Predicting Genetic Richness Across Bird Species

Authors: Juliana Knot, Rufus Rock, Jason Tones, Erin Walsh, Ricky Zhao

Background and Introduction

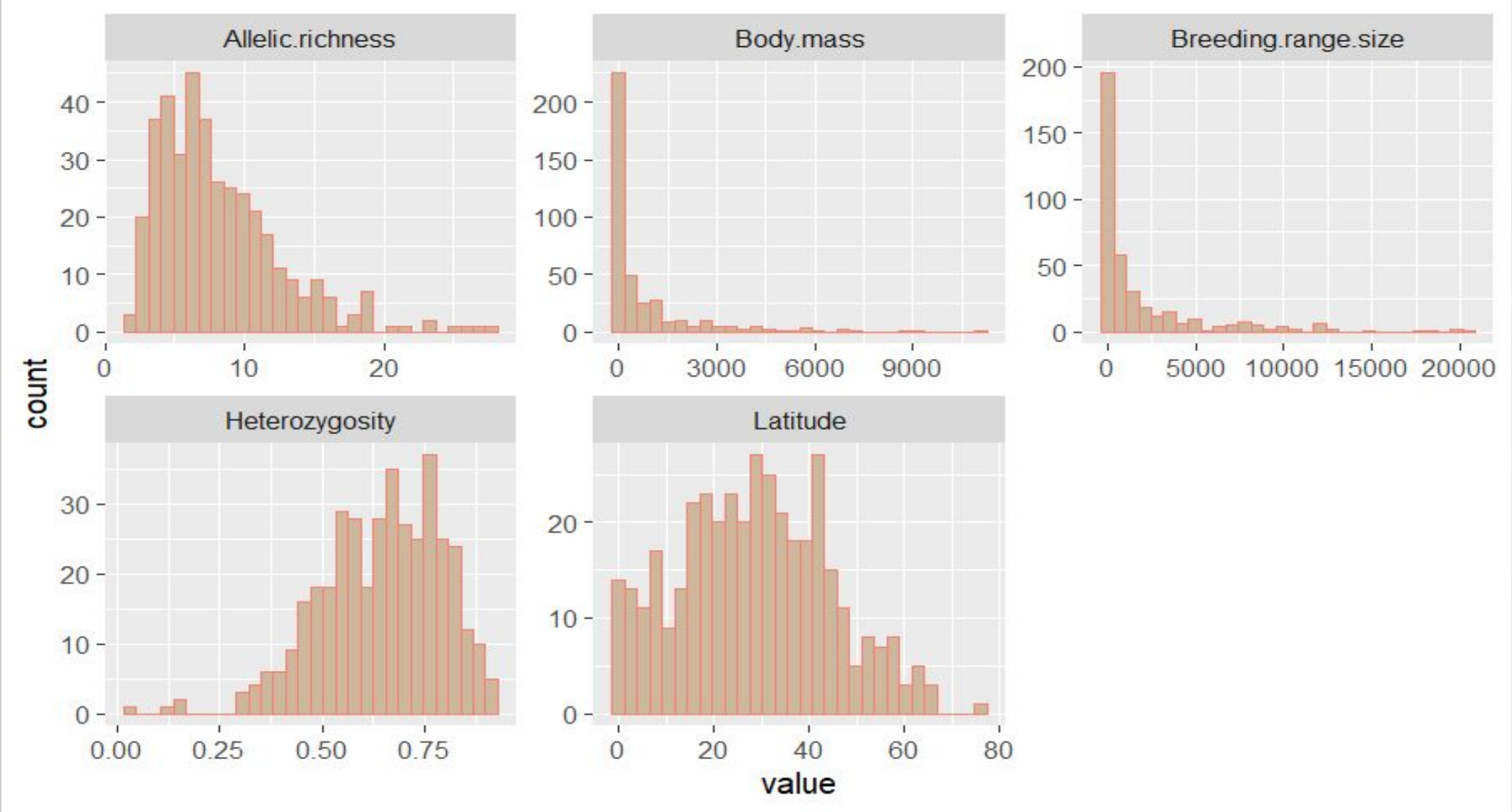
Evolutionary theory suggests that species spread over a wider geographic range should have greater genetic diversity. In this project, we test the extent to which this theory holds for birds. Specifically, we measure different bird species for several variables relating to genes and breeding behavior, and test their ability to predict heterozygosity — a measure of genetic richness for a species.

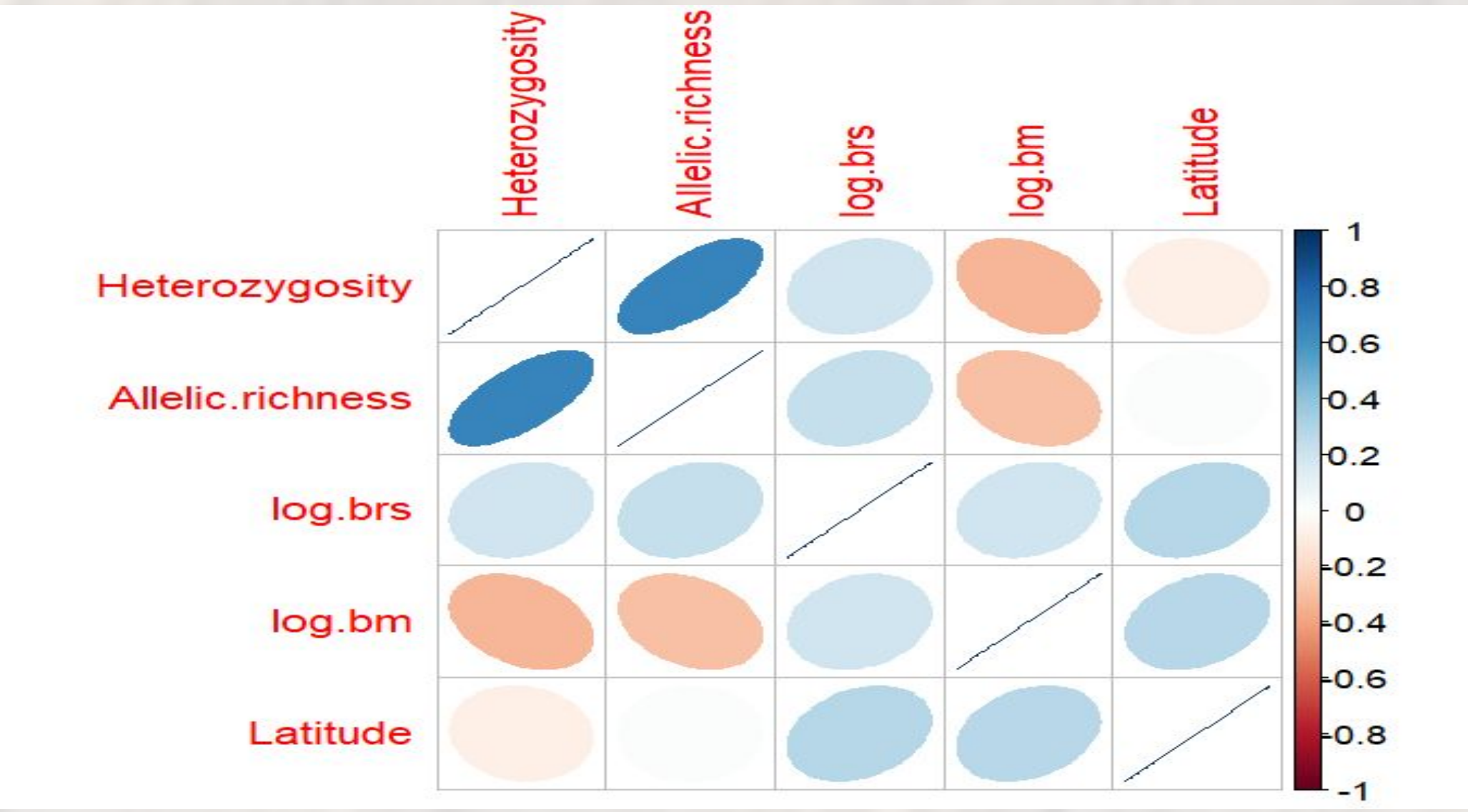
EDA and Data Preprocessing

Our dataset consists of 387 different bird species. For each species we have heterozygosity as a response variable, as well as the following six predictor variables:

Family	The family that the species belongs to
Allele richness	Average number of alleles for each gene
Breeding range size	Measured in units of 10,000 sq km.
Body mass	Average body mass measured in grams
Latitude	Midpoint latitude of breeding range
Migratory status	Binary variable with values Resident or Migratory

Note that two other variables — species name and data reference — were removed from the original data set, as they are statistically uninformative.





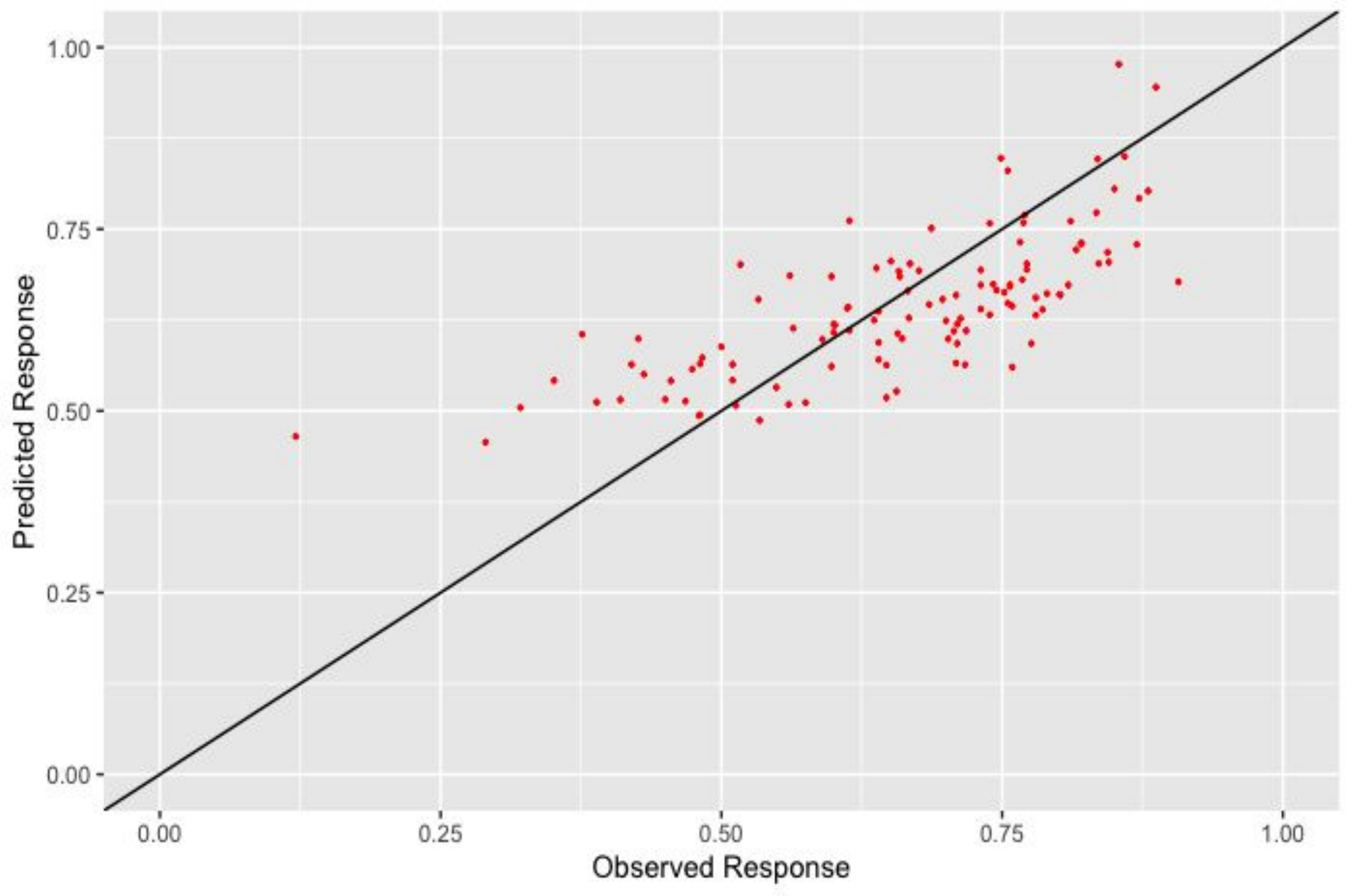
Preliminary data analysis reveals strong right-skewness in body mass and breeding range size, as well as mild right-skewness in allele richness — we mitigate this by performing a log transformation on the former two. The correlation plot on the right shows a strong positive correlation between heterozygosity and allele richness, and a weak to moderate negative correlation between heterozygosity and body mass. There does not appear to be any strong multicollinearity between the quantitative predictor variables.

Methods

We learn the following models: Random Forest, Decision Trees, Linear Regression BSS, SVM, and KNN. For each model we randomly selected 70% of the data to be used as the training set, and the remaining 30% was used as the test set. We used the metric of mean-squared error to assess and compare model quality.

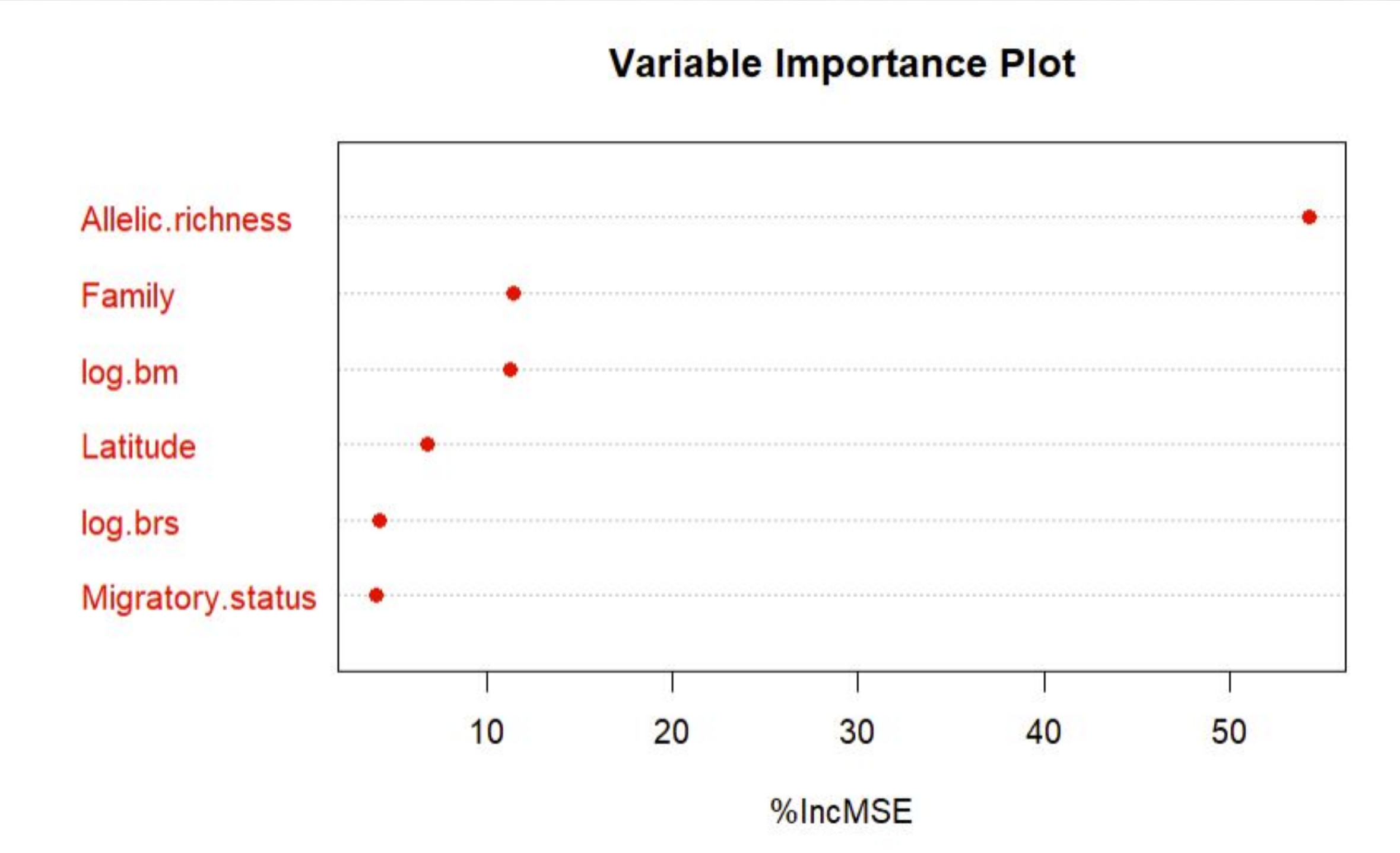
Analysis and Results

Best Subset Selection AIC: Predicted vs. Test Values



The best subset selection model (with penalty based on the Akaike Information Criterion), had the lowest test-set MSE. It retained the variables allelic richness, breeding range size, body mass, latitude and migratory status — all of the original predictor variables except for family.

Random Forest: Variable Importance Plot



Our best-performing machine learning model, Random Forest, ranked allelic richness the most important variable.

Unlike BSS AIC, it considered family the second most important variable.

Model Type	MSE
Linear Regression	0.0112
BSS: BIC	0.0110
BSS: AIC	**0.0105**
Random Forest	0.0111
Decision Tree	0.0117
SVM with linear kernel	0.0117
KNN	0.0118

Conclusions

- Contrary to expectations, migratory status and breeding range are not good indicators of genetic diversity as measured by heterozygosity
- Allelic richness is strongly predictive of heterozygosity
- For the given dataset, machine learning models did not outperform the linear model