Week 5: Spatial Point Pattern Analysis II

Statistical perspective of species distribution modeling

Poisson process

- Homogeneous Poisson process with intensity λ have two properties:
  - $N(A)$ is Poisson distributed with mean $\lambda|A|$, for all A
  - Condition on $N(A)$, the n points are independent and uniformly distributed in A
- model for complete spatial randomness and null model in a statistical analysis.

Non-Poisson Process

- Poisson cluster processes
  - Hierarchical processes: Parent Poisson processes with offspring point processes
  - Example, Matern cluster process, each parent has Poisson(μ) number of offsprings uniformly distributed around the parent.
- Cox processes - A Poisson process with a random intensity function - Example, log-Gaussian Cox processes (LGCP) in which intensity $\lambda(u)$ is a Gaussian random field.

Species distribution modeling

Species distribution models (SDMs) estimate the relationship between species records at sites and the environmental and/or spatial characteristics of those sites. See Figure 1 for commonly used SDM methods.

Absence and background data

- Background data are not attempting to guess at absence locations, but rather to characterize environments in the study region.
- Background data establishes the environmental domain of the study or regions where a specie should habitat but hasn't, whilst presence data should establish under which conditions a species is more likely to be present than on average.
- A closely related concept is "pseudo-absences", which is also used for generating the non-presence class. In this case, researchers try to guess where absences might occur they may sample the whole region except at presence locations, or they might sample at places unlikely to be suitable for the species.

Regression models

- Converting a SDM problem into a generalized linear regression (GLM) or generalized additive models (GAM)

$$\eta(s) = \alpha + \sum_{p=1}^{P} \beta_i X_i(s) + \epsilon(s)$$

where $\eta(\cdot) = \frac{P(s)}{1-P(s)}$ is the logit link function, X_i is the $i-$th variable and $\epsilon(\cdot) \sim N(\mu, \sigma)$ is Gaussian distributed residuals.

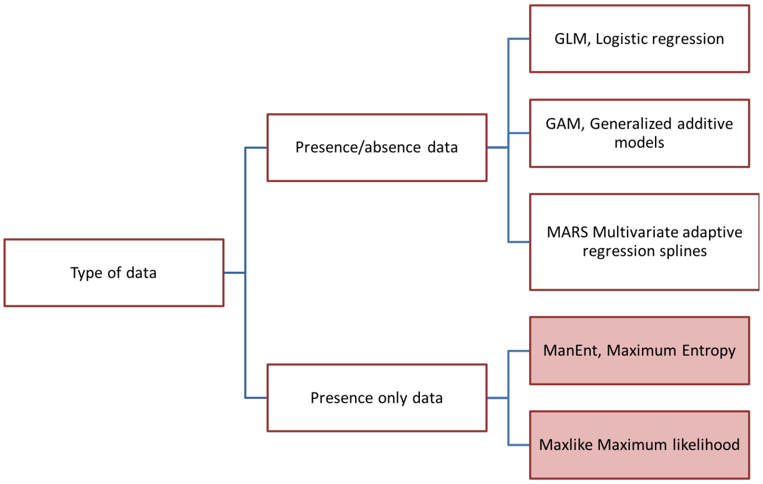
Type of data
Presence/absence data
Presence only data
GLM, Logistic regression
GAM, Generalized additive models
MARS Multivariate adaptive regression splines
ManEnt, Maximum Entropy
Maxlike Maximum likelihood

Figure 1: SDM methods

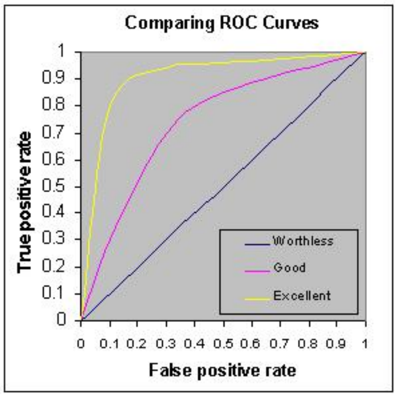

Figure 2: ROC curves and AUC

- The diagnostic tools (e.g., AIC, k-fold methods) avaiable in GLM can be applied to here

How good are the models

- Confusion matrix

	+	-
+	True Positive	Faluse Positive
-	False Positive	True Negative

- Kappa statistics: Scales between 0 and 1; >0.7 good, 0.4 – 0.7 fair, <0.4 poor
- ROC curves and AUC: 0.8 good, 0.6 – 0.8 fair, 0.5 random, <0.6 poor

MaxEnt

- One of the most widely used SDM framework with consistently competitive predictive performance, developed by: Phillips, Steven J., Robert P. Anderson, and Robert E. Schapire. "Maximum entropy modeling of species geographic distributions." Ecological modelling 190, no. 3-4 (2006): 231-259.
- **Premise:** the best approximation of a distribution is determined by maximum entropy, subject to constraints on it's moments. It agrees with everything that is known, but carefully avoids assuming anything that is not known (Jaynes 1990)
- The distribution π assign non-negative probability $\pi(x)$ to each point x, and these probabilities sum to 1. Our approximation of π is also a probability distribution, and we denote it $\hat{\pi}$. The entropy of $\hat{\pi}$ is defined as (Phillips et al. 2006):

$$H(\hat{\pi}(x)) = -\sum_{x \in X} \hat{\pi}(x) ln \hat{\pi}(x)$$

Geographic models

- Distance-based
- Convex Hull
- Geographic circles
- Interpolation
- Voronoi diagram

Reading:

Phillips et al. 2006: Maximum entropy modeling of species geographic distributions

https://cran.r-project.org/web/packages/dismo/vignettes/sdm.pdf