Committee on Evolutionary Biology
University of Chicago
1025 E. 57th Street
Culver Hall 402
Chicago, IL 60637
psmits@uchicago.edu

February 29, 2016

Editor
Systematic Biology

Dear Editor,

Please find enclosed the manuscript entitled "How cryptic is cryptic diversity? Machine learning approaches to classifying morphological variation in the Pacific Pond Turtle (*Emys marmorata*)" which we are submitting for consideration at the journal *Systematic Biology*. This manuscript is authored by myself (Peter D Smits), Kenneth D Angielczyk, James F Parham, and Bryan Stuart. This paper was previously submitted to this journal with manuscript ID USYB-2015-207.

We believe our analysis and results are of interest to the community as we present a non-standard approach for evaluating the intensity of cryptic diversity and the efficacy of different classification schemes.

Morphological and molecular attempts to classify the turtle species *Emys marmorata* have produced three different potential sub-species or species divisions. In this study we compared these hypotheses of how to subdivide *Emys marmorata* using multiple machine learning approaches. We analyzed over 500 specimens of this species from over its entire geographic range. Additionally, instead of classification accuracy, we use a metric of the ratio of true-positive and false-positive rates for both in-sample and out-of-sample model performance. Finally, in order to test whether our approaches were valid, we compared these results to those from two other independent datasets: the subspecies of *Trachemys stripta* and seven morphologically distinct turtle species within the same subfamily.

Our results demonstrate, that while our approaches can very precisely identify morphologically distinct groupings, the morphological diversity of the *Emy marmorata* species "complex" is inconsistent with any of the proposed classification schemes. We propose that these results may be due to the possibility that genetic differentiation is not associated with plastron shape variation *below* the species level and/or that local selective pressures (e.g., from hydrological regime) overwhelm morphological differentiation. Additionally, a reconsideration of the methods used to delimit *E. "pallida,"* the lack of barriers to gene flow, the strong evidence for widespread admixture between lineages, and the fact that plastron shape can be used to differentiate other emydine species suggest that its lack of

diagnosability most likely reflects the non-distinctiveness of this proposed taxon.

The reviews and comments from the Editor, Associate Editor, an anonymous reviewer, and Julien Claude were very helpful and we believe the manuscript has been much improved for them. Our responses to those reviewers follow.

Thank you for considering our work. Please send all correspondence regarding this manuscript to me via my email address (psmits@uchicago.edu).

encl: Article, figures, tables.

Editor

My main concern first struck me at Line 147, which I think is the first mention of plastron shape. At that point, I realized it wasnt clear to me what aspects of shell shape were used in previous studies. Was plastron shape considered by Holland or Seeliger? Basically, Id just like to know more about how the data used in this study were used in previous studies (if at all). Why use plastron shape rather than other features? Why not the triangular inguinal scales instead (or as well)? Without a clear connection to previous studies, the relevance of the plastron data for this question is not obvious. I presume that it is and Im just ignorant, but if we accept this for publication, I wont be the only ignorant reader, so this should be made very clear.

1) Line 41: "idetifying", "endagered" - typos
2) Line 73: principle should be "principal" here.
3) Line 123: Should be "Holland (1992)"
4) Line 127: "souther" - typo
5) Line 166: The binning schemes have to be clearly described. They are called "Mito 1", "Mito 2", etc., in Figure 4. What are these? Reviewer 1 points this out as well.
6) Line 216: "The results of this analysis produces" - subject-verb agreement issue here
7) Line 245: "the maximum number of features were included" - subject-verb agreement
8) Line 246: Hyphenate "lowest-ranked feature"
9) Line 247: Delete comma after "feature".
10) Line 251: This is not an appropriate place to define the acronym "AUC", because you have already used it a few times. You need to introduce it earlier (reviewer 1 agrees).
11) Line 270: "out of sample" should be "out-of-sample"
12) Line 274: "effected" should be "affected" here
13) Line 304: "with out" should be "without"

Corrections made where appropriate.

14) Lines 299-305: As you note, figure 2 shows that the solution with two groupings has the greatest mean gap statistic, so I think its misleading to cite Fig. 2 when asserting that "Comparison of gap statistic values from PAM clustering show that the optimal, minimal number of clusters is most likely one". I understand the argument about the lack of a geographic signal, but unless I am severely misunderstanding something (always possible), Fig. 2, by itself, tells me two groupings is best. This needs to be clarified.

15) Line 308: Should be "best-performing models"
16) Line 309: "as possible 6)" – ???
17) Line 328: Delete comma after "observed" here.
18) Line 329-330: "AUC valuesor estimating testing data set membership, does not indicate" - subject-verb agreement issue here
19) Line 341: No hyphen needed on "morphologically distinct species"
20) Line 342: "of the both the" - typo

Corrections made where appropriate.

21) Line 343: What are "ROC scores"? This acronym has not been introduced, as far as I can tell.

The methods section and text in general have been adjusted to introduce "ROC", "AUC", and similar terms earlier.

22) Line 344: "which is contrast to" - Poor wording here, or something.

Correction made.

23) Line 350: Again, I disagree with this; it is not what Figure 2 shows. Now, the geographical argument may be a good one, but it is accessory and is being invoked a posteriori to reject two groupings despite what Figure 2 shows.

24) Line 360: Again, why use plastron shape? If it was used by others to delineate two (or more) species within this group, o.k., but right now that is not clear.

25) Line 373: "Nevertheless, it is important to note that plastron shape is an extremely effective method for differentiating members of the other seven species we investigated" - Doesnt this sort of run counter to the statement in line 364 ("Plastron shape does not seem to preserve a strong phylogenetic signal at the interspecific level in emydine turtles")? Or is it that plastron is fine for classification even though it lacks strong phylogenetic signal?

26) Table 2 caption: "asterix" should be "asterisk".
27) Line 379: "iin" typo
28) Line 401: "morphologyand" - typo

29) Line 415: No need for a hyphen in "morphologically diagnosable"
30) Line 416: "that out" should be "that our"

Corrections made where appropriate.

31) Line 419: As I often tell my student, you should use quotes sparingly. This is the third quote from Carstens et al. (2013), and that just strikes me as somewhat lazy writing. I think simply paraphrasing Carstens et al. (2013) would be adequate in all three cases. Carstens is a fine writer, but I dont feel that these quotes are particularly amazing crystallizations of a crucial concept that need to be repeated verbatim.

Quotation removed and text re-enforced.

32) Line 432: "The comparisons with other emydine species, suggests" - Watch subject-verb agreement, and delete comma after species.
33) Line 437: No need to capitalize "pallida" here.

Corrections made where appropriate.

Anonymous reviewer # 1

I have two primary criticisms of the paper. First, the presentation of some material is not always conducive to a readers full understanding of this study. Aside from numerous typos or grammatical issues (see below), in some cases the material presented is incomplete, or not given enough context for the reader to understand. The major issues for me were as follows: the Results section contains multiple numbers in parentheses - are these supposed to refer to figure numbers? It would seem like they are, and that the authors just neglected to put "Fig." in front, but they dont show up in the proper order, so I cant be sure. "AUC" first shows up in line 212 but is not defined until 251; it should be defined at its first appearance. Table 1 provides abbreviations for a series of "Schemes", but I cant find in the text where such schemes are explicitly described. The authors should provide this information ("Morph 1" corresponds to X classification in Y reference, for example).

Figure 3 purports to show the geographic distribution of groups arrived at for a certain solution; however, Im pretty sure that I dont see 354 points, which would correspond to all of the specimens in the data set. And this made me question my earlier assumption that the input for the grouping methods consisted of data for individual turtles. If the input matrices consisted of something else, like means for a given locality, the authors should state this. And it does make

sense that the clustering should be performed on mean values; otherwise, it seems like some of the first clusters to come out, at least of the unsupervised approaches, would just separate males and females. So the authors should describe how they handled the issue of sexual dimorphism in plastron shape (well-known in many turtle species) in E. marmorata.

My second criticism has to do with the choice of data set for the authors "clear-cut example". In one view, this study doesnt need a "clear-cut" example for comparison at all. This is the case if the question of the study is: "Is there evidence for morphological differentiation in the plastra of E. marmorata, consistent with some previously-proposed classifications?" But on another view, an appropriate comparator is necessary. This is the case if the question of the study is: "Is there as much evidence in the plastra of E. marmorata for grouping members of this species into multiple subspecies, as there is in other turtle species which have already been grouped into multiple subspecies?". And of course, a comparator is also necessary to answer the question: "Can the methods used in this paper ever produce a taxonomically-appropriate clustering?"
The authors clearly intend the latter for their study. They almost could not have chosen a set of taxa better-suited to successfully finding differences (maybe not at all within a given subfamily of turtles). The species they have chosen differ in pretty much all of the major factors that are known to affect the plastra of turtles: size, plastral kinesis, habitat, and degree of phylogenetic relatedness. Perhaps this is why they chose the taxa that they did, but if so, then Im not sure that the results are particularly germane to the rest of the analyses. Of course the methods will produce an appropriate grouping for that set of taxa - I cannot imagine any published methods not doing so. So these results dont really provide a useful comparator with the E. marmorata groupings; they just show that the methods used arent utterly abysmal.
Instead, I think that the authors need to be able to answer the second question that I proposed above. They need to be able to show that these methods can successfully group turtle plastra which show the same magnitude of variation that is expected in subspecies or recently-diverged species or cryptic species in related groups. Grouping species within Emydinae wont provide this. Im not sure what the best comparisons would be. Perhaps Chrysemys subspecies, or C. picta versus C. dorsalis. Perhaps Terrapene species or subspecies. But the authors should pick and justify an appropriate comparison. And if the methods cannot successfully group these taxa either, then the relevance of their failure to group E. marmorata will be put in its proper context.
The paper is otherwise strong. Again, I think the authors did an excel-

lent job of synthesizing results from multiple methods, and the study is important (and, needless to say, appropriate for Systematic Biology). But I think the authors need to clean up their presentation and choose a more appropriate comparison or validation of the relevance of their results.

Minor editorial comments follow. Line numbers are provided, along with suggested corrections:
15 - what is meant by an "appropriate" signal?
41 - "identifying"
59 - "identification of images" (otherwise the sentence asserts that handwritten zip codes are a type of automated imagine identification)
62 - "estimated; they"
127 - "southern"
150 - "the proposed species of Spinks et al."
160 - can you provide numbers for sexes as well?
247 - "feature remained"
274 - "affected"
291 - can you provide numbers of specimens and sexes here as well?
304 - "without"
335 - "sets"
379 - "in"
384 - to be fair, streamlining is probably not relevant for T. ornata, and possibly much less so for G. insculpta and maybe even G. muhlenbergii
391 - "among"
401 - "morphology and"
404 - well, geography, at least, was certainly brought to bear on the classification, yes? Even if life history, morphology, and behavior were not.
416 - "our"
432 - "species suggest"
437 - "pallida"
582 - "2015"
583 - provide issue and pages, or "in press"

Corrections made where appropriate.

Reviewer # 2 Julien Claude

0. non-significant remarks
-l. 41 idetifying -> identifying
-l 189. update your citation in R.
-l. 212. The explanation of AUC is coming late, you should avoid us-

ing abbreviation if the reader do not understand its meaning, at least you should give the meaning of each letter as soon a you are using the abbreviation.
-l 379: iin -> in
-explain which package/functions you have been using for LDA, random forest, and polynomial regression.

Changes made where appropriate. In particular, the explanation of AUC is moved earlier in the methods along with a few other aspects of terminology.

Major remarks:

1. Interrogation concerning the choice of the methods

1.1. The authors are using three supervised techniques but only one unsupervised technique to test or evaluate the clustering pattern in their turtle. Three of these methods (LDA, and PAM) have been used with morphometrics (PAM in Claude 2008: Morphometrics with R), there is a good example of regression trees applied to Kangaroo measurements in Faraway 2006 (Extending the linear model in R). I am not sure for multinomial logistic regression, but see my remark below : it may not be the best

1.2. I am surprised that only PAM was used as unique non supervised method (with the use of the gap stat to select the number of group). This is certainly one of the simplest method with the kmeans, but there are at least two other methodologies that the authors could have selected because they have considerable advantages on PAM: the one of geneland (see Guillot et al., 2012, Systematic Biology) that has the advantage to consider geography or not and may work with the data of the authors (since they are spatialized; and Mclust (Fraley and Rafteri, 2002) which allows different model of covariance and variances within groups. l. 46, the work of Guillot et al. 2012 could be cited for species delimitation. l. 198, the use of PAM is not new in morphometrics, I have been using it in 2008 (maybe you should cite this example).

PCA is technically an unsupervised method, it just isn't a clustering method.

1.3. l. 77 you say "Most previous morphometric studies did not assess which amongst a set of alternative classification hypotheses was optimal."; It is maybe true, but some have done the job, and I gave some possibilities in R and packages in Claude 2008. Here again you are omitting Guillot et al., 2012.

1.4- why using only 1 unsupervised techniques, and three supervised techniques (on which two present certain weaknesses according to your data) ?

As far as I understand, the paper is not intented to compare the behaviour of supervised techniques, but it is using several ones... By looking at the results of bootstrap for AUC, it seems that LDA is the method that is clearly the less sensitive to sampling artefact (it is really clear for the trial with the 7 species), and in average it seems LDA reach a better score than others. I would consider only this in the ms, keeping the two other for appendices. It is rather easy to understand why: in the LDA the metric used is the mahalanobis distance : this allows somehow to disregard variation that can be due to intragroup variation. This does not seem to be the case for the two other methods. As stated by the authors, PCs are considered as independant features by forest models, and this may not be wonderful... indeed PC scores depend on the ordination of individuals in the shape space and therefore of the sampling (they can be influenced by allometry, grouping structures, etc) and the superimposition... they are not really independant variables and not measuring intergroup variation (even if intergroup variation is well structured).

-> use mclust and or geneland instead or with PAM, put in appendix the methods/results that may have behaviroral problems

Principal components are considered independent features in all models where they are used as predictors of any variable; this happens in random forest, multinomial logistic regression, the linear formulation of LDA, all of them. If random forest has problems because of this fact, all methods have problems because of this fact.

Also PCs are, by definition, independent descriptors of the observed variation. Principal component regression is an extremely common technique for dimenstional reduction and is essentially what we're doing here. Our large multivariate dataset is not ammenable to regression, so we use PCA to get independent/uncorrelated predictors of class. This approach is very kosher see *Elements of Statistical Learning* by Tibshiri et al. for evidence.

We are not perporting that species evolve along PC axes or that groups are defined by axes; we just state that groups may represent sections of morphospace defined by regions of each PC. The PCs are just means of reducing dimensions and putting everything in a nice coordinate system; these coordinates then help us identify regions of morphospace that may correspond to a group of interest.

Additionally, we do not agree with your suggestion to move the other methods to the appendix. That undermines the entire point of this paper.

2. Interrogation regarding the nature of the data.

2.1.- l 188. explain whether you work on raw procrustes coordinates or on tangent coordinates. Explain why you are using PCs instead of

tangent coordinates.

2.2.- can the inconclusive results come from the fact that measurement variation is great by comparison to interindividual variation ?Nowhere you estimate measurement error by comparion to interindividual variation in your data set (see Yezerinac et al., 1992 Syst. Bioloy; Claude et al., 2003 BJLS). How many percents of shape variation is it ? from my experience, measuring landmark on plastron in 2 D can produce an important error rate; is it your case ? are we able to recognize individuals ? I guess that the methods you are using are certainly sensitive to measurement error; but this is not estimated in your ms.

l. 356. This can be also because intragroup variation is high by comparison to intergroup variation for the set of landmarks you selected (allometry, developmental noise, etc. . .). Maybe shape sex dimorphism is stronger than intergroup difference. Why not focusing only on males or females (in the intro you say that sex shape dimorphism has not been evaluated)?

2.3. - Why do you symmetrise your individuals ?Actually asymmetry could not be geographically structured, I do not see clearly what kind of problems you could have with degrees of freedom. In symmetrising as you do, I hardly understand how you can compute the Riemanian distance ? does that mean that your individuals were superimposed on the basis of half the paired coordinates ? In doing so; are you considering that the landmarks on the symmetry axis are not constrained by the symmetry ? How much fluctuating asymetry do you have by comparison to measurement error and interindividual variation ?

-> estimate the importance of asymmetry, digitization error, and sex dimorphism

3. Number of PCs retained in LDA and other supervised techniques. I really do not understand why the authors restricted the number of PCs to 6. In exploring the results obtained, it seems that the best AUC is usually reached for the maximal number of PC retained; I guess it would be better if you increase the number of PCs, why not using all the non Null PCs (you can follow Chiari & Claude approach 2012 -Compte rendus Biologie), there are 354 observations, 13 "operating" landmarks (-> at the end you have less that 26 non null PCs, which is considerably less than the number of observation - the maximal number of expected groups). In general, you can reasonnably use all of the variables rather than to reduce the dimensionality of your dataset. remark than in PAM, you are using variation in the whole shape space since you input the shape distance between two observations.

-> use the selection of PCs that give the best cross validated classification rather than a small set

All these methods involve the estimation of minimum 1 parameter per PC used as a predictor of class. We restricted the number of PCs because we wanted a minimum sample size per PC so that our models were not overfit and bias towards the in-sample data; this would make out-of-sample predictions much weaker. However, following your suggestion, we have increased the number of features used per model. We've also included centroid size as a predictor of class.

4. Group numbers and PAM

Pam and marmorata study. according to fig 2., l. 299, the best clustering is for 2, not 1. You can explain later why 1 should be prefered on 2. 2 may come from the presence of a strong sex shape dimorphism but this is not evaluated.I do not understand how the vertical bars were obtained for the gap stat. The resampling should be explained here. But if this resampling correspond to the sampling error for the gap stat: note that according to your results: you can hesitate between 1, 2, 3, 4, 5 and 8 according to your plot, these partitions do not significanly differ amont each others. This part is rather inconclusive; rather than it clearly conclude for 1 group.

5. differences between well and wrongly assigned groups.

l 276 to 282; I do not understand why the authors use a complex strategy to estimate differences between these groups, and the way it is run is not completely clear. Why not using Manova and look at the Pillai values and stat ?

-> Better explain that part or make it simpler.

This section has been removed because it did not contribute to the over all text.

6. Problems in considering that PCs are "independant" shape variables.

Note that grouping structure should not necessarily follow the direction of eigenvectors. The direction of eigenvectors depends on many things and is not necessarily determined by the groups. The direction may depends on allometries, or different number of observations within every groups and eventually different sex ratio within groups. Allometry is not necessarily orthogonal to group differentiation (because groups may differ in their allometry patterns - see for instance Claude, 2013-). The presence of allometry is not evaluated in your data set; the importance of shape sex dimorphism is not evaluated.

-> avoid interpreting that a differentiation on PC 1 is of greatest importance than elsewhere. PC1 may represent just 15% of total shape variation (there is still 85% of remaing variation, and group structure is rarely colinear to PCs)

-> investigate the role of allometry

7. why not including size in the analyses ?

You are not including size (even if you are using only adults) in your supervised learning method, but you recognize that postnatal ontogeny can be important. However, incorporating size in your analyses may increase the differentiation between groups (in LDA, considering the fact that size can covary with shape; may influence ordination of individuals greatly). Allometry may be present in your data set and may obscure group differentiation (see Chaval et al., Mammalia, 2015).

-> perform the same analysis with size

Size has now been included as a predictor of class in all analyses.

8. intro and background too long. . . , machine learning section in methods too long

This part can be reduced, the readership of systematic biology is aware about clustering and partitioning techniques. the readership is also familiar with geometric morphometrics and problems regarding species delimitation. The readership knows also how a LDA works.

Adjustments made as requested.

9. omitted references. . .

I was surprised that none of the publication that I have authored were mentionned in the text, while they could have been considered. As I have no competive interests with the authors, I thought it rather strange. Among these figure the selection of number of PCs for LDA, the use of unsupervised techniques, and the use of the combination between geography and morphology; which may be important to consider for improving this version.

No harm or foul was meant. Citations have bene updated to reflect your request.