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 Evolution Biology, University of Chicago
[2]Department of Geology, Field Museum of Natural History
[3]Department of Geological Sciences, California State University – Fullerton

July 3, 2013

Abstract

1 Introduction

Cryptic diversity is when taxa were only first deliminated via molecular means and were not or cannot deliminated via morphological identification CITATION. The discovery of this previously unknown diversity has

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

[*] psmits@uchicago.edu
[†] kangielczyk@fieldmuseum.org
[‡] jparham@fullerton.edu

can provide corroboration for subspecific assignment, and if it is possible to determine the best classification hypothesis amongst a few.

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). More recently, *E. marmorata* was divided into four subgroups based on mitochrondial DNA: a northern clade, a southern clade, and two central Californian clades (Spinks and Shaffer, 2005; Spinks et al., 2010).

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.

2 Materials and Methods

2.1 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.

2.2 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 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 of shape were calculated. This was done using the `shapes` package for R (Dryden, 2013; R Core Team, 2013).

2.3 Machine learning analyses

2.3.1 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 ρ (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 partioning around mediods (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 mediods 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).

2.3.2 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.

In order to prevent over fitting each machine learning model, tuning parameters were estimated using 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 PACKGE.

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. 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.

3 Results

3.1 Geometric morphometrics

3.2 Machine learning analyses

3.2.1 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?. Increasing the number of clusters does appear to improve the gap statistic enough to merit comparison.

3.2.2 Supervised learning

For all classification schemes, the optimal random forest model based on recursive feature selection by maximizing AUC ROC of the model was one with many features (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, (Fig. 4)

4 Discussion

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.

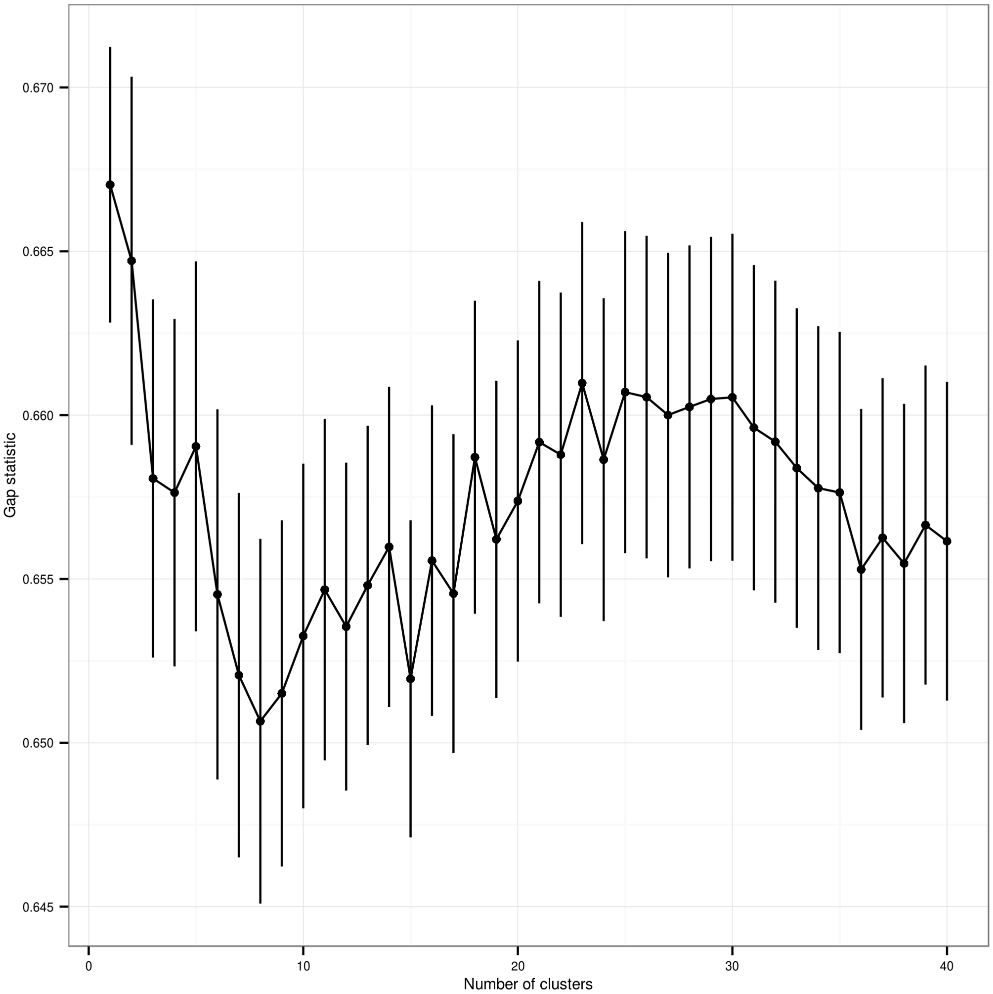

Figure 1: Gap statistic values for PAM clustering results for the ρ dissimliarity matrix of plastron shape. Error bars are standard errors estimated via 500 bootstrap samples.

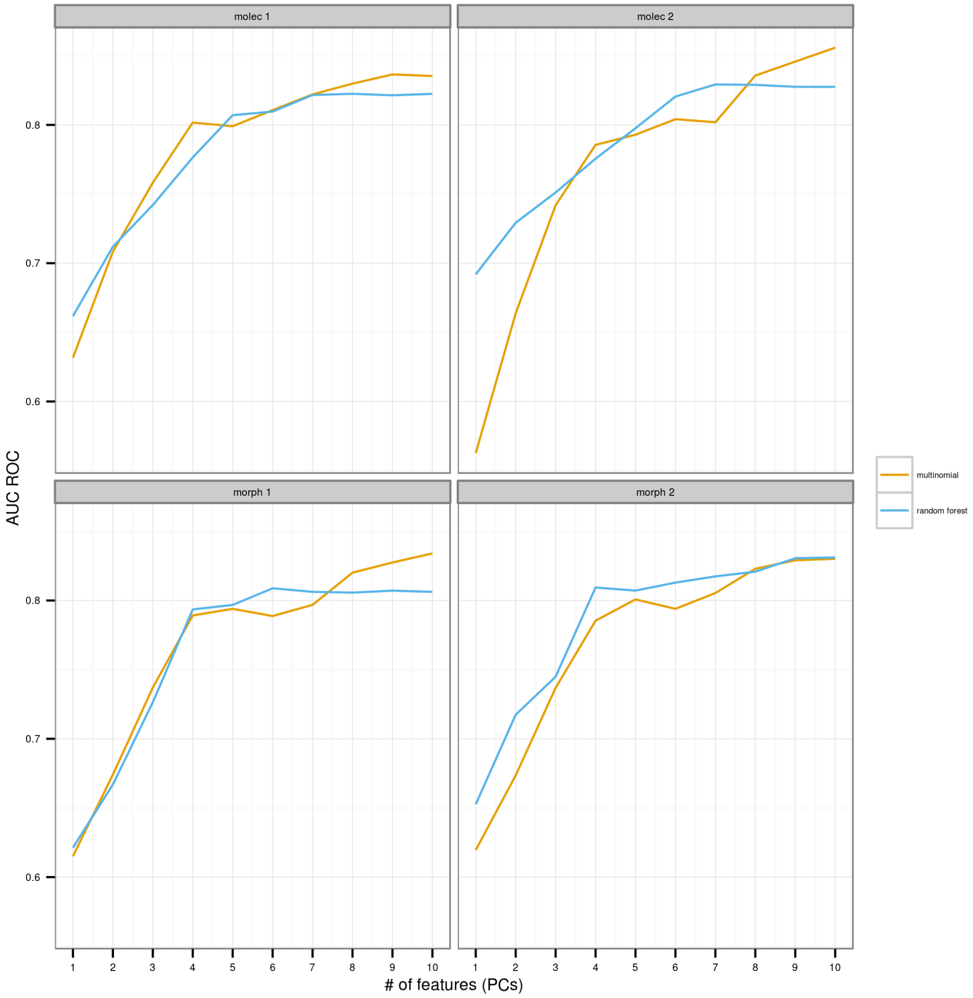

Figure 2: Effect of increasing the number of PCs as features, or predictors, of classification of plastra for all four classification schemes. As the number of PCs increase, AUC ROC increases until eventually leveling off. Both multinomial logisitic regression and random forest models are illustrated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

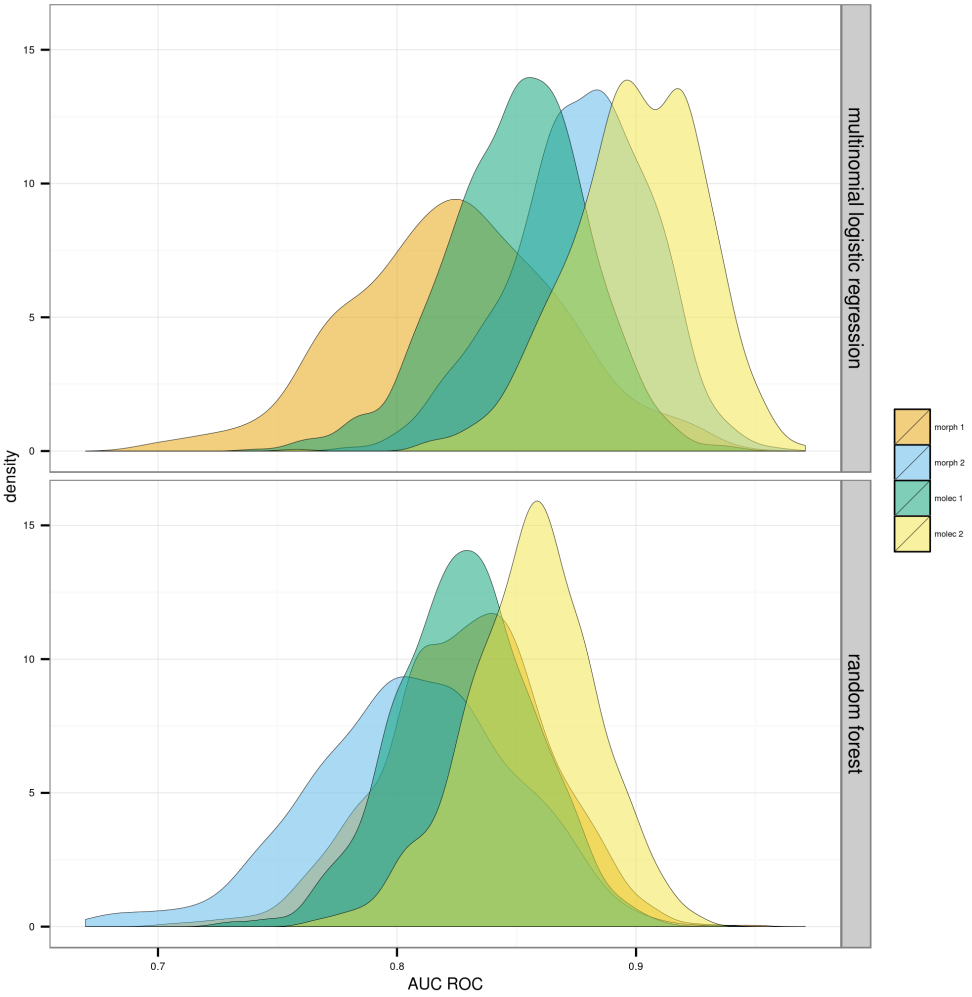

Figure 3: Density estimates of AUC ROC values of predictions of the testing dataset of plastra from 1000 bootstrap resamples. The top facet corresponds to values using the optimal multinomial logistic regression model, as chosen by minimum AICc value. The bottom facet corresponds to the values using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

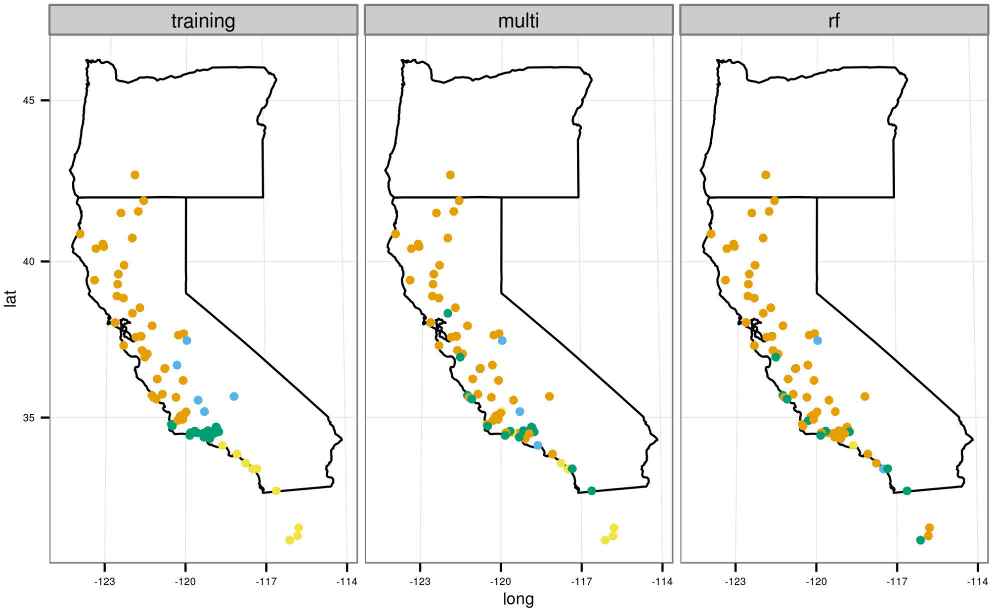

Figure 4: Comparison between reference classification of testing data set and the estimated classifications based on the selected multinomial logistic regression and random forest models, from left to right respectively. Classification corresponds to the four classes as suggested by the hypothesis of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