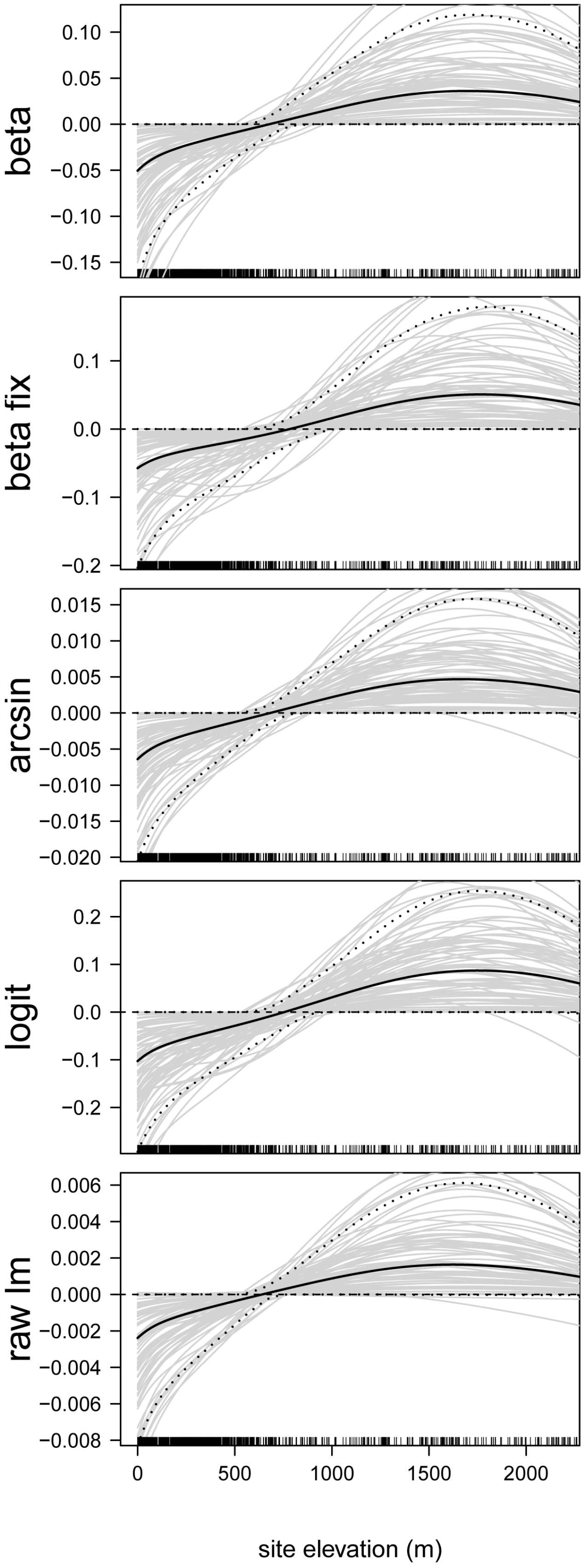

Figure 7. Function estimates for the site elevation. Analysis of the NLA Data. The five panels contain the function estimates for the site elevation (computed from 100 bootstrap samples). In case of beta regression, estimates present the effects of the site elevation on the mean parameter μ. Black lines correspond to the mean and the 0.05 and 0.95 quantiles of the function estimates. For reasons of interpretability, the range of the x-axes was restricted to the lower 95% of the sample values.

doi:10.1371/journal.pone.0061623.g007

objective functions (here, the beta log-likelihood) in an iterative fashion.

The most important feature of gamboostLSS is its ability to carry out variable selection during the fitting process. This is accomplished by (a) assessing the individual fits of each predictor variable, and by (b) updating only the coefficient of the best-fitting predictor variable in each iteration. Also, when using gamboostLSS to fit a beta regression model, variable selection is carried out successively for both the mean model (3) and the precision model (4). After a finite number of iterations, the algorithm is stopped, so that the final model only contains the subset of best-fit predictor variables. A schematic overview of boosted beta regression is as follows:

1. Set the initial values for β and γ to zero and start iterating.
2. In each iteration….

(a)…keep the current value of the estimate of γ fixed and consider the mean model (3) only. Select the predictor variable $\mathbf{X}_i^*$ leading to the best improvement of the beta log-likelihood and update the estimate of the coefficient β_i^* corresponding to $\mathbf{X}_i^*$.

(b)…keep the current value of the estimate of β fixed and consider the precision model (4) only. Select the predictor variable $\mathbf{X}_j^*$ leading to the best improvement of the beta log-likelihood and update the estimate of the coefficient γ_j^* corresponding to $\mathbf{X}_j^*$.

3. Repeat steps (a) and (b) until the stopping iteration of the algorithm (denoted by m_{stop}) is reached.

Analogously to the original gamboostLSS algorithm described in [18], boosted beta regression uses the gradient of the beta log-likelihood to compute the estimates of β_i^* and γ_j^* in steps (a) and (b). A technical description of boosted beta regression is given in the Supporting Information.

The variable selection mechanism in (a) and (b) is fundamentally different from the Newton-type method proposed in [6], which updates the *whole* vectors β and γ in each iteration. Specifically, the initial model in step 1 (with $\beta = \gamma = 0$) does not depend on any of the predictor variables. As a consequence, only the predictor variables selected in (a) and (b) will contribute to the final model fit. Because variable selection and parameter estimation are carried out simultaneously in step 2 (cf. [18]), boosted beta regression results in a variable selection process that is more stable than classical methods such as stepwise selection.

An important question is how to choose the stopping iteration of gamboostLSS. Usually, the stopping iteration of a boosting algorithm is chosen such that prediction accuracy of the model becomes highest [23]. For gamboostLSS, this is accomplished by using cross-validation techniques [18]. Note that it is possible to increase flexiblity of the algorithm by using two different stopping iterations for the mean and the precision models (see [32] for details). Because the benefits of a two-dimensional stopping strategy are usually small [18], we will not consider this method in our numerical studies.

Nonlinear Predictor-Response Relationships

An attractive feature of gradient boosting (and therefore also of boosted beta regression) is that the linear predictors $\mathbf{X}^T\beta$ and $\mathbf{X}^T\gamma$

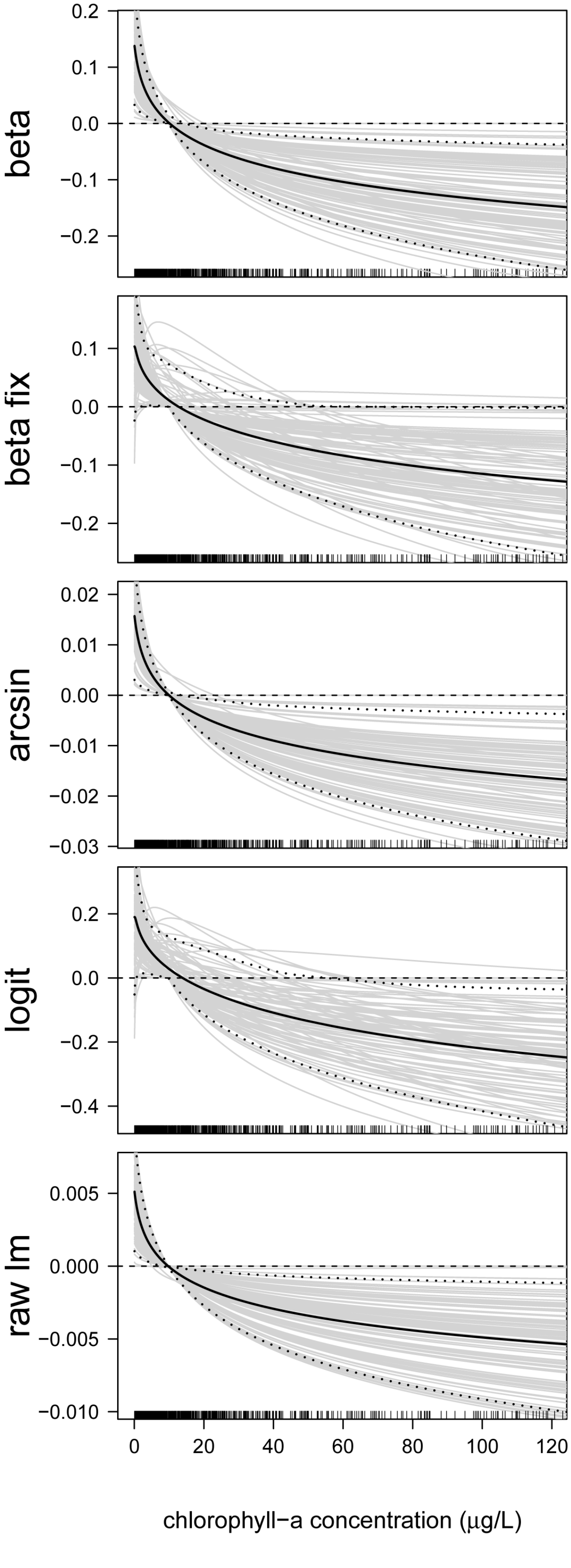

Figure 8. Function estimates for the chlorophyll- *a* concentration. Analysis of the NLA Data. The five panels contain the function estimates for the chlorophyll- *a* concentration (computed from 100 bootstrap samples). In case of beta regression, estimates present the effects of the chlorophyll- *a* concentration on the mean parameter μ. Black lines correspond to the mean and the 0.05 and 0.95 quantiles of the function estimates. For reasons of interpretability, the range of the x-axes was restricted to the lower 95% of the sample values.
doi:10.1371/journal.pone.0061623.g008

can be easily replaced by more flexible predictors of the form.

$$g(\mathbb{E}(Y|\mathbf{X})) = f_1(\mathbf{X}_1) + \ldots + f_p(\mathbf{X}_p) =: \eta_\mu , \qquad (5)$$

$$\tilde{g}(\phi(\mathbf{X})) = \tilde{f}_1(\mathbf{X}_1) + \ldots + \tilde{f}_p(\mathbf{X}_p) =: \eta_\phi , \qquad (6)$$

where f_i and $\tilde{f}_j$, $i,j = 1,\ldots,p$, are arbitrary differentiable nonlinear functions. Analogously to the well-established generalized additive modeling and GAMLSS approaches [19,33,34], equations (5) and (6) extend classical beta regression by incorporating nonlinear predictor-response relationships. Note that the forms of the functions f_i and $\tilde{f}_j$ are determined automatically by gamboostLSS, where estimation of f_i and $\tilde{f}_j$ is accomplished using penalized regression splines ("P-splines", [35]). This approach is a major difference to the method by Simas et al. [17], who considered parametric nonlinear functions with pre-specified functional forms. P-spline estimates allow for easy inspection and easy visualization of predictor effects. In addition, by decomposing P-spline estimators into a linear part and a nonlinear part, it is possible to automatically select among linear and smooth nonlinear modeling alternatives for the same predictor variables [24]. Consequently, if competing modeling alternatives are used as base-learners in boosted beta regression, the model will typically contain a subset of predictors with nonlinear predictor-response relationships and another subset with smooth nonlinear relationships. Technical details on P-spline base-learners are given in [24] and [36].

Concerning the analysis of the NLA data, we proceed as follows: The functions f_i and $\tilde{f}_j$ corresponding to *continuous predictors* are modeled using one-dimensional P-spline estimators [35,36]. Moreover, we follow the strategy by Kneib et al. [24] and decompose P-spline base-learners into a set of linear base-learners and another set of smooth nonlinear base-learners. This strategy results, for example, in linear effects of the mean site depth and in nonlinear effects of the total nitrogen concentration on EPHEptax (see the next section for details). *Categorical predictors* (such as Köppen-Geiger climate regions) are modeled using dummy coded binary variables. Hence the resulting estimates for categorical predictors have the same interpretation as in classical linear models.

To account for spatial dependency between neighboring lakes, we specify smooth surface functions quantifying *spatial predictor effects*. These functions depend on the coordinates of the site locations and are added to the other functions specified in (5) and (6) (cf. [15,24]). To estimate the shapes of the surface functions, we use P-spline tensor product surfaces depending on the NAD83 coordinates of the lakes. Thus, denoting the longitude and latitude coordinates by $\mathbf{X}_{\mathrm{Lon}}$ and $\mathbf{X}_{\mathrm{Lat}}$, respectively, the spatial effects become smooth surfaces $f_{\mathrm{sp,mean}}(\mathbf{X}_{\mathrm{Lon}},\mathbf{X}_{\mathrm{Lat}})$ and $\tilde{f}_{\mathrm{sp,precision}}(\mathbf{X}_{\mathrm{Lon}},\mathbf{X}_{\mathrm{Lat}})$ depending on the bivariate "predictor" variable $(\mathbf{X}_{\mathrm{Lon}},\mathbf{X}_{\mathrm{Lat}})$. Note that $f_{\mathrm{sp,mean}}$ and $\tilde{f}_{\mathrm{sp,precision}}$ can be conveniently interpreted as realizations of a spatially correlated

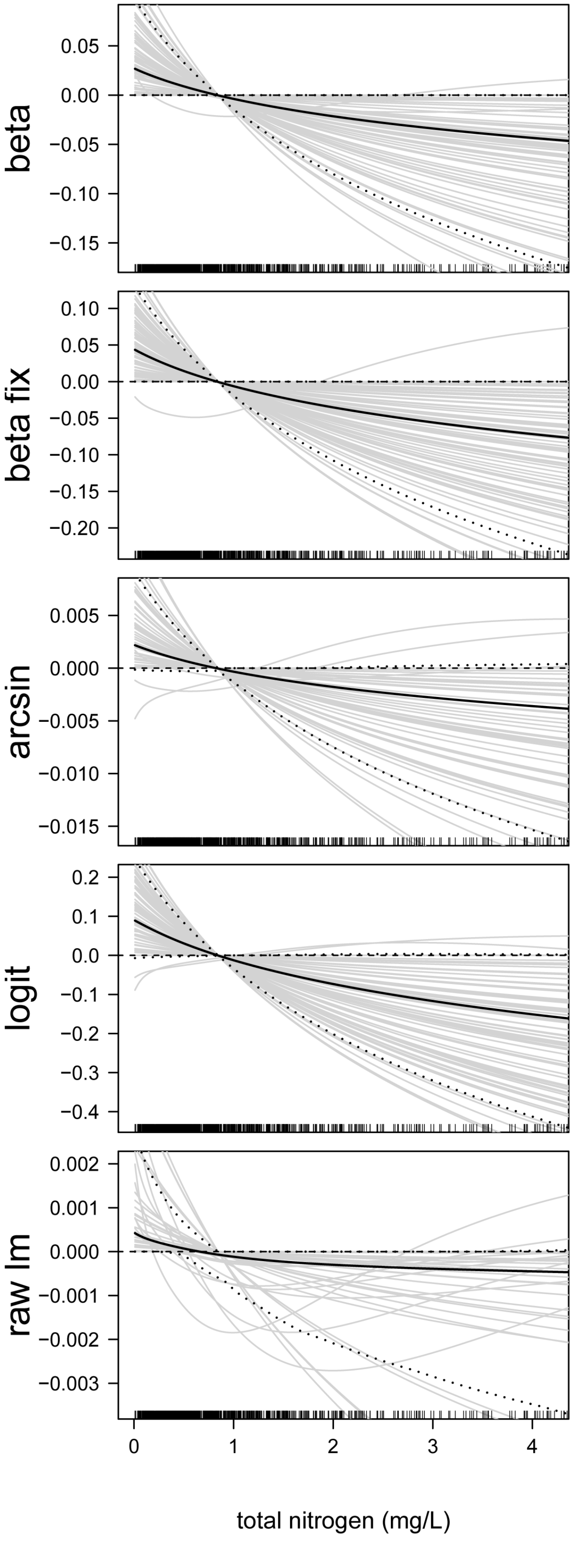

Figure 9. Function estimates for the total nitrogen concentration. Analysis of the NLA Data. The five panels contain the function estimates for the total nitrogen concentration (computed from 100 bootstrap samples). In case of beta regression, estimates present the effects of the total nitrogen concentration on the mean parameter μ. Black lines correspond to the mean and the 0.05 and 0.95 quantiles of the function estimates. For reasons of interpretability, the range of the x-axes was restricted to the lower 95% of the sample values.
doi:10.1371/journal.pone.0061623.g009

stochastic process [15]. Again, we refer to Kneib et al. [24] for technical details.

Results

In the following we will use boosted beta regression to model biological condition in lakes in the conterminous U.S. The outcome considered in our study is the percentage of benthic macroinvertebrate taxa in the order Ephemeroptera (EPHEptax). In addition to analyzing boosted beta regression, we compare the new method to conventional approaches such as OLS regression with transformed response. The first subsection starts with a description of the study design and the NLA database. Statistical analysis results are presented in the second subsection.

The NLA Database

Statistical analysis is based on data from the 2007 U.S. National Lakes Assessment program (NLA), during which 1,157 lakes were sampled in the summer from across the conterminous U.S. (see Figure 2).

Littoral zone sampling consisted of ten randomly selected quadrates around each lake littoral zone that were combined into a single sample. In each sample, benthic macroinvertebrates were collected and physical habitat assessed. Habitat condition was measured by visual estimates of riparian vegetation condition, shoreline substrate (at the water's edge), fish cover, aquatic macrophytes, and littoral bottom substrate in the samples. In addition, human disturbance or presence was estimated in each sample by identifying human activities (e.g., docks, roads, buildings, etc.) in the water or in adjacent riparian areas. Sampling at the deepest point of the lake (index site) included all other biological and chemical measures. Water column profiles of temperature, dissolved oxygen, conductivity, and pH were taken using a multi-probe sonde.

The NLA database also contains estimates of lake drainage conditions (e.g., land use/land cover, precipitation, elevation). To provide estimates of lake connectivity and colonization sources, we calculated the number and total surface areal coverage of other lakes (from NHDplus, [37]) within a 1km and 20km radius of the sampling location. We further calculated geographic distance to and surface area of the nearest lake and nearest large lake (i.e., $>1\text{km}^2$ surface area) within the NHDplus data set. Finally, we classified sites into major drainage basin-climate regions by intersecting the NHDplus HUC2 to which a site resided and the main climates of the Köppen-Geiger Climate Classification [38]. This resulted in 37 basin-climate regions. Four regions had too few sites ($<$ 6 sites) and were combined into nearby regions leaving a total of 33 drainage basin-climate regions.

Statistical analysis was based on a sample of 994 lakes that contained no missing values in any of the predictor variables. Altogether, 78 predictor variables were used for statistical analysis. Predictors with a highly right-skewed distribution were log transformed before fitting models for EPHEptax. The full list of predictor variables is given in the Supporting Information.

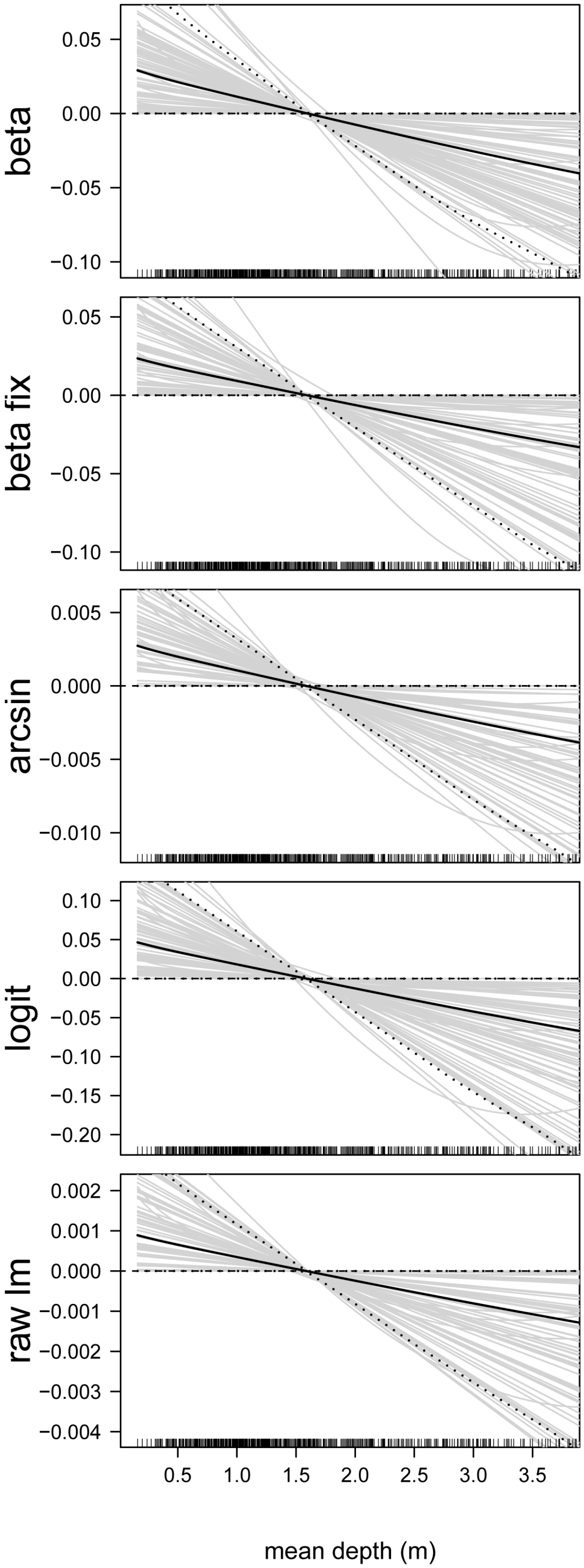

Figure 10. Function estimates for mean site depth. Analysis of the NLA Data. The five panels contain the function estimates for the mean depth at the sites (computed from 100 bootstrap samples). In case of beta regression, estimates present the effects of the depth on the mean parameter μ. Black lines correspond to the mean and the 0.05 and 0.95 quantiles of the function estimates. For reasons of interpretability, the range of the x-axes was restricted to the lower 95% of the sample values.
doi:10.1371/journal.pone.0061623.g010

Statistical Analysis and Results

In a first step, we used graphical checks to analyze the response transformations discussed in the methods section of the paper. Figure 3 presents normal quantile-quantile plots for the arcsine-transformed, the logit-transformed and for the untransformed EPHEptax values (panels (a) to (c)). Neither transformation worked well, as the transformed EPHEptax values clearly do not follow a normal distribution. In addition, the inclusion of lakes with zero percentages seemed to be problematic because their values in the quantile-quantile plots did not match well with EPHEptax values that were larger than zero (see the horizontal accumulation of points in panels (a) to (c)). In contrast, EPHEptax was well approximated by a beta distributed random variable (Figure 3(d)). This result suggested that boosted beta regression is an adequate method for modeling EPHEptax.

We next investigated the performance of boosted beta regression by comparing the new method to classical response transformation models. In a first step, we generated 100 random samples of size n from the NLA data by drawing observations with replacement (bootstrapping, [39]). The 100 bootstrap samples can be interpreted as independent random samples from the empirical distribution of $(Y,\mathbf{X})$ and can therefore be used to compute empirical confidence intervals for effect estimates and performance measures. Next, boosted beta regression was applied to each of the 100 bootstrap samples. This strategy resulted in 100 model fits for EPHEptax. To compare boosted beta regression to other modeling approaches, we additionally fitted an arcsine square root transformed model, a logit transformed model, and an OLS model without response transformation to the same 100 bootstrap samples. In addition, we fitted a beta regression model with fixed precision parameter ϕ. To allow for a fair comparison of the modeling approaches, we applied the same gradient boosting strategy for the latter models as the one used for boosted beta regression. Specifically, we allowed for the same nonlinear predictor-response relationships as those discussed in the methods section of the paper. The stopping iterations for the models were determined by applying 25-fold bootstrap cross-validation [40] to the 100 samples. All computations were carried out using the **mboost** and **gamboostLSS** packages of the statistical software R [41,42].

Analysis of Model Performance

To evaluate the overall performance of the modeling approaches, we calculated the generalized R^2 criterion [43] from the 100 model fits. The R^2 criterion relates the log-likelihood of a fitted model to the corresponding log-likelihood of a "null" model containing no predictor variables. It can therefore be used as a goodness-of-fit criterion that measures the improvement of the fitted model over the null model. Boosted beta regression explained the data best, while the transformation models performed worse than beta regression on average (Figure 4, left panel). As expected, OLS regression was the worst model in terms of goodness-of-fit. Also, beta regression with a fixed precision

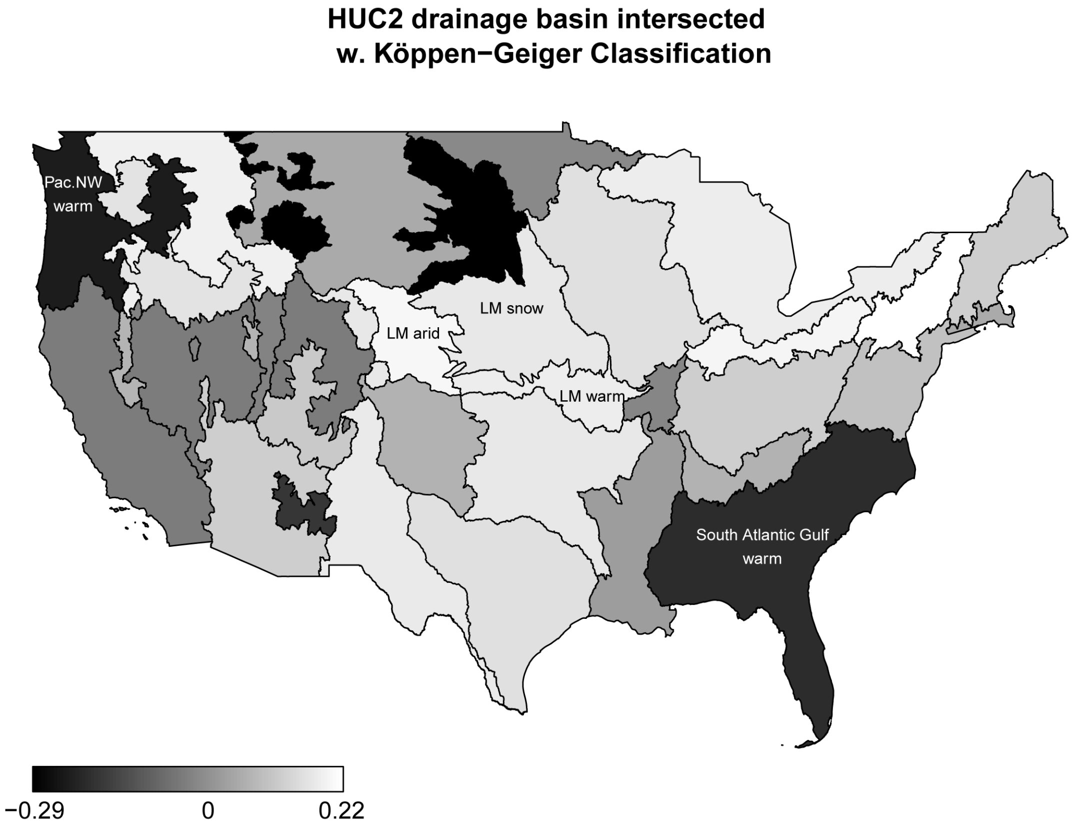

Figure 11. Effects of basin-climate regions. Estimated effects of basin-climate regions on EPHEptax, as obtained from applying boosted beta regression to 100 bootstrap samples of the NLA data. The figure presents median effect estimates for the mean parameter μ computed from the 100 model fits (LM = Lower Missouri).
doi:10.1371/journal.pone.0061623.g011

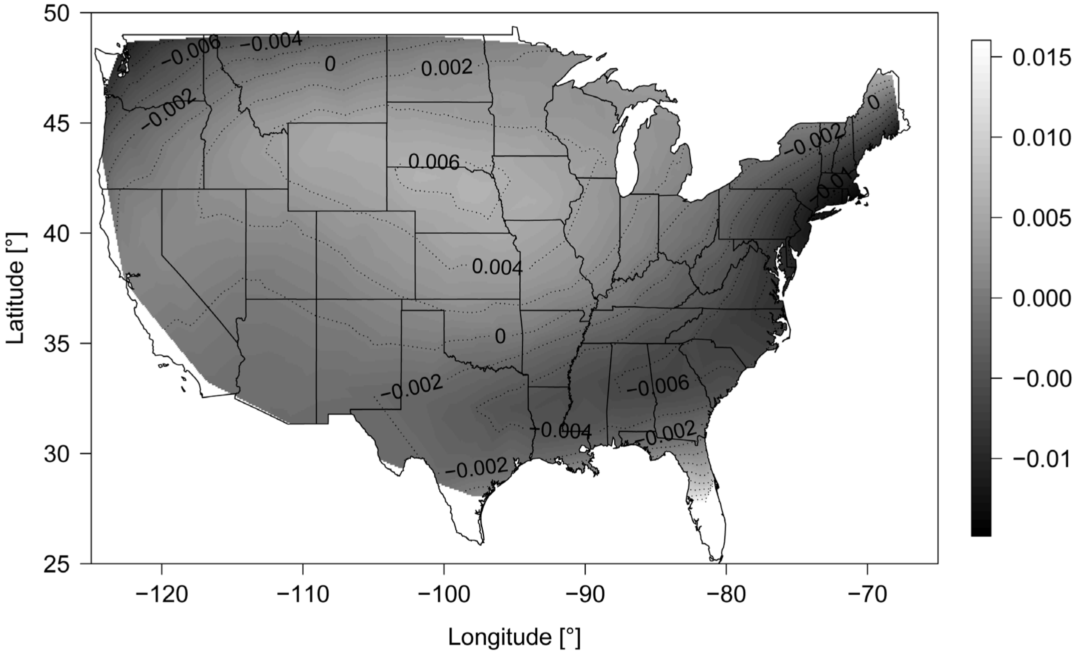

Figure 12. Estimated spatial surface function. Estimated spatial surface function $f_{\mathrm{sp,mean}}(\mathbf{X}_{\mathrm{Lon}},\mathbf{X}_{\mathrm{Lat}})$ for the mean parameter μ in boosted beta regression. The figure presents the median spatial surface obtained from the 100 bootstrap samples.
doi:10.1371/journal.pone.0061623.g012

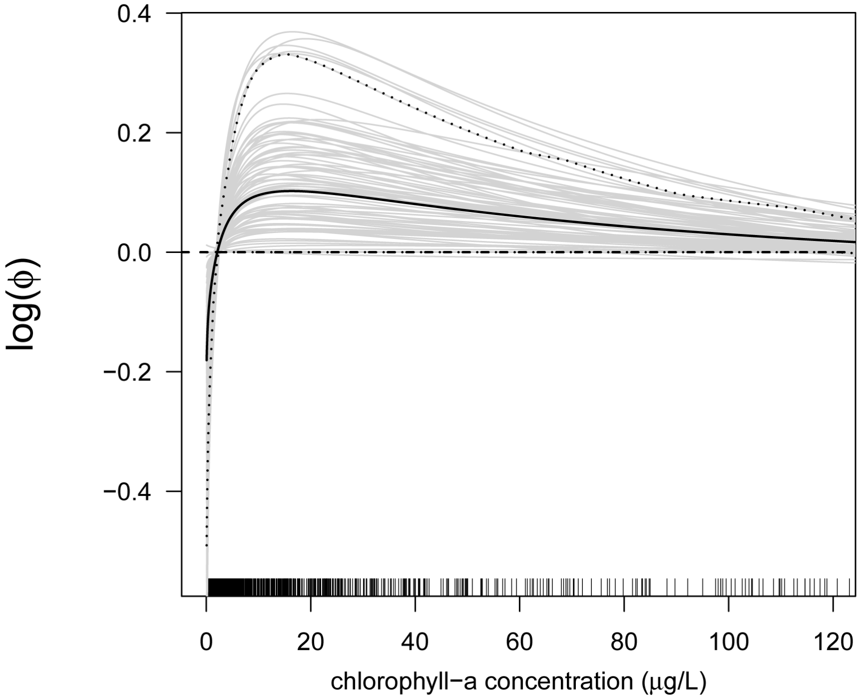

Figure 13. Effect of chlorophyll- *a* concentration on the precision parameter. Analysis of the NLA Data. The figure contains the estimated effect of the chlorophyll- *a* concentration on the logarithm of the precision parameter ϕ (computed from 100 bootstrap samples). Black lines correspond to the mean and the 0.05 and 0.95 quantiles of the function estimates. For reasons of interpretability, the range of the x-axis was restricted to the lower 95% of the sample values.
doi:10.1371/journal.pone.0061623.g013

parameter ϕ resulted in a worse model fit than boosted beta regression with a flexible precision parameter.

In addition to evaluating the goodness-of-fit of the models, we investigated the *predictive* performance of the modeling approaches. This issue is of interest in many ecological applications, where regression models are often used to obtain predictions of future or unseen response values. In case of the NLA data, for example, a boosted beta regression model could be used to predict the EPHEptax values of unsurveyed lakes based on values of the predictor variables measured at the site locations. Note that the R^2 values presented in the left panel of Figure 4 cannot be used to evaluate the predictive performance of the models, because the same data were used to fit the models and to compute the R^2 values. The latter values would therefore be too optimistic for measuring prediction accuracy.

To obtain unbiased estimates of prediction accuracy, we used the models obtained from the 100 bootstrap samples and computed predictions for EPHEptax from the 100 respective sets of out-of-bootstrap observations. In other words, predictions were computed from those observations that were not part of the bootstrap samples and that were therefore not involved in model fitting (bootstrap cross-validation, [40]). The predictions and the true EPHEptax values of the 100 out-of-bootstrap data sets were then used to compute 100 predictive R^2 values.

The cross-validated R^2 values suggest that boosted beta regression leads to the highest predictive R^2 values among the modeling approaches (Figure 4, right panel). Wilcoxon signed rank tests on the differences in R^2 values between boosted beta regression and the other approaches resulted in highly significant results (Bonferroni-adjusted p-values <0.001). Moreover, predictive performance increased when the precision parameter ϕ of a beta regression model was regressed to the predictor variables.

Summarizing the results presented in Figure 4, boosted beta regression outperformed the other modeling approaches for EPHEptax in terms of both goodness-of-fit and prediction accuracy.

Selection Rates of Modeling Approaches

Each modeling approach incorporated approximately 15 linear predictor effects and approximately 10 nonlinear predictor effects on average (Figure 5). Moreover, the percentage of non-linear predictor effects was highest on average in the boosted beta regression model. This result further demonstrated the flexibility of the proposed algorithm.

Analysis of Predictor-Response Relationships

In the next step, we analyzed the estimated effects sizes and the functional forms of the predictor-response relationships. First consider the predictor-response relationships of the continuous predictor variables. By way of example, we present the estimates of the following predictors: proportion of developed land in catchment (Figure 6), site elevation (in m, Figure 7), chlorophyll-*a* concentration (in μg/L, Figure 8), total nitrogen concentration (in mg/L, Figure 9), and mean depth at the sites (in m, Figure 10).

Figures 6, 7, 8, 9, 10 illustrate that all five modeling approaches resulted in very similar estimates of predictor-response relationships. For example, all analyses indicated a negative non-linear relationship between EPHEptax and the proportion of developed land in a basin (Figure 6). All models showed a rapid decrease with EPHEptax up to about 3%, after which the decreasing trend lessened. For example, it is seen from the upper panel of Figure 6 (boosted beta regression model), that the average effect of EPHEptax decreased by approximately 0.05 if the proportion of developed land in a basin increased from 0% to 3%. Consequently, the odds of the proportion of the assemblage as Ephemeroptera decreased by the factor $\exp(-0.05)$ ($\approx 5\%$) in this range. Although the confidence bands in Figure 6 encompass the zero line (and can therefore not be considered "statistically significant"), the negative pattern associated with the amount of developed land in a basin was observed for the majority of the 100 bootstrap samples. Such sensitivity to developed land (i.e., urbanization) has been shown for benthic macroinvertebrates in streams (e.g., [44]) with recent thresholds reported as low as 1.5% to 3.0% [16]. Thus, Ephemeroptera in lakes appear to be as sensitive to urban development as analogous taxa in streams.

The elevation of the sites had an inverted U-shaped effect on EPHEptax (Figure 7), where, for example, low values ranging from 0m to 800m resulted in EPHEptax values that were below average. This "humped-shaped" pattern between diversity and altitude has been previously shown for littoral benthic macroinvertebrates (e.g., [45]) and is likely a result of numerous factors that co-vary with altitude (e.g., temperate) or that affect dispersal [46,47].

Figure 8 suggests that the effect of the chlorophyll- *a* concentration on EPHEptax is distinctly nonlinear, with large values leading to below-average values of EPHEptax. Chlorophyll-*a* is often used to indicate impairment of aquatic systems with high levels indicating eutrophication (e.g., [25]). As such, species richness and diversity of littoral benthic macroinvertebrates declines with chloropyll- *a* [48]. Because Ephemeroptera are sensitive taxa, they may be disproportionately affected by higher chlorophyll- *a* levels. Additionally, high levels of Chlorophyll- *a* reduces mayfly secondary production [49], which would possibly further reduce their presence.

The total nitrogen concentration had a pronounced negative effect on EPHEptax (Figure 9). Total nitrogen is an important environmental factor related to littoral benthic macroinvertebrate community structure [50] and negatively relates to macroinvertebrate diversity [48]. Moreover, loss of Ephemeroptera has been reported in lakes where total nitrogen surpasses a threshold [51].

Average depth at a sampling station had a negative linear effect on EPHEptax (Figure 10). Littoral benthic macroinvertebrates often show a marked effect of depth (e.g., [52,53]). The negative pattern in our study suggests Ephemeroptera prefer shallower habitats in lakes. Note that the variability of the estimates was large for all analyzed models, implying that the uncertainty about the association of the average depth at a sampling station with EPHEptax was large as well.

Next consider the effects of the basin-climate regions (obtained from intersecting the NHDplus HUC2 to which lake sites resided with the main climates of the Köppen-Geiger Climate Classification). Figure 11 shows that basin-climate regions have a relatively strong effect on EPHEptax. Consider, for example, the Lower Missouri/arid, Lower Missouri/snow and Lower Missouri/warm temperate regions (i.e., the Lower Missouri watershed including the Northern Plains and Temperate Plains). The average coefficient estimates of these regions were 0.20, 0.17, and 0.18, respectively, implying that the odds of the proportion of the assemblage as Ephemeroptera increased by the factors $\exp(0.2) \approx 1.22$, $\exp(0.17) \approx 1.19$ and $\exp(0.18) \approx 1.20$, respectively. Conversely, the dark regions in Figure 11 correspond to climate-basin regions with negative effect estimates. For example, in the South Atlantic-Gulf Region/warm temperate region (effect estimate $= -0.20$) and the Pacific Northwest Region/warm temperate region (effect estimate $= -0.23$), the odds of the proportion of the assemblage as Ephemeroptera decreased by the factors $\exp(-0.2) \approx 0.82$ and $\exp(-0.23) \approx 0.79$, respectively. Comparing the effects of the basin-climate regions (Figure 11) to the magnitude of the predictor-response relationships shown in Figures 6, 7, 8, 9, 10, it is seen that basin-climate regions are by far the most important predictors for EPHEptax. Geospatial regions, based on environmentally similar characteristics (e.g., ecoregions), have had mixed success in accounting for variation in benthic macroinvertebrates (see, e.g., [54,55]). Our results, even though we defined regions more broadly than ecoregions, suggest an importance of regional differences in the Ephemeropteran portion of lentic benthic macroinvertebrate assemblages.

Finally consider the estimated spatial surface function $f_{\mathrm{sp,mean}}(\mathbf{X}_{\mathrm{Lon}}, \mathbf{X}_{\mathrm{Lat}})$ for the mean parameter μ of the boosted beta regression model. Figure 12 suggests that effect estimates are below average at the Eastern Coast and in the Pacific Northwest region of the U.S. Conversely, they are above average in the Mid-Western Region. Because there is little systematic variation in Figure 12, these results can possibly be explained by boundary effects that are often observed when using estimators based on P-spline tensor products. Note that $f_{\mathrm{sp,mean}}(\mathbf{X}_{\mathrm{Lon}}, \mathbf{X}_{\mathrm{Lat}})$ corresponds to the realization of a residual stochastic process that cannot be explained by the predictor variables. Alternatively, the variations in Figure 12 and could be due to unmeasured predictors.

By way of example, we also present the effect of the chlorophyll-*a* concentration on the logarithm of the precision parameter ϕ (Figure 13). Because ϕ is inversely related to the variance of EPHEptax (which is given by $\mu(1-\mu)/(1+\phi)$), Figure 13 implies that very low chlorophyll- *a* concentration levels tend to increase the variance of EPHEptax. The variation decreases until levels of $15\mu g/L$ are reached. For chlorophyll- *a* concentration levels larger than $15\mu g/L$, the variance of EPHEptax increases again.

In summary, the results presented in Figures 6, 7, 8, 9, 10 suggest that all five modeling approaches resulted in similar functional patterns. On the other hand, the generalized R^2 values presented in Figure 4 clearly suggest that the magnitude of the predictor effects (and therefore the contribution of each preditor on EPHEptax) is best captured by boosted beta regression.

Discussion

Linear regression with normally distributed errors is arguably the most prominent analysis tool in applied statistics. The popularity of linear regression is based on the fact that random variations in observed data can often be approximated by a normal distribution with constant variance. If the response variable in a regression model is a rate or percentage, however, the normal approximation is no longer appropriate. For this reason, and because the analysis of percentages is an important issue in many fields of research, developing statistically valid analysis tools for percentage data is of high practical interest.

Several approaches to remedy the problems with linear regression have been proposed in the literature. The first approach is to transform the percentage response and to hope that linear regression with the transformed response will result in (approximately) normally distributed errors with constant variance. This is, for example, the rationale of arcsine square root transformation. As shown in the methods section of the paper, however, response transformation models may result in poor model fits because the normal approximation often fails. Based on the results obtained from the NLA data, we agree with Warton & Hui [2] cautioning use of the arcsine square root transformation in ecological research. The second approach to model percentage outcomes is logistic regression [2]. As demonstrated in [9] and [11], logistic regression models can be generalized to deal with overdispersed data and flexible variance structures (e.g., by using beta-binomial and quasi-likelihood models). Note, however, that logistic regression is only appropriate if the response is based on a binomial distribution (with the raw counts x and N being available).

In this paper, we have proposed *boosted beta regression*, which is a flexible alternative to logistic regression and response transformation models. Because beta regression is a generalization of logit regression to situations where the dependent variable is a proportion [29], our modeling approach is appropriate in both the binomial and the non-binomial case. Moreover, if compared to classical estimation techniques for beta regression [17], boosted beta regression has the advantage that nonlinear effects can be estimated without pre-specifying the functional forms of the predictor-response relationships. This implies that not only the mean but also the variance of a beta distributed response variable can be modeled in a highly flexible way. Specifically, our numerical results suggest that regressing the precision parameter of the model on the covariates leads to a notably better model fit than when the precision parameter is kept constant. In addition to incorporating nonlinear predictor-response relationships, boosted beta regression accounts for spatial correlation in both the mean and the variance structure of the model. Clearly, this issue is important if observation units have a neighborhood structure and may therefore influence each other.

A key aspect of boosted beta regression is its ability to carry out variable selection during the fitting process. This implies that only a small subset of the available predictor variables is included in the final model for the percentage response. Variable selection is of high practical interest in applications where large amounts of potentially important predictor variables are available. Consequently, one is often interested in determining the most informative predictor variables and in discarding those predictors that have a negligible effect on the response. In case of the NLA data, for example, boosted beta regression selected only 15 informative predictor variables out of the total set of 78 available predictors.

It is important to note that the variable selection mechanism in boosted beta regression is fundamentally different from earlier approaches to selecting predictor variables in statistical regression models. For example, because beta regression is a member of GAMLSS model class, one could alternatively fit the model using maximum likelihood techniques and apply AIC-based methods for selecting informative predictor variables (as implemented in the R add-on package **gamlss**, [31]). This strategy, however, requires the model to be fitted multiple times. In contrast, our new method is based on boosting methodology and is therefore able to incorporate variable selection already into the model fitting process. Note that the sets of predictor variables selected for the mean and precision submodels do not have to be identical. For example, boosted beta regression allows for detecting factors that only manifest in the variance (but not in the mean) of the response (cf. [29]).

A challenging problem when modeling percentage outcomes is the inclusion of the boundary values "0%" and "100%". This is because the density of a beta distributed random variable is not defined at the boundary values 0 and 1. In case of the NLA data quantile-quantile plots suggested that zero percentages could be well incorporated into boosted beta regression if a small constant was added to these values (cf. [29]). If this strategy fails, or if the percentage of zero values is large among the observations of a data set, it may alternatively be worth fitting an extra model for the zero observations ("beta inflated regression", [56]). Similar to zero-inflated models for count data [32], this approach could also be incorporated into boosted beta regression. We plan to address this issue in a future paper.

Supporting Information

Text S1 This document provides technical details on boosted beta regression, as well as the full list of predictor variables used for the analysis of the NLA data.
(PDF)

Acknowledgments

We thank Charles Hawkins, Western Center for Monitoring and Assessment of Freshwater Ecosystems at Utah State University for providing several predictor variables. The views expressed in this article are those of the co-authors, and do not necessarily reflect the official views of the U.S. EPA or U.S. government. Use of trade, product or firm names does not imply endorsement by the U.S. Government.

Author Contributions

Conceived and designed the experiments: MS FW KOM RM NF AM. Analyzed the data: MS FW KOM RM NF AM. Wrote the paper: MS FW KOM RM NF AM.

References

1. Girao LC, Lopes AV, Tabarelli M, Bruna EM (2007) Changes in tree reproductive traits reduce functional diversity in a fragmented atlantic forest landscape. PLOS ONE 2: e908.
2. Warton DI, Hui FKC (2011) The arcsine is asinine: The analysis of proportions in ecology. Ecology 92: 3–10.
3. Laliberte E, Adair EC, Hobbie SE (2012) Estimating litter decomposition rate in single-pool models using nonlinear beta regression. PLOS ONE 7: e45140.
4. Schlegel P, Havenhand JN, Gillings MR, Williamson JE (2012) Individual variability in reproductive success determines winners and losers under ocean acidification: A case study with sea urchins. PLOS ONE 7: e53118.
5. Papke LE, Wooldridge JM (1996) Econometric methods for fractional response variables with an application to 401(k) plan participation rates. Journal of Applied Econometrics 11: 619–632.
6. Ferrari SLP, Cribari-Neto F (2004) Beta regression for modelling rates and proportions. Journal of Applied Statistics 31: 799–815.
7. Hunger M, Baumert J, Holle R (2011) Analysis of SF-6D index data: Is beta regression appropriate? Value in Health 14: 759–767.
8. Seow WJ, Pesatori AC, Dimont E, Farmer PB, Albetti B, et al. (2012) Urinary benzene biomarkers and DNA methylation in Bulgarian petrochemical workers: Study findings and comparison of linear and beta regression models. PLOS ONE 7: e50471.
9. Kieschnick R, McCullough BD (2003) Regression analysis of variates observed on (0, 1): Percentages, proportions and fractions. Statistical Modelling 3: 193–213.
10. Cribari-Neto F, Zeileis A (2010) Beta regression in R. Journal of Statistical Software 34: Issue 2.
11. Richards SA (2008) Dealing with overdispersed count data in applied ecology. Journal of Applied Ecology 45: 218–227.
12. Jonsson MT, Thor G (2012) Estimating coextinction risks from epidemic tree death: Affiliate lichen communities among diseased host tree populations of Fraxinus excelsior. PLOS ONE 7: e45701.
13. Peterson EE, Urquhart NS (2006) 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.
14. Gelfand AE (2007) Guest editorial: Spatial and spatio-temporal modeling in environmental and ecological statistics. Environmental and Ecological Statistics 14: 191–192.
15. Schmid M, Hothorn T, Maloney KO,Weller DE, Potapov S (2011) Geoadditive regression modelling of stream biological condition. Environmental and Ecological Statistics 18: 709–733.
16. 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 in Ecology and Evolution 3: 116–128.
17. Simas AB, Barreto-Souza W, Rocha AV (2010) Improved estimators for a general class of beta regression models. Computational Statistics & Data Analysis 54: 348–366.
18. Mayr A, Fenske N, Hofner B, Kneib T, Schmid M (2012) Generalized additive models for location, scale and shape for high dimensional data - a exible

approach based on boosting. Journal of the Royal Statistical Society, Series C 61: 403–427.

19. Rigby RA, Stasinopoulos DM (2005) Generalized additive models for location, scale and shape (with discussion). Applied Statistics 54: 507–554.
20. Ripley BD (2004) Selecting amongst large classes of models. In: Adams N, Crowder M, Hand DJ, Stephens D, editors, Methods and Models in Statistics, London: Imperial College Press. 155–170.
21. Whittingham MJ, Stephens PA, Bradbury RB, Freckleton RP (2006) Why do we still use stepwise modelling in ecology and behaviour? Journal of Animal Ecology 75: 1182–1189.
22. Friedman JH (2001) Greedy function approximation: A gradient boosting machine. The Annals of Statistics 29: 1189–1232.
23. Bühlmann P, Hothorn T (2007) Boosting algorithms: Regularization, prediction and model fitting (with discussion). Statistical Science 22: 477–522.
24. Kneib T, Hothorn T, Tutz G (2009) Variable selection and model choice in geoadditive regression models. Biometrics 65: 626–634.
25. US Environmental Protection Agency (USEPA) (2009) National Lakes Assessment: A Collaborative Survey of the Nation's Lakes. Washington, D.C.: U.S. Environmental Protection Agency; Office of Water and Office of Research and Development. EPA 841-R-09–001.
26. Karr JR, Chu EW (1999) Restoring Life in Running Waters: Better Biological Monitoring. Washington D.C.: Island Press.
27. Barbour MT, Gerritsen J, Snyder BD, Stribling JB (1999) Rapid Bioassessment Protocols for Use in Streams and Wadeable Rivers: Periphyton, Benthic Macroinvertebrates and Fish. Washington, D.C.: U.S. Environmental Protection Agency; Office of Water, 2 edition. EPA 841-B-99–002.
28. Maloney KO, Feminella JW (2006) Evaluation of single- and multi-metric benthic macroinvertebrate indicators of catchment disturbance over time at the Fort Benning Military Installation, Georgia, USA. Ecological Indicators 6: 469–484.
29. Smithson M, Verkuilen J (2006) A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. Psychological Methods 11: 54–71.
30. Cox C (1996) Nonlinear quasi-likelihood models: Applications to continuous proportions. Computational Statistics & Data Analysis 21: 449–461.
31. Stasinopoulos DM, Rigby RA (2007) Generalized additive models for location scale and shape (GAMLSS) in R. Journal of Statistical Software 23: Issue 7.
32. Schmid M, Potapov S, Pfahlberg A, Hothorn T (2010) Estimation and regularization techniques for regression models with multidimensional prediction functions. Statistics and Computing 20: 139–150.
33. Hastie T, Tibshirani R (1990) Generalized Additive Models. London: Chapman & Hall.
34. Wood S (2006) Generalized Additive Models: An Introduction with R. Boca Raton: Chapman & Hall/CRC.
35. Eilers PHC, Marx BD (1996) Flexible smoothing with B-splines and penalties. Statistical Science 11: 89–121.
36. Schmid M, Hothorn T (2008) Boosting additive models using component-wise P-splines. Computational Statistics & Data Analysis 53: 298–311.
37. McKay L, Bondelid T, Rea A, Johnston C, Moore R, et al. (2012) NHDPlus Version 2: User Guide. Available: ftp://ftp.horizon-systems.com/NHDPlus/NHDPlusV21/Documentation.
38. Kottek M, Grieser J, Beck C, Rudolf B, Rubel F (2006) World map of the Köppen-Geiger climate classification updated. Meteorologische Zeitschrift 15: 259–263.
39. Davison A, Hinkley D (1997) Bootstrap Methods and their Application. Cambridge: Cambridge University Press.
40. Hothorn T, Leisch F, Zeileis A, Hornik K (2005) The design and analysis of benchmark experiments. Journal of Computational & Graphical Statistics 14: 675–699.
41. Hothorn T, Buhlmann P, Kneib T, Schmid M, Hofner B (2012) mboost: Model-Based Boosting. Available: http://CRAN.R-project.org/package = mboost. R package version 2.2–1.
42. Hofner B, Mayr A, Fenske N, Schmid M (2012) gamboostLSS: Boosting Methods for GAMLSS Models. Available: https://r-forge.r-project.org/projects/gamboostlss. R package version 1.1–0.
43. Maddala GS (1983) Limited-Dependent and Qualitative Variables in Econometrics. Cambridge: Cambridge University Press.
44. Walsh CJ, Roy AH, Feminella JW, Cottingham PD, Groffman PM, et al. (2005) The urban stream syndrome: Current knowledge and the search for a cure. Journal of the North American Benthological Society 24: 706–723.
45. Mendoza G, Catalan J (2010) Lake macroinvertebrates and the altitudinal environmental gradient in the Pyrenees. Hydrobiologia 648: 51–72.
46. Rahbek C (1997) The relationship among area, elevation, and regional species richness in neotropical birds. The American Naturalist 149: 875–902.
47. Nyman M, Korhola A, Brooks SJ (2005) The distribution and diversity of Chironomidae (Insecta: Diptera) in western Finnish Lapland, with special emphasis on shallow lakes. Global Ecology and Biogeography 14: 137–153.
48. Brodersen KR, Dall PC, Lindegaard C (1998) The fauna in the upper stony littoral of Danish lakes: Macroinvertebrates as trophic indicators. Freshwater Biology 39: 577–592.
49. Welch C, Vodopich D (1989) Production by Hexagenia limbata in a warm-water reservoir and its association with chlorophyll content of the water column. Hydrobiologia 185: 183–193.
50. Heino J (2000) Lentic macroinvertebrate assemblage structure along gradients in spatial heterogeneity, habitat size and water chemistry. Hydrobiologia 418: 229–242.
51. Kilgour B, Clarkin C, Morton W, Baldwin R (2008) Inuence of nutrients in water and sediments on the spatial distributions of benthos in Lake Simcoe. Journal of Great Lakes Research 34: 365–376.
52. Harper PP, Cloutier L (1986) Spatial structure of the insect community of a small dimictic lake in the Laurentians (Quebec). Internationale Revue der gesamten Hydrobiologie und Hydrographie 71: 655–685.
53. Rossaro B, Marziali L, Cardoso AC, Solimini A, Free G, et al. (2007) A biotic index using benthic macroinvertebrates for Italian lakes. Ecological Indicators 7: 412–429.
54. Feminella JW (2000) Correspondence between stream macroinvertebrate assemblages and 4 ecoregions of the southeastern USA. Journal of the North American Benthological Society 19: 442–461.
55. Hawkins C, Norris R, Gerritsen J, Hughes R, Jackson S, et al. (2000) Evaluation of the use of landscape classifications for the prediction of freshwater biota: synthesis and recommendations. Journal of the North American Benthological Society 19: 541–556.
56. Stasinopoulos M, Rigby B (2012) gamlss.dist: Distributions to be Used for GAMLSS Modelling. Available: http://CRAN.R-project.org/package = gamlss.dist. R package version 4.2–0.