Mônica Canaan Carvalho[1+], Lucas Rezende Gomide[1], Rubens Manoel dos Santos[1], José Roberto Soares Scolforo[1], Luís Marcelo Tavares de Carvalho[1], José Márcio de Mello[1]

MODELING ECOLOGICAL NICHE OF TREE SPECIES IN BRAZILIAN TROPICAL AREA

Keywords:
Artificial Neural Networks
Phytogeography
Random Forest

ABSTRACT: Modeling of the ecological niche of vegetal species is useful for understanding the species-environment relationship, for prediction of responses to climate changes and for correct reforestation programs and establishment of plantation's recommendation. The objective of this work was to establish a model for the distribution of four tree species (*Casearia sylvestris, Copaifera langsdorffii, Croton floribundus* and *Tapirira guianensis*), widely used in reforestation projects in the state of Minas Gerais, Brazil. In addition, we analyzed the relationship between environmental characteristics and the occurrence of species and tested the performance of Random Forest and Artificial Neural Networks as modeling methods. These methods were evaluated by their overall accuracy, sensitivity, specificity, Kappa, true skill statistic and the area under the receiver operating curve. The results showed the species *Casearia sylvestris, Copaifera langsdorffii* and *Tapirira guianensis* widely occurring in the state of Minas Gerais, including a broad range of environmental variables. *Croton floribundus* had restricted occurrence in the southern state, showing narrow environmental variation. The resulting algorithms demonstrated greater performance when modeling restricted geographic and environmental species, as well as species occurring with high prevalence in data. The algorithm Random Forest performed better for distribution modeling of all species, although the results varied for each metric and species. The maps generated had acceptable metrics and are supported by and ecological information obtained from other sources, constituting a useful tool to understand the ecology and biogeography of the target species.

Histórico:
Recebido 10/02/2017
Aceito 22/05/2017

MODELAGEM DO NICHO ECOLÓGICOS DE ESPÉCIES ARBÓREAS EM UMA ÁREA TROPICAL BRASILEIRA

Palavras chave:
Fitogeografia
Random Forest
Redes Neurais Artificiais

RESUMO: A modelagem de nicho ecológico de uma espécie é útil para a compreensão da relação espécie-ambiente, para a previsão do comportamento frente às alterações climáticas e para a indicação correta em reflorestamentos e estabelecimento de plantações. O objetivo foi modelar a distribuição de quatro espécies arbóreas amplamente utilizadas em projetos de reflorestamento no estado de Minas Gerais (*Casearia sylvestris*, *Copaifera langsdorffii*, *Croton floribundus* e *Tapirira guianensis*). Como complemento, o objetivo foi analisar a relação entre as características ambientais e a ocorrência de espécies e testar o desempenho das técnicas random forest e redes neurais artificiais como métodos de modelagem. Estes métodos foram avaliados pelas métricas de acurácia global, sensibilidade, especificidade, kappa, true skill statistic e área sob a curva. Verificou-se que as espécies *Casearia sylvestris*, *Copaifera langsdorffii* e *Tapirira guianensis* apresentaram ampla área de ocorrência no estado Minas Gerais, cobrindo ampla gama de variáveis ambientais. Já Croton floribundus demonstrou ocorrência restrita do sul do estado, mostrando estreita variação ambiental. Os resultados dos algoritmos demonstraram maior desempenho na modelagem de espécies geograficamente e ambientalmente restritas, bem como espécies com alta prevalência em dados de ocorrência. O algoritmo random forest alcançou melhor desempenho na modelagem da distribuição de todas as espécies, embora os resultados variem para cada métrica e espécie. Os mapas gerados possuem métricas aceitáveis e são apoiadas por informações ecológicas obtidas em outras fontes, constituindo uma ferramenta útil no entendimento de sua ecologia e biogeografia.

+Correspondência:
monicacanaan@gmail.com

DOI:
10.1590/01047760201723022308

[1] Federal University of Lavras - Lavras, Minas Gerais, Brazil

INTRODUCTION

The interest in describing and understanding geographic and environmental distribution of species is a very old concern (GRINNELL, 1917; HUTCHINSON, 1957). Predominantly, the species distribution is limited by the multidimensional ecological niche of occupancy space (MACARTHUR, 1972), restricted due to several factors, such as climate, soil, disturbances and biotic factors. These factors act on different spatial scales and act as filters to determine which species are more suitable to remain alive through time in the local community (WHITTAKER, 1967; TER STEEGE; ZAGT, 2002). Identifying ecological niches of vegetal species across environmental gradients contributes for the comprehension of forest diversity and ecology (RATTER et al., 2003; OLIVEIRA-FILHO et al., 2005), as well as to understanding potential responses from vegetal species to climate changes (MAIORANO et al., 2013; WANG et al., 2016). Furthermore, it provides better insight of environmental requirements for each species, which is helpful in ecological restoration projects and establishment of plantations (COELHO et al. 2016).

Moreover, in the last 20 years, many methods have been developed to understand and estimate the ecological niche. These methods' main principle is based on relations between known occurrences and environmental conditions. Ecological niche modeling is relevant and actual. This method is useful to understand the species biogeography and its potential occurrence through the development of maps as results. These techniques are widely applied for different goals, such as: conservation of rare or endangered species (QUEIROZ et al., 2012; HAMILTON et al., 2015); identification of climate change impacts (CHUN; LEE, 2013; GWITIRA et al., 2014); reintroduction of species (HIRZEL et al., 2002; MEINERI et al., 2015); identifying of potential areas for invasive species (VACLAVIK; MEENTEMEYER, 2009; GALLIEN et al., 2012). Several techniques for ecological niche modeling are available and can be classified in two models groups: 1) classical statistical and 2) non-classical statistical. The second group is composed by methods like the Random Forest and Artificial Neural Networks which usually has demonstrated superior performance in many studies (ELITH et al., 2006; LORENA et al., 2011; POUTEAU et al., 2012). The advantages are the ability to work with correlated predictors, nonlinear relationships and noisy data. These characteristics are essential to improve the performance and reduce errors in ecological modeling (GARZÓN et al., 2006).

Nowadays, there are some gaps concerning native tree species and their environmental preferences in the tropical area. This viewpoint restricts the suggestion of correct species in forest restoration programs within natural areas (COELHO et al., 2016). According to Lima et al. 2009, the ecological and silvicultural procedures work together to guide the best strategy to recover the damaged systems and guarantee their sustainability in the future. In this context, the main objective was to model the ecological niche of 4 woody species, widely used in reforestation projects: *Casearia sylvestris*, *Copaifera langsdorffii*, *Croton floribundus* and *Tapirira guianensis*. On the other side, was the association among the deep descriptive analysis between environmental characteristics and each species' occurrence. Finally, we tested the performance of the Random Forest and Artificial Neural Networks as modeling methods.

MATERIAL AND METHODS

Study area

The Brazilian tropical area in this study comprises the Minas Gerais state covering 586.53 km^2 (Figure 1). Due to the large area, a wide range of altitude (between 40 and 2,600 meters) and eight climate classes according to the Thornthwaite index (CARVALHO et al., 2008) compose the study. This climatic index gradient decreases from the south to the north of the state. In general, the regions of higher altitudes are characterized by humid and super-humid climates and the regions of lower altitudes by sub-humid to semi-arid climates. The vegetation distribution covers three biomes: Cerrado, Mata Atlântica and Caatinga. The Cerrado area covers 57% of the area of the state (central-western region), the Mata Atlântica 41% (eastern region) and Caatinga only 2% (north and west region) (IEF, 2015). Furthermore, these biomes comprise the phytophysiognomies: ombrófila, estacional decidual, estacional semi-decidual, veredas, campo, campo rupestre, campo cerrado, cerrado and cerradão. The predominant soil class is the latosol with spots of neosol litholic, argisol and cambisol.

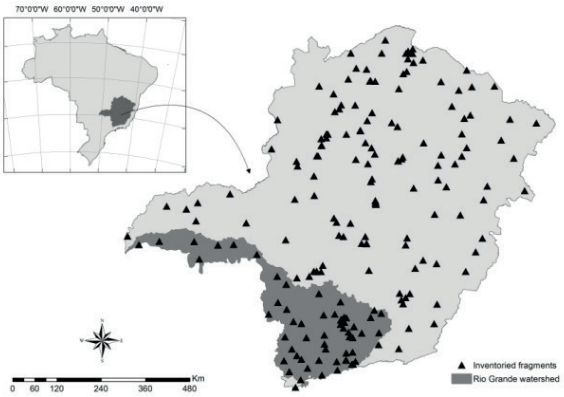

FIGURE 1 Location of study area and inventoried fragments.

Selected species and occurrence data

According to the Forest Inventory of Minas Gerais (SCOLFORO; CARVALHO, 2008), the selected species are native trees and abundant in the study area. In addition, they are largely used in reforestation programs and provide woody products as well as secondary forest products, like oils and resins, and may generate economic return under sustainable management. These species are *Casearia sylvestris* (Salicaceae), *Copaifera langsdorffii* (Leguminosae), *Croton floribundus* (Euphorbiaceae) and *Tapirira guianensis* (Anacardiaceae), which are usually found in riparian forests of the Cerrado as well as in Mata Atlântica. Oliveira-Filho and Ratter (1995) described the location area of *Casearia sylvestris, Copaifera langsdorffi* and *Tapirira guianensis* throughout forest galleries, connecting Amazônia and Mata Atlântica. *Croton floribundus* is most common in primary or secondary remnant areas of the semi-deciduous tropical forest (OLIVEIRA-FILHO et al., 2006).

The data is derived from 197 areas of native vegetation (Figure 1) from the Forest Inventories of Minas Gerais (SCOLFORO; CARVALHO, 2008) and the Rio Grande watershed project. These fragments were chosen according to the physiognomy and spatial distribution of each project scope and provided only natural occurrences and accurate geographical position. The diameter at breast height (dbh) of individual trees was measured, only trees with a diameter greater than 5 cm were considered and identified. Finally, the occurrence species data were based in the fragment location. This procedure was considered instead of plot level because of the low environmental resolution data, which covers larges areas. We defined the fragment centerpoint to extract environmental variables and species occurrence. Boolean variable was used to indicate the presence (1) or absence (0) of each species and for categorical variables. The example of data used to train the algorithms can be partially in the Table 1.

Environmental data

We used 12 environmental variables associated to climate, topography and soils. Climatic variables mean annual temperature, temperature seasonality, maximum temperature, minimum temperature, mean annual precipitation, precipitation seasonality, precipitation during the dry and rainy months and altitude were obtained from Worldclim database (HIJMANS et al., 2005). This base is widely used in ecological modeling studies (ELITH et al., 2006; LORENA et al., 2011) and derived from historical series (1950-2000) by global climate data interpolation with spatial resolution equals to 1 km. Topography (slope) and soil data (soil class and water regime) were obtained from the Ecological Economic Zoning of Minas Gerais (CARVALHO et al., 2008). Soil classes are categorical comprehending 11 soil classes (argisol, cambisol, espodosol, gleysol, latosol, luvisol, fluvisol, litholicneosol, quartzipsament, nitosoland and planosol), 4 soil water regime classes (xeric, aquic, udic and ustic) and 4 slope classes (plane or soft wavy, wavy, strong wavy and mountainous). Geographical projection, pixel size and spatial extent were similar for all variables with 1 km of spatial resolution. The spatial reference system adopted was the South America Albers Conic Equal, Datum SAD69.

Algorithms

Random Forest (RF) and Artificial Neural Networks (ANNs) were tested to predict the ecological niche of the tree species. The first one, proposed by Breiman (2001), is a combination method of classifiers (ensemble) decision trees. These trees are built by Random Forest using the CART algorithm (classification and regression trees). This algorithm divides a set of heterogeneous data (root) into homogeneous sub-set classes (leaves), generating classification rules based on attributes (nodes). The criterion for data partition is based on information gain. The mathematical procedure consists in decreasing the data set entropy after split for some selected attribute. The Random Forest build decision trees under different sets of training (bootstrap). Usually, every split decision chose randomly *m* attributes and the direction to growth. Finally, the gradient is quantified based on entropy gain and the tree is created. Each decision tree will have its classification. Then the Random Forest classification defines the final classes according to a rank for most voted trees. This method has been applied in ecological studies (CUTLER et al., 2007; PRASAD et al., 2006) offering powerful alternatives to traditional parametric and semiparametric

TABLE 1 Example of the data used in the modeling, where X - latitude; Y - longitude; T° C - mean annual temperature; Alt-altitude; P(mm) - mean annual precipitation.

Location		Numeric inputs				Categorical inputs						Output
X	Y	T°C	Alt	P(mm)	...	latosol	...	udic	...	soft_wavy	...	Occurrence
1741950	1601327	20.7	665	1220	...	0	...	1	...	0	...	1
1813670	1728537	20.8	666	1272	...	0	...	1	...	1	...	0
1354313	1509267	19.7	832	1302	...	1	...	0	...	1	...	1

statistical methods with high accuracy and the ability to model complex interactions between variables.

The algorithm parameters were set after previous tests to improve its efficiency. We established 100 decision trees to be created without length and pruning performance constraints. The number of attributes (*m*) used to create each tree was defined by the equation 1 (FRANK et al., 2016), where *n* represents the total number of available variables.

$$m = \log_2(n) + 1 \quad [1]$$

Artificial Neural Networks are techniques inspired on the structure, processing method and learning ability of the brain. They are composed for connected neurons (simple processing units) disposed in one or more layers. Each neuron is connected to one or more units through weighted connections, which simulate the biological synapses. Each input layer neuron receives x_i independent variables values. Random weights w_i are also given to these values and the sum of these weights w_i and their attributes x_i are input values for the activation function. The neurons output of the initial layer act as input to the following layer neurons and so on. The output of the last neuron layer is the final Artificial Neural Network (ANN) classification y_i. The ANN learning consists to of adjusting the weights to approximate the outputs of the ANN to the desired outputs known from the training data d_i. The most used learning algorithm is backpropagation, which propagates the error of the final layer for the initial layers through error derivatives. The structure and operation of ANNs is well discussed by a number of authors (FAUSETT, 1994; HAYKIN, 1994; OZESMI et al., 2006).

We used a multilayer-perceptron type of Artificial Neural Network with 12 independent variables (9 numeric and 3 categorical) and one output (presence or absence). The optimal parameters (number of hidden layers and neurons within them, learning rate, *momentum* and activation function) were determined empirically by creating multiple networks, with all other parameters held constant. By trial and error the following parameters range were tested: 1-2 hidden layers, 1-15 neurons in each hidden layer, 0.01-0.9- learning rate and *momentum*. Threshold function was chosen as activation function since it limits the output between 0 or 1 (presence or absence). Network performance can be sensitive to the random initial weight values set prior to training (OZESMI et al., 2006). For this reason, 10 networks were run based on the same architecture, after resetting the initial weights, and performance was assessed based on the averaged predictions across all runs by the area under the receiver operating curve (AUC). Behind these tests the Artificial Neural Network formation was determined in two hidden layers of processing, with 15 and 2 neurons respectively. The learning rate and *momentum* term were established at 0.3 and 0.2.

Algorithms assessment and application

The occurrence species and environmental variables data were divided into two groups: 1) training and 2) validation (predictive validity). Phillips and Dudik (2008) suggest this strategy to check the method's accuracy and usually adopt 70% and 30%, for training and validation, respectively. Cross-validation was applied to resort all limited areas to train and validate models. Three folds were formed from 197 points. For each technique, the examples from 2 folds were then used to train a classifier, which was evaluated in the remaining fold. This process was repeated 3 times, using at each cycle a different fold for validate. The experiment was conducted with 10 replicates and algorithm measures were obtained by mean of each replicate for validation set. We applied paired t-test statistic (95% confidence) between means of algorithms' measures for its comparison within each species. The software Weka (FRANK et al., 2016) was used to training, evaluation and implementation of Random Forest and Artificial Neural Network.

We applied six metrics to assess the predictive algorithm accuracy, such as: i) Overall accuracy – is the percentage of correctly classified data (equation 2); ii) Sensitivity – the probability of a occurrence species presence be correctly predicted; (equation 3), iii) Specificity - the probability of occurrence species absence be correctly predicted (equation 4); iv) Cohen's Kappa – is widely used to measure the correctly predict occurrence rates (equation 5); v) True skill statistic (TSS) (equation 6) and vi) Area under the receiver operating curve (AUC).

Kappa's statistics ranges from -1 to +1, where +1 indicates perfect agreement and values of zero or less indicate a performance no better than random. The advantages of Kappa are its simplicity, the fact that both commission and omission errors are accounted for in one parameter. However, Kappa's statistics is dependent on prevalence data, introducing bias and statistical artefacts to estimate models' accuracy (Mouton et al. 2010). Due this fact we also applied the true skill statistic (TSS) which is a Kappa variation and avoid this prevalence dependency (ALLOUCHE et al., 2006). The AUC was used to evaluate the classifiers effectiveness in the classification of presence/absence data. To construct a ROC curve we ploted the 1-specificity on the x-axis and sensitivity on the y-axis. These metrics were derived from the confusion matrix, where *a*, is the number of presences correctly classified; *b*, the number of points where the specie was not found but the algorithm predicted its presence; *c*, the number of points where the specie was found but the algorithm predicted its absence; *d*, the number of absences correctly classified; *n*, the total number of data.

$$\text{Overall accuracy} = \frac{a+d}{n} \qquad [2]$$

$$\text{Sensitivity} = \frac{a}{a+c} \qquad [3]$$

$$\text{Specificity} = \frac{d}{b+d} \qquad [4]$$

$$\text{Cohen's Kappa} = \frac{\left(\frac{a+d}{n}\right) - \frac{(a+b)(a+c)+(c+d)(d+b)}{n^2}}{1 - \frac{(a+b)(a+c)+(c+d)(d+b)}{n^2}} \qquad [5]$$

$$\text{TSS} = \frac{ad-bc}{(a+c)(b+d)} \qquad [6]$$

RESULTS AND DISCUSSION

Occurrence species areas

The natural geographical distribution of *Casearia sylvestris*, *Copaifera langsdorffii* and *Tapirira guianensis* demonstrated a wide occurrence throughout the whole state area (Figure 2). On the other hand, *Croton floribundus* showed limited geographic distribution with high occurrence in the southern area and only 24% of the total inventoried areas. The number of areas with occurrence of *Casearia sylvestris*, *Copaifera langsdorffi*, *Croton floribundus* and *Tapirira guianensis,* was of 109, 130, 48 and 92, respectively. The absence occurrence was of 88, 67, 149 and 105 following the same order.

The occurrence of species has been attributed to a range of factors, including mainly environmental attributes. The soil variable is responsible to explain more than 60% of the presence areas, where latosol is predominant. Oppositely, cambisol and litholic neosol are able to explain just 10% and 8%, respectively. The moisture soil regime contains 44% occurrence of udic (the water moves through the soil in all months) and 53% ustic (seasonal drainage) for *Casearia sylvestris* and *Tapirira guianensis*. Furthermore, *Copaifera langsdorffi* showed preference for ustic soils (62%), whereas *Croton floribundus* showed greater occurrence in udic soils (72%).

Topography is another variable used to distinguish species habitats. The slope attributes (plane and soft wavy) were most observed in the presence areas. However, the sloping areas (33% wavy and 10% strong wavy) were most preferred by *Croton floribundus,* which concentrates in the southern areas. This specie had a habitat preference for higher elevations (average 940 m) and rainfall (1,484.7 mm mean per year) despite its preference for mild temperatures (19.5°C). *Casearia sylvestris* and *Tapirira guianensis* had similar habitat according to thevariables observed. However, *Tapirira guianensis* showed slightly more tolerance for areas with higher temperature and less rainfall. The intensity of all topographic aspects can model each specie occurrence. *Copaifera langsdorffi, for example,* tolerates low altitudes (758.7 meters), higher temperatures (21.4° C) and 1,361.6 mm/year rainfall on average. It is possible to evaluate the pressure from geoclimatic variables that affects the species, as shown in Table 2.

Among the standard deviation of geoclimatic variables (environmental variation), *Croton floribundus* has a more restricted ecological niche and can survive properly only in a narrow range of environmental conditions. While the widespread species, *Casearia sylvestris*, *Copaifera langsdorffii* and *Tapirira guianensis,* take another strategy to get success in a large geographic and environmental space.

Regarding the interaction of environmental aspects, the dispersion seed strategy adopted by each species and its successional group also influences their natural occurrence (OLIVEIRA-FILHO et al., 2006; SILVESTRINI; SANTOS, 2015). *Croton floribundus* seasonally produces large number of seeds, however, this specie has an autochoric dispersion, which reduces

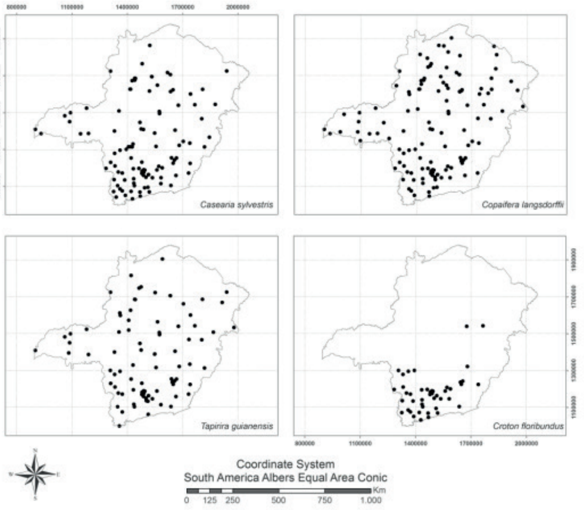

FIGURE 2 Occurrence of all target species quantified from inventory.

TABLE 2 Environmental Information of natural occurrence species with the mean values and their standard deviation (in brackets), for temperature (T°C) and rain precipitation (P mm).

Environmental variables		*Casearia sylvestris*	*Copaifera langsdorffii*	*Croton floribundus*	*Tapirira guianensis*
Altitude (m)		808.55 (252.28)	758.75 (219.53)	940.56 (163.17)	808.39 (259.50)
T (°C)	maximum	28.96 (2.19)	29.55 (1.93)	27.57 (1.22)	29.02 (2.18)
	mean	20.82 (19.96)	21.42 (1.75)	19.49 (1.16)	20.94 (2.37)
	minimum	10.21 (2.15)	10.96 (1.94)	8.86 (1.52)	10.63 (2.22)
P (mm)	maximum	272.55 (34.50)	263.78 (38.21)	286.71 (25.16)	272.97 (38.09)
	minimum	13.10 (6.89)	10.91 (6.80)	16.67 (4.75)	12.46 (7.30)
	anual mean	1377.14 (189.26)	1306.12 (208.14)	1484.75 (114.49)	1361.65 (200.33)

the seed dispersal range. Complementarily, the seed germination only occurs in sites with aspecific range of temperature variation (VÁLIO; SCARPA, 2001).

Casearia sylvestris and *Tapirira guianensis* are widespread in the Brazilian territory. The distribution pattern extends from the Amazonian to the Atlantic forests through middle lands in Brazil (OLIVEIRA; RATTER, 1995). According to literature, these species does not grow on swampy ground, nor excessively drained sites, but they are able to survive in annually flooded areas (RATTER et al., 2003; OLIVEIRA-FILHO et al., 2005). Just as *Croton floribundus*, these species occur mainly in primary successional stage forests, being shade intolerant species. However, its dispersal is zoochoric allowing a wider dispersion (AQUINO; BARBOSA, 2009). *Copaifera langsdorffii* was the most widespread species in our study area. It has a large distribution range in South America, which includes areas from northern Argentina, southern Bolivia and the Brazilian Savanna (OLIVEIRA-FILHO; RATTER, 1995; RATTER et al., 2003). Moreover, the seed tolerance to high temperatures (SOUZA et al., 2015), its zoochoric dispersal (SEBBENN et al., 2011) and its shade tolerance (AQUINO; BARBOSA, 2009) corroborates with its wide geographic distribution.

Algorithms assessment

Regarding the overall accuracy, the RF algorithm numerically surpassed ANN for all evaluated species, but with significant differences only for *Croton floribundus* (Table 3). It was observed that the overall accuracy varied according to the species modeled, as observed in previous studies (SEGURADO; ARAÚJO, 2004; ELITH et al., 2006). *Croton floribundus* obtained the highest percentage of correctly classified data compared to other species, with an accuracy of 90.4% of data achieved by RF. It is possible to verify the metrics for each species achieved by RF and ANN algorithms in Table 3.

In general, the overall accuracy values achieved are within the range obtained in previous studies (FUKUDA et al., 2013). Overall accuracy larger than 90% is commonly found in modeling work using large database from satellite images (GARZÓN et al., 2006; WANG et al., 2016). The high performance of *Croton floribundus* is related to its concentrate area of occurrence. According to Stockwell and Peterson (2002) and Segurando and Araújo (2004), species with a widespread occupancy area show greater overall errors.

The probability of correct presence prediction (sensitivity) was higher for *Copaifera langsdorffii* and *Croton floribundus* (0.88 and 0.80 respectively). *Tapirira guianensis* and *Casearia sylvestris* obtained only 0.74 of this metric when modeled by ANN, which was inferior to RF. The algorithms' performance ranged between species although without differences by the t-test (95%). On the other hand, the specificity values obtained by the algorithms for *Copaifera langsdorffii* were the smallest among the species studied (0.53 for both methods). *Croton floribundus* again showed better metrics, with statistically superior performance achieved by RF with 0.94. In comparison with ANN this algorithm also achieved superior performance according to the t test for *Tapirira guianensis*, which classified correctly 69% of absences. The same pattern was observed for *Casearia sylvestris*, but without differences between the algorithms.

Copaifera langsdorffii achieved a high rate of correctly classified presences and therefore high overall accuracy. This fact is related to the high presence (130) against absence (67) in the database. Usually, when the database favors some occupancy pattern, the results should be overestimated. The high number of presences induced the overestimation for occurrence locals by the algorithms tested. This panorama coincides with observations that sensitivity (true positive rate) was higher for widespread species and lowers for restricted-range species, while specificity (true negative rate) was lower for widespread species and higher for restricted-range species (SEGURADO; ARAÚJO, 2004, MOUNTON et al., 2010).

The high sensitivity and specificity values for *Croton floribundus* indicates the algorithms skills to distingue the

TABLE 3 Statistical results to assess the predictive accuracy of Random Forest (RF) and Artificial Neural Networks (ANN) algorithms.

Species	Algorithm	Overall accuracy (%)	Specificity	Sensitivity	AUC	Kappa	TSS
Casearia sylvestris	RF	70.61 (4.11)	0.67 (0.09)	0.73 (0.07)	0.79 (0.04)	0.41 (0.08)	0.41 (0.08)
	ANN	69.24 (5.49)	0.64 (0.12)	0.74 (0.08)	0.77 (0.06)	0.37 (0.11)	0.37 (0.13)
Copaifera langsdorffii	RF	76.03 (3.30)	0.53 (0.11)	0.88 (0.04)	0.79 (0.05)	0.43 (0.09)	0.41 (0.09)
	ANN	71.73 (5.32)	0.53 (0.13)	0.81 (0.08)	0.75 (0.06)	0.35 (0.12)	0.34 (0.12)
Croton floribundus	RF	90.40 (2.81)**	0.94 (0.03)**	0.80 (0.09)	0.96 (0.02)**	0.74 (0.08)	0.74 (0.09)
	ANN	82.69 (4.44)	0.82 (0.06)	0.84 (0.10)	0.88 (0.03)	0.59 (0.09)	0.67 (0.09)
Tapirira guianensis	RF	64.97 (4.66)	0.69 (0.07)**	0.60 (0.09)	0.72 (0.05)**	0.29 (0.10)	0.29 (0.09)
	ANN	61.27 (4.40)	0.50 (0.14)	0.74 (0.14)	0.64 (0.05)	0.24 (0.09)	0.26 (0.09)

() - standard deviations, ** mean values for significant difference between methods (α=0.05)

presence or absence classes correctly. Therefore, it is possible to assume a more restricted and homogeneous ecological niche for this species when compared to others. The result is corroborated with the real and limited geographical cover areas for this species. *Casearia sylvestris* and *Tapirira guianensis* became a hard task for all algorithms to split the presence/absence classes. The widespread species distribution is usually arduous work and similar results were found by Segurado and Araújo (2004), Cluter et al. (2007) and Wang et al. (2016). On the other side, *Copaifera langsdorffi* assumed the opposite tends because the large number of presences data which amplify the sensitivity. Moreover, the overall accuracy of the models is influenced by the prevalence of data.

Lower imbalance between classes resulted in worse accuracy as observed in *Casearia sylvestris* and *Tapirira guianensis*. This result also can be explained due to similarities between the occurrence data. Data for absence and presence sites may have been similar, which promoted a hard task to dissociate tendencies for these species. In many cases, the absence of species is usually not defined only by environmental characteristics, but scattering factors and human colonization history (PULLIAM, 2000).

The AUC results from RF demonstrated high performance for *Casearia sylvestris*, *Copaifera langsdorffii* and *Tapirira guianensis* (0.7< AUC <0.8) and excellent performance for *Croton floribundus* (0.9< AUC <1). The same pattern was observed for ANN except for *Tapirira guianensis* (0.6< AUC <0.7). In all cases, the RF was numerically superior than ANN in terms of AUC. According to Phillips and Dudik (2008), models with AUC values above 0.75 are considered potentially useful.

Kappa and TSS results showed similar tendency and suggest that all model outputs are no randomly effect (Kappa and TSS >0). Once again, the algorithms showed higher perform for *Croton floribundus* and lower for *Tapirira guianensis*. The assessment classification proposed by Monserud and Leemans (1992) indicated a poor performance for both algorithms except for *Croton floribundus*. Cluter et al. (2007) employed RF to model four invasive species (*Verbascum thapsus*, *Urtica dioica*, *Cirsium vulgare* and *Marrubium vulgare*) distribution, with 8,251 occurrence data. The mean Kappa values obtained were ranging from 0.607 to 0.809 and the authors considered excellent performance for all species. The high number of observations was associated to the high performance achieved. Moisen et al. (2006) while modeling 13 trees species derived from 1,930 sample plots obtained average of 0.87 for AUC while the Kappa ranged from 0.19 to 0.75. Motloung et al. (2014), modeling the distribution of 15 trees species obtained an overall TSS of 0.15, sensitivityof 0.80 and specificity of 0.35. The hypothesis to understanding the algorithm performance is associated to environmental gradients and species, not only the size of the database. The experimental area from the present work includes a huge landscape with diverse gradients.

The algorithms' performance is essential to understand the method's limitations to predict occurrence species. In others words, this statistical analysis supports the results to be spatial and geographical validated into the maps. The superiority of the RF against ANN was clear, although this difference, in some instances, was not confirmed by the t-test. This high performance was also supported by different authors (ELITH et al., 2006; GARZON et al., 2006; LORENA et al., 2011). Fukuda et al. (2013) suggested RF as an accurate method to model the species distribution when compared with other 6 algorithms. However, they found inconsistencies between different performance measures, showing that different models may obtain a high score on a particular aspect and perform worse on other aspects. Rodrigues-Galiano et al. (2015) applied Artificial Neural Networks, Random Forest, Regression Trees and Support Vector Machines to model mineral prospectivity and also corroborate with this same trend.

The superior performance of RF can be linked to the fact that tree-based (discrete) models (like RF) may be able to distinguish presence and absence cases better than models with continuous outputs such as Support Vector Machines and ANN (FUKUDA et al. 2013). Furthermore, the ANN training is complex and requires definition of their structure and parameters set, which is a time consuming task. Moreover, the results are strongly sensitive to parameters' variations (RODRIGUES-GALIANO et al., 2015).

Ecological niche predicted maps

Instead, based on algorithm performance, we applied only RF to predict the potential distribution of all species. The predicted ecological niches are showed in Figure 3. The percentage of cover area was of 19% for *Croton floribundus* (117,654 km^2), 54% for *Casearia sylvestris* (327,677 km^2), 57% for *Tapirira guianensis* (343,538 km^2) and 80% for *Copaifera langsdorffii* (486,851 km^2) of the Minas Gerais state.

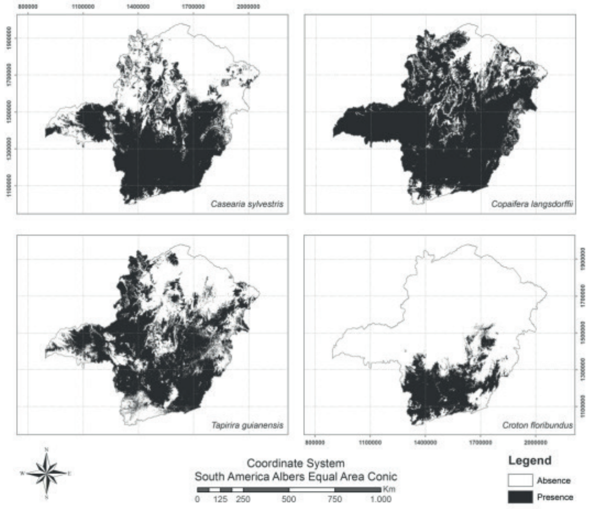

FIGURE 3 Potential distribution predicted by Random Forest for species *Casearia sylvestris*, *Copaifera langsdorffii*, *Croton floribundus* and *Tapirira guianensis* in Minas Gerais state.

The geographical area of *Copaifera langsdorffii* (Figure 3) is comprehends almost the entire map studied, except in the extreme parts, such as the southern and northern of Minas Gerais state. This wide distribution in the state was expected, since the species is considered of generalist habitat, occurring in various forest formations in the Cerrado and Mata Atlântica, provinces included in the study area (LOPES et al., 2012). In contrast, we recognized a short clustering area for *Croton floribundus* despite the literature also describing it as a widespread pioneer species (SILVESTRINI et al., 2015). However, their natural occurrence in the state cover a small environmental variation when compare with other species. *Casearia sylvestris* and *Tapirira guianensis* covered a similar predicted area (Figure 3). This suggests their niche is overlapping itself and the conservation and managements rules must be associated.

Raes (2012) suggested that partial model including artificial boundaries does not reflect the real species occurrence. Probably, the data from a larger geographical area (full model), i.e. South America, would increase our species predicted areas from Minas Gerais state. We understand that species doesn´t follow the geographical state boundary and it should "check-mate" our model predicting area, but we also believe that by using only presence data the overestimation of area is possible.

There are a plenty of uses for these potential species distribution maps, which includes greater scientific knowledge about biogeography, evolutionary ecology and conservation. The ecological niche modeling allows the extrapolation of this information to a geographical plan, being a tool of great practical value for many goals and decision support. This may be particularly important when choosing the right species for reforestation plans and to guarantee environmental suitability (HIDALGO et al., 2008; COELHO et al., 2016). This strategy may increase the success and feasibility of reforestation. The method can be associated to future reforestation plans when using scenarios of climate projection. Climate changes are an important deal to predicting impact in forests and for selecting suitable tree species to match future climates for afforestation and restoration (WANG et al., 2016). These maps can be also used to conservation purposes in case of rare or endangered species (MCCUNE, 2016), seed collection purposes (BREED et al., 2013) and when finding new populations (WILLIAMS et al., 2009).

CONCLUSIONS

Casearia sylvestris, *Copaifera langsdorffii* and *Tapirira guianensis* are widespread in the Minas Gerais state, covering a broad range of environmental conditions. *Croton floribundus* exhibited restricted geographic occurrence in the south of the state, showing a narrow environmental variation. Modeling ecological niche strategy is more affected when environmental and occurrence species data doesn´t have defined gradients. The algorithm Random Forest performed better for distribution modeling of all species, although the results varied for each metric and species. The species distribution predicted maps are powerful output to guide reforestation programs and biogeography studies in tropical areas.

ACKNOWLEDGMENTS

We thank the Energy Company of Minas Gerais (CEMIG) for the financial support of CEMIG GT-456 project.

REFERENCES

AQUINO, C.; BARBOSA, L. M. Classes sucessionais e síndromes de dispersão de espécies Arbóreas e arbustivas existentes em vegetação ciliar remanescente (conchal, sp), como subsídio para avaliar o potencial do fragmento como fonte de propágulos para Enriquecimento de áreas revegetadas no rio Mogi-Guaçu, SP. **Árvore** v. 33, p. 349-358, 2009.

ALLOUCHE, O.; TSOAR, A.; KADMON, R. 2006. Assessing the accuracy of species distribution models: prevalence, kappa and true skill statistic (TSS). **Journal of applied ecology**, v. 43, p. 1223-1232, 2006.

BREED M, F.; STEAD, M. G.; OTTEWELL, K. M; GARDNER, M. G.; LOWE, A. J. 2013. Which provenance and where? Seed sourcing strategies for revegetation in a changing environment. **Conservation Genetics**, v. 14, p. 1-10, 2013.

BREIMAN, L. 2001. Random Forests. **Machine Learning**, v. 45, p. 5–32, 2001.

CARVALHO, L. G; OLIVEIRA, M.; ALVES, M.; VIANELLO, R.; SEDIYAMA, G.; NETO, P.; DANTAS A. Clima. In: SCOLFORO, J.; CARVALHO, L. M.; OLIVEIRA, A. (Eds). **Zoneamento ecológico-econômico do estado de Minas Gerais: Componentes geofísico e biótico**, Lavras: Editora UFLA. 2008. 161p.

CHUN, J. H.; LEE, C. B. 2013. Assessing the Effects of Climate Change on the Geographic Distribution of Pinus densiflora in Korea using Ecological Niche Model. **KJAFM**, v. 14, p. 219-233, 2013.

CLUTER, D. R.; EDWARDS, T. C.; BEARD, K. H.; CLUTER, A.; HESS, K. T.; GIBSON, J. C. Random Forest for classification in ecology. **Ecology**, v. 11, p. 2783-2792, 2007.

COELHO, G. L.; TAVARES, L. M; GOMIDE, L. R. Modelagem preditiva de distribuição de espécies pioneiras no Estado de Minas Gerais. **Pesquisa Agropecuária Brasileira**, v. 51, p. 207-214, 2016.

ELITH, J. et al. Novel methods improve prediction of species' distributions from occurrence data. **Ecography**, v. 29, p.129-151, 2016.

FAUSETT, L. **Fundamentals of Neural Networks Architectures. Algorithms and Applications**, USA: Prentice Hall. 1994. 461p.

FRANK, E.; HALL, M. A.; WITTEN, I. The WEKA Workbench**. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques"**, Morgan Kaufmann, Fourth Edition, 2016.

FUKUDA, S.; BAETS, B.; WAEGEMAN, W.; VERWAEREN, J. Habitat prediction and knowledge extraction for spawning European grayling (*Thymallus thymallus* L.) using a broad range of species distribution models. **Environmental modelling & software**, v. 47, p. 1-6, 2013.

GALLIEN, L.; DOUZET, R.; PRATTE, S.; ZIMMERMANN, N. E.; THUILLER, W. Invasive species distribution models – how violating the equilibrium assumption can create new insights. **Global Ecology and Biogeography**, v. 21, p.1126–1136, 2012.

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 *Pinus sylvestris* L. in the Iberian Peninsula. **Ecological modelling**, v. 197, p. 383 – 393, 2006.

GRINNELL, J. Field tests of theories concerning distributional control. **The American Naturalist**, v. 51, p.115-128, 1917.

GWITIRA, I.; MURWIRA, A.; SHEKEDE, M. D.; MASOCHA, M.; CHAPANO, C. Precipitation of the warmest quarter and temperature of the warmest month are key to understanding the effect of climate change on plant species diversity in Southern African savannah. **African journal of ecology**, v. 52, p. 209-216, 2014.

HAMILTON, S. H.; POLLINO, C. A.; JAKEMAN, A. J. Habitat suitability modelling of rare species using Bayesian networks: Model evaluation under limited data. **Ecological modelling**, v. 299, p. 64-78, 2015

HAYKIN, S. **Neural Networks: A Comprehensive Foundation**, USA: Prentice Hall. 1994. 768p.

HIDALGO, P. J.; MARÍN, J. M.; QUIJADA, J.; MOREIRA, J. M. A spatial distribution model of cork oak (Quercus suber) in southwestern Spain: A suitable tool for reforestation. **Forest Ecology and Management**, v. 255, p. 25-34, 2008.

HIJMANS, R. J.; CAMERON, S. E.; PARRA, J. L.; JONES, P. G.; JARVIS, A. 2005. Very high resolution interpolated climate surfaces for global land areas. **International journal of climatology**, v. 25, p.1965-1978, 2005.

HIRZEL, A. H.; HAUSSER, J.; CHESSEL, D.; PERRIN, N. Ecological-niche factor analysis: how to compute habitat-suitability maps without absence data? **Ecology**, v. 83, p. 2027-2036, 2002.

Hutchinson, G. E. Concluding remarks. In: **Cold Spring Harbour Symposium on Quantitative Biology**, n. 22, p.415-427, 1957.

INSTITUTO ESTADUAL DE FLORESTAS - IEF. **Cobertura Vegetal de Minas Gerais**. Disponível em: <http://www.ief.mg.gov.br/florestas>. Accessed in 15 jan. 2015.

LIMA, J.; SANTANA, D.; NAPPO, M. Comportamento inicial de espécies na revegetação de mata de galeria na fazenda Mandaguari, em Indianópolis, MG. **Árvore**, v. 33, p. 685-694, 2009.

LOPES, S. F.; SCHIAVINI, I.; OLIVEIRA, A. P.; VALE, V. S. An Ecological Comparison of Floristic Composition in Seasonal Semideciduous Forest in Southeast Brazil: Implications for Conservation. **International Journal of Forestry Research**, v. 2012, p. 1-15, 2012.

LORENA, A.; JACINTHO, L.; SIQUEIRA, M.; DE GIOVANNI, R.; LOHMANN, G.; CARVALHO, A.; YAMAMOTO, M. Comparing machine learning classifiers in potential distribution modelling. **Expert Systems with Applications**, v. 38, p.5268-5275, 2011.

MACARTHUR, R. H. **Geographical Ecology: Patterns in the Distribution of Species.** New York: Princeton University Press. 1972. 269 p.

MAIORANO, L. et al. Building the niche through time: using 13,000 years of data to predict the effects of climate change on three tree species in Europe. **Global Ecology and Biogeography**, v. 22, p. 302-317, 2013.

MCCUNE, J. L. Species distribution models predict rare species occurences despite significant effects of landscape context. **Journal of Applied Ecology**, v. 53, n. 6, p. 1871-1879, 2016.

MEINERI, E.; DEVILLE, A. S.; GRÉMILLET, D.; GAUTHIER-CLERC, M.; BECHET, A. Combining correlative and mechanistic habitat suitability models to improve ecological compensation. **Biological Reviews**, v. 90, p. 314-329, 2015.

MOISEN, G. G.; FREEMAN, E. A.; BLACKARD, J. A.; FRESCINO, T. S.; ZIMMERMAN, N. E.; EDWARDS, T. C. Predicting tree species presence and basal área in Utah: a comparison of stochastic gradiente boosting, generalized additive models, and tree-based methods. **Ecological modelling**, v. 199, p. 176-187, 2006.

MONSERUD, R. A.; LEEMANS, R. The comparison of global vegetation maps. **Ecological modelling**, v. 62, p. 275-293, 1992.

MOTLOUNG, R. F.; ROBERTSON, M. P.; ROUGET, M.; WILSON, J. R. U. Forestry trial data can be used to evaluate climate-based species distribution models in predicting tree invasions. **NeoBiota**, v. 20, p. 31-48, 2014.

MOUNTON, A. M.; BAETS, B.; GOETHALS, P. L. M. Ecological relevance of performance criteria for species distribution models. **Ecological modelling**, v. 221, p. 1995-2002, 2010.

OLIVEIRA-FILHO, A. T.;RATTER, J. A. A study of the origin of central brazilian forests by the analysis of plant species distribution paterns. **Edinburgh Journal of Botany**, v. 2, p. 141-194, 1995.

OLIVEIRA-FILHO, A. T.; NETO, E. T; CARVALHO, W. A. C.; WERNECK, M.; BRINA, A. E.; VIDAL, C. V.; REZENDE, S. C.; PEREIRA, J. A. A. Análise florística do compartimento arbóreo de áreas de floresta atlântica sensu lato na região das bacias do leste. **Rodriguésia**, v. 56, p. 185-235, 2005.

OLIVEIRA-FILHO, A. T.; JARENKOW, J. A.; RODAL, M.J N. Floristic relationships of seasonally dry forests of eastern South America based on tree species distribution patterns. **SYSTEMATICS ASSOCIATION SPECIAL VOLUME**, v. 69, p. 159, 2006.

OZESMI, U.; TAN, C. O.; OZESMI, S. L.; ROBERTSON, R.J. Generalizability of artificial neural network models in ecological applications: Predicting nest occurrence and breeding success of the red-winged blackbird *Agelaius phoeniceus*. **Ecological Modelling**, v. 195, p. 94-104, 2006.

PHILLIPS, S. J.; DUDIK, M. Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. **Ecography**, v. 31, p. 161–175, 2008.

POUTEAU, R.; MEYER, J. Y.; TAPUTUARAI, R.; STOLL, B. Support Vector machines to map rare and endangered native plants in Pacific islands forests. **Ecological Informatics**, v. 9, p. 37-46, 2012.

PRASAD, A.M.; IVERSON, L.R., LIAW, A. Newer classification and regression tree techniques: bagging and random forests for ecological prediction. **Ecosystems**, v. 9, p. 181–199, 2006.

PULLIAM, H. R. On the relationship between niche and distribution. **Ecology letters**, v. 3, p. 349-361, 2000.

QUEIROZ, T. F.; BAUGHMAN, C.; BAUGHMAN, O.; GARA, M.; WILLIAMS, N. Species Distribution Modeling for Conservation of Rare, Edaphic Endemic Plants in White River alley, Nevada. **Natural Areas Journal**, v. 32, p. 149-158, 2012.

RAES, N. Partial versus Full Species Distribution Models. **Nat. Conserv.**, v. 10, p. 127 – 138, 2012.

RATTER, J. A.; BRIDGEWATER, S.; RIBEIRO, J. F. Analysis of the floristic composition of the Brazilian cerrado vegetation III: comparison of the woody vegetation of 376 areas. **Edinburgh Journal of botany**, v. 60, p. 57-109, 2003.

RODRIGUEZ-GALIANO, V.; SANCHEZ-CASTILLO, M.; CHICA-OLMO, M.; CHICA-RIVAS, M. Machine learning predictive models for mineral prospectivity: an evaluation of neural networks, Random Forest, regression trees and support vector machines. **Ore Geology Reviews**, v. 71, p. 804-818, 2015.

SCOLFORO, J. R. S.; CARVALHO, L. M. T. (Eds). **Mapeamento e Inventário da Flora Nativa e dos Reflorestamentos de Minas Gerais.** Lavras: Editora UFLA. 2008. 287 p.

SEBBENN, A. M.; CARVALHO, A. C. M.; FREITAS, M. L. M.; MORAES, S. M. B.; GAINO, A. P. S. C.; SILVA, J. M.; JOLIVET, C.; MORAES, M. L. T. Low levels of realized seed and pollen gene flow and strong spatial genetic structure in a small, isolated and fragmented population of the tropical tree *Copaifera langsdorffii* Desf. **Heredity**, v. 106, p. 134–145, 2011.

SEGURADO, P.; ARAÚJO, M. B. An evaluation of methods for modelling species distributions. **Journal of Biogeography**, v. 31, p. 1555-1568, 2004.

SILVESTRINI, M.; MCCAULEY, D. E.; ZUCCHI, M. I.; SANTOS, F. A. M. How do gap dynamics and colonization of a human disturbed area affect genetic diversity and structure of a pioneer tropical tree species?. **Forest Ecology and Management**, v. 344, p. 38-52, 2015.

SILVESTRINI, M.; SANTOS, F. A. M. Variation in the population structure between a natural and a human modified forest for a pioneer tropical tree species not restricted to large gaps. **Ecology and evolution**, v. 5, p. 2420-2432, 2015.

SOUZA, M. L.; SILVA, D. R. P; FANTECELLE, L. B.; LEMOS FILHO, J. P. Key factors affecting seed germination of *Copaifera langsdorffii*, a Neotropical tree. **Acta Botanica Brasilica**, v. 29, p. 473-477, 2015.

STOCKWELL, D. R. B.; PETERSON, A. T. Effects of sample size on accuracy of species distribution models. **Ecological modelling**, v. 148, p. 1-13, 2002.

TERSTEEGE, H.; ZAGT, R. Density and diversity. **Nature**, v. 417, p. 689–699, 2002.

VACLAVIK, T.; MEENTEMEYER, R. K. Invasive species distribution modeling (iSDM): Are absence data and dispersal constraints needed to predict actual distributions? **Ecological modelling**, v. 220, p. 3248-3258, 2009.

VALIO, I. F. M.; SCARPA, F. M. Germination of seeds of tropical pioneer species under controlled and natural conditions. **Brazilian Journal of Botany**, v. 24, p. 79-84, 2001.

WANG, T.; WANG, G.; INNES, J.; NITSCHKE, C.; KANG, H. Climatic niche models and their consensus projections for future climates for four major forest tree species in the Asia–Pacific region. **Forest Ecology and Management**, v. 360, p. 357-366, 2016.

WHITTAKER, R. H. Gradient analysis of vegetation. **Biological reviews**, v. 42, p. 207-264, 1967.

WILLIAMS, J. N.; SEO, C.; THORNE, J.; NELSON, J. K.; ERWIN, S.; O'BRIEN, J. M.; SCHWARTZ. Using species distribution models to predict new occurrences for rare plants. **Diversity and Distributions**, v. 15, p. 565–576, 2009.