

Modelagem da Suscetibilidade a Movimentos de Massa em Taludes Ferroviários com Random Forest: Uma Abordagem com Índices Morfométricos e Ambientais

Felipe Pacheco Silva

Pós-doutorando, Universidade de São Paulo, São Paulo, Brasil, felipe_pacheco@usp.br

Max Levin

Bolsista de Iniciação Científica, Universidade de São Paulo, São Paulo, Brasil, max.levin@usp.br

Luiz Felipe Goulart Fiscina

Doutorando, Universidade de São Paulo, São Paulo, Brasil, luizfiscina@usp.br

Gabriela Vitelli

Mestranda, Universidade de São Paulo, São Paulo, Brasil, gabrielavitelli@usp.br

Marcos Massao Futai

Professor Titular, Universidade de São Paulo, São Paulo, Brasil, futai@usp.br

RESUMO: Os escorregamentos de terra representam uma das principais ameaças à integridade de infraestruturas lineares no mundo e em especial nas regiões tropicais quando se trata de Brasil. Este estudo aplicou o algoritmo Random Forest para modelar a suscetibilidade a escorregamentos em um trecho de 41 km da Estrada de Ferro Carajás (EFC), no estado do Maranhão, Brasil. A metodologia integrou doze variáveis condicionantes derivadas de Modelos Digitais de Elevação (MDE) e dados ambientais, com destaque para declividade, NDVI, precipitação e distância a falhas geológicas. Os dados foram organizados em ambiente SIG, e o modelo foi treinado com 7.500 amostras positivas e 10.000 negativas, obtidas a partir de inventário de escorregamentos e áreas estáveis. A validação cruzada estratificada e a análise da curva ROC ($AUC = 0,94$) indicaram alta capacidade preditiva. O mapa final de suscetibilidade destacou setores com maior risco, especialmente próximos a taludes ferroviários, onde se concentram as classes "Muito Alta" e "Extrema". A abordagem mostrou-se eficiente para identificar áreas críticas e variáveis mais influentes, sendo uma ferramenta útil para o planejamento e mitigação de riscos geotécnicos. A abordagem adotada permite identificar a importância relativa de cada variável na previsão da suscetibilidade, além de gerar mapas probabilísticos de áreas propensas a movimentos de massa.

PALAVRAS-CHAVE: Suscetibilidade a escorregamentos, Infraestrutura ferroviária, Modelos Digitais de Elevação (MDE), Random Forest, SIG (Sistema de Informação Geográfica), Variáveis morfométricas e ambientais.

ABSTRACT: Landslides are among the main threats to the integrity of linear infrastructure in tropical regions. This study applied the Random Forest algorithm to model landslide susceptibility along a 41 km segment of the Carajás Railway (EFC), in Maranhão, Brazil. The methodology integrated twelve conditioning variables derived from Digital Elevation Models (DEM) and environmental data, highlighting Declividade, NDVI, rainfall, and distance to faults. Data were processed in a GIS environment, and the model was trained using 7,500 positive and 10,000 negative samples, based on a landslide inventory and stable areas. Stratified cross-validation and ROC curve analysis ($AUC = 0.94$) indicated high predictive performance. The final susceptibility map revealed high-risk areas, particularly near railway embankments, where "Very High" and "Extreme" classes are concentrated. The approach proved effective in identifying both critical zones and the most influential variables, providing a valuable tool for geotechnical risk planning and mitigation. This approach enables the identification of the relative importance of each predictor in estimating susceptibility, while also generating probabilistic maps of areas prone to mass movement.

KEYWORDS: Landslide susceptibility, Machine learning, Railway infrastructure, Digital Elevation Models (DEMs), Geographic Information Systems (GIS), Morphometric and environmental variables.

1 INTRODUÇÃO

A análise de suscetibilidade a escorregamentos tem sido amplamente desenvolvida desde os anos 1970, com crescente destaque na literatura científica nas últimas décadas, refletindo a preocupação global com os impactos geodinâmicos (Reichenbach *et al.*, 2018). A suscetibilidade a escorregamentos corresponde à estimativa espacial da probabilidade de ocorrência desses eventos, com base em fatores condicionantes do terreno, sem considerar sua frequência temporal (Fell *et al.*, 2008; Cascini, 2008).

O mapeamento da suscetibilidade integra registros históricos e variáveis ambientais para identificar áreas potencialmente instáveis, incluindo aquelas afetadas por escorregamentos originados externamente à área de estudo (Bonini *et al.*, 2025). Tal abordagem constitui uma ferramenta fundamental para o planejamento territorial e a gestão de riscos geotécnicos (Pandey *et al.*, 2019; Gorsevski *et al.*, 2006).

Este estudo analisa a distribuição e os condicionantes dos escorregamentos ao longo de um trecho da Estrada de Ferro Carajás (EFC) (Figura 1), inserida em uma superfície planáltica de baixa altitude do Planalto Dissecado Gurupi-Grajaú (IBGE, 2011; Rodrigues de Sousa *et al.*, 2025). Para modelagem da suscetibilidade, será empregado o algoritmo Random Forest, reconhecido pela alta acurácia em contextos geográficos diversos.

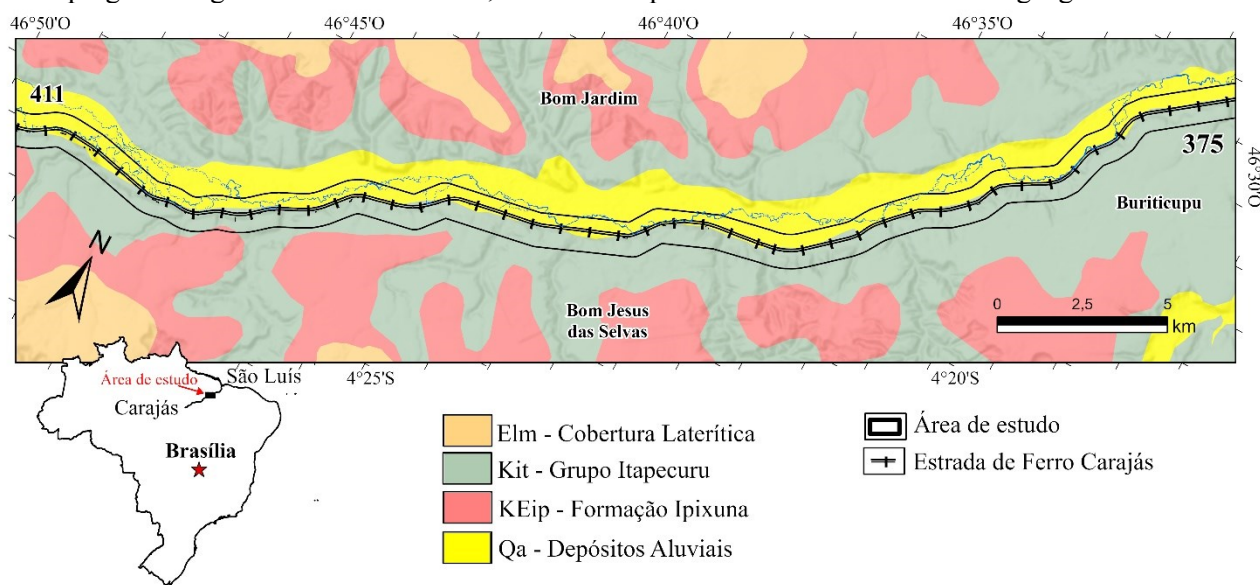


Figura 1. Mapa geológico simplificado do trecho da Estrada de Ferro Carajás (EFC) entre os km 375 a 411, localizado entre os municípios de Buriticupu, Bom Jardim e Bom Jesus das Selvas, no estado do Maranhão.

2 MATERIAIS E MÉTODOS

O processo metodológico para a geração do mapa de suscetibilidade a escorregamentos deste estudo (Figura 2), compreende as seguintes etapas: (a) importação e pré-processamento dos dados geoespaciais; (b) extração dos atributos preditores e divisão dos dados em conjuntos de treino (70%) e teste (30%); (c) aplicação do modelo Random Forest para predição contínua e análise da importância das variáveis; (d) validação do modelo com métricas e curva ROC; e (e) elaboração do mapa final de suscetibilidade. Os dados foram processados no *software* R version 4.5.0 (R Core Team, 2025) e rasterizados e reamostrados com a mesma resolução no ambiente ArcGIS Pro 3.2 (ESRI, 2023).

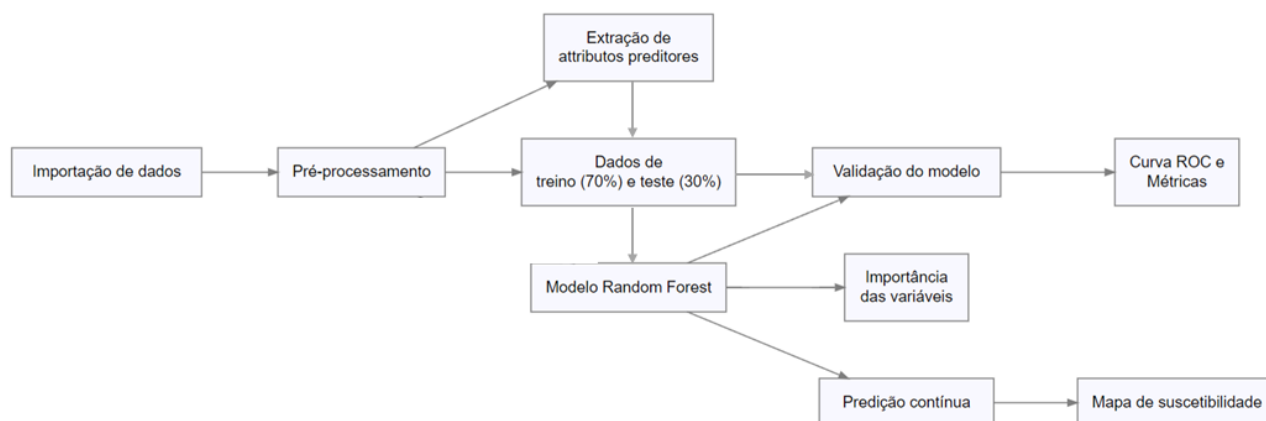


Figura 2. Metodologia adotada neste estudo.

2.1 Aquisição de dados

Foram utilizados dados vetoriais oriundos do inventário de escorregamentos, totalizando 7.500 amostras positivas associadas a 75 eventos mapeados, além de 10 mil amostras negativas correspondentes a áreas estáveis. Os fatores condicionantes foram extraídos e processados a partir de camadas raster geradas com base no Modelo Digital de Elevação ALOS PALSAR (resolução de 12,5 m; JAXA, 2021), bem como de dados ambientais auxiliares. Entre estes, destaca-se a precipitação média anual, obtida a partir do modelo CHIRPS por meio da plataforma Climate Engine (Huntington et al., 2017), acessada em 5 de novembro de 2024 via <https://climateengine.org>. As amostras foram coletadas em um trecho da Estrada de Ferro Carajás (EFC), inserido em um setor geotécnico com histórico relevante de instabilidades superficiais.

A base de dados pontual foi balanceada por subamostragem da classe majoritária (ausência de escorregamentos), com posterior divisão espacial estratificada em conjuntos de treino (70%) e teste (30%). As amostras foram associadas a uma malha regular de 1.000 m para garantir independência espacial entre os grupos.

Foram utilizados doze preditores ambientais derivados de Modelos Digitais de Elevação e dados temáticos, representando variáveis morfométricas, hidrológicas e ecológicas: Curvatura Tangencial (Florinsky, 2011), Curvatura do Perfil (Profile Curvature), declividade, Índice de Umidade Topográfica (Topographic Wetness Index – TWI) (Sørensen et al., 2006), Altura Acima da Drenagem Mais Próxima (Height Above Nearest Drainage – HAND) (Santos et al., 2020), Potência Erosivo do Canal (Stream Power Index – SPI) (Florinsky, 2011), distância até rodovias, distância até rios, distância até falhas geológicas, Orientação do terreno (Aspect), precipitação média anual (Huntington et al., 2017) e o Índice de Vegetação por Diferença Normalizada (Normalized Difference Vegetation Index – NDVI). Esses preditores foram selecionados com o objetivo de caracterizar diferentes aspectos do relevo e do ambiente físico, refletindo processos relacionados à morfologia, dinâmica de vertentes e canais fluviais, bem como à cobertura vegetal e ao regime hídrico da área de estudo. A combinação dessas variáveis permite uma avaliação abrangente dos fatores condicionantes de instabilidades superficiais ao longo da ferrovia. Para evitar redundâncias e garantir a independência relativa entre os preditores, foi realizada uma análise de correlação entre as variáveis ambientais. A matriz resultante orientou a exclusão de fatores altamente correlacionados, contribuindo para a robustez do modelo e a mitigação de efeitos de multicolinearidade.

2.2 Modelo Random Forest

O algoritmo RF é um método de aprendizado de conjunto, que combina vários modelos de árvore de decisão (Breiman 2001). Esse modelo seleciona aleatoriamente porcentagens dos dados de entrada e os separa em subconjuntos. O algoritmo RF pode ser usado para problemas de classificação e regressão. Neste estudo, adotamos a opção de regressão, devido à resposta numérica numéricas geradas (Kim et al. 2018). A modelagem foi conduzida com validação cruzada estratificada ($k=5$) e ajuste do hiperparâmetro $mtry$. O modelo foi treinado com 1.000 árvores e possibilitou a quantificação da importância relativa dos preditores. A saída contínua da probabilidade de ocorrência de escorregamentos foi espacializada, resultando no mapa de suscetibilidade final.

2.3 Avaliação da precisão do modelo

A acurácia do modelo foi validada com dados independentes. As métricas obtidas incluíram acurácia global, precisão, sensibilidade, especificidade e F1-score. A área sob a curva ROC (AUC) também foi utilizada como medida de desempenho discriminativo. O modelo apresentou elevada capacidade preditiva, conforme demonstrado pelas métricas e curva ROC.

3 ANALISES E RESULTADOS

3.1 Análise dos fatores condicionantes

A análise de multicolinearidade foi conduzida por meio do cálculo do Fator de Inflação da Variância (VIF), com base nas variáveis preditoras utilizadas no modelo. Inicialmente, foram avaliados 13 fatores condicionantes, dos quais dois (NDWI e NDVI) apresentaram VIFs elevados (> 10), indicando colinearidade significativa. Após a remoção do NDWI, os demais fatores apresentaram $VIF < 5$, valor considerado aceitável para exclusão de dependência linear significativa entre os preditores (Tabela 1). As variáveis com os maiores VIFs após a filtragem foram Curvatura Tangencial e Declividade, ambas ainda dentro dos limites toleráveis, indicando ausência de multicolinearidade crítica no conjunto final de variáveis.

Tabela 1. Diagnóstico de multicolinearidade entre variáveis independentes.

Tipo de dado	Fatores condicionantes	Teste de Colinearidade	
		VIF	VIF (após remoção)
Características geomorfológicas	Declividade	1,87	1,87
	Aspecto	1,12	1,11
	Curvatura Tangencial	2,06	2,06
	Curvatura	1,94	1,94
	TWI	1,85	1,83
Características hidrológicas	HAND	1,85	1,78
	SPI	1,77	1,77
	Distancia de rios	1,14	1,14
Características geológicas	Precipitação	1,07	1,05
	Distancia de falhas	1,11	1,08
Uso e cobertura do solo	NDWI	13,50	removido
	NDVI	12,93	1,12
	Distancia de rodovias	1,40	1,38

A matriz de correlação evidencia relações estatísticas relevantes entre os fatores topográficos, hidrológicos, de uso da terra e de infraestrutura, com impacto direto na modelagem de suscetibilidade (Figura 3). A declividade apresenta correlação moderada e negativa com o TWI ($r = -0,52$), indicando menor acúmulo de umidade em encostas íngremes. Com o SPI, a correlação é moderada e positiva ($r = 0,50$), refletindo maior energia do fluxo e potencial erosivo em áreas inclinadas. A variável HAND mostra correlações fracas, porém positivas, com declividade ($r = 0,24$) e SPI ($r = 0,21$), e moderada e negativa com TWI ($r = -0,45$), sugerindo menor saturação hídrica em áreas mais elevadas. Esses padrões destacam a influência da posição topográfica na dinâmica do escoamento e saturação.

3.2 Importância das variáveis

A Figura 4 apresenta a importância relativa das variáveis preditoras no modelo Random Forest, com base na métrica de redução média da acurácia (Mean Decrease Accuracy). Essa abordagem avalia o impacto de cada variável na performance do modelo ao comparar a acurácia obtida com os dados originais e com os valores da variável embaralhados aleatoriamente. A diferença média entre esses cenários reflete a relevância da variável na capacidade preditiva do modelo — valores mais altos indicam maior influência na previsão. Os resultados indicam que NDVI, precipitação e distância às falhas são os preditores mais relevantes, evidenciando a influência significativa de fatores ambientais e estruturais na modelagem de suscetibilidade.



Variáveis topográficas e hidrológicas, como distância aos rios, HAND e TWI, também demonstram alta contribuição. Por outro lado, SPI, curvatura e distância às rodovias apresentaram menor impacto sobre a acurácia, sugerindo influência preditiva secundária no contexto analisado.

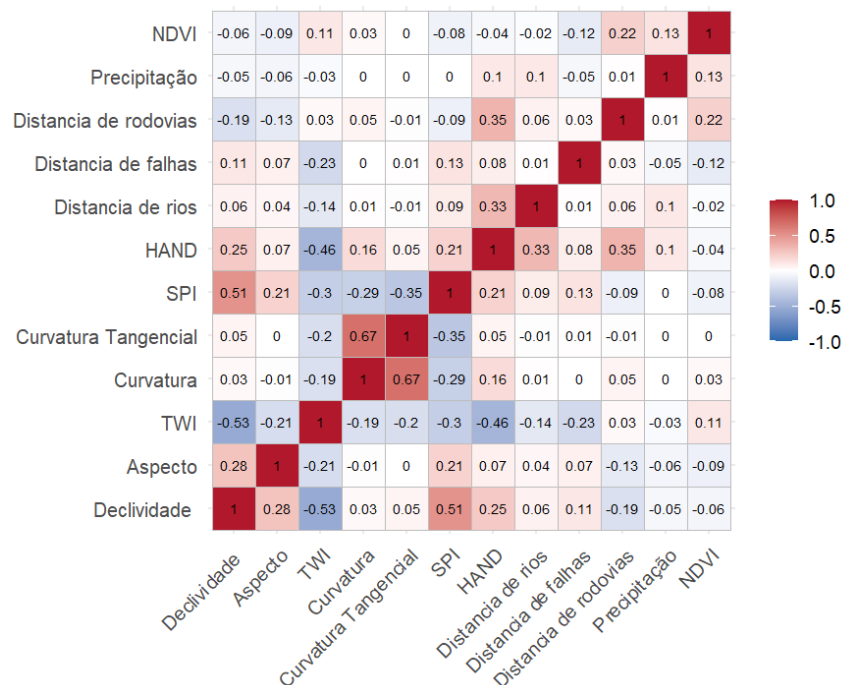


Figura 3. Correlação entre os fatores preditores utilizados na análise de escorregamentos.

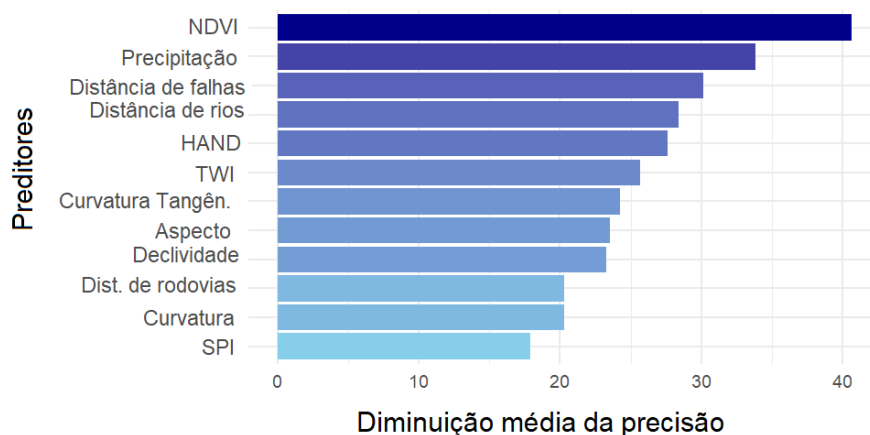


Figura 4. Importância relativa das variáveis preditoras segundo o modelo Random Forest. A métrica utilizada é a redução média do Gini.

3.3 Mapas de suscetibilidade a escorregamentos de terra

Os resultados do modelo de suscetibilidade a escorregamentos são apresentados na Figura 5, que exhibe a distribuição espacial das classes ao longo do trecho ferroviário analisado. As classes variam de "Muito Baixa" a "Extrema", permitindo identificar trechos com maior potencial de instabilidade. Observa-se uma concentração significativa das classes mais elevadas nas proximidades dos taludes ferroviários, especialmente nos trechos destacados, como os entornos dos km 408, 398, 379 e 375, onde há predomínio das classes "Muito Alta" e "Extrema".

Com base nos produtos gerados, a Figura 6 apresenta a distribuição espacial da suscetibilidade a escorregamentos ao longo de um trecho da ferrovia, evidenciando setores com elevada concentração de áreas classificadas como de alto risco. A classificação da suscetibilidade foi dividida em cinco classes: "Muito Baixa", "Baixa", "Moderada", "Muito Alta" e "Extrema". Observa-se

predominância das classes "Muito Baixa" (36,5%) e "Extrema" (38,7%) em termos de área total mapeada (Figura 6a). A análise da distribuição dos escorregamentos observados (Figura 6b) revela que 69,9% dos eventos ocorreram na classe "Extrema", seguida por 18,2% na classe "Muito Alta", demonstrando forte correlação entre os registros empíricos e as zonas de maior suscetibilidade.

A Figura 6c sintetiza graficamente a relação entre a área ocupada por cada classe e a frequência relativa de escorregamentos, destacando a desproporção entre extensão territorial e concentração de eventos. A classe "Extrema", embora represente menos de 40% da área, concentra mais de dois terços dos escorregamentos, indicando sua elevada criticidade. Já a classe "Muito Alta" cobre 10,6% da área, mas responde por 18,2% dos registros. Dessa forma, evidencia-se que as classes "Muito Alta" e "Extrema" concentram mais de 70% dos escorregamentos, embora representem cerca de 40% da área total mapeada.

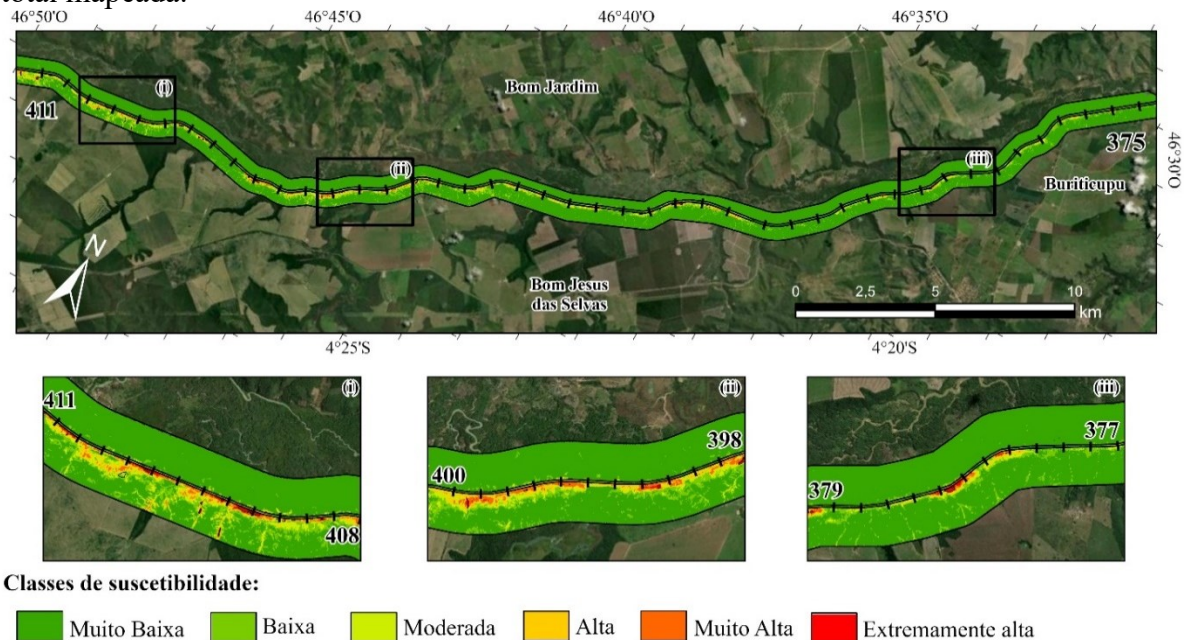


Figura 5. Mapa de suscetibilidade a escorregamentos ao longo da ferrovia. As classes variam de "Muito Baixa" a "Extrema",

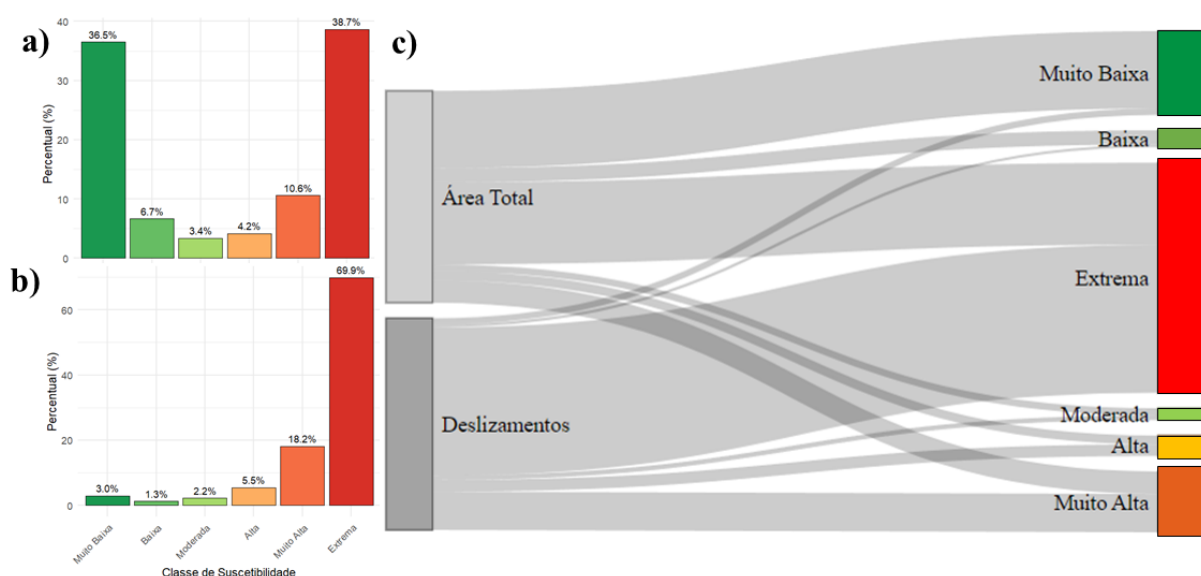


Figura 6. Distribuição das classes de suscetibilidade a escorregamentos. (a) Proporção da área total atribuída a cada classe de suscetibilidade; (b) Proporção das ocorrências de escorregamentos registradas em cada classe; (c) Diagrama de Sankey ilustrando a relação entre a área total, os deslizamentos observados e as classes de suscetibilidade, evidenciando a concentração dos eventos nas classes de maior risco.

3.4 Avaliação do desempenho do modelo

A curva ROC (Figura 7) confirma o bom desempenho do modelo preditivo, com $AUC = 0,94$, indicando alta sensibilidade e especificidade. No entanto, essa métrica, embora útil, não garante por si só a eficácia da avaliação espacial da suscetibilidade (Adnan *et al.*, 2020; Rossi *et al.*, 2010). Entre as limitações, destaca-se a heterogeneidade dos dados de entrada, com fatores derivados de fontes com diferentes resoluções e escalas. Por exemplo, a precipitação média anual foi obtida a partir do modelo CHIRPS, com resolução espacial de 5 km, contrastando com os fatores morfométricos derivados de dados topográficos de maior detalhamento. Isso pode afetar a coerência espacial dos resultados. Adicionalmente, a seleção dos fatores condicionantes é suscetível à subjetividade, apesar do uso do VIF para minimizar multicolinearidade. A limitação do inventário de escorregamentos também pode impactar a robustez do modelo.

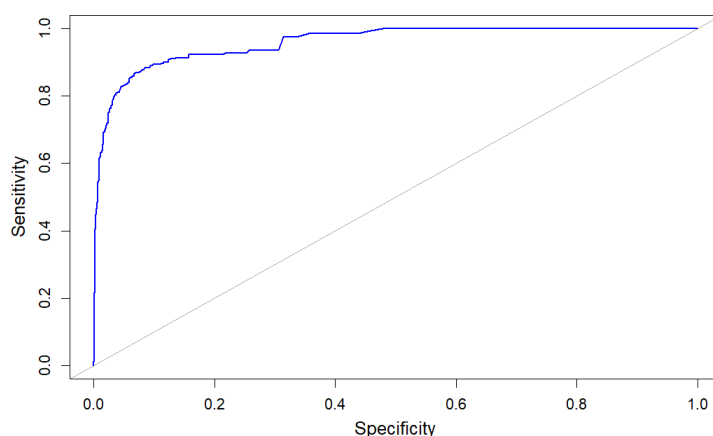


Figura 7. Curva ROC do modelo Random Forest, com $AUC = 0,94$. A curva representa a capacidade discriminante do modelo, com melhor desempenho quanto mais próxima do canto superior esquerdo.

4 CONCLUSÃO

Este estudo demonstrou a eficácia do modelo Random Forest na predição da suscetibilidade a escorregamentos ao longo de um trecho da Estrada de Ferro Carajás (EFC), localizado no estado do Maranhão. A aplicação do algoritmo, combinada com a análise de 12 variáveis ambientais e morfométricas, permitiu identificar áreas com maior propensão à instabilidade superficial. O desempenho estatístico do modelo foi considerado elevado, com $AUC = 0,94$, indicando alta capacidade discriminativa entre áreas com e sem ocorrência de escorregamentos. A elevada correlação entre os escorregamentos observados e as classes de maior suscetibilidade reforça a confiabilidade do mapeamento produzido.

A metodologia adotada contemplou rigorosas etapas de tratamento e validação dos dados, incluindo a análise de multicolinearidade, balanceamento das amostras e validação cruzada estratificada. Fatores como declividade, NDVI, precipitação média anual e distância a falhas estruturais apresentaram maior importância preditiva. Ainda que o modelo tenha alcançado bons resultados, algumas limitações foram observadas, como a heterogeneidade nas escalas dos dados de entrada e a limitação do inventário de escorregamentos utilizado no treinamento. Esses aspectos podem afetar a precisão espacial e a representatividade dos resultados.

Apesar dessas limitações, a abordagem adotada demonstrou grande potencial como ferramenta de apoio à gestão de riscos geotécnicos em corredores de infraestrutura lineares. O mapeamento de suscetibilidade gerado fornece subsídios valiosos para o planejamento de intervenções preventivas, a priorização de monitoramentos e a elaboração de estratégias de mitigação de riscos. Recomenda-se, para estudos futuros, a ampliação da base de dados históricos e o uso de modelos complementares que incorporem variabilidade temporal, visando ao aprimoramento contínuo da análise de áreas suscetíveis a movimentos de massa.

AGRADECIMENTOS



Agradecemos ao GeoInfra da Escola Politécnica da USP pelo suporte técnico e institucional essencial ao desenvolvimento desta pesquisa.

REFERÊNCIAS BIBLIOGRÁFICAS

- Bonini, J. E., Parenti, C., Grassi, F., Mancini, F., Vieira, B. C., & Soldati, M. (2025). Assessing landslide susceptibility and dynamics at cultural heritage sites by integrating machine learning techniques and persistent scatterer interferometry. *Geomorphology*, 469, 109522. <https://doi.org/10.1016/j.geomorph.2024.109522>
- Breiman L. 2001. Random forests. *Mach. Learn.* 45(1):5–32.
- Cascini, L. (2008). Applicability of landslide susceptibility and hazard zoning at different scales. *Engineering Geology*, 102(3–4), 164–177. <https://doi.org/10.1016/j.enggeo.2008.03.016>
- Gorsevski, P. V., Gessler, P. E., Boll, J., Elliot, W. J., & Foltz, R. B. (2006). Spatially and temporally distributed modeling of landslide susceptibility. *Geomorphology*, 80(3–4), 178–198. <https://doi.org/10.1016/j.geomorph.2006.02.011>
- ESRI, 2023. *ArcGIS Pro* [Versão 3.2.0]. Environmental Systems Research Institute, Inc., Redlands, CA.
- Fell, R., Corominas, J., Bonnard, C., Cascini, L., Leroi, E., & Savage, W. Z. (2008). Guidelines for landslide susceptibility, hazard and risk zoning for land use planning. *Engineering Geology*, 102(3–4), 85–98. <https://doi.org/10.1016/j.enggeo.2008.03.022>
- Florinsky, Igor, 2011. *Digital Terrain Analysis in Soil Science and Geology*, first ed. Academic Press, ISBN-13: 978-0123850362.
- Huntington JL, Hegewisch KC, Daudert B, Morton CG, Abatzoglou JT, McEvoy DJ, Erickson T (2017) Climate Engine: Cloud Computing and Visualization of Climate and Remote Sensing Data for Advanced Natural Resource Monitoring and Process Understanding. *Bulletin of the American Meteorological Society* 98:2397–2410. <https://doi.org/10.1175/BAMS-D-15-00324.1>
- JAXA, Japan Aerospace Exploration Agency, 2021. *ALOS PALSAR Radiometrically Terrain Corrected (RTC) Digital Elevation Model (12.5m resolution)*. Alaska Satellite Facility, NASA DAAC. Disponível em: <https://asf.alaska.edu>.
- Kim JC, Lee S, Jung HS, Lee S. 2018. Landslide susceptibility mapping using random forest and boosted tree models in Pyeong-Chang, Korea. *Geocarto Int.* 33(9):1000–1015.
- R Core Team (2025). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. <<https://www.R-project.org/>>
- Pandey, V. K., Sharma, K. K., Pourghasemi, H. R., & Bandooni, S. K. (2019). Sedimentological characteristics and application of machine learning techniques for landslide susceptibility modelling along the highway corridor Nahan to Rajgarh (Himachal Pradesh), India. *Catena*, 182, 104150. <https://doi.org/10.1016/j.catena.2019.104150>
- Reichenbach, P., Rossi, M., Malamud, B. D., Mihir, M., & Guzzetti, F. (2018). A review of statistically-based landslide susceptibility models. *Earth-Science Reviews*, 180, 60–91. <https://doi.org/10.1016/j.earscirev.2018.03.001>
- Rodrigues de Sousa, M.T., Silva, F.P., Goulart Fiscina, L.F., Cordeiro, W.B., Futai, M.M., Analysis of land use and coverage and its relationships with Declividade instability events on the Carajás Railway: The case of the Pindaré River Basin in the Brazilian Amazon biome, *Journal of South American Earth Sciences*, <https://doi.org/10.1016/j.jsames.2025.105373>
- Santos, E. D. S., Pinheiro, H. S. K., & Gallo Junior, H. (2021). Height Above the Nearest Drainage to Predict Flooding Areas in São Luiz do Paraitinga, São Paulo. *Floresta e Ambiente*, 28(2). <https://doi.org/10.1590/2179-8087-floram-2020-0070>
- Sørensen, R., Zinko, U., Seibert, J., 2006. On the calculation of the topographic wetness index: evaluation of different methods based on field observations. *Hydrol. Earth Syst. Sci.* 10, 101–112