

How cryptic is cryptic diversity? Machine learning approaches to classifying morphological variation in *Emys marmorata* (Testudinoidea, Emydidae).

Peter D Smits¹, Kenneth D Angielczyk^{1,2}, and James F Parham³

¹Committee on Evolutionary Biology, University of Chicago

²Integrative Research Center, Field Museum of Natural History

³Department of Geological Sciences, California State University – Fullerton

December 3, 2014

Corresponding author: Peter D Smits, Committee on Evolutionary Biology, University of Chicago, 1025 E. 57th Street, Culver Hall 402, Chicago, IL, 60637, USA; E-mail: psmits@uchicago.edu

Abstract

2

INTRODUCTION

4

MATERIALS AND METHODS

Specimens and sampling

- 6 We collected landmark-based morphometric data from SAMPLE adult *E. marmorata* museum specimens. These specimens include both newly sampled individuals and those sampled in
8 previous studies of plastral shape variation (Angielczyk and Feldman 2013; Angielczyk et al. 2011; Angielczyk and Sheets 2007).
- 10 Specimen classification was based on known specimen geographic information which was recorded from museum collection information. When precise latitude and longitude information
12 was not available it was estimated from whatever locality information was present. Because the specimens used to define the subclades in Spinks and Shaffer (2005) and Spinks et al. (2010)
14 were not available for study, all specimen classifications were based solely on this geographic

information and not from explicit assignment in previous studies. Instead, classification was based on matching museum locality data with the geographic boundaries of the molecularly-defined clades of Spinks and Shaffer (2005) and Spinks et al. (2010). Because the exact barriers between different biogeographic regions are unknown and unclear, two assignments for both the morphologically and molecularly based hypotheses were used. Each morphologically based hypothesis had three classes, while each molecular-based had four classes. In total, each specimen was given four different classifications.

Geometric morphometrics

Following previous work on plastral variation (Angielczyk and Feldman 2013; Angielczyk et al. 2011; Angielczyk and Sheets 2007), 19 landmarks were digitized using TpsDig 2.04 (Rohlf 2005). These landmarks were chosen to maximize the description of general plastral variation (Fig. 1). 17 of these landmarks are at the endpoints or intersection of the keratinous plastral scutes that cover the plastron. 12 of these landmarks were chosen to be symmetrical across the axis of symmetry and, in order to prevent degrees of freedom and other concerns (Klingenberg et al. 2002), prior to analysis these landmarks were reflected across the axis of symmetry (i.e. midline) and the average position of each symmetrical pair was used. In cases where damage or incompleteness prevented symmetric landmarks from being determined, only the single member of the pair was used. Analysis was conducted on the resulting “half” plastra. Plastral landmark configurations were superimposed using generalized Procrustes analysis (Dryden and Mardia 1998) after which, the principal components (PC) of shape were calculated. This was done using the `shapes` package for R (Dryden 2013; R Core Team 2013).

Machine learning analyses

Unsupervised learning.— In order to preserve the relationship between all landmark configurations in shape space, the dissimilarity between observations was measured using Kendall’s Riemannian shape distance or ρ (Dryden and Mardia 1998; Kendall 1984). This metric was chosen because shape space, or the set of all possible shape configurations following Procrustes superimposition, is a Riemannian manifold and thus non-Euclidean (Dryden and Mardia 1998). ρ varies between 0 and $\pi/2$ when there is no reflection invariance, which should not be a concern in the case of the half plastral landmark configurations used in the study.

The ρ dissimilarity matrix was divisively clustered using partitioning around medoids clustering (PAM), a method similar to k -means clustering except that instead of minimizing the sum of squared Euclidean distances between observations and centroids, the sum of squared dissimilarities between observations and medoids is minimized (Kaufman and Rousseeuw 1990). Because the optimal number of clusters of shape configurations in the study was unknown, being possibly three, four, or some other value, clustering solutions were estimated with the number of clusters varied between one and NUMBER. Clustering solutions were

compared using the gap statistic, which is a measure of goodness of clustering (Tibshirani et al. 2001). The gap statistic is defined

$$Gap_n(k) = E_n^*[\log(W_k)] - \log(W_k)$$

where W_k is

$$W_k = \sum_{r=1}^k \frac{1}{2n_r} \left(\sum_{i,i' \in C_r} d_{ii'} \right)$$

$d_{ii'}$ is the dispersion of the clustering solution or the sum of the pairwise dissimilarities between observations in each cluster and their respective medoids (C) for all clusters r . This value is averaged and compared to the expected dispersion (E_n^*) of a sample n from a reference distribution. In this case, the reference distribution was estimated from NUMBER resamples.

This analysis was conducted using the `cluster` package for R (Maechler et al. 2013).

Supervised learning.— Three different supervised learning, or classification, approaches were used: linear discriminate analysis, multinomial logistic regression, and random forests. Linear discriminiate analysis, also known as canonical variate analysis, is commonly used in studies of geometric morphometric data (Mitteroecker and Bookstein 2011; Zelditch et al. 2004). The other two methods, however, are not. Each of these three methods has a different interpretation and reveal very different aspects of the data. In all cases, the optimal number of PCs used as predictors was chosen via max within-sample AUC value, explained below.

Linear discriminate analysis (LDA) attempts to find a linear combination of predictors to best model two or more classes. LDA is very similar to PCA except that instead of finding the linear combination of features that maximize the amount of explained variance in the data, LDA maximizes the differences between classes. The results of this analysis produces a transformation matrix by which the original features can be transformed to reflect the best discrimination between the classes. In this analysis, LDA was applied on the eigenscores from a subset of the total number of PCs, ranging from two to NUMBER in increasing order of complexity. In total, this produced nine different LDA scaling matrices.

Multinomial logistic regression is an extension of logistic regression, where instead of a binary response there are three or more response classes (Venables and Ripley 2002). Similar to the odds ratios calculated from the coefficients of a logistic regression, the relative risk of a classification can be determined from the coefficients of the model. Similar to LDA, the optimal number of PCs as predictors was determined by comparing within-sample AUC values across multiple models.

Random forest models are an extension of classification and regression trees (CART) (Breiman 2001; Breiman et al. 1984). The goal of CARTs are to use a series of different features to

84 estimate the class of an observation. In top-down induction of decision trees for each member
of a given set of predictor variables, attribute value test are used to estimate the differences
86 between classes. This process, called recursive partitioning, is then repeated on each subset.
The recursion continues until the resulting observations all share the same class or no
88 more meaningful partitions are possible. The resulting model is a tree structure by which
observations are classified at each intersection via the estimated cutoff points from the
90 attribute tests made during model fitting.

In a random forest model, many CARTs are built from a random subsample of both the
92 features and the observations. This process is then repeated many times and the parameters
of the final model was chosen as the mode of estimates from the distribution of CARTs
94 (Breiman 2001). In addition to classifying the observations, this procedure allows for the
features to be ranked in order of importance. This is a generally important property that
96 is useful for many other studies which want to describe and model the differences between
classes and the relative importance of different predictors.

98 In this analysis, random forest model parameters were estimated from 1000 subtrees. The
best set of predictors necessary for each classification scheme was estimated using a recursive
100 feature selection algorithm was used to choose the optimal number of PCs to include based
on the AUC of the model. Following the backwards selection algorithm implemented in `caret`
102 (Kuhn and Johnson 2013), the maximum number of features were included in the initial
model, their importance ranked, and the AUC of the model calculated. The lowest ranked
104 feature was then removed, and the AUC of the model recalculated. This was repeated until
only one feature, remained. Because PCs were kept in order of importance and not in relation
106 to the amount of variance each PC described, these means that the PCs are not included in
ascending eigenvalue.

108 In classification studies, like this one, a common metric of performance is area under the
receiver operating characteristic curve (AUC). AUC is an estimate of the relationship between
110 the false positive and true positive rates, as opposed to just the true positive rate (accuracy).
This relationship is especially useful in cases where misclassification needs to be minimized
112 just as much as accurate classification, as in this study. AUC ranges between 0.5 and 1, with
0.5 indicating classification no better than random and 1 indicating perfect classification
114 (Hastie et al. 2009). AUC was used for both LDA and random forest models while multinomial
logistic model selection was done using the likelihood based AICc (Burnham and Anderson
116 2002).

The standard AUC calculation is defined for of binary classification, however in this application
118 there are multiple categories. The alternative calculation used here follows an all-against-one
strategy where the individual AUC values for each class versus all others are averaged to
120 produce a multiclass AUC (Hand and Till 2001).

The ultimate measure of model fit is accurately predicting the values of unobserved samples
122 (Hastie et al. 2009; Kuhn and Johnson 2013). Within-sample performance is inherently biased
upwards, so model evaluation requires overcoming this bias. With very large sample sizes,

as in this study, part of the sample can be used as the “training set” and the remainder acts as the “testing set.” The former is used for fitting the model while the later is used for measuring model performance. This is called model generalization. In this analysis, 75% of samples were used as the training set while the remaining 25% were used as the testing set.

In order to estimate confidence intervals on the out-of-sample AUC values, a nonparametric bootstrap was performed where the true and estimated classifications were resampled with replacement. This was done 1000 times.

It is common for some out of sample observations to be misclassified. This misclassification may be due to the model not accurately representing shape variance, systematic differences between the training and test sets, or systematic differences between the accurately and inaccurately classified samples. Testing and training sets are determined completely at random within each class and with respect to shape. Results were not effected by changes in testing or training set assignment.

To determine if there were systematic differences between the correctly and incorrectly classified samples, the multivariate centroids of the correct and incorrect groups were compared to what would be expected by random chance. The group labels were permuted and the difference between the new centroids was calculated. This was repeated 1000 times. The number of permutations less than the empirical difference were counted and divided by 1000, giving a p -value. Significant results indicate that correctly and incorrectly classified specimens are systematically different. This was done only when there were 10 or of that

RESULTS

Unsupervised learning

Comparison of gap statistic values from PAM clustering show that the best number of clusters is 1. This means that there is no means of naturally partitioning plastron shape into distinct subgroups.

Supervised learning

For multinomial logistic regression, the best number of predictors was always the first four PCs. This was not the case for either of the LDA or random forest approaches which varied from two to six different PCs. Note that the best random forest models were determined via recursive feature selection, so PCs were not included in order of percent variance explained.

For both the LDA and multinomial logistic regression models, increasing model complexity increased cumulative percent variation necessary to best model the differing classification schemes (Fig. 2). The high variance in optimal number of PCs used as predictors between schemes may represent either ease of fitting that particular model or that the data was too noising to accurately fit any model beyond the simplest.

As part of fitting a random forest model, a ranking of variable importance is also determined. Interestingly, the the order of variable importance is not the same as the order of principal components (Fig. 3). This means that the principal components that best describe the differences between the various classes are not aligned with the principal components which describe the largest amount of variance. This result would be the case if variation between classes was extremely fine grained and not a part of the principal function or form of the plastron, which makes sense given that the plastron is involved in both protection and aquadynamics and not mate choice (Germano and Bury 2009; Holland 1992; Lubcke and Wilson 2007; Rivera 2008).

The observed AUC values for all of the optimal models are very low, values near 0.5 indicate that a model is no better than completely random assignment TABLE OR FIGURE. This means that in very few cases are any of the proposed classification schemes are generally poor descriptors of the observed variation. It appears that the data set is overwhelmed by noise, making any accurate classifications difficult at best. This observation is cemented with the generalizations of the models to the testing data set.

AUC values from model generalization, or estimating testing data set membership, does not indicate a clear “best” classification scheme (Fig. 4). While the scheme with two species has the greatest AUC point estimate for each modeling approach, this scheme is not significantly greater from any other except in some limited cases (e.g. LDA).

Mean AUC values for all model generalizations are, as expected, less than from the training set. This is expected because the low initial AUC values indicate little classification ability compared random, and the generalization to new data would only lead to loss in accuracy.

Differences in mean shape between correctly and incorrectly classified observations from test set were frequently statistically significant, though there are exceptions. The frequency of these results, however, is important because this means that the different models are poor predictors of class membership. This may be because differences in plastron shape do not align with the any of the hypothesized classification schemes.

DISCUSSION

Generalization and the unsupervised learning results are surprisingly congruous.

Plastron shape indicates no optimal classification scheme. A possible explanation for this that the genetic divergence associated with (sub)speciation is either not based on plastron morphology or local selective pressures due to hydrological regime overwhelming any possible morphological divergence.

(a) random forest

Scheme	Class	distance	P(distance - simulated > 0)
sh1	CCR	1.59	0.77
	marm	2.06	0.87
sh2	CCR	1.81	0.88
	marm	2.16	1.00
sh3	CCR	2.37	0.94
	marm	2.37	0.99
sh4	marm	2.07	1.00
	pall	2.13	0.99
sh5	marm	1.91	0.85
	pall	2.00	0.94
spinks	1	1.79	0.40
	3	3.30	0.97

(b) multinomial logistic regression

Scheme	Class	distance	P(distance - simulated > 0)
sh1	CCR	2.06	1.00
	marm	2.22	0.93
sh2	CCR	2.50	1.00
	marm	2.60	1.00
sh3	CCR	2.39	0.99
	marm	2.24	0.98
sh4	marm	2.23	1.00
	pall	2.15	1.00
sh5	marm	2.43	0.97
	pall	2.60	1.00
spinks	1	2.96	0.92
	3	3.18	0.99

(c) linear discriminate analysis

Scheme	Class	distance	P(distance - simulated > 0)
sh1	CCR	2.07	1.00
	marm	2.22	1.00
sh2	CCR	2.20	1.00
	marm	1.87	0.98
sh3	CCR	2.75	0.98
	marm	2.36	0.47
sh4	marm	2.23	1.00
	pall	2.15	1.00
sh5	marm	2.43	0.96
	pall	2.60	1.00
spinks	1	2.96	0.90
	3	3.33	1.00

Table 1: Results of comparisons between correctly and incorrectly classified observations from the testing data set. For each scheme, the classifications with at least 10 observations were tested. This was done for each of the three modeling techniques included in this study.

BIBLIOGRAPHY

- Angielczyk, K. D. and C. R. Feldman. 2013. Are diminutive turtles miniaturized? The ontogeny of plastron shape in emydine turtles. *Biological Journal of the Linnean Society* 108:727–755.
- Angielczyk, K. D., C. R. Feldman, and G. R. Miller. 2011. Adaptive evolution of plastron shape in emydine turtles. *Evolution* 65:377–394.
- Angielczyk, K. D. and H. D. Sheets. 2007. Investigation of simulated tectonic deformation in fossils using geometric morphometrics. *Paleobiology* 33:125–148.
- Breiman, L. 2001. Random Forests. *Machine Learning* 45:5–32.
- Breiman, L., J. Friedman, C. J. Stone, and R. A. Olshen. 1984. Classification and regression trees. Wadsworth International Group, Belmont.
- Burnham, K. P. and D. R. Anderson. 2002. Model selection and multi-model inference: a practical information-theoretic approach. 2nd ed. Springer, New York.
- Dryden, I. L. 2013. shapes: Statistical shape analysis. R package version 1.1-8.
- Dryden, I. L. and K. Y. Mardia. 1998. Statistical shape analysis. Wiley, New York.
- Germano, D. J. and R. B. Bury. 2009. Variation in body size, growth, and population structure of *Actinemys marmorata* from lentic and lotic habitats in Southern Oregon. *Journal of Herpetology* 43:510–520.
- Hand, D. J. and R. J. Till. 2001. A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems. *Machine Learning* 45:171–186.
- Hastie, T., R. Tibshirani, and J. Friedman. 2009. The elements of statistical learning: data mining, inference, and prediction. 2nd ed. Springer, New York.
- Holland, D. C. 1992. Level and pattern in morphological variation: a phylogeographic study of the western pond turtle (*Clemmys marmorata*). Ph.D. thesis University of Southwestern Louisiana.
- Kaufman, L. and P. J. Rousseeuw. 1990. Finding groups in data : an introduction to cluster analysis. Wiley, New York.
- Kendall, D. G. 1984. Shape Manifolds, Procrustean Metrics, and Complex Projective Spaces. *Bulletin of the London Mathematical Society* 16:81–121.
- Klingenberg, C. P., M. Barluenga, and A. Meyer. 2002. Shape analysis of symmetric structures: quantifying variation among individuals and asymmetry. *Evolution* 56:1909–1920.
- Kuhn, M. and K. Johnson. 2013. Applied predictive modeling. Springer, New York, NY.

- 224 Lubcke, G. M. and D. S. Wilson. 2007. Variation in shell morphology of the Western Pond
Turtle (*Actinemys marmorata* Baird and Giarard) from three aquatic habitats in Northern
226 California. *Journal of Herpetology* 41:107–114.
- Maechler, M., P. Rousseeuw, A. Struyf, M. Hubert, and K. Hornik. 2013. *cluster: Cluster
228 Analysis Basics and Extensions*. R package version 1.14.4.
- Mitteroecker, P. and F. Bookstein. 2011. Linear Discrimination, Ordination, and the Visual-
230 ization of Selection Gradients in Modern Morphometrics. *Evolutionary Biology* 38:100–114.
- R Core Team. 2013. *R: A Language and Environment for Statistical Computing*. R Foundation
232 for Statistical Computing Vienna, Austria.
- Rivera, G. 2008. Ecomorphological variation in shell shape of the freshwater turtle *Pseudemys*
234 *concinna* inhabiting different aquatic flow regimes. *Integrative and comparative biology*
48:769–87.
- 236 Rohlf, F. J. 2005. *TpsDig* 2.04.
- Spinks, P. Q. and H. B. Shaffer. 2005. Range-wide molecular analysis of the western pond
238 turtle (*Emys marmorata*): cryptic variation, isolation by distance, and their conservation
implications. *Molecular ecology* 14:2047–64.
- 240 Spinks, P. Q., R. C. Thomson, and H. B. Shaffer. 2010. Nuclear gene phylogeography reveals
the historical legacy of an ancient inland sea on lineages of the western pond turtle, *Emys*
242 *marmorata* in California. *Molecular ecology* 19:542–56.
- Tibshirani, R., G. Walther, and T. Hastie. 2001. Estimating the number of clusters in a
244 data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical
Methodology)* 63:411–423.
- 246 Venables, W. and B. D. Ripley. 2002. *Modern applied statistics with S*. 4th ed. Springer, New
York.
- 248 Zelditch, M. L., D. L. Swiderski, and H. D. Sheets. 2004. *Geometric morphometrics for
biologists: a primer*. Elsevier Academic Press, Amsterdam.

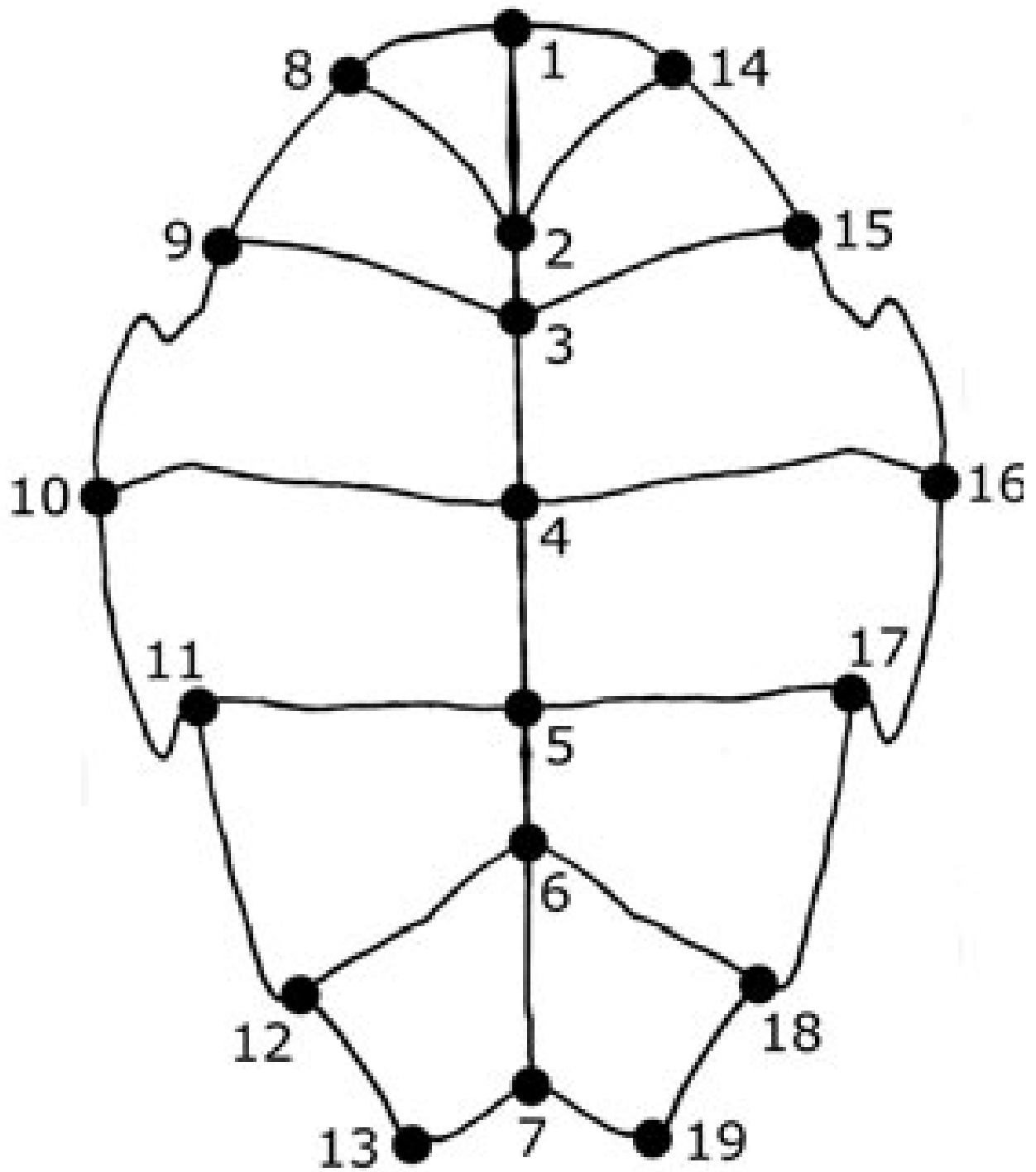


Figure 1: Depiction of general plastral shape of *E. marmorata* and position of the 19 landmark used in this study. Anterior is towards the top of the figure.

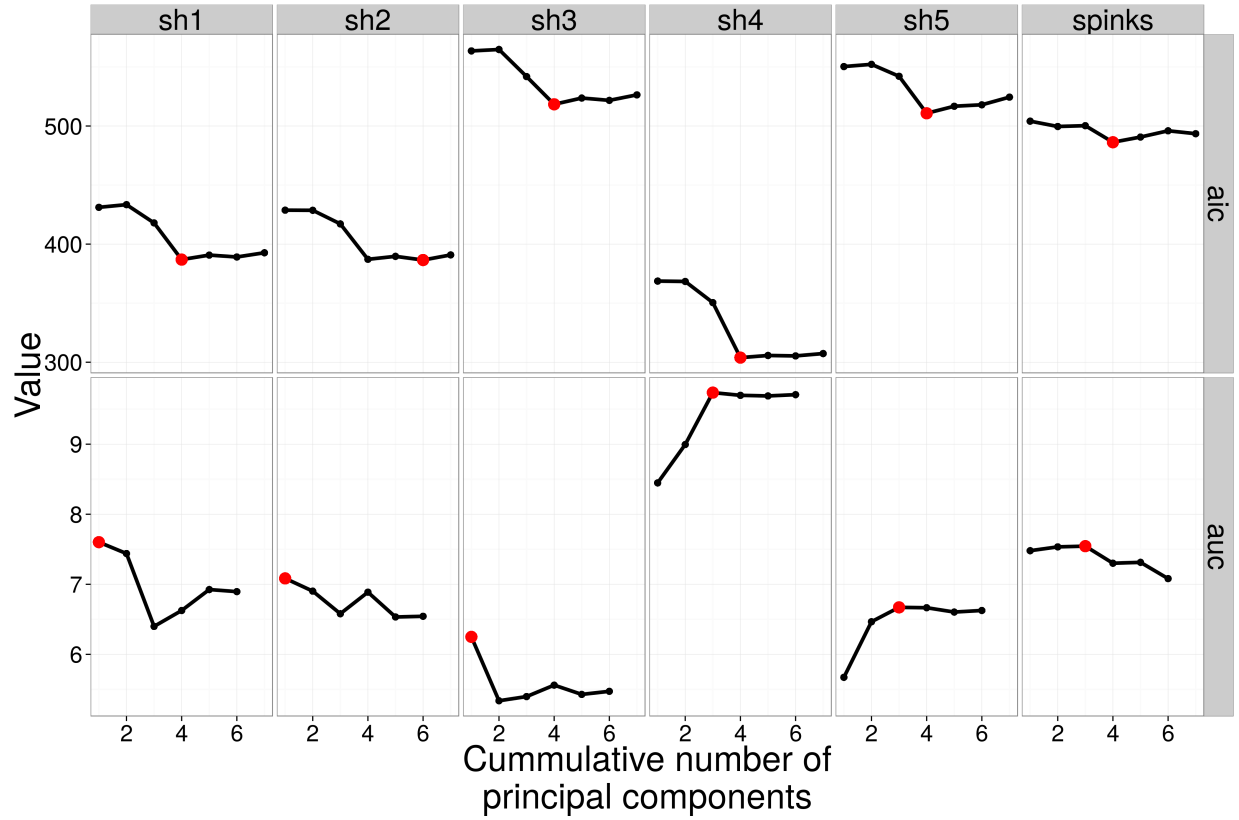


Figure 2: Graphical representation of the AICc and AUC values from model selection for multinomial logistic regression and linear discriminate analysis, respectively. AICc model selection is based on lowest AICc value while AUC model selection is based on greatest AUC value. The horizontal axis corresponds to the cumulative number of axes included in the model of interest. A red dot corresponds to the AICc or AUC best model for that classification scheme.

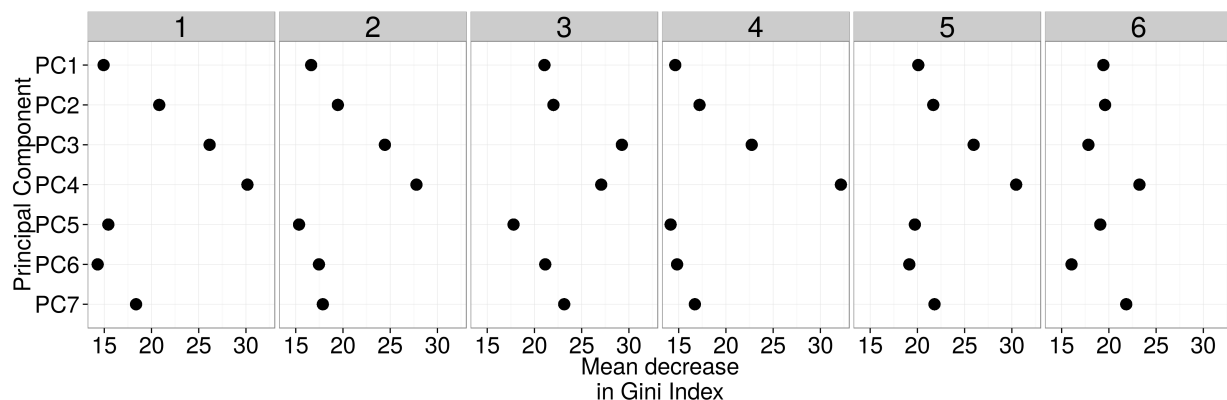


Figure 3: Variable importance from the random forest models for each of the six classification schemes. Importance is measured as the mean decrease in Gini Index, which is a measure of the strength by which that variable determines CART structure. Indices that are farther to the right indicate greater variable importance.

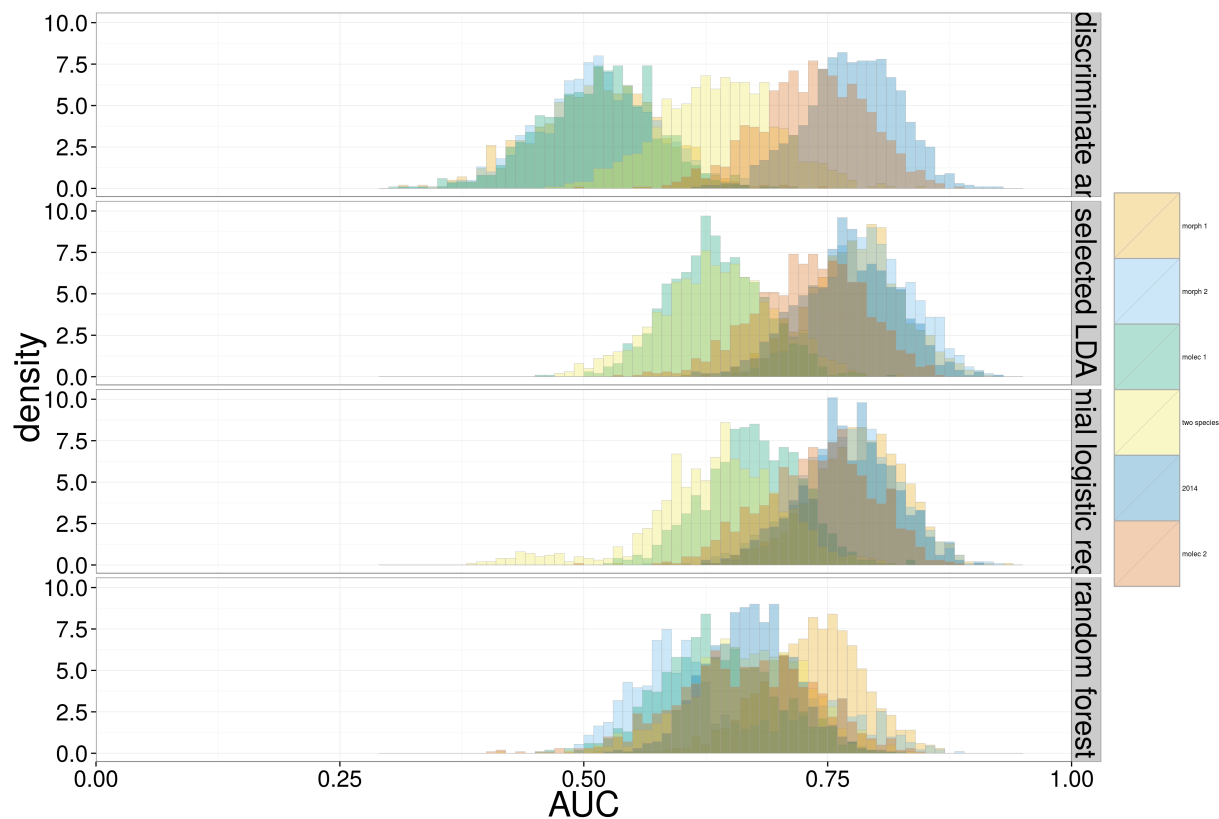


Figure 4: Bootstrap distributions for generalized AUC values for each of the classification schemes. Each row corresponds to a different modeling approach: LDA, LDA using best variables from random forest, multinomial logistic regression, and random forest. Each distribution corresponds to 1000 bootstrap replicates.