REVIEW

Ecological Monographs, 90(4), 2020, e01422

A translucent box: interpretable machine learning in ecology

Tim C. D. Lucas [1]

Big Data Institute, University of Oxford, Old Road Campus, Oxford OX3 7LF United Kingdom

Citation: Lucas, T. C. D. 2020. A translucent box: interpretable machine learning in ecology. Ecological Monographs 90(4):e01422. 10.1002/ecm.1422

Abstract. Machine learning has become popular in ecology but its use has remained restricted to predicting, rather than understanding, the natural world. Many researchers consider machine learning algorithms to be a black box. These models can, however, with careful examination, be used to inform our understanding of the world. They are translucent boxes. Furthermore, the interpretation of these models can be an important step in building confidence in a model or in a specific prediction from a model. Here I review a number of techniques for interpreting machine learning models at the level of the system, the variable, and the individual prediction as well as methods for handling non-independent data. I also discuss the limits of interpretability for different methods and demonstrate these approaches using a case example of understanding litter sizes in mammals.

Key words: *interpretable machine learning; machine learning; model interpretation; phylogenetic regression; random effects; Random Forest.*

Introduction

Machine learning in ecology

Machine learning methods are a collection of techniques that focus on making accurate predictions from data (Breiman et al. 2001, Crisci et al. 2012, Domingos 2012). It differs from the broader field of statistics in two respects: (1) the estimation of parameters that relate to the real world is less emphasized and (2) the driver of the predictions are expected to be the data rather than expert opinion and careful selection of plausible mechanistic models (Breiman et al. 2001, Domingos 2012). High-level machine learning libraries that streamline the full machine learning pipeline (Phillips and Dudík 2008, Pedregosa et al. 2011, Thuiller et al. 2016, Kuhn et al. 2017) have made machine learning easy to use. These techniques have therefore become popular, particularly in the fields of species distribution modeling (Elith et al. 2006, Phillips and Dudík 2008, Thuiller et al. 2016, Golding et al. 2018, Gobeyn et al. 2019) and species identification from images or acoustic detectors (Shamir et al. 2014, Xue et al. 2017, Mac Aodha et al. 2018, Wäldchen and Mäder 2018, Fairbrass et al. 2019). Other uses include any study where prediction, rather than inference, is the focus, such as predicting the conservation status of species (Bland et al. 2015) and predicting behavioral states (Browning et al. 2018). However, machine learning methods have a reputation for being a black box, inscrutable and mindlessly applied.

This reputation is not totally unfounded with a number of factors making machine learning models difficult to interpret. First, they are often nonparametric. They therefore estimate nonlinear relationships between covariates and response variables that are not summarized by a small number of interpretable parameters instead containing huge numbers of parameters often without estimates of uncertainty (Domingos 2012). Second, they often fit deep interactions between covariates (Lunetta et al. 2004). Even simple, two-way interactions in linear models cause confusion (Engqvist 2005, Lamina et al. 2012) and deep, nonlinear interactions are difficult to visualize or understand. Third, fitting machine learning models is often stochastic (Breiman 2001, Glorot and Bengio 2010) and fitting the same model with different starting values may give a totally different set of fitted parameters (though perhaps with similar predictive performance). However, while interpreting machine learning models can be difficult, there is plenty of insight to be gained by appraising these models at the

Manuscript received 27 February 2020; revised 8 May 2020; accepted 22 May 2020. Corresponding Editor: Hannah L. Buckley.
[1] E-mail: timcdlucas@gmail.com

Box 1 The classification of machine learning models used in this review

Parametric, statistical models

This group includes many models commonly used by biologists. They are parametric because their functional form (the shapes that the relationships between covariates and response variables can take) is defined in advance. They are statistical because they will include some kind of likelihood function that makes the model probabilistic. Therefore generalized linear models are included in this category. A common technique to improve prediction is regularization that biases parameter estimates (toward zero in the case of a linear model) to give a simpler model and avoid overfitting. Methods for regularization of linear models include the LASSO and other penalties (Tibshirani 1996, Zou and Hastie 2005), as used by maxent for example (Phillips and Dudík 2008). Other methods include stepwise selection (Hocking 1976) or Bayesian regularization (Park and Casella 2008, Liu et al. 2018).

Nonparametric, statistical models

These models are fitted in a formal statistical framework as above but the functional form is not defined in advance. Instead, flexible curves are fitted. This group includes splines (and GAMs that combine splines and other linear terms) and Gaussian process regression (Rasmussen 2004). These methods have principled uncertainty estimates due to being statistical. Furthermore, while the nonparametric components are often not represented by a small number of interpretable parameters, they are often controlled, and regularized, by a small number of hyperparameters. If these hyperparameters are fitted in a hierarchical framework (as is common) then they are can be interpreted with associated uncertainty.

Nonparametric, non-statistical models

These methods encompass many more algorithmic methods (Crisci et al. 2012) such as decision trees, ensembles of trees like Random Forest (Breiman 2001) and boosted regression trees (Friedman 2001, Elith et al. 2008), neural networks (Ripley 2007), support vector machines (Cortes and Vapnik 1995), and nearest neighbor (Altman 1992). Each model has benefits but the variety of non-statistical machine learning methods is reviewed elsewhere (Crisci et al. 2012). These methods are not fully probabilistic, and often have large numbers of parameters that do not have uncertainty estimates. Uncertainty estimates for predictions typically have to be computed using bootstrapping. Regularization is achieved in these models in many different ways but in all cases the aim is the force the model to fit a smoother relationship between covariates and response values.

level of the system, the variable and the individual prediction as well as by considering autocorrelative structure in the data.

Given the black box reputation, one might wonder why we should bother interpreting machine learning models; if the predictions are good, then the objective has been achieved. However, any predictions that may be used to make decisions (i.e., any predictions of interest) should be examined carefully. Particular examples of this include predictions used for conservation policy or health care (Vayena et al. 2018). Careless predictions can have severe effects on the entity for which the predictions are being made (an endangered species or a person at risk of a disease, for example) and can more generally erode trust between modelers, policy makers, and other stakeholders. In regulated fields such as healthcare, these considerations come with legal backing. Appraising models as part of model verification has been the primary driver of research in the field of interpretable machine learning (Ribeiro et al. 2016*c*, Molnar 2018). Broadly, we might want to carefully quantify how accurate the predictions from our models are, consider whether the fitted relationships between covariates and response variables are plausible and understand why an individual prediction is given.

However, there are further reasons to interpret machine learning models that apply to fields that are further removed from policy decisions (Elith and Leathwick 2009). The same traits that make machine learning models good at prediction and difficult to interpret also makes them potentially useful in exploratory analysis before more formal statistical modeling (Nosek et al. 2012, Gelman and Loken 2014, Zhao and Hastie 2019). First, they can be useful as a global-level baseline to compare how well a mechanistic model performs. Second, the nonparametric nature of many machine learning models means they can discover nonlinear relationships between covariates and response variables and interactions between covariates without specifying them a priori as would be required in more statistical modeling. It is also worth noting that standard statistical models are often not as interpretable as they seem; understanding the results from a statistical model is made more difficult in the presence of collinearity between covariates or when nature's true model is not in the set of models being considered (Yao et al. 2017,

Lyddon et al. 2018). Therefore, in some cases it might be better to fit a more predictive model and sacrifice some, but not all, interpretability. Alternatively, it might be useful to use a highly predictive model to generate hypotheses that could then be tested in a more formal statistical framework (Zhao and Hastie 2019).

An overview of machine learning

Throughout this paper, the working definition of machine learning is any model where the focus is prediction rather than the building of models that directly aim to reflect reality. Supervised learning is a subfield of machine learning and is the archetypal modeling found in biology. The analyst has some response data and some covariates and the task is to predict the response data. Therefore models such as generalized linear models, mixed-effects models and time series modeling would come under supervised learning. If the response variable is continuous, supervised learning is referred to as regression; if the response variable is categorical, with two or more categories, the task is referred to as classification. While there are many different ways you could classify machine learning models, one that is useful for discussions of interpretability is to split models into three groups: (1) parametric, statistical models, (2) nonparametric, statistical models, and (3) non-statistical, nonparametric models (Box 1). In this review, I define statistical models as being those fitted in a maximum likelihood or Bayesian framework with using a likelihood function such that the loss for any given prediction can be given as a probability. If a model with a binomial likelihood predicts that, for a given data point, 50% of trials should be successes, and the observed data contain two out of two successes, the probability of this data, given the prediction, is 0.25. Grounding these models in probability, and the inclusion of explicit noise terms, means that principled uncertainty distributions of both the parameters and predictions can be derived. In contrast, I define non-statistical models as those that are fitted using non-likelihood based objective functions. Often the only way to derive uncertainty intervals from these models is bootstrapping.

The focus on predictive performance and the prevalence of nonparametric models in machine learning means that some concepts that are marginalized in typical ecological statistics become very important. The first is out-of-sample performance. With simple, parametric models, the goodness of fit of a model can be examined by looking at how far the data used to fit the model fall from the model predictions (R^2 of a regression, for example). However, with a nonparametric model, or a model with many covariates, you can easily fit a model that perfectly fits the data it was fitted to. However, predictions from such a model will often be poor as the model was fitting to the noise in the data rather than extracting useful signal; this is termed overfitting. To measure goodness of fit, we need to test how well the model predicts data it has never seen before. This out-of-sample performance can be tested by holding back a portion of the data, or by splitting the data into k chunks and fitting the model k times each time holding out one chunk. While out-of-sample performance is a measure of goodness of fit that is less affected by overfitting, we still need a way to make our models fit the signal rather than noise in the data. This is achieved by regularization. Regularization encompasses many techniques that force the model to be simpler or fit less complicated relationships. The strength of regularization is typically controlled by hyperparameters and often the only way to choose these parameters is to try a range of values and then select the values that provide the best out-of-sample performance. Finally, it is worth noting that there are some differences in nomenclature between the machine learning literature and the statistics literature. Notably, covariates are referred to as features, response data are referred to as labels and data points are referred to as instances. A major shift in the statistical analysis of ecological and evolutionary data in recent decades is the acknowledgement that observational, biological data rarely conform to assumptions of independence due to phylogeny (Felsenstein 1985, Ives and Zhu 2006), space (Diggle et al. 1998, Redding et al. 2017), time (Ives and Zhu 2006), or other categorical, grouping variables (Bolker et al. 2009, Harrison et al. 2018). This issue of autocorrelation is largely under-appreciated in the machine learning literature and only recently have random effects been explicitly built into non-statistical machine learning models (Eo and Cho 2014, Hajjem et al. 2014, 2017, Miller et al. 2017). Most machine learning models make some assumption of independence and estimates of out-of-sample predictive ability can be biased if cross-validation is used without accounting for autocorrelation. There are, however, a number of strategies to mitigate biases caused by autocorrelation and for gaining insight into the random effects themselves. These include simple methods such as including typical covariates to model random effects or preprocessing the data to reduce autocorrelation (Elith et al. 2010). These methods will be examined in more detail in the body of the review.

In this review, I demonstrate how machine learning models can be interpreted and visualized to understand the global properties of the modeled system, the fitted relationships between covariates and response variables and why a particular prediction was made, using methods that are applicable to a wide variety of machine learning models. I then demonstrate methods for handling and interpreting autocorrelation in the data. I demonstrate these methods using an illustrative analysis in which I fit predictive models of litter size to the PanTHERIA data set (Jones et al. 2009), which contains mammalian life history traits. The full analysis is included as a reproducible R (R Core Team 2017) script that reads data directly from online repositories (Data S1: lucas_translucent_full_analysis.R) and is also available online (see *Data Availability*).

Example Analysis

Data

IIn this illustrative analysis, I will examine potential factors relating to the average litter size of mammal species. The PanTHERIA database contains mammalian life history traits collected from the published literature (Jones et al. 2009). Overall, it contains 5,416 species and data on 35 traits, complimented by a further 15 variables calculated by intersecting IUCN shapefiles for each species and remotely sensed geographic data. There are large amounts of missing data for many of the life history traits and these gaps were filled with median imputation as this method is both simple and conservative. As the response variable, I use litter size with a log $(x + 1)$ transform due to the strong left skew and presence of zeroes. As each data row represents a species, the data may not be independent; species with more recent common ancestors may have more similar life history traits. Most analyses of this type of data (Felsenstein 1985, Ives and Zhu 2006, Pellissier et al. 2012, Ferguson-Gow et al. 2014, Gay et al. 2014) would use phylogenetic regression, which includes an estimated phylogeny, converted to a covariance matrix, as a random effect (Magnusson et al. 2017, Orme et al. 2018). Methods for handling non-independence while using machine learning models are demonstrated in the section *Handling Non-Independent Data*.

Model fitting

II fitted four classes of model (with variations) to the data: a linear model with a priori variable selection, a regularized linear model, a statistical, nonparametric Gaussian process model, and a non-statistical Random Forest model. I used fivefold cross-validation to test model accuracy and select hyperparameters. Given the very different levels of flexibility in the models, this out-of-sample test of accuracy is important and given the non-statistical nature of the Random Forest, statistical, within-sample model comparisons such as AIC are not possible. All models were fitted using caret (Kuhn et al. 2017) in R (R Core Team 2017). One major benefit of caret is that most of the procedures presented later for interpreting the models are immediately useable with over 200 machine learning models including up-to-date implementations of various models such as xgboost, h2o, and keras (Chen and Guestrin 2016, Allaire and Chollet 2018, LeDell et al. 2018).

A priori variable selection.—The standard approach for modeling in ecology and comparative biology is to carefully select a relatively small set of covariates based on a priori knowledge of the system (Whittingham et al. 2006). This process ensures that all variables are reasonably likely to be causally important, reduces overfitting, and keeps the number of parameters small. As a baseline model, I fitted a linear model, selecting covariates that the literature suggests are related to litter size. I chose body size (Leutenegger 1979, Tuomi 1980), gestation length (Okkens et al. 1993, Bielby et al. 2007), metabolic rate (White and Seymour 2004), litters per year (White and Seymour 2004), and longevity (Zammuto 1986, Wilkinson and South 2002).

Statistical, parametric models.—If we have many covariates relative to sample size and have minimal a priori knowledge of the system, we may wish to include all the covariates in a linear model but regularize the coefficients. Similarly, if we want to include many interactions or transformed variables (as in maxent [Phillips and Dudík 2008] for example), the number of covariates can grow rapidly and regularization becomes vital. This approach is also sensible if we care more about prediction than about unbiased estimates of parameters. The simplest regularized linear models are ridge regression (Hoerl and Kennard 1970), which includes a penalty on the square of the coefficients, and LASSO (Tibshirani 1996), which penalizes the absolute value of the coefficients and therefore more strongly penalizes smaller values. For the PanTHERIA analysis, I fitted an elastic net, a common model that blends both the ridge and the LASSO penalty. The total strength of the penalty, and the relative contribution of the two penalties were selected using cross-validation. Fig. 1A shows how predictive performance varies with hyperparameter values and we can see that the best performing model had mostly a LASSO penalty and an intermediate amount of regularization.

Nonparametric, statistical models.—Given the parametric nature of the elastic net model, the way to include nonlinear responses and interactions is to define them manually before model fitting. This, however, still imposes important restrictions as it is difficult to know which nonlinear functions are potentially useful and the model is still ultimately constrained by the effects we can think of to include (typically polynomial terms, log and exponential transforms, and sine transforms). In contrast, nonparametric models like Gaussian processes (Rasmussen 2004) or splines (Hastie and Tibshirani 1986) require no pre-specification of functional forms and instead the overall flexibility of the model is controlled with a hyperparameter. Given their statistical nature, the uncertainty estimates around predictions are a natural part of the model and should be well calibrated even if we extrapolate far from the data. For the PanTHERIA analysis, I have fitted a Gaussian process model with a radial basis kernel (Karatzoglou et al. 2004), selecting the scale hyperparameter using cross-validation. We can

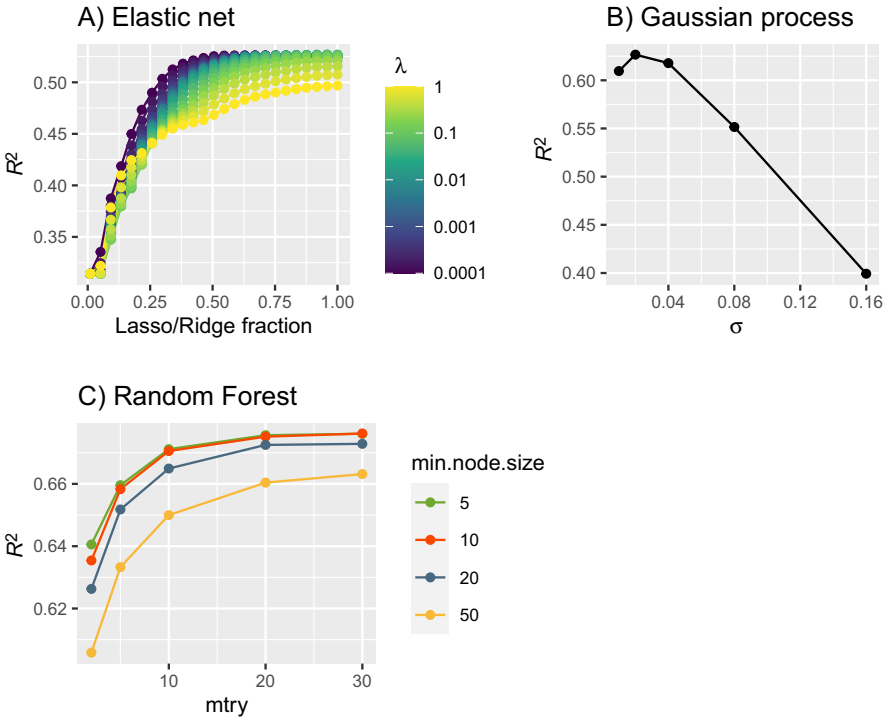

FIG. 1. Model performance against hyperparameter values for each model. Each plot shows the hyperparameter(s) along the *x*-axis with color and model performance (bigger is better) on the *y*-axis. (A) The elastic net model with the proportion of LASSO and ridge regularization (0 is ridge regression, 1 is LASSO regression) shown on the *x*-axis and the overall strength of the regularization shown in color (yellow is strong regularization). (B) The Gaussian process model with the scale parameter sigma on the *x*-axis (large values are smoother, more regularized relationships). (C) The Random Forest model: mtry on the *x*-axis determines the number of randomly chosen covariates considered for splits at each node (lower values give stronger regularization). The minimum node size (shown with color) controls the size of leaf nodes (higher values are stronger regularization).

see (Fig. 1B) that a relatively short scale length, which allows quite nonlinear relationships, had the best model performance.

Nonparametric, non-statistical models.—Finally, I fitted a Random Forest model (Breiman 2001, Wright and Ziegler 2015) as an example of a non-statistical, non-parametric model. I selected Random Forest as they tend to be easy to use, with few hyperparameters, and are robust to overfitting. A Random Forest is an ensemble of decision trees with each tree being fitted to a random bootstrap sample of the input data and each split in each tree selecting from a random sample of the covariates. Random Forests using the ranger package (Wright and Ziegler 2015) via caret have three hyperparameters. Split rule, which determines how the decision tree splits are chosen, was set to "variance." The maximum number of data points at a leaf (min.node.size), which can be used to prevent overfitting was selected by cross-validation. The number of randomly selected covariates to be considered at each split was also selected by cross-validation. We can see that high values of the number of randomly selected covariates and low values of minimum node size had the best performing hyperparameters (Fig. 1B).

Global Properties

The first level at which we can interpret the model is the global level; what do the fitted models tell us about the system as a whole? One global property of interest is how predictable the system is. This can be assessed using scatter plots of observed vs. out-of-sample predictions (Fig. 2) as well as metrics such as r^2 (the proportion of the variance of the held-out observed values explained by the predictions) or the root mean squared error (Table 1). Random Forests are effective here as they are fast to fit, robust and need relatively little tuning. If a Random Forest has poor predictive performance then it is likely that either vital covariates are missing from the data set or that the response is in fact very noisy. The Random Forest model fitted here has fairly good predictive performance (the points in Fig. 2C cluster relatively closely to the one-one line) with an r^2 of 0.68. We can be fairly sure that this trait is not noisy as the evolutionary consequences of litter size are large. Therefore we are probably missing some important covariates, which is a useful result gained solely by interpreting the model at the global level.

It can be seen that species with very large litters, are predicted quite poorly by all models (Fig. 2). Machine

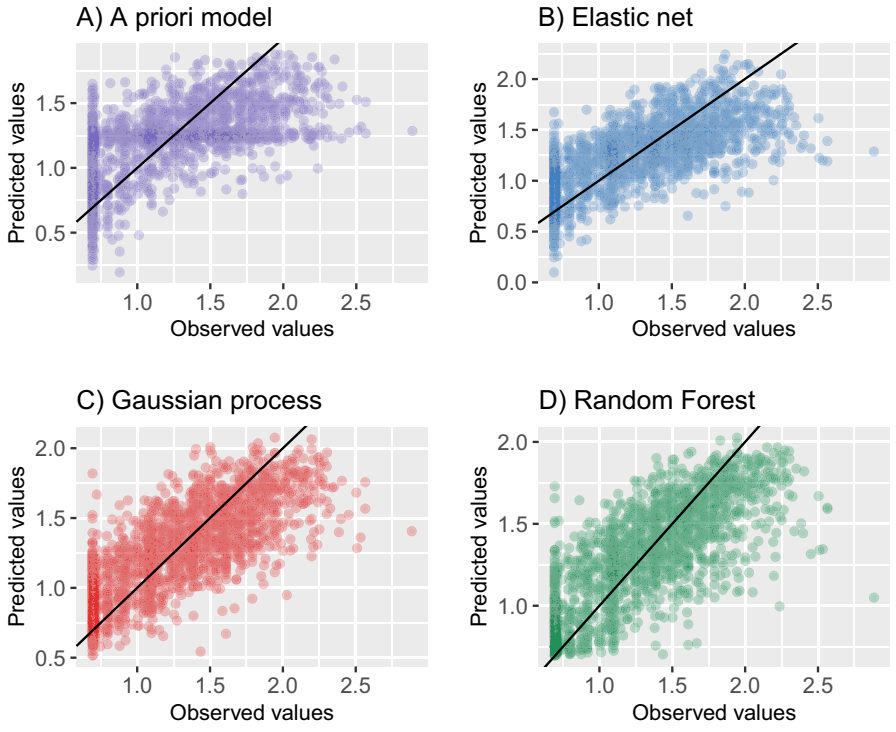

FIG. 2. Predicted vs. observed litter sizes for (A) the a priori selected model, (B) the elastic net model, (C) the Gaussian process model, and (D) the Random Forest model. The black line is the one-to-one line showing the correct predictions. The models are fitted under fivefold cross-validation such that the data being predicted is not used at all during model fitting. We can clearly see that the a priori model is not flexible enough to make good predictions resulting in artefacts such as many data points being predicted at very low values. We can also see in panel D that Random Forests are unable to extrapolate beyond the range of the data (no predictions are below the one-to-one line in the bottom left corner).

TABLE 1. R^2 for all models fitted.

Model	R^2
A priori linear	0.34
Elastic net	0.53
Gaussian process	0.63
Random Forest	0.68
Random Forest with genus	0.70
A priori phylogenetic	0.72
Regularized phylogenetic	0.74
Stacked generalization	0.72
Random Forest with phylogenetic distance	0.81

learning models are known to extrapolate poorly, both at the edges of the range of the response variable and at the extremes of any given covariate (Elith et al. 2010). This is a general issue with nonparametric models and especially an issue with tree-based models like Random Forest. The predictions from nonparametric models are very reliant on data in nearby regions of covariate space. However, there is typically less data at the edges of covariate space and so predictive performance tends to be low. Tree-based methods are, in addition, unable to make predictions higher or lower than the range of the data, which also tends to make their predictive performance of extreme values poor. Analysts should be careful when interpreting predictions or relationships in these areas of model extrapolation.

We can also use predictive performance of machine learning models to scale our expectations for how well a more statistical or mechanistic model fits the data. Here, the linear model with a priori variable selection (Fig. 2A) is unable to accurately model litter size and performed considerably worse than the other three models. Furthermore, the Gaussian process model and Random Forest model performed considerably better than the elastic net model (Table 1). This suggests that some of the covariates included in the elastic net model, but not in the a priori model were useful and that there are nonlinearities or interactions that are important but were not included in either the a priori model or the elastic net model. This is not a suggestion to go back and add these variables to our a priori model. This would amount to severe data snooping and would bias any significance tests performed on the a priori model (White 2000, Gelman and Loken 2014). The suggestion of missing complexity in the a priori model is evidence that the model is misspecified and therefore care should be taken when interpreting even this simple linear

model (Maldonado and Greenland 1993, Lyddon et al. 2018).

We can also attempt to interpret the hyperparameters of our models to try to understand something about the complexity of the system. For the elastic net model, the regularization parameter and the number of non-zero coefficients give us some idea of the system's complexity (Fig. 1A); if very few variables are retained and we get good predictive performance this suggests a simple system. Here we have 0.03 as the selected regularization hyperparameter and only one coefficient being forced to zero. This gives some evidence that this is a complex system not easily explained by a few covariates. Similarly, the length scale, sigma, in the Gaussian process model is a crude measure of complexity, with small values implying that the functional relationships are highly nonlinear (Fig. 1B).

Finally, the Random Forest model has two hyperparameters (Fig. 1C); the number of randomly selected covariates to be examined at each split in each tree and the minimum node size, i.e., the maximum number of data points that can be in a leaf node of a tree. The minimum node size protects against overfitting and gives an indication of how much noise relative to signal there is. Here, the smallest value of minimum node size tested gets elected, which implies there is not much noise in the data relative to signal. The selected value for the number of randomly selected covariates was 20. The number of randomly selected covariates can be difficult to interpret and depends on the number of covariates included in the model. Very small values imply little or no interactions between covariates while intermediate or high values indicate that there are interactions between covariates. However, large values like the 20 selected here do not indicate interaction depths of 20. Instead, it more likely implies that there are many uninformative covariates and so 20 covariates are needed to avoid splits that consider no useful covariates.

Variable-Level Properties

The second level at which we can try to interpret the model is that of the covariate. We can examine variable importance (Oppel et al. 2009), importance of interactions between pairs of covariates and start to examine the functional responses of covariates. It is important, however, to remember that these models are not designed for inference; the following methods should be thought of as hypothesis generation and more formal, subsequent tests (on a different data set) would be needed to confirm relationships between covariates and the response variable. Table 2 shows the top five most important variables as determined by the three models (Oppel et al. 2009). These importance measures are not in absolute units so they are scaled such that the most important covariate has a value of 100. For the regularized linear model, variable importance is given simply by the magnitude of the regression coefficients (i.e., ignoring the sign) and these raw values might be more useful than the scaled importance values. We can see that gestation length comes top for all three models and that latitude and potential evapotranspiration rate in the species' range (PET) are prominent in all three as well. Fitting multiple models and searching for consistency is one useful way to increase confidence in results (as in Appelhans et al. 2015). The fact that gestation length is found to be important also highlights the issue of causality; it is not clear which direction causality flows between gestation length and litter size. Do large litter sizes force gestation length to be small or does short gestation length allow large litters? It could also be true that causality flows in different directions in different species.

Caret provides an easy interface for getting variable importance measures for many model types; however, the calculations being performed differ between models. For some models, including Random Forest, there are different methods for calculating variable importance (Oppel et al. 2009, Wright et al. 2016, Basu et al. 2018, Seifert et al. 2019) and some are more correct than others, especially in the presence of categorical variables. Here I have used the default variable importance given by caret; given there are no categorical variables this is acceptably unbiased. If more accurate variable importance measures are needed, a related model, conditional inference forests (Hothorn et al. 2006) or the maxstat split rule in ranger (Wright et al. 2017), should be used instead.

It is also worth noting that the replicability of variable importance differs between model types and depends on the data set. For example, repeatedly fitting a neural network to these data gives very different results each time. In contrast, linear models, Gaussian processes, and Random Forest generally give the same results each time.

Table 2. Variable importance values.

Variable	Importance
Elastic net model	
Gestation length	100
Adult body mass	70.67
Mid range latitude	67.42
Minimum range latitude	62.73
Potential evapotranspiration rate (range mean)	61.68
Gaussian process model	
Gestation length	100
Adult body mass	70.67
Mid range latitude	67.42
Minimum range latitude	62.73
Potential evapotranspiration rate (range mean)	61.68
Random Forest model	
Gestation length	100
Adult body mass	58.065
Adult forearm length	26.915
Mid range latitude	25.800
Potential evapotranspiration rate (range mean)	22.957

Furthermore, variable importance in the presence of collinearity is less reliable and less interpretable (Dormann et al. 2013). Given two collinear variables, some models such as Random Forest will share the variable importance between them potentially masking an important variable. In contrast, other models such as stepwise regression might put all the variable importance into one variable with no guarantee that the correct variable is selected.

Once some important covariates have been identified, it is useful to examine the shape of the relationship between covariate and response. The simplest way to do this is a partial dependence plot (PDP; Friedman 2001). This plot is created by evaluating the model $n \times m$ times, with all but one covariates taking their values from the n data points and the covariate of interest taking m equally spaced values. The mean response for each value of the covariate of interest is then plotted. For the regularized linear model all the responses are, by definition, linear so a PDP is not particularly useful but is included (Fig. 3A) for reference and comparison. The PDPs for gestation length for the Gaussian process and Random Forest models are shown in Fig. 3B, C. It can be seen that neither response is linear and are both decreasing for low values of gestation length. However, the PDP for the Gaussian process model is increasing at high values of gestation length and is similar to a squared curve. In contrast, the Random Forest model is flat at high values of gestation length.

While PDPs are computed as the mean of the response over the data set, the variable importance measures calculated above are evaluated over all training data. There can therefore be a mismatch where a PDP looks flat while the variable importance is high. Relatedly, the PDP gives no information on interactions because only one curve is plotted. To address these issues we can calculate the interaction importance for each covariate. Table 3 shows the interaction importance values for the Gaussian process and R. This value is given by decomposing the prediction function into contributions from just the focal covariate, contributions from everything except the focal covariate and contributions that rely on both the focal covariate and non-focal covariates together (Friedman and Popescu 2008).

Once we have identified covariates with important interactions we can use individual conditional expectation (ICE) plots. Like PDPs, ICE plots calculate the predicted response value across a range of the focal covariate. However, instead of averaging over the data set, they plot one curve for each data point (Fig. 4A–C).

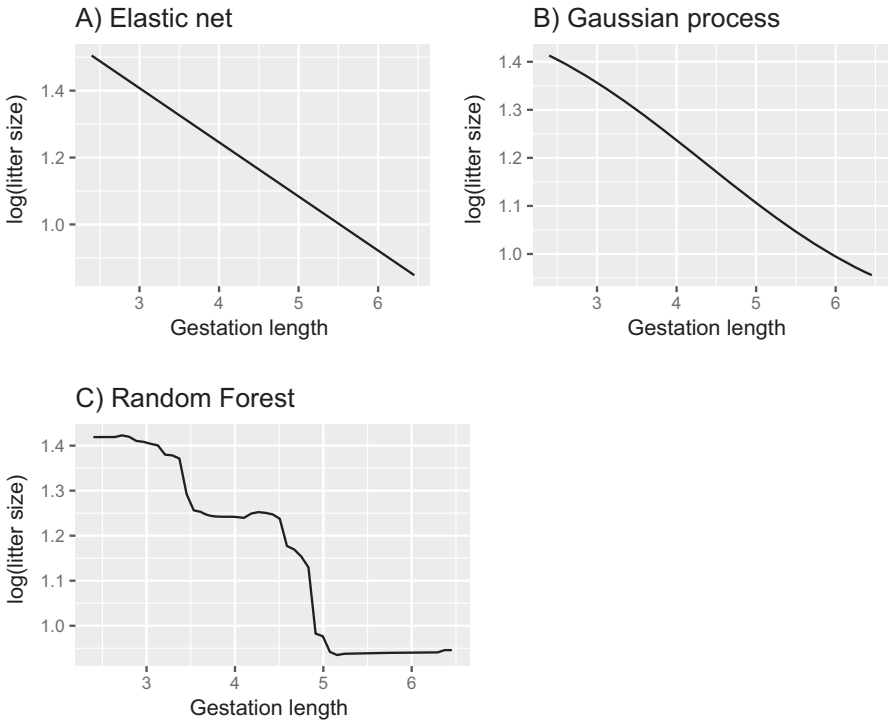

Fig. 3. Partial dependence plot (PDP) for gestation length against log litter size in (A) the elastic net model, (B) the Gaussian process model, and (C) the Random Forest model. These plots show the predicted log litter size for each value of gestation length. Predictions are made at the covariate values of every data point, varying gestation length for each. The final curve is the mean of all n predictions at each value of gestation length. The elastic net model can only fit linear relationships, the Gaussian process model fits smooth nonlinear relationships while the Random Forest model fits more blocky relationships due to the decision trees that underpins it. All three models estimate a broadly decreasing trend. Litter size and gestation length are in arbitrary units.

Table 3. Interaction strengths.

Variable	Interaction importance
Gaussian process model	
Adult body mass	0.20
Age at eye opening	0.20
Terrestriality	0.20
Dispersal age	0.16
Maximum longevity	0.11
Random Forest model	
Adult body mass	0.34
Gestation Length	0.23
Terrestriality	0.15
Maximum range latitude	0.10
Potential evapotranspiration rate (range mean)	0.10

In these plots, we can start to see that the response curve differs depending on what value the other covariates take. As the number of data points increases, these plots can get very busy and so clustering the curves is useful (Fig. 4D, E). Here we can clearly see the range of responses that exist for a single covariate, with gestation length having a negative relationship with litter size in many cases but a flatter relationship in others.

Gaussian process models and Random Forests implicitly consider deep interactions that become increasingly difficult to interpret. However, if we can identify important two-way interactions we can start to interpret these. We can find the interaction strength between two features in a similar fashion to finding variable importance. We can consider the 2D PDP of two covariates (Fig. 5) and calculate what proportion of the curve is explained by the sum of the two 1D PDPs (Fig. 3). We can therefore take one covariate that we know has strong interactions ("gestation length" as seen in Table 3) and calculate the two-way interaction strength between that covariate and all other covariates (Table 4). Finally, once important interactions have been identified, the 2D PDP can be examined to determine the shape of that interaction (Fig. 5). Looking at the 2D PDP of gestation length and PET for the Gaussian process model we can see that for most values of gestation length we have a decreasing relationship between PET and litter size. However, at high values of gestation length the relationship becomes slightly U-shaped. In the Random Forest model, there is a decreasing relationship between gestation length and litter size for all values of PET.

Data-Point-Level Properties

The third level at which we can try to interpret models is that of the individual prediction (Ribeiro et al. 2016*b, c*, Lundberg and Lee 2017, Pedersen and Benesty 2018). Model interpretation at a single point is a much easier task than interpreting the global model because at a small enough scale the response curve is either flat or monotonically increasing or decreasing, so complex nonlinear curves do not need to be considered.

However, it is difficult to examine the model at all data points. Therefore we must focus our analysis on a few, interesting or important points. Points with the highest or lowest predicted values may tell us something about what factors made these points receive extreme predictions. Alternatively, we might be more interested in a subset of points for an external reason. In the PanTHERIA example we might be particularly interested in one taxonomic group. Alternatively, we might want to interpret predictions that are going to be used directly in a conservation program for example.

The method Local Individual Model Evaluation (LIME) examines the behavior of a model at a point by permuting the covariates slightly around the point and making predictions from the model over this new data set (Ribeiro et al. 2016*b, c*, Lundberg and Lee 2017, Pedersen and Benesty 2018). Then a simple, interpretable model, such as ridge regression, is fitted to this data set. As we do not need to consider nonlinear relationships, this simpler model should accurately describe the behavior at the local scale.

In Fig. 6, we can see the outputs of a LIME analysis for the two data points with the highest predicted litter size according to the Gaussian process and Random Forest model. The simple model is a ridge regression model with the 10 covariates with highest weights being plotted. The "explanation fit" values given in the plot are the R^2 of how well the simple, interpretable models explain the predictions of the more complex models. In all cases, the simpler model explains around 60% of the variation in the predictions of the permuted data. It is important to note that as we know the true litter size values for these species we can see that the top-predicted data points are not actually the species with the highest litter size (Fig. 2). This reminds us not to interpret this LIME analysis as "what factors imply the biggest litter size" but rather "why are these particular species predicted as having large litters." We can, however, interpret these outputs alongside the variable importance estimates in Table 2.

If we examine the species shown in the left panel of Fig. 6, we can see that dispersal age is the most important factor in determining why this species has been predicted as having a large litter size. As the bars show negative weights this suggests that it is because the dispersal age of these species is small that they are predicted as having large litter sizes. Indeed, both of these species are in the lowest 3% of dispersal ages among the species in this data set. Another way of phrasing this (that can be more useful for predictions in the center of the range of response values) is that, if the species had larger values for dispersal age, we would expect them to have smaller predicted litter sizes. Similarly, the variables with positive weights indicate that these species have large

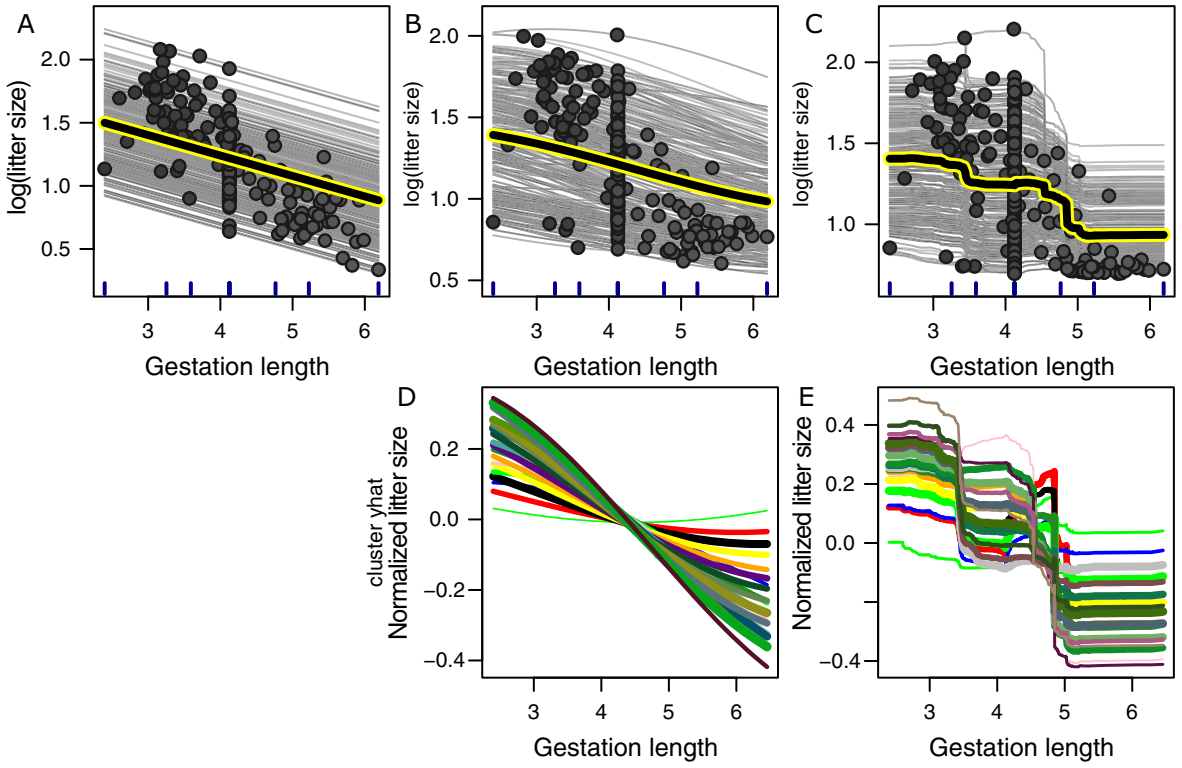

FIG. 4. (A–C) Individual conditional expectation (ICE) plots for gestation length against log litter size in (A) the elastic net model, (B) the Gaussian process model, and (C) the Random Forest model. Each gray curve is the relationship between gestation length and litter size evaluated at one of the data points (only 10% of data points are plotted). The yellow curves are the median of these curves and are therefore similar to PDPs as seen in Fig. 3. The differences in the shapes of the curves are interaction effects; the relationship between gestation length and log litter size changes depending on the values of the other covariates. (D, E) Clustered ICE plot for gestation length against normalized litter size in (D) the Gaussian process model and (E) the Random Forest model. When the number of data points used in the ICE plots gets large it becomes difficult to see individual curves. Clustered ICE plots are created by subtracting the mean of each curve and then clustering them using *k*-means. This normalization step forces all the linear relationships in panel A to be identical and so the clustered plot is not shown.

predicted litter sizes because these covariates are large or that if these covariates were larger, the predicted litter sizes would be larger.

Handling Non-Independent Data

The PanTHERIA data set is an example of a data set that strongly violates assumptions of independence. The autocorrelation here arises due to common ancestry of species; two species that recently diverged from a common ancestor are likely to be more similar than species whose common ancestor is in the deep past. This autocorrelation is typically handled with a phylogenetic random effect (Felsenstein 1985, Ives and Zhu 2006, Pellissier et al. 2012, Ferguson-Gow et al. 2014, Gay et al. 2014) while other sources of autocorrelation such as time or space can be similarly handled with an appropriate random effects term (Diggle et al. 1998, Ives and Zhu 2006, Redding et al. 2017). Categorical random effects can be used to model a wide variety of sources of autocorrelation such as multiple samples from a single individual, study site, or lab (Bolker et al. 2009, Harrison et al. 2018).

Including random effects within parametric or nonparametric statistical models is entirely possible with flexible modeling packages such as Stan, INLA, TMB, or Greta (Rue et al. 2009, Kristensen et al. 2016, Stan Development Team 2016, Golding 2019). As a simple demonstration, I fitted a phylogenetic linear model with INLA (Rue et al. 2009) using the a priori selected covariates ($r^2 = 0.72$) and a phylogenetic linear model using all covariates and strong regularizing priors ($r^2 = 0.74$). For both models, the species-level phylogenetic terms correlate strongly with the species-level fitted value suggesting the phylogenetic effects are important components of the models. This interpretation is supported by the fact that the out-of-sample r^2 values are much higher in the phylogenetic models ($r^2 = 0.72$ and 0.74) than those of the non-phylogenetic models ($r^2 = 0.34$ and 0.53).

However, combining random effects with nonparametric, non-statistical models is more difficult. While these models are starting to be developed (Sela and

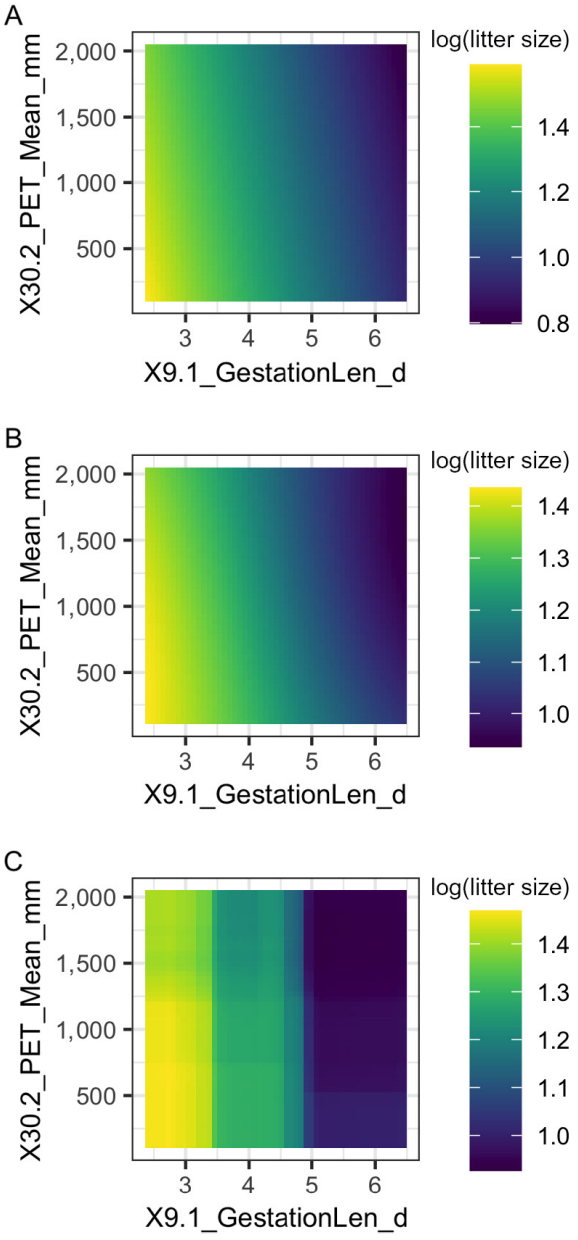

FIG. 5. Two-dimensional PDP plot for gestation length and potential evapotranspiration rate (PET) in (A) the elastic net model, (B) the Gaussian process model, and (C) the Random Forest model. These plots are created in the same way as the one-dimensional PDP plots except now we are varying two covariates (gestation length and PET). As before, we evaluate the models at a range of the values for the covariates of interest and at the covariate values from every data point and then take the mean. In the elastic net model, we can see that the relationship is linear with both covariates and the slopes do not change (i.e., there is no interaction). In the Gaussian process model, we can see that for most values of gestation length we have a decreasing relationship between PET and litter size. However, at high values of gestation length the relationship becomes slightly U-shaped.

TABLE 4. Specific interaction strengths between gestation length and other variables.

Variable	Interaction importance
Gaussian process model	
Estimated body mass at weaning	0.54
Social group size	0.22
Weaning body mass	0.20
Home range	0.17
Age at first birth	0.16
Random Forest model	
Adult forearm length	0.14
Adult body mass	0.12
Weaning age	0.11
Mid range latitude	0.08
Terrestriality	0.07

Notes: Estimated Body Mass at Weaning contains values for species with missing data estimated by the PanTHERIA authors. The values were estimated using a GLM and body mass, neonate body mass and litters per year.

Simonoff 2011, Eo and Cho 2014, Hajjem et al. 2014, 2017, Miller et al. 2017, Ngufor et al. 2019), they are not available in R packages, are only implemented for a small subset of non-statistical models, and do not necessarily benefit from the computational improvements implemented in the most up-to-date packages (Wright and Ziegler 2015, Chen and Guestrin 2016). Therefore, generic methods for handling random effects, that can be used with any machine learning algorithm, are useful. The naïve approach to including random effects within machine learning models would be to simply include them as covariates; categorical random effects as categorical covariates, space or time as continuous variables for example. Interpretation of these variables could then be approached using the methods described in *Variable Level Properties*. However, to understand when this approach is or is not appropriate, we have to examine three factors as to why these effects are not just included as fixed effects in typical mixed-effects models.

First, we expect to extrapolate continuous random effects and expect unseen categories during prediction when using categorical random effects. Many non-statistical machine learning models extrapolate poorly, for example tree-based models will predict a flat response curve outside the range of the data. For an effect such as space, this is undesirable and we would instead typically wish the spatial prediction to return to the mean of the data (Rasmussen 2004, Hengl et al. 2018). Predicting unseen categories of a categorical variable presents problems as well. A categorical variable might often be encoded as a full-rank dummy variable (one dummy variable less than the number of categories) and unseen categories would be implicitly predicted using the fitted value for the first category. This is not how we would wish the model to behave.

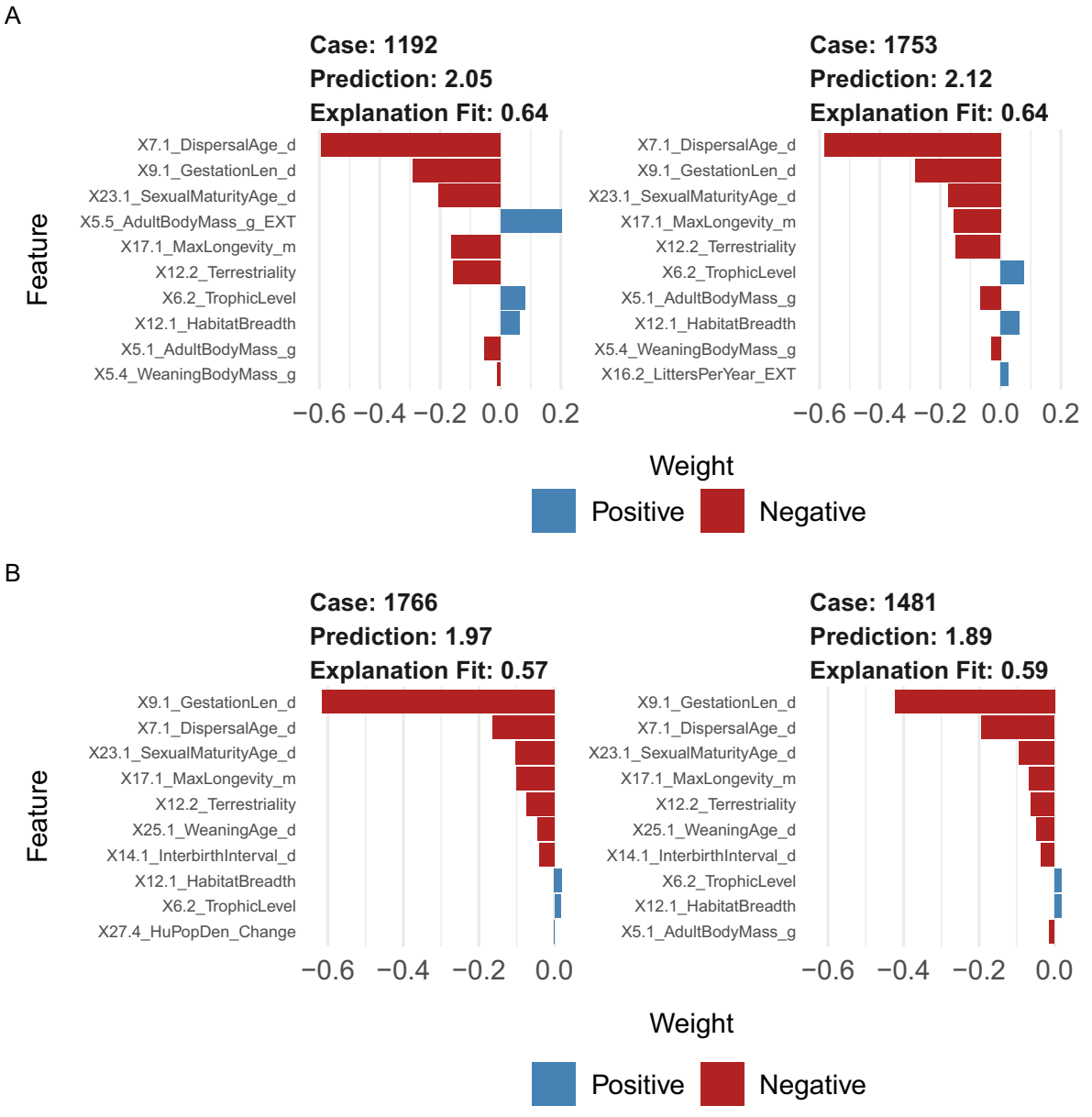

FIG. 6. LIME analysis of predictions of five points from (A) the Gaussian process model and (B) the Random Forest model. The analysis proceeds by taking a data point (the case number), permuting the covariate values around that point, evaluating the model at each permuted covariate value, and then fitting a simple ridge regression model to this new data set. The explanation of fit values is the R^2 for how well the simple model explains the predictions of the more complex models. The absolute value of each bar indicates how important that variable is for this specific prediction. Bars that show negative weights indicate that it is because the value of that covariate is small that high litter sizes are predicted. The bars with positive weights indicate that it is because these values are large that high litter sizes are predicted.

Second, we often have many categories and little data per category in a categorical random effect and wish to share information across groups. This low number of data points per parameter can be reframed as a regularization problem. The regularization can be seen explicitly in the Bayesian formulation of random effects models (hierarchical models) where the random parameters are regularized by a zero centered prior, the strength of which is in turn learned from the data (Simpson et al. 2017).

Finally, random effects are often included as a way to control for autocorrelation rather than being part of the desired predictive model. For example, if all future predictions are to be for unseen categories of a categorical random effect or if all spatial predictions are to be made far from data, then we might want to construct our

model simply so that the model is unbiased by these autocorrelations rather than using them directly in predictions. Similarly, if the data collection was biased with respect to a random effect, we might want to control for this without wanting to use this effect in predictions. For example, if data was collected by different labs or with different protocols, we might want to control for this effect but then predict the latent effect. If the presence of a species is measured using different methods (camera trapping, visual surveys, etc.) we might want to control for this, but we aim to predict the latent state "species presence," not "species presence as measured by camera trapping."

Given these issues, we can consider how to include random effects into the Random Forest model and then examine the results when these are applied to the PanTHERIA analysis. While we are continuing to use Random Forest as our example model, these methods are applicable to most machine learning methods. One way of including phylogenetic information in an analysis is to treat a taxonomic level such as genus as a categorical random effect. While this is less principled than properly including the phylogeny, it is simple. This method also allows a demonstration of categorical random effects.

If we use genus as a categorical random effect to encapsulate some phylogenetic information, the first issue is that we must be careful that the software does not automatically encode the data as a full-rank dummy variable. While less-than-full-rank form would cause identifiability issues with the intercept in a linear model, the random columns and greedy splitting during tree building means we do not have to worry about identifiability in this context while using Random Forest.

The second issue above was that of regularization. Random Forest is natively regularized by the bootstrap aggregation, and the complexity of the model is further controlled by hyper parameters as in Fig. 1C. The new model can therefore be fitted in the same way as the old model. However, given that I added many covariates when creating the dummy variables, I increased the range of values considered for the number of randomly selected covariates.

The final consideration above was the case where we expect all predictions to be made on new categories. In the case of Random Forest, the above methods are suitable. However, given a model that cannot regularize as effectively or if our data set was quite small, we might have wanted to control for genus without including it as a covariate in the model at all. In this case, we can simply weight the data so that each genus is equally represented. Alternatively, we could weight the data so that each genus is represented proportionally to the number of species in each genus in the full prediction set (all mammals for example). Many models in caret accept a weight argument so this is a fairly general solution.

I obtained an r^2 value of 0.70 for the model that used genus encoded as a dummy variables. As this is marginally better than the Random Forest model without genus as a covariate, this provides some evidence that phylogenetic effects are present. The best hyperparameters were 500 for the number of randomly selected covariates (which implies that many of the genera are not very useful on their own) and five for the minimum node size, which is the same as the model without genus as a covariate.

If, however, we wish to include the full phylogeny in our model, we need different methods. The first method is to include all the phylogenetic information in covariates (Hengl et al. 2018). Given the data set of 2,143 data points, we can do this by defining 2,143 new covariates that measure the phylogenetic distance between data points. That is, the first new covariate is the phylogenetic distance between every data point and the first data point. This is then repeated to create 2,143 new covariates. This method is relatively new but is general and can work with any machine learning algorithm. However, interpretation of the strength of the phylogeny will be difficult as it is encoded as 2,143 different covariates. Fitting a Random Forest to this augmented data set was the best performing model out of all tested and gave an r^2 of 0.81.

The second method involves fitting multiple machine learning models and then using phylogenetic regression to "stack" them. We fit a number of machine learning algorithms and make out-of-sample predictions within the cross-validation framework. We then fit a phylogenetic mixed-effects model using the out-of-sample predictions as covariates and constraining the regression coefficients to be positive. This method is likely to be very effective at prediction and the phylogenetic component of the regression is interpretable as it would be in any normal phylogenetic regression. However, this method only corrects for the biases from autocorrelated data after the machine learning models are fitted; while it may still be possible to interpret the machine learning models as we have done previously, the computed nonlinear relationships remain biased. I fitted this model using the three original models (elastic net, Gaussian process regression and Random Forest) and fitted a hierarchical phylogenetic mixed-effects model using INLA (Rue et al. 2009). I obtained a cross-validated r^2 of 0.72. Fitting the model on all the data yielded a posterior mean of 0.03 for the standard deviation of the phylogenetic random effect.

While I cannot demonstrate the handling of spatial or temporal autocorrelation with this data set, the methods described here are equally applicable (Elith and Leathwick 2009). In a method analogous to using genus as a categorical variable, space can be split into regions and the region used as a categorical variable (Appelhans et al. 2015). This approach is commonly used with predefined spatial units such as countries. Another common approach with spatial data is "thinning" and is conceptually similar to the weighting method for categorical data (Elith et al. 2010). In its simplest form, thinning, involves

removing data points so that each spatial pixel has at most one data instance (Elith et al. 2010, Verbruggen et al. 2013). This is equivalent to treating the pixel as a categorical variable and subsampling until each pixel is equally represented (noting that each pixel is represented equally in the prediction data set, i.e., once). Also note that, in the context of presence-only data, this is equivalent to weighting the data but with presence–absence data or continuous response data, weighting is a better way to include all the data. More subtle methods involve removing data based on the local density (Verbruggen et al. 2013). In this method, a kernel bandwidth is chosen either a priori or by cross-validation, then data is probabilistically removed based on the density of data geographically near it. Again, weighting the data may be more principled.

Temporal effects are easier to handle as they are one dimensional with causation only able to occur in one direction. Furthermore, they have been studied in detail in the machine learning literature (Jeong et al. 2008). For regular time series, we can typically include covariates created from the lagged response variable while for irregular time series we can create covariates like "mean response within X units of time previous to this datapoint."

Software

The accessibility to ecologists of machine learning models and methods for interpreting them is largely dependent on the availability of free, user-friendly software. Packages such as caret and the more recent tidymodels allow R users to fit many models within a unified interface (Kuhn et al. 2017, Kuhn and Wickham 2020). The package scikit-learn provides a similar framework for python users. These frameworks also provide functions for tuning hyperparameters, splitting data, and other tasks vital for effective modeling. The specific analyses and visualizations are possible using a number of R packages such as iml, icebox, or pdp for partial dependance plots and lime for LIME analyses (Goldstein et al. 2015, Ribeiro et al. 2016*a*, Greenwell 2017, Molnar et al. 2018). The package dalex, available for python and R, provides pdp, ICE plots, and LIME analyses (Biecek 2018).

Future Directions and Conclusions

It is clear that machine learning is continuing to grow in popularity in ecology. However, it currently remains used almost solely for purely predictive purposes. The full potential of these methods is therefore not being realized. For ecologists to get the most out of machine learning methods, they must be more clear about the purposes of their analyses: is a well-defined hypothesis being tested, is a data set being explored for potential relationships to drive hypothesis generation, or is prediction the main focus? This clarity makes it possible to be clear about the trade-offs in any statistical analysis and to use the most effective tools given the desired outcomes. Using simple linear models is often not optimal if discovery of relationships or predictions are the aim; if a formal hypothesis is being tested, Random Forests are unlikely to be the best choice. Finally, being clear about the aims allows sensible planning on how data will be used in the longer term. If the aim is to discover some relationships and then formally test them, the best use of a given data set may be to split it and use half for discovery and half for hypothesis testing (Nosek et al. 2012, Gelman and Loken 2014). This workflow would not occur to an analyst who was unclear about their task.

While the methods here have been generic machine learning methods, there are a number of approaches for combining mechanistic models and nonparametric models. These include using a mechanistic model as the mean function of a Gaussian process (Rasmussen 2004) or using a mechanistic model as a regularizing prior for a nonparametric model (Lyddon et al. 2018). These methods have great potential for combining the interpretability of mechanistic models and the interpolative predictive ability of nonparametric machine learning models.

Finally, as with all modeling, interpretation of machine learning models requires human input. While many algorithms are objectively tested for various modeling properties, very few have been tested for their ability to aid the human interpreter. Studies that do specifically test this aspect are very welcome (Bastani et al. 2017). This algorithm–psychology interface is an important area of future research.

To conclude, in this review I have demonstrated a number of methods for interpreting machine learning models. These methods were used to examine global properties of the models, describe the ways in which individual variables relate to the response variable, quantify what factors drove individual predictions and finally to explore the autocorrelation structure within the data. Interpreting machine learning models has two major benefits: first, predictions from a better understood model can be better defended in practical applications; second, machine learning methods can be used for exploratory statistics and hypothesis generation. All models lie on a continuum of interpretability and complexity. These models are not as easy to interpret as carefully constrained statistical models. However, this is because they instead allow greater predictive accuracy and allow the analyst to discover relationships they had not hypothesized a priori. The interpretation of these models needs to be done carefully and requires user interaction and exploration in a way that is less vital in interpreting standard statistical models. To this end, machine learning models tend to be better at exploration and hypothesis generation while more robust statistical methods are still needed to formally test most hypotheses. The power here lies in the way a scientist defines their questions and uses machine learning alongside other methods.

Acknowledgments

Thanks go to Katrina Ross, Penelope Hancock, and the Malaria Atlas Project journal club for comments on earlier drafts of the paper. I acknowledge Tristan Cordier and an anonymous reviewer for their useful and considered comments. I would also like to thank Bure Park Nature Reserve, Bicester, where I drafted this manuscript on many walks getting my son to sleep. T. Lucas was supported by grants from the Bill and Melinda Gates Foundation.

Literature Cited

Allaire, J. J., and F. Chollet. 2018. keras: R interface to 'Keras'. R package version 2.2.4. https://CRAN.R-project.org/package=keras

Altman, N. S. 1992. An introduction to kernel and nearest-neighbor nonparametric regression. American Statistician 46:175–185.

Appelhans, T., E. Mwangomo, D. R. Hardy, A. Hemp, and T. Nauss. 2015. Evaluating machine learning approaches for the interpolation of monthly air temperature at Mt. Kilimanjaro, Tanzania. Spatial Statistics 14:91–113.

Bastani, O., C. Kim, and H. Bastani. 2017. Interpreting black-box models via model extraction. arXiv preprint 1705.08504

Basu, S., K. Kumbier, J. B. Brown, and B. Yu. 2018. Iterative random forests to discover predictive and stable high-order interactions. Proceedings of the National Academy of Sciences USA 115:1943–1948.

Biecek, P. 2018. Dalex: Explainers for complex predictive models in R. Journal of Machine Learning Research 19:1–5.

Bielby, J., G. M. Mace, O. R. P. Bininda-Emonds, M. Cardillo, J. L. Gittleman, K. E. Jones, C. D. L. Orme, and A. Purvis. 2007. The fast-slow continuum in mammalian life history: an empirical reevaluation. American Naturalist 169:748–757.

Bland, L. M., B. E. N. Collen, C. David, L. Orme, and J. O. N. Bielby. 2015. Predicting the conservation status of data-deficient species. Conservation Biology 29:250–259.

Bolker, B. M., M. E. Brooks, C. J. Clark, S. W. Geange, J. R. Poulsen, M. H. H. Stevens, and J.-S.- S. White. 2009. Generalized linear mixed models: a practical guide for ecology and evolution. Trends in Ecology & Evolution 24:127–135.

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

Breiman, L., et al. 2001. Statistical modeling: the two cultures (with comments and a rejoinder by the author). Statistical Science 16:199–231.

Browning, E., M. Bolton, E. Owen, A. Shoji, T. Guilford, and R. Freeman. 2018. Predicting animal behaviour using deep learning: GPS data alone accurately predict diving in seabirds. Methods in Ecology and Evolution 9:681–692.

Chen, T., and C. Guestrin. 2016. Xgboost: A scalable tree boosting system. Pages 785–794 *in* Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, New York, New York, USA.

Cortes, C., and V. Vapnik. 1995. Support-vector networks. Machine Learning 20:273–297.

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.

Diggle, P. J., J. A. Tawn, and R. A. Moyeed. 1998. Model-based geostatistics. Journal of the Royal Statistical Society: Series C (Applied Statistics) 47:299–350.

Domingos, P. 2012. A few useful things to know about machine learning. Communications of the ACM 55:78–87.

Dormann, C. F., et al. 2013. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography 36:27–46.

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

Elith, J., M. Kearney, and S. Phillips. 2010. The art of modelling range-shifting species. Methods in Ecology and Evolution 1:330–342.

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 40:677–697.

Elith, J., J. R. Leathwick, and T. Hastie. 2008. A working guide to boosted regression trees. Journal of Animal Ecology 77:802–813.

Engqvist, L. 2005. The mistreatment of covariate interaction terms in linear model analyses of behavioural and evolutionary ecology studies. Animal Behaviour 70:967–971.

Eo, S.-H., and H. J. Cho. 2014. Tree-structured mixed-effects regression modeling for longitudinal data. Journal of Computational and Graphical Statistics 23:740–760.

Fairbrass, A. J., M. Firman, C. Williams, G. J. Brostow, H. Titheridge, and K. E. Jones. 2019. CityNet–deep learning tools for urban ecoacoustic assessment. Methods in Ecology and Evolution 10:186–197.

Felsenstein, J. 1985. Phylogenies and the comparative method. American Naturalist 125:1–15.

Ferguson-Gow, H., S. Sumner, A. F. G. Bourke, and K. E. Jones. 2014. Colony size predicts division of labour in attine ants. Proceedings of the Royal Society B 281:20141411.

Friedman, J. H. 2001. Greedy function approximation: a gradient boosting machine. Annals of Statistics 29:1189–1232.

Friedman, J. H., and B. E. Popescu. 2008. Predictive learning via rule ensembles. Annals of Applied Statistics 2:916–954.

Gay, N., K. J. Olival, S. Bumrungsri, B. Siriaroonrat, M. Bourgarel, and S. Morand. 2014. Parasite and viral species richness of Southeast Asian bats: fragmentation of area distribution matters. International Journal for Parasitology: Parasites and Wildlife 3:161–170.

Gelman, A., and E. Loken. 2014. The statistical crisis in science: data-dependent analysis—a "garden of forking paths"—explains why many statistically significant comparisons don't hold up. American Scientist 102:460–466.

Glorot, X., and Y. Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. Pages 249–256 *in* Y. W. Teh and M. Titterington, editors. Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. JMLR.org, Sardinia, Italy.

Gobeyn, S., A. M. Mouton, A. F. Cord, A. Kaim, M. Volk, and P. L. M. Goethals. 2019. Evolutionary algorithms for species distribution modelling: a review in the context of machine learning. Ecological Modelling 392:179–195.

Golding, N. 2019. greta: simple and scalable statistical modelling in R. Journal of Open Source Software 4:1601.

Golding, N., T. A. August, T. C. D. Lucas, D. J. Gavaghan, E. E. van Loon, and G. McInerny. 2018. The zoon R package for reproducible and shareable species distribution modelling. Methods in Ecology and Evolution 9:260–268.

Goldstein, A., A. Kapelner, J. Bleich, and E. Pitkin. 2015. Peeking inside the black box: visualizing statistical learning with plots of individual conditional expectation. Journal of Computational and Graphical Statistics 24:44–65.

Greenwell, B. M. 2017. pdp: An R package for constructing partial dependence plots. R Journal 9:421–436.

Hajjem, A., F. Bellavance, and D. Larocque. 2014. Mixed-effects random forest for clustered data. Journal of Statistical Computation and Simulation 84:1313–1328.

Hajjem, A., D. Larocque, and F. Bellavance. 2017. Generalized mixed effects regression trees. Statistics & Probability Letters 126:114–118.

Harrison, X. A., L. Donaldson, M. E. Correa-Cano, J. Evans, D. N. Fisher, C. E. D. Goodwin, B. S. Robinson, D. J. Hodgson, and R. Inger. 2018. A brief introduction to mixed effects modelling and multi-model inference in ecology. PeerJ 6: e4794.

Hastie, T., and R. Tibshirani. 1986. Generalized additive models. Statistical Science 1:297–310.

Hengl, T., M. Nussbaum, M. N. Wright, G. B. M. Heuvelink, and B. Gräler. 2018. Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. PeerJ 6:e5518.

Hocking, R. R. 1976. A biometrics invited paper. The analysis and selection of variables in linear regression. Biometrics 32:1–49.

Hoerl, A. E., and R. W. Kennard. 1970. Ridge regression: biased estimation for nonorthogonal problems. Technometrics 12:55–67.

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

Ives, A. R., and J. Zhu. 2006. Statistics for correlated data: phylogenies, space, and time. Ecological Applications 16:20–32.

Jeong, K.-S., D.-K. Kim, J.-M. Jung, M.-C. Kim, and G.-J. Joo. 2008. Non-linear autoregressive modelling by temporal recurrent neural networks for the prediction of freshwater phytoplankton dynamics. Ecological Modelling 211:292–300.

Jones, K. E., et al. 2009. PanTHERIA: a species-level database of life history, ecology, and geography of extant and recently extinct mammals. Ecology 90:2648.

Karatzoglou, A., A. Smola, K. Hornik, and A. Zeileis. 2004. kernlab—an S4 package for kernel methods in R. Journal of Statistical Software 11:1–20.

Kristensen, K., A. Nielsen, C. W. Berg, H. Skaug, and B. M. Bell. 2016. TMB: automatic differentiation and Laplace approximation. Journal of Statistical Software 70:1–21.

Kuhn, M., et al. 2017.caret: classification and regression training. R package version 6.0-76. https://CRAN.R-project.org/package=caret

Kuhn, M., and H. Wickham. 2020. tidymodels: easily install and load the 'Tidymodels' packages. R package version 0.1.0. https://CRAN.R-project.org/package=tidymodels

Lamina, C., G. Sturm, B. Kollerits, and F. Kronenberg. 2012. Visualizing interaction effects: a proposal for presentation and interpretation. Journal of Clinical Epidemiology 65:855–862.

LeDell, E., et al. 2018. h2o: R interface for 'H2O'. R package version 3.20.0.8. https://CRAN.R-project.org/package=h2o

Leutenegger, W. 1979. Evolution of litter size in primates. American Naturalist 114:525–531.

Liu, C., Y. Yang, H. Bondell, and R. Martin. 2018. Bayesian inference in high-dimensional linear models using an empirical correlation-adaptive prior. arXiv preprint 1810.00739.

Lundberg, S. M., and S.-I. Lee. 2017. A unified approach to interpreting model predictions. Pages 4765–4774 *in* I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors. Advances in neural information processing systems 30. Curran Associates, Inc., Red Hook, New York, USA.

Lunetta, K. L., L. Brooke Hayward, J. Segal, and P. Van Eerdewegh. 2004. Screening large-scale association study data: exploiting interactions using Random Forests. BMC Genetics 5:32.

Lyddon, S. P., S. G. Walker, and C. C. Holmes. 2018. Nonparametric learning from Bayesian models with randomized objective functions. arXiv preprint 1806.11544.

Mac Mac Aodha, O., et al. 2018. Bat detective–deep learning tools for bat acoustic signal detection. PLoS Computational Biology 14:e1005995.

Magnusson, A., H. J. Skaug, A. Nielsen, C. W. Berg, K. Kristensen, M. Maechler, K. J. van Bentham, B. M. Bolker, and M. E. Brooks. 2017. glmmtmb: Generalized linear mixed models using a template model builder. R package version 0.1, 3. https://github.com/glmmTMB/glmmTMB

Maldonado, G., and S. Greenland. 1993. Interpreting model coefficients when the true model form is unknown. Epidemiology 4:310–318.

Miller, P. J., D. B. McArtor, and G. H. Lubke. 2017. A gradient boosting machine for hierarchically clustered data. Multivariate Behavioral Research 52:117.

Molnar, C. 2018. Interpretable machine learning. https://christophm.github.io/interpretable-ml-book/

Molnar, C., B. Bischl, and G. Casalicchio. 2018. iml: an R package for interpretable machine learning. Journal of Open Source Software 3:786.

Ngufor, C., H. Van Houten, B. S. Caffo, N. D. Shah, and R. G. McCoy. 2019. Mixed effect machine learning: a framework for predicting longitudinal change in hemoglobin A1c. Journal of Biomedical Informatics 89:56–67.

Nosek, B. A., J. R. Spies, and M. Motyl. 2012. Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. Perspectives on Psychological Science 7:615–631.

Okkens, A. C., T. W. M. Hekerman, J. W. A. De Vogel, and B. Van Haaften. 1993. Influence of litter size and breed on variation in length of gestation in the dog. Veterinary Quarterly 15:160–161.

Oppel, S., C. Strobl, and F. Huettmann. 2009. Alternative methods to quantify variable importance in ecology. Technical Report Number 65. Department of Statistics, University of Munich. https://epub.ub.uni-muenchen.de/10992/1/Oppeletal_techreport.pdf

Orme, D., R. Freckleton, G. Thomas, T. Petzoldt, S. Fritz, N. Isaac, and W. Pearse. 2018. caper: comparative analyses of phylogenetics and evolution in R. R package version 1.0.1. https://CRAN.R-project.org/package=caper

Park, T., and G. Casella. 2008. The Bayesian LASSO. Journal of the American Statistical Association 103:681–686.

Pedersen, T. L., and M. Benesty. 2018. lime: local interpretable model-agnostic explanations. R package version 0.4.0. https://CRAN.R-project.org/package=lime

Pedregosa, F., et al. 2011. Scikit-learn: machine learning in python. Journal of Machine Learning Research 12:2825–2830.

Pellissier, L., K. Fiedler, C. Ndribe, A. Dubuis, J.-N. Pradervand, A. Guisan, and S. Rasmann. 2012. Shifts in species richness, herbivore specialization, and plant resistance along elevation gradients. Ecology and Evolution 2:1818–1825.

Phillips, S. J., and M. Dudík. 2008. Modeling of species distributions with maxent: new extensions and a comprehensive evaluation. Ecography 31:161–175.

R Core Team. 2017. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Rasmussen, C. E. 2004. Gaussian processes in machine learning. Pages 63–71 *in* O. Bousquet, U. von Luxburg, and G. Rätsch, editors. Advanced lectures on machine learning. Springer, New York, New York, USA.

Redding, D. W., T. C. D. Lucas, T. M. Blackburn, and K. E. Jones. 2017. Evaluating Bayesian spatial methods for modelling species distributions with clumped and restricted occurrence data. PLoS One 12:e0187602.

Ribeiro, M. T., S. Singh, and C. Guestrin. 2016*a*. Lime: explaining the predictions of any machine learning classifier. https://github.com/marcotcr/lime

Ribeiro, M. T., S. Singh, and C. Guestrin. 2016*b*. Nothing else matters: model-agnostic explanations by identifying prediction invariance. arXiv preprint 1611.05817.

Ribeiro, M. T., S. Singh, and C. Guestrin. 2016*c*. Why should I trust you? Explaining the predictions of any classifier. Pages 1135–1144 *in* Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, New York, New York, USA

Ripley, B. D. 2007. Pattern recognition and neural networks. Cambridge University Press, Cambridge, UK.

Rue, H., S. Martino, and N. Chopin. 2009. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 71:319–392.

Seifert, S., S. Gundlach, and S. Szymczak. 2019. Surrogate minimal depth as an importance measure for variables in Random Forests. Bioinformatics 35:3663–3671.

Sela, R. J., and J. S. Simonoff. 2011. REEMtree: regression trees with random effects. R package version 0.90.3. http://pages.stern.nyu.edu/~jsimonof/REEMtree/

Shamir, L., C. Yerby, R. Simpson, A. M. von Benda-Beckmann, P. Tyack, F. Samarra, P. Miller, and J. Wallin. 2014. Classification of large acoustic datasets using machine learning and crowdsourcing: application to whale calls. Journal of the Acoustical Society of America 135:953–962.

Simpson, D., H. Rue, A. Riebler, T. G. Martins, and S. H. Sørbye. 2017. Penalising model component complexity: a principled, practical approach to constructing priors. Statistical Science 32:1–28.

Stan Development Team. 2016. RStan: the R interface to Stan. R package version 2.14.1. http://mc-stan.org/

Thuiller, W., D. Georges, R. Engler, and F. Breiner. 2016. Package 'biomod2'. https://cran.r-project.org/package=biomod2

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

Tuomi, J. 1980. Mammalian reproductive strategies: a generalized relation of litter size to body size. Oecologia 45: 39–44.

Vayena, E., A. Blasimme, and I. G. Cohen. 2018. Machine learning in medicine: addressing ethical challenges. PLoS Medicine 15:e1002689.

Verbruggen, H., L. Tyberghein, G. S. Belton, F. Mineur, A. Jueterbock, G. Hoarau, C. F. D. Gurgel, and O. De Clerck. 2013. Improving transferability of introduced species' distribution models: new tools to forecast the spread of a highly invasive seaweed. PLoS One 8:e68337.

Wäldchen, J., and P. Mäder. 2018. Machine learning for image based species identification. Methods in Ecology and Evolution 9:2216–2225.

White, H. 2000. A reality check for data snooping. Econometrica 68:1097–1126.

White, C. R., and R. S. Seymour. 2004. Does basal metabolic rate contain a useful signal? Mammalian BMR allometry and correlations with a selection of physiological, ecological, and life-history variables. Physiological and Biochemical Zoology 77:929–941.

Whittingham, M. J., P. A. Stephens, R. B. Bradbury, and R. P. Freckleton. 2006. Why do we still use stepwise modelling in ecology and behaviour? Journal of Animal Ecology 75:1182–1189.

Wilkinson, G. S., and J. M. South. 2002. Life history, ecology and longevity in bats. Aging Cell 1:124–131.

Wright, M. N., T. Dankowski, and A. Ziegler. 2017. Unbiased split variable selection for random survival forests using maximally selected rank statistics. Statistics in Medicine 36:1272–1284.

Wright, M. N., and A. Ziegler. 2015. Ranger: a fast implementation of Random Forests for high dimensional data in C++ and R. arXiv preprint 1508.04409.

Wright, M. N., A. Ziegler, and I. R. König. 2016. Do little interactions get lost in dark random forests? BMC Bioinformatics 17:145.

Xue, Y., T. Wang, and A. K. Skidmore. 2017. Automatic counting of large mammals from very high resolution panchromatic satellite imagery. Remote Sensing 9:878.

Yao, Y., A. Vehtari, D. Simpson, and A. Gelman. 2017. Using stacking to average Bayesian predictive distributions. arXiv preprint 1704.02030.

Zammuto, R. M. 1986. Life histories of birds: clutch size, longevity, and body mass among North American game birds. Canadian Journal of Zoology 64:2739–2749.

Zhao, Q., and T. Hastie. 2019. Causal interpretations of black-box models. Journal of Business & Economic Statistics 1–10. https://doi.org/10.1080/07350015.2019.1624293

Zou, H., and T. Hastie. 2005. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67:301–320.

Supporting Information

Additional supporting information may be found online at: http://onlinelibrary.wiley.com/doi/10.1002/ecm.1422/full

Data Availability

The full analysis is included in Data S1 and is also available from Zenodo: https://doi.org/10.5281/zenodo.3840994