

# MODELAGEM DA DISTRIBUIÇÃO POTENCIAL DE ESPÉCIES ARBÓREAS NA BACIA DO RIO GRANDE PARA PROGRAMAS DE REVITALIZAÇÃO

**Mônica Canaan Carvalho**

Engenheira Florestal, Mestre em Engenharia Florestal, UFLA, Universidade Federal de Lavras, Departamento de Ciências Florestais DCF / UFLA, Caixa Postal 3037, CEP 37200-000, Lavras MG.  
monicacanaan@yahoo.com.br

**Tayrine Vieira Martins**

Engenheira Florestal, UFLA, Universidade Federal de Lavras, Departamento de Ciências Florestais DCF / UFLA, Caixa Postal 3037, CEP 37200-000, Lavras MG. tatavmartins@gmail.com

**Nathalia de Paiva Mendonça**

Engenheira Florestal,, UFLA, Universidade Federal de Lavras, Departamento de Ciências Florestais DCF / UFLA, Caixa Postal 3037, CEP 37200-000, Lavras MG. nathalia.p.mendonca@hotmail.com

**Lucas Rezende Gomide**

Engenheiro Florestal, Doutor em Engenharia Florestal, UFLA, Universidade Federal de Lavras, Departamento de Ciências Florestais DCF / UFLA, Caixa Postal 3037, CEP 37200-000, Lavras MG. lucasgomide@dcf.ufla.br

**Rubens Manoel dos Santos**

Biólogo, Doutor em Engenharia Florestal, UFLA, Universidade Federal de Lavras, Departamento de Ciências Florestais DCF / UFLA, Caixa Postal 3037, CEP 37200-000, Lavras MG. rubensmanoel@dcf.ufla.br

**José Márcio de Mello**

Engenheiro Florestal, Doutor em Engenharia Florestal, UFLA, Universidade Federal de Lavras, Departamento de Ciências Florestais DCF / UFLA, Caixa Postal 3037, CEP 37200-000, Lavras MG. josemarcio@dcf.ufla.br

## RESUMO

O presente trabalho teve como objetivo aplicar o algoritmo *Random Forest* na modelagem da distribuição potencial de 4 espécies arbóreas abundantes na bacia do Rio Grande. Utilizou-se os dados coletados em 27 fragmentos dentro da bacia, nos quais foram computadas a presença/ausência de cada espécie. As variáveis ambientais utilizadas na modelagem foram relativas à temperatura, precipitação, altitude, solo e relevo, totalizando 25 variáveis abióticas. O treinamento e avaliação do algoritmo se deu por validação cruzada (3 folds), na qual foram avaliadas as métricas: área abaixo da curva (AUC), porcentagem de dados classificados corretamente e taxa de verdadeiros positivos para ambas as classes. Em termos de AUC, o algoritmo *Random Forest* apresentou desempenho moderado para a espécie *Psychotria vellosiana* e bom desempenho para as espécies *Amaioua intermedia*, *Cabralea canjerana*, *Myrcia splendens* e *Psychotria vellosiana*. Apesar dos bons resultados para todas as espécies, seu desempenho mostrou ser altamente influenciado pelo desbalanceamento das classes.

**PALAVRAS-CHAVE:** Fitogeografia, Nicho Fundamental, *Random Forest*.

**Área principal:** AG & MA PO na agricultura e meio ambiente

## ABSTRACT

This study aimed to apply the algorithm *Random Forest* in potential distribution modeling of four abundant tree species in the Rio Grande watershed. We used data collected on 27 fragments within the watershed, which were computed in the presence / absence of each species. Environmental variables used in the modeling were derived from temperature, rainfall, altitude, soil and relief, totaling 25 abiotic variables. The training and evaluating the algorithm is given by cross-validation (3 folds), in which the metrics were evaluated: area under the curve (AUC), percentage of correctly classified data and true positive rate for both classes. In terms of AUC, the *Random Forest* algorithm presented moderate performance for *Psychotria vellosiana* and good performance for *Amaioua intermedia*, *Cabralea canjerana* and *Myrcia splendens*. Despite the good results for all species, their performance proved to be highly influenced by the imbalance of classes.

**KEYWORDS:** Phytogeography, Fundamental Niche, *Random Forest*

**Main Area:** AG & MA OR in agriculture and environment

## 1. Introdução

A modelagem da distribuição de espécies é um método da biogeografia que estima a distribuição potencial de determinada espécie baseada em características ambientais. Sua fundamentação ecológica é baseada no conceito de nicho (Hutchinson, 1957), definido um espaço de  $n$  dimensões, limitado pelas condições bióticas e abióticas necessárias para sobrevivência e reprodução de cada espécie.

Este método associa dados de ocorrência da espécie (coordenadas geográficas) com variáveis ambientais (obtidas através de imagens de Sensoriamento Remoto e dados de um SIG), predizendo ambientes potencialmente adequados, onde em teoria, uma população possa se manter viável (Anderson e Peterson, 2003). Devido à complexidade da natureza e na obtenção de condições bióticas (interações entre espécies, mecanismos de dispersão, barreiras históricas, recursos, entre outros), estes dados não são empregados de maneira representativa na modelagem. Resultando em um ambiente predito de ocorrência da espécie muito menos restrito que a distribuição real desta (potencial), visto que considera apenas algumas características abióticas e pressupõe que as condições bióticas são satisfatórias. Este ambiente na ecologia é denominado nicho fundamental, e compreende o espaço geográfico onde as variáveis utilizadas na modelagem sustentam a ocorrência da espécie.

A estimativa do nicho fundamental através da modelagem, auxilia em estudos sobre ambientes potenciais para espécies invasoras (Peterson et al., 2006), conservação de espécies raras ou ameaçadas (Cluter et al., 2004; Marmion et al., 2009), estimativa de área prioritárias para conservação (Chen, 2009), reintrodução de espécies (Hirzel et al., 2002), impactos de mudanças climáticas (Araújo et al., 2008), entre outros. Em âmbito florestal, a modelagem da distribuição de espécies pode constituir uma ferramenta prática e precisa em planos de revitalização de áreas desflorestadas. A partir de levantamentos fitossociológicos e de variáveis ambientais importantes para a ocorrência da espécie, é possível indicar, com base nos dados modelados, qual espécie estará apta para re(ocupar) determinado local.

Diversos são os métodos disponíveis para modelar a distribuição de espécies (Elith et al., 2006; Williams et al., 2009). Dentre os quais, destaca-se o algoritmo *Random Forest* (Breiman, 2001), método da área de aprendizagem de máquina, que consegue aprender a partir de exemplos e fazer predições para novos dados. 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 (Garzon et al., 2006; Cluter et al., 2007; Lorena et al., 2011; Fukuda et al., 2013).

Diante das devastadoras ações antropogênicas sobre o ecossistema florestal e seus impactos negativos sobre os serviços ecológicos prestados, a utilização da modelagem da distribuição de espécies auxilia na proteção e conservação da biodiversidade, bem como na escolha das espécies na restauração de ecossistemas. Com o objetivo de auxiliar em planos de revegetação, o presente estudo irá desenvolver um modelo preditivo, utilizando o *Random Forest*, da distribuição de 4 espécies arbóreas nativas e abundantes para a bacia hidrográfica do Rio Grande, no estado de Minas Gerais, Brasil.

## 2. Material e Métodos

### 2.1 Área de estudo

A região a ser modelada, bem como a localização dos fragmentos inventariados, situa-se na bacia hidrográfica do Rio Grande, no território mineiro. A bacia compreende a região Sul e

Triângulo Mineiro. Apresenta uma área de 86.110,02 km<sup>2</sup>, referente à 14,68% do território de Minas Gerais (Figura 1). Ocupa uma região com as maiores altitudes do estado, variando de 300 até 2.600 m. A parte Sul da bacia está inserida em áreas de maiores cotas altimétricas, enquanto o Triângulo Mineiro é caracterizado por 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 1) (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).

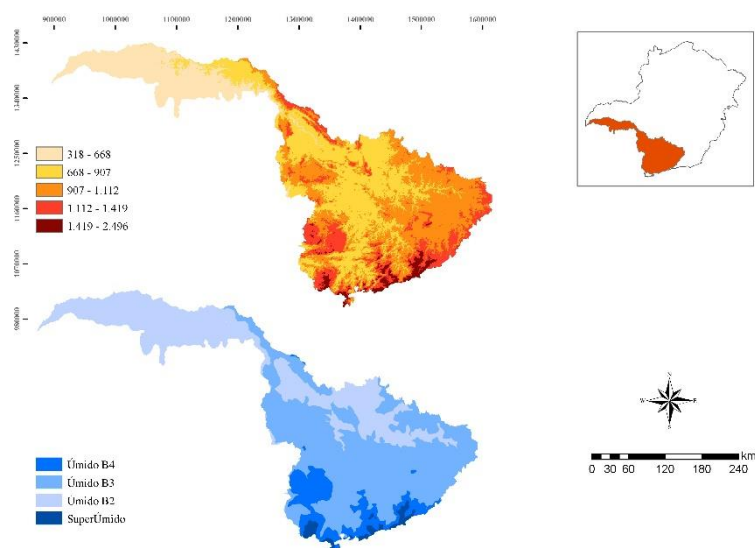


Figura 1 Classes altimétricas e zoneamento climático da bacia hidrográfica do Rio Grande, respectivamente

## 2.2 Dados de ocorrência

Os dados de ocorrência das espécies são provenientes de 27 fragmentos inventariados nos anos de 2013/2014 pelo projeto “Modelo Fitogeográfico da Bacia do Rio Grande”, realizado pela UFLA com o apoio da Companhia Energética de Minas Gerais (CEMIG) (Figura 2). A amostragem dos fragmentos dentro da bacia procurou compreender as diferentes variações nas classes de altitude e clima, além de priorizar a representatividade das matas ciliares. Em todos os fragmentos inventariados, foram considerados apenas os indivíduos com CAP (circunferência a altura do peito) maior que 15,7 cm.

A seleção das espécies foi embasada nos dados obtidos do inventário. Procurou-se espécies nativas e de grande abundância, potenciais para restauração de matas ciliares dentro da bacia.

Foram selecionadas quatro espécies arbóreas: *Amaioua intermedia*, *Cabralea canjerana*, *Myrcia splendens* e *Psychotria vellosiana*.

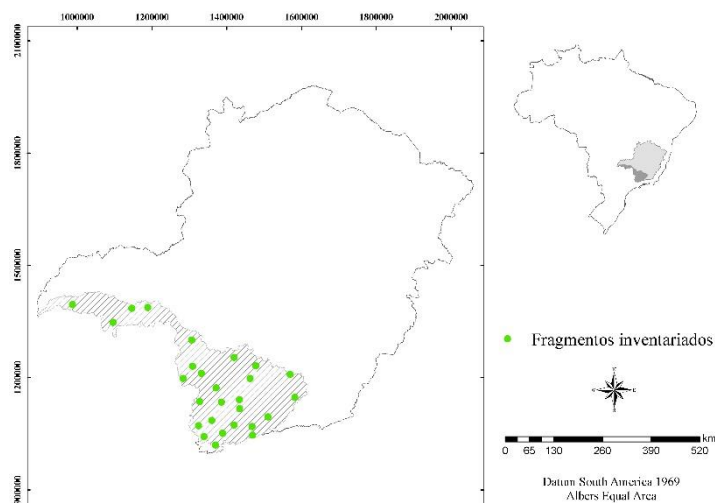


Figura 2 Distribuição dos fragmentos utilizados na modelagem na bacia do Rio Grande

Os dados do inventário de cada fragmento foram trabalhados como sendo uma única unidade amostral, independentemente do número de parcelas empregadas e área do fragmento, sendo representado por seu ponto central. Neste ponto foram extraídas informações sobre as condições ambientais, e computada a presença (1) ou ausência (0) das quatro espécies selecionadas para cada fragmento.

### 2.3 Variáveis ambientais

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) e estão relacionadas com o solo (tipo, textura, teor de matéria orgânica e regime hídrico) e relevo.

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.

### 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 o algoritmo CART (*Classification And Regression Trees*). 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 entropia do conjunto de dados quando submetido à divisão de acordo com um atributo.

O algoritmo *Random Forest* constrói as árvores de decisão 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 treinamento, avaliação e aplicação do algoritmo para cada espécie foi realizado no software Weka (Hall et al., 2009). Diante de alguns testes realizados, determinou-se o número de 40 árvores de decisão a serem geradas, sem restrição de comprimento e sem a realização de poda. O número de atributos utilizados em cada árvore foi definido pela fórmula (1), onde  $n$  equivale ao número total de variáveis disponíveis:

$$NumFeatures = \log_2(n) + 1 \quad (1)$$

## 2.5 Treinamento, avaliação e aplicação do algoritmo

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. O conjunto de treinamento é utilizado no ajuste do algoritmo e o conjunto de teste é empregado na avaliação da capacidade preditiva do modelo com base em dados não utilizados para o ajuste (validação preditiva). Em geral esses conjuntos representam 70% e 30%, respectivamente, do conjunto total dos dados disponíveis (Phillips, Anderson e Schapire, 2006).

Diante do número reduzido de observações de ocorrência da espécie, optou-se por utilizar a validação cruzada. Desta forma, os dados constituídos pelos 27 pontos, serão divididos em três subconjuntos (*folds*) de tamanho aproximado. O treinamento e avaliação serão realizados com três iterações, em qual cada iteração um subconjunto é retirado do treinamento e utilizado para avaliar a capacidade de generalização do algoritmo. A avaliação geral do algoritmo é obtida pela média de cada métrica e iteração.

Todas as métricas de avaliação são provenientes da matriz de confusão, sendo possível identificar o número de instâncias classificadas corretamente e erroneamente para cada classe (presença/ausência). A área abaixo da curva (*Receiver Operating Characteristic*), denominada AUC (*Area Under the Curve*) é uma métrica de avaliação vem sendo amplamente empregada em estudos comparativos na modelagem da distribuição de espécies (Elith et al., 2006; Williams et al., 2009; Lotena et al., 2010; Terribile, Diniz-Filho e Marco Jr., 2010; Pouteau et al., 2012) e mensura a habilidade do modelo em discriminar locais de presença e ausência, fornecendo uma medida da capacidade discriminativa do classificador. Além deste método, serão avaliados a porcentagem de dados classificados corretamente e taxa de verdadeiros positivos para cada classe.

Após treinar e avaliar o *Random Forest* para cada espécie, um novo conjunto de dados foi apresentado ao algoritmo para gerar as predições de ocorrência da espécie em toda a bacia. Este

conjunto é constituído das mesmas variáveis ambientais utilizadas na modelagem, com extensão geográfica compreendendo toda a bacia.

### 3. Resultados e Discussão

De acordo com os 27 fragmentos inventariados, três das quatro espécies não ocupam o Triângulo Mineiro, sendo elas *Amaioua intermedia*, *Cabranea canjerana* e *Psychotria vellosiana*. A espécie *Myrcia splendens* apresentou distribuição ao longo de toda a bacia, demonstrando maior generalidade de ambientes. Dentre os 27 pontos de ocorrência, *Amaioua intermedia*, *Cabranea canjerana*, *Myrcia splendens* e *Psychotria vellosiana* ocorreram, respectivamente, em 17, 19, 24 e 17 fragmentos (Figura 3). O conjunto total dos dados dividido em classes de presença e ausência, para cada espécie, pode ser visualizado na tabela 1.

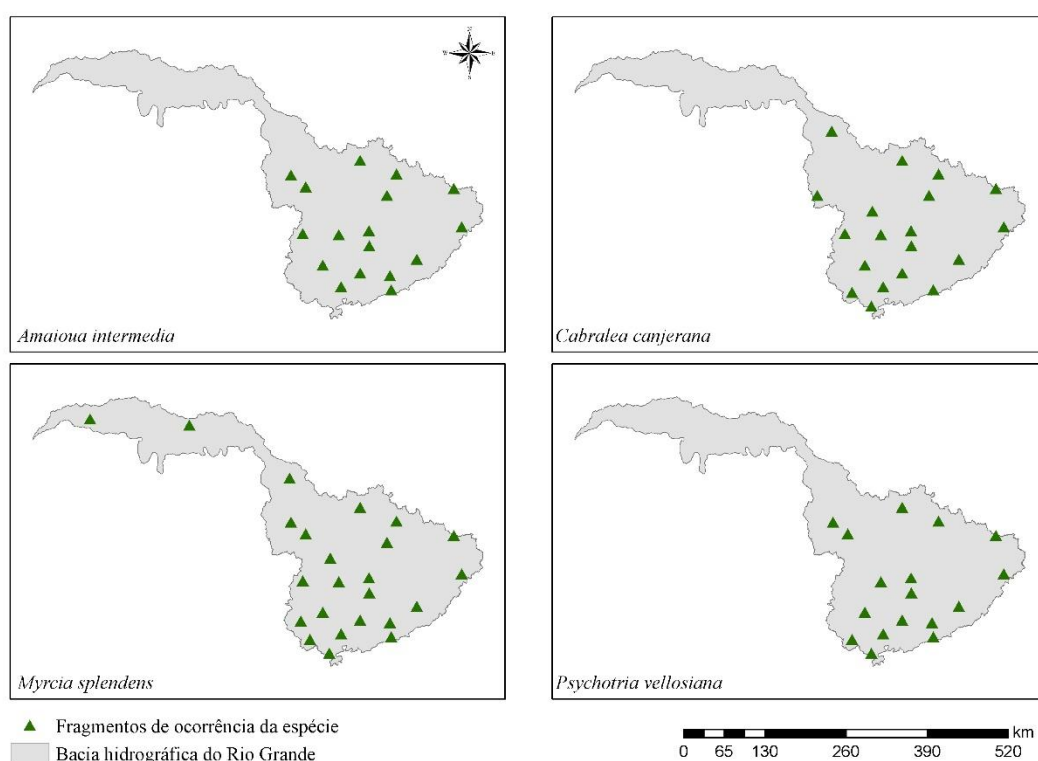


Figura 3 Ocorrência das espécies *Amaioua intermedia*, *Cabranea canjerana*, *Myrcia splendens* e *Psychotria vellosiana* de acordo com os fragmentos inventariados na bacia do Rio Grande

Os resultados obtidos pela validação cruzada podem ser visualizados na Tabela 2. De acordo com a classificação de Metz (1986), o *Random Forest* demonstrou desempenho mediano (AUC entre 0,7 e 0,8) para *Psychotria vellosiana*, e bom desempenho ( $0,8 < \text{AUC} < 0,9$ ) para as demais espécies. Os valores de AUC apresentados pelo *Random Forest*, apesar do pequeno número de dados, estão dentro da faixa de valores encontrados por outras pesquisas de modelagem da distribuição de plantas (Cutler et al., 2007; Dutra e Carvalho, 2008; Marmion et al., 2009).

Tabela 1 Número de presenças e ausências, para cada espécie, de acordo com os 27 fragmentos inventariados

Espécie	Número de presenças	Número de ausências
<i>Amaioua intermedia</i>	17	10
<i>Cabranea canjerana</i>	19	8

<i>Myrcia splendens</i>	24	3
<i>Psychotria vellosiana</i>	17	10

A espécie que obteve melhor classificação pelo *Random Forest* foi *Amaioua intermedia*. Apresentou maior porcentagem dos dados classificados corretamente e maior área abaixo da curva (88,89% e 0,87, respectivamente). O algoritmo conseguiu classificar todas as presenças corretamente, porém errou em 30% na classificação de locais de ausência (3 em 10 ausências possíveis). A espécie *Cabranea canjerana* também foi bem classificada pelo algoritmo, com AUC de 0,85 e 85,18% dos dados classificados corretamente. Dos 19 pontos de presença, 18 foram classificados como tal; e dos 8 pontos de ausência, 5 foram classificados corretamente.

Apesar do bom valor de AUC (0,81) e da porcentagem dos dados classificados corretamente (85,18%), a distribuição da espécie *Myrcia splendens* não foi corretamente aprendida pelo classificador. Das 24 presenças, o *Random Forest* acerta 23, porém, não consegue classificar nenhuma das 3 ausências corretamente. Esta dificuldade está diretamente relacionada com o grande desbalanceamento das classes de presença e ausência e da grande generalidade da espécie. O desbalanceamento entre as classes provoca uma superestimação da classe abundante em detrimento da classe desfavorecida, resultando em um resultado enviesado (Chen, Liaw e Breiman, 2004). Por se tratar de uma espécie generalista, com ampla distribuição, os dados de ausência apresentados ao algoritmo podem ter sido similares a alguns dados de presença, dificultando sua diferenciação. Em muitos casos a ausência da espécie não se deve às características ambientais, e sim fatores de dispersão e histórico de ocupação (Pulliam, 2000).

Tabela 2 Métricas de avaliação obtidas através da validação cruzada sobre o conjunto total dos dados pelo *Random Forest*, apresentadas para as 4 espécies modeladas

Espécie	% classificados corretamente	Taxa de verdadeiros positivos (ausência)	Taxa de verdadeiros positivos (presença)	AUC
<i>Amaioua intermedia</i>	88,89	0,70	1,00	0,87
<i>Cabranea canjerana</i>	85,18	0,63	0,95	0,85
<i>Myrcia splendens</i>	85,18	0,00	0,96	0,81
<i>Psychotria vellosiana</i>	70,37	0,70	0,71	0,76

O menor valor de AUC (0,76) apresentado pelo *Random Forest* foi para a espécie *Psychotria vellosiana*. 70,37% dos dados foram classificados corretamente, sendo que 3, em 10 ausências, foram classificadas como presença e 5, em 17 presenças, classificadas como ausência. Estes erros podem estar relacionados às inconsistências na base de dados. Como dito anteriormente, locais de presença e de ausência podem apresentar características ambientais semelhantes, confundindo o algoritmo. Dados de ausência da espécie são de difícil interpretação, já que a ausência da espécie em determinado local pode estar relacionada à outros fatores, que não os do ambiente.

Os mapas de predição da distribuição potencial das espécies são apresentados na figura 4. Nota-se que as espécies *Amaioua intermedia* e *Psychotria vellosiana*, as quais apresentaram mesmo número de presenças (17) e ocorrência semelhante, apontam uma distribuição potencial similar. Ambas, potencialmente ocupam quase toda parte centro-sul da bacia (68,84% e 65,48%, respectivamente), com exceção, principalmente, do Triângulo Mineiro. A espécie *Myrcia splendens* obteve uma distribuição potencial ao longo de quase toda a bacia, ocupando 93,49% do seu território. Fato este que corrobora com sua generalidade na escolha de ambientes. A distribuição potencial da espécie *Cabralea canjerana* predita pelo *Random Forest*, vai de encontro com a ocorrência observada nos fragmentos. Sua ocupação potencial dentro da bacia corresponde a 70,63% do território, excluindo as regiões do noroeste da bacia.

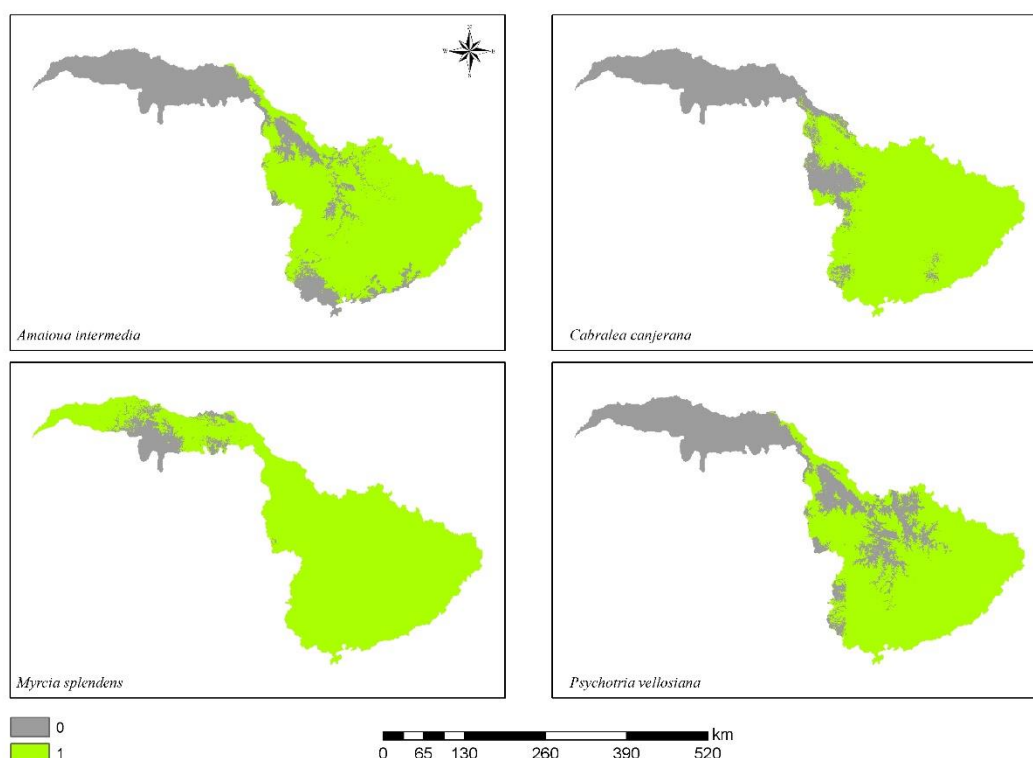


Figura 4 Distribuição potencial das 4 espécies arbóreas modeladas pelo algoritmo *Random Forest*

Verifica-se, para todas as espécies, que a área potencial de ocupação destas corresponde a mais de 50% do território da bacia. Essa grande área predita demonstra o alto potencial de uso destas espécies em planos de revitalização dentro da bacia. Porém, representa um espaço superestimado, já que foi utilizado apenas variáveis ambientais (nicho fundamental) e os dados apresentam alto desbalanceamento em favor dos pontos de presença.

#### 4. Conclusões

Em termos de AUC, o algoritmo *Random Forest* apresentou desempenho moderado para a espécie *Psychotria vellosiana* e bom desempenho para as espécies *Amaioua intermedia*, *Cabralea canjerana*, *Myrcia splendens* e *Psychotria vellosiana*. O desbalanceamento entre as classes de ocorrência de cada espécie prejudicou o aprendizado do algoritmo, fazendo com que o método identifique locais de presença mais facilmente, em detrimento dos locais de ausência. Mesmo assim, na maioria dos casos, o algoritmo pode ser aplicado na modelagem da distribuição destas



espécies. A área predita pelo *Random Forest* é similar à ocorrência apresentada por estas espécies nos fragmentos inventariados.

## 5. Agradecimentos

Os autores agradecem o apoio financeiro da Companhia Energética de Minas Gerais (CEMIG) e a Universidade Federal de Lavras na condução do projeto CEMIG GT456.

## 6. Referências Bibliográficas

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

ARAÚJO, M. B. et al. Exposure of European biodiversity to changes in human induced pressures. **Environmental Science & Policy**, n. 11, p. 38-45, 2008

BREIMAN, L. **Random Forest**. *Machine Learning*, n. 45, p. 5 - 32, 2001.

CARVALHO, L. G. et al. Clima. In: SCOLFORO, J. R. S.; CARVALHO, L. M. T.; OLIVEIRA, A. D. (Ed.). **Zonamento ecológico-econômico do estado de Minas Gerais: Componentes geofísico e biótico**. Lavras - Editora UFLA, 161 p., 2008.

CHEN, C.; LIAW A.; BREIMAN, L. **Using random forest to learn imbalanced data**. Technical Report 666. Statistics Department, University of California, Berkeley, 2004

CHEN, Y.H. Conservation biogeography of the snake family Colubridae of China. **North-Western Journal of Zoology**, n. 5, p. 251-262, 2009

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

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

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

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

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**, n. 197, p. 383 - 393, 2006.

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.H. et al. Ecological-niche factor analysis: how to compute habitat-suitability maps without absence data? **Ecology**, n. 83, p. 2027-2036, 2002

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

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

HALL, M. et al. **The WEKA Data Mining Software: An Update**; SIGKDD Explorations, Volume 11, Issue 1, 2009

MARMION, M. et al. Evaluation of consensus methods in predictive species distribution modelling. **Diversity and Distributions**, n. 15, p. 59–69, 2009

PETERSON, A. T. et al.. Native-range ecology and invasive potential of *Cricetomys* in North America. **Journal of Mammalogy**, n. 87, p. 427-432, 2006

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

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

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

TERRIBILE, LC.; DINIZ-FILHO, JAF; DE MARCO JR., P. How many studies are necessary to compare niche-based models for geographic distributions? Inductive reasoning may fail at the end. **Brazilian Journal of Biology**, n. 2, v. 70, p. 263-269, 2010.

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