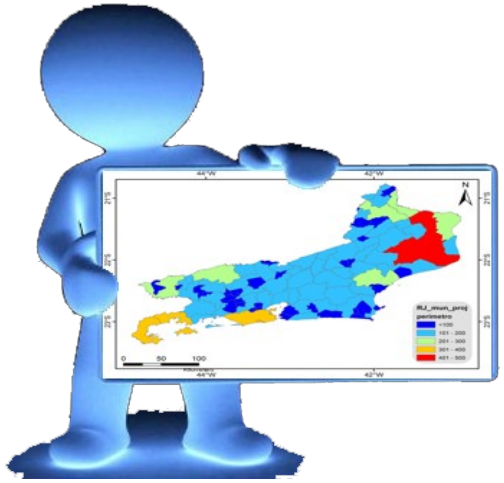
INSTITUTO INTERNACIONAL PARA SUSTENTABILIDADE

INTERNATIONAL INSTITUTE FOR SUSTAINABILITY

IIS

Cartografia & Saúde: Análise geoespacial como ferramenta aplicada na parasitologia

Modelagem de Nicho Ecológico



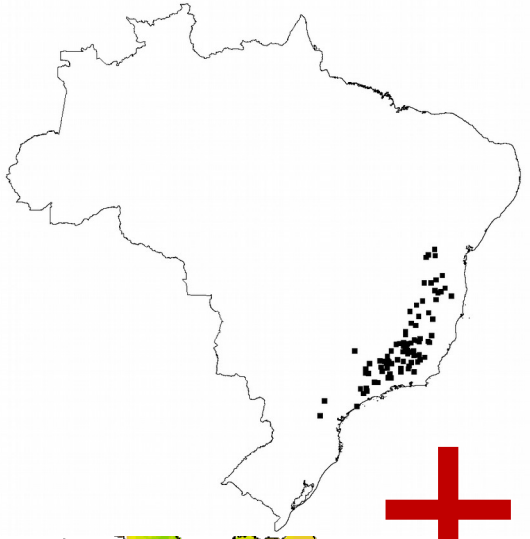
Diogo S. B. Rocha

Validação de Modelos

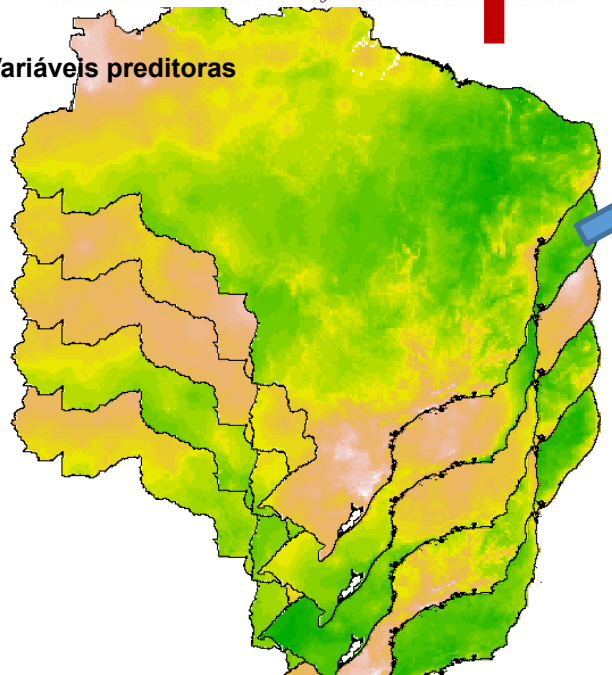
Calibrando e validando um Modelo

Dados de entrada

Registros de ocorrência



Variáveis preditoras



Precipitação

Temperatura

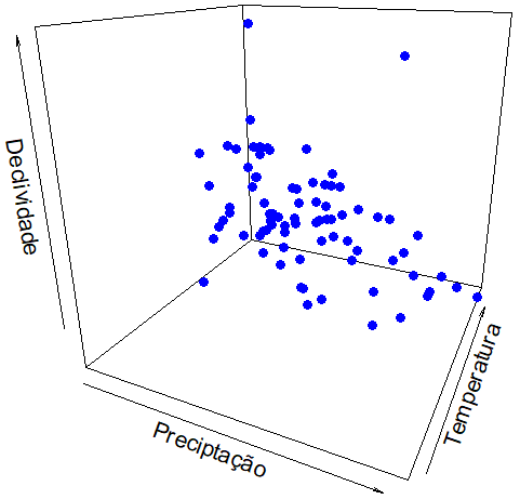
Topografia

Variável X

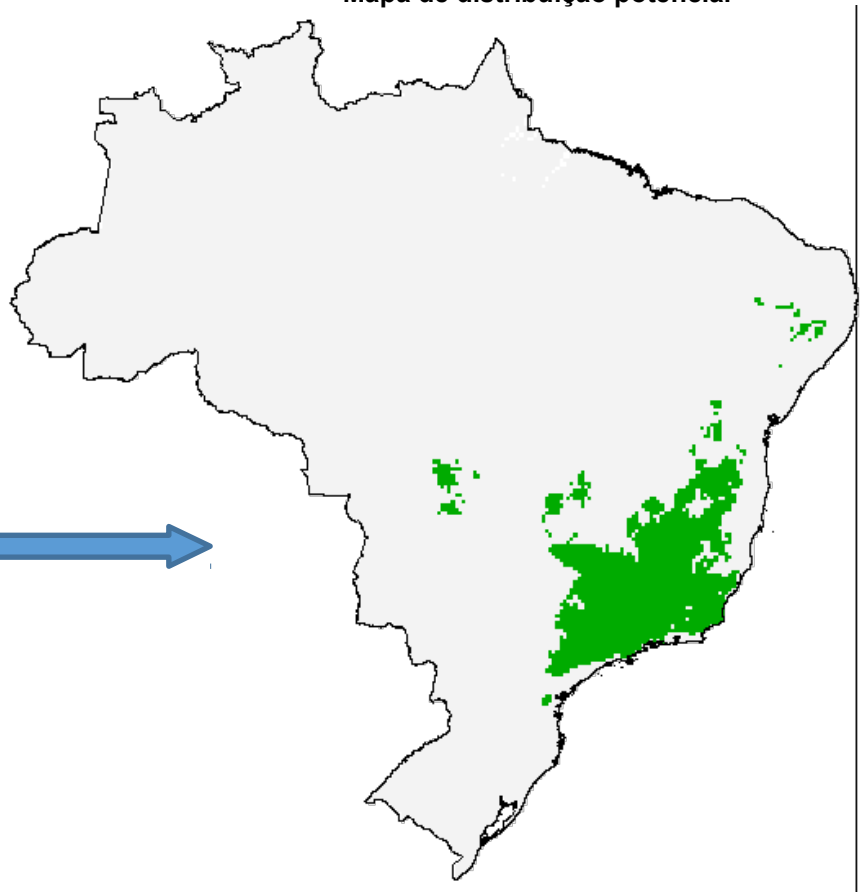
Variável Y

Algoritmos de modelagem

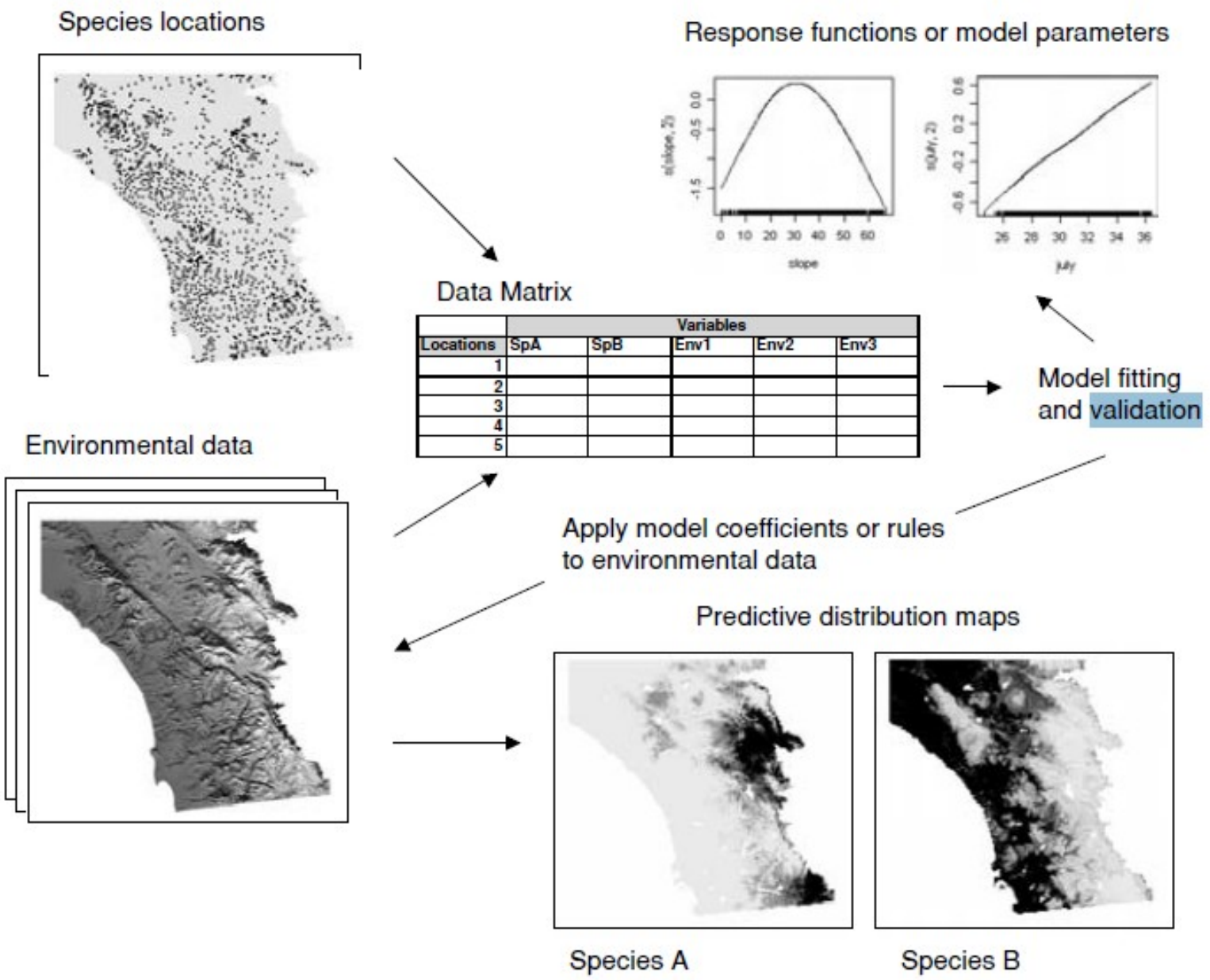
(Bioclim, GLM, GAM, ANN, GARP, MAxEnt, etc.)



Mapa de distribuição potencial

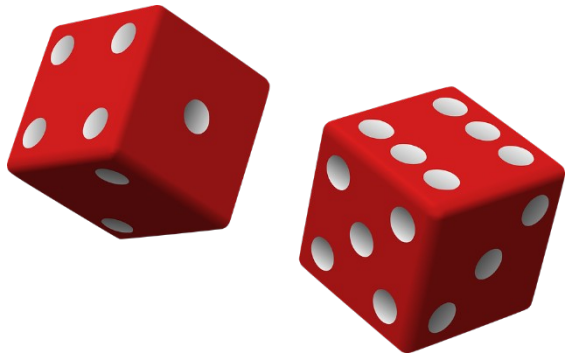


Calibrando e validando um Modelo



Como sabemos se um modelo é bom?

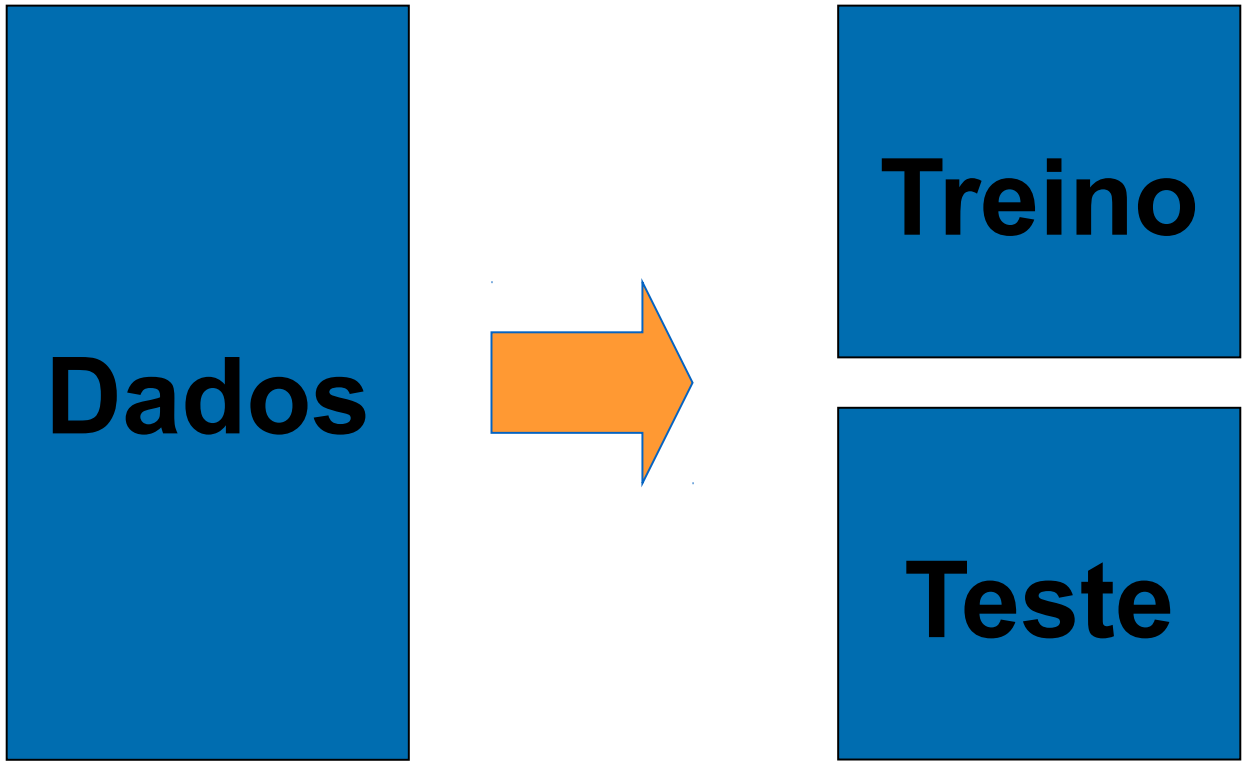
- Retorno ao campo em busca de novos registros



- Consulta ao especialista na biogeografia da espécie modelada

Ou ....

... fazemos uma partição dos dados em conjunto de treino (ajuste) e teste do modelo.



Mas antes: algumas definições!

Dados de treino: registros de ocorrência de espécies que serão utilizados para rodar o modelo para a espécie de interesse.

Dados de teste: registros de ocorrência que não foram utilizados no treino do modelo mas que serão utilizados para testar o modelo gerado (pelos dados de treino), referente a mesma espécie.

Mas antes: algumas definições!

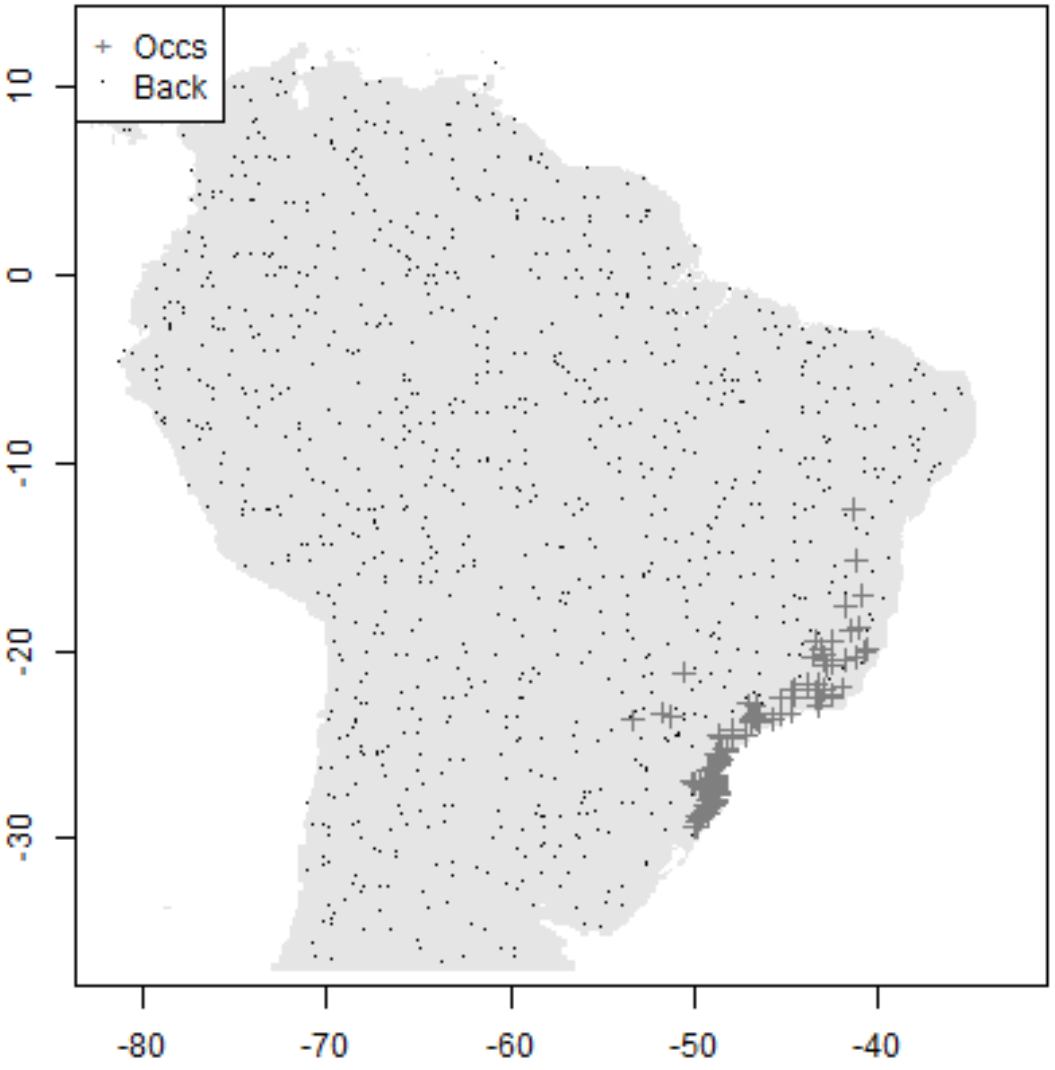
Dados de presença – pontos de ocorrência da espécie

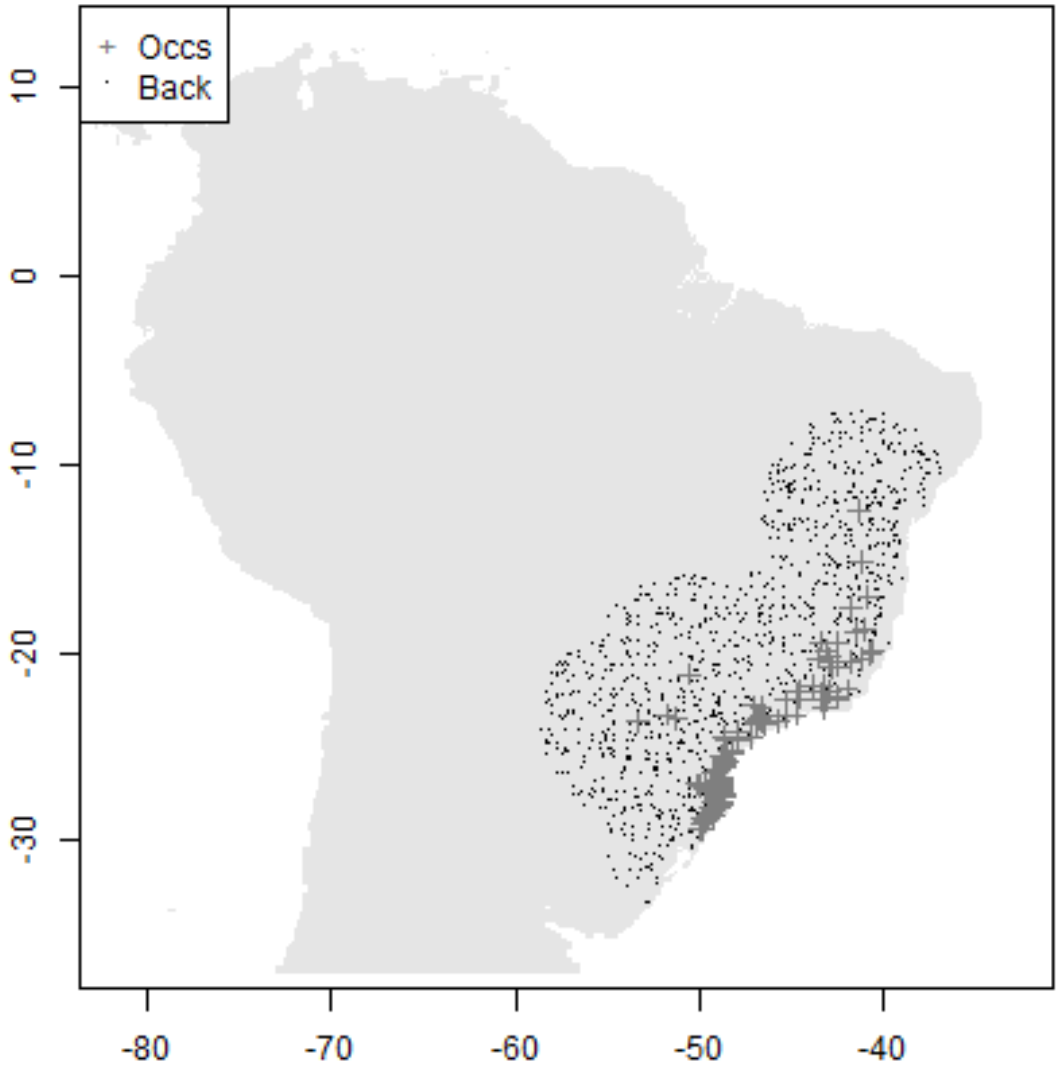
Dados de ausência – registros de ausência da espécie, caso não estejam disponíveis,

é necessário gerar um conjunto de pseudoausências.

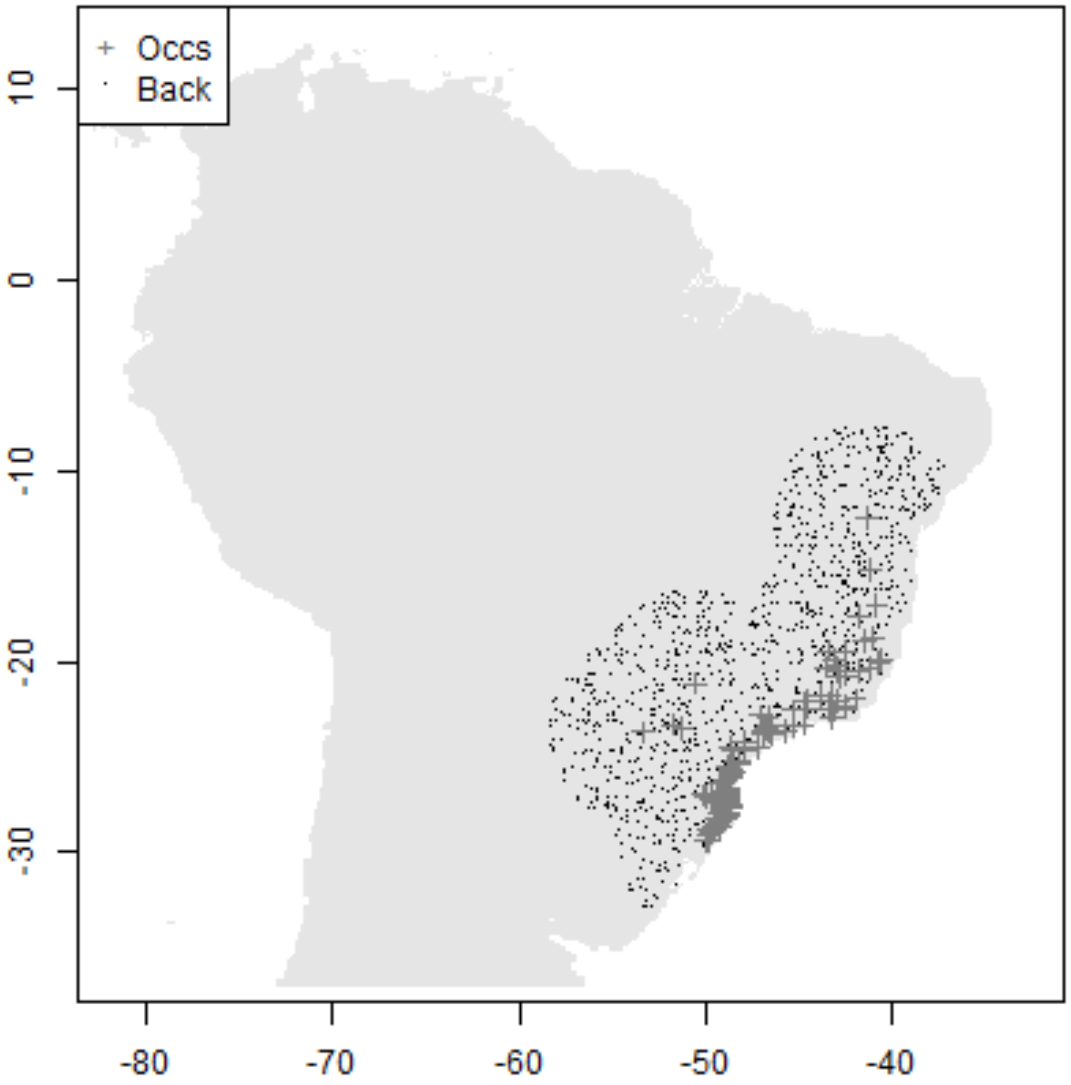
Onde?

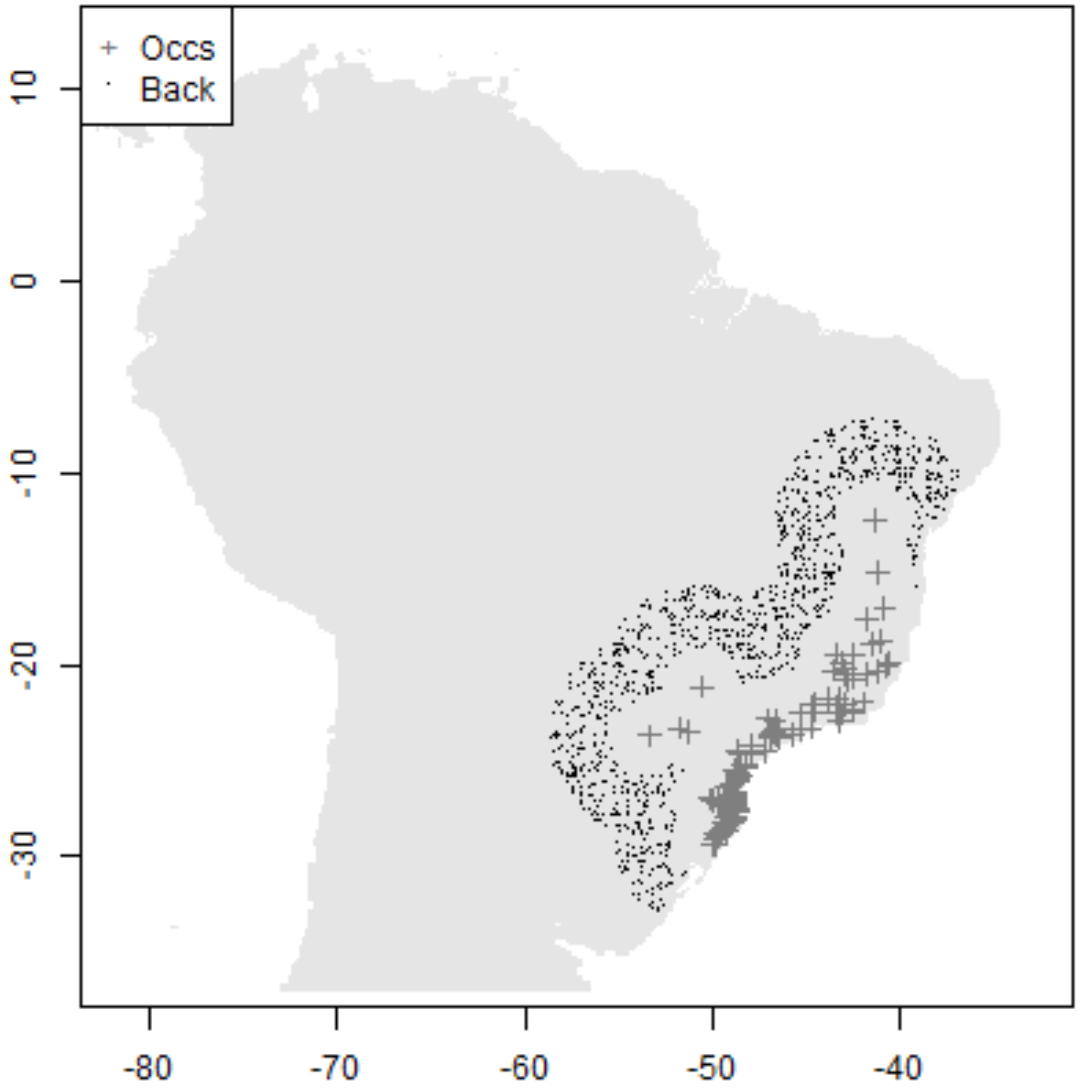
Mas antes: algumas definições!





Mas antes: algumas definições!





Então, para validar um modelo gerado é preciso:

1. Gerar o(s) conjunto(s) de treino e teste

2. Gerar modelo(s) com o(s) conjunto(s) de dados de treino, e

3. Sobrepor o conjunto de teste ao modelo gerado pelo conjunto de treino e quantificar os erros através de uma matriz de confusão (vamos ver isso já já)

E como eu gero esses conjuntos de dados?

Redes

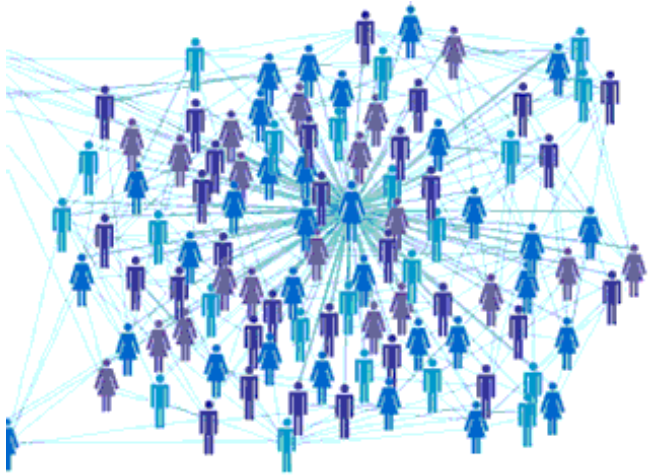
GBIF
Global Biodiversity Information Facility

speciesLink

Literatura acadêmica especializada

Vai ao campo!

Pesquisadores

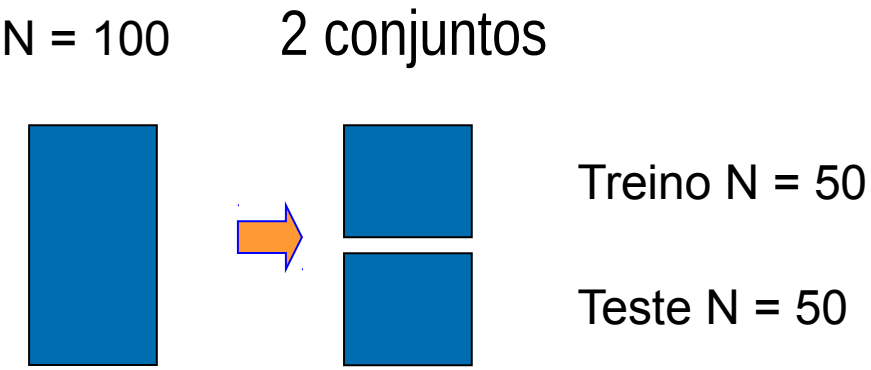


E ainda: como eu gero esses conjuntos de dados?

Para gerar um **conjunto de teste**: Existem pelo menos duas formas de se fazer isso:

a. Coletar novos dados (voltar ao campo)

b. Dividir o dados originais em conjuntos (treino e teste) antes de realizar a modelagem



editado de Peterson & Martinez (2005)

Mas não precisamos ficar com apenas dois conjuntos (um de treino e um de teste)

Os dados podem ser divididos em vários conjuntos de treino e teste (podemos realizar várias partições/divisões nos dados). Isto é feito para:

- calcular a variabilidade dos resultados (e.g. média ± desvio padrão)
- avaliar a qualidade dos pontos
- comparar melhor os resultados de diferentes algoritmos

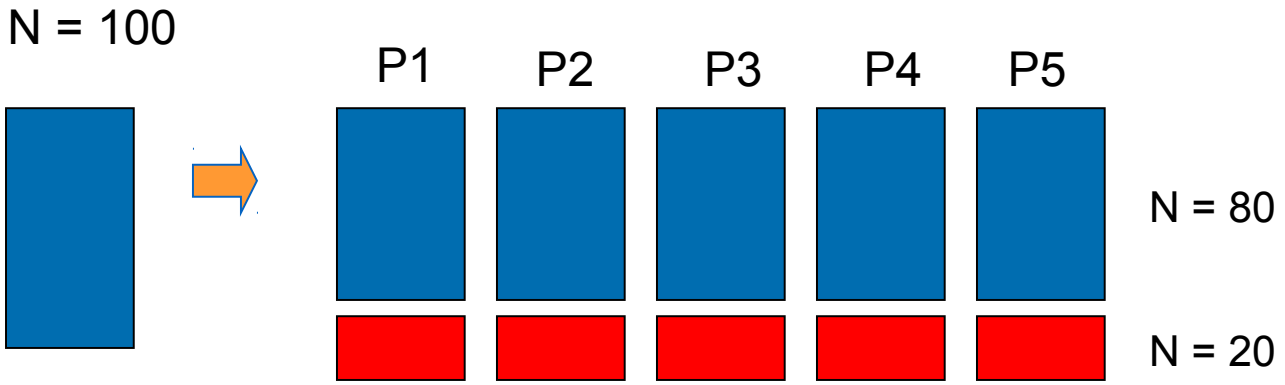
Tipos de partições de dados:

1. Com reposição (ex: bootstrap)
2. Sem reposição (ex: crossvalidate ou validação cruzada)

A escolha de um método irá depender do número de pontos que você possui.

Bootstrap

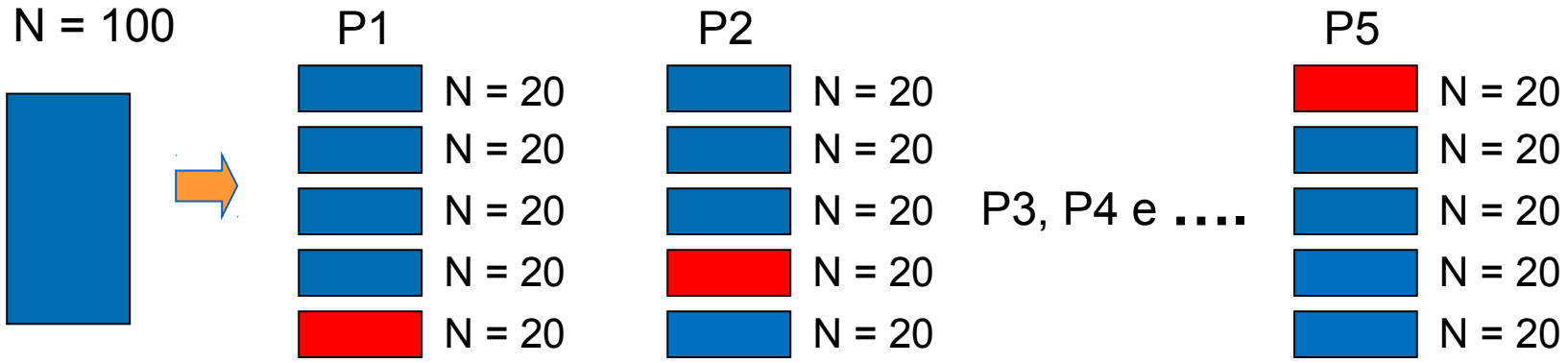
Realizar 5 partições/divisões = 80% treino e 20% teste



Terminamos com 5 conjuntos de dados sendo Treino N = 80 e Teste N = 20

Cross-validation

Realizar 5 partições/divisões iguais nos dados



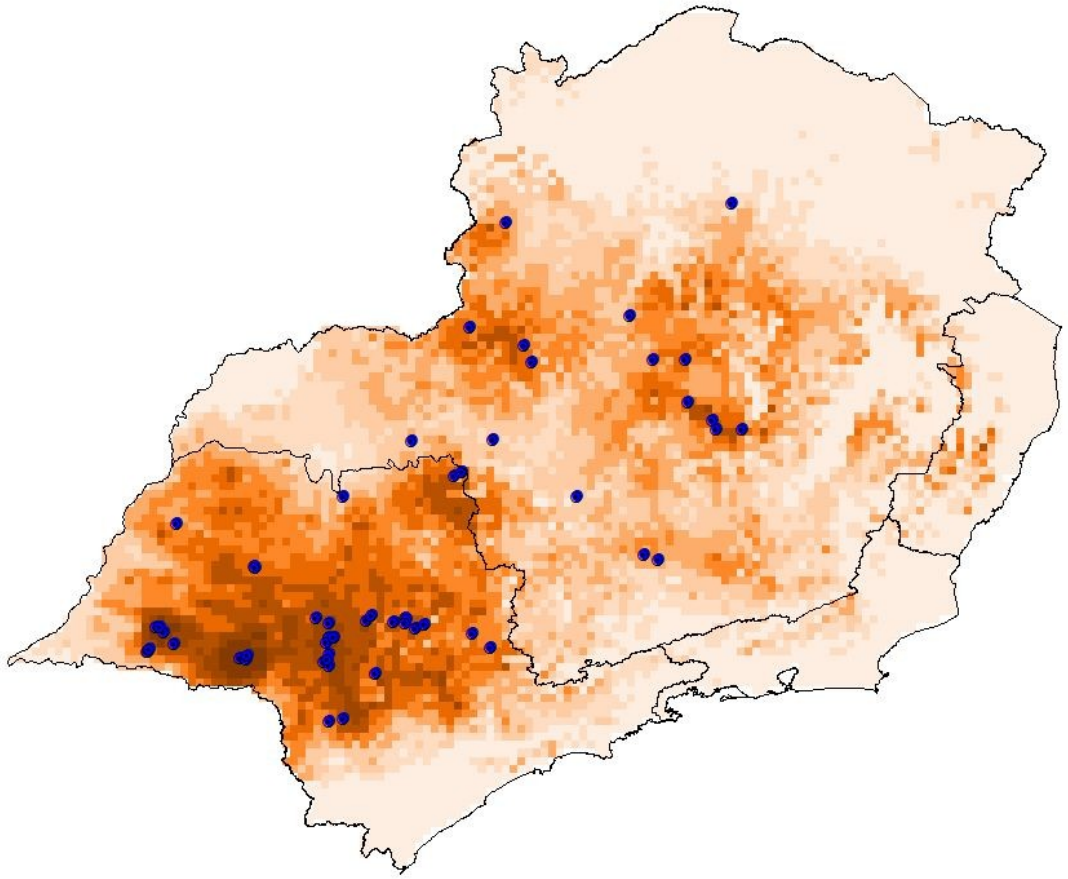
Terminamos com 5 conjuntos de dados sendo Treino N = 80 e Teste N = 20

Para avaliar a qualidade do modelo gerado é preciso:

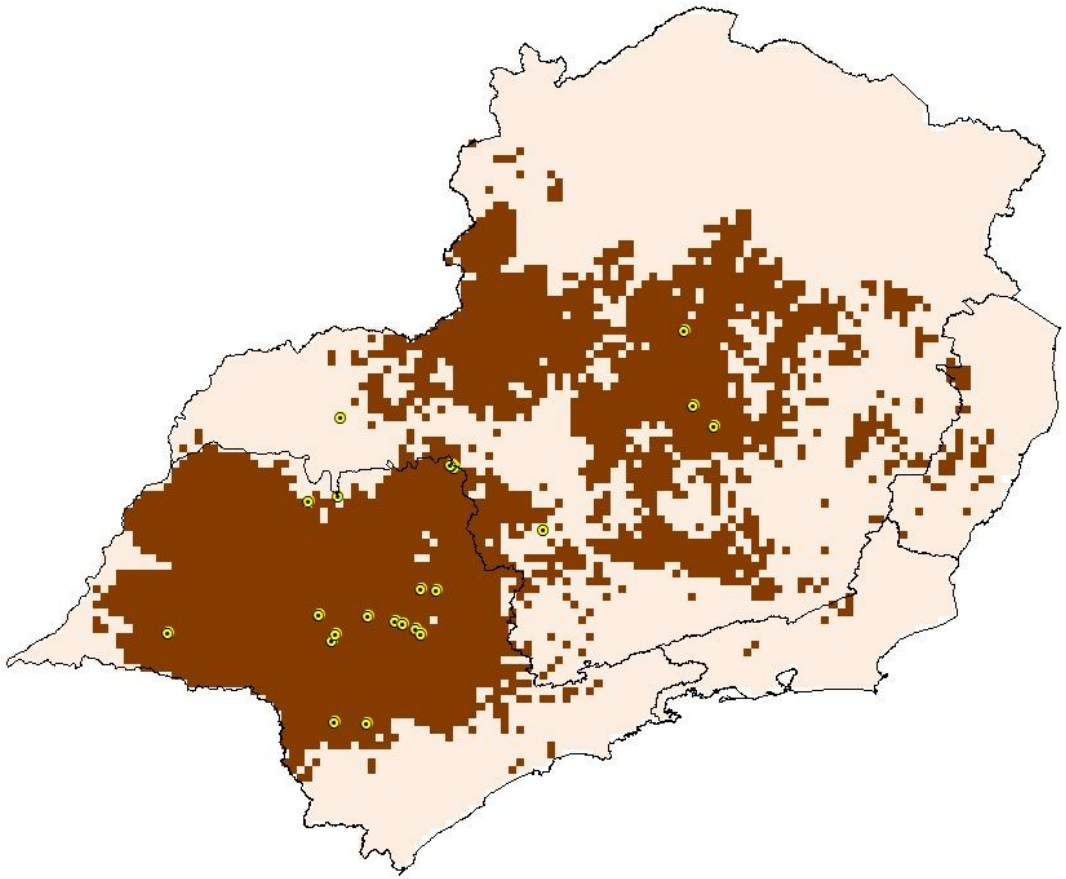
1. Ter um conjunto de teste. Não é aceitável testar um modelo a partir dos pontos que o geraram! Isso não faria sentido!

2. Quantificar os componentes de erro através de uma **matriz de confusão** sobrepondo os pontos de teste ao modelo gerado pelo conjunto de treino

Gerar um modelo com o conjunto de dados de treino



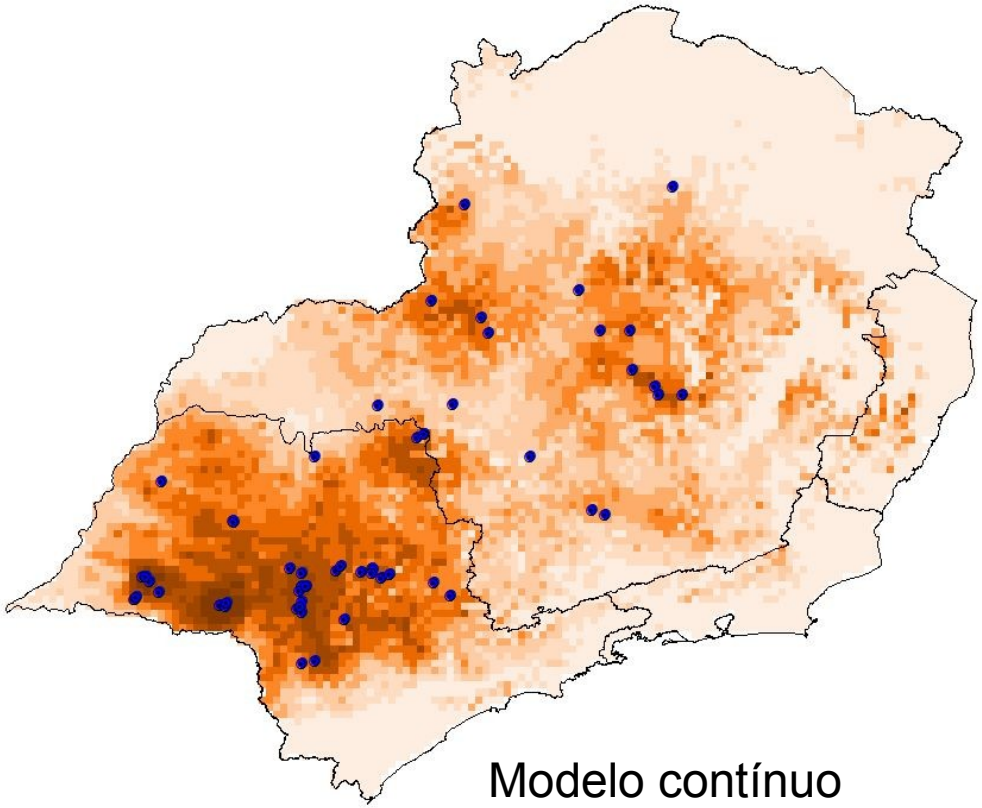
Testar o modelo com o conjunto de teste



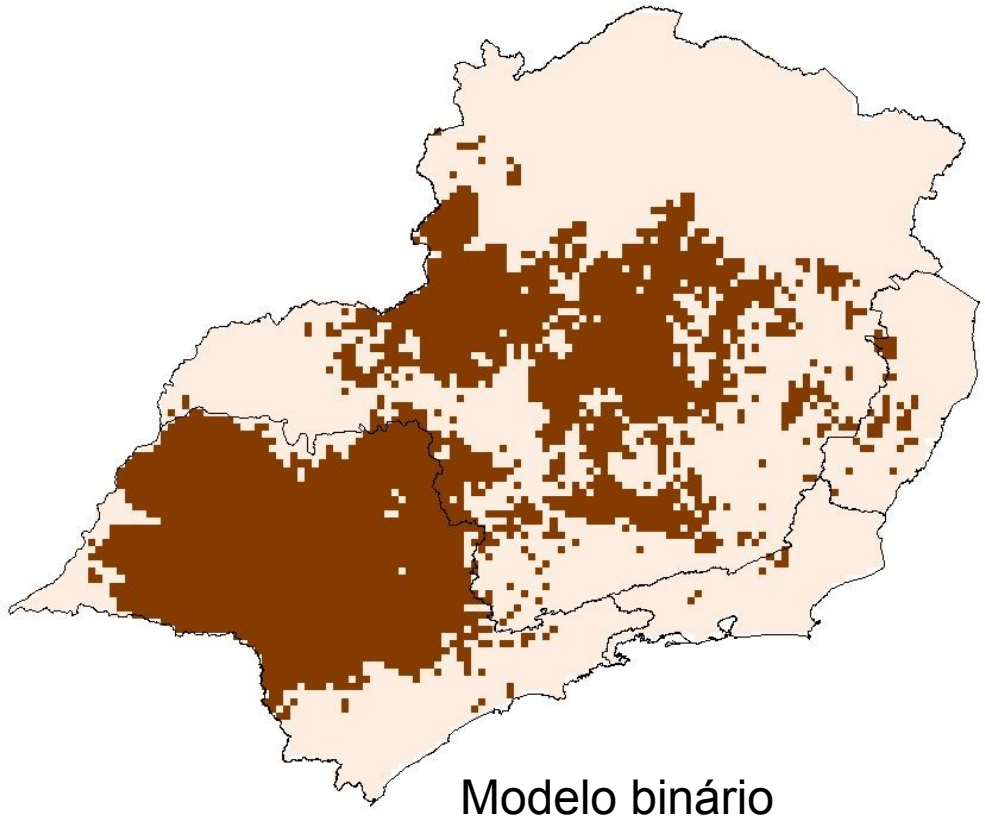
Antes é preciso aplicar um limiar de corte (threshold) ao modelo

Para isso é preciso estabelecer um limite de corte (*threshold*). Um valor de adequabilidade ambiental (*suitability*) a partir do qual será considerada presença provável para a espécie.

No exemplo abaixo, vamos adotar o limiar 0.3 como *threshold*. Então todos os pixels da área de estudo cujo valor de A.A. for superior a 0.3 será considerado como área de presença predita e receberá o valor 1. Consequentemente os pixels com valor de A.A. inferiores a 0.3 receberão o valor 0.



Modelo contínuo



Modelo binário

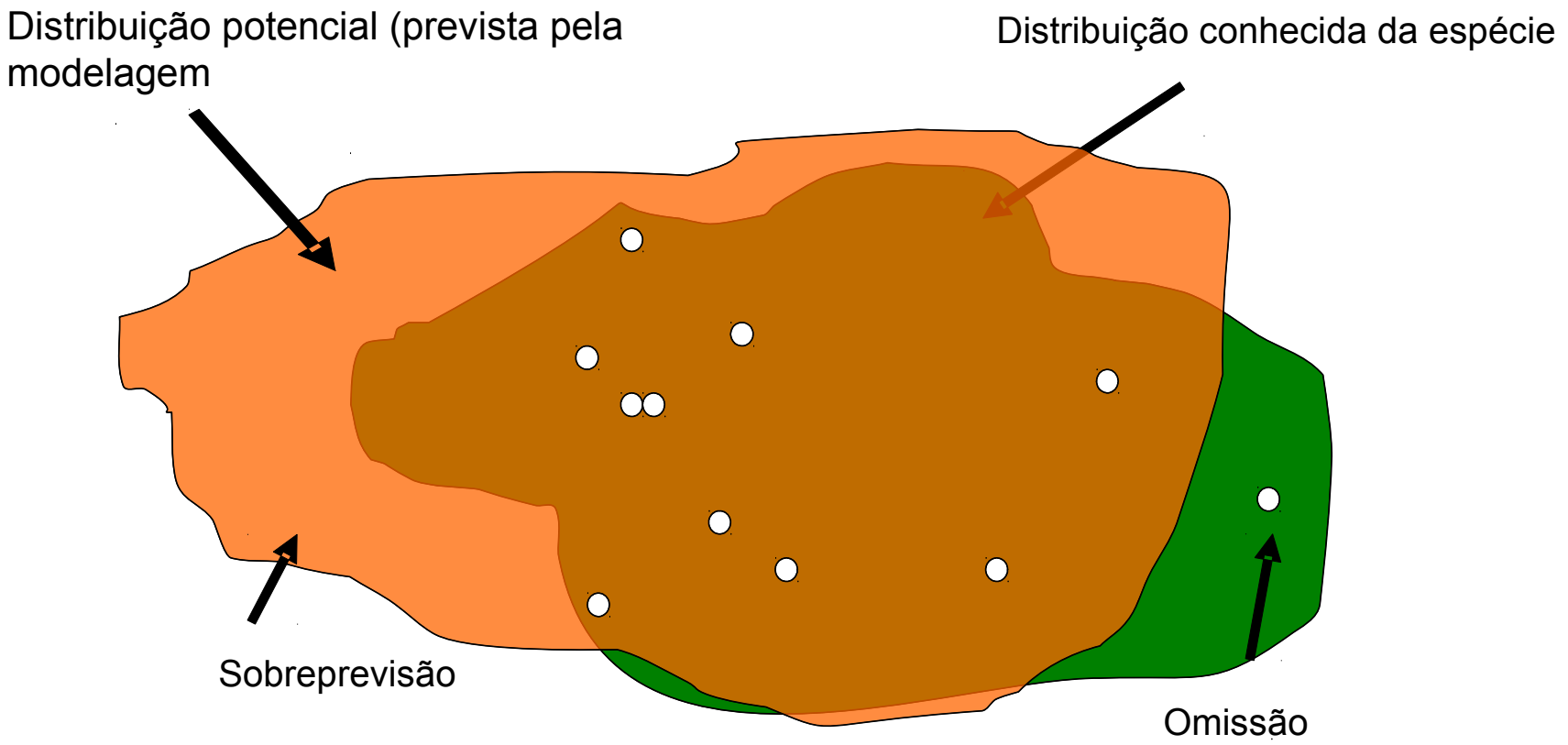
Como fazemos isso? Boas revisões dos limites de corte (thresholds) utilizados (Liu et al. 2005 e Pearson et al. 2007 e Liu et al. 2013):

1. Fixed cumulative value 1 (valor fixo em 1% de A.A.)
2. Fixed cumulative value 5 (valor fixo em 5% de A.A.)
3. Fixed cumulative value 10 (valor fixo em 10% de A.A.)
4. Minimum training presence (omissão = 0% dos pontos de treino)
5. 10 percentile training presence (omissão = 10% dos pontos de treino)
6. Equal training sensitivity and specificity (omissão e comissão iguais dos pontos de treino)
7. Maximum training sensitivity plus specificity (menor omissão de treino na menor área preditiva)
8. Equal test sensitivity and specificity (omissão e comissão iguais dos pontos de teste)
9. Maximum test sensitivity plus specificity (menor omissão de teste na menor área preditiva)

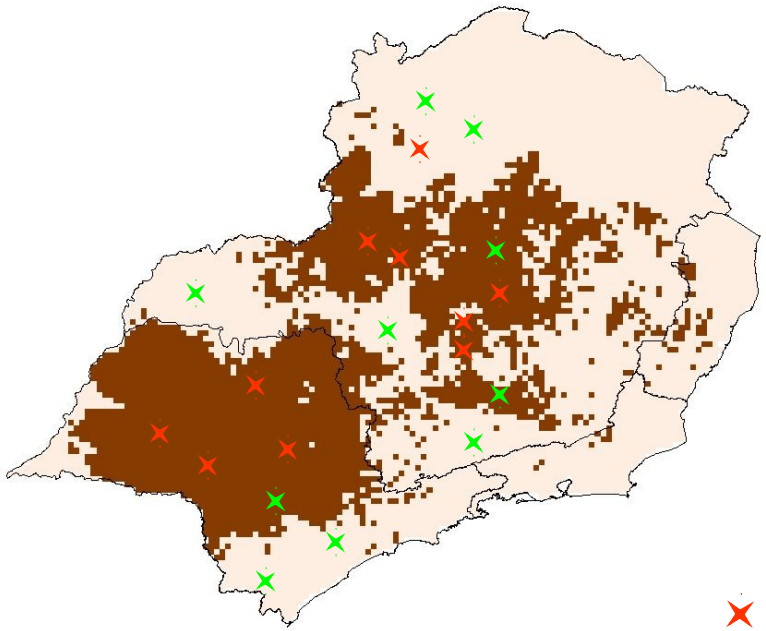
Mas e depois, como contabilizamos os erros e acertos do teste?

Há dois tipos de erro em modelagem

Erro de omissão e de sobreprevisão (*comission*)



Matriz de Confusão



	REGISTRO DE PRESENÇA	REGISTRO DE AUSÊNCIA
PRESENÇA PREDITA	9	3
AUSÊNCIA PREDITA	1	7

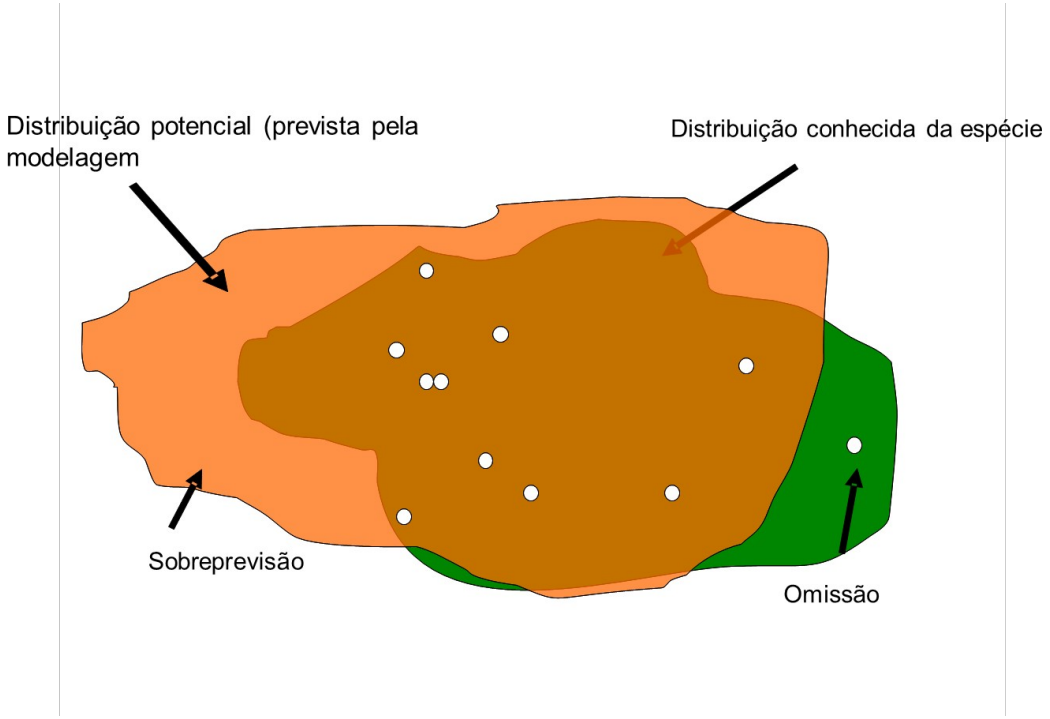
registros de presença

registros de ausência

Matriz de Confusão

	REGISTRO DE PRESENÇA	REGISTRO DE AUSÊNCIA
PRESENÇA PREDITA	A	B
AUSÊNCIA PREDITA	C	D

A e D são acertos
B e C são erros



$$Sensibilidade = \frac{A}{(A+C)}$$

$$Especificidade = \frac{D}{(B+D)}$$

$$Sobreprevisão = \frac{B}{(B+D)}$$

$$omissão = \frac{C}{(A+C)}$$

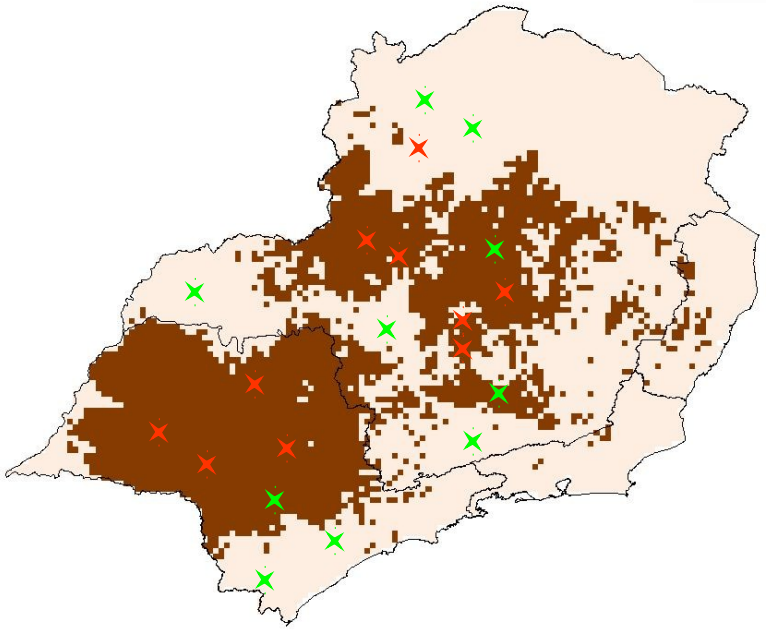
$$Acurácia^{*} = \frac{A+D}{A+B+C+D}$$

Taxas de acertos

Taxas de erros

Matriz de Confusão

	REGISTRO DE PRESENÇA	REGISTRO DE AUSÊNCIA
PRESENÇA PREDITA	9	3
AUSÊNCIA PREDITA	1	7



$Sensibilidade = \frac{A}{(A+C)}$	0.9
$Especificidade = \frac{D}{(B+D)}$	0.7
$Sobreprevisão = \frac{B}{(B+D)}$	0.3
$omissão = \frac{C}{(A+C)}$	0.1
$Acurácia^{*} = \frac{A+D}{A+B+C+D}$	0.8

x registros de presença

x registros de ausência

Qual o significado dos dois tipos de erros no processo de modelagem?

Erro de omissão: no geral, o erro de omissão é considerado um erro verdadeiro. Contudo, algumas vezes um registro de presença pode não ser correto. Isso pode acontecer em algumas circunstâncias, tais como:

1. A identificação da espécie (taxa) está errada.

2. Erro de georeferenciamento.

3. Um registro da espécie encontrado fora do seu habitat natural (indivíduos em trânsito ou introduzidos).

E qual o significado do erro de sobreprevisão?

Erro de comissão ou sobreprevisão: este pode ou não ser um erro, de qualquer forma, não é considerado um erro “grave”. A previsão de ocorrência em áreas onde as espécies não tem registro confirmado pode ser causada por diferentes fatores:

1. A área é habitável pela espécie mas o esforço amostral não foi suficiente para detectá-la.

2. A área é habitável para a espécie mas fatores históricos ou ecológicos (barreiras geográficas, capacidade de dispersão) ou bióticos (competição, predação) impediram a espécie de chegar ou de se estabelecer na região.

3. A área é inabitável mesmo, o que seria o verdadeiro erro de sobreprevisão.

Outras estatísticas para validar os modelos gerados

TSS: True Skill Statistic

- TSS = (sensibilidade + especificidade) – 1.
- TSS = ((A/A+C) + (D/B+D))-1

	REGISTRO DE PRESENÇA	REGISTRO DE AUSÊNCIA
PRESENÇA PREDITA	A	B
AUSÊNCIA PREDITA	C	D

Utilizando nosso exemplo anterior temos:

- TSS = (9/(9+1) + (7/(3+7)) – 1
- TSS = (0.9+0.7) – 1
- TSS = 0.6 (modelo bom!)

$$Sensibilidade = \frac{A}{(A+C)}$$

$$Especificidade = \frac{D}{(B+D)}$$

Obs: A TSS pode variar de -1 a 1. Quanto mais próximo de 1 melhor é o modelo. No geral, acima de 0.6 considera-se um bom ajuste do modelo aos dados. Entre 0.2 – 0.6 um ajuste regular e abaixo de 0.2, um ajuste ruim.

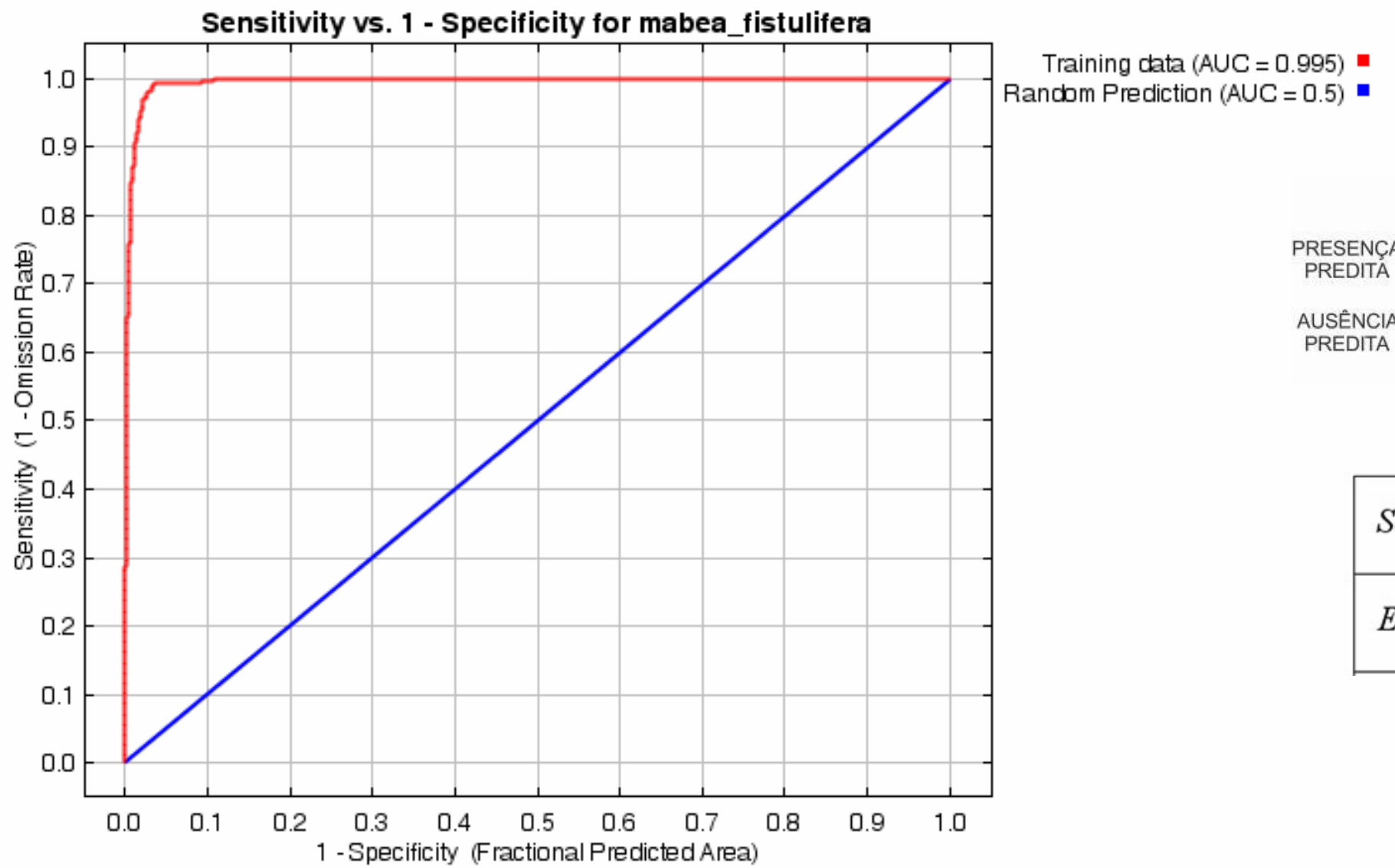
Outras estatísticas para validar os modelos gerados

Análise ROC (cálculo da área sob a curva – AUC): avalia a performance do modelo através do valor representado pela área sob a curva ROC (AUC).

É obtida plotando-se a sensibilidade no eixo y e o valor 1-especificidade no eixo x. Quanto mais próximo de 1 for a área sob a curva, mais distante o resultado do modelo é da previsão aleatória, ou seja, melhor o desempenho do modelo.

Obs: este valor pode ser usado para comparações entre diferentes algoritmos porque independe de um limiar de corte específico.

Análise ROC (cálculo da área sob a curva – AUC):



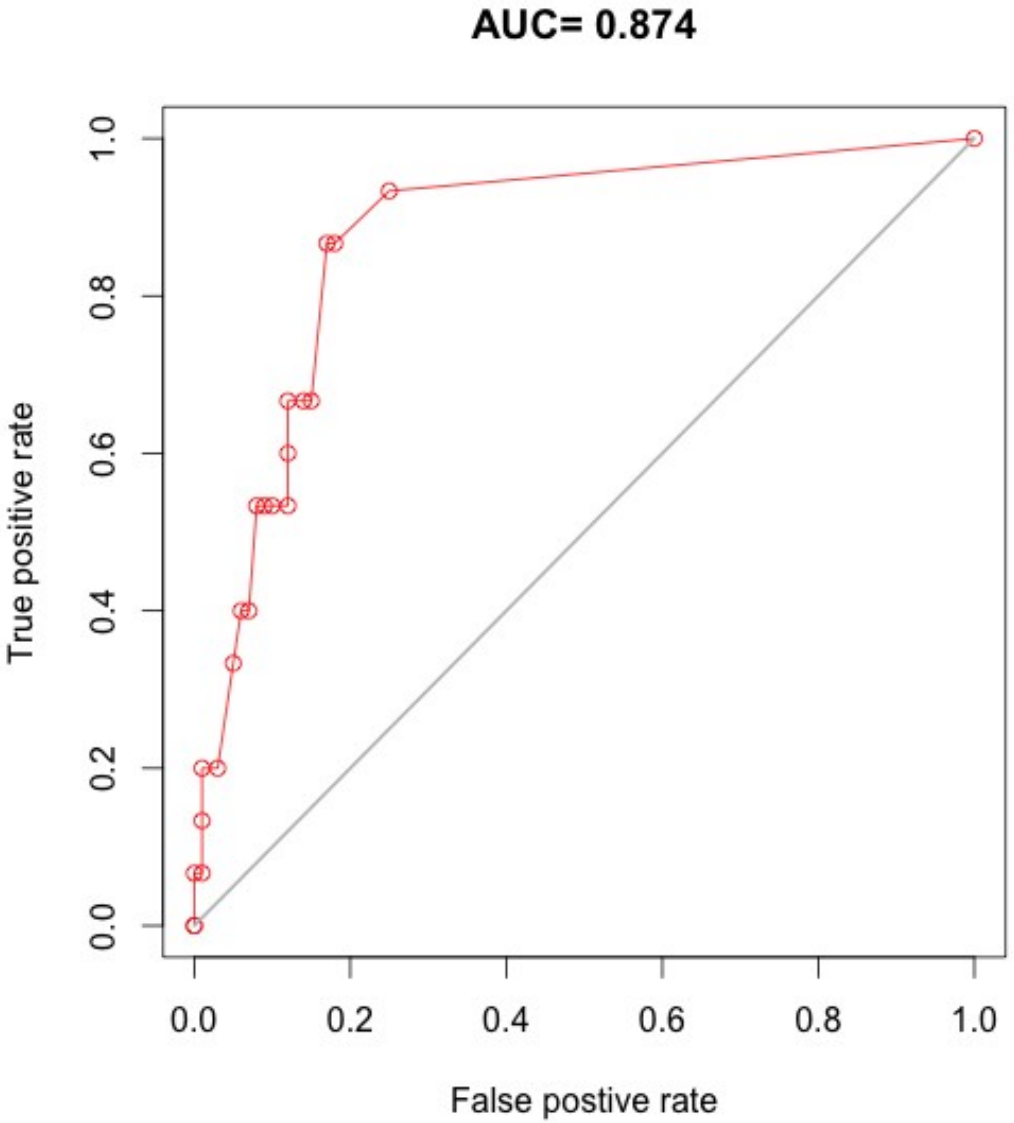
	REGISTRO DE PRESENÇA	REGISTRO DE AUSÊNCIA
PRESENÇA PREDITA	A	B
AUSÊNCIA PREDITA	C	D

$$Sensibilidade = \frac{A}{(A+C)}$$

$$Especificidade = \frac{D}{(B+D)}$$

Análise ROC (cálculo da área sob a curva – AUC):

```
plot(e,"AUC")
```



A pesar de ser muito utilizada no passado, hoje tem caído em desuso:

- Dá igual importância aos erros de omissão e comissão
- Varia com a prevalência da espécie, espécies mais especialistas têm AUC maiores porque acertar as ausências é fácil.

VALIDAÇÃO DOS MODELOS – Considerações importantes

3. **Avaliação do especialista na biogeografia da espécie modelada**

4. **Teste de campo, nada substitui esta validação!!**

VALIDAÇÃO DOS MODELOS – Considerações importantes

Avaliação do especialista na biogeografia da espécie modelada

Teste de campo, nada substitui esta validação!!

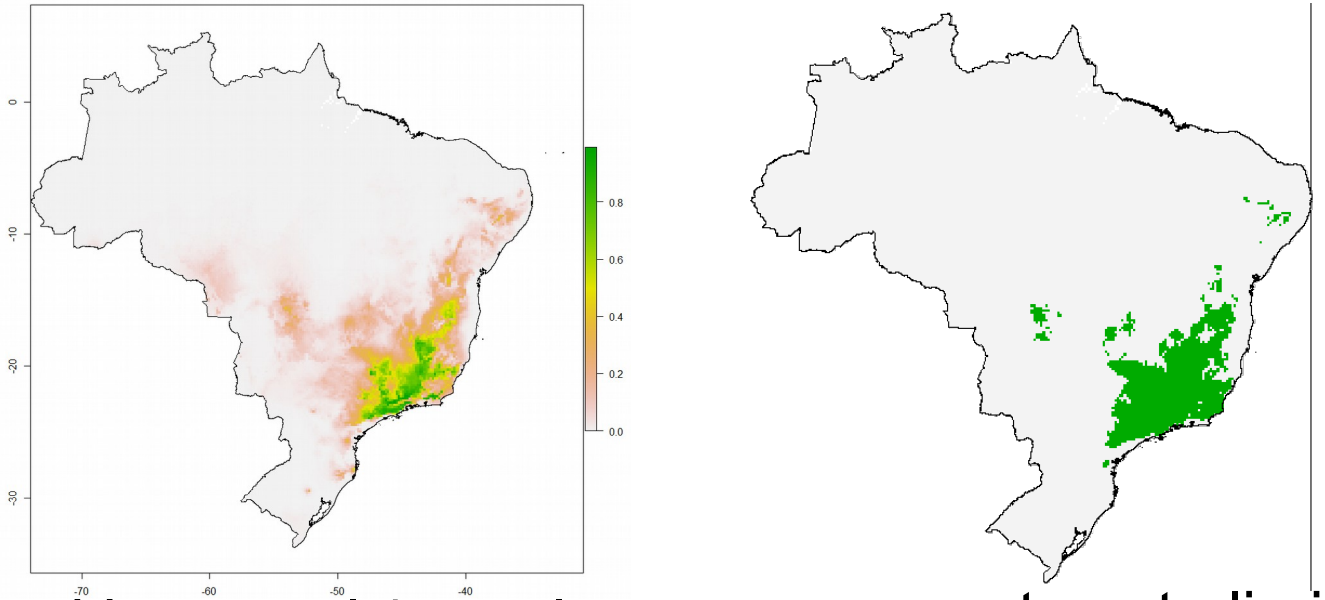
VALIDAÇÃO DOS MODELOS – Considerações importantes

Todo o trabalho para validar os modelos tem por objetivo especificar o quanto o seu modelo foi bom em predizer os pontos de teste (independentes), ou seja, o quanto o seu modelo foi robusto.

Ao se fazer diversas partições nos dados pode-se dizer: **“foram feitas x partições aleatórias dos dados e o valor da estatística de validação externa (teste independente) médio foi de 0.91 +/- 0.04”.**

VALIDAÇÃO DOS MODELOS – Considerações importantes

A validação depende do estabelecimento de um limite de corte. **Este valor representa o limiar entre presença (provável) e ausência (provável) que é conhecido como *threshold***.



Dica: para definir/justificar qual e o valor que representa este limiar deve-se pensar (alinhar) isto com a questão/pergunta a ser respondida pela modelagem e com a qualidade dos dados de origem. Ou seja, que tipo de erro você quer priorizar em relação aos dois tipos de erro associado ao processo (omissão ou comissão)? Ou não, caso você não queira priorizar nenhum deles, e sim trata-los igualmente.

Análises pós-modelagem

- Depende da pergunta inicial
- Consideração de variáveis que não entraram na modelagem: uso da terra, cobertura, etc.
- Interações bióticas
- Modelos multi-espécie
- Projeção no tempo e no espaço

O MNE não é o fim!

Resumindo...

1. Definir a pergunta;
2. Estabelecer a abrangência geográfica/ambiental do estudo;
3. Levantar os dados bióticos e abióticos referentes a pergunta. Verificar se a qualidade e a quantidade dos dados bióticos e abióticos são compatíveis e suficientes;
4. Selecionar quais dados (bióticos e abióticos) serão usados no projeto;
5. Escolher o(s) algoritmo(s) para modelagem;
6. Fazer o desenho amostral do modelo para a avaliação.