How cryptic is cryptic diversity? Machine learning approaches to fine scale variation in the morphology of *Emys marmorata*.

April 4, 2013

1 Methods

No-free lunch theorem. Try lots of things because we don't understand everything.

1.1 Unsupervised

Underlying structure in data?

1.1.1 Gap-based clustering

Comparison of gap statistic results for partitioning around medoids (PAM) divisive clustering. Confidence intervals are determined via bootstrap. The higher the gap statistic, the better the clustering result. Standard errors of the gap statistic were estimated from 100 resamples.

1.1.2 Evidence Accumulation Clustering

Choosing an optimal number of partitions is hard, which is why gap-based cluster selection was used above. An alternative method is to look at the co-occurrence frequency, that is how frequently any two samples occur in the same partition. Repeating this process over and over again creates the frequency, or "vote" for how the data set should be partitioned and which specimens should be in the same cluster.

EAC was originally devised using k-means clustering, but I've extended it to use PAM clustering instead. The hope is to determine underlying structure in the data given a wide enough partition range and a high enough number of iterations. Dissimilarity based EAC was performed using a range of 1 though 200 possible partitions and based on 10,000 iterations.

1.2 Supervised

How well does data conform to predetermined structure?

- multinomial logistic regression (Fig. 4)
- feed-forward neural networks (Fig. 5)
- random forests (Fig. 6)

2 Preliminary results

2.1 Unsupervised

Comparison of gap statistic over a very wide range of plausible partitions indicates that as the number of partitions increase, there is a marginal increasing gap statistic until approximately 40 after which there is a marginal decrease in gap statistic (Fig. 1). It is notable that the standard errors around the gap statistic values are very large, and the marginal increases in gap statistic with an increased number of partitions may not be important. Additionally, all gap statiic values are within 0.0064 of each other meaning that there is little over all consensus for how many clusters are present when comparing gap statistics.

Dissimilarity based EAC estimated approximately 5 optimal partitions (Table 1).

	tmorph.dac
1	722
2	1
3	1
4	1
5	1

Table 1: Number of specimens assigned to optimal number of partitions as determined by dissimilarity based EAC. Each column corresponds to a different partition, with the number assigned directly below it.

2.2 Supervised

I'm currently holding back on showing the tables because there are a lot of them. I'll show the confusion matrix for the multinomial logicistic regressions because they are easy (Tables 2, 3, 4, 5)

	CCR	marm	pall
CCR	40	14	11
marm	25	74	2
pall	3	1	10

Table 2: Confusion matrix for sh1 classification scheme. Rows correspond to predicted class. Columns correspond to reference class.

	CCR	marm	pall
CCR	44	12	10
marm	17	82	6
pall	2	1	7

Table 3: Confusion matrix for sh2 classification scheme. Rows correspond to predicted class. Columns correspond to reference class.

	CCR	marm	pall	SJ
CCR	37	11	9	2
marm	18	59	3	10
pall	7	2	10	0
SJ	1	4	1	7

Table 4: Confusion matrix for sh3 classification scheme. Rows correspond to predicted class. Columns correspond to reference class.

But I can show the comparison of the predictive accuracies (Fig. 3).

In the interest of space/time, I'm only going to display results from one of the selected models from each method for each classification scheme.

3 Future

1. remove juvelines
2. look into multinomial logistic mixed-effects models
3. other unsupervised methods, though that might have to wait for a follow up paper
    - bayesian nonparametrics for categorical data

In general, misclassification seems to be at random. This is interesting for a few reasons. The assignments based on geography do not seem to bias results. All the taxa are extremely similar, though there are differences, hence the 70%

	1	2	3	4
1	107	9	13	13
2	1	8	0	0
3	4	0	8	3
4	4	0	2	8

Table 5: Confusion matrix for spinks classification scheme. Rows correspond to predicted class. Columns correspond to reference class.

accuracy. None of the classification schemes seems necessary better than any of the others, though sh3 is probably the best (4 classes) over all because it is the least "random". I think there is a summary class specific statistic that better explains this (sensitivity? specificity? detection rate?).

4 Miscellaneous affairs

4.1 Evolution 2013

How cryptic is cryptic diversity? Machine learning approaches to fine scale variation in the morphology of *Emys marmorata*.

4.2 Grants

4.2.1 Hinds Fund

Resubmit previous application with the results from this study in the prelim-results section?

Get network for one turtle? This should take long, just requires time.

4.2.2 Paleontological Society

Why not? Need to wait till next year. Should probably join PaleoSoc anyway.

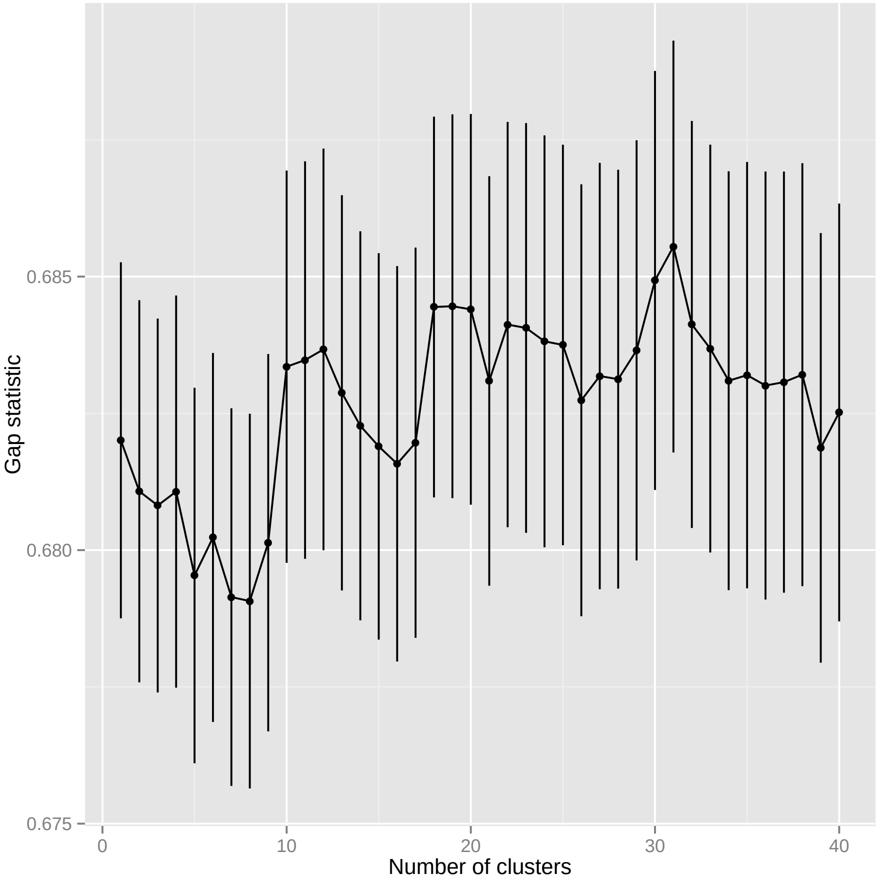


Figure 1: Gap statistic values for multiple PAM-based clustering configurations of the Riemmanian shape distances of the *Emys marmorata* plastra. Higher values indicate greater clustering. Standard errors are estimated from 100 bootstrap resamples.

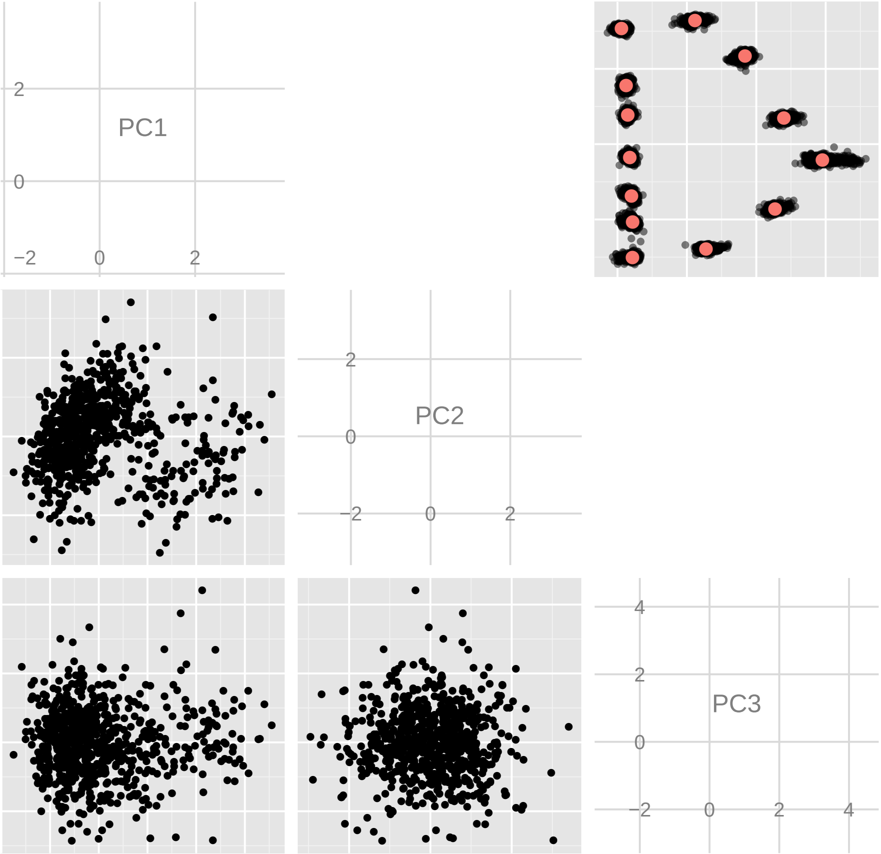


Figure 2: Visualization of PCA of *E. marmorata*. The lower triangle is the pairwise comparison of the first three pricipal components. The upper left corner is the comparison of landmark dispersion for all specimens compared to the mean shape in red.

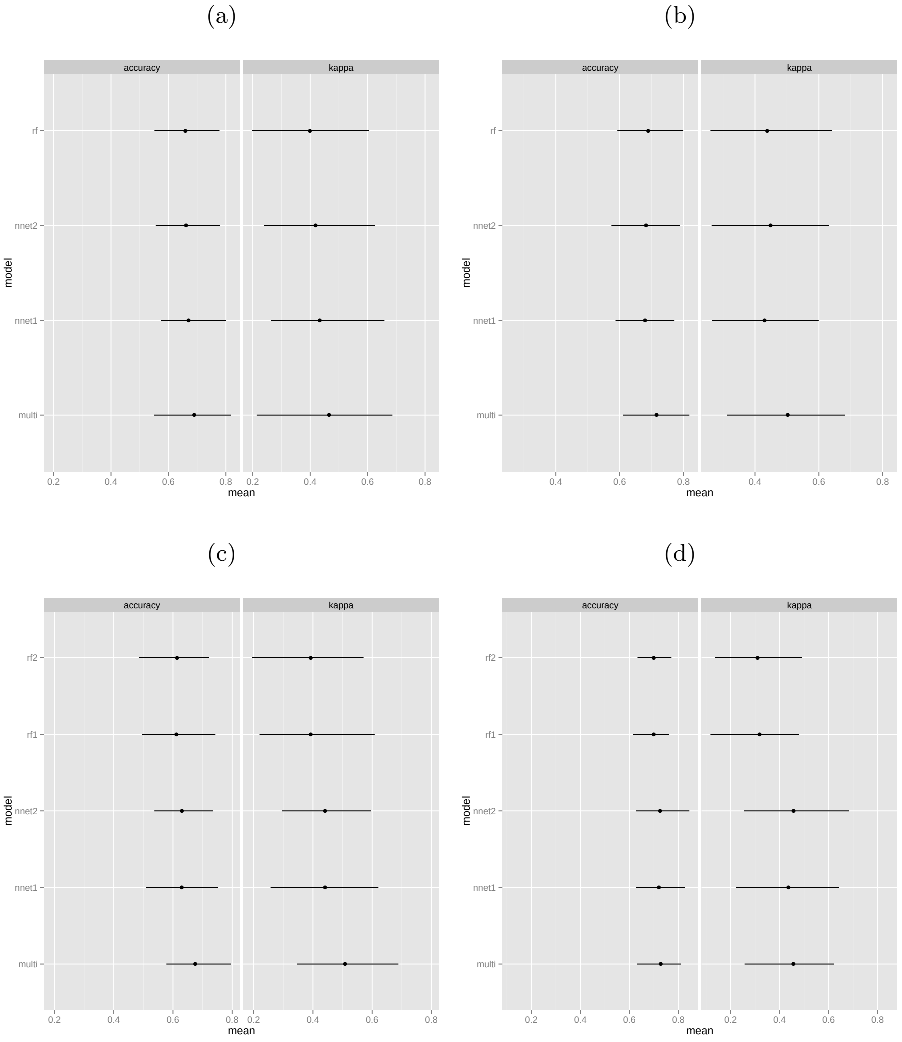
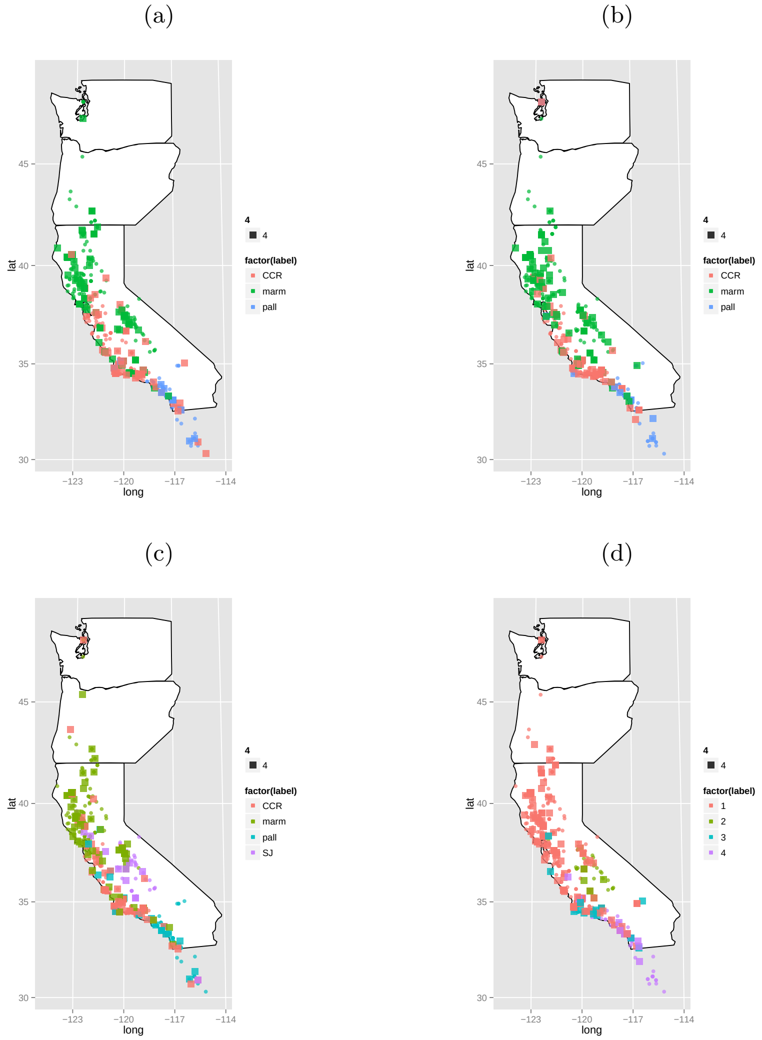


Figure 3: Comparison of resampling distributions of training set accuracy and kappa statistics for the selected models of each classification scheme. 3a: sh1 classification scheme. 3b: sh2 classification scheme. 3c: sh3 classification scheme. 3d: spinks classification scheme.

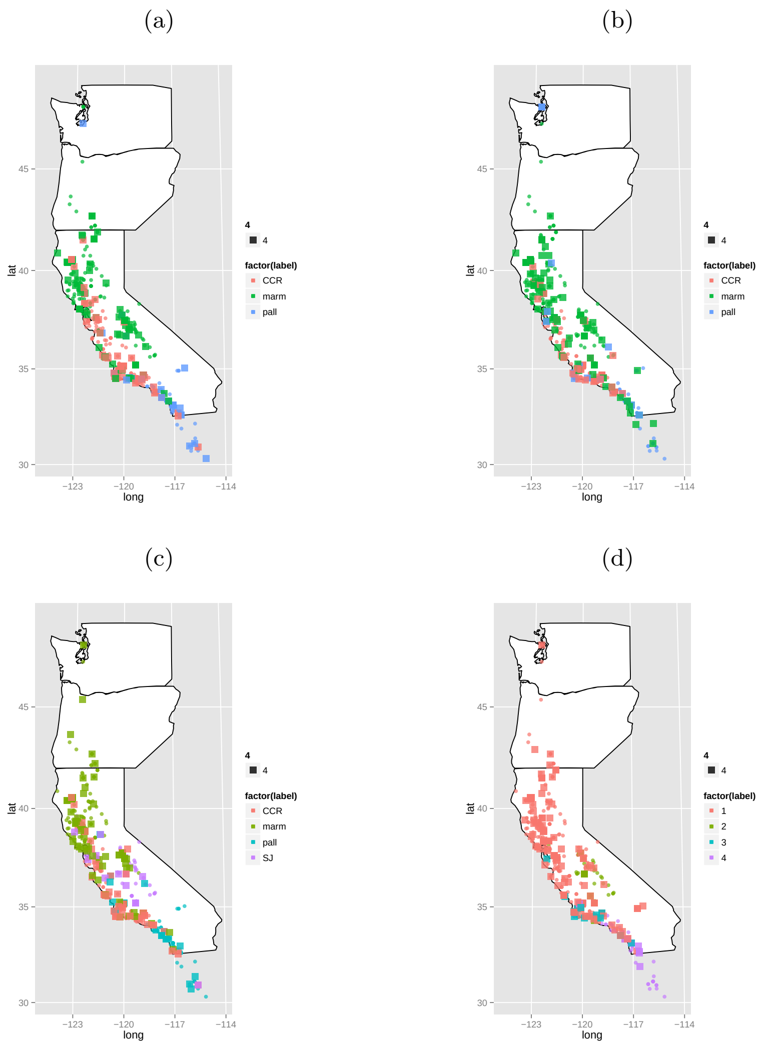

(a)

(b)

(c)

(d)

Figure 4: Geographic position of all turtles sampled. Both training and testing observations are plotted. Training set observations are circles while testing observations are larger squares. Testing set observations are classified based on a multinomial logistic regression model. 4a: sh1 classification scheme. 4b: sh2 classification scheme. 4c: sh3 classification scheme. 4d: spinks classification scheme.

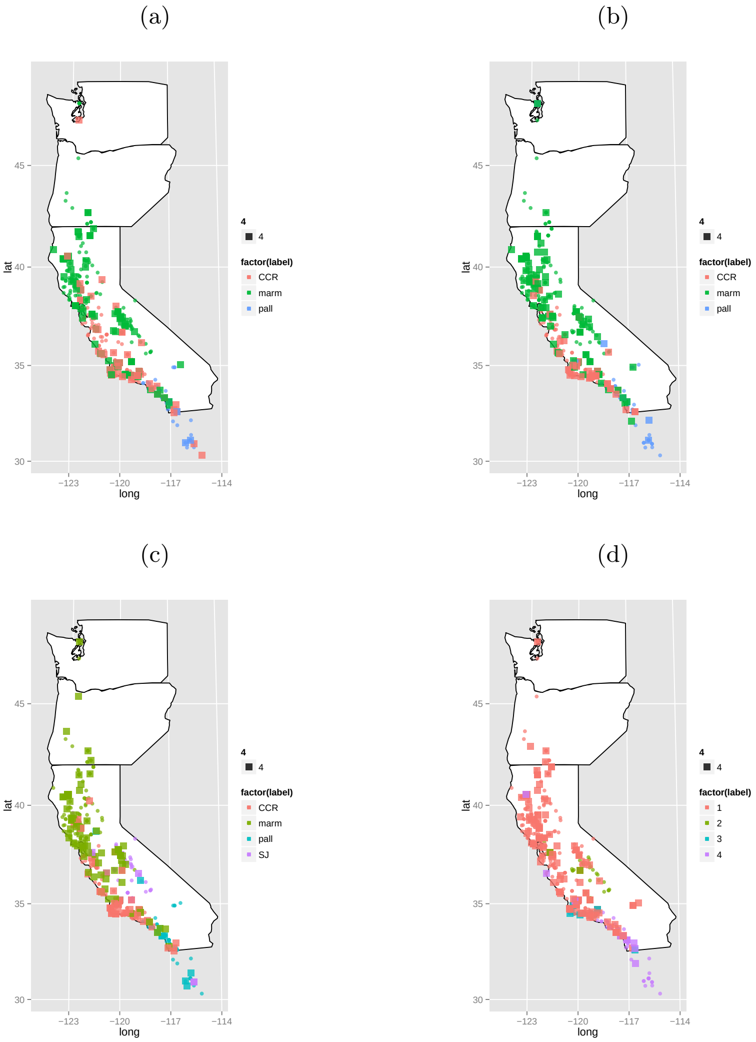

(a)

(b)

(c)

(d)

Figure 5: Geographic position of all turtles sampled. Both training and testing observations are plotted. Training set observations are circles while testing observations are larger squares. Testing set observations are classified based on a feed-forward single layer neural network model. 5a: sh1 classification scheme. 5b: sh2 classification scheme. 5c: sh3 classification scheme. 5d: spinks classification scheme.


(a)

(b)

(c)

(d)

Figure 6: Geographic position of all turtles sampled. Both training and testing observations are plotted. Training set observations are circles while testing observations are larger squares. Testing set observations are classified based on a random forest model. 6a: sh1 classification scheme. 6b: sh2 classification scheme. 6c: sh3 classification scheme. 6d: spinks classification scheme.