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

Peter D Smits[1], Kenneth D Angielczyk[2], James F Parham[3]

[1]Committee on Evolution Biology, University of Chicago, [2]Department of Geology, Field Museum of Natural History, [3]Department of Geological Sciences, California State University – Fullerton

May 10, 2013

Cryptic diversity

How much of cryptic diversity is just a function of sample size?

Emys marmorata

wikimedia

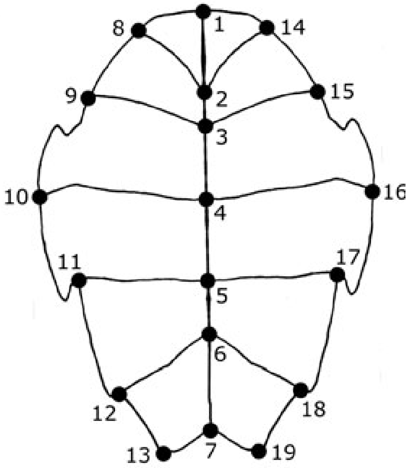
Morphological hypotheses

Phylogenetic hypotheses

Number of subspecies and where they occur.

Methods: morphometrics

- plastral ("belly") shape
- landmarks averaged across bilat axis
- total 13 landmarks, 7 on bilat axis, 6 off
- geographic information known/inferred



Methods: unsupervised learning

Fancy way of saying clustering or density estimation.

Partitioning around mediods (PAM) compared with “gap” statistic.

(dissimilarity based) Evidence accumulation clustering

Methods: supervised learning

Fancy way of saying classification and regression.

Here, features (principal components) predict class (subspecific assignment).

Multnomial logistic regression

Random forests

Methods: model training and selection

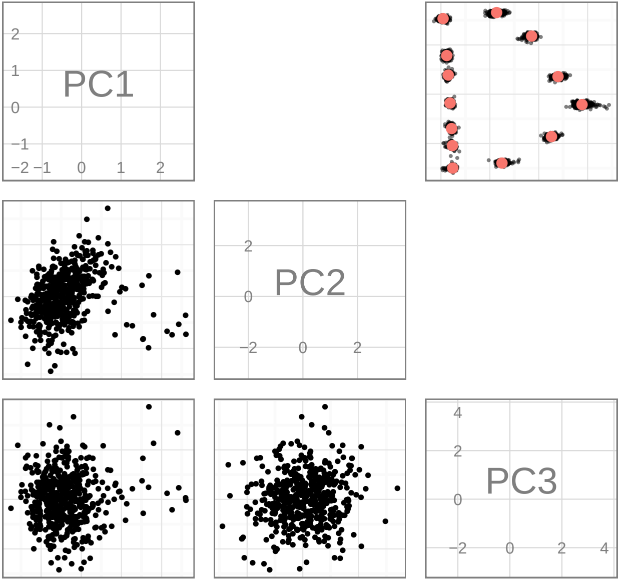
Unknown appropriate number of features to “best” predict class. Want to minimize false positive, while maximizing true positive.

Split data set, 75-25, training and testing.

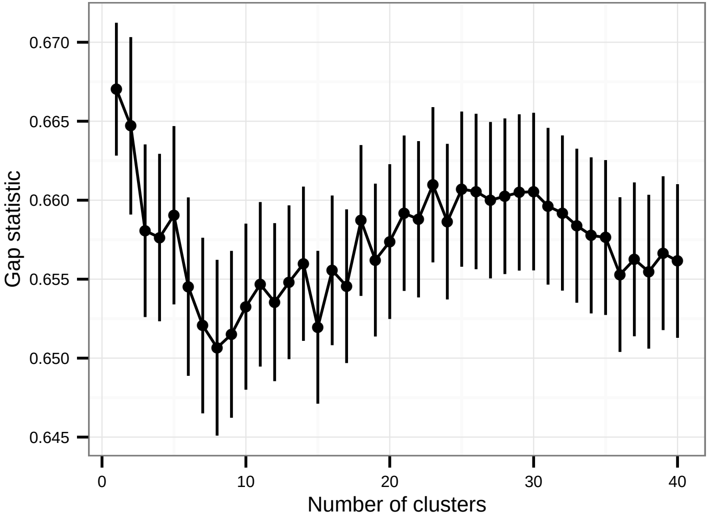
Tuning parameters via grid-search. Uncertainty via 10-fold cross-validation.

Multinomial logistic model selected via min AICc. Random forest model via max AUC ROC.

Results: mophometrics

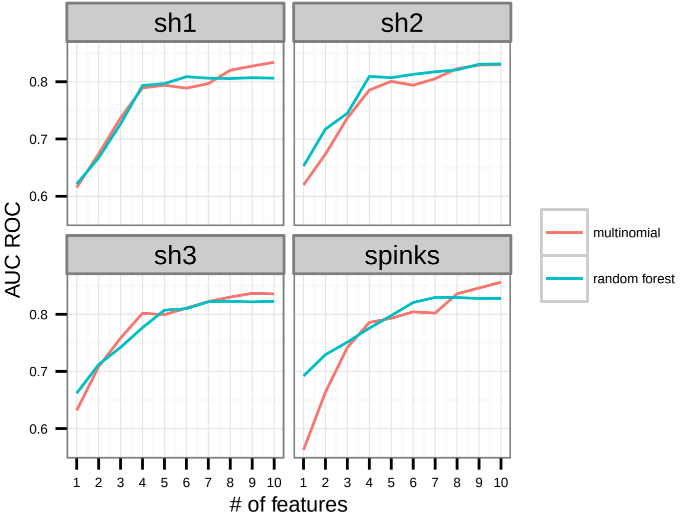


Results: gap clustering



Results: d-EAC

ROC

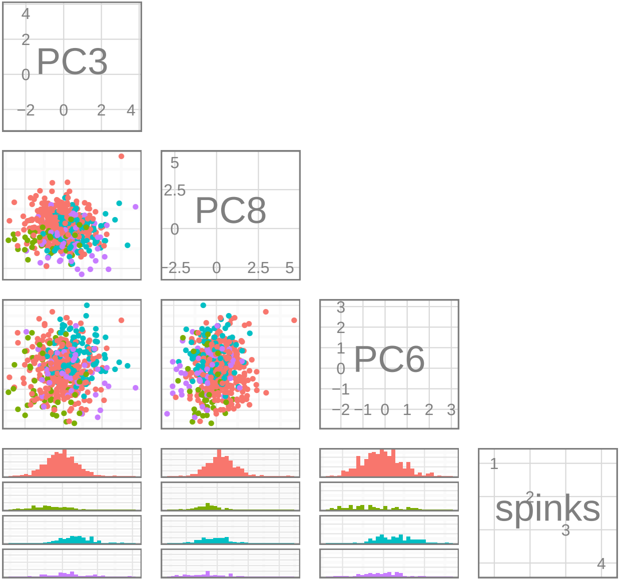


sh1
sh2
sh3
spinks
AUC ROC
0.6
0.7
0.8
1 2 3 4 5 6 7 8 9 10
# of features
multinomial
random forest

Are the AUC ROC values meaningful?

Best classification scheme

Variable importance



Future

Acknowledgements

- Ben Frable, Dallas Krentzel, Michael Foote
- COLLECTIONS