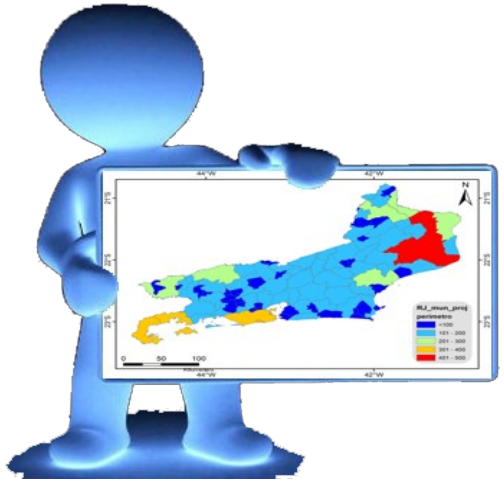
INSTITUTO INTERNACIONAL PARA SUSTENTABILIDADE

INTERNATIONAL INSTITUTE FOR SUSTAINABILITY

IIS

Cartografia & Saúde: Análise geoespacial como ferramenta aplicada na parasitologia

Modelagem de Nicho Ecológico



Diogo S. B. Rocha

TIPOS DE ALGORITMOS DE MODELAGEM

Envelopes BioClimáticos	Ajustes Estatísticos	Int. Artificial e Busca

AQUARIO

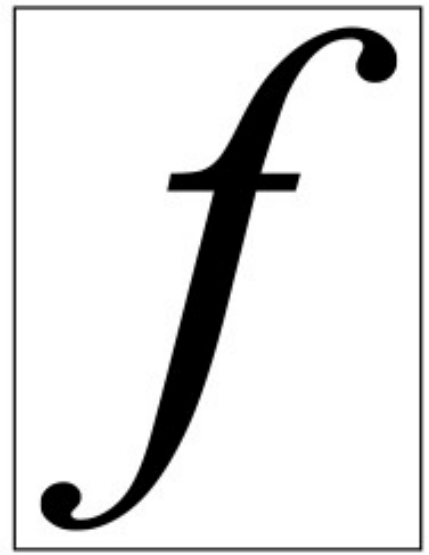
CAIXA PRETA

COFRE

Mod. Distribuição de Espécies

- Bioclim
- Dist. Euclideana
- Dist. Mahalanobis
- Dist. Gower
- ENFA

Envelopes e Distância



- GLM
- GAM
- FDA
- GBM
- MARS

Ajuste Estatístico

- GARP
- MaxEnt
- Random Forest
- Redes Neurais

IA e Busca

Algoritmos de modelagem

- Um dos primeiros algoritmos de modelagem: BioClim.
- Lembram do nicho ecológico?

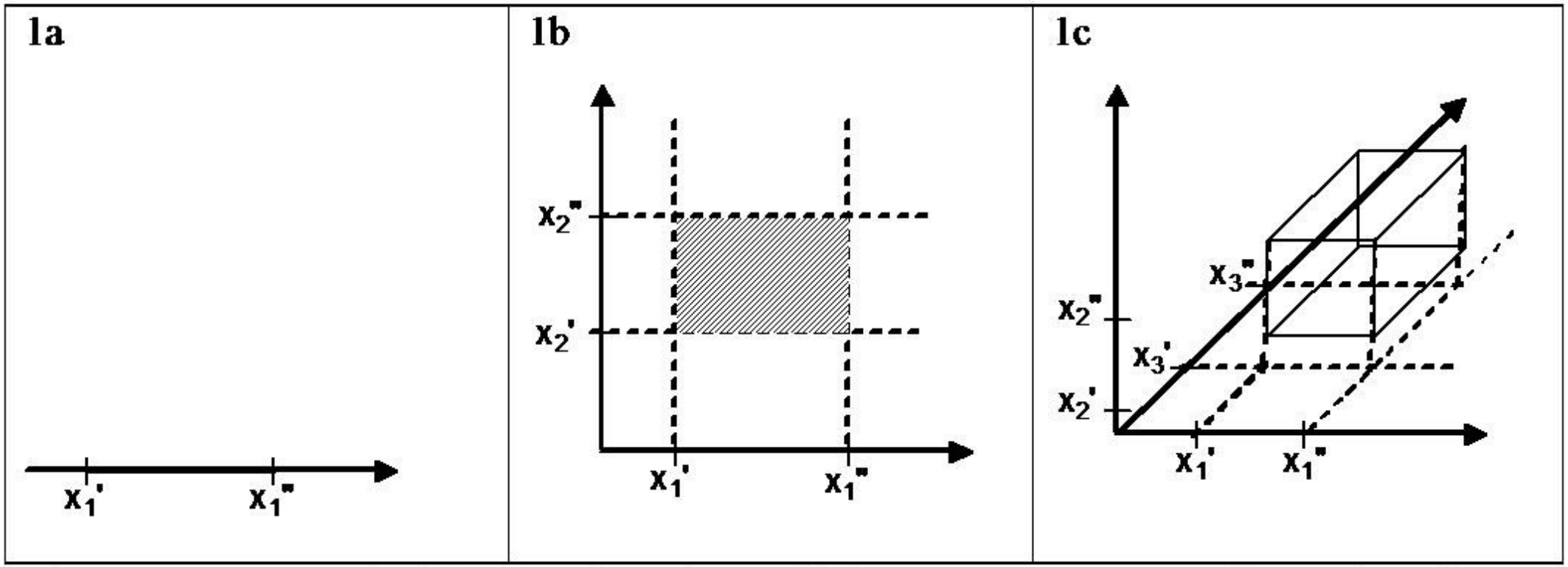


Figura 1: Esquema da definição de nicho ecológico proposta por Hutchinson (1957), para uma (1a), duas (1b) e três (1c) dimensões (variáveis).

Modelagem baseada no nicho

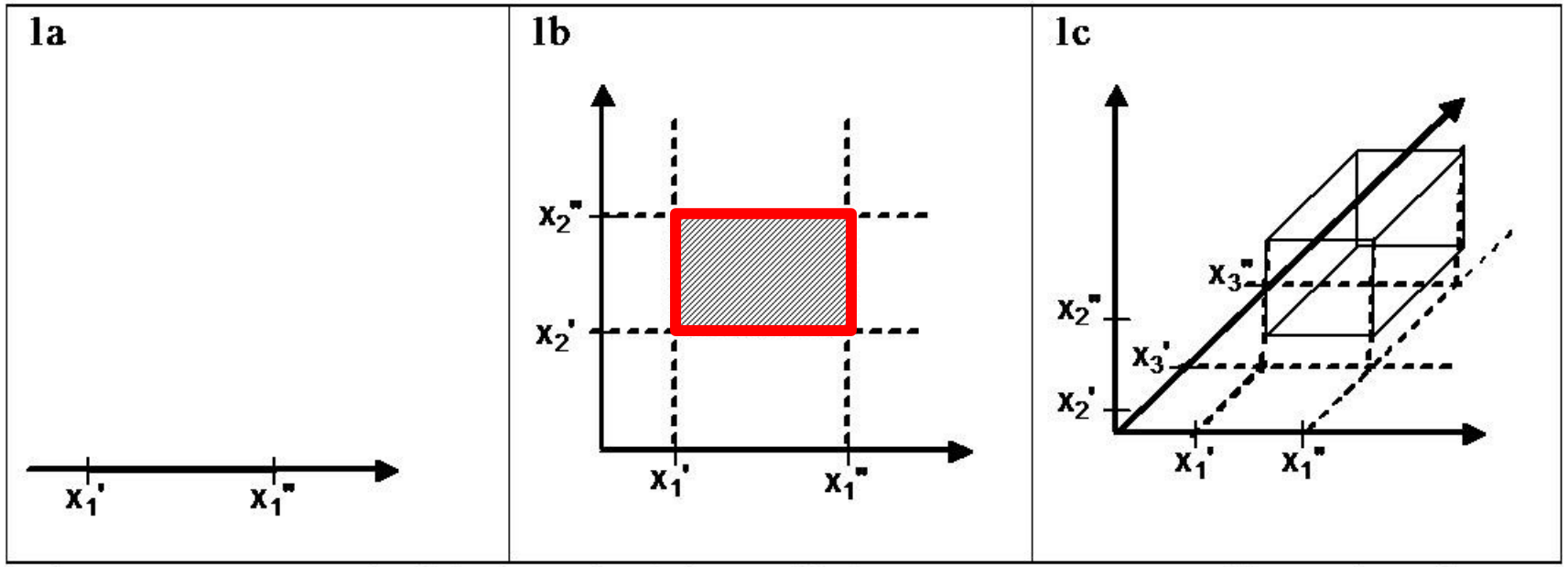


Figura 1: Esquema da definição de nicho ecológico proposta por Hutchinson (1957), para uma (1a), duas (1b) e três (1c) dimensões (variáveis).

Várias formas de se calcular o nicho...algoritmos

Modelagem baseada no nicho

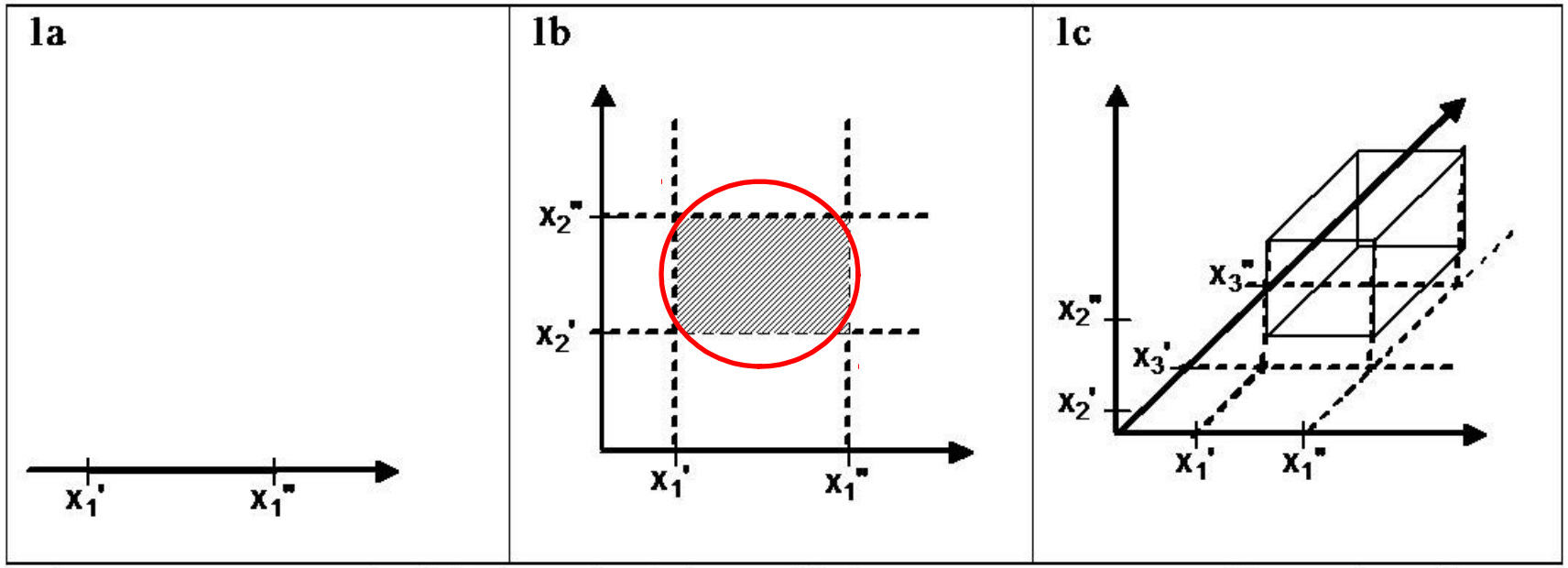


Figura 1: Esquema da definição de nicho ecológico proposta por Hutchinson (1957), para uma (1a), duas (1b) e três (1c) dimensões (variáveis).

Modelagem baseada no nicho

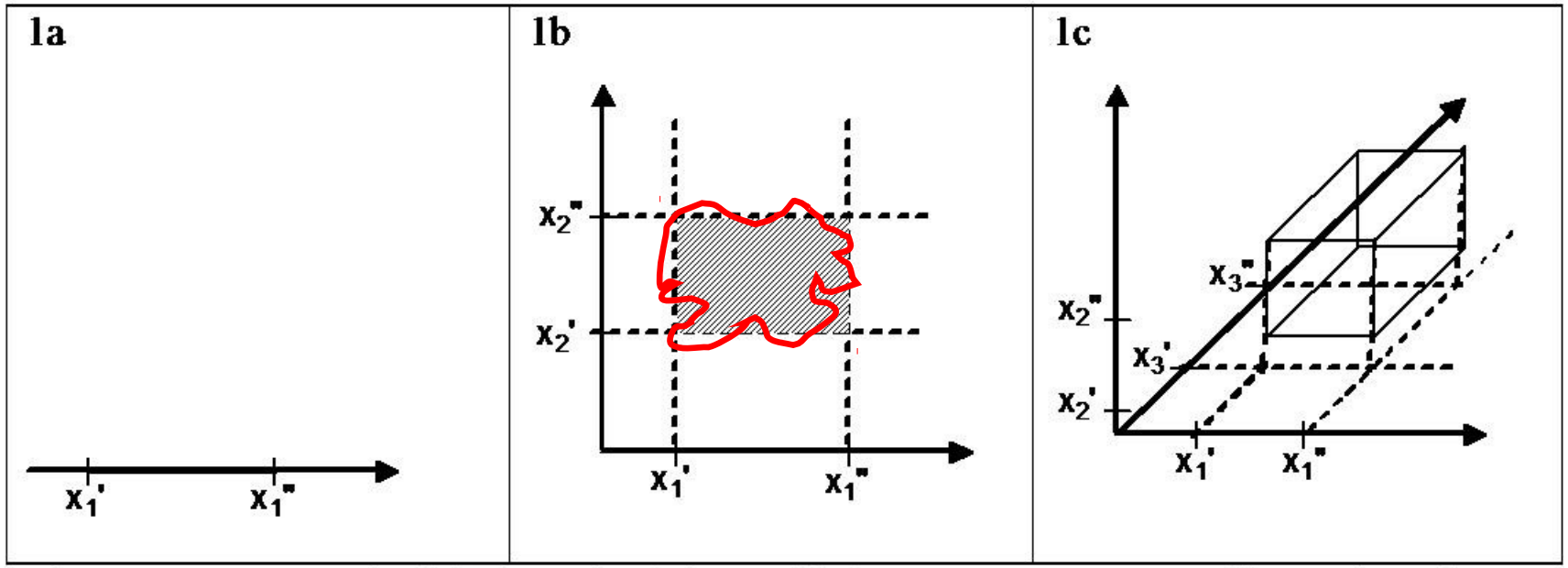
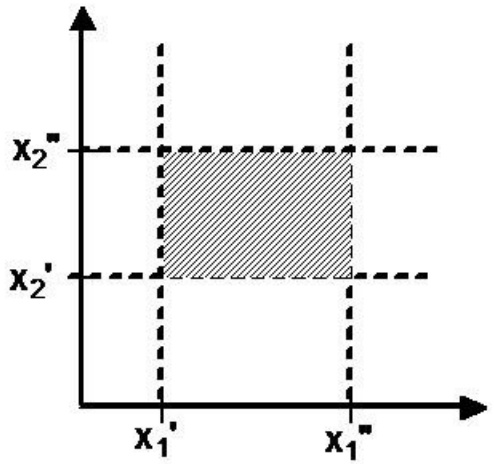


Figura 1: Esquema da definição de nicho ecológico proposta por Hutchinson (1957), para uma (1a), duas (1b) e três (1c) dimensões (variáveis).

Envelopes BioClimáticos

- Modelos de **baixa** complexidade
- Baseiam-se em mecanismos ecológicos conhecidos (e.g. tolerância termal)
- **Não** exigem dados de ausência
- **Fácil** interpretação e comunicação
- Em geral possuem **baixo** ajuste aos dados observados
- Possuem **alta** "transferibilidade"



Envelopes Bioclimáticos

- Para cada variável ambiental: média e o desvio padrão, valores máximo e mínimo.

Cada pixel pode ser classificado como:

- Habitável: se todos os valores ambientais estiverem dentro do envelope calculado -> 1
- Tolerável: se um ou mais valores ambientais estiverem fora do envelope da média e desvio padrão mas dentro dos limites máximo e mínimo ->0.5
- Inabitável: se um ou mais valores associados estiverem fora dos valores limites máximos e mínimos das variáveis ambientais. ->0

  (modelo categórico)

Envelopes Bioclimáticos

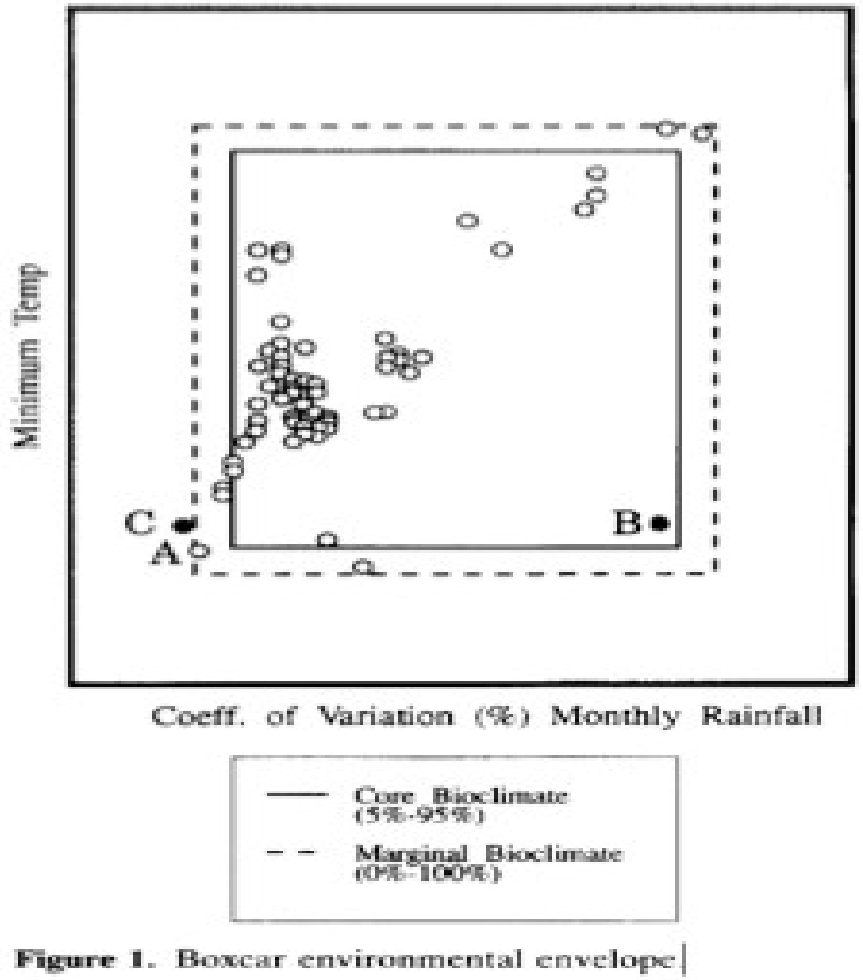
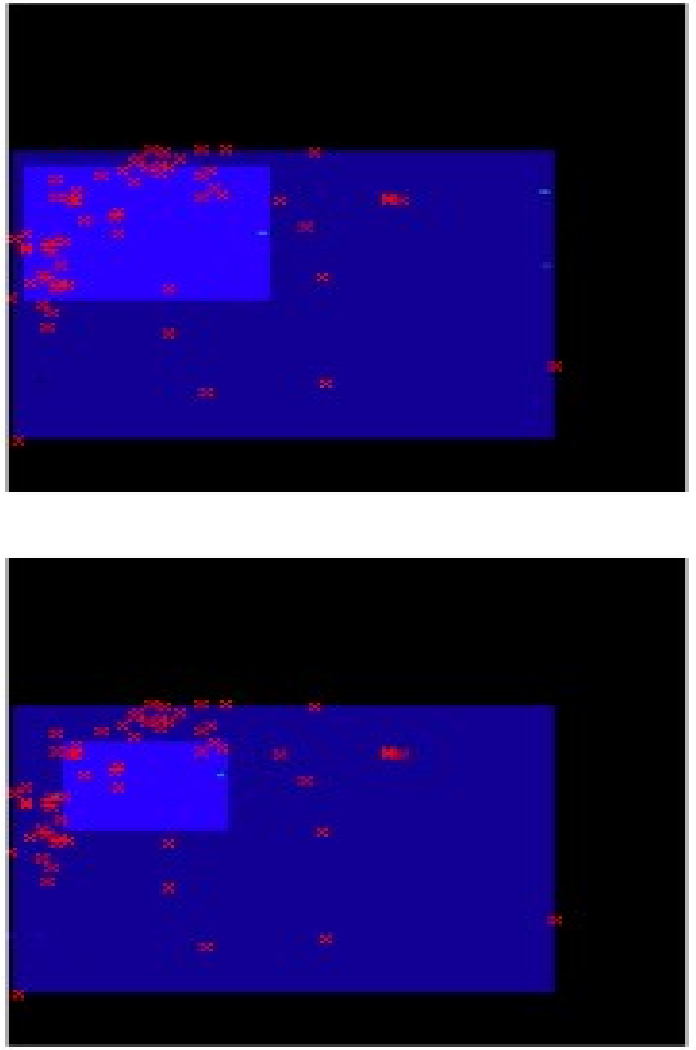


Figure 1. Boxcar environmental envelope

Retirado de Carpenter et al. 1993

BIOCLIM

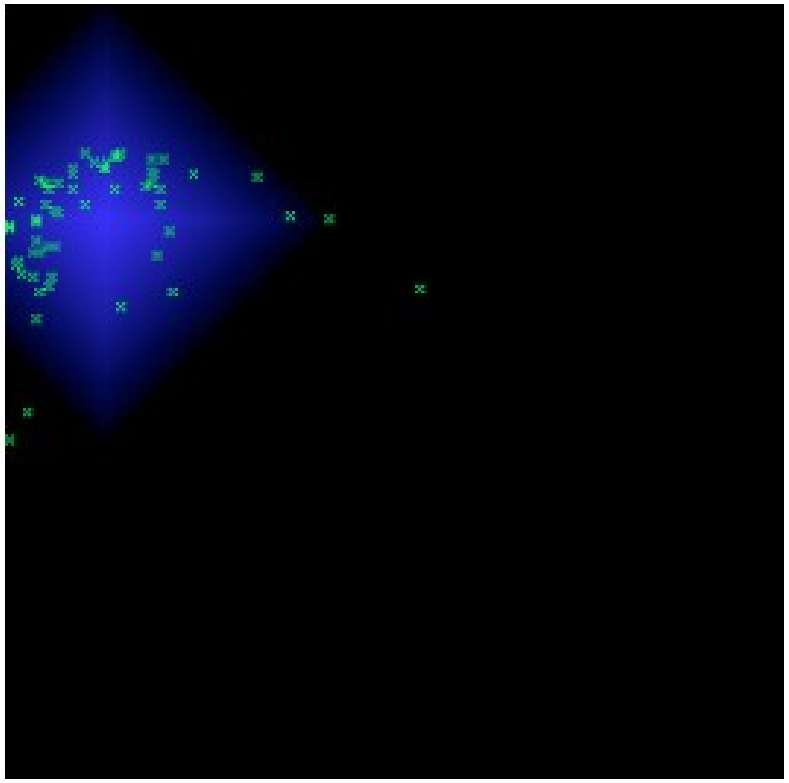


Retirado de http://openmodeller.sourceforge.net/

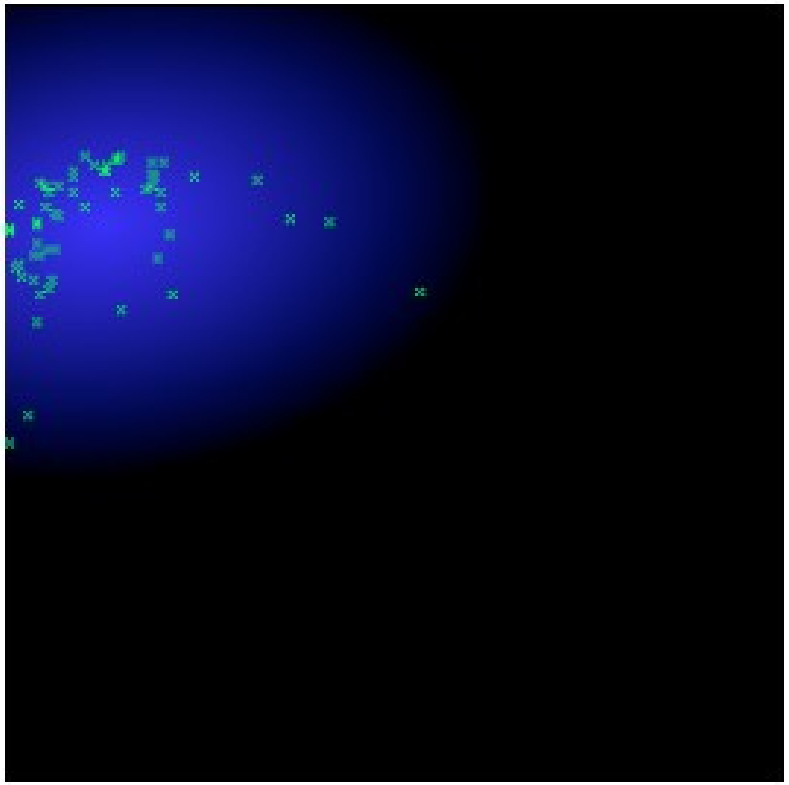
BIOCLIM e BIOCLIM cortado pelo desvio-padrão

Distância Ambiental

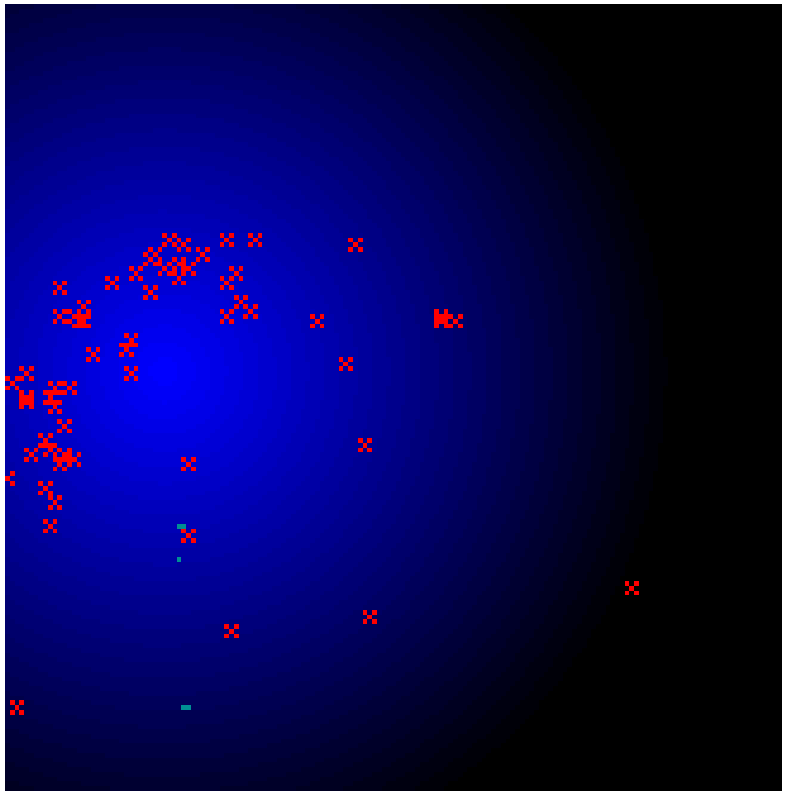
Calculada para o centroide ambiental da distribuição



Gower - DOMAIN

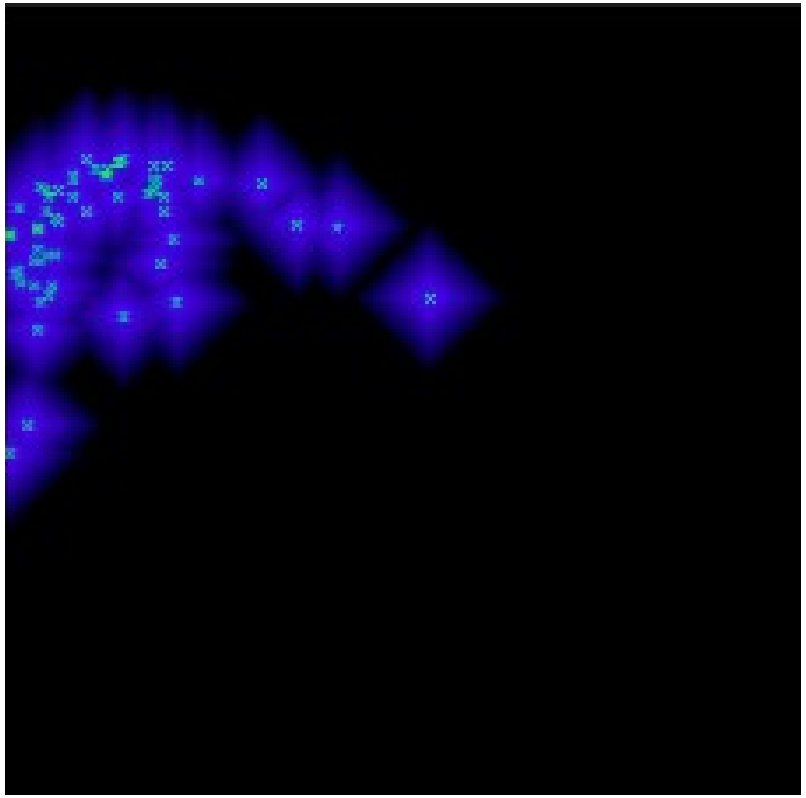


Mahalanobis

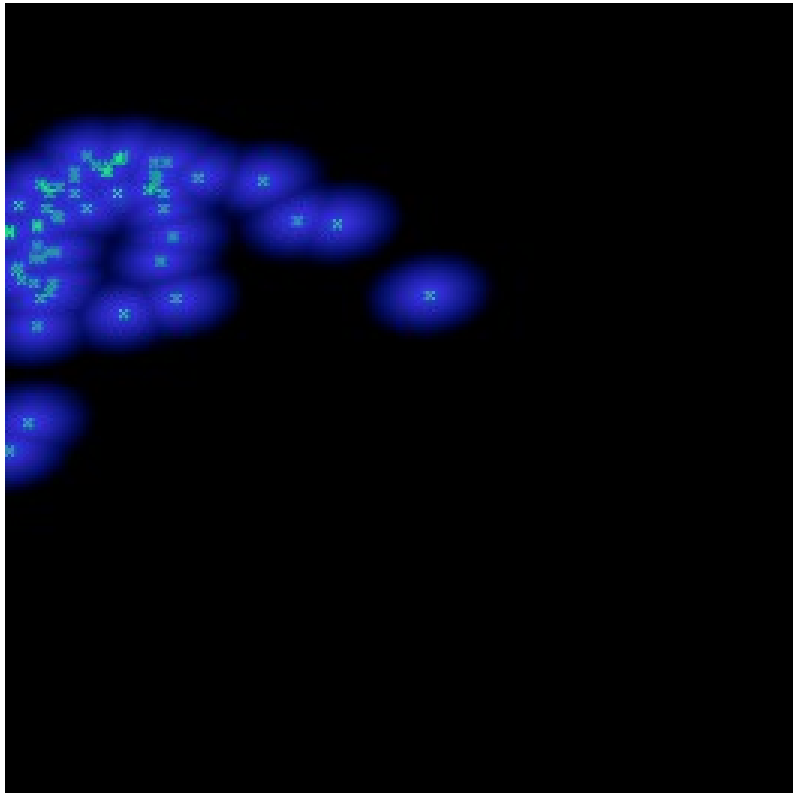


Euclidiana

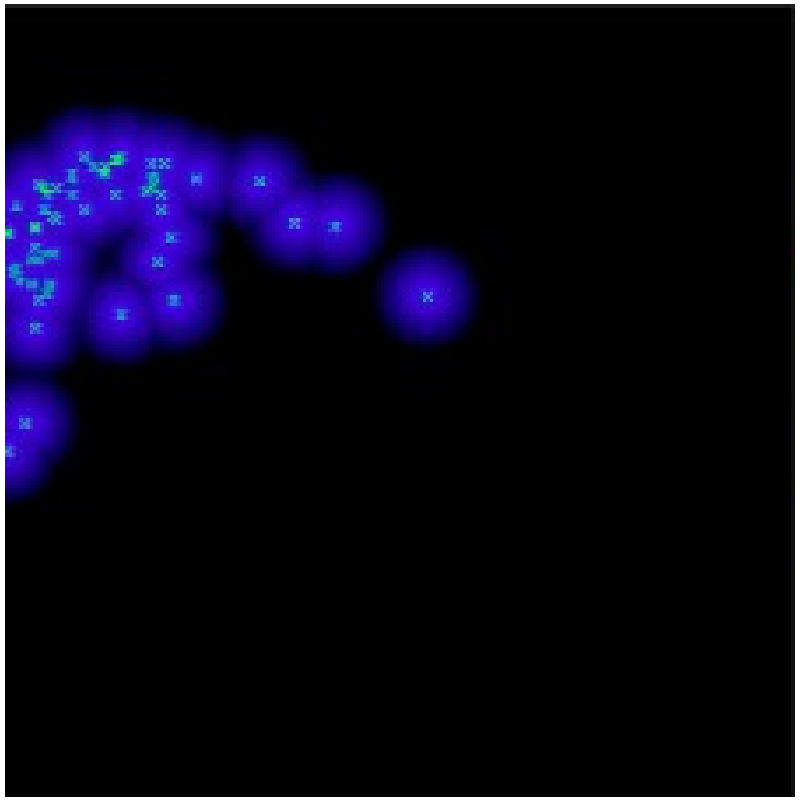
Distância Ambiental



Métrica Gower - DOMAIN



Métrica Mahalanobis



Métrica Euclidiana

Ajustes Estatísticos

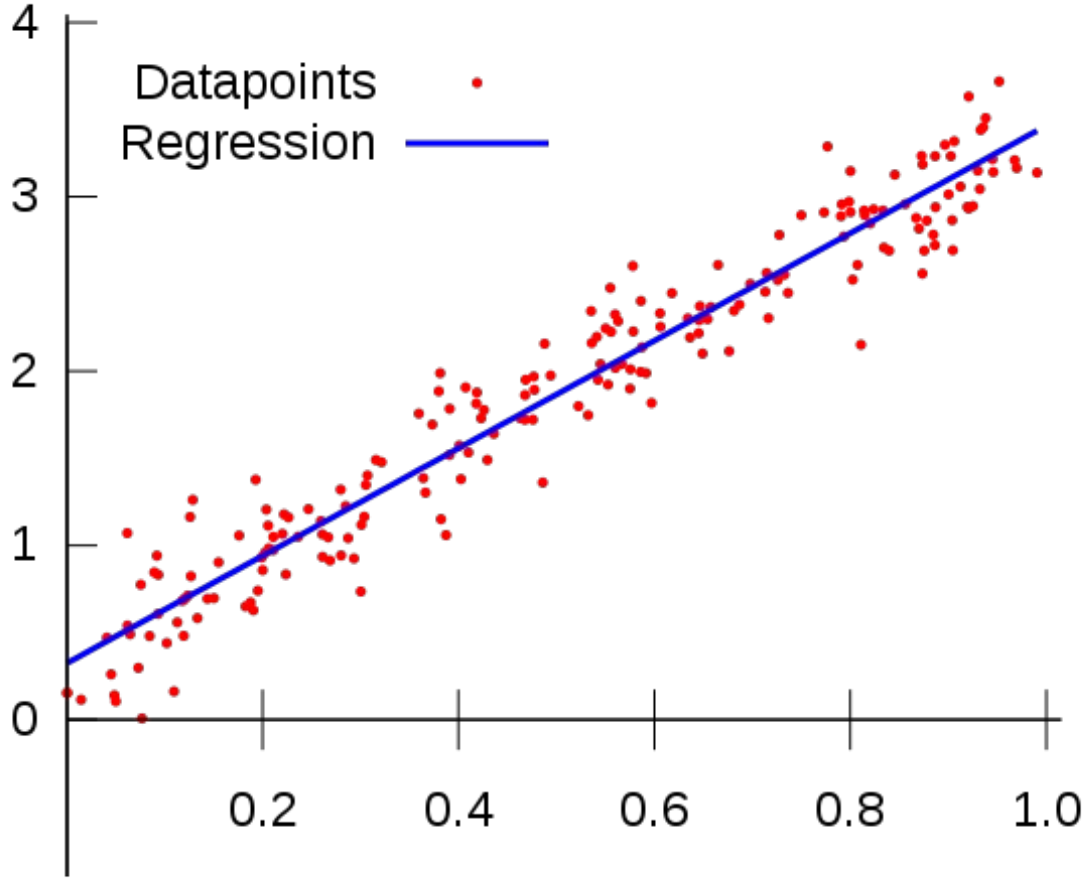
- Modelos de **média** complexidade
- Alguns baseiam-se em mecanismos ecológicos possivelmente aceitos (relações lineares e interações)
- Exigem ausências **verdadeiras**
- **Alguns** podem ser interpretados
- Em geral apresentam **bons** ajustes aos dados observados
- **Boa** "transferibilidade"

Rangel, T.F. 2011. Workshop Modelagem

Ajustes Estatísticos

R**egressão linear**: método para se estimar o valor de uma variável *y*, dados os valores de algumas outras variáveis *x*.

A regressão linear é chamada "linear" porque se considera que a relação da resposta às variáveis é uma função linear de alguns parâmetros.



Modelos de ajuste estatístico

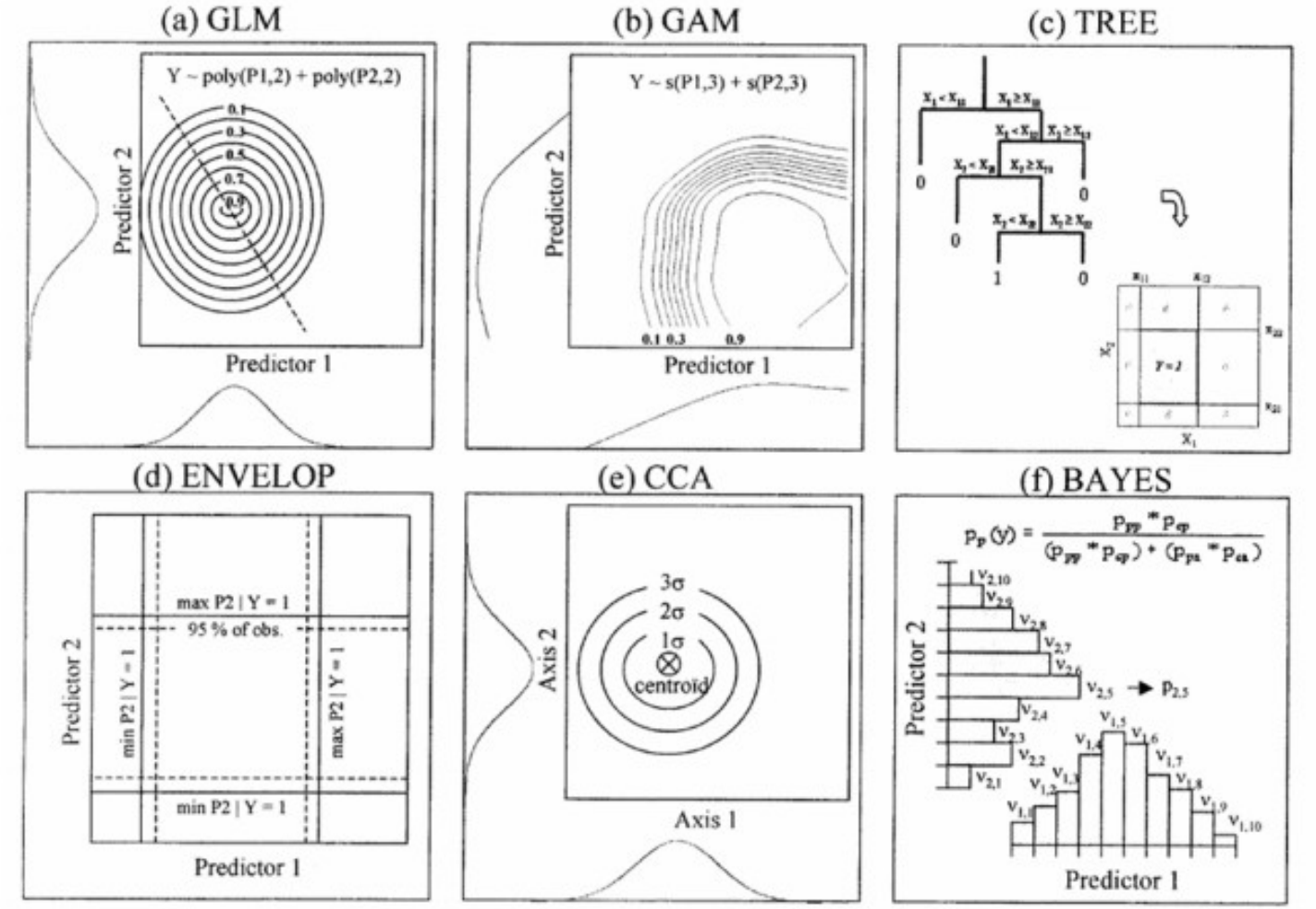


Fig. 6. Examples of response curves for different statistical approaches used to model distribution of plants and vegetation. (a) Generalized linear model with second order polynomial terms; (b) generalized additive model with smoothed spline functions; (c) classification tree; (d) environmental envelope of the BIOCLIM type; (e) canonical correspondence analysis; (f) Bayesian modeling according to Aspinall (1992); p_p = posterior probability of presence of the modeled species, p_{pp} = a priori probability of presence, p_{pa} = a priori probability of absence, p_{cp} = product of *conditional* probability of presence of the various predictor classes, p_{ca} = product of *conditional* probability of absence of the various predictor classes.

- Assumem ausências verdadeiras
- Média complexidade
- Bom ajuste em geral
- Boa “transferibilidade”
- Interpretação relativamente fácil: relações lineares

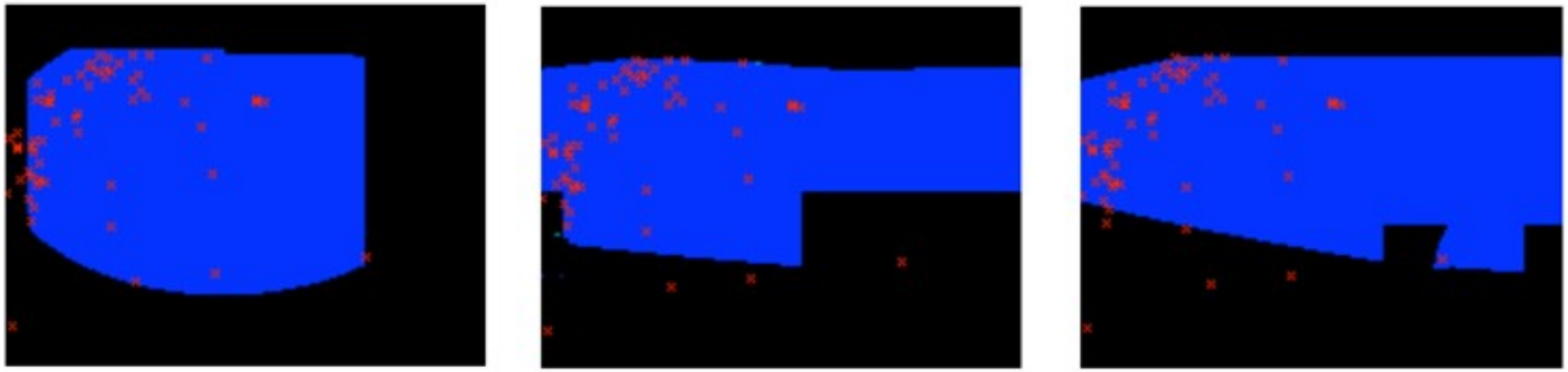
Retirado de Guisan e Zimmermann 2000

Inteligência Artificial e Busca

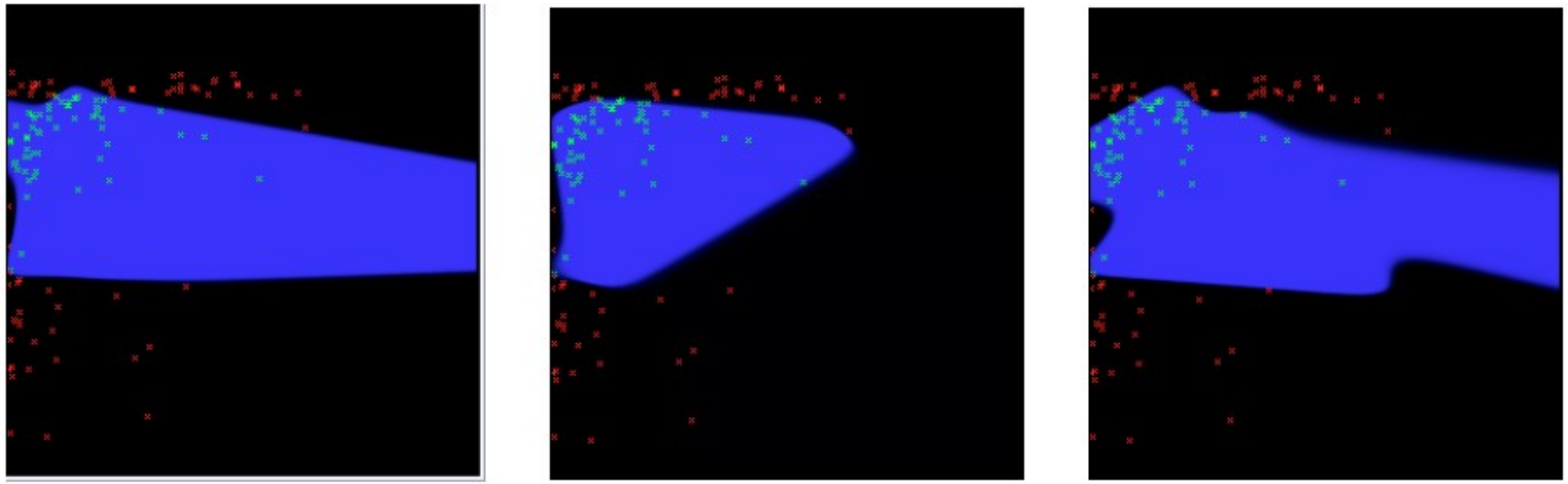
- Modelos de **altíssima** complexidade
- **Não** são baseados em nenhum mecanismos ecológicos
- Exigem **ausências verdadeiras**
- **Não** podem ser interpretados
- Apresentam **excelentes** ajustes aos dados observados
- **Péssima** "transferibilidade"

Rangel, T.F. 2011. Workshop Modelagem

INTELIGÊNCIA ARTIFICIAL



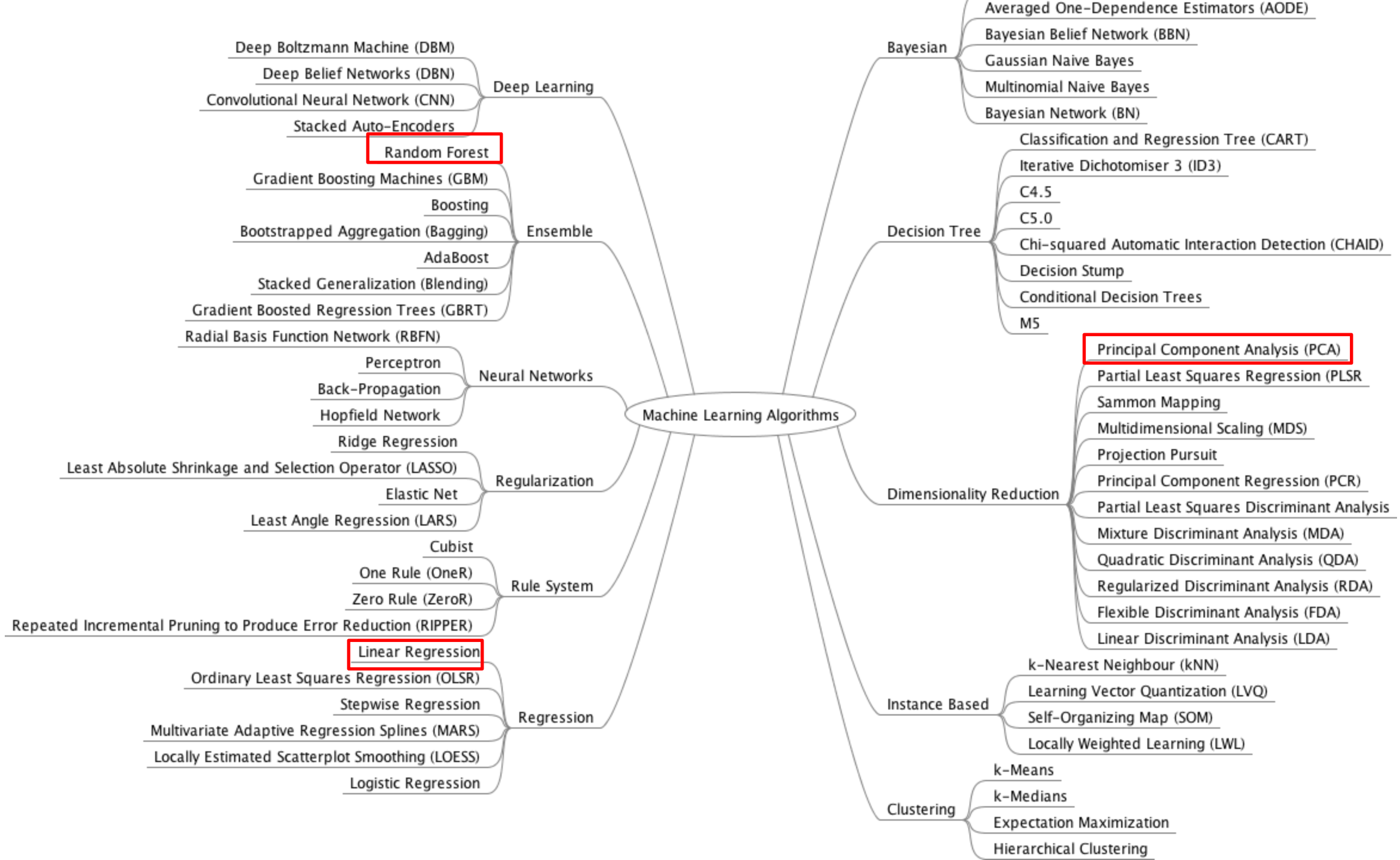
GARP – modelo não determinístico – mesmos dados iniciais, diferentes iterações



Rede neural – modelo não determinístico – mesmos dados iniciais, diferentes iterações

Algoritmos de modelagem

Aprendizagem de máquina


Machine Learning Algorithms
Bayesian
Averaged One-Dependence Estimators (AODE)
Bayesian Belief Network (BBN)
Gaussian Naive Bayes
Multinomial Naive Bayes
Bayesian Network (BN)
Decision Tree
Classification and Regression Tree (CART)
Iterative Dichotomiser 3 (ID3)
C4.5
C5.0
Chi-squared Automatic Interaction Detection (CHAID)
Decision Stump
Conditional Decision Trees
M5
Dimensionality Reduction
Principal Component Analysis (PCA)
Partial Least Squares Regression (PLSR
Sammon Mapping
Multidimensional Scaling (MDS)
Projection Pursuit
Principal Component Regression (PCR)
Partial Least Squares Discriminant Analysis
Mixture Discriminant Analysis (MDA)
Quadratic Discriminant Analysis (QDA)
Regularized Discriminant Analysis (RDA)
Flexible Discriminant Analysis (FDA)
Linear Discriminant Analysis (LDA)
Instance Based
k-Nearest Neighbour (kNN)
Learning Vector Quantization (LVQ)
Self-Organizing Map (SOM)
Locally Weighted Learning (LWL)
Clustering
k-Means
k-Medians
Expectation Maximization
Hierarchical Clustering
Deep Learning
Deep Boltzmann Machine (DBM)
Deep Belief Networks (DBN)
Convolutional Neural Network (CNN)
Stacked Auto-Encoders
Ensemble
Random Forest
Gradient Boosting Machines (GBM)
Boosting
Bootstrapped Aggregation (Bagging)
AdaBoost
Stacked Generalization (Blending)
Gradient Boosted Regression Trees (GBRT)
Neural Networks
Radial Basis Function Network (RBFN)
Perceptron
Back-Propagation
Hopfield Network
Regularization
Ridge Regression
Least Absolute Shrinkage and Selection Operator (LASSO)
Elastic Net
Least Angle Regression (LARS)
Rule System
Cubist
One Rule (OneR)
Zero Rule (ZeroR)
Repeated Incremental Pruning to Produce Error Reduction (RIPPER)
Regression
Linear Regression
Ordinary Least Squares Regression (OLSR)
Stepwise Regression
Multivariate Adaptive Regression Splines (MARS)
Locally Estimated Scatterplot Smoothing (LOESS)
Logistic Regression

Aprendizagem de Máquina

Trabalharemos dois grupos de algoritmos principalmente:

1) Agrupamento de algoritmos pelo estilo de aprendizagem.
   Existem três tipos de aprendizagem:
   a) supervisionada
   b) não supervisionada
   c) semi-supervisionada

2) agrupamento de algoritmos por semelhança na forma ou função.
   Existem vários tipos, alguns são:
   d) Regressão
   e) Classificação

Mas a maioria é hibrida, podendo ter diferentes funções associadas a um ou mais destes tipos de aprendizagem.

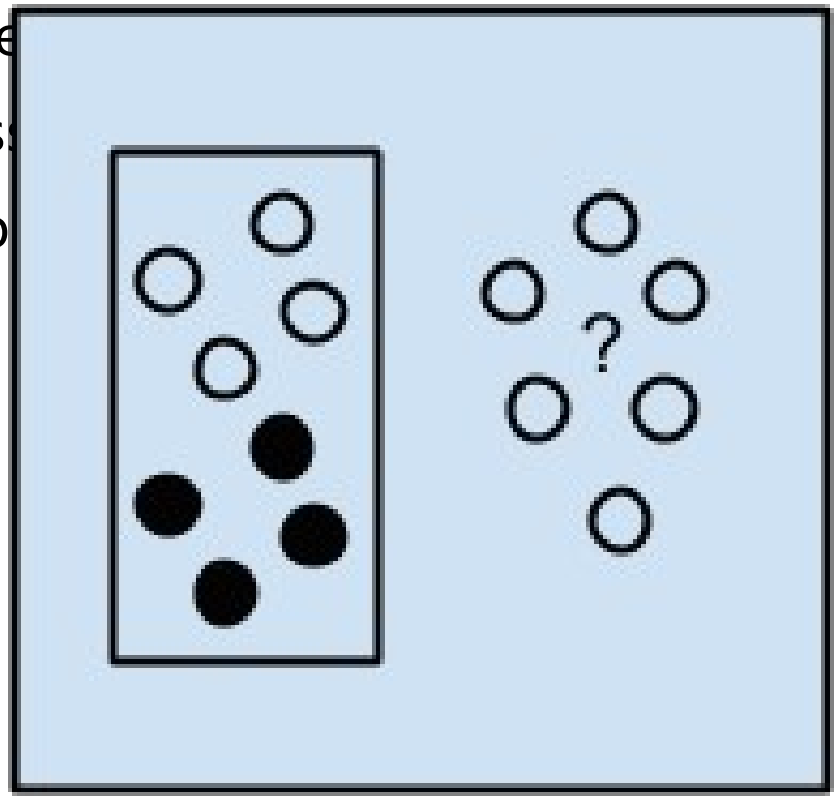
Aprendizagem supervisionada:

Dados de entrada (treinamento) são conhecidos (ex presença)

O modelo é gerado através de um processo de treinamento no qual se faz previsões que são corrigidas quando estão erradas (conjunto de treino e conjunto de teste). O processo continua até que o modelo alcance o nível desejado de pre

Exemplos: algoritmos que aplicam a classificação e regres

Incluem Regressão Logística e Rede Neural de Propagação



Supervised Learning Algorithms

Aprendizagem **não** supervisionada:

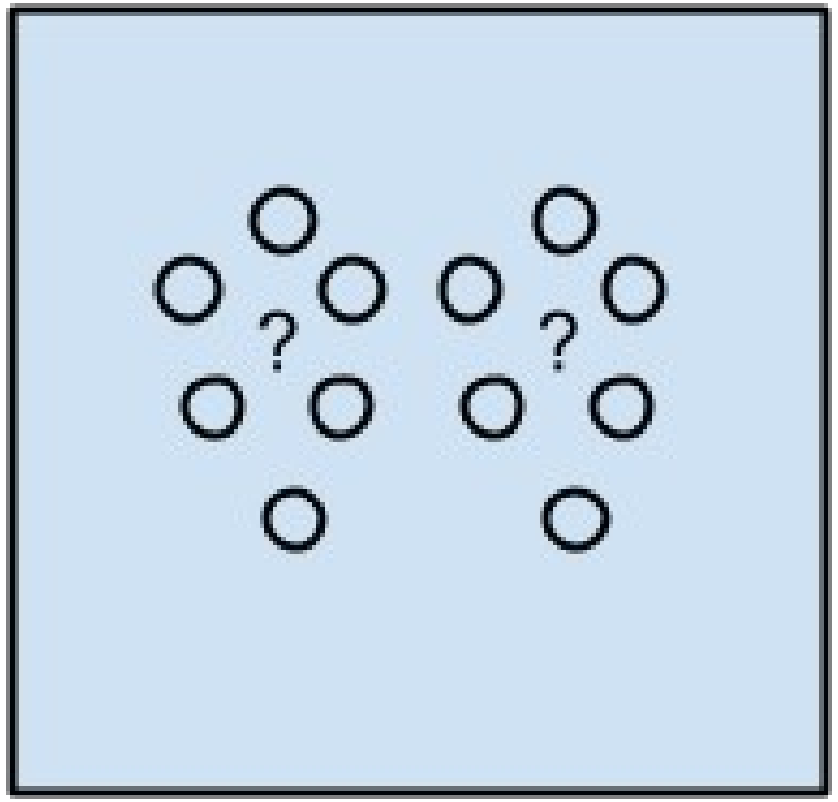
Dados de entrada não são marcados e não existe um resultado conhecido.

O modelo é gerado através da dedução das estruturas presentes nos dados de entrada, extração de regras. Pode ser através de um processo matemático para reduzir sistematicamente a redundância, ou pode ser para organizar os dados por semelhança entre eles.

Exemplos:

- clustering,
- redução de dimensionalidade (PCA)
- e regras de associação.

Incluem algoritmo Apriori e K-Means.



Unsupervised Learning Algorithms

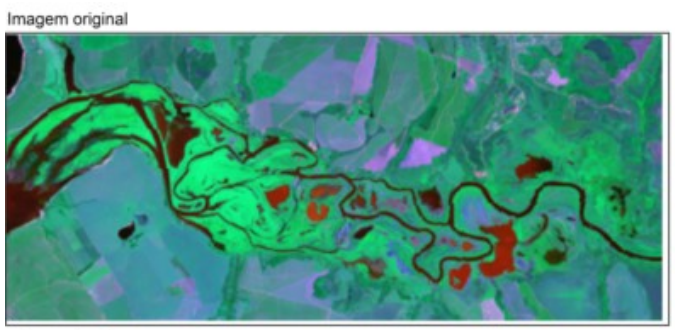
Aprendizagem semi-supervisionada

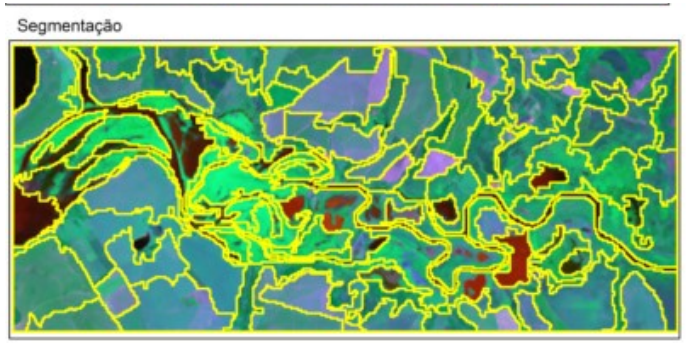
Dados de entrada misturam exemplos rotulados e não rotulados.

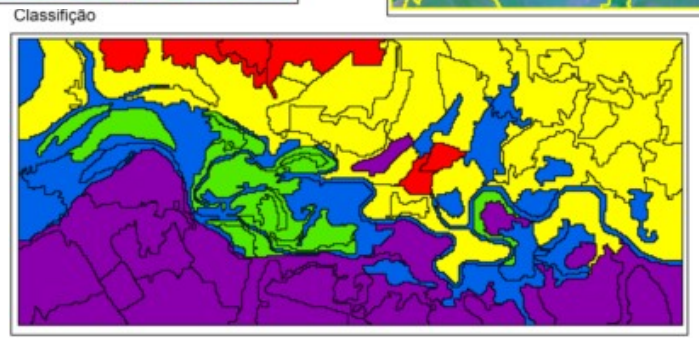
Há um problema de predição desejado mas o modelo deve aprender as estruturas para organizar os dados, bem como fazer previsões.

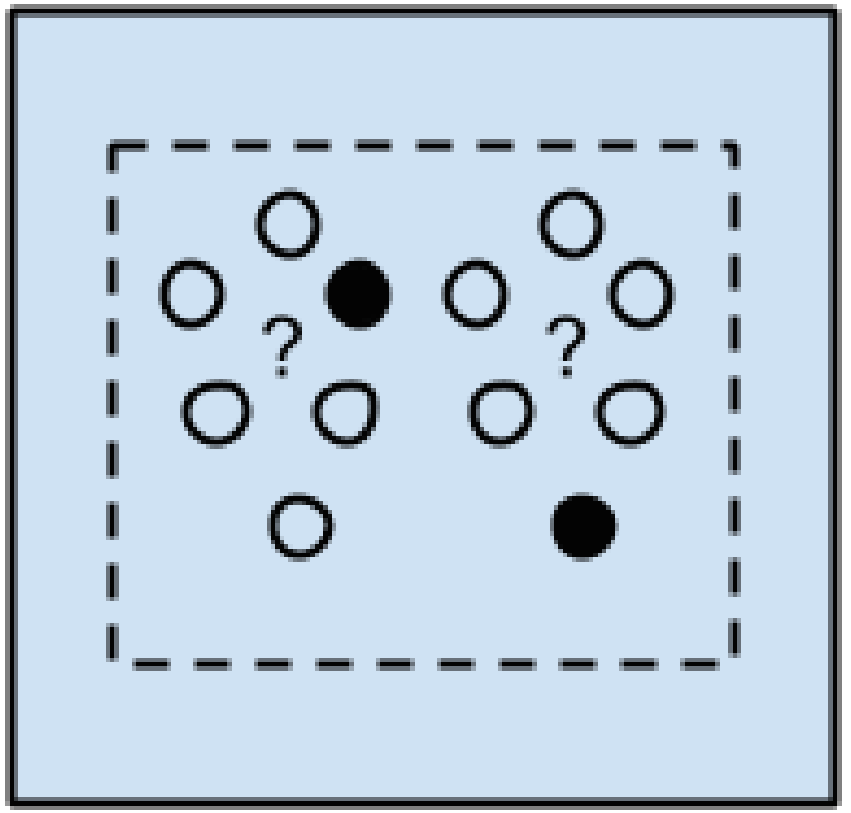
Exemplo: classificação e regressão.

Não tem algoritmos exclusivo, são extensões para outros métodos flexíveis que fazem suposições sobre como modelar os dados não rotulados.









Tema quente em áreas como a classificação de imagens onde existem grandes conjuntos de dados com poucos exemplos rotulados (conhecidos).

Algoritmos também podem ser agrupados por similaridade de forma ou função

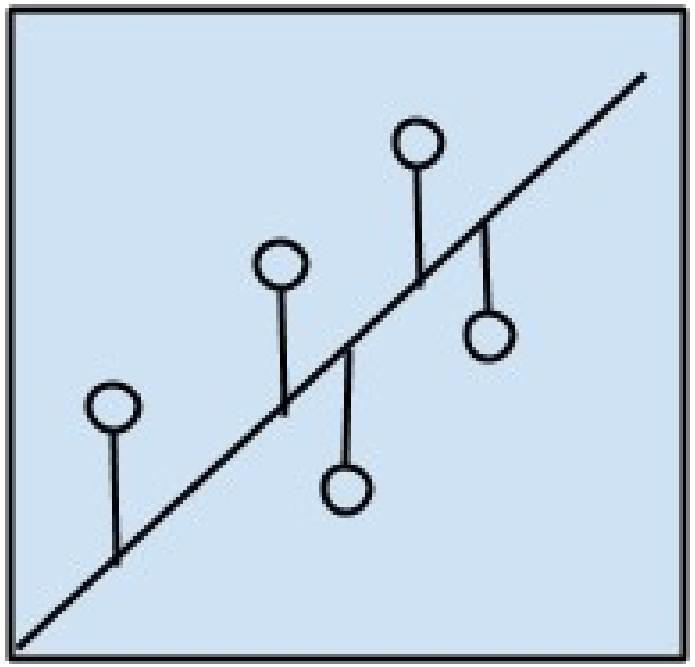
Por exemplo, os métodos baseados em árvore de classificação, métodos inspirados em rede neural, métodos de regressão.

Regressão: Técnica que permite explorar e inferir a relação de uma variável dependente (variável de resposta) com variáveis independentes específicas (variáveis explicatórias) do modelo.

Métodos de regressão são muito comuns na estatística e foram confinados em aprendizado de máquina estatística.

Os algoritmos de regressão mais populares são:

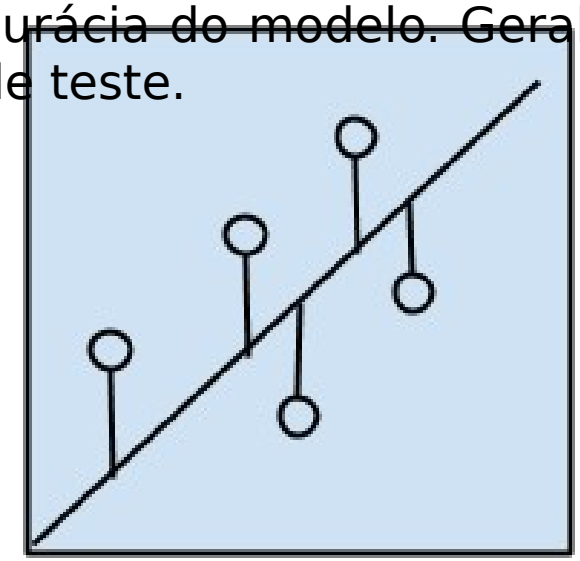
- Regressão linear
- regressão logística
- Regressão stepwise
- Mínimos Quadrados Ordinários Regressão (OLS)
- Multivariada Adaptive Regressão Splines (MARS)
- Scatterplot localmente estimado Smoothing (LOESS)



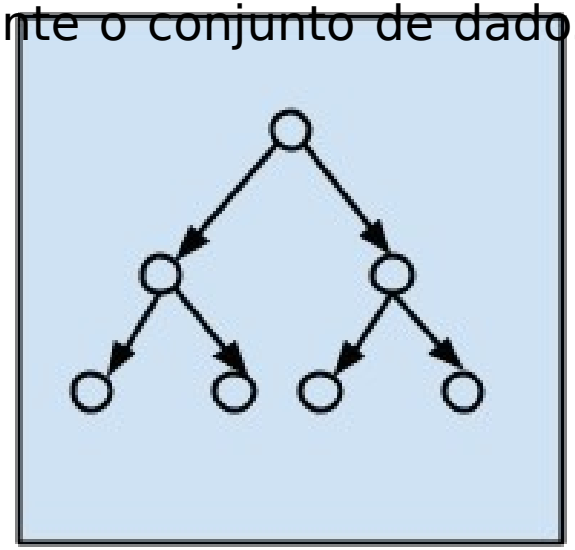
Regression Algorithms

Random Forest: técnica geral de florestas de decisão aleatória, que é um método de aprendizagem conjunto para a classificação, regressão e outras tarefas, que operam através da construção de um grande número de árvores de decisão a partir dos dados de treino e gera um resultado baseado nas classes (classificação) ou em uma média de predição (regressão) das árvores individuais. A ideia do método consiste em um "ensacamento", a fim de construir uma coleção de árvores de decisão com variância controlada. É uma classificação de classificações, dai o nome Floresta (conjunto de árvores)!!

Classificação: Dada uma coleção de registros (conjunto de treinamento) – cada registro contém um conjunto de atributos, e um dos atributos é a classe (ex: presença ou ausência). Encontra um modelo para o atributo classe como uma função dos valores de outros atributos. Objetivo: a classe deve ser atribuída tão acuradamente quanto possível para novos registros. – Um conjunto de teste (test set) é usado para determinar a acurácia do modelo. Geralmente o conjunto de dados é dividido em conjunto de treinamento e conjunto de teste.



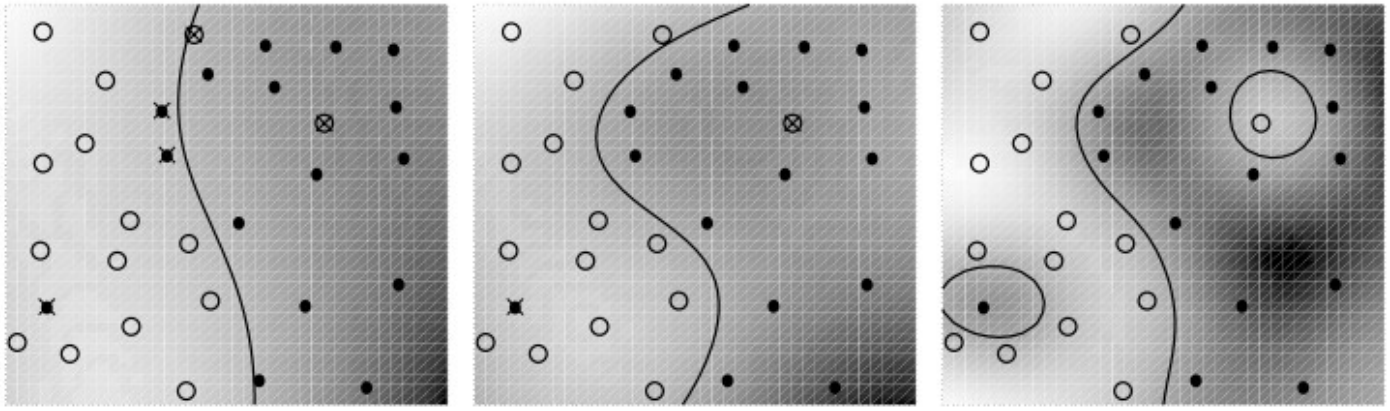
Regression Algorithms

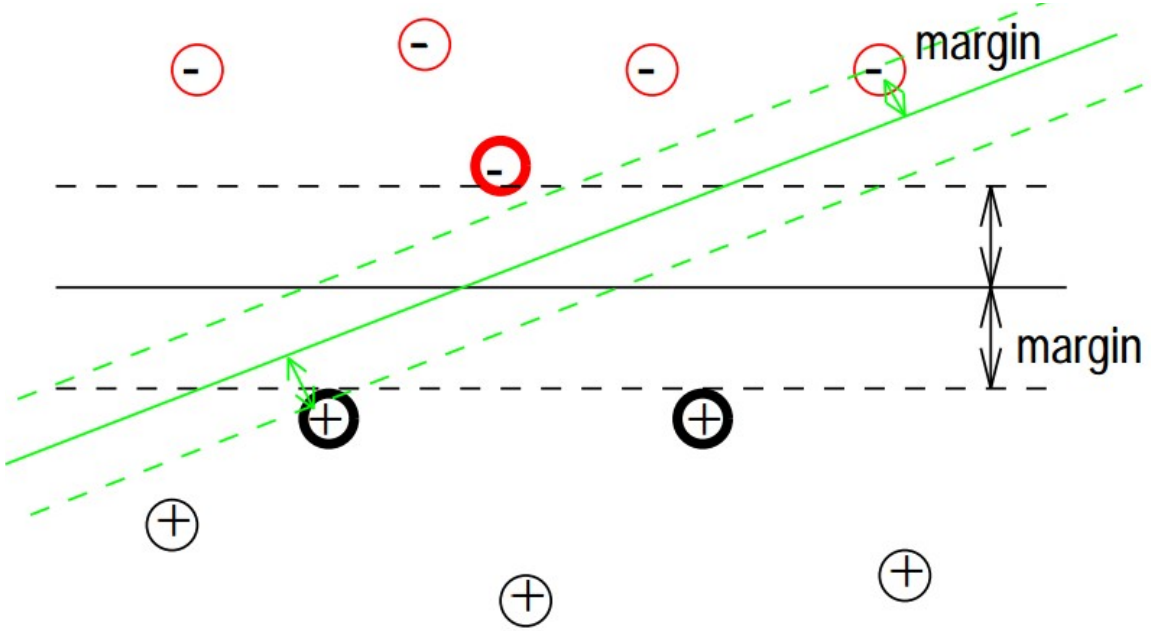


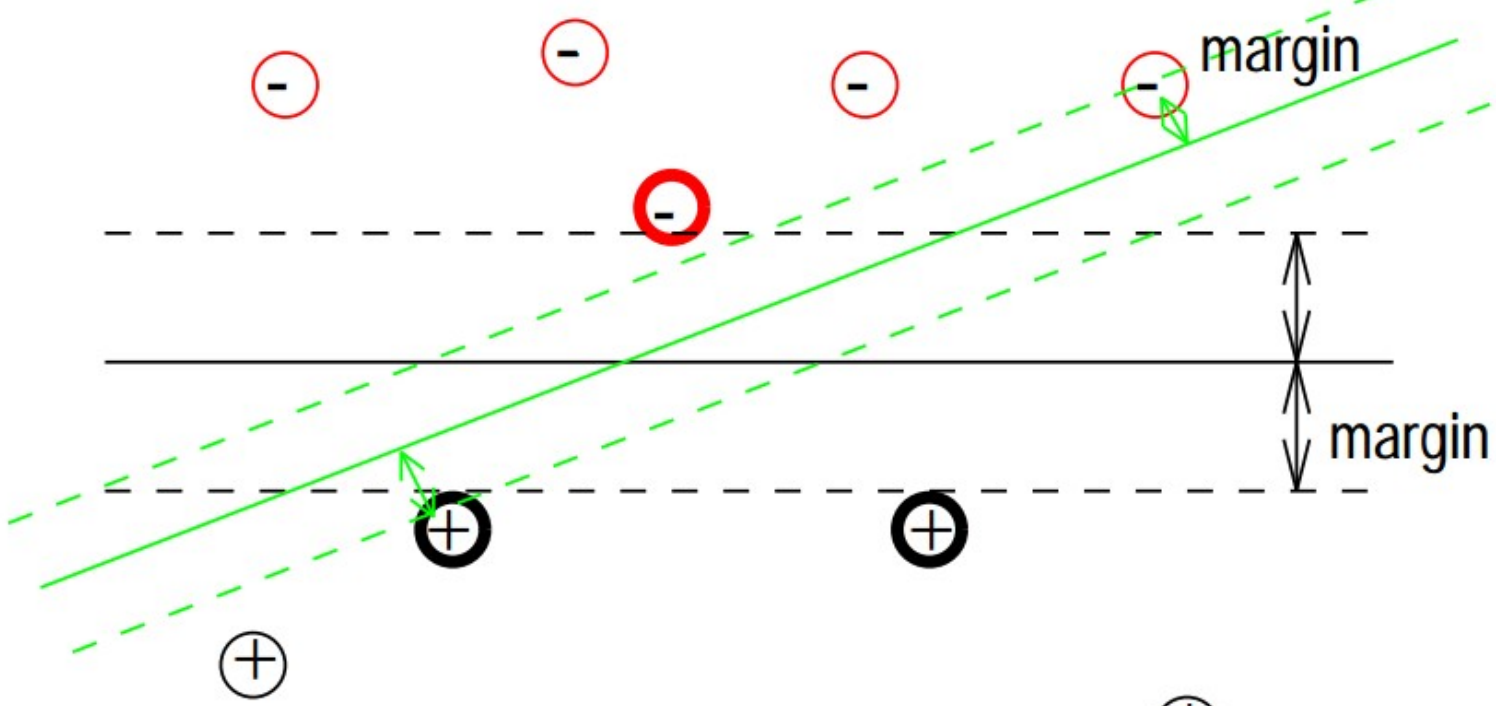
Decision Tree Algorithms

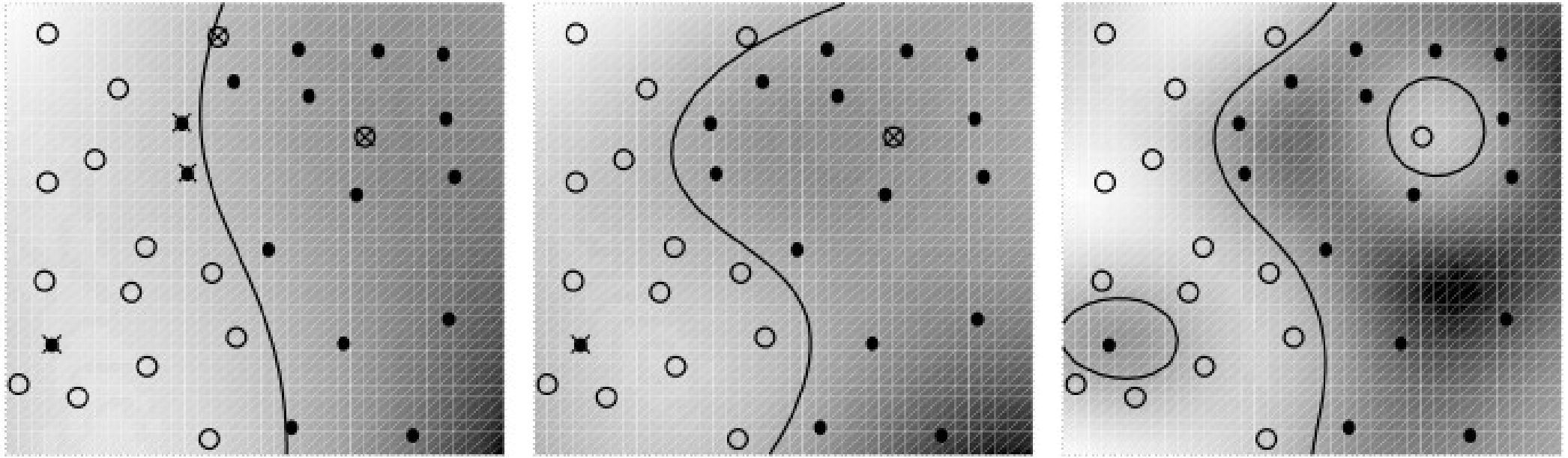
(SVM *support vector machine*) é um conceito na ciência da computação para um conjunto de métodos de aprendizado supervisionado que analisam os dados e reconhecem padrões e usa classificação e análise de regressão. O SVM padrão toma como entrada um conjunto de dados e prediz, para cada entrada dada, qual de duas possíveis classes a entrada faz parte, o que faz do SVM um classificador linear binário não probabilístico.

Ex: suponha que a gente tenha 100 registros de presença e 100 registros de ausência de uma determinada espécie. Esses dados são fornecidos para o algoritmo que os classifica (+=presença e -= ausência). Agora damos um conjunto novo de dados (sem dizer se é presença ou ausência) e esperamos que o algoritmo nos diga se é presença ou ausência. Uma vez treinado o algoritmo, esta predição pode ser feita para uma superfície nova (nossa área de estudo) e é baseada em um hiperplano calculado pelo algoritmo (regressão) que vai passar pelos vetores de suporte (registros que melhor separam as duas classes), o valor de cada pixel será dado pela distância do valor do pixel em relação ao hiperplano (métrica euclidiana por exemplo).






margin
margin



MaxEnt

(Maximum Entropy)

- Segue o princípio de máxima entropia. Busca a distribuição mais uniforme possível que se ajuste às restrições (condições ambientais nos pontos de presença = 1)
  - ‘Importância’ de cada variável ambiental.
  - Altamente usado
  - Boa performance
  - “Caixa preta” até pouco tempo atrás

Diversity and Distributions, (Diversity Distrib.) (2011) **17**, 43–57

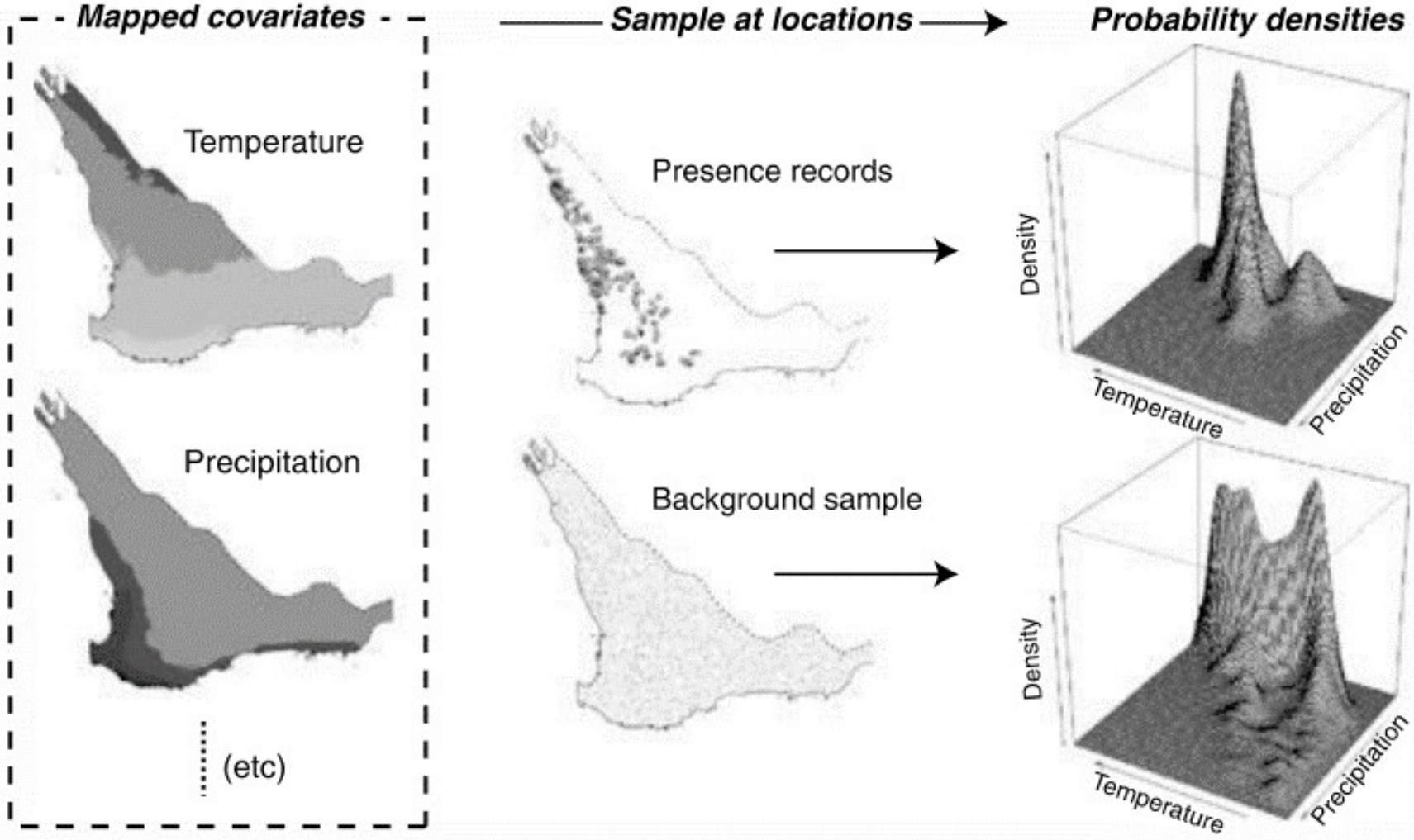
BIODIVERSITY RESEARCH

A statistical explanation of MaxEnt for ecologists

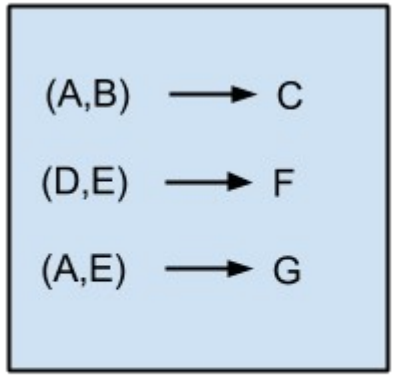
Jane Elith[1]*, Steven J. Phillips[2], Trevor Hastie[3], Miroslav Dudík[4], Yung En Chee[1] and Colin J. Yates[5]

O Maxent é uma técnica de aprendizado de máquina que combina estatística, máxima entropia e métodos Bayesianos cuja finalidade é estimar as distribuições de probabilidade de máxima entropia sujeita a restrições dadas pela informação ambiental associada aos registros de presença e ao background (área de estudo).

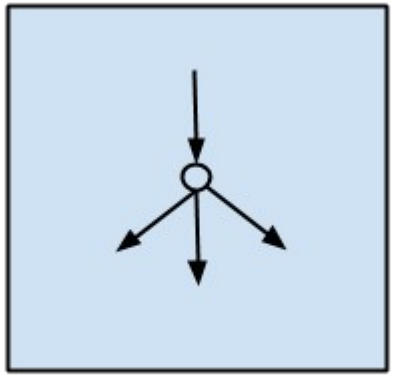
Observações inesperadas tem informações superiores às observações esperadas



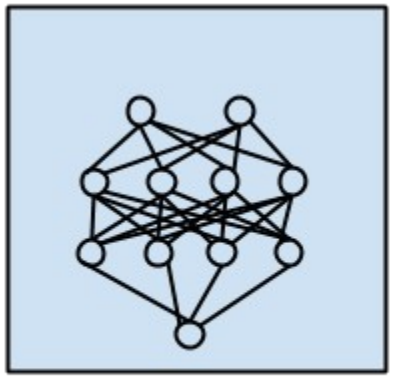
E mais alguns....



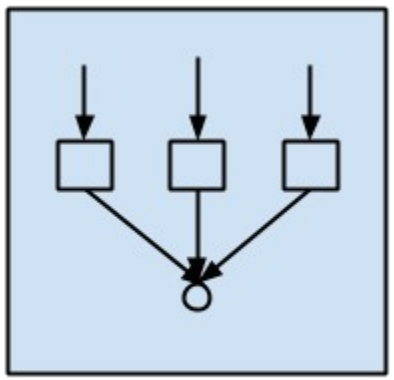
Association Rule Learning Algorithms



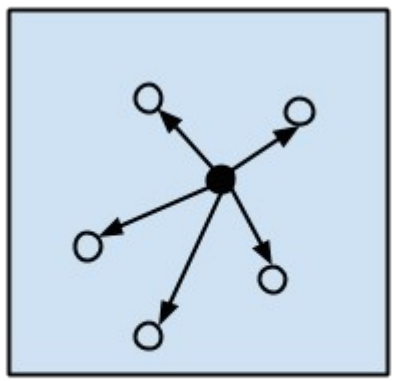
Artificial Neural Network Algorithms



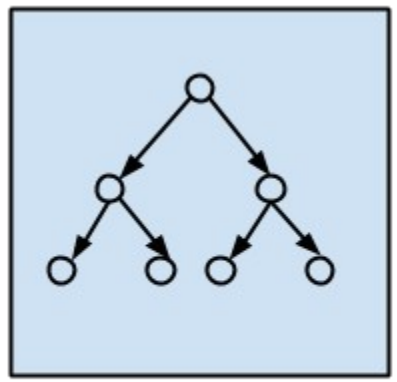
Deep Learning Algorithms



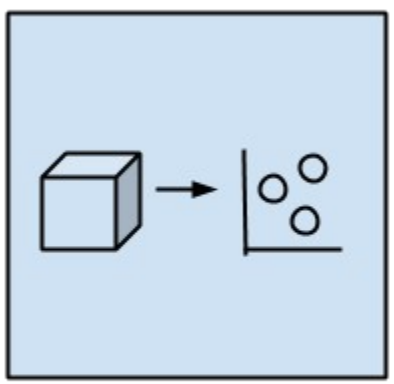
Ensemble Algorithms



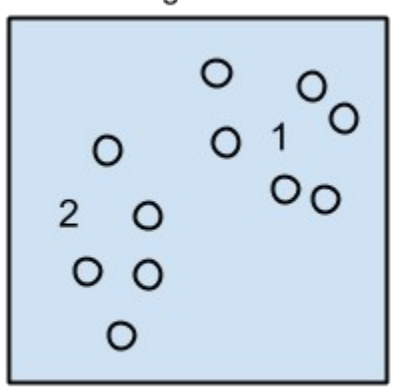
Instance-based Algorithms



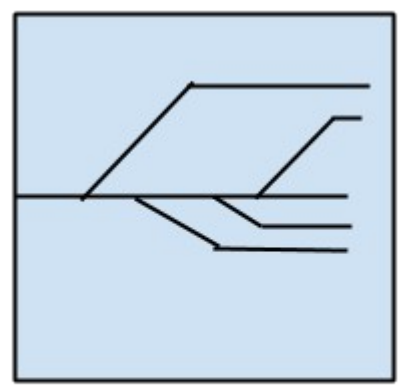
Decision Tree Algorithms



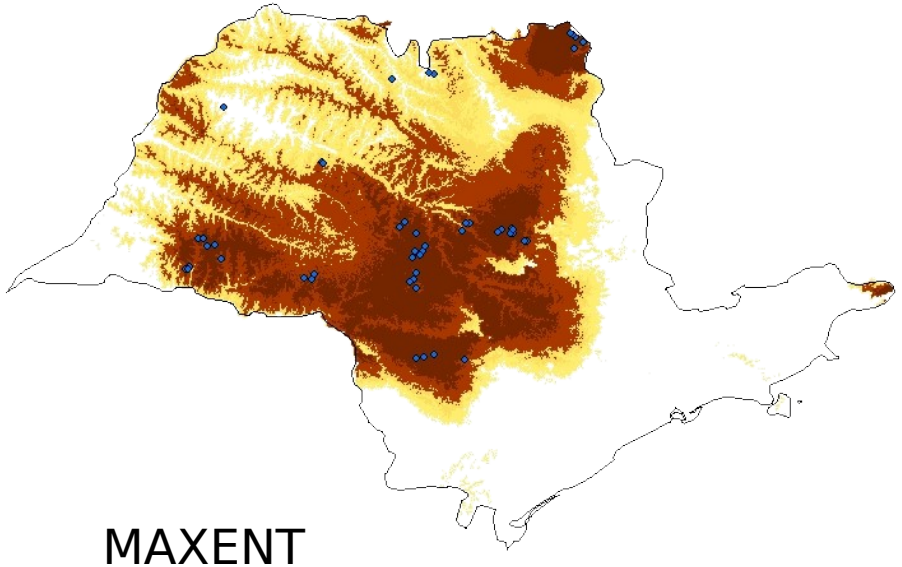
Dimensional Reduction Algorithms



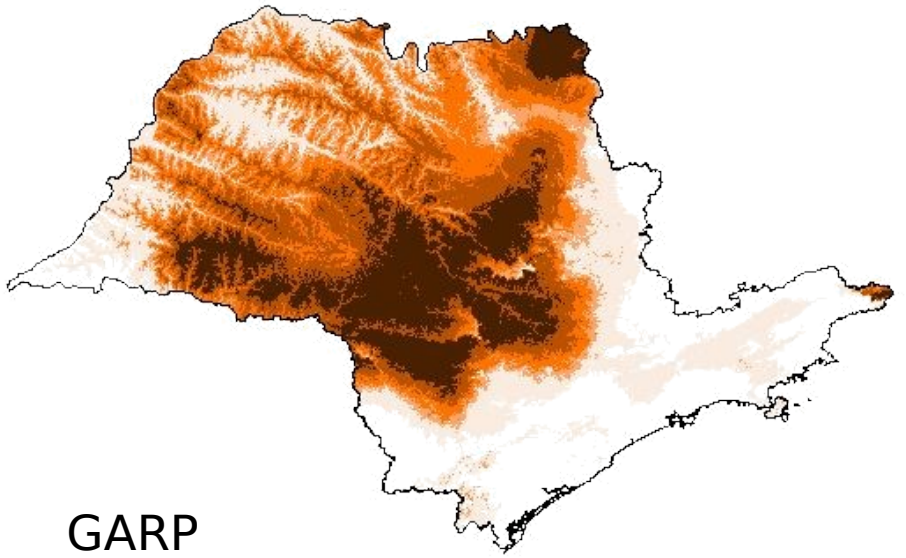
Clustering Algorithms



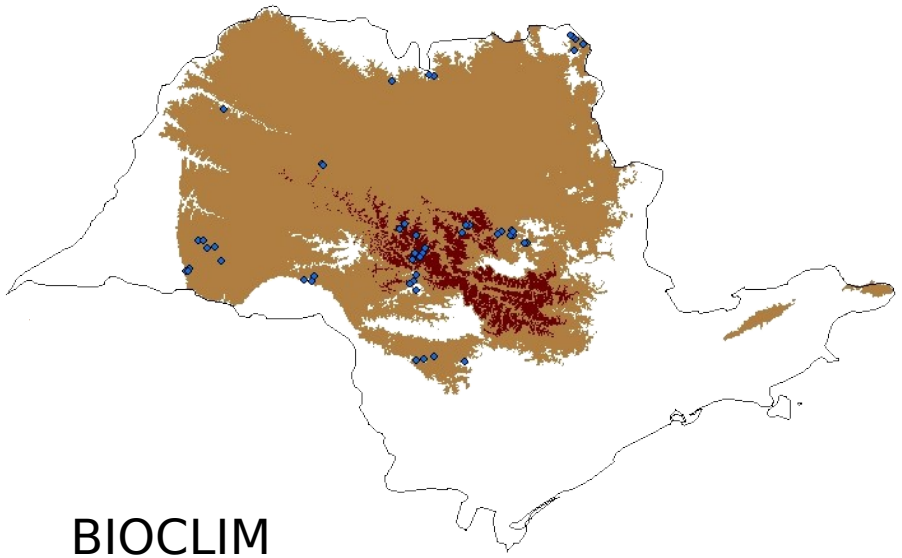
Regularization Algorithms



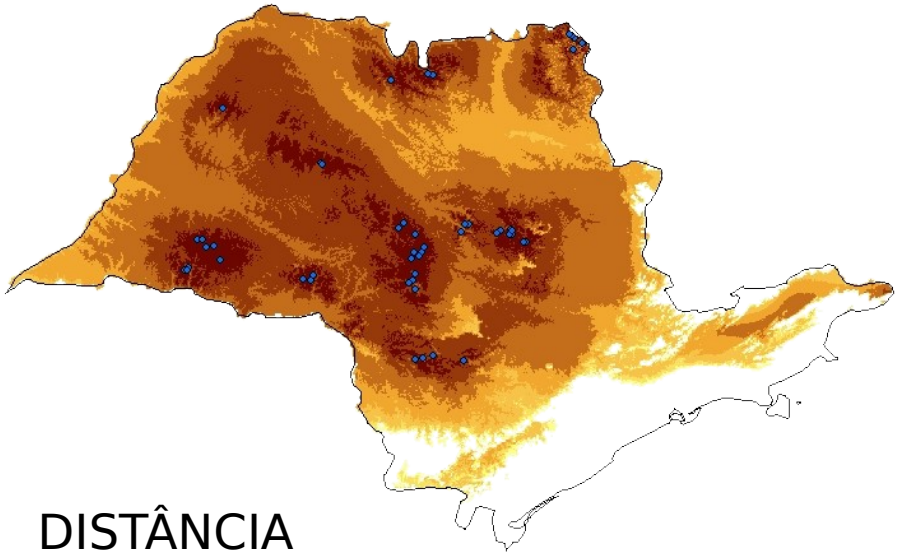
MAXENT



GARP



BIOCLIM



DISTÂNCIA
AMBIENTAL