esa ECOSPHERE

EMERGING TECHNOLOGIES

Statistically reinforced machine learning for nonlinear patterns and variable interactions

MASAHIRO RYO[1,2,]† AND MATTHIAS C. RILLIG[1,2]

[1]*Institute of Biology, Freie Universität Berlin, D-14195 Berlin Germany*
[2]*Berlin-Brandenburg Institute of Advanced Biodiversity Research, D-14195 Berlin Germany*

Citation: Ryo, M., and M. C. Rillig. 2017. Statistically reinforced machine learning for nonlinear patterns and variable interactions. Ecosphere 8(11):e01976. 10.1002/ecs2.1976

Abstract. Most statistical models assume linearity and few variable interactions, even though real-world ecological patterns often result from nonlinear and highly interactive processes. We here introduce a set of novel empirical modeling techniques which can address this mismatch: statistically reinforced machine learning. We demonstrate the behaviors of three techniques (conditional inference tree, model-based tree, and permutation-based random forest) by analyzing an artificially generated example dataset that contains patterns based on nonlinearity and variable interactions. The results show the potential of statistically reinforced machine learning algorithms to detect nonlinear relationships and higher-order interactions. Estimation reliability for any technique, however, depended on sample size. The applications of statistically reinforced machine learning approaches would be particularly beneficial for investigating (1) novel patterns for which shapes cannot be assumed a priori, (2) higher-order interactions which are often overlooked in parametric statistics, (3) context dependency where patterns change depending on other conditions, (4) significance and effect sizes of variables while taking nonlinearity and variable interactions into account, and (5) a hypothesis using parametric statistics after identifying patterns using statistically reinforced machine learning techniques.

Key words: biodiversity; context dependency; data-driven; ecological surprises; higher-order interactions; machine learning; macroecology; microbial ecology; multiple stressors; novel ecosystems; random forest; statistically reinforced machine learning.

Received 26 April 2017; revised 8 August 2017; accepted 12 September 2017. Corresponding Editor: Judy Cushing.

† **E-mail:** masahiroryo@gmail.com

INTRODUCTION

Ecological patterns structured by nonlinear and interactive processes are ubiquitously found at any ecological scales from individual to ecosystem in terrestrial, freshwater, and marine systems (e.g., Dodds et al. 2010, van Nes et al. 2016, Soranno et al. 2014, see also examples in Table 1). Nonlinearity and interactions among factors driving an ecological system often cause unexpected changes in the system's state (Peters et al. 2007, Scheffer 2009, Soranno et al. 2014), for example, as seen in alternative stable states (van Nes et al. 2016). Understanding and modeling such nonlinear interaction dynamics inevitably will make ecological systems more predictable (Urban et al. 2016, Mayfield and Stouffer 2017).

Yet, analyzing complex patterns with statistical modeling is not so easy. Statistical models per se are capable of analyzing nonlinear relationships and variable interactions. However, the design of a statistical model is dependent heavily on how the user views the system. Model designs directly affect system interpretation (e.g., Gilbert

Table 1. Representative ecological examples of nonlinearity and variable interaction.

Feature	Category	Ecological context	Short description	Type(s) of ecology	Reference example
Nonlinearity	Threshold	No effect concentration (cf. critical load)	A critical concentration/magnitude for an external force's effect appears	Ecotoxicology, pollution ecology	Iwasaki and Ormerod (2012)
		Triggering cue	An external forcing event that stimulates an organism's behavior such as reproduction and dispersal	Organismal ecology	Andrade-Linares et al. (2016)
		Tipping point (cf. regime shift, alternative stable state)	A critical level that abruptly changes a system's state	Ecosystem ecology	van Nes et al. (2016)
	Polynomial (unimodal or bimodal)	Physiological adjustment	A systemic response of an organism to an external force	Organismal ecology	Thomas et al. (2012)
		Intermediate disturbance hypothesis	The hypothesis that local species diversity is maximized by disturbances of intermediate frequency/magnitude	Community ecology, ecosystem ecology	Wilkinson (1999)
		Latitude–species diversity relationship	A pattern that species diversity is the highest around the intermediate latitude range (either unimodal/bimodal)	Macroecology	Chaudhary et al. (2016)
Variable interaction	Conditional	Context dependency	The situation that patterns of a specific relationship vary according to other variables' conditions	Ecosystem ecology, community ecology	Tonkin et al. (2016)
		Trait–environmental relationship	Responses of an organism to an external force depend on its traits	Organismal ecology	Hunter et al. (2014)
		Priming (cf. predictive response strategy)	An improved reaction of an organism to an external force following a preceding event	Organismal ecology, community ecology	Rillig et al. (2015*a*)
	Mutual	Biotic interaction	Effects that organisms in a community have on one another, including competition, exploitation, and mutualism	Community ecology	Bastolla et al. (2009)
		Hierarchical interaction (cf. cross-scale interaction)	Patterns and processes that interact across different scales	Macroecology, ecosystem ecology	Soranno et al. (2014)
		Multidimensionality	The concept that a combinatory effect of multiple forces determines its consequence.	Theoretical ecology	Pickett and Cadenasso (2002)

Notes: For these examples, partial dependence plots with permutation-based random forest models are suitable for detecting significant nonlinear patterns. Conditional inference and model-based decision tree models are suited for inferring variable interactions.

and Bennett 2010, Cumming 2016). For practical convenience, the vast majority of statistical models rely on linearity, few variable interactions, and data transformation for normality in ecology. The subjectivity in statistical model design is perhaps one of the major obstacles to advance our understanding of nonlinear interaction dynamics in ecological systems.

Machine learning, in contrast to statistical modeling, explores the structure of the target system without pre-assumption on data (Breiman 2001*a*, Recknagel 2001). Machine learning algorithms aim to build an empirical model that maximizes predictability. Most of them automatically and thoroughly examine possible nonlinear relationships and higher-order interactions (greater than two-way). They generally explain ecological patterns more accurately than statistical models (Olden et al. 2008, Crisci et al. 2012, Thessen 2016).

Applications of machine learning models in ecology have been emerging since a turning point around 2006 (Fig. 1). The two most cited articles applying machine learning in ecology are in the field of biogeography, papers on species

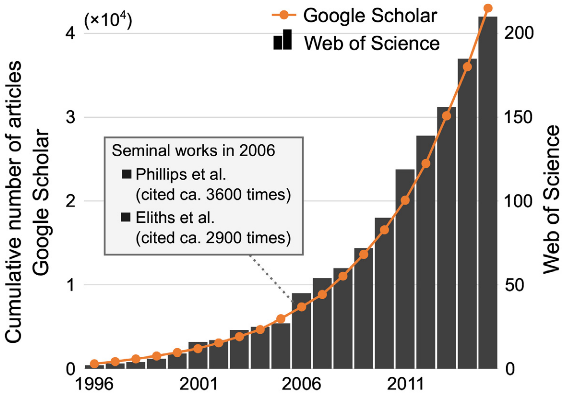

Fig. 1. Increasing applications of machine learning in ecology during the last two decades. Data were obtained by searching for "machine learning" as topic in "ecology" category in the Web of Science (Clarivate Analytics) and for "machine learning" and "ecology" in Google Scholar in October 2016.

distribution modeling by Elith et al. (2006) and Phillips et al. (2006). Machine learning algorithms have demonstrated high accuracy in predicting ecological patterns, such as for data on species diversity (e.g., Olden et al. 2008) and distributions (e.g., Elith and Leathwick 2009). Machine learning is expected to discover novel patterns which provide an opportunity to speculate about underlying mechanisms in fields where prior knowledge is still minimal and hypotheses have not been clearly developed (Hochachka et al. 2007): Examples include aspects of biodiversity science (Kelling et al. 2009, Rillig et al. 2015*b*), macrosystem ecology (Heffernan et al. 2014, Levy et al. 2014), and microbial ecology (Baldi and Brunak 2001). Previous studies and reviews on ecological applications of machine learning have been instrumental in spreading an appreciation of machine learning approaches (e.g., Olden et al. 2008, Crisci et al. 2012, Thessen 2016).

A new movement that develops algorithms merging the two relevant approaches—statistical modeling and machine learning—has been occurring since the mid-2000s. As there is no widely accepted term to lump such algorithms together, we here call them "statistically reinforced machine learning (SML)." Note that an established similar term, the so-called statistical learning exists, but this refers in a much broader sense to any statistical techniques for understanding data (Hastie et al. 2009, James et al. 2013). The fundamental aims of statistical modeling and machine learning essentially differ from each other as being theory-driven (hypothesis-testing) and data-driven (information-searching), respectively. Statistically reinforced machine learning merges both techniques and assumptions. They, therefore, are highly attractive because of a high potential to assist researchers to test hypotheses with the help of artificial intelligence.

Here, we introduce the concept of SML and some of the well-established algorithms that fit to the concept: conditional inference tree, model-based tree, and permutation-based random forest. They share many features because the latter two algorithms were developed based on the first one. We show behaviors of these algorithms together with a linear regression by analyzing an artificially generated dataset. We also test whether the performances of the algorithms are sensitive to sample size. It is important to bear in mind that we do not intend to rank the performances of these three algorithms because they were developed for different aims with different benefits. In particular, random forest algorithms are an advanced form of single tree algorithms to better solve model over-fitting and therefore should perform more accurately than the others.

Methods

Statistically reinforced machine learning

We define "statistically reinforced machine learning (SML)" as a set of machine learning algorithms which take nonlinear associations and higher-order interactions into account automatically, while testing statistical significance and thereby conducting variable selection. Technically, SML can include semi-parametric modeling approaches that require users to specify both a finite-dimensional vector of parameters and an infinite-dimensional vector function. Practical comparisons of representative features of statistical modeling, SML, and machine learning are summarized in Table 2.

The need for reinforcement of machine learning algorithms by statistics has long been recognized (e.g., Mingers 1987, Olden and Jackson 2002). This realization has changed the fundamental premise of machine learning from the idea that all information can be valuable into the idea that variable

Table 2. Practical comparison of representative features of statistical modeling, statistically reinforced machine learning (SML), and machine learning in typical cases.

Parameters	Statistical modeling	SML	Machine learning
Fundamental concept	Hypothesis-testing	Data-mining and hypothesis-testing	Data-mining
Variety of methods available	Diverse	(still) Few	Diverse
Interpretability of modeling structure and procedure	High	Moderate	Low
Statistical inference on the effect of a predictor	Capable	Capable	Incapable
Analyzing nonlinear relationships and higher-order interactions of variables	Need to be explicitly designed	Automatically assumed	Automatically assumed
Modeling accuracy	Often lower than the others	Moderate–high	Moderate–very high

Note: Note that exceptional cases exist because the features depend also on data and model structure.

selection based on statistical significance can improve prediction accuracy (Hothorn et al. 2006*b*, Hapfelmeier and Ulm 2013). Conditional inference tree, developed by Hothorn et al. (2006*b*), is seen as the cornerstone work for SML.

*Conditional inference tree.—*Conditional inference tree is a decision tree model (also known as classification and regression trees; Breiman et al. 1984) that is among the most frequently used machine learning algorithms. Decision tree models explain the variance in a response variable by recursively splitting the data into more homogeneous groups using a combination of predictors. At each split, the data are divided into two groups according to a threshold value of one of the predictors so that the variance after the split decreases from the variance before the split. The splitting procedure is repeatedly applied for each of the split groups until it achieves a criterion: One of the most used criteria is the moment when the model maximizes predictive performance while conducting the fewest splits (Breiman et al. 1984).

Conditional inference trees (ctree; Hothorn et al. 2006*a*, *b*) have incorporated a statistical inference framework to split the data based on statistical tests instead of maximizing prediction power. This improvement solved two fundamental problems of the traditional tree models (Hothorn et al. 2006*b*). First, this model decides whether the data should be divided at each split or not based on statistical significance to avoid model over-fitting (Table 3), while traditional decision tree models can often select predictors for splitting even though their effects are not statistically significant and thus can over-fit. Statistical significance of all predictors is tested independently using different statistical tests according to the combination of types of predictor and response variables (Table 3). Second, ctree model selects a predictor at each split without selection bias, while traditional decision tree models tend to preferentially select predictors in the order binary < categorical < numeric as splitting opportunities simply increase in this order (Hothorn et al. 2006*b*). Traditional decision tree models do not possess these important features, and these caveats have not been acknowledged in ecological studies since decision tree models were first introduced by De'ath and Fabricius (2000). The detailed procedure of ctree modeling is available in Appendix S1. These improvements, to avoid model over-fitting and variable selection bias, are also kept in the following algorithms: model-based tree and permutation-based random forest.

*Model-based tree.—*Another derived form of decision tree models, the model-based trees (mobtree;

Table 3. Statistical tests and test statistics used in conditional inference tree models (Hothorn et al. 2006*a*).

		Response variable	
		Numerical	Categorical
Predictor	Numerical	Pearson's correlation (*t*)	Kruskal–Wallis (χ^2)
	Categorical	Kruskal–Wallis (χ^2)	Cochran–Mantel–Haenszel (χ^2)

Notes: Pearson correlation tests whether the correlation coefficient is equal to 0 based on a *t*-distribution, Kruskal–Wallis test is to test whether samples that belong to different categories originate from the same distribution based on a chi-square distribution, and Cochran–Mantel–Haenszel test is to test whether the relative proportions of one variable are independent of the other variables based on a chi-square distribution.

Zeileis et al. 2008), couples the features of parametric statistical models such as generalized linear models and decision tree models. Mobtree first requires explicitly specifying how predictor–response relationships should be modeled parametrically (e.g., linear model) using a few predictors. It then automatically searches for other important predictors, which can significantly influence the parameter values of the relationships (more details in Appendix S1) based on the M-fluctuation test (Zeileis and Hornik 2007). This method is particularly useful to test a hypothesis that patterns in a relationship between a response variable and some of the predictors are altered by other predictors. For instance, Campetella et al. (2011) found out that a plant trait–environmental relationship (more specifically, the parameters of a linear regression between specific leaf area and inclination of field site) varies according to the age of forest succession at a field site.

Permutation-based random forest.—Random forest models (Breiman 2001*b*) construct a predictive model and estimate the relative importance of predictors; they are acknowledged as one of the most accurate machine learning algorithms to date (Douglas et al. 2011, Crisci et al. 2012). This algorithm first generates a large number of decision tree models that use diverse combinations of predictors and thresholds to explain datasets which are generated for the individual trees by sampling from the original data with replacement. Then, it takes an overall average of these tree models' outputs as a prediction (so-called ensemble/consensus modeling). The relative importance of predictors is usually measured by evaluating how much each predictor contributes to increasing model accuracy.

Recently proposed SML random forest algorithms can evaluate statistical significance of the predictors based on a permutation approach. The permutation approach generates a large number of random forest models to obtain the probability distributions of the relative importance measures of the predictors. Then, it quantifies how rarely the original relative importance measure of each predictor is obtained by chance (more details in Appendix S1). This permutation technique demands high computational performance unachievable using a typical laptop about a decade ago. Although we do not show this here, building a random forest model using only statistically significant predictors chosen with the permutation-based random forest approach generally improves accuracy (Hapfelmeier and Ulm 2013). This is another advantage of this method. An earlier work to employ a permutation approach to another machine learning algorithm within the context of SML is seen in Olden and Jackson (2002).

Generating an artificial dataset

As one of our primary aims is to assess reliability of the SML algorithms, we analyze an artificially generated dataset for which we know which and how predictors are correlated to the response variable, instead of studying an actual ecological dataset. We herein assume an ecological pattern that is structured by multiple influences, in which factors nonlinearly affect another factor and where there are factor interactions (in relevance to Table 2). Mimicking a field monitoring record, we also add considerable random noise to blur the associations. Samples are independent and identically distributed without spatial and temporal autocorrelation. The pattern can be placed in any ecological context including species distribution and abundance, species richness, community composition, and ecosystem function after de-trending autocorrelation.

Let us imagine this scenario where the goal is to understand primary production: Primary production initially increases after a disturbance and then gradually decreases during the successional period (i.e., unimodal pattern; e.g., Campbell et al. 2004), increases with an increased carbon dioxide concentration only when it is the limiting factor (i.e., positive linearity with a threshold), and such patterns may be context-dependent given a combination of climate and soil classifications (e.g., Cleveland et al. 2011). Note that we are not going to discuss results from this viewpoint. Rather, this is just an example to help readers imagine how the data structure can be considered in an ecological context.

We generated an artificial dataset of a numeric response variable Y and 16 predictors: three binary ones ($x1$–$x3$); three categorical ones ($x4$–$x6$); and 10 numeric ones ($x7$–$x16$). Binary predictors ($x1$–$x3$) were generated by random sampling from binomial distributions whose probabilities are 0.5. Categorical predictors ($x4$–$x6$) were randomly sampled from 3, 4, and 5 categories, respectively.

For each numeric predictor (*x*7–*x*16), random sampling was conducted from one of the probability distributions: a uniform distribution (ranging from 0 to 10), a normal distribution (mean = 10, standard deviation [SD] = 3), and a Poisson distribution (λ = 5).

Relevant to the above example of primary production, *Y* was generated in association with the predictors of *x*7, *x*8, and *x*1**x*4**x*9 (* interaction; Fig. 2). The predictor *x*7 structured a unimodal relationship with *Y* using a cosine curve function having a peak of 5. The predictor *x*8 affected *Y* positively linearly with a slope of 0.6 until exceeding a threshold (*x*8 = 10). The predictor *x*9 affected *Y* negatively linearly with a slope of −0.5 when *x*1 = 1 and *x*4 ≠ "low" (*x*4 ε {low, moderate, high}}) as a three-way interaction. Finally, random noise following the Gaussian distribution (mean = 0 and SD = 3) was added to *Y*, which contributed to approximately 50% of the total variance of *Y*. The remaining 11 predictors were not associated with *Y*.

Model performance assessment

We tested whether the above-described SML algorithms changed the detectability of statistical significance of the associated predictors and the interactions (*x*1**x*4**x*9, *x*7, and *x*8) as a function of different sample sizes (*n* = 100, 200, and 300).

For mobtree, we focused on whether the interactive effect of *x*1 and *x*4 on the *Y*–*x*9 relationship is detected when an association of *x*9 and *Y* is presumed. For permutation-based random forest, we qualitatively assessed whether the designed

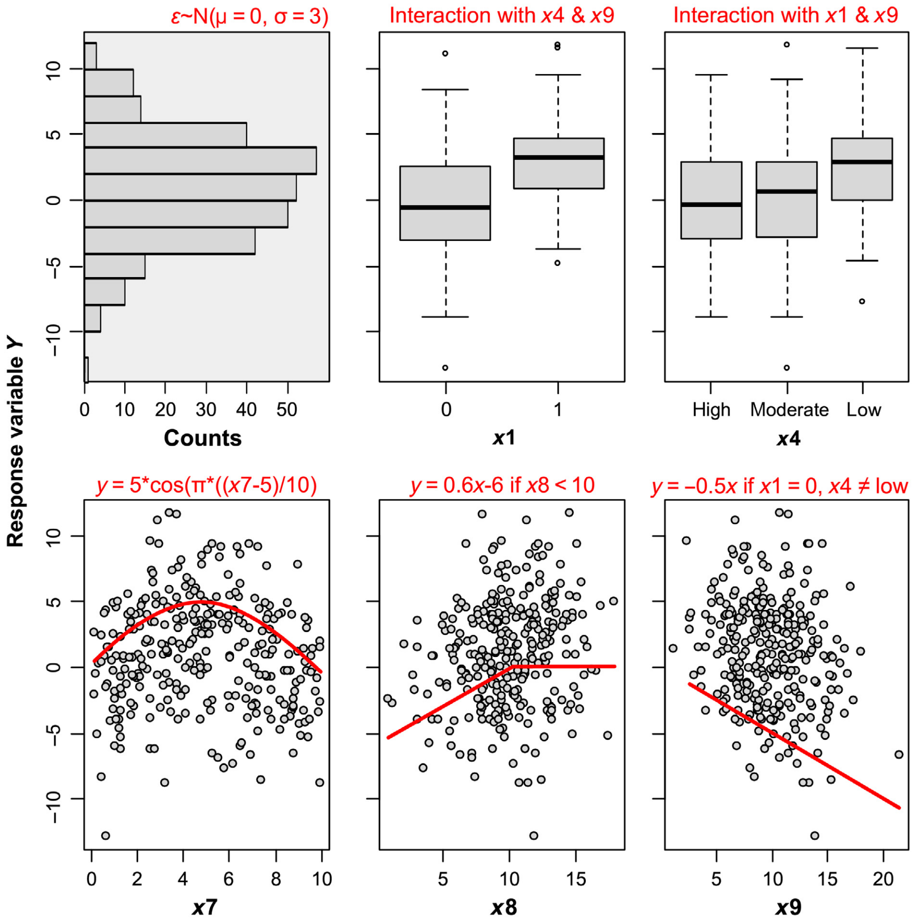

Fig. 2. Artificially generated data structure (*n* = 300 and 16 predictors, *x*1–*x*16). Only predictors that were designed to associate with the response *Y* are shown. The modeled relationships between predictors (*x*7–*x*9) and the response *Y* are depicted with the curves and denoted in red color.

patterns of $x7$, $x8$, and $x1^*x4^*x9$ are found, using partial dependence plots. Partial dependence plots are a technique used to visualize specified predictor–response relationships that are empirically modeled (not limited to SML). Partial dependence plots visualize the dependence of response variable on predictors while taking the effects of the other predictors into account (Hastie et al. 2009). This technique is especially useful for visually assessing the shapes of some interesting relationships but not for quantifying the effect sizes (values in y-axis do not indicate the effect size).

We also applied a linear regression model to compare with the SML models. A model includes the unimodal curve as a second-degree polynomial term, the three-way interactions (also the nested two-way interactions), and the other predictors: $Y \sim f \quad (x7 + x7^2 + x1^*x4^*x9 + x2 + x3 + x5 + \cdots + x16) + \varepsilon$. A stepwise best-model selection was conducted for both models based on Akaike's information criterion. Akaike's information criterion aims to select the best model in terms of maximizing predictability, which is equivalent to the objective of machine learning.

The R script for the entire process including data generation and analysis is available at github (see https://github.com/masahiroryo/R_Statistically-reinforced-machine-learning). We run ctree and mobtree models with R package party (Strobl et al. 2007, 2008) and permutation-based random forest models with the R script modified from Hapfelmeier et al. (2014) in R Statistical Computing (R Development Core Team 2016). For partial dependence plots, we used the R package mlr (Bischl et al. 2016). Note that we used Bonferroni correction for all the SML models, following the convention of these methods. For simplicity, no data transformation was performed. We structured the random forest algorithm with 300 tree models after confirming this number is sufficient to stabilize the results and with 2000 permutations for estimating statistical significance of predictors.

Results

The number of branches in ctree models increased along with enhancing detectability of the designed three-way interactions ($x1^*x4^*x9$ with negative linear $x9$–Y pattern) by increasing sample sizes of the dataset from 100 to 200 and 300 (Fig. 3). $x1$ appeared at the first nodes in the models regardless of sample size. $x4$ appeared under the path of $x1 = 0$, and $x9$ appeared under the path of $x4$ being either high or moderate when $n = 200$ and 300. The model structures were identical between $n = 200$ and 300, differing marginally in the threshold values for $x9$. Significance of the designed associations of unimodal and threshold-linear patterns ($x7$ and $x8$, respectively) was not inferred in any sample sizes.

The mobtree models inferred that negative $x9$–Y relationships were significant only in specific conditions of $x1$, $x4$ (as designed interactions), and $x7$ (designed unimodal pattern) when $n =$ 200 and 300 (Fig. 4). This means that the three-way interactions of $x1^*x4^*x9$ were incorporated as a part of the model structure. Similar to ctree models, the model structures were identical between $n = 200$ and 300, differing marginally in the threshold values for $x7$. In the case of $n = 300$, the model inferred the significant negative $x9$–Y relationship conditional only to $x7$, which was not designed (i.e., false detection; node 9 in Fig. 4c). When $n = 100$, the model did not indicate any predictors that influence the parameters of the negative $x9$–Y relationship (Fig. 4a).

The permutation-based random forest models enhanced detectability of predictors associated with the response variable by increasing sample size (Fig. 5). As in the ctree models, a part of the three-way interaction, $x1$, was consistently selected regardless of sample size. The significances of another interaction factor ($x4$) and the unimodal pattern ($x7$) were detected when $n = 200$ and 300. The last part of the three-way interaction ($x9$) was detected only when $n = 300$. By increasing sample size, these signals became stronger: The values of the relative importance of these predictors increased and became more distinct from the values of the other predictors which stayed near 0.

Partial dependence plots depicted associations of predictors ($x7$ as unimodal, $x8$ as linear-threshold, $x9$ as negative linear conditional to $x1$ and $x4$) with the response variable modeled with the permutation-based random forest models (Fig. 6). The unimodal curve of the $x7$–Y relationship with the peak at $x7 = 5$ (Fig. 2) was adequately modeled when $n = 200$ and 300 (Fig. 6b, c). The conditional positive linear $x8$–Y relationship was

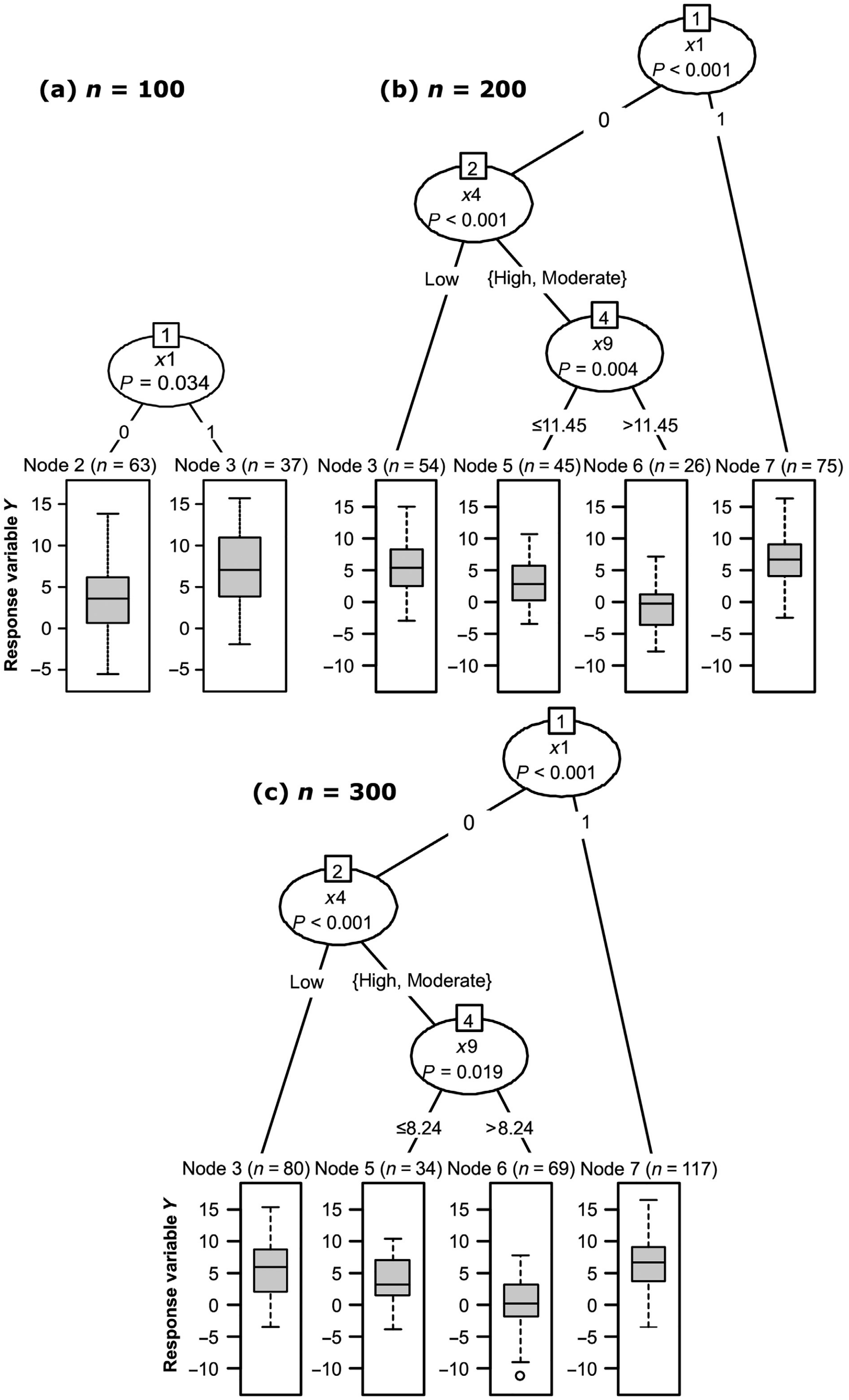

Fig. 3. Conditional inference trees indicated the designed interactions among $x1$, $x4$, and $x9$ (n = 200, 300).

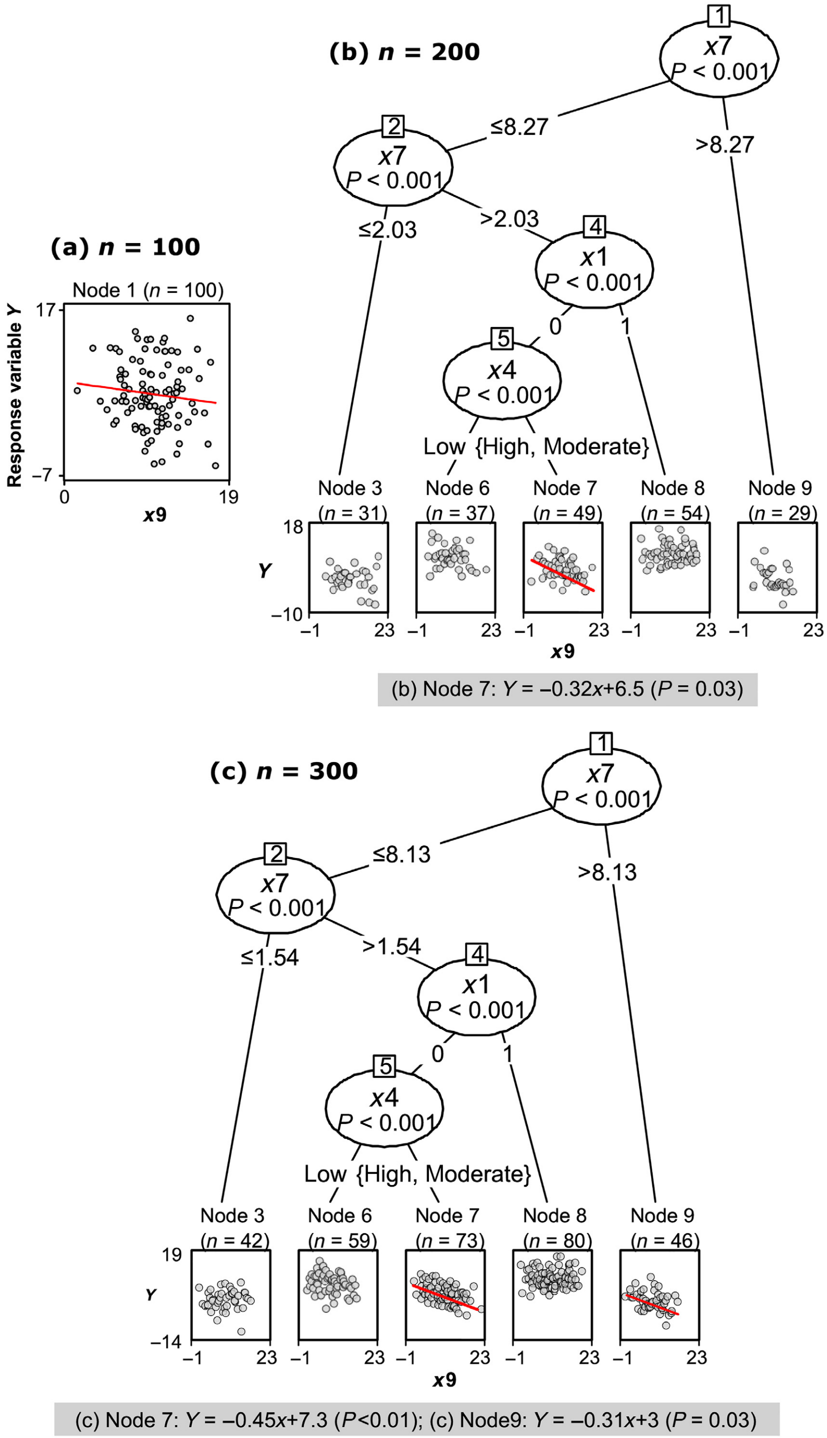

Fig. 4. Model-based trees that presumed $x9$–Y linear relationships indicated the designed interactions among $x1$, $x4$, and $x9$ together with the effect of $x7$ (n = 200, 300). Lines colored in red indicate significant relationships ($P < 0.05$).

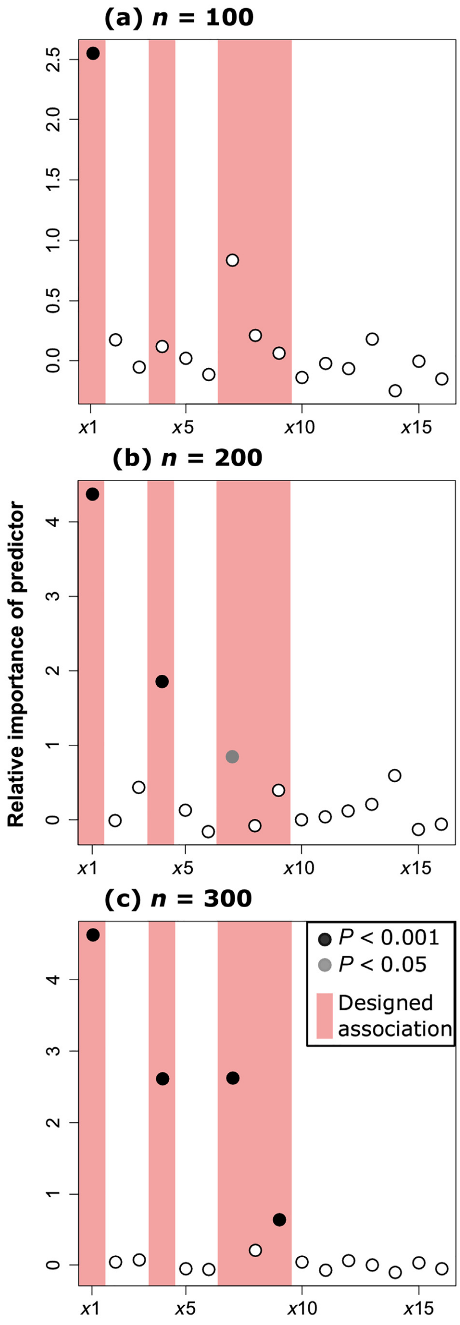

Fig. 5. Relative importance of predictors estimated with permutation-based random forest models. By increasing the sample size, the models detected more predictors that were designed to associate with the response.

most closely captured when *n* = 100, although the models did not support the significance for any sample size (Figs. 5, 6). The negative linear *x*9–*Y* relationship was most closely resembled by the gradual slope when *n* = 300 (Fig. 6c), while a sudden shift around *x*9 = 10 was estimated when *n* = 100 and 200 (Fig. 6a, b). The three-way interactions of *x*1**x*4**x*9 were seemingly suggested when *n* = 200 and 300 (Fig. 6e, f), while two-way interactions of *x*1 and *x*9 were inferred when *n* = 100 (Fig. 6d).

The SML models selected only a subset of predictors which were designed to associate with the response variable (*x*1, *x*4, *x*7, *x*8, and *x*9), while the linear regression model selected predictors that were not designed to influence the dependent variable when *n* = 100 and 200 (i.e., Type I error). The linear regression models did not infer the three-way interactions (*x*1**x*4**x*9) for any sample size, although two-way interactions within parts of the three-way interactions (i.e., *x*1**x*4 and *x*4**x*9) were found only when *n* = 200. The second-degree polynomial curve of *x*7 was consistently significant. *X*8 was significant when *n* = 300.

Discussion

We here introduced the concept of and a practical guide to SML approaches in ecological studies (Table 2) and demonstrated their usefulness for analyzing nonlinearity and higher-order interactions (Table 1) by analyzing an artificially generated dataset. Statistically reinforced machine learning approaches can be used for classification problems, while we demonstrated a regression problem. Moreover, this article introduced only a small fraction of machine learning algorithms, while many more algorithms have been progressively developed (Domingos 2012). This means that other algorithms can be more appropriate for specific ecological studies than the examples we introduced here.

Although machine learning techniques themselves are attractive as a possible alternative to statistical modeling in terms of high accuracy and the abilities of pattern-mining and variable selection, they cannot infer statistical significance. This may discourage ecologists from using them in hypothesis-testing studies. Cutler et al. (2007), who first introduced the original random forest (Breiman 2001*b*) to ecology, also argued this point as follows: “Random forest is not a tool for traditional statistical inference. It is not suitable for ANOVA or hypothesis testing. It does not compute *P* values… (page 2792).” This is,

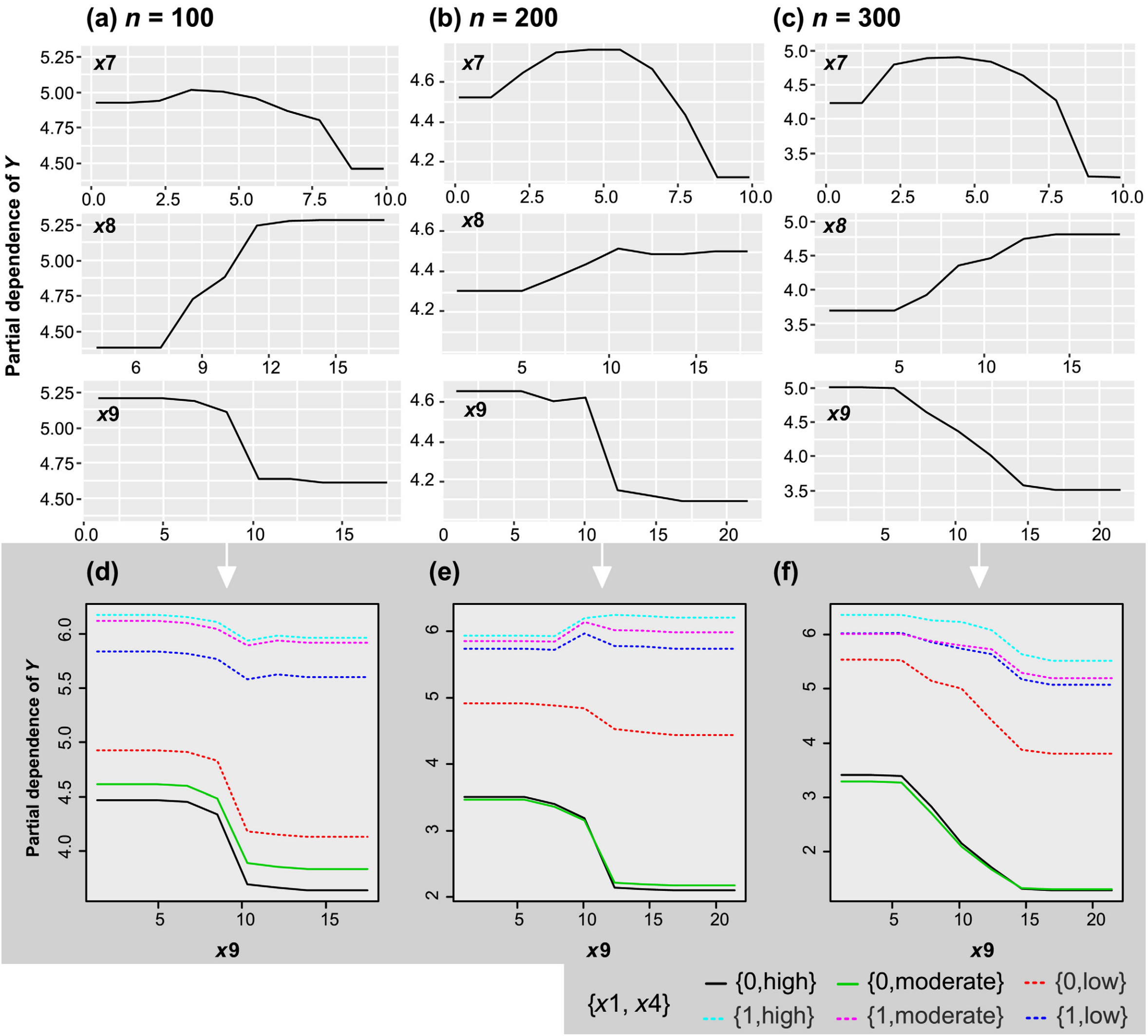

Fig. 6. Partial dependence plots of Y in association with $x7$, $x8$, and $x9$ estimated with permutation-based random forests (a–c). For $x9$–Y relationships, the three-way interactions of $x9$ with $x1$ and $x4$ are further investigated, and the negative $x9$–Y relationship conditional to $x1$ and $x4$ became apparent (d–f).

however, not true for the permutation-based random forest (Hapfelmeier and Ulm 2013).

We found that the SML approach requires an adequate sample size to guarantee both statistical significance and pattern detection. The performances of the linear regression and SML approaches depend on sample size, but they showed different dependencies on sample size. Statistically reinforced machine learning approaches detected more patterns and obtained more statistical significance with increasing sample size, while the linear regression models did not follow this trend but reduced Type I error (Table 4).

The partial dependence plots (Fig. 6) demonstrated that the unimodal curve with the peak value and the conditional negative linear effect of $x9$ can be adequately captured when the sample size is satisfactory (n = 300 in this study). The trend of $x8$, at least its positive linear effect, was somewhat captured, but this was not significant. This is probably because of the large random noise compared to the $x8$ influence.

Table 4. Model performances in detectability of the designed X–Y associations based on P-value.

	Conditional inference tree			Model-based tree			Permutation-based random forest			Linear model with AIC		
Sample size	100	200	300	100	200	300	100	200	300	100	200	300
$x1$	*	***	***		***	***	***	***	***		***	
$x4$		***	***		***	***		***	***			***
$x7$					***	***		*	***	**		
$x8$												**
$x9$		**	*		(presumed)				***		**	**
$(x7)^2$	*n*	*n*	*n*	*n*	*y/n*	*y/n*	*n*	*y*	*y*	***	***	***
$x1^*x4$	*n*	*y*	*y*	*n*	*y*	*y*	*n*	*y/n*	*y/n*		**	
$x1^*x9$	*n*	*y*	*y*	*n*	*y*	*y*	*n*	*y/n*	*y/n*			
$x4^*x9$	*n*	*y*	*y*	*n*	*y*	*y*	*n*	*y/n*	*y/n*		**	
$x1^*x4^*x9$	*n*	*y*	*y*	*n*	*y*	*y*	*n*	*y/n*	*y/n*			
False	0	0	0	0	1	0	1	1	0	4	1	0

Notes: AIC, Akaike's information criterion. Qualitative assessment was also done for some cases based on Figs. 3–6, whether a model indicates the unimodal curve $(x7)^2$ and the variable interactions among $x1$, $x4$, and $x9$ (y as inferable; n as not inferable; y/n as inferable if the association is presumed or interaction detection algorithm is used). The number of predictors falsely selected with statistical significance ($P < 0.05$) is counted ("False" as false positive).

$^*P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$.

As the result suggested, SML algorithms may perform better than linear regressions to find higher-order interactions. The linear regression models, even though they included the higher-order interactions in their model structure, did not capture them (Table 4). By contrast, ctrees and mobtrees visually suggested these interactions without explicitly designing any variable interactions (Figs. 3, 4). Partial dependence plots with random forests can indicate higher-order interactions of nonlinear relationships (Fig. 6d–f). However, such interactions cannot be seen unless they are specified in partial dependence plots. We may have a risk to just interpret the relationships by only confirming the $x9$–Y relationship (Fig. 6a–c). While currently lacking the ability to quantify statistical significance of higher-order interactions using random forests, recently such techniques are emerging (see Basu et al. 2017, a technique that might be able to indicate more than five-way interactions).

Caution is required when interpreting the structure of tree models. Because of the rule of recursive dichotomous separation, ctree and mobtree algorithms are not ideal tools to represent continuous gradual patterns such as curve. For the ctree models, the dichotomous separation of $x9$ (node 4 in Fig. 3b, c) should not be interpreted as the $x9$–Y relationship changing discontinuously. The mobtree models separated $x7$ into three value ranges (nodes 1 and 2 in Fig. 4b, c), although $x7$ was designed as a unimodal peak. Another caveat for interpretation is that connected nodes do not necessarily represent variable interactions (Winham et al. 2012). For example, $x7$ does not actually interact with the others, although the $x7$ node is apparently connected to some others (Fig. 4b, c). A solution for this is to cope with knowledge about $x7$: If the independence of the $x7$ effect from the $x1^*x4^*x9$ is known, we can also build a multiple regression that better represents the designed relationships (Fig. 7).

The most important caveat is that SMLs are not the tools that automatically reveal what the user wants, but artificial intelligence algorithms that assist the user to interpret how the data are structured in a nonlinear and interactive manner. Together with SML approaches, ecological knowledge and theory are surely required. Further applications are needed to better clarify in which circumstances SMLs are more appropriate than machine learning and statistics, and we already present an initial comparison among these (Table 2). These approaches should be used complementarily, depending on study aim and dataset characteristics (Breiman 2001*a*). Even so, we believe that SML tools offer a unique and attractive empirical modeling opportunity for ecological studies. Aiming to further stimulate applications of SML approaches, we introduce five applications in ecological contexts as follows:

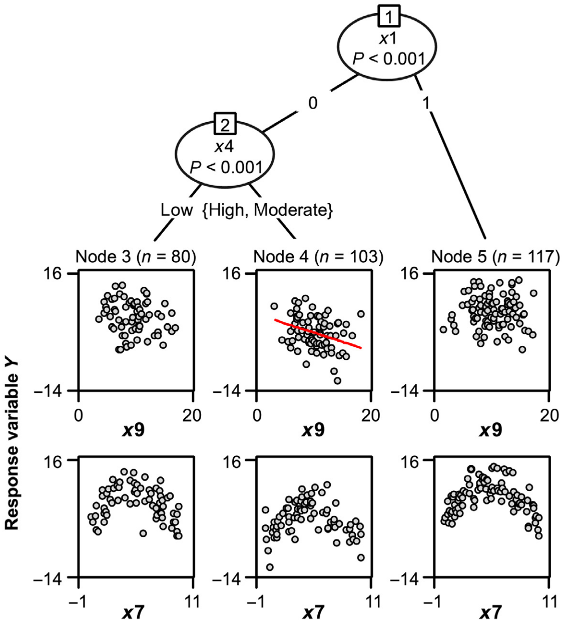

Fig. 7. Model-based tree that presumed linear relationships of *x*9–*Y* and *x*7–*Y* (*n* = 300) indicated the unimodal curve of the *x*7–*Y* relationship (see Fig. 4c for comparison). This indicates that adequate a priori variable selection by expert knowledge enhances the ability of machine learning.

Novel patterns

Statistical designs and assumptions based on ecological knowledge can bias our ways of interpreting observed patterns. For example, using a Kernel density estimation (i.e., a nonlinear non-parametric approach), Chaudhary et al. (2016) showed that the latitudinal gradient in species richness is "bimodal" with a dip near the equator for marine species, while it has been generally believed to be unimodal. The authors also mentioned that the pattern may further change due to climate change. Another concern is seen in the emergence of novel ecosystems that drive patterns and processes which have never occurred previously within a given biome (Hobbs et al. 2006). In a changing climate and with increasing anthropogenic stressors, we may need to reshape classical knowledge and theory on patterns.

Statistically reinforced machine learning is also a promising approach in fields where prior knowledge is still minimal and hypotheses have not been clearly developed (Hochachka et al. 2007) such as biodiversity science (Kelling et al. 2009, Rillig et al. 2015*b*), macrosystem ecology (Heffernan et al. 2014, Levy et al. 2014), and microbial ecology (Baldi and Brunak 2001). Pattern-mining approaches such as using permutation-based random forests in combination with partial dependence plots (Fig. 6) provide an opportunity to objectively see patterns in observed data without subjective constraints by statistical assumptions, knowledge, and theories (e.g., Bergmann et al. 2017).

Higher-order variable interactions

The presence of higher-order variable interactions in ecological communities causes unpredictable nonlinear dynamics (Billick and Case 1994). Assuming linearity and additivity in statistical models implicitly means that more complex non-additive higher-order interactions are assumed to be negligible or absent. The validity of this assumption, however, is rarely known. Mayfield and Stouffer (2017) revealed that the consideration of higher-order interactions is inevitable to better predict species' performances in communities. Macroecology recognizes that an interplay among factors at multiple hierarchical scales significantly contributes to structuring complex nonlinear patterns, which are often unexpected and unpredictable (cross-scale interaction; Peters et al. 2007, Soranno et al. 2014).

Context dependency

Some ecological patterns as well as processes emerge only in particular conditions or vary along an environmental gradient (e.g., Tonkin et al. 2016). For example, latitude–richness relationships can be either positive, negative, or there can be no trend, depending on, for example, length of the latitudinal gradient examined and regions (e.g., Heino 2011). Burkepile and Hay (2006) suggested that the effects of human alteration of food webs and nutrient availability on primary producers vary among latitudes and primary producers, and with the inherent productivity of ecosystems. A response of an organism to an external forcing may depend on the surrounding environmental condition (Hunter et al. 2014) and its past experience (priming; Rillig et al. 2015*a*, Andrade-Linares et al. 2016, Ryo et al. 2017). Searching variable interactions with the mobtree algorithm can identify such context dependencies (Fig. 4). Furthermore, these approaches can be used to integrate seemingly contradictory patterns (e.g., Heino 2011) in a merged dataset by exploring factors which differentiate patterns.

Prediction power as an alternative to statistical significance and effect size

Statistical models can evaluate statistical significance and effect size of predictors based on a pre-defined assumption on data. Therefore, they might miss important associations that apparently do not exist. In addition, assuming linearity on nonlinear relationships often estimates effect size inappropriately (Gilbert and Bennett 2010). Therefore, the structure of a statistical model can be inappropriately assumed when patterns and underlying mechanisms are not well-known, which, in turn, SML can contribute to investigating such fields with non-manipulative data. Permutation-based random forests can directly evaluate the significance and importance of each predictor in terms of prediction power, which is a more honest indicator to assess variable importance than parametric statistical significance and effect size (Hapfelmeier and Ulm 2013; Fig. 5). The algorithm can evaluate importance of predictors, while automatically taking nonlinearity and variable interactions into account.

Statistically reinforced machine learning + statistical modeling.—Statistically reinforced machine learning approaches are used for confirming patterns (e.g., if linear assumption is valid or not and interactions) and variable selection before designing statistical models. Delgado-Baquerizo et al. (2016) used a permutation-based random forest as a priori variable selection and then applied a structural equation model to confirm their hypotheses. We consider that generalized additive models can help evaluate statistical significance of higher-order interactions that are suggested by SML models.

Acknowledgments

The work was funded by the German Federal Ministry of Education and Research (BMBF) within the Collaborative Project "Bridging in Biodiversity Science (BIBS)" (funding number 01LC1501A). We thank M. Thomas for sharing his knowledge and helpful comments on an earlier version of the manuscript. We also thank two anonymous reviewers for constructive comments.

Literature Cited

Andrade-Linares, D. R., S. D. Veresoglou, and M. C. Rillig. 2016. Temperature priming and memory in soil filamentous fungi. Fungal Ecology 21:10–15.

Baldi, P., and S. Brunak. 2001. Bioinformatics: the machine learning approach. Second edition. The MIT Press, London, UK.

Bastolla, U., M. A. Fortuna, A. Pascual-García, A. Ferrera, B. Luque, and J. Bascompte. 2009. The architecture of mutualistic networks minimizes competition and increases biodiversity. Nature 458:1018–1020.

Basu, S., J. B. Brown, K. Kumbier, and B. Yu. 2017. iterative Random Forests: stable recommendation of high-order interactions among biomolecules. arXiv.org, arXiv:1706.08457v2.

Bergmann, J., M. Ryo, D. Prati, S. Hempel, and M. C. Rillig. 2017. Root traits are more than analogues of leaf traits: the case for diaspore mass. New Phytologist. https://doi.org/10.1111/nph.14748, *in press.*

Billick, I., and T. J. Case. 1994. Higher order interactions in ecological communities: What are they and how can they be detected? Ecology 75:1529–1543.

Bischl, B., M. Lang, L. Kotthoff, J. Schiffner, J. Richter, E. Studerus, G. Casalicchio, and Z. Jones. 2016. mlr: Machine learning in R. Journal of Machine Learning Research 17:1–5.

Breiman, L. 2001*a*. Statistical modeling: the two cultures. Statistical Science 16:199–231.

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

Breiman, L., J. Friedman, C. J. Stone, and R. A. Olshen. 1984. Classification and regression trees. Chapman and Hall/CRC, London, UK.

Burkepile, D. E., and M. E. Hay. 2006. Herbivore vs. nutrient control of marine primary producers: context-dependent effects. Ecology 87:3128–3139.

Campbell, J. L., O. J. Sun, and B. E. Law. 2004. Disturbance and net ecosystem production across three climatically distinct forest landscapes. Global Biogeochemical Cycles 18:1–11.

Campetella, G., Z. Botta-Dukát, C. Wellstein, R. Canullo, S. Gatto, S. Chelli, L. Mucina, and S. Bartha. 2011. Patterns of plant trait–environment relationships along a forest succession chronosequence. Agriculture, Ecosystems and Environment 145: 38–48.

Chaudhary, C., H. Saeedi, and M. J. Costello. 2016. Bimodality of latitudinal gradients in marine species richness. Trends in Ecology and Evolution 31: 670–676.

Cleveland, C. C., et al. 2011. Relationships among net primary productivity, nutrients and climate in tropical rain forest: a pan-tropical analysis. Ecology Letters 14:939–947.

Crisci, C., B. Ghattas, and G. Perera. 2012. A review of supervised machine learning algorithms and their applications to ecological data. Ecological Modelling 240:113–122.

Cumming, G. S. 2016. Heterarchies: reconciling networks and hierarchies. Trends in Ecology and Evolution 31:622–632.

Cutler, D. R., T. C. Edwards, K. H. Beard, A. Cutler, K. T. Hess, J. Gibson, and J. J. Lawler. 2007. Random forests for classification in ecology. Ecology 88:2783–2792.

De'ath, G., and K. E. Fabricius. 2000. Classification and regression trees: a powerful yet simple technique for ecological data analysis. Ecology 81:3178–3192.

Delgado-Baquerizo, M., F. T. Maestre, P. B. Reich, T. C. Jeffries, J. J. Gaitan, D. Encinar, M. Berdugo, C. D. Campbell, and B. K. Singh. 2016. Microbial diversity drives multifunctionality in terrestrial ecosystems. Nature Communications 7:10541–10548.

Dodds, W. K., W. H. Clements, K. Gido, R. H. Hilderbrand, and R. S. King. 2010. Thresholds, breakpoints, and nonlinearity in freshwaters as related to management. Journal of the North American Benthological Society 29:988–997.

Domingos, P. 2012. A few useful things to know about machine learning. Communications of the ACM 55:78.

Douglas, P. K., S. Harris, A. Yuille, and M. S. Cohen. 2011. Performance comparison of machine learning algorithms and number of independent components used in fMRI decoding of belief vs. disbelief. NeuroImage 56:544–553.

Elith, J., and J. R. Leathwick. 2009. Species distribution models: ecological explanation and prediction across space and time. Annual Review of Ecology, Evolution, and Systematics 40:415–436.

Elith, J., et al. 2006. Novel methods improve prediction of species' distributions from occurrence data. Ecography 29:129–151.

Gilbert, B., and J. R. Bennett. 2010. Partitioning variation in ecological communities: Do the numbers add up? Journal of Applied Ecology 47:1071–1082.

Hapfelmeier, A., T. Hothorn, K. Ulm, and C. Strobl. 2014. A new variable importance measure for random forests with missing data. Statistics and Computing 24:21–34.

Hapfelmeier, A., and K. Ulm. 2013. A new variable selection approach using Random Forests. Computational Statistics and Data Analysis 60:50–69.

Hastie, T., R. Tibshirani, and J. Friedman. 2009. The elements of statistical learning: data mining, inference, and prediction. Second edition. Springer, Berlin, Germany.

Heffernan, J. B., et al. 2014. Macrosystems ecology: understanding ecological patterns and processes at continental scales. Frontiers in Ecology and the Environment 12:5–14.

Heino, J. 2011. A macroecological perspective of diversity patterns in the freshwater realm. Freshwater Biology 56:1703–1722.

Hobbs, R. J., et al. 2006. Novel ecosystems: theoretical and management aspects of the new ecological world order. Global Ecology and Biogeography 15:1–7.

Hochachka, W. M., R. Caruana, D. Fink, A. Munson, M. Riedewald, D. Sorokina, and S. Kelling. 2007. Data-mining discovery of pattern and process in ecological systems. Journal of Wildlife Management 71:2427–2437.

Hothorn, T., K. Hornik, M. A. van de Wiel, and A. Zeileis. 2006*a*. A Lego system for conditional inference. American Statistician 60:257–263.

Hothorn, T., K. Hornik, and A. Zeileis. 2006*b*. Unbiased recursive partitioning: a conditional inference framework. Journal of Computational and Graphical Statistics 15:651–674.

Hunter, P. J., G. R. Teakle, and G. D. Bending. 2014. Root traits and microbial community interactions in relation to phosphorus availability and acquisition, with particular reference to Brassica. Frontiers in Plant Science 5:27.

Iwasaki, Y., and S. J. Ormerod. 2012. Estimating safe concentrations of trace metals from inter-continental field data on river macroinvertebrates. Environmental Pollution 166:182–186.

James, G., D. Witten, T. Hastie, and R. Tibshirani. 2013. An introduction to statistical learning: with applications in R. Springer, Berlin, Germany.

Kelling, S., W. M. Hochachka, D. Fink, M. Riedewald, R. Caruana, G. Ballard, and G. Hooker. 2009. Data-intensive science: a new paradigm for biodiversity studies. BioScience 59:613–620.

Levy, O., et al. 2014. Approaches to advance scientific understanding of macrosystems ecology. Frontiers in Ecology and the Environment 12:15–23.

Mayfield, M. M., and D. B. Stouffer. 2017. Higher-order interactions capture unexplained complexity in diverse communities. Nature Ecology & Evolution 1:1–7.

Mingers, J. 1987. Expert systems – rule induction with statistical data. Journal of the Operations Research Society 38:39–47.

Olden, J. D., and D. A. Jackson. 2002. Illuminating the "black box": a randomization approach for understanding variable contributions in artificial neuronal networks. Ecological Modelling 154:135–150.

Olden, J. D., J. J. Lawler, and N. L. Poff. 2008. Machine learning methods without tears: a primer for ecologists. Quarterly Review of Biology 83:171–193.

Peters, D. P. C., B. T. Bestelmeyer, and M. G. Turner. 2007. Cross-scale interactions and changing pattern-process relationships: consequences for system dynamics. Ecosystems 10:790–796.

Phillips, S. J., R. P. Anderson, and R. E. Schapire. 2006. Maximum entropy modeling of species geographic distributions. Ecological Modelling 190:231–259.

Pickett, S. T. A., and M. L. Cadenasso. 2002. The ecosystem as a multidimensional concept: meaning, model, and metaphor. Ecosystems 5:1–10.

R Development Core Team. 2016. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Recknagel, F. 2001. Applications of machine learning to ecological modelling. Ecological Modelling 146: 303–310.

Rillig, M. C., J. Rolff, B. Tietjen, J. Wehner, and D. R. Andrade-Linares. 2015*a*. Community priming—Effects of sequential stressors on microbial assemblages. FEMS Microbiology Ecology 91:1–7.

Rillig, M. C., et al. 2015*b*. Biodiversity research: data without theory–theory without data. Frontiers in Ecology and Evolution 3:1–4.

Ryo, M., C. Yoshimura, and Y. Iwasaki. 2017. Importance of antecedent environmental conditions in modeling species distributions. Ecography. https://doi.org/10.1111/ecog.02925, *in press.*

Scheffer, M. 2009. Critical transitions in nature and society. Princeton Study Complex. Princeton University Press, Princeton, New Jersey, USA.

Soranno, P. A., et al. 2014. Cross–scale interactions: quantifying multi-scaled cause-effect relationships in macrosystems. Frontiers in Ecology and the Environment 12:65–73.

Strobl, C., A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis. 2008. Conditional variable importance for random forests. BMC Bioinformatics 9:307.

Strobl, C., A.-L. Boulesteix, A. Zeileis, and T. Hothorn. 2007. Bias in random forest variable importance measures: illustrations, sources and a solution. BMC Bioinformatics 8:25.

Thessen, A. E. 2016. Adoption of machine learning techniques in ecology and earth science. One Ecosystem 1:e8621.

Thomas, M. K., C. T. Kremer, C. A. Klausmeier, and E. Litchman. 2012. A global pattern of thermal adaptation in marine phytoplankton. Science 338: 1085–1088.

Tonkin, J. D., J. Heino, A. Sundermann, P. Haase, and S. C. Jähnig. 2016. Context dependency in biodiversity patterns of central German stream metacommunities. Freshwater Biology 61:607–620.

Urban, M. C., et al. 2016. Improving the forecast for biodiversity under climate change. Science 353: aad8466.

van Nes, E. H., B. M. S. Arani, A. Staal, B. van der Bolt, B. M. Flores, S. Bathiany, and M. Scheffer. 2016. What do you mean, tipping point? Trends in Ecology and Evolution 31:902–904.

Wilkinson, D. M. 1999. The disturbing history of intermediate disturbance. Oikos 84:145–147.

Winham, S. J., C. L. Colby, R. R. Freimuth, X. Wang, M. de Andrade, M. Huebner, and J. M. Biernacka. 2012. SNP interaction detection with Random Forests in high-dimensional genetic data. BMC Bioinformatics 13:164.

Zeileis, A., and K. Hornik. 2007. Generalized M-fluctuation tests for parameter instability. Statistica Neerlandica 61:488–508.

Zeileis, A., T. Hothorn, and K. Hornik. 2008. Model-based recursive partitioning. Journal of Computational and Graphical Statistics 17:492–514.

Supporting Information

Additional Supporting Information may be found online at: http://onlinelibrary.wiley.com/doi/10.1002/ecs2.1976/full