Nativa, Sinop, v. 7, n. 5, p. 600-606, set./out. 2019.
Pesquisas Agrárias e Ambientais
DOI: http://dx.doi.org/10.31413/nativa.v7i5.7214
http://www.ufmt.br/nativa

ISSN: 2318-7670

Algoritmos de aprendizagem de máquina na modelagem da distribuição potencial de habitats de espécies arbóreas

Mônica Canaan CARVALHO[1], Luciano Cavalcante de Jesus FRANÇA[2]*, Isáira Leite e LOPES[2], Laís Almeida ARAÚJO[2], José Marcio de MELLO[2], Lucas Rezende GOMIDE[2]

[1]Instituto Federal do Sudeste de Minas Gerais, Barbacena, MG, Brasil.
[2]Programa de Pós-Graduação em Engenharia Florestal, Universidade Federal de Lavras, Lavras, MG, Brasil.
*E-mail: lucianodejesus@florestal.eng.br

Recebido em setembro/2018; Aceito em janeiro/2019.

RESUMO: O estudo teve como objetivo avaliar três métodos de aprendizagem de máquina (árvore de decisão-J48, *random forest* e redes neurais artificias), na modelagem da distribuição de dez espécies arbóreas mais abundantes em uma sub-bacia do rio São Francisco (MG). Utilizaram-se dados provenientes do Inventário Florestal de Minas, com total de 77 fragmentos amostrados e 2.234 parcelas, nas quais foram computadas a presença/ausência de cada espécie. Empregaram-se 12 variáveis ambientais categóricas procedentes do Zoneamento Ecológico Econômico de Minas Gerais (ZEE/MG), além de variáveis relacionadas ao balanço hídrico do solo (evapotranspiração atual e potencial, aridez e índice *alpha*). A parametrização dos três algoritmos para as dez espécies selecionadas foi feita com o auxílio do algoritmo *cv parameter* do *software* WEKA. Os resultados mostram que os algoritmos testados apresentaram desempenhos estatisticamente iguais em 60% das espécies arbóreas. Os algoritmos *random forest* e *multilayer perceptron* foram estatisticamente iguais para a espécie *Eugenia dysenterica*, sendo superiores ao algoritmo J48. Contudo, o algoritmo *random forest* foi superior aos demais para as três espécies do gênero *Qualea*. Conclui-se que o algoritmo *random forest* apresentou-se como o mais robusto para a modelagem da distribuição potencial de habitat de espécies arbóreas.
Palavras-chave: inteligência artificial; árvore de decisão; *random forest*; redes neurais artificiais.

Machine learning algorithms for modeling the potential distribution habitat of tree species

ABSTRACT: The aim of the present study was to evaluate three methods of machine learning (decision tree-J48, random forest and artificial neural networks) to model the potential habitat distribution of the ten most abundant tree species of the São Francisco river watershed. The presence/absence tree species data were from 77 fragments sampled with 2,234 plots. We used 12 categorical environmental variables from the Economic Ecological Zoning of Minas Gerais (ZEE/MG), as well as variables related to soil water balance (current and potential evapotranspiration, aridity and alpha index). The parameterization of the three algorithms was done with cv parameter algorithm of the WEKA software. The results showed the applied algorithms were statistically similar for 60% of the tree species. The random forest and multilayer perceptron algorithms were statistically similar considering the *Eugenia dysenterica* and superior to J48 algorithm. However, the random forest algorithm was superior to the other for the three species of *Qualea genera*. The conclusion is the random forest was the most robust model for the potential distribution habitat of tree species.
Keywords: artificial intelligence; decision trees; random forest; artificial neural networks.

1. INTRODUÇÃO

As alterações no ambiente resultantes da ação do homem têm colocado em risco a distribuição de espécies arbóreas no planeta. A fragmentação de habitat, mudanças no uso da terra e as mudanças climáticas ameaçam a existência e perpetuação delas. Por outro lado, a necessidade crescente de proteção e restauração dos ecossistemas florestais demandam novas tecnologias capazes de entender as relações entre as características do meio ambiente e a ocorrência de espécies (GIANNINI et al., 2012; EHRLÉN; MORRIS, 2015; MORENO-FERNÁNDEZ et al., 2016).

Atualmente, estudos vêm sendo desenvolvidos para prever áreas de desmatamento (SOUZA; MARCO JR, 2014), ambientes favoráveis à invasão de plantas exóticas (CANESSA et al., 2018) e impactos das mudanças climáticas sobre a distribuição de espécies ameaçadas de extinção (QIN et al., 2017) por meio da Modelagem da Distribuição de Espécies (MDE). Esta metodologia é relevante para conservação, ecologia e manejo florestal (HENDERSON et al., 2014; MATEO et al., 2018), visto que direciona tomadas de decisão e implementação de medidas de gestão, de modo a auxiliar na seleção de áreas para conservação ou proteção de espécies, assegurando que estas não sejam enquadradas em categoria de extinção (COSTA et al., 2018).

O ponto inicial da MDE é o uso de coordenadas geográficas precisas dos dados de ocorrência/ausência das espécies, em conjunto com o uso de variáveis climáticas e ambientais, como precipitação, temperatura, relevo, dentre outras. Após o uso de diversos métodos é possível realizar predições espaciais do habitat mais adequado para uma determinada espécie em análise (CHAKRABORTY et al., 2016).

Devido à complexidade do processo de modelagem, no que tange previsões confiáveis, diferentes abordagens de

algoritmos e métodos têm sido aplicadas, como exemplo, os métodos estatísticos (modelos lineares generalizados – GLM e modelos aditivos generalizados – GAM) e de aprendizagem de máquina (redes neurais artificiais – RNA, *suport vector machine* – SVM, árvores de decisão – CART, *random forest* – RF e entropia máxima – MAXENT), conforme observado nos trabalhos de Paglia et al. (2012); Merow et al. (2014); García-Callejas; Araújo (2016). Pesquisas comprovam que o desempenho dos algoritmos varia de acordo com os dados referentes às espécies e sua distribuição espacial (ROBERTSON et al., 2003; CARVALHO et al., 2017). Não há um consenso de qual o melhor método, já que existem variações dessa natureza, o que decorre em uma lacuna sobre qual a melhor técnica de modelagem e qual algoritmo possui desempenho superior.

Neste sentido, este estudo tem como objetivo avaliar três métodos de aprendizagem de máquina (Árvore de Decisão, *Random Forest* e Redes Neurais Artificiais) na modelagem da distribuição de dez espécies arbóreas mais abundantes em uma sub-bacia hidrográfica do rio São Francisco. Concomitante a este objetivo, pretende-se entender quais são os fatores ambientais que estão mais correlacionados com a distribuição de cada espécie.

2. MATERIAL E MÉTODOS

2.1. Área de estudo e dados de ocorrência

A área amostrada compreende a bacia hidrográfica do rio das Velhas, sub-bacia do rio São Francisco, localizada no estado de Minas Gerais, Brasil. Abrange as regiões do Alto e Médio São Francisco, com área de drenagem de aproximadamente 14.155 km² (Figura 1).

O parâmetro de seleção das espécies arbóreas a serem modeladas baseou-se no critério das 10 espécies com maior abundância total dentro da bacia. Os dados empregados foram provenientes do Inventário Florestal de Minas Gerais (SCOLFORO et al., 2006), realizado entre os anos de 2006 - 2008. Nesse sentido, um conjunto de 77 remanescentes florestais foram selecionados, com área variável entre 1,39 a 85.431 hectares, totalizando 2.234 parcelas. A suficiência amostral foi determinada pelo método da Regressão Linear de Platô para cada fitofisionomia presente no estado de Minas Gerais. Para realizar o ajuste da regressão e calcular a suficiência amostral, as parcelas dentro de cada fragmento foram sorteadas aleatoriamente 30 vezes. Em cada sorteio, calculava-se a frequência acumulada (FA) dessa combinação. Ao final dos sorteios, extraía-se a média de FA e calculava-se ainda a área acumulada referente às parcelas do levantamento florestal. A partir desse ponto, aplicou-se a regressão linear platô, obtendo-se seus parâmetros e o ponto de encontro entre as duas regressões. Para todas as fitofisionomias o coeficiente de determinação foi superior a 70% (SCOLFORO et al., 2006)

Devido à resolução espacial dos dados empregados para as variáveis ambientais, optou-se por trabalhar os dados de ocorrência em nível de fragmento ao invés de parcela, visto que a suficiência amostral foi atingida. Assim, foram extraídos os valores das variáveis ambientais no ponto centroide de cada fragmento, bem como atribuído o valor de presença (1) ou ausência (0). Desta forma, o conjunto de treinamento dos algoritmos foi constituído por 77 observações por espécie, nas quais estão disponíveis os valores das variáveis independentes e a ocorrência da espécie, por meio de valores binários (0 – ausência / 1 – presença). Trabalhou-se com classes de presença/ausência desbalanceadas, sendo o número de presença variável entre 30 e 47 de acordo com cada espécie.

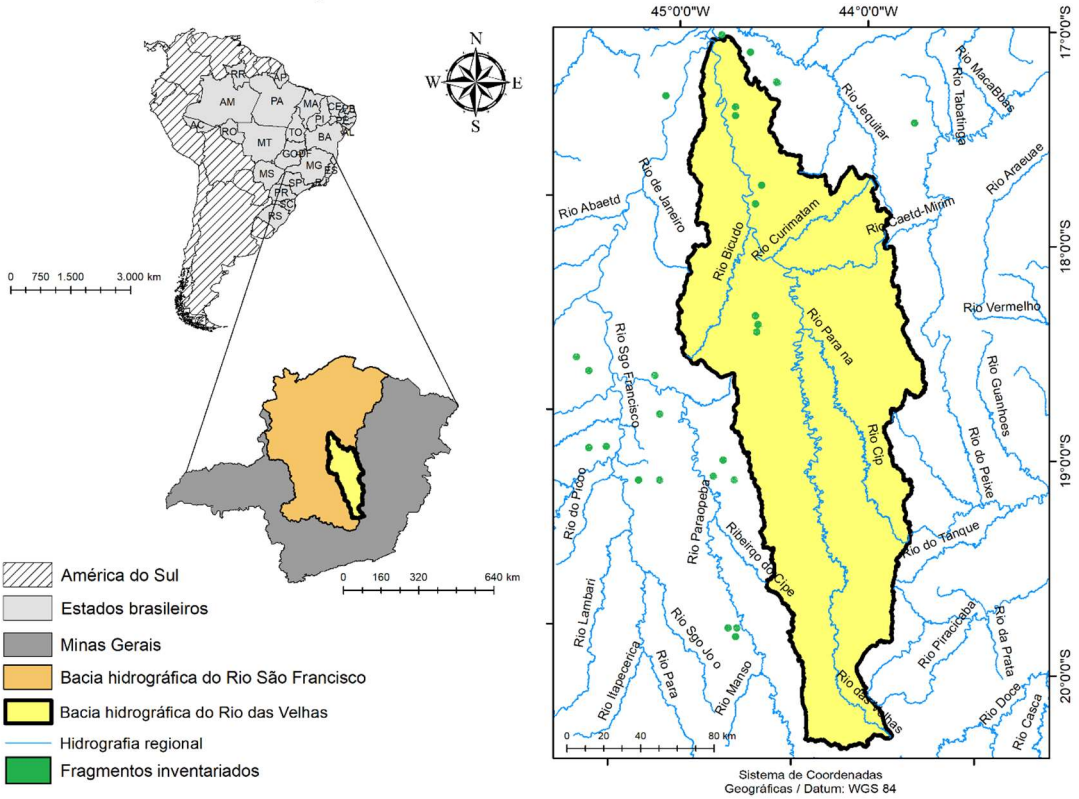

Figura 1. Mapa de localização da área de estudo (Bacia do rio das Velhas), no contexto da América do Sul, Brasil, Minas Gerais e na bacia hidrográfica do Rio São Francisco.

Figure 1. Location map of the study area (Velhas river Basin), in the context of South America, Brazil, Minas Gerais and the São Francisco River Basin.

2.2. Variáveis Ambientais

No total, um conjunto inicial de 39 variáveis independentes foi compilado. Destas, 20 variáveis ambientais foram selecionadas provenientes da base de dados do *World Clim* (HIJMANS et al., 2005), com resolução inicial de 1 km, empregaram-se ainda 12 variáveis ambientais categóricas procedentes do Zoneamento Ecológico Econômico de Minas Gerais (ZEE/MG), com resolução espacial variando entre 30 a 270 metros (SCOLFORO; CARVALHO; OLIVEIRA, 2008). Utilizaram-se também 4 variáveis relacionadas ao balanço hídrico do solo (evapotranspiração atual e potencial, aridez e índice alpha) oriundas da base de dados CGIAR-CSI (TRABUCCO; ZOMER, 2010) com resolução espacial original de 1 km. Outras 2 variáveis foram consideradas devido seu poder de síntese de condições ambientais, sendo elas latitude e longitude. Além destas, o tipo de solo foi utilizado como 1 variável, a partir de dados da classificação do mapa de solos de Minas Gerais (SEMAD, 2010). Todas as variáveis, quando necessário, foram transformadas em formato *raster* com resolução de 270 m e projetadas para o sistema de coordenadas *South America Albers Equal Area Conic*. Essas variáveis assumem valores numéricos ou categóricos.

De posse do conjunto de dados de treinamento contendo as 39 variáveis ambientais, aplicou-se então o algoritmo de seleção de atributos *Correlation-based feature selection* (CFS). A seleção foi realizada utilizando o método de validação cruzada com a formação de 10 subconjuntos. Foi adotado o critério de seleção como sendo as 4 variáveis ambientais mais escolhidas pelo algoritmo nos dez subconjuntos

2.3. Processo de modelagem

As etapas do processo de modelagem variam de acordo com o objetivo do trabalho, algoritmos, base de dados e *software* utilizado. Assim, foi desenvolvido um fluxograma (Figura 2) adaptado de Garzón et al. (2006), para representar e ordenar as etapas da modelagem utilizada nesta pesquisa.

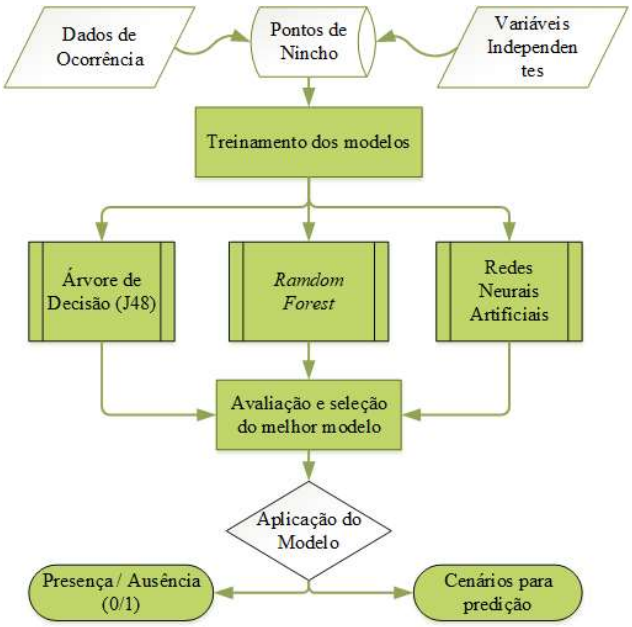

Figura 1. Fluxograma das etapas metodológicas para modelagem de distribuição de espécies florestais.
Figure 1. Flowchart of the methodological steps for modeling the distribution of forest species.

O *software* utilizado neste artigo para o treinamento e aplicação dos algoritmos foi o WEKA –*Waika to Environment for Knowledge Analysis* (GARNER, 1995). Neste estudo optou-se por testar Árvore de Decisão (QUINLAN, 1993), devido à sua simplicidade e legibilidade; *random forest* (BREIMAN, 2001), pelos bons resultados apresentados em outras pesquisas (INZA et al., 2009; BHERING et al., 2016); e Redes Neurais artificiais pela robustez e habilidade em lidar com dados muito complexos.

A parametrização dos três algoritmos para as dez espécies selecionadas foi feita com o auxílio do algoritmo *cvparameter* implementado no *software* WEKA. É um meta-classificador que testa vários valores (pré-definidos) para diferentes parâmetros de cada algoritmo. Para o algoritmo *J48* (Árvore de Decisão) foram avaliados os parâmetros *seed* (1 a 10), *numFolds* (1 a 10) e *confidence Factor* (0,1 a 0,9). No algoritmo *Random Forest* foram testados os parâmetros *num Trees* (1 a 15) e *seed* (1 a 10). Para o algoritmo *Multilayer Perceptron (*redes neurais artificiais*)* testou-se os parâmetros *hidden Layers* (0 a 2), *earning rate* (0,1 a 0,9) e *momentum* (0,1 a 0,9). Para cada espécie foi estabelecido o número de 10 repetições por algoritmo, sendo que a avaliação foi feita por validação cruzada, com 10 repetições. As configurações que apresentaram maior AUC (*area under the curve*) em cada algoritmo foram selecionadas para nova aplicação na base de dados de treinamento.

Após a parametrização e escolha das configurações otimizadas para cada algoritmo por espécie, os modelos foram aplicados novamente no conjunto de treinamento. Foi realizado um experimento, utilizando 10 iterações e validação cruzada com 10 sub-amostragens, em que comparou-se os valores da métrica área abaixo da curva ROC (AUC - *area unde rthe curve ROC*) obtidos por cada modelo (JIMÉNEZ-VALVERDE, 2011). Utilizou-se o teste estatístico T-pareado a 95% de confiança ou probabilidade, já que se trata da mesma base de dados.

2.4. Seleção dos atributos principais

A fim de diminuir o número de atributos, complexidade dos modelos e determinar quais variáveis ambientais são mais representativas na distribuição de determinada espécie, foi aplicado um algoritmo de seleção (*CfsSubsetEval*) implementado no WEKA. Este algoritmo primeiramente calcula uma matriz de correlação entre as variáveis ambientais e ocorrência, além de uma matriz de correlação entre as variáveis ambientais.

Em seguida calcula o mérito (*score*) para cada subconjunto formado utilizando a equação 1. Nesta equação, o numerador pode ser interpretado como o poder preditivo do subconjunto de atributos e o denominador como o grau de redundância existente entre os atributos.

Neste sentido, o *correlation-based feature selection* (CFS) começa com um conjunto vazio de atributos e utiliza a heurística *best-first-search* como algoritmo de busca, na qual o critério de parada é 5 subconjuntos consecutivos que não melhoram o mérito calculado pelo algoritmo.

$$\text{Mérito(S)} = \frac{k \times \overline{r_{ac}}}{\sqrt{k + k(k-1)\overline{r_{aa}}}} \qquad \text{(Eq. 1)}$$

em que: k = número de atributos; $\overline{r_{ac}}$ = média de correlação entre atributo-classe; $\overline{r_{aa}}$ = média de correlação entre atributo-atributo.

3. RESULTADOS

Após a análise dos dados do inventário, chegou-se ao resultado de que as dez espécies mais abundantes na bacia do São Francisco, são: *Anadenanthera colubrina* (Vell.) Brenan, *Eugenia dysenterica* DC., *Hymenaea stignocarpa* Mart. Ex Hayne, *Lafoensia vandelliana* Cham. & Schltdl., *Magonia pubescens* S. St.-Hil., *Pouteria ramiflora* (Mart.) Radlk., *Qualea grandiflora* Mart., *Qualea multiflora* Mart., *Qualea parviflora* Mart. e *Terminalia fagifolia* Mart. Os resultados da seleção de variáveis por espécie são apresentados na Tabela 1, e os valores da métrica de avaliação AUC para as diferentes técnicas por espécie, são apresentados na Tabela 2.

Tabela 1. Porcentagem de seleção dos atributos utilizando o algoritmo CFS com validação cruzada.
Table 1. Percentage of attribute selection using the CFS algorithm with cross-validation.

N	Variáveis Ambientais	Espécies arbóreas									
		1	2	3	4	5	6	7	8	9	10
1	Altitude	100	100	100	100	100	100	100	100	100	0
2	Aridez	40	0	0	0	0	0	0	0	0	0
3	Declividade	0	0	0	0	0	10	0	0	0	0
4	Clima (Classe Thornthwaite)	0	0	0	0	0	0	0	0	0	0
5	Erodibilidade	0	0	0	0	0	0	0	0	0	0
6	Evapotranspiração atual	20	50	0	0	90	0	0	0	10	0
7	Evapotranspiração potencial	10	0	0	0	0	0	0	0	0	10
8	Fitofisionomias	10	100	20	0	0	0	50	90	40	10
9	Grau de conservação da vegetação	0	0	0	0	0	0	0	10	0	0
10	Grau de erosão	0	0	0	0	0	0	0	0	0	0
11	Grau de exposição do solo	50	0	0	0	0	0	0	0	0	0
12	Índice alpha	0	0	0	0	100	0	0	0	0	0
13	Intensidade da chuva	0	0	10	0	0	0	0	0	0	0
14	Isotermalidade	40	0	0	0	0	10	0	0	0	40
15	Lâmina explotável	0	0	0	0	0	0	0	0	0	0
16	Latitude	0	0	0	0	0	0	0	10	0	100
17	Longitude	80	0	0	0	0	0	0	20	0	30
18	Variação média diurna da temperatura	0	0	0	0	0	0	0	0	0	10
19	Precipitação anual	10	0	0	0	0	10	0	50	0	0
20	Precipitação do mês seco	0	0	0	40	0	10	0	0	0	10
21	Precipitação do mês úmido	20	60	0	0	0	0	20	0	0	0
22	Precipitação do trimestre frio	60	10	20	10	0	0	10	0	0	80
23	Precipitação do trimestre quente	0	10	0	0	30	0	0	0	0	20
24	Precipitação do trimestre seco	50	90	20	70	10	0	60	40	50	0
25	Precipitação do trimestre úmido	60	10	0	0	0	10	0	0	0	0
26	Qualidade da água	0	0	30	0	80	0	0	0	0	0
27	Rendimento específico	0	0	0	0	0	0	0	0	0	0
28	Sazonalidade da precipitação	0	100	100	0	10	0	40	40	30	10
29	Sazonalidade da temperatura	0	0	0	0	0	0	0	0	0	100
30	Taxa de decomposição da matéria orgânica	0	0	0	0	30	0	0	0	0	0
31	Temperatura anual média	0	0	0	0	0	10	0	0	0	10
32	Temperatura máxima do mês mais quente	0	0	30	0	50	0	10	60	10	0
33	Temperatura média do trimestre frio	0	90	80	0	10	0	70	0	80	0
34	Temperatura média do trimestre quente	90	0	0	0	0	0	0	0	0	0
35	Temperatura média do trimestre seco	0	0	0	0	0	0	0	0	0	20
36	Temperatura média do trimestre úmido	60	0	0	0	0	10	0	0	0	0
37	Temperatura mínima do mês mais frio	40	10	0	0	30	50	0	50	10	100
38	Tipo de solo	90	0	0	50	20	100	0	10	0	90
39	Variação anual da temperatura	0	10	0	90	20	0	20	0	20	80

Em que: (1) *Anadenanthera colubrina* (Vell.) Brenan*;* (2) *Eugenia dysenterica DC.*; (3) *Hymenaea stignocarpa* Mart. ex Hayne ;(4) *Lafoensia vandelliana* Cham. & Schltdl.; (5) *Magonia pubescens* S. St.-Hil.; (6) *Pouteria ramiflora* (Mart.) Radlk.; (7) *Qualea grandiflora* Mart.; (8) *Qualea multiflora* Mart.; (9) *Qualea parviflora* Mart.; (10) *Terminalia fagifolia* Mart.

Do conjunto total de 39 variáveis, 34 variáveis foram selecionadas pela metodologia CFS ao menos uma vez para alguma espécie, e 5 variáveis não foram selecionadas nenhuma vez para nenhuma das espécies, sendo elas: Clima (classes Thornthwaite), rendimento específico, lâmina explotável (relacionados a disponibilidade de água superficial e subterrânea), grau de erosão e erodibilidade. A altitude foi a variável que mais se destacou entre as demais, sendo selecionada em 100% das repetições para 9 espécies (Tabela 1).

Entre as dez espécies arbóreas modeladas, em seis delas os algoritmos apresentaram desempenhos (AUC) estatisticamente iguais. Para a espécie *Eugenia dysenterica DC.* os algoritmos *random forest* e *multi layer perceptron* apresentaram desempenho superior ao J48, porém estatisticamente iguais entre si. Para as espécies *Qualea grandiflora Mart., Qualea multiflora Mart.* e *Qualea parviflora Mart.* o modelo *random forest* obteve uma diferença significativa frente aos modelos testados (Tabela 2).

Tabela 2. Resultado do teste T-pareado (0,05) entre os valores de AUC obtidos pelos três modelos nas dez espécies arbóreas.

Table 2. Results of the T-paired test (0.05) between the AUC values obtained by the three models in the ten tree species.

Espécies	*Area Under the Curve* (AUC)		
	J48	*Random Forest*	*MultilayerPerceptron*
Anadenanthera colubrina	0,73	0,73	0,80
Eugenia dysenterica	0,72	0,96*	0,89*
Hymenaea stignocarpa	0,90	0,96	0,85
Lafoensia vandelliana	0,80	0,75	0,75
Magonia pubescens	0,70	0,78	0,75
Pouteria ramiflora	0,68	0,75	0,78
Qualea grandiflora	0,85	0,97*	0,94
Qualea multiflora	0,70	0,84*	0,78
Qualea parviflora	0,82	0,94*	0,87
Terminaliafagifolia	0,69	0,70	0,79

*Diferença estatística de acordo com o teste T-pareado a 0,05 de significância.

Após a aplicação do teste T-pareado, os algoritmos que obtiveram melhor desempenho (maior AUC) por espécie, foram então aplicados para modelar a distribuição destas espécies arbóreas na sub-bacia do rio das Velhas. Os resultados desta aplicação por espécie podem ser visualizados na Figura 3.

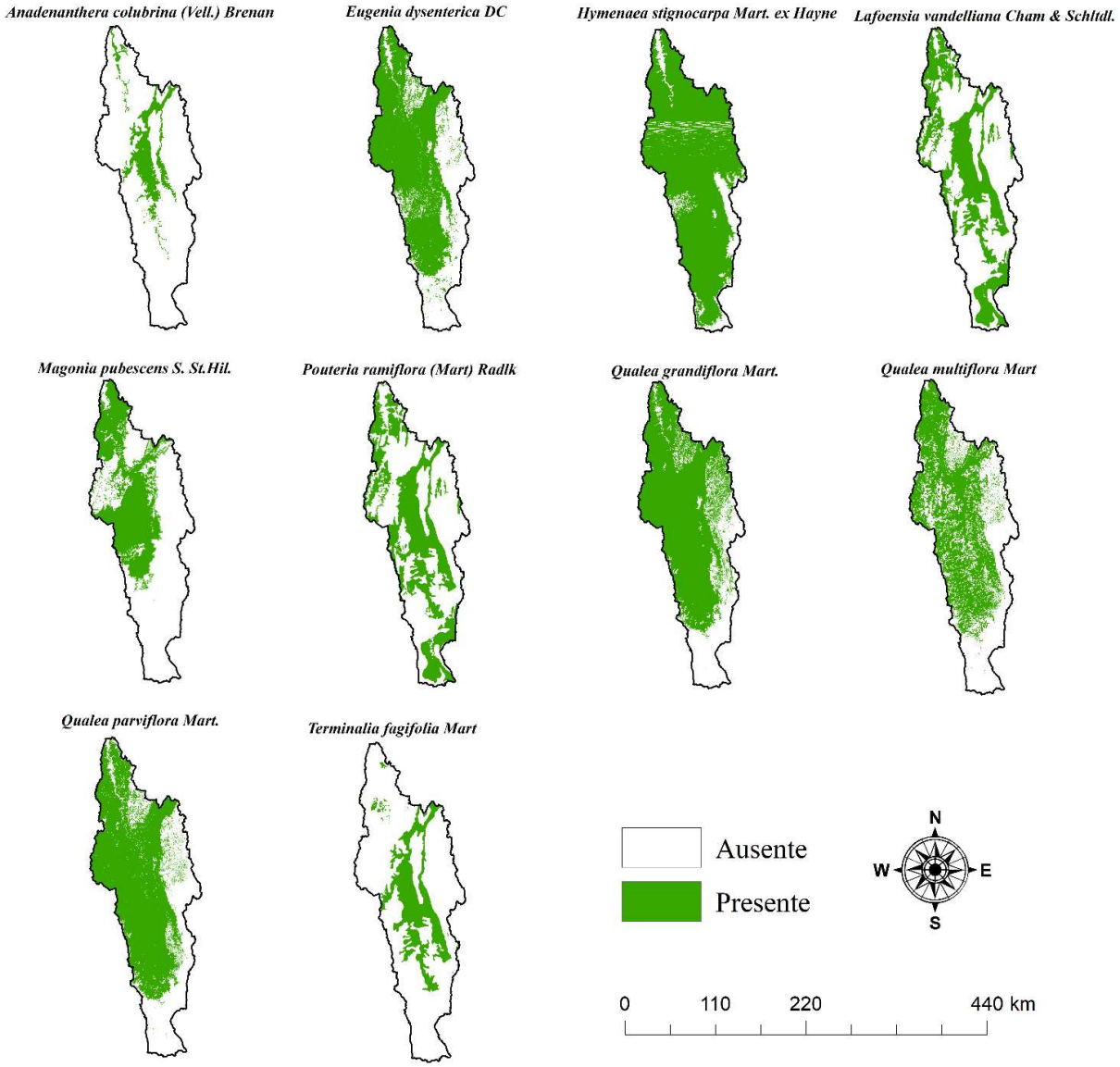

Figura 3. Mapa da distribuição potencial das espécies arbóreas florestais.
Figure 3. Map of potential distribution of forest tree species.

4. DISCUSSÃO

A distribuição das espécies resulta de uma série de fatores, como as características ambientais. A variável altitude, por exemplo, correlacionou-se com 9 das 10 espécies analisadas, seguido das variáveis precipitação dos trimestres seco, sazonalidade da precipitação e tipo de solo (8, 7 e 6 espécies, respectivamente). Um comportamento semelhante foi evidenciado por Callegaro et al. (2018), em que a maior parte das espécies mais abundantes foram influenciadas pela altitude. Chakraborty et al. (2016), ao avaliarem a distribuição de quatros espécies florestais, em relação aos impactos das mudanças climáticas, encontraram que a variável elevação tem forte relevância por aumentar significativamente a acurácia do modelo de distribuição. Também constataram que a tipologia de solos, como variável preditora da distribuição de espécies, influencia na melhoria do desempenho do modelo (WAN et al., 2017). Comportamento este que foi constatado para somente 3 espécies, principalmente devido ao fato de que a tipologia de solos empregada é uma variável categórica com 13 classes, as quais foram amostradas desbalanceadamente de acordo com os 77 fragmentos inventariados, levando a uma baixa caracterização das preferências de solos das espécies estudadas. Em relação à precipitação, Amissah et al., (2014), identificaram este fator como o principal na distribuição de espécies arbóreas.

Os resultados observados indicam que o desempenho de cada algoritmo está intrinsecamente relacionado ao tipo de distribuição da espécie modelada, como também pode ser constatado em estudos anteriores por Segurado; Araújo (2006); Elith et al. (2006); Pearson et al. (2006). Apesar de alguns modelos serem estatisticamente superiores frente aos demais para determinadas espécies, não há consenso de um método superior para todas as circunstâncias.

Lorena et al. (2011), comparando 9 classificadores de aprendizado de máquina na modelagem da distribuição de 35 espécies vegetais da família Bignoniaceae, obteve resultados satisfatórios com o algoritmo *Random Forest*. O método apresentou melhor desempenho em 29 espécies, dentre as 35 testadas. Nesta pesquisa o desempenho desse algoritmo também foi notável, sendo superior em 6 das 10 espécies modeladas. Em ambas as pesquisas o modelo apresentou desempenho estável (baixa variação entre os valores de AUC). Carvalho et al. (2017) compararam o *Random Forest* e as Redes Neurais Artificiais para modelar o nicho ecológico de quatro espécies florestais. O método *Random Forest* apresentou melhor desempenho para a modelagem de distribuição de todas as espécies.

5. CONCLUSÕES

Os resultados obtidos nesta pesquisa, assim como em outros estudos, demonstram que o desempenho dos modelos está intrinsecamente relacionado à espécie modelada. No entanto, dentre os algoritmos testados, o *Random Forest* surge como uma opção robusta para a modelagem da distribuição de espécies. Dentre as variáveis testadas, altitude, precipitação dos trimestres seco, sazonalidade da precipitação e tipo de solo destacaram-se como as variáveis que mais influenciam a distribuição das espécies arbóreas dentro da área de estudo.

Diante das ações antropogênicas sobre os ecossistemas florestais naturais e, seus impactos negativos sobre os serviços ecológicos prestados, a utilização da modelagem da distribuição de espécies pode auxiliar em estratégias para proteção, conservação e restauração da biodiversidade.

6. REFERÊNCIAS

AMISSAH, L.; MOHREN, G. M.; BONGERS, F.; HAWTHORNE, W. D.; POORTER, L. Rainfall and temperature affect tree species distribution in Ghana. **Journal of Tropical Ecology**, Cambridge, v. 30, n. 5, p. 435-446, 2014. DOI: http://dx.doi.org/10.1017/S026646741400025X

BHERING, S. B.; CHAGAS, C. S.; JUNIOR, W. C.; PEREIRA.; FILHO, B. C.; PINHEIRO, H. S. K. Mapeamento digital de areia, argila e carbono orgânico por modelos *Random Forest* sob diferentes resoluções espaciais. **Pesquisa Agropecuária Brasileira**, Brasília, v. 51, n. 9, p. 1359-1370, 2016. DOI: http://dx.doi.org/10.1590/S0100-204X2016000900035

BREIMAN, L. *Random forests*. **Machine Learning**, Boston, v.45, n.1, p.5-32, 2001.

CALLEGARO, R. N.; ARAUJO, M. M.; LONGHI, S. J.; ANDRZEJEWSKI, C. Influência de fatores ambientais sobre espécies vegetais em florestas estacionais para uso potencial em restauração. **Nativa**, Sinop, v. 6, n. 1, p. 91-99, 2018. DOI: http://dx.doi.org/10.31413/nativa.v6i1.4728

CANESSA, R.; SALDAÑA, A.; RÍOS, R. S.; GIANOLI, E. Functional trait variation predicts distribution of alien plant species across the light gradient in a temperate rainforest. **Perspectives in Plant Ecology, Evolution and Systematics**, Jena, v. 32, p. 49-55, 2018. DOI: https://dx.doi.org/10.1016/j.ppees.2018.04.002

CARVALHO, M. C.; GOMIDE, L. R.; SANTOS, R. M.; SCOLFORO, J. S.; CARVALHO, L. M. T.; MELLO, J. M. Modeling ecological niche of tree species in brazilian tropical area. **Cerne**, Lavras, v. 23, n. 2, p.229-240, jun. 2017. DOI: http://dx.doi.org/10.1590/01047760201723022308

CHAKRABORTY, A.; JOSHI, P. K.; SACHDEVA, K. Predicting distribution of major forest tree species to potential impacts of climate change in the central Himalayan region. **Ecological Engineering**, Oxford, v. 97, p. 593-609, 2016. DOI: http://dx.doi.org/10.1016/j.ecoleng.2016.10.006

COSTA, D. P.; COUTO, G. P.; SIQUEIRA, M. F.; CHURCHILL, S. P. Bryofloristic affinities betwwen Itatiaia National Park and tropical Andrean countries. **Phytotaxa**, Nova Zelândia, v. 346, n. 3, p. 203-220, 2018. DOI: https://doi.org/10.11646/phytotaxa.346.3.1

EHRLÉN, J.; MORRIS, W. F. Predicting changes in the distribution and abundance of species under environmental change. **Ecology Letters**, Oxford, v. 18, n. 3, p. 303-314, 2015. DOI: https://dx.doi.org/10.1111/ele.12410

ELITH, J.; GRAHAM, C. H.; ANDERSON, R. P.; DUDÍK, M.; FERRIER, S.; GUISAN, A.; HIJMANS, R.; HUETTMANN, F.; LEATHWICK, J.R.; LEHMANN, A.; LI, J.; LOHMANN, L. G.; LOISELLE, B. A.; MANION, G.; MORITZ, C.; NAKAMURA, M.; NAKAZAWA, Y.; OVERTON, J. McC.; PETERSON, T.; PHILLIPS, S. J.; RICHARDSON, K.; SCACHETTI-PEREIRA, R.; SCHAPIRE, R.; SOBERÓN, J.; WILLIAMNS, S.; WISZ, M. S.; ZIMMERMANN, N. E. Novel methods to improve prediction of species distributions from occurrence data. **Ecography**, Copenhagen, v. 29, n. 2, p. 129-151, 2006. DOI: https://dx.doi.org/10.1111/j.2006.0906-7590.04596.x

GARCÍA-CALLEJAS, D.; ARAÚJO, M. B. The effects of model and data complexity on predictions from species distributions models. **Ecological modelling**, Amsterdam, v. 326, p. 4-12, 2016. DOI: https://dx.doi.org/10.1016/j.ecolmodel.2015.06.002

GARNER, S. R. **Weka: the Waikato environment for knowledge analysis.** In: Proc. of the New Zealand Computer Science Research Students Conference, p. 57-64, 1995. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.54.3371>.

GARZÓN, M. B.; BLAZEK, R.; NETELER, M.; DIOS, R.S.; OLLERO, H.S.; FURLANELLO, C. Predicting habitat suitability with machine learning models: The potential area of *Pinussylvestris* L. um the Iberian Peninsula. **Ecological Modelling**, Amsterdam, v. 197, n. 3-4, p. 383-393, 2006. DOI: https://dx.doi.org/10.1016/j.ecolmodel.2006.03.015

GIANNINI, T. C.; SIQUEIRA, M. F.; ACOSTA, A. L.; BARRETO, F. C. C.; SARAIVA, A. M.; ALVES-DOS-SANTOS, I. Desafios atuais na modelagem preditiva de distribuição de espécies. **Rodriguésia**, Rio de Janeiro, v. 63, n. 3, p. 733-749, 2012. DOI: http://dx.doi.org/10.1590/S2175-78602012000300017

HENDERSON, E. B.; OHMANN, J. L.; GREGORY, M. J.; ROBERTS, H. M.; ZALD, H. Species distribution modelling for plant communities: stacked single species or multivariate modelling approaches?. **Applied vegetation science**, v. 17, n. 3, p. 516-527, 2014. DOI: https://dx.doi.org/10.1111/avsc.12085

HIJMANS, R. J.; 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**, Chichester, v. 25, n. 15, p. 1965-1978, 2005. DOI: https://dx.doi.org/10.1002/joc.1276

INZA, I.; CALVO, B.; ARMANANZAS, R.; BENGOETXEA, E.; LARRANAGA, P.; LOZANO, J. A. Machine Learnign: An Indispensable Tool in Bioinformatics. **Bioinformatics Methods in Clinical Research,** p. 25-48, 2009. DOI: https://dx.doi.org/10.1007/978-1-60327-194-3_2

JIMÉNEZ-VALVERDE, A. Insights into the area under the receiver operating characteristic curve (AUC) as a discrimination measure in species distribution modelling. **Global Ecology and Biogeography**, v. 21, n.4, p. 498 – 507, 2011. DOI: https://doi.org/10.1111/j.1466-8238.2011.00683.x

LORENA, A. C.; JACINTHO, L. F. O.; SIQUEIRA, M. F.; GIOVANNI, R.; LOHMANN, L. G.; CARVALHO, A. P. L. F.; YAMAMOTO, M. Comparing machine learning classifiers in potential distribution modelling. **Expert Systems with Applications,** New York, v. 38, n. 5, p. 5268-5275, 2011. DOI: https://dx.doi.org/10.1016/j.eswa.2010.10.031

MATEO, R. G.; GASTÓN, A.; AROCA-FERNANDEZ, M.J.; SAURA, S.; GARCÍA-VIÑAS, J. I. Optimization of forest sampling strategies for woody plant species distribution modelling at the landscape scale. **Forest Ecology and Management**, Amsterdam, v. 410, p. 104-113, 2018. DOI: https://dx.doi.org/10.1016/j.foreco.2017.12.046

MEROW, C.; SMITH, M. J.; EDWARDS, T. C.; GUISAN, A. McMAHON, S. M.; NORMAND, S.; THUILLER, W.; WUEST, R. O.; ZIMMERMANN, N. E.; ELITH, J. What do we gain from simplicity versus complexity in species distribution models? **Ecography**, Copenhagen, v. 37, n. 12, p. 1267-1281, 2014. DOI: https://dx.doi.org/10.1111/ecog.00845

MORENO-FERNÁNDEZ, D.; Space–time modeling of changes in the abundance and distribution of tree species. **Forest Ecology and Management**, Amsterdam, v. 372, p. 206-216, 2016. DOI: https://dx.doi.org/10.1016/j.foreco.2016.04.024

PAGLIA, A. P.; REZENDE, D. T.; KOCH, I.; KORTZ, A. R.; DONATTI, C. Modelos de distribuição de espécies em estratégias para a conservação da biodiversidade e para adaptação baseada em ecossistemas frente a mudanças climáticas. **Natureza & Conservação**, Curitiba, v. 10, n. 2, p. 231-234, 2012. DOI: http://dx.doi.org/10.4322/natcon.2012.031

QIN, A.; LIU, B.; GUO, Q.; BUSSMANN, R. W.; MA, F.; JIAN, Z.; XU, G.; PEI, S. Maxent modeling for predicting impacts of climate change on the potential distribution of *Thuja sutchuenens* is Franch., an extremely endangered conifer from southwestern China. **Global Ecology and Conservation**, v. 10, p. 139-146, 2017.DOI: https://dx.doi.org/10.1016/j.gecco.2017.02.004

QUINLAN, J. R. **C4.5: programs for Machine Learning.** Elsevier: 1993. 302 p.

ROBERTSON, M. P.; PETER, C. I.; VILLET, M.; RIPLEY, B.S. Comparing models for predicting species' potential distributions: a case study using correlative and mechanism predictive modelling techniques. **Ecological Modelling**, Amsterdam, v. 164, n. 2-3, p. 153-167, 2003. DOI: https://dx.doi.org/10.1016/S0304-3800(03)00028-0

SCOLFORO, J. R. S. **Biometria florestal: modelos de crescimento e produção florestal**. Lavras: FAEPE–UFLA, 2006. 393 p.

SCOLFORO, J. R. S.; CARVALHO, L. M. T.; OLIVEIRA, A. D. **Zoneamento ecológico-econômico do Estado de Minas Gerais: componentes geofísico e biótico**. Lavras: Editora UFLA, 2008. 161 p.

SEGURADO, P.; ARAÚJO, M. B. An evaluation method for modelling species distributions. **Journal of Biogeography**, Oxford, v. 31, n. 10, p. 1555-1568, 2004.DOI: https://dx.doi.org/10.1111/j.1365-2699.2004.01076.x

SOUZA, R. A.; MARCO J. R. P. The use of species distribution models to predict the spatial distribution of deforestation in the western Brazilian Amazon. **Ecological Modelling**, Amsterdam, v. 291, p. 250-259, 2014. DOI: https://dx.doi.org/10.1016/j.ecolmodel.2014.07.007

TRABUCCO, A.; ZOMER, R. J. **Global Soil Water Balance Geospatial Database.** CGIAR Consortium for Spatial Information, 2010.

WAN, J. Z.; WANG, C. J.; YU, F. H. Modeling impacts of human footprint and soil variability on the potential distribution of invasive plant species in different biomes. **Acta Oecologica**, Paris, v. 85, 141-149, 2017. DOI: https://dx.doi.org/10.1016/j.actao.2017.10.008