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

Peter D Smits[1], Kenneth D Angielczyk[1,2], and James F Parham[3]

[1]Committee on Evolutionary Biology, University of Chicago
[2]Integrative Research Center, Field Museum of Natural History
[3]Department of Geological Sciences, California State University – Fullerton

December 22, 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, sampling, morphometrics

6 We collected landmark-based morphometric data from 354 adult *E. marmorata* museum specimens. These specimens are a subset of those included in Angielczyk and Sheets (2007), 8 Angielczyk et al. (2011), and Angielczyk and Feldman (2013) and represents adult individuals. We chose to focus on adults because significant changes in plastron shape occur over the 10 course of ontogeny in *E. mamorata* and other emydines (Angielczyk and Feldman 2013). We assigned a classification to each specimen for the different binning schemes based on 12 geographic occurrence data recorded in museum collection archives. When precise latitude and longitude information was not available we estimated it from whatever locality information 14 was present. Because the specimens sampled to obtain the genetic data used to define the

subclades were not available for study, all specimen classifications were based solely on the geographic information, not explicit assignment in previous studies. Because the exact barriers between different biogeographic regions are unknown and unclear, we represented each hypothesis with two different schemes; se we compared a total of six different schemes.

Following previous work on plastron variation (Angielczyk and Feldman 2013; Angielczyk et al. 2011; Angielczyk and Sheets 2007), we used TpsDig 2.04 (Rohlf 2005) to digitize 19 landmarks 1). Seventeen of the landmarks are at the endpoints or intersection of the keratinous plastral scutes that cover the platron. Twelve of the landmarks were symmetrical across the axis of symmetry and, in order to prevent degrees of freedom and other concerns (Klingenberg et al. 2002), we reflected these landmarks across the axis of symmetry (i.e. midline) prior to analysis and used the average position of each symmetrical pair. In cases where damage or incompleteness prevented symmetric landmarks from being determined, we used only the single member of the pair. We conducted all subsequent analyses on the resulting "half" plastra. We superimposed the plastral landmark configurations using generalized Procrustes analysis (Dryden and Mardia 1998), after which, we calculated the principal components (PC) of shape 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, we measured the dissimilarity between observations using Kendall's Riemannian shape distance or ρ (Dryden and Mardia 1998; Kendall 1984). We chose this metric 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.

We divisively cluster the ρ dissimilarity matrix using partitioning around mediods 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 mediods 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, we estimated clustering solutions in which the number of clusters varied between one and eight. We compated clustering solutions using the gap statistic, which is a measure of goodness of clustering (Tibshirani et al. 2001).

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

Supervised learning.— We used three different supervised learning, or classification, approaches: linear discriminate analysis, multinomial logistic regression, and random forests. Linear discriminate 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 they reveal very different aspects of the data. In all cases, the optimal number of PCs used as predictors was chosen via maximum within-sample AUC value, explained below.

Linear discriminate analysis (LDA) attempts to find a linear combination of predictors that 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 study, we applied LDA to 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 the LDA, we determined the optimal number of PCs as predictors 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 estimate the class of an observation (specimen). In top-down induction of decision trees for each member of a given set of predictor variables, attribute value tests are used to estimate the differences 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 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 attribute tests made during model fitting.

In a random forest model, many CARTs are built from a random subsample of both the features and the observations (specimens). This process is then repeated many times and the parameters of the final model are chosen as the mode of estimates from the distribution of CARTs (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 is useful for many other studies in which the goal is to describe and model the differences between classes and the relative importance of different predictors.

In this analysis, we used 1000 subtrees to estimate the random forest model parameters. We estimated the best set of predictors necessary for each classification scheme was estimated using a recursive feature selection algorithm, and we chose the optimal number of PCs to include based on the AUC of the model. Following the backwards selection algorithm implemented in `caret` (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 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 to the amount of variance each PC described, these means that the PCs are not included in the order of ascending eigenvalue.

In classification studies, such as this one, a common metric of performance is area under the receiver operating characteristic curve (AUC). AUC is an estimate of the relationship between 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 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 (Hastie et al. 2009).

The standard AUC calculation is defined for binary classifications, however in this application there are multiple categories. The alternative calculation that we used follows an all-against-one strategy where the individual AUC values for each class versus all others are averaged to produce a multiclass AUC (Hand and Till 2001). To estimate confidence intervals on the out-of-sample AUC values, we performed a nonparametric bootstrap in which the true and estimated classifications were resampled with replacement. This was done 1000 times.

The ultimate measure of model fit is accurately predicting the values of unobserved samples (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 where as the later is used for measuring model performance, and this process is called model generalization. In this analysis, we used 75% of samples as the training set while the remaining 25% were used as the testing set.

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, we compared the multivariate centroids of the correct and incorrect groups to what would be expected by random. The group labels were permuted 1000 times and the difference between the new centroids was calculated. The number of permutations less than the empirical difference divided by 1000 gives a p-value for the test. Significant results indicate that correctly and incorrectly classified specimens are systematically different. This was done only for classes where there were 10 or more observations.

Results

Unsupervised learning

Comparison of gap statistic values from PAM clustering show that the best number of clusters is most likely one (Fig. 2). There is some ambiguity in choice because, although it is not statistically different from a solution with only one group, the solution with two groupings does have the greatest mean gap statistic. However, there is no geographical signal in the results of this clustering solution (Fig. 3). Because of this, we assert that this means that there is no means of naturally partitioning plastron shape into distinct subgroups.

Supervised learning

AUC–based model selection revealed some important patterns of variation and congruence between the classification schemes and the actual data. Generally, as many PCs as possible were included as predictors for the best models of each of the classification schemes (Fig. 4). 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. 4). That almost all models were as complex as possible indicates that the differences between the different groups within each classification scheme are very small.

As part of fitting a random forest model, a ranking of variable importance also is determined. Interestingly, the order of variable importance is not the same as the order of decreasing explained variance per principal component (Fig. 5). 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. Another way of phrasing this is that the variance describing the differences between the classes does not align with the major axes of variance (i.e. the PCs). This result would be the case if variation between classes was extremely fine grained and not a part of the principal form or function of the plastron, which makes sense given that the plastron is involved in both protection and hydrodynamics and not mate choice (Germano and Bury 2009; Holland 1992; Lubcke and Wilson 2007; Rivera 2008). Moreover, this result is congruous with the results of the AUC–based model selection for the multinomial logistic regression and LDA models.

Observed AUC values for all of the optimal models are not exceptionally high; values near 0.5 indicate that a model is no better than completely random assignment (Fig. 4). This means that in all but a few cases the different 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.

Mean AUC values for the model generalizations, in most cases, are approximately equal to the observed AUC values from the training data set (Table 1). The cases in which the AUC from the generalizations is less than the observed, indicate poor model fit and a poor classification scheme. AUC values from model generalization, or estimating testing data set

membership, does not indicate a clear "best" classification scheme (Fig. 6). Although the scheme with two species has the greatest AUC point estimate for each modeling approach, this scheme is not significantly greater than any other except in some limited cases (e.g. LDA, Table 2). Differences in mean shape between correctly and incorrectly classified observations from test set frequently were statistically significant, though there are exceptions. Again, this test was to determine if the mean shapes were statistically different or not. The frequency of these results, however, is important because it 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

The results of this study indicate that there is no clear grouping of plastra shapes in *E. marmorata*.

The unsupervised learning results which indicate only a single group of observations being optimal is congruous with the results from the generalizations of the supervised learning models. The classification schemes used in the supervised learning models correspond, loosely, to unsupervised learning solutions with multiple groups. Because unsupervised learning solutions with multiple groups are poor descriptors of the observed variation, it is important to see this reinforced by the supervised learning results.

The results from fitting the various supervised learning models for each of the classification scheme generally shows that no one scheme is "best." 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.

Both the low AUC values (< 0.9) and the significant difference between the correctly and incorrectly classified observations support the conclusion that none of the hypothesized classification schemes are good descriptors of the observed plastral variation.

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 symetric structures: quantifying variation among individuals and asymmetry. Evolution 56:1909–1920.

Kuhn, M. and K. Johnson. 2013. Applied predictive modeling. Springer, New York, NY.

Lubcke, G. M. and D. S. Wilson. 2007. Variation in shell morphology of the Western Pond Turtle (Actinemys marmorata Baird and Giarard) from three aquativ habitats in Northern California. Journal of Herpetology 41:107–114.

Maechler, M., P. Rousseeuw, A. Struyf, M. Hubert, and K. Hornik. 2013. cluster: Cluster Analysis Basics and Extensions. R package version 1.14.4.

Mitteroecker, P. and F. Bookstein. 2011. Linear Discrimination, Ordination, and the Visualization 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 for Statistical Computing Vienna, Austria.

Rivera, G. 2008. Ecomorphological variation in shell shape of the freshwater turtle Pseudemys concinna inhabiting different aquatic flow regimes. Integrative and comparative biology 48:769–87.

Rohlf, F. J. 2005. TpsDig 2.04.

Spinks, P. Q. and H. B. Shaffer. 2005. Range-wide molecular analysis of the western pond turtle (Emys marmorata): cryptic variation, isolation by distance, and their conservation implications. Molecular ecology 14:2047–64.

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 marmorata in California. Molecular ecology 19:542–56.

Tibshirani, R., G. Walther, and T. Hastie. 2001. Estimating the number of clusters in a data set via the gap statistic. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 63:411–423.

Venables, W. and B. D. Ripley. 2002. Modern applied statistics with S. 4th ed. Springer, New York.

Zelditch, M. L., D. L. Swiderski, and H. D. Sheets. 2004. Geometric morphometrics for biologists: a primer. Elsevier Academic Press, Amsterdam.

	random forest		multinomial logistic regression		linear discriminate analysis	
Scheme	Observed	Generalized	Observed	Generalized	Observed	Geneeralized
sh1	0.63	0.73	0.75	0.79	0.75	0.80
sh2	0.61	0.58	0.76	0.77	0.76	0.77
sh3	0.63	0.62	0.75	0.63	0.75	0.63
sh4	0.77	0.67	0.80	0.64	0.80	0.63
sh5	0.56	0.67	0.74	0.62	0.74	0.77
spinks	0.56	0.64	0.71	0.74	0.71	0.73

Table 1: AUC values for the best model of each classification scheme for both the observed (training) data and the generalized (testing) data. Results from all three different supervised learning approaches are shown here. AUC values range between 0.5 and 1.

(a) random forest

Scheme	P(best - other > 0)
sh1	*
sh2	0.79
sh3	0.89
sh4	0.79
sh5	0.82
spinks	0.79

(b) multinomial logistic regression

Scheme	P(best - other > 0)
sh1	*
sh2	0.55
sh3	0.94
sh4	0.96
sh5	0.57
spinks	0.69

(c) linear discriminate analysis

Scheme	P(best - other > 0)
sh1	1
sh2	1
sh3	1
sh4	0.96
sh5	*
spinks	0.73

Table 2: Results of bootstrap comparisons between the scheme with the highest mean AUC value and all other schemes. An asterix indicates the best scheme. This was done for each of the three modeling techniques included in this study. Probabilities are the percent of comparisons that are greater than the observed difference in means.

(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 3: 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.

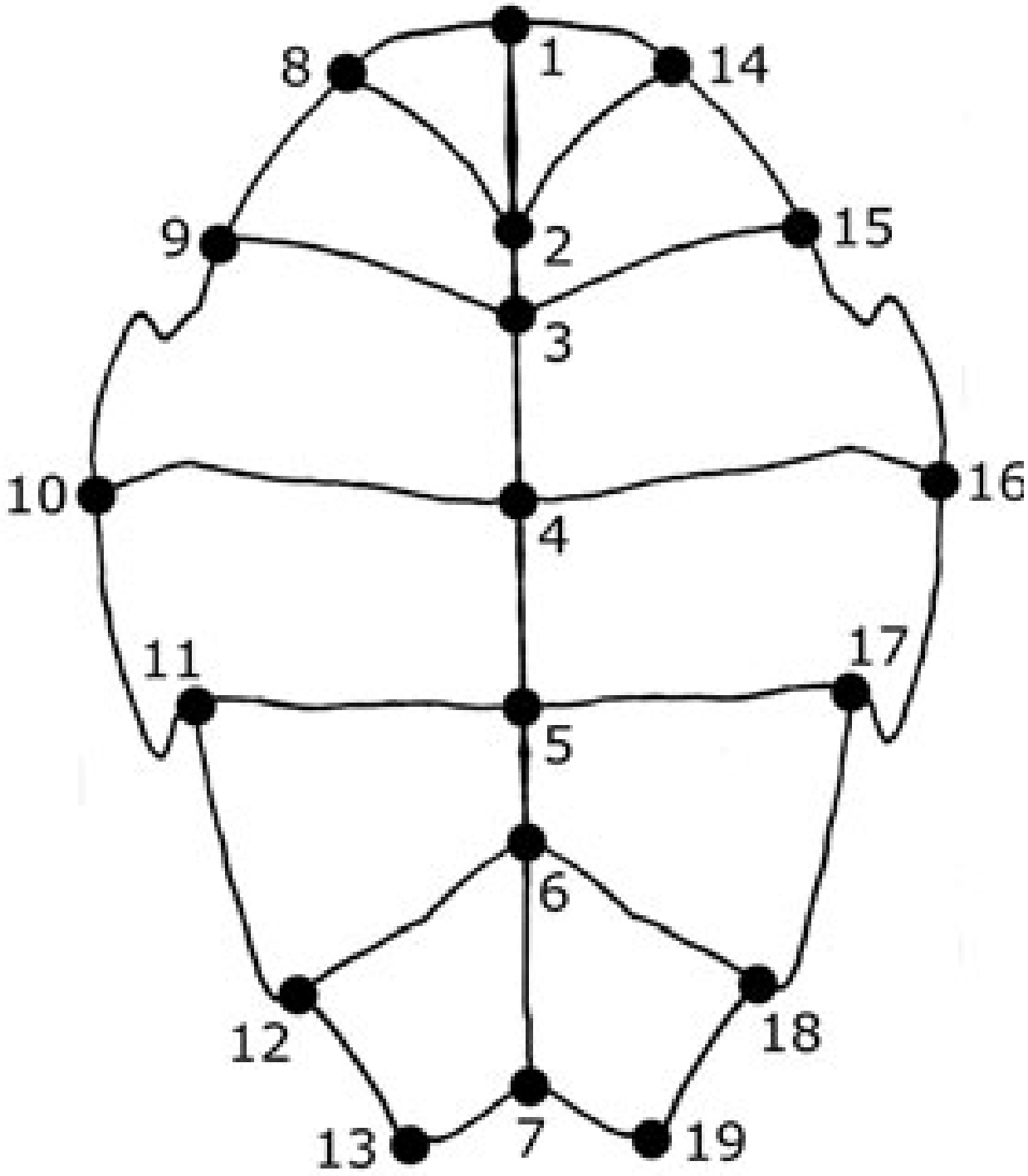

Figure 1: Depiction of general plastral shape of *E. marmorata* and position of the 19 landmark used in this study. Anterior is towards the top of the figure.

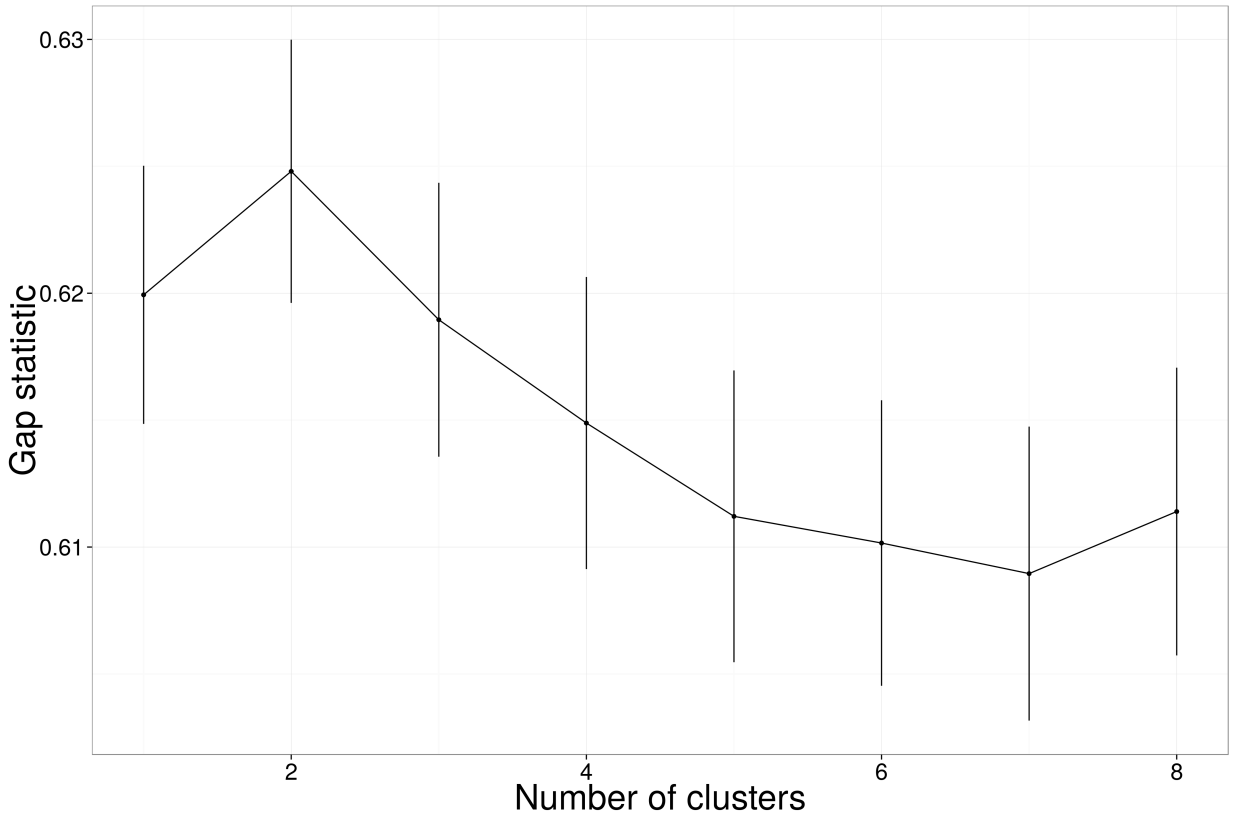

Figure 2: Results from PAM clustering of the Riemannian shape distance for 8 different number of clusters. Vertical lines are 1 standard deviation of the mean determined from 500 resamples.

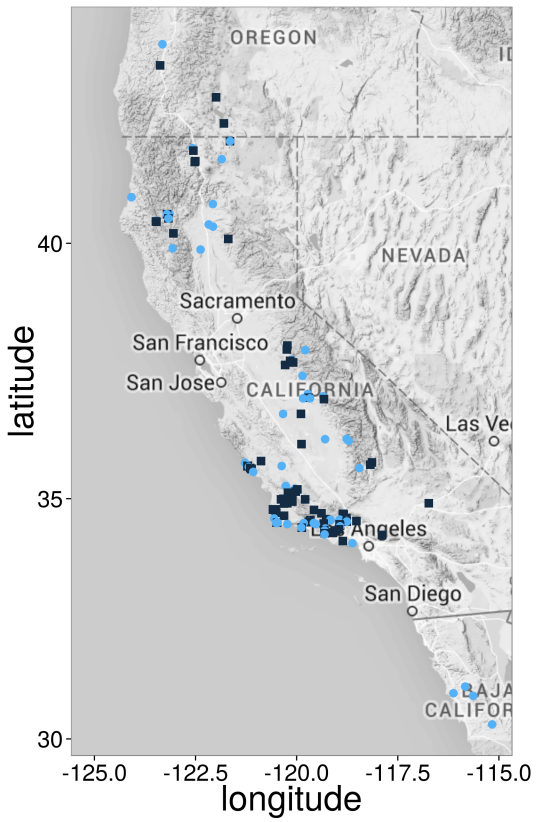

Figure 3: Comparison of geographic distribution of clustered observations from the 2 clustering PAM solution. Colour and shape correspond to each of the groups. There is clearly no geographic signal in the data.

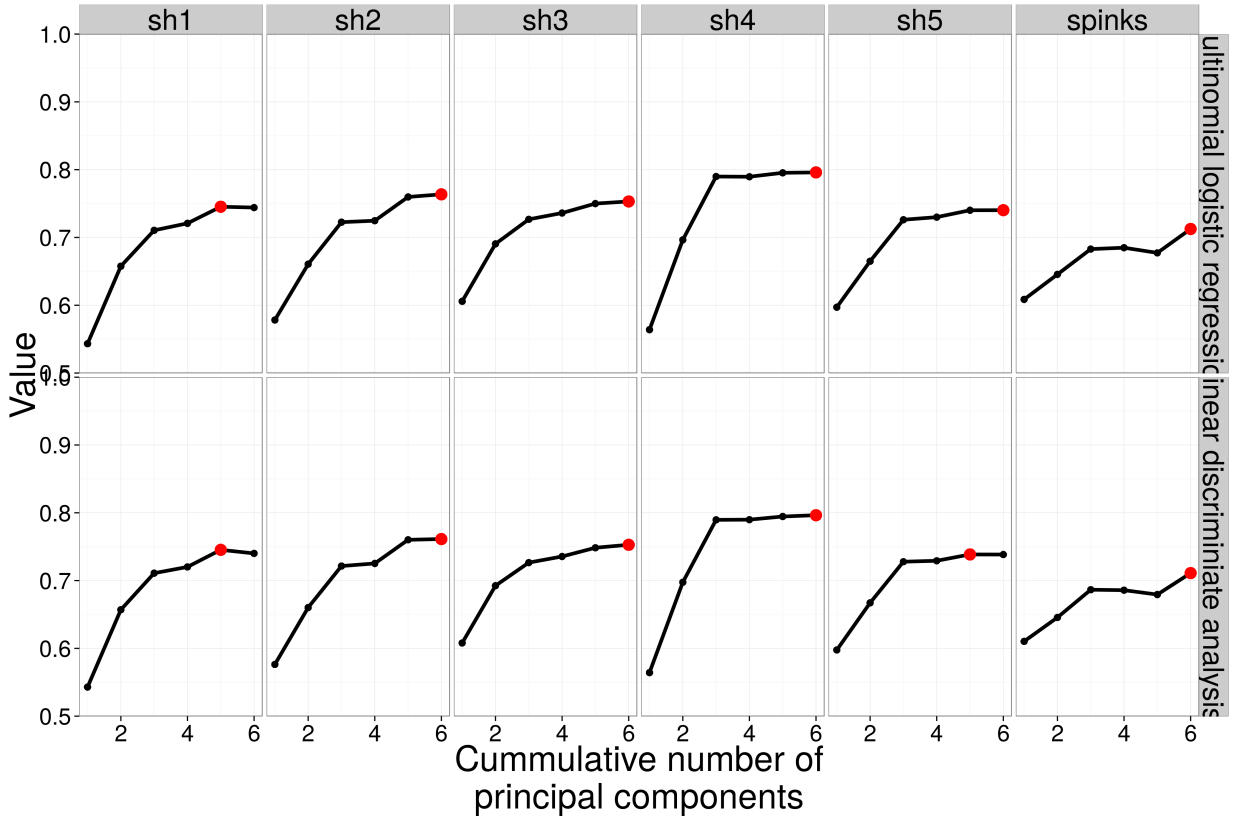

Figure 4: Graphical representation of the AUC values from model selection for multinomial logigistic regression and linear discriminate analysis, respecitively. AUC model selection is based on greatest AUC value. The horizontal axis corresponds to the cummulative number of axes included in the model of interest. A red dot corresponds to the AUC best model for that classification scheme.

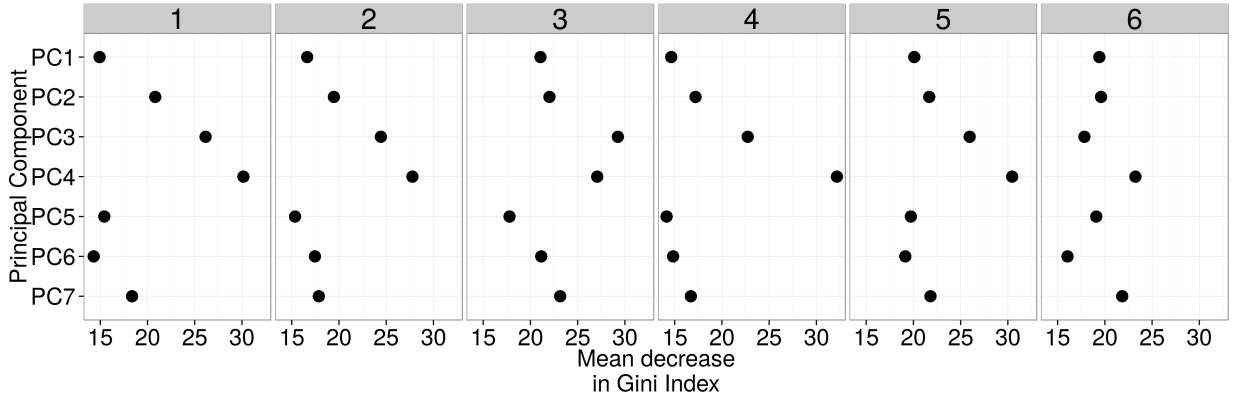

Figure 5: Variable importance from the random forest models for each of the six classification schemes. Importance is measured as the mean decrease in Gini Index, which is a measure of the strength by which that variable determines CART structure. Indices that are farther to the right indicate greater variable importance.

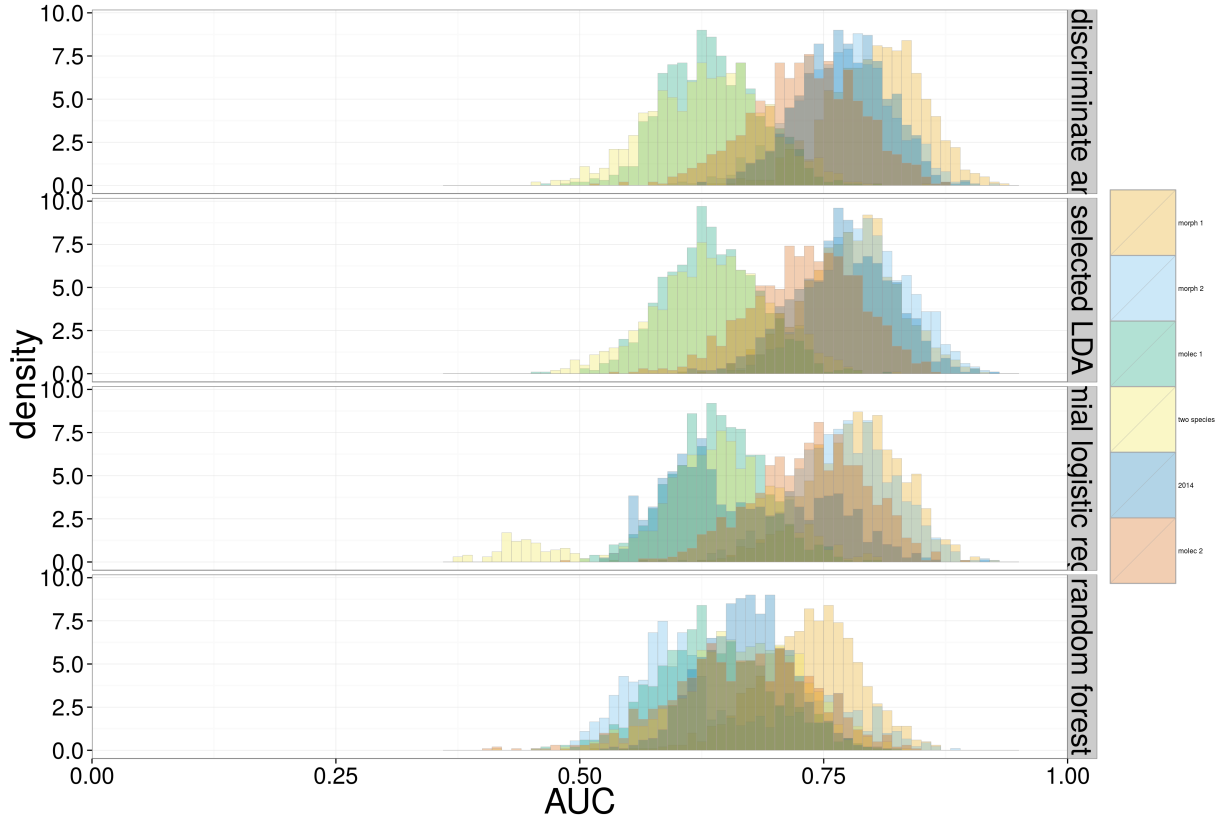

Figure 6: Bootstrap distributions for genearlized AUC values for each of the classification schemes. Each row corresponds to a different modeling approach: LDA, LDA using best variables from random forest, multinomial logistic regression, and random forest. Each distribution corresponds to 1000 bootstrap replicates.