

Modelagem da propensão à antropização no Morro São Pedro (Porto Alegre/RS) com Aprendizado de Máquina: Uma abordagem híbrida baseada em árvores

Modeling Anthropization Propensity in Morro São Pedro (Porto Alegre/RS) using Machine Learning: A Hybrid Tree-Based Approach

Andressa Garcia Fontana¹, Marcos Wellausen Dias de Freitas², Cláudio Wilson Mendes Júnior³, Carina Richardt de Carvalho¹ & Francisco Hélder Fernandes do Amaral⁴

¹Universidade Federal do Rio Grande do Sul (UFRGS), Programa de Pós-Graduação em Sensoriamento Remoto/PPGSR, Porto Alegre, Brasil. E-mail: andressagfontana94@gmail.com, carinardecavalho@gmail.com.

²Universidade Federal do Rio Grande do Sul (UFRGS), Departamento de Geografia, Porto Alegre, Brasil. E-mail: marcoswfreitas@gmail.com.

³Universidade Federal do Rio Grande do Sul (UFRGS), Departamento de Geodésia, Porto Alegre, Brasil. E-mail: claudio.mendes@ufrgs.br.

⁴Universidade Federal do Pará (UFPA), Departamento de Geografia, Belém, Brasil. E-mail: helter.fernandes@unesp.br.

Resumo: O mapeamento da antropização é essencial para compreender as mudanças ambientais e facilitar o planejamento territorial. Nesta pesquisa, o problema central foi analisar os impactos da atividade humana no Morro São Pedro, em Porto Alegre. O objetivo principal foi identificar como a paisagem local foi alterada ao longo do tempo. Para tal, foram utilizados três algoritmos de aprendizagem automática: *Parallel Random Forest* (PRF), *Regularized Random Forest* (RRF) e *XGBoost*. Os resultados indicaram que o modelo PRF foi o mais preciso entre os testados. Essa descoberta sugere que a presença de rodovias é um fator primário que influencia as mudanças no ambiente natural. A proximidade dessas estradas destacou-se como o fator mais significativo em todos os modelos, sugerindo que a construção e o uso de rodovias desempenham um papel fundamental na transformação da região. Em conclusão, os resultados deste estudo destacam a importância de incorporar os impactos das rodovias no planejamento territorial e ambiental, particularmente no que diz respeito à conservação dos ecossistemas e à gestão dos recursos naturais.

Palavras-chave: planejamento territorial; ecossistemas; recursos naturais.

Abstract: Mapping anthropization is essential for comprehending environmental changes and facilitating territorial planning. In this research, the central problem was to analyze the impacts of human activity on Morro São Pedro, Porto Alegre. The main objective was to identify how the local landscape has changed over time. To this end, three machine learning algorithms were employed: *Parallel Random Forest* (PRF), *Regularized Random Forest* (RRF), and *XGBoost*. The results indicated that the PRF model was the most accurate among those tested. This finding suggests that the presence of highways is a primary factor influencing changes in the natural environment. The proximity of these roads stood out as the most significant factor across all models, suggesting that the construction and use of highways play a pivotal role in the transformation of the region. In conclusion, the results of this study highlight the importance of incorporating the impacts of highways into territorial and environmental planning, particularly with regard to ecosystem conservation and natural resource management.

Keywords: territorial planning; ecosystems; natural resources.

1. Introdução

A valorização das paisagens é um tema cada vez mais estudado, devido a sua crescente importância em diversos campos do conhecimento, como planejamento urbano (Eilola *et al.*, 2023), gestão ambiental (Yang *et al.*, 2022), ou implementação de comunicação e infraestrutura (Zhang *et al.*, 2022). Além disso, nas últimas três décadas, a análise da antropização ou das transformações impostas pela sociedade humana sobre a paisagem natural tornou-se um parâmetro importante na busca pelo desenvolvimento socioambiental sustentável, em especial nas áreas sob influência de centros urbanos densamente povoados e com histórico de expansão ocupacional não planejada (Bertrand, 1984).

A proximidade dos setores censitários de 2010, que representam as favelas e comunidades urbanas caracterizadas por ocupações irregulares, infraestrutura precária, elevado adensamento populacional e vulnerabilidade social, intensifica a pressão sobre o ambiente (IBGE, 2010). A busca por moradia e recursos básicos, associada à falta de alternativas, impulsiona a expansão desordenada de favelas, promovendo a ocupação de áreas de risco, a degradação de ecossistemas e o uso inadequado de recursos naturais como água e solo.

Nesse contexto, as análises espaciais que empregam Sistemas de Informação Geográfica (SIG) podem oferecer uma contribuição significativa para o conhecimento científico acerca dos impactos da antropização (Palmer, 2019). Desta forma, o presente estudo possui o objetivo de analisar as paisagens antrópicas, que são definidas como ambientes que sofreram transformações substanciais em decorrência da atividade humana.

De acordo com Christoforetti (1999), as paisagens são o resultado da interdependência entre processos naturais e sociais, formando um sistema complexo e dinâmico que pode ser analisado em diferentes níveis ecológicos e epistemológicos. A antropização, por sua vez, molda a paisagem através de atividades como a urbanização, a agricultura e a industrialização, alterando sua estrutura, dinâmica e organização. Essas paisagens, construídas ou modificadas, refletem as relações socioeconômicas e culturais de uma sociedade (Sotchava, 1977; Coelho, 2001).

A aplicação de algoritmos de aprendizado de máquina (*Machine Learning* - ML) e a consideração de indicadores geoespaciais que influenciam a transformação da paisagem permitem a identificação de áreas mais propensas à antropização. Essa abordagem contribui para uma compreensão mais profunda da dinâmica espacial, da ocupação humana e dos impactos ambientais (Palmer, 2019).

É essencial desenvolver metodologias amplas e multidisciplinares que abordem os problemas ambientais causados pela ação humana (García-Ayllón, 2017). Embora os métodos tradicionais se baseiem em análises espaciais, eles frequentemente falham em estabelecer uma correlação precisa entre as alterações no território e os níveis de degradação ambiental (García-Ayllón, 2018).

As técnicas ML se mostram uma ferramenta eficaz para a modelagem de problemas ambientais. Ao automatizar a construção de modelos analíticos, elas superam, em muitos casos, a precisão de métodos tradicionais (Zhang *et al.*, 2019; Nsengiyumva & Valentino, 2020; Prasad *et al.*, 2021; Xie *et al.*, 2021; Sakaa *et al.*, 2022). Baseadas em *Artificial Intelligence* (IA), as técnicas de ML são capazes de aprender continuamente com os dados, o que as torna eficientes para a análise de dados, uma vez que identificam

padrões complexos e tomam decisões com menor intervenção humana (Rahmati *et al.*, 2017; Mohammed *et al.*, 2020; Elbeltagi *et al.*, 2022).

Historicamente, áreas próximas a drenagens são mais suscetíveis à ocupação humana, e os rios e córregos em áreas urbanas, em particular, atraem diversos interesses econômicos que intensificam a pressão sobre o uso do solo. Essa ocupação, aliada à remoção da vegetação nativa, especialmente das matas ciliares, favorece a erosão do solo, o que, por sua vez, agrava o assoreamento dos cursos d'água e compromete a estabilidade de encostas. Como resultado, há uma degradação dos ecossistemas aquáticos e um aumento do risco de desastres naturais, como os deslizamentos de terra (Christofidis *et al.*, 2019). A quantidade de pessoas por setor censitário é um indicador fundamental para compreender os processos de antropização e seus impactos ambientais. Em áreas com alta densidade populacional, a pressão antrópica é maior, resultando na impermeabilização do solo, desmatamento, degradação dos recursos hídricos, produção de resíduos sólidos, emissão de poluentes atmosféricos e poluição sonora.

Este estudo teve como objetivo mapear e analisar a dinâmica da antropização no Morro São Pedro (MSP), uma área de grande importância ecológica que enfrenta um crescente processo de urbanização. Para isso, foram utilizados dados de sensoriamento remoto e algoritmos de ML, especificamente Floresta Aleatória Paralela (*Parallel Random Forest* - PRF), Floresta Aleatória Regularizada (*Regularized Random Forest* - RRF) e XGBoost, para modelar a propensão à ocupação antrópica na região. A combinação dos resultados desses modelos permitiu gerar mapas de propensão à antropização mais precisos e detalhados.

Esses mapas podem servir como ferramenta estratégica para o planejamento ambiental e a gestão territorial do MSP. Ao identificar as áreas mais vulneráveis à ocupação desordenada, torna-se possível implementar medidas de conservação e recuperação ambiental mais eficientes, contribuindo para a preservação da biodiversidade local e a melhoria da qualidade de vida da população.

2. Área de Estudo

O MSP está situado entre os paralelos 30°08'S e 30°12'S e os meridianos 51°05'W e 51°07'W, localizado na zona sul de Porto Alegre (RS). Possui 1.259 ha, atingindo 289 metros de altitude, destacando-se como um dos pontos mais altos desse município (Hasenack *et al.*, 2008). Esse morro abrange diferentes bairros da zona sul de Porto Alegre, incluindo Pitinga, Lomba do Pinheiro, São Caetano, Restinga, Lageado, Chapéu do Sol e Boa Vista do Sul, configurando um importante corredor ecológico na região (Fig. 1).

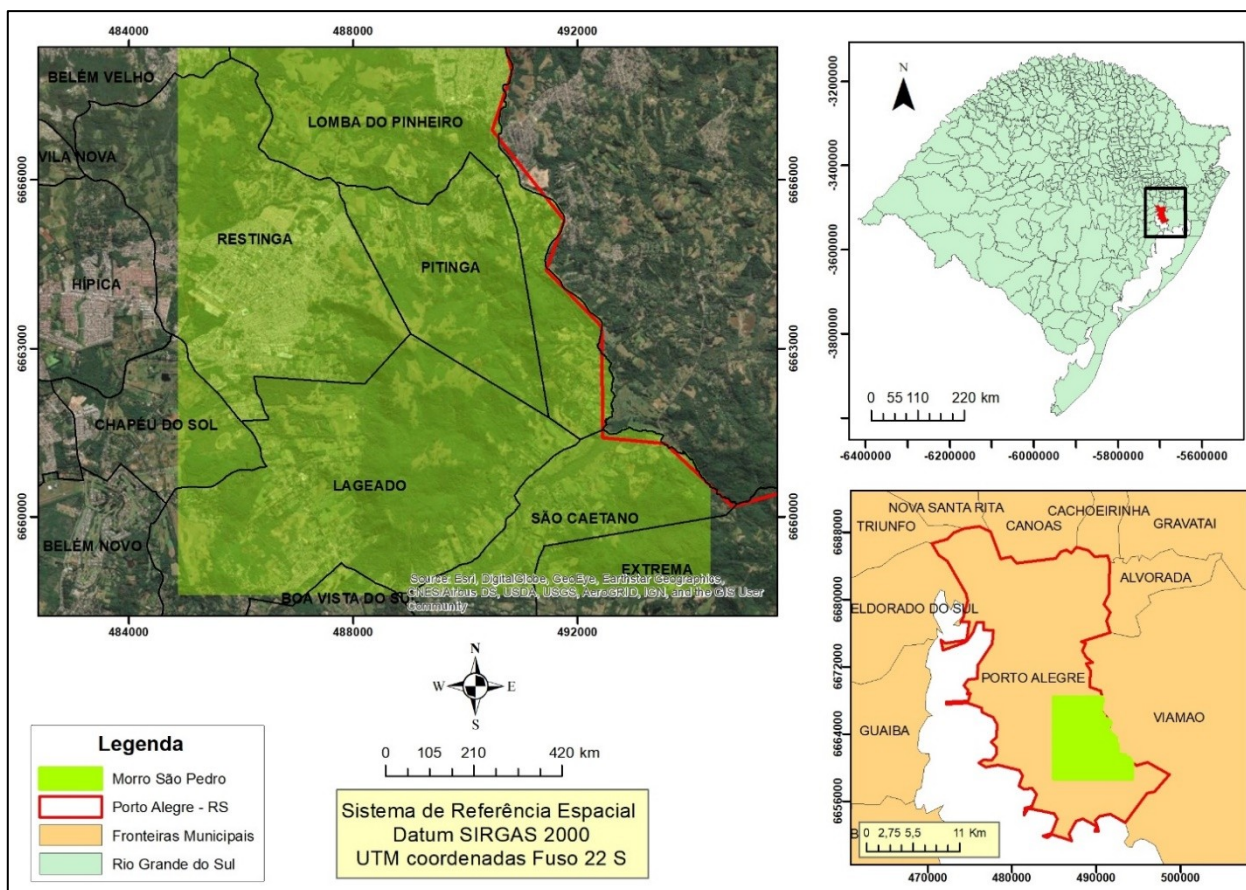


Figura 1. Localização do Morro São Pedro, no município de Porto Alegre, estado do Rio Grande do Sul.

Figure 1. Location of Morro São Pedro, in the municipality of Porto Alegre, state of Rio Grande do Sul.

O MSP atua como um reservatório natural, abastecendo os arroios Salso e Lami. As águas dessas nascentes contribuem significativamente para a bacia do Rio Guaíba, desempenhando um papel fundamental no equilíbrio do ecossistema local (Rambo, 2000).

3. Materiais e Métodos

Para mapear e analisar a propensão à antropização no MSP, foram empregadas técnicas de ML e dados de sensoriamento remoto. A metodologia envolveu a elaboração de mapas de uso e cobertura da terra (*Land Use and Land Cover - LULC*) e a utilização de doze fatores condicionantes (Fig. 2 e 3), classificados em quatro categorias: Topografia (altitude, declividade, geomorfologia, unidades *geomorphons* e HAND) (Gouveia & Ross, 2019), Hidrologia (proximidade dos rios e HAND) (Goerl *et al.*, 2017), Morfoestrutura (geologia e perfis de solo) (Silva & Bacani, 2017) e Influência Humana (proximidade das áreas desmatadas, proximidade das favelas e comunidades urbanas de 2010, número de pessoas por setor e proximidade das rodovias) (Souza & Vieira, 2018).

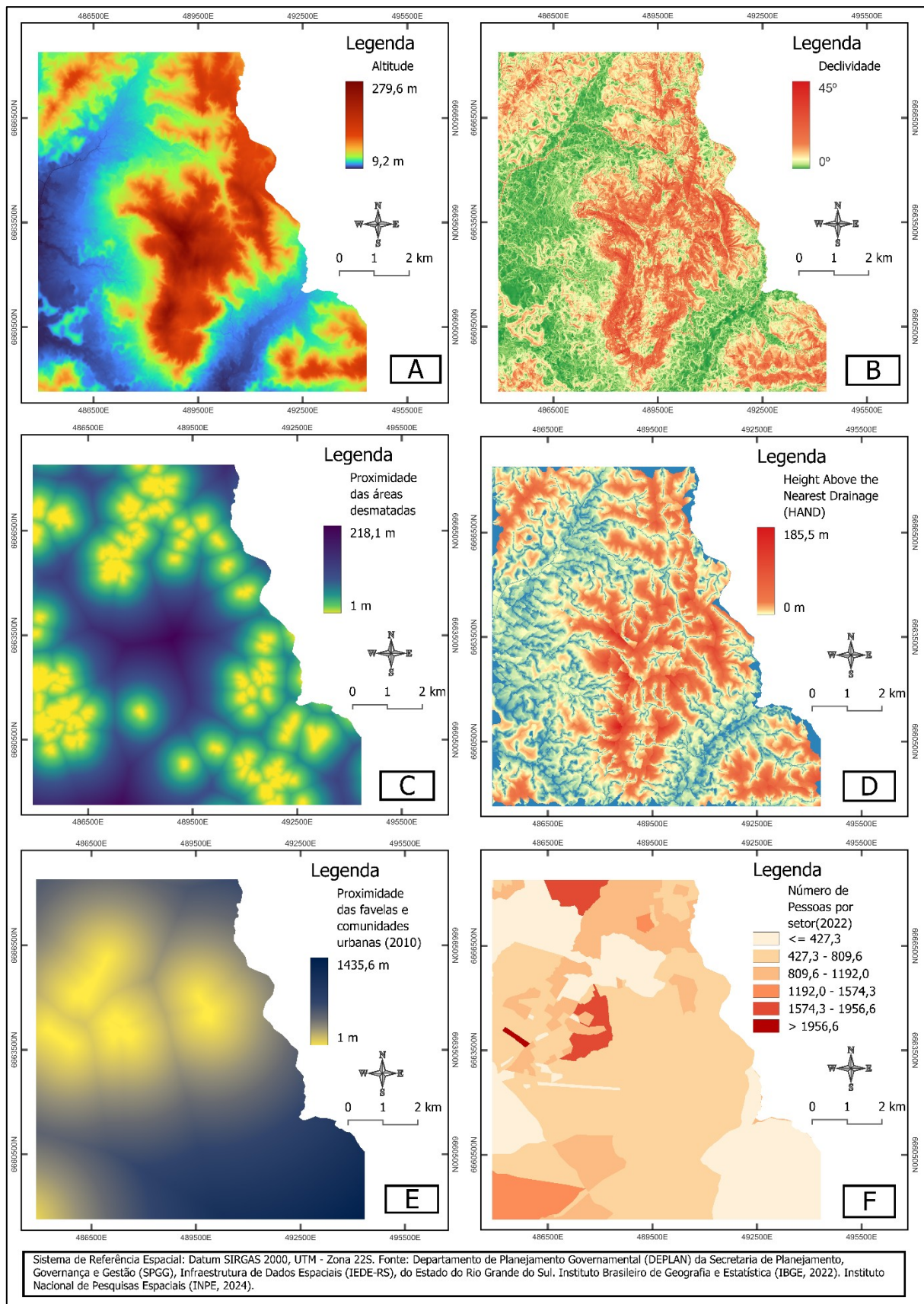


Figura 2. Mapas de variáveis ambientais e socioeconômicas da área de estudo.

Figure 2. Maps of environmental and socioeconomic variables in the study area.

A Figura 3 apresenta um conjunto de seis mapas temáticos que, em conjunto, constituem uma análise integrada da paisagem da RMPA. Os mapas detalham aspectos geocientíficos fundamentais, especificamente: (A) a proximidade com a rede de drenagem; (B) a proximidade com a malha viária; (C) as formas do relevo (*Geomorphons*); (D) a geologia; (E) a distribuição dos tipos de solo; e (F) as unidades geomorfológicas.

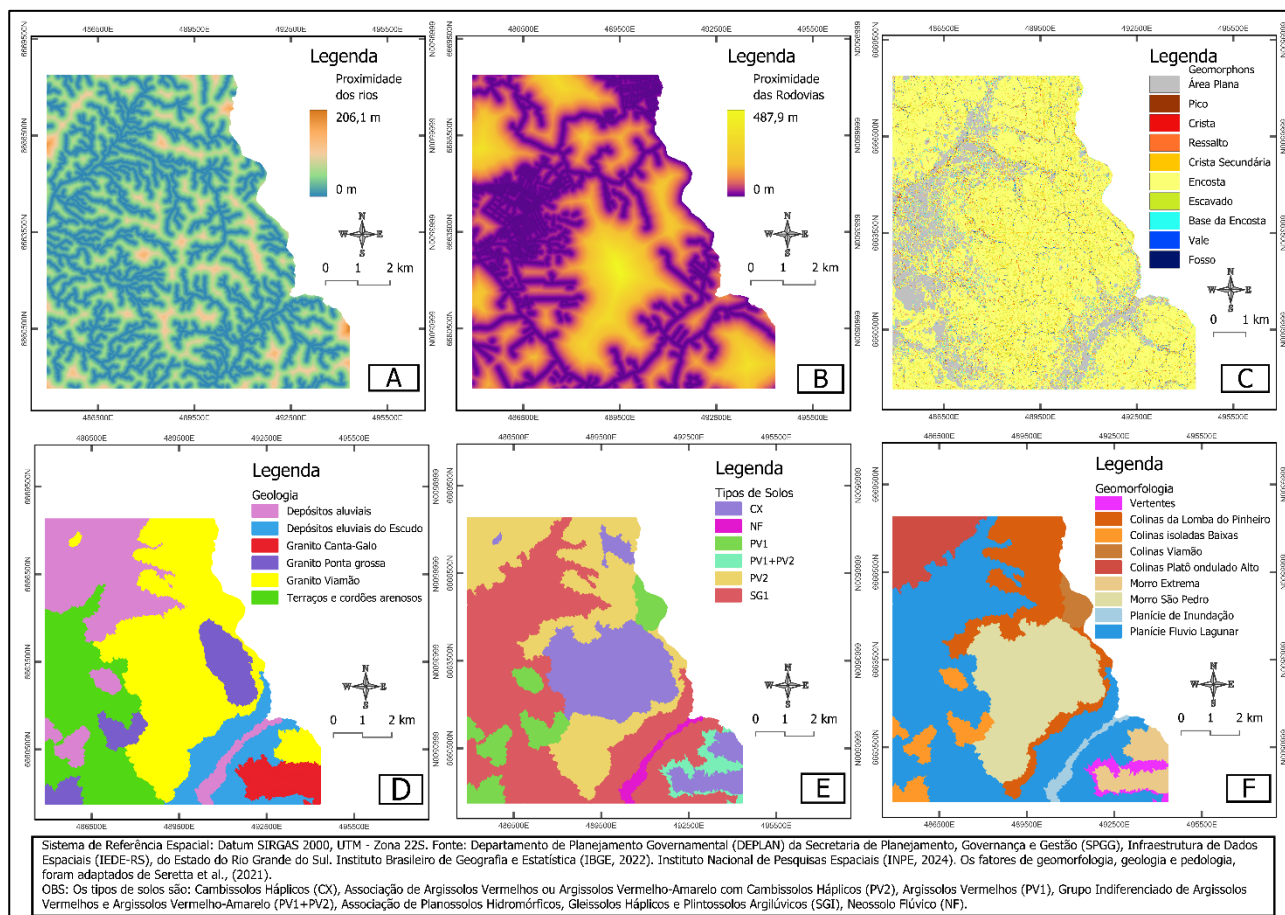


Figura 3. Mapas de variáveis do relevo da área de estudo.

Figure 3. Maps of relief variables in the study area.

A partir dessa base de dados, foram aplicados três algoritmos de ML: PRF, RRF e XGBoost e, posteriormente, foi realizada uma análise da importância dos fatores condicionantes utilizados nesses algoritmos, para identificar quais exercem maior influência na propensão à antropização. A metodologia está detalhada no fluxograma da Figura 4.

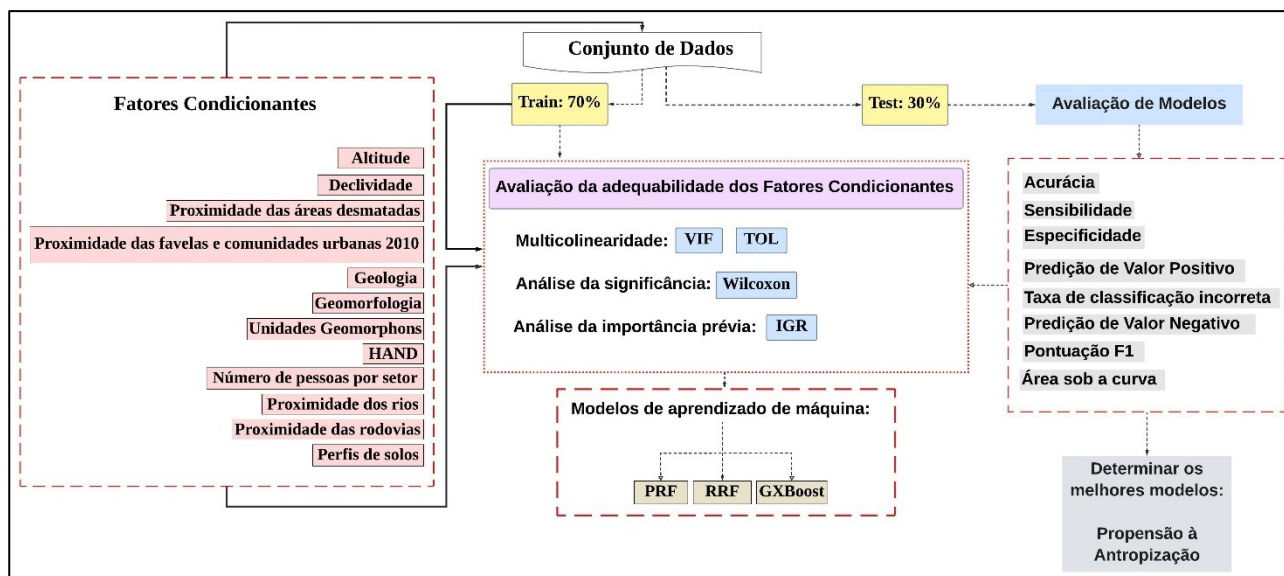


Figura 4. Fluxograma metodológico.

Figure 4. Methodological flowchart.

Os mapas de LULC foram gerados utilizando imagens de alta resolução espacial. Para o ano de 2010, foram empregadas imagens *RapidEye*, com resolução de 5 metros, enquanto que para 2024, utilizaram-se imagens *PlanetScope*, com resolução de 3 metros. A classificação dos mapas de LULC foi realizada por meio da Análise de Imagens Orientada a Objetos (*Geographic Object-Based Image Analysis - GEOBIA*), utilizando o *software eCognition*. Nesse processo, as imagens foram inicialmente segmentadas em objetos homogêneos, utilizando o algoritmo de segmentação multi-resolução.

Posteriormente, foram selecionadas amostras de treinamento para cada classe de cobertura da terra da área de estudo. A partir dessas amostras, foram extraídos atributos espectrais e espaciais dos segmentos (objetos). Por fim, estes objetos foram classificados, utilizando o algoritmo Floresta Aleatória (*Random Forest - RF*), que, com base nos atributos extraídos, atribuiu-se uma classe a cada objeto.

A fim de identificar as áreas mais suscetíveis à antropização, foi realizada uma análise temporal das mudanças de uso e cobertura da terra, comparando os mapas do MSP de 2010 e 2024. Através da ferramenta álgebra de mapas no *software* QGIS, foram quantificadas e espacializadas as transições entre as diferentes classes de cobertura, permitindo mapear as áreas que apresentaram maior dinâmica e, conseqüentemente, maior propensão à urbanização e outras formas de ocupação antrópica.

Para identificar as áreas com maior propensão à antropização, foi realizado um processo de amostragem estratificada utilizando o *software* livre QGIS versão 3.34.6. Inicialmente, foram vetorizadas manualmente as áreas que sofreram alguma forma de alteração antrópica entre 2010 e 2024, criando um *GeoPackage*. Em seguida, foram gerados aleatoriamente 3.000 pontos dentro dessas áreas, garantindo assim uma representação adequada das diferentes classes de uso antrópico. Como grupo de controle, foi criado outro conjunto de 3.000 pontos em áreas que mantiveram sua cobertura natural ao longo do período, permitindo comparar as características das áreas com e sem ocorrência de antropização.

Dessa forma, foram coletados dados de 6.000 pontos geográficos, divididos igualmente entre áreas urbanas e áreas naturais no MSP. Para cada ponto, foram

registradas as coordenadas geográficas (x e y). Com o objetivo de identificar os fatores que influenciam a ocupação humana, foi desenvolvido um modelo de regressão logística para prever a probabilidade de um local ser urbanizado. Nesse modelo, cada local recebeu uma pontuação de 0 (área natural) ou 1 (área urbanizada), representando a presença ou ausência de atividades antrópicas. Para facilitar a análise, os dados referentes a esses fatores foram processados e transformados em um formato padronizado. Os dados foram convertidos e reamostrados para resolução espacial de 3 metros, representados no sistema de coordenadas UTM, zona 22 Sul.

A variação do relevo (Fig. 5), especialmente a altitude e a declividade, afeta fatores como evaporação, temperatura do solo, umidade do solo e radiação solar. Essas condições influenciam a suscetibilidade dos solos a processos erosivos e outros tipos de degradação, como deslizamentos, tornando-os mais vulneráveis (Fenta *et al.*, 2020).

Os fatores topográficos declividade e curvatura do relevo foram calculados utilizando o software livre QGIS versão 3.34.6, a partir de um Modelo Digital de Terreno (MDT) de Porto Alegre (RS) com resolução espacial de 1 metro, gerado por perfilhamento a laser aerotransportado em 2010 (SMAMUS, 2022). As características dos perfis de solos, em conjunto com a geologia e a geomorfologia da região, definem as condições naturais do ambiente e influenciam diretamente a sua vulnerabilidade aos processos de degradação ambiental. Solos com baixa fertilidade, elevada suscetibilidade à erosão, textura arenosa ou baixa profundidade, por exemplo, são mais propensos a se degradar quando submetidos a práticas de uso do solo intensivas, como a agricultura e a urbanização (Fenta *et al.*, 2020).

A geologia local exerce um papel fundamental na definição da capacidade de suporte do terreno para construções e infraestrutura, além de influenciar a disponibilidade de recursos minerais e hídricos. Neste estudo, foram utilizados dados vetoriais de pedologia e geologia do Diagnóstico Ambiental de Porto Alegre (escala 1:15.000) (Hasenack *et al.*, 2008) para analisar a relação entre as características geológicas e a ocupação do solo. A caracterização geomorfológica foi obtida a partir de mapa geomorfológico na escala 1:50.000, produzido por (Moura *et al.*, 2013). Além disso, foram utilizados dados de unidades do relevo (*geomorphons*) da área de estudo, que estão relacionadas com o escoamento superficial e a deposição de sedimentos. Estes dados foram derivados do MDT de Porto Alegre (2010), com a ferramenta *geomorphon* do *software* GRASS GIS.

A análise hidrológica foi realizada utilizando a modelo HAND (*Height Above the Nearest Drainage*), que emprega uma abordagem topo-hidrológica para realinhar a altimetria de acordo com a rede de drenagem. Para processamento do modelo HAND, foi utilizado um *software* desenvolvido em linguagem C++ pelos pesquisadores do Instituto Nacional de Pesquisas Espaciais (INPE) (Rennó *et al.*, 2008). Ao calcular a altura acima da drenagem mais próxima, essa ferramenta permite identificar áreas com maior potencial de concentração de água e, conseqüentemente, com maior risco de processos erosivos.

A relação entre o valor obtido por HAND e o nível de saturação hídrica do solo é bem estabelecida na literatura (Rennó *et al.*, 2008), permitindo discriminar zonas de instabilidade hídrica e de maior susceptibilidade à erosão e deposição de sedimentos. O modelo HAND é eficaz para delimitar várzeas e áreas inundáveis e, ao identificar áreas com maior declive, auxilia na avaliação do potencial de erosão hídrica, sendo fundamental para o planejamento ambiental. No entanto, para uma análise completa da dinâmica ambiental dessas áreas, é necessário integrar o HAND com outras informações, como tipo de solo, cobertura vegetal, regime hidrológico e histórico do uso da terra.

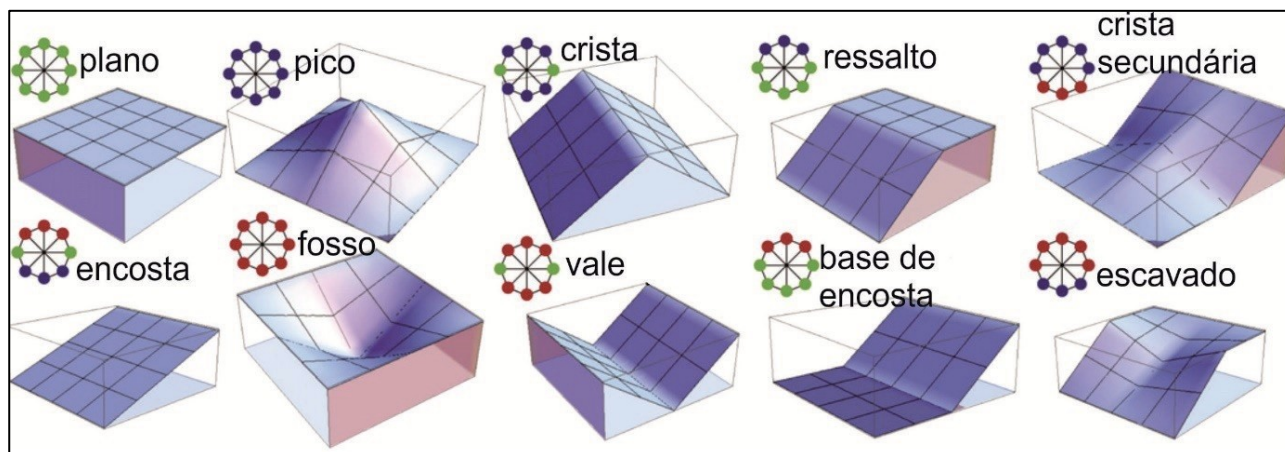


Figura 5. Elementos de relevo obtidos dos *geomorphons*. Fonte: Adaptado de Jasiewicz; Stepinski (2013).

Figure 5. Relief elements obtained from *geomorphons*. Source: Adapted from Jasiewicz; Stepinski (2013).

Neste estudo, a camada de proximidade das drenagens foi gerada utilizando a ferramenta de proximidade do QGIS. A densidade populacional por setor censitário foi calculada dividindo o número total de habitantes da área de estudo pela área de cada setor, conforme os dados do IBGE (2010). O arquivo georreferenciado (*shapefile*) resultante foi convertido para dados matriciais (QGIS), utilizando o método de interpolação (*Inverse Distance Weighting* - IDW), com uma resolução espacial de 3 metros, adequada para a escala do estudo e para identificar padrões de concentração populacional. A distância euclidiana entre os centroides dos setores censitários e a rede viária foi calculada para avaliar a influência da proximidade de estradas na ocupação do solo.

Foram utilizados dados vetoriais da rede rodoviária e intraurbana, obtidos do *OpenStreetMap* (<https://www.openstreetmap.org/>), e a base vetorial foi corrigida para eliminar inconsistências e erros. Em seguida, foi convertida para dados matriciais, facilitando a realização de análises espaciais. Utilizando o software QGIS, calculou-se a distância euclidiana de cada ponto à rede viária, gerando uma nova camada matricial, que representa a proximidade a estradas.

Para quantificar a influência do desmatamento na dinâmica da vegetação, foi calculada a distância de cada ponto amostral à área desmatada mais próxima. Utilizou-se a base vetorial do Projeto de Monitoramento do Desmatamento na Amazônia Legal por Satélite (PRODES), do Instituto Nacional de Pesquisas Espaciais (INPE), contendo os polígonos de desmatamento para o período de 2014 a 2024 (INPE, 2024).

Para gerar o fator de proximidade, os dados vetoriais dos setores censitários do IBGE (2022) foram unidos aos dados de favelas e comunidades urbanas de 2010. O arquivo vetorial resultante dessa junção foi convertido para formato matricial, com resolução espacial de 3m, considerando a escala dos dados e a necessidade de capturar detalhes da distribuição espacial de favelas e comunidades urbanas. Utilizando o algoritmo de distância euclidiana, foi calculada a distância de cada pixel da imagem à área mais próxima de favelas e comunidades urbanas.

3.1 Avaliação da adequabilidade dos Fatores Condicionantes

Nesta pesquisa, foi utilizada a análise de multicolinearidade, um método de análise estatística utilizado para identificar a dependência linear entre dois ou mais fatores condicionantes dentro de um modelo preditivo (Dormann *et al.*, 2013). Quando fatores geoestruturais estão fortemente correlacionados entre si, isso pode comprometer a validade do modelo e reduzir a precisão dos resultados. Essa alta correlação entre fatores condicionantes faz com que o modelo tenha dificuldade em distinguir a contribuição individual de cada fator, levando a estimativas menos confiáveis (Darabi *et al.*, 2021).

Para avaliar e mitigar os efeitos da multicolinearidade, são comumente utilizadas as métricas de Fator de Inflação da Variância (*Variance Inflation Factor* - VIF) e Tolerância (*Tolerance* - TOL) (Band *et al.*, 2020). O VIF mede o quanto a variância de um coeficiente de regressão é inflacionada devido à colinearidade entre as variáveis independentes: valores de VIF superiores a 5 ou 10 geralmente indicam um problema de multicolinearidade significativa (Band *et al.*, 2020). Já a TOL, que é o inverso do VIF, reflete a proporção de variabilidade de uma variável que não é explicada por outras variáveis nos modelos: assim, valores de TOL inferiores a 0,1 ou 0,2 também indicam alta multicolinearidade (Zhu & Huang, 2006). Recomenda-se a remoção dos fatores condicionantes com alta multicolinearidade (Whang *et al.*, 2020). As métricas de TOL e VIF foram calculadas usando as equações a seguir:

$$TOL = 1 - R_j^2 \quad \text{Eq. 1}$$

$$VIF = \frac{1}{TOL} \quad \text{Eq. 2}$$

onde R_j^2 representa o valor de regressão de j em relação a outras variáveis diferentes em um conjunto de dados.

Para verificar a existência de representatividade entre os fatores condicionantes e amostras de propensão à antropização, esta pesquisa utilizou o teste de Wilcoxon para duas amostras independentes que consiste em um teste não paramétrico, utilizado para comparar as medianas de duas amostras quando as suposições do teste t não são atendidas, sendo aplicado para identificar diferenças significativas entre os grupos. O resultado é baseado no valor de W , que representa a soma das posições (*ranks*) de um dos grupos. O p -valor indica a probabilidade de obter um resultado tão extremo ou mais, considerando que não há diferença entre os grupos. Se o p -valor for menor que 5% (ou seja, $p < 0,05$), é rejeitada a hipótese de que os grupos são iguais e conclui-se que há uma diferença significativa entre eles (Wilcoxon, 1945).

A seleção de atributos é uma etapa crucial no desenvolvimento de modelos de classificação binária. Neste estudo, utilizou-se a Razão de Ganho de Informação (*Information Gain Ratio* - IGR) como critério de avaliação de atributos. O IGR, uma medida de importância de atributos, quantifica a redução na entropia de um conjunto de dados após a divisão por um determinado atributo. A escolha do IGR se justifica por sua capacidade de lidar com atributos com muitos valores distintos e por sua ampla utilização em algoritmos de aprendizado de máquina, como árvores de decisão (Shahzad *et al.*, 2022). No contexto deste estudo, o IGR foi empregado para identificar os fatores condicionantes que mais contribuem para a probabilidade de ocorrência de antropização,

ou seja, aqueles que maximizam a redução da incerteza na classificação das áreas como antropizadas ou não (Saha *et al.*, 2021).

Fatores condicionantes com valores de IGR próximos a zero indicam uma baixa correlação com a variável alvo e, conseqüentemente, uma baixa capacidade de discriminar entre as classes. A inclusão desses atributos no modelo pode levar ao sobreajustamento e à diminuição da generalização do modelo. Neste estudo, adotou-se um limiar de 0,05 para a exclusão de atributos, considerando que valores abaixo desse limiar indicam uma contribuição negligível para a redução da entropia. Essa prática visa otimizar o desempenho preditivo do modelo. O IGR foi calculado utilizando a Equação 3:

$$\text{Gain ratio } (x, Z) = \frac{\text{Entropy}(z) - \sum_{i=1}^n \frac{|z_i|}{z} \text{Entropy}(z_i)}{- \sum_{i=1}^n \frac{|z_i|}{z} \log \frac{|z_i|}{z}} \quad \text{Eq. 3}$$

onde, a partir do ponto de treinamento Z com $Z_i1 = 1, 2, 3, \dots, n$ subconjuntos, o atributo x pertence.

3.2 Algoritmos utilizados na modelagem espacial

Deng (2012) propôs um novo método de seleção de recursos que utiliza um único modelo de conjunto para analisar cada elemento dos dados de treinamento. Essa abordagem difere do método de vários conjuntos proposto por Tuv (2009), onde múltiplos modelos são combinados. Nessa pesquisa, cada nó da árvore de decisão examina individualmente cada exemplo de treinamento. O RRF é uma variação do RF que utiliza um critério de divisão mais sofisticado para construir as árvores de decisão (Deng, 2013). Ao invés do tradicional ganho de informação, o RRF emprega o ganho de informação regularizado (GainR) (X_i, v) (Eq. 4):

$$\text{Gain}_R(X_i, v) = \begin{cases} \lambda \times \text{Gain}(X_i, v) & i \notin F \\ \text{Gain}(X_i, v) & i \in F \end{cases} \quad \text{Eq. 4}$$

onde F representa o conjunto de recursos usados para dividir os nós da árvore, $\lambda \in [0, 1]$ caracteriza o coeficiente de penalidade e $i \notin F$ representa o coeficiente penalizado pelo i-ésimo recurso do nó de divisão v. Aqui, λ significa uma penalidade maior.

No RRF, a regularização mínima ocorre quando $\lambda = 1$ e é chamada de RRF (1). Em RRF (1), a seleção do subconjunto de características é conhecida como o subconjunto menor regularizado, o que indica regularização mínima de RRF. A regularização em árvores de decisão é uma técnica utilizada para evitar que o modelo se torne excessivamente complexo e se ajuste demais aos dados de treinamento (*overfitting*). Este método reduz o uso de características irrelevantes ou ruidosas, evitando regras complexas que prejudicam a generalização do modelo. Além disso, a regularização ajuda a evitar interações espúrias, ou seja, relações entre variáveis que parecem existir nos dados de treinamento, mas não são significativas no mundo real. A afirmação de Deng (2013) indica que, em um contexto mais amplo, a ideia de penalizar coeficientes para selecionar características mais relevantes se aplica também a algoritmos como árvores de

decisão. Ao ajustar os hiperparâmetros da árvore, ele está indiretamente controlando a importância que cada característica tem na divisão dos nós.

Nesta pesquisa, também foi utilizado o algoritmo XGBoost, considerado excelente para aprendizagem em conjunto (Wang *et al.*, 2015; Hosseiny *et al.*, 2020). Comparado com outros modelos de conjunto, pode melhorar a robustez do modelo através da introdução de termos regulares e amostragem de colunas e quando cada árvore seleciona o ponto de divisão, uma estratégia de paralelização irá ser adotado para melhorar a velocidade de corrida do modelo (Ji *et al.*, 2019; Budholiya *et al.*, 2022). Comparado com outros modelos de conjunto, o algoritmo XGBoost pode melhorar a robustez do modelo através da introdução de termos regulares e amostragem de colunas e quando cada árvore seleciona o ponto de divisão, uma estratégia de paralelização irá ser adotada para melhorar a velocidade de corrida do modelo (Budholiya *et al.*, 2022).

A aprendizagem integrada consiste em combinar múltiplas técnicas de aprendizado de máquina, como árvores de decisão, redes neurais e métodos de ensemble, para construir modelos mais robustos e com melhor capacidade de generalização. O XGBoost, proposto por Chen & Guestrin (2016), é um exemplo proeminente dessa abordagem. Esse algoritmo de *boosting*, baseado em árvores de decisão, constrói um modelo aditivo forte a partir de uma sequência de árvores de decisão fracas. Cada nova árvore é treinada para corrigir os erros da árvore anterior, utilizando um mecanismo de *gradient boosting*. Essa abordagem sequencial, combinada com técnicas de regularização, permite que o XGBoost obtenha alta precisão e evite o *overfitting*.

O XGBoost utiliza Árvores de Classificação e Regressão (*Classification and Regression Trees* – CART) como classificadores base e, em cada iteração, constrói as árvores de forma a minimizar uma função de perda específica. A amostra de entrada de cada árvore é ponderada de acordo com os erros cometidos pelas árvores anteriores, atribuindo maior importância às observações mal classificadas. Essa estratégia permite que o modelo se concentre nas instâncias mais difíceis de classificar, melhorando seu desempenho geral. Supondo que o modelo XGBoost seja composto por K CARTs, ele pode ser expresso pela Equação 5:

$$\hat{y}_i = \phi(x_i) = \sum_{k=1}^K f_k(x_i), f_k \in F \quad \text{Eq. 5}$$

onde $\hat{y}_i$ representa o valor previsto, f_k caracteriza a k-ésima árvore, $f_k(x_i)$ é a pontuação da i-ésima amostra na k-ésima árvore, K é o número total de amostras, x_i são os i-ésimos dados de entrada e F são todos os CARTs possíveis.

O XGBoost é semelhante à maioria dos modelos de aprendizado de máquina e seu objetivo pode ser a soma da função de perda e do termo regular, controlando respectivamente a precisão e a complexidade do modelo. As equações específicas são as seguintes:

$$L(\phi) = l(\phi) + \Omega(\phi) = \sum_{i=1}^n l(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k) \quad \text{Eq. 6}$$

onde

$$\Omega(f) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T \omega_j^2 \quad \text{Eq. 7}$$

onde y_i é o valor observado e l é uma função de perda, usada principalmente para medir a diferença entre y_i e $\hat{y}_i$. O segundo termo Ω é a regularização de prazo para penalizar a complexidade do modelo, evitando assim sobreajuste; γ representa a complexidade de cada folha; T significa o número total de folhas da árvore de decisão; λ é um parâmetro de compromisso aplicado principalmente para dimensionamento da pena; e w_j representa a pontuação na j -ésima folha.

O terceiro algoritmo utilizado nesta pesquisa foi o PRF e, de acordo com Band *et al.* (2020), os algoritmos tradicionais de RF apresentam limitações ao lidar com grandes volumes de dados. Para superar essas restrições, foi desenvolvida uma versão paralelizada do RF, conhecida como PRF. Enquanto o RF tradicional utiliza um conjunto de múltiplas árvores de decisão para tarefas de classificação, a PRF distribui o processamento dessas árvores em paralelo, otimizando o desempenho e reduzindo o tempo computacional necessário para a análise.

No PRF, o RF é segmentado em múltiplas subflorestas, que são processadas em paralelo. Posteriormente, essas subflorestas são reunidas para formar uma floresta maior no núcleo do processo. A análise no PRF é realizada em duas etapas: a primeira é o mapeamento, que gera pares de chave-valor, onde cada chave representa o índice de um dado específico; a segunda é a redução, que examina esses pares para obter o resultado final (Yang *et al.*, 2022).

O resultado final é uma floresta de decisão consolidada F_{final} , que representa a combinação de todas as subflorestas. A floresta final F_{final} é então utilizada para fazer previsões sobre novos dados. Por fim, cada árvore na floresta realiza uma previsão e a decisão final é baseada na agregação das previsões de todas as árvores (Band *et al.*, 2020). A combinação de múltiplas árvores de decisão melhora a robustez e a generalização do modelo. Como cada árvore pode capturar diferentes padrões dos dados, a agregação das suas previsões reduz o risco de *overfitting* e melhora a precisão do modelo.

3.3 Validação dos algoritmos utilizados na modelagem espacial

Avaliar a precisão dos mapas que preveem a degradação ambiental é fundamental para compreender a eficácia dos modelos computacionais na identificação de áreas naturais propensas à transformação em regiões de atividades humanas. Esses modelos, que empregam algoritmos de aprendizado de máquina para classificar áreas degradadas e não degradadas, são ferramentas eficazes para o planejamento ambiental.

Entretanto, é crucial avaliar a eficácia desses modelos por meio de métricas como exatidão e precisão, levando em consideração também as limitações dos dados e dos algoritmos utilizados. Estudos como o de Mousavi *et al.* (2017), têm investigado essa temática, ressaltando a importância de uma avaliação rigorosa para assegurar a confiabilidade dessas previsões. Neste estudo, foram utilizadas diversas métricas estatísticas aplicadas aos conjuntos de dados de validação, os quais representam 30% das amostras de propensão e não propensão. Essas métricas podem ser classificadas em dois perfis de abordagem avaliativa: métricas independentes e métricas dependentes de corte (Darabi *et al.*, 2021).

As métricas estatísticas dependentes de corte empregadas nesta pesquisa foram derivadas diretamente das matrizes de contingência binárias específicas para cada

modelo de ML. É importante salientar que as matrizes são compostas pelos valores de verdadeiro positivo (*True Positive* - TP) e verdadeiro negativo (*True Negative* - TN), os quais representam, respectivamente, falso positivo (*False Positive*) e falso negativo (*False Negative*). No contexto da propensão à degradação, os valores de TP e TN correspondem ao número de pixels corretamente identificados como áreas propensas e não propensas à migração de uma condição natural para uma condição de antropização, respectivamente. Por outro lado, os valores de FP e FN referem-se aos pixels que foram erroneamente catalogados, sendo FP aqueles identificados incorretamente como propensos quando, na verdade, não são, e FN aqueles rotulados incorretamente como não propensos à antropização quando, na realidade, são propensos (Kadavi *et al.*, 2018).

A partir das matrizes, foram estimadas as seguintes métricas: sensibilidade (*sensitivity* - SST), especificidade (*specificity* - SPF), valor preditivo positivo (*positive predictive value* - PPV), valor preditivo negativo (*negative predictive value* - NPV), F1 Score e taxa de classificação incorreta (*misclassification rate* - MR) (Darabi *et al.*, 2021). Para as métricas dependentes de corte, com exceção da MR, resultados elevados indicam um melhor desempenho dos modelos, enquanto valores mais baixos sinalizam um desempenho inferior (Band *et al.*, 2020).

A Área Sob a Curva de Características Operacionais do Receptor (*Area Under the Curve of the Receiver Operating Characteristic* - AUC-ROC) é uma das principais métricas independentes de corte, destacando-se por sua robustez e recorrente aplicabilidade na avaliação da precisão de modelos binários de ML, utilizados na previsão de fenômenos como erosão do solo, inundações e deslizamentos de terra (Kadavi *et al.*, 2018; Motevalli *et al.*, 2019; Nhu *et al.*, 2020; Bag *et al.*, 2022).

A curva ROC é construída por meio da plotagem da sensibilidade (taxa de verdadeiros positivos – TPR) no eixo Y em relação à taxa de falsos positivos ($1 - \text{especificidade}$ – FPR) no eixo X, permitindo uma avaliação precisa dos modelos ao mensurar sua capacidade de diferenciar corretamente entre classes positivas e negativas em distintos limiares de decisão. Esse procedimento considera a eficiência do modelo em minimizar erros e maximizar acertos (Phan *et al.*, 2019). Um valor de AUC-ROC próximo de 1 indica um desempenho excelente, enquanto valores próximos de 0,5 sugerem uma performance equivalente ao acaso, o que pode caracterizar o modelo como ineficiente ou até mesmo inadequado para a tarefa proposta (Band *et al.*, 2020).

4. Resultados e Discussão

Os dados dos mapeamentos de LULC de 2010 e 2024 confirmam que o MSP passou por um processo acelerado de antropização nesse período, resultando em impactos diretos sobre a biodiversidade e os recursos hídricos da região. O aumento de aproximadamente 100 hectares na área urbana e 65 hectares em áreas classificadas como Campo Antropizado, conforme os dados quantitativos apresentados na Tabela 1, indica uma pressão crescente sobre os recursos naturais e a biodiversidade local.

Tabela 1. Áreas das classes de LULC ao longo da série histórica no MSP.

Table 1. Areas of LULC classes throughout the historical series in MSP.

Código de classificação	Classes	Área (ha)		Variação (%) (2010 a 2024)
		2010	2024	
1	Floresta	3.854,46	3.742,03	-2,92%
2	Campo Nativo	178,38	205,06	14,96%
3	Campo Antropizado	1.380,92	1.445,28	4,66%
4	Silvicultura	243,25	208,79	-14,17%
5	Agricultura	251,16	211,16	-15,93%
6	Urbano	1.294,56	1.394,70	7,74%
7	Água	83,76	83,76	0

A presença das comunidades indígenas Mbyá-Guarani, que empregam práticas de agricultura sustentável e gestão tradicional dos recursos naturais, aliada à implementação do Plano Diretor de Desenvolvimento Urbano e Ambiental (PDDUA) e de suas diretrizes voltadas à proteção de áreas verdes, tem favorecido a conservação da cobertura vegetal original. A regeneração natural, associada às iniciativas das comunidades e do poder público, evidencia que o MSP constitui um sistema socioambiental complexo, no qual a conservação ambiental e o desenvolvimento urbano podem coexistir.

A análise de multicolinearidade foi conduzida com o objetivo de avaliar a independência das variáveis explicativas na modelagem da propensão à antropização no MSP, empregando os valores de TOL e VIF. Os resultados obtidos são apresentados na Tabela 2.

Tabela 2. Resultados de colinearidade entre os fatores condicionantes.

Table 2. Results of collinearity between the conditioning factors.

Fator	TOL	VIF
Altitude	0,204	4,891
Declividade	0,554	1,806
Proximidade das áreas desmatadas	0,796	1,257
Proximidade das favelas e comunidades urbanas de 2010	0,652	1,534
Geologia	0,400	2,503
Geomorfologia	0,288	3,479
Unidades <i>geomorphons</i>	0,827	1,209
HAND	0,330	3,033
Número de pessoas por setor	0,793	1,261
Proximidade dos Rios	0,569	1,758
Proximidade das Rodovias	0,597	1,675
Perfis de Solos	0,377	2,654

A análise de multicolinearidade revelou que as variáveis do modelo apresentaram um baixo grau de correlação entre si. Os valores de TOL variaram de 0,204 a 0,827, enquanto os valores de VIF oscilaram entre 1,209 e 4,891. Embora a variável Altitude tenha apresentado o maior valor de VIF, este permaneceu abaixo do limiar de 5, comumente adotado na literatura como indicativo de multicolinearidade severa. Conforme Whang *et al.* (2020), valores de VIF inferiores a esse limite não requerem a exclusão da variável do modelo.

A variável unidades *geomorphons*, por sua vez, apresentou o menor valor de VIF (1,209) e o maior valor de TOL (0,827) entre todas as variáveis do modelo. Esses resultados indicam que essa variável não apresenta alta correlação com as demais variáveis independentes, o que sugere que sua influência sobre a variável dependente pode ser estimada com maior precisão. A Declividade e HAND apresentaram valores de VIF de 1,806 e 3,033, respectivamente, indicando uma correlação moderada com as demais variáveis do modelo. Embora esses valores sejam superiores ao VIF da variável unidades *geomorphons*, permanecem abaixo do limiar de 5.

A relação entre Declividade e HAND não compromete significativamente a interpretação dos coeficientes de regressão, mas indica que a influência isolada de cada uma dessas variáveis sobre a variável dependente pode ser ligeiramente superestimada. A análise do IGR, realizada anteriormente, revelou que a variável Proximidade das Rodovias foi o fator condicionante com a maior contribuição para explicar a suscetibilidade a posteriori, apresentando um valor de IGR de 0,103. A construção de rodovias frequentemente envolve a remoção da vegetação nativa e a movimentação de grandes volumes de terra, expondo o solo à ação dos agentes erosivos, como a água da chuva e o vento. Posteriormente, o fator de Número de Pessoas por Setor apresentou um IGR de 0,089, destacando-se como o segundo fator de maior relevância. Outros fatores, como Declividade, Geomorfologia e HAND, apresentaram valores de IGR próximos, variando entre 0,082 e 0,087, indicando certo grau de influência na descrição das áreas mais ou menos propensas à antropização (Tab. 3).

Tabela 3. Relevância dos fatores condicionantes.

Table 3. Relevance of the conditioning factors.

Fatores Condicionantes	IGR
Proximidade das Rodovias	0,103
Número de pessoas por setor	0,089
Declividade	0,087
Geomorfologia	0,086
Altitude	0,083
HAND	0,082
Proximidade das favelas e comunidades urbanas de 2010	0,081
Proximidade dos Rios	0,079
Unidades <i>geomorphons</i>	0,073
Proximidade das áreas desmatadas	0,069
Perfis de solo	0,063
Geologia	0,06

Contudo, os fatores "Proximidade das Áreas Desmatadas", "Proximidade das Favelas e Comunidades Urbanas de 2010" e "Proximidade dos Rios" apresentaram valores de IGR ligeiramente inferiores, porém ainda significativos, variando de 0,069 a 0,081. Os fatores "Geologia" e "Solo" também demonstraram valores relativamente baixos de IGR, com 0,060 e 0,063, respectivamente, sugerindo uma influência limitada na modelagem. Por outro lado, o menor valor de IGR foi associado ao fator "Geologia" (0,060), indicando menor relevância na descrição da propensão ao processo de antropização, embora permaneça acima de qualquer limiar de exclusão sugerido por

Whang *et al.* (2020). Dessa forma, todos os fatores foram mantidos no processo de modelagem.

A análise de Wilcoxon revelou diferenças estatisticamente significativas entre áreas antropizadas e não antropizadas para diversos fatores. Esses resultados corroboram a hipótese de que a antropização exerce uma influência substancial sobre o ambiente, ressaltando a importância da consideração desses fatores na determinação da propensão à antropização (Tab. 4).

Dessa forma, os fatores "Altitude", "Declividade", "Geologia", "Geomorfologia", "HAND", "Proximidade das Rodovias" e "Solo" destacaram-se por apresentarem valores de P ($< 2.2e-16$), indicando que as diferenças entre as condições de propensão à antropização e a preservação das feições naturais são altamente significativas. Assim, esses fatores desempenharam um papel fundamental na identificação das áreas mais vulneráveis à ação antrópica.

A expansão da rede viária é um dos principais motores da ocupação e transformação do espaço, acelerando processos de urbanização e fragmentação ambiental (Tehrany *et al.*, 2017). Além disso, a crescente acessibilidade proporcionada pelas estradas intensifica a ocupação do solo, promovendo a conversão de áreas naturais em espaços construídos e a fragmentação de habitats, o que compromete a sustentabilidade ambiental a longo prazo (Shuster *et al.*, 2005). Nesse sentido, as rodovias formam uma rede que facilita a expansão urbana, agrícola e industrial, gerando uma pressão crescente sobre o uso da terra e resultando em desmatamento, conversão de áreas naturais e fragmentação de habitats, com consequências significativas para a biodiversidade (Shuster *et al.*, 2005).

Adicionalmente, os fatores "Proximidade das Áreas Desmatadas", "Proximidade das Favelas e Comunidades Urbanas de 2010", "Unidades *Geomorphons*" e "Número de Pessoas por Setor" também apresentaram diferenças estatisticamente significativas, com valores de P inferiores a 0,05. Esses fatores refletem a influência antrópica, evidenciando que a proximidade com áreas previamente modificadas pelo homem, como regiões desmatadas e assentamentos irregulares, desempenha um papel relevante na fragilização do MSP.

Tabela 4. Avaliação da significância pelo teste de Wilcoxon.

Table 4. Evaluation of significance using the Wilcoxon test

Fatores Condicionantes	P-VALOR - Wilcoxon
Altitude	$< 2.2e-16$
Declividade	$< 2.2e-16$
Proximidade das áreas desmatadas	7,96E-05
Proximidade das favelas e comunidades urbanas de 2010	7.49e-12
Geologia	$< 2.2e-16$
Geomorfologia	$< 2.2e-16$
Unidades <i>geomorphons</i>	4,25E-11
HAND	$< 2.2e-16$
Número de pessoas por setor	2,48E-07
Proximidade dos Rios	0.02597
Proximidade das Rodovias	$< 2.2e-16$
Solo	$< 2.2e-16$

A "Proximidade dos Rios", embora estatisticamente significativa ($P = 0,02597$), teve sua influência atenuada pela heterogeneidade espacial das características do MSP. Essa heterogeneidade, conforme evidenciado por Kang *et al.* (2019), exerce uma influência complexa sobre a distribuição das áreas de várzea, o que limita o poder explicativo desse fator na modelagem da propensão à antropização.

4.1 Importância dos Fatores Condicionantes

Os fatores condicionantes na previsão dos cenários foram avaliados considerando a interação entre suas características espaciais e os algoritmos RRF, PRF e XGBoost. A análise das amostras de treinamento possibilitou a quantificação da variação do impacto de cada fator, destacando a importância de considerar as particularidades de cada algoritmo na interpretação dos resultados. A "Proximidade de Rodovias" destacou-se como o fator mais determinante em todos os modelos, apresentando o valor máximo de importância em cada um deles. Esse resultado consistente sugere que áreas situadas próximas a essas infraestruturas possuem alta propensão a alterações em sua cobertura original, possivelmente em função das perturbações no solo e das mudanças no acesso à região, o que facilita a ocupação humana e a intensificação das atividades antrópicas (Tab. 5).

Tabela 5. Impacto dos fatores nos modelos de mapeamento da propensão à antropização.

Table 5: Impact of factors on anthropization propensity mapping models.

Fator	RRF	PRF	XGBoost	Média	Mediana	Mínima	Máxima
Altitude	0,384	0,378	0,102	0,288	0,378	0,102	0,384
Declividade	0,434	0,333	0,113	0,293	0,333	0,113	0,434
Proximidade das áreas desmatadas	0,515	0,444	0,113	0,357	0,444	0,113	0,515
Proximidade das favelas e comunidades urbanas de 2010	0,535	0,489	0,128	0,384	0,489	0,128	0,535
Geologia	0,121	0,089	0,010	0,073	0,089	0,010	0,121
Geomorfologia	0,172	0,156	0,021	0,116	0,156	0,021	0,172
Unidades <i>geomorphons</i>	0,010	0,030	0,007	0,016	0,010	0,007	0,030
HAND	0,242	0,244	0,094	0,193	0,242	0,094	0,244
Número de pessoas por setor	0,333	0,333	0,063	0,243	0,333	0,063	0,333
Proximidade dos rios	0,414	0,333	0,114	0,287	0,333	0,114	0,414
Proximidade das rodovias	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Solo	0,152	0,111	0,028	0,097	0,111	0,028	0,152

Outro fator de destaque foi a "Proximidade das Favelas e Comunidades Urbanas de 2010", que apresentou alta importância nos modelos RRF (0,535) e PRF (0,489), enquanto no modelo XGBoost mostrou-se menos relevante (0,128). A "Proximidade das Áreas Desmatadas" seguiu um padrão semelhante, com elevados valores de importância nos modelos RRF (0,515) e PRF (0,444), porém com impacto reduzido no XGBoost (0,113). Além disso, a "Declividade" revelou-se um fator significativo, especialmente no modelo RRF (0,434), seguido pelo PRF (0,333) e pelo XGBoost (0,113).

Esse comportamento evidencia a relevância da inclinação do terreno na propensão à degradação das feições naturais, uma vez que áreas com maior declividade, quando submetidas a ações antrópicas, tendem a favorecer o escoamento superficial, intensificando a ocorrência de processos erosivos. Esses processos, por sua vez, podem desencadear movimentações gravitacionais de massa, ampliando os impactos ambientais nessas regiões. A "Altitude", outro fator topográfico, também apresentou importância significativa nos três modelos, sendo mais relevante no RRF (0,384) e no PRF (0,378) do que no XGBoost (0,102). Da mesma forma, a importância da "Geologia" e da "Geomorfologia" variou entre os modelos, destacando-se no RRF (0,121) e no PRF (0,089), enquanto no XGBoost seu impacto foi consideravelmente menor (0,010).

No entanto, os valores de importância foram consistentemente baixos em todos os casos, indicando que esses fatores, embora possam contribuir marginalmente para a previsão da degradação, não se configuram como os principais determinantes desse processo. A "Declividade" e a "Densidade Populacional" demonstraram uma relação indireta com a expansão antrópica, especialmente nos modelos RRF e PRF. Os fatores "HAND" e "Número de Pessoas por Setor" apresentaram moderada relevância, particularmente no RRF (0,242) e no PRF (0,333), enquanto no XGBoost suas importâncias foram reduzidas (0,094 para HAND e 0,063 para Número de Pessoas por Setor). Esses fatores sugerem uma relação indireta com a maior propensão à expansão das ações antrópicas no MSP, possivelmente refletindo condições específicas de uso e ocupação do solo em áreas com diferentes densidades populacionais e padrões de drenagem.

Por outro lado, as "Unidades *Geomorphons*" e "Solo" apresentaram as menores importâncias em todos os modelos. A variável "Solo" demonstrou pouca relevância, com importâncias variando de 0,028 a 0,152, sugerindo que, embora o tipo de solo seja um fator relevante, seu impacto é inferior em comparação a outros condicionantes, como topografia e proximidade de rodovias. Em síntese, a modelagem por meio dos diferentes algoritmos evidencia que, embora todos os fatores condicionantes contribuam para a previsão da propensão à antropização, aqueles relacionados à proximidade de rodovias, declividade e interferências humanas, como urbanização e desmatamento, foram consistentemente os mais relevantes. A variação entre os modelos também ressalta a importância de considerar diferentes abordagens na seleção do algoritmo de ML mais adequado, a fim de capturar nuances específicas da área de estudo e garantir previsões mais precisas.

Na análise comparativa dos modelos PRF, RRF e XGBoost, foram utilizados diversos indicadores de desempenho, derivados das matrizes de contingência, para avaliar a eficácia de cada modelo na previsão da propensão à antropização de áreas naturais no MSP. O PRF e o RRF apresentaram desempenhos muito próximos em várias métricas, com o PRF ligeiramente superior em termos de acurácia (0,827 contra 0,826) e o RRF com uma pequena vantagem em sensibilidade (0,803 contra 0,802) (Tab. 6). Ambos os modelos demonstraram alta eficácia geral na previsão da propensão à antropização das áreas naturais, evidenciando sua robustez na modelagem dos dados. Em contraste, o XGBoost apresentou uma acurácia inferior (0,807), refletindo uma leve diminuição em sua capacidade de prever corretamente os casos positivos e negativos em comparação aos outros dois modelos.

Tabela 6. Validação dos modelos usando o conjunto de 30% das amostras.

Table 6. Validation of the models using the 30% sample set.

Métrica	PRF	RRF	XGBoost
Acurácia	0,827	0,826	0,807
Sensibilidade	0,802	0,803	0,784
Especificidade	0,852	0,850	0,831
Predição de Valor Positivo	0,844	0,842	0,822
Predição de Valor Negativo	0,811	0,812	0,794
Taxa de classificação incorreta - <i>Misclassification Rate</i> (MR)	0,172	0,173	0,192
Pontuação F1	0,8314	0,830	0,812
Área sob a curva - <i>Area under the curve</i> (AUC)	0,910	0,904	0,881

A sensibilidade do PRF (0,802) e do RRF (0,803) foi levemente superior à do XGBoost (0,784), o que sugere que os dois primeiros modelos são mais eficazes na identificação de eventos verdadeiros positivos, ou seja, na detecção correta de áreas propensas à antropização. Por outro lado, o XGBoost apresentou uma especificidade menor (0,831) em comparação com o PRF (0,852) e o RRF (0,850), indicando uma maior probabilidade de produzir falsos positivos.

Em termos de valores preditivos positivos (*Pos Pred Value*), os três modelos apresentaram desempenho semelhante, com o PRF alcançando 0,844, o RRF com 0,842 e o XGBoost com um valor ligeiramente inferior a 0,822. Esses resultados indicam que todos os modelos conseguem, em grande parte, prever corretamente as áreas propensas à antropização. Da mesma forma, os valores preditivos negativos (*Neg Pred Value*) também se mostraram próximos: 0,811 para o PRF, 0,812 para o RRF e 0,794 para o XGBoost, sugerindo que os modelos têm uma alta capacidade de prever corretamente as áreas não propensas à antropização das áreas naturais.

Uma métrica importante a ser destacada é a taxa de erro de classificação (MR), que mede a proporção de previsões incorretas realizadas pelo modelo. Nesse contexto, o PRF apresentou a menor taxa de erro (0,172), seguido pelo RRF (0,173), enquanto o XGBoost exibiu uma taxa de erro mais elevada (0,192). Esses resultados reforçam a ideia de que, embora o XGBoost seja uma ferramenta poderosa, os modelos PRF e RRF demonstraram maior precisão na modelagem nesta aplicação específica.

O F1 Score, que combina a precisão e a sensibilidade, novamente favoreceu os modelos PRF (0,8314) e RRF (0,830), indicando que esses modelos equilibram bem a taxa de verdadeiros positivos e falsos negativos. O XGBoost, com um F1 Score de 0,812, apresentou um desempenho inferior, sugerindo que ele pode ser menos eficaz em cenários onde é necessário maximizar tanto a precisão quanto a recuperação simultaneamente. Finalmente, a AUC é uma métrica crítica para avaliar a capacidade discriminativa dos modelos (Fig. 6). O PRF apresentou o melhor valor de AUC (0,910), indicando que possui a maior capacidade de distinguir entre áreas suscetíveis e não suscetíveis. O RRF apresentou um AUC próximo, de 0,904, demonstrando também alta eficácia. Em contraste, o XGBoost teve um desempenho inferior (0,881), o que reforça a ideia de uma menor eficiência em capturar todas as nuances dos dados (Tab. 6).

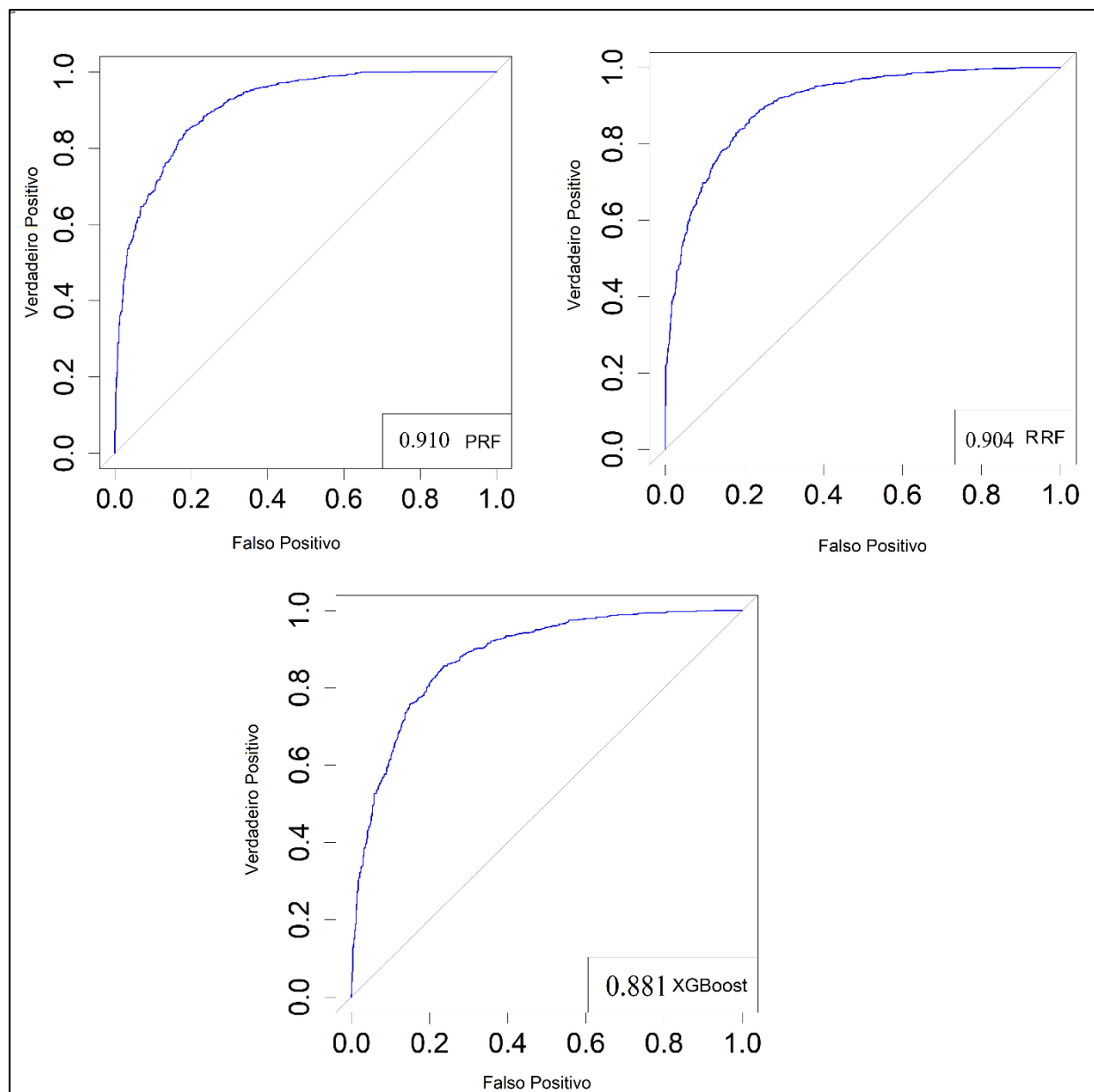


Figura 6. AUC-ROC da métrica de validação dos modelos.

Figure 6. AUC-ROC of the model validation metric.

Assim, é importante ressaltar que os modelos PRF e RRF se destacaram pela alta acurácia, sensibilidade e AUC, tornando-se opções mais robustas para a modelagem da propensão à antropização. O XGBoost mostrou-se inferior, especialmente em termos de acurácia e capacidade de discriminação. Essas diferenças de desempenho podem estar relacionadas à maneira como cada modelo lida com os dados e suas estruturas específicas. No entanto, todos os modelos analisados oferecem *insights* valiosos para a previsão de áreas propensas à antropização das áreas naturais.

Com base nos resultados obtidos utilizando os modelos PRF, RRF e XGBoost para a estimativa da propensão à antropização das feições naturais, entendida aqui como as áreas mais suscetíveis a sofrer intervenções humanas, independentemente do interesse ou finalidade atribuída, as áreas foram classificadas em seis categorias de propensão: muito baixa, baixa, intermediária a baixa, intermediária a alta, alta e muito alta (Tab. 7). Áreas com alta declividade e baixa acessibilidade tendem a apresentar menor

vulnerabilidade, enquanto regiões planas e próximas a centros urbanos são mais suscetíveis à ocupação humana e às atividades antrópicas.

Tabela 7. Volume de área da propensão à antropização.

Table 7. Volume of area prone to anthropization.

Categoria	PRF (km²)	RRF (km²)	XGBoost (km²)	PRF (%)	RRF (%)	XGBoost (%)
Muito baixa propensão	17,48	17,93	11,76	24,09	24,68	16,19
Baixa propensão	11,43	11,46	12,77	15,76	15,78	17,58
Intermediária a baixa propensão	11,75	11,14	11,92	16,20	15,34	16,41
Intermediária a alta propensão	12,02	11,60	12,23	16,57	15,96	16,83
Alta propensão	10,50	11,07	11,79	14,47	15,24	16,23
Muito alta propensão	9,38	9,45	12,18	12,92	13,00	16,76

Nos modelos PRF e RRF, a classe de 'propensão muito baixa' foi predominante, abrangendo a maior área em ambos os casos, com uma extensão de cerca de 17,5 km². Esses resultados indicam que uma parcela significativa do MSP apresenta um risco mínimo de antropização, sugerindo a existência de extensas áreas com características naturais preservadas, como terrenos estáveis e regiões de baixa acessibilidade (Fig. 7).

A classe "baixa propensão" apresentou valores semelhantes nos três modelos, com 11,43 km² no PRF, 11,46 km² no RRF e 12,77 km² no XGBoost. Essa similaridade indica uma alta consistência entre os modelos na identificação de áreas com risco moderadamente baixo de antropização, corroborando a tendência de menor vulnerabilidade em determinadas porções do morro. A topografia acidentada, a presença de vegetação densa e a distância em relação a centros urbanos podem ser alguns dos fatores que contribuem para essa menor suscetibilidade.

Na classe denominada "intermediária a baixa propensão", os resultados evidenciam pequenas variações entre os modelos. O PRF abrange uma área de 11,75 km², enquanto o RRF cobre 11,14 km² e o modelo XGBoost cobre 11,92 km². Essas diferenças sutis indicam que os modelos demonstram consistência na identificação de áreas com risco intermediário de antropização, situando-se entre as classes de baixa e alta propensão. A pequena variação na extensão da classe "baixa propensão" entre os modelos indica uma alta concordância na identificação de áreas com risco intermediário de antropização, que se situam entre as classes de baixa e alta propensão. Essa consistência observada entre os modelos reforça a confiabilidade dos resultados e sugere a existência de um consenso sobre a localização dessas áreas. Fatores como a topografia suave, a presença de vegetação e a distância em relação a centros urbanos podem ser considerados elementos que contribuem para essa classificação intermediária.

A classe "intermediária a alta propensão" apresenta resultados comparáveis entre os modelos analisados. O PRF abrange uma área de 12,02 km², enquanto o modelo de RRF cobre 11,60 km² e o modelo XGBoost abrange 12,23 km². Essa faixa intermediária a alta indica a presença de áreas sujeitas a um risco significativo de antropização, localizadas predominantemente em encostas menos íngremes e em zonas de transição dentro do morro.

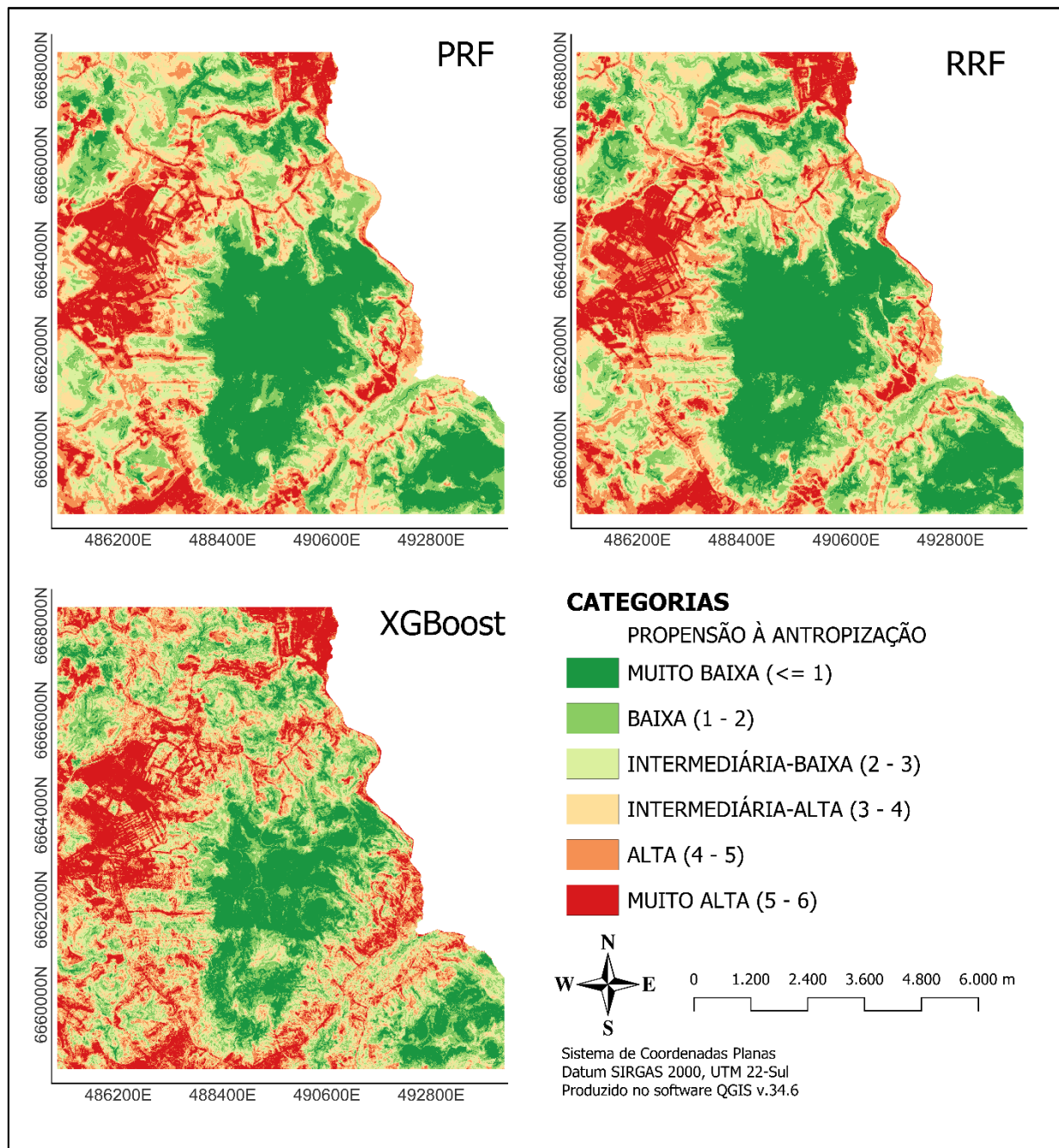


Figura 7. Mapa de propensão à antropização pelos modelos PRF, RRF e XGBoost.

Figure 7. Map of propensity to anthropization using the PRF, RRF, and XGBoost models.

Nas classes de "alta propensão" e "muito alta propensão", os modelos evidenciam uma crescente propensão à antropização. No PRF, as áreas classificadas como alta e muito alta propensão ocupam 10,50 km² e 9,38 km², respectivamente. O RRF apresenta 11,07 km² para alta propensão e 9,45 km² para muito alta propensão, enquanto o modelo XGBoost revela 11,79 km² para alta propensão e 12,18 km² para muito alta propensão. Os resultados indicam que o risco de antropização é mais elevado em áreas específicas do morro, concentrando-se principalmente na região Oeste e nas várzeas.

Com base nos resultados apresentados, o PRF se destacou como o mais eficiente, apresentando uma acurácia de 0,827, sensibilidade de 0,802, especificidade de 0,852 e F1 Score de 0,8314. Além disso, o PRF obteve a maior AUC, alcançando 0,910,

evidenciando-se como o modelo mais robusto na distinção entre as classes. Esses resultados são consistentes com a literatura, na qual modelos baseados em *ensemble learning* com paralelização, como o PRF, são reconhecidos por sua capacidade de lidar com grandes volumes de dados e alta dimensionalidade, além de proporcionarem uma melhor generalização em comparação com outros métodos (Band *et al.*, 2020).

O RRF apresentou desempenho muito semelhante ao do PRF, com uma acurácia de 0,826, sensibilidade de 0,803 e especificidade de 0,850. A principal diferença entre os dois modelos reside na aplicação da regularização no RRF, que tem como objetivo controlar o sobreajuste, especialmente em cenários com um grande número de variáveis correlacionadas. De acordo com Deng *et al.* (2012), o RRF é eficaz na abordagem de problemas de multicolinearidade, uma característica frequentemente observada em dados ambientais. Ademais, o RRF se beneficiou da regularização para aprimorar a estabilidade das predições, como evidenciado por seu F1 Score de 0,830 e AUC de 0,904, ambos próximos aos valores apresentados pelo PRF.

O modelo XGBoost, embora amplamente reconhecido por seu desempenho superior em diversas competições de aprendizado de máquina e em aplicações práticas, obteve resultados inferiores neste estudo. Com uma acurácia de 0,807, sensibilidade de 0,784 e AUC de 0,881, o XGBoost se revelou o menos eficiente entre os três modelos analisados. Essa performance inferior pode ser atribuída ao fato de que o XGBoost utiliza o método de *boosting* para ajustar sucessivamente os erros cometidos por modelos anteriores. Embora esse ajuste possa resultar em uma maior precisão em certos cenários, pode ser menos eficaz quando o modelo se torna sensível aos hiperparâmetros ou quando há grande variabilidade nas características dos dados.

Linh *et al.* (2022) destacam que, embora o XGBoost seja um modelo poderoso, sua performance pode ser negativamente impactada se não for ajustado adequadamente, especialmente em contextos nos quais o *overfitting* é uma preocupação. Além disso, o XGBoost demanda uma configuração mais cuidadosa dos parâmetros, como a taxa de aprendizado e a profundidade máxima das árvores, para alcançar um desempenho ideal. Essa necessidade de ajuste minucioso pode ter contribuído para os resultados ligeiramente inferiores observados neste estudo.

Do ponto de vista teórico, os métodos de *ensemble learning* baseados em *bagging*, como o PRF e o RRF, tendem a ser mais estáveis e a generalizar melhor em dados ruidosos ou não balanceados, o que é uma característica importante em problemas ambientais, como a modelagem da propensão à antropização (Janizadeh *et al.*, 2022). Por outro lado, embora o XGBoost seja eficaz em muitos cenários, ele pode ser mais sensível a características específicas dos dados e à escolha de hiperparâmetros, o que pode explicar seus resultados mais modestos.

Assim, é importante ressaltar que o PRF se destacou como o mais eficiente, apresentando a melhor capacidade preditiva em termos de AUC, acurácia e F1 Score. O RRF o seguiu de perto, demonstrando também um excelente desempenho, especialmente em relação à regularização e ao controle do *overfitting*. O modelo XGBoost, embora reconhecido como uma ferramenta potente, mostrou-se menos adequado para esta aplicação específica, provavelmente devido à sua sensibilidade em relação aos ajustes de parâmetros.

5. Conclusão

Este estudo utilizou três algoritmos de aprendizado de máquina – PRF, RRF e XGBoost – para avaliar a propensão à antropização no MSP, com o objetivo de identificar o modelo mais eficaz para a delimitação de áreas com influência antrópica. Os resultados demonstraram que a aplicação de técnicas de aprendizado de máquina em questões ambientais, como a degradação do meio ambiente nativo é essencial. Esses métodos otimizaram o tempo envolvido na modelagem, alcançando altos níveis de precisão na previsão das áreas mais propensas à antropização.

Através da validação com um conjunto representativo de 30% das amostras, foi possível identificar que o modelo PRF se destacou como o mais eficaz na previsão das áreas propensas à antropização. Essa descoberta contribui para o avanço da pesquisa na modelagem preditiva de impactos ambientais e fornece uma base sólida para a implementação de estratégias de conservação e manejo sustentável no MSP.

A principal contribuição deste estudo foi a aplicação e o desenvolvimento de abordagens híbridas paralelas e de regularização para estimar a propensão à antropização no MSP. Ademais, a importância dos fatores condicionantes na previsão da conversão de áreas naturais em áreas antropizadas variou conforme o algoritmo empregado. A análise das amostras de treinamento, após a imputação e o processamento pelos modelos RRF, PRF e XGBoost, revelou uma sinergia entre determinados fatores, como “Declividade” e “Distância de rodovias”, especialmente quando o algoritmo RRF foi empregado. Por outro lado, o XGBoost atribuiu maior peso à variável “Distância dos centros urbanos”. Esses achados sugerem que a escolha do algoritmo de aprendizado de máquina pode impactar significativamente a interpretação dos resultados e, conseqüentemente, a formulação de estratégias de conservação.

A pesquisa revelou que a “Proximidade de rodovias” foi o fator mais determinante em todos os modelos analisados. Esse achado indica que a construção de rodovias não apenas fragmenta habitats, mas também facilita o acesso a atividades agrícolas e extrativistas, além de intensificar a pressão urbanizadora, aumentando a vulnerabilidade dessas áreas à conversão. Embora variáveis como topografia e tipo de solo também tenham contribuído para explicar a distribuição espacial da propensão à alteração, a influência da proximidade de rodovias mostrou-se significativamente superior. A geologia e a geomorfologia tiveram uma contribuição marginal na previsão da degradação antrópica, apresentando valores de importância até uma ordem de magnitude inferiores em comparação à proximidade de rodovias.

A “Declividade” e a “Densidade populacional” tiveram uma importância moderada na previsão da expansão antrópica, especialmente nos modelos RRF e PRF. Enquanto áreas com maior declividade tendem a ser menos atrativas para a urbanização devido a restrições topográficas, regiões com alta densidade populacional podem apresentar uma demanda crescente por expansão urbana. Contudo, a menor relevância desses fatores no modelo XGBoost sugere que variáveis como a proximidade de rodovias e a disponibilidade de infraestrutura podem exercer um papel mais determinante, especialmente em áreas com menor pressão de desenvolvimento. Esses resultados reforçam que a expansão antrópica é um processo dinâmico e multifatorial, caracterizado por interações não lineares entre diferentes variáveis ambientais e socioeconômicas.

Nesse contexto, as características intrínsecas do ambiente físico desempenham um papel secundário em relação aos processos induzidos pela ocupação humana. No entanto, vale destacar que a resolução dos dados utilizados neste estudo pode ter limitado a detecção de influências geológicas em pequenas escalas. Dessa forma,

pesquisas futuras poderiam se beneficiar do uso de dados com maior resolução espacial, o que possibilitaria uma avaliação mais precisa da importância relativa desses fatores e potencialmente revelaria padrões não detectados neste estudo.

A matriz de contingência possibilitou uma avaliação comparativa rigorosa entre os modelos PRF, RRF e XGBoost, quantificando seus desempenhos na previsão da propensão à antropização. Os resultados obtidos contribuem para o avanço do conhecimento científico na modelagem ambiental e oferecem subsídios para a seleção de abordagens mais adequadas em estudos futuros, auxiliando na tomada de decisões para o planejamento territorial e a gestão ambiental.

Dessa forma, destaca-se que o PRF apresentou o melhor desempenho, alcançando os maiores valores de AUC, acurácia e F1 Score, evidenciando sua superioridade preditiva. O RRF também obteve um desempenho expressivo, sobressaindo-se especialmente pela capacidade de regularização e controle do *overfitting*. Já o XGBoost, mesmo sendo uma ferramenta robusta, mostrou-se menos adequado para esta aplicação específica, possivelmente devido à sua maior sensibilidade aos ajustes de hiperparâmetros.

O mapeamento dos processos antrópicos é fundamental, pois revela não apenas a extensão da intervenção humana na paisagem, mas também a dinâmica complexa que impulsiona a vulnerabilidade ambiental, oferecendo, assim, um roteiro indispensável para o planejamento sustentável e a formulação de estratégias de conservação eficazes.

Agradecimentos: Agradecemos à Universidade Federal do Rio Grande do Sul (UFRGS) e à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo apoio financeiro concedido, que possibilitou a concretização deste estudo.

Contribuições dos Autores: Concepção, M.W.D.F. e C.W.M.J.; metodologia, A.G.F., F.H.F.A. e M.W.D.F.; software, F.H.F.A. e C.R.C.; validação, A.G.F., F.H.F.A. e C.R.C.; análise formal, C.W.M.J.; pesquisa, A.G.F., F.H.F.A., C.R.C. e M.W.D.F.; recursos, M.W.D.F.; preparação de dados, F.H.F.A., C.R.C. e A.G.F.; escrita do artigo, A.G.F. e F.H.F.A.; revisão, C.W.M.J.; supervisão, M.W.D.F. e C.W.M.J. Todos os autores leram e concordaram com a versão publicada do manuscrito.

Conflito de Interesse: Os autores declaram não haver conflito de interesse.

Referências

- Bag, R., Mondal, I., Dehbozorgi, M., Bank, S.P., Das, D.N., Bandyopadhyay, J., Pham, Q.B., Al-Quraishi, A.M.F. & Nguyen, X.C. 2022. Modelling and mapping of soil erosion susceptibility using machine learning in a tropical hot sub-humid environment. *Journal of Cleaner Production*, 364: 132428. <https://doi.org/10.1016/j.jclepro.2022.132428>
- Band, S.S., Janizadeh, S., Pal, S.C., Saha, A., Chakraborty, R., Melesse, A.M. & Mosavi, A. 2020. Flash Flood Susceptibility Modeling Using New Approaches of Hybrid and Ensemble Tree-Based Machine Learning Algorithms. *Remote Sensing*, 12(21): 3568. <https://doi.org/10.3390/rs12213568>
- Bertrand, G. 1984. Les géographes français et leurs paysages. *Annales de Géographie*, 93(516): 218-229.
- Budholiya, K., Shrivastava, S.K. & Sharma, V. 2022. An optimized XGBoost based diagnostic system for effective prediction of heart disease. *Journal of King Saud University - Computer and Information Sciences*, 34(7): 4514-4523. <https://doi.org/10.1016/j.jksuci.2020.10.013>
- Cao, J., Zhang, Z., Du, J., Zhang, L., Song, Y. & Sun, G. 2020. Multi-geohazards susceptibility mapping based on machine learning - a case study in Jiuzhaigou, China. *Natural Hazards*, 102(3): 851-871. <https://doi.org/10.1007/s11069-020-03927-8>
- Chen, T. & Guestrin, C. 2016. XGBoost: A Scalable Tree Boosting System. In: THE 22ND ACM SIGKDD INTERNATIONAL CONFERENCE, 2016. *Anais...* New York, ACM, p. 785-794.

- Christofidis, D., Assumpção, R.S.F.V. & Kligerman, D.C. 2019. A evolução histórica da drenagem urbana: da drenagem tradicional à sintonia com a natureza. *Saúde Debate*, 43(3): 94-108. <https://doi.org/10.1590/0103-11042019S307>
- Christofoletti, A. 1999. *Modelagem de sistemas ambientais*. São Paulo: Edgard Blucher, 256p.
- Coelho, M.C.N. 2001. Impactos ambientais em áreas urbanas – teorias, conceitos e métodos de pesquisa. In: Guerra, A.J.T.; Cunha, S.B. (org.). *Impactos ambientais urbanos no Brasil*. Rio de Janeiro: Bertrand Brasil, p. 19-45.
- Darabi, H., Rahmati, O., Naghibi, S.A., Mohammadi, F., Ahmadisharaf, E., Kalantari, Z. & Bui, D.T. 2021. Development of a novel hybrid multi-boosting neural network model for spatial prediction of urban flood. *Geocarto International*, 37(19): 5716-5741. <https://doi.org/10.1080/10106049.2021.1920629>
- Deng, H. & Runger, G. 2012. Feature selection via regularized trees. In: INTERNATIONAL JOINT CONFERENCE ON NEURAL NETWORKS, 2012, Brisbane, Austrália. *Anais...* New York, IEEE.
- Deng, H. & Runger, G. 2013. Gene selection with guided regularized random forest. *Pattern Recognition*, 46(12): 3483-3489. <https://doi.org/10.1016/j.patcog.2013.05.018>
- Dormann, C.F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J.R.G., Gruber, B., Lafourcade, B., Leitão, P.J., Münkemüller, T., McClean, C., Osborne, P.E., Reineking, B., Schröder, B., Skidmore, A.K., Zurell, D. & Lautenbach, S. 2013. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, 36(1): 27-46. <https://doi.org/10.1111/j.1600-0587.2012.07348.x>
- Eilola, S., Jaalama, K., Kangassalo, P., Nummi, P., Staffans, A. & Fagerholm, N. 2023. 3D Visualisations for Communicative Urban and Landscape Planning: What Systematic Mapping of Academic Literature Can Tell Us of Their Potential? *Landscape Urban Planning*, 234: 104716. <https://doi.org/10.1016/j.landurbplan.2023.104716>
- Elbeltagi, A., Pande, C.B., Kouadri, S. & Islam, A.R.M.T. 2022. Applications of various data-driven models for the prediction of groundwater quality index in the Akot basin, Maharashtra, India. *Environmental Science and Pollution Research*, 29: 17591–17605. <https://doi.org/10.1007/s11356-021-17064-7>
- Fenta, A.A., Tsunekawa, A., Haregeweyn, N., Poesen, J., Tsubo, M., Borrelli, P., Panagos, P., Vanmaercke, M., Broeckx, J., Yasuda, H., Kawai, T. & Kurosaki, Y. 2020. Land susceptibility to water and wind erosion risks in the East Africa region. *Science of the Total Environment*, 703: 135016. <https://doi.org/10.1016/j.scitotenv.2019.135016>
- García-Ayllón, S. 2017. Diagnosis of Complex Coastal Ecological Systems: Environmental GIS Analysis of a Highly Stressed Mediterranean Lagoon through Spatiotemporal Indicators. *Ecological Indicators*, 83: 451-462. <https://doi.org/10.1016/j.ecolind.2017.08.015>
- García-Ayllón, S. 2018. Predictive Diagnosis of Agricultural Periurban Areas Based on Territorial Indicators: Comparative Landscape Trends of the So-Called “Orchard of Europe”. *Sustainability*, 10(6): 1820. <https://doi.org/10.3390/su10061820>
- Goerl, R.F., Michel, G.P. & Kobiyama, M. 2017. Mapeamento de áreas suscetíveis à inundação com o modelo HAND e análise do seu desempenho em diferentes resoluções espaciais. *Revista Brasileira de Cartografia*, 69(1): 61-69. <https://doi.org/10.14393/rbcv69n1-44032>
- Hasenack, H. et al. 2008. *Diagnóstico ambiental de porto alegre: geologia, solos, drenagem, vegetação/ocupação e paisagem*. Porto Alegre, Secretaria Municipal do Meio Ambiente.
- Hosseiny, H., Nazari, F., Smith, V. & Nataraj, C. 2020. A framework for modeling flood depth using a hybrid of hydraulics and machine learning. *Scientific Reports*, 10: 8222. <https://doi.org/10.1038/s41598-020-65232-5>
- IBGE. Instituto Brasileiro de Geografia e Estatística. 2010. *Censo 2010 mostra as características territoriais dos aglomerados subnormais e suas diferenças das demais áreas das cidades*. Disponível em: <https://agenciadenoticias.ibge.gov.br/agencia-sala-de-imprensa/2013-agencia-de-noticias/releases/14504-asi-censo-2010-mostra-as-caracteristicas-territoriais-dos-aglomerados-subnormais-e-suas-diferencas-das-demas-areas-das-cidades>. Acesso em: 15 set. 2024.
- IBGE. Instituto Brasileiro de Geografia e Estatística. 2022. *Censo 2022*. Disponível em: <https://basedosdados.org/dataset/08a1546e-251f-4546-9fe0-b1e6ab2b203d?table=c5c43530-0c76-4226-b67c-9e0d73b66d12>. Acesso em: 12 set. 2024.

Instituto Nacional de Pesquisas Espaciais (INPE). 2024. *Projeto de Monitoramento do Desmatamento na Amazônia Legal por Satélite (PRODES)*. São José dos Campos, SP: INPE. Disponível em: <http://www.obt.inpe.br/prodes/>. Acesso em: 15 jan. 2025.

Janizadeh, S., Vafakhah, M., Kapelan, Z. & Mobarghaee Dinan, N. 2022. Hybrid XGBoost model with various Bayesian hyperparameter optimization algorithms for flood hazard susceptibility modeling. *Geocarto International*, 37(25): 8273-8292. <https://doi.org/10.1080/10106049.2021.1996641>

Jasiewicz, J. & Stepinski, T. F. 2013. Geomorphons a Pattern Recognition Approach to Classification and Mapping of Landforms. *Geomorphology*, 182: 147-156. <https://doi.org/10.1016/j.geomorph.2012.11.005>

Ji, M. *et al.* 2019. An improved gradient boosting decision tree algorithm. In: CHINESE AUTOMATION CONGRESS (CAC), 2019, Hangzhou. *Anais...* IEEE, p. 446-451.

Kadavi, P.R., Lee, C-W. & Lee, S. 2018. Application of Ensemble-Based Machine Learning Models to Landslide Susceptibility Mapping. *Remote Sensing*, 10(8): 1252. <https://doi.org/10.3390/rs10081252>

Kang, L., Jiang, S., Hu, X. & Li, C. 2019. Evaluation of return period and risk in bivariate non-stationary flood frequency analysis. *Water*, 11(1): 79. <https://doi.org/10.3390/w11010079>

Linh, N.T.T., Pandey, M., Janizadeh, S., Bhunia, G.S., Norouzi, A., Ali, S., Pham, Q.B., Anh, D.T., & Ahmadi, K. 2022. Flood susceptibility modeling based on new hybrid intelligence model: optimization of XGBoost model using GA metaheuristic algorithm. *Advances in Space Research*, 69(9): 3301-3318. <https://doi.org/10.1016/j.asr.2022.02.027>

Mohammed, I., Marshall, M., Bie, K., Estes, L., & Nelson, A. 2020. A blended census and multiscale remote sensing approach to probabilistic cropland mapping in complex landscapes. *ISPRS Journal of Photogrammetry and Remote Sensing*, 161: 233-245. <https://doi.org/10.1016/j.isprsjprs.2020.01.024>

Moroz-Caccia Gouveia, I.C. & Ross, J.L.S. 2019. Fragilidade Ambiental: uma proposta de aplicação de Geomorphons para a variável relevo. *Revista do Departamento de Geografia*, 37: 123-136. <https://doi.org/10.11606/rdg.v37i0.151030>

Motevalli, A., Naghibi, A. S., Hashemi, H., Berndtsson, R., Pradhan, B. & Gholami, V. 2019. Inverse method using boosted regression tree and k-nearest neighbor to quantify effects of point and non-point source nitrate pollution in groundwater. *Journal of Cleaner Production*, 228: 1248-1263. <https://doi.org/10.1016/j.jclepro.2019.04.293>

Moura, N., Hasenack, H. & Silva, L. 2013. *Mapa geomorfológico dos municípios de Porto Alegre, Viamão e Alvorada-RS*. Porto Alegre: UFRGS-IB-Centro de Ecologia, 1 v.

Mousavi, S.M., Golkarian, A., Amir Naghibi, S., Kalantar, B. & Pradhan, B. 2017. GIS-based groundwater spring potential mapping using data mining boosted regression tree and probabilistic frequency ratio models in Iran. *AIMS Geosciences*, 3(1): 91-115. <https://doi.org/10.3934/geosci.2017.1.91>

Nhu, V.H., Mohammadi, A., Shahabi, H., Ahmad, B.B., Al-Ansari, N., Shirzadi, A., Clague, J.J., Jaafari, A., Chen, W. & Nguyen, H. 2020. Landslide Susceptibility Mapping Using Machine Learning Algorithms and Remote Sensing Data in a Tropical Environment. *International Journal of Environmental Research and Public Health*, 17(14):4933. <https://doi.org/10.3390/ijerph17144933>

Nsengiyumva, J.B. & Valentino, R. 2020. Predicting landslide susceptibility and risks using GIS-based machine learning simulations, case of upper Nyabarongo catchment. *Geomatics, Natural Hazards and Risk*, 11(1): 1250-1277. <https://doi.org/10.1080/19475705.2020.1785555>

Palmer, J. F. 2019. The Contribution of a GIS-Based Landscape Assessment Model to a Scientifically Rigorous Approach to Visual Impact Assessment. *Landscape and Urban Planning*, 189: 80-90. <https://doi.org/10.1016/j.landurbplan.2019.03.005>

Phan, A., Ha, D.N., Man, C.D., Nguyen, T.T., Bui, H.Q. & Nguyen, T.T.N. 2019. Rapid assessment of flood inundation and damaged rice area in Red River Delta from Sentinel 1A imagery. *Remote Sensing*, 11(17): 2034. <https://doi.org/10.3390/rs11172034>

Prasad, M., Dubey, C.P., Joshi, K.B., & Tiwari, V.M. 2021. Crustal density and susceptibility structure beneath Achankovil shear zone, India. *Lithosphere*, n. Special (6): 6017801. <https://doi.org/10.2113/2021/6017801>

Rahmati, O., Tahmasebipour, N., Haghizadeh, A., Pourghasemi, H.R. & Feizizadeh, B. 2017. Evaluation of different machine learning models for predicting and mapping the susceptibility of gully erosion. *Geomorphology*, (298): 118-137. <https://doi.org/10.1016/j.geomorph.2017.09.006>

Rambo, B. 2000. *A Fisionomia do Rio Grande do Sul: Ensaio de Monografia*. São Leopoldo, Editora UNISINOS.

Rennó, C., Nobre, A.D., Cuartas, L. A., Soares, J.V., Hodnett, M.G. & Tomasella, J. 2008. HAND, a new terrain descriptor using SRTM-DEM: Mapping terra-firme rainforest environments in Amazonia. *Remote Sensing of Environment*, 112(9): 3469-3481. <https://doi.org/10.1016/j.rse.2008.03.018>

Saha, S., Roy, J., Pradhan, B. & Hembram, T.K. 2021. Hybrid ensemble machine learning approaches for landslide susceptibility mapping using different sampling ratios at East Sikkim Himalayan, India. *Advances in Space Research*, 68(7): 2819-2840. <https://doi.org/10.1016/j.asr.2021.05.018>

Sakaa, B., Elbeltagi, A., Boudibi, S., Chaffai, H., Islam, A.B.M.T., Kulimushi, L.C., Choudhari, P., Hani, A., Brouziyne, Y. & Wong Y.J. 2022. Water quality index modeling using random forest and improved SMO algorithm for support vector machine in Saf-Saf river basin. *Environmental Science and Pollution Research*, 29(32): 48491-48508. <https://doi.org/10.1007/s11356-022-18644-x>

Shahzad, N., Ding, X., & Abbas, S. 2022. A comparative assessment of machine learning models for landslide susceptibility mapping in the rugged terrain of northern Pakistan. *Applied Sciences*, 12(5): 2280. <https://doi.org/10.3390/app12052280>

Shuster, W.D., Bonta, J., Thurston, H., Warnemuende, E., & Smith, D.R. 2005. Impacts of impervious surface on watershed hydrology: A review. *Urban Water Journal*, 2(4): 263-275. <https://doi.org/10.1080/15730620500386529>

Silva, L.F. & Bacani, V.M. 2017. Análise da fragilidade ambiental e das áreas de preservação permanente da bacia hidrográfica do córrego fundo, município de Aquidauana-MS. *Caderno de Geografia*, 27(50): 119-137. <https://doi.org/10.5752/p.2318-2962.2017v27n49p264>

SMAMUS. Secretaria Municipal do Meio Ambiente, Urbanismo e Sustentabilidade da Prefeitura de Porto Alegre. 2022. *Mapas Digitais da SMAMUS*. Disponível em: <https://prefeitura.poa.br/carta-de-servicos/mapasdigitais-da-smamus>. Acesso em: 10 jun. 2025.

Sotchava, V.B. 1977. *O Estudo de Geossistemas*. São Paulo, 16p. Métodos em Questão, 16. Instituto de Geografia, Universidade de São Paulo.

Souza, A.C. & Vieira, R.S. 2018. Mapeamento da vulnerabilidade ambiental por antropização em uma bacia hidrográfica do Semiárido brasileiro. *Revista Brasileira de Geografia Física*, 11(4): 1198-1215.

Tehrany, M.S., Shabani, F. & Jebur, M.N. 2017. GIS-based spatial prediction of flood prone areas using standalone frequency ratio, logistic regression, weight of evidence and their ensemble techniques. *Geomatics, Natural Hazards and Risk*, 8(2): 1538-1561. <https://doi.org/10.1080/19475705.2017.1362038>

Tuv, E., Borisov, A., Runger, G. & Torkkola, K. 2009. Feature selection with ensembles, artificial variables, and redundancy elimination. *Journal of Machine Learning Research*, 10(45): 1341-1366.

Wang, Z., Lai, C., Chen, X., Yang, B., Zhao, S. & Bai, X., 2015. Flood hazard risk assessment model based on random forest. *Journal of Hydrology*, 527: 1130-1141. <https://doi.org/10.1016/j.jhydrol.2015.06.008>

Whang, S.E. & Lee, Jae-Gil. 2020. Data collection and quality challenges for deep learning. *Proceedings of the VLDB Endowment*, 13(12): 3429-3432. <https://doi.org/10.14778/3415478.3415562>

Wilcoxon, F. 1945. Some uses of statistics in plant pathology. *Biometrics Bulletin*, 1(4): 41-45. <https://doi.org/10.2307/3002011>

Xie, W., Li, X., Jian, W., Yang, Y., Liu, H., Robledo, L.F. & Nie, W. 2021. A Novel Hybrid Method for Landslide Susceptibility Mapping-Based GeoDetector and Machine Learning Cluster: A Case of Xiaojin County, China. *ISPRS International Journal of Geo-Information*, 10(2): 93. <https://doi.org/10.3390/ijgi10020093>

Yang, H., Leng, Q., Xiao, Y. & Chen, W. 2022. Investigating the Impact of Urban Landscape Composition and Configuration on PM2.5 Concentration under the LCZ Scheme: A Case Study in Nanchang, China. *Sustainable Cities and Society*, 84: 104006. <https://doi.org/10.1016/j.scs.2022.104006>

Zhang, G., Wang, M. & Liu, K. 2019. Forest Fire Susceptibility Modeling Using a Convolutional Neural Network for Yunnan Province of China. *International Journal of Disaster Risk Science*, 10: 386-403. <https://doi.org/10.1007/s13753-019-00233-1>

Zhang, X., Wang, X., Zhang, C. & Zhai, J. 2022. Development of a Cross-Scale Landscape Infrastructure Network Guided by the New Jiangnan Watertown Urbanism: A Case Study of the Ecological Green Integration Demonstration Zone in the Yangtze River Delta, China. *Ecological Indicators*, 143: 109317. <https://doi.org/10.1016/j.ecolind.2022.109317>

Zhu, L. & Huang, J. 2006. GIS-based logistic regression method for landslide susceptibility mapping in regional scale. *Journal of Zhejiang University-SCIENCE A*, 7(12): 2007-2017. <https://doi.org/10.1631/jzus.2006.A2007>