# Modelagem da suscetibilidade de queimadas no bioma Pantanal

Ana Paula F L Menezes
UTFPR - Campus Dois Vizinhos
apaulaflm@gmail.com

Gabriel Stankevix Soares
UTFPR - Campus Dois Vizinhos
gabriel.stankevix@gmail.com

Heron Carlos Gonçalves
UTFPR - Campus Dois Vizinhos
heroncarlos67@gmail.com

Isabela Fernanda Capetti
UTFPR - Campus Dois Vizinhos
isabelacapeti@gmail.com

***Abstract*—Com impactos ambientais e sociais, as ocorrências de queimadas vem aumentando nos últimos dois anos conforme focos monitorados desde 1998 pelo Instituto Nacional de Pesquisas Espaciais (INPE). Esses episódios de queima da vegetação nativa para desenvolvimento rural, urbano, na agricultura e pecuária aumentam a produção de poluentes atmosféricos e acabam alterando a biodiversidade brasileira. A medida risco de fogo (RF) mensura a probabilidade de ocorrência de focos de incêndio e é baseada em dados meteorológicos diários obtidos por sensoriamento remoto espacial desenvolvida pelo INPE. O Brasil é composto pelos biomas Amazônia, Cerrado, Caatinga, Pantanal, Mata Atlântica e Pampa. Quando aplicada a métrica RF em dados reais geoespaciais meteorológicos e de queimadas dos anos 2020 e 2021, segmentados por biomas e considerando variáveis exógenas temporais como mês e hora, foram observados comportamentos distintos em cada um deles. Junto aos resultados dos testes estatísticos e análises de tendências, que apontam correlação com maior quantidade de variáveis que compõem este estudo, e o fato do Pantanal ter tido quase um terço da sua área devastada pelas queimadas em 2020, este bioma foi selecionado para aplicação de métodos de regressão baseados em árvores de decisão para predizer seus valores de RF. O erro médio quadrático resultante foi de 0.13, reafirmando a importância de considerar variáveis exógenas temporais, permitindo predizer a suscetibilidade de queimadas no Pantanal.**

.

## I. Introdução

O Pantanal é um dos seis biomas presentes no Brasil e uma das maiores planícies inundáveis contínuas de água doce do mundo [1]. O fogo faz parte da dinâmica natural do Pantanal e de qualquer ecossistema, pois permite renovação das pastagens nativas e favorece o crescimento de muitas espécies. Entretanto, a prática inadequada sem técnicas de controle colocam em risco a conservação do bioma [2]. No ano de 2020, foram registrados 22.116 focos de queimadas no Pantanal, sendo que, somente entre os meses de janeiro a agosto, foram contabilizados 10.153 [3].

Diante da necessidade de distribuir e quantificar recursos para prevenção de incêndios e mitigar os impactos destes eventos, a medida risco de fogo foi desenvolvida em 2002 pelo Centro de Previsão de Tempo e Estudos Climáticos (CPTEC), que faz parte do INPE. A RF foi construída após análise de várias ocorrências de queimadas no país e indica quão propícia a vegetação está para ser queimada [4].

Esse artigo se propõe a apresentar uma nova modelagem da RF para o bioma Pantanal, contemplando diferentes variáveis além daquelas utilizadas pelo INPE. Considerando tanto dados reais meteorológicos quanto geoespaciais e indicativos de tempo, em cada bioma separadamente, as relações encontradas foram divergentes indicando uma nova estratégia de variáveis para composição de modelos individuais. As relações entre esses dados foram avaliadas para o ecossistema pantaneiro com aplicação de métodos de regressão.

### A. Revisão de Literatura

Em 1972, Soares criou a chamada Fórmula de Monte Alegre (FMA) [5] baseada em incêndios ocorridos numa fazenda localizada no interior do estado do Paraná. A FMA é uma fórmula simples que considera a umidade relativa do ar e a precipitação. Como uma melhoria desse método, que possui um desajuste causado pela mudança de regimes de chuvas interferindo no seu desempenho, foi desenvolvida em 2005 a Fórmula de Monte Alegre Alterada (FMA+). Nessa fórmula são usados fatores como umidade relativa do ar, número de dias sem chuva e velocidade do vento [6]. Essa melhoria trouxe um desempenho superior à FMA, porém, ainda não considera itens importantes para incidência de focos de queimadas como temperatura nem tipo de vegetação.

A equipe do CPTEC desenvolveu em 2002 a medida risco de fogo [4] que é usada até os dias atuais para quantificar a suscetibilidade de ocorrência de queimadas. O cálculo inicial foi baseado em valores ajustados de precipitação, temperatura, umidade relativa do ar e tipo de vegetação do local. Posteriormente passaram a considerar ocorrência de focos de queimadas detectados pelos satélites além de latitude e longitude. Entretanto, fatores como velocidade do vento, umidade do solo e características específicas de cada ecossistema não são utilizados.

Ao aplicar o método de regressão logística em variáveis geoespaciais, indicativas de condições climáticas e fatores socioeconômicos, Guo [7] mapeou a probabilidade de risco do fogo na província de Fujian na China e classificou como

áreas mais propícias àquelas de baixas elevações, sinalizando que padrões geográficos também devem ser considerados no contexto de queimadas. Diante dessa motivação, as variáveis consideradas nesse presente artigo foram selecionadas especificamente para o bioma pantaneiro com objetivo de obter melhor desempenho que o modelo usado pelo INPE.

Este artigo está organizado da seguinte maneira: seção II são descritos as fontes de dados, as ferramentas e algoritmos utilizados e seus delineamentos; na seção III estão sumarizadas as considerações dos resultados encontrados.

## II. Métodos

### A. Base de dados

Este artigo fez uso de 3 conjuntos de dados. O primeiro é *Queimadas* que refere-se a dados de queimadas que ocorreram no território brasileiro no período do mês de Abril de 2020 ao mês de Abril de 2021, contemplando 219.455 registros. Os objetos geométricos que realizam a representação utilizam POINT para seu registro. Dentre os principais campos existentes na tabela podemos citar: riscofogo, frp, bioma e uf. Fonte de Dados: BDQueimadas[8]. O segundo é *Reservas*, que são dados das reservas indígenas localizadas no território brasileiro, contempla 50 reservas indígenas. Os objetos geométricos que realizam a representação utilizam MULTIPOLYGON para seu registro. Dentre os principais campos existentes na tabela podemos citar: nome da reserva, etnia dos povos indígenas pertencentes as reservas, município, uf. Fonte de Dados: Funai-Terras Indigenas[9]. E último *Meteorologia* que são dados das condições climáticas no território brasileiro no período de 2020 a 2021. Este conjunto de dados possui 237.168 registros e não contempla objetos geométricos. Dentre os principais campos existentes na tabela podemos citar: umidade relativa, temperatura, vento, uf, município. A disponibilização dos dados foi feita através de 27 arquivos classificados por UF. Fonte de Dados: Instituo Nacional de Meteorologia[10].

A base de dados final gerada para o desenvolvimento deste estudo possui duas tabelas. A primeira, Queimadas Brasil Reservas, é o resultado da junção entre a base de queimadas e a base de reservas indígenas, por meio de uma função de operação topológica: Contains(geometry reserva, geometry queimadas). Sendo assim, esta tabela contém todos os registros de queimadas e apenas as reservas que possuem pontos de queimadas nas suas áreas. Foi criado um campo do tipo flag, chamado flg_q_r, para indicar a presença ou não de pontos de queimadas nas reservas indígenas. A segunda, Meteorologia, possui exclusivamente os dados meteorológicos. Esta tabela, combinada com a Queimadas Brasil Reservas, formou o conjunto de dados utilizado para a análise exploratória dos dados (EAD) e o desenvolvimento dos modelos.

### B. Modelagem

A modelagem de regressão preditiva tenta encontrar regras para prever os valores de uma ou mais variáveis dado um conjunto de dados objetivos(target) dos valores de outras variáveis no conjunto de dados de entrada, conhecidas como Features. Existem algumas técnicas de modelagem e as mais comuns são regressão linear, rede neural e modelos de árvore de decisão. Os algoritmos que foram desenvolvidos sobre essas técnicas de modelagem surgiram por meio de pesquisas metodológicas em várias áreas, incluindo a estatística, reconhecimento de padrões e aprendizado de máquina [11].

Desta forma, foi desenvolvida uma configuração unificada em ambiente Python, para um modelo regressão linear e construção de variados modelos baseados em árvore de decisão que processam os dados para seleção do melhor modelo de predição de risco de fogo.

No processo de análise exploratória e preparação dos dados, foram realizados testes estatísticos como a correlação de Pearson entre 43 variáveis, onde foram constatados níveis de correlação moderados positivamente e negativamente com variável objetivo risco fogo. Selecionado um conjunto de 9 potenciais variáveis independentes para o desenvolvimento dos modelos.

Para a definição dos modelos fez-se uma análise de linearidade e não linearidade, onde identificou-se que a variável objetivo Risco Fogo atuava em intervalo probabilístico de 0 a 1, não podendo ser extrapolada e que se desenvolve não linearmente ao longo do tempo. Dado que, os modelos baseados em arvore de decisão suportam não linearidade, ao contrário de modelos de regressão linear simples, e que as potenciais features possuem graus de relacionamento significantes, por exemplo, a correlação de avg-umd-ar e avg-temp-ar de 0.70 [12], decidiu-se pela construção de modelos de regressão baseados em arvore de decisão majoritariamente. Todavia, para fins de estudo e comparação manteve-se o modelo de regressão linear.

### C. Modelo de regressão linear

Análise de Regressão é uma das técnicas mais populares para construção de modelos de predição [13]. O modelo de regressão OLS (Ordinary Least of Squares), escolhe parâmetros de uma função linear dado um conjunto de variáveis explicativas pelo princípio dos mínimos quadrados. Este método minimiza a soma dos quadrados das diferenças entre a variável dependente (observada) e aqueles previstos pela função linear da variável independente [11].

No estudo para predição do RF foi utilizado um modelo OLS para regressão linear simples.

### D. Modelos baseados em árvore de decisão

Conforme estudo de [13], Árvore de decisão, Florestas Aleatórias (Random Forest) e Boosting estão entre as top 20 ferramentas e soluções de aprendizado de maquina utilizados por cientistas de dados. Estes três métodos são bem similares pois aplicam o conceito de árvore de decisão como parte da solução. Uma regressão linear, de acordo com [12], falha em situações onde a relação de uma feature e o target é não linear ou as features interagem entre si. Uma árvore de decisão divide os dados várias vezes de acordo com valores de corte de cada feature. Por meio desta divisão, são criados diferentes subconjuntos de dados onde cada instância pertence

a um subconjunto. Os subconjuntos finais são chamados de nós terminais ou folha, os intermediários de nós internos ou divididos. A predição de cada nó folha ocorre pelo resultado médio dos dados de treinamento de cada nó.

O modelo de Florestas Aleatórias, conforme [14], é uma combinação de preditores de árvore de decisão por meio de regras majoritárias, como um vetor amostral aleatório independente, para tomar decisões e obter o resultado final por meio de uma votação na tomada de decisão. O modelo de Grandient Boosting Regressor (GBR) também é baseado em árvore de decisão, porém diferencia-se em 3 elementos[15]: o uso de uma função de perda para ser otimizada, um weak learner (árvore de decisão) para realizar as predições e um modelo aditivo para adicionar uma nova weak learner para minimizar a função de perda.

Neste estudo, foram desenvolvidos modelos para Árvore de Decisão, Florestas Aleatórias e GBR, com o objetivo de analisar o processo de aprendizagem e evolução dos resultados preditos por cada modelo, comparando cada metodologia.

### *E. Critério seleção de Modelo*

Para a construção dos modelos testados, as seguintes variáveis foram selecionadas das 9 potenciais : i.diasemchuv (quantidade de dias sem chuva no bioma), ii.mes (mês corrente do monitoramento), iii.quadrimestre(quadrimestre correspondente ao mês do monitoramento), iv.avg_pressao_atm (pressão média atmosférica em milibar), v.avg_umd_ar (percentual da média da umidade do ar). As variáveis foram padronizadas antes das avaliações por meio do método MinMaxScaler.

Como critério de comparação e definição de qual modelo mais adequado, foi usado o erro médio quadrático (RMSE). O RMSE é a média das diferenças entre valores preditos e os reais ao quadrado. Quanto menor for esse valor, melhor é a precisão preditiva do modelo. Os conjuntos de dados de treino e teste foram aplicados aos métodos de regressão. A Floresta Aleatória resultou em RMSE = 0.13 nos dados de teste e foi o modelo selecionado.

De acordo com o modelo de Floresta Aleatória, o teste de importância de variáveis para predição do RF resultou como principais valores: 26.24% para quadrimestre seguida da média da umidade do ar com 21.63%.

## III. Discussão

O risco de fogo pode apoiar na tomada de decisão e planejamento da estratégia dos órgãos responsáveis pela fiscalização e controle dos focos de queimadas. De diferentes origens, os dados utilizados na escolha do modelo foram combinados e tratados para aplicação de teste de correlação. Os dados geoespaciais foram relacionados e derivaram uma nova variável que não foi significativa para a sequência da modelagem, porém, eles são importantes para evidenciar numérica e graficamente a ocorrência de pontos de queimadas, não somente no Pantanal mas em todos os demais biomas.

Foram construídos modelos de regressão baseados em árvore de decisão e regressão linear, avaliando o relacionamento da variável risco fogo com as demais features, otimizando hiperparâmetros e comparando os índices de RMSE. Os resultados foram bem promissores após a otimização dos hiperparâmetros e seleção de variáveis com maior grau de significância. Conforme o desenvolvimento do treinamento e teste para o modelo Floresta Aleatória, os resultados de predição foram muito bons. Dado um certo erro, RMSE de 0.13, este modelo consegue predizer valores de risco fogo de maneira satisfatória durante períodos de alta variação dos seus índices.

Embora o modelo de Floresta Aleatória tenha obtido o melhor resultado, modelos mais simples e tradicionais como árvore de decisão com otimização de parâmetros obtiveram resultados equivalentes. A regressão linear, obtive uma RMSE um pouco pior devido a menor precisão durante períodos de maior variação do risco fogo, o que evidencia a importância dos fatores na construção das árvores de decisão. Como uma próxima etapa, seria aplicação de um método de predição baseado em redes neurais do risco fogo, comparando os resultados destas três técnicas de modelagem.

## References


[1] E. Vicente and N. Guedes, "Organophosphate poisoning of hyacinth macaws in the southern pantanal, brazil," *Nature - Scientific Reports*, vol. 11, no. 3, 2021.

[2] J. A. Marengo, A. P. Cunha, L. A. Cuartas, K. R. Deusdará Leal, E. Broedel, M. E. Seluchi, C. M. Michelin, C. F. De Praga Baião, E. Chuchón Ângulo, E. K. Almeida, M. L. Kazmierczak, N. P. A. Mateus, R. C. Silva, and F. Bender, "Extreme drought in the brazilian pantanal in 2019–2020: Characterization, causes, and impacts," *Frontiers in Water*, vol. 3, p. 13, 2021.

[3] INPE, "Monitoramento dos focos ativos por bioma," 2020. [Online]. Available: https://queimadas.dgi.inpe.br/queimadas/portal-static/estatisticas_estados/

[4] A.Setzer, R.Sismanoglu, and J.G.M.Santos, "MÉtodo do cÁlculo do risco de fogo do programa do inpe - versÃo 11, junho/2019," *Instituto Nacional de Pesquisas Espaciais - INPE*, 2019.

[5] V.R.Soares, "Determinação de um índice de perigo de incêndio para a região centro paranaense brasil." Tese de Mestrado.

[6] J. N. e Ronaldo Soares e Antonio Batista, "Fma+ um novo índice de perigo de incêncios floretais para o estado do paraná brasil," *FLORESTA*, vol. 36, no. 1, 2006. [Online]. Available: https://revistas.ufpr.br/floresta/article/view/5509

[7] F.-T. Guo, Z. Su, G. Wang, L. Sun, F. Lin, and A. Liu, "Wildfire ignition in the forests of southeast china: Identifying drivers and spatial distribution to predict wildfire likelihood," *Applied Geography*, vol. 66, pp. 12–21, 01 2016.

[8] BDQueimadas. (2021, April) Programa queimadas. [Online]. Available: https://queimadas.dgi.inpe.br/queimadas/bdqueimadas#tabela-de-atributos

[9] Funai. (2021, April) Terras indígenas. [Online]. Available: http://www.funai.gov.br/index.php/shape

[10] INMET. (2021, April) Dados históricos. [Online]. Available: https://portal.inmet.gov.br/dadoshistoricos

[11] K. Yeturu, *Handbook of Statistcs*. Elsevier, 2020, vol. 43.

[12] C. Molnar. (2021, Jan) A guide for making black box models explainable. [Online]. Available: https://christophm.github.io/interpretable-ml-book/tree.html

[13] kdnuggets. (2019, April) Ranking de ferramentas data science. [Online]. Available: https://www.kdnuggets.com/2019/04/top-data-science-machine-learning-methods-2018-2019.html

[14] L. Breiman, "Random forests," p. 5–32, 2001.

[15] J. Brownlee. (2020, Aug) A gentle introduction to the gradient boosting algorithm for machine learning. [Online]. Available: https://machinelearningmastery.com/gentle-introduction-gradient-boosting-algorithm-machine-learning/


## IV. Anexos