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

June 12, 2013

Cryptic diversity

Crytic species are species delimitated 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

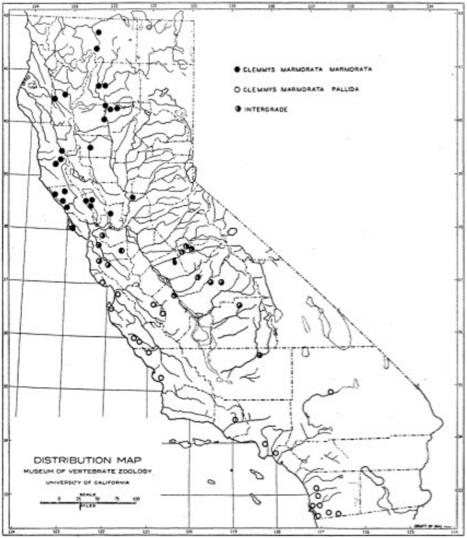
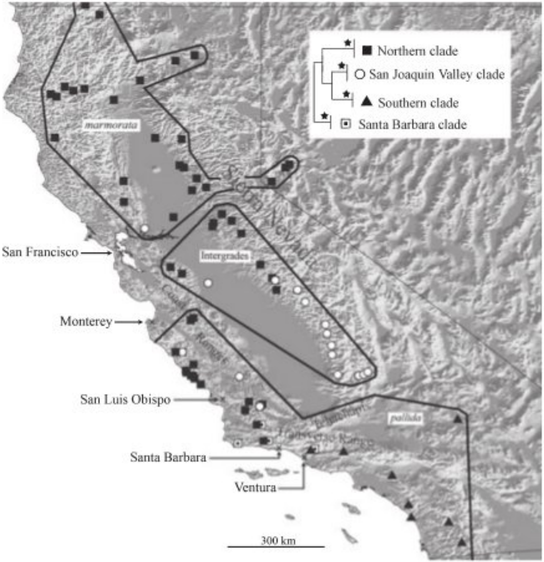


Fig. 4. California localities from which specimens have been examined.

Seeliger 1945 *Copeia*

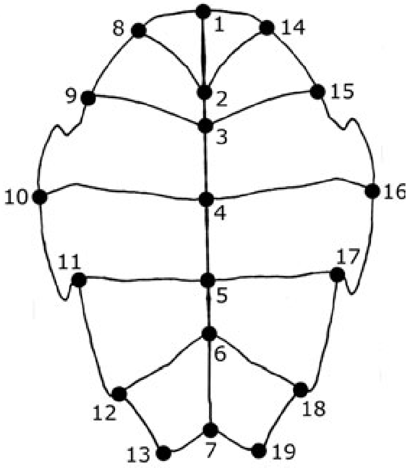
Phylogenetic hypotheses



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 mediods (PAM) compared with “gap” statistic.

Supervised learning

Fancy way of saying classification (and regression).

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

Multinomial logistic regression and 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.

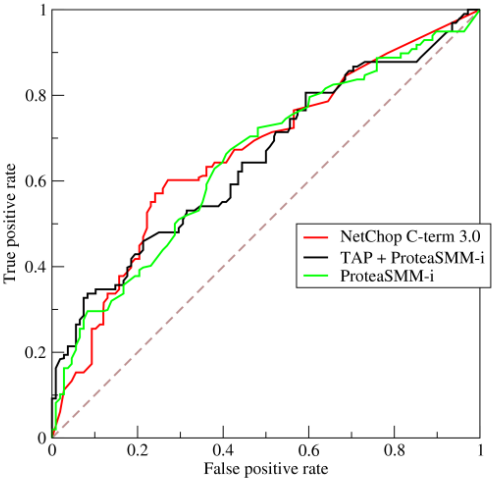
Best multinomial logistic model selected via min AICc. Best random forest model via max AUC ROC.

ROC and confusion matrices

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

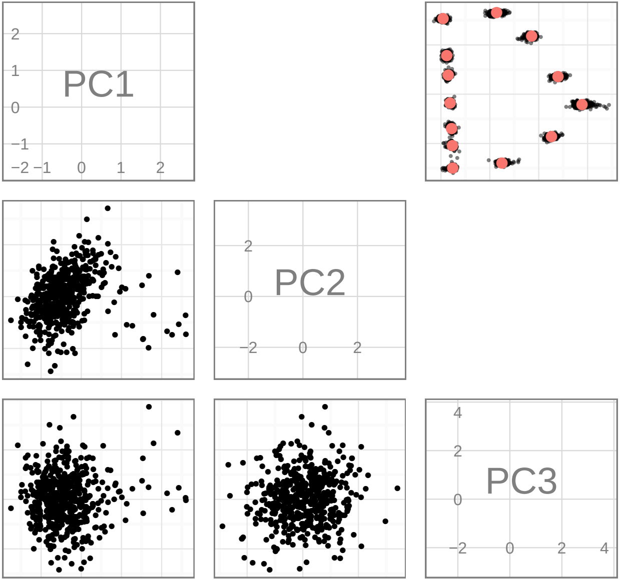
ROC

- 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*)

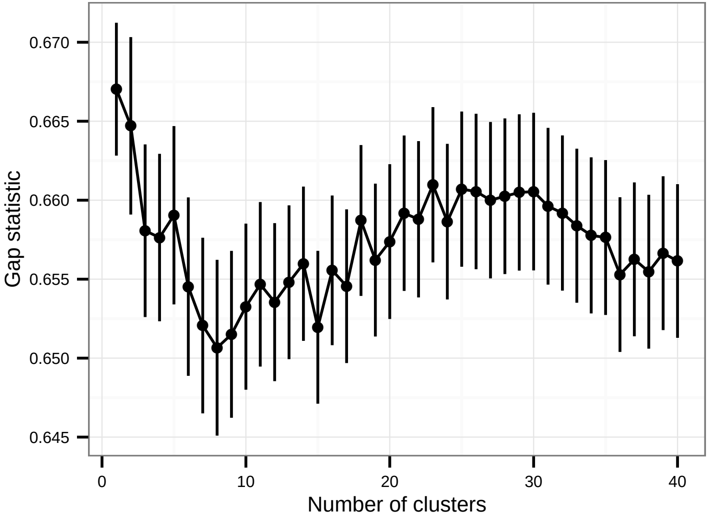


wikimedia

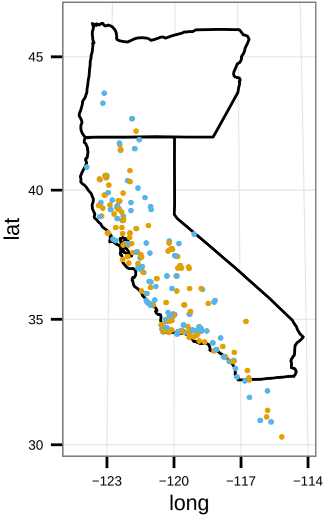
Results: mophometrics



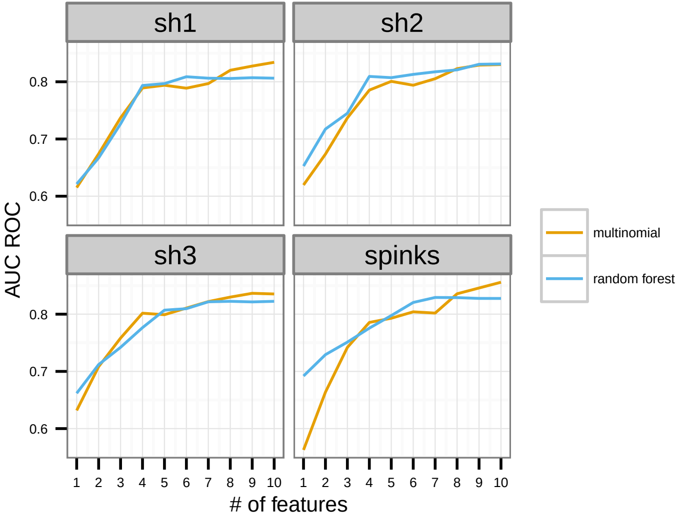
Results: gap clustering



Second best cluster

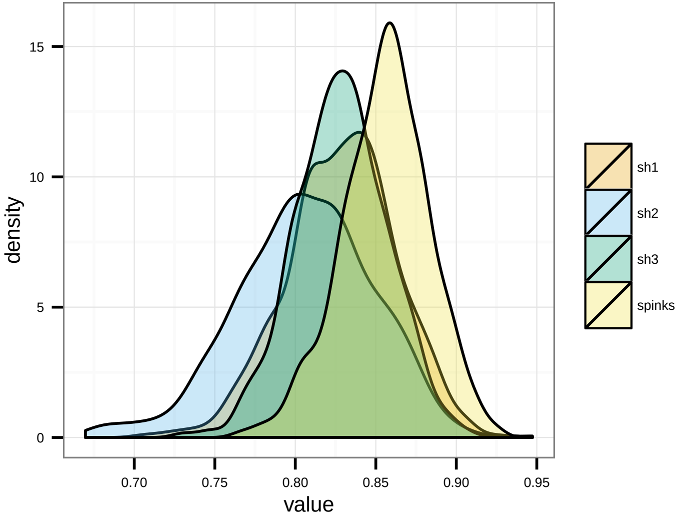


ROC



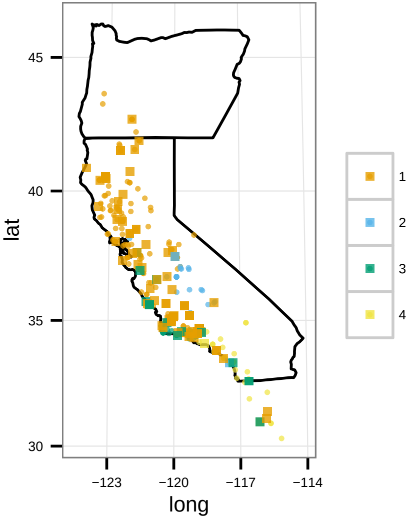
sh1
sh2
sh3
spinks
AUC ROC
0.6
0.7
0.8
1 2 3 4 5 6 7 8 9 10
# of features
multinomial
random forest

Generalize

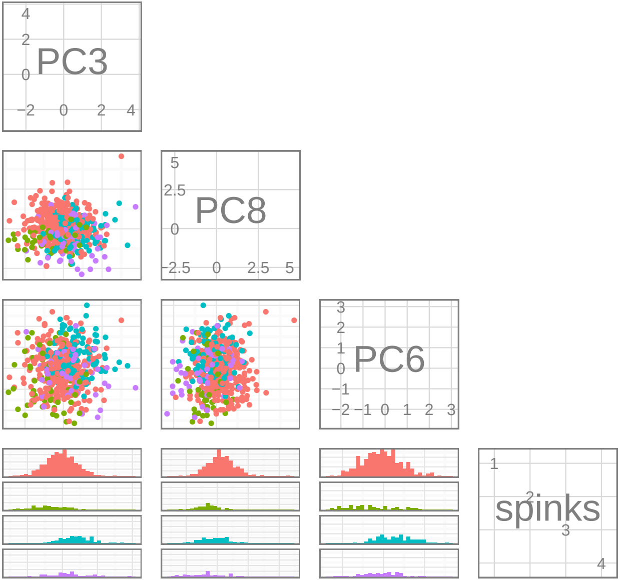


15
10
5
0
density
0.70
0.75
0.80
0.85
0.90
0.95
value
sh1
sh2
sh3
spinks

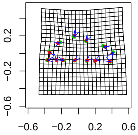
Best classification scheme?

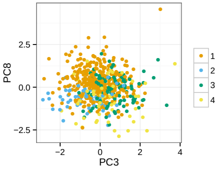


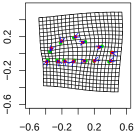
Variable importance



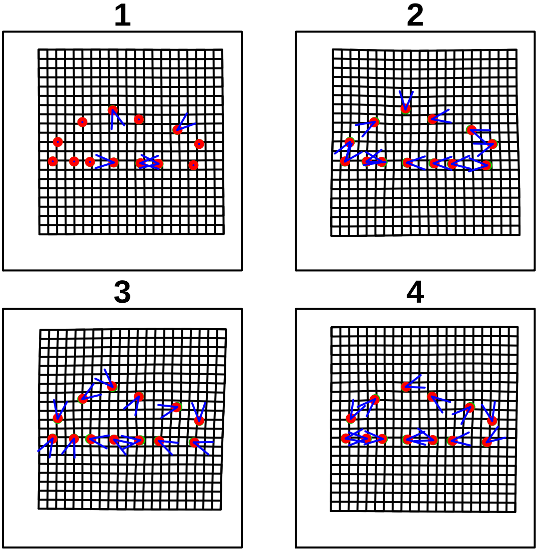
Shape across most important features







Mean shape of classes



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
- COLLECTIONS
- FUNDING AGENCIES