MÔNICA CANAAN CARVALHO

MODELAGEM PREDITIVA DA DISTRIBUIÇÃO POTENCIAL DE ESPÉCIES ARBÓREAS NA BACIA HIDROGRÁFICA DO RIO GRANDE, MG

LAVRAS – MG
2015

MÔNICA CANAAN CARVALHO

MODELAGEM PREDITIVA DA DISTRIBUIÇÃO POTENCIAL DE ESPÉCIES ARBÓREAS NA BACIA HIDROGRÁFICA DO RIO GRANDE, MG

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Ciências Florestais, área de concentração em Manejo Florestal, para a obtenção do título de Mestre.

Orientador

Dr. Lucas Rezende Gomide

LAVRAS – MG

2015

Ficha catalográfica elaborada pelo Sistema de Geração de Ficha Catalográfica da Biblioteca Universitária da UFLA, com dados informados pelo(a) próprio(a) autor(a).

Carvalho, Mônica Canaan.
Modelagem preditiva da distribuição potencial de espécies arbóreas na bacia hidrográfica do rio Grande, MG / Mônica Canaan Carvalho. – Lavras : UFLA, 2015.
87 p. : il.

Dissertação (mestrado acadêmico)–Universidade Federal de Lavras, 2015.
Orientador(a): Lucas Rezende Gomide.
Bibliografia.

1. Árvore de decisão. 2. Fitogeografia. 3. *Maxent*. 4. Redes neurais artificiais. 5. *Random Forest*. I. Universidade Federal de Lavras. II. Título.

MÔNICA CANAAN CARVALHO

MODELAGEM PREDITIVA DA DISTRIBUIÇÃO POTENCIAL DE ESPÉCIES ARBÓREAS NA BACIA HIDROGRÁFICA DO RIO GRANDE, MG

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Ciências Florestais, área de concentração em Manejo Florestal, para a obtenção do título de Mestre.

APROVADA em 20 de fevereiro de 2015.

Dra. Gleyce Campos Dutra	UFVJM
Dr. Luciano Teixeira de Oliveira	UFLA

Dr. Lucas Rezende Gomide
Orientador

LAVRAS – MG

2015

AGRADECIMENTOS

Sempre, em primeiro lugar, a Deus. Pelo presente da vida e todas as dádivas concedidas.

Ao meu pai, Jacinto, por desempenhar de forma admirável seu papel em minha vida. Por todo amor, respeito, apoio e, principalmente, por ser o meu exemplo de caráter e profissionalismo. À minha mãe, Dodora. Por razões indescritíveis que brevemente tento citar aqui. Pelo amor ilimitado, compreensão, amizade, companheirismo e ensinamentos. Por saber levar a vida de uma maneira leve e por me fazer feliz com pequenas coisas. À Juliana, por ser minha irmã mais velha e ter me dado os melhores exemplos a serem seguidos. Por ter estado ao meu lado todos os anos da minha vida.

Ao meu orientador, Lucas R. Gomide, pela convivência, paciência e orientação.

Aos amigos de longa data. Àqueles muitas vezes negligenciados pela falta de tempo. Agradeço pelas melhores risadas, pela intimidade compartilhada, conselhos e compreensão. Em especial Leandro, Priscilla, Aline e Júlia.

Aos amigos da UFLA e principalmente do Lemaf, por fazerem da UFLA um lugar agradável para se passar tanto tempo. Pelas festas e trabalhos, pelos ensinamentos e conselhos, pelas conversas na cozinha. Ao Bruno (Toró) e Henrique (Moreira) por serem meus "irmãos" nestes dois anos. À Nathália (Natita), pela grande amizade e ensinamentos no ArcGis. Aos companheiros do laboratório de P.O (Matheus, Nathália e Tayrine) pela convivência diária e conhecimento compartilhado. A todos do laboratório do Passarinho (Carol's, Lizi, Guilherme) por sempre me receberem de portas abertas. À Thaisa e Thaís pelo companheirismo ao longo destes dois anos. Ao Luciano (Bodinho) pelos ensinamentos e mensagens positivas.

Aos professores Rubens Manoel dos Santos, Luiz Marcelo Tavares, Bruno Groenner, William Lacerda e José Márcio. Pelas ideias, explicações e contribuições essenciais ao trabalho. Às funcionárias Beth, Gláucia, Raísa e Simone (*in memoriam*) pelos favores prestados, desabafos e conversas agradáveis.

Agradeço à CEMIG, pelo apoio financeiro na coleta dos dados florestais na bacia do Rio Grande. Ao Inventário Florestal de Minas Gerais (IFMG), em parceria com a UFLA, pela disponibilização dos dados dos 169 fragmentos inventariados.

RESUMO

O presente estudo tem como objetivo principal comparar o desempenho de quatro algoritmos de aprendizagem de máquina (árvore de decisão, *Random Forest*, redes neurais artificiais e *Maxent*) na modelagem da distribuição de espécies arbóreas em Minas Gerais. Para este fim, utilizou-se os dados proveniente de 197 fragmentos inventariados em Minas Gerais e 25 variáveis relacionadas ao clima, topografia e solo. A capacidade preditiva dos algoritmos foi avaliada pela métrica *Area Under the Curve* (AUC), obtida através da validação cruzada (10%) e por um conjunto de dados de teste independente (30%). Os resultados obtidos pela validação cruzada foram testados pelo teste estatístico T-pareado, com 95% de confiança. Foram avaliados dois conjuntos de atributos abióticos na modelagem, sendo o primeiro formado por todas as 25 variáveis abióticas disponíveis e o segundo, pelas 10 variáveis com maior ganho de informação. As espécies com grande abundância e ampla distribuição selecionadas no estado de Minas Gerais são *Casearia sylvestris, Copaifera langsdorffii, Croton floribundis* e *Tapirira guianensis*. Para todas estas espécies, os algoritmos não apresentaram melhora significativa em seu desempenho quando modelados com os atributos pré-selecionados. De acordo com a validação cruzada, a maioria das espécies não apresentou diferença significativa entre a capacidade preditiva dos algoritmos árvore de decisão, *Random Forest* e redes neurais. Entretanto, o *Random Forest* demonstrou AUC numericamente superior na maioria dos casos. Já para a validação com o conjunto de teste, o *Random Forest* superou todos os algoritmos, inclusive o *Maxent*, para todas as espécies. A área predita pelo *Random Forest* foi menor do que a área predita pelo *Maxent* quando utilizado o limiar mínimo de adequabilidade ambiental presente no conjunto de treinamento; e menor quando a adequabilidade ambiental é reclassificada adotando o limiar 0,5. De acordo com as métricas de avaliação e os mapas obtidos para cada espécie, o *Random Forest* se mostrou um algoritmo com grande potencial para a modelagem da distribuição de espécies.

Palavras-chave: Fitogeografia. *Random Forest*. Árvore de decisão. Redes neurais Artificiais. *Maxent*.

ABSTRACT

The present study had the main objective of comparing the performance of four machine learning algorithms (Decision Tree, Random Forest, Artificial Neural Networks and Maxent) in modeling the distribution of tree species in the state of Minas Gerais, Brazil. To this end, we used data from 197 inventoried fragments in Minas Gerais and 25 variables related to climate, topography and soil. The predictive capacity of the algorithms was evaluated by measuring the Area Under the Curve (AUC) obtained by cross-validation (10%) and by a set of independent test data (30%). The results obtained by the cross-validation were tested by the T-matched statistical test, with 95% confidence. We evaluated two sets of abiotic attributes in the modeling, the first was formed by all 25 abiotic variables available and the second by 10 variables with the highest information gain. The species *Casearia sylvestris*, *Copaifera langsdorffii*, *Croton floribundis* and *Tapirira guianensis* were selected according to their high abundance and wide distribution in the state. For all these species, the algorithms showed no significant improvement in performance when modeled. According to the cross-validation, most species showed no significant difference between the predictive capacity of the Decision tree, Random Forest and Artificial Neural Networks. However, Random Forest demonstrated numerically superior AUC in most cases. The Random Forest was the superior of all tested algorithms, including the Maxent, when the validation set was run. The area predicted by the Random Forest was smaller than that predicted by Maxent when using the minimum limit of environmental suitability present in the training set; and smaller when the environmental suitability is reclassified, adopting the limit of 0.5. According to the evaluation metrics and maps obtained for each species, the Random Forest algorithm showed great potential for modeling species distributions.

Keywords: Phytogeography. Random Forest. Decision Tree. Artificial Neural Networks. Maxent.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

A existência humana tem colocado em risco a distribuição de espécies arbóreas no planeta. A ocorrência e a perpetuação dessas espécies vêm sendo ameaçadas pela perda de seus *habitats* naturais ocasionada em sua grande maioria pela ação do homem.

A perda de ecossistemas florestais afeta negativamente diversos aspectos ambientais, como a biodiversidade da fauna e flora, a qualidade do solo, ciclo hidrológico, microclima, etc. Logo, surge a necessidade crescente de proteção e restauração desses ecossistemas, demandando novas tecnologias, capazes de entender as relações entre as características do meio ambiente e a ocorrência de espécies. O entendimento destas relações é imprescindível para se realizar, com sucesso, a restauração florestal de áreas modificadas pelo homem.

Técnicas que tentam entender e replicar a relação entre variáveis bióticas e abióticas do meio ambiente e a ocorrência de determinada espécie, estão inseridas no campo da modelagem preditiva da distribuição de espécies. O processo de modelagem relaciona dados de ocorrência de espécies (coordenadas geográficas) com variáveis ambientais (obtidas através de imagens de Sensoriamento Remoto e dados de um SIG), conseguindo predizer ambientes potencialmente adequados, onde, em teoria, uma população possa se manter viável. O resultado da modelagem é então projetado em um mapa, indicando as regiões com distribuição potencial da espécie (ANDERSON; LEW; PETERSON, 2003).

Os modelos de predição de *habitat* constituem uma ferramenta prática e precisa em planos de revitalização de áreas desflorestadas. A partir de levantamentos fitossociológicos e de variáveis ambientais da região de interesse, é possível indicar qual espécie estará apta para re(ocupar) determinado local.

O desenvolvimento do Sistema de Informação Geográfica (SIG) e Sensoriamento Remoto e a evolução de técnicas computacionais foram substanciais para o aprimoramento e aplicabilidade da modelagem da distribuição de espécies. Atualmente existem diversas fontes de dados ambientais e de ocorrência, técnicas de processamento dos dados e modelos disponíveis, tornando o processo de modelagem cada vez mais preciso.

Diversos cuidados devem ser tomados ao se modelar a distribuição de uma espécie. A escolha das variáveis ambientais deve ser baseada nos aspectos ecológicos da espécie, podendo ser empregada uma seleção das variáveis principais a fim de diminuir ruídos na modelagem. Os registros de ocorrência devem ser precisos e representativos da distribuição real da espécie. Além da qualidade, é preciso obter quantidade de dados suficiente para treinamento e teste dos modelos gerados. Os algoritmos devem ser escolhidos de acordo com o tipo e complexidade dos dados.

Estudos comparativos entre diferentes algoritmos demonstram que o desempenho dos métodos varia para cada espécie modelada, não sendo possível identificar a técnica mais precisa para a modelagem da distribuição de qualquer espécie (ELITH et al., 2006; WILLIAMS et al., 2009). Portanto, para se obter um bom desempenho na predição de *habitats*, faz-se necessário, além da qualidade dos dados utilizados, a comparação entre diferentes métodos de modelagem.

Dentre os diversos tipos de modelos disponíveis para a modelagem preditiva de distribuição de espécies, os algoritmos da área de aprendizagem de máquina vêm surgindo como uma boa alternativa, diante do bom desempenho obtido em diversas pesquisas. Esses algoritmos são capazes de sintetizar funções de regressão e classificação baseados nos dados disponíveis e extrair conhecimento de dados previamente observados, fazendo predições com base em novos dados. Dentre os métodos de aprendizagem de máquina, destacam-se

a árvore de decisão, redes neurais artificiais, algoritmo genético (GARP), *Suport Vector Machine, Random Forest* e *Maxent.*

Assim, diante do disposto, o presente estudo tem como objetivo principal comparar o desempenho de quatro algoritmos da área de aprendizagem de máquina (árvore de decisão, *Random Forest*, redes neurais e *Maxent*) na modelagem preditiva da distribuição potencial de espécies arbóreas em Minas Gerais. Para isso, pretende-se:

a) Selecionar quatro espécies arbóreas abundantes, com ampla distribuição e potencial para revitalização de matas ciliares em Minas Gerais;
b) Compreender o comportamento da distribuição das espécies em relação às variáveis abióticas e selecionar as mais representativas segundo o método ganho de informação;
c) Verificar se existe diferença (em termos de AUC) no desempenho dos modelos ao selecionar os atributos abióticos pelo método Ganho de informação;
d) Avaliar a capacidade preditiva dos modelos, em termos de AUC, através de validação cruzada e com um conjunto de teste independente;
e) Selecionar o modelo com maior capacidade preditiva para cada espécie e aplicá-lo na bacia hidrográfica do rio Grande;
f) Comparar a distribuição potencial predita pelo algoritmo mais preciso de presença e ausência com a distribuição gerada pelo *Maxent*, utilizando dois diferentes limiares para sua reclassificação.

2 REFERENCIAL TEÓRICO

2.1 Fitogeografia

Com o intuito de fornecer respostas sobre a ocorrência da biodiversidade e sobre os padrões de distribuição desta na superfície da Terra, surgiu o estudo da biogeografia. A fitogeografia é um ramo da biogeografia que estuda a distribuição geográfica do reino vegetal, seja ao nível de espécie ou até mesmo de grandes formações vegetais.

De acordo com Rizzini (1997), a fitogeografia integra o estudo da estrutura das comunidades vegetais, fatores ambientais, padrões de distribuição, migração e dispersão dos vegetais, abrangendo diversas disciplinas como a botânica, ecologia, fitossociologia e sistemática.

Por meio do estudo dos padrões de ocorrência relacionados à diferentes variáveis, é possível identificar fatores que condicionam ou limitam a existência de determinada espécie (ou grupo) em locais específicos. Estes estudos podem ser realizados em diferentes escalas (local, regional ou global), e demandam grande variedade de metodologias, envolvendo diversas áreas do conhecimento científico.

A fitogeografia estabelece importantes relações entre variáveis abióticas do ambiente (como temperatura, altitude, precipitação, tipo de solo, etc.) e a ocorrência de determinado bioma, fitofisionomia, família, gênero ou espécie. Com o entendimento destas relações é possível delimitar, geograficamente, áreas de ocorrência real ou potencial de determinada espécie ou formação vegetal.

De acordo com o Instituto Brasileiro de Geografia e Estatística - IBGE (2004), o estado de Minas Gerais está inserido no domínio de três biomas brasileiros: Cerrado, Mata Atlântica e Caatinga, este último em menores porções de terra. O bioma Cerrado abrange 57% da extensão territorial do estado,

localizado na região centro-ocidental, o bioma Atlântico ocupa cerca de 41% do estado mineiro e encontra-se distribuído na porção oriental. Já o bioma Caatinga, está restrito ao norte do estado e triângulo mineiro, abrangendo apenas 2% do estado de Minas Gerais.

Devido à heterogeneidade ambiental compreendida por estes biomas, principalmente o Cerrado e Atlântico, existem espécies melhores adaptadas às certas regiões, o que indica a existência de padrões fitogeográficos regionais dentro dos biomas e regiões floristicamente distintas, originando as diferentes fitofisionomias.

O bioma Cerrado compreende uma faixa contínua nos estados de Minas Gerais, Mato Grosso do Sul, Goiás, Tocantins, Bahia, Maranhão e Piauí e ainda algumas áreas disjuntas por outros estados, como em São Paulo. Abrange um amplo espectro de condições ambientais, sendo a variação da temperatura média anual de 18 a 28 °C e a oscilação da precipitação anual entre 800 e 2.000 mm, com intensa escassez de chuva na estação seca (abril - setembro). Os tipos de solos são também muito diversos, sendo em grande maioria, distróficos, com baixa disponibilidade de cálcio e magnésio, alta concentração de alumínio e bem drenados (DURIGAN et al., 2003). Além de um extenso gradiente latitudinal, o Cerrado varia em altitude do nível do mar (estado do Maranhão) até cerca de 1.600 m nas mais elevadas áreas do Planalto Central e Cadeia do Espinhaço. Sua vegetação é bastante variada, apresentando fisionomias que vão desde formações florestais com 12 a 15 m de dossel à savânicas e campestres, com arbustos esparsos (RATTER; RIBEIRO; BRIDGEWATER, 1997).

O bioma Cerrado compreende grande extensão territorial em Minas Gerais, retratando um gradiente fisionômico que abrange as áreas de Campo, Campo Rupestre, Campo Cerrado, Cerrado *Sensu Stricto*, Cerradão e Vereda (SCOLFORO; MELLO; SILVA, 2008).

Heringer et al. (1977) acreditam ser o Cerrado, o progenitor dos biomas Mata Atlântica e Amazônia, devido à grande quantidade de espécies em comum que este bioma apresenta com os outros. Devido à sua heterogeneidade ambiental e sua localização entre diferentes biomas, o Cerrado apresenta alta diversidade de espécies. Ratter et al. (1996), ao estudar 98 fragmentos dentro do bioma Cerrado, encontrou 534 espécies, das quais 30% ocorreram em um único fragmento; 5% ocorreram em mais de 49 locais e nenhuma esteve presente em todos os fragmentos.

Ainda dentro deste bioma, encontram-se as matas de galeria situadas ao longo dos cursos d'água. Estas matas estão presentes em solos mais ricos, formando grandes áreas de florestas mesofíticas, nas quais estão presente maior número de espécies arbóreas. Estas áreas geralmente apresentam uma maior tendência de espécies endêmicas e dominantes e menor diversidade florística quando comparadas com áreas de Cerrado típico.

Ratter, Ribeiro e Bridgewater (2003), estudando a vegetação arbórea em 376 áreas de Cerrado, destacam que fatores como a deficiência hídrica do solo e temperatura média, que tendem a aumentar na direção sudeste-nordeste, estão relacionados com a distribuição das espécies dentro do bioma. Ainda de acordo com estes autores, o tipo de solo é o fator que mais determina a diferenciação florística e estrutural entre comunidades do Cerrado.

O bioma Mata Atlântica vem sendo duramente fragmentado ao longo de anos, desde a descoberta do Brasil. Atualmente está presente em 17 estados brasileiros através de manchas de vegetação e *hotspots*. A Mata Atlântica engloba vários ecossistemas florestais, associada aos ecossistemas costeiros, enseadas, fozes de rios, baías, restingas, baixadas arenosas, florestas mistas de araucárias, campos de altitude e rupestres. Dentro do estado de Minas Gerais, está localizada principalmente nas regiões leste e sul, sendo preferenciais

ambientes de maiores altitudes, associadas principalmente às fitofisionomias Floresta Estacional Semidecidual e Ombrófila.

De acordo com Oliveira Filho e Ratter (2000), sua vegetação é altamente influenciada pela distância do oceano, seguido do regime de distribuição de chuvas, altitude e da duração da estação de seca. De maneira geral, os solos da Mata Atlântica são muito lixiviados, ácidos e distróficos, porém mesmo em solos com baixa fertilidade não é observado sintomas de deficiência nutricional, devido à decomposição da matéria orgânica proveniente da serrapilheira (SILVA et al., 2007).

Em Minas Gerais, o domínio atlântico está situado, em sua grande maioria, em latossolos, principalmente as Florestas Ombrófilas. Já as Florestas Estacionais Semideciduais encontram-se distribuídas entre argissolos, cambissolos, neossolos e latossolos (SCOLFORO; MELLO; SILVA, 2008).

Segundo Silva et al. (2008b), ao estudarem 25 fragmentos de Florestas Estacionais Semideciduais e oito de Florestas Ombrófilas distribuídos em território mineiro, as cinco espécies mais abundantes na fitofisionomia Semidecidual são (ordem decrescente): *Eremanthus incanus* (Less.) Less., *Copaifera langsdorffii* Desf., *Myrcia splendens* (Sw.) DC., *Tapirira obtusa* (Benth.) J.D. Mitch. e *Tapirira guianensis* Aubl. Para a Floresta Ombrófila as cinco espécies com maior abundância dentro dos fragmentos amostrados são *Psychotria vellosiana* Benth., *Myrsine umbellata* Mart., *Myrceugenia alpigena* (DC.) Landrum, *Myrcia splendens* (Sw.) DC. e *Podocarpus sellowi* Klotzch ex Endl.

O bioma Caatinga ocorre em pequenas porções no território mineiro situadas principalmente no norte e em algumas áreas descontínuas no triângulo mineiro. O clima predominante é de caráter semiárido quente, com altas temperaturas, precipitações escassas e irregulares, com sete a 10 meses de estação seca. A temperatura média anual varia entre 24 e 26 °C e a precipitação

anual fica entre 250 e 1.000 mm. Os solos, de maneira geral, são rasos, argilosos e rochosos ou profundos e arenosos (LIMA, 1981).

Em Minas Gerais, o domínio da Caatinga está representado principalmente pela Floresta Estacional Decidual. De acordo com Mello, Scolforo e Carvalho (2008), estas se estratificam em dois grupos: o primeiro formado pelas matas do Jaíba e Jequitinhonha e o segundo formado pelas matas ciliares das respectivas sub-bacias. De acordo com Silva et al. (2008a), as cinco espécies mais abundantes desta fitofisionomia no território mineiro são: *Handroanthus chrysotrichus* (Mart. ex A. DC.) Mattos, *Poincianella pluviosa* (DC.) L. P. Queiroz, *Anadenanthera colubrina* (Vell.) Brenan, *Myracrodruon urundeuva* Allemão e *Machaerium acutifolium* Vogel. As espécies com maior valor de importância típicas do bioma Caatinga ou Matas Secas, segundo Dutra (2009), são *Schinopsis brasiliensis* Engl. e *Annona leptopetala* (R. E. Fr.) H. Rainer.

Nos dias atuais, é válido ressaltar a importância da área de tecnologia da informação para o aprimoramento dos estudos fitogeográficos. Com o advento dos Sistemas de Informações Geográficas (SIG's) e o avanço das técnicas de Sensoriamento Remoto, se faz possível a utilização de diversas ferramentas para o entendimento dos padrões de distribuição de espécies e sua relação com as características do ambiente. Uma ferramenta que vem sendo amplamente empregada em estudos fitogeográficos é a modelagem de distribuição de espécies. Sua utilização vai além da delimitação de locais de ocorrência de determinada espécie. Com o auxílio da modelagem, é possível identificar áreas potenciais para a sobrevivência e reprodução da espécie, bem como áreas endêmicas, prioritárias para a conservação.

2.2 Modelagem Preditiva da Distribuição Potencial de Espécies

Seja para fins de conservação, restauração ou produção, a questão de como o reino vegetal e animal estão distribuídos na Terra tem sido fonte de estudos para diversos pesquisadores. É certo que fatores climáticos, físicos e biológicos são responsáveis, em diferentes escalas, pela distribuição das espécies no planeta. A partir desta informação, pesquisadores vêm utilizando estes fatores como variáveis de entrada em modelos matemáticos, para a predição de locais que satisfaçam às necessidades da espécie. Este tipo de modelagem é denominado *Modelagem Preditiva de Habitat* ou *Modelagem Preditiva de Distribuição de Espécies* (GIANNINI et al., 2012).

O processo de modelagem de distribuição de espécies tem como objetivo encontrar relações não aleatórias, entre as variáveis ambientais, relevantes para a espécie e seus dados de ocorrência. A modelagem preditiva de *habitat* combina dados de ocorrência de espécies (coordenadas geográficas) com variáveis ambientais e ecológicas (como por exemplo: temperatura, precipitação, altitude, tipo de solo, índices de vegetação, etc.) para realizar a predição de ambientes adequados, onde, em teoria, uma população possa se manter viável. O resultado da modelagem é então projetado em um mapa, indicando as regiões com distribuição potencial da espécie (ANDERSON; LEW; PETERSON, 2003).

A modelagem preditiva de distribuição de espécies, para muitos autores, é conhecida também como modelagem de nicho ecológico. Esta nomeação é devido ao fato de que, independente do algoritmo utilizado, a modelagem é normalmente baseada no conceito de nicho ecológico de uma espécie.

De acordo com Hutchinson (1957), nicho é a soma de todos os fatores ambientais que atuam sobre o organismo e consiste em um *hiperespaço* composto de *n* fatores limitantes (por exemplo, radiação, temperatura e recursos alimentares) à sobrevivência do ser vivo e sua amplitude de tolerância a estes

fatores. Representa o espaço abrangido pela faixa de variação dos fatores ambientais ao qual a espécie é capaz de sobreviver e se reproduzir (GIANNINI et al., 2012).

De acordo com Soberón e Peterson (2005), a distribuição espacial de uma espécie está relacionada a quatro fatores principais:

a) Fatores abióticos – condições ambientais que limitam a capacidade de sobrevivência e reprodução da espécie em determinada região (por exemplo: altitude, inclinação do terreno, fertilidade do solo, pluviosidade, temperatura, etc.);
b) Fatores bióticos – Conjunto de interações com outras espécies que influenciam na sobrevivência da espécie em estudo, como competição, parasitismo, predação, mutualismo, etc;
c) Fatores de acessibilidade – Relacionados à capacidade de dispersão, que refletem quais locais são acessíveis para indivíduos de uma determinada espécie (importante para distinguir distribuição atual e distribuição potencial);
d) Fatores evolucionários – Relacionados com a capacidade de adaptação às novas condições (plasticidade da espécie).

Utilizando apenas os fatores abióticos como condicionadores de um *habitat* viável, obtêm-se locais satisfatórios para a espécie (potencial) e não exatamente locais ocupados pela espécie (real) (SÓBERON; PETERSON, 2005). A principal razão para a utilização desses fatores em detrimento dos demais é a dificuldade de se obter variáveis que representem condições bióticas, cuja interpretação é complexa.

Devido à falta de variáveis ecológicas, o conceito “modelagem de nicho ecológico” tem sido questionado, visto que a distribuição das espécies é um

resultado complexo da ecologia, evolução da espécie e outros fatores biológicos, como sua história evolutiva e sua capacidade de dispersão. De acordo com Pulliam (2000), é possível encontrar indivíduos em locais onde as condições não são inteiramente adequadas à sobrevivência e à reprodução da espécie e, por outro lado, apresentando limitações de sua dispersão, não encontrá-las em locais adequados. Assim, a maioria dos autores que utilizam apenas variáveis abióticas passa a utilizar o termo modelagem da distribuição potencial e não de nicho.

Para o treinamento e validação dos modelos de distribuição de espécies, são necessários, basicamente, dois tipos de dados: 1) dados de ocorrência (coordenadas geográficas dos registros de ocorrência da espécie) e 2) dados abióticos (mapas temáticos que sintetizam as características ambientais mais relevantes da área de estudo).

Os registros de ocorrência da espécie (presença e ausência) são representados por pontos georreferenciados (latitude e longitude) e são utilizados para ajustar e validar os modelos de distribuição de espécies (GUISAN; ZIMMERMANN, 2000). Os dados de ocorrência da espécie podem ser obtidos em coletas de campo, em acervos de coleções de herbários e museus e em bases de dados de biodiversidade disponíveis na internet.

Os dados oriundos de coleções biológicas, museus e até mesmo base de dados *online* podem representar uma fonte de erro para a modelagem, uma vez que a maioria apresenta imprecisão na localização do ponto de coleta. No caso de coleta em campo, deve-se evitar o viés na escolha das áreas a serem amostradas, sem privilegiar locais de fácil acesso ou locais onde já se tem a certeza da existência da espécie, além de incertezas na identificação da espécie.

A partir dos dados de ocorrência, os modelos utilizados na distribuição de espécies podem ser categorizados em *presence-only* (somente dados de presença) e presença/ausência. A maioria dos métodos *presence-only* operam construindo um espaço de *n*-dimensões (*n* variáveis ambientais) de acordo com a

ocorrência da espécie. Locais fora do espaço construído são então classificados como ausência. Já as técnicas que trabalham com presença e ausência, classificam novos locais de acordo com o conhecimento adquirido pelos exemplos de presença e ausência utilizados no ajuste do modelo.

Os dados de ausência da espécie são mais difíceis de obter precisamente. O fato de uma espécie não se encontrar em determinado ambiente nem sempre pode ser considerado como uma falta de recursos necessários a sua sobrevivência. A espécie pode passar despercebida, devido sua baixa detectabilidade (raridade), ou estar ausente devido a razões históricas e barreiras para sua dispersão (PHILLIPS; ANDERSON; SCHAPIRE, 2006).

Devido à dificuldade em se obter dados de ausência, algoritmos que utilizam apenas dados de presença têm sido bastante utilizados. Os modelos *presence-only* mais consolidados em pesquisas ecológicas são DOMAIN (CARPENTER; GILLISON; WINTER, 1993; ELITH et al., 2006), BIOCLIM (BEAUMONT; HUGHES, 2002; BRERETON; BENNETT; MANSERGH, 1995; BUSBY, 1991) e ENFA (HIRZEL et al., 2002).

A utilização de métodos estatísticos na modelagem, como regressão logística, modelos lineares generalizados e técnicas derivadas da inteligência artificial, carece da existência das duas categorias de dados: presença e ausência. Logo, para suprir a falta dos registros de ausência, muitos algoritmos usam dados de pseudoausência para ajustar os modelos. Estes dados são pontos escolhidos aleatoriamente na área de estudo e usados como ausências durante o processo de modelagem.

De acordo com Engler, Guisan e Rechsteiner (2004), a definição dos pontos de pseudoausência deveria seguir alguma estratégia e não ser feita aleatoriamente. Devem-se evitar pontos com valores das características ambientais muito próximos aos dos pontos de presença, reduzindo possíveis ruídos nos dados de entrada. Vanderwal et al. (2009) afirmaram que a qualidade

do modelo é menor, quando os pontos de pseudoausência são definidos em regiões muito restritas ou muito amplas, em relação às presenças registradas.

Trabalhar apenas com dados de presença e ausência para a predição de *habitats* adequados gera grande incerteza quanto à capacidade da espécie em ocupar determinada região. Uma espécie pode estar presente em um local, porém em pequeno número de indivíduos, o que indica que as condições ambientais da região não são tão favoráveis à sua sobrevivência e perpetuação Os dados de abundância da espécie (número de indivíduos) são essenciais quando se deseja trabalhar com a capacidade de ocupação da espécie. São um complemento aos dados de ocorrência, indicando o sucesso da adaptação da espécie em um local específico. Diversas técnicas regressivas são empregadas para a predição da riqueza de espécies, porém a desvantagem é de que o número de espécies é variável para locais ambientalmente semelhantes, sendo altamente relacionado aos fatores bióticos e de dispersão, não incluídos na maioria dos modelos.

O outro tipo de dado necessário para a modelagem da distribuição de espécies são as variáveis ambientais. A coleta de dados ambientais para a modelagem evoluiu na década de 90, quando as imagens de sensoriamento remoto tornaram-se amplamente acessíveis. Concomitante, com o crescimento dos Sistemas de Informação Geográfica e técnicas geoestatísticas, houve uma expansão na disponibilidade de dados ambientais e, consequentemente, no uso de modelos de predição de *habitat*.

Os dados ambientais ou abióticos são geralmente representados por arquivos em formato raster. Estes arquivos consistem em bancos de dados georreferenciados formados por uma matriz de células que contém os valores da variável em questão. O tamanho das células é o que determina a resolução espacial do raster, sendo que, quanto maior o tamanho da célula menor a resolução e vice-versa (GIANNINI et al., 2012).

A resolução espacial dos mapas deve ser compatível com as necessidades da pesquisa, e deve-se ter o cuidado de utilizar dados relevantes e em número suficiente para garantir a geração de modelos, que representem as condições ambientais ótimas para a sobrevivência e reprodução da espécie. De acordo com Gárzon et al. (2006), estudos regionais devem ter sua resolução variando entre 1 a 10 km.

Quando se trabalha com variáveis ambientais, provenientes de diferentes bancos de dados e diferentes resoluções, é necessário realizar uma padronização destes fatores em relação à resolução espacial. Geralmente, esta padronização é realizada de acordo com a menor resolução espacial (*grids* maiores) encontrada na base de dados de variáveis ambientais. No entanto, isto não é uma regra. Ao realizar a padronização pela menor resolução, perde-se informação daquelas variáveis com maior resolução. O aconselhável é não utilizar variáveis ambientais muito destoantes quanto ao tamanho das células do raster.

Diante da grande disponibilidade de dados ambientais, muitos métodos são empregados para comparar o desempenho das diferentes variáveis ambientais na predição de *habitats*, bem como a existência de preditores correlacionados. Portanto, faz-se necessário o uso de uma pré-seleção de preditores ambientais utilizando, por exemplo, diferentes tipos de análise multivariada, teste Jackknife, Árvore de Regressão, redes neurais artificiais, Partição da Variância, meta-heurísticas, dentre outras técnicas.

Devido às características dos métodos de aprendizagem de máquina, como a capacidade de trabalhar com preditores correlacionados, de processar relações não lineares entre os preditores e de processar dados complexos e com ruídos, estes vêm sendo amplamente empregados na modelagem da distribuição de espécies, visto que dados ecológicos apresentam estas dificuldades (GARZÓN et al., 2006).

As técnicas de aprendizagem de máquina (AM) são capazes de aprender a partir de exemplos, extraindo conhecimento dos dados previamente observados e gerar predições com base em novos dados (MITCHELL, 1997). O tipo de inferência lógica utilizada pelos algoritmos de AM é a indução, ou seja, o raciocínio originado em um conceito específico é generalizado para o restante dos dados (RODRIGUES, 2012).

Elith et al. (2006), comparando 16 métodos de modelagem da distribuição de espécies, obtiveram maior eficácia dos modelos baseados nos métodos de aprendizagem de máquina em relação aos modelos bem estabelecidos, como modelos aditivos generalizados (GAM´s) e BIOCLIM. Segurado e Araújo (2004) também encontraram resultados que confirmam o potencial dos métodos de aprendizado de máquina na modelagem da distribuição de espécies, ao comparar sete métodos em 44 espécies de anfíbios.

Dentre os métodos de aprendizagem de máquina mais utilizados na modelagem da distribuição de espécies estão a árvore de decisão, redes neurais artificiais, algoritmo genético, *Suport Vector Machine* (máquinas de vetores de suporte), *Random Forest* (florestas aleatórias) e *Maxent*.

2.2.1 Árvore de decisão

De acordo com Garzón et al. (2006), dentre as técnicas de aprendizagem de máquina, a árvore de decisão foi a primeira a ser empregada na modelagem da distribuição de espécies. Árvore de regressão (atributos contínuos) ou classificação (atributos nominais) são técnicas de aprendizagem de máquina que utilizam a estratégia de *dividir-para-conquistar*, decompondo um problema maior em subproblemas mais simples, de forma recursiva. Dentre as principais vantagens dessa técnica, está a capacidade de trabalhar as relações entre os

dados de entrada, produzindo resultados sob a forma de árvores de decisão de grande simplicidade e legibilidade.

Árvores de decisão desenvolvem modelos através de exemplos (aprendizagem supervisionada) e simulam o processo de abstração humana por meio de uma categorização hierárquica, obtendo regras similares a uma chave de classificação. O algoritmo particiona um conjunto de dados heterogêneo (raiz) em classes homogêneas (folhas), gerando regras de classificação com base em atributos (nós). O critério para a partição dos dados é baseado no ganho de informação que, para classificação, é proveniente da diminuição da entropia do conjunto de dados quando submetido à divisão de acordo com um atributo (Figura 1).

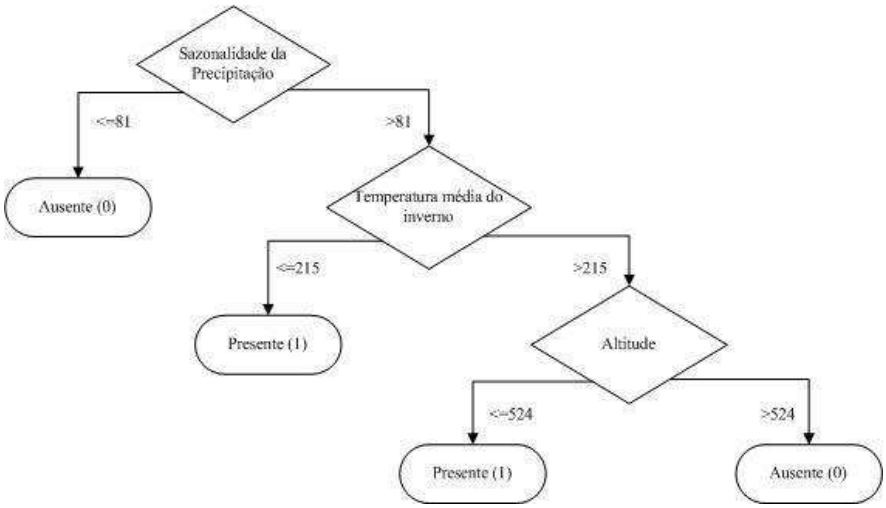

Figura 1 Representação de uma árvore de decisão com regras baseadas em variáveis ambientais

Os objetivos a serem alcançados na construção de uma árvore de decisão são: a) diminuição da entropia (aleatoriedade do atributo); b) a consistência com o conjunto de dados e c) simplicidade (menor número de nós). Quanto maior a

quantidade de nós (onde ocorrem a partição dos dados), maiores as chances de que a árvore gerada esteja demasiadamente ajustada (*overfitting*) ao seu conjunto de treinamento. Este sobreajuste atrapalha a capacidade preditiva do algoritmo com a entrada de novos dados.

Assim, para contornar este problema, alguns algoritmos utilizam a poda da árvore que, assim como na realidade, consiste em cortar algumas ramificações da árvore (nós e folhas). Esta operação pode ser realizada concomitante à construção da árvore, interrompendo seu crescimento ou após a sua construção, com seu tamanho completo. Em ambas as operações, são retiradas as subdivisões que não implicam na diminuição do erro em relação às classes superiores.

Existem diferentes algoritmos de árvore de decisão, dentre os quais destacam-se o CART (*Classification and Regression Trees*) e o C4.5. O algoritmo CART foi proposto por Breiman et al. (1984). É considerado o algoritmo inicial de qualquer árvore de decisão, não realizando qualquer tipo de poda na construção da árvore. Já o algoritmo C4.5 (QUINLAN, 1993) realiza a pós-poda da árvore inicialmente gerada, realizando uma busca na árvore, de baixo para cima, e transformando nós em folhas (divisões finais) aqueles ramos que não apresentam ganho de entropia ou diminuição do erro.

Vayssières, Plant e Allen-Diaz (2000) compararam o desempenho da árvore de decisão com modelos de regressão logística para a modelagem da distribuição de três espécies do gênero *Quercus*. A capacidade preditiva do método da árvore de decisão foi similar às obtidas pelas técnicas de regressão logística. A vantagem do primeiro método é que ele fornece uma visão das relações espécie-ambiente (efeito de todas as variáveis de previsão envolvidas no modelo), permitindo uma melhor interpretação dos resultados, ao contrário das outras técnicas utilizadas.

Com o advento e aprimoramento de outras técnicas da área de aprendizagem de máquina, a utilização do método de árvore de decisão vem decaindo ao longo dos anos em pesquisas ecológicas. Apesar da facilidade de interpretação dos resultados, a capacidade preditiva desta técnica tem sido superada em diversas pesquisas de comparação de métodos para a modelagem da distribuição de espécies (FUKUDA et al., 2013; GÁRZON et al., 2006; TAMVAKIS et al., 2014).

2.2.2 GARP

O GARP, *Genetic Algotithm for Rule-set Prediction*, é um algoritmo desenvolvido precisamente para a modelagem de distribuição de espécies, seguindo os princípios básicos dos Algoritmos Genéticos (STOCKWELL; NOBLE, 1992). Consiste na definição das condições ambientais ótimas de uma espécie através de um conjunto de regras selecionadas por um algoritmo genético. O método realiza uma seleção dessas regras, excluindo regras menos eficientes e gerando novas regras a partir de indivíduos sobreviventes. A principal vantagem do sistema é a capacidade de filtrar e lidar com diversos tipos de erros (STOCKWELL; PETERS, 1999).

Segundo Stockwell e Peters (1999), as etapas de funcionamento do GARP podem ser resumidas da seguinte maneira: No início do procedimento, um conjunto inicial de regras é gerado. A seguir, um laço iterativo seleciona aleatoriamente um conjunto de dados, por amostragem, a partir de metade dos dados disponíveis. O conjunto de regras atuais é avaliado, aplicando-o na base de dados (registros de presença e ausência) amostrada aleatoriamente. As regras mais representativas (de acordo com os critérios adotados) são armazenadas e depois analisadas. Se o grau de convergência aceitável for atingido ou o número máximo de iterações, o procedimento será encerrado, caso contrário, o

procedimento continua. O GARP é muito utilizado por apresentar resultados robustos mesmo com um número pequeno de ocorrências (STOCKWELL; PETERSON, 2002).

Peterson, Papes e Eaton (2007) compararam a capacidade preditiva do algoritmo GARP e *Maxent* em prever a distribuição da ocorrência de espécies em regiões não amostradas. Ambos os métodos geraram mapas que coincidiram com a ocorrência observada das espécies estudadas. A predição produzida pelo GARP abrangeu uma área excessivamente grande quando em comparação com a distribuição real. Os valores da métrica de avaliação AUC, *Area Under the Curve*, obtidos para o GARP e *Maxent*, respectivamente, são 0,608 e 0,733. Este resultado indica menor capacidade preditiva da técnica GARP em relação ao *Maxent*. Esta tendência dos resultados também pode ser observada na pesquisa de Elith et al. (2006) e Phillips, Anderson e Schapire (2006), nas quais a performance do GARP foi ultrapassada por outros métodos, incluindo *Maxent*.

2.2.3 Redes neurais artificiais

As redes neurais artificiais (RNA) foram criadas com base na aprendizagem de sistemas biológicos, representando uma rede complexa de neurônios interconectados. Estes neurônios, do tipo *Perceptron* (ROSENBLAT, 1958), recebem os valores de cada atributo (x_i) bem como o valor de saída esperado. Os valores de entrada recebem um peso aleatório, w_i, que determina a contribuição da entrada x_i na saída do *Perceptron* (Figura 2).

O somatório do produto destes pesos (w_i) e seus respectivos atributos (x_i) é o valor de entrada para a função de ativação, responsável pela classificação final do neurônio. Esta função, que pode ser limiar, sigmoidal, hiperbólica ou semilinear, restringe a amplitude de saída do neurônio e aplica a não linearidade do modelo. O aprendizado do neurônio consiste em ajustar os pesos de cada

entrada para que o valor de saída obtido seja igual ao valor dado pela amostra de treinamento, no sentido de minimizar o erro.

Um único neurônio, do tipo *Perceptron*, consegue separar somente dados linearmente separáveis. A fim de contornar o problema da classificação de dados não linearmente separáveis, surgiu o *Perceptron* de múltipla camada (*Multilayer Perceptron* - RNA). Neste neurônio (ou rede neural), são adicionadas uma ou mais camadas intermediárias, onde são realizados o processamento e extração das características dos dados. Com a implementação de uma ou mais camadas intermediárias, a rede é capaz de representar qualquer função contínua, sendo considerada um aproximador de funções não lineares.

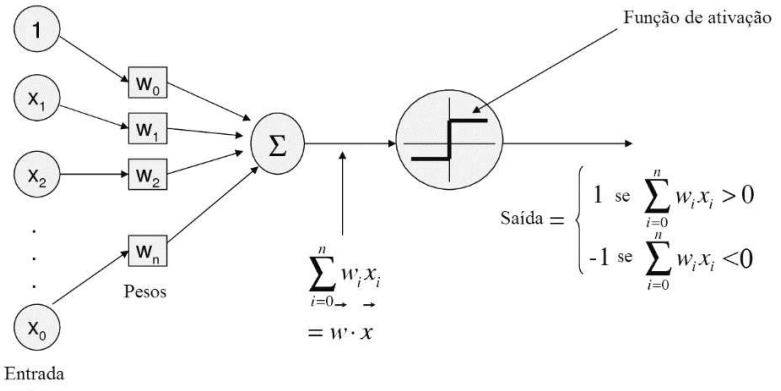

Figura 2 Esquema simplificado de um Perceptron
Fonte Bishop (1995).

O treinamento da rede é realizado de forma supervisionada, ou seja, são apresentados, no conjunto de treinamento, os valores de entrada e saída esperada. O método utilizado para o aprendizado da rede é o de retropropagação (*backpropagation*), o qual consiste na mudança interativa dos pesos sinápticos a fim de minimizar os erros de classificação para as próximas iterações. O

pseudocódigo simplificado do algoritmo *backpropagation* pode ser visto a seguir:

a) Aleatorização dos pesos entre as conexões;
b) Apresentação dos dados de treinamento (entrada e saída);
c) Cálculo da saída de cada neurônio de acordo com os pesos gerados na etapa 1(1ª iteração) ou na etapa 6 (demais iterações);
d) Comparação entre a saída gerada pela rede com a saída esperada e cálculo do erro cometido pela rede para os neurônios da camada de saída;
e) Atualização dos pesos dos neurônios da camada de saída com base no erro calculado na etapa 4;
f) Cálculo do erro das camadas anteriores (até a camada de entrada) com base nos pesos já atualizados das camadas seguintes;
g) Repetir os passos 2, 3, 4, 5 e 6 até chegar o critério de parada do algoritmo, que pode ser um erro mínimo ou número de épocas de treinamento.

Segurado e Araújo (2004) modelaram a distribuição espacial de 44 espécies de anfíbios e répteis em Portugal, utilizando sete técnicas de modelagem, dentre elas árvores de classificação e regressão, redes neurais, modelos lineares generalizados, modelos aditivos generalizados e interpoladores espaciais. Os resultados indicaram uma forte relação entre o desempenho do modelo e o tipo da distribuição espacial da espécie. Na maioria dos casos os modelos baseados em redes neurais apresentaram melhor desempenho, seguido pelos modelos aditivos generalizados.

De acordo com Fukuda et al. (2013), as redes neurais artificiais obtiveram performance satisfatória para modelar a distribuição potencial de uma

espécie de peixe (*Thymallus thymallus* L.) na Europa. Seu desempenho foi superado pelas técnicas *Random Forest*, *Suport Vector Machine* e árvores de decisão, porém foi superior aos modelos generalizados lineares e aditivos. Garzón et al. (2006), ao modelarem a distribuição potencial de *Pinus* comparando redes neurais, árvore de decisão e *Random Forest*, obtiveram acurácia satisfatória para todos os métodos, sendo *Random Forest* um pouco superior às redes neurais, que por sua vez foi superior à árvore de decisão.

2.2.4 *Random Forest*

O algoritmo *Random Forest (*RF*)*, inicialmente proposto por Breiman (2001), é um método de combinação entre classificadores (*ensemble*), neste caso, as árvores de decisão. Estas árvores construídas pelo *Random Forest* são desenvolvidas utilizando a técnica de CART (*Classification And Regression Trees*). As árvores de decisão são construídas sob diferentes conjuntos de treinamento (*bootstrap*), constituídos de *n* instâncias de treinamento escolhidas aleatoriamente na base de dados. Em cada divisão da árvore, *m* atributos são aleatoriamente selecionados para direcionar o crescimento da árvore baseado no ganho de entropia. De maneira geral, o valor de *m* deve ser menor do que o número total de atributos, para que possam ser geradas árvores distintas. Cada árvore de decisão terá sua classificação, sendo contabilizado o número de votos para cada classe. A classificação final do *Random Forest* será a classe que receber mais votos.

O algoritmo *Random Forest* apresenta excelentes características de precisão, generalização para outras amostras que não aquelas em que o classificador foi treinado e capacidade de desempenho satisfatório em amostras menores. Diversos estudos demonstraram o ótimo desempenho do RF, quando comparado com outras técnicas, para a modelagem da distribuição de espécies

raras e invasoras (GARZÓN; DIOS; OLLERO, 2008; PRASAD; IVERSON; LIAW, 2006).

Em diversas pesquisas na área de modelagem da biodiversidade, o algoritmo *Random Forest* tem se mostrado muito superior ao seu algoritmo primário, a árvore de decisão. É apontado como um dos métodos mais robustos para a modelagem da distribuição de espécies em estudos comparativos, superando inclusive outras técnicas de aprendizagem de máquina e métodos de regressão (CLUTER et al., 2007; FUKUDA et al., 2013; GARZON et al., 2006; LORENA et al., 2011).

2.2.5 *Suport Vector Machine*

Outro método empregado em modelos de distribuição de espécies é o SVM (*Suport Vector Machine* – máquina de vetores de suporte), caracterizado por ser um conjunto de técnicas de aprendizagem supervisionada pertencentes à família dos classificadores lineares generalizados. Os SVM´s foram introduzidos inicialmente por Vapnik (1995) e têm superado a maioria dos sistemas em uma ampla variedade de aplicações. Ao invés de tentar aperfeiçoar o desempenho sobre o conjunto de treinamento, os SVM´s tentam minimizar a probabilidade de classificar erroneamente padrões ainda não vistos pela distribuição de probabilidade dos dados (MARCO JÚNIOR; SIQUEIRA, 2009).

O classificador SVM se baseia em um algoritmo capaz de encontrar um hiperplano que maximize a separação entre as classes. Os dados que se encontrarem na margem deste hiperplano constituem os vetores de suporte, que serão os elementos críticos do conjunto de treinamento do classificador. Sendo assim, os modelos gerados pelo SVM dependem apenas de uma parte dos dados que compõem a amostra de treinamento. Esta característica torna esta técnica especialmente interessante para utilização em situações nas quais a

confiabilidade dos dados de entrada é duvidosa ou incompleta, o que é comum em levantamentos da biodiversidade.

O método SVM apresenta diversas vantagens em relação às outras técnicas da aprendizagem de máquina: dificilmente alcança um sobreajuste (*overfitting*) da base de dados de treinamento; produz resultados competitivos em relação aos melhores métodos utilizados em classificação; apresenta excelente desempenho de generalização ao resolver problemas não lineares e séries temporais; trabalha com pequenas bases de dados para o treinamento (TIRELLI; GAMBA; PESSANI, 2012).

No trabalho realizado por Pouteau et al. (2012), a performance do SVM constantemente superou à obtida pelo *Random Forest.* As métricas de avaliação utilizadas foram a *Area Under the Curve* e o Índice Kappa. Tirelli, Gamba e Pessani (2012) modelaram a distribuição de uma espécie de peixe nativa da Itália comparando três técnicas de aprendizagem de máquina (redes neurais, árvore de decisão e SVM). De acordo com os resultados obtidos, o método SVM superou o desempenho das árvores de decisão e foi similar à performance das redes neurais, provando sua aplicabilidade na modelagem da distribuição de espécies.

2.2.6 *Maxent*

Este algoritmo é baseado no princípio de máxima entropia e trabalha com dados de presença e pseudoausência. O *Maxent* ajusta uma distribuição de probabilidades de ocorrência da espécie a partir das variáveis ambientais de entrada, tornando-a mais próxima à distribuição uniforme, e com isso gerando menos incertezas quanto à ocorrência de um evento, alcançando a máxima entropia do sistema. Esta distribuição é ajustada sob a restrição de que os valores esperados para cada variável ambiental estejam de acordo com os valores

empíricos observados nos pontos de ocorrência (MARCO JÚNIOR; SIQUEIRA, 2009). O modelo de saída do *Maxent* para uma determinada espécie é uma superfície contínua de valores entre 0 e 100, em que altos valores indicam maior probabilidade de adequabilidade da espécie ao ambiente.

Este algoritmo apresenta algumas características que explicam sua grande aplicabilidade e alto desempenho na modelagem da distribuição de espécies. Primeiramente, é um método que exige apenas os dados de presença e as variáveis ambientais para toda a área de estudo. Essa característica é determinante para sua alta aplicação em dados ecológicos, já que dados de ausência são de difícil obtenção. Podem ser utilizados dados contínuos e categóricos, sendo o algoritmo capaz de incorporar as interações entre as variáveis de entrada. Os valores de saída do modelo são contínuos, possibilitando uma interpretação mais detalhada sobre a adequabilidade das áreas preditas. O *Maxent* é um método que realiza a generalização, o que é uma vantagem quando se tem um pequeno conjunto de treinamento (PHILLIPS; ANDERSON; SCHAPIRE, 2006).

Terribile, Diniz-Filho e Marco Junior (2010) realizaram a comparação entre o método de Máxima Entropia e o GARP para a modelagem da distribuição de 39 espécies de cobras corais do Novo Mundo. Os valores de AUC obtidos pelo GARP variaram entre 0,923 a 0,999, enquanto que para o *Maxent* variaram entre 0,877 a 0,999. De maneira geral, as diferenças na acurácia dos dois métodos foram pequenas, embora o GARP tenha apresentado melhor desempenho que o *Maxent* para 10 espécies. Os autores concluíram que estes resultados sugerem a necessidade de mais estudos para determinar sob que circunstâncias o desempenho estatístico dos modelos varia. Pearson et al. (2007) compararam os métodos *Maxent* e GARP para prever a distribuição de espécies com pequeno número de registros de ocorrência, constatando que o método de

Máxima Entropia foi melhor do que o GARP quando o número de presença foi menor que 10.

2.2.7 Métricas de Avaliação

Depois de ajustar os modelos, é necessário determinar a utilidade e aplicabilidade de cada um, o que requer uma avaliação do desempenho e da precisão dos mapas gerados. Esta avaliação identificará os pontos fortes e fracos de cada técnica e delimitará a gama de utilizações em que cada um pode ser aplicado. A avaliação de desempenho dos modelos é baseada nas medidas de precisão calculadas a partir de um conjunto de teste independente ou validação cruzada, juntamente com interpretações ecológicas.

Existem diversas métricas que podem ser utilizadas na avaliação dos métodos empregados na modelagem da distribuição de espécies, a maioria delas derivadas da matriz de confusão, representada na Tabela 1 (acurácia, erro de omissão, erro de sobreprevisão, índice Kappa e curva ROC). Pela matriz de confusão é possível identificar o número de instâncias classificados corretamente e erroneamente para cada classe (presença/ausência).

Tabela 1 Exemplo de uma matriz de confusão obtida após a classificação dos dados, sendo a – número de verdadeiros positivos; b – número de falsos positivos; c – número de falsos negativos; d - número de verdadeiros negativos

Predito \ Observado	Presença	Ausência
Presença	*a*	*b*
Ausência	*c*	*d*

Uma medida simples que pode ser derivada da matriz de confusão é a acurácia (1), que determina a proporção de locais corretamente classificados. No entanto, esta métrica vem sendo criticada por atribuir altas precisões para conjuntos com menor número de instâncias (FIELDING; BELL, 1997; MANEL; DIAS; ORMEROD, 1999).

$$Acurácia = \frac{(a+d)}{(a+b+c+d)} \quad (1)$$

Uma métrica de precisão recorrente em estudos para previsões de presença/ausência é o Índice Kappa. Esta medida foi adotada para amenizar o problema de superestimação da acurácia, e corrige a precisão dos modelos de predição pela acurácia esperada de ocorrer ao acaso (2). Este índice varia de -1 a +1, em que +1 indica perfeito acordo entre o observado e predito, e valores perto de 0 ou menos, indicam que a performance do modelo não é melhor do que uma classificação aleatória. O termo *n* indica o somatório de *a, b, c* e *d*.

$$Índice\ Kappa = \frac{\left(\frac{a+d}{n}\right) - \frac{(a+b)(a+c)+(c+d)(d+b)}{n^2}}{1 - \frac{(a+b)(a+c)+(c+d)(d+b)}{n^2}} \quad (2)$$

Apesar de sua ampla utilização, existe uma crítica a respeito desta estatística. O índice Kappa é intrinsecamente dependente da prevalência dos dados, sendo esta dependência uma fonte de viés para a estimação da acurácia (ALLOUCHE; TSOAR; KADMON, 2006).

A técnica da curva ROC (*Receiver Operating Characteristics*) tem sido amplamente utilizada na modelagem da distribuição de espécies, aplicada principalmente em modelos que utilizam dados de presença e ausência (ELITH

et al., 2006; GARZÓN et al., 2006; LORENA et al., 2011). A curva é gerada pela representação gráfica da sensibilidade (verdadeiros positivos) (3) e o complemento da especificidade (1 – especificidade) (falsos positivos) (4), sendo a sensibilidade definida como a proporção de presenças verdadeiras em relação ao total de presenças preditas e a especificidade como a proporção de ausências verdadeiras em relação às ausências preditas (MARCO JÚNIOR; SIQUEIRA, 2009).

$$Sensibilidade = \frac{a}{a+c} \tag{3}$$

$$Especificidade = \frac{d}{d+b} \tag{4}$$

A área abaixo da curva ROC (*Area Under the Curve* – AUC) mensura a habilidade do modelo em discriminar locais onde a espécie está presente e locais onde está ausente, fornecendo uma medida da capacidade discriminativa do classificador. Seu valor varia entre 0 e 1, em que 1 indica uma perfeita discriminação entre locais de presença *versus* ausência. Valores próximos a 0,5 fornecem a informação de que as predições realizadas pelo modelo são aleatórias. Valores abaixo de 0,5 indicam que a performance do modelo é inferior aos valores gerados aleatoriamente (ELITH et al., 2006). Quanto maior a AUC, maior a optimização da sensibilidade em função da especificidade, resultando em maior precisão. A vantagem desta métrica é que ela não é afetada pelo desbalanceamento ou prevalência das classes de presença/ausência, comuns em base de dados biológicos (MCPHERSON; JETZ; ROGERS, 2004).

3 MATERIAL E MÉTODOS

3.1 Descrição da área de estudo

A área de estudo, em que estão distribuídos todos os fragmentos utilizados neste trabalho, compreende todo o estado de Minas Gerais. O território mineiro, com área aproximada de 586.528 km, abrange uma região de relevo acidentado, com altitudes que oscilam entre 40 a 2.600 metros (Figura 3). As menores cotas altimétricas estão situadas na porção Centro-oriental do estado, variando desde 40 a 400 m. Apresenta um gradiente altitudinal no sentido Sul-norte, possuindo o Centro-oeste e Triângulo mineiro, na maior parte de seu território, uma altitude média de 500 metros. A maior parte do estado está inserida em uma faixa de 800 a 1.500 metros de altitude.

De acordo com o balanço hídrico climatológico (CARVALHO et al., 2008), Minas Gerais abrange oito dentre as nove classes de clima existentes de acordo com o índice de umidade de Thornthwaite (Figura 3). Este índice decresce no sentido Sudoeste a Nordeste do estado, sendo, de maneira geral, as regiões de maiores altitudes caracterizadas pelos climas úmidos a superúmido e as regiões de menores cotas altimétricas pelos climas subúmido a semiárido.

O estado de Minas Gerais compreende três domínios vegetacionais: Cerrado, Mata Atlântica e Caatinga. O domínio Cerrado abrange 57% da extensão territorial do estado, localizado na região Centro-ocidental, o domínio Atlântico ocupa 41% do estado mineiro e encontra-se distribuído na porção Oriental. Já o domínio Caatinga, está restrito ao Norte do estado e Triângulo mineiro, abrangendo apenas 2% do estado de Minas Gerais (INSTITUTO ESTADUAL DE FLORESTAS - IEF, 2015). Nestes domínios estão presente as fitofisionomias Florestas Ombrófilas, Florestas Estacionais Deciduais, Florestas

Estacionais Semideciduais, Veredas, Campo, Campo Rupestre, Campo Cerrado, Cerrado e Cerradão.

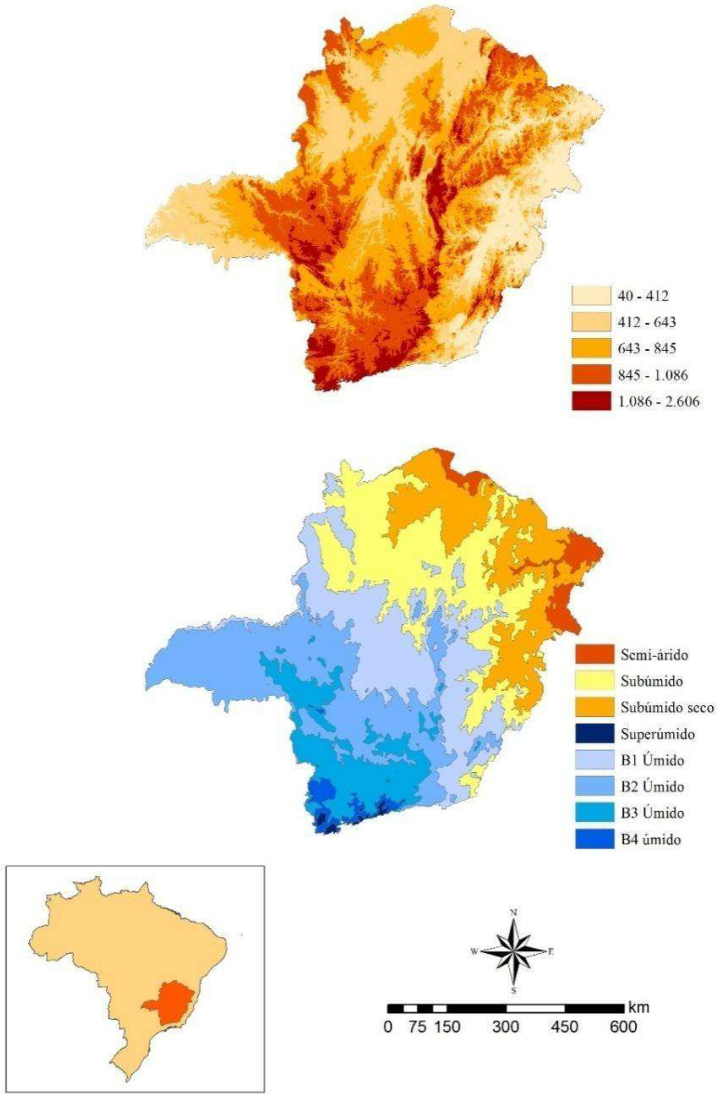

Figura 3 Classes altimétricas e zoneamento climático do estado de Minas Gerais, respectivamente

Devido à grande extensão territorial de Minas Gerais, os resultados obtidos nesta pesquisa (mapas contendo a distribuição potencial de cada espécie) serão aplicados em uma porção do estado, compreendida pela bacia hidrográfica do rio Grande. A bacia está localizada ao Sul do estado, com uma área de 86.110,02 km², representando 14,68% do território de Minas Gerais (Figura 4).

Ocupa uma região com as maiores altitudes, variando de 300 até 2.600 m. A parte Sul da bacia está inserida em áreas de maiores cotas altimétricas e o triângulo mineiro em áreas mais baixas. De acordo com a classificação climática baseada no Índice de Umidade de Thornthwaite (Iu), a parte Leste da Bacia do Rio Grande encontra-se na classe climática Úmido B2, onde o índice de umidade varia entre 40 e 60%. A temperatura média anual varia entre 19 e 20 °C e a precipitação média anual acumulada entre 1.500 a 1.600 mm. A região Central da Bacia é caracterizada como sendo do tipo Úmido B3, com índice de umidade variando entre 60 a 80%, precipitação anual superior a 1.600 mm e temperatura média inferior a 18 °C. Na região Sul da bacia encontra-se duas classes climáticas: Úmido B4 e Superúmido. O tipo de clima Úmido B4 apresenta umidade elevada associada a níveis de temperatura mais baixos, sofrendo influências de regiões serranas, com umidade variando entre 80 a 100%, índice pluviométrico superior a 1.700 mm e temperaturas amenas. Já o clima Superúmido possui índice climático superiores a 100, com temperaturas médias anuais inferiores a 14 °C e precipitação média acumulada superior a 1.750 mm (Figura 4) (CARVALHO et al., 2008).

A bacia do Rio Grande encontra-se inserida em uma faixa de transição entre os biomas Mata Atlântica e Cerrado. A vegetação da Bacia é bem representativa dos dois ambientes, variando entre espécies gramíneo-lenhosas do Cerrado e espécies lenhosas de grande porte das Florestas Estacionais e Florestas Ombrófilas. O tipo de solo predominante na Bacia do Rio Grande é

latossolo, com manchas expressivas de neossolo litólico, argissolo e cambissolo (CARVALHO et al., 2008).

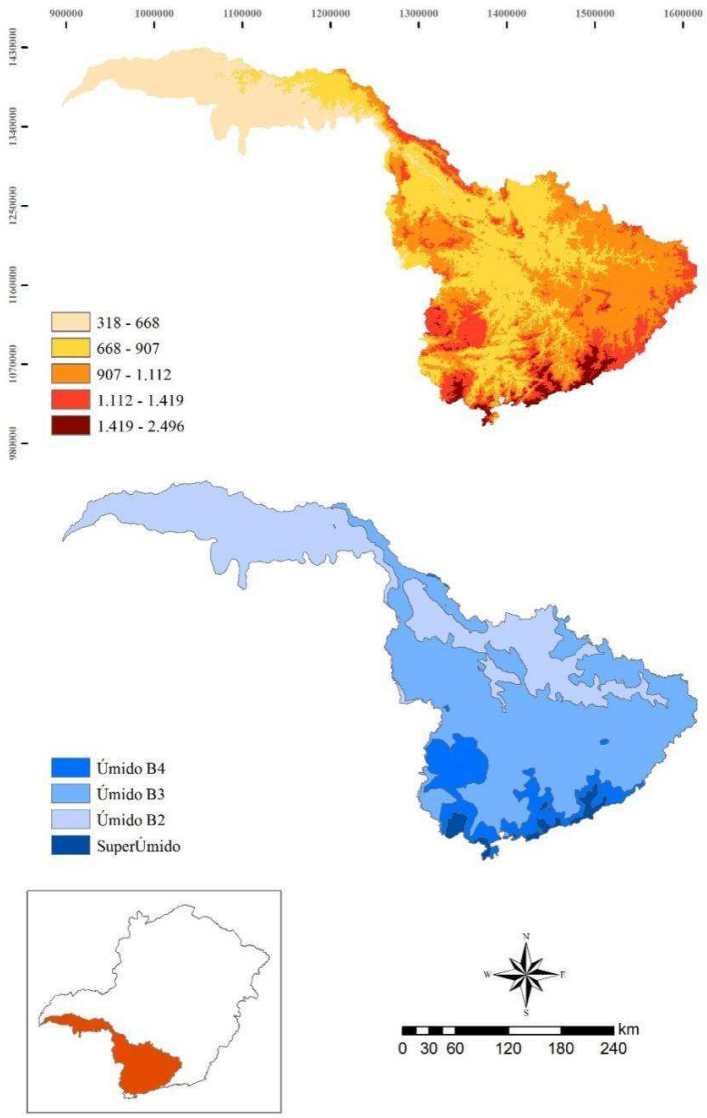

Figura 4 Classes altimétricas e zoneamento climático da bacia hidrográfica do Rio Grande, respectivamente

3.2 Levantamento florístico e seleção das espécies

Os dados de ocorrência das espécies modeladas derivam de 197 fragmentos distribuídos pelo território mineiro, provenientes do Inventário Florestal de Minas Gerais (IFMG) (SCOLFORO; CARVALHO, 2008) e do Inventário realizado na bacia hidrográfica do rio Grande (Figura 5). Para todos os fragmentos inventariados, foi coletado material botânico apenas dos indivíduos com CAP (circunferência a altura do peito) maior que 15,7 cm.

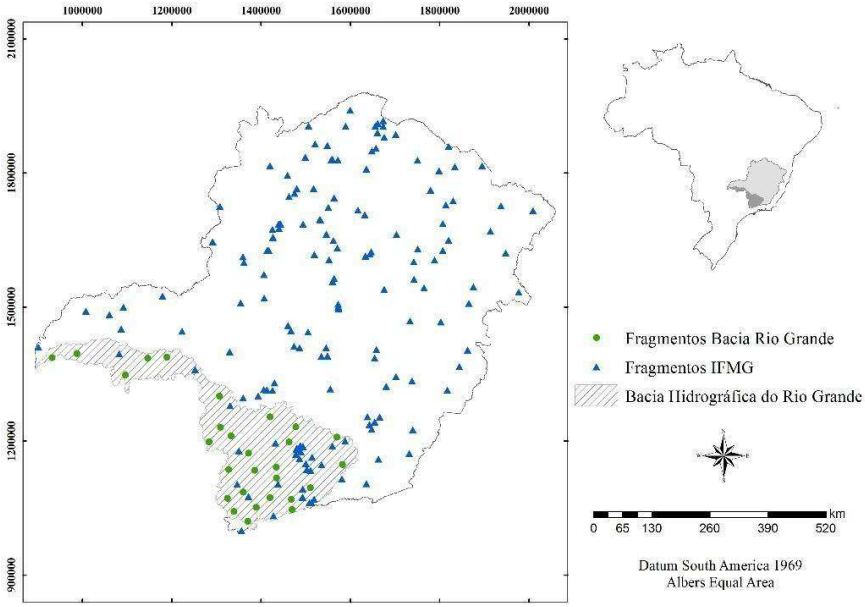

Figura 5 Distribuição dos fragmentos utilizados na modelagem no estado de Minas Gerais

A seleção das espécies foi embasada nos dados obtidos pelos fragmentos inventariados e em estudos fitossociológicos realizados na bacia do Rio Grande (OLIVEIRA FILHO et al., 1995; RODRIGUES et al., 2003; VILELA, 2000;

VILELA et al., 1993). Foram selecionadas quatro espécies arbóreas de grande abundância e distribuição na região e potencial para revegetação de matas ciliares: *Casearia Sylvestris* SW., *Copaifera langsdorffii* Desf., *Croton floribundus* Spreng. e *Tapirira guianesis* Aubl.

3.3 Variáveis abióticas

Foram compiladas 25 variáveis abióticas amplamente reconhecidas em estudos envolvendo modelagem preditiva da distribuição de espécies. Deste total, 20 provêm da base de dados do WorldClim (HIJMANS et al., 2005), sendo dados numéricos e com resolução espacial de 1 km aproximadamente. Entre as variáveis estão altitude e 19 variáveis climáticas, derivadas basicamente da temperatura e precipitação mensal, modeladas geoespacialmente a partir de estações meteorológicas. As demais variáveis são categóricas e oriundas do Zoneamento Ecológico Econômico de Minas Gerais (CARVALHO et al., 2008) (Tabela 2).

O conjunto de variáveis abióticas foi pré-processado de modo que a projeção geográfica, tamanho e alinhamento do *pixel,* e extensão geográfica (ao longo de toda a área de estudo) fossem comuns para todas as variáveis. Optou-se por trabalhar com o sistema de referência espacial *South America Albers Equal Conic,* Datum SAD69. Todas as variáveis abióticas foram reamostradas considerando uma resolução espacial de 1 km.

Tabela 2 Fonte e descrição das variáveis abióticas inicialmente utilizadas neste trabalho

Variáveis abióticas	Fonte	Descrição
Altitude	WorldClim	Altitude (m)
Bio01	WorldClim	Temperatura média anual (°C)
Bio02	WorldClim	Variação mensal média da temperatura (°C)
Bio03	WorldClim	Isotermalidade ((Bio02/Bio07)*100)
Bio04	WorldClim	Sazonalidade da temperatura (desvio padrão*100)(°C)
Bio05	WorldClim	Temperatura máxima do mês mais quente (°C)
Bio06	WorldClim	Temperatura mínima do mês mais frio (°C)
Bio07	WorldClim	Variação anual da temperatura (°C)
Bio08	WorldClim	Temperatura média do trimestre mais úmido (°C)
Bio09	WorldClim	Temperatura média do trimestre mais seco (°C)
Bio10	WorldClim	Temperatura média do trimestre mais quente (°C)
Bio11	WorldClim	Temperatura média do trimestre mais frio (°C)
Bio12	WorldClim	Precipitação anual (mm)
Bio13	WorldClim	Precipitação do mês mais úmido (mm)
Bio14	WorldClim	Precipitação do mês mais seco (mm)
Bio15	WorldClim	Coeficiente de variação da precipitação
Bio16	WorldClim	Precipitação do trimestre mais úmido (mm)
Bio17	WorldClim	Precipitação do trimestre mais seco (mm)
Bio18	WorldClim	Precipitação do trimestre mais quente (mm)
Bio19	WorldClim	Precipitação do trimestre mais frio (mm)
Tipo de solo	ZEE	Argissolo, Cambissolo, Espodossolo, Gleissolo, Latossolo, Luvissolo, Neossolo Flúvico, Neossolo Litólico, Neossolo quartzarênico, Nitossolo, Planossolo
Textura do solo	ZEE	Fina, média, grossa
Teor de M.O. do solo	ZEE	Alto, médio, baixo
Regime hídrico do solo	ZEE	Aquíco, arídico, údico, ústico
Relevo	ZEE	Plano ou suave-ondulado, ondulado, forte ondulado

A fim de selecionar as variáveis mais representativas e diminuir o número de atributos e a complexidade dos modelos, muitas pesquisas utilizam técnicas de seleção de atributos. Neste estudo, a fim de comparação entre resultados, foi aplicado um algoritmo de seleção (*InfoGainAttributeEval*) implementado no *Weka* (WITTEN; FRANK; HALL, 2011). Os três algoritmos (árvore de decisão, *Random Forest* e redes neurais) foram treinados sob dois conjuntos de variáveis abióticas: o primeiro contendo as 25 variáveis; e o segundo contendo os atributos selecionados pelo método do ganho de informação.

O algoritmo *InfoGainAttributeEval* seleciona um subconjunto de atributos baseado no ganho de informação (diminuição da entropia das classes de acordo com os atributos selecionados). A entropia (5) caracteriza o grau de impureza dos dados (falta de homogeneidade das classes), sendo máxima (igual a 1) quando o conjunto de dados é heterogêneo. Logo, dado um conjunto de entrada (S), que pode ter *c* classes distintas, sendo p_i a proporção de dados em S que pertencem à classe *i*, sua entropia é dada por:

$$Entropia(S) = \sum_{i \in C} -p_i \log_2 p_i \tag{5}$$

O ganho de informação (6) para um atributo (A) em relação ao conjunto de dados (S), fornece a medida da diminuição da entropia esperada quando se faz a repartição de S em função de A. Sendo a *Entropia (S)* e *Entropia (A)* a medida de (não) homogeneidade do conjunto S e do conjunto de S particionado pelo atributo (A), respectivamente.

No caso de variáveis numéricas, é utilizado um ponto de referência (qualquer ponto intermediário entre classes diferentes) para realizar a partição do conjunto de dados. Seja P(A) o conjunto de valores possíveis de A; x um

elemento deste conjunto e S_x o subconjunto de S formado pelos dados em que A = x; a entropia de particionar S em função de A é dada pela equação 7.

$$Ganho(S, A) = Entropia(S) - Entropia(A) \quad (6)$$

$$Entropia(A) = \sum_{x \in P(A)} \frac{|S_x|}{|S|} Entropia(S_x) \quad (7)$$

O algoritmo de seleção começa com um conjunto vazio de atributos e utiliza a heurística *Ranker* como algoritmo de busca, que avalia os atributos individualmente e os organiza de acordo com sua ordem de importância (ganho de informação). O parâmetro *numToSelect* (número de atributos a serem selecionados) foi fixado em 10.

3.4 Conjunto de treinamento e teste

Na modelagem preditiva da distribuição de espécies, a base de dados formada pelos pontos de ocorrência e variáveis ambientais, é dividida em dois conjuntos de dados independentes: conjunto de treinamento e conjunto de teste. Em geral, esses conjuntos representam 70% e 30%, respectivamente, do conjunto total dos dados disponíveis (PHILLIPS; ANDERSON; SCHAPIRE, 2006). O conjunto de treinamento é utilizado no ajuste dos modelos e o conjunto de teste é empregado na avaliação da capacidade preditiva dos modelos com base em dados não utilizados para o ajuste (validação preditiva).

Para a formação do conjunto de treinamento, foram selecionados 136 fragmentos (aproximadamente 70% do total dos dados) distribuídos em cinco classes de altitude (Tabela 3). Essa estratificação foi realizada com base no conhecimento de que o gradiente altitudinal está altamente relacionado com o clima, e por conseguinte com as diferentes fisionomias e estruturas florestais

(HERNÁNDEZ et al., 2012; HOMEIER et al., 2010). Diferentes cotas altimétricas propiciam diferentes condições ambientais, como temperatura, umidade do ar, disponibilidade hídrica, exposição aos ventos e características edáficas (CARVALHO, 2005; HOMEIER et al., 2010). A localização dos fragmentos que compõem o conjunto de treinamento pode ser visualizada na Figura 6.

Tabela 3 Estratificação do conjunto de treinamento com base na altitude

Classe	Cotas altimétricas	Area (km²)	Área (%)	Número de Fragmentos
1	< 425	52.651	8,7	10
2	425 - 650	161.113	26,63	37
3	650 - 845	204.680	33,83	46
4	845 - 1085	149.111	24,64	34
5	> 1085	37.559	6,21	9
Total	-	60.5114	100	136

O conjunto de teste compreende 61 fragmentos, distribuídos ao longo do estado de Minas Gerais (Figura 6), sendo formado pelos fragmentos restantes, que não foram selecionados para compor o conjunto de treinamento. As classes altimétricas 1 e 3 não estão representadas neste conjunto devido à falta de fragmentos inventariados nestas cotas de altitude.

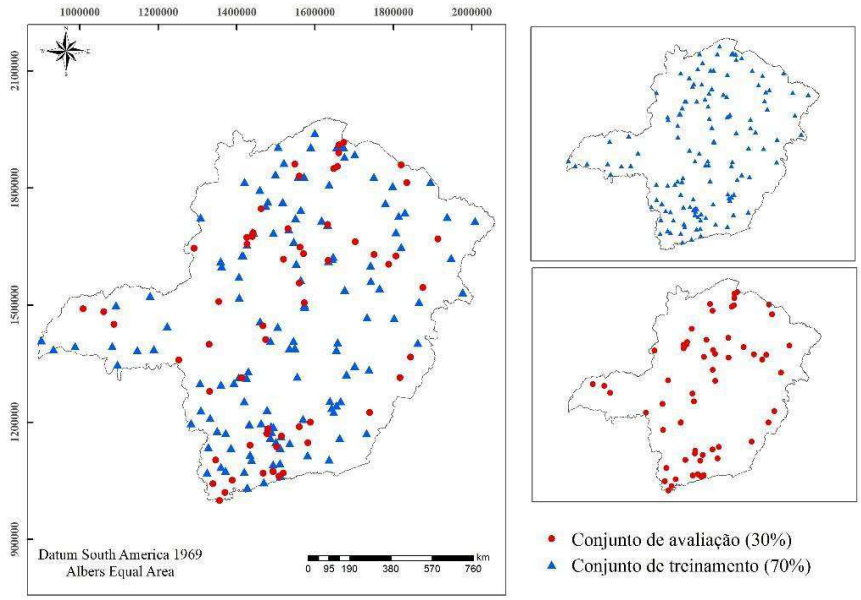

Figura 6 Localização dos fragmentos que compõem o conjunto de treinamento e teste

Devido à resolução espacial das variáveis abióticas utilizadas neste trabalho (1 km), optou-se por trabalhar os dados de ocorrência em nível de fragmento. Adotando esta estratégia pretende-se evitar a geração de dados com ruídos, já que existe a possibilidade de parcelas inseridas em características ambientais similares apresentarem diferentes respostas de ocorrência.

Os dados do inventário de cada fragmento foram trabalhados como sendo uma única unidade amostral, independente do número de parcelas empregadas, sendo representado por seu ponto central. A partir desta etapa, computou-se a presença (1) ou ausência (0) das quatro espécies selecionadas para cada fragmento.

3.5 Algoritmos de aprendizado de máquina

O algoritmo de árvore de decisão (AD) utilizado neste trabalho é C4.5. Foi utilizado o *software* Weka, no qual o algoritmo está implementado sob a denominação de J48. Anterior à sua aplicação na base de dados, foi realizada a parametrização do modelo, com o auxílio do metaclassificador *cvparameter* (também disponível no *software Weka*). Foram avaliados os parâmetros *minNumObj* (2 a 10) e *confidenceFactor* (0,1 a 0,5). O parâmetro *minNumObj* determina o número mínimo de instâncias em cada folha e o *confidenceFactor* analisa a precisão das regras geradas.

O algoritmo *Random Forest* (RF) também foi utilizado no *software Weka.* Para sua parametrização, empregando o metaclassificador *cvparameter*, foram testados os parâmetros *numTrees* (5, 10, 15, 20 e 25), que corresponde ao número de árvores de decisão a serem geradas pelo algoritmo, e *NumFeatures,* que foi calculado em função do número de atributos utilizados na modelagem, pela equação (9).

$$NumFeatures = \log_2(n) + 1 \tag{9}$$

A rede neural artificial (RNA) empregada neste estudo, recebe a denominação de *MultilayerPerceptron* no *software* Weka. Em sua parametrização testou-se os parâmetros *hiddenLayers* (0 a 2), *learningRate* (0,1 a 0,6) e *Momentum* (0,2 a 0,6), com o auxílio do metaclassificador *cvparameter*. *HiddenLayers* identifica o número de camadas ocultas a serem utilizadas na rede neural, enquanto *learningRate* determina a velocidade de aprendizado da rede. O termo *momentum* modifica a regra inicial do algoritmo *backpropagation*, evitando a parada em ótimos locais.

Na parametrização dos três algoritmos acima, para cada espécie, foi estabelecido o número de 500 iterações e 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*) por algoritmo, foram selecionadas para treinamento e avaliação dos algoritmos no presente estudo.

Apesar do *Maxent* ser um método diferente dos algoritmos utilizados neste trabalho (utiliza apenas dados de presença e gera um grande número de pseudoausências), este foi empregado a título de comparação, visto que, atualmente, é o método mais amplamente utilizado na modelagem da distribuição de espécies, apresentando bons resultados (ELITH et al., 2006; TERRIBILE; DINIZ-FILHO; MARCO JÚNIOR, 2010; WILLIAMS et al., 2009).

Neste estudo foi utilizado o *software Maxent* (PHILLIPS; DUDIK; SCHAPITE, 2004), escrito em Java e disponível gratuitamente em http://www.cs.princeton.edu/~schapire/maxent. Optou-se por utilizar o número máximo de 10.000 pontos de pseudoausência (*background points*) e número de iterações igual a 500.

Visto que o *Maxent* gera mapas de probabilidades, representando a adequabilidade ambiental de cada espécie, mapas de presença e ausência foram gerados através da reclassificação dos dados para posterior comparação das áreas de ocorrência. Dois limiares de adequabilidade foram adotados para a reclassificação: valor mínimo de adequabilidade encontrado em um ponto de presença (cruzamento dos pontos de ocorrência do conjunto de treinamento e o modelo gerado), por espécie e valor de adequabilidade igual a 0,5 (abaixo de 0,5 ausência; acima de 0,5 presença) para todas as espécies.

3.6 Avaliação e aplicação dos modelos

Os algoritmos de aprendizado de máquina foram comparados, para cada espécie modelada, com base na área abaixo da curva ROC (*Receiver Operating Characteristic*), denominada AUC (*Area Under the Curve*). Essa métrica de avaliação vem sendo amplamente empregada em estudos comparativos na modelagem da distribuição de espécies (ELITH et al., 2006; LORENA et al., 2010; POUTEAU et al., 2012; TERRIBILE; DINIZ-FILHO; MARCO JUNIOR, 2010; WILLIAMS et al., 2009).

Inicialmente, os valores de AUC, para os algoritmos de presença e ausência (AD, RF e RNA), foram calculados a partir da validação cruzada, sob o conjunto de treinamento. Nesta etapa, além dos valores de AUC, foram utilizadas duas outras métricas de avaliação para discussão do desempenho dos algoritmos: porcentagem dos dados classificados corretamente e taxa de verdadeiros positivos. O método de validação cruzada garante aleatoriedade dos conjuntos de teste, sendo que cada modelo foi avaliado dez vezes com 10% dos dados de treinamento. Por meio deste método de avaliação é possível aplicar testes estatísticos de comparação entre as métricas obtidas. Os valores de AUC provenientes da validação cruzada foram submetidos ao teste t-pareado, com 95% de confiança, para verificar se existe diferença significativa entre o desempenho dos algoritmos testados.

De posse dos resultados obtidos com a validação cruzada, foi realizada escolha do conjunto de variáveis abióticas (base de dados 1 ou e base de dados 2) que apresentou melhores resultados, em termos de AUC, para a maioria dos algoritmos treinados para todas as espécies. Após essa escolha, os algoritmos foram treinados novamente com o conjunto total de treinamento, porém avaliados segundo um único conjunto de dados independentes e nunca utilizados, o conjunto de teste. Ao aplicar este tipo de validação preditiva,

buscou-se avaliar os diferentes algoritmos testados sobre um mesmo conjunto de teste, gerando assim um único valor de AUC. Nessa etapa, à título de comparação, também foi utilizado o algoritmo *Maxent*, treinado e avaliado com os mesmos dados de presença empregados nos algoritmos de presença e ausência, porém, fazendo o uso de pseudoausências.

O algoritmo que apresentou maior AUC por espécie, de acordo com a validação preditiva com o conjunto de teste, foi então empregado para modelar a distribuição potencial das quatro espécies trabalhadas. Devido à magnitude da área do estado de Minas Gerais, optou-se por apresentar apenas a bacia hidrográfica do rio Grande (MG) como objeto final de espacialização das informações geradas.

No caso do *Maxent*, o mapa de adequabilidade de *habitat* (transformado em presença/ausência) foi gerado para todas as espécies, com o intuito de comparar a área de ocorrência obtida por este método (*presence-only*) e o algoritmo de presença-ausência com maior capacidade preditiva.

4 RESULTADOS E DISCUSSÃO

De acordo com os 197 fragmentos inventariados no estado de Minas Gerais, três, das quatros espécies trabalhadas neste estudo, apresentaram uma ampla distribuição pelo território mineiro (Figura 7). São elas: *Casearia sylvestris, Copaifera langsdorffii* e *Tapirira guianensis*. A espécie *Croton floribundus* teve sua ocorrência limitada, principalmente, ao Sul do estado.

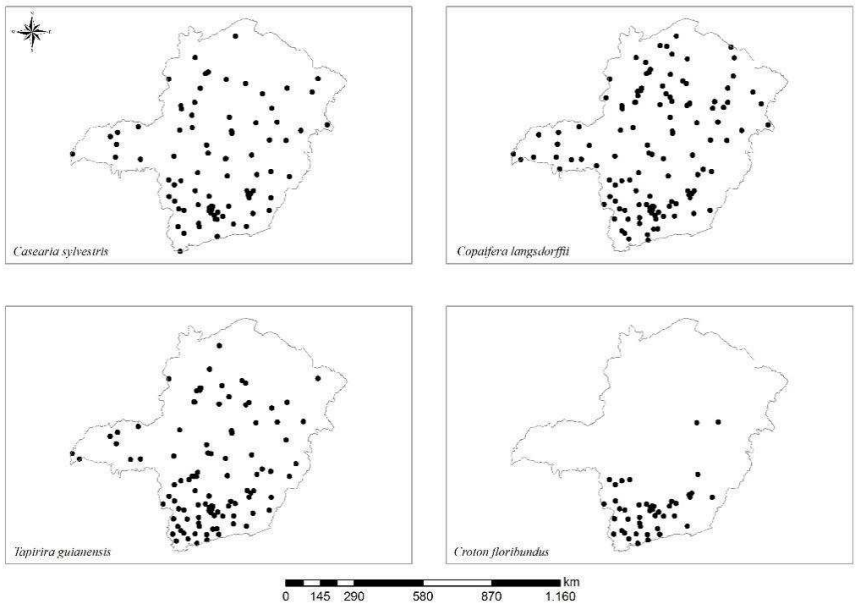

Figura 7 Ocorrência das espécies *Casearia sylvestris, Copaifera langsdorffii* e *Tapirira guianensis* de acordo com os fragmentos inventariados em Minas Gerais

Com relação ao conjunto de treinamento (136 fragmentos), delineado de forma a compreender diferentes classes de altitude, a espécie *Casearia sylvestris* esteve presente em 76 fragmentos, *Copaifera langsdorffii* ocorreu em 93

fragmentos e *Tapirira guianensis* em 70. Baseado nas características ambientais, a ocorrência destas espécies foi similar em alguns aspectos. A faixa de altitude de maior ocorrência destas espécies está entre 350 m e 1.000 m. A maior recorrência está ligada ao tipo de solo latossolo com regime hídrico údico ou ústico. As espécies estiveram presentes em locais com ampla variação de temperatura (16,1 °C até 24,8 °C), principalmente acima dos 19 °C, e variação da precipitação (805 mm a 1.783 mm), estando presentes em maior quantidade na faixa de precipitação entre 1.131 mm e 1.620 mm (Figura 8). A grande variação de características abióticas em relação à ocorrência destas espécies e a distribuição espacial pelo território mineiro, vão de acordo com a literatura (SILVA-LUIZ; PIRANI, 2013), que classificam estas espécies como generalistas. Essas ocorrem em diversos estados brasileiros e em grande abundância em Minas Gerais, além de estarem presentes em todos os domínios fitogeográficos (Amazônia, Caatinga, Cerrado, Mata Atlântica, Pampa e Pantanal).

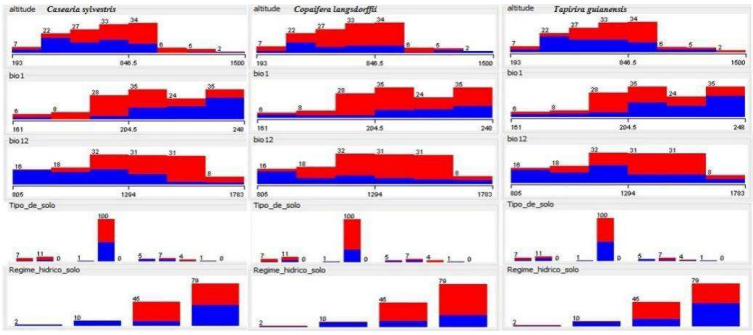

Figura 8 Histograma de ocorrência (no qual vermelho representa presença, e azul ausência) para as espécies *Casearia sylvestris, Copaifera langsdorffi* e *Tapirira guianensis* em relação às variáveis abióticas: altitude, temperatura média anual (bio1), precipitação média anual (bio12), tipo de solo e regime hídrico do solo

A espécie *Croton floribundus,* que ocorreu em apenas 36 dos 136 fragmentos do conjunto de treinamento, teve sua presença atrelada a uma faixa de altitude entre 680 m e 1.000 m. Ocorreu, principalmente, em solos do tipo latossolo údico ou ústico. A espécie esteve foi mais recorrente em fragmentos com temperatura entre 19 °C e 21,9 °C e precipitação variando entre 1.295 mm e 1.620 mm (Figura 9). De acordo com Lorenzi (2008), essa espécie ocorre naturalmente nos estados de Minas Gerais, Rio de Janeiro, São Paulo e Paraná, sendo sua dispersão maior em regiões de altitude e em florestas semideciduais da bacia do Paraná.

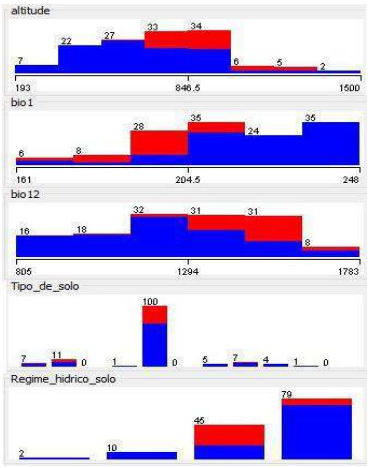

Figura 9 Histograma de ocorrência (em vermelho presença, e em azul ausência) para a espécie *Croton floribundus* em relação às variáveis abióticas: altitude, temperatura média anual (bio1), precipitação média anual (bio12), tipo de solo e regime hídrico do solo

Após a realização da parametrização para os algoritmos árvore de decisão (AD), *Random Forest* (RF) e redes neurais artificiais (RNA), com o

emprego do metaclassificador *cvparameter,* as seguintes configurações, contidas nas Tabelas 4, 5 e 6, foram selecionadas:

Tabela 4 Valores dos parâmetros fator de confiança (confidenceFactor) e mínimo de instâncias na folha (minNumObj) otimizados para a árvore de decisão para as 4 espécies estudadas

Espécies	árvore de decisão	
	Fator de confiança	Mínimo de instâncias na folha
Casearia sylvestris	0,2	6
Copaifera langsdorffii	0,1	2
Croton floribundus	0,1	2
Tapirira guianensis	0,3	6

Tabela 5 Valores dos parâmetros número de árvores (numTrees) e número de atributos em cada árvore (numFeatures) otimizados para o *Random Forest* para as 4 espécies estudadas, sendo a base de dados 1 correspondente às 25 variáveis abióticas e a base de dados 2 correpondente á utilização do método de seleção pelo ganho de informação (10 atributos)

Espécies	*Random Forest*		
	Número de árvores	Número de atributos em cada árvore	
		base de dados 1	base de dados 2
Casearia sylvestris	25	4	3
Copaifera langsdorffii	25	4	3
Croton floribundus	20	4	3
Tapirira guianensis	25	4	3

Tabela 6 Valores dos parâmetros taxa de aprendizagem (learningRate), Momentum e e número de camadas ocultas (hiddenLayers) otimizados para as redes neurais artificiais para as 4 espécies estudadas

Espécies	redes neurais artificiais		
	Taxa de aprendizagem	Momentum	Número de camadas ocultas
Casearia sylvestris	0,5	0,33	2
Copaifera langsdorffii	0,5	0,47	2
Croton floribundus	0,3	0,2	2
Tapirira guianensis	0,4	0,2	2

O resultado da seleção de atributos realizada com base no ganho de informação pode ser visualizado na Tabela 7, em que o número entre parênteses é o ganho de informação obtido após a subdivisão do conjunto de dados pelo atributo correspondente.

Tabela 7 Seleção dos atributos para cada espécie com base no ganho de informação

Espécie	Atributos selecionados
Casearia sylvestris	Bio11 (0,27); Bio01 (0,266); Bio14 (0,252); Bio15 (0,25); Bio19 (0,225); Bio04 (0,225); Bio17 (0,204); Bio12 (0,201); Bio16 (0,2); Bio18 (0,197)
Copaifera langsdorffii	Bio17 (0,196); Bio12 (0,1459); Bio16 (0,093); Regime hídrico do solo (0,0898); Bio19 (0,0867); Bio14 (0,0783); Bio11 (0,0765); Tipo de solo (0,0681); Bio01 (0,0633); Teor de matéria orgânica (0,0181)
Croton floribundus	Bio11 (0,436); Bio09 (0,415); Bio01 (0,377); Bio03 (0,368); Bio06 (0,348); Bio05 (0,335); Bio10 (0,327); Bio04 (0,319); Bio08 (0,312); Bio17 (0,279)
Tapirira guianensis	Bio19 (0,1381); Bio17 (0,1103); Bio12 (0,1042); Bio15 (0,0945); Bio14 (0,0903); Bio18 (0,0898); Bio16 (0,0873); Bio13 (0,086); Regime hídrico do solo (0,0856); Bio05 (0,0852)

De posse das informações sobre a heterogeneidade dos dados de cada espécie relacionada a cada atributo, foi possível verificar que a espécie *Croton floribundus* possui um conjunto de treinamento mais homogeneamente dividido pelos atributos selecionados (maior ganho de informação). Este fato pode ser explicado pela ocorrência limitada (não generalista) da espécie no estado de Minas Gerais, compreendendo uma menor gama de características ambientais, conferindo-lhe um caráter mais especialista.

A grande maioria dos atributos selecionados para esta espécie foi relacionada à temperatura (°C). A variável Bio11 (Temperatura média da estação fria) foi o atributo que melhor dividiu o conjunto de dados em presença e ausência. As presenças se encontraram na faixa entre 13,2 °C a 19,4 °C, principalmente entre 16,3 °C e 17,5 °C. Fragmentos com temperatura média de inverno acima dos 19,4 °C não apresentaram a ocorrência de *Croton floribundus*. A única variável relacionada com a temperatura selecionada pelo método foi a Bio17 (Precipitação do trimestre mais seco), que indica a quantidade mínima de precipitação na estação seca. A ocorrência da espécie se deu acima dos 20 mm, sendo maior em fragmentos com 38 a 56 mm de precipitação no inverno.

As espécies *Copaifera langsdorffii* e *Tapirira guianensis* apresentaram grande heterogeneidade das classes de presença/ausência quando submetidas à divisão de acordo com as variáveis abióticas disponíveis (baixo ganho de informação). Este comportamento pode ser devido à ampla distribuição das espécies no território mineiro, abrangendo grande espectro de características ambientais (generalistas), e a presença e ausência em fragmentos com características abióticas similares.

A maior parte dos atributos selecionados para *Copaifera langsdorffii* e *Croton floribundus* foi proveniente de valores de precipitação, indicando que a ocorrência destas espécies está altamente atrelada a este fator climático. O maior

número de presenças ocorreu na faixa de precipitação mínima da estação seca (Bio17) entre 2 mm e 56 mm, sendo esta variável uma das mais informativas para as duas espécies. Esta informação, em conjunto com a faixa de temperatura média anual de ocorrência das espécies, indica que, apesar destas espécies necessitarem de considerável quantidade de água ao longo do ano, as mesmas admitem uma estação seca bem delimitada, típicas de espécies das florestas semideciduais.

Outra variável selecionada para as duas espécies foi o regime hídrico do solo, indicando a importância da umidade do solo na ocorrência destas espécies, sendo preferenciais os solos úmidos a maior parte do ano ou os intermediários (údico ou ústico). Diversos autores confirmam a generalidade da distribuição de *Tapirira guianensis* e *Copaifera langsdorffii*, porém sempre apresentando preferência por ambientes de solo úmido (LORENZI, 2008; OLIVEIRA FILHO; RATTER, 2000; SILVA JUNIOR, 1997; WALTER; RIBEIRO, 1997).

A espécie *Casearia sylvestris*, apesar de possuir extensa área de ocorrência, apresentou maior homogeneidade dos dados quando submetidos às diferentes variáveis abióticas. Sua ocorrência não foi definida por condições tão específicas como as de *Croton floribundus* nem tão generalistas como as de *Copaifera langsdorffii* e *Tapirira guianensis*. As variáveis abióticas selecionadas de acordo com a homogeneidade dos dados foram, em sua maioria, relacionadas à precipitação.

A Tabela 8 apresenta as médias dos resultados dos três algoritmos (AD, RF e RNA), obtidas por validação cruzada, para a espécie *Casearia sylvestris,* com e sem a seleção das variáveis abióticas mais informativas. Entre parênteses, encontram-se os valores dos desvios relativos a cada média. Para a modelagem utilizando o conjunto completo de atributos (base de dados 1), os algoritmos obtiveram desempenho estatisticamente iguais, de acordo com o teste T-pareado com 95% de confiança para todas as métricas de avaliação. Os algoritmos RF e

RNA apresentaram AUC iguais e numericamente superiores à árvore de decisão (AD). No entanto, o RF mostra ser mais eficiente na classificação dos locais de ocorrência da espécie, visto que apresenta maior taxa de verdadeiros positivos.

Já o desempenho dos algoritmos modelados com a base de dados contendo apenas as variáveis abióticas selecionadas pelo método do ganho de informação (base de dados 2), apresentou diferença significativa em relação a AUC. RF e RNA foram estatisticamente superiores ao algoritmo AD, sendo as métricas de avaliação das RNA's ligeiramente superiores ao RF, demonstrando um melhor desempenho para reconhecer os padrões de distribuição da *Casearia sylvestris*.

Tabela 8 Métricas de avaliação obtidas por validação cruzada do conjunto de treinamento para os algoritmos árvore de decisão (AD), *Random Forest* (RF) e redes neurais artificiais (RNA) para a espécie *Casearia sylvestris*

Casearia sylvestris				
Base de dados	Algoritmos	AUC	% Classificado corretamente	Taxa de verdadeiros positivos
1	AD	0,73 (0,14)	70,40 (12,48)	0,66 (0,22)
	RF	0,80 (0,11)	71,01 (11,61)	0,67 (0,20)
	RNA	0,80 (0,12)	70,41 (10,53)	0,64 (0,22)
2	AD	0,69 (0,13)	67,41 (11,75)	0,62 (0,25)
	RF	0,78 (0,10)*	69,40 (10,20)	0,66 (0,19)
	RNA	0,81 (0,12)*	74,00 (11,59)	0,76 (0,21)

* Significativamente superior ao algoritmo AD de acordo com o teste T-pareado (95% de confiança);
** Significativamente inferior ao algoritmo AD de acordo com o teste T-pareado (95% de confiança).

As métricas de avaliação obtidos para a espécie *Copaifera langsdorffii* (Tabela 9), principalmente a AUC e a taxa de verdadeiros positivos, foram

inferiores às obtidas para *Casearia sylvestris* (Tabela 8). O baixo desempenho dos modelos para esta espécie pode ser explicado pela heterogeneidade de sua base de dados, confirmada pelo método de ganho de informação. Este comportamento é esperado pois, a espécie, é conhecida por sua generalidade na escolha de ambientes (CARVALHO, 2005).

Os algoritmos ajustados com o conjunto de treinamento, contendo as 25 variáveis abióticas (base de dados 1), para *Copaifera langsdorffii,* não apresentaram diferenças significativas em seus desempenhos. Porém, vale ressaltar, que o algoritmo RF apresentou melhores valores de AUC e taxa de verdadeiros positivos que os outros métodos, diferenças estas não significativas diante do teste-t com 95% de confiabilidade (Tabela 9).

Em relação ao conjunto de dados em que foi realizada a seleção de atributos (base de dados 2), o RF apresentou AUC estatisticamente superior ao algoritmo AD e numericamente maior ao RNA. As taxas de verdadeiros positivos apresentadas pelos três métodos foram baixas, indicando a dificuldade em predizer locais de ocorrência da espécie. O algoritmo AD, apesar de obter uma porcentagem dos dados classificados corretamente e ser estatisticamente superior às redes neurais, obteve a menor taxa de verdadeiros positivos. Ou seja, maior dificuldade em predizer corretamente os locais de presença da espécie.

Tabela 9 Métricas de avaliação obtidas por validação cruzada do conjunto de treinamento para os algoritmos árvore de decisão (AD), *Random Forest* (RF) e redes neurais artificiais (RNA) para a espécie Copaifera langsdorffii

Copaifera langsdorffii				
Base de dados	Algoritmos	AUC	% Classificado corretamente	Taxa de verdadeiros positivos
1	AD	0,60 (0,13)	69,00 (11,52)	0,27 (0,22)
	RF	0,68 (0,14)	66,40 (11,01)	0,40 (0,22)
	RNA	0,52 (0,18)	62,35 (11,87)	0,26 (0,20)
2	AD	0,61 (0,11)	74,24 (8,79)	0,23 (0,21)
	RF	0,72 (0,15)*	68,20 (12,05)	0,39 (0,22)
	RNA	0,63 (0,15)	66,87 (10,32)**	0,31 (0,24)

* Significativamente superior ao algoritmo AD de acordo com o teste T-pareado (95% de confiança);
** Significativamente inferior ao algoritmo AD de acordo com o teste T-pareado (95% de confiança).

A Tabela 10 contém os resultados obtidos pelos três algoritmos para a espécie *Croton floribundus*. Os valores apresentados para esta espécie, para todas as métricas de avaliação, são maiores do que os obtidos para as demais. Em conjunto com o método de ganho de informação, esta informação corrobora com a ideia de que *Croton floribundus* apresenta dados mais homogêneos e com relações mais fortes com as variáveis abióticas utilizadas neste trabalho. Para ambas as bases de dados (completa e com seleção de atributos) todos os algoritmos apresentaram resultados estatisticamente semelhantes, sem diferenças significativas para nenhuma métrica de avaliação. Porém, de forma absoluta, mais uma vez, o algoritmo RF apresentou métricas numericamente superiores aos demais métodos.

Tabela 10 Métricas de avaliação obtidas por validação cruzada do conjunto de treinamento para os algoritmos árvore de decisão (AD), *Random Forest* (RF) e redes neurais artificiais (RNA) para a espécie *Croton floribundus*

Croton floribundus				
Base de dados	Algoritmos	AUC	% Classificado corretamente	Taxa de Verdadeiros positivos
1	AD	0,89 (0,09)	84,86 (8,40)	0,86 (0,11)
	RF	0,90 (0,10)	85,23 (10,17)	0,89 (0,10)
	RNA	0,85 (0,11)	79,62 (9,37)	0,83 (0,10)
2	AD	0,88 (0,09)	85,13 (9,07)	0,85 (0,11)
	RF	0,92 (0,07)	85,37 (8,10)	0,89 (0,09)
	RNA	0,90 (0,08)	84,71 (8,44)	0,84 (0,10)

As métricas de avaliação de cada algoritmo para a espécie *Tapirira guianensis* se encontram na Tabela 11. Todos os algoritmos, para ambas as bases de dados, não apresentaram métricas de avaliação significativamente diferentes. Os valores de AUC foram parecidos com os obtidos para a espécie *Copaifera langsdorffii*, porém as taxas de verdadeiros positivos para *Tapirira guianensis* foram maiores, indicando maior habilidade dos métodos em classificar locais de ocorrência para esta espécie. A seleção de atributos para esta espécie contribuiu para um melhor desempenho do método de árvore de decisão, levando em consideração a AUC.

Lorena et al. (2011) compararam o desempenho de vários algoritmos de aprendizagem de máquina na modelagem da distribuição de 35 espécies da família Bignoniaceae. O número de dados de presença variou entre 60 e 193 observações, sendo o número de pseudoausência igual ao número de presença para cada espécie. O algoritmo *Random Forest* apresentou melhor performance (AUC) para 29 espécies. Considerando as demais espécies, seu desempenho ficou entre os melhores. Os algoritmos *Suport Vector Machine* e redes neurais

artificiais também demonstraram bom desempenho para todas as espécies, sendo considerados promissores na modelagem de distribuição potencial de espécies.

Williams et al. (2009), ao modelarem a distribuição de seis espécies vegetais raras, também obtiveram desempenho maior, em termos de AUC, para o *Random Forest*. Os resultados obtidos pelo *Maxent*, apesar de inferiores, seguiram a mesma linha do RF, demonstrando concordância entre as predições. O algoritmo *Random Forest* também apresentou maior AUC ao modelar a distribuição de *Pinus sylvestris* na Península Ibérica. Seu desempenho superou o algoritmo árvore de decisão e redes neurais artificiais (GARZÓN et al., 2006).

Apesar de diversos estudos evidenciarem a boa performance do *Random Forest* frente aos demais métodos, ainda é indicado que se compare diferentes técnicas para a modelagem de acordo com a base de dados utilizada. Estudos que compararam diferentes métodos e base de dados obtiveram resultados diversos, evidenciando que o desempenho de cada algoritmo é muito dependente do conjunto de dados utilizados na modelagem (ELITH et al., 2006; SEGURADO; ARAÚJO, 2004).

Tabela 11 Métricas de avaliação obtidas por validação cruzada do conjunto de treinamento para os algoritmos árvore de decisão (AD), *Random Forest* (RF) e redes neurais artificiais (RNA) para a espécie *Tapirira guianensis*

Tapirira guianensis				
Base de dados	Algoritmos	AUC	% Classificado corretamente	Taxa de Verdadeiros positivos
1	AD	0,60 (0,15)	58,35 (12,48)	0,59 (0,21)
	RF	0,67 (0,15)	61,81 (13,06)	0,62 (0,21)
	RNA	0,65 (0,15)	61,33 (12,71)	0,56 (0,22)
2	AD	0,68 (0,15)	63,63 (13,47)	0,55 (0,24)
	RF	0,63 (0,16)	59,85 (14,18)	0,56 (0,23)
	RNA	0,68 (0,14)	60,60(12,04)	0,57 (0,22)

As médias das métricas de avaliação obtidas pela modelagem com a seleção de atributos (base de dados 2), para todas as espécies, foram similares quando comparados às médias obtidas através da base de dados completa, contendo todas as 25 variáveis abióticas (Tabela 8, 9, 10 e 11). Como essa redução no conjunto original dos atributos não afetou o desempenho dos algoritmos, optou-se por utilizar os algoritmos treinados com a seleção de atributos, para a reavaliação com um conjunto independente e posterior aplicação na bacia hidrográfica do rio Grande, a fim de reduzir o esforço computacional e complexidade dos modelos.

Dutra e Carvalho (2008) testaram quatro alternativas de variáveis ambientais para a modelagem de *Amaioua guianensis* utilizando o *Maxent*. Os conjuntos de atributos foram formados por toda a base de dados disponível, apenas com as variáveis climáticas, com as variáveis selecionadas pelo teste Jackknife e com as variáveis selecionadas pela análise de correspondência canônica (CCA). A performance do modelo utilizando apenas as variáveis climáticas foi a pior, seguida pelo conjunto total de variáveis ambientas. As heurísticas de seleção de atributos apresentaram maior desempenho em termos de AUC, sendo o teste Jacknife ligeiramente superior à CCA para o método e base de dados trabalhada.

Os resultados obtidos pela avaliação com o conjunto de teste independente, podem ser visualizados na Tabela 12. Nesta etapa, todos os algoritmos, inclusive o *Maxent*, foram avaliados com a mesma base de dados, sob o conjunto de teste correspondente a 30% dos 197 fragmentos inventariados.

Tabela 12 Valores da área abaixo da curva ROC (AUC) baseado no conjunto de avaliação para os algoritmos árvore de decisão (AD), *Random Forest* (RF) e redes neurais artificiais (RNA)

Espécie	AUC			
	AD	RF	RNA	*MAXENT*
Casearia sylvestris	0,6943	0,7538	0,6613	0,6756
Copaifera langsdorffii	0,5220	0,8401	0,4313	0,6464
Croton floribundus	0,8878	0,9702	0,8435	0,8883
Tapirira guianensis	0,5472	0,7599	0,6434	0,5797

Dentre os três algoritmos inicialmente trabalhados neste estudo (AD, RF e RNA), o algoritmo *Random Forest* foi superior, em termos de AUC, em todas as quatro espécies de acordo com o conjunto independente de teste. Para as espécies *Casearia sylvestris* e *Croton floribundus* o aumento em AUC foi similar, em torno de 11% em relação aos outros algoritmos. Ainda em relação a essas duas espécies, a melhora da capacidade de distinção entre as classes apresentada pelo *Random Forest* foi maior quando comparada às redes neurais. *Copaifera langsdorffii* foi a espécie que obteve maiores acréscimos na AUC quando modelada com *Random Forest*, com 94% e 60% de aumento em AUC em relação ao RNA e AD, respectivamente. O aumento da capacidade preditiva para a espécie *Tapirira guianensis* foi de 38% em relação à árvore de decisão e de 18% em relação às redes neurais. O desempenho das redes neurais foi inferior para as três espécies, exceto para *Tapirira guianensis*, em que o valor de AUC deste algoritmo foi superior ao AD.

O bom desempenho apresentado pelos métodos do tipo árvore (AD e RF) frente às RNA's, pode estar relacionado com a seleção dos atributos realizada através do método de ganho de informação. Os métodos do tipo árvore, usados neste trabalho, utilizam o mesmo critério da heurística de seleção de atributos empregada, que é a minimização da entropia dos dados. A pré-seleção

dos dados seguindo este critério facilitou o trabalho destes algoritmos, fornecendo somente os dados que mais homogeneamente dividem as classes de presença e ausência.

O método de máxima entropia apresentou considerável performance para todas as espécies, quando comparado aos outros métodos. Com relação à árvore de decisão, o *Maxent* superou seu desempenho, em termos de AUC, para três espécies (*Copaifera langsdorffii, Croton floribundus* e *Tapirira guianensis*). O poder preditivo das redes neurais também foi superado pelo *Maxent* em três, das quatro espécies modeladas. Em média o desempenho das redes neurais foi inferior para todas as espécies, exceto para *Tapirira guianensis*, em que sua performance superou o *Maxent* e árvore de decisão.

De acordo com o conjunto de teste, a capacidade preditiva do *Random Forest* foi ainda mais alta, abrindo uma maior vantagem, em termos de AUC, para os demais modelos utilizados. Para todas as espécies, este algoritmo foi o mais preciso. Diante disso, o *Random Forest* foi selecionado para gerar a distribuição potencial das 4 espécies estudadas na bacia do rio Grande. Sua área de presença e ausência estimada foi comparada às geradas pelo *Maxent*, reclassificado de acordo com o limiar mínimo de adequabilidade ambiental do conjunto de treinamento e com o limiar de adequabilidade ambiental de 0,5. A distribuição potencial predita pelo RF, para cada espécie, pode ser visualizada na Figura 10.

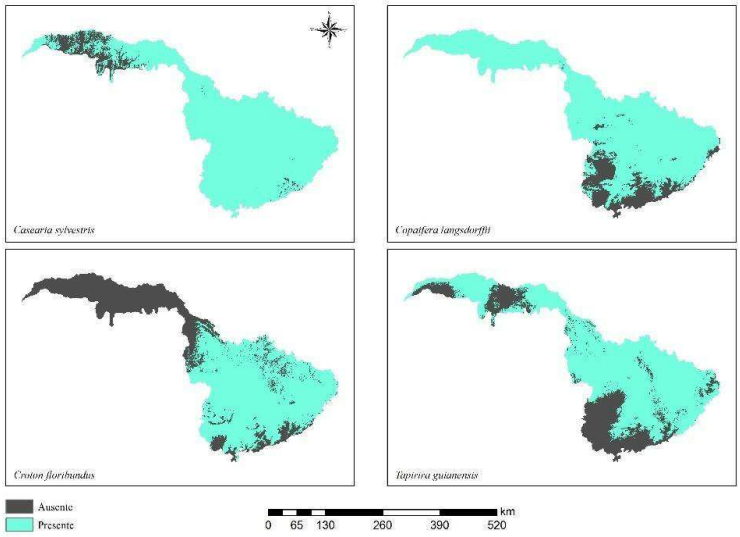

Figura 10 Distribuição potencial das espécies *Casearia sylvestris, Copaifera langsdorffii, Croton floribundus* e *Tapirira guianensis* predita pelo *Random Forest*

Com base nos mapas da distribuição potencial predita pelo RF (Figura 10) e seu respectivo quadro de áreas (Tabela 13), todas as espécies ocuparam mais de 50% da área da bacia. Este resultado confirma o potencial destas para a restauração de áreas desflorestadas na região. A espécie *Casearia sylvestris* ocorreu em grande parte da bacia (92,43%), com exceção do triângulo minero, onde estão localizadas algumas manchas de florestas Estacionais Deciduais. Já a espécie *Copaifera langsdorffii*, que em conjunto com a primeira configuram as espécies de maior distribuição na região, teve sua ausência limitada à região Sul, caracterizada por altos valores de altitude.

As espécies *Croton floribundus* e *Tapirira guianensis* obtiveram menor área de ocupação na bacia do Rio Grande quando comparada às demais. A distribuição potencial de *Croton floribundus,* que representa 65,75% da bacia,

ocorreu principalmente em regiões de grandes altitudes e alta pluviosidade, sendo ausente em todo o Triângulo mineiro. *Tapirira guianensis* apresentou distribuição ao longo de toda a bacia (74,06%), com exceção da parte Sul, com alta altitude, e regiões do Triângulo mineiro.

Tabela 13 Quadro das áreas potenciais de presença (P%) e ausência (A%) em porcentagem relativa à área da bacia do rio Grande, preditas pelo *Random Forest* e *Maxent* (limiar mínimo de presença e limiar de 0,5)

Espécie	*Random Forest*		*Maxent*			
			Limiar mínimo de presença		Limiar 0,5	
	P%	A%	P%	A%	P%	A%
Casearia sylvestris	92,43	7,57	99,98	0,02	75,06	24,94
Copaifera langsdorffii	85,93	14,07	97,53	2,47	61,15	38,85
Croton floribundus	65,75	34,25	77,44	22,56	53,94	46,06
Tapirira guianensis	74,06	25,94	100,00	0,00	60,08	39,92

Algumas pesquisas apontam que somente a utilização de variáveis bioclimáticas para a modelagem de distribuição de espécies (que é o caso, em geral, desta aplicação) consegue explicar a distribuição de uma espécie em grandes escalas (regional, continental e até global), porém, possuindo alto grau de generalização, superestimando a área de ocorrência da espécie (DUTRA; CARVALHO, 2008; GUISAN; ZIMMERMANN, 2000).

Após a geração dos mapas de adequabilidade ambiental por espécie, pelo *Maxent*, estes foram reclassificados pelos dois limiares, anteriormente selecionados para o estudo. O valor mínimo de adequabilidade ambiental encontrado nos dados de presença do conjunto de treinamento foi, em média, de 0,15 para todas as espécies. Os mapas reclassificados de acordo com o limiar mínimo de presença e limiar de 0,5, podem ser visualizados nas Figuras 11 e 12, respectivamente.

;

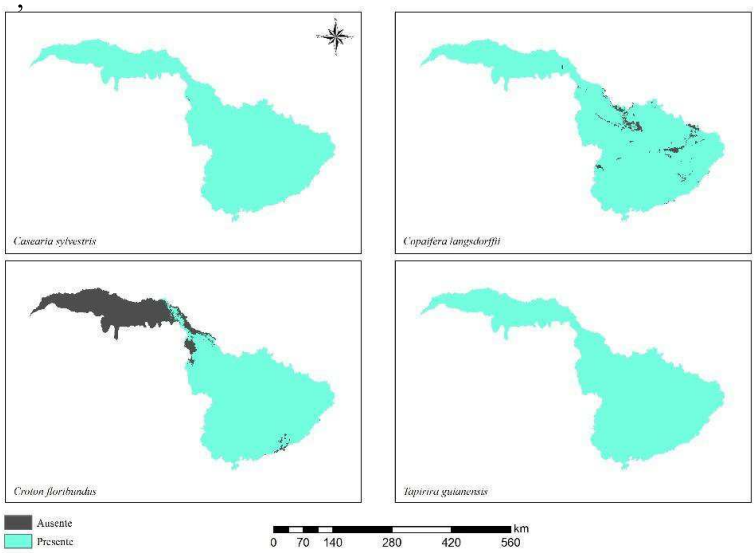

Figura 11 Distribuição potencial das espécies *Casearia sylvestris, Copaifera langsdorffii, Croton floribundus* e *Tapirira guianensis* predita pelo *Maxent*, considerando o valor mínimo de adequabilidade ambiental do conjunto de treinamento para a reclassificação do mapa em presença e ausência

Comparando as áreas de distribuição do RF em relação às áreas obtidas pelo *Maxent* com o limiar mínimo de presença (Tabela 13), verifica-se que o último produz uma maior estimativa da área de ocorrência para todas as espécies. Esta diferença está diretamente relacionada aos tipos de dados utilizados na modelagem. Modelos de *presence-only* (*Maxent*) tendem a produzir maiores áreas de ocorrência quando comparados às técnicas que utilizam dados de presença e ausência (PHILLIPS; ANDERSON; SCHAPIRE, 2006). O limiar de ocorrência utilizado para a geração do mapa binário também interfere diretamente na área predita. Em média, as ocorrências das espécies

apresentaram um valor mínimo de adequabilidade ambiental de 0,15, o que lhes garantiu ampla área de abrangência.

As espécies *Casearia sylvestris*, *Copaifera langsdorffii* e *Tapirira guianensis*, de acordo com este limiar de reclassificação, podem ocupar toda a região da bacia. Essas espécies abrangem, respectivamente, 99,98%, 97,53% e 100% da bacia do Rio Grande. Já a distribuição de *Croton floribundus* (65,75%), que foi similar à produzida pelo RF, fica restrita à parte Centro-sul da bacia.

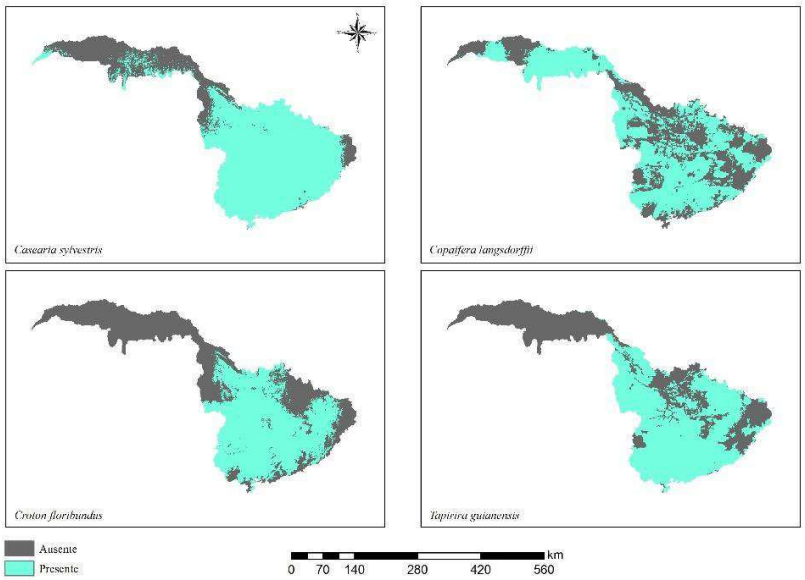

Figura 12 Distribuição potencial das espécies *Casearia sylvestris, Copaifera langsdorffii, Croton floribundus* e *Tapirira guianensis* predita pelo *Maxent*, considerando o valor de adequabilidade ambiental 0,5 para a reclassificação do mapa em presença e ausência

Por meio da reclassificação adotando um limiar de 0,5, o *Maxent* apresentou áreas de distribuição potencial, das quatro espécies modeladas, inferiores às obtidas pelo *Random Forest*. No entanto, diante do disposto, todas

as espécies apresentaram área de abrangência superior a 50% da bacia (Figura 12). A área de ausência de *Casearia sylvestris* e *Croton floribundus* foi maior no Triângulo mineiro, seguindo a mesma tendência da distribuição potencial predita pelo RF. Já a distribuição potencial das espécies *Copaifera langsdorffii* e *Tapirira guianensis* obtidas pelo *Maxent* (com limiar 0,5) foram discrepantes quando comparadas com o RF. Enquanto a distribuição potencial predita pelo RF foi restrita, principalmente, na região Sul da bacia, a distribuição potencial predita para essas duas espécies pelo *Maxent*, foi restrita no Triângulo mineiro e em manchas na região Central da bacia.

5 CONCLUSÕES

As espécies *Casearia sylvestris, Copaifera langsdorffii, Croton Floribundus* e *Tapirira guianensis* que, de acordo com a literatura são potenciais para a revitalização de matas ciliares, apresentaram grande abundância e distribuição no estado de Minas Gerais.

De acordo com o método ganho de informação, a espécie *Croton floribundus* apresentou uma distribuição mais homogeneamente dividida pelos atributos selecionados (maior ganho de informação), confirmando seu caráter especialista. A grande maioria dos atributos selecionados para esta espécie foi relacionada à temperatura (°C). As espécies *Copaifera langsdorffii* e *Tapirira guianensis* apresentaram grande heterogeneidade das classes de presença/ausência (baixo ganho de informação), o que pode ser explicado pela ampla distribuição das espécies no território mineiro, abrangendo grande espectro de características ambientais (generalistas). A espécie *Casearia sylvestris*, apesar de possuir extensa área de ocorrência, apresentou maior homogeneidade dos dados quando submetidos às diferentes variáveis abióticas.

O método de seleção dos atributos abióticos não obteve melhoras significativas em relação ao desempenho dos modelos quando treinados com todos os atributos disponíveis.

De acordo com cada espécie modelada e com a validação cruzada sob o conjunto de treinamento, os algoritmos árvore de decisão *Random Forest* e redes neurais artificiais apresentaram desempenho semelhantes, confirmando a premissa de que a capacidade preditiva de cada algoritmo varia de acordo com a espécie modelada. No entanto, o algoritmo *Random Forest* apresentou performance considerável para todas as espécies. Por meio da avaliação realizada com o conjunto de teste (30%), os resultados obtidos pelo *Random*

Forest superaram todos os outros algoritmos, incluindo o *Maxent*, para todas as espécies modeladas.

O algoritmo *Random Forest*, através das avaliações obtidas pela validação cruzada e com o conjunto de teste, apresentou melhores valores de AUC para todas as espécies. Este algoritmo foi então selecionado para modelar a distribuição potencial dessas espécies na bacia do Rio Grande.

A área predita pelo *Random Forest* foi menor do que a área predita pelo *Maxent* quando utilizado o limiar mínimo de adequabilidade ambiental presente no conjunto de treinamento; e menor quando a adequabilidade ambiental é reclassificada adotando o limiar 0,5.

REFERÊNCIAS

ALLOUCHE, O.; TSOAR, A.; KADMON, R. Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). **Journal of Applied Ecology**, Oxford, v. 43, n. 6, p. 1223-1232, Dec. 2006.

ANDERSON, R. P.; LEW, D.; PETERSON, A. T. Evaluating predictive modeling of species' distributions: Criteria for selecting optimal models. **Ecological Modelling**, Amsterdam, v. 162, n. 3, p. 211-232, Apr. 2003.

BEAUMONT, L. J.; HUGHES, L. Potential changes in the distributions of latitudinally restricted Australian butterfly species in response to climate change. **Global Change Biology**, Oxford, v. 8, n. 10, p. 954-971, Oct. 2002.

BISHOP, M. **Neural networks for pattern recognition Christopher**. Oxford: Oxford University, 1995. 482 p.

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

BREIMAN, L. et al. **Classification and regression trees**. Belmont: Wadsworth, 1984. 358 p.

BRERETON, R.; BENNETT, S.; MANSERGH, I. Enhanced greenhouse climate change and its potential effect on selected fauna of south-eastern Australia: a trend analysis. **Biological Conservation**, Essex, v. 72, n. 3, p. 339-354, July 1995.

BUSBY, J. R. BIOCLIM a bioclimatic analysis and prediction system. In: MARGULES, C. R.; AUSTIN, M. P. (Ed.). **Nature conservation:** cost effective biological surveys and data analysis. Melbourne: CSIRO, 1991. p. 64-68.

CARPENTER, G.; GILLISON, A. N.; WINTER, J. DOMAIN: a flexible modelling procedure for mapping potential distributions of plants and animals. **Biodiversity and Conservation**, Bengaluru, v. 11, n. 6, p. 2239-2274, Dec. 1993.

CARVALHO, D. A. et al. Variações florísticas e estruturais do componente arbóreo de uma floresta ombrófila alto-montana às margens do rio Grande, Bocaina de Minas, MG, Brasil. **Acta Botanica Brasilica**, São Paulo, v. 19, n. 1, p. 91-109, 2005.

CARVALHO, L. G. et al. Clima. In: SCOLFORO, J. R. S.; CARVALHO, L. M. T.; OLIVEIRA, A. D. (Ed.). **Zonemento ecológido-econômico do estado de Minas Gerais:** componentes geofísico e biótico. Lavras: UFLA, 2008. p. 89-101.

CARVALHO, P. E. R. **Copaíba**. Colombo: EMBRAPA Florestas, 2005. 6 p. (Circular Técnica, 114).

CLUTER, R. D. et al. *Random Forest* for classification in ecology. **Ecology**, Durham, v. 88, n. 11, p. 2783-2792, Nov. 2007.

DURIGAN, G. et al. The vegetation of priority areas for cerrado conservation in São Paulo State, Brazil. **Edinburgh Journal of Botany**, Edinburgh, v. 60, n. 3, p. 217-241, Nov. 2003.

DUTRA, G. C. **Modelagem da distribuição geográfica de fitofisionomias no Estado de Minas Gerais**. 2009. 59 p. Tese (Doutorado em Engenharia Florestal) - Universidade Federal de Lavras, Lavras, 2009.

DUTRA, G. C.; CARVALHO, L. M. T. Modelos de distribuição geográfica de Amaioua guianensis Aubl. Em Minas Gerais, Brasil. **Ambiência**, Guarapuava, v. 4, p. 47-55, 2008. Edição especial.

ELITH, J. et al. Novel methods improve prediction of species' distributions from occurrence data. **Ecography**, Copenhagen, v. 29, n. 2, p. 129-151, Apr. 2006.

ENGLER, R.; GUISAN, A.; RECHSTEINER, L. An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. **Journal of Applied Ecology**, Oxford, v. 41, n. 3, p. 263-274, June 2004.

FIELDING, A. H.; BELL, J. F. A review of methods for the assesment pf prediction errors in conservation presence/abscene models. **Environmental Conservation**, Lausanne, v. 24, n. 1, p. 38-49, Mar. 1997.

FUKUDA, S. et al. Habitat prediction and knowledge extraction for spawning European grayling (*Thymallus thymallus* L.) using a broad range of species distribution models. **Enviromental Modelling & Software**, New York, v. 47, n. 1, p. 1-6, Sept. 2013.

GARZON, M. B.; DIOS, R. S.; OLLERO, H. S. Effects of climate change on the distribution of Iberian tree species. **Applied Vegetation Science**, San Francisco, v. 11, n. 2, p. 169-178, Apr. 2008.

GARZÓN, M. B. et al. Predicting habitat suitability with machine learning models: the potential area of *Pinus sylvestris* L. in the Iberian Peninsula. **Ecological Modelling**, Amsterdam, v. 197, n. 3/4, p. 383-393, Aug. 2006.

GIANNINI, T. C. et al. Desafios atuais na modelagem preditiva de distribuição de espécies. **Rodriguésia**, São Paulo, v. 63, n. 3, p. 733-749, 2012. Disponível em: <http://rodriguesia-seer.jbrj.gov.br/index.php/rodriguesia/article/view/339>. Acesso em: 22 out. 2013.

GUISAN, A.; ZIMMERMANN, N. E. Predictive habitat distribution models in ecology. **Ecological Modelling**, Amsterdam, v. 135, n. 2/3, p. 147-186, Dec. 2000.

HERINGER, E. P. et al. A flora do cerrado. In: SIMPÓSIO SOBRE O CERRADO, 4., 1977, São Paulo. **Anais...** São Paulo: USP, 1977. p. 211-232.

HERNÁNDEZ, L. et al. Changes in structure and composition of evergreen forests on na altitudinal gradient in the Venezuelan Guayana Shield. **Revista de Biologia Tropical**, San José, v. 60, n. 1, p. 11-33, 2012.

HIJMANS, R. J. et al. Very high resolution interpolated climate surfaces for global land areas. **International Journal of Climatology**, Chichester, v. 25, n. 15, p. 1965-1978, Nov. 2005.

HIRZEL, A. et al. Ecological-niche factors analysis: how to compute habitat-suitability map without absence data? **Ecology**, Durham, v. 83, n. 8, p. 2027-2036, Aug. 2002.

HOMEIER, J. et al. Tree diversity, forest structure and productivity along altitudinal and topographical gradients in a species-rich Ecuadorian Montane Rain Forest. **Biotropica**, Washington, v. 42, n. 2, p. 140-148, 2010.

HUTCHINSON, G. E. Concluding remarks. In: SYMPOSIUM ON QUANTITATIVE BIOLOGY, 22., 1957, Cold Spring Harbour. **Proceedings...** Cold Spring Harbour, 1957. p. 415-427.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. **Mapa de Biomas do Brasil, primeira aproximação**. Rio de Janeiro, 2004. Disponível em: <http://www.ibge.gov.br>. Acesso em: 10 nov. 2014.

INSTITUTO ESTADUAL DE FLORESTAS. **Cobertura vegetal de Minas Gerais**. Disponível em: <http://www.ief.mg.gov.br/florestas>. Acesso em: 15 jan. 2015.

LIMA, D. A. The caatinga dominium. **Revista Brasileira de Botânica**, São Paulo, v. 4, p. 149-153, 1981.

LORENA, A. C. et al. Comparing machine learning classifiers in potential distribution modelling. **Expert Systems With Applications**, New York, v. 38, n. 5, p. 5268-5275, 2011.

LORENZI, H. **Árvores brasileiras:** manual de identificação e cultivo de plantas arbóreas nativas do Brasil. 5. ed. Nova Odessa: Instituto Plantarum, 2008. v. 1, 368 p.

MANEL, S.; DIAS, J. M.; ORMEROD, S. J. Comparing discriminant analysis, neural networks and logistic regression for predicting species distributions: a case study with a Himalayan river bird. **Ecological Modelling**, Amsterdam, v. 120, n. 2/3, p. 337-347, Aug. 1999.

MARCO JÚNIOR, P.; SIQUEIRA, M. F. Como determinar a distribuição potencial de espécies sob uma abordagem conservacionista? **Megadiversidade**, São Paulo, v. 5, n. 1/2, p. 65-76, 2009.

MCPHERSON, J. M.; JETZ, W.; ROGERS, D. J. The effects os species' range sizes on the accuracy of distribution models: ecological phenomenon or statistical artefact? **Journal of Applied Ecology**, Oxford, v. 41, n. 5, p. 811-823, Oct. 2004.

MELLO, J. M.; SCOLFORO, J. R. S.; CARVALHO, L. M. T. **Floresta estacional decidual:** florística, estrutura, diversidade, similaridade, distribuição diamétrica e de altura, volumetria, tendências de crescimento e manejo florestal. Lavras: UFLA, 2008. 265 p.

MITCHELL, T. M. **Machine learning**. Boston: WCB/McGraw-Hill, 1997. 414 p.

OLIVEIRA FILHO, A. T. et al. **Estudos florísticos e fitossociológicos em remanescentes de matas ciliares do alto e médio rio Grande**. Belo Horizonte: CEMIG, 1995. 27 p.

OLIVEIRA FILHO, A. T.; RATTER, J. A. Padrões florísticos das matas ciliares da região do cerrado e a evolução das paisagens do Brasil central durante o Quaternário Tardio. In: RODRIGUES, R. R.; LEITÃO FILHO, H. F. (Ed.). **Matas ciliares:** conservação e recuperação. São Paulo: EDUSP/FAPESP, 2000. p. 73-89.

PEARSON, R. G. et al. Predicting species distributions form small numbers of ocurrence records: a test case using cryptic geckos in Madagascar. **Journal of Biogeography**, Oxford, v. 34, n. 1, p. 102-117, 2007.

PETERSON, A. T.; PAPES, M.; EATON, M. Transferability and model evaluation in ecological niche modeling: a comparison of GARP and Maxent. **Ecography**, Copenhagen, v. 30, n. 4, p. 550-560, Aug. 2007.

PHILLIPS, S. J.; ANDERSON, R. P.; SCHAPIRE, R. E. Maximum entropy modeling of species geographic distributions. **Ecological Modelling**, Amsterdam, v. 190, n. 3/4, p. 231-259, Jan. 2006.

POUTEAU, R. et al. Support vctor machines to map rare and endangered native plants in Pacific islands forests. **Ecological Informatics**, New York, v. 9, p. 37-46, May 2012.

PRASAD, A. M.; IVERSON, L. R.; LIAW, A. Newer classification and regression tree techniques: bagging and random forests for ecological prediction. **Ecosystems**, New York, v. 9, n. 2, p. 181-199, Mar. 2006.

PULLIAM, H. R. On the relationship between niche and distribution. **Ecology Letters**, Oxford, v. 3, n. 4, p. 349-361, July 2000.

QUINLAN, J. R. **C4.5:** programs for machine learning. San Francisco: M. Kaufmann, 1993. 299 p.

RATTER, J. A.; BRIDGEWATER, S.; RIBEIRO, J. F. Analysis of the floristic composition of the Brazilian Cerrado vegetation III: comparison of the wood vegetation of 376 areas. **Edinburgh Journal of Botany**, Edinburgh, v. 60, n. 1, p. 57-109, Mar. 2003.

RATTER, J. A. et al. Analysis of the floristic composition of the Brazilian Cerrado vegetation II: comparison of the woody vegetation of 98 areas. **Edinburgh Journal of Botany**, Edinburgh, v. 53, n. 2, p. 153-180, July 1996.

RATTER, J. A.; RIBEIRO, J. F.; BRIDGEWATER, S. The Brazilian cerrado vegetation and threats to its biodiversity. **Annals of Botany**, London, v. 80, n. 3, p. 223-230, May 1997.

RIZZINI, C. T. **Tratado de fitogeografia do Brasil**. 2. ed. Rio de Janeiro: Âmbito Cultural, 1997. 747 p.

RODRIGUES, F. A. **Um método de referância para análise de desempenho preditivo de algoritmos de modelagem de distribuição de espécies**. 2012. 150 p. Tese (Doutorado em Ciências) - Escola Politécnica da Universidade de São Paulo, São Paulo, 2012.

RODRIGUES, L. A. et al. Florística e estrutura da comunidade arbórea de um fragmento florestal em Luminárias, MG. **Acta Botânica Brasilica**, São Paulo, v. 17, n. 1, p. 71-87, jan./mar. 2003.

ROSENBLAT, F. The perceptron: a probabilistic model for information storage and organization in the brain. **Psychological Review**, Washington, v. 65, n. 6, p. 386-408, Nov. 1958.

SCOLFORO, J. R. S.; CARVALHO, L. M. T. (Ed.). **Mapeamento e inventário da flora nativa e dos reflorestamentos de Minas Gerais**. Lavras: UFLA, 2008. 287 p.

SCOLFORO, J. R. S.; MELLO, J. M.; SILVA, C. P. C. (Ed.). **Inventário florestal de Minas Gerais:** floresta estacional semidecidual e ombrófila: florística, estrutura, diversidade, similaridade, distribuição diamétrica e de altura, volumetria, tendências de crescimento e áreas aptas para manejo florestal. Lavras: UFLA, 2008. 1029 p.

SEGURADO, P.; ARAÚJO, M. B. An evaluation of methods for modelling species distributions. **Journal of Biogeography**, Oxford, v. 31, n. 10, p. 1555-1568, Oct. 2004.

SILVA, C. P. C. et al. Composição florística na Floresta Estacional Decidual. In: SCOLFORO, J. R. S.; MELLO, J. M.; SILVA, C. P. C. (Ed.). **Inventário florestal de Minas Gerais:** floresta estacional decidual: florística, estrutura, diversidade, similaridade, distribuição diamétrica e de altura, volumetria, tendências de crescimento e manejo florestal. Lavras: UFLA, 2008a. p. 65-85.

SILVA, C. P. C. et al. Composição florística na Floresta Estacional Semidecidual e Floresta Ombrófila. In: SCOLFORO, J. R. S.; MELLO, J. M.; SILVA, C. P. C. (Ed.). **Inventário florestal de Minas Gerais:** floresta estacional semidecidual e ombrófila: florística, estrutura, diversidade, similaridade, distribuição diamétrica e de altura, volumetria, tendências de crescimento e áreas aptas para manejo florestal. Lavras: UFLA, 2008b. p. 193-228.

SILVA, W. G. et al. Relief influence on the spatial distribution of the Atlantic Forest cover on the Ibiúna Plateau, SP. **Brazilian Journal of Biology**, São Carlos, v. 67, n. 3, p. 631-637, Aug. 2007.

SILVA JÚNIOR, M. Relationships between the three communities of the Pitoco, Monjolo and Taquara gallery forest and environmental factors. In: INTERNATIONAL SYMPOSIUM ON ASSESSMENT AND MONITORING OF FORESTS IN TROPICAL DRY REGIONS WITH SPECIAL REFERENCE TO GALLERY FORESTS, 1., 1997, Brasília. **Proceedings…** Brasília: UnB, 1997. p. 287-298.

SILVA-LUIZ, C. L.; PIRANI, J. R. **Lista de espécies da flora do Brasil**. Rio de Janeiro: Jardim Botânico do Rio de Janeiro, 2013. Disponível em: <http://floradobrasil.jbrj.gov.br/jabot/floradobrasil/FB44>. Acesso em: 29 dez. 2014.

SOBERÓN, J.; PETERSON, A. T. Interpretation of models of fundamental ecological niches and species distributional areas. **Biodiversity Informatics**, Lawrence, v. 2, p. 1-10, 2005.

STOCKWELL, D.; PETERS, D. The garp modelling system: problems and solutions to automated spatial prediction. **International Journal of Geographical Information Science**, London, v. 13, n. 2, p. 143-158, 1999.

STOCKWELL, D. R. B.; NOBLE, I. R. Induction of sets of rules form animal distribution data: a robust and informative method of analysis. **Mathematics and Computers in Simulation**, Amsterdam, v. 33, n. 5/6, p. 385-390, Apr. 1992.

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

TAMVAKIS, A. et al. Optimizing biodiversity prediction from abiotic parameters. **Enviromental Modelling & Software**, New York, v. 53, p. 112-120, Mar. 2014.

TERRIBILE, L. C.; DINIZ-FILHO, J. A. F.; MARCO JUNIOR, P. de. How many studies are necessary to compare niche-based models for geographic distributions?: inductive reasoning may fail at the end. **Brazilian Journal of Biology**, São Carlos, v. 70, n. 2, p. 263-269, 2010.

TIRELLI, T.; GAMBA, M.; PESSANI, D. Support vector machines to model presence/absence of Alburnus alburnus alborella (Teleostea, Cyprinidae) in North-Western Italy: comparison with other machine learning techniques. **Revista Comptes Rendus Biologies**, Paris, v. 335, n. 10/11, p. 680-868, Oct./Nov. 2012.

VANDERWAL, J. et al. Selecting pseudoabsence data for presence-only distribution modeling: how far should you stray from what you know? **Ecological Modelling**, Amsterdam, v. 220, n. 4, p. 589-594, Feb. 2009.

VAPNIK, V. **The nature of statistical learning theory**. Berlin: Springer Verlag, 1995. 314 p.

VAYSSIÈRES, M. P.; PLANT, R. E.; ALLEN-DIAZ, B. H. Classification trees: an alternative non-parametric approach for predicting species distributions. **Journal of Vegetation Science**, Knivsta, v. 11, n. 5, p. 679-694, 2000.

VILELA, E. A. Caracterização estrutural de floresta ripária do Alto Rio Grande, em Madre de Deus de Minas, MG. **Cerne**, Lavras, v. 6, n. 2, p. 41-54, 2000.

VILELA, E. A. et al. Espécies de matas ciliares com potencial para estudos de revegetação no alto Rio Grande, Sul de Minas. **Revista Árvore**, Viçosa, MG, v. 17, n. 2, p. 117-128, 1993.

WALTER, B. M. T.; RIBEIRO, J. F. Spatial floristic patterns in galery forests in the Cerrado Region, Brazil. In: IMAÑA-ENCINAS, J.; KLEINN, C. (Ed.). **Proceedings of the international symposium on assessment and monitoring of forests in tropical dry regions with special reference to gallery forests**. Brasília: UnB, 1997. p. 339-349.

WILLIAMS, J. N. et al. Using species distribution models to predict new ocurrences for rare plants. **Diversity and Distributions**, London, v. 15, n. 4, p. 565-576, July 2009.

WITTEN, I. W.; FRANK, E.; HALL, M. A. **Data mining:** practical machine learning tools and techniques. 3rd ed. New York: M. Kaufmann, 2011. 664 p. (The Morgan Kaufmann Series in Data Management Systems).