

# How cryptic is cryptic diversity? Machine learning approaches to plastral variation in *Emys marmorata*.

Peter D Smits<sup>1</sup>, Kenneth D Angielczyk<sup>2</sup>, and James F Parham<sup>3</sup>

<sup>1</sup>Committee on Evolution Biology, University of Chicago

<sup>2</sup>Department of Geology, Field Museum of Natural History

<sup>3</sup>Department of Geological Sciences, California State University – Fullerton

July 14, 2013

**Corresponding author:** Peter D Smits, Committee on Evolutionary Biology, University of Chicago, ADDRESS, STATE, ZIP, USA; E-mail: psmits@uchicago.edu

## Abstract

(Keywords: Testudines, morphology, geometric morphometrics)

Cryptic diversity is when taxa were only first delimited via molecular means and were not or cannot be delimited via morphological identification CITATION. The discovery of this previously unknown diversity has expanded diversity amongst certain taxa CITATION. Additionally, a suite of methods have been developed to more accurately delimit taxa based on molecular divergence CITATION.

Concerns about cryptic diversity are enhanced when taxa are only known from morphology, such as the case of extinct organisms CITATION.

Here, we address the question of how much of cryptic diversity may be a product of sample size as well as methodology used for classifying taxa based solely on morphology. Specifically, we ask if fine scale variation in morphology can provide corroboration for subspecific assignment, and if it is possible to determine the best classification hypothesis amongst a few.

In the analysis of class differences and identification in morphometric analysis, classification methods such as discriminant analysis and canonical variates analysis CITATIONS have been used frequently for understanding how variation in specific morphologies best differentiate classes of taxa. Additionally, there has been work on the application of neural network models in classification schemes of observations especially in the context of automated taxon identification CITATIONS MACLEOD'S BOOK. Here, we used multiple types of machine learning methods, both unsupervised and supervised, in order to understand the best classification scheme of the taxa. Each method provides a series of unique advantages for understanding how to classify taxa, which are discussed below.

In this study, we address the subspecific classification scheme of *Emys marmorata*, or western pond turtle. *E. marmorata* has a distribution from northern Washington State, USA to Baja California, Mexico. Traditionally, *E. marmorata* was classified into three subgroups: the northern *E. marmorata marmorata*, the southern *E. marmorata palida*, and a central Californian intergrade zone (Seeliger 1945). *E. marmorata marmorata* is differentiated from *E. marmorata palida* by the presence of a pair of triangular inguinal plates and darker neck markings. It should be noted that the triangular inguinal plates can sometimes be present in *E. marmorata palida* though they are considerably smaller. More recently, *E. marmorata* was divided into four clades based on mitochondrial DNA: a northern clade, a southern clade, and two central Californian clades (Spinks and Shaffer 2005; Spinks et al. 2010). Nuclear DNA supports two major clades, one northern and one southern, however Spinks et al. (2010) argue that the four clade classification is of greater conservation utility to use the mitochondrial

classification scheme. There is now known morphological differentiation between these clades.

In this study, we apply multiple machine learning approaches to estimate the best classification scheme of *E. marmorata* subspecies based on morphological variation in plastral shape. Because of unclear geographic boundaries between subgroups of *E. marmorata*, we compare two hypotheses of morphology-based classification and two hypotheses of molecular-based classification.

## MATERIALS AND METHODS

### *Specimens*

We collected morphometric data from 524 specimens. Geographic information was recorded from museum collection information. When precise latitude and longitude information was not known for a specimen, it was inferred from whatever locality information was presented.

Specimens were given a class assignment was based on geographic information. Because the exact geographic barriers between different class is unknown and fuzzy, two assignments for both morphological and molecular hypotheses of class were used.

### *Geometric morphometrics*

Following Angielczyk et al. (2011), 19 landmarks were digitized using TpsDig 2.04 (Rohlf 2005). 17 of these landmarks are at the endpoints or intersection of the keratinous plastral scutes that cover the plastron. These landmarks were chosen to maximize the description of plastral variation. 12 of these landmarks are symmetrical across the axis of symmetry and in order to prevent degrees of freedom and other concerns (Klingenberg et al. 2007), these landmarks were reflected across the axis of symmetry and the average position of each

symmetrical pair was used. In cases where damage or incompleteness prevented symetric landmarks from being determined, only the single member of the pair was used. Analysis was then conducted on the resulting “half” plastra.

“Half” plastra 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.*— Because shape space, or configurations after Procrustes superimposition, is a Riemannian manifold (Dryden and Mardia 1998) the dissimilarity between each landmark configuration was measured as the Riemmanian shape distance or  $\rho$  (Dryden and Mardia 1998; Kendall 1984) which should vary between 0 and  $\pi/2$  assuming no reflection invariance.

The dissimilarity matrix of shape was divisivly clustering using partitioning around medioids (PAM) which is analogous 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 medioids is minimized (Kaufman and Rousseeuw 1990). The optimal number of clusters of shape configurations is unknown being possibly three, four, or some other value. Clustering solutions were estimated for between 1 and 40 clusters. Clustering solutions were compared using the gap statistic, which is a measure of goodness of clustering (Tibshirani et al. 2001). Standard errors of the gap statistic for each clustering solution were estimated from 500 bootstrap samples. PAM clustering and gap statistic calculation was conducted using the **cluster** package for R (Maechler et al. 2013).

*Supervised learning.*— The dataset of 524 plastron landmarks was split into training and

testing datasets. The former was used for model fitting (training) and was 75% of the total dataset, split proportionally per class, while the testing dataset was used to estimate the effectiveness of each classification scheme (i.e. performance in the wild).

Two types of supervised learning, or classification, models were fit to the PCs of plastral shape: multinomial logistic regression and random forest. These model types were chosen because of various properties of these models which allow for useful interpretations about the strength and structure of the classification. Multinomial logistic regression models were fit using the `nnet` package for R (Venables and Ripley 2002) while random forest models were fit using the `randomForest` package for R (Liaw and Wiener 2002).

Multinomial logistic regression is an extension of logistic regression, where instead of a binary response it is possible to have three or more response classes CITATION. Effectively, this type of model can be viewed as multiple, simultaneous logistic regression models for each class and the final classification of the observation being the most probable of all the sub-model classifications. From the final model the relative risk of a given classification, with reference to a given class, can be calculated from the coefficients of the features, or predictors. This is similar to the log-odds calculated from the coefficients of a logistic regression.

Random forest models are an extension of classification and regression trees (CART) CITATION. Basically, CARTs are built for random subsamples of both the features of the proposed model and observations. This process is repeated many times, 1000 times here, and the final model is chosen as the mode of the parameter estimates from the distribution of CARTs CITATION. In addition to fitting a classification model, this procedure allows for the features to be ranked in order of importance, means that the variables most important for determining a given classification scheme can be estimated. In the context of predicting class from geometric morphometric data, this identifies the PCs that describe the variation that best distinguishes the different classes.

For both types of supervised learning models, tuning parameters were estimated

using 10 rounds of 10-fold cross-validation (CV) across a grid search of all tuning parameter combinations. Optimal tuning parameter values were selected based on area under the receiver operating characteristic curve (AUC ROC). Multiclass AUC ROC was estimated using the all-against one strategy derived by Hand and Till (2001) in implemented in PROC PACKAGE.

For the multinomial logistic regression models, PCs were added sequentially in order to increase the overall amount of variation in shape included in each model and the final model was that with the lowest AICc (Burnham and Anderson 2002) AKAIKE AND OTHER CITATION. This procedure was used because the optimal number of PCs to include is unknown, and while including all of the PCs of shape would mean that all of the variability in plastron shape would be used to estimate class, this may cause the model to be over fit and not provide an accurate estimate of unsampled plastral variation. The maximum number of PCs allowed to be used as predictors was 10 because of both the number of parameters estimated per model and the necessary sample size needed to estimate that many parameters accurately.

Because random forest models are not fit using maximum likelihood, a recursive feature selection algorithm was used to choose the optimal number of PCs to include based on the AUC ROC of the model. PCs were sequentially added as features until the AUC ROC of the model did not increase. Random forest model parameters were estimated from 1000 subtrees. After each PC was added, 10-fold CV was used to estimate the optimal values of the tuning parameters as well as quantify the uncertainty of each model. Like the multinomial logistic regression models, 10 was the maximum number of PCs that could have been included in the model. The recursive feature selection algorithm used here is that implemented in the `caret` package for R (Kuhn 2013).

The final selected models were then used to estimate the class assignments of the training dataset. Model performance was measured using AUC ROC. A distribution of AUC ROC values were estimated for each classification scheme using 1000 nonparametric bootstrap

resamples of the training dataset.

## RESULTS

### *Geometric morphometrics*

### *Machine learning analyses*

#### *Unsupervised learning.*—

Comparison of the gap statistic values for the different PAM solutions indicates that the optimal number of clusters is 1 (Fig. 1). The second best clustering solution is two clusters, however there is no geographic structure to this classification scheme SUPPLEMENT?. Our dataset does not include enough or detailed information on the sex of each *E. marmorata* specimen, thus it is not possible to determine if this clustering solution corresponds to sexual dimorphism between the observations. Male Emydine turtles are known to have a plastral concavity. Increasing the number of clusters does appear to improve the gap statistic enough to merit comparison.

#### *Supervised learning.*—

The optimal multinomial logistic regression models for all four classification schemes, based on AICc, have all 10 possible PCs as features (Table SUPPLEMENT). The second best models for all classification schemes have 9 PCs as features. The  $\delta$ AICc values between the optimal and second best model range from 2.0639 for the first morphological based classification hypothesis to 19.8349 for the second molecular based classification hypothesis.

While the  $\delta$ AICc value between the optimal and second best model for the first morphological based classification hypothesis is within the range to be considered sufficient

and equally optimal (Burnham and Anderson 2002), for this analysis we chose to use only the absolute AICc best model. While AICc values can not be compared between models with different responses (Burnham and Anderson 2002), we interpret the fact that the optimal model for all classification schemes is the global model as a reason to just use this model structure for all cases. Additionally, by maintaining a single optimal model for all classification schemes, this limits the total number of comparisons between the bootstrap resampled distributions of the AUC ROC values for the testing data set (see below).

The selected number of features in the optimal random forest models for each classification scheme vary much more than in the case of the multinomial logistic regression models (Fig. 2), ranging from 6 for the first morphological based classification hypothesis to 10 for the second morphological based classification hypothesis.

In the case of all models, there is a substantial increase in model performance as measured by AICc for the multinomial logistic models (Tables SUPPLEMENT) or in AUC ROC for the random forest models and illustrated for the multinomial logistic regression models as the number of features increases (Fig. 2).

Results of the bootstrap resamples of the AUC ROC of the generalization of the selected multinomial logistic regression and random forest models demonstrates that one of the molecular classification hypotheses based on Spinks and Shaffer (2005) and Spinks et al. (2010) appears to be the best classification scheme (Fig. 3). The distribution of bootstrapped AUC ROC for the molecular hypothesis is significantly different MANN-WHITNEY U TEST and greater than all of the other classification scheme. What is remarkable is that the best classification hypothesis is identical based on both the multinomial logistic regression and random forest models.

When the classification results of the training set for the optimal classification scheme are compared with the references classes, the higher AUC ROC value of the best multinomial logistic regression model becomes apparent as the classifications are in general much more



similar to the reference (Fig. 4). The best random forest model misclassified many of the observations as the northern clade. This trend is observable but not as exaggerated in the results of the classifications of the multinomial logistic regression model.

This pattern of misclassification may be caused by the differences in mean shape between each of the different classes (Fig. 5). The mean shape of the northern clade is the most similar to the mean shape of the entire dataset (Fig. 57), which may mean that specimens that are closer to the mean shape will be systematically misclassified as the northern clade.

The results of training the random forest model also include the variable importance for best separating the different classes. Recursive feature selection of the best random forest model of the chosen classification scheme indicated that after 7 PCs were included as features, AUC ROC would not increase. Of these 7 features, three are illustrated here (Fig. 6) the first two of which are most important SUPPLEMENT WITH VARIABLE IMPORTANCE INFORMATION?.

The first two most important features, according to the random forest model, describe different aspects of variation (Fig. 7). The third PC, or first most important PC, describes variation in the relative position of landmarks on anterior and posterior portions of the plastron. The eighth PC, or second most important PC, mostly describes variation in landmarks along the midline of the plastron. The major variation along these axes correspond well to the differences between the class means and the mean plastron shape (Fig. 5) where major class differences seems restricted to the relative ballooning or shrinking of the anterior and posterior portions of the plastron together.

## DISCUSSION

The results of this study provide support for the mitochondrial based hypothesis of classification of *E. marmorata* (Spinks and Shaffer 2005; Spinks et al. 2010). This is contrary to the

original classification of *E. marmorata* (Seeliger 1945) and lends credence to the idea that at least some aspect of cryptic diversity is a product either sample size, methodology, or both.

The lack of coherent geographical subclass assignment from PAM based clustering (Fig. 1) as well as the large number of features necessary before plateau in AUC ROC for all models (Fig. 2) can be taken as indicators that the variation between subclasses is extremely fine grained. This is also exemplified by the differences in mean class shape for the final chosen classification scheme (Fig. 7).

Ultimately, it would be optimal to not require such explicit classification hypotheses, especially when concerned about possible cryptic variation in extinct taxa. The method employed in this study, PAM, is rather simple and not model based. Comparison is facilitated by comparison of a summary statistic and bootstrap confidence intervals. A more useful and future avenue would be employing various model based clustering approaches CITATIONS. In this manner, a series of candidate models can be compared via model comparison methods, such as AIC or Bayes factors CITATION, in order to assess the best clustering solution. Of particular note are nonparametric Bayesian approaches to model based clustering CITATIONS. This approach uses a class of flexible priors to allow for the most optimal clustering solution to be decided from the data. Currently, there exists a nonparametric Bayesian clustering method and further development of this approach may prove extremely fruitful for better delimiting taxa solely from morphology.

In this study we have demonstrated that given a large sample size and appropriate methodology, it is possible to determine which of multiple hypothesized classification schemes best explain variation in a taxon. The plastral variation of *E. marmorata* is most consistent with the mitochondrial based hypothesis of Spinks and Shaffer (2005) and Spinks et al. (2010) and not with the original morphology based hypothesis of Seeliger (1945). We have also demonstrated the utility of various machine learning approaches to understanding variation in morphometric data. Specifically, better understanding odds misclassification and identifying

which is the most important for delimiting different classes.

## ACKNOWLEDGEMENTS

PDS would like to thank David Bapst, Michael Foote, Benjamin Frable, and Dallas Krentzel for useful discussion which enhanced the quality of this study.

\*

## References

- Angielczyk, K. D., C. R. Feldman, and G. R. Miller. 2011. Adaptive evolution of plastron shape in emydine turtles. *Evolution* 65:377–394.
- 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.
- 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.
- 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. 2007. Shape analysis of symmetric structures: quantifying variation among individuals and asymmetry. *Evolution* 56:1909–1920.

- Kuhn, M. 2013. caret: Classification and Regression Training. R package version 5.15-61.
- Liaw, A., and M. Wiener. 2002. Classification and regression by randomforest. R News 2:18–22.
- Maechler, M., P. Rousseeuw, A. Struyf, M. Hubert, and K. Hornik. 2013. cluster: Cluster Analysis Basics and Extensions. R package version 1.14.4.
- R Core Team. 2013. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
- Rohlf, F. J. 2005. TpsDig 2.04.
- Seeliger, L. M. 1945. Variation in the Pacific Mud Turtle. Copeia 1945:150–159.
- 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. N., and B. D. Ripley. 2002. Modern Applied Statistics with S. 4th ed. Springer, New York.

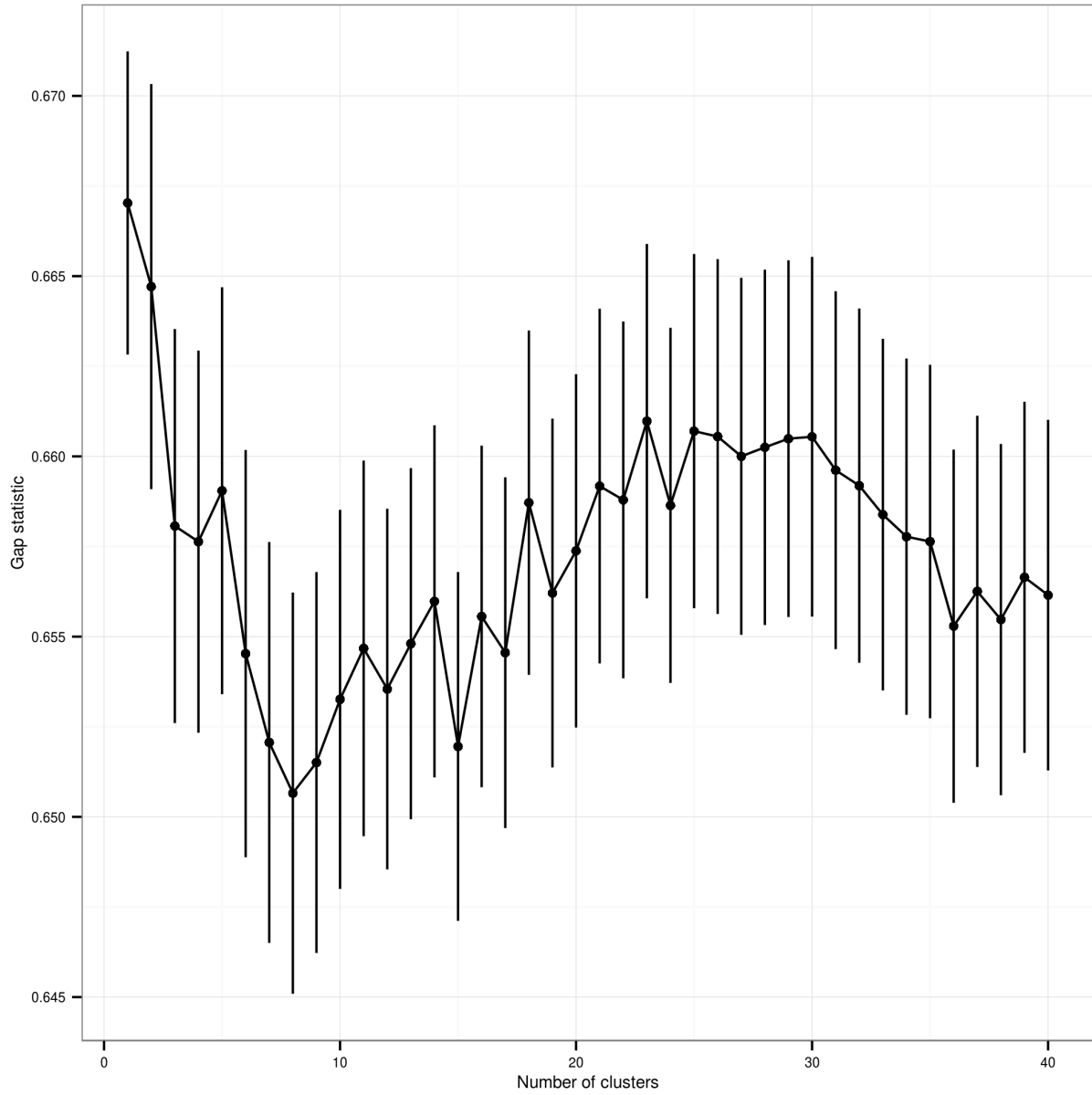


Figure 1: Gap statistic values for PAM clustering results for the  $\rho$  dissimilarity matrix of plastron shape. Error bars are standard errors estimated via 500 bootstrap samples.

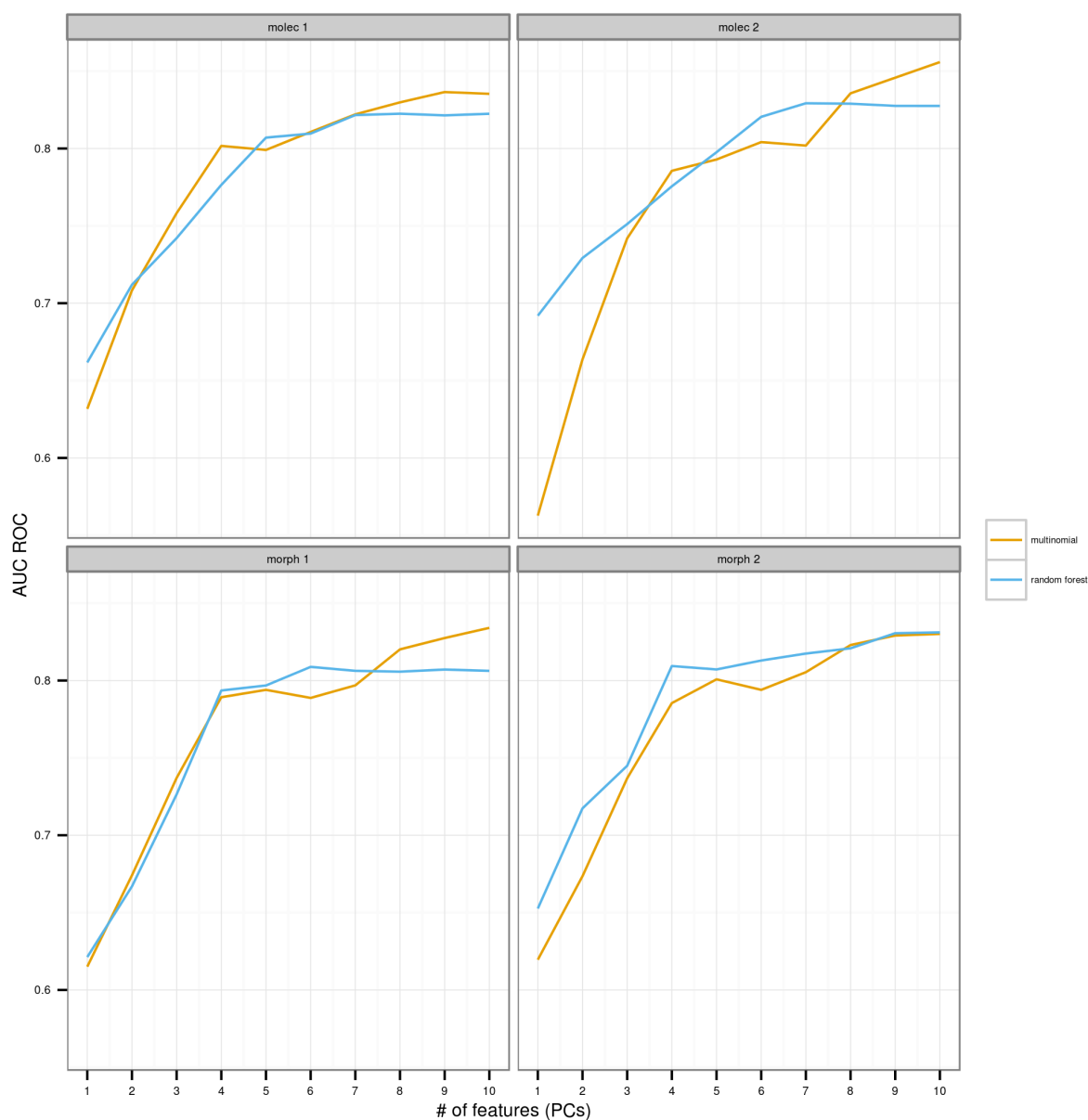


Figure 2: Effect of increasing the number of PCs as features, or predictors, of classification of plastra for all four classification schemes. As the number of PCs increase, AUC ROC increases until eventually leveling off. Both multinomial logistic regression and random forest models are illustrated here, though AUC ROC based model selection was only performed for random forest models.

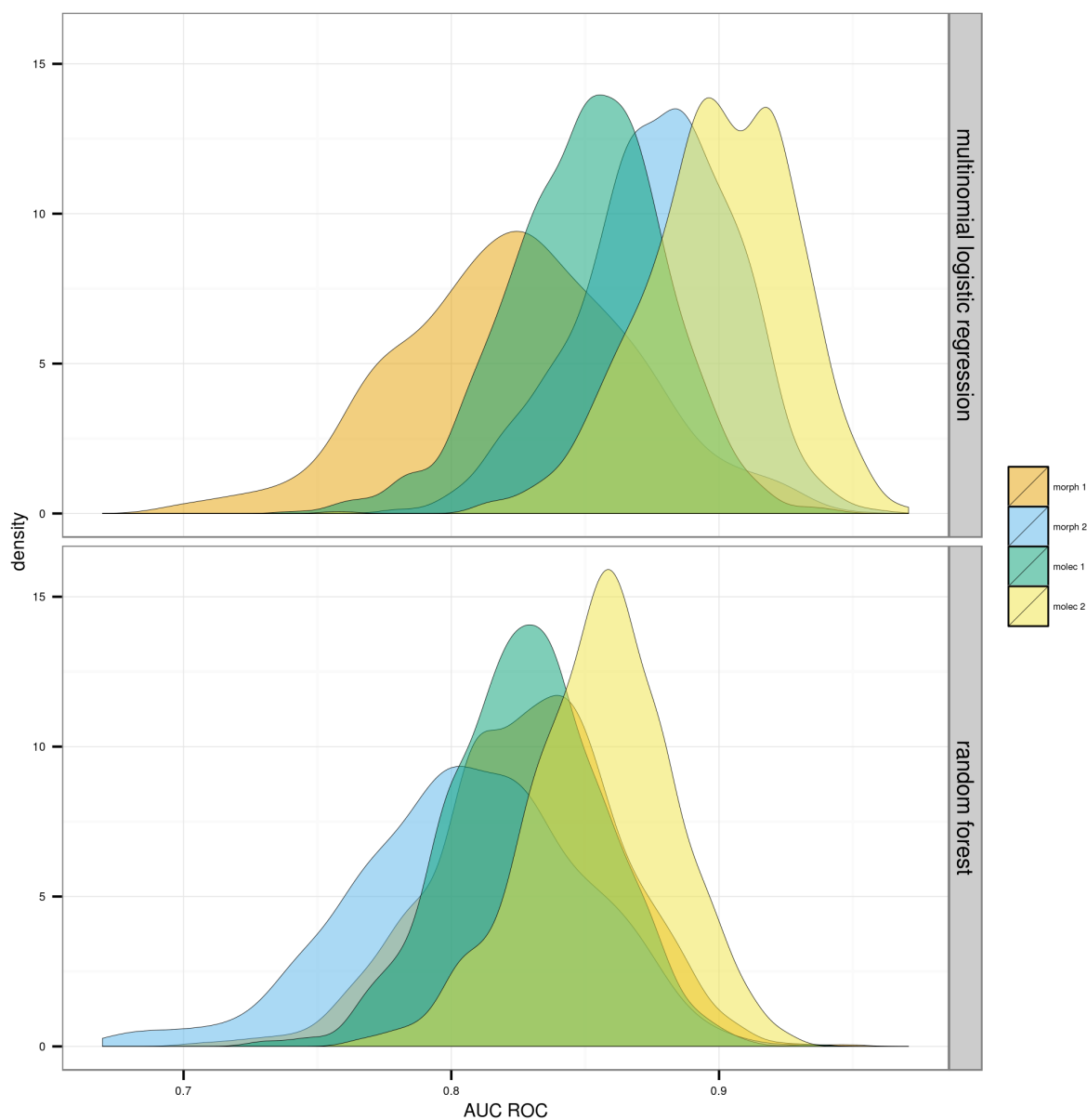


Figure 3: Density estimates of AUC ROC values of predictions of the testing dataset of plastra from 1000 bootstrap resamples. The top facet corresponds to values using the optimal multinomial logistic regression model, as chosen by minimum AICc value. The bottom facet corresponds to the values using the optimal random forest model, as chosen by maximum AUC ROC value.

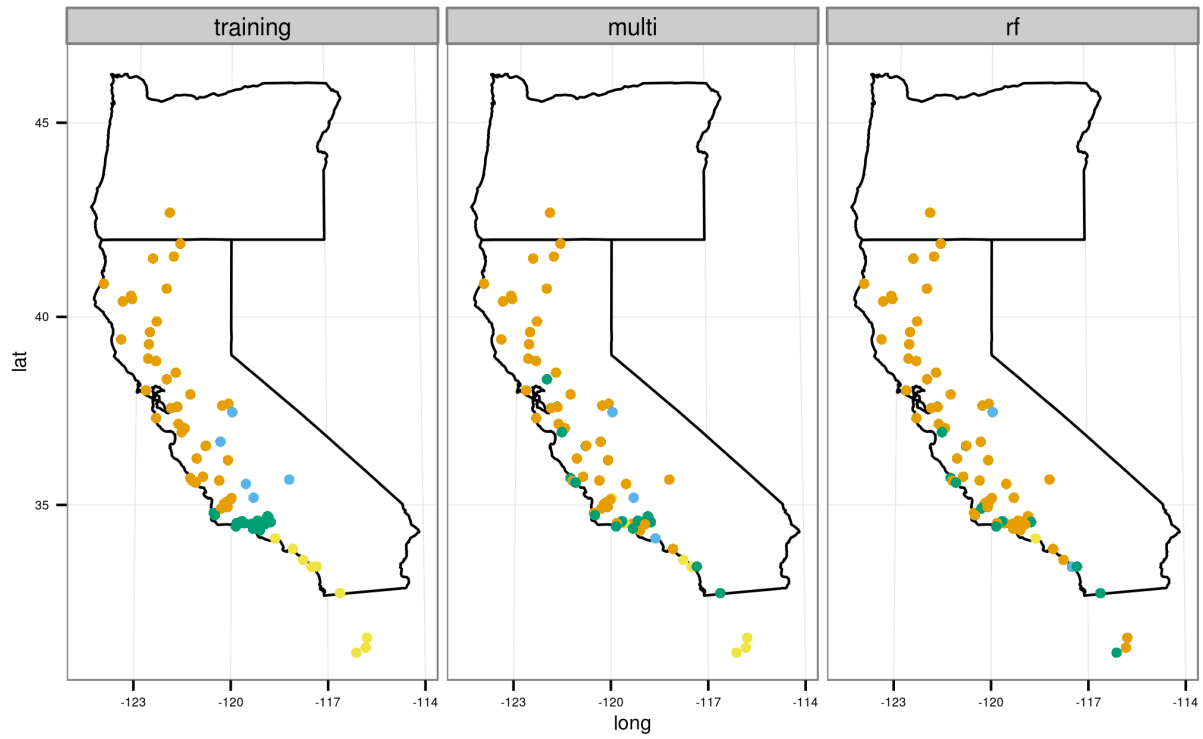


Figure 4: Comparison between reference classification of testing data set and the estimated classifications based on the selected multinomial logistic regression and random forest models, from left to right respectively. Classification corresponds to the four classes as suggested by the hypothesis of Spinks and Shaffer (2005) and Spinks et al. (2010).



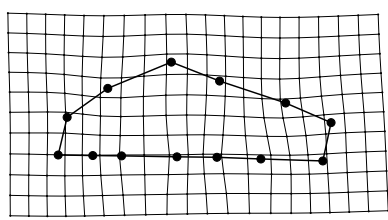
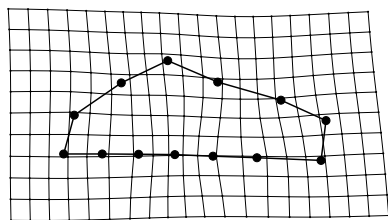
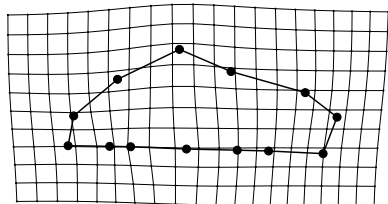
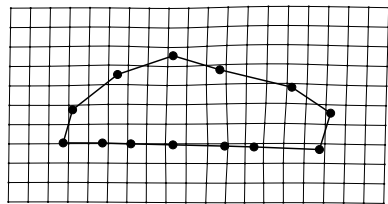


Figure 5:  $j$ +caption text+ $i$

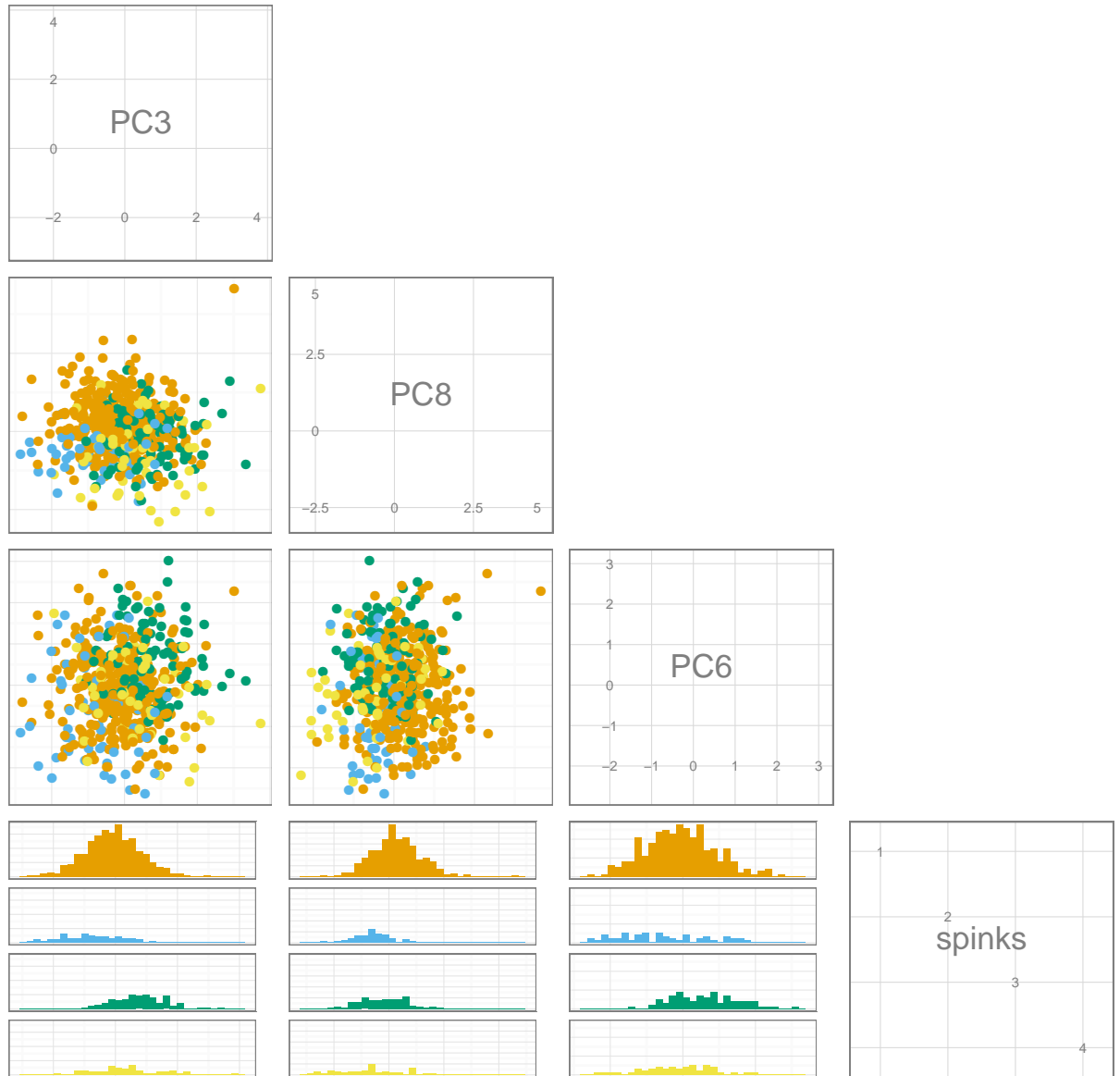


Figure 6: Pairs plot of the first three most important variables of the optimal random forest model of turtle plastral shape. The variables descend in importance from the upper left to the lower right. The observations are colored as in Figures 3 and 4. The bottom row are histograms of classification occurrences along the PCs.

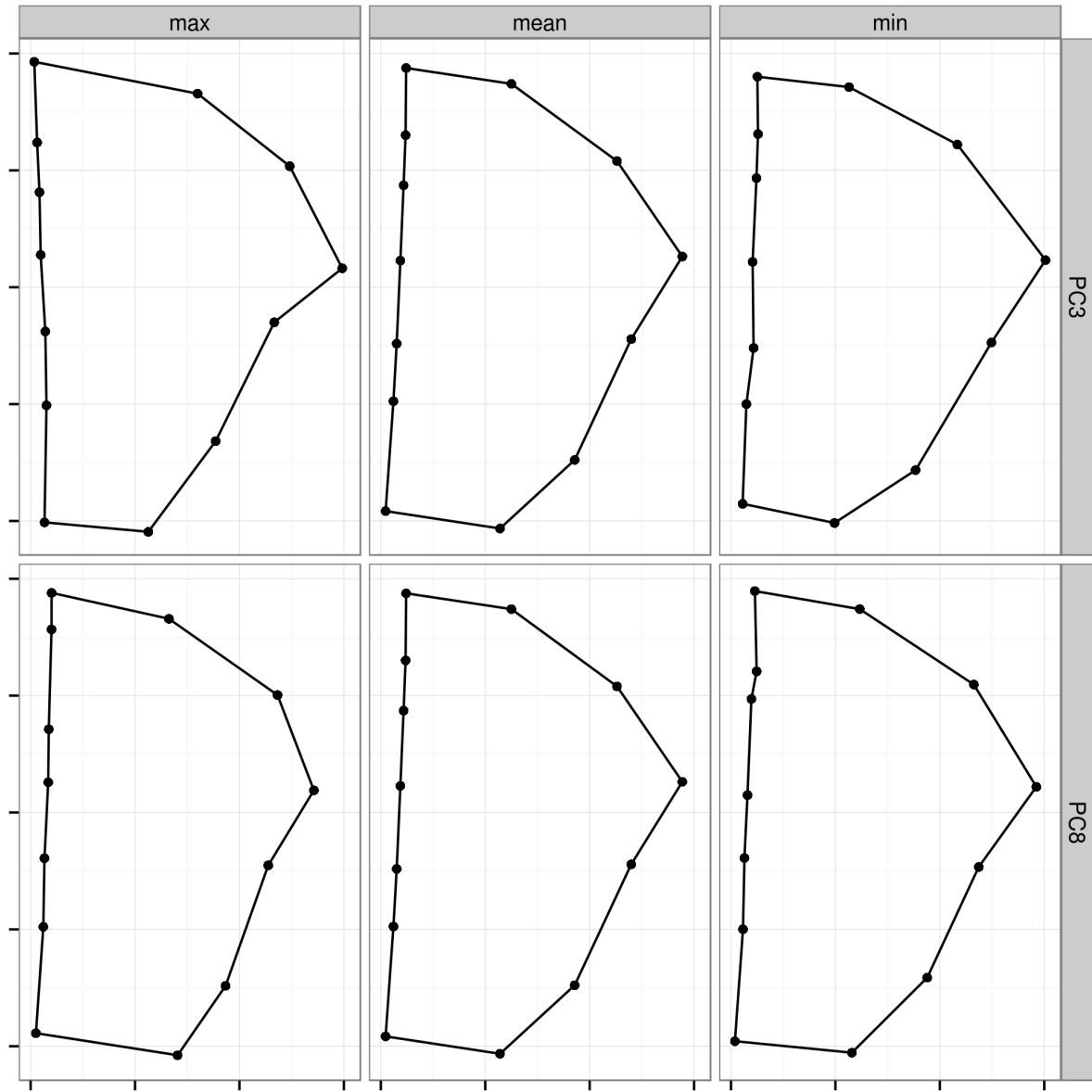


Figure 7: Landmark variation along the two most important features (PCs) based on the final random forest model. The first row corresponds to the third PC and the second corresponds to eighth PC. Landmark configurations are minimum observed on that PC, mean shape, and maximum observed on that PC.