

How cryptic is cryptic diversity? Machine learning approaches to classifying morphological variation in *Emys marmorata* (Testudinoidea, Emydidae).

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 21, 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

Cryptic diversity is the phenomenon where some taxa are believed to be identifiable only based on molecular data. This is concerning because the majority of extant taxa and virtually all extinct taxa are delimited entirely via morphology. Here we address questions about whether it is possible to determine, based on morphology, if one classification hypotheses can be considered better than others in order to determine if possible cryptic

variation is actually cryptic or just a case of extremely fine scale morphological variation.
8 Using a combination of unsupervised and supervised machine learning methods we
demonstrate a suite of approaches for better understanding differences in morphology
10 between classes, the odds of classifying one class relative to another, and what aspects of
morphology best describe the differences between classes. These approaches are applied
12 to the classification of the emydid turtle, *Emys marmorata*. This species has conflicting
hypotheses of the number of meaningful subclades based on either morphological
14 or molecular information. We compared multiple explicit classification hypotheses by
characterizing variation in plastral shape and how it may be identifiably different between
16 classes. By splitting a large dataset of specimens into both training and testing datasets,
we were also able to determine which of the classification hypotheses best corresponded
18 to the observed plastral variation in general. The results from our analysis shows that
the best classification of plastral variation in *Emys marmorata* is in accordance with
20 the molecularly based hypothesis. This demonstrates that, by using alternative methods
for characterizing variability, it is possible to estimate the classification scheme which
22 most agrees with observed variation. Additionally, we demonstrate how it is possible
that not all examples of cryptic variation are truly cryptic and may just be a product
24 of sample size or methodology because of the extremely fine scale variation between the
different classes.

26 (Keywords: Testudines, Emydidae, morphology, geometric morphometrics, random
forests)

28 Cryptic diversity is the phenomenon that not all taxa can be recognized from mor-
phology and can only be delimited using molecular information (Clare 2011; Funk et al. 2012;
30 Pfenninger and Schwenk 2007; Stuart et al. 2006). Concerningly, most extant taxa, and nearly
all extinct taxa, are delimited based solely via morphology. This phenomenon is of great
32 concern when studying variation and diversity dynamics over long periods of time, where
apparent morphological stasis may not actually reflect true diversity (Eldredge and Gould

³⁴ 1972; Gould and Eldredge 1977; Hunt 2008; Van Bocxlaer and Hunt 2013). In the case of
³⁶ endangered or conserved taxa, morphometric approaches for classifying and identifying taxa
and populations of importance would greatly improve the ability to maintain these high risk
groups. Additionally, this could lead to better classification of extinct taxa.

³⁸ Much work has been devoted to species delimitation via sequence difference (Fujita
et al. 2012; Yang and Rannala 2010) while comparatively little has been devoted to introducing
⁴⁰ new methodology for case of purely morphological data (Mitteroecker and Bookstein 2011;
Zelditch et al. 2004). The majority of this effort has focused on identifying differences between
⁴² already identified taxa (Demandt and Bergek 2009; Gaubert et al. 2005; Gündüz et al. 2007;
Polly 2003, 2007; Zelditch et al. 2004) and automated taxon identification (MacLeod 2007).

⁴⁴ Here, we address the question of how can alternative approaches and methodology
improve morphology based classification. From this approach, we ask if it is possible to
⁴⁶ determine which amongst a set of classification hypotheses is best in order to determine if a
case of cryptic diversity is truly cryptic or just a case of extremely fine scaled morphological
⁴⁸ variation.

Background and system

⁵⁰ Machine learning is in many respects just an extension of known statistical methodology
(Hastie et al. 2009) where the emphasis is placed on inferring rules and properties of data in
⁵² order to explain the underlying structure. The basic statistical mechanics are supplemented
by randomization, sorting, and partitioning algorithms and along with the maximization or
⁵⁴ 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
⁵⁶ use in medical research, epidemiology, economics and automated image identification such as
handwritten zip codes (Hastie et al. 2009). Two major classes of machine learning methods are
⁵⁸ unsupervised and supervised learning. Unsupervised learning methods are used with unlabeled

data where the underlying structure is estimated and are analogous to clustering and density
60 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
62 to output are inferred. These are analogous classification and regression models (Breiman
et al. 1984). The application of the alternative approaches used in this study illustrates only
64 a sampling of the various previously derived methods for clustering observations and fitting
classification models.

66 Differences in morphological variation between different classes has previously been
analyzed using methods like linear discriminate analysis and canonical variates analysis
68 (Demandt and Bergek 2009; Gaubert et al. 2005; Gündüz et al. 2007; Mitteroecker and
Bookstein 2011; Polly 2003, 2007; Zelditch et al. 2004). These methods are comparatively
70 straight forward ways of understanding the differences in morphology between classes. Also,
they are very visual methods which aides with the interpretation and presentation of in-
72 formation. Previous studies, however, normally do not compare which amongst a set of
candidate classification hypotheses is better. For example, studies such as those of Caumul
74 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
76 morphology to multiple classification schemes. Of note, however, is the work of Cardini et al.
(2009), which compared morphological variation in marmots at both population, regional, and
78 species levels to determine fidelity between shape each of these different hierarchical levels.
Importantly, however, is that the classification models have not been generalized to testing
80 data and training data accuracy is used almost exclusively as the metric off classification
strength.

82 Here, we used multiple machine learning methods, both unsupervised and supervised,
in order to compare different classification hypotheses. These methods provide different and
84 unique advantages for understanding how to classify taxa, with what accuracy, and what these

classifications are based on. While machine learning methods such as neural networks have
86 been applied to studying shape variation (MacLeod 2007), they have been primarily applied
in the context of automated taxon identification and not in terms of group classification and
88 strength of classification. Additionally, we investigate variation in continuous traits and not
discrete differences between each class, instead focusing on differences in the multivariate
90 quantification of shape. Also, instead of pure classification accuracy, here we used a statistic of
classification strength that reflects the rate at which taxa are both accurately and inaccurately
92 classified (see Methods).

In this study, we investiage the subspecific classification of the western pond turtle,
94 *Emys marmorata*. *E. marmorata* is distributed from northern Washington State, USA
to Baja California, Mexico. Traditionally, *E. marmorata* was classified into three groups:
96 the northern *E. marmorata marmorata*, the southern *E. marmorata palida*, and a central
Californian intergrade zone (Seeliger 1945). *E. marmorata marmorata* is differentiated from
98 *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
100 *E. marmorata palida* though they are considerably smaller.

Previous work on morphological variation in *E. marmorata* has focused, primarily, on
102 differentiation between different populations within a subset of the total species range (Bury
et al. 2010; Germano and Bury 2009; Germano and Rathbun 2008; Lubcke and Wilson 2007)
104 with comparatively little done over the entire species range (Holland 1992). These studies
have focused on how local biotic and abiotic factors may contribute to differences in carapace
106 length (Germano and Bury 2009; Germano and Rathbun 2008; Lubcke and Wilson 2007)
and found that size can vary greatly between different populations.

108 Additionally, there has been found a great deal of evidence for sized-based sexual
dimorphism in *E. marmorata* (Germano and Bury 2009; Holland 1992; Lubcke and Wilson
110 2007) 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
112 between populations (Holland 1992), which makes sense in light of the amount of between
population size difference (Germano and Bury 2009; Lubcke and Wilson 2007). However, the
114 effect of sexual dimorphism on shape, *sensu* Kendall (1977), was not assessed (Germano and
Rathbun 2008; Holland 1992; Lubcke and Wilson 2007).

116 Of particular note is the work of Holland (1992) which compared morphological
differences between and among many populations of *E. marmorata* across the species range.
118 Holland (1992) studied the relative effect of distance versus barriers had in terms of fostering
morphological differentiation in *E. marmorata*. Analyses were performed to determine how
120 different, morphologically, different populations in three different regions of the species range.
Measurements were made from all different aspects of carapace morphology and not just
122 total carapace length.

Holland (1992) concluded that distance was a poor indicator of morphological differen-
124 tiation as opposed to barriers, such as different drainage basins, are probably more important
barriers to reproduction. This conclusion was later echoed by Spinks and Shaffer (2005) via
126 molecular phylogenetic analysis. Additionally, Holland (1992) found that with increasing
amount of barriers and distance, morphological differentiation was observable though the
128 underlying variation required many variables obtain indicating the very fine degree of mor-
phological differentiation between putatively distinct populations. Holland (1992) concluded
130 that *E. marmorata* is best classified as three distinct species as opposed to subspecies: a
northern species, southern species, and Columbia basin species. This classification is similar
132 to Seeliger (1945), except elevated to the species as opposed to subspecific level.

More recently, *E. marmorata* was divided into four clades based on mitochondrial
134 DNA: a northern clade, a southern clade, and eastern and western central Californian clades
(Spinks and Shaffer 2005; Spinks et al. 2010). While nuclear DNA supports two major clades,
136 one northern and one southern, Spinks et al. (2010) argue that the four clade classification is

of greater conservation utility even though the variation between these groups is considered
138 cryptic. While the mitochondrially based classification is considered robust, there is no known
morphological differentiation between these clades.

140 In this study, we attempt to estimate the best classification scheme of *E. marmorata*
based on variation in plastral shape in order to determine if the molecular based hypothesis of
142 Spinks and Shaffer (2005) and Spinks et al. (2010) is actually a case of cryptic diversity or not.
Because of unclear geographic boundaries between subgroups of *E. marmorata*, we compare
144 two hypotheses of morphologically based classification and two hypotheses of molecularly
based classification. We hypothesize that if morphological variation corresponds to class
146 assignment, then it should be possible to determine the best classification hypothesis of *E.*
marmorata from amongst multiple candidate hypotheses. However, if morphological variation
148 variation does not correspond to any classification hypothesis, then supervised learning model
generalization performance will be poor and reflect how variation may not follow along with
150 any of the candidate classification hypotheses.

MATERIALS AND METHODS

152 *Specimens*

We collected landmark-based morphometric data from 524 adult *E. marmorata* museum
154 specimens. These specimens include both newly sampled individuals and those sampled
in previous studies of plastral shape variation (Angielczyk and Feldman 2013; Angielczyk
156 et al. 2011; Angielczyk and Sheets 2007). Specimen classification was based on known
specimen geographic information which was recorded from museum collection information.
158 When precise latitude and longitude information was not available it was estimated from
whatever locality information was present. Because the specimens used to define the subclades

¹⁶⁰ in Spinks and Shaffer (2005) and Spinks et al. (2010) were not available for study, all specimen classifications were based solely on this geographic information and not from ¹⁶² explicit assignment in previous studies. Instead, classification was based on matching museum locality data with the geographic boundaries of the molecularly-defined clades of Spinks ¹⁶⁴ and Shaffer (2005) and Spinks et al. (2010). Because the exact barriers between different biogeographic regions are unknown and unclear, two assignments for both the morphologically ¹⁶⁶ and molecularly based hypotheses were used. Each morphologically based hypothesis had three classes, while each molecular-based had four classes. In total, each specimen was given ¹⁶⁸ four different classifications.

Geometric morphometrics

¹⁷⁰ Following previous work on plastral variation (Angielczyk and Feldman 2013; Angielczyk et al. 2011; Angielczyk and Sheets 2007), 19 landmarks were digitized using TpsDig 2.04 ¹⁷² (Rohlf 2005). These landmarks were chosen to maximize the description of general plastral variation(Fig. 1). 17 of these landmarks are at the endpoints or intersection of the keratinous ¹⁷⁴ plastral scutes that cover the platron. 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 ¹⁷⁶ (Klingenberg et al. 2002), 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 ¹⁷⁸ where damage or incompleteness prevented symmetric landmarks from being determined, only the single member of the pair was used. Analysis was conducted on the resulting “half” ¹⁸⁰ plastra. Plastral landmark configurations were superimposed using generalized Procrustes analysis (Dryden and Mardia 1998) after which, the principal components (PC) of shape ¹⁸² 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 between all landmark configurations in shape space, the dissimilarity between observations was measured using Kendall's Riemannian 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 used in the study.

The ρ dissimilarity matrix was divisively clustered using partitioning around mediods 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 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 clusters varied 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

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

. $d_{ii'}$ is the dispersion of the clustering solution or the sum of the pairwise dissimilarities between observations in each cluster and their respective mediods (C) for all clusters r . This value is averaged and compared to the expected dispersion (E_n^*) of a sample n from a reference distribution. In this case, the reference distribution was estimated from 500 resamples of the

206 dataset while maintaining the original dispersion of the data. This analysis was conducted using the `cluster` package for R (Maechler et al. 2013) using all 524 observations.

208 *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
210 by class, and was used for model fitting. The testing dataset represented the remaining 25% of the total dataset and was used after model fitting to estimate the effectiveness of
212 each classification hypothesis and generalizability of the supervised learning models (i.e. performance in the wild). This split was chosen to allow for a large enough sample size for
214 model fitting while also providing a large enough testing dataset to determine any systematic misclassifications.

216 Three different supervised learning methods were used to model the relationship between plastral shape and class: linear discriminant analysis, multinomial logistic regression
218 and random forest. These methods were chosen because of various properties of these methods which allow for useful interpretations about the quality and structure of the classification.

220 Linear discriminant analysis (LDA) is a frequently applied method for characterizing the primary differences in morphology between different classes (Mitteroecker and Bookstein
222 2011; Zelditch et al. 2004). This method attempts to find a linear combination of predictors to best model two or more classes. LDA is very similar to PCA except that instead of finding
224 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
226 transformation matrix by which the original features can be transformed to reflect the best discrimination between the classes. Like other supervised learning methods, LDA can also be
228 used for predictive analysis on testing data. LDA was done using the `MASS` package for R (Venables and Ripley 2002).

230 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).

232 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
234 of all the constituent model results. Similar to the odds ratios calculated from the coefficients
of a logistic regression, the relative risk of a classification with reference to a baseline class
236 can be determined from the coefficients of the model. Multinomial logistic regression models
were fit using the `nnet` package for R (Venables and Ripley 2002)

238 Random forest models are an extension of classification and regression trees (CART)
(Breiman 2001; Breiman et al. 1984). Because this study relies on classification models,
240 CARTs are explained with reference to classification but the approach is equally valid for
regression. The goal of CARTs are to use a series of different features to estimate the final
242 class. In top-down induction of decision trees for each member of a given set of predictor
variables, attribute value test are used to estimate the differences between classes. This
244 process is then repeated on each subset, called recursive partitioning. The recursion continues
until the resulting observations all share the same class or no more meaningful partitions
246 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
248 fitting.

In a random forest model, many CARTs are built from a random subsample of both the
250 features and the observations. This process is then repeated many times and the parameters
of the final model was chosen as the mode of estimates from the distribution of CARTs
252 (Breiman 2001). 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
254 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
256 amount of variation in the entire dataset. This is a generally important property that is

useful for many other studies which want to describe and model the differences between
258 classes and the relative importance different features. Random forest models were fit using
the `randomForest` package for R (Liaw and Wiener 2002).

260 The supervised learning models used here, except LDA, have tuning parameters which
help to increase the generalizability of the model and prevent them from being overfit. For the
262 supervised learning models fit in this study, tuning parameters were estimated via 10 rounds
of 10-fold cross-validation (CV) across a grid search of all tuning parameter combinations.
264 Optimal tuning parameter values were selected based on area under the receiver operating
characteristic (ROC) curve. The area under the multiclass ROC curves was estimated using
266 the all-against one strategy derived by Hand and Till (2001). This tuning process was
implemented following the default grid search implemented in the `caret` package for R (Kuhn
268 2013).

ROC is a confusion matrix (Table 1) statistic that is a descriptor of the relationship
270 between the false positive rate (FPR , Eq. 1) of a classification model and the true positive
rate (TPR , Eq. 2) of a classification model (Hastie et al. 2009). The area under the ROC
272 curve (AUC) is a summary statistic of the quality of the classification and varies between
0.5 and 1, with values of 0.5 indicating a model that classifies no better than random and a
274 value of 1 indicating perfect classification (Hastie et al. 2009). AUC can be used as a model
selection criterion for classification models and is especially useful in cases where some if not
276 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 (Hastie et al. 2009). It is important to note that,
278 unlike AICc, AUC is not calculated with reference to the complexity of the model.

$$FPR = \frac{FP}{FP + TN} \quad (1)$$

$$TPR = \frac{TP}{TP + FN} \quad (2)$$

280 LDA was applied on the eigenscores from a subset of the total number of PCs, ranging
from two to 10 in increasing order of complexity. In total, this produced nine different LDA
282 scaling matrices. From this set, the best number of PCs used to estimate the LDA scaling
matrix were chosen. As LDA is not “fit” via maximum likelihood, the final combination of
284 number of PCs and LDA scaling matrix chosen was that with the greatest AUC value from
the training set.

286 For the multinomial logistic regression models, 10 different models were fit each having
sequentially more PCs as predictors in order to have models representing different levels of
288 overall amount of shape variation and to estimate how much was necessary and sufficient
to best estimate class. The maximum number of PCs allowed as predictors was 10 because
290 of both the large 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
292 AICc (Akaike 1974; Burnham and Anderson 2002; Hurvich and Tsai 1989). AICc is a model
selection criterion where the model with lowest AICc has the fairest variance–bias tradeoff
294 (Burnham and Anderson 2002). Model selection was performed in this manner because the
optimal number of PCs to use as predictors was not known *a priori*, and while including all
296 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
298 plastral variation. In addition to the AICc of each model the Δ AICc and Akaike weights are
also reported. Δ AICc values are the difference in AICc between the AICc best model and that
300 model while Akaike weights are a transformation of the AICc of a model with relation to all
other models being compared and measures the relative amount of information explained by
302 that model compared to all other models (Burnham and Anderson 2002).

304 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. Following the

306 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
308 model calculated. The lowest ranked feature was then removed, and the AUC of the model
recalculated. This was repeated until only one feature, remained. Similar to the multinomial
310 logistic regression models described above, the maximum number of PCs that could have
been included in the model was 10. After each PC was removed , 10-fold CV was used to
312 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. Because
314 PCs were kept in order of importance and not in relation to the amount of variance each PC
described, this means that the exact PCs included in each model do not correspond to the
316 PCs in each of the 10 multinomial logistic regressions models.

The final selected models were then used to estimate the class assignments of the
318 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
320 for each classification scheme via 1000 nonparametric bootstrap resamples of the training
dataset. The difference in distributions was assessed using pairwise Mann-Whitney U tests.

322

RESULTS

Geometric morphometrics

324 The results of the PCA of the total dataset of *E. marmorata* pastral landmarks configurations
demonstrates no clear or obvious groupings (Fig. 2). The first three PCs, which represent
326 45.29% of the total variation, are a cloud of points with no structure. Additionally, individual
landmark variation is mostly circular around each landmark with some more elliptical
328 variation observed along some midline landmarks and the most lateral landmark (Fig. 2).

However, it is important to note that Procrustes based superimposition attempts to evenly
330 distribute variance around the mean shape (Zelditch et al. 2004) and this observation should
be considered cursory at best.

332 The first two PCs appear to describe principally variation in the lateral margin of the
palstra, from a pointed medial edge to a more rounded and blunt edge (Fig. 3). Landmark 10
334 (Fig. 1), which appears to be the most variable along these axes (Fig. 2 and 3), is positioned
on the bridge between the plastron and the carapace. Over ontogeny, this is an area that
336 deepens dorsoventrally and when the plastron was projected into two dimensions it created
the effect of mediolateral movement. Lateral landmark variation along the first PC seemed
338 concentrated in the posterior portion of the plastra with additional variance observed in
midline landmarks (Fig. 3). This variance in midline landmarks was most likely caused by
340 the fact that plastral scutes frequently do not line up perfectly. Along PC 2, lateral variation
appeared to be concentrated in the anterior portion o the plastra (Fig. 3).

342 When the eigenscores of the first two PCs are compared to the centroid sizes of the
observations, a significant correlation is observed between centroid size and the first PC
344 (Spearman's ρ : -0.56, p -value: 7.51×10^{-44}) but not between centroid size and the second
PC (Spearman's ρ : 0.03, p -value: 0.55). This difference is clearly observable in in plots of
346 the data (Fig. 4).

Machine learning analyses

348 *Unsupervised learning.—*

Comparison of gap statistic values for the range of PAM solutions indicates that the
350 optimal number of clusters is one (Fig. 5). The next best clustering solution had only two
clusters, however there is no geographic structure to this classification scheme, with members
352 of these clusters being seemingly randomly distributed (Fig. 6). Importantly, these clusters do

not conform to the northern and southern groups from the nuclear DNA hypothesis (Spinks
354 et al. 2010).

Sex information was only available for 399 of the 524 turtles. A χ^2 test of the relation-
356 ship between sex observation and cluster assignment from PAM with two clusters showed that
there was no significant relationship between cluster assignment and sex observation (χ^2 : 1.12,
358 df: 1, *p*-value: 0.29, Table 2). This results is interesting because while sexual dimorphism has
been observed in linear measures and mass estimates of *E. marmorata* (Germano and Bury
360 2009; Holland 1992; Lubcke and Wilson 2007), this results demonstrates that this dimorphism
may not translate into differences in shape. Interestingly, male emydid turtles are known to
362 have a plastral concavity which may influence landmark position along the midline. However,
the plastral concavity of *E. marmorata* males is considered less pronounced than in other
364 emydid turtles.

The gap statistic values for both three and four clusters are much lower than for one
366 and two and are statistically identical. Interestingly, other solutions with a much greater
number of clusters have relatively high gap statistic values as well though these are also
368 not significantly different. Increasing the number of clusters does appear to improve the gap
statistic enough compared to the best clustering solution to merit detailed discussion.

370 *Supervised learning*.—

The optimal number of PCs used for LDA, as determined by highest ROC score, for
372 three of the four classification schemes had all 10 possible PCs (Fig. 7). These were both of
the morphological based classification hypotheses and the second molecular hypothesis. LDA
374 of the PCs of the first molecular hypothesis found that, based on ROC, only the first 9 PCs
were necessary to best discriminate between the classes (Fig. 7). The first 9 PCs describe
376 83.23% of total variation in plastral shape, while the first 10 PCs describe 86.54% of the
variation.

378 The AICc best multinomial logistic regression model for three of the four classification
schemes had the first 9 PCs as features (Tables 3, 4, and 5). The second molecularly based
380 classification hypothesis included all 10 possible PCs as predictors (Tables 6). The ΔAICc
values between the optimal and second best model range from 1.18 for the first morphological
382 based classification hypothesis to 26.51 for the second molecular based classification hypothesis
(Tables 3, 4, 5, and 6).

384 While the ΔAICc value between the optimal and second best model for the first
morphological and first molecular based classification hypothesis was within the range to
386 be considered 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
388 with different responses (Burnham and Anderson 2002), we interpret the fact that the
 ΔAICc best model in these cases is the simpler model and that the optimal model for
390 three of the classification schemes having the same number of predictors as reasons to use
only the AICc best model for all cases. Additionally, by using a single model for each of
392 the classification hypotheses, this limits the number of comparisons between the bootstrap
resampled distributions of the AUC values for the testing dataset (see below).

394 The selected number of features in the final random forest model for each classification
scheme was very simpler to the model selection results for the LDA-based classification and
396 the multinomial logistic regression models, ranging from 9 for the second morphological based
classification hypothesis and both molecular based classification hypotheses to 10 for the first
398 morphological based classification hypothesis (Fig. 7).

400 In the case of all models, there is a substantial increase in model performance as
measured by AICc for the multinomial logistic models (Tables 3, 4, 5, and 6) or in AUC for
the LDA-based predictions and random forest models and illustrated for the multinomial
402 logistic regression models as the number of features increases (Fig. 7).

The results from the generalization of the selected supervised learning models, mea-

404 sured by the distributions of the bootstrapped AUC values of the testing dataset, show
that a molecular classification hypotheses was the best overall classification scheme (Fig.
406 8). Remarkably, the best classification hypothesis was the second molecular classification
hypothesis based on the LDA-based predictions, the multinomial logistic regression and
408 random forest models. For both methods, the distribution of bootstrapped AUC for the
molecular hypothesis was significantly greater than all of the other classification schemes
410 (Tables 7, 8 and 9).

When the classification results of the training set for the best classification scheme
412 based on the generalization results are compared with the references classes, the higher
AUC value of the best results from LDA and the best multinomial logistic regression model
414 compared to the best random forest model can be observed as the classifications are much
closer to the reference classes (Fig. 9). The best random forest model misclassified many of the
416 observations as the northern clade instead of the correct class. This pattern of misclassification
is observable but not as exaggerated in the LDA-based classifications and those from the
418 multinomial logistic regression model (Fig. 9).

This pattern of misclassification may have been caused by the subtle differences in
420 mean shape between each of the different classes (Fig. 10). The mean shape of the northern
clade is the most similar to the mean shape of the entire dataset (Fig. 10a), which may
422 indicate that specimens that are closer to the mean shape will be systematically misclassified
as the northern clade.

424 The results of fitting the final random forest model also include the variable importance
for best separating the different classes. The selected random forest model for the best
426 classification scheme had 9 PCs as features. The PCs included as features in the final random
forest model, in descending order of importance, were PCs 3, 2, 1, 6, 5, 10, 9, 8 and 4. Of these
428 9 features, the first three are illustrated here (Fig. 11) in descending order of importance.

The first two most important features describe different aspects of variation (Fig. 12).

430 The third and most important PC describes variation roundedness of the medial portion of the
plaстрон, both the anterior and posterior portions of the plaстрон. Additionally, the relative
432 position of the landmarks along the midline varies greatly along PC3 (Fig. 12). This PC
represents 12.19% of total variation. The second and second most important PC is described
434 above and principally described variation in landmarks along the lateral and anterior margin
of the plaстрон. This PC represents 12.78% of total variation. The major variations along
436 these axes correspond well to the differences between the mean shape of each class (Fig.
10) where major class differences seem based on the relative ballooning or shrinking of the
438 anterior and posterior portions of the plaстрон together along with differential “pinching”
of the midline landmarks. It is important to note that the first two most important axes,
440 according to the random forest model, were not correlated with size. This can be interpreted
to mean that class membership is principally not related to turtle size. This interpretation
442 makes sense in light of previous work showing that body size in *E. marmorata* can vary
widely across the entire species range (Germano and Bury 2009; Holland 1992; Lubcke and
444 Wilson 2007).

The relative risk values for classification from the multinomial logistic regression
446 model, based on the three most important PCs, demonstrate that individual axes contribute
to classification differently and that given multiple features the odds of determining the
448 correct classification increase (Fig. 13). The first most important axis contributes strongly
to classifying both the western and southern groups while changes along the second most
450 important axes contribute very little to increasing the odds of classification for all but the
eastern group. This is observable from the class histograms of PC 3 and 2 (Fig. 12). Changes
452 along the first and third most important axes contribute more obviously to increasing the
odds of correctly identifying the class of an observation, a result that is observable in both
454 the relative risk (Fig. 13) and the different class histograms of the PCs (Fig. 12).

The graphical results from the LDA of the training dataset for models of the second

⁴⁵⁶ molecular classification scheme agree with the subtle distinctions between the different classes
⁴⁵⁷ (Fig. 14). There is no clear distinction in terms of multivariate space between the four different
⁴⁵⁸ classes. Instead, across all three axes there is substantial overlap as indicated by both the
scatter of the points in space and the distribution of observations along each axis.

⁴⁶⁰ DISCUSSION

The results of this study support the mitochondrial based classification hypothesis of *E.*
⁴⁶¹ *marmorata* (Spinks and Shaffer 2005; Spinks et al. 2010). This is contrary to the original
classification of *E. marmorata* (Holland 1992; Seeliger 1945) and lends credence to the idea
⁴⁶⁴ that at least some aspect of cryptic diversity is a product of sample size, methodology, or
both.

⁴⁶⁶ The lack of coherent geographical subclass assignment from PAM clustering (Fig. 5)
as well as the large number of features necessary before no increase in AUC for all models
⁴⁶⁸ (Fig. 7) indicates that the morphological variation between classes is extremely fine grained.
This was also exemplified by the small differences between mean class shapes of the final
⁴⁷⁰ chosen classification scheme (Fig. 12).

The approaches presented here for supervised learning analysis of the landmark varia-
⁴⁷² tion 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
⁴⁷⁴ simultaneously, and not along individual PCs, including all shape variation in each model
might increase model complexity beyond a reasonable level for the sample size and possibly the
⁴⁷⁶ necessary complexity to accurately model the response. Additionally, because only individual
PCs are used as features in the models, this does not accurately represent shape evolution
⁴⁷⁸ and how exactly different classes might be evolving in relation to each other. However, this
compromise is not without its advantages. Because both AICc and AUC values improved

480 rapidly with increased model complexity (Fig. 7), this helped demonstrate how fine scale
the actual variation between classes was. The variable importance information from the
482 random forest models was extremely useful for understanding what aspects shape variance
contributed most to differentiating the classes and in what order as opposed purely in the
484 order of largest variance (Fig. 11 and 12). Importantly, the order of variable importance
demonstrated that the two most important axes (PC 2 and PC 3) for determining class
486 membership were not correlated with size, unlike the third most important axis (PC 1),
indicating that class membership is not strongly related to size as opposed to being related to
488 actual shape differences. Additionally, the relative risk values from the multinomial logistic
regression models demonstrate that a single PC is probably not sufficient for estimating the
490 class of an observation, but that given a set of PCs this classification would be more accurate
(Fig. 13).

492 Ultimately, it would be useful to not require such explicit classification hypotheses,
especially when concerned about possible cryptic variation in extinct taxa. The only unsuper-
494 vised method employed in this study, PAM, is rather simple and not model based. A more
useful approach would be to employ various model based clustering approaches (Fraley and
496 Raftery 2002; Van Bocxlaer and Hunt 2013; Zhong and Ghosh 2003). In this manner, a series
of candidate models can be compared via model comparison methods, such as AIC or Bayes
498 factors (Fraley and Raftery 2002), in order to asses the best clustering solution. Here we
focused on the results and utility of supervised methods because they are both more powerful
500 and hypothesis driven (Hastie et al. 2009). Because there are two alternative classification
schemes for *E. marmorata*, it was most appropriate to compare these two hypotheses and
502 estimate which one most accurately reflected the variation. Future work would be to explore
and derive unsupervised methods which corroborate these results.

504 In this study we have demonstrated that, using alternative methodology to that
which is most frequently applied, it is possible to determine which classification scheme best

506 matches 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
508 Spinks and Shaffer (2005) and Spinks et al. (2010) and not with the original morphology based
hypothesis of Holland (1992); Seeliger (1945). We have also demonstrated the utility of various
510 machine learning approaches to understanding the structure of variation in morphometric
data. Specifically, methods for better understanding misclassification and identifying which is
512 the most important for delimiting different classes. These methods represent new applications
which may be important for future studies on class-based morphological comparison and
514 variation, both in the context of cryptic diversity and with known classifications.

ACKNOWLEDGEMENTS

516 PDS would like to thank David Bapst, Michael Foote, Benjamin Frable, and Dallas Krentzel
for useful discussion which enhanced the quality of this study. For access to emydine specimens,
518 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);
520 P. Collins (SBMNH); B. Hollingsworth (SDMNH); C. Bell and R. Burroughs (TMM); T.
LaDuc and R. Burroughs (TNHC); P. Holroyd (UCMP); R. Symonds (UMZC); J. Buskirk.
522 We are greatful to S. Sweet for field assistance and the California Department of Fish and
Game for permits. Much of the data collection was funded by NSF DBI-0306158 (to KDA).

*

524

References

- 526 Akaike, H. 1974. A new look at the statistical model identification. IEEE Transactions on
Automatic Control 19:716–723.

- ⁵²⁸Angielczyk, 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.
- ⁵³⁰Angielczyk, K. D., C. R. Feldman, and G. R. Miller. 2011. Adaptive evolution of plastron shape in emydine turtles. *Evolution* 65:377–394.
- ⁵³²Angielczyk, K. D. and H. D. Sheets. 2007. Investigation of simulated tectonic deformation in fossils using geometric morphometrics. *Paleobiology* 33:125–148.
- ⁵³⁴Breiman, 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.
- Burnham, K. P. and D. R. Anderson. 2002. Model selection and multi-model inference: a practical information-theoretic approach. 2nd ed. Springer, New York.
- ⁵³⁸Bury, 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.
- ⁵⁴²Cardini, 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.
- 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.
- ⁵⁴⁸Clare, E. L. 2011. Cryptic species? Patterns of maternal and paternal gene flow in eight neotropical bats. *PloS one* 6:e21460.

- ⁵⁵⁰Demandt, M. H. and S. Bergek. 2009. Identification of cyprinid hybrids by using geometric morphometrics and microsatellites. *Journal of Applied Ichthyology* 25:695–701.
- ⁵⁵²Dryden, 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.
- ⁵⁵⁴Eldredge, N. and S. J. Gould. 1972. Punctuated equilibria: an alternative to phyletic gradualism. Pages 82–115 *in* *Models in Paleobiology* (T. J. M. Schopf, ed.). Freeman Cooper, San Francisco.
- ⁵⁵⁶Fraley, C. and A. E. Raftery. 2002. Model-Based Clustering, Discriminant Analysis, and Density Estimation. *Journal of the American Statistical Association* 97:611–631.
- Fujita, M. K., A. D. Leaché, F. T. Burbrink, J. a. McGuire, and C. Moritz. 2012. Coalescent-based species delimitation in an integrative taxonomy. *Trends in ecology & evolution* 27:480–8.
- ⁵⁶⁰Funk, 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 genets (Carnivora, Viverridae, Genetta spp.) from the southern African subregion. *Biological Journal of the Linnean Society* 86:11–33.
- ⁵⁶²Germano, 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.
- ⁵⁶⁸Germano, 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. Yeniyurt, 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.

Holm, S. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 6:65–70.

Hunt, G. 2008. Gradual or pulsed evolution: when should punctuational explanations be preferred? *Paleobiology* 34:360–377.

Hurvich, C. M. and C.-L. Tsai. 1989. Regression and time series model selection in small samples. *Biometrika* 76:297–307.

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.

- ⁵⁹⁴Kendall, D. G. 1984. Shape Manifolds, Procrustean Metrics, and Complex Projective Spaces. Bulletin of the London Mathematical Society 16:81–121.
- ⁵⁹⁶Klingenberg, C. P., M. Barluenga, and A. Meyer. 2002. Shape analysis of symmetric structures: quantifying variation among individuals and asymmetry. Evolution 56:1909–1920.
- ⁵⁹⁸Kuhn, 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.
- ⁶⁰⁰Lubcke, G. M. and D. S. Wilson. 2007. Variation in shell morphology of the Western Pond Turtle (*Actinemys marmorata* Baird and Girard) 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.
- ⁶⁰⁴Maechler, M., P. Rousseeuw, A. Struyf, M. Hubert, and K. Hornik. 2013. cluster: Cluster Analysis Basics and Extensions. R package version 1.14.4.
- ⁶⁰⁶Mitteroecker, P. and F. Bookstein. 2011. Linear Discrimination, Ordination, and the Visualization of Selection Gradients in Modern Morphometrics. Evolutionary Biology 38:100–114.
- Pfenninger, M. and K. Schwenk. 2007. Cryptic animal species are homogeneously distributed among taxa and biogeographical regions. BMC evolutionary biology 7:121.
- ⁶¹⁰Polly, P. D. 2003. Paleophylogeography of *Sorex araneus*: molar shape as a morphological marker for fossil shrews. Mammalia 68:233–243.
- ⁶¹²Polly, 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
616 for Statistical Computing Vienna, Austria.

Rohlf, F. J. 2005. TpsDig 2.04.

618 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
620 turtle (*Emys marmorata*): cryptic variation, isolation by distance, and their conservation
implications. Molecular ecology 14:2047–64.

622 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*
624 *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
626 by sympatric lineages of Southeast Asian forest frogs. Biology letters 2:470–4.

Tibshirani, R., G. Walther, and T. Hastie. 2001. Estimating the number of clusters in a
628 data set via the gap statistic. Journal of the Royal Statistical Society: Series B (Statistical
Methodology) 63:411–423.

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

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

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

⁶³⁶Zelditch, M. L., D. L. Swiderski, and H. D. Sheets. 2004. Geometric morphometrics for biologists: a primer. Elsevier Academic Press, Amsterdam.

⁶³⁸Zhong, S. and J. Ghosh. 2003. A unified framework for model-based clustering. *The Journal of Machine Learning Research* 4:1001–1037.

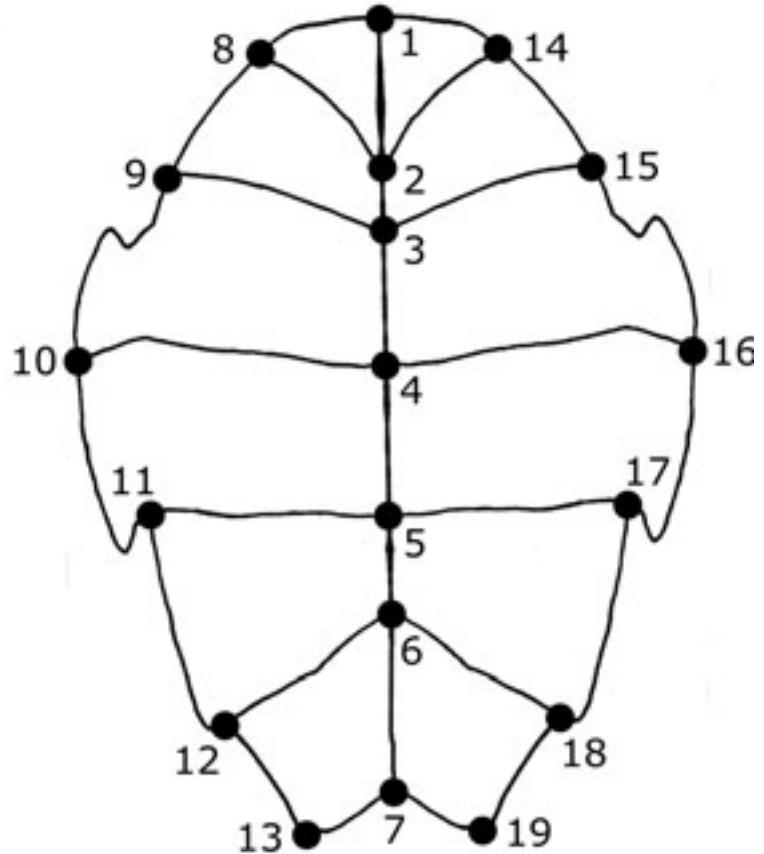


Figure 1: Depiction of general plastral shape of *E. marmorata* and position of the 19 landmark used in this study. Anterior is towards the top of the figure.

		Predicted class	
		1	0
Actual class	1	TRUE POSITIVE	FALSE NEGATIVE
	0	FALSE POSITIVE	TRUE NEGATIVE

Table 1: Example confusion matrix. The columns correspond to the predicted class of an observation, while the rows correspond to the actual class of that observation. Depending on the type match between the prediction and reality, four different outcomes are possible: true positive (TP), false negative (FN), false positive (FP), and true negative (TN). These four quantities are used for calculating all confusion matrix statistics. Each of these values is an integer and the sum of the number of occurrences of that event during classification.

	F	M	tot
1	101	112	213
2	99	87	186
tot	200	199	399

Table 2: Tabular comparison between sex observation and cluster assignement from PAM with two clusters. This number of clusters was chosen because it represented the second best clustering solution as determined via gap statistic comparison (Fig. 5). χ^2 analysis of this contingency table showed that there is no relationship between sex observation and cluster assignment (χ^2 : 1.12, df: 1, *p*-value: 0.29).

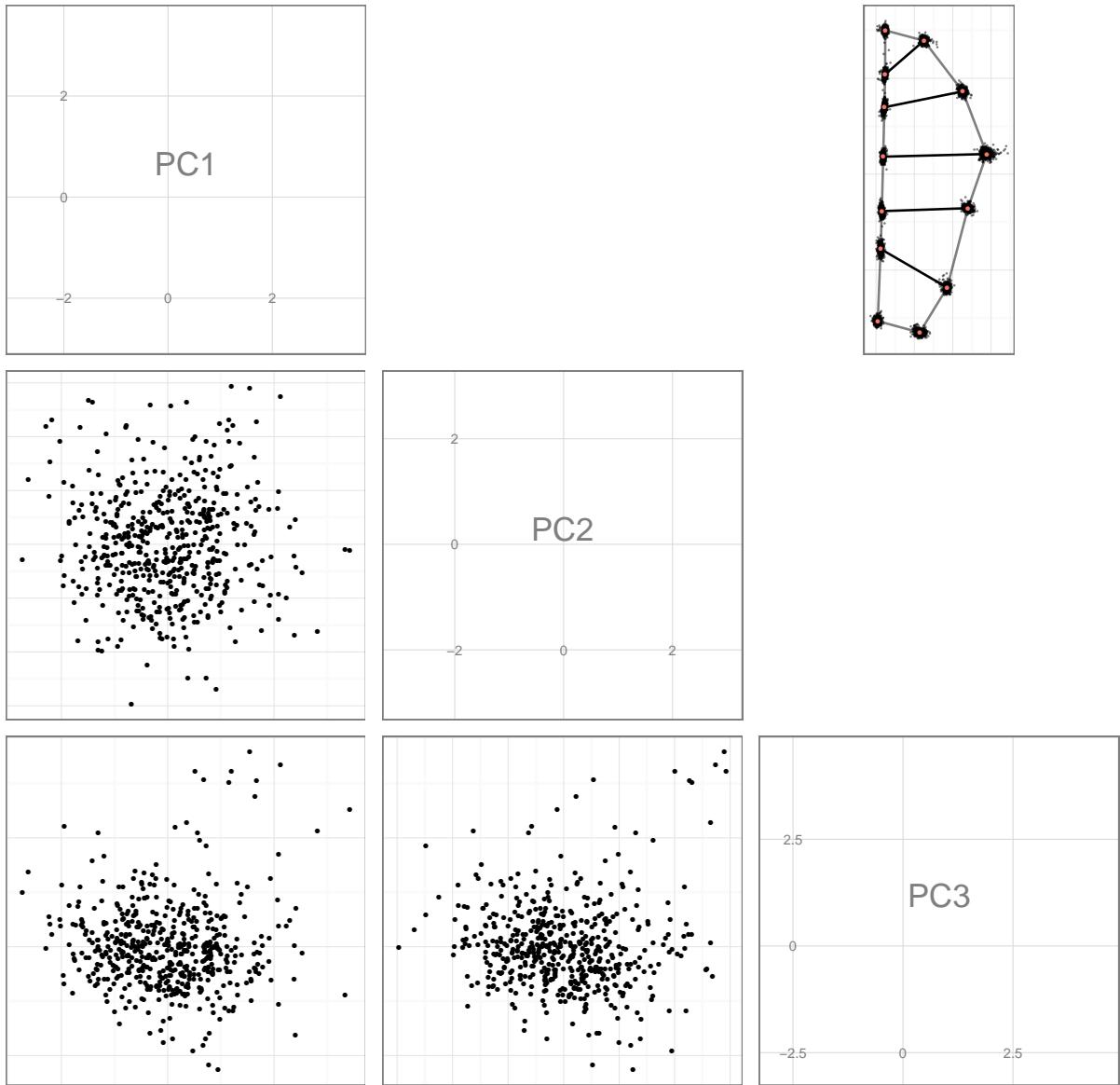


Figure 2: Results from PCA of the Procrustes superimposed “half” plastral landmarks. Depicted here are the for three PCs (lower triangle) and the mean shape with observed variance around each point (upper right). The first three PCs account for total 45.2924805932624% of the variance in plastral shape.

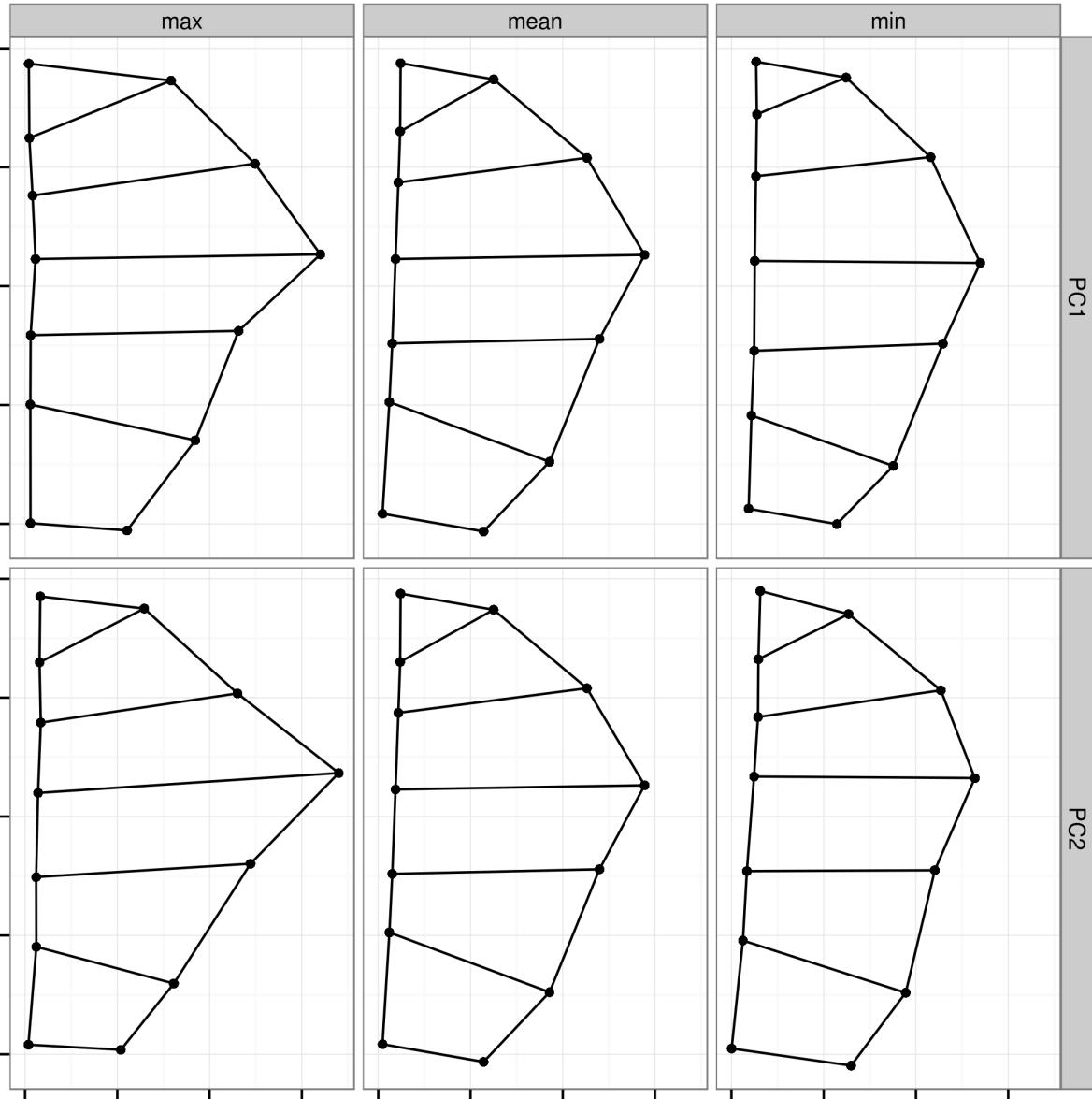


Figure 3: Landmark variation along the first two PCs of the Procrustes superimposed “half” plastral landmarks. The first row corresponds to variation along the first PC, while the second row corresponds to the second PC. The left most column represents the observation with the highest eigenscore along that PC, while the right most column represents the observation with the lowest eigenscore. The middle column, for both rows, is the mean plastral shape for all observations. The first PC represents 20.32% of the total variation in plastral shape while PC represents 12.78% of the variance.

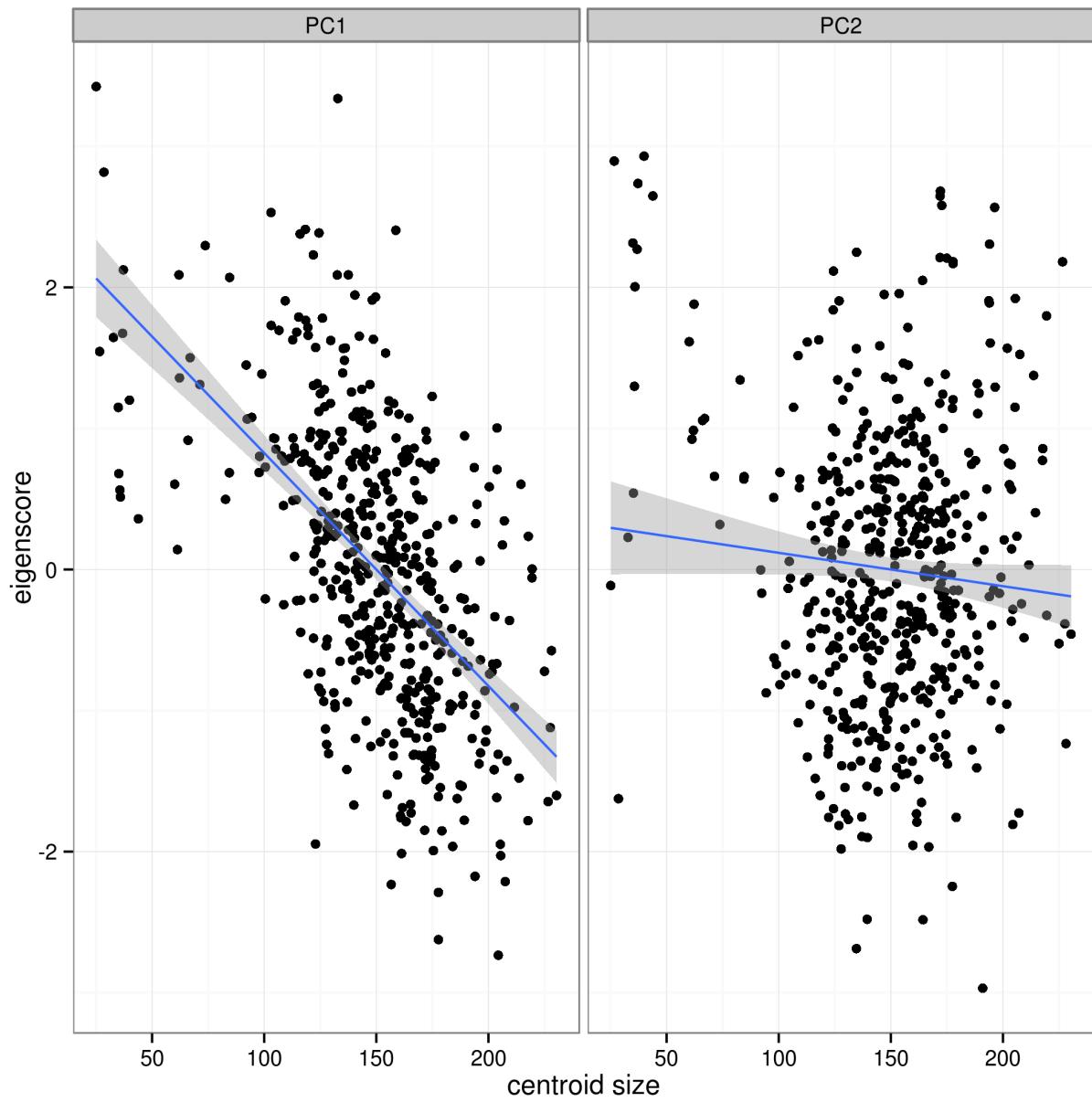


Figure 4: Comparison of the centroid size of each observation to the eigenscore of that observation from the PCA of plastral shape. The first two PCs are illustrated from right to left. The regression line is fit via ordinary least squares and is presented to illustrate the difference in correlation between the first two axes. The shaded regions represent 95% confidence intervals on the parameters of the regression line. Centroid size is significantly correlated with the first PC but not the second.

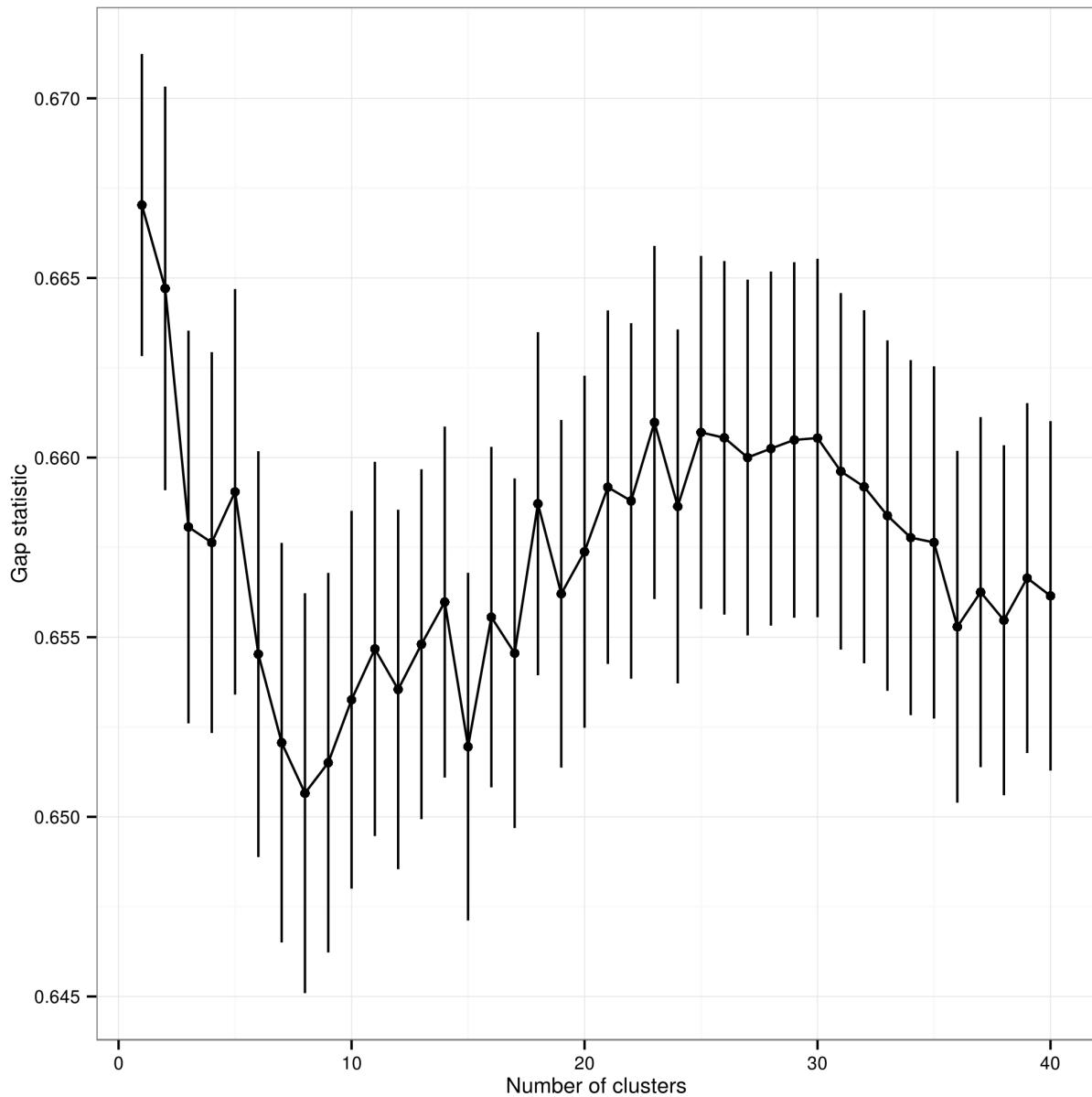


Figure 5: Gap statistic values for PAM clustering results for the ρ dissimilarity matrix of plastron shape. Error bars are standard errors estimated via 500 bootstrap samples.

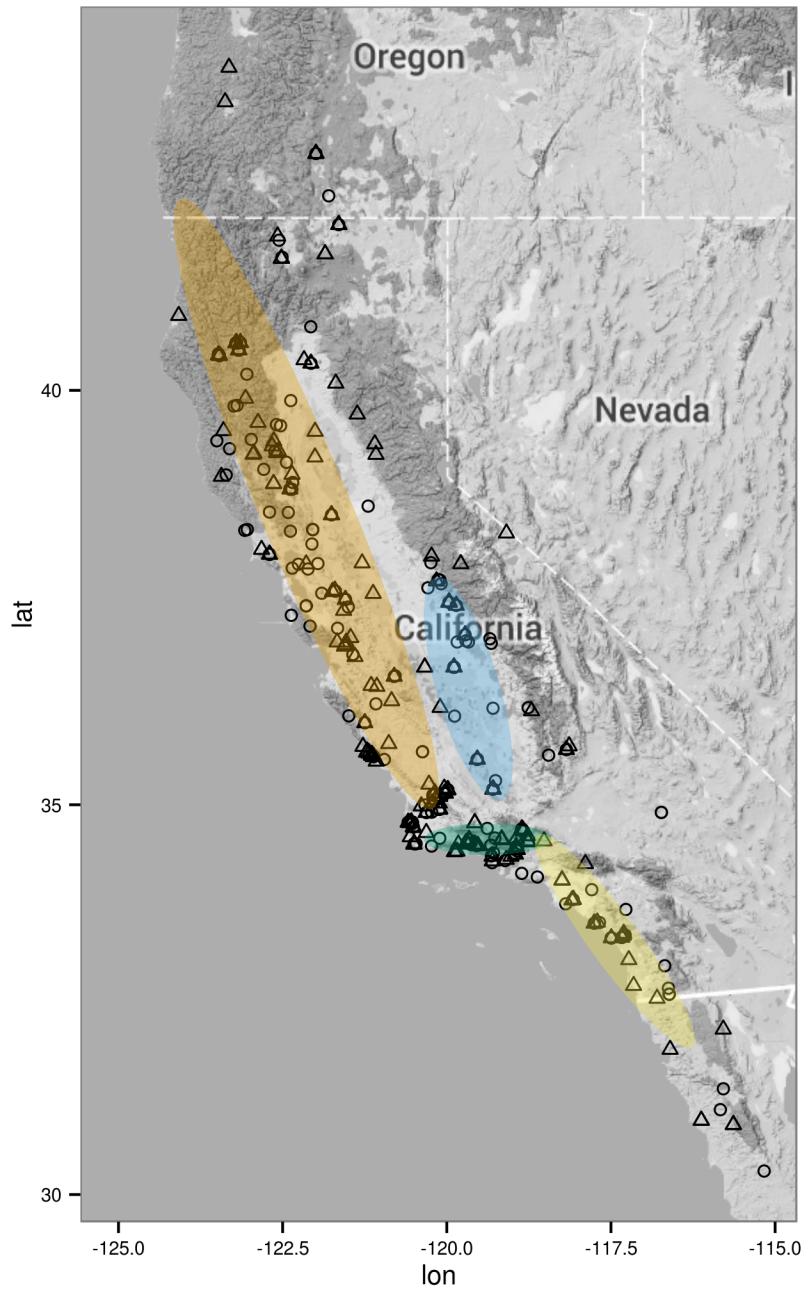


Figure 6: Clustering solution for PAM with two medioids for the entire set of observed *E. marmorata*. Clustering was based entirely on the ρ dissimilarity matrix of “half” plastral landmark configurations following Procrustes superimposition. Point shapes correspond to the two clusters while the colored ellipses correspond to 95% confidence ellipses of the four groups from Spinks and Shaffer (2005) and Spinks et al. (2010).

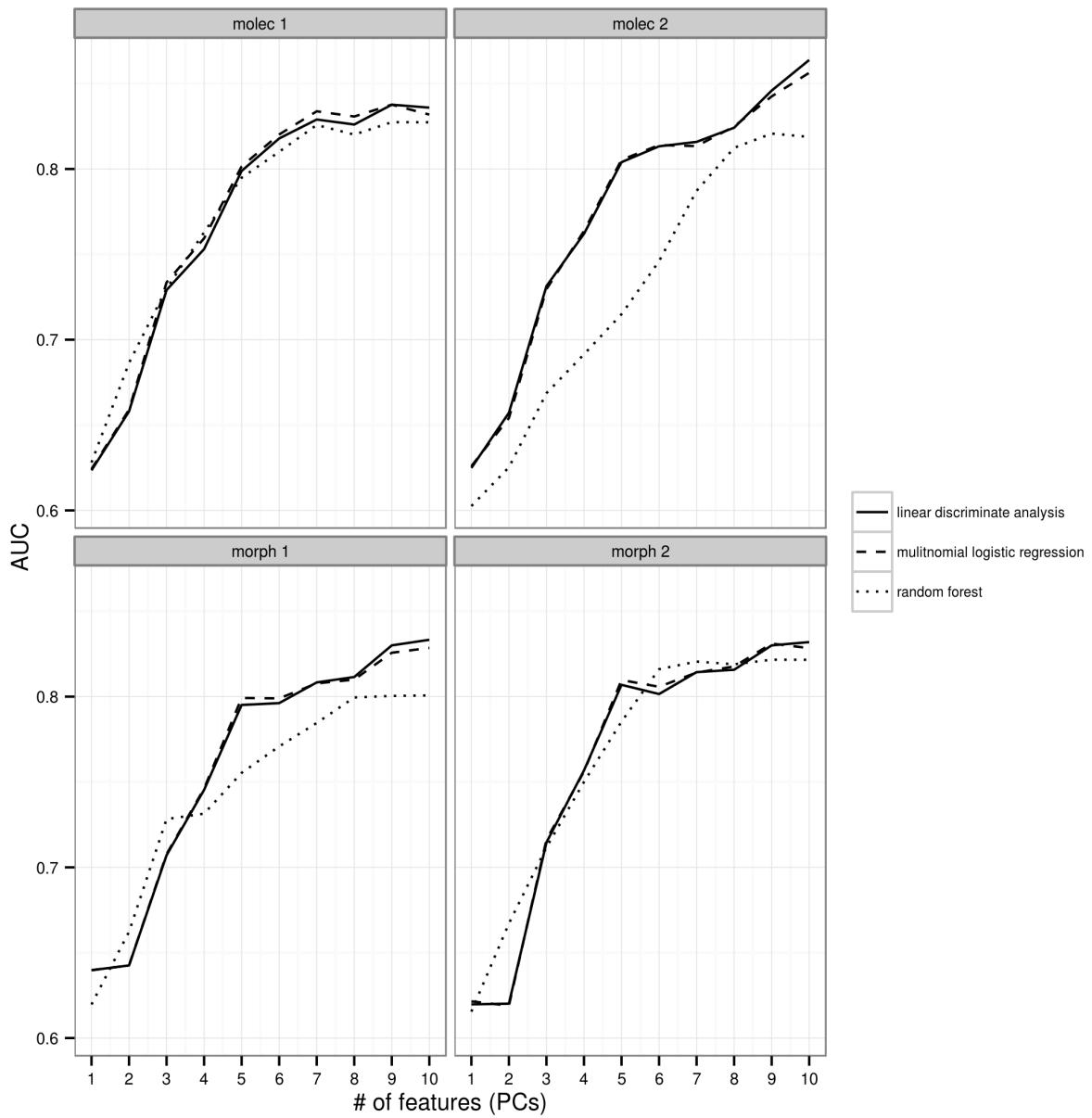


Figure 7: Effect of increasing the number of PCs as features, or predictors, of classification of plastra for all four classification schemes. As the total number of features increase, AUC increases until eventually leveling off. LDA-based classification, multinomial logistic regression and random forest models are illustrated here, though AUC based model selection was only performed for the LDA-based classification and the random forest models.

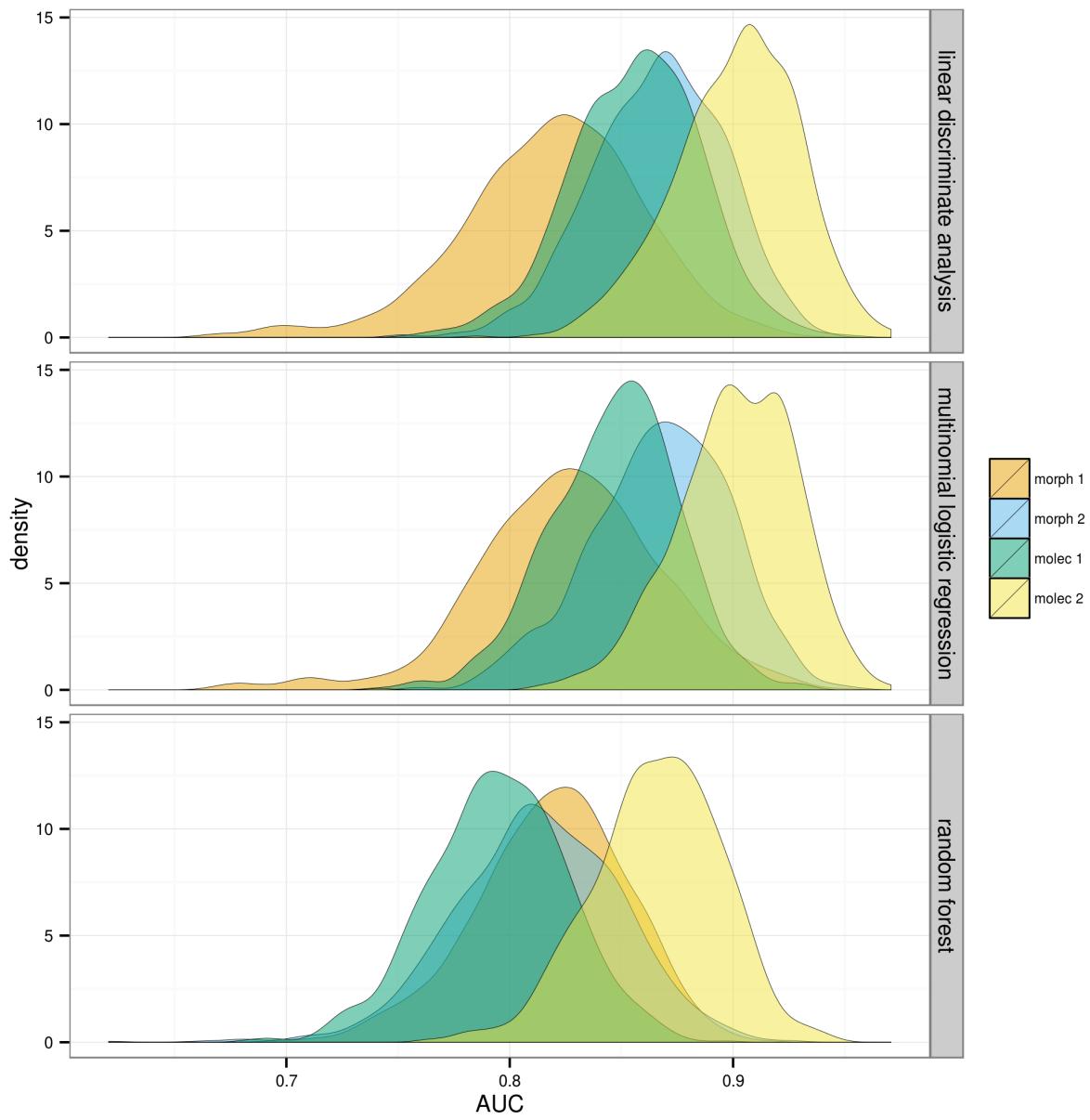


Figure 8: Density estimates of AUC values of predictions of the testing dataset of plastra from 1000 bootstrap resamples. The top facet corresponds to values using the best LDA-based classifications of the eigenscores of shape, as chosen by maximum AUC. The middle facet corresponds to values using the 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.

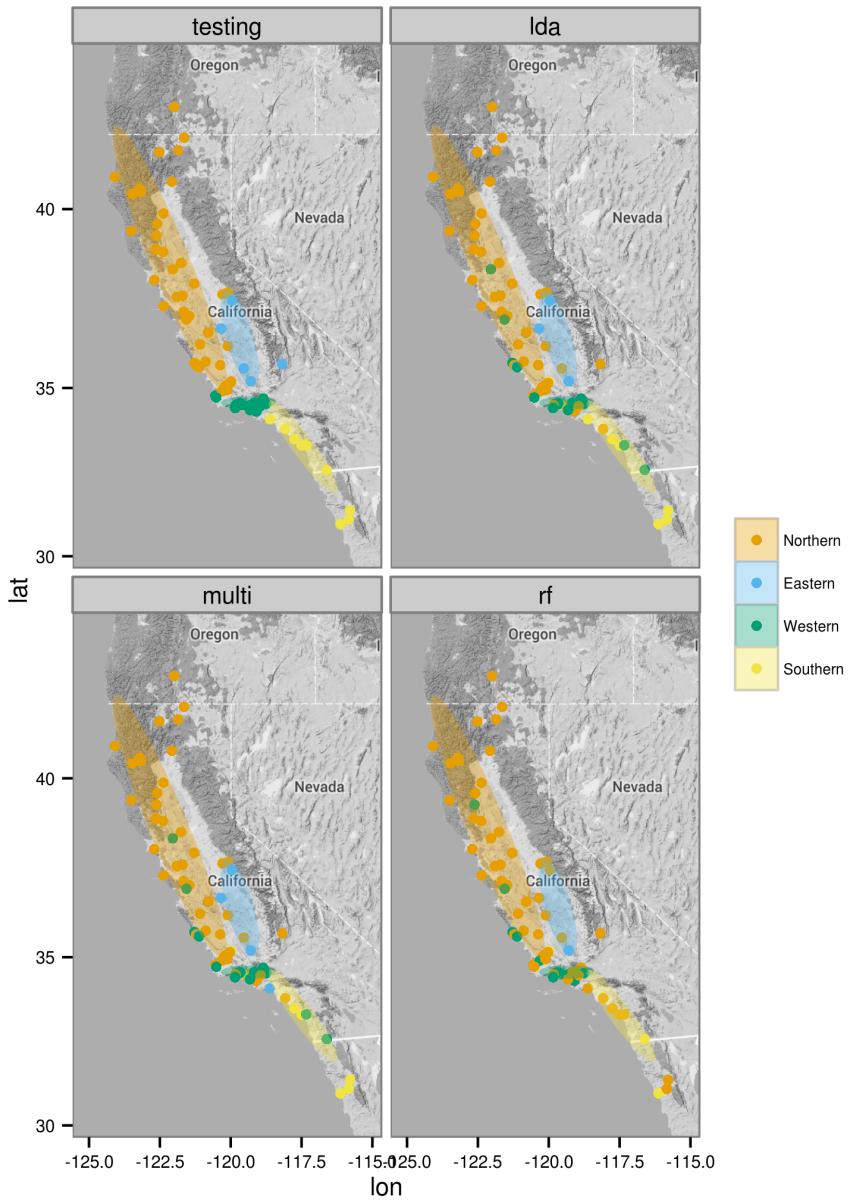


Figure 9: Comparison between reference classification of testing data set (upper left) and the estimated classifications based on the selected LDA-based classification (lda, upper right), multinomial logistic regression (multi, lower left) and random forest models (rf, lower right). Classification corresponds to the four classes as suggested by the hypothesis of Spinks and Shaffer (2005) and Spinks et al. (2010). 95% confidence ellipses for the four classes are depicted and calculated from the initial assignments from the total dataset.

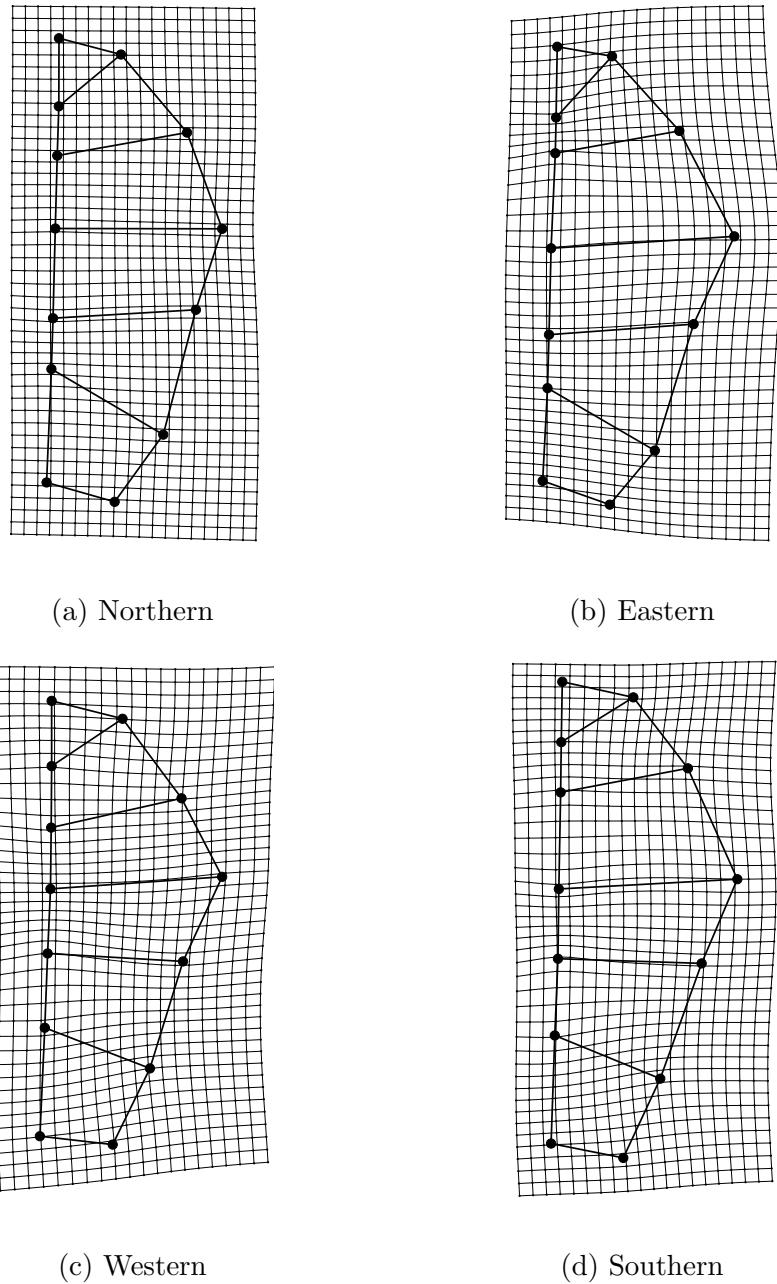


Figure 10: Thin-plate splines for each of the four classes from the best classification hypothesis based on the generalization results (Fig. 8). The four different classes are labeled according to the biogeographic groups as depicted in figure 9. The deformations are depicted with 2x magnification from base.

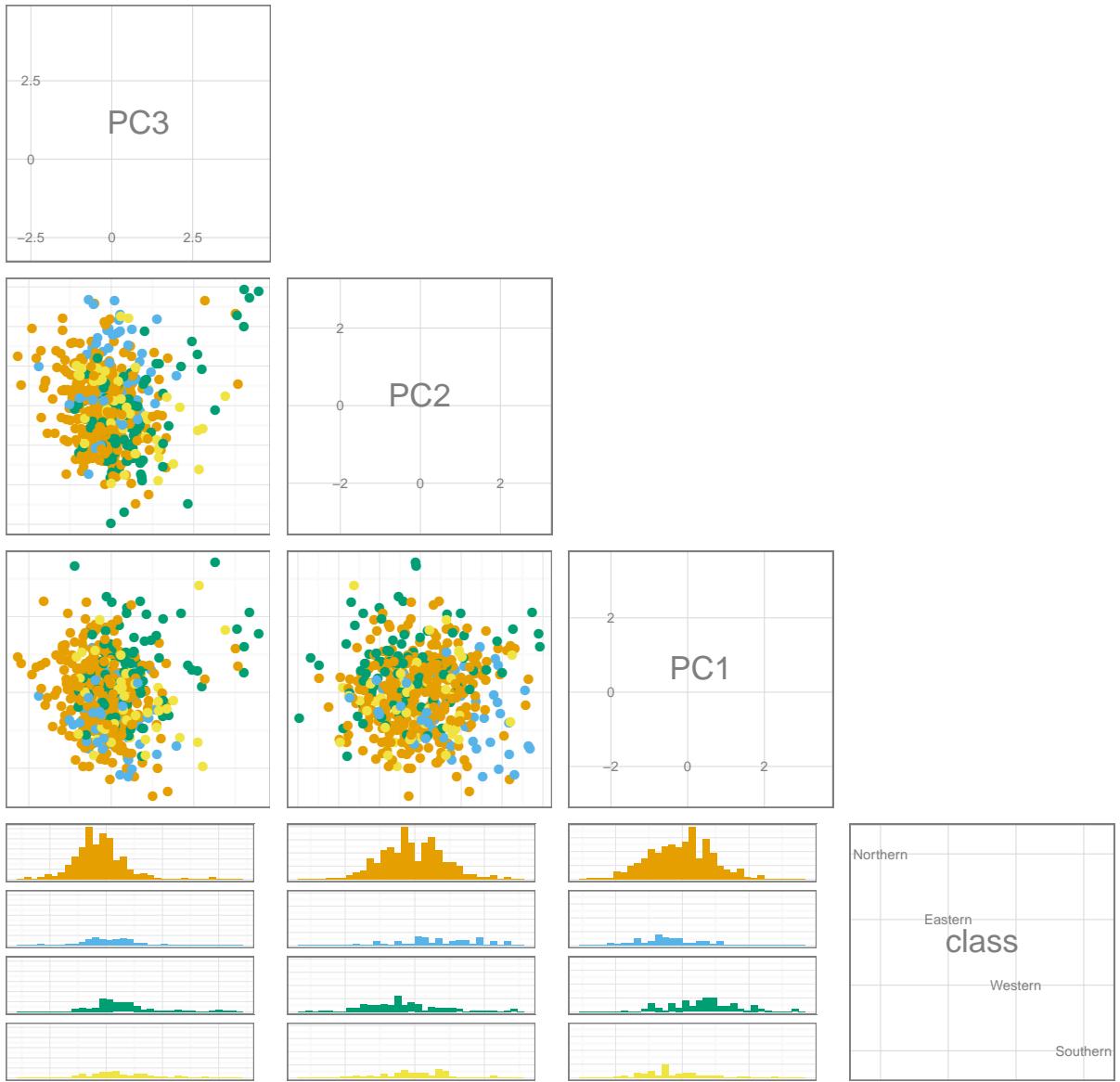


Figure 11: Pairs plot of the first three most important variables of the optimal random forest model of turtle plastral shape. The variables descend in importance from the upper left to the lower right. The observations are colored as in figure 9. The bottom row are histograms of classification occurrences along the PCs.

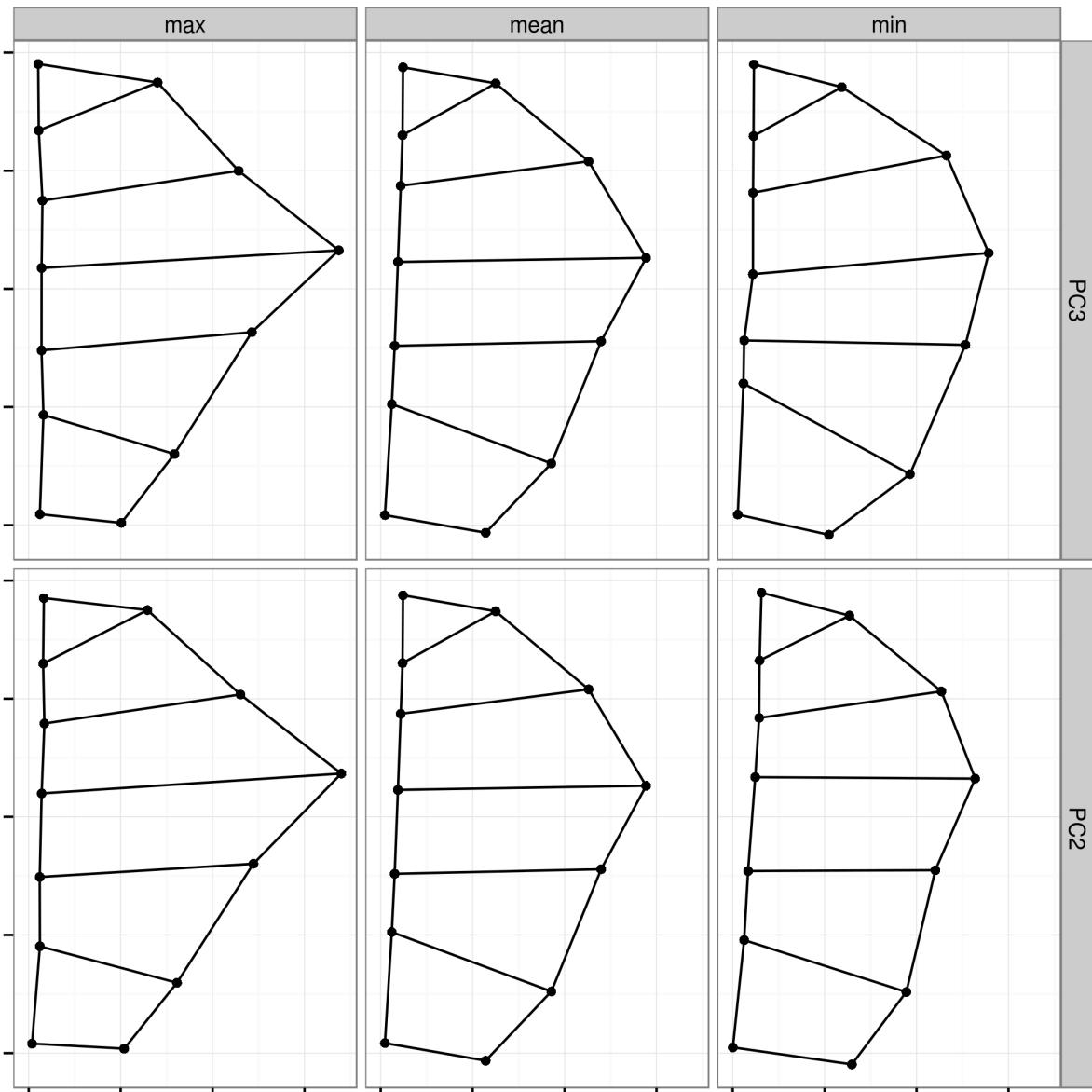


Figure 12: Landmark variation along the two most important features (PCs) based on the final random forest model. The first row corresponds to the third PC and the second corresponds to the second PC. Landmark configurations are minimum observed on that PC, mean shape, and maximum observed on that PC.

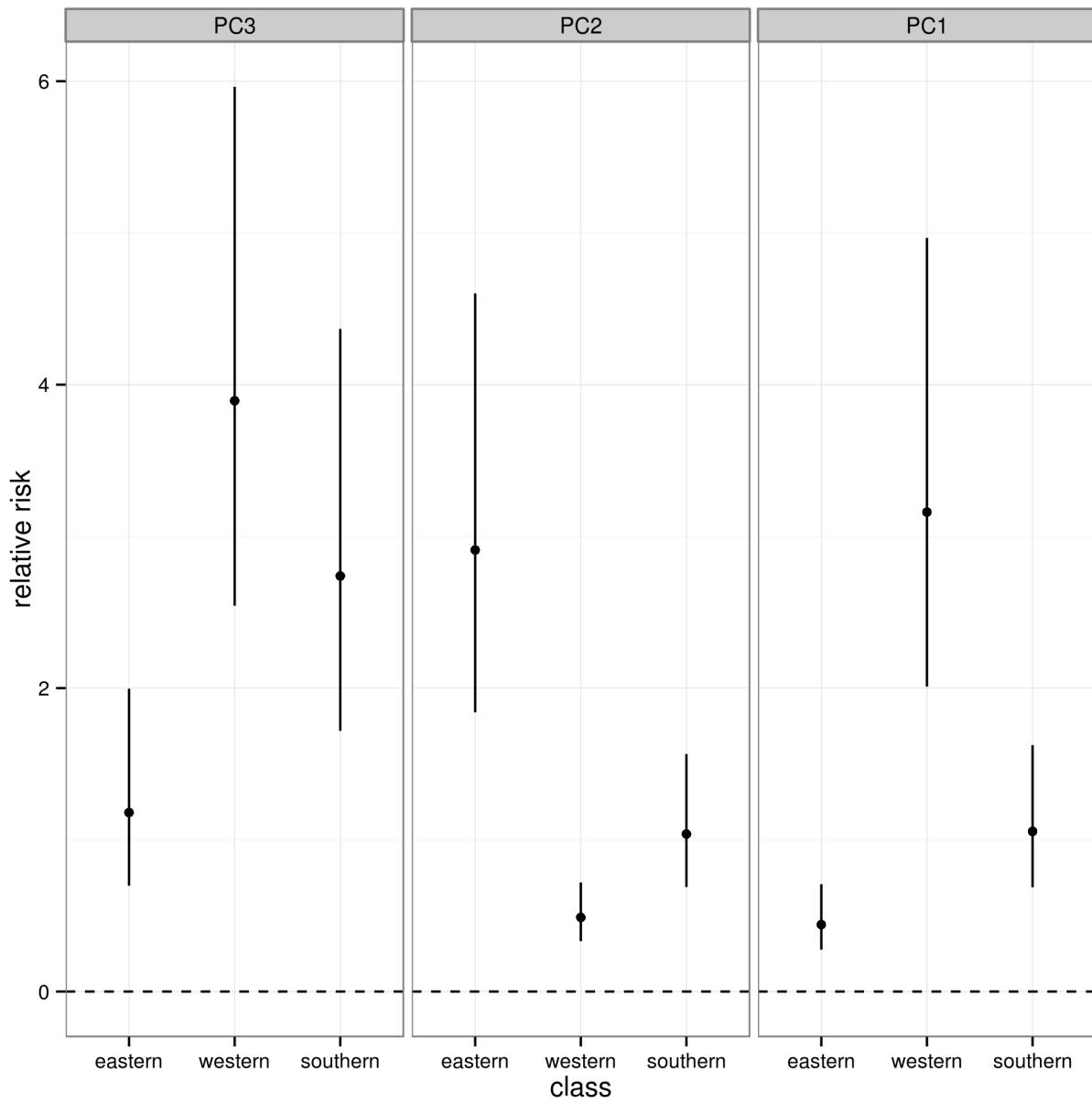


Figure 13: Forest plot of the relative risk, with 95% confidence intervals, of classifying a give specimen based on the first three most important variables according to the random forest model. Relative risk values are calculated from the coefficients of the multinomial logistic regression model. All risks are relative to the northern group from Spinks and Shaffer (2005); Spinks et al. (2010). Variable importance is from left to right.

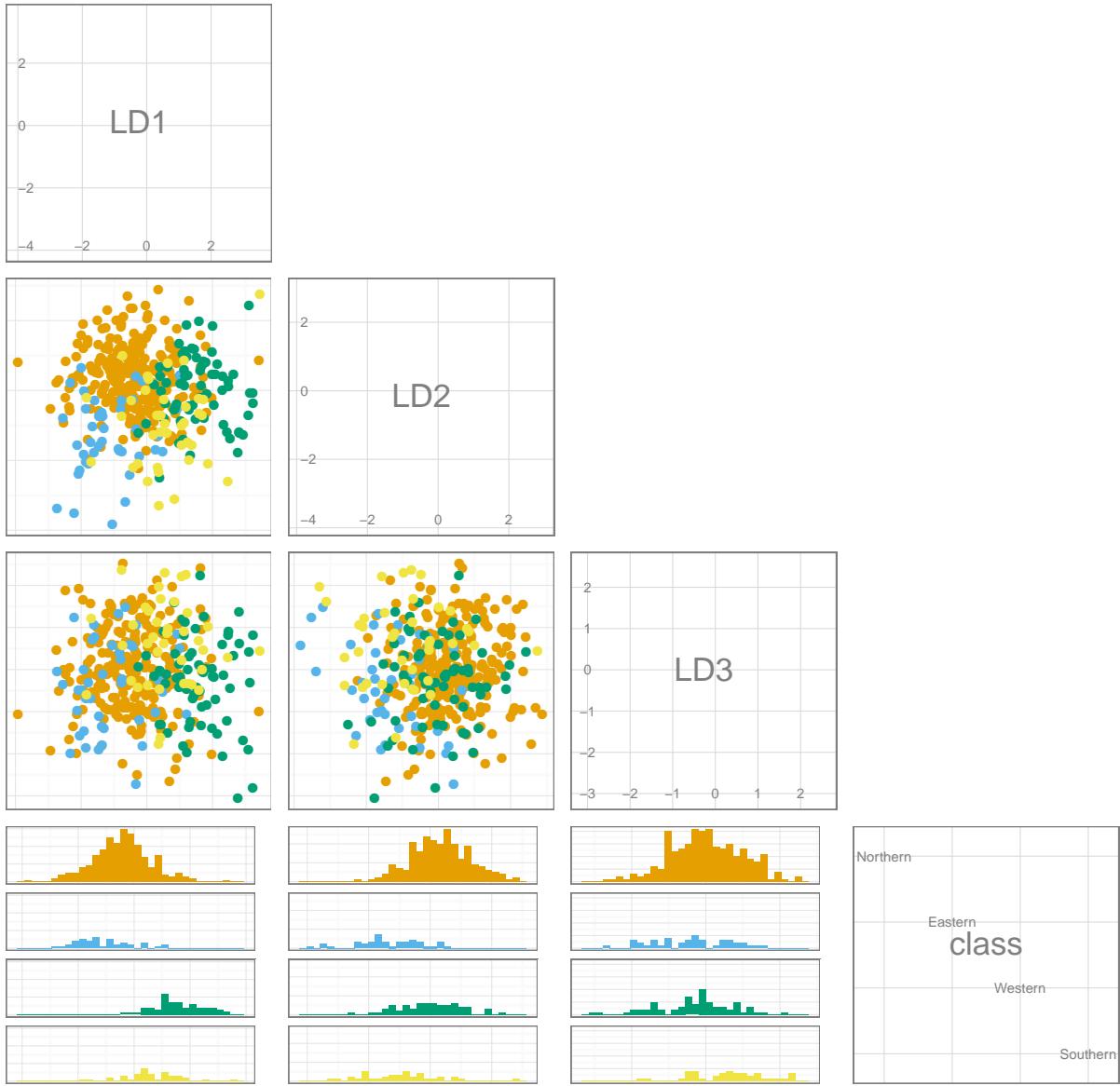


Figure 14: Pairs plots of the three discriminant axes from the linear discriminant analysis of the eigenscores from the first 10 PCs of plastral shape. The observations figured are from the training data set used for all models for the second molecular classification hypothesis based on Spinks and Shaffer (2005) and Spinks et al. (2010). Observations are colored as in Fig. 9.

(Intercept)	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	df	logLik	AICc	delta	weight
+	+	+	+	+	+	+	+	+	+	+	20.00	-250.00	542.26	0.00	0.64
+	+	+	+	+	+	+	+	+	+	+	22.00	-248.35	543.43	1.18	0.36
+	+	+	+	+	+	+	+	+	+	+	16.00	-261.94	557.33	15.07	0.00
+	+	+	+	+	+	+	+	+	+	+	18.00	-259.99	557.82	15.56	0.00
+	+	+	+	+	+	+	+	+	+	+	14.00	-275.68	580.48	38.22	0.00
+	+	+	+	+	+	+	+	+	+	+	12.00	-281.10	587.03	44.77	0.00
+	+	+	+	+	+	+	+	+	+	+	10.00	-305.55	631.68	89.43	0.00
+	+	+	+	+	+	+	+	+	+	+	8.00	-318.48	653.34	111.09	0.00
+	+	+	+	+	+	+	+	+	+	+	6.00	-344.14	700.49	158.24	0.00
+	+	+	+	+	+	+	+	+	+	+	4.00	-346.80	701.71	159.45	0.00

Table 3: Model selection table for the multinomial logistic regression models of the first morphologically based classification hypothesis. This classification hypothesis corresponds to “morph 1” also depicted in figures 7 and 8. This hypothesis is based on Seeliger (1945). The column “delta” corresponds to the ΔAICc values of each model, while “weights” correspond to the Akaike weight of that model relative to all others.

(Intercept)	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	df	logLik	AICc	delta	weight
+	+	+	+	+	+	+	+	+	+	+	20.00	-245.15	532.56	0.00	0.83
+	+	+	+	+	+	+	+	+	+	+	22.00	-244.53	535.79	3.23	0.17
+	+	+	+	+	+	+	+	+	+	+	18.00	-254.69	547.21	14.64	0.00
+	+	+	+	+	+	+	+	+	+	+	16.00	-258.00	549.45	16.88	0.00
+	+	+	+	+	+	+	+	+	+	+	14.00	-268.69	566.49	33.93	0.00
+	+	+	+	+	+	+	+	+	+	+	12.00	-271.30	567.42	34.86	0.00
+	+	+	+	+	+	+	+	+	+	+	10.00	-298.53	617.64	85.07	0.00
+	+	+	+	+	+	+	+	+	+	+	8.00	-314.50	645.37	112.81	0.00
+	+	+	+	+	+	+	+	+	+	+	6.00	-342.94	698.10	165.53	0.00
+	+	+	+	+	+	+	+	+	+	+	4.00	-349.55	707.20	174.64	0.00

Table 4: Model selection table for the multinomial logistic regression models of the first morphologically based classification hypothesis. This classification hypothesis corresponds to “morph 2” also depicted in figures 7 and 8. This hypothesis is based on Seeliger (1945). The column “delta” corresponds to the ΔAICc values of each model, while “weights” correspond to the Akaike weight of that model relative to all others.

(Intercept)	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	df	logLik	AICc	delta	weight
+	+	+	+	+	+	+	+	+	+	+	30.00	-303.61	672.34	0.00	0.77
+	+	+	+	+	+	+	+	+	+	+	33.00	-301.25	674.74	2.41	0.23
+	+	+	+	+	+	+	+	+	+	+	27.00	-314.28	686.70	14.36	0.00
+	+	+	+	+	+	+	+	+	+	+	24.00	-318.22	687.70	15.37	0.00
+	+	+	+	+	+	+	+	+	+	+	21.00	-335.11	714.71	42.37	0.00
+	+	+	+	+	+	+	+	+	+	+	18.00	-353.04	743.91	71.57	0.00
+	+	+	+	+	+	+	+	+	+	+	15.00	-385.20	801.67	129.34	0.00
+	+	+	+	+	+	+	+	+	+	+	12.00	-397.69	820.21	147.87	0.00
+	+	+	+	+	+	+	+	+	+	+	9.00	-437.13	892.73	220.39	0.00
+	+	+	+	+	+	+	+	+	+	+	6.00	-451.19	914.60	242.27	0.00

Table 5: Model selection table for the multinomial logistic regression models of the first morphologically based classification hypothesis. This classification hypothesis corresponds to “molec 1” also depicted in figures 7 and 8. This hypothesis is based on Seeliger (1945). The column “delta” corresponds to the ΔAICc values of each model, while “weights” correspond to the Akaike weight of that model relative to all others.

(Intercept)	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	df	logLik	AICc	delta	weight
+	+	+	+	+	+	+	+	+	+	+	33.00	-251.73	575.67	0.00	1.00
+	+	+	+	+	+	+	+	+	+	+	30.00	-268.54	602.18	26.51	0.00
+	+	+	+	+	+	+	+	+	+	+	27.00	-283.99	626.10	50.43	0.00
+	+	+	+	+	+	+	+	+	+	+	24.00	-295.61	642.46	66.78	0.00
+	+	+	+	+	+	+	+	+	+	+	21.00	-302.50	649.48	73.81	0.00
+	+	+	+	+	+	+	+	+	+	+	18.00	-316.59	671.00	95.32	0.00
+	+	+	+	+	+	+	+	+	+	+	15.00	-340.84	712.95	137.27	0.00
+	+	+	+	+	+	+	+	+	+	+	12.00	-353.01	730.84	155.17	0.00
+	+	+	+	+	+	+	+	+	+	+	9.00	-378.16	774.78	199.11	0.00
+	+	+	+	+	+	+	+	+	+	+	6.00	-395.71	803.64	227.97	0.00

Table 6: Model selection table for the multinomial logistic regression models of the first morphologically based classification hypothesis. This classification hypothesis corresponds to “molec 2” also depicted in figures 7 and 8. This hypothesis is based on Seeliger (1945). The column “delta” corresponds to the ΔAICc values of each model, while “weights” correspond to the Akaike weight of that model relative to all others.

	morph 1	morph 2	molec 1	molec 2
morph 1				
morph 2		0.00		
molec 1		0.00	0.00	
molec 2		0.00	0.00	0.00

Table 7: Results from pairwise Mann-Whitney U test between the AUC distributions of the generalizations of the LDA-based classification from the first 10 PCs of plastral shape. Labels correspond to those in Figure 8. Values of 0 correspond to p-values lower than 0.01. P-values were corrected for multiple comparison using the Holm method (Holm 1979).

	morph 1	morph 2	molec 1	molec 2
morph 1				
morph 2		0.00		
molec 1		0.00	0.00	
molec 2		0.00	0.00	0.00

Table 8: Results from pairwise Mann-Whitney U test between the AUC distributions of the generalizations of the multinomial logistic regression models. Labels correspond to those in Figure 8. Values of 0 correspond to p-values lower than 0.01. P-values were corrected for multiple comparison using the Holm method (Holm 1979).

	morph 1	morph 2	molec 1	molec 2
morph 1				
morph 2		0.00		
molec 1	0.00		0.00	
molec 2	0.00	0.00	0.00	

Table 9: Results from pairwise Mann-Whitney U test between the AUC distributions of the generalizations of the random forest models. Labels correspond to those in Figure 8. Values of 0 correspond to p-values lower than 0.01. P-values were corrected for multiple comparison using the Holm method (Holm 1979).