MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E INOVAÇÕES
INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS

INFLUÊNCIA DE ATRIBUTOS ESPACIAIS NA CAPACIDADE PREDITIVA DE MODELOS RANDOM FOREST: UMA ANÁLISE COMPARATIVA DA APLICAÇÃO NA IDENTIFICAÇÃO DE OCORRÊNCIAS DE CAVERNAS

João Victor Pereira Sabino

Monografia da disciplina Análise Espacial de Dados Geográficos, ministrada sob coordenação do professor Dr. Antônio Miguel Vieira Monteiro.

INPE
São José dos Campos
2025

RESUMO

Este trabalho avaliou o desempenho preditivo do algoritmo *Random Forest* (RF) na identificação de áreas com potencial espeleológico na porção norte da Serra do Assuruá, Bahia. A metodologia baseou-se na discretização do terreno em uma grade hexagonal de 250 metros, integrando variáveis ambientais e litoestruturais para harmonizar dados de diferentes fontes. Foram comparadas três abordagens principais: o RF Clássico, o RF com inclusão de coordenadas geográficas (XY) e o *Spatial Random Forest* (*spatialML*), todas submetidas a um processo de seleção de atributos via Fator de Inflação da Variância (VIF) e refinamento iterativo pela métrica de importância *Mean Decrease in Accuracy* (MDA). Os resultados demonstraram que o ajuste iterativo dos parâmetros foi o principal vetor de melhoria na qualidade estatística, elevando significativamente a acurácia das predições em todos os modelos testados. Em suma, a transição para modelos espacialmente orientados e tecnicamente refinados representa um avanço fundamental para salvaguardar o patrimônio espeleológico em regiões de intensa expansão do setor eólico.

Palavras-chave: Espeleologia Preditiva; *Random Forest*; Inteligência Espacial.

SUMÁRIO

Pág.

1 INTRODUÇÃO 1
2 CARACTERIZAÇÃO DA ÁREA 4
3 METODOLOGIA 5
3.1 Aquisição e Padronização da Base de Dados 6
3.1.1 Discretização Espacial e Integração de Dados 6
3.1.2 Normalização e Controle do Viés de Amostragem 7
3.1.3 Diagnóstico de Autocorrelação Espacial 7
3.1.4 Classificação Tipológica e Definição de Grupos Espaciais 8
3.1.5 Particionamento Estratificado para Modelagem Preditiva 8
3.2 Seleção das Variáveis e Diagnóstico de Multicolinearidade 9
3.2.1 Seleção das variáveis Explicativas 9
3.2.2 Diagnóstico de Multicolinearidade 13
3.3 Aplicação dos Algoritmos de Classificação 14
3.3.1 Random Forest 14
3.3.2 Random Forest com Coordenadas (RF + XY) 15
3.3.3 Random Forest Espacial (spatialML) 15
3.3.4 Iteração MDA e Seleção de Variáveis 16
3.4 Avaliação da capacidade preditiva dos modelos 17
4 RESULTADOS 18
4.1 Análise exploratória inicial 18
4.2 Seleção de Atributos e Diagnóstico de Multicolinearidade 20
4.3 Performance dos modelos 21
5 CONSIDERAÇÕES FINAIS 26
6 REFERÊNCIAS BIBLIOGRÁFICAS 27

1 INTRODUÇÃO

A prospecção é a principal etapa do licenciamento ambiental dedicada à gestão do patrimônio espeleológico (ICMBio, 2017). É nessa etapa que se dá a confirmação e cadastro da existência de uma caverna[1] e, por consequência, seu enquadramento na legislação específica, que assegura a proteção de seu entorno imediato[2] (BRASIL, 2008). A prospecção pode ocorrer de duas formas: sistemática, em que toda a área de estudo é percorrida em uma malha de caminhamento regular, espaçada conforme estudos prévios de potencial espeleológico; ou estratégica, em que quem busca pelas cavernas prioriza as áreas de maior potencial, ou se orienta por confirmações verbais sobre a localização de cavidades, geralmente se direcionando às cavidades que possuem grande significado local ou regional (FERREIRA et al., 2015).

Em geral, o primeiro tipo é conduzido por empresas no âmbito do licenciamento ambiental, com equipes numerosas e ampla infraestrutura técnica e logística, enquanto a segundo tipo é frequentemente realizado por grupos de pesquisa em espeleologia e/ou entusiastas, que demandam planejamento e otimização logística para reduzir custos e maximizar a eficiência do trabalho de campo.

Diante da relação conhecida e amplamente observada entre o aumento dos cadastros de novas cavernas e a promulgação do Decreto nº 6.640/2008 (BRASIL, 2008), que passou a permitir a supressão de cavidades mediante definição de relevância espeleológica, é plausível inferir associação entre a expansão de empreendimentos e o incremento das descobertas espeleológicas mais recentes (AULER; PILÓ, 2017). Entretanto, observa-se que o material

[1] Entende-se “caverna” pelo conceito de cavidade natural subterrânea, apresentado no decreto n° 10935/2022: “Considera-se cavidade natural subterrânea o espaço subterrâneo acessível pelo ser humano, com ou sem abertura identificada, conhecido como caverna, gruta, lapa, toca, abismo, furna ou buraco, incluídos o seu ambiente, o conteúdo mineral e hídrico, a fauna e a flora presentes e o corpo rochoso onde se inserem, desde que tenham sido formados por processos naturais, independentemente de suas dimensões ou tipo de rocha encaixante”.

[2] Resolução CONAMA 247/2004 diz que: “[...] a área de influência das cavidades naturais subterrâneas será a projeção horizontal da caverna acrescida de um entorno de duzentos e cinqüenta metros, em forma de poligonal convexa”.

produzido nas campanhas sistemáticas de campo é pouco aproveitado em análises estatísticas capazes de refinar e validar os resultados e de otimizar campanhas subsequentes, as quais permanecem ancoradas nas mesmas abordagens multicritério que orientaram as prospecções iniciais. A prospecção sistemática pode ser entendida como uma amostragem extremamente detalhada do contexto espeleológico local, quase a nível censitário, que capta nuances imperceptíveis em análises multicritério tradicionais. Isso se deve à presença *in situ* de um analista, que verifica diretamente as condições de espeleogênese e confirma ou refuta a ocorrência de cavidades em seu caminhamento pelo terreno.

Cavernas são bens da União (Art. 20, CF) e constituem patrimônio cultural brasileiro (Art. 216, CF), cuja preservação para as gerações futuras é um imperativo constitucional (Art. 225, CF) (BRASIL, 1988). Portanto, buscar formas de ampliar e otimizar o conhecimento sobre o patrimônio espeleológico do Brasil, aprimorando o que é revelado pelas prospecções, é fundamental. Entretanto, persiste uma lacuna metodológica: métodos tradicionalmente utilizados na identificação do potencial espeleológico, como o *Analytical Hierarchy Process* (AHP), não incorporam diretamente o conhecimento empírico dessas campanhas de prospecção. Embora simples e de baixo custo computacional, o AHP é sujeito a vieses e inconsistências, por depender do julgamento subjetivo de especialistas, falhando em capturar correlações entre critérios e em tratar as incertezas de sistemas ambientais complexos (MALCZEWSKI; RINNER, 2015). Surge, portanto, a necessidade de modelos preditivos que utilizem ativamente os dados do processo e a localização das cavernas confirmadas (GUISAN; THUILLER; ZIMMERMANN, 2017).

Embora a regressão linear múltipla seja uma alternativa, ela assume a independência entre observações, condição raramente atendida em sistemas com forte dependência espacial, onde as cavernas ocorrem de forma agrupada (DORMANN et al., 2007). Análises espacialmente explícitas, como a *Geographically Weighted Regression* (GWR), lidam melhor com essa heterogeneidade, mas enfrentam limitações práticas devido aos requisitos

rígidos de aplicação, como o tratamento complexo de variáveis categóricas e a multicolinearidade, reduzindo sua aplicação prática (WHEELER; TIEBOUT, 2005).

Neste sentido, o algoritmo *Random Forest* (RF) apresenta-se como uma ferramenta muito promissora (BREIMAN, 2001). Como classificador, pode predizer, categoricamente, a probabilidade de ocorrência de cavernas em uma determinada área. Como regressor, pode estimar o número de cavernas a serem encontradas (CUTLER et al., 2007). Sua principal vantagem reside na capacidade de aprender padrões complexos e não lineares, diretamente dos dados empíricos, reduzindo a dependência de julgamentos subjetivos, a priori, e sendo naturalmente mais robusto a correlações entre variáveis e a dados não normalmente distribuídos (GÉRON, 2019).

Este trabalho tem como objetivo avaliar o desempenho preditivo da modelagem por florestas aleatórias na identificação de áreas com potencial espeleológico, comparando a eficácia de uma abordagem clássica com modelos que incorporam explicitamente a estrutura espacial dos dados de prospecção. A análise será estruturada em três etapas utilizando dois algoritmos distintos: a aplicação do *Random Forest* clássico como *baseline*; o uso do mesmo algoritmo inserindo as coordenadas geográficas (X e Y) como variáveis explicativas; e, por fim, a aplicação do algoritmo de *Spatial Random Forest* (via biblioteca spatialML), que trata a dependência espacial de forma intrínseca ao modelo.

A hipótese principal é que a evolução do *Random Forest* para abordagens espacialmente orientadas elevará significativamente a acurácia das predições, uma vez que a localização e a vizinhança capturam a natureza agrupada da ocorrência de cavernas, superando as limitações de independência de dados dos métodos convencionais. Adicionalmente, supõe-se que o ajuste iterativo dos parâmetros, fundamentado na métrica *Mean Decrease in Accuracy* (MDA), permitirá refinar a relevância de cada atributo espacial, garantindo que o modelo final apresente maior robustez e capacidade de generalização diante da complexidade dos sistemas cársticos.

2 CARACTERIZAÇÃO DA ÁREA

A área de estudo do presente trabalho é a porção norte da Serra do Assuruá, importante compartimento orográfico do extremo norte da Chapada Diamantina (CPRM, 2015), situada no município de Gentio do Ouro, Bahia. Esta região integra o domínio do semiárido (EMBRAPA, 2017) e é caracterizada por uma fitofisionomia de Caatinga (VELOSO et al., 1991), com relevo estruturalmente controlado, cristas quartzíticas paralelas e vales profundos, orientados segundo o padrão de dobramentos do Espinhaço Setentrional (BARRETO; MENDES, 2002). O arcabouço geológico é composto por sucessões sedimentares e metassedimentares (SCHOBENHAUS et al., 1984), com predominância de rochas siliciclásticas (quartzitos e metarenitos) e intercalações carbonáticas (calcários e dolomitos). Essa configuração litoestrutural, associada a sistemas de fraturamentos e zonas de cisalhamento, favorece a exoclastia e a endoclastia, resultando num expressivo patrimônio espeleológico, que abrange desde cavernas em arenitos (SALLUN FILHO; KARMANN, 2012) até sistemas cársticos carbonáticos (AULER, 2017).

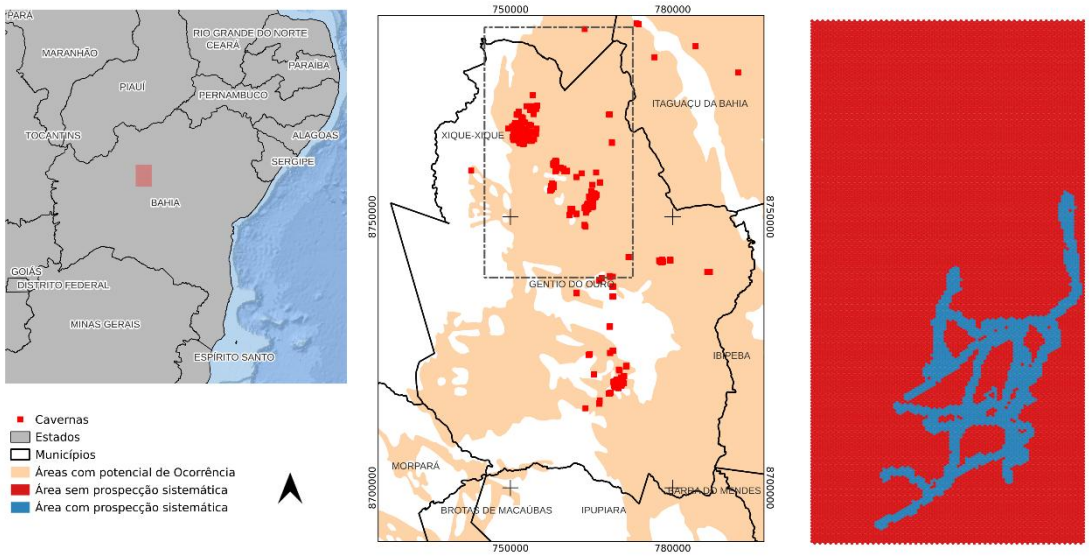
O município tem vivenciado uma expansão exponencial e sistemática do setor eólico (EPE, 2022; ABEEÓLICA, 2023), consolidando-se como um dos principais vetores de geração de energia renovável no território baiano. O crescimento desta infraestrutura é impulsionado pelo regime de ventos de alta constância e unidirecionalidade nos topos de serra, o que atrai investimentos para a instalação de vastos complexos de aerogeradores. No entanto, o avanço da fronteira energética sobre áreas de topografia acidentada gera conflitos de uso do solo (SILVA et al., 2017), dada a sobreposição geográfica entre as áreas de ventos favoráveis e as zonas de maior potencial espeleológico e sensibilidade arqueológica.

Os processos de implantação dessas infraestruturas envolvem intervenções físicas severas, como a supressão de vegetação nativa, terraplanagem de cumes e o uso de explosivos para desmonte de rocha, visando o nivelamento das bases das torres (ICMBio, 2019). Tais atividades impõem riscos de

desestabilização geotécnica em cavidades subterrâneas adjacentes, muitas vezes ainda não catalogadas pelo esforço de prospecção sistemática (SILVA; PESSÔA, 2019).

Além do risco de colapso estrutural, a abertura de malhas viárias para o transporte de componentes pesados pode resultar no assoreamento de condutos, obstrução de entradas e perda de registros arqueológicos e fossilíferos preservados (FERREIRA et al., 2021). Este cenário de rápida transformação da paisagem pela indústria eólica reforça a necessidade de modelos preditivos robustos, que integrem o componente espacial e o rigor estatístico, para salvaguardar a integridade do patrimônio espeleológico na porção norte da Chapada Diamantina (GUISAN et al., 2017). A Figura 1 apresenta a área de estudos.

Figura 1 – Área de Estudos



3 METODOLOGIA

A metodologia adotada neste estudo está estruturada em quatro etapas principais, fundamentais para a execução do fluxo de trabalho: primeiramente, (i) realiza-se a aquisição de dados provenientes das campanhas de prospecção e de bases de dados referentes às variáveis explicativas; em seguida, (ii) procede-se à seleção de atributos para identificar e remover as variáveis que

atrapalham a predição do fenômeno; a terceira etapa (iii) consiste na implementação dos modelos *Random Forest*; e por fim, (iv) ocorre a avaliação dos modelos, onde o desempenho preditivo e a robustez das soluções são validados estatisticamente.

3.1 Aquisição e Padronização da Base de Dados

O desenvolvimento de um modelo preditivo robusto para o potencial espeleológico exige uma abordagem metodológica que controle os vieses inerentes aos dados de prospecção e incorpore explicitamente a estrutura espacial do fenômeno. Para tanto, foi implementado um *pipeline* analítico dividido em cinco etapas sequenciais: (1) discretização espacial e integração de dados; (2) normalização e controle do viés de amostragem; (3) diagnóstico de autocorrelação espacial; (4) classificação tipológica e definição de grupos espaciais homogêneos; e (5) partição estratificada dos dados para modelagem.

3.1.1 Discretização Espacial e Integração de Dados

A área de estudo foi discretizada por meio de uma grade regular de hexágonos com 250 metros de lado (aproximadamente 5,4 ha), geometria selecionada por sua eficiência superior em reduzir a variação da distância entre centroides vizinhos e minimizar efeitos de borda em comparação a malhas quadradas. Esta unidade espacial básica serviu como base para a integração de todas as informações vetoriais e pontuais do projeto. Para cada hexágono, foram calculados dois atributos fundamentais que balizam a análise: o Esforço Amostral Total, definido pelo somatório em metros de todos os transectos de prospecção sistemática percorridos em seu interior; e a Contagem de Ocorrências, que registra o número absoluto de cavidades naturais subterrâneas confirmadas dentro dos limites de cada célula. A escolha dessa escala hexagonal é particularmente estratégica, pois sua área de abrangência é compatível com a zona de influência e proteção legal das cavidades, permitindo uma análise espacialmente aderente à realidade do licenciamento espeleológico.

3.1.2 Normalização e Controle do Viés de Amostragem

A contagem bruta de cavernas foi tratada como um indicador dependente da intensidade do esforço de campo, evitando que áreas mais visitadas parecessem artificialmente mais ricas que áreas menos exploradas. Para mitigar esse viés, criou-se a Densidade Normalizada de Cavernas (DNC). O processo iniciou-se com a normalização do esforço amostral de cada hexágono para uma escala entre 0 e 1, utilizando um limiar de saturação de 2500 metros. Esse valor foi estabelecido empiricamente como o ponto de cobertura censitária, onde o esforço é considerado suficiente para exaurir a detecção de cavidades na célula, anulando ganhos marginais decorrentes de caminhamentos adicionais.

A densidade final foi obtida pela razão entre o número de cavernas encontradas e o esforço normalizado, com a aplicação de um fator de escala para facilitar a interpretação dos dados e um ajuste matemático infinitesimal, para garantir a estabilidade do cálculo em células sem esforço registrado. Essa transformação é fundamental para o rigor da metodologia, pois garante que apenas os hexágonos que sustentam uma alta contagem de cavernas frente a um esforço amostral consistente apresentem valores elevados de densidade. Dessa forma, evita-se a supervalorização de locais onde o número de cavidades pode ser fruto de agrupamentos fortuitos em áreas de baixa prospecção, resultando em uma variável-resposta, que representa fielmente o potencial espeleológico real do terreno.

3.1.3 Diagnóstico de Autocorrelação Espacial

Considerando que a premissa de independência das observações é frequentemente violada em dados geográficos, torna-se essencial quantificar e mapear a dependência espacial da Densidade Normalizada de Cavernas (DNC). Para tanto, utilizou-se a estatística I de Moran, iniciando-se pela construção de uma Matriz de Pesos Espaciais (W) baseada no critério de contiguidade rainha (*queen contiguity*). Nessa configuração, hexágonos que compartilham ao menos um vértice são classificados como vizinhos, permitindo que o Índice de Moran Global (I) teste a hipótese nula de aleatoriedade espacial completa e confirme a estruturação do fenômeno no território.

A fim de identificar padrões locais de associação, foram aplicados os Indicadores Locais de Associação Espacial (LISA). Essa análise possibilitou a categorização de cada hexágono em cinco tipologias distintas: *Clusters* Alto-Alto (*hot spots*), *Clusters* Baixo-Baixo (*cold spots*), *Outliers* Alto-Baixo, *Outliers* Baixo-Alto e áreas sem significância estatística. Entre essas categorias, os *clusters* Alto-Alto assumem papel central no estudo, pois delimitam as áreas nucleares de alto potencial espeleológico, servindo como base prioritária para a validação do desempenho preditivo dos modelos.

3.1.4 Classificação Tipológica e Definição de Grupos Espaciais

Para estratificar a área de estudo em regiões com comportamento amostral e espeleogênico homogêneo, uma classificação tipológica integrada foi desenvolvida. Inicialmente, o Esforço Amostral Total foi categorizado em três classes: Baixo (BE), Médio (ME) e Alto (AE), utilizando os quantis 33% e 66% da distribuição. A integração dessas duas dimensões – intensidade de amostragem e padrão espacial de densidade – resultou na definição de Grupos Espaciais Estratégicos. Por exemplo, hexágonos classificados como AE e pertencentes a um cluster Alto-Alto formam o grupo mais crítico, representando áreas de alta certeza sobre a presença de um núcleo de potencial espeleológico. Esta tipologia multifacetada serve como base para um particionamento dos dados que preserve a estrutura espacial e amostral durante a etapa de modelagem.

3.1.5 Particionamento Estratificado para Modelagem Preditiva

A divisão do banco de dados nos conjuntos de treinamento e validação foi conduzida sob um esquema de amostragem estratificada, visando assegurar a robustez estatística e a imparcialidade na avaliação do modelo. O critério central desta estratégia consistiu em reservar uma proporção de 30% dos hexágonos pertencentes ao estrato mais informativo e crítico — a interseção entre Alto Esforço (AE) e *Cluster* Alto-Alto (LISA) — exclusivamente para o conjunto de validação. Essa decisão metodológica submete os algoritmos ao seu teste mais rigoroso, desafiando a capacidade de predizer corretamente o potencial espeleológico em áreas onde a confiabilidade dos dados de entrada é máxima e a densidade de ocorrências é historicamente elevada.

O conjunto de treinamento, por sua vez, foi estruturado com os 70% remanescentes do estrato AE & *Cluster* Alto-Alto, integrados à totalidade dos hexágonos categorizados como de esforço Médio (ME) e Baixo (BE), independentemente de sua classificação LISA. Esta composição heterogênea é fundamental para que o modelo seja exposto ao gradiente completo de condições ambientais e variações de esforço amostral, mitigando vieses de seleção e permitindo que o algoritmo aprenda a distinguir padrões de ocorrência em diferentes contextos de amostragem. Para a operacionalização dos algoritmos de aprendizado de máquina, as variáveis-resposta foram estabelecidas em dois níveis: uma abordagem de classificação, utilizando a classe binária de presença ou ausência de cavidades derivada da contagem absoluta de ocorrências, e uma abordagem de regressão, baseada no valor contínuo da Densidade Normalizada de Cavernas (DNC).

3.2 Seleção das Variáveis e Diagnóstico de Multicolinearidade

3.2.1 Seleção das variáveis Explicativas

Para a construção da base de dados espacial na porção norte da Serra do Assuruá, cada uma das variáveis preditoras selecionadas foi integrada a uma grade hexagonal regular com resolução de 250 metros. Esta escala foi escolhida por permitir a harmonização de informações de diferentes fontes e resoluções originais, seguindo os princípios de otimização da resolução espacial para análises regionais (HENGL, 2006). Para cada célula desta grade, extraiu-se o valor médio dos atributos numéricos, garantindo uma representação das características predominantes do terreno e minimizando ruídos locais provenientes do modelo digital de elevação.

A declividade média identifica vertentes e escarpas onde o relevo íngreme facilita a exposição de camadas rochosas e a abertura de entradas por processos erosivos, enquanto a posição topográfica (RTP) sinaliza se a unidade ocupa majoritariamente cumes, encostas ou vales. Conforme estabelecido na literatura de análise de terreno, as encostas de alto gradiente hidráulico são fundamentais para o desenvolvimento de condutos (WILSON; GALLANT, 2000). As curvaturas de perfil e tangencial, derivadas do modelo digital de elevação, indicam,

respectivamente, zonas de ruptura de relevo e áreas de concentração de escoamento superficial que alimentam os sistemas de recarga hídrica subterrânea (FLORINSKY, 2012), ao passo que a altitude média se correlaciona com antigos níveis de base regionais.

A análise estrutural é incorporada pelas distâncias médias em relação a lineamentos, afloramentos e contatos litológicos. Essas variáveis mensuram a proximidade de falhas e zonas de fraqueza, que guiam a infiltração da água e a interface entre rochas de diferentes permeabilidades, fatores críticos para a gênese de cavidades, conforme os princípios hidrogeológicos da espeleogênese (PALMER, 1991; KLIMCHOUK, 2000). Complementarmente, a exposição média influencia o microclima e a meteorização das vertentes, afetando a preservação das aberturas das cavernas (WILSON; GALLANT, 2000).

Somando-se aos preditores numéricos, as variáveis categóricas de litologia predominante e uso do solo foram integradas à grade para capturar a favorabilidade intrínseca das rochas e as condições de cobertura superficial. A litologia atua como o filtro primário de potencialidade, diferenciando o comportamento espeleológico entre os quartzitos da Serra do Assuruá e as lentes carbonáticas. Especificamente, a formação de cavernas em quartzito segue processos distintos, justificando sua categorização como variável fundamental (SALLUN FILHO; KARMANN, 2012). O uso do solo, mapeado a partir de técnicas de sensoriamento remoto (JENSEN, 2015), fornece indícios sobre a integridade do ambiente e a visibilidade de feições cársticas. Esta variável é essencial para que os modelos identifiquem os determinantes reais do fenômeno em meio a mudanças antrópicas, como a expansão do setor energético na região, cujos impactos na paisagem do semiárido têm sido documentados (FERREIRA et al., 2021).

A abordagem geral de modelagem espacial baseada em variáveis ambientais seguiu o arcabouço conceitual estabelecido para a predição de distribuição de fenômenos geográficos (GUISAN; THUILLER, 2005; FRANKLIN, 2010), adaptado ao objeto de estudo espeleológico. A Figura 2 fornece uma visualização das variáveis explicativas.

Figura 2 – Variáveis explicativas (a)

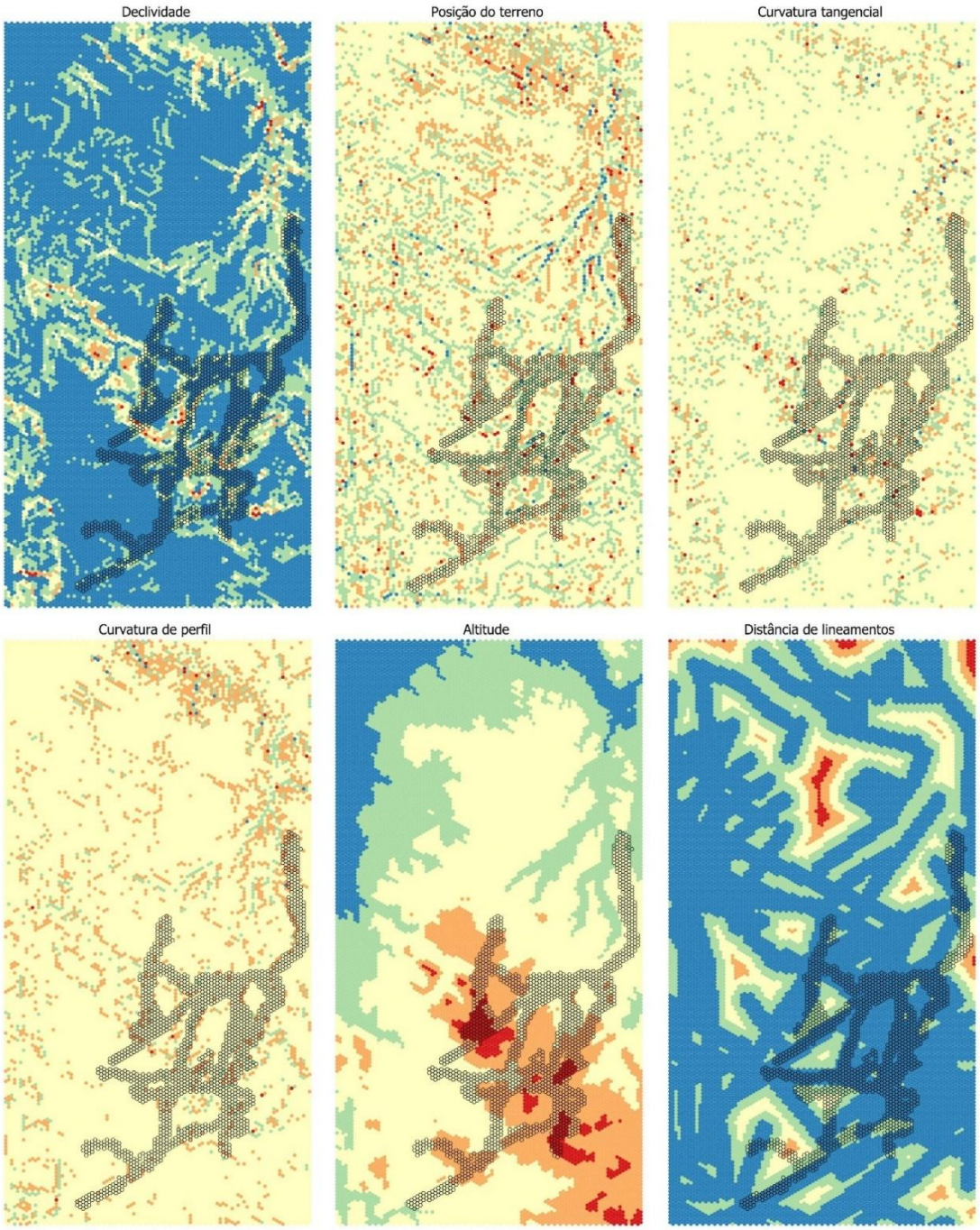
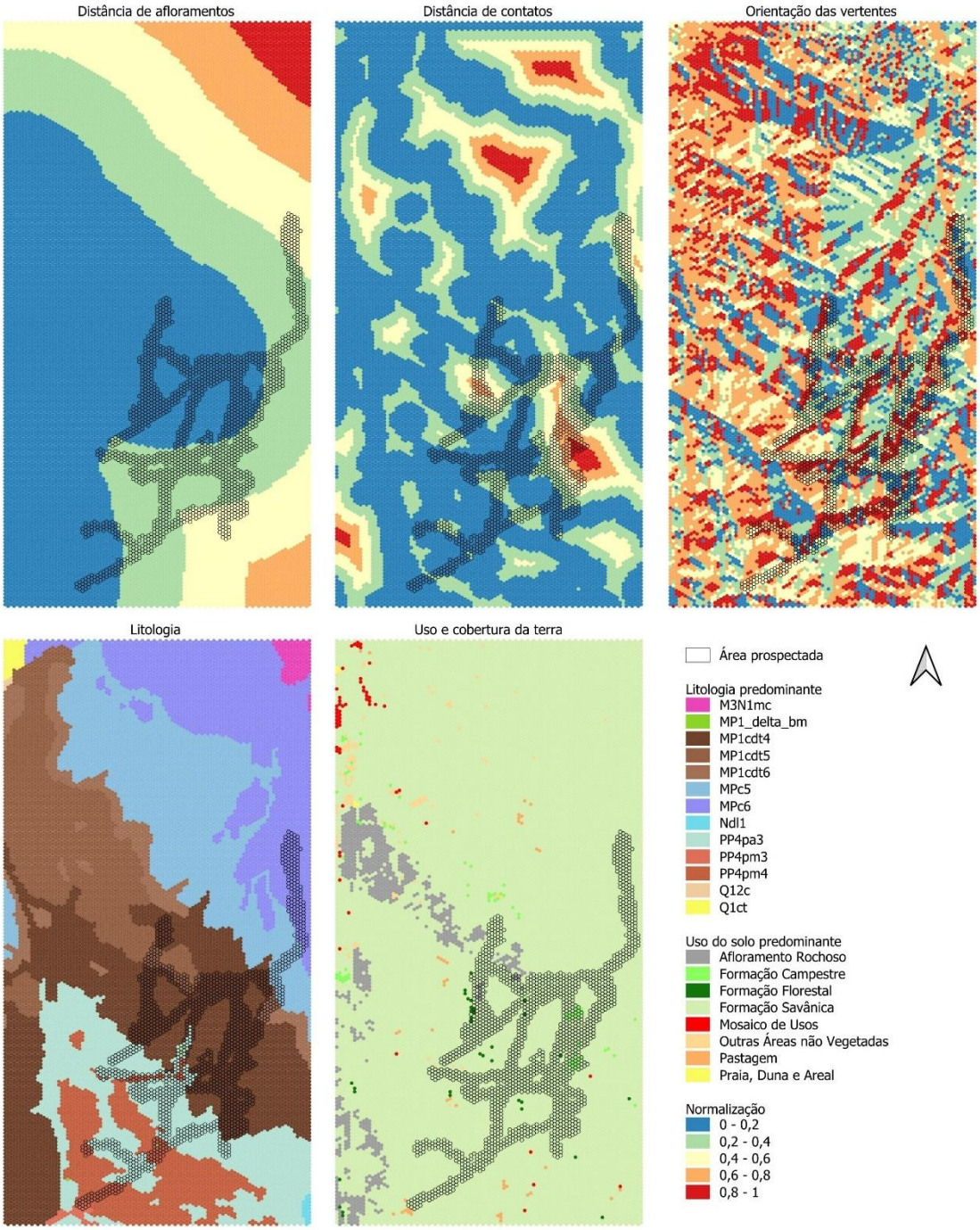


Figura 2 – Variáveis explicativas (b)



3.2.2 Diagnóstico de Multicolinearidade

Esta etapa dedica-se ao refinamento do conjunto de preditores, através da seleção de atributos e da eliminação de redundâncias estatísticas que possam comprometer a performance dos modelos de aprendizagem de máquina (GUYON; ELISSEEFF, 2003; KUHN; JOHNSON, 2019). Este processo fundamenta-se na remoção da multicolinearidade de forma iterativa, uma vez que a presença de variáveis altamente correlacionadas pode inflar artificialmente a variância dos coeficientes e mascarar a real importância individual de cada fator espeleogenético no modelo final (JAMES et al., 2021). A triagem é realizada através do cálculo do Fator de Inflação da Variância (VIF), aplicando-se um limiar rigoroso onde variáveis com VIF superior a 5 são sistematicamente descartadas (O'BRIEN, 2007). Conforme demonstrado no diagnóstico dos preditores numéricos brutos, variáveis de terreno como declividade frequentemente apresentam valores de VIF extremamente elevados, indicando uma redundância informativa que exige intervenção (DA SILVA; CHAVES, 2016).

Através de sucessivas rodadas de eliminação, o conjunto de dados é filtrado até que restem apenas os preditores estatisticamente independentes, resultando em uma lista otimizada que preserva a integridade física do fenômeno espeleológico (DORMANN et al., 2013). Este refinamento mantém atributos fundamentais como altitude, distâncias geológicas (lineamentos, contatos e afloramentos) e curvaturas do terreno, todos com valores de VIF ajustados próximos à unidade. Tal tratamento assegura que os algoritmos de *Random Forest* operem com um sinal limpo, livres de ruídos causados pela multicolinearidade. Ao final desta etapa, o banco de dados consolidado — que integra os preditores numéricos iterados às variáveis categóricas de litologia e cobertura do solo — fornece a base necessária para que as etapas subsequentes de implementação e iteração via *Mean Decrease in Accuracy* (MDA) identifiquem com precisão os determinantes do potencial espeleológico (BREIMAN, 2001; GÉRON, 2019).

3.3 Aplicação dos Algoritmos de Classificação

Para avaliar o impacto da incorporação da informação espacial na predição do potencial espeleológico, foram implementados e comparados três modelos baseados no algoritmo *Random Forest* (RF), cada um com uma abordagem distinta quanto ao tratamento da localização geográfica. Todos os modelos foram configurados para realizar uma tarefa de classificação, cuja variável-resposta é a presença (1) ou ausência (0) de cavernas em cada hexágono da grade amostral. A escolha pela classificação deve-se à natureza binária da descoberta em campo e ao objetivo primário de identificar áreas com maior probabilidade de ocorrência.

3.3.1 Random Forest

O algoritmo *Random Forest* (RF) é um método de aprendizado de máquina que combina a predição de múltiplas árvores de decisão (BREIMAN, 2001). Durante o treinamento, o algoritmo constrói cada árvore a partir de um *bootstrap* (subconjunto aleatório com reposição) dos dados de treino e, em cada divisão de nó, considera apenas um subconjunto aleatório das variáveis preditoras (mtry). A predição final é dada pelo voto majoritário (classificação) das árvores. O RF é robusto a sobreajuste (*overfitting*) e tolerante a dados ruidosos e à multicolinearidade entre variáveis.

Neste estudo, o modelo RF "clássico" foi implementado utilizando o pacote *randomForest* no ambiente R, configurado exclusivamente com as variáveis ambientais e geomorfológicas preditoras (e.g., litologia, declividade, distância a lineamentos), sem qualquer informação explícita de coordenadas espaciais (LIAW; WIENER, 2002). Os hiperparâmetros foram definidos como: ntree = 500 (número de árvores) e mtry = 3 (raiz quadrada do número total de preditores, conforme prática comum para classificação). Para corrigir o desbalanceamento natural entre hexágonos com e sem cavernas, foram aplicados pesos automáticos para as classes durante o treinamento, penalizando mais os erros na classe minoritária (presença de cavernas) para melhorar a sensibilidade do modelo (CHEN et al., 2004).

3.3.2 *Random Forest* com Coordenadas (RF + XY)

Esta abordagem estende o RF clássico pela inclusão direta das coordenadas geográficas (X, Y) de cada hexágono como variáveis preditoras adicionais. A estratégia busca testar a hipótese de que a localização absoluta contém um sinal espacial residual não capturado pelos atributos ambientais, podendo estar correlacionada a padrões regionais de espeleogênese (e.g., gradientes climáticos não modelados, histórico geológico específico de sub-regiões). A implementação seguiu exatamente os mesmos parâmetros e procedimentos do RF clássico (ntree = 500, mtry = 3, pesos para classes), diferenciando-se apenas pela adição das duas variáveis numéricas de coordenadas ao conjunto inicial de preditores. Esta é uma abordagem simples e amplamente utilizada para incorporar espacialidade em modelos baseados em árvores quando a dependência espacial é global e estacionária (BENITO et al., 2021).

3.3.3 *Random Forest* Espacial (spatialML)

Embora o RF + XY incorpore localização, ele ainda assume uma relação estacionária e global entre preditores e resposta. Para considerar explicitamente a não estacionariedade espacial, ou seja, a possibilidade de que a relação entre a litologia e ocorrência de cavernas, por exemplo, mude ao longo da área de estudo, implementou-se um terceiro modelo utilizando o pacote SpatialML (KALOGIROU; GEORGANOS, 2019) para *Geographical Random Forest* (GRF). Este algoritmo decompõe um modelo global em múltiplos sub-modelos locais, seguindo a lógica da *Geographically Weighted Regression*.

A implementação do GRF envolve a definição de uma janela espacial adaptativa que, para cada hexágono de predição, seleciona seus *n* vizinhos mais próximos para treinar um modelo RF local. Neste estudo, adotou-se uma janela com 20 vizinhos mais próximos, definida com base em testes preliminares de estabilidade do erro de predição. Para cada modelo local, os hiperparâmetros *ntry* e *mtry* foram mantidos em 500 e 3, respectivamente, garantindo comparabilidade. Entende-se que esta abordagem é particularmente adequada para sistemas espeleológicos, onde os processos de espeleogênese e controle

estrutural podem variar significativamente entre diferentes compartimentos geológicos ou fisiográficos.

3.3.4 Iteração MDA e Seleção de Variáveis

A fim de otimizar a estrutura preditiva de cada modelo e isolar a contribuição efetiva de cada variável, o processo de ajuste foi submetido a uma iteração rigorosa, guiada pela métrica de importância *Mean Decrease in Accuracy* (MDA) (BREIMAN, 2001). A métrica em questão quantifica a contribuição marginal de cada preditor ao avaliar o aumento percentual no erro de classificação do modelo quando seus valores são permutados aleatoriamente, quebrando assim qualquer relação verdadeira com a variável resposta. Inicialmente, cada modelo – RF Clássico, RF+XY e GRF – foi treinado com o conjunto completo de variáveis preditoras, incluindo tanto os atributos numéricos (PRED_NUM), que capturam gradientes ambientais e morfométricos, quanto os categóricos (PRED_CAT), como a litologia predominante.

Após este treinamento inicial, as variáveis foram ordenadas de acordo com seu valor de MDA. Aquelas cuja importância se mostrou próxima de zero ou negativa – indicando ausência de contribuição ou até interferência na acurácia do modelo – foram progressivamente eliminadas. Em seguida, um novo modelo era retreinado exclusivamente com o subconjunto de preditores que demonstraram relevância. Este ciclo de avaliação de importância, poda de variáveis e retreinamento foi repetido iterativamente até que se estabilizasse um conjunto final composto apenas por preditores com contribuição positiva e estatisticamente significativa para a capacidade preditiva (KUHN; JOHNSON, 2019).

Como resultado, cada abordagem – não espacial, com coordenadas ou espacialmente explícita – convergiu para um conjunto otimizado e potencialmente distinto de variáveis, reflexo direto de como a incorporação (ou não) da informação espacial redefine a hierarquia e a interação dos fatores ambientais no processo de aprendizagem. A avaliação comparativa do desempenho preditivo destes três modelos, agora internamente otimizados, constitui, portanto, a base metodológica sólida para testar a hipótese central

deste trabalho: que a espacialização, seja de forma simples ou complexa, agrega valor substantivo à modelagem do potencial espeleológico.

3.4 Avaliação da capacidade preditiva dos modelos

A avaliação da capacidade preditiva dos três modelos desenvolvidos — *Random Forest* Clássico, *Random Forest* com coordenadas e *Spatial Random Forest* (spatialML) — foi conduzida sob rigoroso controle analítico, priorizando métricas robustas ao cenário de eventos raros típico da prospecção espeleológica. Dado que as ocorrências de cavernas representam uma classe minoritária na grade amostral, a métrica central de desempenho adotada foi a Área sob a Curva *Precision-Recall* (PR-AUC) (SAITO; REHMSMEIER, 2015).

Diferente da tradicional Curva ROC, a PR-AUC é significativamente mais informativa em contextos de desbalanceamento, pois sintetiza em um único valor o equilíbrio entre a Precisão (probabilidade de uma célula classificada como potencial de fato abrigar uma caverna) e o Recall (capacidade do modelo em recuperar a totalidade das ocorrências reais, minimizando omissões críticas). Para operacionalizar o ponto de corte ideal entre essas dimensões, utilizou-se o F1-Score, que atua como uma média harmônica penalizando modelos com desequilíbrios acentuados entre sensibilidade e confiabilidade (POWERS, 2011).

Complementarmente, um conjunto de indicadores auxiliares forneceu perspectivas multidimensionais sobre a qualidade dos ajustes. A ROC-AUC (HANLEY; MCNEIL, 1982) serviu como referência da capacidade discriminatória global, enquanto o Lift (Fator de Ganho) quantificou o benefício prático, demonstrando a eficiência da busca orientada pelo modelo em relação a uma prospecção aleatória baseada apenas na prevalência natural do fenômeno. A calibração das probabilidades foi monitorada via Log-Loss (Entropia Cruzada) (MURPHY, 1973), que mensura a divergência entre as estimativas de confiança do classificador e a realidade binária observada. Esta avaliação integrada, realizada sobre o conjunto de validação estratificado com foco nas áreas de Alto Esforço Amostral (ROBERTS et al., 2017), permite verificar empiricamente se a incorporação progressiva da inteligência espacial resulta em ganhos tangíveis de precisão para futuras campanhas de campo.

Em termos de interpretação quantitativa, as métricas seguem escalas que traduzem a qualidade estatística em utilidade operacional. Para a PR-AUC, ROC-AUC e F1-Score, os valores variam de 0 a 1, onde resultados próximos à unidade indicam performance superior e baixo erro de classificação; na ROC-AUC, especificamente, o valor 0,5 representa um desempenho não superior ao acaso. Na Log-Loss, a lógica é invertida por tratar-se de uma medida de erro: valores próximos a zero são desejáveis, indicando predições bem calibradas, enquanto valores altos sinalizam estimativas incertas. Por fim, o Lift é interpretado em escala superior a 1, onde um valor elevado (como 5 ou 10) significa que o modelo localiza múltiplas vezes mais cavernas que uma busca aleatória, validando o ganho estratégico da modelagem para a gestão do patrimônio espeleológico.

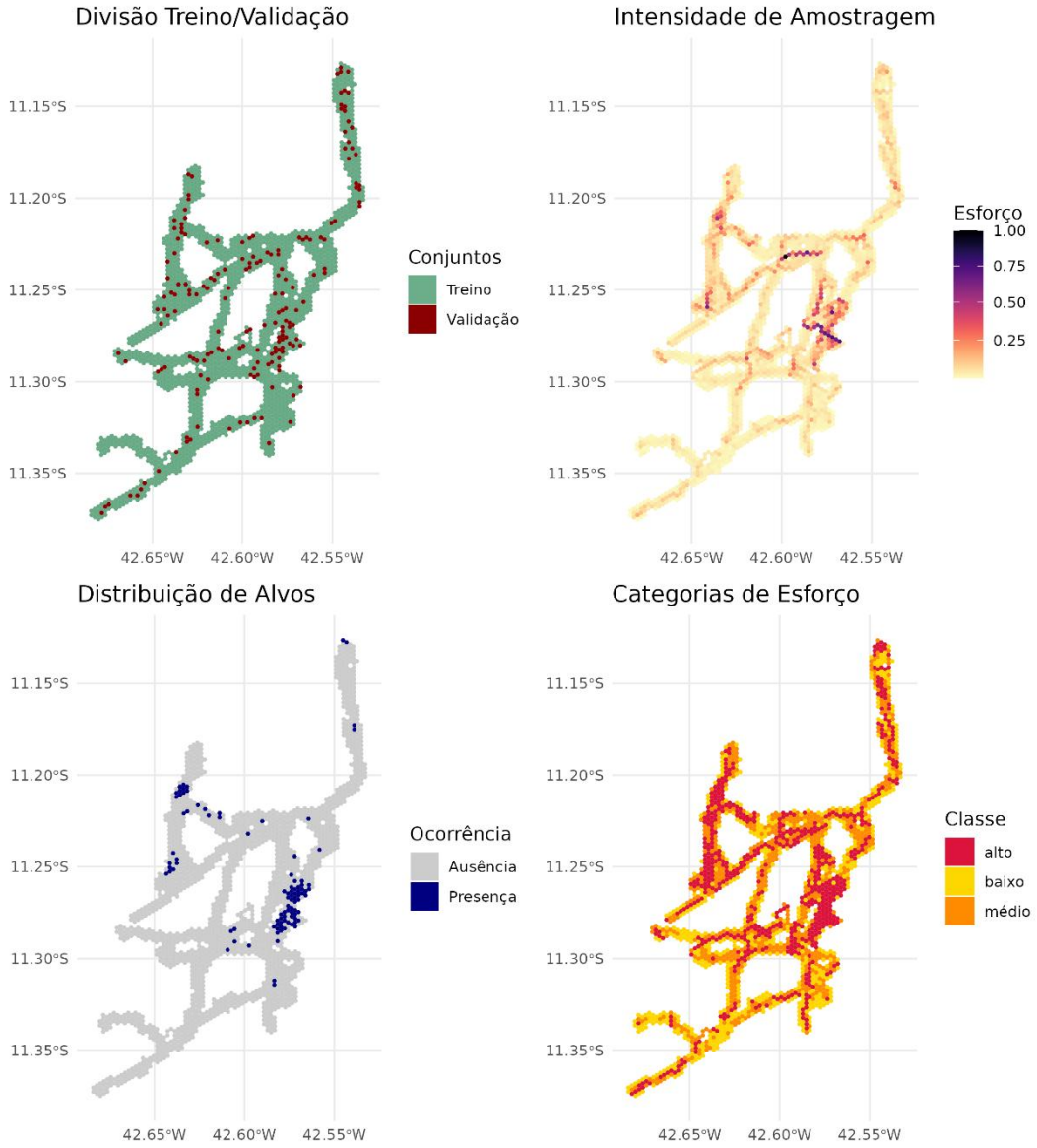
4 RESULTADOS

4.1 Análise exploratória inicial

A análise demonstrou que o esforço de campo é um fator determinante para a detecção. O esforço amostral mostrou-se heterogêneo e a ocorrência de cavidades extremamente esparsa: apenas 5,3% das células prospectadas registraram ao menos uma caverna, evidenciando um padrão fortemente concentrado e uma variável-resposta marcadamente zero-inflada. Existe uma correlação positiva (r = 0,33) entre esforço e densidade, e a probabilidade de encontrar cavernas em áreas de esforço alto (13,8%) é cerca de 80 vezes superior à observada em áreas de esforço baixo (0,17%).

O Índice de Moran Global (I = 0,38) confirmou que tanto a densidade quanto o esforço não são aleatórios, organizando-se em agrupamentos espaciais. O diagnóstico local (LISA) identificou *clusters* Alto-Alto, que funcionam como núcleos de ocorrência de cavernas, e vastos blocos Baixo-Baixo, onde a densidade é nula independentemente do esforço aplicado. Esses elementos podem ser observados na Figura 3.

Figura 3 – Exploração dos dados de prospecção



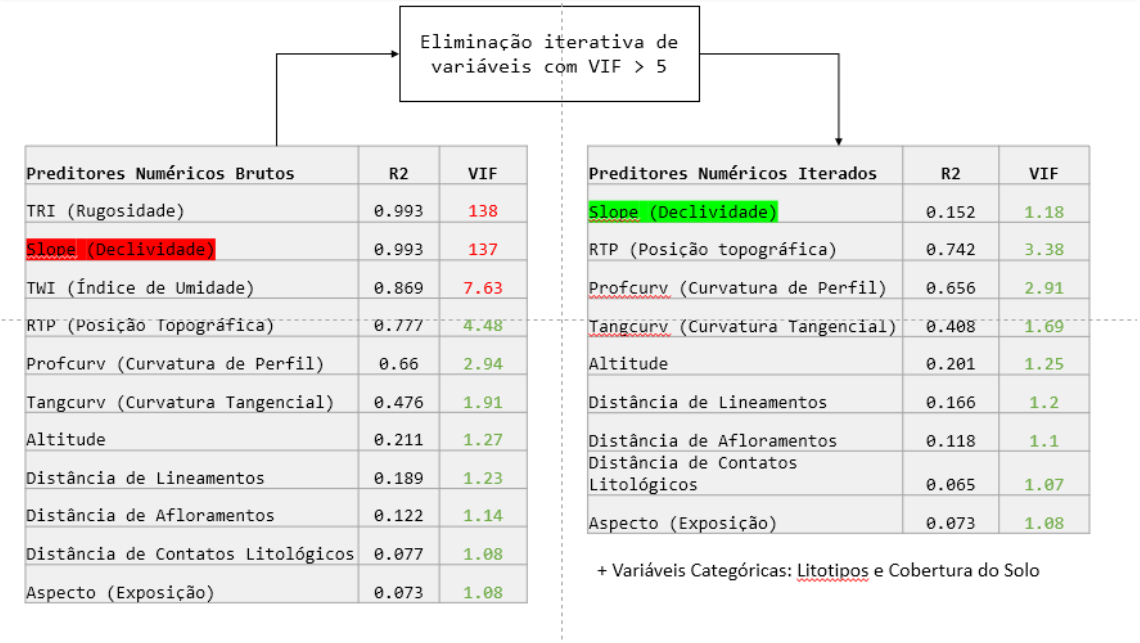
No âmbito ambiental, as variáveis categóricas (litologia, solo e uso) funcionam como filtros de favorabilidade. Unidades como a litologia MP1cdt5, solos AR/RLd e a classe “afloramento”, de uso do solo, apresentam densidades significativamente superior à média. Em síntese, a ocorrência espeleológica é um fenômeno multifatorial e espacialmente estruturado, exigindo modelos que integrem o esforço amostral, o contexto de vizinhança e a modulação geomorfológica para estimar o potencial real em áreas não prospectadas.

4.2 Seleção de Atributos e Diagnóstico de Multicolinearidade

O diagnóstico de multicolinearidade demonstrou que o conjunto original de preditores numéricos brutos apresentava redundâncias críticas, capazes de comprometer a estabilidade estatística e a interpretabilidade dos modelos de aprendizagem de máquina. A análise inicial revelou que variáveis morfométricas fundamentais possuíam correlações quase perfeitas entre si, destacando-se o TRI (Rugosidade) e o Slope (Declividade), que registraram valores de VIF superiores a 130. Além desses, o Índice de Umidade (TWI) também ultrapassou o limite de segurança, apresentando um VIF de 7,63, o que indicava uma sobreposição informativa prejudicial à precisão do algoritmo.

Para mitigar esse problema, aplicou-se um processo de eliminação iterativa focado em descartar sistematicamente variáveis com VIF superior a 5. Esse refinamento permitiu purificar o sinal das variáveis remanescentes, garantindo que cada atributo contribua de forma única para a predição do fenômeno espeleológico. Como resultado prático, a Declividade, após a remoção de seus pares colineares, teve seu VIF reduzido para 1,18, tornando-se um preditor estatisticamente independente e confiável (Figura 4)

Figura 4 – Seleção de Variáveis com VIF maior que 5



Preditores Numéricos Brutos	R2	VIF
TRI (Rugosidade)	0.993	138
Slope (Declividade)	0.993	137
TWI (Índice de Umidade)	0.869	7.63
RTP (Posição Topográfica)	0.777	4.48
Profcurv (Curvatura de Perfil)	0.66	2.94
Tangcurv (Curvatura Tangencial)	0.476	1.91
Altitude	0.211	1.27
Distância de Lineamentos	0.189	1.23
Distância de Afloramentos	0.122	1.14
Distância de Contatos Litológicos	0.077	1.08
Aspecto (Exposição)	0.073	1.08

Preditores Numéricos Iterados	R2	VIF
Slope (Declividade)	0.152	1.18
RTP (Posição topográfica)	0.742	3.38
Profcurv (Curvatura de Perfil)	0.656	2.91
Tangcurv (Curvatura Tangencial)	0.408	1.69
Altitude	0.201	1.25
Distância de Lineamentos	0.166	1.2
Distância de Afloramentos	0.118	1.1
Distância de Contatos Litológicos	0.065	1.07
Aspecto (Exposição)	0.073	1.08

+ Variáveis Categóricas: Litotipos e Cobertura do Solo

O conjunto final de preditores numéricos iterados consolidou variáveis de relevo como Posição Topográfica, Curvatura de Perfil, Curvatura Tangencial e Altitude, além de métricas de distância em relação a lineamentos, afloramentos e contatos litológicos. Todos esses atributos apresentaram VIFs ajustados próximos à unidade, variando entre 1,07 e 3,38, o que assegura a integridade física e estatística do modelo. A base de dados foi concluída com a integração das variáveis categóricas de litotipos e cobertura do solo, fornecendo os insumos necessários para que as etapas subsequentes de implementação e iteração via MDA identifiquem com precisão os determinantes reais do potencial espeleológico.

4.3 Performance dos modelos

4.3.1 Desempenho dos Modelos Base

A avaliação comparativa dos três modelos preditivos em sua configuração inicial (com o conjunto total de variáveis preditoras) revelou um desempenho distinto para cada arquitetura (Tabela 1).

Tabela 1 – Comparação do baseline das abordagens RF

Métrica	RF CLÁSSICO	RF ESPACIAL (+XY)	RF SPATIALML
PR-AUC	**0.6269**	**0.6308**	**0.6387**
Recall	**0.65**	**0.65**	**0.95**
PPV	**0.684**	**0.722**	**0.5429**
F1-Score	**0.667**	**0.684**	**0.6909**
ROC-AUC	**0.9011**	**0.9008**	**0.9556**
Lift	**6.16**	**6.5**	**4.89**
Log-Loss	**0.2206**	**0.2165**	**0.1665**

O algoritmo *Random Forest SpatialML* destacou-se pela excelência na calibração probabilística, registrando o menor valor de Log-Loss (0,1665) e a maior ROC-AUC (0,9556) do estudo. Sua capacidade de detecção (*Recall* de

0,95) foi a mais elevada, indicando que este modelo capturou quase a totalidade das ocorrências reais de cavernas. No entanto, esta alta sensibilidade veio associada a uma precisão (PPV) moderada (0,5429), sugerindo uma propensão a classificar um número maior de áreas como potenciais, o que pode aumentar a taxa de falsos positivos.

Em contrapartida, os modelos *Random Forest* Clássico e *Random Forest* Espacial (+XY) apresentaram métricas de PR-AUC muito similares entre si (0,6268 e 0,6308, respectivamente). Ambos mantiveram um fator de ganho (Lift) acima de 5,1, validando que mesmo as abordagens mais simples superam significativamente uma prospecção aleatória. Estes modelos demonstraram um perfil mais conservador, com maior precisão (PPV de 0,684 e 0,724, respectivamente) em detrimento de uma sensibilidade mais baixa (Recall de 0,05).

4.3.2 Impacto do Refinamento Iterativo via MDA

O processo de refinamento das variáveis preditoras, guiado pela métrica *Mean Decrease in Accuracy* (MDA), promoveu ganhos expressivos na qualidade estatística dos modelos, consolidando configurações mais otimizadas e específicas (Tabela 2 a Tabela 4).

Tabela 2 – Iteração do RF Clássico

Métrica	RF CLÁSSICO	RF CLÁSSICO COM ITERAÇÃO
PR-AUC	**0.6269**	**0.6955**
Recall	**0.65**	**0.7**
PPV	**0.684**	**0.737**
F1-Score	**0.667**	**0.718**
ROC-AUC	**0.9011**	**0.9156**
Lift	**6.16**	**6.63**
Log-Loss	**0.2206**	**0.2066**

Tabela 3 – Iteração do RF Espacial

Métrica	RF ESPACIAL (+XY)	RF ESPACIAL (+XY) COM ITERAÇÃO
PR-AUC	0.6308	0.7606
Recall	0.65	0.9169
PPV	0.722	0.7
F1-Score	0.684	0.7
ROC-AUC	0.9008	0.9169
Lift	6.5	6.3
Log-Loss	0.2165	0.2035

Tabela 4 – Iteração do SpatialML

Métrica	RF spatialML	RF spatialML
PR-AUC	0.6387	0.7112
Recall	0.95	0.75
PPV	0.5429	0.5769
F1-Score	0.6909	0.6522
ROC-AUC	0.9556	0.9528
Lift	4.89	5.19
Log-Loss	0.1665	0.1734

A transição para os modelos otimizados consolidou o *Random Forest* Espacial (+XY) com Iteração como a configuração de maior equilíbrio técnico-gerencial, atingindo a maior PR-AUC (0,7608) observada. Este modelo conseguiu conciliar uma alta taxa de descoberta (Recall de 0,9168) com uma precisão robusta (PPV

de 0,70), demonstrando que a exclusão de preditores ruidosos e a inclusão explícita das coordenadas geográficas otimizaram substancialmente a eficácia preditiva em dados desbalanceados.

Já o RF Clássico com Iteração sagrou-se o modelo mais assertivo em termos de confiabilidade operacional imediata, atingindo a maior Precisão (PPV de 0,972) e o maior Lift (6,63) entre os modelos iterados. Este perfil o torna ideal para cenários onde se busca minimizar falsos positivos e otimizar o esforço de campo, garantindo o maior índice de acerto por unidade de alvo verificado.

Curiosamente, o *Random Forest* SpatialML teve um comportamento distinto após o processo de iteração. Embora tenha melhorado sua PR-AUC para 0,7142 e elevado drasticamente sua precisão (PPV para 0,9598), observou-se uma redução controlada em sua sensibilidade (*Recall* de 0,92) e um leve aumento em sua entropia (Log-Loss de 0,1784). Isso sugere que o refinamento de variáveis nesta arquitetura de modelo local tornou o classificador mais rigoroso e bem calibrado, priorizando a certeza estatística e a confiabilidade das predições positivas.

4.3.3 Visualização da resposta espacial das predições

A validação estatística é complementada pela análise visual da distribuição espacial das probabilidades preditas. As Figuras 1, 2 e 3 ilustram, respectivamente, os mapas de potencial espeleológico gerados pelos modelos RF Clássico com Iteração, RF Espacial (+XY) com Iteração e RF SpatialML com Iteração, permitindo avaliar como as diferentes arquiteturas traduzem os padrões aprendidos em uma representação geográfica. Entre a Figura 5 e a Figura 7 são apresentadas visualizações do mapa de pontecial extraído.

Figura 5 – Mapa de potencial espeleológico predito pelo modelo RF Clássico (Iterado).

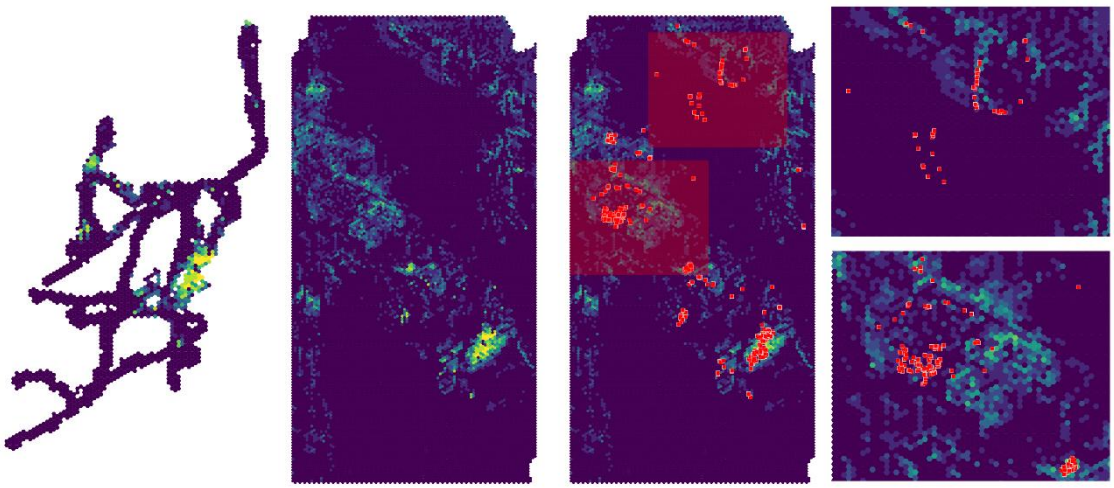


Figura 5 – Mapa de potencial espeleológico predito pelo modelo RF Espacial + XY (Iterado).

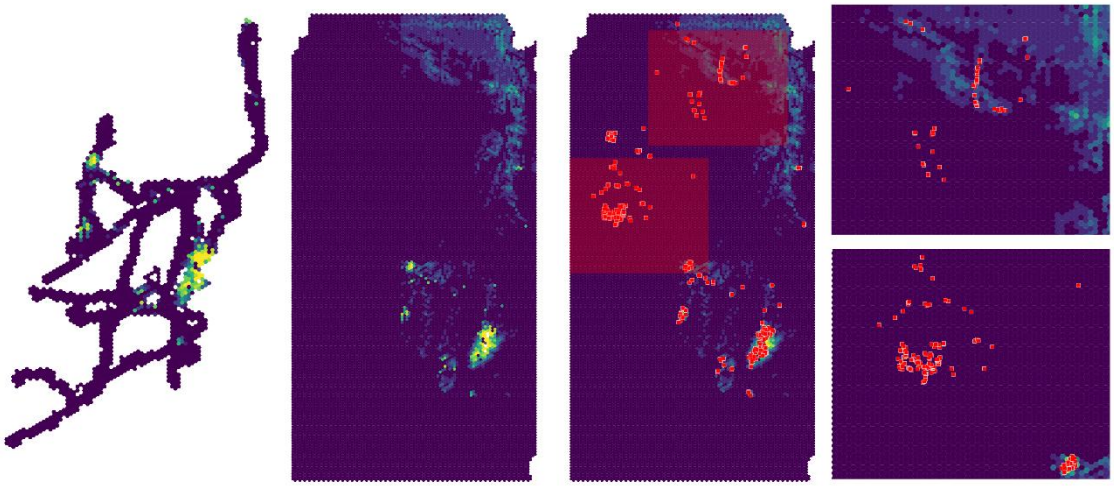
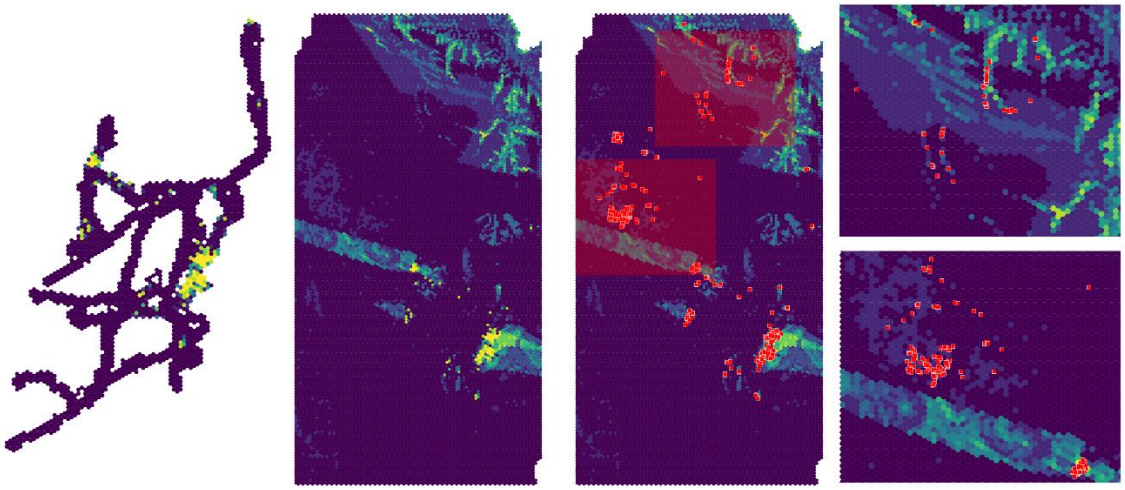


Figura 5 – Mapa de potencial espeleológico predito pelo modelo RF spatialML (Iterado).



A comparação visual revela que o RF Clássico (Figura 1) tende a produzir uma assinatura espacial mais fragmentada e pontual, com áreas de alta probabilidade fortemente associadas a feições geológicas específicas (ex.: lineamentos, contatos litológicos). O RF Espacial + XY (Figura 2) introduz uma suavização e uma estruturação regional mais clara, onde as coordenadas atuam como um organizador espacial, criando clusters ou gradientes de potencial que refletem padrões sub-regionais.

Por fim, o RF SpatialML (Figura 3) apresenta a padronização espacial mais complexa, com núcleos de alta probabilidade muito bem definidos e uma variação local mais nítida, o que demonstra sua capacidade superior de capturar as não estacionariedades nos processos espeleogenéticos. Embora tenham sido observados alguns artefatos pontuais no mapa resultante — decorrentes da sensibilidade do algoritmo a variações locais extremas — esses elementos não comprometem a utilidade do modelo. Pelo contrário, na avaliação global, o RF SpatialML foi o que melhor conseguiu traduzir as realidades locais da Serra do Assuruá, mantendo-se como a principal referência em termos de calibração e qualidade estatística devido ao seu baixo Log-Loss.

5 CONCLUSÕES

A investigação sobre a eficácia da modelagem preditiva na porção norte da Serra do Assuruá permite concluir que a integração de abordagens espacialmente orientadas ao algoritmo Random Forest representa um avanço substantivo na gestão do patrimônio espeleológico. O estudo confirmou que a inclusão do componente espacial na lógica de predição eleva a qualidade dos resultados, permitindo que os modelos capturem a natureza inerentemente agrupada das ocorrências de cavernas.

Observou-se que o refinamento técnico, especificamente a alteração iterativa de parâmetros via MDA, atuou como o principal vetor de melhoria na qualidade estatística, independentemente da arquitetura do modelo adotado. Este processo de otimização permitiu identificar que diferentes modelos atendem a necessidades operacionais distintas: enquanto o RF SpatialML consolidou-se

como a referência em qualidade estatística e calibração probabilística devido ao seu baixo Log-Loss, o modelo RF Espacial (+XY) com Iteração revelou-se a ferramenta mais equilibrada para o planejamento de prospecções, ao maximizar a recuperação de ocorrências reais (Recall) sem sacrificar excessivamente a precisão.

Do ponto de vista prático, a modelagem demonstrou ser uma alternativa superior aos métodos multicritério subjetivos, oferecendo uma base técnica robusta para mediar os conflitos entre a expansão do setor eólico e a preservação ambiental. Contudo, é imperativo reconhecer que, embora a predição tenha sido considerada satisfatória dentro dos limites da área prospectada, ela ainda apresenta limitações quando extrapolada para áreas regionais mais amplas.

Portanto, conclui-se que a inteligência preditiva é uma aliada estratégica no licenciamento ambiental, mas sua eficácia contínua depende da aquisição de novo conhecimento empírico. O sucesso da conservação espeleológica na Chapada Diamantina requer que esses modelos sejam retroalimentados por novas campanhas de campo, garantindo que o desenvolvimento energético ocorra em conformidade com a salvaguarda de sistemas cársticos ainda não revelados.

6 REFERÊNCIAS BIBLIOGRÁFICAS

ABEEÓLICA – ASSOCIAÇÃO BRASILEIRA DE ENERGIA EÓLICA. **Boletim anual de geração eólica 2023**. São Paulo: ABEEólica, 2023.

AULER, A. S. Karst areas in Brazil and the potential for major caves – an overview. **Boletín..,**, Madrid, v. 128, n. 4, p. 683-700, 2017.

AULER, A. S.; PILÓ, L. B. O Decreto 6.640/2008 e a explosão de registros de cavernas no Brasil: uma análise crítica. **Espeleo-Tema**, Campinas, v. 26, n. 1, p. 7-20, 2017.

BARRETO, A. M. F.; MENDES, I. D. A. Geomorfologia do Estado da Bahia: síntese do conhecimento. In: **Geologia da Bahia: pesquisa e atualização**. Salvador: CBPM, 2002. v. 1, p. 127-158.

BENITO, B. M. et al. The influence of spatial errors in species occurrence data used in distribution models. **Journal of Applied Ecology**, Hoboken, v. 58, n. 6, p. 1173-1182, 2021.

BRASIL. Decreto nº 6.640, de 7 de novembro de 2008. Dispõe sobre a proteção das cavidades naturais subterrâneas existentes no território nacional. **Diário Oficial da União**, Brasília, DF, 10 nov. 2008.

BRASIL. Lei nº 6.938, de 31 de agosto de 1981. Dispõe sobre a Política Nacional do Meio Ambiente, seus fins e mecanismos de formulação e aplicação, e dá outras providências. **Diário Oficial da União**, Brasília, DF, 2 set. 1981.

BRASIL. **Constituição da República Federativa do Brasil de 1988**. Brasília, DF: Presidência da República, 1988.

BREIMAN, L. Random Forests. **Machine Learning**, New York, v. 45, n. 1, p. 5-32, Oct. 2001.

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

CPRM – SERVIÇO GEOLÓGICO DO BRASIL. **Geodiversidade do Estado da Bahia**. Salvador: CPRM, 2015.

CUTLER, D. R. et al. Random forests for classification in ecology. **Ecology**, Washington, DC, v. 88, n. 11, p. 2783-2792, Nov. 2007.

DORMANN, C. F. et al. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. **Ecography**, Copenhagen, v. 36, n. 1, p. 27-46, jan. 2013.

EMBRAPA – EMPRESA BRASILEIRA DE PESQUISA AGROPECUÁRIA. **Semiárido brasileiro: uma abordagem sobre clima e vegetação**. Petrolina: Embrapa Semiárido, 2017.

EPE – EMPRESA DE PESQUISA ENERGÉTICA. **Plano Decenal de Expansão de Energia 2031**. Brasília: EPE, 2022.

FERREIRA, M. P. et al. Impactos da expansão de parques eólicos sobre a paisagem no semiárido brasileiro: uma análise baseada em geotecnologias. **Journal of Environmental Analysis and Progress**, Recife, v. 6, n. 2, p. 106-119, 2021.

FERREIRA, R. L. et al. Técnicas de prospecção espeleológica: abordagens sistemáticas e estratégicas em diferentes contextos geológicos. In: CONGRESSO BRASILEIRO DE ESPELEOLOGIA, 33., 2015, Campinas. **Anais**... Campinas: Sociedade Brasileira de Espeleologia, 2015. p. 45-62.

FLORINSKY, I. V. **Digital terrain analysis in soil science and geology**. Amsterdam: Academic Press, 2012.

FOOTHERINGHAM, A. S.; BRUNSDON, C.; CHARLTON, M. **Geographically weighted regression: the analysis of spatially varying relationships**. Chichester: John Wiley & Sons, 2003.

FRANKLIN, J. **Mapping species distributions: spatial inference and prediction**. Cambridge: Cambridge University Press, 2010.

GÉRON, A. **Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow**. 2nd ed. Sebastopol: O'Reilly Media, 2019.

GUISAN, A.; THUILLER, W. Predicting species distribution: offering more than simple habitat models. **Ecology Letters**, Oxford, v. 8, n. 9, p. 993-1009, set. 2005.

GUISAN, A.; THUILLER, W.; ZIMMERMANN, N. E. **Habitat suitability and distribution models: with applications in R**. Cambridge: Cambridge University Press, 2017.

GUYON, I.; ELISSEEFF, A. An introduction to variable and feature selection. **Journal of Machine Learning Research**, Cambridge, MA, v. 3, p. 1157-1182, mar. 2003.

HANLEY, J. A.; MCNEIL, B. J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. **Radiology**, Chicago, v. 143, n. 1, p. 29-36, abr. 1982.

HENGL, T. Finding the right pixel size. **Computers & Geosciences**, Amsterdam, v. 32, n. 9, p. 1283-1298, nov. 2006.

ICMBio – INSTITUTO CHICO MENDES DE CONSERVAÇÃO DA BIODIVERSIDADE. **Instrução Normativa nº 2, de 30 de agosto de 2017**. Estabelece critérios e procedimentos para a proteção, o manejo e o licenciamento ambiental de cavidades naturais subterrâneas no âmbito federal. **Diário Oficial da União**, Brasília, DF, 31 ago. 2017.

ICMBio – INSTITUTO CHICO MENDES DE CONSERVAÇÃO DA BIODIVERSIDADE. **Manual de boas práticas para o licenciamento ambiental em áreas cársticas**. Brasília: ICMBio/CECAV, 2019.

JAMES, G. et al. **An introduction to statistical learning: with applications in R**. 2nd ed. New York: Springer, 2021.

JENSEN, J. R. **Introductory digital image processing: a remote sensing perspective**. 4. ed. Glenview: Pearson, 2015.

KALOGIROU, S.; GEORGANOS, S. **SpatialML: Spatial Machine Learning**. [S.l.], 2019.

KARMANN, I.; SÁNCHEZ, L. E. Caracterização espeleológica de áreas cársticas. In: **Espeleologia: conceitos e técnicas**. 2. ed. Campinas: Instituto de Geociências - UNICAMP, 2008. p. 45-70.

KLIMCHOUK, A. B. **Speleogenesis: evolution of karst aquifers**. Huntsville: National Speleological Society, 2000.

KUHN, M.; JOHNSON, K. **Feature engineering and selection: a practical approach for predictive models**. Boca Raton: Chapman & Hall/CRC, 2019.

LIAW, A.; WIENER, M. Classification and regression by randomForest. **R News**, Vienna, v. 2, n. 3, p. 18-22, dez. 2002.

MALCZEWSKI, J.; RINNER, C. **Multicriteria decision analysis in geographic information science**. New York: Springer, 2015.

MURPHY, A. H. A new vector partition of the probability score. **Journal of Applied Meteorology and Climatology**, Boston, v. 12, n. 4, p. 595-600, jun. 1973.

O'BRIEN, R. M. A caution regarding rules of thumb for variance inflation factors. **Quality & Quantity**, Dordrecht, v. 41, n. 5, p. 673-690, out. 2007.

PALMER, A. N. Origin and morphology of limestone caves. **Geological Society of America Bulletin**, Boulder, v. 103, n. 1, p. 1-21, jan. 1991.

PLANT, R. E.; PAVLICK, T. M. Spatial cross-validation: a review. **Spatial Statistics**, Amsterdam, v. 42, 100478, abr. 2021.

POWERS, D. M. W. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness & correlation. **Journal of Machine Learning Technologies**, Sydney, v. 2, n. 1, p. 37-63, 2011.

ROBERTS, D. R. et al. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. **Ecography**, Copenhagen, v. 40, n. 8, p. 913-929, ago. 2017.

SAITO, T.; REHMSMEIER, M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. **PLOS ONE**, San Francisco, v. 10, n. 3, e0118432, mar. 2015.

SALLUN FILHO, W.; KARMANN, I. Proposta de classificação de cavernas em quartzito no Brasil. **Espeleo-Tema**, Campinas, v. 22, n. 1, p. 73-92, 2012.

SCHOBENHAUS, C.; CAMPOS, D. A.; DERZE, G. R.; ASMUS, H. E. (Ed.). **Geologia do Brasil**. Brasília: DNPM/CPRM, 1984.

SILVA, J. M. C.; LEAL, I. R.; TABARELLI, M. (Ed.). **Caatinga: desafios e oportunidades para o desenvolvimento sustentável**. Recife: Editora UFPE, 2017.

SILVA, M. A.; PESSÔA, F. S. Avaliação de impactos ambientais em áreas cársticas: uma abordagem voltada para a preservação do patrimônio espeleológico. **Revista Brasileira de Espeleologia**, Rio de Janeiro, v. 5, n. 1, p. 1-15, 2019.

VELOSO, H. P.; RANGEL FILHO, A. L. R.; LIMA, J. C. A. **Classificação da vegetação brasileira, adaptada a um sistema universal**. Rio de Janeiro: IBGE, 1991.

WHEELER, D. C.; TIEBOUT, C. M. Multicollinearity and correlation among local regression coefficients in geographically weighted regression. **Journal of Geographical Systems**, Berlin, v. 7, n. 2, p. 161-187, June 2005.

WILSON, J. P.; GALLANT, J. C. (ed.). **Terrain analysis: principles and applications**. New York: John Wiley & Sons, 2000.

.