Avaliação do desempenho de modelos de distribuição potencial da espécie *Wunderlichia azulenzis*

Alline Zagnoli Villela Motta [1]
Sollano Rabelo Braga[1]
Nathalia Drummond Marques da Silva[1]
Cristiano Christofaro[1]

[1] Universidade Federal dos Vales do Jequitinhonha e Mucuri- UFVJM
Caixa Postal 34 - 39100-000 - Diamantina - MG, Brasil
{allinezvm, sollanorb, christofaro}@gmail.com; nathalia.florestal@yahoo.com.br

Abstract. Potential distribution models, when allowing the occurrence mapping of species, can be a powerful tool for conservation of natural resources programs. The objective of this study is to evaluate the performance of many modeling algorithms utilizing distribution data of the *Wunderlichia azulenzis* species. The species is listed in the Ministry of Environment's National list of endangered species of flora in the Caatinga biome. Two groups of algorithms, classified according to two types of entry data (presence and absence), were evaluated using the Area Under the Curve - AUC. From the registered occurrences for the species on database Global Biodiversity Information Facility – GBIF, and utilizing six temperature and precipitation variables selected from the Worldclim project, species distribution maps were created. Six different algorithms were used to create the distribution maps of the species. The Mahalanobis Distance (0,978) and the Random Forest (0,0993) algorithms presented the greatest AUC values among its respective groups, while the Bioclim (0,931) and General Linear Model - GLM (0,807) algorithms presented the lowest values. The algorithms that are a part of the group of models that use only presence registers (Bioclim, Domain and Mahalanobis Distance) were considered efficient.

Palavras-chave: specie distribution modelling, potential distribution, modelagem de distribuição de espécies, distribuição potencial, caatinga

1. Introdução

A Caatinga é o único bioma exclusivamente brasileiro, apresentando elevada biodiversidade, endemismo e heterogeneidade. Com uma área aproximada de 850 mil km^2, ocupa cerca de 11% do território nacional, englobando de forma contínua parte dos estados do Maranhão, Piauí, Ceará, Rio Grande do Norte, Paraíba, Pernambuco, Alagoas, Sergipe, Bahia (região Nordeste do Brasil) e parte do Norte de Minas Gerais (Região Sudeste do Brasil) (MMA, 2016). Contudo, essa enorme biodiversidade vem sendo ameaçada por alterações ambientais causadas pelo homem, como a perda e fragmentação de habitats e as mudanças climáticas, todas com consequências diretas sobre a distribuição das espécies (Mittermeier *et al.*, 2005). Essas ameaças crescentes demandam abordagens que permitam adquirir ou aprofundar o conhecimento existente sobre as espécies e auxiliar em sua proteção e conservação.

Os modelos de distribuição de espécies (*Species Distribution Models* - SDMs), também chamados de modelos de nicho ecológico ou modelos de envelope bioclimático, são ferramentas úteis para complementar a informação sobre a distribuição geográfica das espécies ao longo do tempo (Elith *et al*., 2006) assim como para o gerenciamento de recursos naturais. Atualmente, os SDMs constituem um dos campos de pesquisa mais ativos na ecologia, sendo aplicados em diversos estudos, como analisar a dinâmica de distribuição das espécies em cenários de mudanças climáticas passadas, análise dos padrões de riqueza de espécies, dentre outros (Lima-Ribeiro *et al.*, 2012).

A modelagem preditiva de distribuição de espécies consiste em um processamento computacional que combina dados de ocorrência de uma ou mais espécies com variáveis ambientais, permitindo levantar as condições ambientais requeridas pelas espécies (Anderson *et al*., 2003). Nesse contexto, os SDM podem ser utilizados como uma importante abordagem conservacionista, permitindo caracterizar a distribuição potencial de espécies da flora

ameaçadas de extinção nos biomas brasileiros, avaliar a efetividade das unidades de conservação já existentes em relação a esse grupo, além de contribuírem para o estabelecimento de novas áreas prioritárias para a conservação das espécies ameaçadas.

O objetivo desse trabalho foi realizar a modelagem preditiva da espécie *Wunderlichia azulensis*, criticamente ameaçada de extinção e típica do bioma caatinga, a fim de avaliar a influência das variáveis ambientais e o desempenho dos algoritmos de modelagem.

2. Materiais e Métodos

2.1 Área de Estudo.

O nome "Caatinga" é de origem Tupi-Guarani e significa "floresta branca". Composta de árvores e arbustos baixos com algumas características xerofíticas, a Caatinga é caracterizada como floresta arbórea ou arbustiva. Comparada a outras formações brasileiras, apresenta a mais alta radiação solar, baixa nebulosidades, a mais alta temperatura média anual, as mais baixas taxas de umidade relativa, evapotranspiração mais elevada e precipitações mais baixas e irregulares. Possui temperaturas médias anuais muito elevadas com valores entre 26 a 28ºC (Prado, 2003).

Apesar da sua importância, o bioma tem sido desmatado de forma acelerada, principalmente nos últimos anos, devido principalmente ao consumo de lenha nativa, explorada de forma ilegal e insustentável, para fins domésticos e indústrias, ao sobrepastoreio e a conversão para pastagens e agricultura.

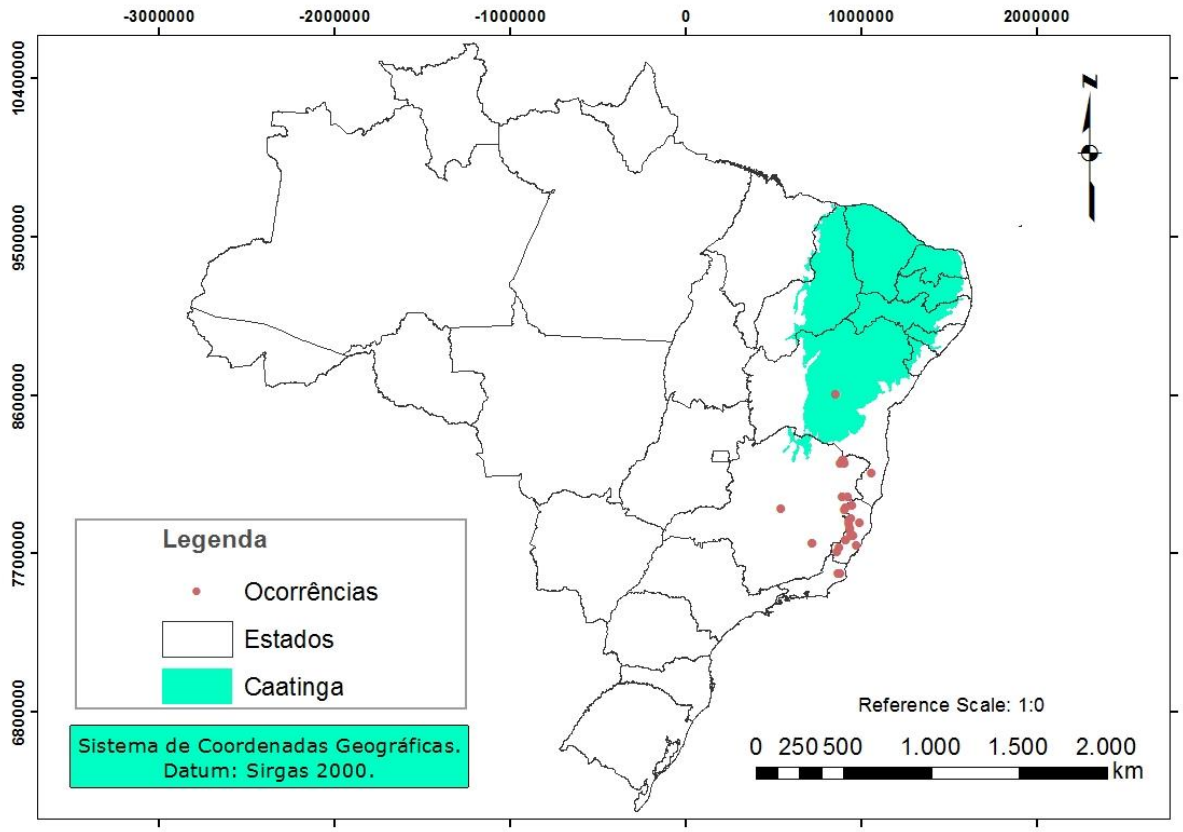

Figura 1. Mapa de localização do bioma Caatinga no Brasil e ocorrências da espécie *Wunderlichia azulenzis*.

2.2 Seleção da espécie

Optou-se por realizar a modelagem preditiva com espécies criticamente ameaçadas de extinção no bioma Caatinga segundo a Lista Nacional de Espécies Ameaçadas de Extinção (MMA, 2008).

A espécie *Wundelichia azulenzis* é uma espécie arbustiva deciduifólia, típica de ambientes rupestres, atingindo 2 a 10 metros, pertencente à família *Asteraceae*. A espécie consta na lista de espécies ameaçadas de extinção no bioma Caatinga (MMA, 2008), apresentando, dentre as espécies ameaçadas da Caatinga, o maior número de registros no banco de dados *Global Biodiversity Information Facility* - GBIF.

A espécie foi coletada pela primeira vez por J. G. Kuhlmann (número coleção 6616, 1943/05/12, RB), e descrita por Barroso e Maguire (1973), com poucos registros seguintes. Apresenta registros em afloramentos rochosos na Bahia, Minas Gerais e Espirito Santo (Souza-Buturi 2013b), com padrão de distribuição associado a locais com precipitação anual inferior a 1.200 mm.

Apesar de ter sido registrado em três estados brasileiros, sua ocorrência é limitada a afloramentos rochosos isolados com condições climáticas igualmente limitadas o que explica em parte o risco atual de extinção.

2.3 Variáveis ambientais

Primeiramente, foi feita uma busca na literatura a fim de identificar possíveis variáveis que melhor explicassem a distribuição da espécie selecionada. A seguir, as variáveis climáticas e bioclimáticas foram selecionadas do projeto *Worldclim* (Hijmans *et al.*, 2005) na resolução de 30 Arc segundos (aproximadamente 1km).

2.4 Modelagem a Avaliação

As ocorrências georreferenciadas da espécie obtidas no banco de dados GBIF foram associadas aos dados ambientais.

Existem diversos algoritmos utilizados na realização da modelagem de distribuição, que podem ser separados em dois grandes grupos de acordo com os tipos de dados de entrada. O primeiro grupo consiste em modelos que utilizam apenas registros de presença e o segundo grupo em modelos que utilizam dados de presença e ausência. No primeiro grupo, podemos enquadrar os *Profile methods* (métodos de perfil) *Bioclim* (envelopes bioclimáticos), *Domain* e Distância de Mahalanobis. No segundo grupo estão os *Machine learning methods* (métodos de aprendizagem-automática) *Random forest* e *Support vector machines* - *SVM*, bem como o método de *Logistic Regression* (regressão logística) *Generalized Linear Model* - *GLM* (Modelo Linear Generalizado).

O primeiro conjunto de algoritmos testados foram os métodos de perfil. Neste conjunto foram analisados os métodos *Bioclim*, *Domain* e Distância de Mahalanobis. O algoritmo *Bioclim*, amplamente utilizado para a modelagem de distribuição de espécies, consiste em um "envelope bioclimático" clássico. Este algoritmo calcula a semelhança de uma localização comparando os valores das variáveis ambientais em qualquer local com uma distribuição percentual dos valores em locais conhecidos de ocorrência.

O algoritmo Distância de Mahalanobis é baseado nas correlações entre variáveis com as quais distintos padrões podem ser identificados e analisados. Ele normaliza os valores das variáveis ambientais e calcula a distância entre as condições ambientais para cada ponto de ocorrência selecionando a menor distância (distância mínima).

O algoritmo *Domain* utiliza da similaridade métrica na qual a predição de adequabilidade é calculada pela distância mínima do espaço ambiental para cada ponto de presença e, assim, consegue gerar predições mais amplas que se aproximam do nicho fundamental das espécies.

O segundo conjunto de algoritmos testados incluiu os métodos de *Machine learning* bem como o método de *Logistic Regression.* Neste conjunto estão o *Random Forest*, o *Support*

Vector Machines e o GLM. O algoritmo *Random Forest* é um tipo de método de aprendizagem de conjunto que consiste de uma coleção de classificadores estruturados em árvores de decisão, que por sua vez são utilizadas na classificação de novos objetos. *Random Forest* (Breiman, 2001b) é uma extensão da Classificação (classificar se a espécie está presente ou não) e *Regression trees*.

O algoritmo *Support Vector Machines* caracteriza-se por ser um conjunto de métodos de aprendizagem supervisionado pertencente à família dos classificadores lineares generalizados. A teoria do SVM preconiza a minimização do *risco estrutural*, ou seja, a probabilidade de classificar errado padrões ainda não vistos pela distribuição de probabilidade dos dados. Os modelos gerados pela SVM dependem apenas de um subconjunto de dados de treino e utilizam só os dados mais informativos para gerar os SDM's. Isto faz com que essa técnica seja interessante para utilizar em situações onde os dados de entrada (registros de ocorrência da espécie e/ou variáveis ambientais) são duvidosos ou incompletos (Junior e Siqueira, 2009).

O GLM investiga a relação entre uma variável resposta e uma ou mais variáveis preditoras. É uma técnica complexa que precisa de dados de ausência e presença e um número maior de dados, além de ser considerada mais adequada para modelar a distribuição real (Kamino, 2009).

Os modelos foram avaliados a partir da análise da curva *Receiver Operating Characteristic – ROC* (Características Operacionais do Receptor), considerando a *Area Under the Curve – AUC* (Área Abaixo da Curva). A curva ROC é a relação entre a sensitividade (proporção de presenças corretamente preditas) e o erro de comissão (ou taxa de falsos positivos) (Franklin *et al*., 2009). Valores de AUC de 0,5 indicam um modelo aleatório, onde a proporção de locais preditos corretamente é igual à proporção de lugares preditos incorretamente. Valores de AUC mais próximos de 1 informam uma prevalência da sensitividade sobre os erros de comissão. Valores entre 0,8 e 0,9 indicam bons modelos e valores acima de 0,9 indicam modelos ótimos (Thuiler *et al*., 2005). Todas as análises foram realizadas no programa R (R *Development Core Team*, 2015).

3. Resultados e Discussão

Foram levantadas 33 ocorrências para a espécie estudada. As variáveis selecionadas no *WorldClim* foram: Temperatura média anual (bio1), Precipitação anual (bio12), Precipitação da estação úmida (bio16), Precipitação da estação seca (bio17), Temperatura máxima do mês mais quente (bio5) e Oscilação térmica anual (bio7).

3.1 Avaliação dos Métodos de Perfil

De acordo com a Tabela 1, o método da Distância de Mahalanobis foi o mais eficiente quando comparado aos outros dois métodos com o mais elevado AUC para os dados de treino. O segundo método mais eficiente foi o *Domain* e o menos eficiente foi o *Bioclim* gerando o modelo com o valor de AUC mais baixo.

Tabela 1. Valores de AUC dos modelos gerados pelos três métodos de perfil testados para a espécie *Wunderlichia azulensis.*

Métodos de Perfil	AUC	Classificação do modelo*
Mahalanobis	0,978	Ótimo
Domain	0,961	Ótimo
Bioclim	0,931	Ótimo

*Classificação segundo Thuiler *et al*., 2005

O mapa da **Figura 2a** foi gerado pelo método da Distância de Mahalanobis e mostra os locais onde existe adequabilidade ambiental para a espécie. A distribuição obtida neste

modelo está de acordo com a literatura (Souza-Buturi, 2013). Desta forma, o modelo foi considerado altamente eficiente, contudo restrito aos dados de origem. Esta restrição é mais forte no modelo gerado pelo algoritmo *Bioclim* (**Figura 2c**), onde a predição aponta em maior parte as ocorrências na região Sudeste, incluindo alguns trechos do sul da região sudeste. Embora geralmente o *Bioclim* não seja tão bom como alguns outros métodos de modelagem ainda é usado, entre outras razões porque o algoritmo é fácil de compreender e, portanto, útil no ensino de modelagem de distribuição de espécies (Hijmans e Elith, 2015).

É possível perceber, por meio da **Figura 2b**, que a distribuição da espécie gerada pelo método *Domain* aponta ocorrências nas regiões Sudeste, Centro-Oeste e Nordeste, abrangendo ainda países da América do Sul e Central. Levando-se em consideração que esta distribuição gerada pelo modelo não corresponde as características associadas à real distribuição da espécie, mesmo com alto valor de AUC, o modelo não foi considerado eficiente.

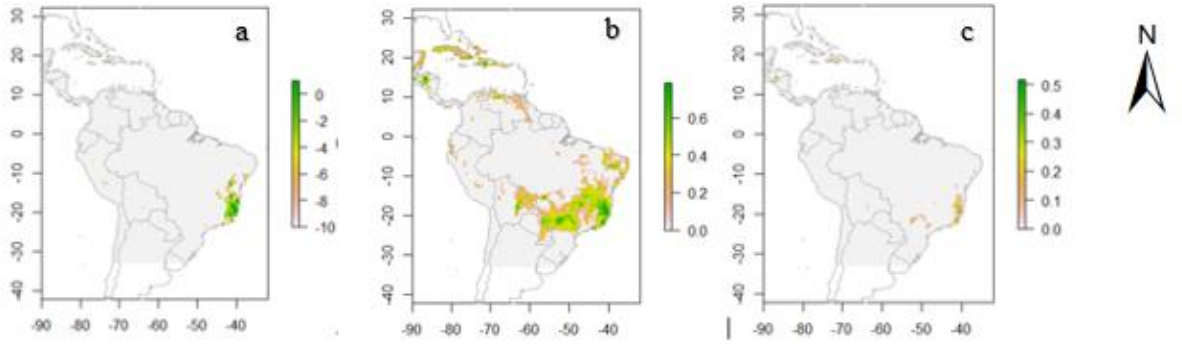

Figura 2. Predição potencial para a ocorrência de *Wunderlichia azulensis*, utilizando os métodos de perfil. (a) Distância de Mahalanobis. (b) Domain. (c) Bioclim.

3.2 Métodos de aprendizagem automática

De acordo com a Tabela 2, o método *Random Forest* foi mais eficiente que o método *Support Vector Machine* uma vez que apresentou maior AUC.

Tabela 2. Valores de AUC dos modelos gerados pelos dois métodos de aprendizagem-automática testados.

Métodos de aprendizagem automática	AUC	Classificação do modelo*
Random Forest	0,993	Ótimo
Support Vector Machine	0,959	Ótimo

*Classificação segundo Thuiler *et al.*, 2005

Observa-se pela **Figura 3a** que a distribuição da espécie gerada pelo método *Random Forest* é bem mais restritiva que a gerada pelo método SVM (**Figura 3b),** concentrando a maior parte da distribuição da espécie na região Sudeste. Desta forma, o modelo apresenta certo grau de confiabilidade, pois as ocorrências registradas encontram-se na mesma região. O modelo gerado pelo método SVM apresentou ampla distribuição com maior concentração nas regiões Nordeste, Sudeste e Centro-Oeste. O modelo não foi considerado tão eficiente, apesar de apresentar alto valor de AUC, uma vez que a área de predição está em desacordo com as características ambientais das ocorrências registradas da espécie na literatura.

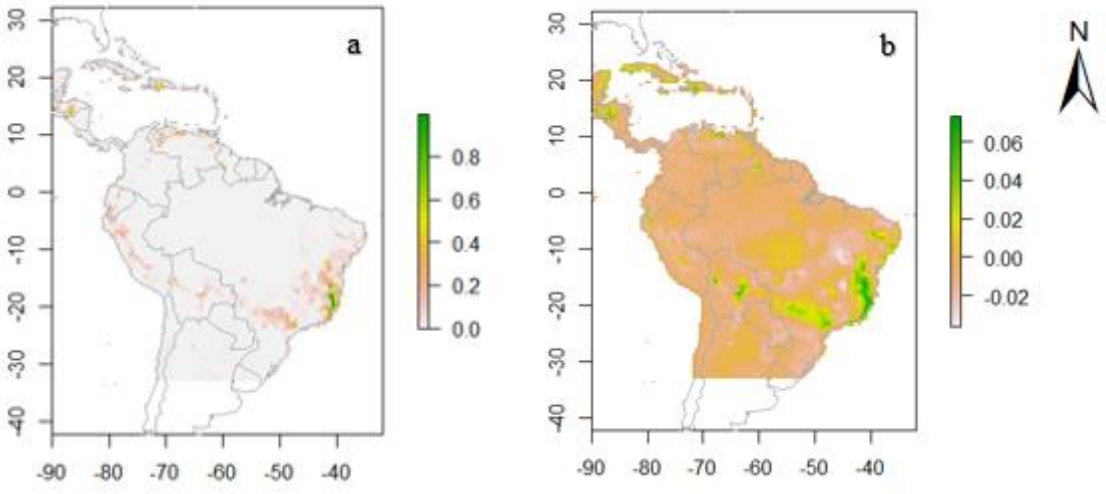

Figura 3. Predição potencial para ocorrência de *Wunderlichia azulensis*, utilizando os métodos de aprendizagem automática. (a) Random forest. (b) Support vector machines.

3.3 Método de Regressão Logística

O Modelo Linear Generalizado apresentou valor de AUC de 0,807, sendo considerado um bom modelo. É possível observar por meio da **Figura 4** que o Modelo Linear Generalizado resultou em uma ampla distribuição da espécie *W. azulensis,* com ocorrências da região Sudeste até a região Norte. O modelo não foi considerado eficiente uma vez que a predição estava abrangendo áreas que não representam as características ambientais de ocorrência da espécie.

Isto pode ser explicado uma vez que o GLM é uma técnica complexa que precisa de ausência (no presente estudo foram simulados os dados de ausência) e presença e um número maior de dados (Kamino, 2009).

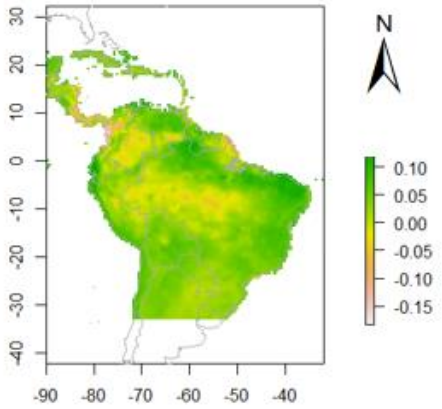

Figura 4. Predição potencial para a ocorrência *Wunderlichia azulensis*, utilizando o modelo linear generalizado.

4. Conclusão

Dentre os métodos de perfil testados, o método da Distância de Mahalanobis gerou o modelo mais eficiente uma vez que a distribuição gerada para a espécie coincidiu com a distribuição atualmente disponível na literatura. O modelo gerado pelo método *Bioclim*, apesar de apresentar menor AUC, pode ser considerado mais eficiente que o modelo gerado pelo método *Domain* pois sua área de predição foi mais restrita aos ambientes onde ocorreram os registros da espécie.

Entre os métodos de aprendizagem automática o *Random Forest* foi mais eficiente com AUC mais elevado e coincidência com a distribuição disponível na literatura. O modelo linear generalizado, apresentou um AUC acima de 0,8 indicando ser um bom modelo, porém, sua predição não representa a distribuição das ocorrências da espécie em questão.

Os algoritmos que fazem parte do grupo de modelos que utilizam apenas registros de presença (*Bioclim*, *Domain* e Distância de Mahalanobis) foram considerados eficientes.

Agradecimentos

Os autores agradecem o apoio financeiro promovido pela Universidade Federal dos Vales do Jequitinhonha e Mucuri, CNPQ e FAPEMIG.

Referências Bibliográficas

Anderson, R.P.; Lew, D. & Peterson, A.T. Evaluating predictive models of species distributions: criteria for selecting optimal models. **Ecological Modelling,** vol. 162, p. 211-232. 2003.

Breiman, L., 2001b. Random Forests. Machine Learning vol. 45, p. 5-32.

Elith J, Graham C. H, Anderson R. P, Dudík M, Ferrier S, et al. Novel methods improve prediction of species' distributions from occurrence data. **Ecography**, vol. 29, p. 129–51. 2006.

Franklin, J., Wejnert, K. E., Hathaway, S. A., Rochester, C.J. & Fisher, R. N. Effect of species rarity on the accuracy of species distribution models for reptiles and amphibians in southern California. **Diversity and Distributions,** vol. 15, p. 167-177. 2009.

Hijmans, J. R.; Cameron, S. E.; Parra, J. L.; Jones, P. G. & Jarvis, A. Very high resolution interpolated climate surfaces for global land areas. **International Journal of Climatology**, vol. 25, p. 1965-1978. 2005.

Junior P. de M., Siqueira de M. F. Como determinar a distribuição potencial de espécies sob um abordagem conservacionista? **Megadiversidade**, vol. 5, p. 1-2. 2009.

Kamino, L. H. Y. Modelos de distribuição geográfica potencial: aplicação com plantas ameaçadas de extinção da floresta atlântica. Belo Horizonte. 2009.

Lima-Ribeiro, Matheus de Souza; Diniz-Filho, José F. A. Modelando a distribuição geográfica das espécies no passado: uma abordagem promissora em Paleoecologia. **Revista Brasileira de Paleontologia**, vol. 15, p. 371-385. 2012.

Mauad L. P., Buturi F. O. de S., Souza T. P. Nascimento M. T., Brga J. M. A. New distribution record and implications for conservation of the endangered Wunderlichia azulensis Maguire & G.M. Barroso (Asteraceae: Wunderlichieae). Check List. p. 706-708, 2014.

Mittermeier, C. G., Gil, P. R., Hoffmann, M., Pilgrim, J., Brooks, T., Lamourex, J. & Fonseca, G. A. B. **Hotspots Revisitados. As regiões biologicamente mais ricas e ameaçadas do planeta.** Belo Horizonte, Conservação Internacional do Brasil. 2005.

MMA. 2008. Instrução Normativa n o. 6, de 23 de setembro de 2008. **Lista Nacional de Espécies da Flora Ameaçadas de Extinção.** Imprensa Oficial. Brasília.

MMA. Disponível em: <http://www.mma.gov.br/biomas/caatinga>. Acesso em: 03.out.2016.

Prado, D.E. As Caatingas da América do Sul. In: Leal, R.I.; Tabarelli, M.; Silva, J.M.C. da. **Ecologia e conservação da Caatinga**. Recife: Ed. Universitária da UFPE, 2003. 823p.

Robert J. Hijmans and Jane Elith. **Species distribution modeling with R.** 2015.

Souza-buturi, F.O. *Wunderlichia* em **Lista de Espécies da Flora do Brasil**. Jardim Botânico do Rio de Janeiro. Disponível em: <http://floradobrasil.jbrj.gov.br/jabot/floradobrasil/FB5542>. Acesso em: 03.out.2016.

Thuiller W, Richardson D. M, Pyšek P, Midgley G. F, Hughes G. O, Rouget M. Niche-based modelling as a tool for predicting the risk of alien plant invasions at a global scale. **Glob Change Biol**, vol. 11, p. 2234-2250. 2005.