REVISTA TÓPICOS
ISSN 2965-6672 | Qualis A2

MODELAGEM PREDITIVA DE RISCO DE FOCOS DE INCÊNDIO NO ESTADO DE SÃO PAULO UTILIZANDO MACHINE LEARNING E GOOGLE EARTH ENGINE

PREDICTIVE MODELING OF WILDFIRE RISK IN THE STATE OF SÃO PAULO USING MACHINE LEARNING AND GOOGLE EARTH ENGINE

Ciências Exatas e da Terra, Engenharias • 31/08/2026

REGISTRO DOI: 10.70773/revistatopicos/788066925

Mario Benini Martinelli[1]

Gustavo Henrique Romão[2]

Hygor Santiago Lara[3]

RESUMO

O estudo desenvolve um modelo preditivo de risco de focos de incêndio para o estado de São Paulo, Brasil, utilizando técnicas de aprendizado de máquina e a plataforma Google Earth Engine. A pesquisa combina variáveis meteorológicas de reanálise climática (ERA5-Land), sensoriamento remoto (Sentinel-2), dados topográficos (NASADEM), indicadores hídricos (CHIRPS, GLDAS) e informações de cobertura do solo (MapBiomas) para um conjunto de dados com 47.149 registros de focos de fogo e não-fogo entre 2020 e 2024. Três algoritmos de classificação — Random Forest, Extreme Gradient Boosting e Regressão Logística — são comparados por meio de duas estratégias de validação, temporal (Leave-One-Year-Out) e espacial (Spatial K-Fold), sendo os fatores mais relevantes identificados com o auxílio de SHAP Values. Os resultados indicam que o modelo Random Forest apresenta o melhor desempenho nas duas validações, com AUC-ROC médio de 0,852 e 0,845, respectivamente. As análises exploratória e de interpretabilidade demonstram que as ocorrências de incêndio no estado estão primariamente associadas ao déficit hídrico, destacando-se a umidade relativa do ar, a precipitação acumulada de 30 dias e a radiação solar como as variáveis mais importantes. O estudo evidencia ainda maior concentração espacial e temporal dos focos nas regiões norte do estado e nos meses de agosto e setembro, associados a áreas agrícolas de cana-de-açúcar e mosaico de usos. Conclui-se que a integração entre dados de sensoriamento remoto, reanálise climática e algoritmos baseados em árvores oferece resultados robustos para o monitoramento e a prevenção de incêndios florestais na região.

Palavras-chave: Incêndio florestal; Aprendizado de máquina; Google Earth Engine; Random Forest; Sensoriamento Remoto.

ABSTRACT

The study develops a predictive wildfire risk model for the state of São Paulo, Brazil, using machine learning techniques and the Google Earth Engine platform. The research combines meteorological variables from climate reanalysis (ERA5-Land), remote sensing (Sentinel-2), topographic data (NASADEM), hydrological indicators (CHIRPS, GLDAS), and land cover information (MapBiomas) for a dataset comprising 47,149 fire and non-fire records between 2020 and 2024. Three classification algorithms — Random Forest, Extreme Gradient Boosting, and Logistic Regression — are compared through two validation strategies, temporal (Leave-One-Year-Out) and spatial (Spatial K-Fold), with the most relevant factors identified through the use of SHAP Values. The results indicate that the Random Forest model achieves the best performance in both validation strategies, with an average AUC-ROC of 0.852 and 0.845, respectively. The exploratory and interpretability analyses demonstrate that fire occurrences in the state are primarily associated with water deficit conditions, with relative humidity, 30-day accumulated precipitation, and solar radiation standing out as the most important variables. The study further reveals greater spatial and temporal concentration of fire occurrences in the northern regions of the state and during the months of August and September, associated with sugarcane agricultural areas and mixed-use mosaics. It is concluded that the integration of remote sensing data, climate reanalysis, and tree-based algorithms offers robust results for wildfire monitoring and prevention in the region.

Keywords: Wildfire; Machine learning; Google Earth Engine; Random Forest; Remote Sensing.

1. INTRODUÇÃO

Os incêndios florestais são fenômenos que ocorrem em diversas regiões do planeta e são responsáveis por grandes impactos nos ecossistemas locais. Esses impactos podem ser tanto na própria biodiversidade da região com a devastação de habitats naturais de fauna e flora, quanto em áreas urbanas com os incêndios alcançando a população e gerando risco para a segurança das pessoas. A própria atividade humana e o contexto atual de mudanças climáticas acabam tornando essa situação cada vez mais crítica (Ejaz; Choudhury, 2025).

O aumento de ocorrências e da gravidade desses incêndios deixa clara a necessidade do desenvolvimento de sistemas responsáveis por prever esses fenômenos e tentar antecipar as regiões que podem ser afetadas. Para esses novos sistemas, a utilização de variáveis climáticas e de terreno, por exemplo, tem gerado ótimos resultados quando combinados com algoritmos de aprendizado de máquina (Jain et al., 2020).

No Brasil, a combinação de períodos de seca, uso do fogo como prática agrícola e mudanças climáticas tem gerado ocorrências de incêndios que são cada vez mais graves, impactando negativamente a biodiversidade do país e a população próxima dessas ocorrências. Isso evidencia a crescente demanda por políticas que atuem na prevenção dessas catástrofes (Pivello et al., 2021). Os dados de estudos de área queimada no bioma Cerrado mostraram que os incêndios impactaram cerca de 40% de todo o bioma entre os anos 1985 e 2022, sendo que essas ocorrências tiveram como principal fator de ignição a atividade humana (Arruda et al., 2024).

No estado de São Paulo (Brasil), a quantidade de ocorrências de focos de fogo tem aumentado, mesmo não sendo um dos maiores

estados brasileiros. De acordo com dados do Instituto Nacional de Pesquisas Espaciais (INPE), o ano de 2024 apresentou a maior quantidade de ocorrências desses focos no estado desde 1998 (ano do início da análise realizada pelo INPE), chegando a 8700 registros (INPE, 2026b). Devido ao estado de São Paulo ter grande variedade no tipo de cobertura do solo, como Mata Atlântica, Cerrado e plantio de algumas espécies, o estudo de toda a área do estado acaba tendo grande importância na busca do entendimento dessa dinâmica e trazer informações que auxiliem na prevenção de danos graves à fauna, flora e população local.

Mesmo com a crescente necessidade de desenvolvimento desses sistemas de predição para amenizar os potenciais riscos dos incêndios, a quantidade dos estudos voltados especificamente para o estado de São Paulo utilizando dados de sensoriamento remoto e algoritmos de aprendizado de máquina (ML) ainda não é muito expressiva. Após análise da bibliografia de pesquisas pelo mundo que se baseiam em sensoriamento remoto para mapear incêndios, verifica-se o aumento de publicações nos últimos anos e que a concentração dessas publicações está localizada em regiões como Estados Unidos, Espanha e China (Guiop-Servan et al., 2025). Apesar do Brasil aparecer em quarto lugar em quantidade de publicações, os estudos não apresentam foco na predição de risco e no estado de São Paulo. O que pode ser verificado é uma maior quantidade de trabalhos focados na Amazônia e no Cerrado como um todo, o que acaba gerando essa lacuna em publicações do tipo para o estado de São Paulo.

Diante do contexto apresentado, este trabalho propõe o desenvolvimento de um modelo preditivo de risco de focos de incêndio para o estado de São Paulo. Para isso, através do *Google*

Earth Engine (GEE) foram extraídas variáveis meteorológicas provenientes de reanálise climática (ERA5-Land), sensoriamento remoto (Sentinel-2), dados topográficos (NASADEM), indicadores hídricos (CHIRPS, GLDAS) e informações de cobertura do solo (MapBiomas). Três algoritmos (*Random Forest, Extreme Gradient Boosting* e Regressão Logística) foram comparados com validações temporal (*Leave-One-Year-Out*) e espacial (*Spatial K-Fold*). A identificação dos principais fatores para a ocorrência dos focos de fogo foi aprofundada pelo uso de SHAP *Values*, o que auxiliou na interpretação dos resultados.

2. REFERENCIAL TEÓRICO

A elevada disponibilidade de dados de sensoriamento remoto e o avanço na tecnologia de computação em nuvem auxiliaram no crescimento do número das aplicações de ML no setor de predição de risco de incêndios florestais, principalmente nos últimos anos. Dentre essas aplicações, o algoritmo *Random Forest* (RF) apresenta maior desempenho para prever ocorrências de fogo e também para prever possíveis áreas queimadas após um incêndio (Jain et al., 2020). Além disso, a utilização de um conjunto de algoritmos diferentes (como RF, *Extreme Gradient Boosting*, Regressão Logística e redes neurais) no mesmo trabalho para comparar os seus desempenhos aparece em diversas regiões. Essas comparações demonstram que os modelos baseados em árvores de decisão têm melhores resultados que os modelos baseados em métodos lineares, já que as variáveis ambientais apresentam relações não-lineares com a ocorrência de incêndios (Ejaz; Choudhury, 2025).

Selecionar as variáveis preditoras mais relevantes é uma etapa crucial na elaboração dos modelos de ML. O Índice de Vegetação por

Diferença Normalizada (NDVI) é uma variável conhecida nos estudos sobre incêndios florestais e em muitos casos é o fator com a maior contribuição para prever ocorrências de fogo nas florestas (Piralilou et al., 2022). Outro método eficiente além do uso de índices de vegetação capturados por imagens de satélite, é a utilização de variáveis climáticas provenientes de conjuntos de dados conhecidos, como o ERA5-Land, sendo responsável por fornecer dados atmosféricos e dados sobre a situação hídrica de uma região. O uso desses dados em conjunto com o uso das imagens de satélite em plataformas como o GEE, tem provado ser um mecanismo eficiente para o monitoramento de regiões inteiras que sofrem com incêndios e com mudanças de cobertura do solo (Lasaponara et al., 2022). O GEE apresenta uma grande biblioteca de dados públicos e de fácil acesso que cobrem diversas áreas do planeta e também proporciona ferramentas para o desenvolvimento de modelos de ML na própria plataforma (Choi; Yun; Chae, 2025).

Em relação à interpretação dos resultados obtidos por modelos de ML, verifica-se um avanço na metodologia nos últimos anos. O uso de Inteligência Artificial Explicável (XAI), em particular SHAP *Values* (*Shapley Additive Explanations*) para esse tipo de tarefa passou a ser relevante (Lundberg; Lee, 2017). A utilização de SHAP *Values* para interpretar os fatores mais relevantes para ocorrências de fogo apresenta as variáveis relacionadas ao déficit hídrico como sendo as mais importantes na predição, enquanto outras variáveis, como NDVI, complementam o conjunto dos fatores de predição (Kondylatos et al., 2022). Esse método de interpretação tem a capacidade de elencar a contribuição de cada uma das variáveis analisadas e permite uma análise dos limites relevantes dentro dos valores de cada variável. Estudos que comparam um conjunto de modelos de ML e utilizam SHAP *Values* para interpretação dos

resultados no contexto de mapeamento de riscos de incêndios não apenas conseguem avaliar a importância dos fatores de predição, como também possibilitam um maior entendimento dessas variáveis como fatores de risco (Iban; Aksu, 2024).

A seleção das variáveis mais importantes para a predição de fogo acontece antes da modelagem e tem grande impacto no desempenho e na interpretabilidade dos modelos de ML. Em estudos que utilizam dados ambientais, como é o caso deste trabalho, é comum que as variáveis demonstrem distribuição assimétrica e não-normal, o que faz com que a escolha por testes não-paramétricos seja a mais indicada para verificar a diferença entre os grupos de variáveis. Para esse cenário, o teste Mann-Whitney U é o teste utilizado para comparar se duas amostras são da mesma distribuição (Sundjaja; Shrestha; Krishan, 2023). Esse teste converte os valores em postos (*ranks*), comparando a soma dos *ranks* dos grupos. Para complementar os resultados do teste, utiliza-se como medida de tamanho de efeito a correlação biserial de postos (*rank-biserial correlation*), que quantifica a dominância estocástica entre os grupos em uma escala de −1 a +1: valores (r) próximos de zero indicam ausência de diferença prática, enquanto valores próximos de ±1 indicam dominância completa de um grupo sobre o outro (Kerby, 2014).Convencionalmente, valores de $|r| \geq 0{,}10$ são interpretados como efeito pequeno, $|r| \geq 0{,}30$ como efeito médio e $|r| \geq 0{,}50$ como efeito grande. O conceito de tamanho de efeito (*effect size*) é fundamental para a interpretação dos resultados, pois o valor-p (p-*value*) indica apenas a probabilidade de se observar um resultado tão extremo quanto o obtido sob a hipótese nula, não demonstra qual de fato é a relevância da diferença que foi detectada (Fritz; Morris; Richler, 2012).

Além disso, utiliza-se o coeficiente de correlação de Spearman para analisar as relações entre as variáveis contínuas, que é uma medida não-paramétrica que também funciona a partir dos ranks dos dados e também é utilizada quando há distribuição não-normal dos dados (Jain et al., 2020). Valores da correlação que sejam próximos de -1 ou 1 indicam forte correlação negativa ou positiva, respectivamente, enquanto valores próximos de zero indicam baixa correlação. Na etapa de análise exploratória, a correlação de Spearman é amplamente utilizada para identificar pares de variáveis colineares, já que a utilização dessas variáveis em conjunto pode comprometer os resultados dos modelos (Ejaz; Choudhury, 2025).

A dimensão espacial dos dados de ocorrência de fogo torna necessária a verificação da presença de autocorrelação espacial, ou seja, a tendência de observações geograficamente próximas serem mais similares entre si do que observações distantes. O índice de Moran I (Moran's I) é uma estatística amplamente utilizada para mensurar a autocorrelação espacial global de uma variável (Moran, 1950). Seus valores variam entre −1 e +1, onde valores positivos indicam agrupamento espacial (*clustering*) de valores similares, valores negativos indicam dispersão, e valores próximos de zero indicam distribuição espacialmente aleatória. A detecção de autocorrelação espacial significativa nos dados tem implicação direta sobre a estratégia de validação dos modelos, pois viola a premissa de independência entre observações assumida pela validação cruzada aleatória convencional (Pascolini-Campbell et al., 2025).

Além da utilização da correlação de Spearman para tentar entender se alguma das variáveis coletadas deveria ser retirada do conjunto de dados final para os modelos de ML, utilizou-se a comparação

entre três métricas de importância de variáveis: a importância baseada na impureza de Gini, a importância por permutação e SHAP *Values*. A importância baseada na impureza de Gini mensura a contribuição média de cada variável na redução da impureza dos nós de divisão ao longo de todas as árvores do modelo, sendo uma medida intrínseca aos algoritmos baseados em árvores de decisão (Breiman, 2001). A importância por permutação complementa essa análise ao avaliar o impacto de cada variável no desempenho preditivo do modelo: os valores de cada variável são embaralhados aleatoriamente, quebrando sua relação com a variável resposta, e a degradação resultante na métrica de avaliação é registrada como medida de importância. A comparação entre os três ranqueamentos permitiu identificar variáveis com contribuição consistentemente baixa em todos os critérios, as quais foram candidatas ao descarte. Essa estratégia reduz o risco de eliminar variáveis relevantes com base em um único critério potencialmente enviesado (Fisher; Rudin; Dominici, 2019).

O *Random Forest* (RF) é um método de aprendizado de máquina baseado em conjunto de árvores de decisão. O algoritmo constrói múltiplas árvores de decisão durante o treinamento, utilizando um subconjunto aleatório de variáveis em cada nó de divisão. A predição final é obtida pela agregação das previsões individuais das árvores, o que evita a geração de sobreajuste (*overfitting*) e possibilita a captura de interações não-lineares entre as variáveis (Breiman, 2001). O *Extreme Gradient Boosting* (XGBoost) é um algoritmo de gradient boosting que constrói árvores de forma sequencial, onde cada nova árvore é ajustada para corrigir os erros residuais da anterior, minimizando uma função de perda por meio de gradiente descendente (Chen; Guestrin, 2016). Já a Regressão Logística, é um modelo linear generalizado que estima a probabilidade de

pertencimento a uma classe por meio de uma função sigmoide aplicada à combinação linear das variáveis preditoras (Hosmer; Lemeshow; Sturdivant, 2013).

A avaliação da capacidade de generalização dos modelos em dados espaciotemporais requer estratégias de validação que respeitem a estrutura de dependência dos dados. A validação cruzada aleatória convencional (*random k-fold*) tende a superestimar o desempenho dos modelos quando aplicada a dados com autocorrelação espacial ou temporal, pois pontos próximos podem aparecer simultaneamente nos conjuntos de treino e teste (Roberts et al., 2017). O *Spatial K-Fold Cross-Validation* é uma estratégia que divide os dados em blocos de espaços próximos, garantindo separação geográfica entre os conjuntos de treino e teste e produzindo estimativas de desempenho mais realistas para dados com dependência espacial (Pohjankukka et al., 2017; Valavi et al., 2019). O *Leave-One-Year-Out* (LOYO) é uma estratégia análoga na dimensão temporal: em cada iteração, os dados de um único ano são retidos para teste, enquanto os dados de todos os demais anos são usados para treino, avaliando assim a capacidade do modelo de generalizar para condições climáticas e de uso do solo não observadas durante o ajuste (Sweet et al., 2023).

A performance dos modelos foi avaliada por três métricas. A Área sob a Curva ROC (AUC-ROC) mede a capacidade discriminatória do modelo ao longo de todos os limites de classificação possíveis, variando de 0,5 (equivalente a classificação aleatória) a 1,0 (classificação perfeita) (Fawcett, 2006). A Precisão (*Precision*) representa a proporção de predições positivas que são verdadeiramente positivas, enquanto o *Recall*, também denominado sensibilidade ou taxa de verdadeiros positivos, indica a proporção de

casos positivos reais que foram corretamente identificados pelo modelo.

3. METODOLOGIA

3.1. Área de Estudo

Neste trabalho, o estado de São Paulo, localizado na região sudeste do Brasil, foi a região definida como área de estudo. Os limites do seu território foram obtidos a partir do banco de dados Global Administrative Unit Layers (GAUL, nível 1), que é disponibilizado pela *Food and Agriculture Organization of the United Nations* (FAO) e foi acessado através do GEE. Apenas áreas terrestres foram consideradas, já que superfícies aquáticas não são de interesse neste estudo. Para essa restrição, foi utilizado o mascaramento dessas superfícies aquáticas no GEE com o produto *Global Surface Water* (GSW 1.4) disponibilizado pelo *Joint Research Centre* (JRC), em que os pixels com presença de registro de água foram excluídos. O sistema de referência selecionado para a projeção dos dados foi o EPSG:4326.

3.2. Registros de Focos de Fogo

Os registros de focos de fogo utilizados no trabalho foram coletados através do Programa de Monitoramento de Queimadas do INPE (INPE, 2026a). Estes dados correspondem ao período de 2020 a 2024 e foram filtrados para conter apenas ocorrências no estado de São Paulo. Os focos de fogo registrados pelo programa englobam ocorrências de fogo com aproximadamente 30 metros de extensão por 1 metro de largura, no mínimo, e são registradas quando aparecem em um pixel das imagens, que foram captadas pelo satélite de referência AQUA_M-T (INPE, 2026c).

3.3. Geração de Pontos de Ausência de Fogo

Já com os registros de focos de fogo coletados, foi necessário gerar uma amostra de registros em que não há presença desses focos. Os registros de não-fogo foram gerados aleatoriamente dentro dos limites do estado de São Paulo, em uma proporção de um para um em relação aos registros de fogo registrados a cada ano, separadamente. Com o intuito de evitar pontos falsos de não-fogo, como pontos localizados em rios, utilizou-se o mascaramento pelo produto GSW dessas regiões de corpos d'água. Para cada um dos pontos gerados foi atribuída uma data aleatória dentro do respectivo ano.

3.4. Variáveis Preditoras

Todas as variáveis preditoras foram extraídas no GEE para cada ponto (coordenada geográfica) do conjunto de dados, tanto para os registros de foco de fogo quanto para os registros de não-fogo, através de sensoriamento remoto e produtos de reanálise climática. A extração final foi realizada na escala de 100 metros, na projeção EPSG:4326. A Tabela 1 apresenta todas as variáveis coletadas.

Tabela 1. Estrutura final do dataset

Descrição	Fonte	Resolução	Unidade
Elevação	NASADEM	30 m	m
Declividade	NASADEM	30 m	°
Orientação de encosta	NASADEM	30 m	°
Cobertura vegetal	MODIS MOD44B	250 m/anual	% (0–100)

Índice de vegetação	Sentinel-2 SR	10 m	adimensional (−1 a 1)
Índice de umidade da vegetação	Sentinel-2 SR	20 m	adimensional (−1 a 1)
Temperatura do ar (2 m)	ERA5-Land	~9 km	°C
Temperatura de superfície	ERA5-Land	~9 km	°C
Velocidade do vento (10 m)	ERA5-Land	~9 km	m/s
Umidade relativa	ERA5-Land	~9 km	% (0–100)
Radiação solar	ERA5-Land	~9 km	J/m²
Precipitação acumulada (30 dias)	CHIRPS Daily	~5,5 km	mm
Umidade do solo	GLDAS-2.1 NOAH	~27 km	kg/m²
Distância a assentamentos	GHSL 2020	100 m	m
Uso e cobertura do solo	MapBiomas Col. 9	30 m	Código categórico

Fonte: elaboração própria a partir de dados coletados no GEE.

4. RESULTADOS

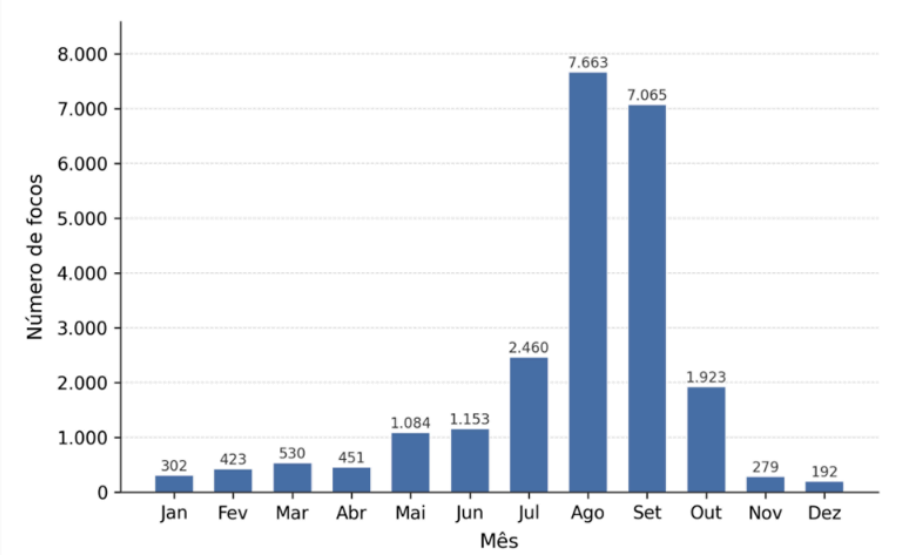
Antes da etapa de utilização de modelos, suas validações e métricas, foi realizada uma análise exploratória em cima dos dados coletados, com o objetivo de verificar a qualidade desses dados, observar a sazonalidade das ocorrências coletadas, identificar padrões entre ocorrência de focos de fogo e tipo de vegetação local, verificar a

distribuição. O dataset final é composto por 47149 registros igualmente distribuídos entre ocorrências de fogo (23525 registros ou 49,9% do total) e ocorrências de não-fogo (23624 registros ou 50,1% do total). das variáveis e compará-las entre si para entender possíveis colinearidades e elencar as melhores variáveis para o estudo.

4.1. Distribuição Temporal dos Focos de Fogo

Dentro do período analisado de 2020 à 2024, temos que o ano com maior número de ocorrências de focos de fogo foi o ano de 2024, chegando à 8700 registros, e o ano com menor número de ocorrências foi o ano de 2022, chegando à 1592 registros. Os anos de 2020, 2021 e 2023 tiveram 6113, 5460 e 1660 ocorrências, respectivamente. De acordo com os dados publicados pelo INPE (INPE, 2026b) sobre a série histórica de focos ativos do estado de São Paulo entre 1998 e 2024, o ano de 2024 apresentou a maior quantidade de ocorrências, superando o ano de 2010, que apresentava a máxima de 7291 ocorrências. Já os anos de 2022 e 2023 apresentaram os menores números de ocorrências de todo o período histórico analisado pelo INPE. Quando é feita a análise da distribuição de ocorrências de incêndio pelos meses do ano (Figura 1) é possível verificar que dois meses claramente se destacam dos demais: agosto e setembro.

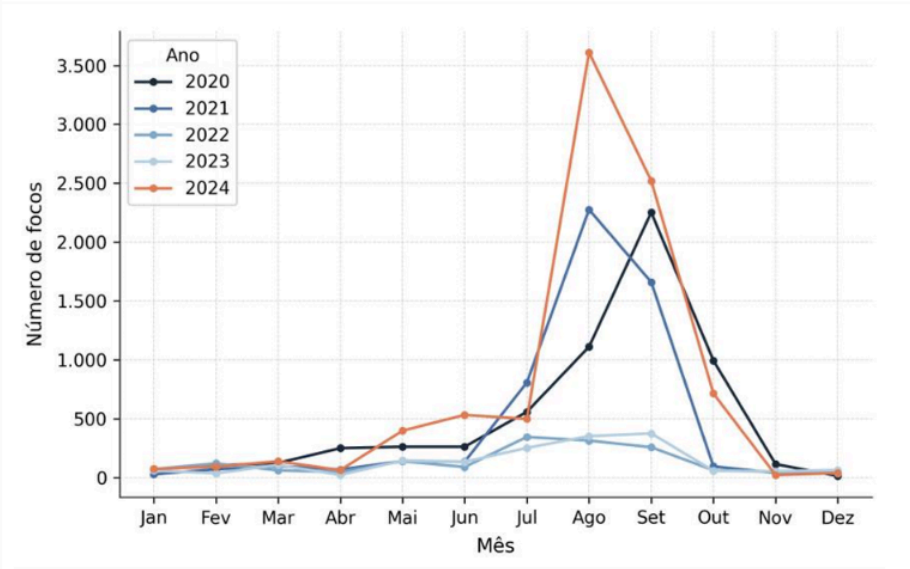
Figura 1. Distribuição Mensal dos Focos de Incêndio (2020 - 2024)



Fonte: Elaboração própria a partir de dados coletados do INPE.

Esse salto no número de ocorrências é explicado pelas características do clima no período: esses dois meses são caracterizados historicamente por serem períodos de seca. No entanto, em 2024 esses valores ultrapassaram marcas históricas. Apenas no mês de agosto de 2024 foram identificados 3610 focos de fogo (Figura 2), o que representa mais do que a soma dos totais de registros de fogo dos anos de 2022 e 2023.

Figura 2. Distribuição Mensal dos Focos de Incêndio por Ano (2020 - 2024)



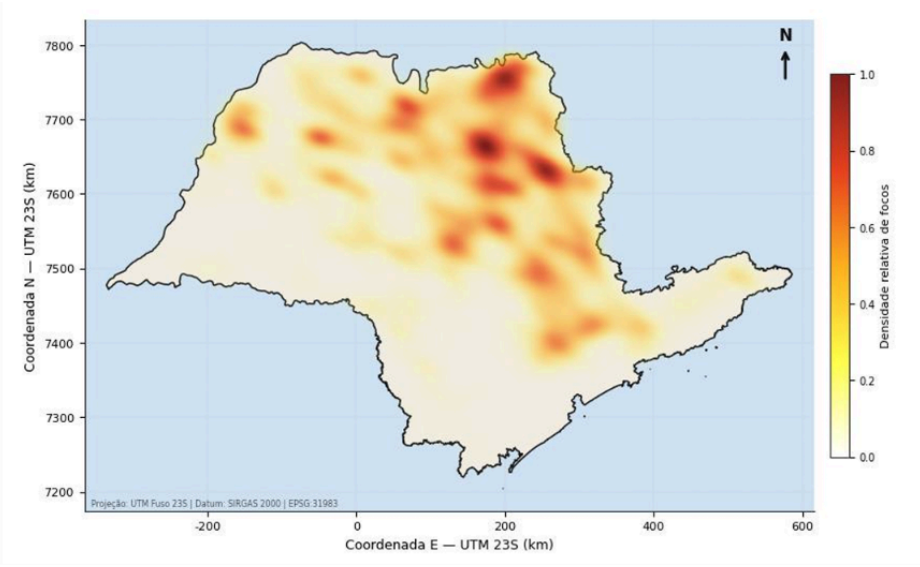
Fonte: Elaboração própria a partir de dados coletados do INPE.

De acordo com o que foi apresentado até o momento, verifica-se a presença de sazonalidade das ocorrências de focos de fogo pelo estado de São Paulo. Porém, apenas como esclarecimento, para os registros de não-fogo gerados aleatoriamente essa sazonalidade não é verificada, os registros mensais para este tipo de ocorrência apresentam quantidades uniformes dentro dos anos de referência.

4.2. Cobertura e Uso do Solo

Além da sazonalidade, as ocorrências de focos de fogo também apresentam outro tipo de padrão. Quando dividimos o estado de São Paulo por suas mesorregiões, verifica-se que essas ocorrências se concentram em Araçatuba, São José do Rio Preto, Ribeirão Preto, Araraquara, Piracicaba, Campinas, Macro Metropolitana Paulista, Metropolitana de São Paulo e Vale do Paraíba Paulista (Figura 3).

Figura 3. Densidade Relativa de Focos de Fogo no Estado de São Paulo, Brasil (2020 - 2024)

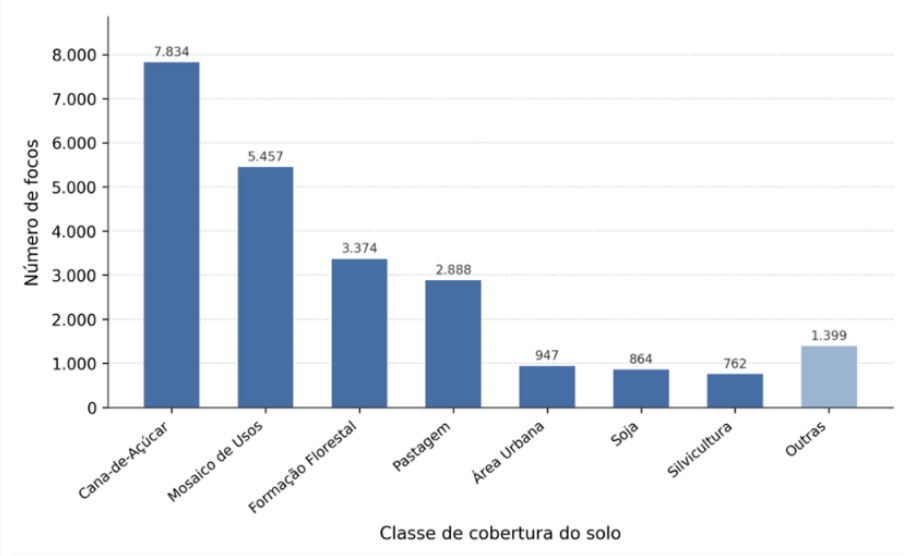


Fonte: Elaboração própria a partir de dados coletados do INPE.

O motivo desses focos se concentrarem em sua maioria nessas mesorregiões do estado de São Paulo mencionadas acima, com exceção de Macro Metropolitana Paulista, Metropolitana de São

Paulo e Vale do Paraíba Paulista, é devido ao fato de que são mesorregiões caracterizadas por grande presença de áreas de agricultura. A rede global MapBiomas realiza o levantamento de como as terras de alguns países estão sendo utilizadas e disponibiliza essas informações através de uma base de dados, que foi coletada através do GEE, em que os diferentes usos de terra são divididos por classes. Com esses dados, verifica-se que a maior parte dos focos de fogo analisados pertencem à classe “Cana-de-Açúcar” (Figura 4), aproximadamente 33,3% do total de ocorrências, com a classe “Mosaico de Usos” (classe que representa o conjunto de áreas de pequenos agricultores e que engloba agricultura, pasto e vegetação nativa em recuperação) aparecendo em segundo lugar em relação à participação do total dos focos de fogo, aproximadamente 23,2% do total de ocorrências.

Figura 4. Ocorrências de Fogo por Classe de Cobertura



Fonte: Elaboração própria a partir dos dados coletados do INPE e MapBiomas.

Em terceiro lugar com maior participação no número total de registros de focos de fogo coletados está a classe “Formação Florestal”, representando 14,3% do total de ocorrências. Essa é uma informação preocupante quando se discute a saúde do meio

ambiente. Focos de fogo em áreas da classe “Formação Florestal” indicam queima de florestas nativas e perda da biodiversidade local.

Para finalizar essa análise do contexto da cobertura do solo dos dados coletados, é importante ressaltar a presença relativamente elevada de focos de fogo em áreas de classe “Área Urbana”, cerca de 4,0% das ocorrências totais. Isso pode ser resultado de focos de fogo em vegetação urbana, como parques, terrenos baldios e margens de rodovias. No entanto, essa participação também pode ser explicada pela desatualização da área urbana registrada ao longo do período analisado, já que para os dados coletados do ano de 2024 foi utilizada a coleção de dados do ano 2023 do MapBiomas para a atribuição dessas classes de cobertura.

4.3. Análise Univariada das Variáveis Preditoras

Após uma análise univariada das variáveis numéricas preditoras coletadas, observou-se que as ocorrências de focos de fogo no estado de São Paulo são primariamente fenômenos relacionados ao déficit hídrico. As variáveis que apresentaram maior diferença relativa nas médias dos valores para fogo e não-fogo foram umidade relativa (diferença de 23,6 pontos percentuais), demonstrando que os focos de incêndio ocorrem em condições em que o ar é mais seco; precipitação acumulada nos últimos 30 dias (diferença de 66,9 mm), quase quatro vezes mais chuvas em pontos de não-fogo; umidade do solo (diferença de 82 kg/m²), solo é mais seco onde há fogo; NDVI com valor médio de 0,439 para fogo e 0,546 para não-fogo; e NDWI com valor médio de 0,037 para fogo e 0,141 para não-fogo. As condições meteorológicas de maior temperatura e ventos mais fortes aparecem em segundo plano, apresentando uma diferença moderada entre as ocorrências de fogo e não-fogo. Já as

variáveis de topografia e proximidade humana apresentaram pouca diferença entre as médias.

A distribuição das variáveis aponta que a normalidade acaba sendo uma exceção, já que a grande maioria das variáveis apresentam comportamento de não-normalidade. Devido à esse comportamento observado de assimetria da distribuição das variáveis e do tamanho amostral consideravelmente grande (quase 50 mil registros), foram utilizados o teste não-paramétrico Mann-Whitney U (Mann; Whitney, 1947) e o p-valor foi complementado pela correlação *Rank-Biserial* como medida de *effect size* (Kerby, 2014). Todas as variáveis coletadas demonstraram diferença significativa ($p < 0,05$), o que já era esperado pelo tamanho da amostra de dados coletada, o que torna a análise do effect size o principal fator de interpretação das variáveis. A umidade relativa do ar e a precipitação acumulada dos últimos 30 dias apresentaram os maiores efeitos, enquanto as variáveis de declividade, elevação, distância a assentamentos e orientação apresentaram *effect size* negligenciáveis neste primeiro teste estatístico.

4.4. Matriz de Correlação

Com a matriz de correlação de Spearman foi observado que os dois pares de variáveis (NDVI com NDWI e Temperatura do ar com Temperatura de superfície) demonstraram alta correlação ($|r| >= 0,75$), o que serviu como um dos fatores da remoção das variáveis NDWI e Temperatura de superfície dos modelos de ML finais.

4.5. Autocorrelação Espacial

A verificação de uma possível correlação espacial foi realizada através do índice de Moran’s I. Esse teste apontou que os dados

coletados demonstraram a existência de *clustering* espacial (0,9994), o que estabeleceu para este trabalho a necessidade da utilização do *Spatial K-Fold* como estratégia complementar da validação do modelo de ML utilizado.

4.6. Seleção de Variáveis

Para selecionar as variáveis mais importantes para este estudo, foram utilizados três métodos complementares: importância baseada na impureza de Gini (BREIMAN, 2001), importância por permutação (Fisher; Rudin; Dominici, 2019) e SHAP *Values* (Lundberg; Lee, 2017). A utilização desses três métodos possibilitou uma análise mais confiável e apresentou consistentemente quais variáveis são as mais importantes (variáveis que se posicionaram no topo dos *rankings* dos três métodos).

As variáveis umidade relativa do ar, precipitação acumulada de 30 dias e radiação solar ocuparam consistentemente as três primeiras posições, seguidas por elevação e temperatura do ar. As variáveis de declividade e orientação de encosta apresentaram importância negligenciável nos três métodos e foram removidas do modelo final. Além da remoção dessas duas variáveis, esses três métodos somados à análise anterior da correlação de Spearman foram os fatores para a decisão da remoção também das variáveis NDWI e temperatura de superfície.

4.7. Algoritmos

Neste trabalho foram utilizados três algoritmos de ML para a classificação de risco de fogo no estado de São Paulo: RF, XGBoost e Regressão Logística. Os modelos foram treinados com hiperparâmetros padrão da biblioteca scikit-learn/XGBoost, com

exceção do número de estimadores fixado em 200 para RF e XGBoost. A otimização de hiperparâmetros não foi realizada. O pré-processamento das variáveis foi adaptado para cada algoritmo. Para os modelos RF e XGBoost não foi realizada nenhuma normalização ou escalonamento (Breiman, 2001). Já para o modelo de Regressão Logística as variáveis foram padronizadas através do *Standard Scaler.* Em relação à codificação da variável de cobertura do solo, foi utilizado o *Label Encoding* para o RF (grupos de 0 a 6), foi utilizado o próprio suporte nativo para variáveis categóricas do XGBoost e foi utilizado *One-Hot Encoding* para a variável de cobertura do solo (a partir dos grupos definidos pelo *Label Encoding*).

4.8. Estratégia de Validação

A escolha do uso da estratégia LOYO vem com a possibilidade de validar o modelo para datas futuras, já que simula o cenário real de treinar com dados históricos para realizar a predição em anos que ainda não foram observados. A análise de autocorrelação espacial apresentada na seção 4.5 deste trabalho apontou para a necessidade de uma validação espacial do modelo através da utilização do *Spatial K-Fold*. Para a realização dessa validação, o estado de São Paulo foi dividido em quatro regiões: Norte Leste, Norte Oeste, Sul Leste e Sul Oeste. Essa separação em quatro regiões se demonstrou a divisão com maior equilíbrio de pontos e garantindo uma quantidade mínima aceitável de pontos por bloco. Os dois blocos do Norte possuem a maior quantidade de focos de fogo, fenômeno já observado anteriormente na seção 4.2 deste trabalho.

5. CONCLUSÃO

Os três modelos utilizados neste trabalho apresentaram resultados satisfatórios para a predição de focos de fogo no estado de São Paulo. Como pode ser visto pela Figura 5, o modelo RF obteve o melhor desempenho nas duas estratégias de validação: na validação LOYO seu AUC-ROC médio para os cinco anos testados foi de 0,852 aproximadamente; e na validação *Spatial K-Fold* seu AUC-ROC médio para as quatro regiões foi de 0,845 aproximadamente. Em seguida, vem o modelo XGBoost com valores médios de AUC-ROC de 0,845 e 0,832 aproximadamente para as estratégias de validação LOYO e *Spatial K-Fold*, respectivamente. O modelo de Regressão Logística demonstrou o menor desempenho entre os três modelos, com valores médios de AUC-ROC de 0,827 e 0,831 aproximadamente para as estratégias de validação LOYO e *Spatial K-Fold*, respectivamente.

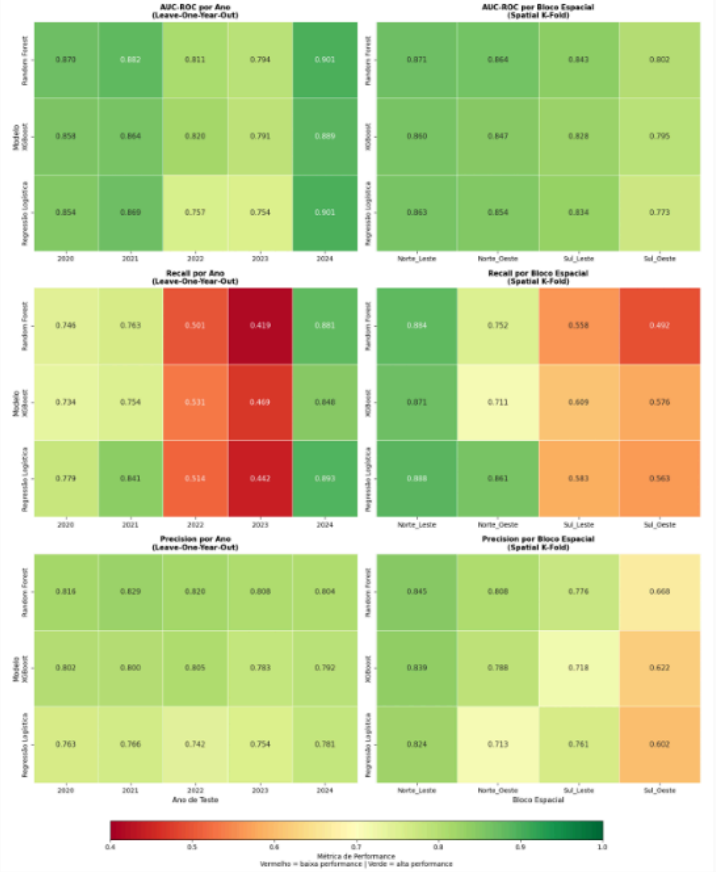
A análise por ano através da estratégia LOYO revelou uma variação considerável da performance de todos os três modelos. Os anos de 2022 e 2023 foram anos atípicos e registraram as menores quantidades de registros de fogo de todo o período histórico entre 1998 e 2024 (Fonte INPEb), apresentando em 2023 um AUC-ROC mínimo de 0,794 e *Recall* mínimo de 0,419 para o modelo RF, um alto contraste se comparado com os valores de 2024 (AUC-ROC de 0,901 e *Recall* de 0,881), o ano com a melhor performance observada para os três modelos. Essa queda dos registros de fogo nos anos de 2022 e 2023 possivelmente está associada às mudanças climáticas derivadas do fenômeno *La Niña*, principalmente no comportamento de precipitações na região do estado de São Paulo.

A análise por regiões do estado de São Paulo através da estratégia de *Spatial K-Fold* demonstrou que a performance dos modelos foi superior para os dois blocos da região Norte, com AUC-ROC de 0,871

e 0,864 para as regiões Norte Leste e Norte Oeste, respectivamente, no modelo RF. Isso está conectado ao fato de que a região Norte do estado possui a maior quantidade de registros de fogo de todo o estado de São Paulo, o que faz com que os modelos façam uma generalização com menor precisão nas regiões com menor quantidade de registro de fogo, como é o caso dos dois blocos da região Sul deste trabalho.

A comparação entre as métricas de *Recall* e *Precision* revelou que nos três modelos os valores de *Precision* são maiores que os valores de *Recall* na maioria dos casos, ou seja, os modelos predizem focos de fogo com alta confiança, mas não conseguem detectar uma parcela dos focos reais que acontecem, principalmente em anos que o comportamento climático foge do padrão, como observado em 2022 e 2023.

Figura 5. Performance por estratégia de validação dos três modelos nas métricas AUC ROC, Recall e Precision



Fonte: Elaboração própria a partir dos resultados dos modelos dos dados coletados.

REFERÊNCIAS BIBLIOGRÁFICAS

BREIMAN, L. **Random forests.** Machine Learning, [S. l.], v. 45, p. 5–32, 2001.

CHEN, T.; GUESTRIN, C. **XGBoost:** a scalable tree boosting system. In: ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 22., 2016, San Francisco. Anais [...]. New York: ACM, 2016. p. 785–794.

CHOI, J.; YUN, Y.; CHAE, H. **Forest fire risk prediction in South Korea using Google Earth Engine:** comparison of machine learning models. Land, [S. l.], v. 14, n. 1155, 2025.

DA SILVA ARRUDA, V. L. et al. **Assessing four decades of fire behavior dynamics in the Cerrado biome (1985 to 2022).** Fire

Ecology, [S. l.], v. 20, n. 64, 2024.

EJAZ, N.; CHOUDHURY, S. **A comprehensive survey of the machine learning pipeline for wildfire risk prediction and assessment.** Ecological Informatics, [S. l.], p. 103325, 2025.

FAWCETT, T. **An introduction to ROC analysis.** Pattern Recognition Letters, [S. l.], v. 27, n. 8, p. 861–874, 2006.

FISHER, A.; RUDIN, C.; DOMINICI, F. **All models are wrong, but many are useful:** learning a variable's importance by studying an entire class of prediction models simultaneously. Journal of Machine Learning Research, [S. l.], v. 20, n. 177, p. 1–81, 2019.

FRITZ, C. O.; MORRIS, P. E.; RICHLER, J. J. **Effect size estimates:** current use, calculations, and interpretation. Journal of Experimental Psychology: General, [S. l.], v. 141, n. 1, p. 2–18, 2012.

GUIOP-SERVAN, R. E. et al. **Remote sensing for wildfire mapping:** a comprehensive review of advances, platforms, and algorithms. Fire, [S. l.], v. 8, n. 316, 2025.

HOSMER, D. W.; LEMESHOW, S.; STURDIVANT, R. X. **Applied logistic regression.** 3. ed. Hoboken: Wiley, 2013.

IBAN, M. C.; AKSU, O. **SHAP-driven explainable artificial intelligence framework for wildfire susceptibility mapping using MODIS active fire pixels:** an in-depth interpretation of contributing factors in Izmir, Türkiye. Remote Sensing, [S. l.], v. 16, n. 2842, 2024.

INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS (INPE). **Banco de Dados de Queimadas (BDQueimadas):** exportar dados. São José

dos Campos: INPE, 2026a. Disponível em: https://terrabrasilis.dpi.inpe.br/queimadas/bdqueimadas/#exportar-dados. Acesso em: 27 jan. 2026.

INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS (INPE). **Programa Queimadas:** estatísticas de focos por estado. São José dos Campos: INPE, 2026b. Disponível em: https://terrabrasilis.dpi.inpe.br/queimadas/situacao-atual/estatisticas/estatisticas_estados/. Acesso em: 1 jun. 2026.

INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS (INPE). **Programa Queimadas:** perguntas frequentes (FAQ). São José dos Campos: INPE, 2026c. Disponível em: https://terrabrasilis.dpi.inpe.br/queimadas/portal/pages/secao_informacoes/faq/index.html. Acesso em: 1 jun. 2026.

JAIN, P. et al. **A review of machine learning applications in wildfire science and management.** Environmental Reviews, [S. l.], v. 28, n. 4, p. 478–505, 2020.

KERBY, D. S. **The simple difference formula: an approach to teaching nonparametric correlation.** Comprehensive Psychology, [S. l.], v. 3, p. 1–9, 2014.

KONDYLATOS, S. et al. **Wildfire danger prediction and understanding with deep learning.** Geophysical Research Letters, [S. l.], v. 49, n. 17, e2022GL099368, 2022.

LASAPONARA, R. et al. **On the use of Sentinel-2 NDVI time series and Google Earth Engine to detect land-use/land-cover changes in fire-affected areas.** Remote Sensing, [S. l.], v. 14, n. 19, p. 4723, 2022.

LUNDBERG, S. M.; LEE, S.-I. **A unified approach to interpreting model predictions.** In: ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS (NeurIPS), 31., 2017, Long Beach. Anais [...]. Long Beach: Curran Associates, 2017. p. 4766–4777. Disponível em: https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf. Acesso em: 9 jun. 2026.

MANN, H. B.; WHITNEY, D. R. **On a test of whether one of two random variables is stochastically larger than the other.** Annals of Mathematical Statistics, [S. l.], v. 18, n. 1, p. 50–60, 1947.

MORAN, P. A. P. **Notes on continuous stochastic phenomena.** Biometrika, [S. l.], v. 37, n. 1/2, p. 17–23, 1950.

PASCOLINI-CAMPBELL, M. et al. **Assessment of spatial autocorrelation and scalability in fine-scale wildfire random forest prediction models.** Scientific Reports, [S. l.], v. 15, n. 21504, 2025.

PIRALILOU, S. T. et al. **A Google Earth Engine approach for wildfire susceptibility prediction fusion with remote sensing data of different spatial resolutions.** Remote Sensing, [S. l.], v. 14, n. 3, p. 672, 2022.

PIVELLO, V. R. et al. **Understanding Brazil's catastrophic fires:** causes, consequences and policy needed to prevent future tragedies. Perspectives in Ecology and Conservation, [S. l.], v. 19, n. 3, p. 233–255, 2021.

POHJANKUKKA, J. et al. **Estimating the prediction performance of spatial models via spatial k-fold cross validation.** International

Journal of Geographical Information Science, [S. l.], v. 31, n. 10, p. 2001–2019, 2017.

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

SUNDJAJA, J. H.; SHRESTHA, R.; KRISHAN, K. **McNemar and Mann-Whitney U tests.** In: STATPEARLS [Internet]. Treasure Island: StatPearls Publishing, 2023. Disponível em: https://www.ncbi.nlm.nih.gov/books/NBK560699/. Acesso em: 9 jun. 2026.

SWEET, L. B. et al. **Cross-validation strategy impacts the performance and interpretation of machine learning models.** Artificial Intelligence for the Earth Systems, [S. l.], v. 2, n. 4, 2023.

VALAVI, R. et al. **blockCV:** an R package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models. Methods in Ecology and Evolution, [S. l.], v. 10, n. 2, p. 225–232, 2019.

---

[1] Bacharel em Economia pela Faculdade de Campinas (FACAMP) e em Engenharia Mecânica pela Pontifícia Universidade Católica de Campinas (PUC-Campinas). E-mail: acesse o artigo original para visualizar o e-mail

[2] Bacharel em Ciência da Computação, Universidade Federal de São João del Rei. E-mail: acesse o artigo original para visualizar o e-mail

[3] Doutor em Engenharia Mecânica, Universidade de Campinas. E-mail: acesse o artigo original para visualizar o e-mail