

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

September 4, 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 (Bickford et al. 2007; Clare
30 2011; Funk et al. 2012; Pfenninger and Schwenk 2007; Schilck-Steiner et al. 2007; Stuart
et al. 2006). Concerningly, most extant taxa, and nearly all extinct taxa, are delimited based
32 solely via morphology. This phenomenon is of great concern when studying variation and
diversity dynamics over long periods of time, where apparent morphological stasis may not

³⁴ actually reflect true diversity (Eldredge and Gould 1972; Gould and Eldredge 1977; Hunt 2008;
Van Bocxlaer and Hunt 2013). In the case of endangered or conserved taxa, morphometric
³⁶ 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
³⁸ classification of extinct taxa.

Much work has been devoted to species delimitation via sequence difference (Fujita
⁴⁰ et al. 2012; Yang and Rannala 2010) while comparatively little has been devoted to introducing
new methodology for case of purely morphological data (Mitteroecker and Bookstein 2011;
⁴² Zelditch et al. 2004). The majority of this effort has focused on identifying differences between
already identified taxa (Demandt and Bergek 2009; Gaubert et al. 2005; Gündüz et al. 2007;
⁴⁴ Polly 2003, 2007; Zelditch et al. 2004) and automated taxon identification (MacLeod 2007).

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

⁵⁰ *Background and system*

Machine learning is in many respects just an extension of known statistical methodology
⁵² (Hastie et al. 2009) where the emphasis is placed on inferring rules and properties of data in
order to explain the underlying structure. The basic statistical mechanics are supplemented
⁵⁴ by randomization, sorting, and partitioning algorithms and along with the maximization or
minimization of summary statistics in order to best estimate a general model for all data,
⁵⁶ 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
⁵⁸ handwritten zip codes (Hastie et al. 2009). Two major classes of machine learning methods are

unsupervised and supervised learning. Unsupervised learning methods are used with unlabeled
60 data where the underlying structure is estimated and are analogous to clustering and density
estimation methods (Kaufman and Rousseeuw 1990). Supervised learning methods are used
62 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 classification and regression models (Breiman
64 et al. 1984). The application of the alternative approaches used in this study illustrates only
a sampling of the various previously derived methods for clustering observations and fitting
66 classification models.

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

Here, we used multiple machine learning methods, both unsupervised and supervised,
84 in order to compare different classification hypotheses. These methods provide different and

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

94 In this study, we investiage the subspecific classification of the western pond turtle,
Emys marmorata. *E. marmorata* is distributed from northern Washington State, USA
96 to Baja California, Mexico. Traditionally, *E. marmorata* was classified into three groups:
the northern *E. marmorata marmorata*, the southern *E. marmorata palida*, and a central
98 Californian intergrade zone (Seeliger 1945). *E. marmorata marmorata* is differentiated from
E. marmorata palida by the presence of a pair of triangular inguinal plates and darker neck
100 markings. It should be noted that the triangular inguinal plates can sometimes be present in
E. marmorata palida though they are considerably smaller.

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

110 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

2007) with males being on average larger than females based on total carapace length and
112 other linear measurements. However, the quality of size as a classifier of sex can vary greatly
between populations (Holland 1992), which makes sense in light of the amount of between
114 population size difference (Germano and Bury 2009; Lubcke and Wilson 2007). However, the
effect of sexual dimorphism on shape, *sensu* Kendall (1977), was not assessed (Germano and
116 Rathbun 2008; Holland 1992; Lubcke and Wilson 2007).

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

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

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

one northern and one southern, Spinks et al. (2010) argue that the four clade classification is
138 of greater conservation utility even though the variation between these groups is considered
cryptic. While the mitochondrially based classification is considered robust, there is no known
140 morphological differentiation between these clades.

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

152 MATERIALS AND METHODS

Specimens

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

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

¹⁷⁰ *Geometric morphometrics*

Following previous work on plastral variation (Angielczyk and Feldman 2013; Angielczyk
¹⁷² et al. 2011; Angielczyk and Sheets 2007), 19 landmarks were digitized using TpsDig 2.04
(Rohlf 2005). These landmarks were chosen to maximize the description of general plastral
¹⁷⁴ variation(Fig. 1). 17 of these landmarks are at the endpoints or intersection of the keratinous
plastral scutes that cover the platron. 12 of these landmarks were chosen to be symmetrical
¹⁷⁶ across the axis of symmetry and, in order to prevent degrees of freedom and other concerns
(Klingenberg et al. 2002), prior to analysis these landmarks were reflected across the axis of
¹⁷⁸ symmetry (i.e. midline) and the average position of each symmetrical pair was used. In cases
where damage or incompleteness prevented symmetric landmarks from being determined,
¹⁸⁰ only the single member of the pair was used. Analysis was conducted on the resulting “half”
plastron. Plastral landmark configurations were superimposed using generalized Procrustes
¹⁸² analysis (Dryden and Mardia 1998) after which, the principal components (PC) of shape
were calculated. This was done using the **shapes** package for R (Dryden 2013; R Core Team
¹⁸⁴ 2013).

Machine learning analyses

¹⁸⁶ *Unsupervised learning.*— In order to preserve the relationship between all landmark configurations in shape space, the dissimilarity between observations was measured using Kendall's
¹⁸⁸ Riemanninan shape distance or ρ (Dryden and Mardia 1998; Kendall 1984). This metric was chosen because shape space, or the set of all possible shape configurations following
¹⁹⁰ 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
¹⁹² 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
¹⁹⁴ 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
¹⁹⁶ dissimilarities between observations and mediods is minimized (Kaufman and Rousseeuw 1990). Because the optimal number of clusters of shape configurations in the study was
¹⁹⁸ 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
²⁰⁰ using the gap statistic, which is a measure of goodness of clustering (Tibshirani et al. 2001).

The gap statistic is defined

$$Gap_n(k) = E_n^*[\log(W_k)] - \log(W_k)$$

²⁰² where W_k is

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

²⁰⁴ . $d_{ii'}$ 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
²⁰⁶ distribution. In this case, the reference distribution was estimated from 500 resamples of the

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

Supervised learning.— The total dataset of 524 observations was split into training and testing
210 datasets. The training dataset represented 75% of the total dataset, split proportionally
by class, and was used for model fitting. The testing dataset represented the remaining
212 25% of the total dataset and was used after model fitting to estimate the effectiveness of
each classification hypothesis and generalizability of the supervised learning models (i.e.
214 performance in the wild). This split was chosen to allow for a large enough sample size for
model fitting while also providing a large enough testing dataset to determine any systematic
216 misclassifications.

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

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

Multinomial logistic regression is an extension of logistic regression, where instead

232 of a binary response there are three or more response classes (Venables and Ripley 2002).
Effectively, this type of model can be viewed as multiple, simultaneous logistic regression
234 models for each class and the final classification of the observation being the most probable
of all the constituent model results. Similar to the odds ratios calculated from the coefficients
236 of a logistic regression, the relative risk of a classification with reference to a baseline class
can be determined from the coefficients of the model. Multinomial logistic regression models
238 were fit using the `nnet` package for R (Venables and Ripley 2002)

Random forest models are an extension of classification and regression trees (CART)
240 (Breiman 2001; Breiman et al. 1984). Because this study relies on classification models,
CARTs are explained with reference to classification but the approach is equally valid for
242 regression. The goal of CARTs are to use a series of different features to estimate the final
class. In top-down induction of decision trees for each member of a given set of predictor
244 variables, attribute value test are used to estimate the differences between classes. This
process is then repeated on each subset, called recursive partitioning. The recursion continues
246 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
248 each intersection via the estimated cutoff points from the attribute tests made during model
fitting.

250 In a random forest model, many CARTs are built from a random subsample of both the
features and the observations. This process is then repeated many times and the parameters
252 of the final model was chosen as the mode of estimates from the distribution of CARTs
(Breiman 2001). In addition to fitting a classification model, this procedure allows for the
254 features to be ranked in order of importance. In the context of this study, this means that
the PCs most important for describing the difference between classes can be estimated, and
256 thus illustrate the most important variation amongst classes as opposed to just the greatest
amount of variation in the entire dataset. This is a generally important property that is

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

The supervised learning models used here, except LDA, have tuning parameters which
262 help to increase the generalizability of the model and prevent them from being overfit. For the
supervised learning models fit in this study, tuning parameters were estimated via 10 rounds
264 of 10-fold cross-validation (CV) across a grid search of all tuning parameter combinations.

Optimal tuning parameter values were selected based on area under the receiver operating
266 characteristic (ROC) curve. The area under the multiclass ROC curves was estimated using
the all-against one strategy derived by Hand and Till (2001). This tuning process was
268 implemented following the default grid search implemented in the `caret` package for R (Kuhn
2013).

270 ROC is a confusion matrix (Table 1) statistic that is a descriptor the relationship
between the false positive rate (FPR , Eq. 1) of a classification model and the true positive
272 rate (TPR , Eq. 2) of a classification model (Hastie et al. 2009). The area under the ROC
curve (AUC) is a summary statistic of the quality of the classification and varies between
274 0.5 and 1, with values of 0.5 indicating a model that classifies no better than random and a
value of 1 indicating perfect classification (Hastie et al. 2009). AUC can be used as a model
276 selection criterion for classification models and is especially useful in cases where some if not
all of the models in question were not fit via maximum likelihood where a criterion such as
278 AICc (see below) or similar can be used (Hastie et al. 2009). It is important to note that,
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
282 from two to 10 in increasing order of complexity. In total, this produced nine different LDA
scaling matrices. From this set, the best number of PCs used to estimate the LDA scaling
284 matrix were chosen. As LDA is not “fit” via maximum likelihood, the final combination of
number of PCs and LDA scaling matrix chosen was that with the greatest AUC value from
286 the training set.

For the multinomial logistic regression models, 10 different models were fit each having
288 sequentially more PCs as predictors in order to have models representing different levels of
overall amount of shape variation and to estimate how much was necessary and sufficient
290 to best estimate class. The maximum number of PCs allowed as predictors was 10 because
of both the large number of parameters estimated per model and the necessary sample size
292 needed to estimate that many parameters accurately. The final model was that with the lowest
AICc (Akaike 1974; Burnham and Anderson 2002; Hurvich and Tsai 1989). AICc is a model
294 selection criterion where the model with lowest AICc has the fairest variance–bias tradeoff
(Burnham and Anderson 2002). Model selection was performed in this manner because the
296 optimal number of PCs to use as predictors was not known *a priori*, and while including all
of the PCs of shape would mean that all shape variability would be used to estimate class,
298 this may cause the model to be overfit and not provide an accurate estimate of unsampled
plastral variation. In addition to the AICc of each model the Δ AICc and Akaike weights are
300 also reported. Δ AICc values are the difference in AICc between the AICc best model and that
model while Akaike weights are a transformation of the AICc of a model with relation to all
302 other models being compared and measures the relative amount of information explained by
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
306 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
308 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
310 recalculated. This was repeated until only one feature, remained. Similar to the multinomial
logistic regression models described above, the maximum number of PCs that could have
312 been included in the model was 10. After each PC was removed , 10-fold CV was used to
estimate the optimal values of the tuning parameters as well as quantify the uncertainty of
314 each model. Random forest model parameters were estimated from 1000 subtrees. Because
PCs were kept in order of importance and not in relation to the amount of variance each PC
316 described, this means that the exact PCs included in each model do not correspond to the
PCs in each of the 10 multinomial logistic regressions models.

318 The final selected models were then used to estimate the class assignments of the
training dataset. Model generality for both methods for all four classification schemes was
320 measured using the AUC of the assignments. A distribution of AUC values was estimated
for each classification scheme via 1000 nonparametric bootstrap resamples of the training
322 dataset. The difference in distributions was assessed using pairwise Mann-Whitney U tests.

RESULTS

324 *Geometric morphometrics*

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

³³⁰ However, it is important to note that Procrustes based superimposition attempts to evenly
³³¹ distribute variance around the mean shape (Zelditch et al. 2004) and this observation should
³³² be considered cursory at best.

The first two PCs appear to describe principally variation in the lateral margin of the
³³⁴ palstra, from a pointed medial edge to a more rounded and blunt edge (Fig. 3). Landmark 10
³³⁵ (Fig. 1), which appears to be the most variable along these axes (Fig. 2 and 3), is positioned
³³⁶ on the bridge between the plastron and the carapace. Over ontogeny, this is an area that
³³⁷ deepens dorsoventrally and when the plastron was projected into two dimensions it created
³³⁸ the effect of mediolateral movement. Lateral landmark variation along the first PC seemed
³³⁹ concentrated in the posterior portion of the plastra with additional variance observed in
³⁴⁰ midline landmarks (Fig. 3). This variance in midline landmarks was most likely caused by
³⁴¹ the fact that plastral scutes frequently do not line up perfectly. Along PC 2, lateral variation
³⁴² appeared to be concentrated in the anterior portion o the plastra (Fig. 3).

When the eigenscores of the first two PCs are compared to the centroid sizes of the
³⁴⁴ observations, a significant correlation is observed between centroid size and the first PC
³⁴⁵ (Spearman's ρ : -0.56, p -value: 7.51×10^{-44}) but not between centroid size and the second
³⁴⁶ PC (Spearman's ρ : 0.03, p -value: 0.55). This difference is clearly observable in in plots of
the data (Fig. 4).

³⁴⁸ *Machine learning analyses*

Unsupervised learning.—

³⁵⁰ Comparison of gap statistic values for the range of PAM solutions indicates that the
³⁵¹ optimal number of clusters is one (Fig. 5). The next best clustering solution had only two
³⁵² clusters, however there is no geographic structure to this classification scheme, with members
of these clusters being seemingly randomly distributed (Fig. 6). Importantly, these clusters do

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

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

³⁶⁶ The gap statistic values for both three and four clusters are much lower than for one and two and are statistically identical. Interestingly, other solutions with a much greater
³⁶⁸ number of clusters have relatively high gap statistic values as well though these are also not significantly different. Increasing the number of clusters does appear to improve the gap
³⁷⁰ statistic enough compared to the best clustering solution to merit detailed discussion.

Supervised learning.—

³⁷² The optimal number of PCs used for LDA, as determined by highest ROC score, for three of the four classification schemes had all 10 possible PCs (Fig. 7). These were both of
³⁷⁴ the morphological based classification hypotheses and the second molecular hypothesis. LDA of the PCs of the first molecular hypothesis found that, based on ROC, only the first 9 PCs
³⁷⁶ were necessary to best discriminate between the classes (Fig. 7). The first 9 PCs describe 83.23% of total variation in plastral shape, while the first 10 PCs describe 86.54% of the
³⁷⁸ variation.

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

While the ΔAICc value between the optimal and second best model for the first
386 morphological and first molecular based classification hypothesis was within the range to
be considered equally optimal (Burnham and Anderson 2002), for this analysis we chose to
388 use only the AICc best model. While AICc values can not be compared between models
with different responses (Burnham and Anderson 2002), we interpret the fact that the
390 ΔAICc best model in these cases is the simpler model and that the optimal model for
three of the classification schemes having the same number of predictors as reasons to use
392 only the AICc best model for all cases. Additionally, by using a single model for each of
the classification hypotheses, this limits the number of comparisons between the bootstrap
394 resampled distributions of the AUC values for the testing dataset (see below).

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

In the case of all models, there is a substantial increase in model performance as
400 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).

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

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

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

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

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

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

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

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

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

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

DISCUSSION

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

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

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

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

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

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

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

516

ACKNOWLEDGEMENTS

PDS would like to thank David Bapst, Michael Foote, Benjamin Frable, and Dallas Krentzel
518 for useful discussion which enhanced the quality of this study. For access to emydine specimens,
we thank: J. Vindum and R. Drewes (CAS); A. Resetar (FMNH); R. Feeney (LACM); C.
520 Austin (LSUMNS); S. Sweet (MSE); J. McGuire and C. Conroy (MVZ); A. Wynn (NMNH);
P. Collins (SBMNH); B. Hollingsworth (SDMNH); C. Bell and R. Burroughs (TMM); T.
522 LaDuc and R. Burroughs (TNHC); P. Holroyd (UCMP); R. Symonds (UMZC); J. Buskirk.
We are greatful to S. Sweet for field assistance and the California Department of Fish and
524 Game for permits. Much of the data collection was funded by NSF DBI-0306158 (to KDA).

*

526

References

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

- Angielczyk, K. D. and C. R. Feldman. 2013. Are diminutive turtles miniaturized? The ontogeny
530 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
532 in emydine turtles. *Evolution* 65:377–394.
- Angielczyk, K. D. and H. D. Sheets. 2007. Investigation of simulated tectonic deformation in
534 fossils using geometric morphometrics. *Paleobiology* 33:125–148.
- Bickford, D., D. J. Lohman, N. S. Sodhi, P. K. L. Ng, R. Meier, K. Winker, K. K. Ingram, and
536 I. Das. 2007. Cryptic species as a window on diversity and conservation. *Trends in ecology &*
evolution 22:148–55.
- 538 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
540 trees. Wadsworth International Group, Belmont.
- Burnham, K. P. and D. R. Anderson. 2002. Model selection and multi-model inference: a
542 practical information-theoretic approach. 2nd ed. Springer, New York.
- Bury, R. B., D. J. Germano, and G. W. Bury. 2010. Population Structure and Growth of the
544 Turtle *Actinemys marmorata* from the KlamathSiskiyou Ecoregion: Age, Not Size, Matters.
Copeia 2010:443–451.
- 546 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
548 from the Vancouver Island marmot. *Ethology Ecology & Evolution* 21:209–223.
- Caumul, R. and P. D. Polly. 2005. Phylogenetic and environmental components of morphological
550 variation: skull, mandible, and molar shape in marmots (Marmota, Rodentia). *Evolution; international journal of organic evolution* 59:2460–72.

- 552 Clare, E. L. 2011. Cryptic species? Patterns of maternal and paternal gene flow in eight
neotropical bats. *PLoS one* 6:e21460.
- 554 Demandt, M. H. and S. Bergek. 2009. Identification of cyprinid hybrids by using geometric
morphometrics and microsatellites. *Journal of Applied Ichthyology* 25:695–701.
- 556 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.
- 558 Eldredge, 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
560 Francisco.
- Fraley, C. and A. E. Raftery. 2002. Model-Based Clustering, Discriminant Analysis, and Density
562 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
564 species delimitation in an integrative taxonomy. *Trends in ecology & evolution* 27:480–8.
- Funk, W. C., M. Caminer, and S. R. Ron. 2012. High levels of cryptic species diversity uncovered
566 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
568 of cryptic hybridization revealed using an integrative approach: a case study on genets
(Carnivora, Viverridae, Genetta spp.) from the southern African subregion. *Biological Journal
570 of the Linnean Society* 86:11–33.
- Germano, D. J. and R. B. Bury. 2009. Variation in body size, growth, and population structure
572 of *Actinemys marmorata* from lentic and lotic habitats in Southern Oregon. *Journal of
Herpetology* 43:510–520.

⁵⁷⁴ 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*
⁵⁷⁶ *Conservation and Biology* 7:188–194.

Gould, S. J. and N. Eldredge. 1977. Punctuated equilibria: the tempo and mode of evolution
⁵⁷⁸ reconsidered. *Paleobiology* 3:115–151.

Gündüz, I., M. Jaarola, C. Tez, C. Yeniyurt, P. D. Polly, and J. B. Searle. 2007. Multigenic
⁵⁸⁰ 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.

⁵⁸² 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.

⁵⁸⁴ Hastie, T., R. Tibshirani, and J. Friedman. 2009. *The elements of statistical learning: data mining, inference, and prediction*. 2nd ed. Springer, New York.

⁵⁸⁶ 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
⁵⁸⁸ 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
⁵⁹⁴ 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.
- ⁵⁹⁸Kendall, D. G. 1984. Shape Manifolds, Procrustean Metrics, and Complex Projective Spaces. *Bulletin of the London Mathematical Society* 16:81–121.
- ⁶⁰⁰Klingenberg, C. P., M. Barluenga, and A. Meyer. 2002. Shape analysis of symmetric structures: quantifying variation among individuals and asymmetry. *Evolution* 56:1909–1920.
- ⁶⁰²Kuhn, 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.
- ⁶⁰⁴Lubcke, 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.
- ⁶⁰⁶MacLeod, 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.
- ⁶¹⁰Mitteroecker, 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.
- ⁶¹⁴Polly, P. D. 2003. Paleophylogeography of *Sorex araneus*: molar shape as a morphological marker for fossil shrews. *Mammalia* 68:233–243.
- ⁶¹⁶Polly, 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
620 for Statistical Computing Vienna, Austria.

Rohlf, F. J. 2005. TpsDig 2.04.

622 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.
624 Trends in ecology & evolution 22:391–392.

Seeliger, L. M. 1945. Variation in the Pacific Mud Turtle. Copeia 1945:150–159.

626 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
628 implications. Molecular ecology 14:2047–64.

Spinks, P. Q., R. C. Thomson, and H. B. Shaffer. 2010. Nuclear gene phylogeography reveals
630 the historical legacy of an ancient inland sea on lineages of the western pond turtle, *Emys*
marmorata in California. Molecular ecology 19:542–56.

632 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.

634 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
636 Methodology) 63:411–423.

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

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

Yang, Z. and B. Rannala. 2010. Bayesian species delimitation using multilocus sequence data.
642 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:
644 a primer. Elsevier Academic Press, Amsterdam.

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

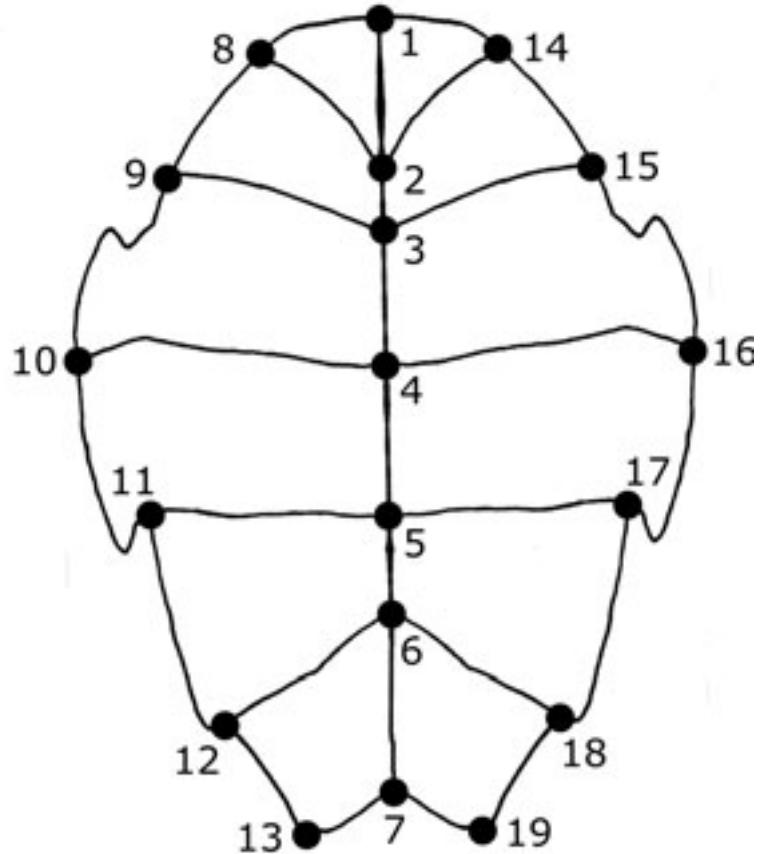


Figure 1: Depiction of general plastral shape of *E. marmorata* and position of the 19 landmark used in this study. Anterior is towards the top of the figure.

		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).

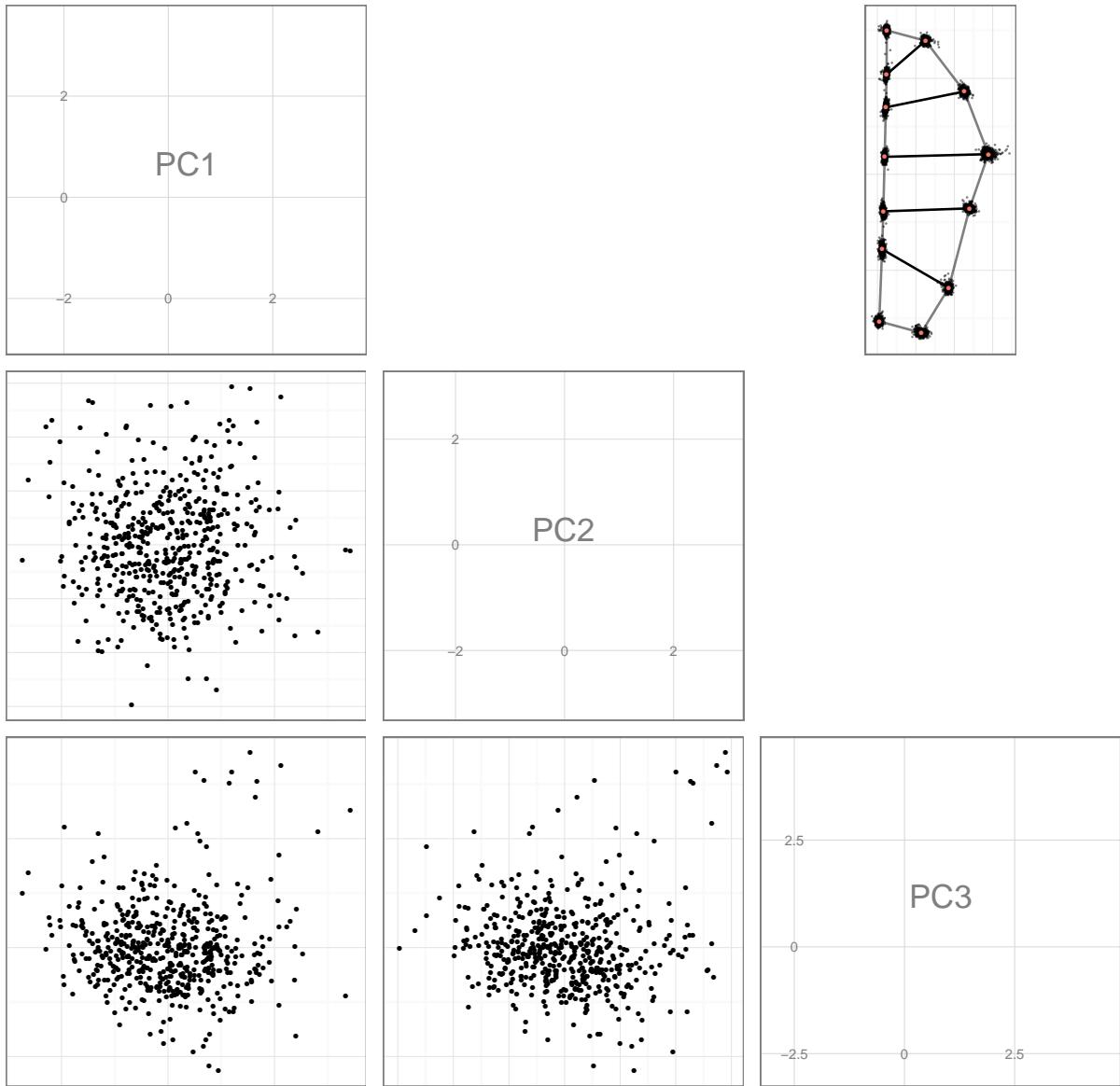


Figure 2: Results from PCA of the Procrustes superimposed “half” plastral landmarks. Depicted here are the for three PCs (lower triangle) and the mean shape with observed variance around each point (upper right). The first three PCs account for total 45.2924805932624% of the variance in plastral shape.

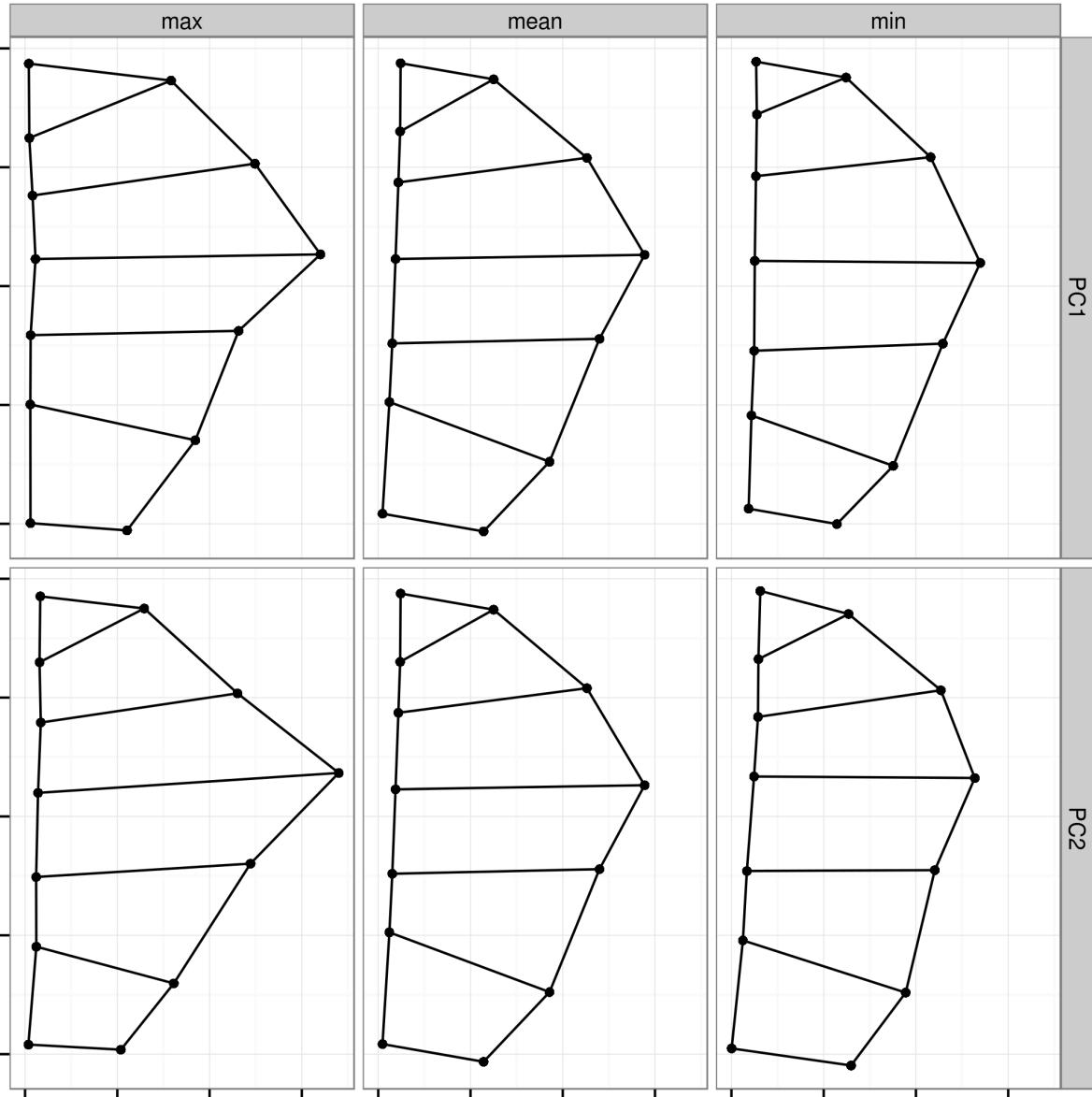


Figure 3: Landmark variation along the first two PCs of the Procrustes superimposed “half” plastral landmarks. The first row corresponds to variation along the first PC, while the second row corresponds to the second PC. The left most column represents the observation with the highest eigenscore along 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.

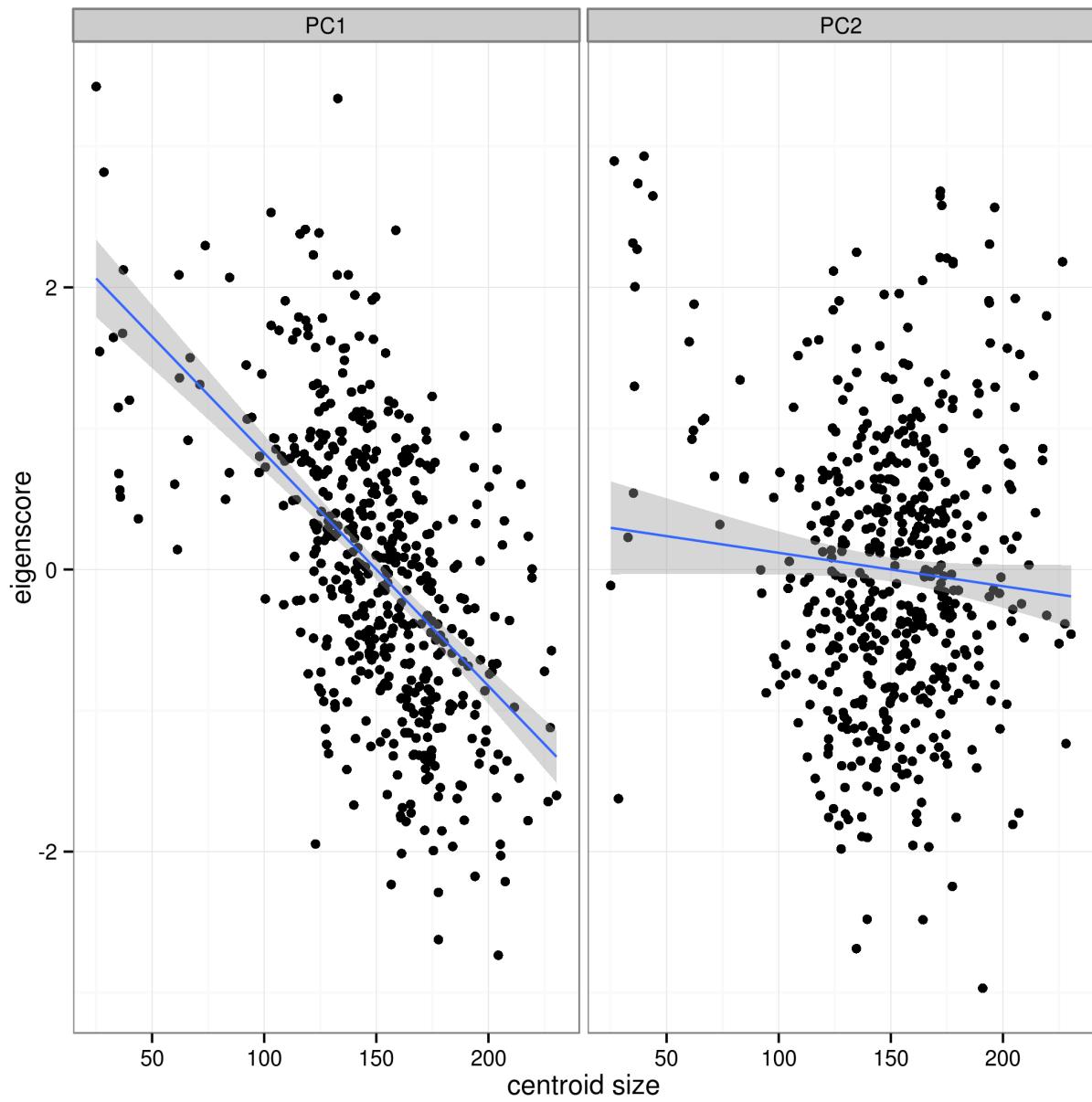


Figure 4: Comparison of the centroid size of each observation to the eigenscore of that observation from the PCA of plastral shape. The first two PCs are illustrated from right to left. The regression line is fit via ordinary least squares and is presented to illustrate the difference in correlation 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.

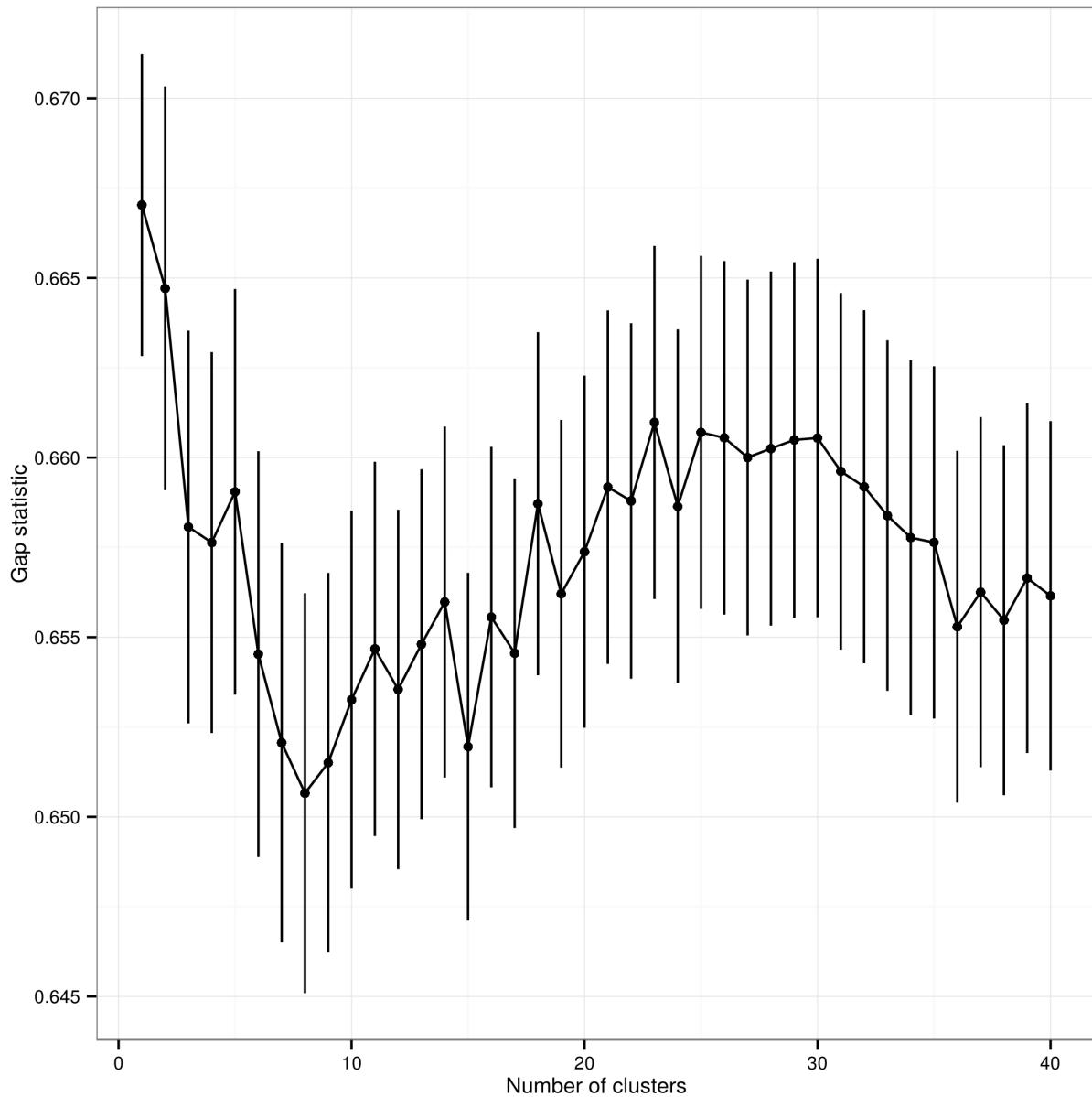


Figure 5: Gap statistic values for PAM clustering results for the ρ dissimilarity matrix of plastron shape. Error bars are standard errors estimated via 500 bootstrap samples.

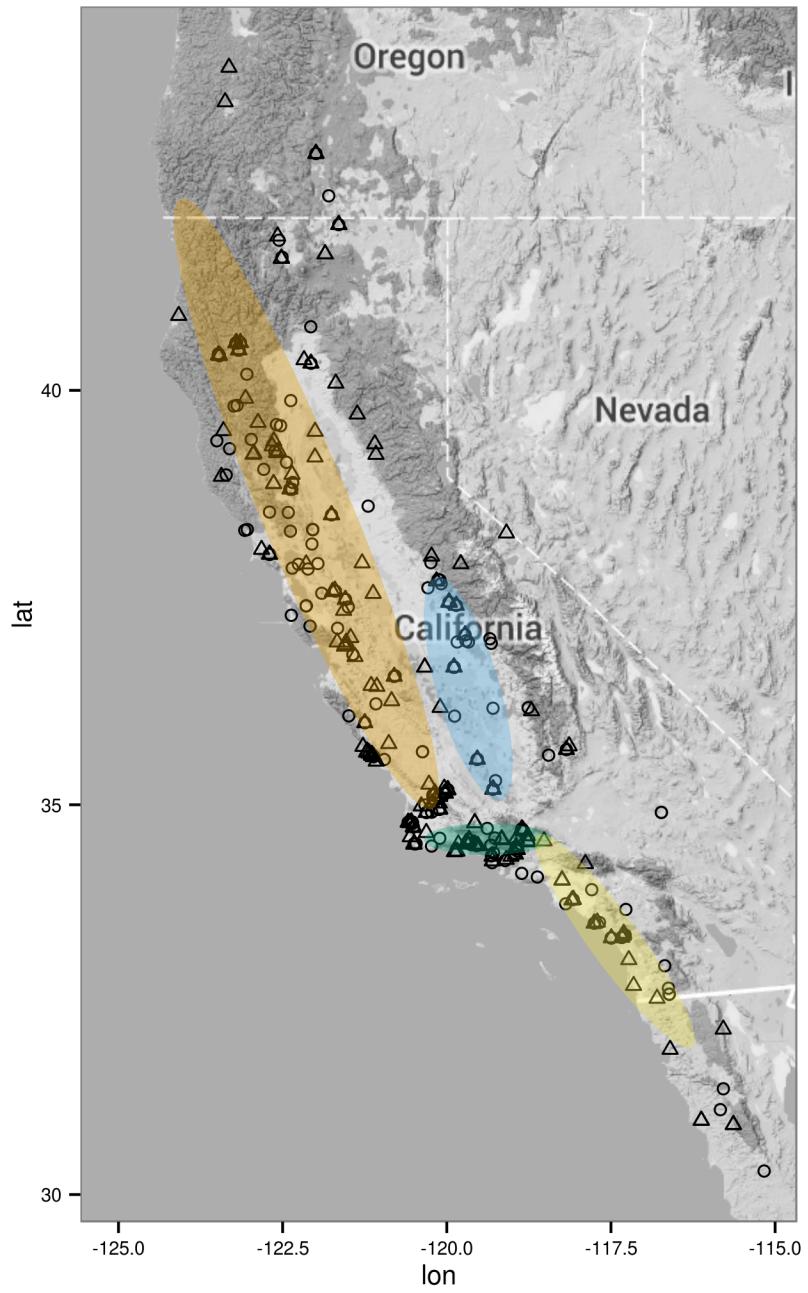


Figure 6: Clustering solution for PAM with two medioids for the entire set of observed *E. marmorata*. Clustering was based entirely on the ρ dissimilarity matrix of “half” plastral landmark configurations following Procrustes superimposition. Point shapes correspond to the two clusters while the colored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

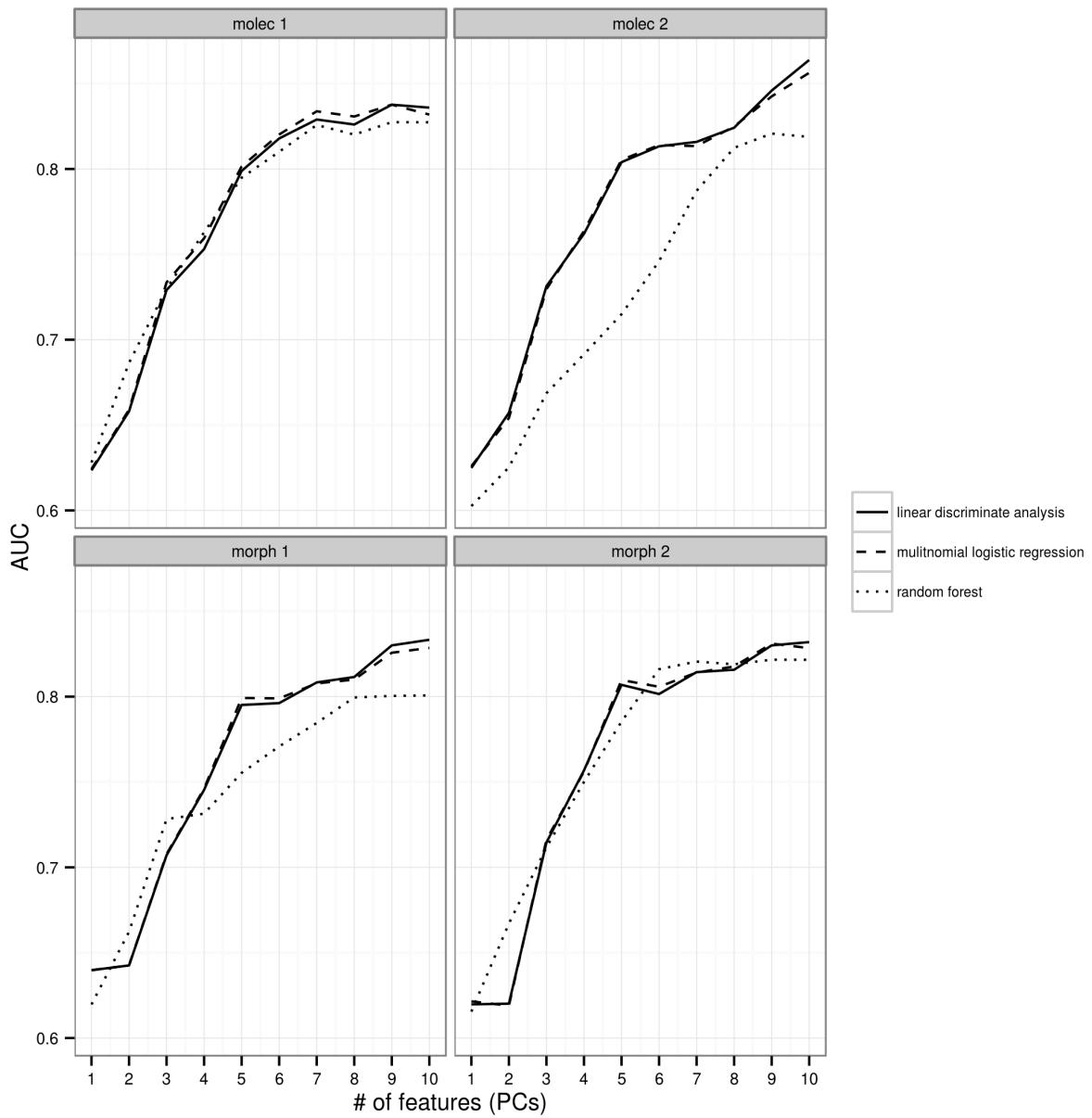


Figure 7: Effect of increasing the number of PCs as features, or predictors, of classification of plastra for all four classification schemes. As the total number of features increase, AUC increases until eventually leveling off. LDA-based classification, multinomial logistic regression and random forest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

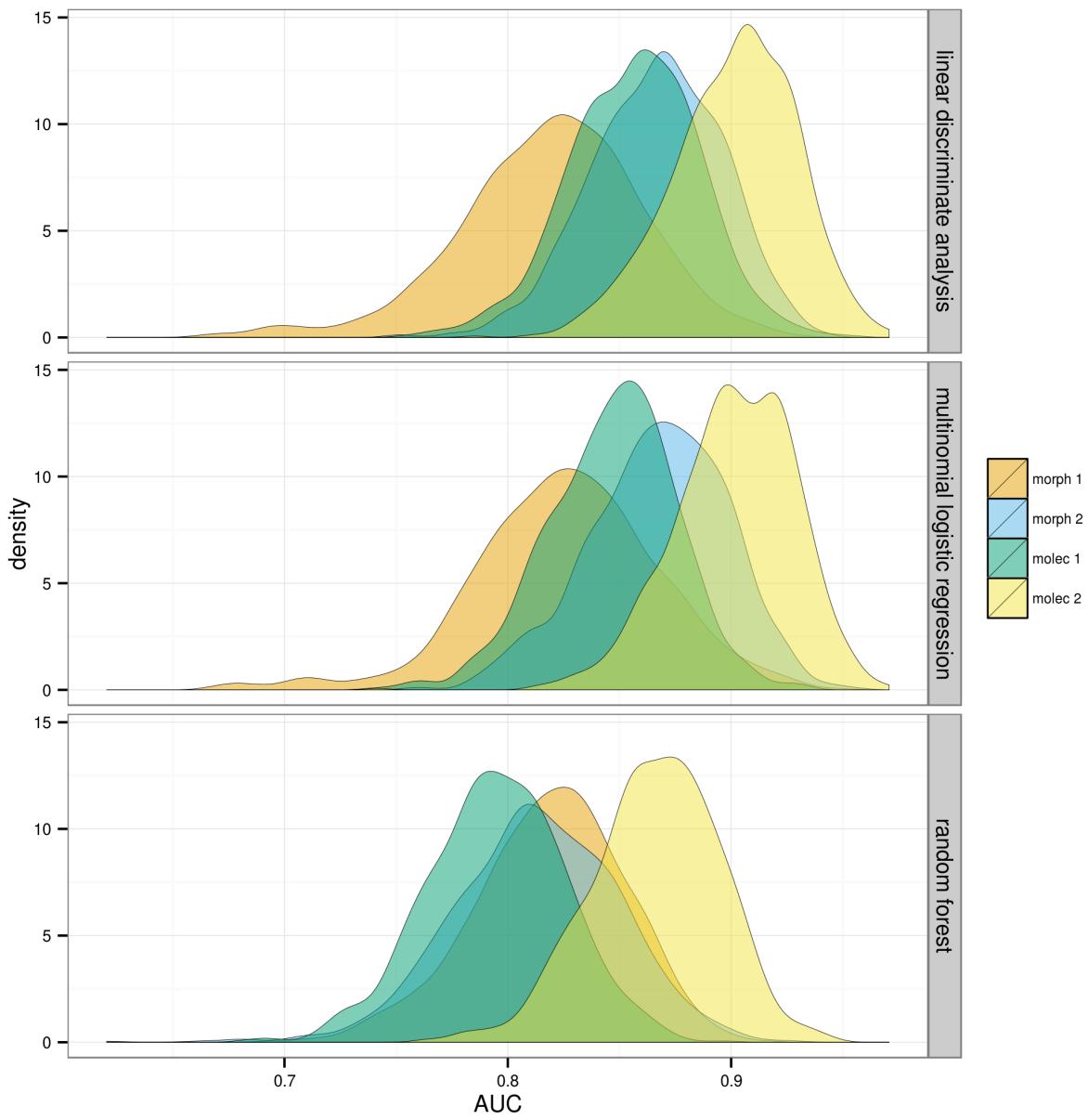


Figure 8: Density estimates of AUC values of predictions of the testing dataset of plastra from 1000 bootstrap resamples. The top facet corresponds to values using the best LDA-based classifications of the eigenscores of shape, as chosen by maximum AUC. The middle facet corresponds to values using the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

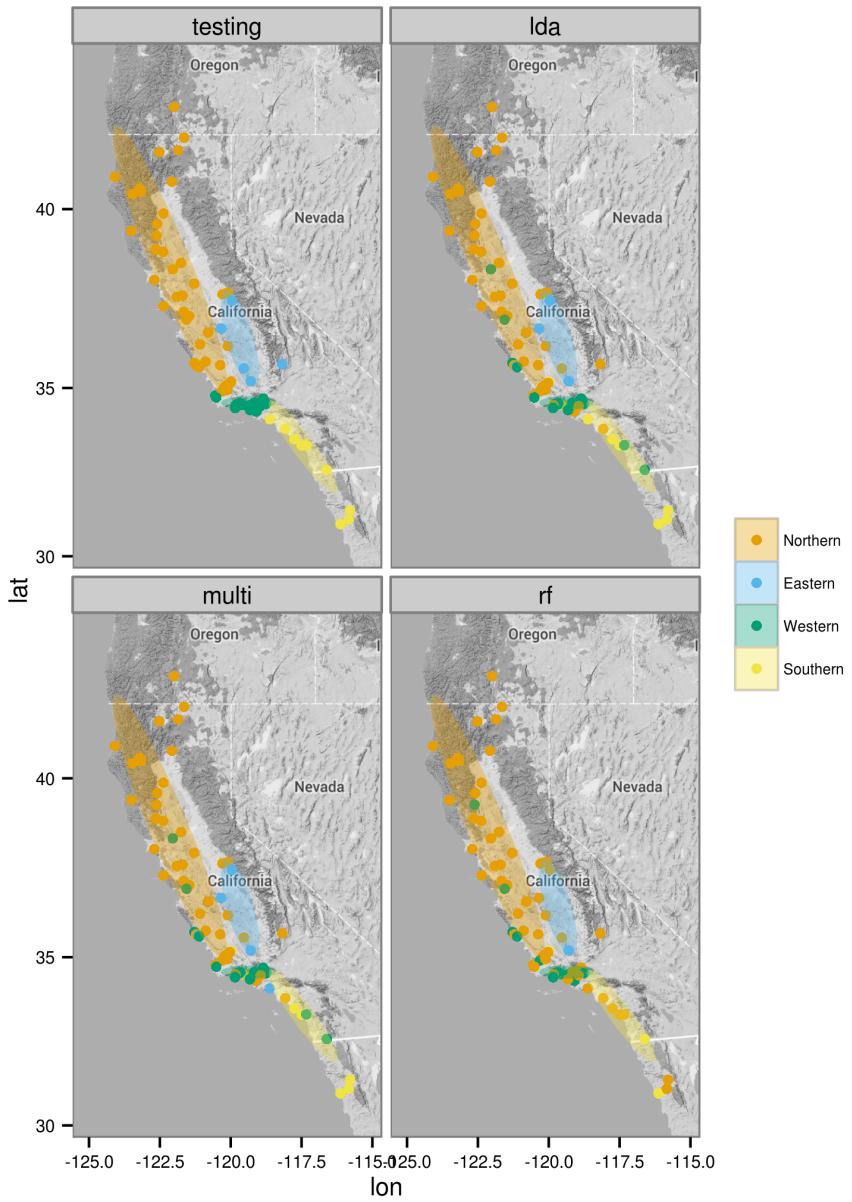


Figure 9: Comparison between reference classification of testing data set (upper left) and the estimated classifications based on the selected LDA-based classification (lda, upper right), multinomial logistic regression (multi, lower left) and random forest models (rf, lower right). Classification corresponds 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.

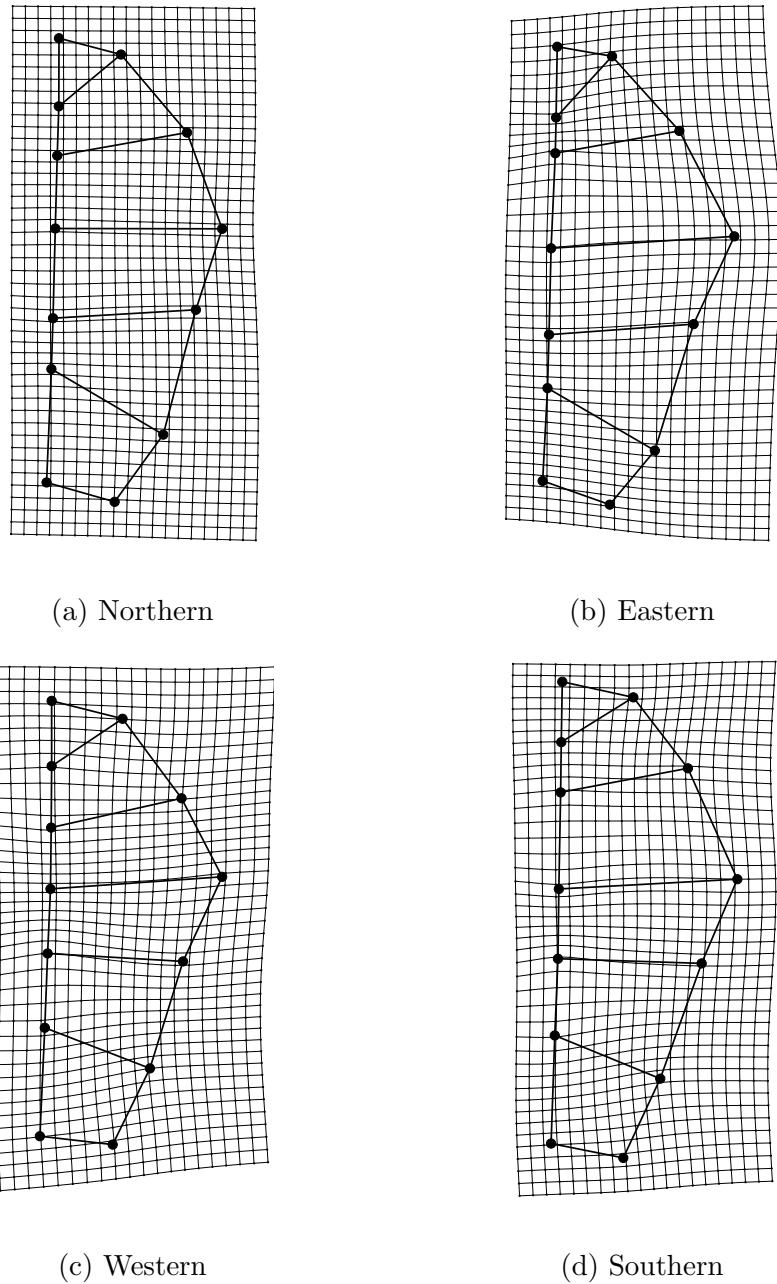


Figure 10: Thin-plate splines for each of the four classes from the best classification hypothesis based on the generalization results (Fig. 8). The four different classes are labeled according to the biogeographic groups as depicted in figure 9. The deformations are depicted with 2x magnification from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

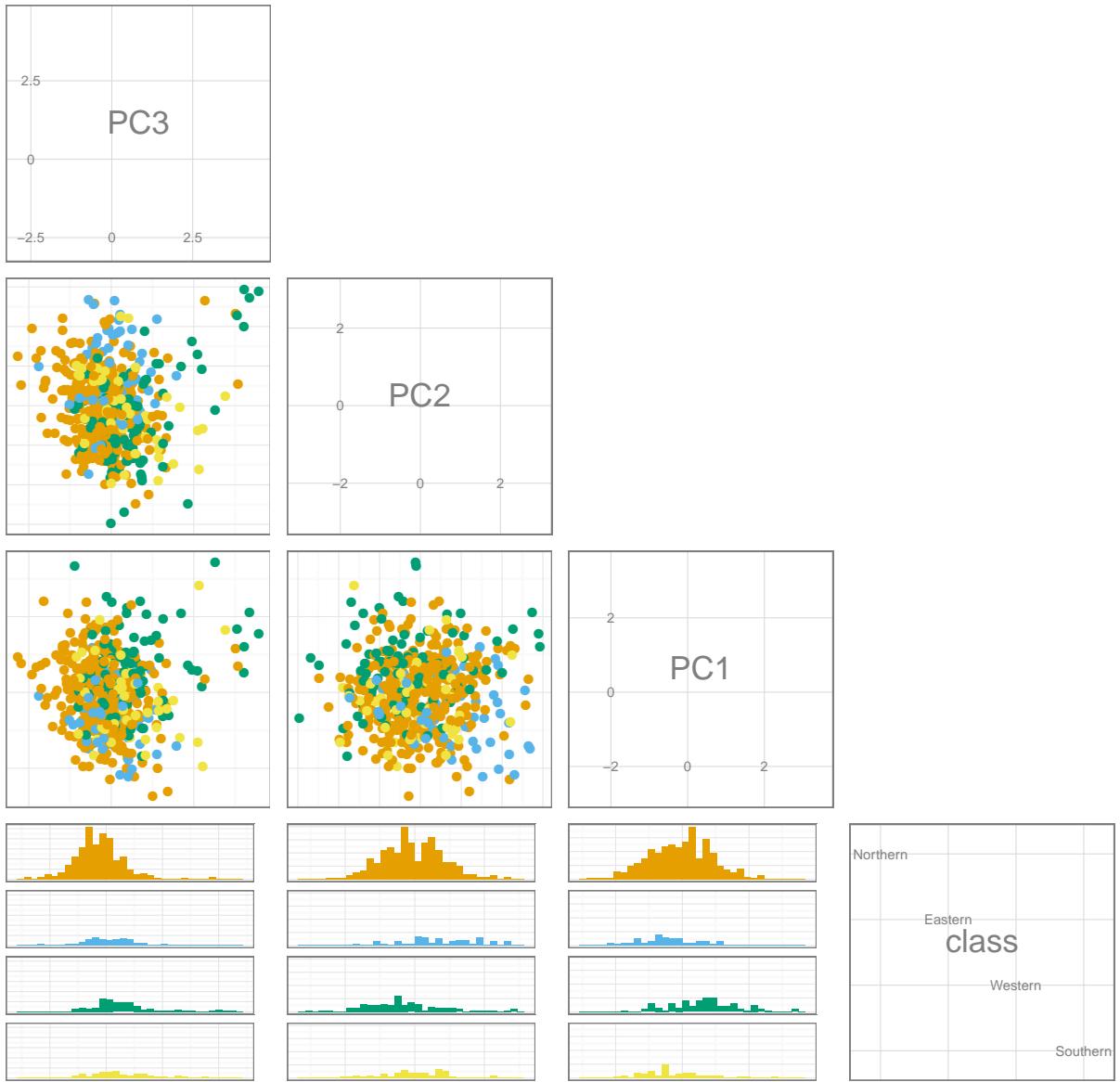


Figure 11: Pairs plot of the first three most important variables of the optimal random forest model of turtle plastral shape. The variables descend in importance from the upper left to the lower right. The observations are colored as in figure 9. The bottom row are histograms of classification occurrences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

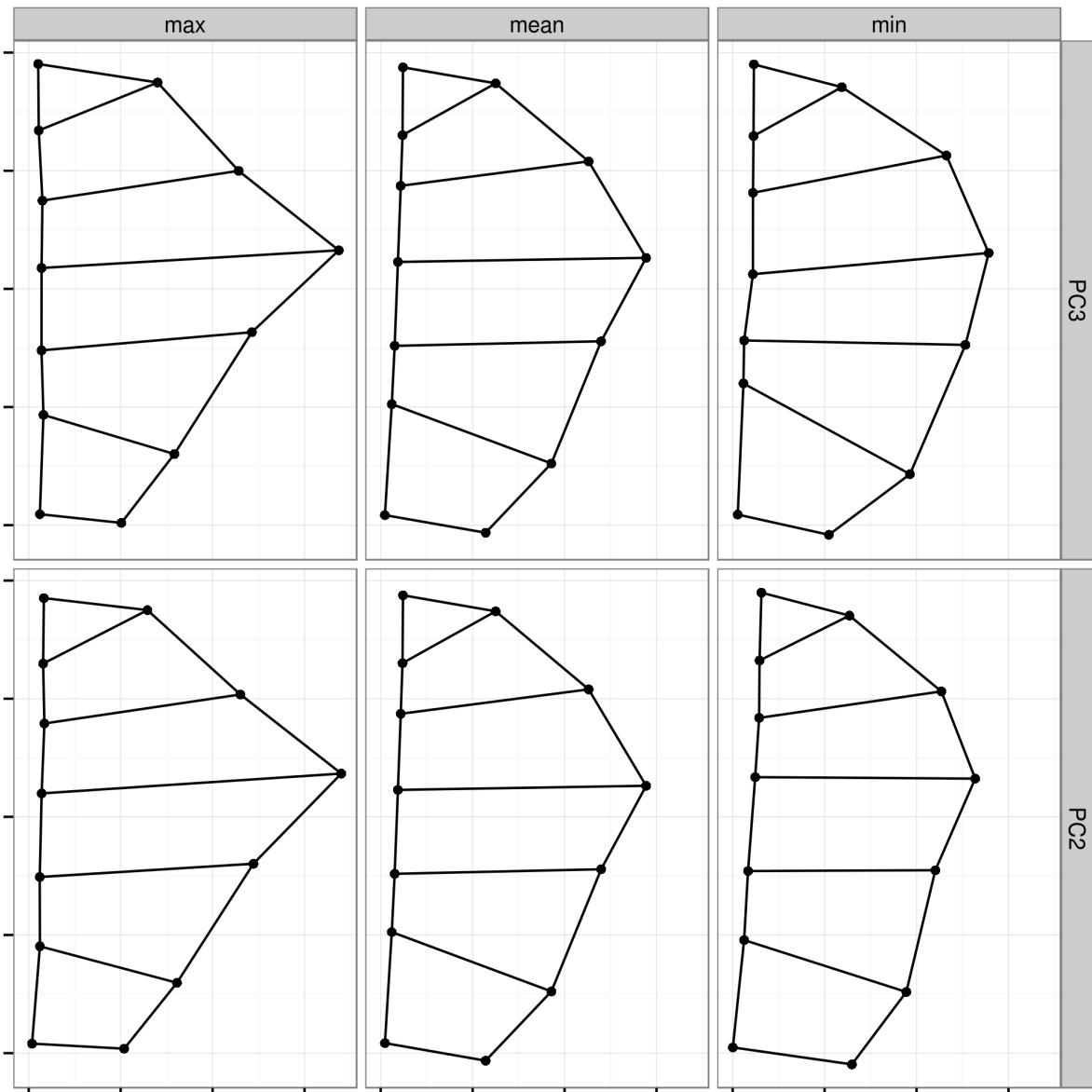


Figure 12: Landmark variation along the two most important features (PCs) based on the final random forest model. The first row corresponds to the third PC and the second corresponds to the second PC. Landmark configurations are minimum observed on that PC, mean shape, and maximum observed on that PC.

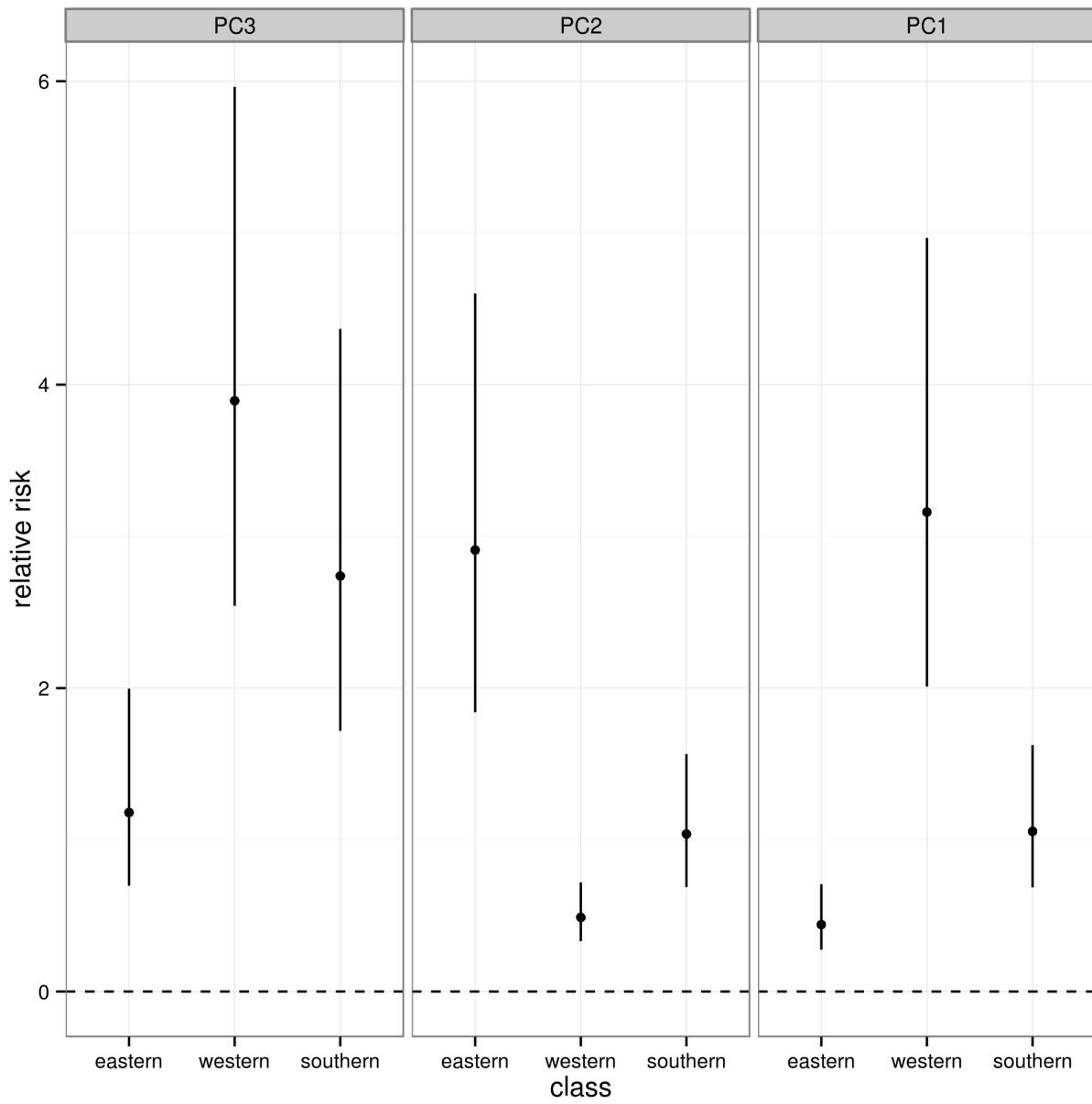


Figure 13: Forest plot of the relative risk, with 95% confidence intervals, of classifying a give specimen based on the first three most important variables according to the random forest model. Relative risk values are calculated from the coefficients of the multinomial logistic regression model. All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

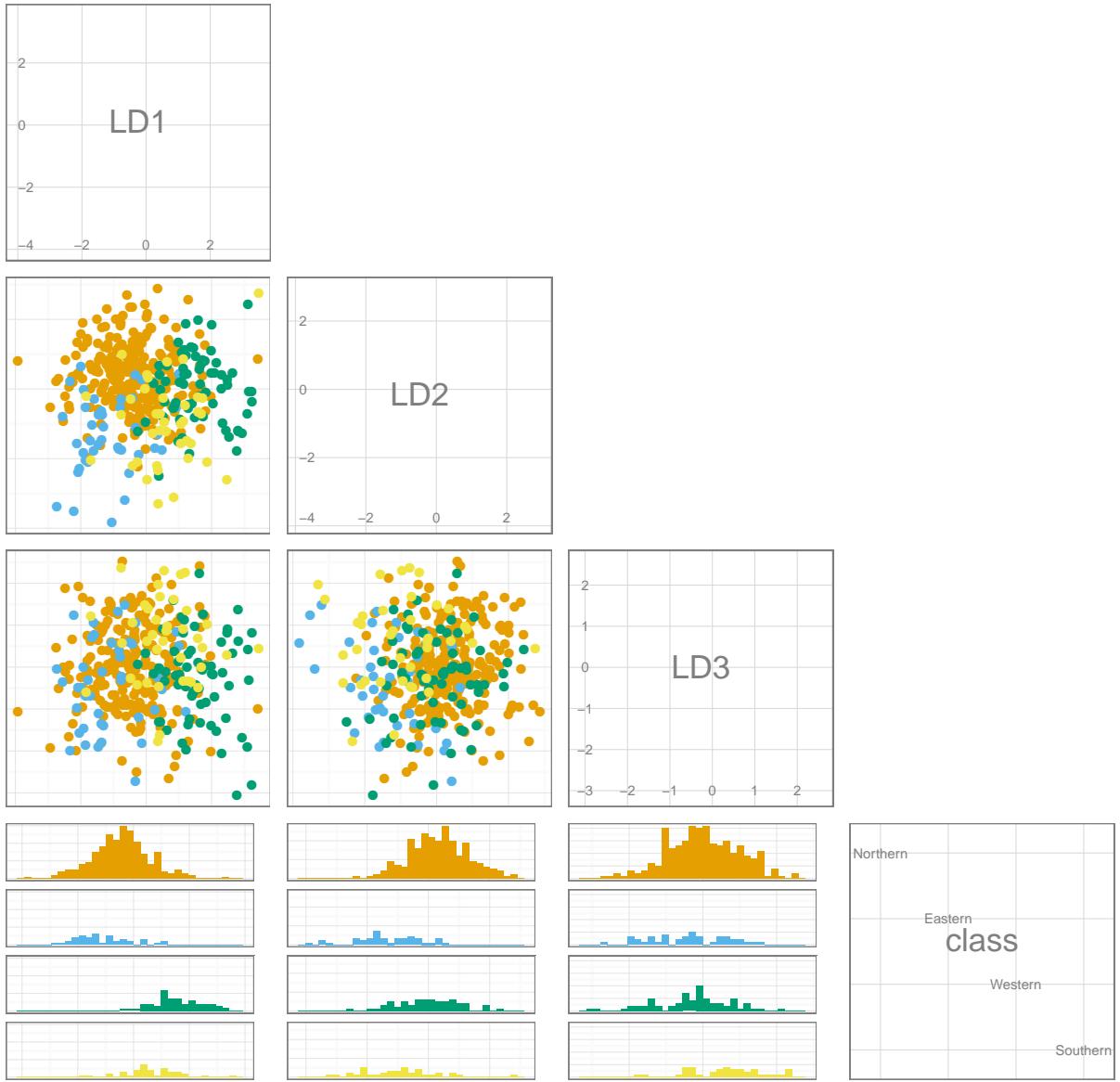


Figure 14: Pairs plots of the three discriminant axes from the linear discriminant analysis of the eigenscores from the first 10 PCs of plastral shape. The observations figured are from the training data set used for all models for the second molecular classification hypothesis based on Spinks and Shaffer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