

How cryptic is cryptic diversity? Machine learning approaches to classifying morphological variation in the Pacific Pond Turtle (*Emys marmorata*)

Peter D Smits¹, Kenneth D Angielczyk^{1,2}, and James F Parham³

¹Committee on Evolutionary Biology, University of Chicago

²Integrative Research Center, Field Museum of Natural History

³Department of Geological Sciences, California State University – Fullerton

August 26, 2015

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

ABSTRACT NEEDS TO BE WRITTEN

INTRODUCTION

Molecular systematics has repeatedly demonstrated the existence of cryptic species that can only be diagnosed using genetic data (Stuart et al. 2006; Bickford et al. 2007; Schilck-Steiner et al. 2007; Pfenninger and Schwenk 2007; Clare 2011; Funk et al. 2012). In attempts to streamline the documentation of biodiversity, several methods of species delimitation that rely almost entirely on genetic data have recently been proposed (Yang and Rannala 2010). Although strong caveats on the utility of these methods have been raised (Bauer et al. 2000; Carstens et al. 2013), they are nevertheless being used to name species (Leaché and Fujita 2010; Spinks et al. 2014).

The majority of extant taxa, and almost all extinct taxa, are delimited by morphology alone. This disjunction complicates interpretations of variation and diversity in deep time, as apparent morphological stasis may not reflect the true underlying diversity (Eldredge and Gould 1972; Gould and Eldredge 1977; Hunt 2008; Van Bocxlaer and Hunt 2013). Similarly, for many museum specimens of extant taxa (e.g. those preserved in formalin), it is difficult to acquire the genetic data needed for non-morphological species delimitation methods.

These considerations have sparked interest in whether geometric morphometric analyses can capture fine-scale variation that can be used for identifying cryptic species. Most such studies focus on using morphometrics to discover differences between taxa that were identified by other means (Polly 2003; Zelditch et al. 2004; Gaubert et al. 2005; Gündüz et al. 2007; Polly 2007; Demandt and Bergek 2009). Additionally, there has been a fair amount of work on automated taxon identification and classification of taxa into groups (Baylac et al. 2003; Dobigny et al. 2003; MacLeod 2007; van den Brink and Bokma 2011). In cases where genetic data are ambiguous or lacking for many samples, morphometric approaches could help identify cryptic species. This would make the task of identifying and maintaining endangered or conserved groups much easier and could contribute to improved classifications of extinct taxa and populations.

Here, we investigate the morphometric identification of cryptic species using machine-learning approaches. In particular, we ask whether it is possible to determine which amongst a set of classification hypotheses is best and examine the implications of the results for a recently proposed set of cryptic turtle species.

Background and study system

Machine learning is an extension of known statistical methodology (Hastie et al. 2009) that emphasizes high predictive accuracy and generality at the expense of the interpretability of individual parameters. The basic statistical mechanics are supplemented by randomization, sorting, and partitioning algorithms, along with the maximization or minimization of summary statistics, in order to best estimate a general model for all data, both sampled and unsampled (Hastie et al. 2009). Machine learning approaches have found use in medical research, epidemiology, economics and automated image identification such as handwritten zip codes (Hastie et al. 2009). The two major classes of machine learning methods are unsupervised and supervised learning. Unsupervised learning methods are used with unlabeled data where the underlying structure is estimated, and they are analogous to clustering and density estimation methods (Kaufman and Rousseeuw 1990). Supervised learning methods are used with labeled data where the final output of data is known and the rules for going from input to output are inferred. These are analogous to classification and regression models (Breiman et al. 1984). Our application of the alternative approaches used in this study illustrates only a sampling of the various previously-derived methods for clustering observations and fitting classification models.

Geometric morphometric approaches to identifying differences in morphological variation between different classes, including cryptic species, mostly have used methods like linear discriminate analysis and canonical variates analysis (Polly 2003; Zelditch et al. 2004; Gaubert et al. 2005; Gündüz et al. 2007; Polly 2007; Francoy et al. 2009; Sztencel-Jabonka et al. 2009; Mitrovski-Bogdanovic et al. 2013). Because of their similarity to multivariate approaches like principle components analysis (PCA), these methods are comparatively straightforward ways of understanding the differences in morphology between classes. They also benefit from producing results that can be easily visualized, which aids in the interpretation and

presentation of data and results. Most previous morphometric studies did not assess which amongst a set of alternative classification hypotheses was optimal. For example, studies such as those of Caumul and Polly (2005) and Polly (2007) focused on comparing different aspects of morphology and their fidelity to a classification scheme instead of comparing the fidelity of one aspect of morphology to multiple classification schemes. In this context, the study of Cardini et al. (2009) is noteworthy because they compared morphological variation in marmots at the population, regional, and species level and determined the fidelity of shape to divisions at each of these levels.

Here, we used multiple machine learning methods, both unsupervised and supervised, to compare different classification hypotheses. These methods provide different and unique advantages for understanding how to classify taxa, as well as the accuracy of the resulting classifications. Although machine learning methods such as neural networks have been applied to studying shape variation (Baylac et al. 2003; Dobigny et al. 2003; MacLeod 2007; van den Brink and Bokma 2011), including in the context of automated taxon identification and classification of groups, the number of cases remains limited. In the current study, we not only consider pure classification accuracy but also use a statistic of classification strength that reflects the rate at which taxa are both accurately and inaccurately classified.

We analyzed the problem of whether there are distinct subspecies or cryptic species within the western pond turtle, *Emys marmorata* BAIRD CITATIONS (formerly *Clemmys marmorata*; see Feldman and Parham 2002). *Emys marmorata* is distributed from northern Washington State, USA to Baja California, Mexico. Traditionally, *E. marmorata* was classified into two named subspecies: the northern *E. marmorata marmorata* and the southern *Emys marmorata pallida* (Seeliger 1945), with a central Californian intergrade zone in between. *Emys marmorata marmorata* is differentiated from *E. marmorata pallida* by the presence of a pair of triangular inguinal scales and darker neck markings. The triangular inguinal plates can sometimes be present in *E. marmorata pallida* though they are considerably smaller. Seeliger (1945) did not formally include the Baja California populations of *E. marmorata* in either taxon, implying the existence of a third distinct but unnamed subspecies.

Previous work on morphological variation in *E. marmorata* has focused primarily on differentiation between populations over a portion of the species' total range (Lubcke and Wilson 2007; Germano and Rathbun 2008; Germano and Bury 2009; Bury et al. 2010); comparatively few studies have included specimens from across the entire range (Holland 1992). Most of these studies considered how local biotic and abiotic factors may contribute to differences in carapace length and found that size can vary greatly between different populations (Lubcke and Wilson 2007; Germano and Rathbun 2008; Germano and Bury 2009). There also has been interest in size-based sexual dimorphism in *E. marmorata* (Holland 1992; Lubcke and Wilson 2007; Germano and Bury 2009), with males being on average larger than females based on total carapace length and other linear measurements. However, the quality of size as a classifier of sex can vary greatly between populations (Holland 1992) because of the amount of size differences among populations (Lubcke and Wilson 2007; Germano and Bury 2009). However, the effect of sexual dimorphism on shape, *sensu* Kendall (1977), has not

been assessed (Holland 1992; Lubcke and Wilson 2007; Germano and Rathbun 2008).

Of particular importance in the context of cryptic diversity in *E. marmorata* is the morphometric analysis of carapace shape carried out by Holland (1992), who compared populations of *E. marmorata* from three areas of its range. This study concluded that geographic distance was a poor indicator of morphological differentiation, and instead hypothesized that geographic features such as breaks between different drainage basins are probably more important barriers to dispersal and interbreeding. Additionally, (Holland 1992) suggested that morphological differences were more pronounced as the magnitude of barriers and distance increased, but this variation required many variables to adequately capture, implying only very subtle morphological differentiation between putatively distinct populations. Finally, Holland concluded that *E. marmorata* is best classified as three distinct species: a northern species, southern species, and a Columbia Basin species. This classification is similar to that of Seeliger (1945), except elevated to the species level and without recognition of a distinct Baja species.

More recently, the phylogeography of *E. marmorata* and the possibility of cryptic diversity was investigated using molecular data (Spinks and Shaffer 2005; Spinks et al. 2010, 2014). Based on mitochondrial DNA, Spinks and Shaffer (2005) recognized four subclades within *E. marmorata*, a northern clade, a San Joaquin Valley clade, a Santa Barbara clade, and a southern clade. Analyses with nuclear DNA (Spinks et al. 2010) with single-nucleotide polymorphism (SNP) data suggest a primarily north–south division in *E. marmorata*, although the dataset differed from that of Spinks and Shaffer (2005) in the location of this break point. These studies discussed the potential taxonomic implications of their results, with Spinks et al. (2014) going so far as to strongly advocate for the recognition of at least two species (*E. marmorata* and *E. pallida*), and a possible third based on populations in Baja California. However, they did not discuss in detail the morphological characters that would help to diagnose these species beyond those specified by Seeliger (1945). Given that these characters are somewhat variable within the proposed species, and that Holland (1992) described shell shape variation that might be consistent with this taxonomy, a geometric morphometric analysis of shell shape might provide a reliable way to diagnose groups (whether species or subspecies) within *E. marmorata*.

In this study, we attempt to estimate the best classification scheme of *E. marmorata* based on variation in plastron (ventral shell) shape in order to determine whether this character is consistent with any of the past divisions based on other morphological features or molecular data. We are particularly interested in whether any aspect of plastron shape can be used to reliably diagnose Spinks et al. (2014) proposed species, and if so, the nature of that shape variation.

Because of unclear geographic boundaries between subgroups of *E. marmorata*, we compare multiple hypotheses of morphologically– and molecularly–based classification. We hypothesize that if morphological variation corresponds to class assignment, it should be possible to determine the best classification hypothesis of *E. marmorata* from amongst multiple candidate hypotheses. However, if morphological variation does not correspond to any of the standing

hypothesis, then supervised learning model generalization performance will be poor.

MATERIALS AND METHODS

Specimens, sampling, morphometrics

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), Angielczyk et al. (2011), and Angielczyk and Feldman (2013), essentially representing the adult individuals included in those papers. We chose to focus on adults because significant changes in plastron shape occur over the course of ontogeny in *E. marmorata* and other emydines (Angielczyk and Feldman 2013). PETER NOTE – HOW WAS SOMETHING CLASSIFIED AS AN ADULT?

We assigned a classification to each specimen for the different binning schemes based on geographic occurrence data recorded in museum collection archives. When precise latitude and longitude information were not available we estimated them from locality information. Because Spinks and Shaffer (2005), Spinks et al. (2010), and Spinks et al. (2014) did not use vouchered specimens we were not able to examine the sample individuals. Therefore our 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 schemes for a total of six different schemes. The schemes differed based on where geographic boundaries were assigned. This changes the classification of certain individuals near the boundaries between groups, providing a test of the robustness of the classification schemes.

Following previous work on plastron variation (Angielczyk and Sheets 2007; Angielczyk et al. 2011; Angielczyk and Feldman 2013), we used TpsDig 2.04 (Rohlf 2005) to digitize 19 landmarks (Fig. 1). Seventeen of the landmarks are at the endpoints or intersection of the keratinous plastral scutes that cover the plastron. Twelve of the landmarks were symmetrical across the axis of symmetry and, in order to prevent issues related to degrees of freedom and other similar 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 (R Core Team 2013; Dryden 2013).

Machine learning analyses

174 *Unsupervised learning.*— In order to preserve the relationship between all landmark configurations in shape space, we measured the dissimilarity between observations using Kendall's
176 Riemannian shape distance or ρ (Kendall 1984; Dryden and Mardia 1998). We chose this metric because shape space, or the set of all possible shape configurations following Procrustes
178 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
180 a concern in the case of the half plastral landmark configurations used in the study.

We divisively clustered the shape dissimilarity matrix using partitioning around mediods
182 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
184 dissimilarities between observations and mediods is minimized (Kaufman and Rousseeuw 1990). Because the optimal number of clusters of shape configurations in the study was
186 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 compared clustering solutions
188 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).

190 *Supervised learning.*— We used three different supervised learning, or classification, approaches: linear discriminate analysis, multinomial logistic regression, and random forests. Linear
192 discriminate analysis, also known as canonical variate analysis, is commonly used in studies of geometric morphometric data (Zelditch et al. 2004; Mitteroecker and Bookstein 2011).
194 The other two methods, however, are not. In all cases, the optimal number of PCs used as predictors was chosen via maximum within-sample AUC value, explained below.

196 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
198 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
200 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
202 subset of the total number of PCs, ranging from two to 6 in increasing order of complexity. In total, this produced nine different LDA scaling matrices.

204 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
206 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.

208 Random forest models are an extension of classification and regression trees (CART) (Breiman et al. 1984; Breiman 2001). The goal of CARTs is to use a series of different features (i.e.
210 predictors) to estimate the class of an observation. In top-down induction of decision trees for each member of a given set of predictor variables, attribute value tests are used to estimate
212 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 useful property for 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 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 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, 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 such as this study where misclassification needs to be minimized just as much as an accurate classification must be obtained. 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 our 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, a process 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 the individual specimen class assignments to the testing or training sets.

To determine if there were systematic differences in plastron shape between the correctly and incorrectly classified samples, we used a permutation test to estimate if the dissimilarity between the correctly and incorrectly classified individuals was significantly greater than random. The group labels were permuted 1000 times and the distance 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.

Comparison with clear-cut example

In addition to the *E. marmorata/pallida* case, we also analyzed a set of seven morphologically distinct species in order to test if the approaches used here actually work. The seven species analyzed were *Emys blandingii*, *Terrapene coahuila*, *Clemmys guttata*, *Glyptemys insculpta*, *Glyptemys muhlenbergii*, *Emys orbicularis*, and *Terrapene ornata* with a total of XXX specimens analyzed in total. Just as in the *E. marmorata/pallida* example, we analyzed plastron shape variation. The support for a classification scheme with seven categories was evaluated using random forest, linear discriminate analysis, and multinomial logistic regression where the number of PCs used as predictors ranged from 2 to 11. Data was split into training and testing sets and model performance for both sets was evaluated using the AUC metric, just as above.

RESULTS

Unsupervised learning

Comparison of gap statistic values from PAM clustering show that the optimal, minimal 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 there is no means of naturally partitioning plastron shape into distinct subgroups with out reference to external information.

Supervised learning

AUC-based model selection revealed some important patterns of variation and congruence between the classification schemes and the actual data. Generally, the best performing models tended to include as many PCs as possible (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. That almost all LDA and multinomial logistic regression 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 the PCs (5). This means that the variance describing the differences between the classes does not align with the major axes of variance (i.e. the PCs). Such a result is expected if variation between classes was extremely fine grained and not a part of the principal form or function of the plastron. Moreover, this result is congruous with the results from 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 (4). In most 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 (6).

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.

COMPARISON WITH CLEAR-CUT EXAMPLE

In contrast with the above results, the analysis of the clearly distinct species SUPPLEMENT demonstrates almost perfect classification of the both the in-sample and out-of-sample datasets (Table 4). These results demonstrate that if there were strong distinctions between the states of the classification schemes, the methods used here would be able to recover this.

DISCUSSION

The results of this study indicate that there is no clear grouping of *E. marmorata* based on plastron shape.

The unsupervised learning results indicate only a single group of observations being optimal, which 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 generally supported by the supervised learning results.

The results from fitting the various supervised learning models to each of the classification schemes generally shows that no one scheme is “best.” Possible explanations include that the genetic differentiation is not associated with plastron shape variation and/or that local selective pressures (e.g. from hydrological regime) overwhelm morphological differentiation. This makes sense given that the plastron is involved in both protection and hydrodynamics, and not necessary mate choice CITATIONS and that shell shape in *E. marmorata* is known to vary among populations inhabiting water bodies with different flow regimes CITATIONS. Plastron shape does not seem to preserve a strong phylogenetic signal at the interspecific level in emydine turtles (Angielczyk et al. 2011), and our current results suggest that this may be the case for phylogeographic signal within emydine species as well. Stated another way, plastron shape may be a poor morphological marker for phylogenetic and phylogeographic studies, despite the ease with which its shape can be quantified. A final possibility (explored below) is that the proposed classification schemes themselves do not represent significant evolutionary lineages.

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.

Is there more than one species of Western Pond Turtle?

The lack of morphological support for the distinctiveness of *E. pallida* does not, on its own, preclude the recognition of this taxon. However, this apparent lack of congruence does prompt a reexamination of the methods and concepts that led to that taxonomic revision. In other words, before we can assess the significance of the morphological non-diagnosability, it is essential to evaluate the methods and concepts that led to taxonomic revision.

Spinks et al. (2014) elevated *E. pallida* based on a Bayesian species delimitation analysis of SNP data using BPP (Yang and Rannala 2010). However, Spinks et al. (2014) did not heed the caveats about species delimitation methods raised by Carstens et al. (2013). In addition to specifically addressing the shortcomings of validation methods such as BPP that rely on guide trees and so should be interpreted with caution, Carstens et al. (2013) also strongly emphasize that “Inferences regarding species boundaries based on genetic data alone are likely inadequate, and species delimitation should be conducted with consideration of the life history, geographical distribution, morphology and behaviour (where applicable) of the focal

system. . .” These caveats evoke the development of the Unified Species Concept (Dayrat
2005; De Queiroz 2007), Integrative Taxonomy (Padial 2010), and other pluralist
approaches to species delimitation. None of these considerations were brought to bear on the
E. marmorata system until now, and in doing so we find the proposal that *E. pallida* is a
distinct species to be lacking.

In particular, the natural history and geographical distribution of *E. marmorata* and *E. pallida*
make the recognition of these taxa implausible. The data from Spinks et al. (2014) show
extensive introgression and admixture in Central California, which is expected because there
are no significant barriers to gene flow in this region. Combined with the well-demonstrated
ability for testudinoid turtles, including emydids and even *Emys*, to hybridize (e.g. Buskirk
et al. 2005; Spinks and Shaffer 2009; Parham et al. 2013) it is hard to imagine how *E.*
marmorata and *E. pallida* could maintain their integrity in the face of such admixture.
Because the geography, natural history, and demonstrated genetic admixture of *E. marmorata*
conflict with the recognition of *E. pallida*, we hypothesize that our inability to classify the
morphological data by proposed species is because *E. pallida* is not a distinct species. We
agree with Carstens et al. (2013) that “the inferences drawn from species delimitation studies
should be conservative, for in most contexts it is better to fail to delimit species than it is to
falsely delimit entities that do not represent actual evolutionary lineages.”

We fully agree with Spinks et al. (2014) that *E. marmorata* (*sensu lato*) is a species deserving
of strong conservation efforts, we do not wish to trivialize this need. Moreover, the genetic
diversity uncovered by the analysis of Spinks et al. (2014) should be explicitly accounted for
in any conservation plan. Given the apparent lack of morphological distinction, however, we
consider that this diversity should be considered Evolutionary Significant Units or Distinct
Population Segments instead of distinct species.

Finally, it is important to note that the data and analyses we present do not let us definitively
say whether the apparent lack of morphological divergence within *E. marmorata* truly reflects
the presence of a single species, or if it is an artifact of plastron shape being a poor morpho-
logical marker for phylogenetic and phylogeographic divergences. Tests of both our preferred
conclusion (*E. marmorata* as a single species) and that Spinks et al. (2014) should include
additional tests of species delimitation using different methods and search for corroborating
evidence as suggested by Carstens et al. (2013). From a morphological standpoint, support
for the validity of “*E. Pallida*” may come from other aspects of morphology, such as carapace
shape or other features that might be more informative. Likewise, further investigation of the
phylogeographic utility of plastron shape in other turtle species will help to clarify whether
the lack of differentiation seen in *E. marmorata* is typical or an unusual case.

*

Acknowledgements Data collection for this project was supported in part by NSF DBI-0306158
(to KDA). G. Miller assisted with data collection and her participation in this research was
supported by NSF REU DBI-0353797 (to R. Mooi of CAS). For access to emydine specimens,

404 we thank: J. Vindum and R. Drewes (CAS); A. Resetar (FMNH); R. Feeney (LACM); C.
Austin (LSUMNS); S. Sweet (MSE); J. McGuire and C. Conroy (MVZ); A. Wynn (NMNH);
406 P. Collins (SBMNH); B. Hollingsworth (SDMNH); P. Holroyd (UCMP). We are grateful for
S. Sweet for field assistance and the California Department of Fish and Game for permits.

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.
- Bauer, A. M., J. F. Parham, R. M. Brown, B. L. Stuart, L. Grismer, T. J. Papenfuss, W. Bohme, J. M. Savage, S. Carranza, J. L. Grismer, P. Wagner, A. Schmitz, N. B. Ananjeva, and R. F. Inger. 2000. Availability of new Bayesian-delimited gecko names and the importance of character-based species descriptions. *Proceedings of the Royal Society B: Biological Sciences* 278:490–492.
- Baylac, M., C. Villemant, and G. Simbolotti. 2003. Combining geometric morphometrics with pattern recognition for the investigation of species complexes. *Biological Journal of the Linnean Society* 80:89–98.
- Bickford, D., D. J. Lohman, N. S. Sodhi, P. K. L. Ng, R. Meier, K. Winker, K. K. Ingram, and I. Das. 2007. Cryptic species as a window on diversity and conservation. *Trends in ecology & evolution* 22:148–55.
- 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.
- Bury, R. B., D. J. Germano, and G. W. Bury. 2010. Population Structure and Growth of the Turtle *Actinemys marmorata* from the KlamathSiskiyou Ecoregion: Age, Not Size, Matters. *Copeia* 2010:443–451.
- Buskirk, S. W., J. F. Parham, and C. R. Feldman. 2005. On the hybridisation between two distantly related Asian turtles (Testudines: *Scalia* x *Mauremys*). *Salamandra* 41:21–26.
- Cardini, A., D. Nagorsen, P. O’Higgins, P. D. Polly, R. W. Thorington Jr, and P. Tongiorgi. 2009. Detecting biological distinctiveness using geometric morphometrics: an example case from the Vancouver Island marmot. *Ethology Ecology & Evolution* 21:209–223.
- Carstens, B. C., T. a. Pelletier, N. M. Reid, and J. D. Satler. 2013. How to fail at species delimitation. *Molecular ecology* 22:4369–83.
- Caumul, R. and P. D. Polly. 2005. Phylogenetic and environmental components of morphological variation: skull, mandible, and molar shape in marmots (*Marmota*, Rodentia). *Evolution; international journal of organic evolution* 59:2460–72.

Clare, E. L. 2011. Cryptic species? Patterns of maternal and paternal gene flow in eight
neotropical bats. *PloS one* 6:e21460.

Dayrat, B. 2005. Towards integrative taxonomy. *Biological Journal of the Linnean Society*
85:407–415.

De Queiroz, K. 2007. Species concepts and species delimitation. *Systematic Biology* 56:879–86.

Demandt, M. H. and S. Bergek. 2009. Identification of cyprinid hybrids by using geometric
morphometrics and microsatellites. *Journal of Applied Ichthyology* 25:695–701.

Dobigny, G., L. Granjon, V. Aniskin, K. Ba, and V. Voloboulev. 2003. A new sigling species
of *Taterillus* (Muridae, Gerbillinae) from West Africa. *Mammalian Biology* 68:299–316.

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.

Eldredge, N. and S. J. Gould. 1972. Punctuated equilibria: an alternative to phyletic gradual-
ism. Pages 82–115 *in* *Models in Paleobiology* (T. J. M. Schopf, ed.). Freeman Cooper, San
Francisco.

Feldman, C. R. and J. F. Parham. 2002. Molecular phylogenetics of emydine turtles: taxonomic
revision and the evolution of shell kinesis. *Molecular Phylogenetics and Evolution* 22:388–98.

Francoy, T. M., R. A. O. Silva, P. Nunes-Silva, C. Menezes, and V. L. Imperatriz-Fonseca.
2009. Gender identification of five genera of stingless bees (Apidae, Meliponini) based on
wing morphology. *Genetics and molecular research* 8:207–214.

Funk, W. C., M. Caminer, and S. R. Ron. 2012. High levels of cryptic species diversity
uncovered in Amazonian frogs. *Proceedings of the Royal Society B: Biological Sciences*
279:1806–14.

Gaubert, P., P. J. Taylor, C. a. Fernandes, M. W. Bruford, and G. Veron. 2005. Patterns
of cryptic hybridization revealed using an integrative approach: a case study on genetids
(Carnivora, Viverridae, *Genetta* spp.) from the southern African subregion. *Biological*
Journal of the Linnean Society 86:11–33.

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.

Germano, D. J. and G. B. Rathbun. 2008. Growth, population structure, and reproduction of
western pond turtles (*Actinemys marmorata*) on the Central Coast of California. *Chelonian*
Conservation and Biology 7:188–194.

Gould, S. J. and N. Eldredge. 1977. Punctuated equilibria: the tempo and mode of evolution
reconsidered. *Paleobiology* 3:115–151.

- Gündüz, I., M. Jaarola, C. Tez, C. Yenyurt, P. D. Polly, and J. B. Searle. 2007. Multigenic and morphometric differentiation of ground squirrels (*Spermophilus*, *Sciuridae*, *Rodentia*) in Turkey, with a description of a new species. *Molecular phylogenetics and evolution* 43:916–35.
- 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.
- Hunt, G. 2008. Gradual or pulsed evolution: when should punctuational explanations be preferred? *Paleobiology* 34:360–377.
- Kaufman, L. and P. J. Rousseeuw. 1990. Finding groups in data : an introduction to cluster analysis. Wiley, New York.
- Kendall, D. G. 1977. The diffusion of shape. *Advances in Applied Probability* 9:428–430.
- 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 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.
- Kuhn, M. and K. Johnson. 2013. Applied predictive modeling. Springer, New York, NY.
- Leaché, A. D. and M. K. Fujita. 2010. Bayesian species delimitation in West African forest geckos (*Hemidactylus fasciatus*). *Proceedings. Biological sciences / The Royal Society* 277:3071–7.
- Lubcke, G. M. and D. S. Wilson. 2007. Variation in shell morphology of the Western Pond Turtle (*Actinemys marmorata* Baird and Giarard) from three aquatic habitats in Northern California. *Journal of Herpetology* 41:107–114.
- MacLeod, N. 2007. Automated taxon identification in systematics: theory, approaches and applications. CRC Press, Boca Raton.
- Maechler, M., P. Rousseeuw, A. Struyf, M. Hubert, and K. Hornik. 2013. cluster: Cluster Analysis Basics and Extensions. R package version 1.14.4.
- Mitrovski-Bogdanovic, A., A. Petrovic, M. Mitrovic, A. Ivanovic, V. Žikic, P. Starý, C. Vorburger, and v. Tomanovic. 2013. Identification of two cryptic species within the *Praon*

abjectum group (Hymenoptera: Braconidae: Aphidiinae) using molecular markers and geometric morphometrics. *Annals of the entomological society of America* 106:170–180.

Mitteroecker, P. and F. Bookstein. 2011. Linear Discrimination, Ordination, and the Visualization of Selection Gradients in Modern Morphometrics. *Evolutionary Biology* 38:100–114.

Parham, J. F., T. J. Papenfuss, P. P. V. Dijk, B. S. Wilson, C. Marte, L. R. Schettino, and W. Brian Simison. 2013. Genetic introgression and hybridization in Antillean freshwater turtles (*Trachemys*) revealed by coalescent analyses of mitochondrial and cloned nuclear markers. *Molecular phylogenetics and evolution* 67:176–87.

Pfenninger, M. and K. Schwenk. 2007. Cryptic animal species are homogeneously distributed among taxa and biogeographical regions. *BMC evolutionary biology* 7:121.

Polly, P. D. 2003. Paleophylogeography of *Sorex araneus*: molar shape as a morphological marker for fossil shrews. *Mammalia* 68:233–243.

Polly, P. D. 2007. Phylogeographic differentiation in *Sorex araneus*: morphology in relation to geography and karyotype. *Russian Journal of Theriology* 6:73–84.

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.

Schilck-Steiner, B. C., B. Seifert, C. Stauffer, E. Christian, R. H. Crozier, and F. M. Steiner. 2007. Without morphology, cryptic species stay in taxonomic crypsis following discovery. *Trends in ecology & evolution* 22:391–392.

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. and H. B. Shaffer. 2009. Conflicting mitochondrial and nuclear phylogenies for the widely disjunct *Emys* (Testudines: Emydidae) species complex, and what they tell us about biogeography and hybridization. *Systematic biology* 58:1–20.

Spinks, P. Q., R. C. Thomson, and H. Bradley Shaffer. 2014. The advantages of going large: genome wide SNPs clarify the complex population history and systematics of the threatened western pond turtle. *Molecular Ecology* Pages n/a–n/a.

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.

Stuart, B. L., R. F. Inger, and H. K. Voris. 2006. High level of cryptic species diversity revealed by sympatric lineages of Southeast Asian forest frogs. *Biology letters* 2:470–4.

Sztencel-Jabonka, A., G. Jones, and W. Bogdanowicz. 2009. Skull Morphology of Two Cryptic Bat Species: *Pipistrellus pipistrellus* and *P. pygmaeus* A 3D Geometric Morphometrics Approach with Landmark Reconstruction. *Acta Chiropterologica* 11:113–126.

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.

Van Bocxlaer, B. and G. Hunt. 2013. Morphological stasis in an ongoing gastropod radiation from Lake Malawi. *Proceedings of the National Academy of Sciences* .

van den Brink, V. and F. Bokma. 2011. Morphometric shape analysis using learning vector quantization neural networks an example distinguishing two microtine vole species. *Annales Zoologici Fennici* 48:359–364.

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

Yang, Z. and B. Rannala. 2010. Bayesian species delimitation using multilocus sequence data. *Proceedings of the National Academy of Sciences* 107:9264–9.

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

Scheme	random forest		multinomial logistic regression		linear discriminate analysis	
	Observed	Generalized	Observed	Generalized	Observed	Generalized
Morph 1	0.63	0.73	0.75	0.79	0.75	0.80
Morph 2	0.61	0.58	0.76	0.77	0.76	0.77
Mito 1	0.63	0.62	0.75	0.63	0.75	0.63
Mito 2	0.77	0.67	0.80	0.64	0.80	0.63
Mito 3	0.56	0.64	0.71	0.74	0.71	0.73
Nuclear	0.56	0.67	0.74	0.62	0.74	0.77

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)
Morph 1	*
Morph 2	0.79
Mito 1	0.89
Mito 2	0.82
Mito 3	0.79
Nuclear	0.79

(b) multinomial logistic regression

Scheme	P(best - other > 0)
Morph 1	*
Morph 2	0.55
Mito 1	0.94
Mito 2	0.57
Mito 3	0.69
Nuclear	0.96

(c) linear discriminate analysis

Scheme	P(best - other > 0)
Morph 1	1
Morph 2	1
Mito 1	1
Mito 2	*
Mito 3	0.73
Nuclear	0.96

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)
Morph 1	CCR	1.59	0.77
	marm	2.06	0.87
Morph 2	CCR	1.81	0.88
	marm	2.16	1.00
Mito 1	CCR	2.37	0.94
	marm	2.37	0.99
Mito 2	marm	1.91	0.85
	pall	2.00	0.94
Mito 3	1	1.79	0.40
	3	3.30	0.97
Nuclear	marm	2.07	1.00
	pall	2.13	0.99

(b) multinomial logistic regression

Scheme	Class	distance	P(distance - simulated > 0)
Morph 1	CCR	2.06	1.00
	marm	2.22	0.93
Morph 2	CCR	2.50	1.00
	marm	2.60	1.00
Mito 1	CCR	2.39	0.99
	marm	2.24	0.98
Mito 2	marm	2.43	0.97
	pall	2.60	1.00
Mito 3	1	2.96	0.92
	3	3.18	0.99
Nuclear	marm	2.23	1.00
	pall	2.15	1.00

(c) linear discriminate analysis

Scheme	Class	distance	P(distance - simulated > 0)
Morph 1	CCR	2.07	1.00
	marm	2.22	1.00
Morph 2	CCR	2.20	1.00
	marm	1.87	0.98
Mito 1	CCR	2.75	0.98
	marm	2.36	0.47
Mito 2	marm	2.43	0.96
	pall	2.60	1.00
Mito 3	1	2.96	0.90
	3	3.33	1.00
Nuclear	marm	2.23	1.00
	pall	2.15	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.

	# predictors	In-sample AUC	Out-of-sample AUC
random forest	11	1.000	0.999
linear discriminate analysis	10	1.000	1.000
multinomial logistic regression	6	1.000	0.998

Table 4: Results from classification model estimates of the secondary, multi-species dataset. Comparison of in-sample and out-of-sample AUC of the best performing model, along with the number of predictors. AUC values range between 0.5 (no better than random) and 1 (perfect classification).

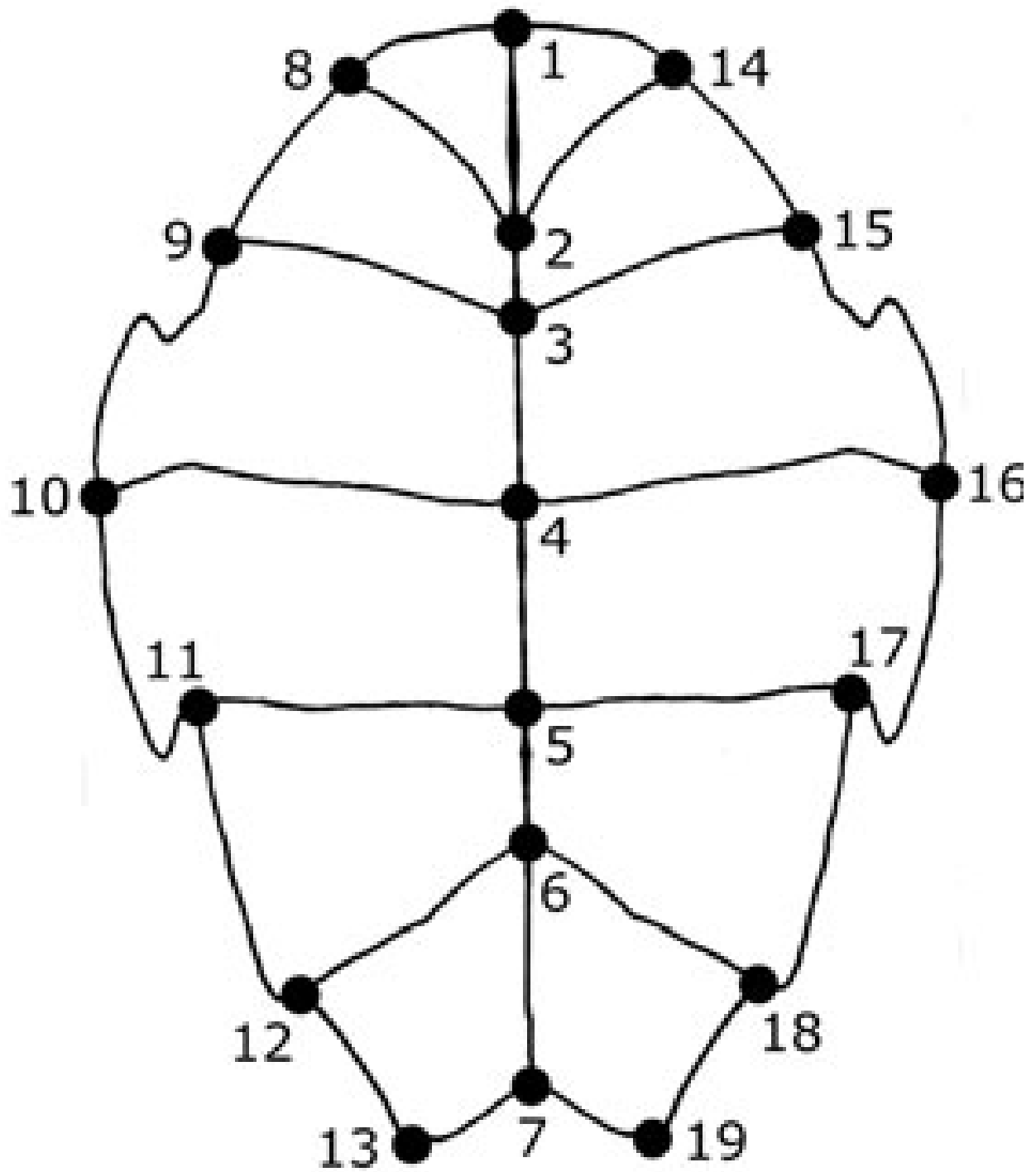


Figure 1: Depiction of general plastral shape of *E. marmorata* and position of the 19 landmark used in this study. Anterior is towards the top of the figure.

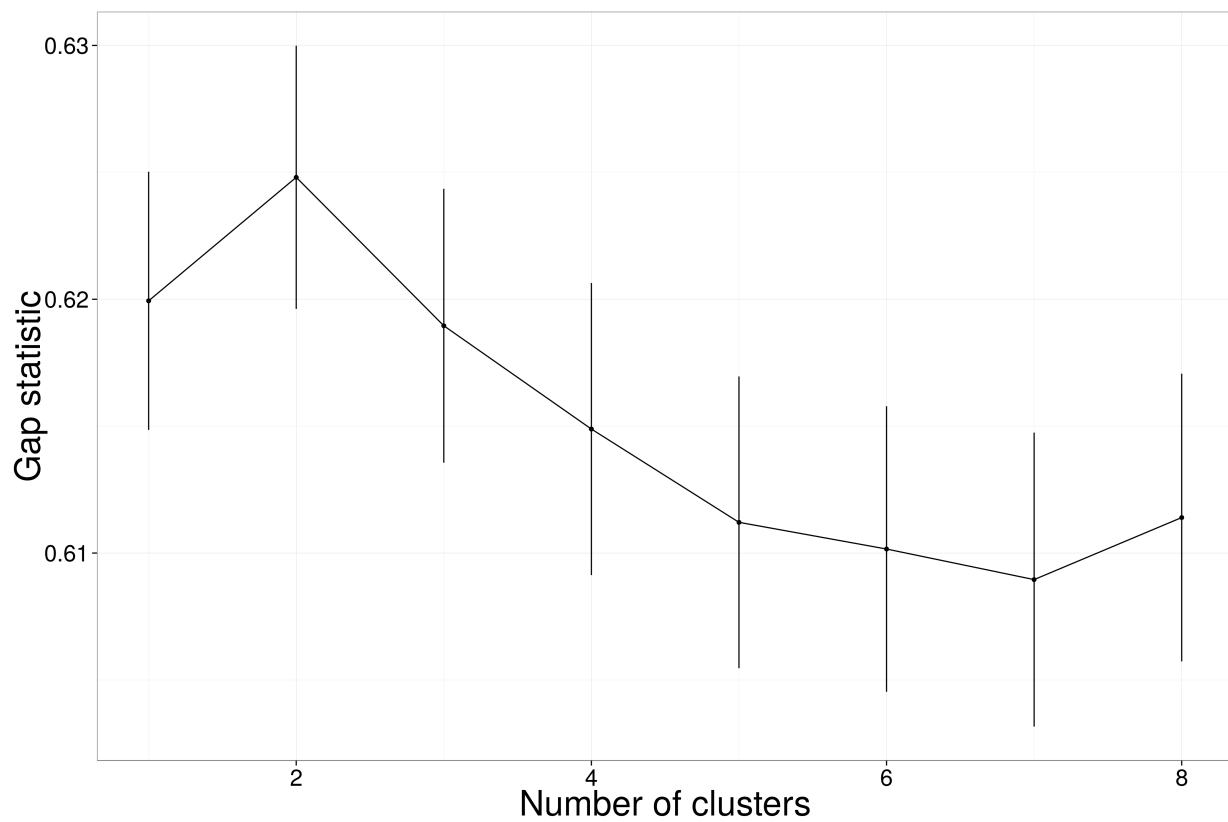


Figure 2: Results from PAM clustering of the Riemannian shape distance for 8 different number of clusters. Vertical lines are 1 standard deviation of the mean determined from 500 resamples.

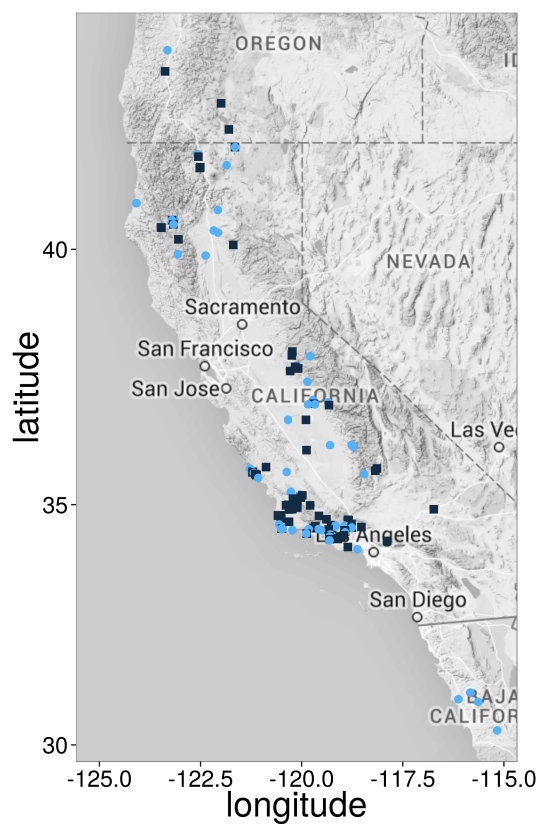


Figure 3: Comparison of geographic distribution of clustered observations from the 2 clustering PAM solution. Colour and shape correspond to each of the groups. There is clearly no geographic signal in the data.

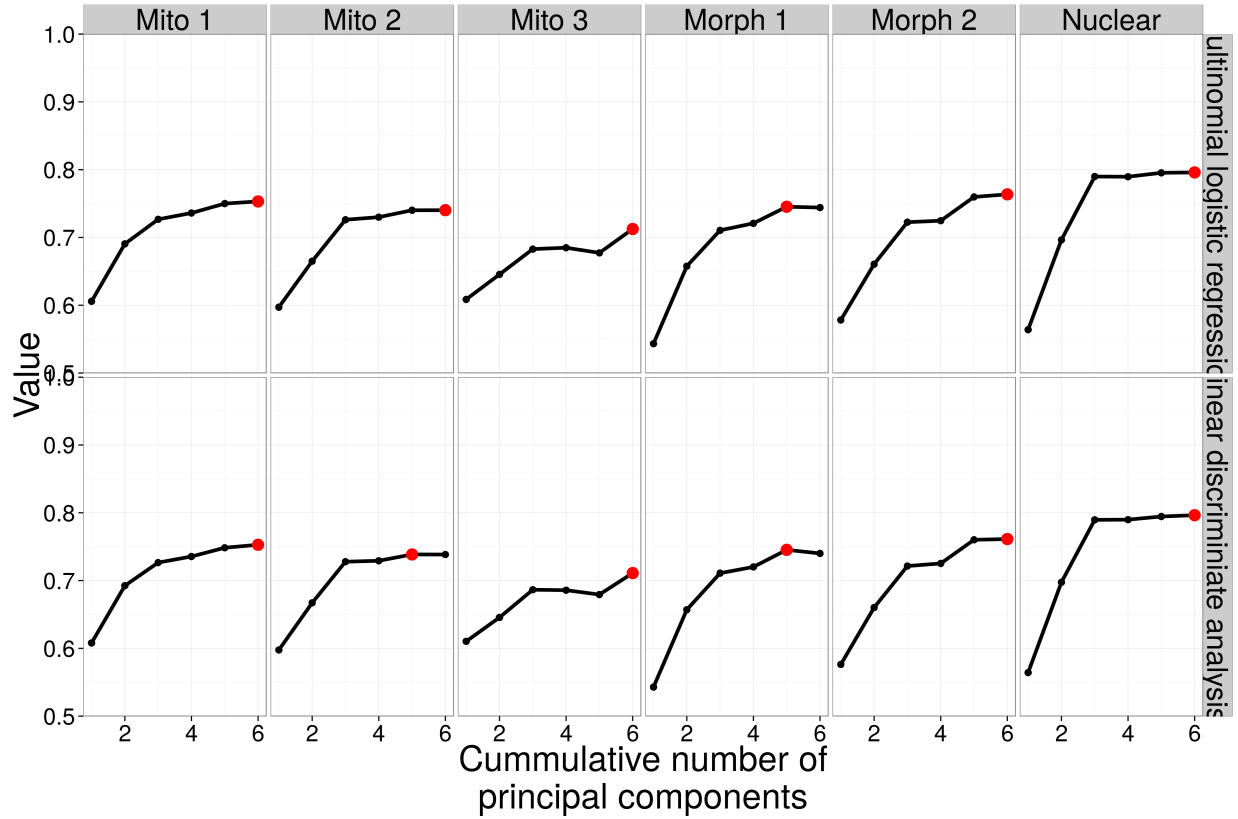


Figure 4: Graphical representation of the AUC values from model selection for multinomial logistic regression and linear discriminate analysis, respectively. AUC model selection is based on greatest AUC value. The horizontal axis corresponds to the cumulative number of axes included in the model of interest.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

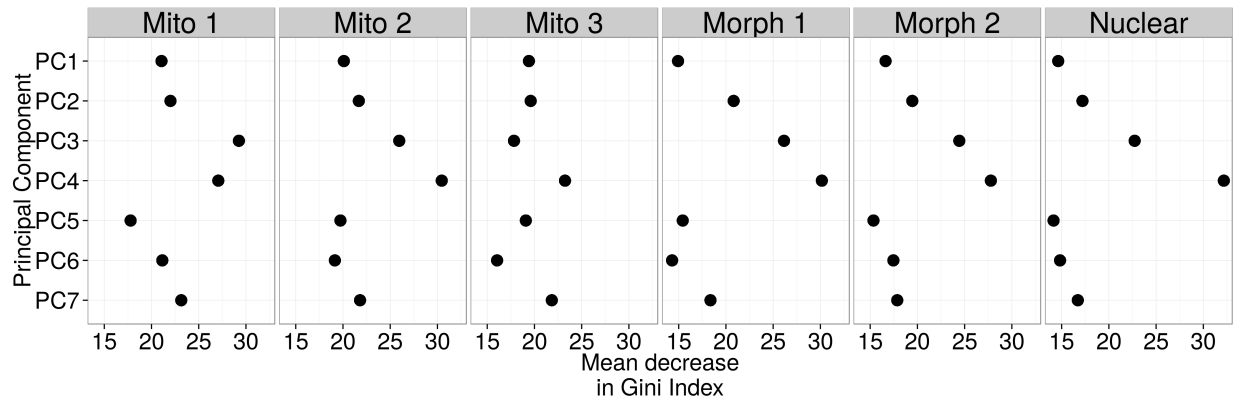


Figure 5: Variable importance from the random forest models for each of the six classification schemes. Importance is measured as the mean decrease in Gini Index, which is a measure of the strength by which that variable determines CART structure. Indices that are farther to the right indicate greater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

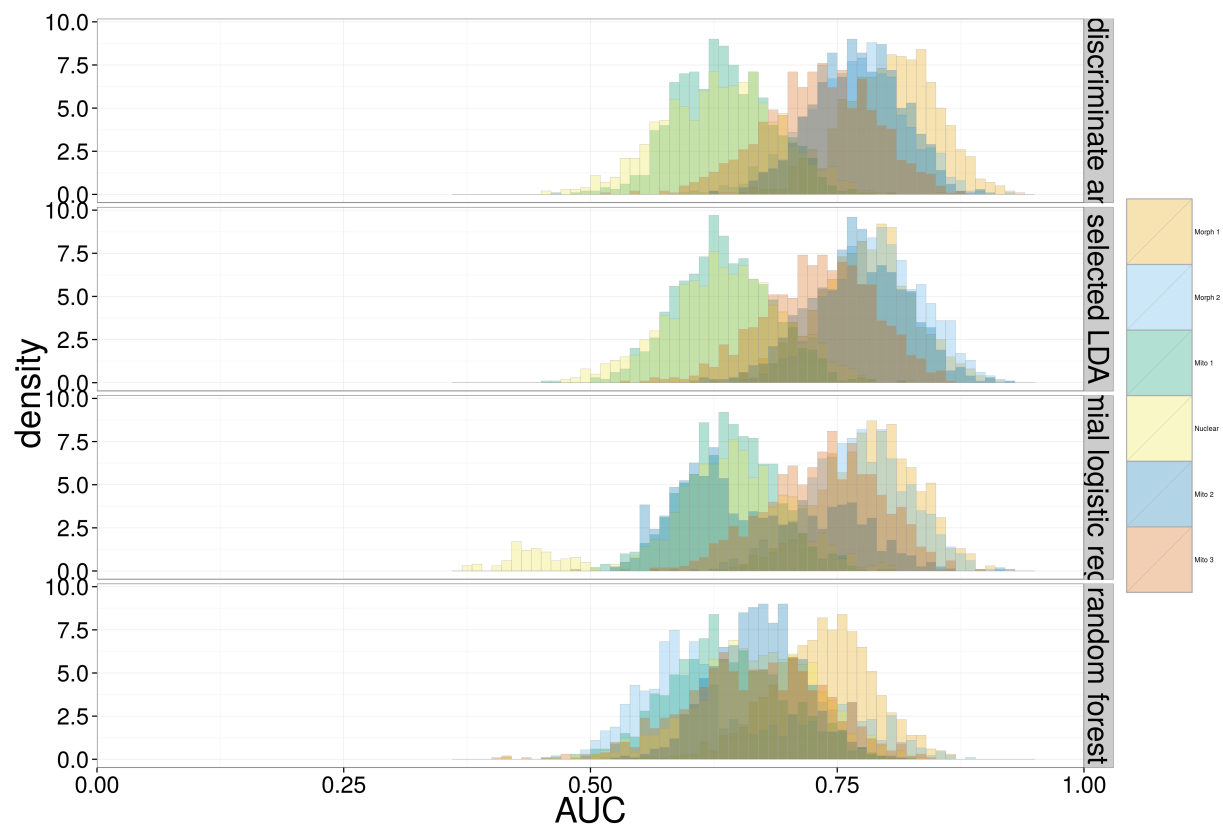


Figure 6: Bootstrap distributions for generalized AUC values for each of the classification schemes. Each row corresponds to a different modeling approach: LDA, LDA using best variables from random forest, multinomial logistic regression, and random forest. Each distribution corresponds to 1000 bootstrap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