

# Methods and Tools for Spatial Modeling

Elliott Hazen, Ben Best, Jason Roberts, Patrick Halpin

Society for Marine Mammalogy Conference  
Ecological Modeling Workshop  
October 11, 2009




## Our Focus / Biases

- Disciplinary
  - Space (and Time)
  - Environment + Prey
  - Statistics & prediction

## Our Focus

- Disciplinary
  - Space (and Time)
  - Environment + Prey
  - Statistics to prediction
- Flowchart

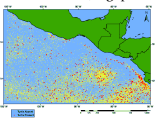


```

            graph LR
            A[Data inputs  
• Distribution  
• Predictors] --> B[Statistical model fitting]
            B --> C[Habitat prediction models]
            
```

## Modeling habitat (overview)

Distribution data, e.g. presence/absence



Presence only data, e.g.

- Vessels of opportunity
- Hydrophones

Presence / absence, density data

- Survey sightings w/ effort
- Bycatch

Event based data

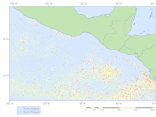
- Focal follows
- Short term tag data

Movement data

- Short and long term tag data

## Modeling habitat (overview)

Distribution data, e.g. presence/absence



**In situ data**

- Continuous data – surface sensors, fisheries acoustics, ADCP
- Station data – CTDs, trawls

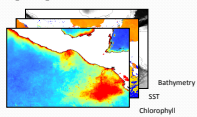
Remotely sensed data

- SST, SSH, Chl - Data centers

Physical / spatial data

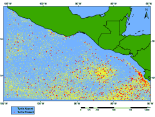
- Bathymetry, distance from feature (e.g. slope, shore, break, front)

Sampled predictive data

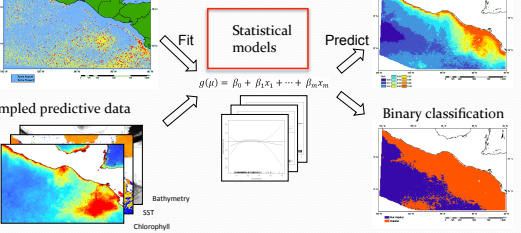


## Modeling habitat (overview)

Distribution data, e.g. presence/absence



Probability of occurrence predicted from environmental covariates



Fit

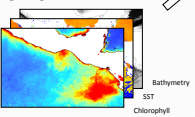
Statistical models

Predict

$$g(x) = \beta_0 + \beta_1 x_1 + \dots + \beta_n x_n$$

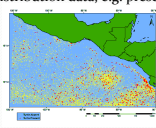
Binary classification

Sampled predictive data



## Model selection

Distribution data, e.g. presence/absence



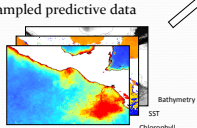
Statistical models

$$g(\mu) = \beta_0 + \beta_1 x_1 + \dots + \beta_m x_m$$

Akaike's Information Criterion → presence - s(SST) + s(prey density)

Bayesian Information Criterion → "Final" model

Sampled predictive data



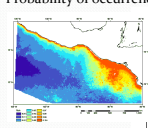
Bathymetry  
SST  
Chlorophyll

AIC =  $2k - 2\ln(L)$   
BIC =  $2\ln(L) + k\ln(n)$

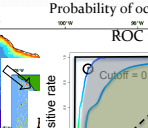
$n$  - sample size  
 $k$  - number of parameters  
 $L$  - Likelihood function

## Evaluating habitat models

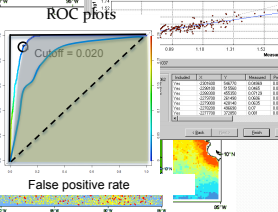
Probability of occurrence



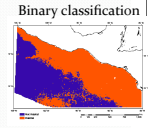
Probability of occurrence



ROC plots



Binary classification



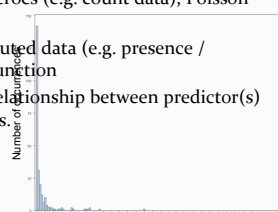
Cross-validation

## Types of statistical models

- There are many, and constantly changing / growing
- Correlation/Regression techniques – GLMs, GAMs (Austin 2002), Mixed models (Wood 2006), regression trees & random forests (Breiman 2001)
- Ordination – Multivariate dimensional scaling, e.g. CCAs (Guisan et al. 1999),
- Maximum Entropy models – species distributions “closest to uniform” (Phillips et al. 2006)
- Recent reviews of modeling approaches (Redfern et al. 2006, Elith et al. 2006, Dormann et al. 2007, Aarts et al. 2008)

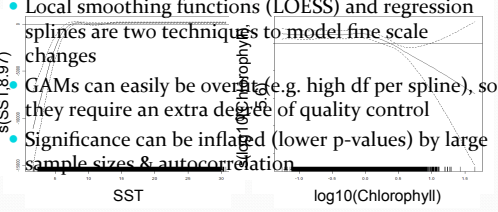
## Generalized Linear Models

- GLMs are an extension of linear models,  $y \sim f(x_1, x_2, \dots, x_n) + \epsilon$  using MLE and a link function
- For data with many zeroes (e.g. count data), Poisson – log link
- For binomially distributed data (e.g. presence / absence) - logit link function
- Relies upon a linear relationship between predictor(s) and response variables.



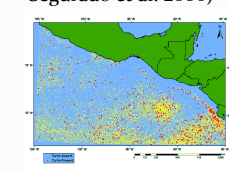

## Generalized Additive Models

- GAMs can use a combination of parametric and non-parametric functions ( $y \sim A + f(x_1) + f(x_2) + \dots + f(x_n) + \epsilon$ )
- Local smoothing functions (LOESS) and regression splines are two techniques to model fine scale changes
- GAMs can easily be overfitted (e.g. high df per spline), so they require an extra degree of quality control
- Significance can be inflated (lower p-values) by large sample sizes & autocorrelation



## Spatial autocorrelation

- Why do we care?
- Model assumption is independence of data points
- Spatial autocorrelation may bias model results (see Segurado et al. 2006)

Albatross track leaving French Frigate Shoals  
From OBIS-SEAMAP (Shaffer et al. 2005)

## Spatial autocorrelation

- Ways to test for it
  - Geary's C (0 to 2)
  - Moran's I (-1 to 1)
- Many methods to model it (Dormann et al. 2007)
  - Autocovariate regression & spatial eigenvectors
  - Generalized least squares (GLS), GLMMs, GEEs
  - Partial Mantel's tests (Legendre and Legendre 1998)

## Useful software packages

- MATLAB / IDL – multipurpose scientific programming language; PERL
- WinBUGS – toolset for bayesian analysis
- EcoPath / EcoSim – Mass balance models
- R / S+ / SAS – statistical programming language
- Python – scripting language used by Arc
- ArcGIS Desktop – Geographic Information System
  - Model builder
  - Hawth's tools, Biomapper, MGET toolbox

## What is MGET?

<http://code.env.duke.edu/projects/mget>

- A collection of geoprocessing tools for marine ecology
  - Oceanographic data management and analysis
  - Habitat modeling, connectivity modeling, statistics
  - Highly modular; designed to be used in many scenarios
  - Emphasis on batch processing and interoperability
- Free, open source software
- Written in Python, R, MATLAB, C#, and C++
- Minimum requirements: Win XP, Python 2.4
- ArcGIS 9.1 or later currently needed for many tools
- ArcGIS and Windows are only non-free requirements

## MGET interface in ArcGIS

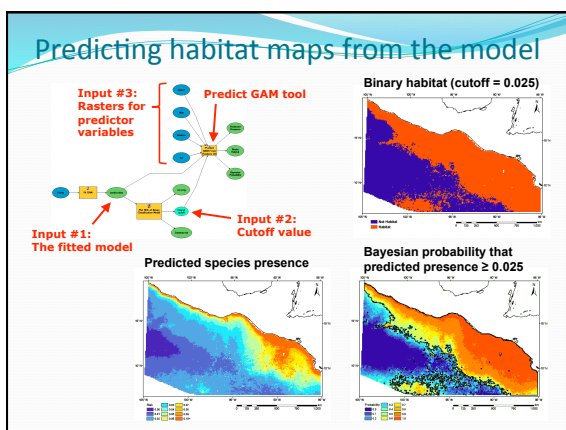
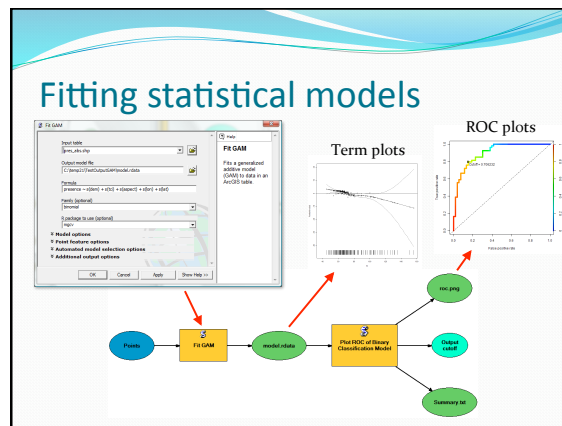
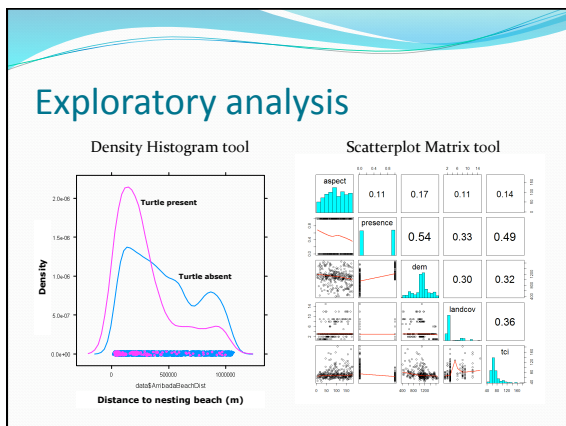
- Expand the toolbox to find the tools
- Double-click tools to execute directly, or drag to geoprocessing models to create a workflow

## Simplified workflow

MGET includes tools that assist with all of these steps

## MGET statistics tools

- Lots of tools, many more planned
- Built from Ben Best's ArcRStats / HabMod projects
- Tools require the R statistics program to be installed on your computer



### Acknowledgements

A special thanks to the many developers of the open source software that MGET is built upon! Also, folks that have helped with this talk:

Sara Maxwell, MGEL lab, Ecological Modeling workshop committee, and many others

Thanks to our funders:

### For more information

Download MGET:  
<http://code.env.duke.edu/projects/mget>

Email us:  
[jason.roberts@duke.edu](mailto:jason.roberts@duke.edu), [bbest@duke.edu](mailto:bbest@duke.edu),  
[elliott.hazen@duke.edu](mailto:elliott.hazen@duke.edu)

Intro to habitat modeling:  
 Guisan, A., Zimmermann, N.E. (2000) Predictive habitat distribution models in ecology. *Ecological Modelling* 135, 147–186.

Thanks for attending!

### Autoregressive / Mixed models

- Autoregressive models incorporate a spatial covariance matrix (Vc) in the error term.
- Mixed models (GLMMs and GAMMs)
  - Can model random effects (e.g. tag deployment) and spatial autocorrelations in within-group errors for sequential data points.
  - Example (Hazen et al. 2009): Humpback whale surface feeding  $\sim f(\text{environmental data, prey metrics}) + \text{random(whale)} + \text{AR}_1 \text{ correlation structure}$ .

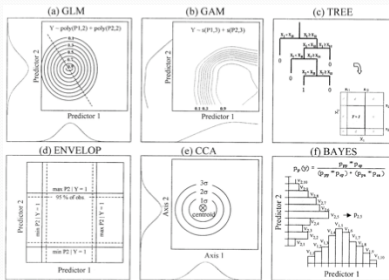
## Data types

- Sightings – presence / absence, density
  - Binary response, zero inflated, many techniques
- Acoustic hydrophones – presence only
  - To be discussed later
- Tag data / focal follows – behavioral state/event
  - State models, movement models
- Vessels of opportunity – effort?
  - Presence only models

## Predictor variables

- Location – bathymetry, distance from feature
- *In situ* oceanography / mooring / remotely sensed data
- Prey data – trawls, stomach contents, fisheries acoustics

## Statistical Models



Guisan & Zimmermann (2000)

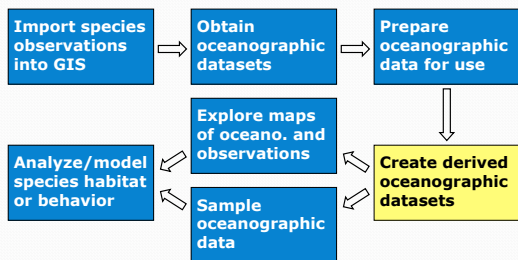
## Correlative methods

- Linear vs. Additive models
  - Advantages to each
  - Assumptions – pseudo-absences,
  - Generalization – poisson distributed data (ZIP)

## Introductions to the Software

- RTFM
- ArcGIS
  - Good, commercial help (+ video)
  - Training.ESRI.com
- Python
  - DiveIntoPython.org – free book
- R
  - A Beginner's Guide to R – free Springer book

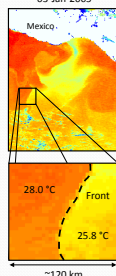
## Simplified workflow



● Marine Geospatial Ecology Tool  
● Connectivity Analysis  
● Connectivity  
● Data Management  
● Data Products  
● Oceanographic Analysis  
● Spatial Analysis  
● Statistics

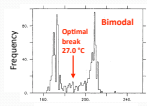
## Identifying SST fronts

AVHRR Daytime SST  
03-Jan-2005




Cayula and Cornillon (1992) edge detection algorithm

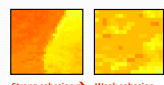
**Step 1: Histogram analysis**



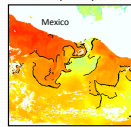
ArcGIS model



**Step 2: Spatial cohesion test**



Example output

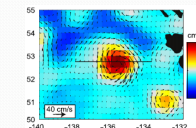


~120 km

● Marine Geospatial Ecology Tool  
● Connectivity Analysis  
● Connectivity  
● Data Management  
● Data Products  
● Oceanographic Analysis  
● Spatial Analysis  
● Statistics

## Identifying geostrophic eddies

Available in  
**MGET 0.8**



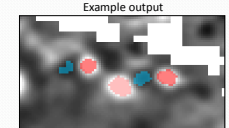
SSH anomaly

$$u = -\frac{g}{f} \frac{\partial h}{\partial y}, \quad v = \frac{g}{f} \frac{\partial h}{\partial x}$$

$$\omega = \frac{\partial v}{\partial x} - \frac{\partial u}{\partial y}, \quad s_n = \frac{\partial u}{\partial x} - \frac{\partial v}{\partial y}, \quad s_s = \frac{\partial v}{\partial x} - \frac{\partial u}{\partial y}$$

$$W = s_n^2 + s_s^2 = \omega^2$$

Example output



Aviso DT-MSLA 27-Jan-1993  
Red: Anticyclonic Blue: Cyclonic

Negative  $W$  at eddy core

## Our Biases

- Disciplinary
  - Space (and Time)
  - Environment + Prey
  - Prediction
- Toolset
  - ArcGIS
  - Python
  - R