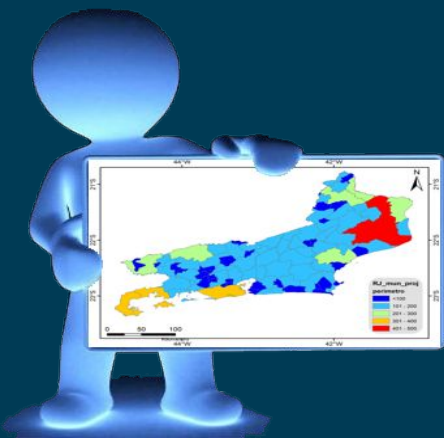


Tópicos Especiais em RGV II

Modelagem de adequabilidade ambiental



Diogo S. B. Rocha

Validação de Modelos



Registros de ocorrência

Dados de entrada

Variáveis preditoras



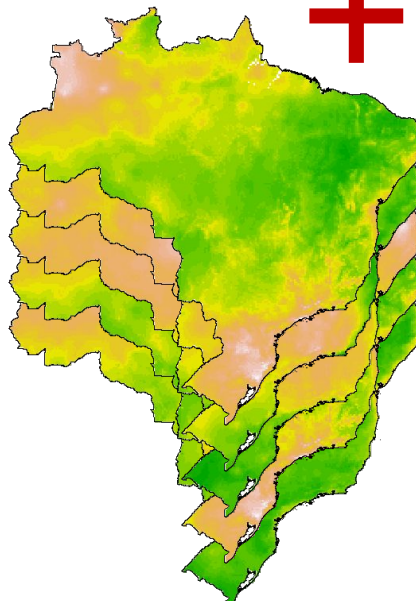
Precipitação

Temperatura

Topografia

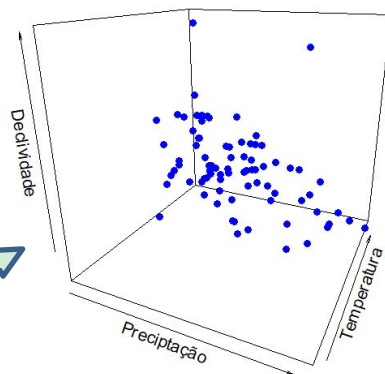
Variável X

Variável Y

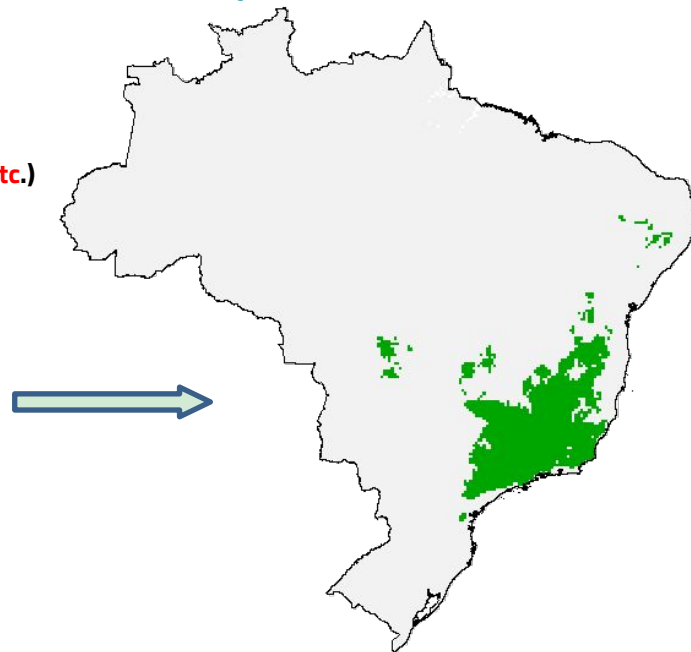


Algoritmos de modelagem

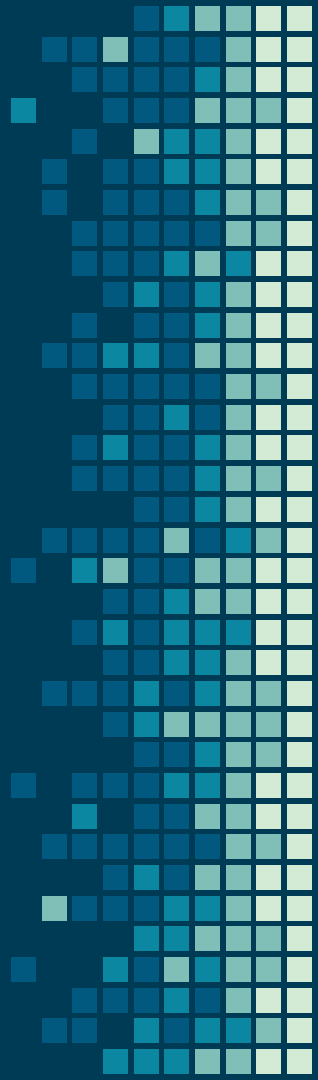
(Bioclim, GLM, GAM, ANN, M_AxEnt, etc.)



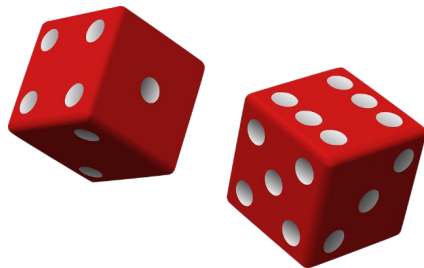
Mapa de distribuição potencial



Como sabemos se
um modelo é bom?



- Retorno ao campo

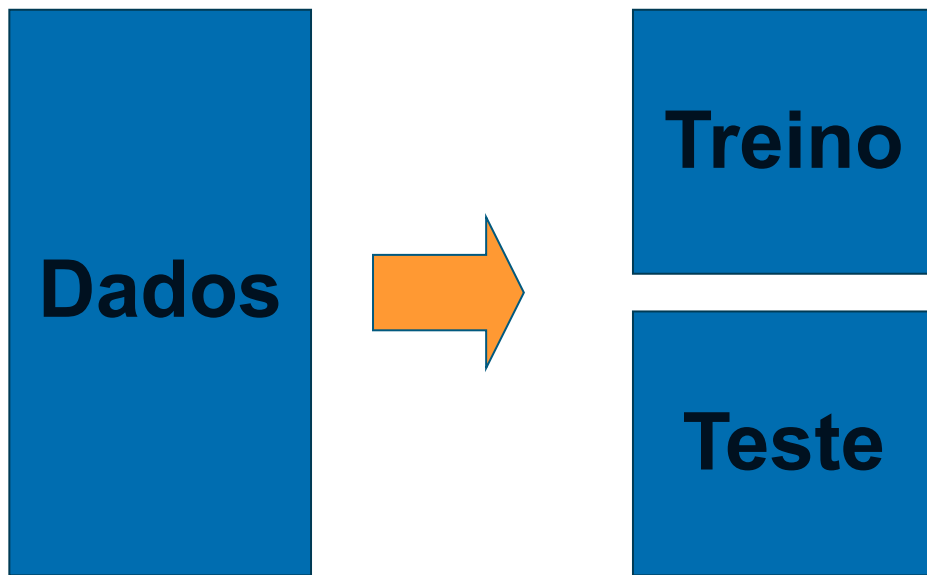


- Consulta ao especialista na biogeografia da espécie modelada



Ou ...

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



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.



Algumas definições!

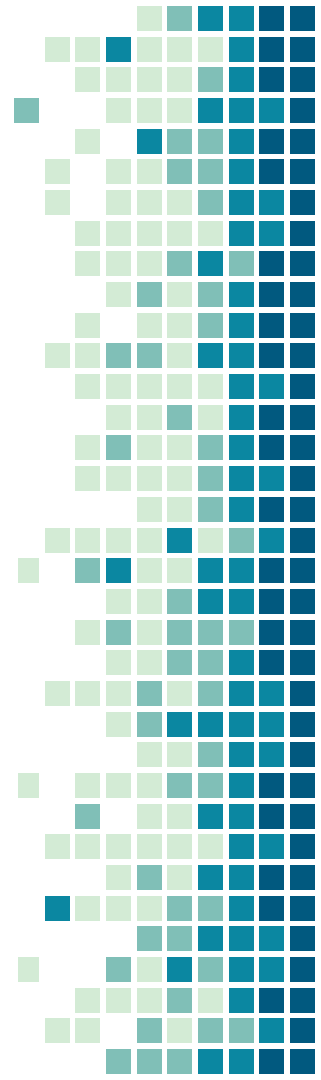
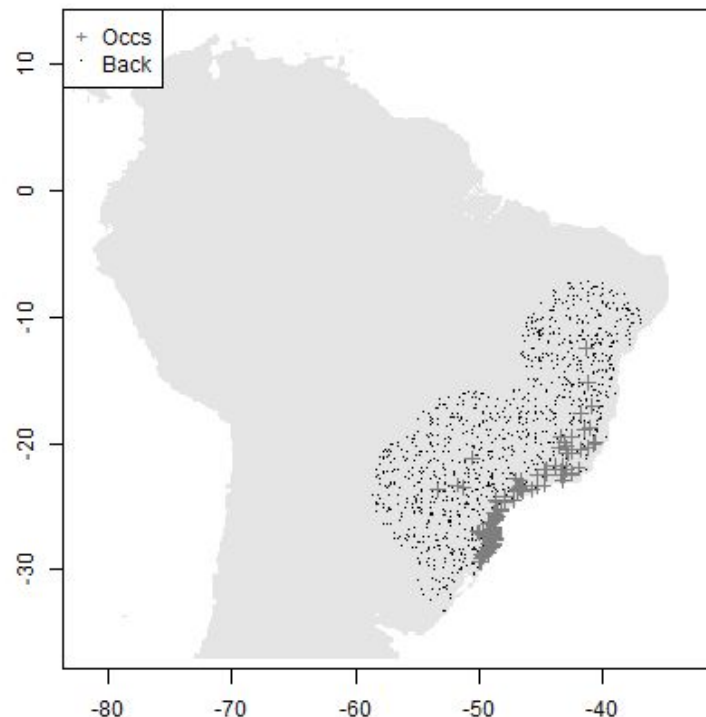
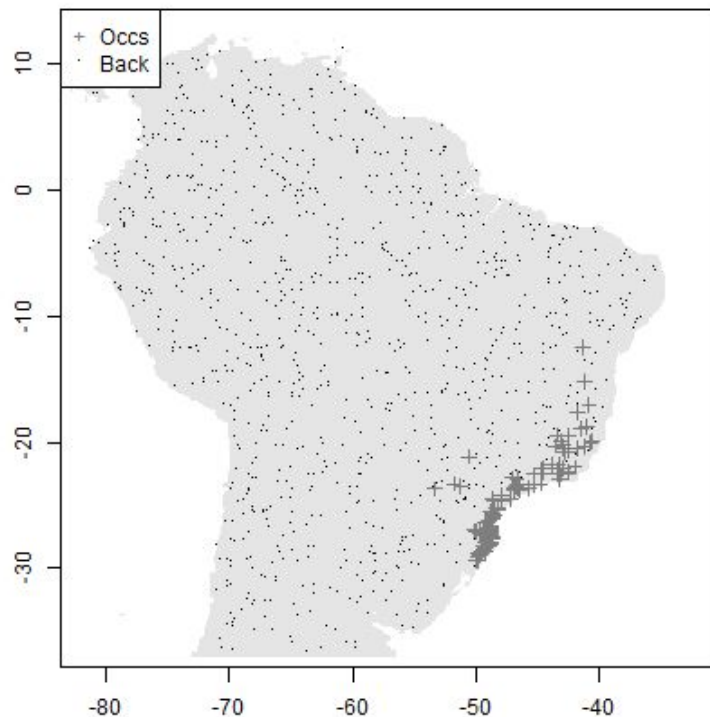
Dados de presença – pontos de ocorrência da espécie

Dados de ausência – registros de ausência da espécie, caso não estejam disponíveis, é necessário gerar um conjunto de pseudoausências.

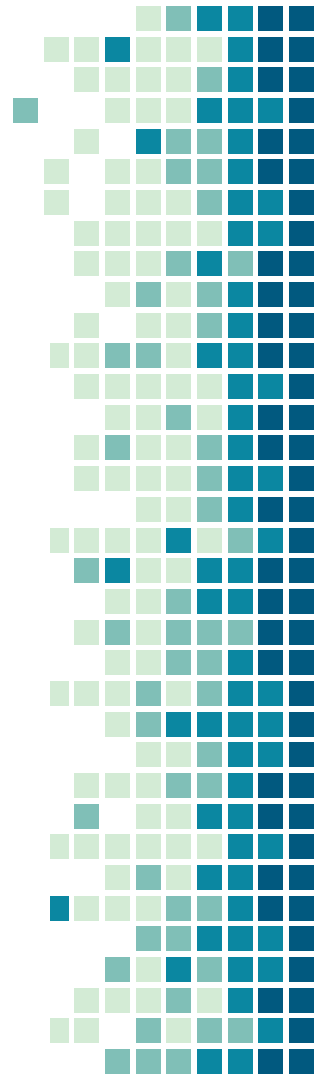
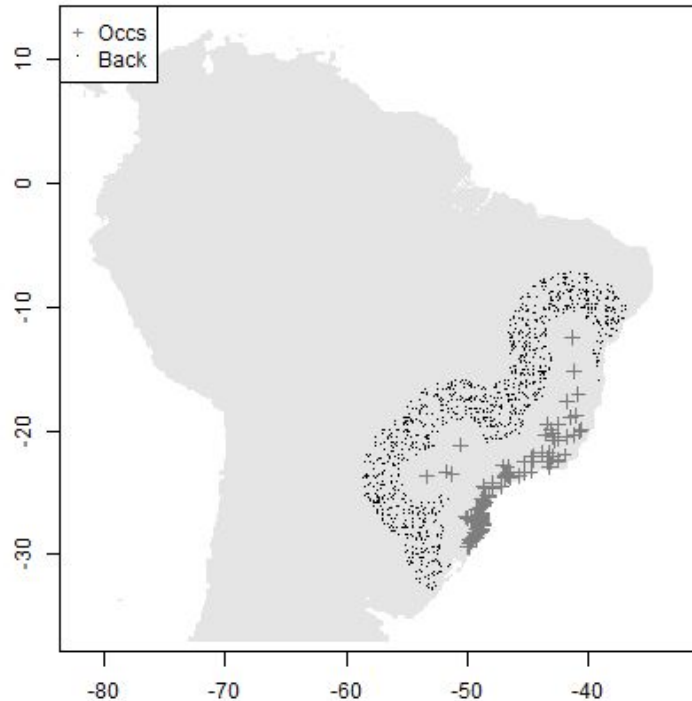
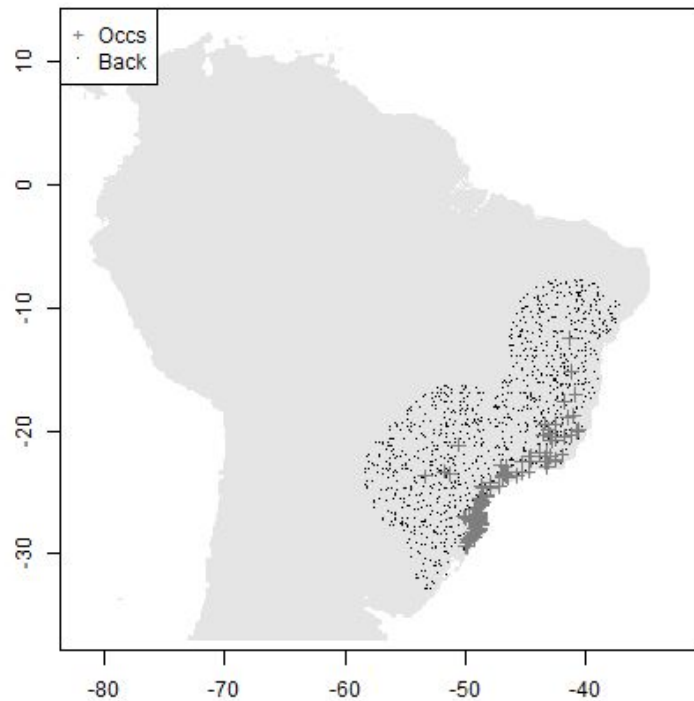
Onde?



Área de calibração do modelo



Área de calibração do modelo



Então, para validar um modelo 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
3. Sobrepor o conjunto de teste ao modelo e quantificar os erros através de uma matriz de confusão

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



Mas não precisamos ficar com apenas dois conjuntos

Os dados podem ser divididos em vários conjuntos de treino e teste

Com isso, podemos:

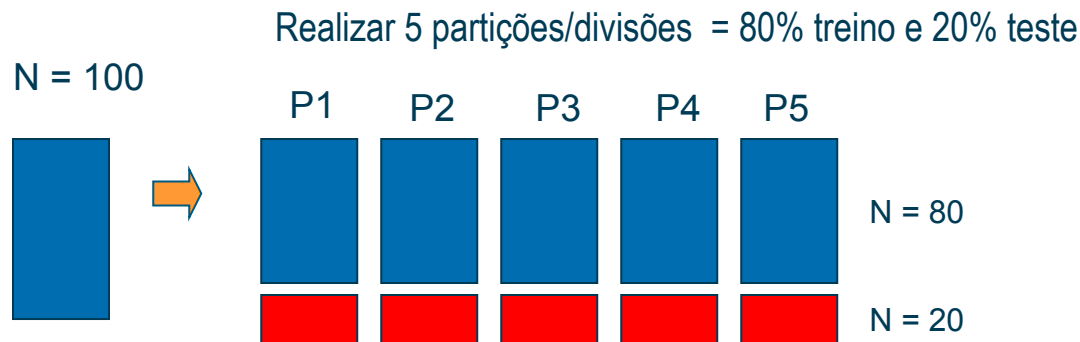
- calcular a variabilidade dos resultados (média \pm desvio padrão)
- avaliar a qualidade dos pontos
- comparar melhor os resultados de diferentes algoritmos

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.

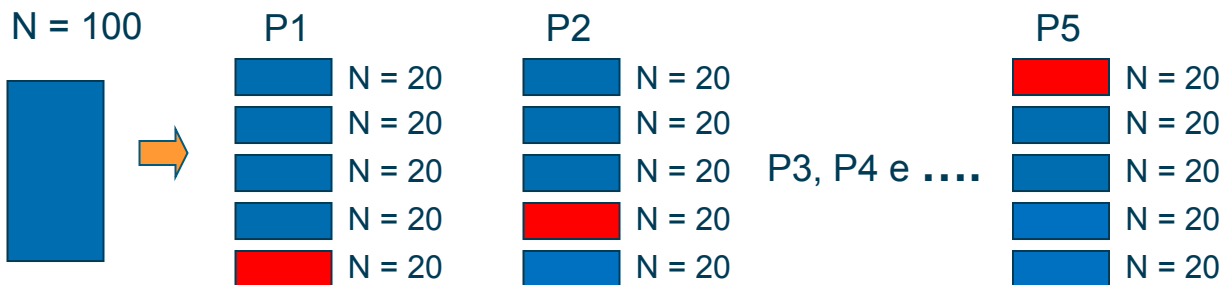
Bootstrap



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

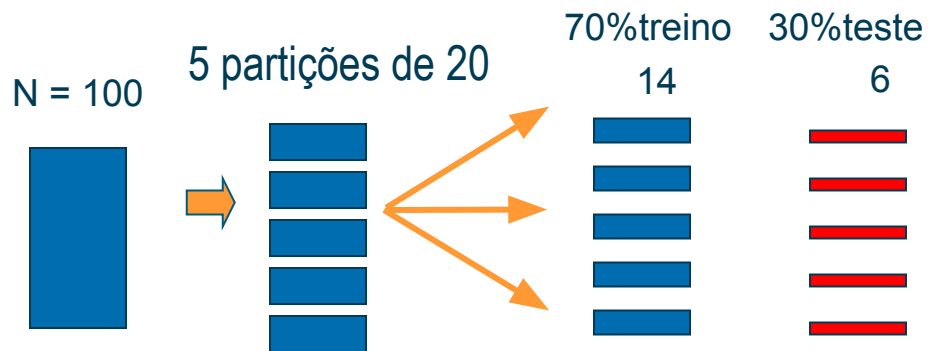
Cross-validation

Realizar 5 partições/divisões iguais nos dados



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

Sub-sample

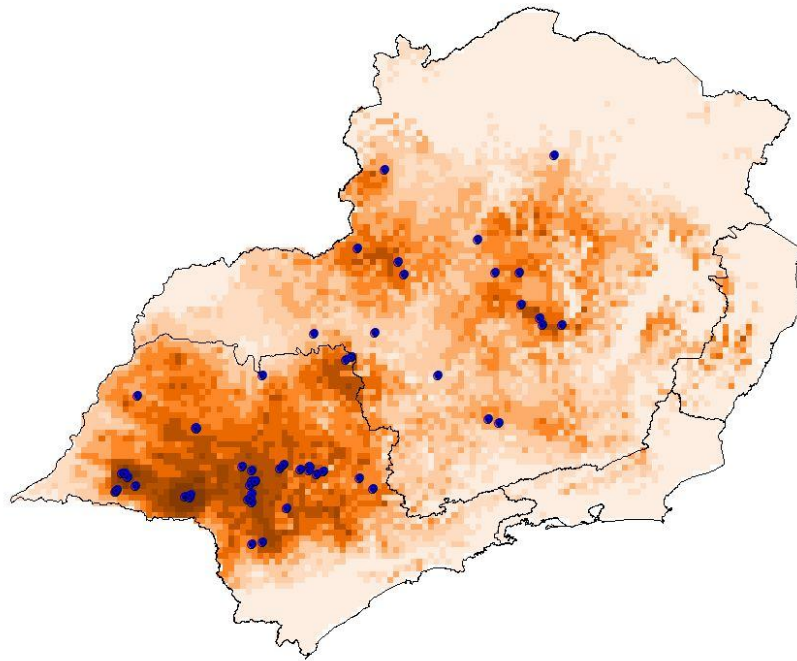


Para avaliar a qualidade do modelo gerado é preciso:

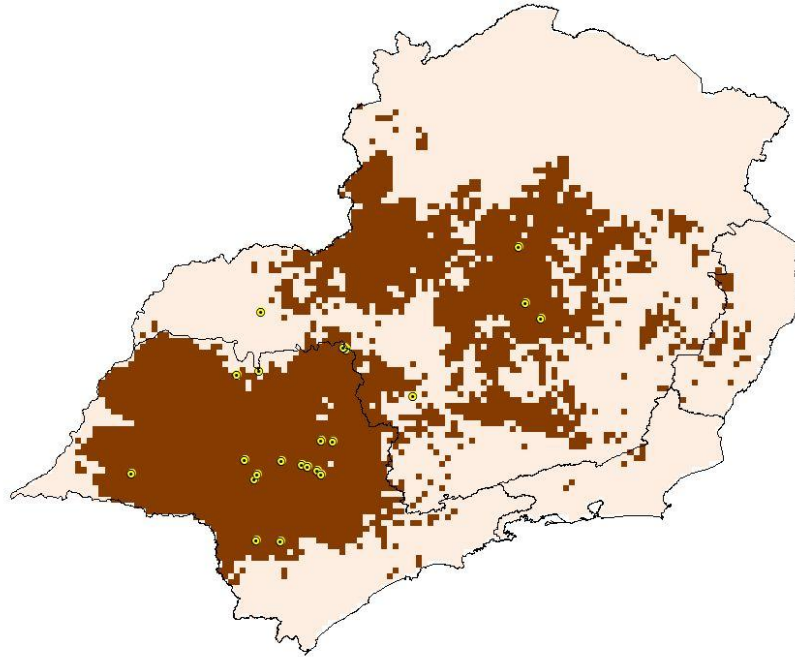
1. Ter um conjunto de teste. Não é aceitável testar um modelo a partir dos pontos que o geraram! Isso não faria sentido!
2. Quantificar os componentes de erro através de uma **matriz de confusão** sobrepondo os pontos de teste ao modelo gerado pelo conjunto de treino



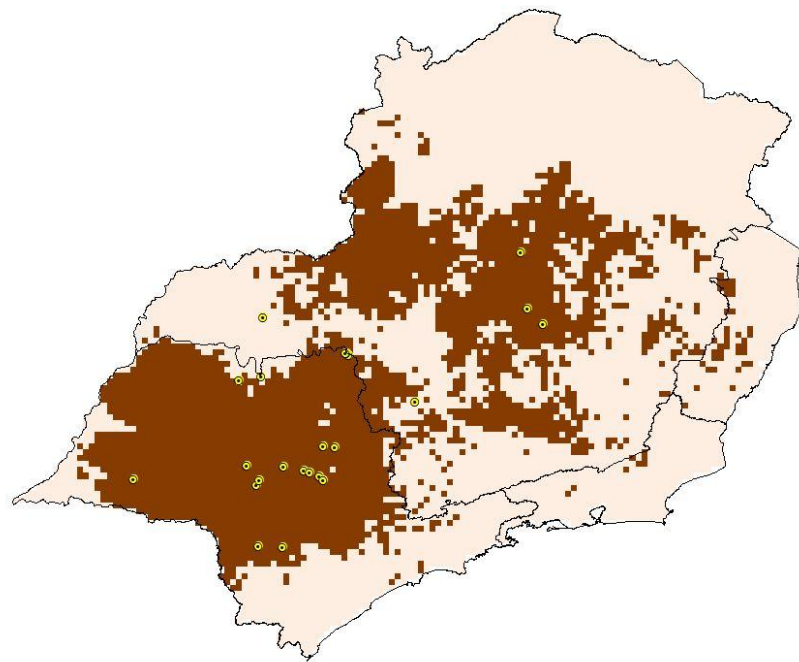
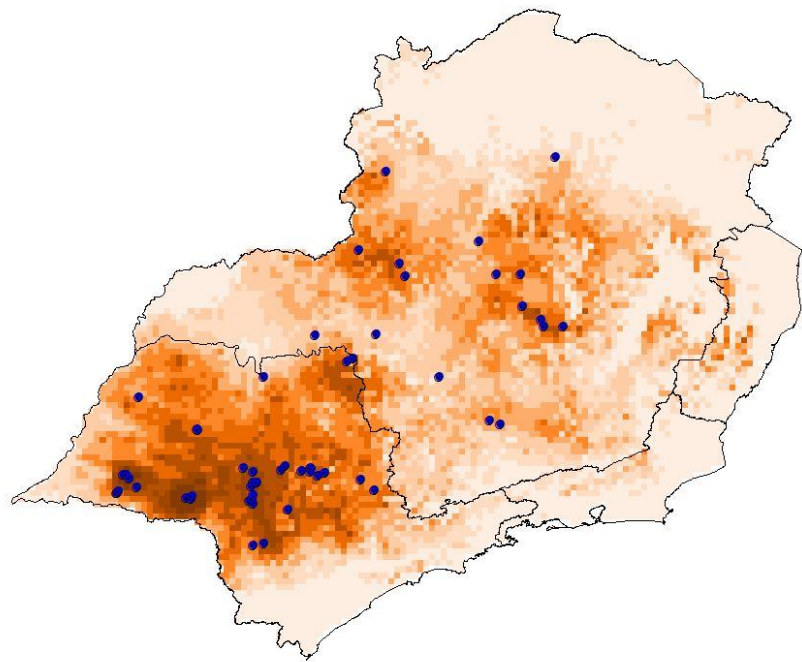
Gerar um modelo com o conjunto de dados de treino



Testar o modelo com o conjunto de teste



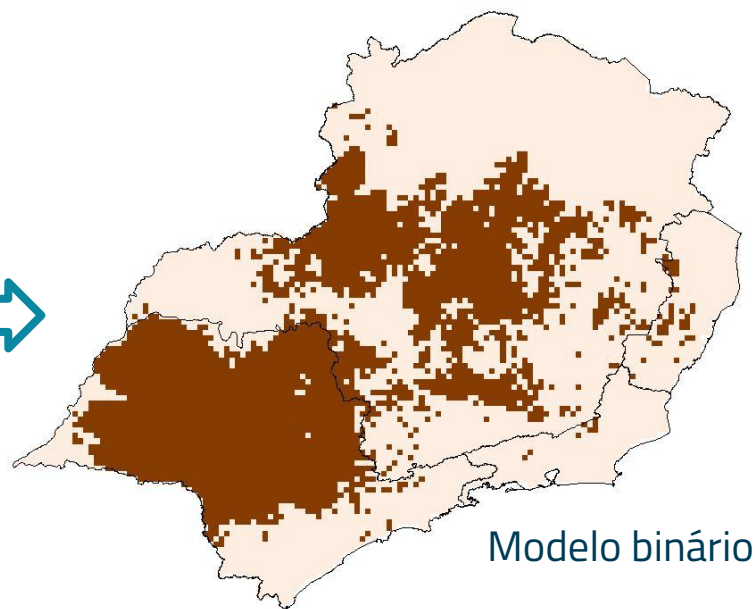
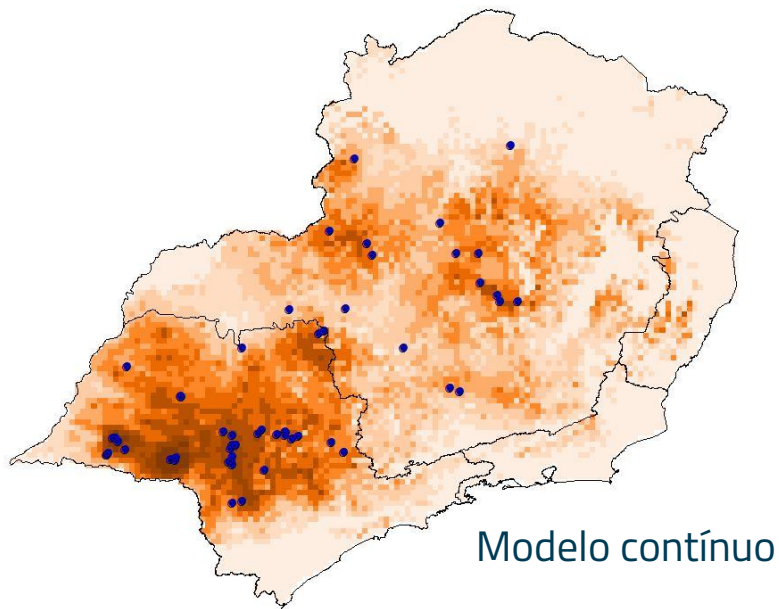
Qual a diferença?



Limiar de corte (threshold)

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.

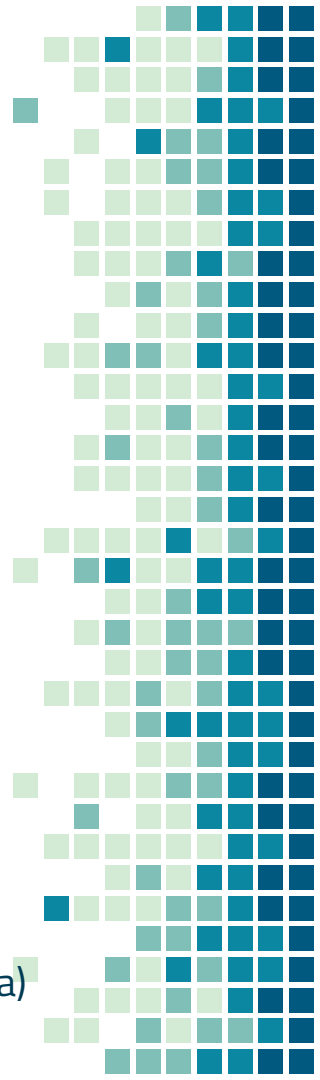


Como fazemos isso?

Boas revisões dos limites de corte (thresholds) utilizados:

(Liu et al. 2005, Pearson et al. 2007, 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)

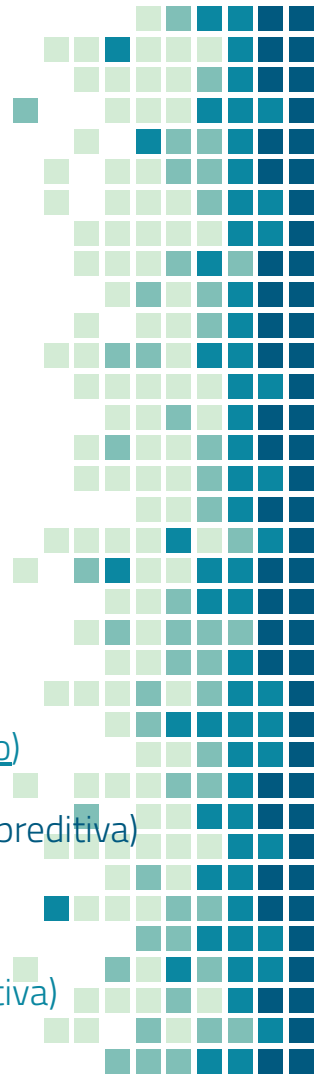


Como fazemos isso?

Boas revisões dos limites de corte (thresholds) utilizados:

(Liu et al. 2005, Pearson et al. 2007, 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)



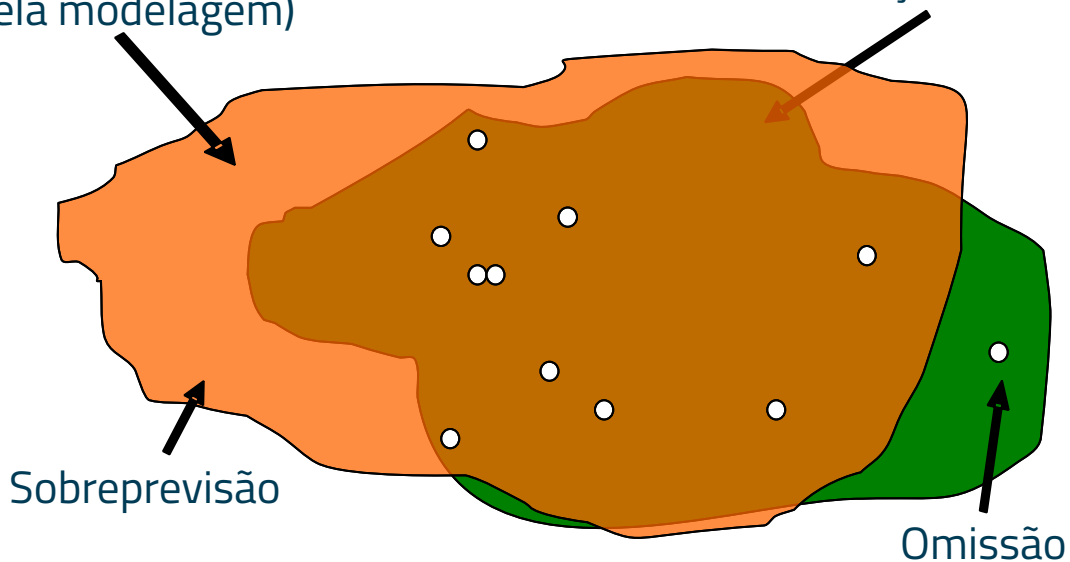
Mas e depois, como contabilizamos os erros e acertos do teste?

Há dois tipos de erro em modelagem

Erro de omissão e de sobreprevisão

Distribuição potencial
(prevista pela modelagem)

Distribuição conhecida da espécie



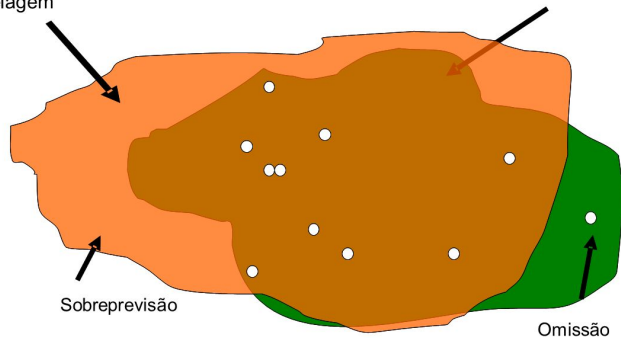
Matriz de Confusão

A e D são acertos
B e C são erros

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

Distribuição potencial (prevista pela modelagem)

Distribuição conhecida da espécie



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

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

$$\text{Sobreprevisão} = \frac{B}{(B + D)}$$

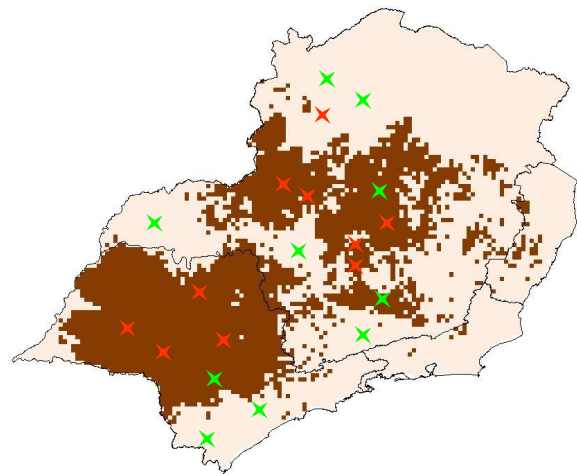
$$\text{omissão} = \frac{C}{(A + C)}$$

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

Taxas de acertos

Taxas de erros

Matriz de Confusão



✗ registros de presença
✗ registros de ausência

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

$Sensibilidade = \frac{A}{(A+C)}$
$Especificidade = \frac{D}{(B+D)}$
$Sobrepresvis\tilde{a}o = \frac{B}{(B+D)}$
$omiss\tilde{a}o = \frac{C}{(A+C)}$
$Acur\acute{a}cia^* = \frac{A+D}{A+B+C+D}$

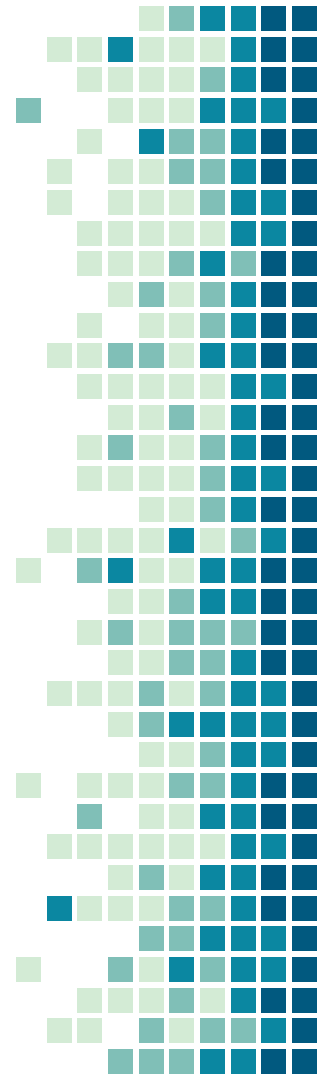
0.9

0.7

0.3

0.1

0.8



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

Erro de omissão: em 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 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 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

True Skill Statistic (TSS)

- $TSS = (\text{sensibilidade} + \text{especificidade}) - 1$
- $TSS = ((A/A+C) + (D/B+D)) - 1$

	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!)

$\text{Sensibilidade} = \frac{A}{(A+C)}$
$\text{Especificidade} = \frac{D}{(B+D)}$

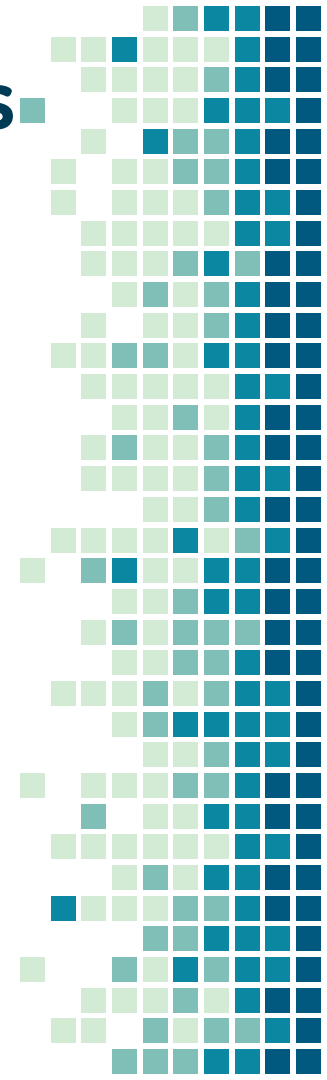
Obs: A TSS pode variar de -1 a 1. Quanto mais próximo de 1 melhor é o modelo. No geral, acima de 0.6 considera-se um bom ajuste do modelo aos dados. Entre 0.2 – 0.6 um ajuste regular e abaixo de 0.2, um ajuste ruim.

Outras estatísticas para validar os modelos

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).

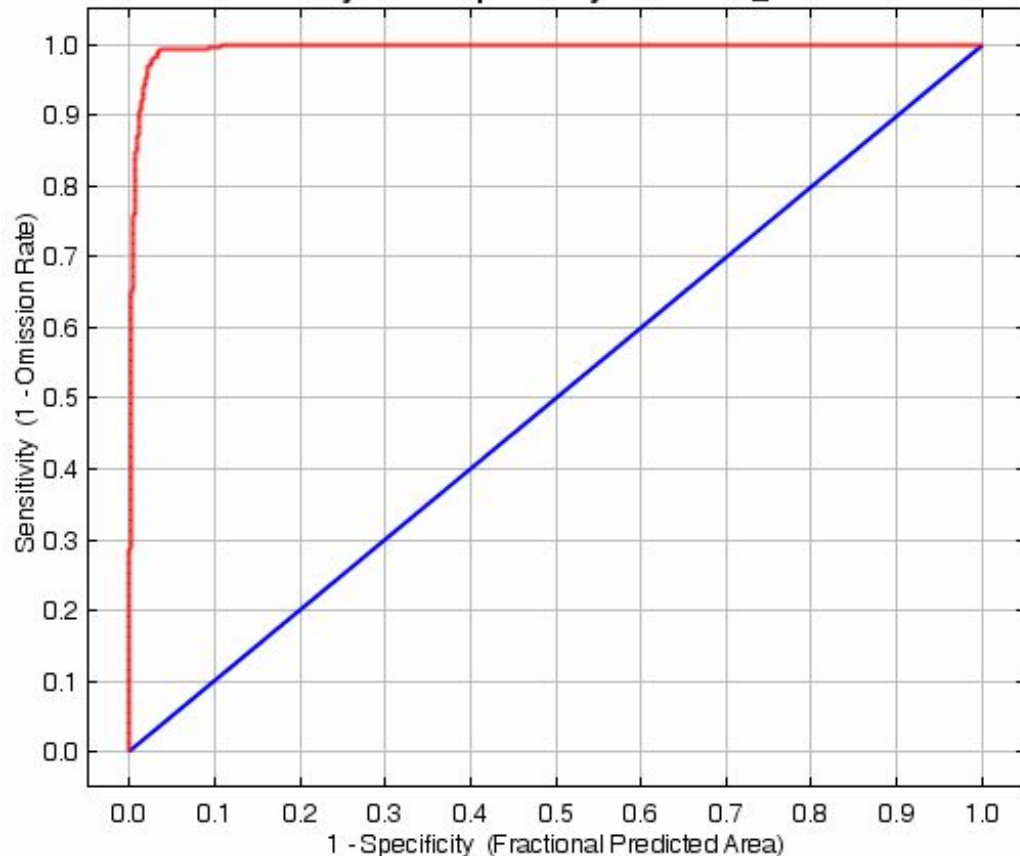
É obtida plotando-se a sensibilidade no eixo y e o valor 1-especificidade no eixo x. Quanto mais próximo de 1 for a área sob a curva, mais distante o resultado do modelo é da previsão aleatória, ou seja, melhor o desempenho do modelo.

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):

Sensitivity vs. 1 - Specificity for mabea_fistulifera



Training data (AUC = 0.995) ■
Random Prediction (AUC = 0.5) ■

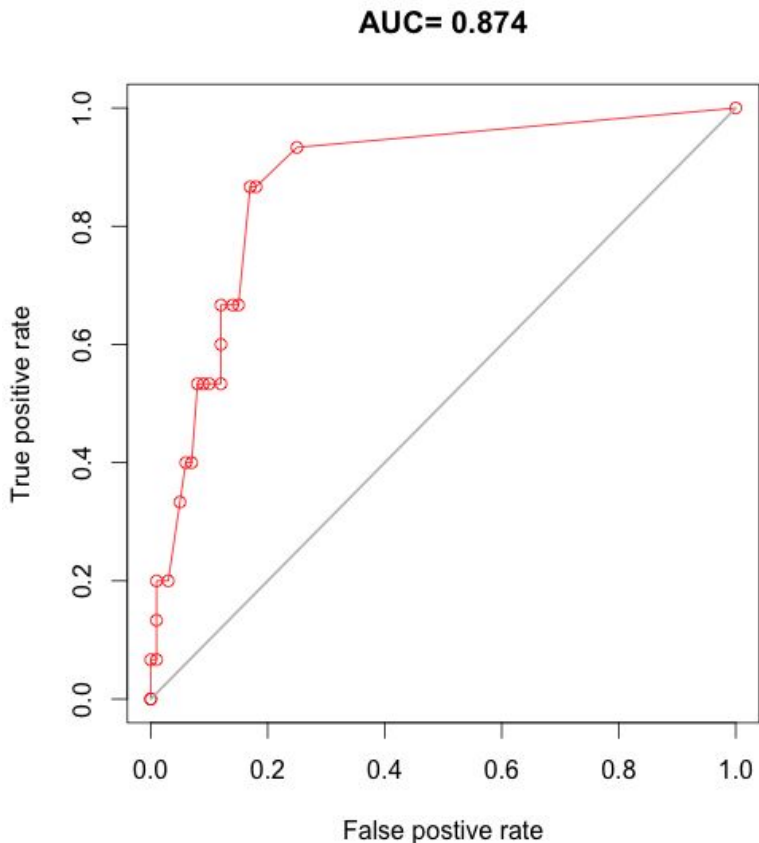
	REGISTRO DE PRESENÇA	REGISTRO DE AUSÊNCIA
PRESENÇA PREDITA	A	B
AUSÊNCIA PREDITA	C	D

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

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

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

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



Apesar 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!!



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!