

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}

¹Committee on Evolution Biology, University of Chicago

²Department of Geology, Field Museum of Natural History

³Department of Geological Sciences, California State University – Fullerton

June 28, 2013

Abstract

2

1 Introduction

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

8 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

†kangelczyk@fieldmuseum.org

‡jparham@fullerton.edu

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

12 In this study, we address the subspecific classification scheme of *Emys*
marmorata, or western pond turtle. *E. marmorata* has a distribution from
14 northern Washington State, USA to Baja California, Mexico. Traditionally,
E. marmorata was classified into three subgroups: the northern *E. marmorata*
16 *marmorata*, the southern *E. marmorata palida*, and a central Californian
intergrade zone (Seeliger, 1945). More recently, *E. marmorata* was divided
18 into four subgroups based on mitochondrial DNA: a northern clade, a southern
clade, and two central Californian clades (Spinks and Shaffer, 2005, 2009).

20 In this study, we apply multiple machine learning approaches to esti-
mate the best classification scheme of *E. marmorata* subspecies based on
22 morphological variation in plastral shape.

2 Materials and Methods

2.1 Specimens

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

28 Specimens were given a class assignment was based on geographic informa-
tion. Because the exact geographic barriers between different class is unknown
and fuzzy, two assignments for both morphological and molecular hypotheses
30 of class were used.
32

2.2 Geometric morphometrics

34 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
36 of the keratinous plastral scutes that cover the plastron. These landmarks were
chosen to maximize the description of plastral variation. 12 of these landmarks
38 are symmetrical across the axis of symmetry and in order to prevent degrees
of freedom and other concerns (Klingenberg et al., 2007), these landmarks
40 were reflected across the axis of symmetry and the average position of each
symmetrical pair was used. In cases where damage or incompleteness prevented

42 symmetric landmarks from being determined, only the single member of the
pair was used. Analysis was then conducted on the resulting “half” plastra.
44 “Half” plastra landmark configurations were superimposed using general-
ized Procrustes analysis (Dryden and Mardia, 1998) after which, the principal
46 components of shape were calculated. This was done using the `shapes` package
for R (Dryden, 2013; R Core Team, 2013).

48 **2.3 Machine learning analyses**

2.3.1 Unsupervised learning

50 PAM was conducted using the `cluster` package for R (Maechler et al., 2013).

2.3.2 Supervised learning

52 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
54 total dataset, split proportionally per class, while the testing dataset was used
to estimate the effectiveness of each classification scheme (i.e. performance in
56 the wild).

Multinomial logistic regression models were fit using the `nnet` package for
58 R (Venables and Ripley, 2002) while random forest models were fit using the
`randomForest` package for R (Liaw and Wiener, 2002).

60 In order to prevent over fitting each machine learning model, tuning
parameters were estimated using 10-fold cross-validation (CV) across a grid
62 search of all tuning parameter combinations. Optimal tuning parameter values
were selected based on area under the receiver operating characteristic curve
64 (AUC ROC). Multiclass AUC ROC was estimated using the all-against one
strategy described in Hand and Till (2001).

66 For the multinomial logistic regression models, PCs were added sequentially
in order to increase the overall amount of variation in shape included in each
68 model and the final model was that with the lowest AICc Burnham and
Anderson (2002). This procedure was used because the optimal number of
70 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
72 estimate class, this may cause the model to be over fit and not provide an
accurate estimate of unsampled plastral variation. The maximum number
74 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
76 estimate that many parameters accurately.

Because random forest models are not fit using maximum likelihood, a
78 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
80 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
82 of the tuning parameters as well as quantify the uncertainty of each model.
Like the multinomial logistic regression models, 10 was the maximum number
84 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
86 R (Kuhn, 2013).

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

3 Results

92 3.1 Geometric morphometrics

3.2 Machine learning analyses

94 3.2.1 Unsupervised learning

3.2.2 Supervised learning

96 4 Discussion

Acknowledgements

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

100 **References**

- 102 Angielczyk, K. D., C. R. Feldman, and G. R. Miller. 2011. Adaptive evolution of plastron shape in emydine turtles. *Evolution* 65:377–394.
- 104 Burnham, K. P., and D. R. Anderson. 2002. Model selection and multi-model inference: a practical information-theoretic approach. 2nd ed. Springer, New York.
- 106 Dryden, I. L. 2013. shapes: Statistical shape analysis. R package version 1.1-8.
- 108 Dryden, I. L., and K. Y. Mardia. 1998. Statistical shape analysis. Wiley, New York.
- 110 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.
- 112 Klingenberg, C. P., M. Barluenga, and A. Meyer. 2007. Shape analysis of symmetric structures: quantifying variation among individuals and asymmetry. *Evolution* 56:1909–1920.
- 116 Kuhn, M. 2013. caret: Classification and Regression Training. R package version 5.15-61.
- 118 Liaw, A., and M. Wiener. 2002. Classification and regression by randomforest. *R News* 2:18–22.
- 120 Maechler, M., P. Rousseeuw, A. Struyf, M. Hubert, and K. Hornik. 2013. cluster: Cluster Analysis Basics and Extensions. R package version 1.14.4.
- 122 R Core Team. 2013. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
- 124 Rohlf, F. J. 2005. TpsDig 2.04.
- Seeliger, L. M. 1945. Variation in the Pacific Mud Turtle. *Copeia* 1945:150–159.
- 126 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.

- 130 ———. 2009. Conflicting mitochondrial and nuclear phylogenies for the
widely disjunct *Emys* (Testudines: Emydidae) species complex, and what
132 they tell us about biogeography and hybridization. *Systematic biology*
58:1–20.
- 134 Venable, W. N., and B. D. Ripley. 2002. *Modern Applied Statistics with S*.
4th ed. Springer, New York.