

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

Peter D Smits¹, Kenneth D Angielczyk², James F Parham³

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

May 17, 2013

Cryptic diversity

Cryptic species are species delimited via molecular means which were not/cannot be identified via morphology.

How much of cryptic diversity is just a function of sample size and/or method?

Emys marmorata



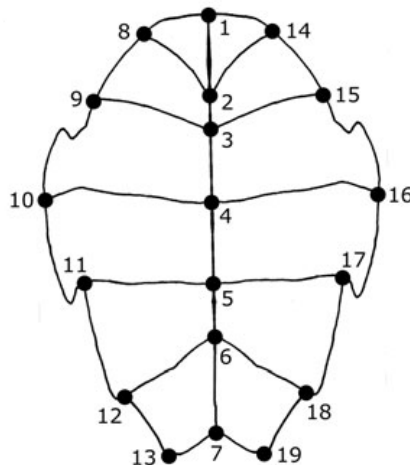
wikimedia

Morphological hypotheses

Phylogenetic hypotheses

Methods: morphometrics

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



Angielczyk *et al.* 2011 *Evolution*

Unsupervised learning

Fancy way of saying clustering or density estimation.

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

(dissimilarity based) Evidence accumulation clustering

Supervised learning

Fancy way of saying classification and regression.

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

Multinomial logistic regression

Random forests

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. Selection via max AUC ROC.

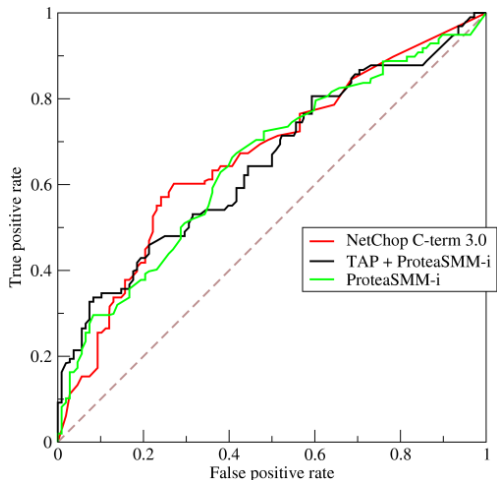
Best multinomial logistic model selected via min AICc. Best random forest model via max AUC ROC.

ROC and confusion matrices

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

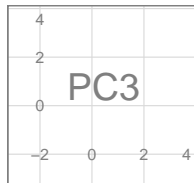
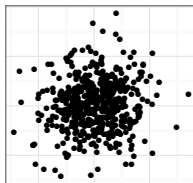
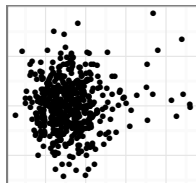
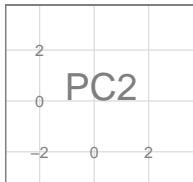
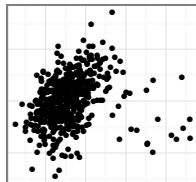
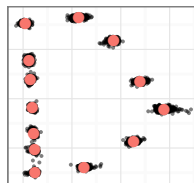
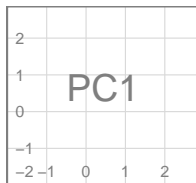
ROC

- ▶ receiver operating characteristic
- ▶ true positive rate or sensitivity: $\frac{TP}{TP+FN}$
- ▶ false positive rate or 1 - specificity: $\frac{FP}{FP+TN}$
- ▶ multiclass, all-against-one (Hand and Till 2001 *Machine Learning*)
- ▶ area under curve
model summary statistic (min 0.5)

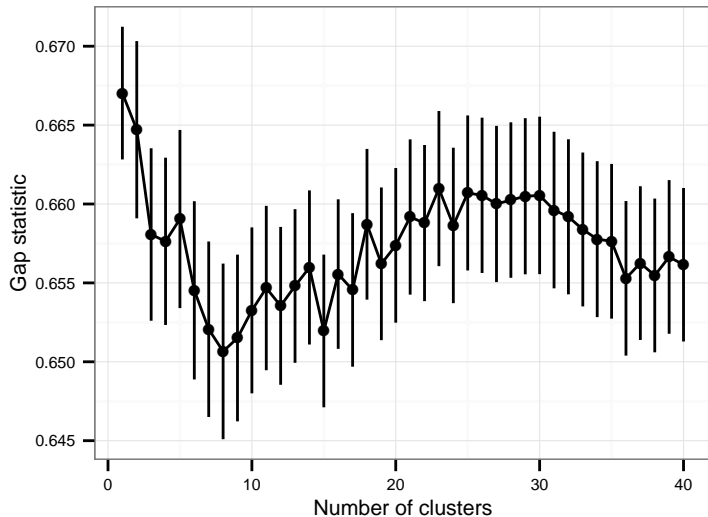


wikimedia

Results: mophometrics

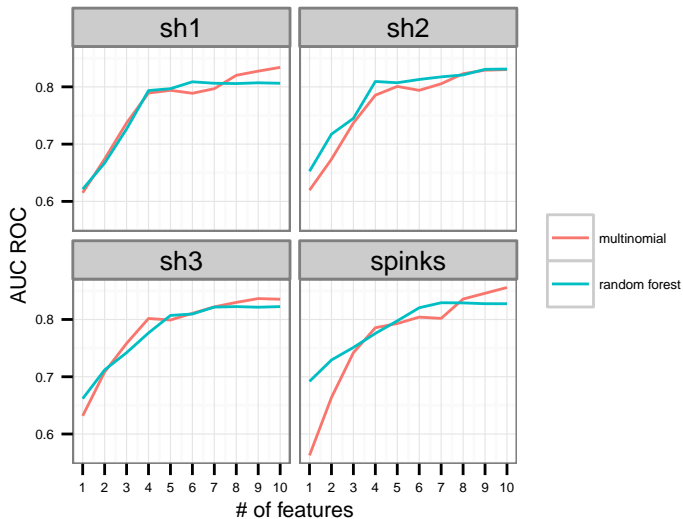


Results: gap clustering



Results: d-EAC

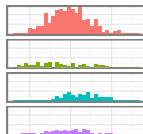
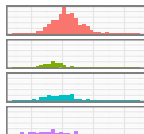
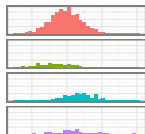
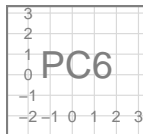
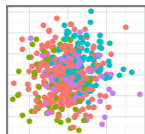
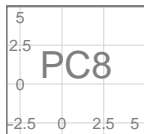
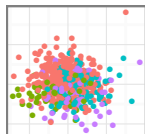
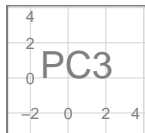
ROC



Are the AUC ROC values meaningful?

Best classification scheme?

Variable importance



Future

Acknowledgements

- ▶ Ben Frable, Dallas Krentzel,
Michael Foote
- ▶ COLLECTIONS
- ▶ FUNDING AGENCIES



The **Field**
Museum

