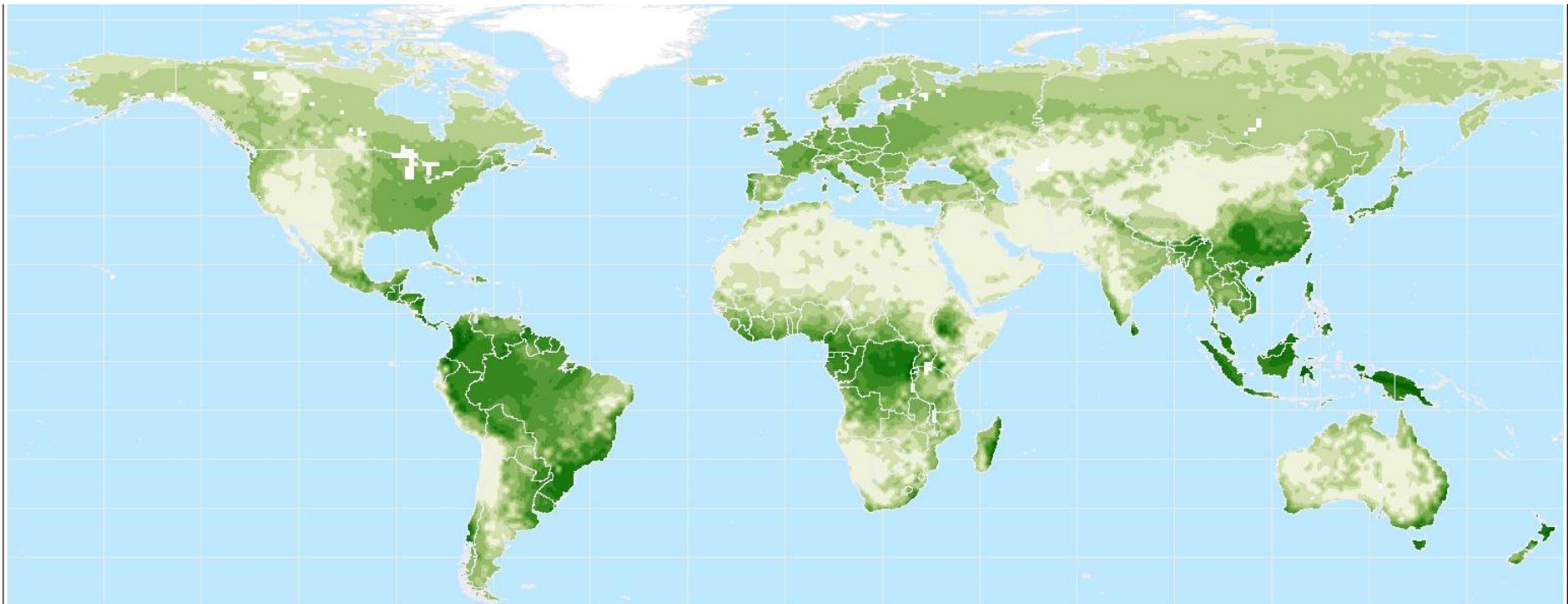# Mapping and modeling species distributions

Department of Botany and Zoology, Masaryk University

Bi9661 Selected issues in Ecology, Autumn 2013

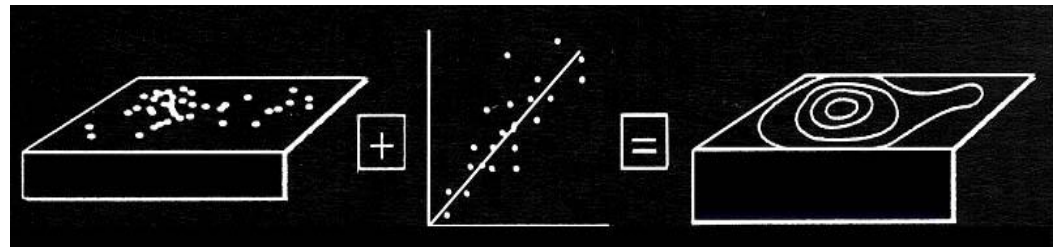Borja Jiménez-Alfaro, PhD

Part 2:
MODELING
MODELING METHODS

SDMs are correlative models

Species distribution is infered from known environmental conditions
* Habitat suitability
* Realized niche?

Johnston, 1993

But models can be also mechanistic (process-based)

Assessing the bio-physiological aspects of a species to generate the conditions in which the species can ideally persist, based on observations made in controlled field or laboratory studies
* The Real NICHE
* Fundamental niche?

TYPES OF (CORRELATIVE) MODELS

We will use the classification of Hijmans and Elith (2013)
for "dismo" package in R

> Species distribution modeling with R
>
> Robert J. Hijmans and Jane Elith
>
> July 8, 2013

Similar classification is provided in Wikipedia

Article	Talk		Read	Edit	View history	Search

Environmental niche modelling

From Wikipedia, the free encyclopedia

Environmental niche modelling, alternatively known as **species distribution modelling**, **(ecological) niche modelling**, **predictive habitat distribution modelling**, and **climate envelope modelling** refers to the process of using computer algorithms to predict the distribution of species in geographic space on the basis of a mathematical representation of their known distribution in environmental space (= realized ecological niche). The environment is in most cases represented by climate data (such as temperature, and precipitation), but other variables such as soil type, water depth, and land cover can also be used. These models allow for interpolating between a limited number of species occurrence and they are used in several research areas in conservation biology, ecology and evolution.

The extent to which such modelled data reflect real-world species distributions will depend on a number of factors, including the nature, complexity, and accuracy of the models used and the quality of the available environmental data layers; the availability of sufficient and reliable species distribution data as model input; and the influence of various factors such as barriers to dispersal, geological history, or biotic interactions, that increase the difference between the realized niche and the fundamental niche. Environmental niche modelling may be considered a part of the discipline of biodiversity informatics.

A brief overview of methods and algorithms

Profile methods

(Presence data)

Bioclim
Domain
Mahalanobis distance

Regression models

(Presence/absence data)

Generalized Linear Models (GLM)
Generalized Additive Models (GAM)

Machine learning

(Presence data)
(+background data)

Boosted Regression Trees (BRT)
Random Forests (RF)
MaxEnt

Profile methods

◆ Only consider presence data

◆ Based on the environmental distance to known sites

◆ Simple algorithms easy to understand

◆ Bioclim, Domain, Mahalabobis

◆ Others: Environmental Niche Factor Analyses (ENFA)

BIOCLIM

The BIOCLIM algorithm – environmental envelope – computes the similarity of a location by comparing the values of environmental variables at any location to a percentile distribution of the values at known locations of occurrence ('training sites')

Nix, H.A. (1986) A biogeographic analysis of Australian elapid snakes. In: Atlas of Elapid Snakes of Australia. (Ed.) R. Longmore, pp. 4-15. Australian Flora and Fauna Series Number 7. Australian Government Publishing Service: Canberra
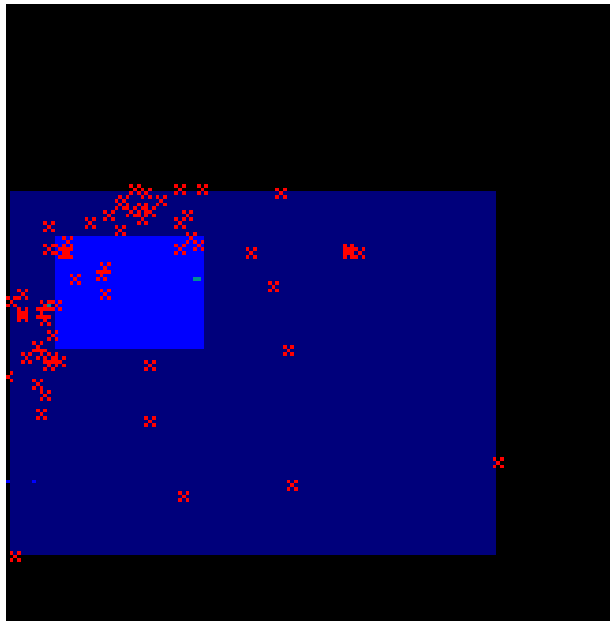
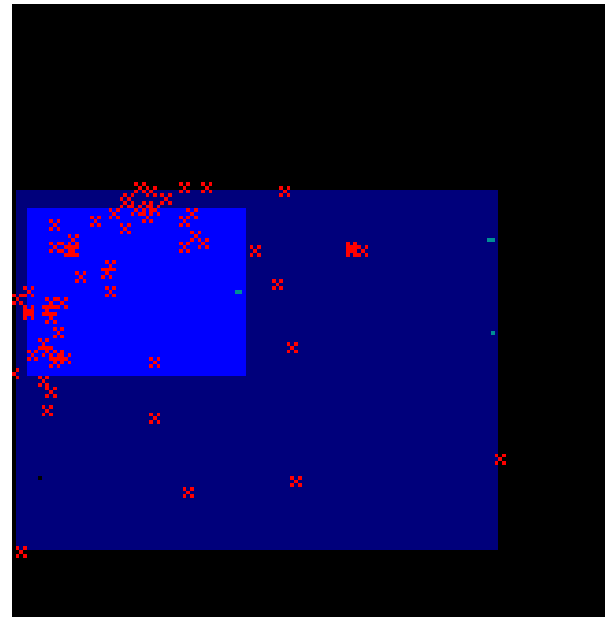Implemented in:
R (dismo), Openmodeller, Diva-GIS

Each variable has its own **envelope** represented by the interval [m - c*s, m + c*s]
The result is a cubic space (**box**) of n dimensions

Example for one species using temperature x precipitation


Cutoff = 0.67

Cutoff = 0.99

DOMAIN

The Domain algorithm is a **distance-based** method based on the **Gower** distance between environmental variables at any location and those at any of the known locations ('training sites'). Results are not a probability of occurrence but a measure of similarity

Carpenter G., A.N. Gillison and J. Winter, 1993. Domain: a flexible modelling procedure for mapping potential distributions of plants and animals. Biodiversity Conservation 2: 667-680
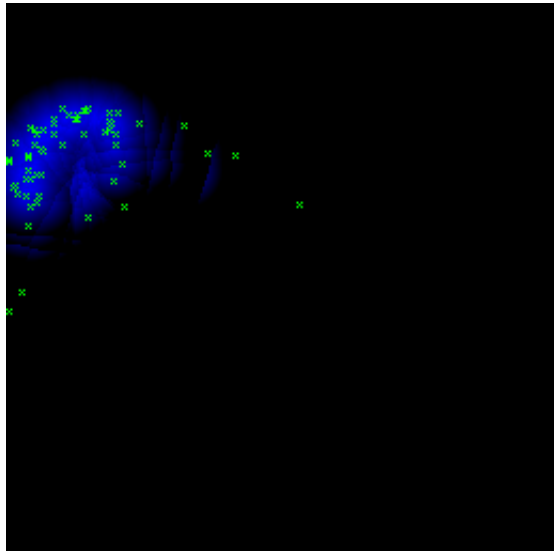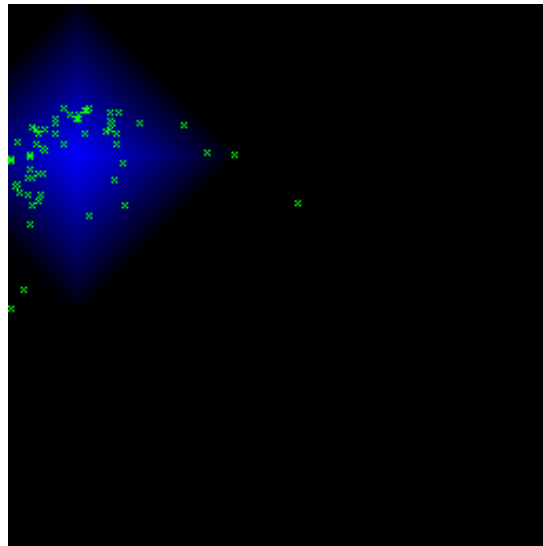
Implemented in:
R (dismo), Openmodeller, Diva-GIS

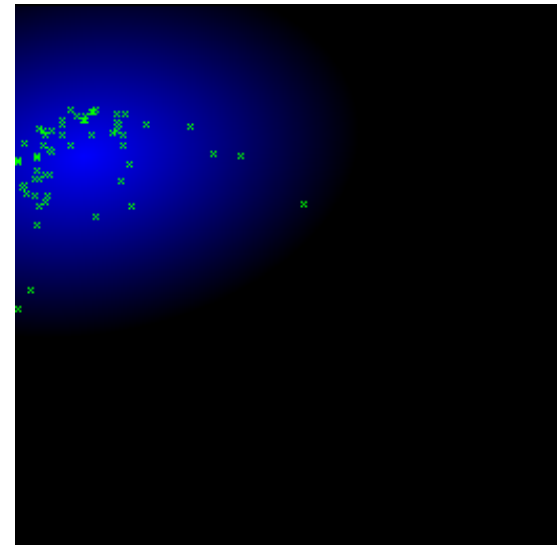# Ecological distance methods – Multivariate similarity
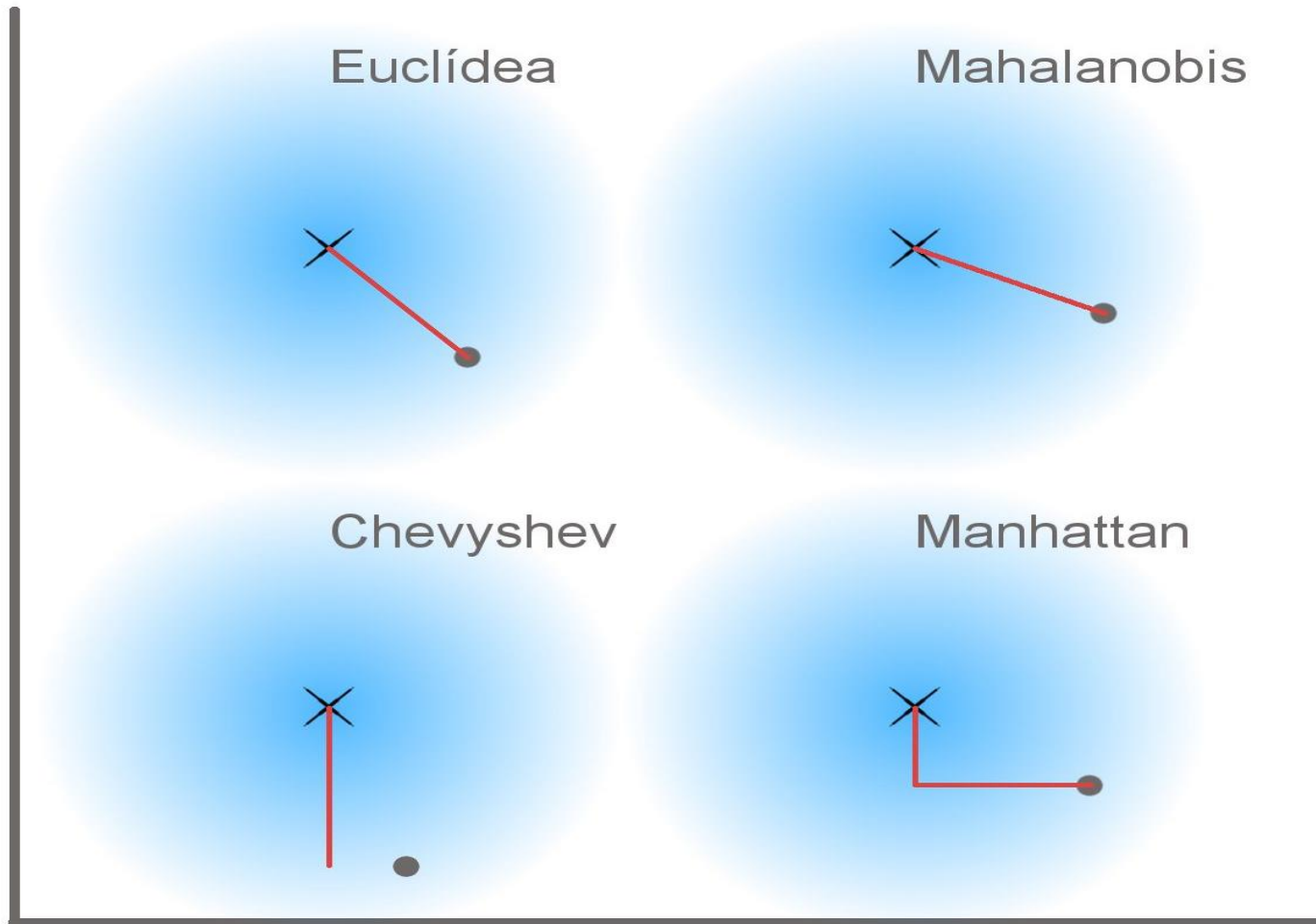
Euclidean distance to the centroid

Gower distance to the centroid

Mahalanobis distance to the centroid

models in the environmental space (temperature x precipitation) generated with the same input (*Thalurania furcata boliviana* localities) but with different parameters (source: openmodeller)

MAHALANOBIS Distance

Unlike Euclidean distance, Mahalanobis distance takes into account the **correlations of the variables** in the data set, and it is not dependent on the scale of measurements. It is based on a descriptive statistic developed by P.C. Mahalanobis in 1936.

Mahalanobis, P.C., 1936. On the generalised distance in statistics. Proceedings of the National Institute of Sciences of India 2: 49-55.

Implemented in:
R (dismo), Openmodeller, IDRISI, ArcView

A comparison of PROFILE ALGORITHMS

(Tsoar et al 2007)

BIODIVERSITY RESEARCH

A comparative evaluation of presence-only methods for modelling species distribution

Asaf Tsoar*, Omri Allouche, Ofer Steinitz, Dotan Rotem and Ronen Kadmon

Department of Evolution, Systematics and Ecology, Institute of Life Sciences, The Hebrew University of Jerusalem, Campus Edmond Safra, Jerusalem 91904, Israel

ABSTRACT

In spite of increasing application of presence-only models in ecology and conservation and the growing number of such models, little is known about the relative performance of different modelling methods, and some of the leading models (e.g. GARP and ENFA) have never been compared with one another. Here we compare the performance of six presence-only models that have been selected to represent an increasing level of model complexity [BIOCLIM, HABITAT, Mahalanobis distance (MD), DOMAIN, ENFA, and GARP] using data on the distribution of 42 species of land snails, nesting birds, and insectivorous bats in Israel. The models were calibrated using data from museum collections and observation databases, and their predictions were evaluated using Cohen's Kappa based on field data collected in a standardized sampling design covering most parts of Israel. Predictive accuracy varied between modelling methods with GARP and MD showing the highest accuracy, BIOCLIM and ENFA showing the lowest accuracy, and HABITAT and DOMAIN showing intermediate accuracy levels. Yet, differences between the various models were relatively small except for GARP and MD that were significantly more accurate than BIOCLIM and ENFA. In spite of large differences among species in prevalence and niche width, neither prevalence nor niche width interacted with the modelling method in determining predictive accuracy. However, species with relatively narrow niches were modelled more accurately than species with wider niches. Differences among species in predictive accuracy were highly consistent over all modelling methods, indicating the need for a better understanding of the ecological and geographical factors that influence the performance of species distribution models.

Keywords

Bats, BIOCLIM, birds, distribution range, DOMAIN, ecological niche models, ENFA, GARP, HABITAT, habitat suitability, Kappa, Mahalanobis distance, predictive maps, prevalence, snails, tolerance.
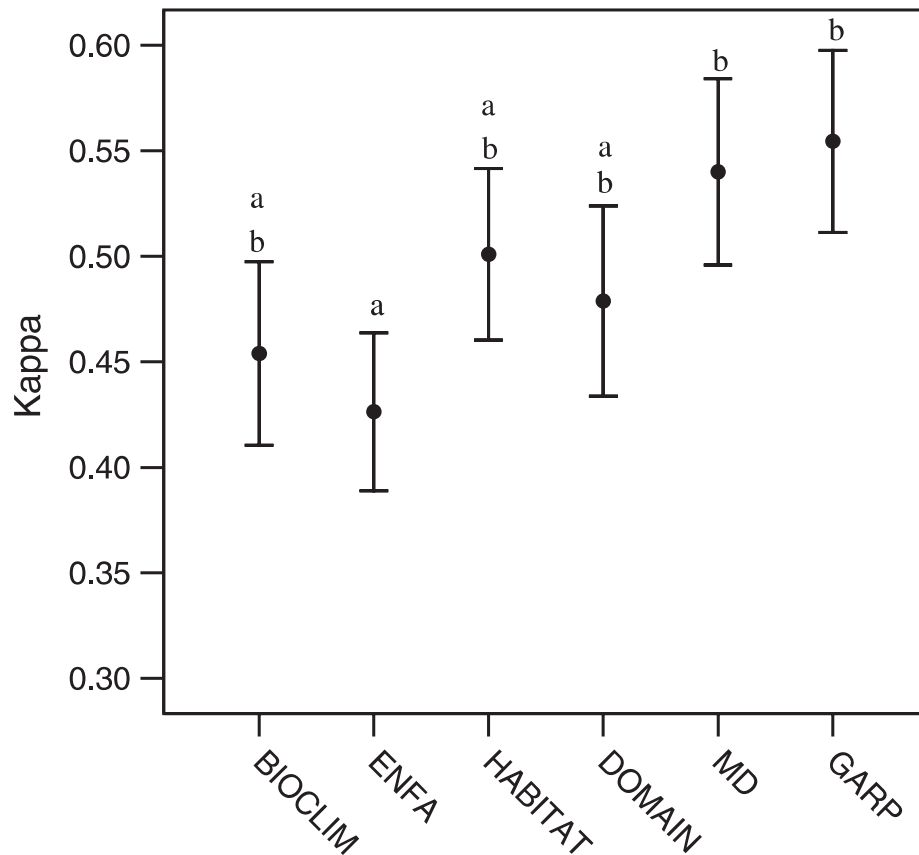
*Correspondence: Asaf Tsoar, Department of Evolution, Systematics and Ecology, Institute of Life Sciences, The Hebrew University of Jerusalem, Campus Edmond Safra, Jerusalem 91904, Israel. Tel.: 972 26586080; Fax: 972 26584655; E-mail: tsoarasaf@pob.huji.ac.il

Figure 2 Differences in Kappa among modelling methods (BIOCLIM, HABITAT, GARP, ENFA, DOMAIN, and MD) when data for all taxa (snails, birds, and bats) are pooled. Error bars represent mean ± 1 standard error. Models sharing the same letters do not differ from each other significantly ($P > 0.05$ following Bonferroni corrections for multiple comparisons).

PROFILE ALGORITHMS

Pros

Models of easy interpretation and representation
Support low number of occurrences
Widely used for large-scale approaches

Cons

Lack of procedures for variable selection
Interactions between variables are not allowed
All the variables have the same weight
Very sensitive to bias data and marginal points

Regression models

◆ Combine presence and absence data

◆ Models are calibrated by extracting the values of locations

◆ Robust methods and classical statistics (probabilistic inference)

◆ GLMs, GAMs (implemented in most stat software)

◆ Other: Multiple Adaptive Regression Splines (MARS)

GENERALIZED LINEAR MODELS (GLM)

Characteristics

- Assumes an underlying distribution for the error, but not necessarily the normal

- Variances (σ^2) are not assumed constant

- The addition of a link function, $g(\mu)$, transforms the mean of the response to a linear function of the predictor variables

- Basic model form:

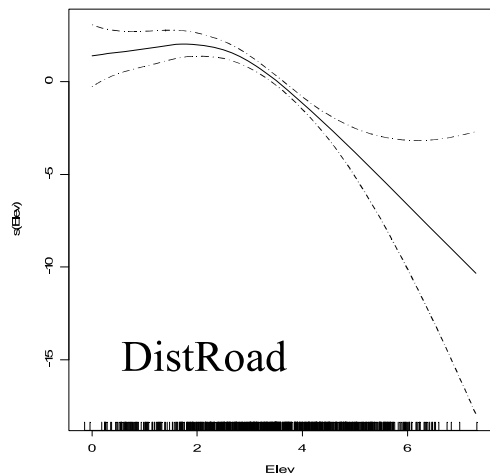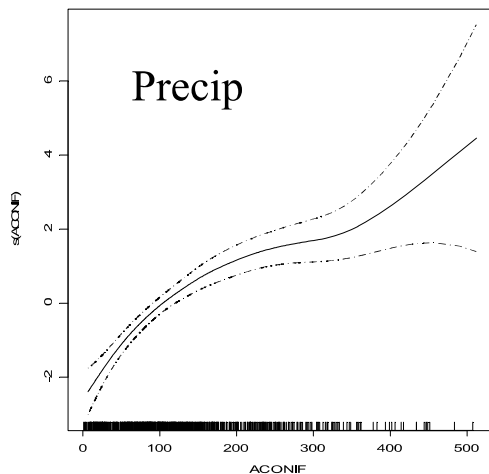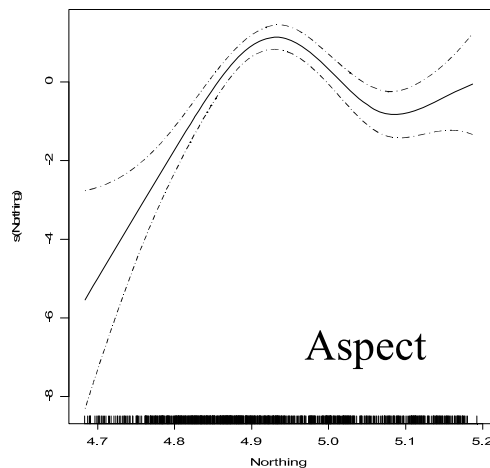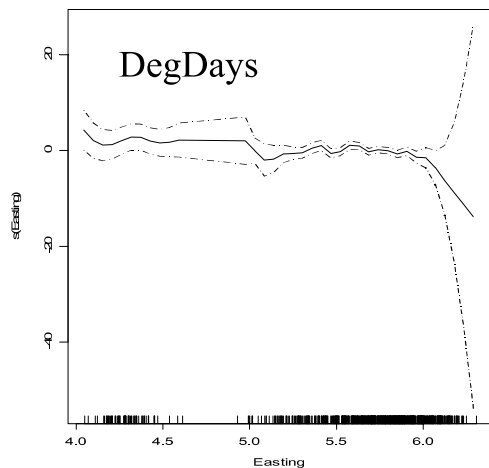$$g(\mu) = \alpha + \sum \beta_i \mathbf{X}_i + \varepsilon_i$$

Example Links	
Normal	$= \mu$
Binomial (logit)	$= \log(\mu)$
Poisson	$= \log[\mu/(1-\mu)]$

GENERALIZED LINEAR MODELS (GLM)

- Algebraic reorganization of the odds ratio rescales the likelihood of event class i to

$$P(y \mid x) = \frac{e^{\alpha + \beta X}}{1 + e^{\alpha + \beta X}}$$

where

$$\alpha + \beta X = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$

- Extends multiple linear regression to data with a binary response variable

GENERALIZED LINEAR MODELS (GLM)

Logistic regression with the "glm" function
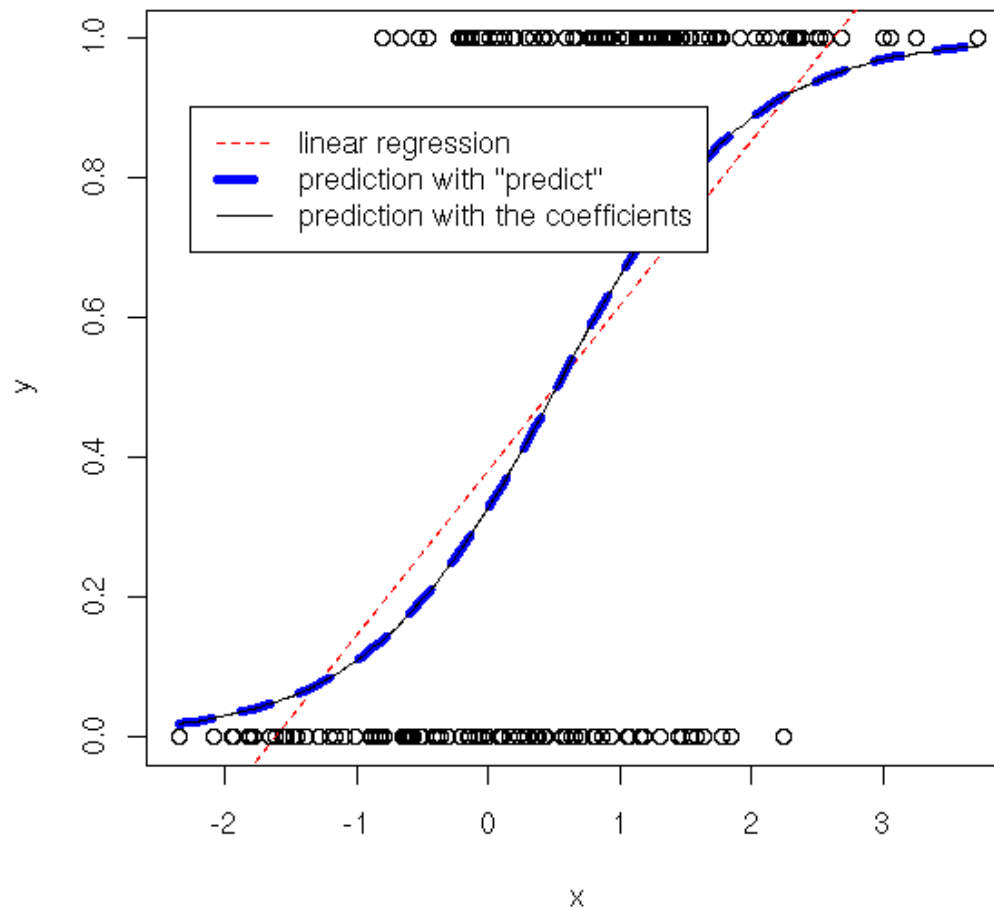
GENERALIZED ADDITIVE MODELS (GAM)

Characteristics

- No underlying distribution is assumed

- Driven entirely by data

- Uses a link function, $g(\mu)$, to transform the mean of the response to a smoothed function of the predictor variables

- Basic model form:

$$g(\mu) = \alpha + \sum s_i(\mathbf{X})_i$$

Little trickier here ...

$P[\text{Lodgepole pine}]$

$\mathbf{y}$

$s_i(x_j)$ Aspect

GENERALIZED ADDITIVE MODELS (GAM)

A Visual Specification Test...

Daily S&P 500 Logarithmic Returns

Probability of positive return tomorrow?

Today's Return

Logisitic predictions in black; Spline fit in blue; GAM fit in red
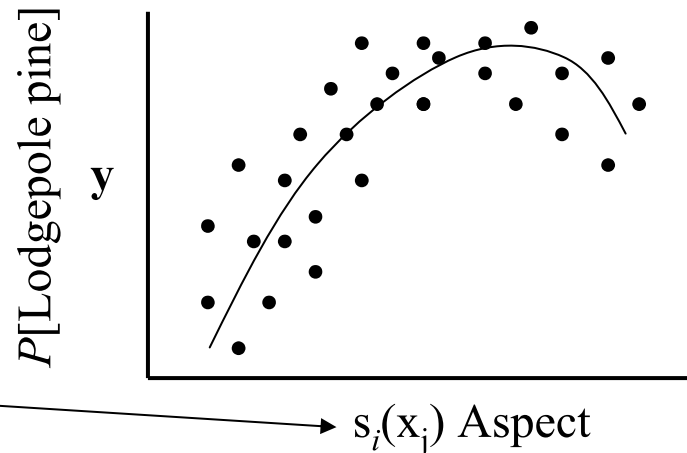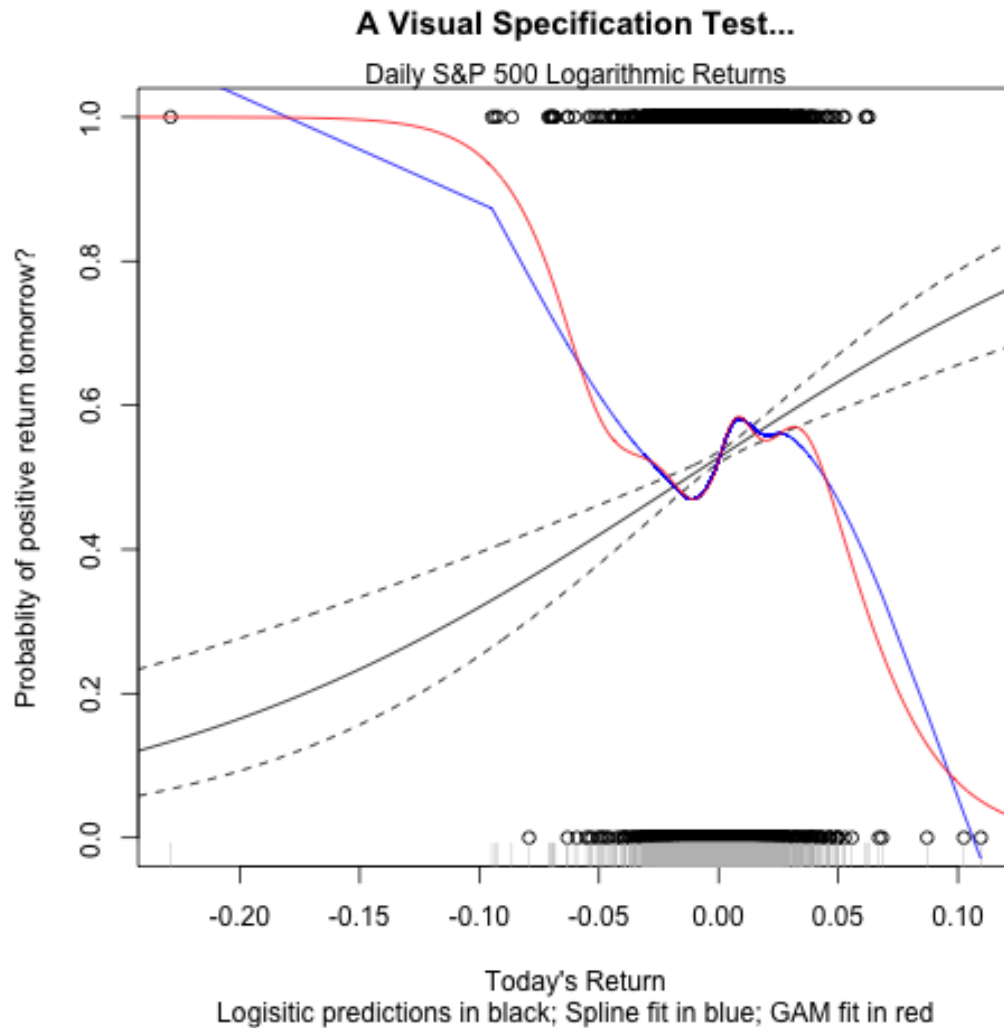
GENERALIZED ADDITIVE MODELS (GAM)

"Semi-parametric" extension of GLM

Very flexible, it can fit very complex models

Main limitation: GAM cannot be used to calculate species response parameters

Good for exploratory analysis of responses

Edwards et al., ECOCHANGE, , Lausanne, Sep 2009

Example of applications in SDMs
Prediction surface for mullein, Lava Beds NM
(From Edwards 2009)
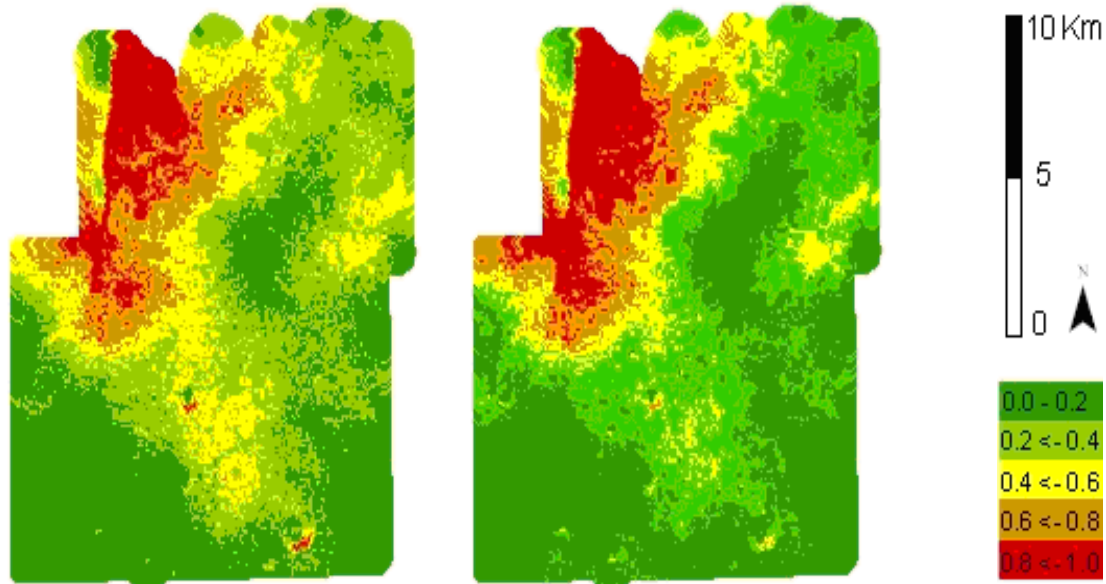
Additive Logistic
Model (GAM)

Logistic Model
(GLM)

Machine learning techniques

◆ They consider presence data and absence or **background** data

◆ Also called ´data mining´methods

◆ Based on information theory and artificial inteligence

◆ Probablistic distributions are infered from incomplete data

◆ Boosted Regresion Trees (RT), Random Forests (RF), MaxEnt

◆ Others: Artificial Neural Networks, Genetic algorithm (GARP)

Why machine learning systems?

SDMs can be considered as a supervised learning problem

In statistical inference you must decide distributional form and parametres

ML methods learn the function **inductively** from training data

Inductive (supervised) machine learning methods are commonly used in artificial inteligence (computer vision, robotics, medical diagnosis, etc)

Boosted Regression Trees

(= "Gradient Boost", "Stochastic Gradient Boosting")
It uses a technique of boosting (a regression-like method based on Machine learning) to combine large numbers of simple tree models adaptively, to optimize predictive performance

Elith, J., J.R. Leathwick and T. Hastie, 2009. A working guide to boosted regression trees. Journal of Animal Ecology 77: 802-81

Implemented in:
R (dismo, gbm)

Boosted Regression Trees

Ensemble models based on a LARGE number of tree models
Computationally intensive methods
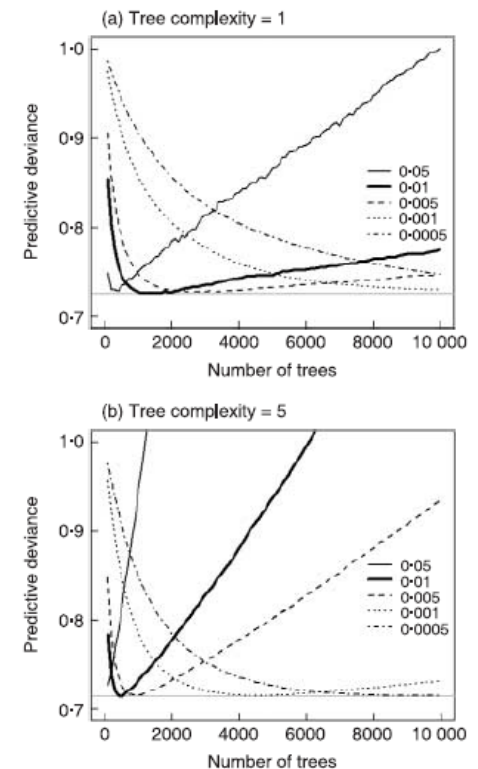
Random Forests

Random Forest (Breiman, 2001) method is an extension of Classification and regression trees (CART; Breiman et al., 1984) It builds a multitude of decision trees at training time and outputting the class that is the mode of the classes output by individual trees

Breiman, L., 2001. Random Forests. Machine Learning 45: 5-32

.

Implemented in:
R (randomForest), openmodeller

Random Forests

Not a real models like in logistic regression
Higher prediction accuracy than ordinary decision trees
No graphical tree model like in classification trees
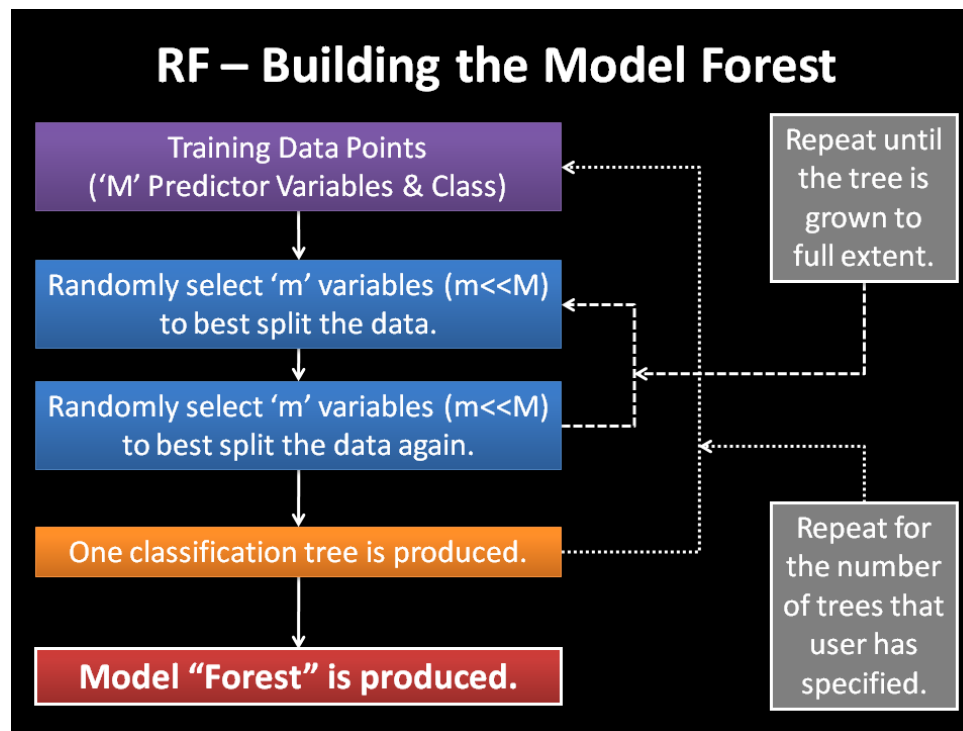
- One measure of predictor relationships is the *variable importance plot*

- Portrays each predictor as a function of mean decrease in accuracy
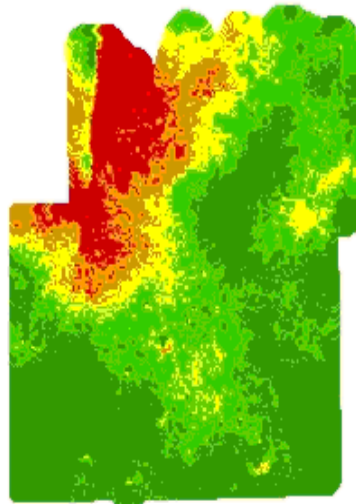
- No agreed on amount of "decrease" as important
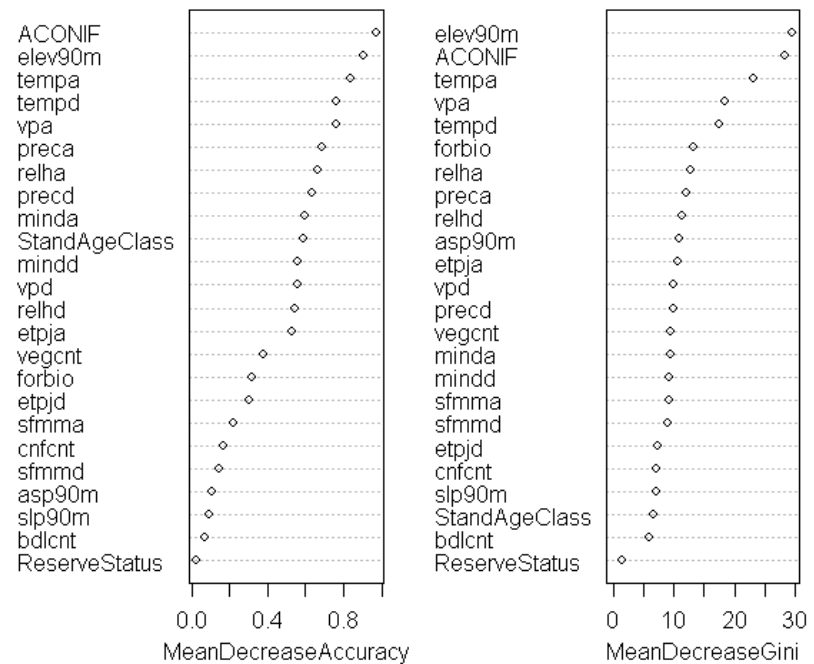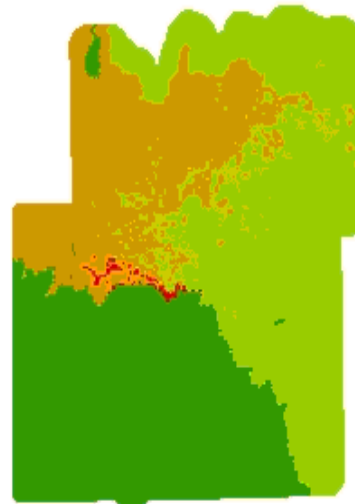

Variable importance plots

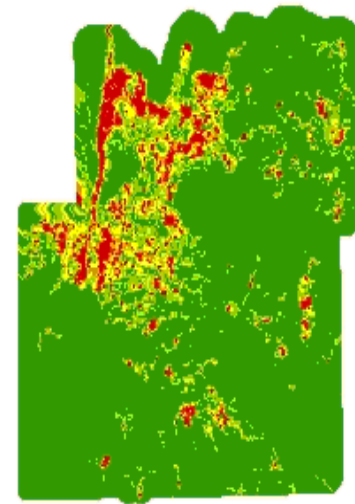Different model forms inevitably lead to different spatial depictions -- you decide!!
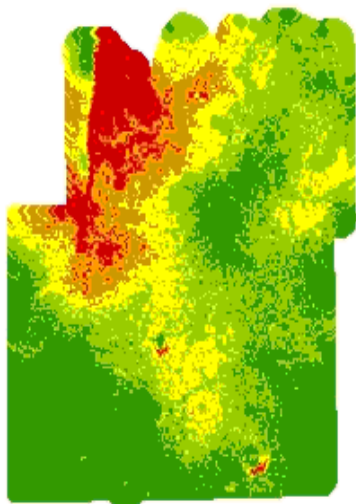
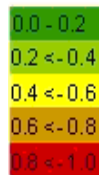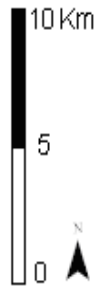Additive Logistic Model (GAM)   Logistic Model (GLM)   Classification Tree Model   Random Forest Model

From Edwards (2009)

	Logistic	GAM	CART	RF
PCC	75.5	75.9	77.6	87.7
Specificity	74.5	74.4	76.8	89.4
Sensitivity	84.1	90.3	86.5	71.0
kappa	0.29	0.32	0.35	0.46
AUC	0.88	0.89	0.85	0.88

From Edwards (2009)

MAXENT

MaxEnt (Maximum Entropy; Phillips et al., 2006) is the most widely used SDM algorithm. Is a general-purpose machine learning method with precise mathematical formulation. It uses presences and absences (generating a **background** layer to fit the model)

Phillips SJ et al. 2006. Maximum entropy modeling of species geographic distributions. Ecological Modelling 190 231-259
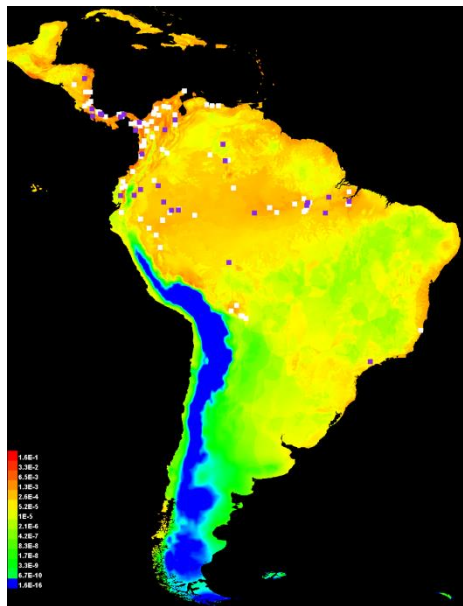
Implemented in:
R (dismo, others), Stand-alone Java program

MAXENT

Main statement: a probability distribution with maximum entropy (the most spread out, closest to uniform), subject to known constrains, is the best approximation of an **unknown distribution.** It combines machine learning with statistical methods
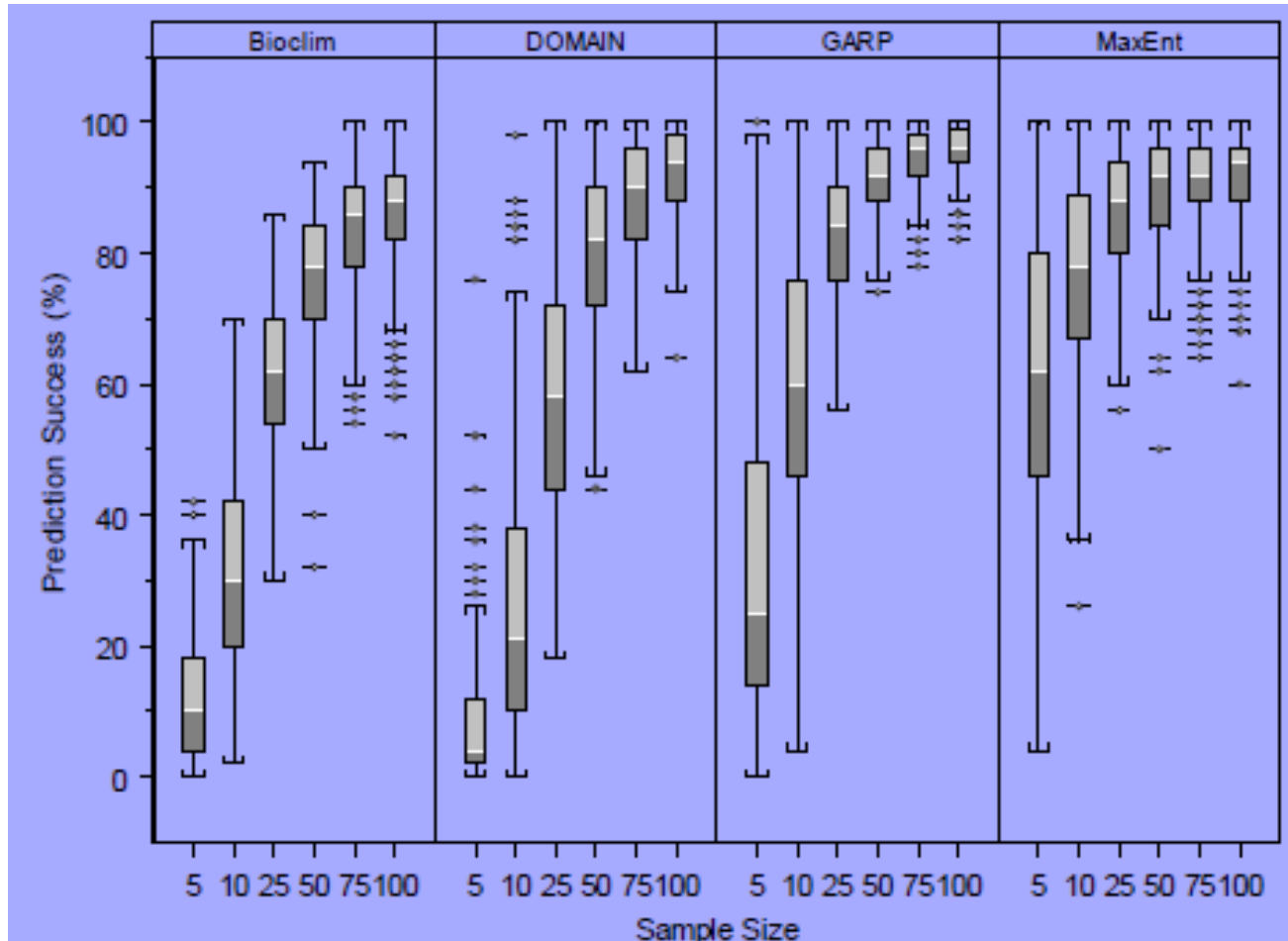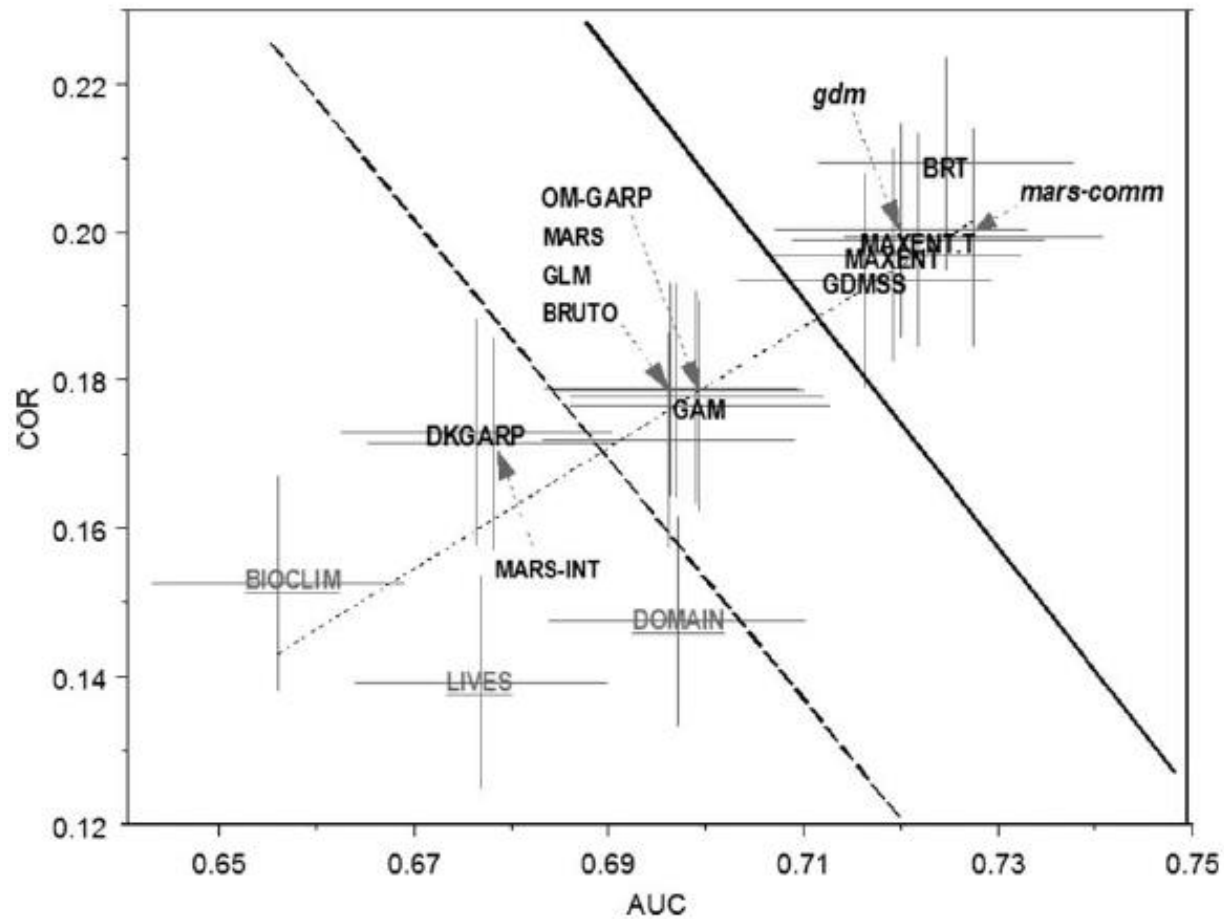
Exponential output
(difficult to intepret)

Logistic
output (0 - 1)
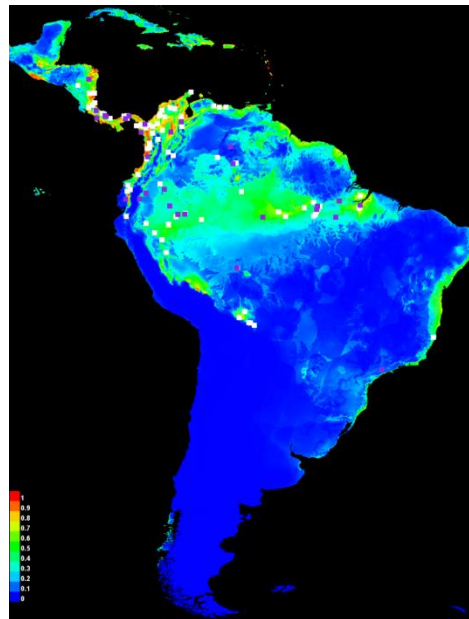
Model comparisons

Elith et al. 2006. Novel methods improve prediction of species' distributions from occurrence data. *Ecography* 29: 129 -151

NEXT WEEK… PRACTICE WITH MAXENT