Modelos de distribuição geográfica de *Amaioua guianensis* Aubl. em Minas Gerais, Brasil

Geographic distribution models of *Amaioua guianensis* Aubl. in Minas Gerais, Brazil

Gleyce Campos Dutra[1]
Luis Marcelo Tavares de Carvalho[2]

Resumo

O objetivo deste trabalho é avaliar o desempenho de 4 modelos utilizando diferentes conjuntos de variáveis ambientais, variando em complexidade, na tentativa de prever a distribuição geográfica de *Amaioua guianensis* Aubl. no Estado de Minas Gerais. Os pontos de ocorrência da espécie foram recuperados no banco de dados TreeAtlan 1.0 para o Estado de Minas Gerais. As bases ambientais utilizadas para o trabalho compreendem coberturas climáticas relacionadas com temperatura e precipitação, dados relativos ao relevo, distância do oceano, índices de vegetação do sensor MODIS, tipo de solo e litologia. Para a modelagem de distribuição da espécie foi utilizado o algoritmo de Máxima Entropia (Maxent). Quatro alternativas de conjuntos de variáveis, foram gerados: com toda a base de dados, só com as variáveis bioclimáticas, com as variáveis selecionadas por meio da CCA e com as variáveis selecionadas por meio de uma análise prévia do teste Jackknife para todas as variáveis. A análise do desempenho dos modelos foi feita utilizado a área sob a curva ROC e taxas de omissão extrínsecas. As simulações demonstram que a seleção de variáveis ambientais mais relevantes para uma determinada espécie produz modelos mais acurados.

Palavras-chave: modelagem de distribuição de espécies; Maxent; *Amaioua guianensis*.

Abstract

The objective of this work is to evaluate the performance of 4 models using different environmental datasets in an attempt to predict the geographic distribution of *Amaioua guianensis* Aubl. in the State of Minas

1 Departamento de Ciências Florestais – Universidade Federal de Lavras (UFLA), Caixa Postal 3037 - CEP 37200-000 – Lavras – MG – Brasil; e-mail: gleycedutra@yahoo.com.br

2 Departamento de Ciências Florestais – Universidade Federal de Lavras (UFLA), Caixa Postal 3037 - CEP 37200-000 – Lavras – MG – Brasil;e-mail: gleycedutra@yahoo.com.br

Gerais. The specie occurrence points was recovered from the TreeAtlan 1.0 database for Minas Gerais. The environmental datasets used for the work involved climatic coverings related to temperature and precipitation, elevation, distance from the ocean, MODIS vegetation indices, soil classes and litology. The Maximum Entropy method (Maxent) was used for modeling specie distribution. Four alternative environmental datasets was input to Maxent: (a) all variables, bioclimatic variables, variables selected by CCA and variables selected using the Jackknife test applied to all variables. The performance of the models was analyzed using the area under ROC curve and extrinsic omission rates. The simulations show that the selection of environmental variable produced the most accurate results.

Key words: species distribution modelling; Maxent; *Amaioua guianensis.*

Introdução

Amaioua guianensis Aubl. é uma espécies arbórea encontrada nas fitofisionomias: Florestas Estacionais Semideciduais em diferentes faixas altitudinais, Floresta Ombrófila Densa Baixo-Montana e Alto Montana, Floresta de Galeria e Cerrado *sensu lato* (OLIVEIRA-FILHO, 2006). Ocorre principalmente no interior de matas primárias, sendo uma espécie de subbosque, e capoeirões sobre terrenos inclinados de solos arenosos (LORENZI, 2002).

A maior parte das espécies de plantas tropicais ainda não é bem caracterizada em relação a sua distribuição geográfica, tornando a estimativa de sua distribuição geográfica, uma importante ferramenta de análise para dar suporte às políticas de conservação e ao planejamento de estratégias de recuperação de diversas áreas.

Os modelos de previsão de distribuição de espécies receberam uma grande atenção nas duas últimas décadas, dada a sua potencialidade para extrapolar a informação sobre a distribuição de espécies, em áreas onde há a falta dessa informação. É possível gerar mapas que indicam a provável presença e ausência de uma espécie, em função de variáveis ambientais relevantes (SIQUEIRA e PETERSON, 2003). Essas variáveis definem um hiperespaço e a extensão desse espaço, delimitada pelas exigências de determinada espécie, é o seu nicho.

O nicho ecológico fundamental pode ser conceituado como as condições em que a espécie pode existir sem atuação de fatores bióticos limitantes. Indica as respostas dos indivíduos aos parâmetros físicos, como temperatura, precipitação, elevação, geologia, vegetação, inferindo regiões geográficas da aptidão positiva, no qual combinações das variáveis ambientais associadas com a presença observada da espécie podem ser identificadas e projetadas em num especo geográfico (SOBERÓN e PETERSON, 2005).

Desenvolvimentos recentes nos Sistemas de Informação Geográfica (SIG) tornaram possível o armazenamento e análise quantitativa de um grande número de dados espaciais, disponibilizando a informação acerca de variáveis ambientais para várias localidades. Porém a inserção

exagerada de variáveis ambientais na tentativa de entender a distribuição potencial de uma espécie, pode acarretar diversos problemas no processo de modelagem. De acordo com Peterson et al. (2007) o sobre-ajuste (*overfitting*) é provavelmente mais sério em espaços ambientais altamente dimensionais. Sem contar que o custo computacional dos processos de modelagem é proporcional à complexidade dados.

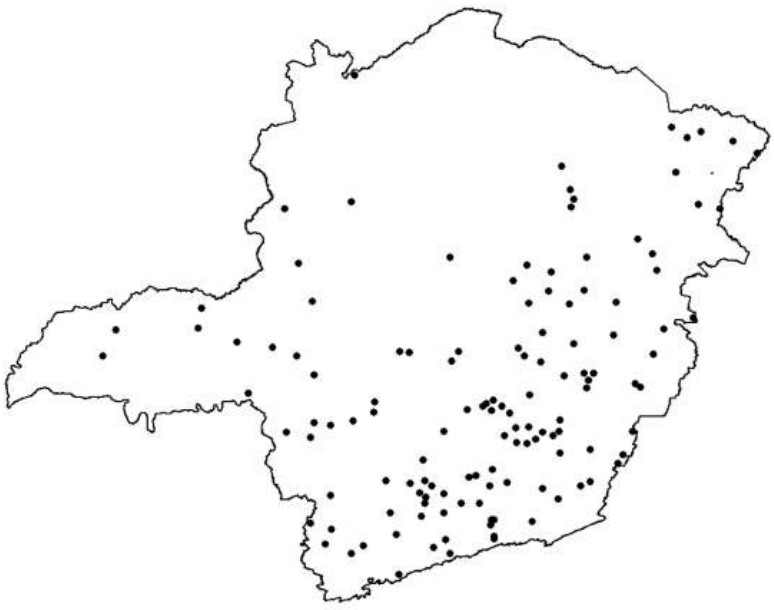
A complexidade da superfície de decisão de um algoritmo é proporcional ao número de parâmetros livres que ele possui, ou seja, as variáveis utilizadas para a delimitação do nicho fundamental de uma espécie. Quando o número de variáveis é grande, o algoritmo tende a se adaptar a detalhes específicos da base de treinamento, o que pode causar uma redução da taxa de acerto. Esse fenômeno é conhecido como sobre-ajuste. Para evitá-lo, é desejável que o classificador seja o mais simples possível, pois assim será dada mais importância às maiores regularidades nos dados e as menores serão ignoradas, pois essas podem ser resultantes de ruídos (CAMPOS, 2001).

O objetivo geral deste trabalho é avaliar o desempenho de quatro modelos utilizando diferentes conjuntos de variáveis ambientais, varian outro sinônimo que ache melhordo em níveis de complexidade, na modelagem da distribuição geográfica de *A. guianensis* Aubl. no Estado de Minas Gerais.

Materiais e Métodos

Os pontos de ocorrência de fragmentos de matas semidecíduas e ombrófilas foram recuperados no banco de dados TreeAtlan 1.0 para o Estado de Minas Gerais (OLIVEIRA-FILHO, 2008), totalizando 130 áreas (Figura 1). Selecionou-se a espécie *A. guianensis*

Figura 1. Registros de ocorrência de *Amaioua guianensis* Aubl.. em Minas Gerais



Aubl. por apresentar alto valor de indicação para essas fitofisionomias segundo metodologia proposta por Dufrêne e Legendre (1997) para o Estado de Minas Gerais, em trabalho realizado por Dutra et al. (2008).

Uma série de variáveis climáticas foi obtida do WorldClim versão 1.4 (HIJMANS et al., 2005), que incluem onze variáveis relacionadas à temperatura e oito relacionadas à precipitação. Foram também incluídos dados relativos ao relevo, distância do oceano, índices de vegetação do sensor MODIS, tipo de solo e litologia. Os dados apresentaram a seguinte estrutura: temperatura média anual, variação diurna média, isotermalidade, sazonalidade de temperatura, temperatura máxima no período mais quente, temperatura mínima no período mais frio, variação de temperatura anual, temperatura média no trimestre mais úmido, temperatura média no trimestre mais seco, temperatura média no trimestre mais quente, temperatura média no trimestre mais frio, precipitação anual, precipitação no período mais úmido, precipitação no período mais seco, sazonalidade da precipitação, precipitação no trimestre mais úmido, precipitação no trimestre mais seco, precipitação no trimestre mais quente, precipitação no trimestre mais frio, modelo de elevação, declividade, aspecto e convexidade do perfil, os índices de vegetação NDVI e EVI de junho de 2003, NDVI e EVI de abril de 2004, distância do oceano, tipo de solo, classes litológicas. Totalizando trinta variáveis. Os dados foram reamostrados para uma resolução de 0,0083 graus (aproximadamente 1 km) e, casos necessários, dados de melhor resolução foram degradados para serem compatíveis com dados em menores resoluções.

Foi realizada uma análise de correspondência canônica (CCA) relacionando a matriz de ocorrência das espécies do banco de dados original com as trinta variáveis ambientais. Por meio de uma rotina de seleção progressiva de variáveis, associada a testes de permutação de Monte Carlo, foi verificada a significância das mesmas, no programa CANOCO versão 4.5. Foram selecionadas 15 variáveis ambientais: isotermalidade, temperatura média no trimestre mais úmido, temperatura média no trimestre mais frio, precipitação anual, precipitação no período mais úmido, sazonalidade da precipitação, precipitação no trimestre mais úmido, precipitação no trimestre mais seco, precipitação no trimestre mais quente, precipitação no trimestre mais frio, modelo de elevação, declividade, NDVI de junho de 2003, distância do oceano, tipo de solo, classes litológicas.

Os pontos de ocorrência de *A. guianensis* Aubl., associadas às bases ambientais, foram usados para modelar a sua distribuição geográfica potencial aplicando o algoritmo de Máxima Entropia (PHILIPS et al., 2006). A máxima entropia (Maxent) é um método para realizar previsões ou inferências a partir de informações incompletas e vem sendo aplicado recentemente na modelagem de distribuição de espécies.

Maxent ajusta a probabilidade da distribuição de ocorrência de uma determinada espécie para um conjunto de *pixels* da região de estudo baseado na idéia de que a melhor explicação para o fenômeno desconhecido é aquela que maximizará a entropia da distribuição de probabilidade (PHILIPS et al., 2006). Apresenta, entre outras, vantagens de se basear em dados

de presença, a possibilidade de utilizar bases ambientais contínuas e discretas e também de incorporar interações entre as diferentes variáveis. Possui, ainda, acrescentam os autores, a vantagem de que a saída do modelo é contínua, permitindo uma fina distinção entre os modelos gerados para diferentes áreas.

Para avaliar a qualidade do modelo geramos um conjunto independente de dados divididos em dois conjuntos (treino e teste) antes de efetuar a modelagem. Em seguida, realizamos a análise da curva característica de operação (ROC) que avalia o desempenho do modelo através de um único valor, que representa a área sob a curva (AUC). A análise ROC é baseada na medida da sensibilidade, que é a taxa de verdadeiros positivos (ausência de erro de omissão) *versus* a especificidade que é a taxa de falso positivo (erro de sobreprevisão).

Foram usados os parâmetros padrões utilizados na versão 2.3, com 30% dos pontos de ocorrência sendo usados como amostras teste e o restante como amostra de treinamento do algoritmo. Este algoritmo ainda possui a opção de rodar um teste Jackknife, que permite estimar a significância de uma variável ambiental individualmente na análise da distribuição da espécie (PHILIPS et al, 2006; SAATCHI et al., 2008). Desta forma, puderam ser excluídas, as variáveis que apresentaram valores menores de ganhos num modelo gerado com todas elas: variação de temperatura anual, aspecto, tipo de solo, EVI e NDVI abr2004, EVI jun2003, convexidade do perfil, precipitação do trimestre mais quente.

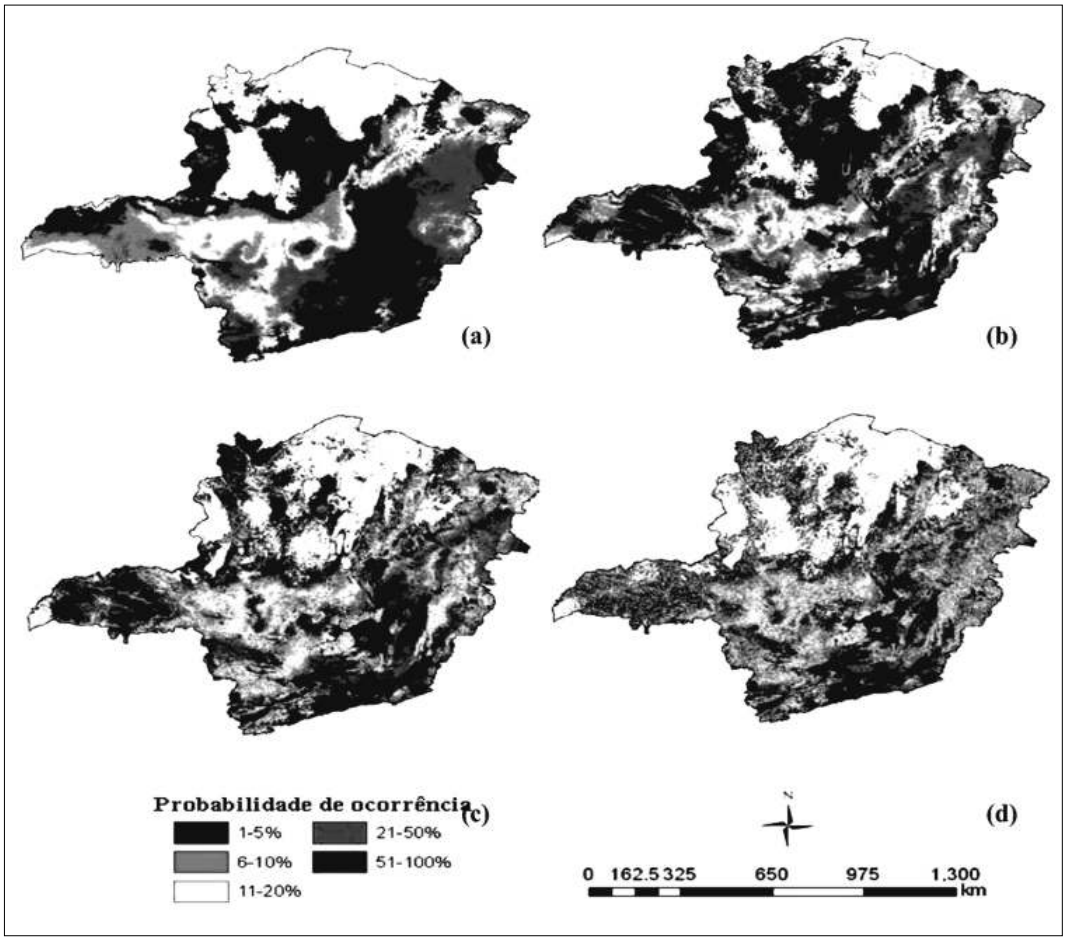
Foram gerados, com diferentes alternativas de conjuntos de variáveis, os seguintes modelos: com toda a base de dados (TUDO), só com as variáveis bioclimáticas (BIOS), com as variáveis selecionadas por meio de uma CCA (CNC), com as variáveis selecionadas por meio de uma análise prévia da base de dados total pelo teste Jackknife no programa Maxent (JKNF).

A validação dos modelos gerados foi feita utilizado o chamado gráfico do receptor-operador (ROC-plot), no qual são representadas as frações dos verdadeiros positivos contra os falsos positivos. A área sob a curva (AUC) é tomada como uma medida de acurácia do modelo e caracteriza o seu desempenho (FIELDING e BELL, 1997; PHILIPS et al., 2006). O Maxent tem uma rotina interna que examina a taxa de omissão extrínseca e a área sob a curva ROC (AUC) para um conjunto de localidades selecionadas aleatoriamente para a amostra teste. A taxa de omissão extrínseca, definida como a fração de locais das amostras teste que coincidem com *pixels* fora da área prevista, foi analisada com um limiar de 10% de probabilidade cumulativa, seguindo metodologia de Saatchi et al. (2008).

Resultados e Discussões

Observando a Figura 2, os modelos Maxent produziram previsões na forma de números reais entre 0 e 100, representando probabilidade de ocorrência acumulativa, ou seja, o valor da *pixel* é a soma das probabilidades daquele *pixel* e todos os outros *pixels* com igual o menor probabilidade, apresentados em porcentagem (PHILIPS et al., 2006). Assim, quanto maior o valor do *pixel*, maior a probabilidade

Figura 2. Previsão de distribuição geográfica potencial de *Amaioua guianensis* Aubl. geradas por quatro conjuntos de variáveis ambientais: a) só com as variáveis bioclimáticas – BIOS; b) com as variáveis selecionadas por meio da CCA – CNC; c) com as variáveis selecionadas por meio de uma análise prévia do Jackknife no programa Maxent – JKNF e d) com toda a base de dados – TUDO.



de ocorrência da espécie. Teoricamente, segundo Saatchi et al. (2008), qualquer *pixel* com probabilidade de ocorrência maior que 1% é considerado adequado para a espécie, mas interessa-se por áreas com maior probabilidade de ocorrência da espécie (>20%).

As previsões do Maxent para *A. guianensis* Aubl.(Figura 2) apontam que áreas com maior potencial de ocorrência, *pixels* com probabilidade de ocorrência superior a 20%, coincidem com áreas classificadas como Domínio Atlântico pelo IBGE (2004). Os modelos ainda extrapolam a extensão de ocorrência para áreas do Domínio do Cerrado, com probabilidade entre 11 e 20%, em regiões de transição dos dois domínios na área centro-leste de Minas Gerais e Triângulo Mineiro, com manchas de ausência em

áreas do Domínio da Caatinga, pois sendo uma espécie de subbosque, ocorre em regiões onde há relativa cobertura arbórea. Registros de ocorrência dessa espécie são encontrados em compilação realizada por Ratter et al. (2003) em áreas de Cerrado em São Paulo, Paraná e sul de Minas Gerais.

Para todos os modelos, as áreas sob a curva ROC (Tabela 1), geradas pelas amostras teste, foram maiores que o valor aleatório (0,5). 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. Esse valor de AUC pode ser resultado da interpretação da probabilidade que o modelo teve ao classificar corretamente a presença e para a dada espécie.

Figura 2d pode-se observar que o Modelo TUDO apresenta muita heterogeneidade entre grupos de *pixels*, isso mostra o esforço do algoritmo em tentar se adaptar à dimensionalidade desse conjunto de dados. Na tabela 1, ainda pode ser observado que este modelo previu uma menor extensão de ocorrência tanto para todas as probabilidades, como também para regiões de alta probabilidade.

Ainda na tabela 1, observa-se que o modelo BIOS apresentou maior extensão de ocorrência. Esse tendência de sobre previsão usando dados bioclimáticos também foi observada por Carnaval e Moritz (2008) para florestas do Domínio Atlântico. Dados bioclimáticos são efetivos ao explicar a distribuição de espécie numa escala de regional a continental ou global, possuindo um

Tabela 1. Resultado das taxas de omissão, a um limiar de 10% de probabilidade, áreas sob a curva ROC, áreas de distribuição potencial e áreas previstas acima de 20% de probabilidade de *Amaioua guianensis* Aubl... em km^2.

MODELO	Taxa Omissão	AUC - teste(treino)	Área prevista (km^2)	Área (>20%)
BIOS	0,204	0,691(0,837)	481008,89	210177,15
CNC	0,179	0,729(0,862)	478171,78	187480,72
JKNF	0,154	0,751(0,888)	478764,17	181973,73
TUDO	0,231	0,738(0,905)	456933,7	164402,17

Observa-se na tabela 1, que a AUC para os casos de treinamento e teste mostraram pouca diferença, sugerindo pouco sobre-ajuste (*overfitting*) nas previsões realizadas pelo algoritmo. Mesmo assim, a maior diferença entre os valores do modelo TUDO sugere que o uso de muitas variáveis prejudica sua performance, ao tentar capturar as variações das variáveis ambientais em relação aos pontos de ocorrência. Na

alto grau de generalização (PEARSON e DAWSON, 2003; GUISAN e ZIMMERMANN, 2000). E apesar de apresentar a pior performance, as classes foram mais homogêneas na figura 2a, afirmando a sua importância como uma primeira aproximação.

As taxas de omissão extrínseca, a 10% de probabilidade acumulativa, foram pequenas (Tabela 1), sugerindo que, somente uma pequena fração das amostras

utilizadas como teste, coincidiu com *pixels* previstos como inadequados para a ocorrência da espécie. Valores semelhantes foram observados por Saatchi et al. (2008).

Observa-se, também, que as opções de modelo com seleção de variáveis (CNC e JKNF – Figura 2b e 2c respectivamente) apresentaram menores taxas de omissão e valores satisfatórios de AUC na análise ROC. A escolha de variáveis por uma avaliação de prévia de todo o conjunto de dados pelo teste Jackknife do Maxent (JKNF) apresentou uma performance mais acurada (AUC = 0,75). Valores apresentados pelo modelo gerado pela seleção de espécies realizadas por uma CCA (CNC) também demonstram um bom desempenho e a menor acurácia se deve ao fato destas variáveis serem selecionadas em uma análise multivariada para um conjunto de espécies e não individualmente para *A. guianensis* Aubl.

Os resultados do teste Jackknife mostraram que em todos os modelos gerados para a espécie, a variável ambiental com ganho mais elevado quando usada isoladamente é a isotermalidade, a qual parece ter a maior informação útil agregada. A isotermalidade é obtida pela razão da variação de temperatura diurna média mensal pela variação de temperatura anual (HIJMANS et al., 2005). Analisando as relações dos modelos com essa variável, observa-se que a espécie se adapta melhor em regiões com menores valores de isotermalidade. A variável ambiental que mais prejudica o ganho, quando é omitida, é a litologia, que parece conter a maioria de informação que não está presente nas outras variáveis. Segundo Pearson e Dawson (2003) o uso de variáveis topográficas, informações sobre cobertura do solo, incluindo índices de vegetação, e tipos de solos, refinam a qualidade da informação obtida pelos modelos e são usados para interpretar-los em escalas regionais a locais, sendo uma importante informação adicional.

Considerações Finais

A seleção de variáveis ambientais mais relevantes para uma determinada espécie, evita problemas com sobre-ajuste do algoritmo e além de diminuírem o custo computacional do processo de modelagem, assim produzindo modelos mais acurados.

Os modelos gerados apenas por variáveis bioclimáticas são muito úteis, como uma primeira aproximação, ao dar uma idéia mais geral da extensão de ocorrência de uma espécie, apesar de sua tendência à superestimativa.

Referências

ANDERSON, R.P. et al. Evaluating predictive models of species' distributions: criteria for selecting optimal models. *Ecol. Model*, n.162, p.211–232, 2003.

CAMPOS, T.E. *Técnicas de seleção de características com aplicações em reconhecimento de faces.* 2001. 160p. Dissertação (Mestrado) - Universidade de São Paulo, São Paulo, SP.

CARNAVAL, A.C.; MORITZ, C. Historical climate modelling predicts patterns of current biodiversity in the Brazilian Atlantic forest. *Journal of Biogeography*, n.35, p.1187–1201, 2008.

DUFRÊNE, M.; LEGENDRE, P. Species assemblages and indicator species: the need for a flexible asymmetrical approah. *Ecological Monograhs*, v.3, n.67, p.345-366, 1997.

DUTRA, G.C. et al. Espécies indicadoras de fitofisionomias no Estado de Minas Gerais. In: CONGRESSO NACIONAL DE BOTÂNICA, 59, 2008, *Resumo*, Natal, 2008. CD-rom

FIELDING, A.H., BELL, J.F. A review of methods for the assessment of prediction errors in conservation presence/absence models. *Env. Conserv*. n.24, p.38–49, 1997

GUISAN, A., ZIMMERMANN, N.E. Predictive habitat distribution models in ecology. *Ecological Modeling*, n. 135, p. 147- 186, 2000.

HIJMANS, R. J. et al. Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, n. 25, p. 1965-1978, 2005.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA – IBGE *Mapas de Biomas do Brasil*. Rio de Janeiro, 2004. Disponível em: <ftp://ftp.ibge.gov.br/Cartas_e_Mapas/Mapas_Murais/> Acesso em: jun 2007.

LORENZI, H. *Árvores brasileiras*: manual de identificação e cultivo de plantas arbóreas nativas do Brasil. Nova Odessa: Editora Plantarum, v.2, 2002.

OLIVEIRA-FILHO, A.T. *Catálogo de árvores nativas de Minas Gerais:* mapeamento e Inventário e dos Reflorestamentos de Minas Gerais. Lavras: Editora UFLA, 2006. 423p.

OLIVEIRA-FILHO, A.T. *TreeAtlan*: flora arbórea da Mata Atlântica e domínios adjacentes: Um banco de dados envolvendo geografia, diversidade e conservação. Disponível em: <http://www.treeatlan.dcf.ufla.br>. Acesso em: maio 2008.

PEARSON, R. G.; DAWSON, T. P. Predicting the impacts of climate change on the distribution of species: are bioclimate envelope models useful? *Global Ecology & Biogeography*, n.12, p.361–371, 2003.

PETERSON et al. Transferability and model evaluation in ecological niche modeling: a comparison of GARP and Maxent. *Ecography*, n.30, p. 550-560, 2007.

PHILIPS, S.J.et al. Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, n.190, p.31-259, 2006

RATHER, A et al. Analysis of the floristic composition of the brazilian Cerrado vegetation: comparison of the woody vegetation of 376 areas. *Journal of Botany*, v.1, n.60, p57–109, 2003.

SAATCHI, S. Modeling distribution of Amazonian tree species and diversity using remote sensing measurements. *Remote Sensing of Environments*, n. 112, p.2000-2017, 2008.

SIQUEIRA, M.F.; PETERSON, A.T. Consequences of global climate change for geographic distributions of Cerrado tree species. *Biota Neotropica*, v. 3, n. 2, p.1-14, 2003.

SOBERÓN, J.; PETERSON, A.T. Interpretation of Models of Fundamental Ecological Niches and Species' Distributional Areas. *Biodiversity Informatics*, v.2, p.1-10, 2005.