Improving range shift predictions: enhancing the power of traits

Output Outline

Created: 7 January 2018 (Tony Cannistra)

Introduction

Species have been responding to recent climate change by tracking their environment in space or time, adapting or acclimating, or facing declines (Parmesan, 2006), but we are largely unable to predict how particular species will respond (Maguire et al., 2015). Extensive documentation of shifts in distribution and seasonal timing (phenology) reveal that responses vary among species markedly in direction and extent (Rapacciuolo et al., 2014). Detailed empirical studies often succeed in identifying functional traits that govern climate change responses (e.g., Adrian et al., 2006) and consequently functional ecology has been rapidly gaining prominence in climate change ecology (Buckley & Kingsolver, 2013). However, attempts to use traits to predict the relative magnitude of responses among species generally identify traits that are significant, but weak, predictors of climate change responses (Estrada et al., 2016; MacLean & Beissinger, 2017). Across studies, species traits were found to account for ~16% of the among species variation in range shifts and ~42% of the among species variation in phenological shifts (Buckley & Kingsolver, 2013). How can we close the discrepancy between traits predicting responses well in detailed studies but poorly in broad studies? What statistical techniques will allow us to generalize the importance of traits in mediating climate change responses?

Addressing such questions is imperative for anticipating and adapting to the biological impacts of climate change. Indeed, traits are already being used to predict species' sensitivity to climate change in vulnerability frameworks (Williams et al., 2008; Foden et al., 2013; Pacifici et al., 2015). However, the frameworks remain largely untested. Moreover, the traits included in vulnerability analyses differ sufficiently that different frameworks predict dramatically different rankings of vulnerability (Wheatley et al., 2017). Improving the ability of traits to predict climate change responses is necessary for robust vulnerability analyses.

Most attempts to use species' traits to predict the magnitude of their climate change responses rely on linear regression (Buckley & Kingsolver, 2013; MacLean & Beissinger, 2017), yet detailed empirical studies often reveal non-linear relationships between traits and their function (Stenseth & Mysterud, 2002). Unimodal relationships and thresholds are common. For example, extreme diet specialization may drive a species' to track the range shift of a food item (Altermatt, 2010; Diamond et al., 2011), but reducing diet specialization only slightly may alleviate the need for a species' to track its food. Diet generalization could facilitate species' moving to capitalize on newly climatically suitable habitat, yielding a unimodal relationship between diet specialization and the magnitude of range shifts. Likewise, low dispersal ability may prevent a species' tracking its environmental niche (Schloss et al., 2012), but the threshold of dispersal ability may be relatively low for allowing species to niche track. Can statistical techniques that allow for non-linear relationships between traits and species' responses improve our predictive ability?

The universe of modeling techniques is vast, and offers ecologists a wealth of tools to assess whether these nonlinear interactions between traits and range shifts can predict future responses. Standard approaches to capture variable interactions and nonlinearities in linear regressions (such as the explicit inclusion of interacting variables or polynomial expansion) rely on prior knowledge or model selection techniques to determine which variables to select. Other model types, through optimization of model parameters, capture functional forms in natural data which are not attainable using brute-force linear techniques. Their long-standing use in statistics and in the rapidly-developing field of machine learning has provided not only statistically-robust and efficient methods for model fit but also methods for inspecting the predictions of these often opaque models.

The majority of the use of these more complex models by ecologists has been by researchers with a more computational background [Olden et al., 2008], and thus many studies of ecological processes with complex, interacting relationships have not been approached using these methods. There have been studies using advanced machine learning methods in other biological domains (e.g. to identify novel zoonotic disease reservoirs [Han et al., 2015]), and some ecologists have used methods like neural networks [Olden et al., 2008], boosted regression trees [Elith et al., 2008], and random forest classifiers [Cutler et al., 2007] to study complex ecological relationships. Largely, however, these methods have not seen wide use in an ecological context.

Aside from the technical challenges of understanding and deploying these models, another factor contributing to their slow adoption is their relative lack of interpretability. Due to the formulation of models which are able to capture nonlinearities and complex relationships and the optimization procedures used to train them, there are rarely clear coefficients to inspect for an understanding of a model's learned correlations. In this paper we assess two aspects of machine learning-based models as applied to the challenge of predicting species' range shifts. First, we consider whether several models which are able to capture nonlinear relationships can outperform linear models in their predictive ability. Second, we assess the trustworthiness of these predictions using recent developments in model interpretability techniques: does the basis for a model's predictions represent rigorous and supported ecological theory?

Results summary?

Methods

This approach brings together functional trait data, nonlinear modeling approaches, and a framework for interpreting model predictions to determine whether traits can play a more powerful role in predicting ecological responses to climate change.

Modeling Approach

The pool of modeling techniques chosen for this analysis represents three classes of learning algorithms: regularized linear regression, kernel-based regression, and tree-based regression. Regularization is a modification to traditional generalized linear regression which limits the complexity of the learned model to avoid overfitting to the data (Hastie et al., 2009). Several types of regularization exist; for the purposes of this experiment, we chose to use a "ridge"-regularized linear model, which imposes a penalty on the magnitude of each learned coefficient. The cumulative effect of this regularization procedure is a set of coefficients which both minimize prediction error on the training data and prevent overfitting. These coefficients can be interpreted explicitly as with ordinary least squares regression.

While regularization reduces overfitting when compared to a standard least-squares linear fit, regularized linear models are still not able to capture nonlinearities among or interactions between predictive variables. To remedy this, two additional classes of models are employed in our analysis: kernel-based regression and tree-based regression. A “kernel” is a function which projects a set of input data, often into a high-dimensional space, to allow for the linear “separability” of the data for the purposes of classification or regression (Hastie et al., 2009). One kernel-based method employed herein uses a radial basis function (or squared exponential) kernel applied to the training data and fits a ridge-regularized linear model to this transformed input. As a result of this transformation the learned coefficients, while regularized, are not immediately interpretable.

We also evaluate a kernel-based technique known as a support vector machine (SVM). This popular learning method can be formulated for regression, is robust to outliers, and can capture nonlinearities and variable interactions through a similar radial basis function kernel as in the Kernel Ridge approach. Similarly to the Kernel Ridge method, the SVM regressor does not have interpretable coefficients. Finally we train a random forest regression algorithm to evaluate the performance of tree-based methods. All of these models are implemented using in the Python programming language using the scikit-learn software package. All code for this project is available on GitHub in Jupyter notebooks, a “literate programming” format which combines text and executable code viewable in a web browser.

For comparison to the original analyses upon which this work is based, we train an ordinary least squares regression model, which assumes linear relationships and no variable interaction. However we diverge from these prior experiments to follow a common predictive modeling paradigm: while these original analyses were mostly conducted in a single-variable framework (that is, to assess the effect size of M different potential predictive variables M models were trained, each model containing only 1 variable), we include all variables in these analyses to enable the models to capture variable interactions.

Evaluation

A critical element to evaluating predictive models is ensuring that performance evaluation is conducted on data which was not used to fit the model. To assess the predictive performance of our models we employ a k-fold cross validation scheme combined with a squared error loss function. This cross validation technique partitions data points into $k = 10$ subsets; $k - 1$ subsets are used to fit the model (the “training set”), reserving one subset for testing model performance. A mean squared error loss is then computed for the model prediction of range shift magnitude in the reserved test data. This process is repeated k times. A paired t-test is used to assess statistical significance of model model performance using mean error values from cross validation.

Model Interpretation

The core of any basic regression analysis is typically an inspection of the significance of the coefficients of a fitted model. The kernel methods employed herein (Kernel Ridge, SVM) are better equipped to fit complex nonlinear relationships and model interacting variables than standard linear regression, but do not expose any sort of interpretable coefficients. To address this, we utilize the Shapley additive feature value method, proposed in Lundberg and Lee [2017]. Shapley values are computed for each variable by treating the explanation of a given model's prediction as a model in and of itself—values are computed by training an additive method derived from cooperative game theory to learn each variable's contribution to a model's prediction. Explanations are generated from each prediction in the model training set to identify the most important variables during training, and are averaged for each feature across all training examples to generate a whole-model feature importance scale. We perform this procedure for each of the training sets generated by the leave-one-out cross validation scheme described above to compute average feature importance values and their standard deviations. To compare all of the learning techniques, we use either mean regression coefficients (for OLS and Ridge regression), mean Shapley variable importance values (Kernel Ridge and SVM), or mean Gini variable importance values (Random Forest) to rank all variables such that each feature has an importance ranking for each of the several regression methods.

Trait and Range Shift Data

We evaluate our approach using datasets associating trait values with observed historical range shifts gathered by Angert et al. (2011). The first dataset contains traits and elevational range shifts for Swiss alpine plants as established by Holzinger et al. (2008) (N = 139), and the second contains similar information for Yosemite small mammals established by Moritz et al (2008). We remove samples which are missing features, one-hot encode categorical features for regression, and normalize/center the numeric features to have zero mean and unit norm. After this processing the Swiss plants data contains N = 20 species and d = 43 features; the Yosemite mammals dataset contains N = 28 species and d = 26 features.