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

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

[1]Committee on Evolutionary Biology, University of Chicago

[2]Department of Geology, Field Museum of Natural History

[3]Department of Geological Sciences, California State University – Fullerton

July 22, 2013

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

Write me please. (Keywords: Testudines, morphology, geometric morphometrics)

Cryptic diversity is the phenomenon that not all taxa can be recognized based on morphology and can only be delimited using molecular information. The most extant taxa, and nearly all extinct taxa, are delimited based solely via morphological analysis. Concerns about how pervasive this phenomenon is and much diversity is actually cryptic have

Much work has been devoted to species delimitation via sequence difference CITATIONS while comparatively little has been devoted for case of morphological data. The

majority of this effort has focused on automated taxon identification CITATIONS or identifying morphological differences amongst already identified taxa CITATIONS.

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

Here, we address two questions about cryptic diversity: how much of cryptic diversity is a product of sample size, and how can methodology impact classification based solely on morphology. Additionally, we ask if fine scale variation in morphology can provide corroboration for various classification hypotheses, and if it is possible to determine which is the best classification hypothesis.

Differences in morphological variation between different classes has previously been analyzed using methods like linear discriminate analysis and canonical variates analysis CITATION ZELDICH. Neural network models have also been introduced and applied in the context of automated taxon identification along with more general applications MACLEOD BOOK. Here, we used multiple alternative machine learning methods, both unsupervised and supervised, in order to compare different classification hypotheses. These different methods provide different and unique advantages for understanding how to classify taxa, with what accuracy, and what these classifications are based on. Additionally, we investigate variation in continuous traits, and do not search for discrete differences between each class, instead focusing on suites of traits together.

The two major cases of machine learning methods, unsupervised and supervised, are essentially extensions of known statistical methods. Unsupervised learning methods are analogous to clustering and density estimation methods, while supervised learning methods are analogous classification and regression models. In both cases, these methods may not be fit via maximum likelihood but may be supplemented by randomization algorithms and the maximization or minimization of summary statistics in order to best estimate a general model for all data, but sampled and unsampled. The expansion of methods used to understand

morphological differentiation

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) UNPUBLISHED MASTERS THESIS. *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.

Previous work on morphological differentiation between subspecies of *E. marmorata* focused on just the known subspecies of Seeliger (1945)

More recently, *E. marmorata* was divided into four clades based on mitochrondial 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 no known morphological differentiation between these clades.

In this study, we attempt 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. We hypotheize that if morphological variation corresponds to subspecific assignment, then it should be possible to differentiate the best classification hypothesis of *E. marmorata* from amongst multiple candidate hypotheses. However, if morphological variation variation does not correspond to any classification hypothesis, then supervised learning model generalization performance will

be extremely poor and reflect how variation does not follow along with classification.

Materials and Methods

Specimens

We collected landmark-based morphometric data from 524 adult *E. marmorata* museum 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 present.

Specimens were classified based on museum record geographic information. The specimens used to define the subclades in Spinks and Shaffer (2005) and Spinks et al. (2010) were not available for study, all classifications were based solely on the geographic information for each specimen and not from explicit assignment in previous studies. Because the exact geographic barriers between different classes are unknown and unclear, two assignments for both the morphological and molecular hypotheses were used. Each morphology-based hypothesis had three classes, while each molecular-based had four classes. In total, each specimen was given four different classifications.

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 platron. These landmarks were chosen to maximize the description of general plastral variation. 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. 2007) 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 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.— In order to preserve the relationship in shape space between all landmark configurations, the dissimilarity between observations was measured using the Riemanninan 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.

The ρ dissimilarity matrix was divisively clustering using partitioning around mediods (PAM), a method which is 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, clustering solutions were estimated with the number of mediods varying between one and 40. 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

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

the dispersion of the clustering solution or the sum of the pairwise dissimilarities ($d_{ii'}$) between observations in each cluster and there respective mediods (C) for all clusters r. E_n^* is the expectation of dispersion from a sample n of a reference distribution. In this case, the reference distribution was estimated from a 500 resamples of the dataset taking into account the original structure of the data. This analysis was conducted using the `cluster` package for R (Maechler et al. 2013) using all 524 plastral landmark configurations.

Supervised learning.— The total dataset of 524 observations was split into training and testing datasets. The training dataset represented 75% of the total dataset, split proportionally per class, and was used for model fitting. The testing dataset represented the remaining 25% of the total dataset and was used to estimate the effectiveness of each classification hypothesis and generalizability of the supervised learning models (i.e. performance in the wild).

Two different supervised learning methods and model types were used to model the relationship between plastral shape and class: multinomial logistic regression and random forest. These models were chosen because of various properties of these models which allow for useful interpretations about the quality and structure of the classification.

Multinomial logistic regression is an extension of logistic regression, where instead of a binary response there are 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 constituent model classifications. From the final model, the relative risk of a possible classification, with

reference to a baseline class, can be calculated from the coefficients of the model. These are similar to the odds ratios calculated from the coefficients of a logistic regression. Multinomial logistic regression models were fit using the `nnet` package for R (Venables and Ripley 2002)

Random forest models are an extension of classification and regression trees (CART) CITATION. In a random forest model, many CARTs are built from a random subsample of both the features and the observations. This process was repeated 1000 times and the parameters of the final model was chosen as the mode of 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. In the context of this study, this means that the PCs most important for describing the difference between classes can be estimated, and thus illustrate the most important variation amongst classes as opposed to just the greatest amount of variation in the entire dataset. This is a generally important property that should be useful for other morphometric studies which want to describe and model the shape differences between different classes and the relative importance of various aspects of variation. Random forest models were fit using the `randomForest` package for R (Liaw and Wiener 2002).

Supervised learning models have tuning parameters which are estimated to increase the genearlizability of the model and prevent them from being overfit. For both types of supervised learning methods, 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 (ROC). The area under the multiclass ROC curves was estimated using the all-against one strategy derived by Hand and Till (2001).

ROC is a type of confusion matrix statistic that is a descriptor the relationship between the false positive rate or 1 - specificity of a classification model and the true positive rate or sensitivity of a classification model CITATION. The area under the ROC curve (AUC) is a

summary statistic of the quality of the classification and varies between 0.5 and 1, with 0.5 indicating a model that classifies no better than random and a value of 1 indicating perfect classification CITATION. AUC can be used as a model selection criterion for classification models and is especially useful in cases where some if not all of the models in question were not fit via maximum likelihood where a criterion such as AICc (see below) or similar can be used CITATION.

For the multinomial logistic regression models, 10 different models were fit with each having sequentially more PCs as predictors in order to have models representing different levels of overall amount of shape variation and estimate how much variation was necessary and sufficient to best estimate class. 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. The final model was that with the lowest AICc (Burnham and Anderson 2002) AKAIKE AND OTHER CITATION. Similar to AUC discussed above, AICc is a model selection criterion where instead of maximizing the ratio of true positive rate to false positive rate the model with lowest AICc has the fairest variance–bias tradeoff (Burnham and Anderson 2002). Model selection was performed in this manner because the optimal number of PCs to use as predictors was not know *a priori*, and while including all of the PCs of shape would mean that all shape variability would be used to estimate class, this may cause the model to be overfit and not provide an accurate estimate of unsampled plastral variation.

Random forest models are not fit using maximum likelihood so AICc based model selection was not possible. Instead, a recursive feature selection algorithm was used to choose the optimal number of PCs to include based on the AUC of the model. PCs were sequentially added as features until the AUC of the model did not increase. Like the multinomial logistic regression models, 10 was the maximum number of PCs that could have been included in the model. 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. Random forest model parameters were estimated from 1000 subtrees. The recursive feature selection algorithm used in this study was 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 generality for both methods for all four classification schemes was measured using the AUC of the assignments. A distribution of AUC values was estimated for each classification scheme via 1000 nonparametric bootstrap resamples of the training dataset.

Results

Geometric morphometrics

The results of the PCA of the total dataset of *E. marmorata* pastral landmarks configurations demonstrates no clear or obvious groupings (Fig. 1). The first three PCs, which represent 55.0075790871272% of the total variation, are a cloud of points. Additionally, individual landmark variation is mostly circular around each landmark with some more elliptical variation observed along some midline landmarks and the most lateral landmark. However, it is important to note that Procustes based superimposition attempts to evenly distribute landmark variance ZELDICH and this observation should be considered cursory and not rigorous.

Machine learning analyses

Unsupervised learning.—

Comparison of the gap statistic values for the range of PAM solutions indicates that the optimal number of clusters is 1 (Fig. 2). The next best clustering solution is two clusters, however there is no geographic structure to this classification scheme, with members of these clusters being seemingly randomly distributed geographically 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 Emyidine turtles are known to have a plastral concavity which may influence landmark position. However, the plastral concavity of *E. mamorata* males is considered less prnounced than in other Emyidine turtles. While we cannot completely rule out sexual dimorphism as the root cause of this observation, we are less concerned with possible effects of sexual dimorphism for the later supervised learning methods. Increasing the number of clusters does appear to improve the gap statistic enough to merit comparison.

Supervised learning.—

The AICc best multinomial logistic regression models, for all four classification schemes, each have the first 10 possible PCs as features (Table SUPPLEMENT). The second best models for all classification schemes had the first 9 PCs as features. The ΔAICc 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 (Table IN SUPPLEMENT?). The first 10 PCs describe 88.6043874205075% of total variation in plastral shape.

While the ΔAICc value between the optimal and second best model for the first morphological based classification hypothesis was within the range to be considered sufficient and equally optimal (Burnham and Anderson 2002), for this analysis we chose to use only the 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 use the AICc best model for all cases. Additionally, by using a single model for all classification schemes, this limits the number of comparisons between the bootstrap resampled distributions of the AUC values for the testing data set (see below).

The selected number of features in the final random forest model for each classification scheme varied much more than in the case of the multinomial logistic regression models (Fig. 3), 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 for the random forest models and illustrated for the multinomial logistic regression models as the number of features increases (Fig. 3).

Results from the generalization of the selected supervised learning models, measured by the distributions of the bootstrapped AUC values of the testing dataset demonstrates that one of the molecular classification hypotheses based on Spinks and Shaffer (2005) and Spinks et al. (2010) was the best overall classification scheme (Fig. 4). For both methods, the distribution of bootstrapped AUC for the molecular hypothesis was significantly greater MANN-WHITNEY U TEST than all of the other classification scheme. Remarkably, the best classification hypothesis was 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 value of the best multinomial logistic regression model compared to the best random forest model can be observed as the classifications are much closer to the reference classes (Fig. 5). The best random forest model misclassified many of the observations as the northern clade instead of the correct class. This systematic misclassification 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. 6). The mean shape of the northern clade is the most similar to the mean shape of the entire dataset (Fig. 6, 8), which may indicate that specimens that are closer to the mean shape will be systematically misclassified as the northern clade.

The results of fitting the final random forest model also include the variable importance for best separating the different classes. The final random forest model of the best classification scheme indicated that after 7 PCs were included as features, AUC would not increase. Of these 7 features, the first three are illustrated here (Fig. 7) in descending order of importance SUPPLEMENT WITH VARIABLE IMPORTANCE INFORMATION?.

The first two most important features describe different aspects of variation (Fig. 8). The third PC, or first most important PC, describes variation in the relative position of landmarks on anterior and posterior portions of the plastron and represents 9.6697% of total variation. The eighth PC, or second most important PC, mostly describes variation in landmarks along the midline of the plastron and represents 4.1061% of total variation. The major variations along these axes correspond well to the differences between the class means and the mean plastron shape (Fig. 6) where major class differences seem based on 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 mitochrondial 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 of sample size, methodology, or both.

The lack of coherent geographical subclass assignment from PAM clustering (Fig. 2) as well as the large number of features necessary before no increase in AUC for all models (Fig. 3) indicates that the morphological variation between subclasses is extremely fine grained. This was also exemplified by the small differences in mean class shape for the final chosen classification scheme (Fig. 8).

The methods presented here for supervised learning analysis of the landmark variation represent a compromise between explicitly modeling all shape variation and preventing models from being overfit and ungeneralizable. While all aspects of shape may be evolving simultaneously, and not along individual PCs, including all PCs as predictors in each model might increase model complexity beyond a reasonable level for the sample size or the necessary complexity to accurately model the response. However, this compromise is not without its advantages. Because the AICc and AUC values improved rapidly with increased model complexity (Fig. 3),k this indicated how fine scale the actual variation between classes was and how necessary including many PCs was. Additionally, the relative risk values from the mulitinomial logistic regression models demonstrate MORE STUFF HERE. Also, the systematic misclassification of the final random forest model and the possible relationship with the similarity the mean shape of the northern clade and the overall mean shape (Fig. 5, 6) is a good indicator of the very fine scale variation between classes.

Ultimately, it would be most useful to not require such explicit classification hypotheses, especially when concerned about possible cryptic variation in extinct taxa. The only unsupervised method employed in this study, PAM, is rather simple and not model based. Comparison here was facilitated via a summary statistic and bootstrapped confidence intervals. A more useful approach 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 asses the best clustering solution.

In this study we have demonstrated that given a large sample size and alternative methodology, it is possible to determine which classification scheme best explains variation in a taxon amongst a set of alternative hypotheses. The observed 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. These methods represent new applications which may be important for future studies interested in class-based morphological comparison and variation, both in the context of cryptic diversity and with known classifications.

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 symetric 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.

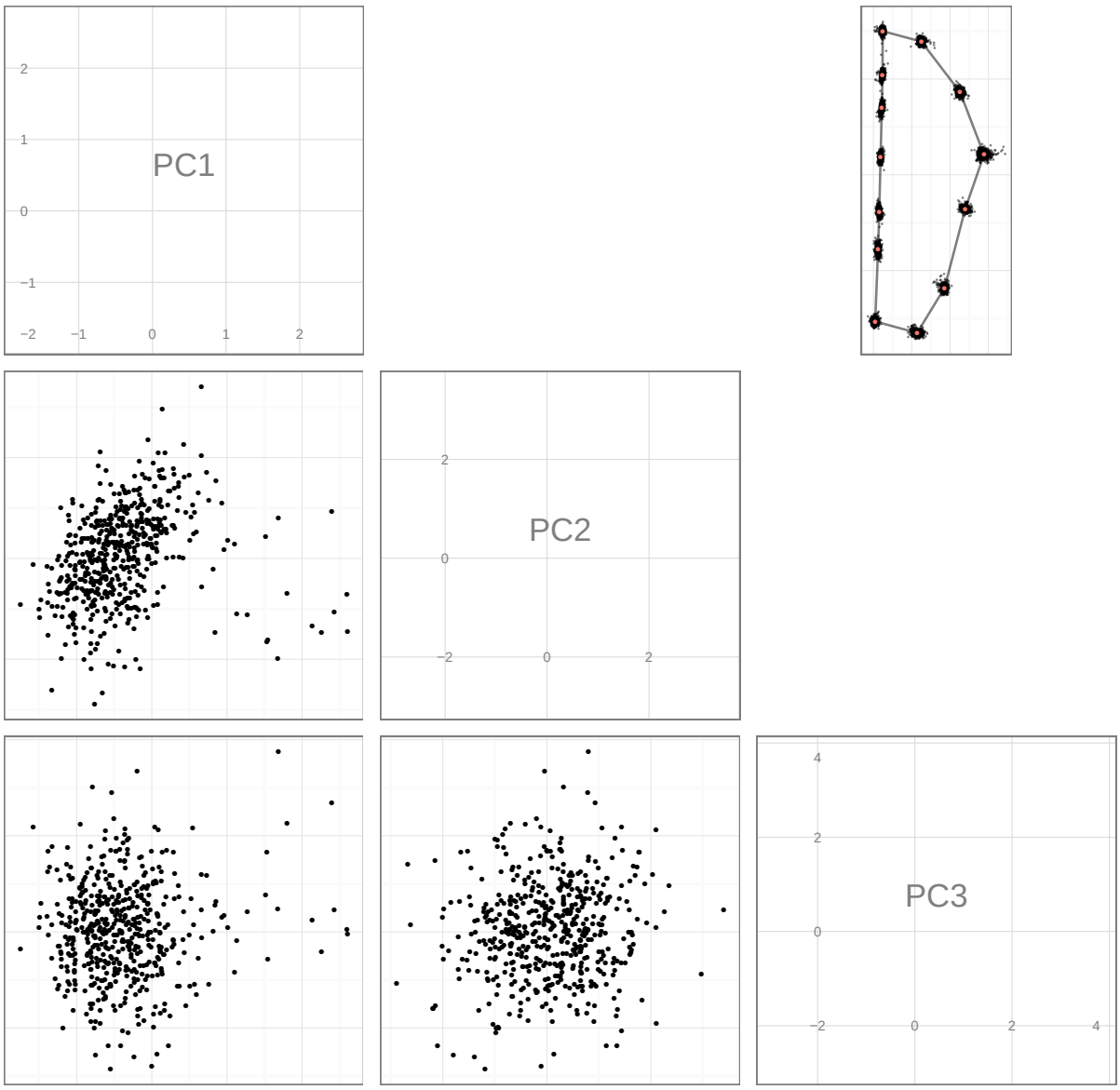

Figure 1

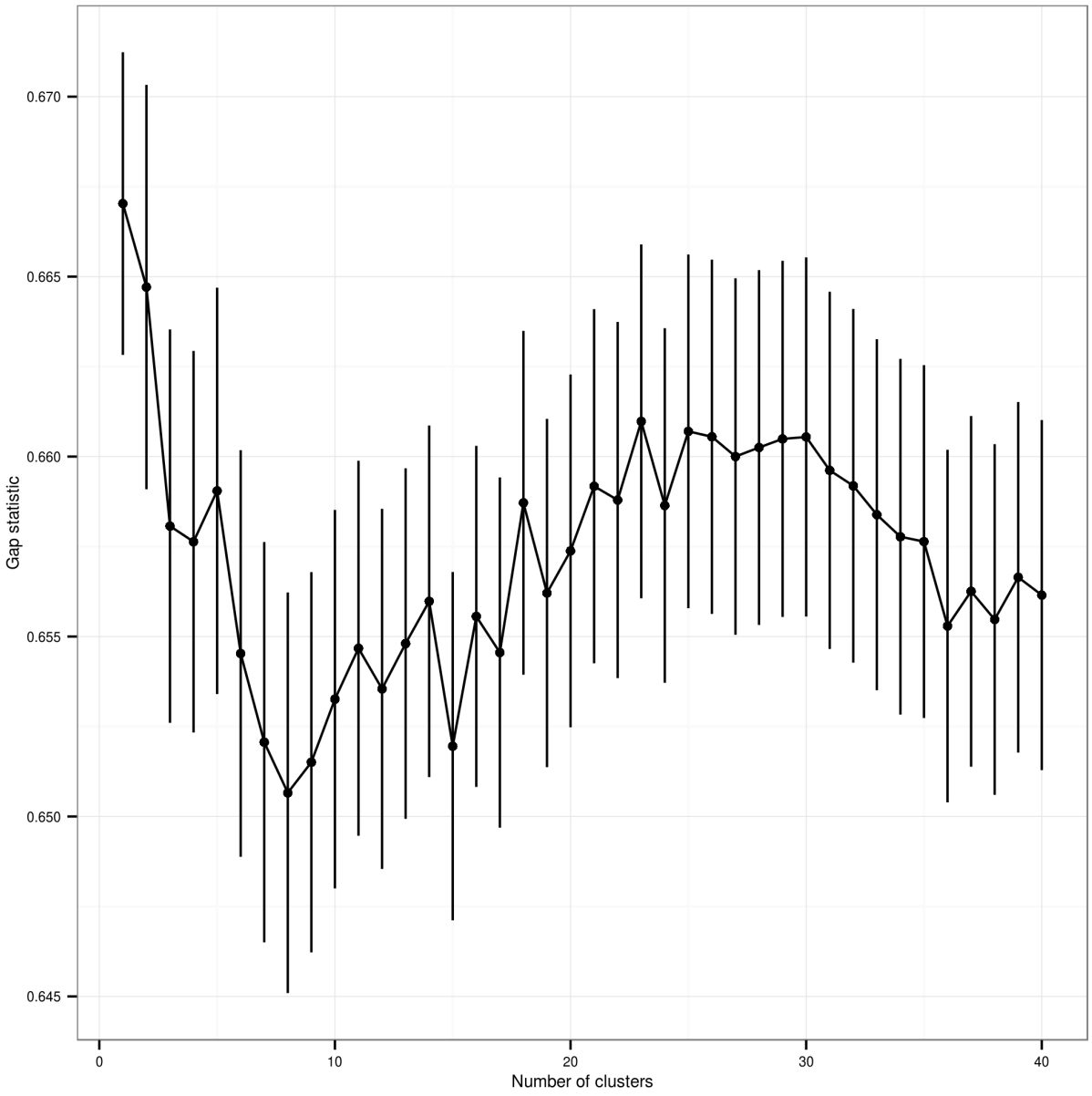

Figure 2: Gap statistic values for PAM clustering results for the ρ dissimliarity matrix of plastron shape. Error bars are standard errors estimated via 500 bootstrap samples.

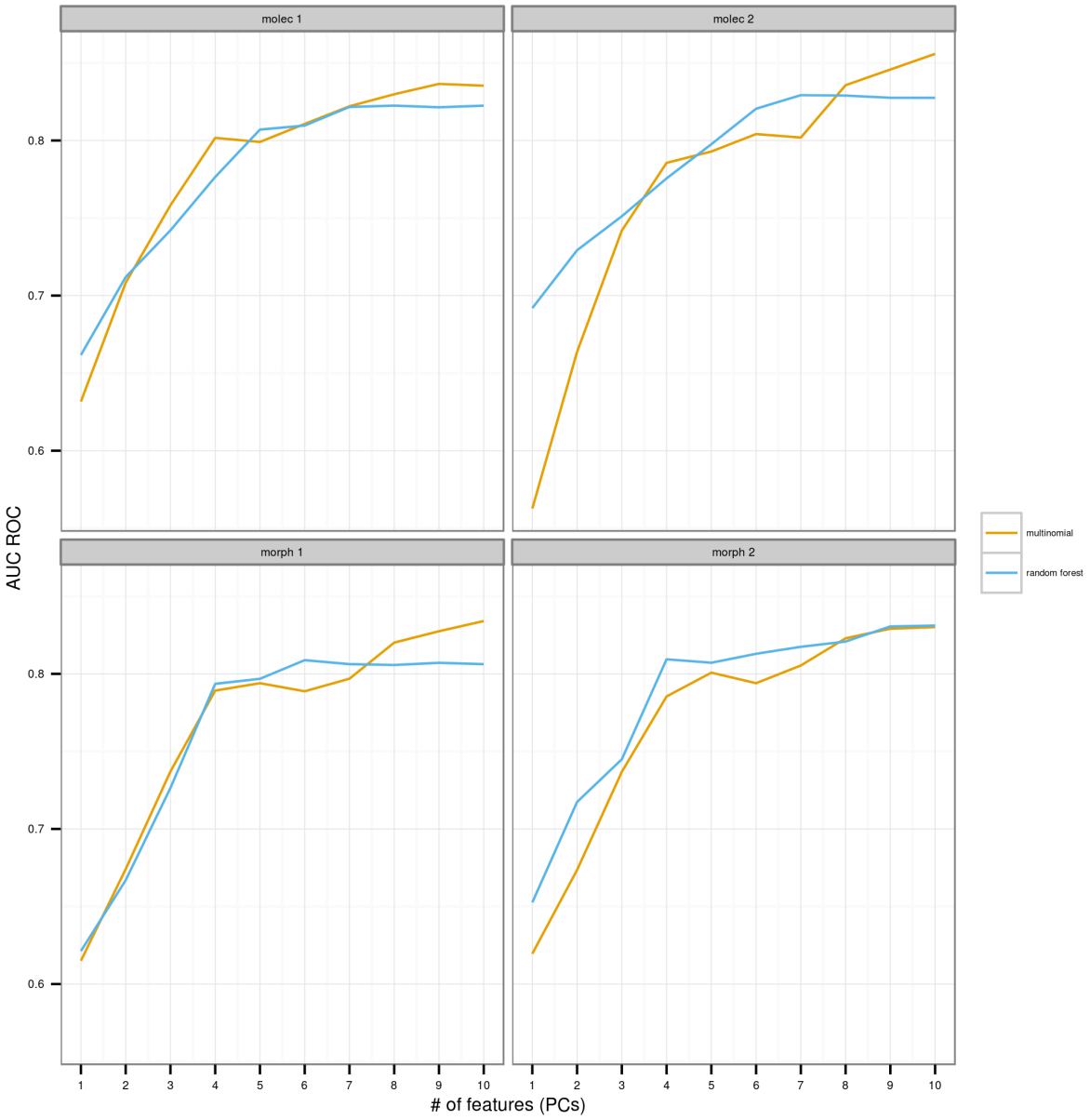

Figure 3: Effect of increasing the number of PCs as features, or predictors, of classification of plastra for all four classification schemes. As the number of PCs increase, AUC increases until eventually leveling off. Both multinomial logisitic regression and random forest models are illustrated here, though AUC based model selection was only performed for random forest models.

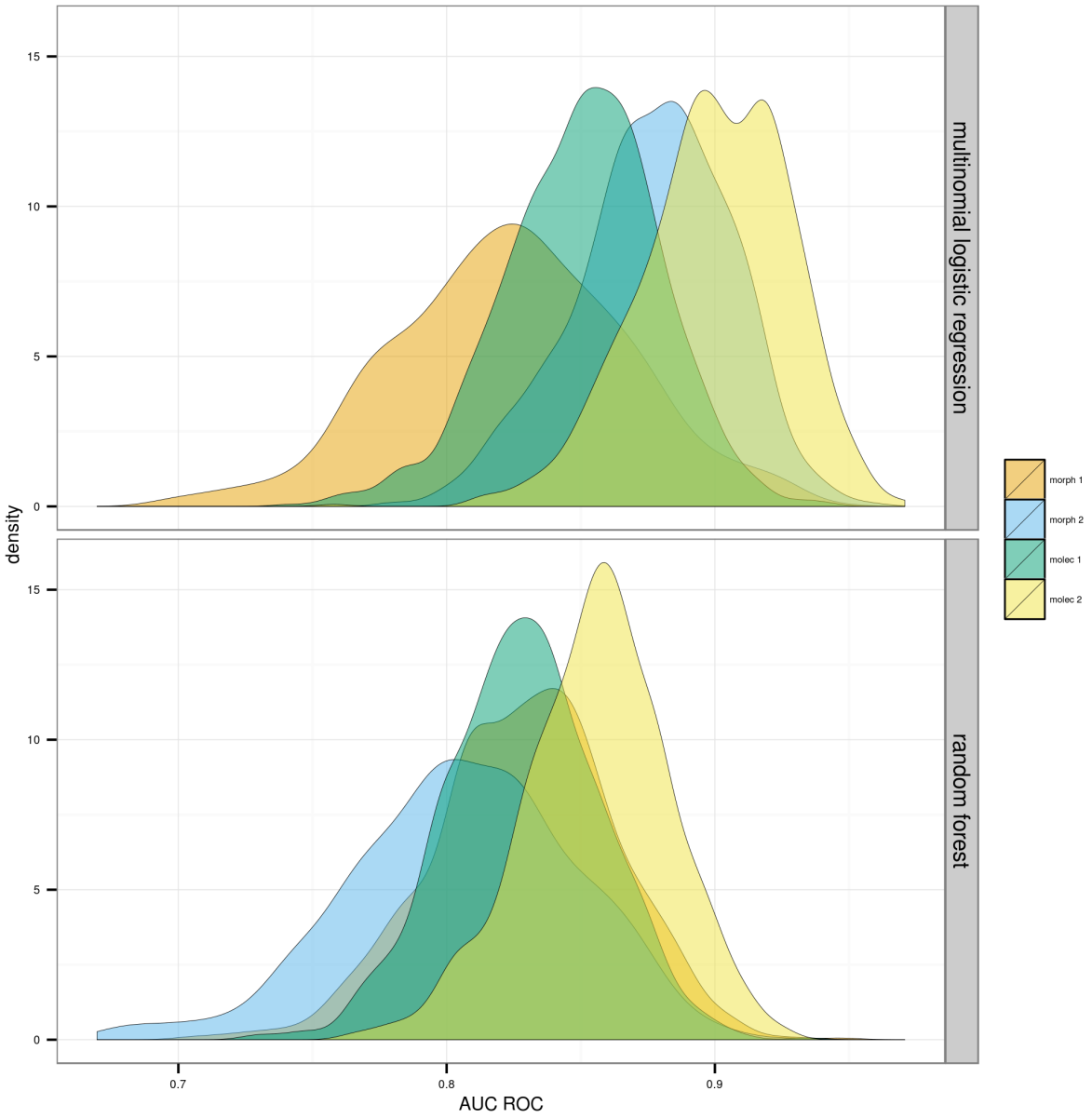

Figure 4: Density estimates of AUC values of predictions of the testing dataset of plastra from 1000 bootstrap resamples. The top facet corresponds to values using the optimal multinomial logistic regression model, as chosen by minimum AICc value. The bottom facet corresponds to the values using the optimal random forest model, as chosen by maximum AUC value.

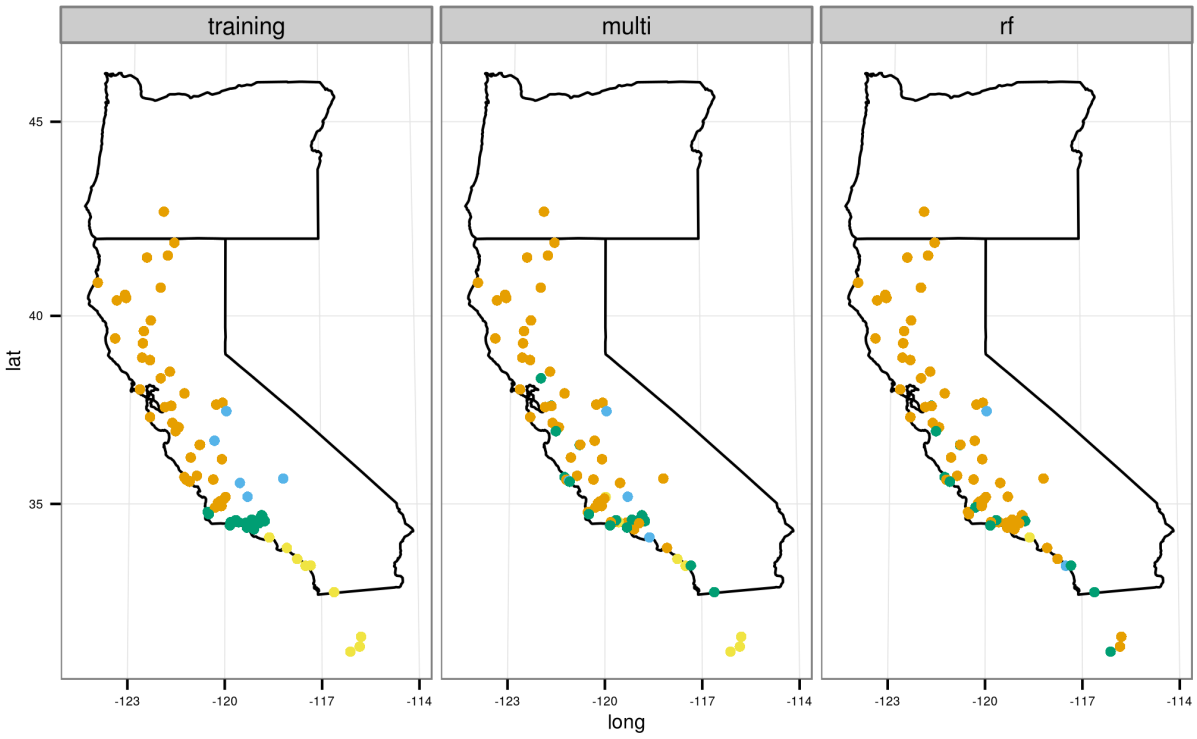

Figure 5: Comparison between reference classification of testing data set and the estimated classifications based on the selected multinomial logistic regression and random forest models, from left to right respectively. Classification corresponds to the four classes as suggested by the hypothesis of Spinks and Shaffer (2005) and Spinks et al. (2010).

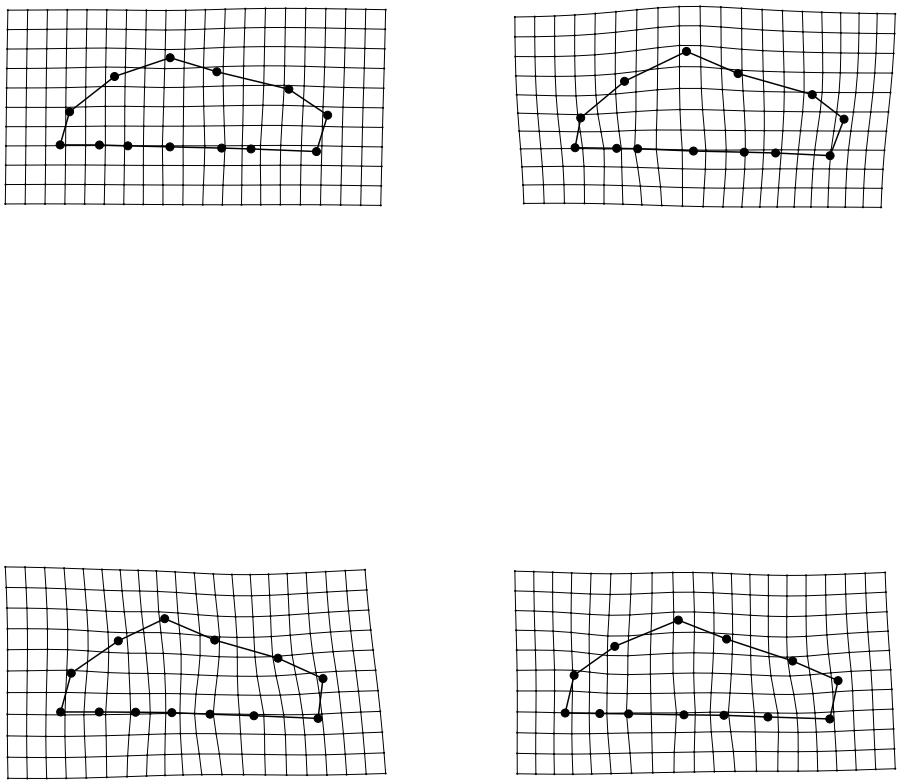

Figure 6: 2x magnification

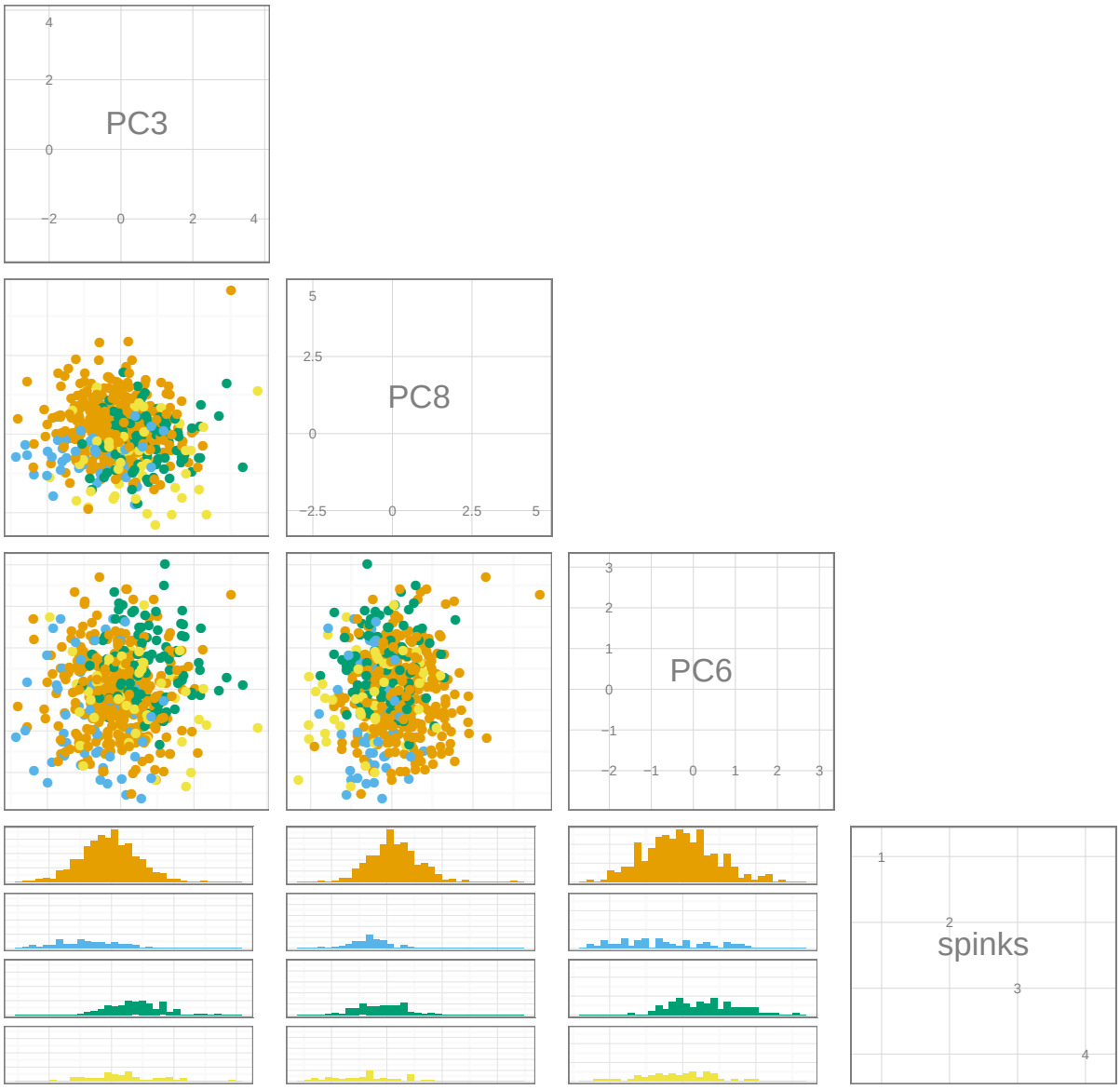

Figure 7: Pairs plot of the first three most important variables of the optimal random forest model of turtle plastral shape. The variables descend in importance from the upper left to the lower right. The observations are colored as in Figures 4 and 5. The bottom row are histograms of classification occurrences along the PCs.

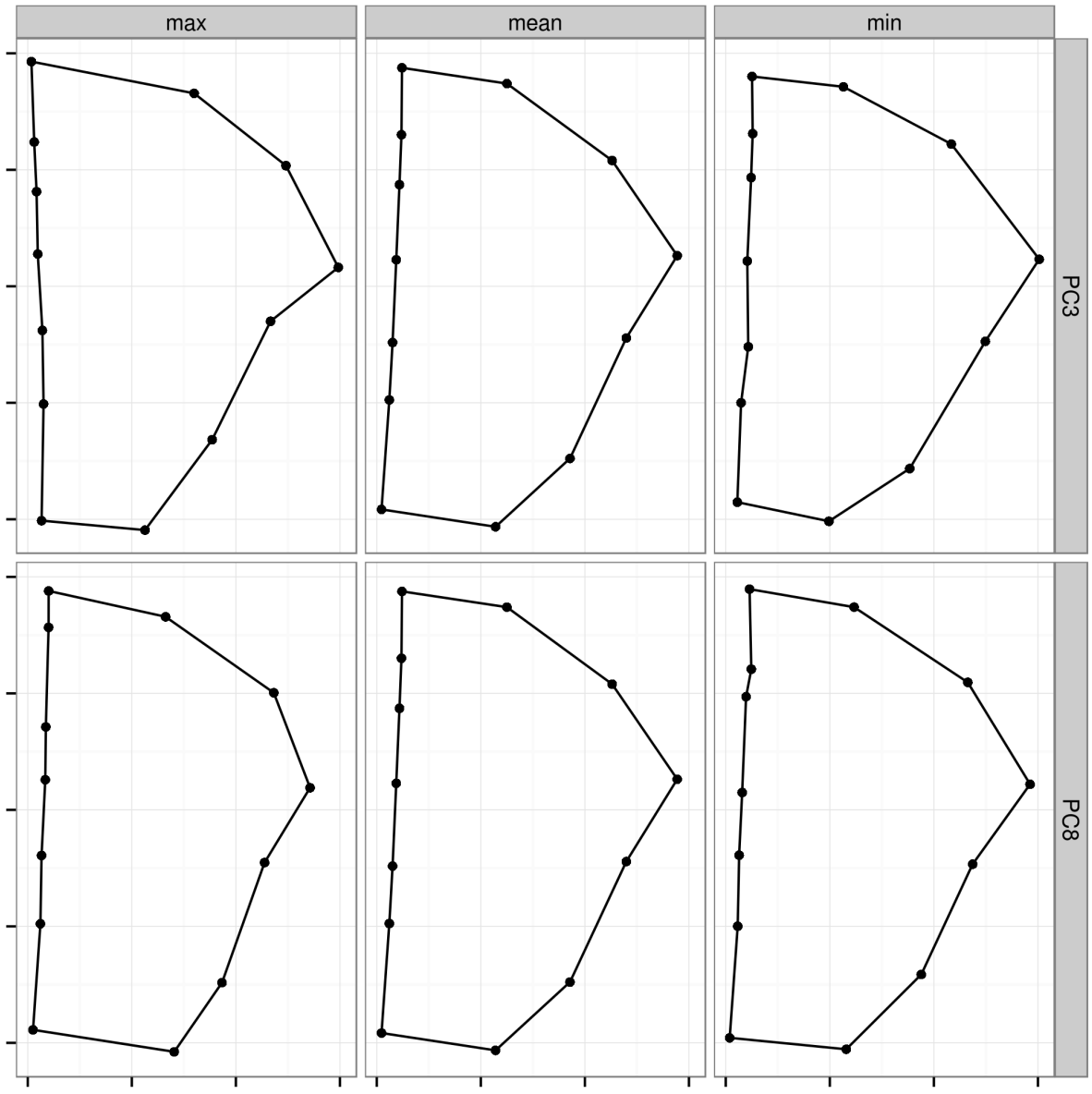

Figure 8: Landmark variation along the two most important features (PCs) based on the final random forest model. The first row corresponds to the third PC and the second corresponds to eighth PC. Landmark configurations are minimum observed on that PC, mean shape, and maximum observed on that PC.