

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

June 13, 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 hypothesis

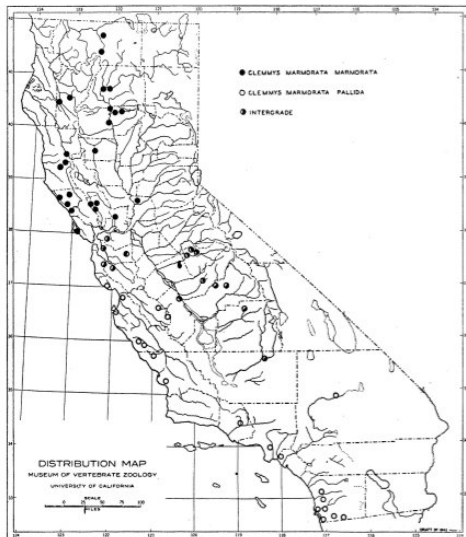
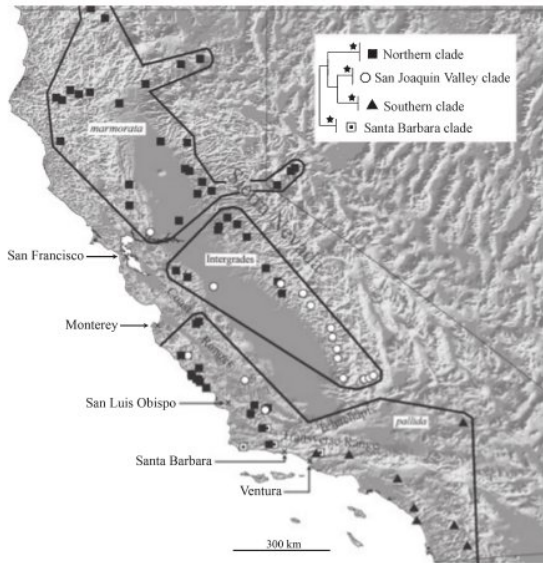


Fig. 4. California localities from which specimens have been examined.

Seeliger 1945 *Copeia*

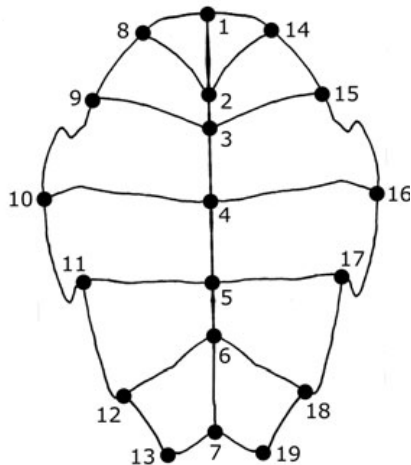
Phylogenetic hypotheses



Spinks *et al.* 2010 *Molec. Ecol.*

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.

Minimize sum of dissimilarities between points and medoids.

“Gap” is analogous to goodness-of-clustering.

Supervised learning

Fancy way of saying classification (and regression).

Features (principal components) predict class (subspecific assignment).

Multinomial logistic regression and random forests.

Model training and selection

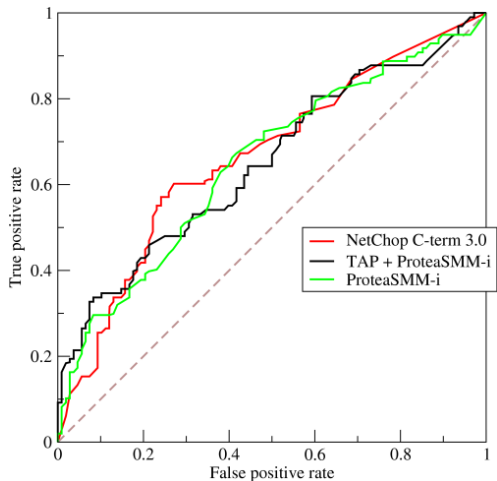
- ▶ split into training and testing sets, 75-25.
- ▶ tuning parameters via grid-search
- ▶ uncertainty via 10-fold CV
- ▶ model selection
 - ▶ multinomial logistic regression: min AICc
 - ▶ random forest: max ROC

ROC and confusion matrices

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

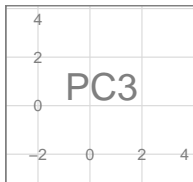
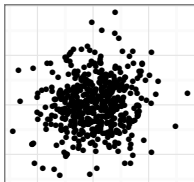
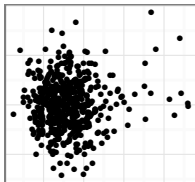
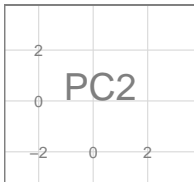
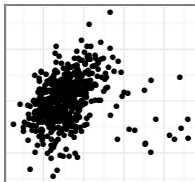
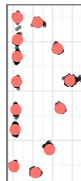
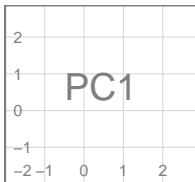
ROC

- ▶ 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*)

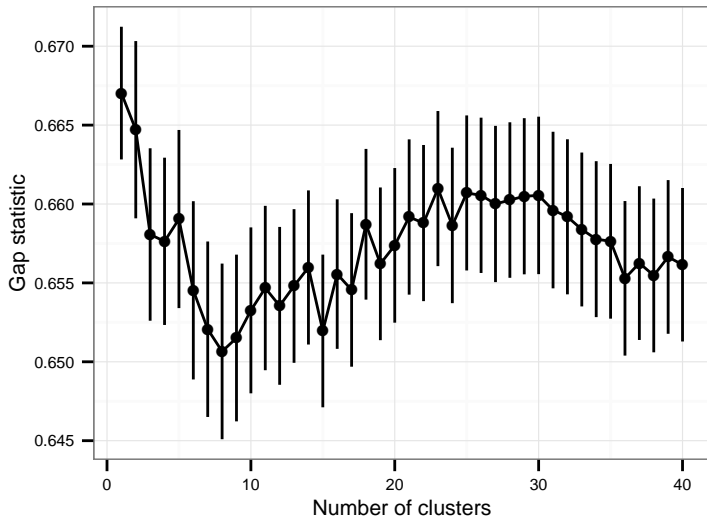


wikimedia

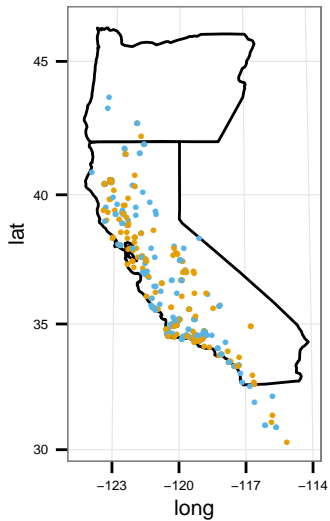
Results: mophometrics



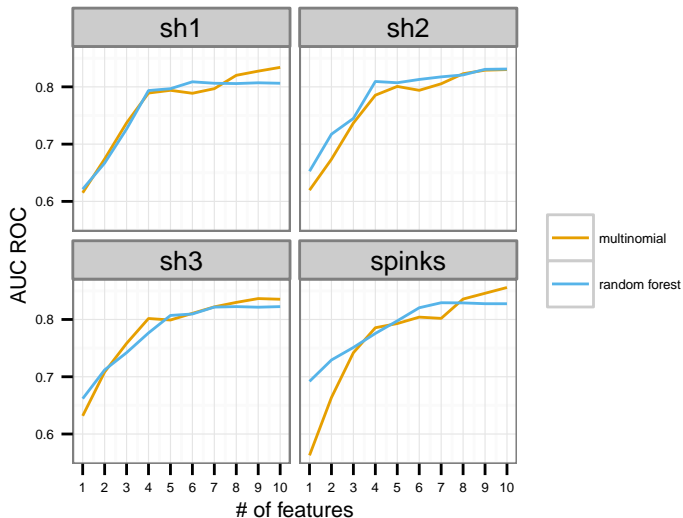
Results: gap clustering



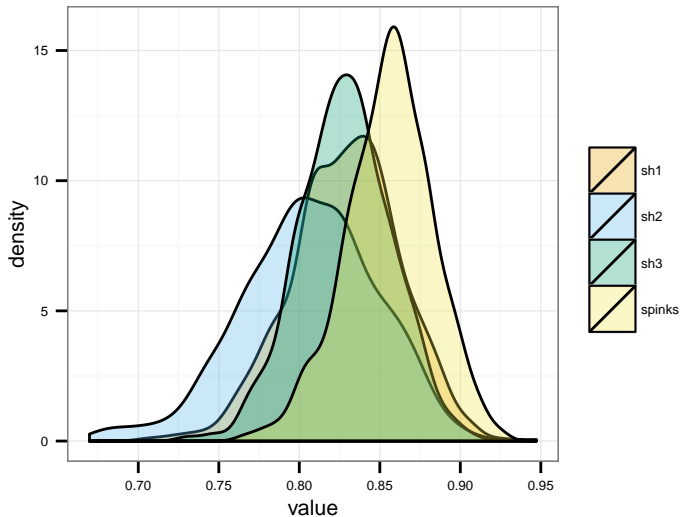
Second best cluster



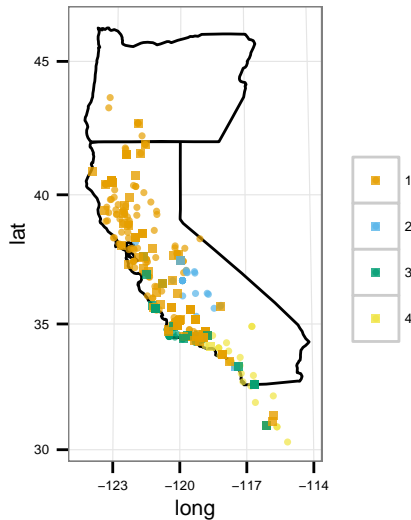
Model selection via ROC



Generalize using best random forest model

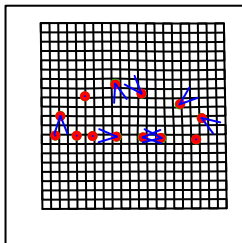


Best classification scheme via RF model results

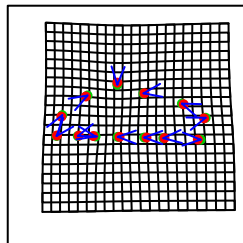


Mean shape of classes

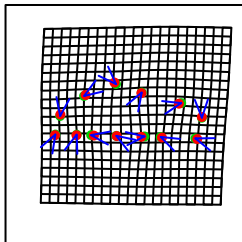
1



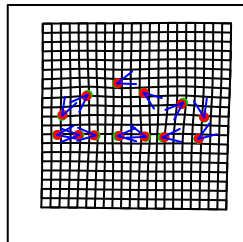
2



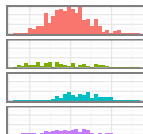
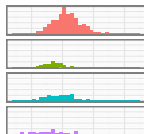
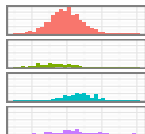
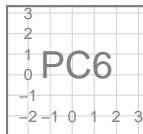
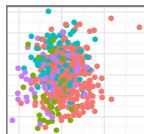
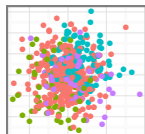
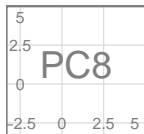
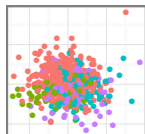
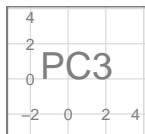
3



4

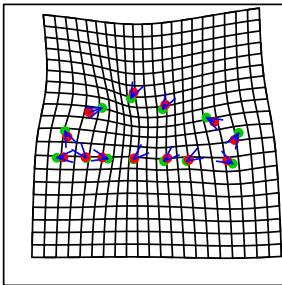


Variable importance of random forest model

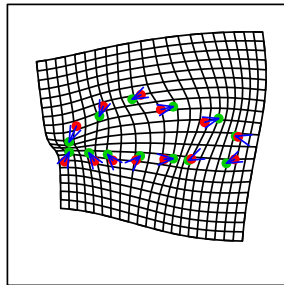


Shape across PC3

min observed

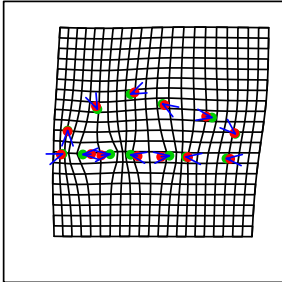


max observed

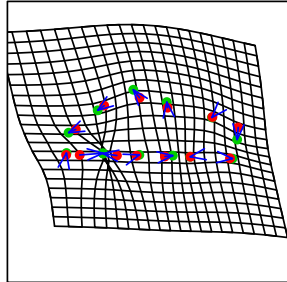


Shape across PC8

min observed



max observed



Future

- ▶ illustration of morphological validation of previously cryptic variation
 - ▶ the concordance is remarkable
 - ▶ large sample sizes can be difficult
- ▶ utility of large data, machine learning methods
- ▶ unsupervised methods for when no explicit hypothesis – nonparametric Bayes
- ▶ cause of interclass variation – local adaptation? pure isolation?

Acknowledgements

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



The **Field**
Museum

