UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA
Programa de Pós-Graduação em Engenharia Urbana

TATIANE FERREIRA OLIVATTO

DINÂMICA DA EROSÃO LINEAR EM ÁREAS URBANAS: INFLUÊNCIA DAS MUDANÇAS DE COBERTURA DO SOLO NAS ÁREAS DE CONTRIBUIÇÃO HÍDRICA

São Carlos-SP
2025

TATIANE FERREIRA OLIVATTO

DINÂMICA DA EROSÃO LINEAR EM ÁREAS URBANAS: INFLUÊNCIA DAS MUDANÇAS DE COBERTURA DO SOLO NAS ÁREAS DE CONTRIBUIÇÃO HÍDRICA UTILIZANDO ALGORITMO DE CLASSIFICAÇÃO INTERPRETÁVEL

Tese apresentada ao Programa de Pós-graduação em Engenharia Urbana, da Universidade Federal de São Carlos, para obtenção do título de Doutora em Engenharia Urbana. Área de concentração:

Orientador: Prof. Dr. José Augusto Di Lollo
Coorientação: Prof. Dr. Orlando Mauricio Quiroz Londoño
Financiamento: CAPES

São Carlos-SP
2025

Olivatto, Tatiane Ferreira

Dinâmica da erosão linear em áreas urbanas: influência das mudanças de cobertura do solo nas áreas de contribuição hídrica utilizando algoritmo de classificação interpretável / Tatiane Ferreira Olivatto -- 2025.
149f.

Tese de Doutorado - Universidade Federal de São Carlos, campus São Carlos, São Carlos
Orientador (a): José Augusto Di Lollo
Banca Examinadora: Rafael de Paula Garcia, Vandoir Bourscheidt, Vitor Eduardo Molina Junior, Fabio Leandro da Silva
Bibliografia

1. Engenharia urbana. 2. Processos erosivos. 3. Aprendizado de máquina. I. Olivatto, Tatiane Ferreira. II. Título.

Ficha catalográfica desenvolvida pela Secretaria Geral de Informática (SIn)

DADOS FORNECIDOS PELO AUTOR

Bibliotecário responsável: Arildo Martins - CRB/8 7180

UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Ciências Exatas e de Tecnologia
Programa de Pós-Graduação em Engenharia Urbana

Folha de Aprovação

Defesa de Tese de Doutorado da candidata Tatiane Ferreira Olivatto, realizada em 19/11/2025.

Comissão Julgadora:

Prof. Dr. José Augusto Di Lollo (UNESP)

Prof. Dr. Rafael de Paula Garcia (UFV)

Prof. Dr. Vandoir Bourscheidt (UFSCar)

Prof. Dr. Vitor Eduardo Molina Junior (UNICAMP)

Prof. Dr. Fabio Leandro da Silva (UFSCar)

O Relatório de Defesa assinado pelos membros da Comissão Julgadora encontra-se arquivado junto ao Programa de Pós-Graduação em Engenharia Urbana.

AGRADECIMENTOS

Em primeiro lugar, ao Prof. Dr. José Augusto Di Lollo, pela orientação, dedicação e pelos ensinamentos que transcendem os limites da Universidade. Meu sincero agradecimento por acreditar em meu potencial e pela motivação fundamental nos momentos de desânimo. Ao Prof. Dr. Orlando Mauricio Quiroz Londoño, coorientador deste trabalho, pela disponibilidade, pelas contribuições técnicas e pelas valiosas reflexões que enriqueceram esta pesquisa ao longo de seu desenvolvimento.

Aos membros da banca examinadora, Prof. Dr. Rafael de Paula Garcia, Prof. Dr. Vandoir Bourscheidt, Prof. Dr. Vitor Eduardo Molina Junior e Prof. Dr. Fabio Leandro Da Silva, pela criteriosa leitura do manuscrito, pelas importantes sugestões e pelo tempo dedicado à avaliação deste trabalho, que contribuíram significativamente para seu aprimoramento.

Aos professores e colaboradores do PPGEU, em especial à secretária Lívia Francisco Arantes de Souza - por ser sempre prestativa e atenciosa – e à docente Elza Luli Miyazaka, pela valiosa parceria profissional e pela amizade que construímos ao longo desta caminhada.

Aos colegas do grupo de pesquisa "Cidades e Pessoas: Conectadas", pelo acolhimento e pelos ricos aprendizados compartilhados. Um agradecimento especial às colegas Priscila Kauana Barelli Forcel e Júlia Neves Andrade, pelo companheirismo e pela colaboração sempre tão presentes.

À minha família, pelo alicerce incondicional. Ao meu marido, Wilson, pela paciência nos momentos de dificuldade com as ferramentas computacionais, e à minha irmã, Talita, por todo suporte emocional e revisão textual.

Por fim, ao apoio financeiro concedido pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) - Código de Financiamento 001 (Nº do Processo 88887.658530/2021-00).

We do not learn from experience.
We learn from reflecting on experience.
(John Dewey, 1938)

RESUMO

Os processos erosivos lineares ocorrem em função de diversos fatores naturais, porém, são acelerados por fatores antrópicos, como as mudanças de uso do solo. A modelagem espacial de processos erosivos é amplamente utilizada no contexto rural para prever perdas e prevenir prejuízos e, no contexto urbano, para embasar o planejamento e gestão do território. Atualmente, as técnicas de modelagem mais difundidas focam em áreas rurais, contemplando principalmente processos erosivos laminares. Os estudos contemplando áreas urbanas e periurbanas, com foco nas erosões lineares, são escassos. Em ambos os casos, o fator de uso e cobertura do solo, quando incorporado aos estudos, não considera dinâmicas temporais de mudanças e é analisado sem considerar a área de contribuição para o fenômeno. Reconhecendo esta lacuna, o objetivo principal da presente tese foi investigar o efeito das transições de cobertura da área de contribuição hídrica no desencadeamento de erosões lineares em áreas urbanas e periurbanas. Para atingir este fim, a pesquisa foi estruturada em três artigos. Fundamentando a hipótese desta tese, o primeiro artigo demonstrou que a variabilidade do fenômeno é predominantemente explicada por variáveis antrópicas (uso e cobertura do solo), as quais tiveram um impacto superior ao de fatores naturais. Por meio de uma revisão de literatura, o segundo artigo identificou que a amostragem de instâncias negativas e a delimitação de suas áreas de influência emergem como aspectos metodológicos críticos para a robustez de modelos de aprendizado de máquina aplicados à análise de processos erosivos. O terceiro artigo investigou os padrões de mudança na cobertura do solo com maior influência na ocorrência de erosão. Para isso, foi realizada uma análise comparativa que incorporou transições de uso do solo em um intervalo de vinte anos (1990-2010), adotando duas abordagens espaciais distintas: (i) um *buffer* de 100 m e (ii) a área de contribuição hídrica de cada feição. Os modelos de árvore de decisão revelaram que as transições mais importantes para a classificação foram aquelas relacionadas a usos agropecuários que se mantiveram estáveis ou se converteram em áreas não vegetadas. A comparação entre as abordagens demonstrou que a área de contribuição hídrica foi mais eficaz na identificação de casos positivos de erosão, sendo a variável de áreas de floresta preservada a mais relevante em seu modelo. Contudo, a estratégia do *buffer* proporcionou um melhor equilíbrio preditivo geral entre as classes. De forma geral, os resultados têm potencial de auxiliar no entendimento e predição de processos erosivos lineares em áreas urbanas, contribuindo no meio técnico-científico e para a sociedade como um todo no sentido de auxiliar na prevenção de danos ambientais, sociais e econômicos.

Palavras-chave: processos erosivos; uso e cobertura do solo; aprendizado de máquina; árvore de decisão; CART.

ABSTRACT

Linear erosive processes result from various natural factors but are accelerated by anthropogenic factors, such as land use changes. The spatial modeling of erosive processes is widely used in rural contexts to predict losses and prevent damages, and in urban contexts to support territorial planning and management. Currently, the most widespread modeling techniques focus on rural areas, primarily addressing sheet erosion. Studies encompassing urban and peri-urban areas, with a focus on linear erosion, are scarce. In both cases, when incorporated into studies, the land use and land cover factor often fails to consider temporal dynamics of change and is analyzed without accounting for the contributing area to the phenomenon. Acknowledging this gap, the main objective of this thesis was to investigate the effect of land cover transitions within the hydrological contributing area on the triggering of linear erosion in urban and peri-urban areas. To achieve this aim, the research was structured into three articles. Laying the groundwork for this thesis's hypothesis, the first article demonstrated that the variability of the phenomenon is predominantly explained by anthropogenic variables (land use and land cover), which had a greater impact than natural factors. Through a literature review, the second article identified that the sampling of negative instances and the delimitation of their areas of influence emerge as critical methodological aspects for the robustness of machine learning models applied to the analysis of erosive processes. The third article investigated the patterns of land cover change with the greatest influence on the occurrence of erosion. For this purpose, a comparative analysis was conducted, incorporating land use transitions over a twenty-year interval (1990-2010) and adopting two distinct spatial approaches: (i) a 100 m *buffer* and (ii) the hydrological contributing area of each feature. Decision tree models revealed that the most important transitions for classification were those related to agricultural uses that remained stable or converted to non-vegetated areas. The comparison between approaches demonstrated that the hydrological contributing area was more effective in identifying positive cases of erosion, with the "preserved forest areas" variable being the most relevant in its model. However, the *buffer* strategy provided a better overall predictive balance between the classes. In summary, the results have the potential to aid in the understanding and prediction of linear erosive processes in urban areas, contributing to the technical-scientific field and to society as a whole by assisting in the prevention of environmental, social, and economic damages.

Keywords: erosive processes; land use and land cover; machine learning; decision tree; CART.

SUMÁRIO

1 APRESENTAÇÃO

Há um consenso entre pesquisadores acerca do impacto significativo que as atividades antrópicas têm no desencadeamento e aceleração de processos erosivos (Guo *et al.*, 2019; Poesen *et al.*, 2003; Dos Santos *et al.*, 2017). Historicamente, muitos pesquisadores focam em investigações relacionadas a atividades que interferem notadamente nesses processos, como erosões laminares em áreas de plantio e processos diversos em áreas de mineração (Djoukbala *et al.*, 2019; Nearing *et al.*, 1989; Tarodiya; Levy, 2021; Wijesundara; Abeysingha; Dissanayake, 2018; Williams, 1975; Williams; Nicks; Arnold, 1985; Wischmeier; Smith, 1965, 1978). Já no que tange aos processos erosivos lineares, entendidos como um processo de desprendimento e arrastamento de partículas de solo causado pela ação da água que, frequentemente, se materializam na forma de ravinas e voçorocas, verificam-se diferentes dinâmicas em áreas rurais e urbanas.

Em áreas rurais, a remoção da cobertura vegetal natural para a implantação de áreas de cultivo e/ou pastagem resulta na redução — e, em alguns casos, na eliminação — do serviço ecossistêmico de interceptação da água da chuva desempenhado pela vegetação. Esse processo compromete a infiltração hídrica no solo, intensifica o escoamento superficial e favorece o desenvolvimento de sulcos, ravinas e voçorocas (Castillo; Gómez, 2016; Reece *et al.*, 2023). De forma geral, os impactos associados incluem a degradação do solo, perda de terras aráveis e aumento da sedimentação em corpos d'água, contribuindo para desequilíbrios ecológicos mais amplos, com reflexos socioeconômicos que se manifestam na diminuição da produtividade agrícola, na vulnerabilidade de comunidades dependentes do uso do solo e na elevação dos custos associados à mitigação e à recuperação de áreas degradadas (Junior *et al.*, 2010; Telles; Guimarães; Dechen, 2011).

Em áreas urbanas este processo também ocorre, contudo, a relação de mudança e impacto podem ser mais intensas, uma vez que o fenômeno de urbanização promove a impermeabilização de cerca de 80% das áreas (Brasil, 1979). No caso das áreas urbanas, o regime de interceptação de água se dá na atuação dos indivíduos de arborização urbana, nas áreas verdes e nos telhados das construções. O regime de infiltração também se traduz na atuação dos indivíduos de arborização urbana e nas áreas verdes, além da capacidade de engolimento das bocas de lobo do sistema de drenagem, quando existente (Yoo *et al.*, 2021).

Contudo, o escoamento superficial configura-se como o principal agente na intensificação da relação entre as transformações do uso e ocupação do solo e o desencadeamento de processos erosivos em áreas urbanas, uma vez que a impermeabilização modifica substancialmente os componentes do ciclo hidrológico, especialmente a infiltração, a evapotranspiração e a geração de escoamento. Nessas áreas, o escoamento superficial e o escoamento subterrâneo associado ao sistema de drenagem ocorrem quase simultaneamente ao início da precipitação, reduzindo significativamente o tempo de concentração da bacia hidrográfica. Essa maior concentração de água em menor intervalo de tempo impacta diretamente o desencadeamento de processos erosivos lineares. (Kuchkarova; Achilova, 2021; Yoo *et al.*, 2021). Particularmente em contextos urbanos, onde a maior parte da população reside atualmente, a erosão linear representa uma ameaça direta às infraestruturas críticas, podendo resultar em obstruções, inundações e comprometimento da eficácia da drenagem, colocando em risco a segurança pública e minando a funcionalidade das estruturas urbanas (Carvalho Junior *et al.*, 2010).

A erosão linear representa um desafio significativo para o desenvolvimento sustentável, intersectando-se com vários Objetivos de Desenvolvimento Sustentável (ODS) estabelecidos pelas Nações Unidas, os quais fornecem um arcabouço essencial para abordar questões globais e promover o progresso econômico, social e ambiental (ONU, 2015). Em particular, a relação entre a erosão linear e os ODS ganha destaque em áreas urbanas, onde os impactos da erosão podem ser exacerbados devido à concentração populacional e à intensificação das atividades humanas.

Ao explorar as interconexões diretas entre a erosão linear e os ODS, destacam-se os ODS 11 e 13, com as premissas de “tornar as cidades e os assentamentos humanos inclusivos, seguros, resilientes e sustentáveis” e “tomar medidas urgentes para combater a mudança climática e seus impactos” (ONU, 2015). Este último, em especial, tem se destacado recentemente devido ao aumento da ocorrência de eventos climáticos extremos. No caso específico das erosões, precipitações mais frequentes e mais intensas (Velastegui-Montoya *et al.*, 2022).

De forma indireta, também é crucial examinar como outros objetivos do desenvolvimento sustentável, como a erradicação da pobreza (ODS 1), a promoção da segurança alimentar (ODS 2), a garantia à vida saudável e bem-estar (ODS 3), a garantia de água limpa e saneamento (ODS 6), a detenção da degradação da terra (ODS 15), entre outros, estão intrinsecamente ligados à mitigação e adaptação aos impactos da erosão linear em áreas urbanas (McElwee *et al.*, 2020; ONU, 2015).

Neste contexto, torna-se evidente que os impactos da erosão linear transcendem o ambiente natural, exercendo uma influência significativa sobre as comunidades humanas. Essa realidade é ainda mais desafiadora para populações vulneráveis, cuja falta de recursos e resiliência as torna particularmente suscetíveis aos efeitos adversos da erosão, acentuando as desigualdades sociais e ressaltando a urgência de abordagens inclusivas e equitativas (Saha *et al.*, 2021).

Diante do exposto, fica evidente a necessidade de uma análise mais aprofundada sobre a dinâmica espacial dos processos erosivos lineares em áreas urbanas. Nesse sentido, esta tese se propõe a investigar a importância dos atributos de mudança de uso do solo na área de contribuição hídrica para o desencadeamento desses processos. Ao compreendermos melhor como tais atributos influenciam a intensidade e a distribuição da erosão linear nas áreas urbanas, poderemos desenvolver estratégias mais eficazes de mitigação e adaptação. Além disso, essa abordagem permitirá subsidiar as políticas de desenvolvimento urbano de forma mais sustentável, alinhando nossos esforços com os ODS supracitados, contribuindo para a construção de comunidades mais resilientes e equitativas.

1.1 QUESTÃO DA PESQUISA E HIPÓTESE

O monitoramento e a predição de processos erosivos em áreas urbanas desempenham um papel crucial na gestão ambiental e no planejamento territorial, especialmente no contexto brasileiro, onde as características climáticas variadas, a complexa formação geológica, o relevo diversificado e os padrões de urbanização intensiva contribuem para a vulnerabilidade dessas áreas (Bonzanini; Lupinacci; Stefanuto, 2022; Ferreira; Martins, 2021; Magalhães *et al.*, 2023). Entre os diversos fatores que desencadeiam esses processos, as mudanças na cobertura do solo, primordialmente relacionadas à atividade humana, emergem como uma preocupação particular (Chaplot *et al.*, 2005; Endres *et al.*, 2006; Guo *et al.*, 2019). É amplamente reconhecido que as alterações causadas pelo homem, como a expansão urbana desordenada e a remoção da vegetação natural, aceleram os processos erosivos, tornando essencial entender e prever esses impactos.

Diante desse cenário, a presente tese busca responder ao problema central de pesquisa: **Como os fenômenos de erosão linear em áreas urbanas e periurbanas são afetados pelas mudanças de cobertura do solo?**. Partindo dessa questão, desenvolveu-se a hipótese de que as mudanças na cobertura do solo são os principais desencadeadores dos processos erosivos em áreas urbanas, e que essas mudanças devem ser cuidadosamente estudadas e incorporadas aos modelos de suscetibilidade à erosão, considerando não apenas a ocorrência local, mas toda a área de contribuição hídrica. Isso implica uma abordagem mais abrangente que não apenas avalia a situação atual da cobertura do solo, mas também considera suas mudanças ao longo do tempo e seu impacto potencial nos processos erosivos.

A pesquisa apresenta relevância científica e social ao considerar as mudanças na cobertura do solo como um fator-chave na erosão urbana e integrá-las à modelos de aprendizado de máquina, avançando o conhecimento existente e proporcionando percepções valiosas para a gestão sustentável do ambiente urbano. Ao incorporar variáveis antropogênicas de forma distribuída, ou seja, considerando a área de contribuição hídrica, a pesquisa oferece uma abordagem mais abrangente em comparação às mais difundidas na literatura, promovendo assim uma gestão mais eficaz e sustentável do ambiente urbano. Além disso, ao abordar aspectos metodológicos relacionados à amostragem de ocorrência e não ocorrência dos fenômenos erosivos, a pesquisa contribui para o aprimoramento das técnicas de modelagem, especialmente em um contexto onde o emprego de técnicas de inteligência artificial, como o aprendizado de máquina, está se tornando cada vez mais comum.

1.2 OBJETIVOS

Analisar a influência das mudanças de cobertura do solo na ocorrência de fenômenos de erosão linear em áreas urbanas, adotando a área de contribuição hídrica como unidade de análise. Visando atingir este objetivo primário, os objetivos específicos deste trabalho são:

I. Identificar e hierarquizar as variáveis antrópicas e naturais correlacionadas ao desencadeamento de processos erosivos lineares, determinando aquelas de maior impacto;

II. Mapear as estratégias de amostragem (em particular de instâncias negativas) utilizadas em estudos sobre erosão linear que empregam aprendizado de máquina, com foco no papel da área de contribuição hídrica; e

III. Operacionar os conhecimentos gerados nos objetivos anteriores para identificar os padrões de mudança de cobertura do solo com maior impacto na ocorrência de erosão linear em áreas urbanas.

1.3 ESTRUTURA DA TESE

A presente tese está estruturada em sete capítulos:

Capítulo 1 – Apresentação geral, questão de pesquisa e objetivos;

Capítulo 2 – Contextualização geral da problemática;

Capítulo 3 – Fundamentação Teórica;

Capítulo 4 – Artigo 1: "*Dynamics of linear erosion processes in urban areas: an approach based on GIS and CART algorithm*" (Dinâmica dos processos erosivos lineares em áreas urbanas: uma abordagem apoiada em SIG e algoritmo CART);

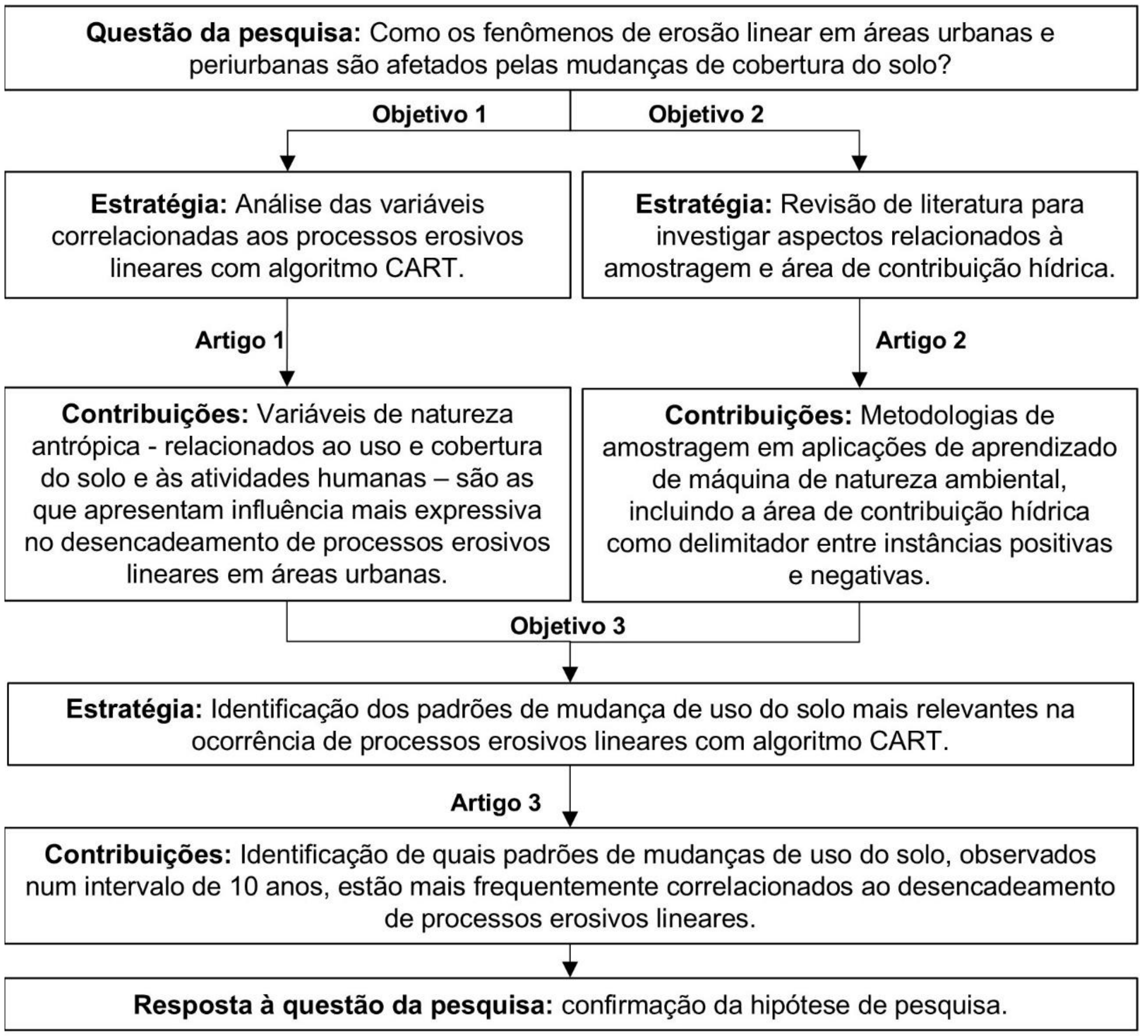
Capítulo 5 – Artigo 2: "*Sampling strategies for machine learning-based linear erosion studies: a review approaching contributing area*" (Estratégias de amostragem para estudos de erosão linear embasados em aprendizado de máquina: uma revisão abordando a área de contribuição);

Capítulo 6 – Artigo 3: "*Unveiling Land Use Drivers of Urban Gully Erosion: An Interpretable Machine Learning Approach*" (Desvendando os fatores de uso do solo na erosão linear urbana: uma abordagem de aprendizado de máquina interpretável); e

Capítulo 7 – Considerações finais.

A Figura 1 apresenta um esquema dos artigos a serem desenvolvidos para responder a cada objetivo específico proposto, além de mostrar as estratégias adotadas e contribuições esperadas em cada um deles.

Figura 1 – Esquema dos artigos da tese e relações com a hipótese de pesquisa



Fonte: Elaborado pela autora.

2 CONTEXTUALIZAÇÃO

2.1 EROSÃO DO SOLO

A erosão do solo consiste em um processo morfogenético responsável pela modelagem da superfície terrestre por meio da remoção e do transporte de partículas, atuando de forma contínua na dinâmica das paisagens (Christofoletti, 1980). Esse processo pode ser intensificado por diferentes agentes naturais, especialmente a água e o vento, sendo a ação da água considerada o principal fator erosivo em ambientes continentais (Bertoni; Lombardi Neto, 1990). Morgan (2009) descreve a erosão como um fenômeno composto por três etapas sucessivas — destacamento, transporte e deposição dos sedimentos — ressaltando sua importância na evolução das formas de relevo.

De modo complementar, Bazzoffi *et al.* (1993) diferenciam a erosão natural, também denominada geológica, que ocorre em áreas não perturbadas e participa da construção das paisagens ao longo do tempo geológico, da erosão acelerada, associada às intervenções antrópicas e responsável por perdas rápidas de solo e intensificação da degradação ambiental. Esses autores destacam ainda o caráter ambíguo do processo erosivo, que pode atuar de forma destrutiva nas áreas de origem dos sedimentos, ao mesmo tempo em que contribui para a formação de ambientes deposicionais férteis, como deltas e planícies aluviais.

A erosão do solo manifesta-se sob diferentes formas, podendo ser classificada conforme seus mecanismos e padrões de ocorrência. A erosão laminar caracteriza-se pela remoção uniforme de camadas superficiais do solo e a erosão linear resulta da concentração do escoamento superficial (Guerra; Jorge, 2018). Outros tipos incluem a erosão intersulcos, a erosão por tubulação (piping erosion - relacionada à remoção subsuperficial de material), e os movimentos de massa, como os deslizamentos.

No contexto da erosão linear, torna-se fundamental distinguir seus diferentes estágios evolutivos, embora essa diferenciação nem sempre seja simples, devido à diversidade de critérios adotados internacionalmente para a classificação dessas feições (Vanmaercke *et al.*, 2016). Estudos de abrangência global realizados por Dube *et al.* (2020) e Kuhn *et al.* (2023) propuseram a identificação de três estágios principais da erosão linear: sulcos (*rills*), ravinas (*ravines*) voçorocas (*gullies*).

Segundo esses autores, os sulcos representam o estágio inicial da erosão por escoamento concentrado, formando canais rasos com profundidade inferior a 0,25 m. Desenvolvem-se em encostas a partir do fluxo superficial da água, que promove o destacamento e o transporte de partículas do solo, podendo ser facilmente eliminados por práticas agrícolas ou pela deposição natural de sedimentos. As ravinas correspondem a um estágio mais avançado, apresentando profundidades superiores a 0,25 m e, na classificação brasileira, frequentemente exibem seção transversal em forma de “U”. Sua formação está associada ao aumento da energia do escoamento concentrado, resultando em perdas significativas de solo e na ramificação de canais.

As voçorocas constituem feições ainda mais profundas e extensas, geralmente com seção transversal em forma de “V” e profundidades superiores a 0,50 m. Essas estruturas frequentemente se conectam ao lençol freático, favorecendo a erosão lateral e intensificando os processos de degradação da paisagem. Em razão de sua magnitude e permanência, as voçorocas representam um estágio avançado da erosão linear, promovendo alterações significativas no relevo e nos ecossistemas. Em contextos específicos, especialmente em regiões áridas e semiáridas, essas feições exercem papel central na dinâmica da paisagem, uma vez que a escassez de cobertura vegetal e a ocorrência de chuvas irregulares e intensas favorecem a formação de canais profundos associados à degradação severa dos solos (Arabameri *et al.*, 2020).

2.2 EROSÃO LINEAR EM ÁREAS URBANAS

As abordagens contemporâneas sobre os processos erosivos, especialmente no contexto do Antropoceno, têm enfatizado sua natureza complexa e multidimensional. Nessa perspectiva, a erosão linear é compreendida não apenas como um fenômeno geomorfológico, mas também como um processo de degradação ambiental resultante da interação entre dinâmicas naturais e ações antrópicas, produzindo impactos ambientais, econômicos e sociais em múltiplas escalas espaço-temporais (Poesen, 2018). Essa concepção reforça a necessidade de análises integradas que considerem simultaneamente os controles físicos da paisagem e as transformações induzidas pelo uso e ocupação do solo.

O desenvolvimento da erosão linear é influenciado por diversos fatores, entre os quais se destacam precipitação, características geomorfométricas, propriedades dos solos e dos materiais de origem; e uso do solo (Guerra, 2012). A intensidade e a frequência das chuvas exercem papel fundamental na configuração dos padrões erosivos, uma vez que eventos pluviométricos concentrados potencializam o escoamento superficial, favorecendo a formação de sulcos e voçorocas. A topografia da paisagem, expressa principalmente pela declividade e pelo comprimento das vertentes, controla a velocidade e a concentração do fluxo hídrico, condicionando a energia disponível para o destacamento e o transporte de partículas. As propriedades dos solos e dos substratos, como textura, estrutura e teor de matéria orgânica, influenciam sua suscetibilidade à erosão, sendo solos pouco estruturados ou arenosos mais propensos ao desagregamento e à mobilização. Adicionalmente, práticas de uso da terra, como agricultura intensiva, desmatamento e urbanização, tendem a agravar esses processos ao reduzir a cobertura vegetal e modificar os padrões naturais de drenagem (Chalise *et al.*, 2019; Falcão *et al.*, 2020; Guerra, 2012).

No espaço urbano, a erosão linear adquire especial relevância em função das profundas alterações impostas à dinâmica hidrológica superficial. A expansão urbana, associada à remoção da vegetação e ao aumento das superfícies impermeáveis, intensifica o escoamento superficial e favorece a concentração dos fluxos, promovendo a formação de voçorocas e ravinas. Esses processos comprometem infraestruturas essenciais, como vias de circulação e sistemas de drenagem, ampliando os riscos à segurança da população e à funcionalidade dos serviços urbanos (Polovina *et al.*, 2021). A centralidade dessa problemática é reforçada pelo fato de que a maior parte da população mundial reside atualmente em áreas urbanizadas, o que torna imprescindível o aprofundamento do conhecimento sobre as formas de erosão linear e seus impactos socioeconômicos e ambientais (Moreno-Monroy *et al.*, 2020; Kuhn *et al.*, 2023).

A literatura aponta de forma consistente o papel das superfícies impermeabilizadas e sistemas de drenagem pluvial deficientes como importantes desencadeadores da erosão linear nas cidades. Ruas pavimentadas, taludes de cortes e aterros, valetas e pontos de lançamento concentrado de águas pluviais frequentemente atuam como caminhos preferenciais de escoamento, favorecendo a iniciação de cabeceiras erosivas e o aprofundamento dos canais, sobretudo em áreas com elevada declividade ou convergência topográfica (Gudiño-Elizondo *et al.*, 2018; Rahmati *et al.*, 2022; Imwangana *et al.*, 2025). Evidências provenientes de cidades do Sul Global indicam que mesmo áreas com suscetibilidade natural moderada podem desenvolver processos erosivos severos quando a hidrologia urbana é intensamente modificada (Zolezzi *et al.*, 2018; Mawe *et al.*, 2024).

De forma geral, a erosão linear em áreas urbanas reflete um descompasso estrutural entre as práticas de uso e ocupação do solo e a capacidade física da paisagem em absorver e redistribuir os fluxos hídricos. Estudos recentes convergem ao indicar que a mitigação efetiva desses processos depende menos de intervenções pontuais e mais da adoção de estratégias integradas (Frankl *et al.*, 2020; Mawe *et al.*, 2025).

Portanto, assumindo que estes processos estão fortemente associados à reorganização dos fluxos hidrológicos induzida pelo uso e ocupação do solo, torna-se imprescindível reconhecer as áreas que efetivamente contribuem para a concentração do escoamento e para o desencadeamento das feições erosivas. A análise dos processos, portanto, exige a delimitação das regiões que funcionam como fontes desses fluxos, permitindo relacionar padrões de urbanização, infraestrutura de drenagem e suscetibilidade geomorfológica. Essa perspectiva conduz à incorporação do conceito de área de contribuição como elemento estruturante da análise espacial dos processos erosivos.

2.3 ÁREA DE CONTRIBUIÇÃO

O conceito de área de contribuição tem origem consolidada na hidrologia, onde designa a porção do território que drena para um ponto específico, contribuindo com o escoamento superficial ou subterrâneo que converge em direção a um curso d’água ou sistema receptor (Rieger, 1998; Mackay; Band, 1998). Essa concepção hidrológica, associada a métodos de modelagem topográfica e à análise de bacias hidrográficas, constituiu a base para a disseminação posterior do termo em outros domínios das ciências ambientais e do planejamento territorial.

Com o avanço das geotecnologias e o aumento da complexidade dos sistemas socioambientais, a noção de área de contribuição passou a ser empregada de forma mais ampla, extrapolando o contexto hidrológico e assumindo novas interpretações conceituais. Em sentido geral, área de contribuição refere-se à região espacial que gera, fornece ou origina um fluxo, recurso ou efeito em direção a um ponto, infraestrutura, ecossistema ou processo territorial (Guan; Sillanpää; Koivusalo, 2015; Mortoja; Yigitcanlar; Mayere, 2020).

No campo do planejamento territorial e ambiental, a área de contribuição passou a ser entendida como um instrumento analítico para identificar as origens espaciais dos fluxos físicos, ecológicos ou socioeconômicos que estruturam o território. Essa perspectiva é particularmente relevante quando se busca compreender a relação entre as formas de ocupação e os processos ambientais, como no caso de áreas que contribuem para recarga hídrica, dispersão de poluentes, ilhas de calor, mobilidade pendular, pressão sobre serviços públicos ou provisão de serviços ecossistêmicos (Cattaneo; Nelson; McMenomy, 2021; Wang *et al.*, 2023). Em todos esses contextos, a área de contribuição expressa a dimensão geradora de um fenômeno, representando a dimensão espacial das interações.

Apesar de sua proximidade terminológica, a área de contribuição difere conceitualmente da área de influência, ainda que ambas frequentemente sejam utilizadas como sinônimos em estudos urbanos e ambientais. Enquanto a área de contribuição corresponde à origem ou fonte de um fluxo ou efeito, a área de influência diz respeito à extensão ou alcance espacial onde os efeitos de determinado processo, estrutura ou intervenção são sentidos (Mortoja; Yigitcanlar; Mayere, 2020). Essa distinção é crucial, pois reflete diferentes perspectivas de análise — uma voltada à geração e outra à propagação de fenômenos territoriais.

A transposição do conceito de área de contribuição do domínio hidrológico para outros campos demandou adaptações metodológicas. Em hidrologia, sua delimitação é realizada a partir de Modelos Digitais de Elevação (MDE) e algoritmos de direção de fluxo (Rieger, 1998; Duvillier *et al.*, 2023). No contexto do planejamento territorial, novas abordagens têm sido incorporadas, como o uso de modelos de interação espacial, redes de mobilidade, clustering espacial, dados de sensoriamento remoto e aprendizado de máquina (ML, sigla do inglês *machine learning*) (Arribas-Bel; García-López; Viladecans-Marsal, 2021; Fang *et al.*, 2025). Essas técnicas permitem identificar áreas que “contribuem” para determinados fenômenos urbanos, ecológicos ou econômicos, de modo dinâmico e baseado em evidências empíricas.

Em aplicações urbanas e ambientais, a área de contribuição pode ser interpretada de diferentes maneiras. Em ecologia da paisagem, por exemplo, ela representa o conjunto de fragmentos que alimentam fluxos ecológicos de espécies ou energia (Foltête; Vuidel, 2017). Em planejamento de transportes, corresponde à área de origem de viagens associadas a uma estação de metrô ou terminal (Aprigliano *et al.*, 2024). No contexto de infraestrutura verde, o conceito se aplica à origem dos benefícios ambientais, como a vegetação que contribui para a redução de temperatura em áreas adjacentes (Semeraro *et al.*, 2021). Em todos os casos, trata-se de identificar o território que fornece uma contribuição mensurável para um processo maior, físico, biológico ou social.

A literatura recente destaca que não existe um método universal para a delimitação de áreas de contribuição, pois sua definição depende do sistema analisado, da escala espacial, da disponibilidade de dados e dos objetivos da análise (Chen *et al.*, 2024). Essa flexibilidade metodológica reflete a própria evolução do conceito, que se desloca de uma visão física e unidimensional para uma abordagem integrada e sistêmica, atualmente incorporada também em modelos de aprendizado de máquina e análises baseadas em dados ambientais (Olivatto; Di Lollo, 2025), alinhada às demandas contemporâneas do planejamento territorial e ambiental.

Em síntese, a área de contribuição representa uma categoria espacial fundamental para a compreensão das relações de causa e origem no território. Sua evolução do campo hidrológico para uma aplicação interdisciplinar revela o esforço em reconhecer os múltiplos fluxos — naturais, infraestruturais e sociais — que configuram a paisagem. Mais do que uma unidade de medida física, trata-se de um conceito analítico que articula fontes, processos e efeitos em diferentes escalas, oferecendo suporte teórico e metodológico ao planejamento voltado à sustentabilidade e à resiliência dos sistemas territoriais.

2.4 AVANÇOS NA ANÁLISE DE PROCESSOS EROSIVOS

A compreensão e o monitoramento dos processos erosivos lineares são fundamentais para o manejo sustentável do solo e para a preservação ambiental, uma vez que o entendimento de suas dinâmicas é condição necessária para a formulação de estratégias eficazes de prevenção e controle. A análise desses processos fornece subsídios essenciais à tomada de decisão e à gestão sustentável dos recursos naturais (Frankl *et al.*, 2020; Telles *et al.*, 2011).

Diversas estratégias têm sido empregadas para a avaliação da erosão do solo, incluindo métodos diretos, como a perfilometria e a coleta de sedimentos, e abordagens indiretas baseadas em estimativas e modelagens computacionais. Entre os modelos mais difundidos destaca-se a Equação Universal de Perda do Solo (USLE, sigla do inglês *Universal Soil Loss Equation*), desenvolvida por Wischmeier e Smith (1978), que estima a perda média anual de solo a partir de variáveis relacionadas à pluviometria, aos solos, à topografia e à cobertura da terra. Posteriormente, essa equação foi revisada, originando a RUSLE (*Revised* USLE), que manteve os fatores fundamentais do modelo original, incorporando procedimentos digitalizados e aprimoramentos na representação dos efeitos da precipitação e do escoamento superficial (Renard *et al.*, 1997). Em complemento, modelos como o *Sediment Delivery Ratio* (SDR) buscam estimar a produção, o transporte e a retenção de sedimentos, sendo disponibilizados em plataformas como o InVEST®, que reúne diferentes modelos espaciais aplicáveis a múltiplas escalas (Borselli *et al.*, 2008; Natural Capital Project, 2025).

Apesar de sua ampla aplicação, esses modelos são fortemente orientados à estimativa do transporte de sedimentos, o que nem sempre corresponde de forma direta às dinâmicas específicas da erosão linear em ambientes urbanos, nas quais a gênese e a evolução das feições erosivas estão frequentemente associadas à reorganização do escoamento superficial e às transformações antrópicas da paisagem.

O avanço das geotecnologias representou um marco nos estudos sobre erosão do solo, ao possibilitar a análise integrada de grandes volumes de dados espaciais. O sensoriamento remoto permite a aquisição de informações orbitais com elevada resolução espacial, espectral e temporal, enquanto os Sistemas de Informação Geográfica (SIG) viabilizam a integração, o tratamento e a análise desses dados em ambiente computacional (Ganasri; Ramesh, 2016). Estudos baseados em modelos empíricos associados a dados de sensoriamento remoto e SIG evidenciam o caráter dinâmico e interconectado da erosão em escala global (Borrelli *et al.*, 2020). Mais recentemente, a incorporação de veículos aéreos não tripulados (VANT) e de técnicas de inteligência artificial (IA) tem ampliado as possibilidades analíticas. Comparações entre estimativas volumétricas de perda de solo obtidas por modelos tradicionais e aquelas derivadas de dados de VANT ou de modelos baseados em IA indicam o elevado potencial dessas novas abordagens frente às técnicas convencionais (Giglio, 2025; Samarinas *et al.*, 2024).

Nesse sentido, a integração de técnicas de inteligência artificial tem promovido avanços significativos nos estudos sobre erosão linear, sobretudo por meio da aplicação de algoritmos de ML. Essas técnicas se destacam em função de sua capacidade de identificar padrões complexos em grandes volumes de dados e de produzir modelos preditivos com elevado desempenho. No campo das ciências ambientais, essas técnicas vêm sendo utilizadas para apoiar análises relacionadas à suscetibilidade a processos naturais, como escorregamentos, erosão hídrica e formação de voçorocas, além de aplicações voltadas à avaliação de riscos e ao planejamento territorial (Pourghasemi *et al*., 2017; Amiri *et al*., 2019).

Diversos estudos indicam que modelos preditivos baseados em ML têm superado modelos que utilizam modelos estatísticos bivariados tradicionais e a própria RUSLE (Garosi *et al.*, 2019; Pourghasemi *et al.*, 2020; Rahmati *et al.*, 2017; Lin; Zhao, 2022). Isso porque enquanto modelos clássicos dependem de pressupostos empíricos, os algoritmos de ML são capazes de identificar padrões complexos a partir dos próprios dados (Liu *et al.*, 2023). Além deste fato, o avanço dessas abordagens também está associado à crescente disponibilidade de bases de dados geoespaciais, oriundas de sensores remotos, modelos digitais de terreno e sistemas de informações geográficas (SIG), bem como ao aumento da capacidade computacional. Esse cenário tem favorecido a adoção de algoritmos capazes de lidar com relações não lineares e interações entre múltiplas variáveis explicativas, superando limitações frequentemente observadas em métodos estatísticos tradicionais (Pourghasemi *et al*., 2017).

Entretanto, embora os resultados obtidos com técnicas de ML sejam, em muitos casos, promissores, sua aplicação em estudos ambientais apresenta desafios específicos. Entre eles, destacam-se as dificuldades relacionadas à qualidade, à representatividade e à escala dos dados utilizados na modelagem, bem como à transferência de modelos entre diferentes contextos geográficos e ambientais

(Rahmati *et al.*, 2017). Esses fatores podem comprometer a robustez dos modelos e limitar sua capacidade de generalização.

Além disso, os processos ambientais são caracterizados por elevada variabilidade espacial e temporal, o que impõe restrições adicionais à aplicação direta de algoritmos desenvolvidos originalmente para outros domínios do conhecimento. A heterogeneidade dos sistemas naturais, associada à influência simultânea de fatores climáticos, geomorfológicos, pedológicos e antrópicos, exige abordagens metodológicas que considerem explicitamente essa complexidade (Amiri *et al.*, 2019). Isso reflete diretamente em limitações quanto generalização dos modelos. Resultados obtidos em determinadas áreas de estudo nem sempre podem ser transferidos diretamente para outros contextos geográficos. Além disso, a definição das variáveis explicativas e a forma de representação das feições erosivas no banco de dados influenciam significativamente o desempenho dos algoritmos (Rahmati *et al.*, 2017).

Nesse contexto, a literatura recente tem enfatizado a necessidade de integrar técnicas de aprendizado de máquina a fundamentos teóricos já consolidados, de modo a garantir que os modelos resultantes sejam não apenas estatisticamente consistentes, mas também coerentes do ponto de vista físico e ambiental (Pourghasemi *et al.*, 2017; Rahmati *et al.*, 2017). Assim, a aplicação dessas ferramentas deve ser acompanhada de uma análise crítica de seus pressupostos, limitações e implicações para a interpretação dos fenômenos estudados.

Outro ponto recorrente na literatura diz respeito à necessidade de interpretar os resultados dos modelos à luz do conhecimento técnico-científico já consolidado. Embora os algoritmos possam indicar a importância relativa dos fatores condicionantes, essa informação deve ser analisada criticamente, de modo a evitar conclusões dissociadas dos processos físicos que controlam a evolução das feições erosivas. Assim, a integração entre modelagem computacional e análise crítica especializada constitui um requisito fundamental para a confiabilidade dos estudos baseados em aprendizado de máquina (Pourghasemi *et al.*, 2017; Liu *et al.*, 2022).

3 FUNDAMENTAÇÃO TEÓRICA

3.1 APRENDIZADO DE MÁQUINA

O aprendizado de máquina (*machine learning* – ML) constitui um subcampo da inteligência artificial voltado ao desenvolvimento de métodos capazes de identificar padrões em dados e realizar previsões ou classificações a partir da experiência, sem que todas as regras de decisão sejam explicitamente programadas. Essas técnicas têm sido amplamente incorporadas a diferentes áreas do conhecimento, impulsionadas pelo aumento da disponibilidade de dados digitais e pelo avanço da capacidade computacional, que viabilizam a análise de conjuntos de dados extensos e complexos (Hastie; Tibshirani; Friedman, 2009; James *et al*., 2013).

De modo geral, os algoritmos de aprendizado de máquina podem ser organizados conforme o tipo de aprendizagem adotado. No aprendizado supervisionado, o modelo é treinado a partir de um conjunto de dados que inclui tanto as variáveis explicativas quanto a variável resposta, permitindo o ajuste de funções capazes de estimar classes ou valores desconhecidos. No aprendizado não supervisionado, por sua vez, não há rótulos previamente definidos, sendo o objetivo identificar estruturas latentes ou agrupamentos naturais nos dados. Já o aprendizado por reforço baseia-se na interação contínua entre o modelo e o ambiente, no qual decisões são aperfeiçoadas progressivamente a partir de mecanismos de recompensa (Hastie; Tibshirani; Friedman, 2009).

Além dessa distinção, os métodos de aprendizado de máquina também podem ser classificados segundo a natureza do problema a ser resolvido. Destacam-se, nesse sentido, os problemas de classificação e de regressão. Nos problemas de classificação, busca-se atribuir cada observação a uma categoria ou classe previamente definida, enquanto nos problemas de regressão o objetivo consiste em estimar valores contínuos a partir de um conjunto de variáveis explicativas. Há ainda abordagens associadas à análise de agrupamentos (*clustering*) e à redução de dimensionalidade, empregadas principalmente na exploração inicial dos dados e na identificação de padrões de similaridade entre as observações (Hastie; Tibshirani; Friedman, 2009; James *et al*., 2013).

O emprego dessas técnicas apresenta vantagens relevantes em comparação a métodos estatísticos tradicionais, sobretudo no que se refere à capacidade de lidar com relações não lineares e com interações complexas entre variáveis. Essa característica torna o aprendizado de máquina particularmente adequado para aplicações em sistemas naturais e socioambientais, nos quais múltiplos fatores atuam de forma simultânea e interdependente (Amiri *et al*., 2019; Pourghasemi *et al*., 2017).

Entretanto, a adoção de algoritmos de ML também impõe desafios metodológicos. O desempenho dos modelos depende fortemente da qualidade, da representatividade e da consistência dos dados utilizados no treinamento. Bases de dados incompletas ou enviesadas podem comprometer a robustez dos resultados e limitar sua capacidade de generalização. Ademais, muitos algoritmos operam como modelos de “caixa-preta”, dificultando a compreensão dos processos internos responsáveis pelas predições geradas (Molnar, 2022).

Outro ponto recorrente na literatura diz respeito à necessidade de interpretar os resultados dos modelos à luz do conhecimento técnico-científico já consolidado. Embora os algoritmos possam indicar a importância relativa das variáveis explicativas, essa informação deve ser analisada criticamente, de modo a evitar conclusões dissociadas dos processos teóricos que fundamentam o fenômeno investigado. Assim, a integração entre modelagem computacional e análise especializada constitui um requisito fundamental para a confiabilidade dos estudos baseados em aprendizado de máquina (Pourghasemi *et al*., 2017; Liu *et al*., 2022).

Nesse sentido, emerge o aprendizado de máquina explicável com o objetivo de viabilizar a aproximação entre os resultados obtidos por modelos computacionais e o conhecimento do especialista. As técnicas de explicabilidade fornecem instrumentos que permitem compreender como e em que medida as variáveis de entrada influenciam as predições dos algoritmos, favorecendo a interpretação dos resultados e sua confrontação com fundamentos teóricos previamente estabelecidos. Ao reduzir o caráter de “caixa-preta” dos modelos, a explicabilidade contribui para a validação conceitual das inferências produzidas e para a incorporação do julgamento técnico-científico no processo analítico (Doshi-Velez; Kim, 2017; Molnar, 2022).

Dessa forma, a aplicação de técnicas de aprendizado de máquina em pesquisas científicas representa uma abordagem promissora, desde que acompanhada de procedimentos rigorosos de validação, seleção criteriosa de variáveis e interpretação fundamentada no arcabouço teórico da área de estudo.

Esses cuidados são essenciais para assegurar que os modelos desenvolvidos reflitam adequadamente a complexidade dos sistemas analisados e possam subsidiar a produção de conhecimento científico confiável.

3.2 APRENDIZADO DE MÁQUINA EXPLICÁVEL

O uso crescente de modelos de aprendizado de máquina em aplicações científicas tem evidenciado a necessidade de compreender não apenas os resultados obtidos, mas também os mecanismos que conduzem às predições geradas. Muitos algoritmos amplamente empregados apresentam elevada complexidade estrutural, o que dificulta a interpretação direta de seu funcionamento interno e de suas relações entre variáveis de entrada e saída. Essa característica tem sido associada ao chamado problema da “caixa-preta”, no qual o modelo fornece respostas precisas, porém pouco transparentes do ponto de vista conceitual (Doshi-Velez; Kim, 2017; Molnar, 2022).

Nesse contexto, desenvolveu-se o campo do aprendizado de máquina explicável (*explainable artificial intelligence* – XAI), cujo objetivo central consiste em ampliar a transparência dos modelos e possibilitar a compreensão dos fatores que influenciam suas predições. A explicabilidade está associada à capacidade de um sistema fornecer informações inteligíveis sobre seu comportamento, permitindo que usuários humanos interpretem e avaliem criticamente os resultados produzidos. Em aplicações científicas, essa característica é particularmente relevante, uma vez que favorece a verificação da coerência entre as respostas do modelo e o conhecimento teórico já estabelecido na área de estudo (Doshi-Velez; Kim, 2017).

As abordagens de explicabilidade podem ser organizadas, de modo geral, em dois grupos principais: modelos intrinsecamente interpretáveis e métodos de interpretação *post hoc*. Os primeiros correspondem a algoritmos cuja própria estrutura possibilita a compreensão direta das relações entre variáveis, como regressões lineares e árvores de decisão simples. Já os métodos *post hoc* são aplicados após o treinamento do modelo, com o propósito de explicar seu comportamento sem alterar sua arquitetura original, sendo especialmente relevantes no caso de modelos mais complexos, como florestas aleatórias e redes neurais artificiais (Molnar, 2022).

Entre os principais objetivos das técnicas de XAI destaca-se a identificação da contribuição relativa das variáveis de entrada para as predições do modelo, permitindo avaliar quais fatores exercem maior influência sobre os resultados obtidos. Essa informação pode ser utilizada tanto para a validação conceitual dos modelos quanto para a detecção de possíveis inconsistências, vieses ou dependências espúrias nos dados utilizados no treinamento. Dessa forma, a explicabilidade contribui para aumentar a confiabilidade dos modelos e para reduzir o risco de interpretações inadequadas baseadas exclusivamente em métricas de desempenho (Doshi-Velez; Kim, 2017; Molnar, 2022).

Outro aspecto relevante refere-se à relação entre explicabilidade e generalização dos modelos. A análise interpretativa dos resultados possibilita verificar se os padrões aprendidos pelo algoritmo são compatíveis com os processos teóricos que fundamentam o fenômeno investigado ou se refletem apenas características específicas do conjunto de dados empregado. Assim, as técnicas de XAI constituem instrumentos auxiliares no controle de sobreajuste (*overfitting*) e na avaliação da robustez conceitual dos modelos, ao permitir que os resultados sejam examinados à luz de critérios científicos externos ao próprio algoritmo (Molnar, 2022).

Diversos estudos apontam para a necessidade de se buscar um equilíbrio entre desempenho preditivo e interpretabilidade, a depender da aplicação considerada (Zhou *et al*., 2018). Um exemplo ilustrativo é o estudo conduzido por Fernández-Delgado *et al*. (2014), no qual 179 classificadores foram avaliados utilizando 121 conjuntos de dados da base UCI. Os resultados indicaram que métodos como florestas aleatórias, máquinas de vetores de suporte, redes neurais e técnicas de *boosting* figuram entre os algoritmos com melhor desempenho médio. Todavia, tais métodos são frequentemente caracterizados como modelos do tipo “caixa-preta”, uma vez que permitem observar apenas as entradas e saídas, sem fornecer uma descrição transparente dos mecanismos internos responsáveis pelas predições.

Em contraposição, os chamados modelos de “caixa-branca” são concebidos de modo a permitir a inspeção de sua estrutura interna, possibilitando a compreensão explícita das decisões produzidas. Entre esses modelos destacam-se as árvores de decisão e os modelos de regressão, cujos parâmetros e regras podem ser diretamente analisados e interpretados (Zhou *et al*., 2018). Entre os algoritmos usualmente considerados interpretáveis incluem-se a regressão linear, a regressão logística e as árvores de decisão, embora essa classificação esteja em constante evolução em função dos avanços metodológicos e das novas ferramentas de visualização e análise (Zhou *et al*., 2018; Murdoch *et al*., 2019; Molnar, 2022).

Cada um desses modelos apresenta limitações específicas. A regressão linear, por exemplo, não é adequada a problemas de classificação. A regressão logística, embora supere essa restrição, pode apresentar dificuldades de interpretação em razão da natureza multiplicativa de seus coeficientes. Ambos os modelos tendem a falhar quando a relação entre variáveis explicativas e resposta é não linear ou quando há interações complexas entre os atributos. As árvores de decisão, por sua vez, mostram-se menos eficientes em cenários dominados por relações lineares simples, porém apresentam elevado potencial para representar interações entre variáveis, além de fornecerem estruturas decisórias facilmente compreensíveis (Molnar, 2022).

Apesar de suas vantagens, a explicabilidade não elimina completamente as limitações associadas aos modelos de aprendizado de máquina. As interpretações fornecidas correspondem, em muitos casos, a aproximações do comportamento real do modelo, estando sujeitas a incertezas e variações conforme o método empregado. Por esse motivo, a literatura ressalta que as técnicas de XAI devem ser utilizadas como ferramentas complementares, e não como substitutos da análise teórica e do conhecimento especializado do pesquisador (Doshi-Velez; Kim, 2017; Molnar, 2022).

Portanto, o potencial central dessa abordagem reside na capacidade de conciliar desempenho preditivo e transparência, fornecendo mecanismos de análise das decisões dos modelos à luz do conhecimento especializado. Nesse contexto, modelos intrinsecamente interpretáveis, como as árvores de decisão, ocupam papel de destaque, por possibilitarem a visualização direta das regras que conduzem às predições. Tal característica justifica a adoção deste algoritmo como objeto de análise neste trabalho, uma vez que alia simplicidade estrutural, capacidade de modelagem de relações não lineares e potencial explicativo, aspectos que serão discutidos na seção seguinte.

3.3 ÁRVORE DE DECISÃO: CART

As árvores de decisão constituem um método de aprendizado supervisionado aplicável tanto a problemas de classificação quanto de regressão. Entre os algoritmos mais difundidos dessa família destaca-se o CART (*Classification and Regression Trees*), proposto por Breiman *et al*. (1984), amplamente utilizado em função de sua simplicidade operacional, eficiência computacional e elevada capacidade interpretativa.

O processo de construção de uma árvore CART baseia-se na partição recursiva binária do conjunto de dados. Em cada etapa, o algoritmo seleciona uma variável explicativa e um ponto de corte que possibilitem dividir os dados em subconjuntos progressivamente mais homogêneos em relação à variável resposta. Essa divisão é orientada por medidas de impureza, cujo objetivo é reduzir a incerteza associada à classificação ou à estimativa da variável dependente.

Em problemas de classificação, a métrica de impureza mais empregada é o índice de Gini, o qual expressa o grau de heterogeneidade das classes em cada nó da árvore. Para um nó t, o índice de Gini é definido por:

$$Gini(t) = 1 - \sum_{i=1}^{C} p_i^2$$

em que $p_i$representa a proporção de observações pertencentes à classe $i$no nó considerado, e Ccorresponde ao número total de classes. A divisão ótima é aquela que resulta no menor valor do índice de Gini, indicando maior pureza dos subconjuntos formados (Breiman *et al*., 1984).

Após a realização das primeiras divisões, a árvore continua a crescer de forma iterativa até que sejam atendidos determinados critérios de parada, tais como a profundidade máxima da árvore ou o número mínimo de amostras por nó. Esses critérios exercem papel fundamental no controle da complexidade do modelo, uma vez que influenciam diretamente o equilíbrio entre sobreajuste (*overfitting*) e subajuste (*underfitting*) (Phalak, 2014). Conforme discutido por Buitinck *et al*. (2013), os principais parâmetros associados a esse controle incluem:

- **Profundidade máxima da árvore (*max_depth*)**, que restringe o número de níveis hierárquicos do modelo. Árvores excessivamente profundas tendem a capturar ruídos presentes nos dados de treinamento, enquanto árvores rasas podem ser incapazes de representar adequadamente a complexidade do problema.
- **Número mínimo de amostras para divisão de um nó (*min_samples_split*)**, que define o tamanho mínimo de um subconjunto para que uma nova partição seja realizada, evitando a criação de ramos muito específicos.
- **Número mínimo de amostras em nós folha (*min_samples_leaf*)**, que assegura um quantitativo mínimo de observações em cada nó terminal, reduzindo a probabilidade de divisões baseadas em poucos dados.
- **Redução mínima de impureza (*min_impurity_decrease*)**, que estabelece um limiar mínimo de ganho na pureza para que uma divisão seja considerada válida, inibindo a criação de ramos pouco informativos.

Além dos critérios de parada, a poda constitui uma etapa relevante no processo de ajuste das árvores de decisão, sendo empregada com o objetivo de limitar a complexidade do modelo e melhorar sua capacidade de generalização. A poda busca eliminar ramos pouco representativos, reduzindo a variância do modelo e minimizando o risco de aprendizado de padrões espúrios presentes nos dados de treinamento (Ahmed; Rizaner; Ulusoy, 2018).

Em tarefas de classificação, destaca-se a técnica de poda por custo-complexidade (*Cost-Complexity Pruning* – CCP), que estabelece uma relação entre o erro de classificação e a complexidade estrutural da árvore (Mingers, 1989). Nesse método, a árvore é inicialmente expandida até seu tamanho máximo, sendo posteriormente simplificada por meio da remoção progressiva de ramos, de acordo com penalizações associadas a complexidade do modelo.

Uma das principais vantagens das árvores de decisão, particularmente do algoritmo CART, reside em sua interpretabilidade. A estrutura hierárquica do modelo permite representar explicitamente as regras de decisão adotadas em cada nó, possibilitando a visualização do processo de classificação de forma intuitiva e compreensível (Molnar, 2020). Essa característica torna o algoritmo especialmente adequado a aplicações em que a transparência dos resultados é um requisito relevante.

Outro aspecto associado à interpretabilidade refere-se à análise da importância das variáveis explicativas. Em geral, essa importância é estimada a partir da contribuição de cada variável para a redução da impureza ao longo das divisões realizadas na árvore. Variáveis que aparecem com maior frequência nos níveis superiores tendem a exercer maior influência sobre o processo decisório do modelo (Sandri; Zuccolotto, 2008). Contudo, apesar dessas vantagens, as árvores de decisão podem apresentar sensibilidade a pequenas variações nos dados de entrada, o que pode resultar na geração de estruturas significativamente distintas a partir de alterações discretas no conjunto de treinamento (Hastie; Tibshirani; Friedman, 2009).

Do ponto de vista computacional, as árvores de decisão caracterizam-se por relativa eficiência no treinamento e baixa exigência de memória, sobretudo em bases de dados de pequeno e médio porte. Entretanto, em situações que envolvem grandes volumes de dados ou variáveis categóricas com muitos níveis, o tempo de construção pode aumentar consideravelmente, uma vez que o algoritmo avalia exaustivamente as possíveis divisões em cada nó (Hastie; Tibshirani; Friedman, 2009).

A ampla disponibilidade de implementações em bibliotecas computacionais modernas, como o *scikit-learn* (Pedregosa *et al*., 2011), tem contribuído para a difusão do algoritmo CART em diferentes áreas do conhecimento. Essas ferramentas disponibilizam mecanismos flexíveis para ajuste de parâmetros, definição de critérios de parada, aplicação de poda e escolha de métricas de avaliação, favorecendo a reprodutibilidade dos experimentos e a adaptação do modelo a distintos contextos de pesquisa.

3.3.1 Métricas de avaliação

A avaliação do desempenho de modelos de aprendizado de máquina constitui uma etapa central do processo de modelagem, pois permite quantificar a capacidade preditiva dos algoritmos e comparar diferentes estratégias metodológicas. As métricas empregadas variam conforme a natureza do problema analisado, sendo usualmente distintas para tarefas de classificação e de regressão (Hastie; Tibshirani; Friedman, 2009; James *et al*., 2013).

Em problemas de classificação, a análise do desempenho baseia-se, em geral, em medidas derivadas da matriz de confusão, que relaciona as classes previstas pelo modelo com as classes observadas. Entre as métricas mais utilizadas destaca-se a acurácia, definida como a proporção de predições corretas em relação ao total de observações. Embora amplamente empregada, essa medida pode ser insuficiente em situações nas quais as classes são desbalanceadas, uma vez que resultados elevados de acurácia podem mascarar dificuldades do modelo em identificar corretamente a classe minoritária (James *et al*., 2013; Powers, 2011).

Outras métricas relevantes incluem a sensibilidade (*recall*), associada à capacidade do modelo em reconhecer corretamente os casos positivos, e a precisão (*precision*), que expressa a proporção de predições positivas que são efetivamente corretas. A medida F1, definida como a média harmônica entre precisão e sensibilidade, constitui um indicador sintético particularmente útil quando se busca equilibrar esses dois aspectos do desempenho. A especificidade, por sua vez, representa a proporção de casos negativos corretamente classificados, sendo empregada em conjunto com a sensibilidade para caracterizar o comportamento do classificador (Powers, 2011).

Além dessas medidas pontuais, a curva ROC (*Receiver Operating Characteristic*) e a área sob a curva (AUC) são amplamente utilizadas para avaliar o desempenho global de modelos de classificação. A curva ROC expressa a relação entre a taxa de verdadeiros positivos e a taxa de falsos positivos para diferentes limiares de decisão, enquanto a AUC fornece uma medida agregada dessa relação, permitindo a comparação entre distintos classificadores independentemente do ponto de corte adotado (Fawcett, 2006). Valores de AUC mais elevados indicam maior capacidade discriminatória do modelo.

Para problemas de regressão, a avaliação do desempenho baseia-se em métricas que quantificam a discrepância entre os valores observados e os valores estimados pelo modelo. Entre as medidas mais empregadas destacam-se o erro quadrático médio (MSE), a raiz do erro quadrático médio (RMSE) e o erro absoluto médio (MAE). Essas métricas fornecem estimativas diretas da magnitude dos erros de predição, sendo particularmente úteis para comparar diferentes modelos ajustados a um mesmo conjunto de dados. O coeficiente de determinação (R^2) também é utilizado para expressar a proporção da variância explicada pelo modelo, embora sua interpretação deva ser realizada com cautela em contextos de elevada complexidade dos dados (Hastie; Tibshirani; Friedman, 2009; James *et al*., 2013).

A escolha das métricas de avaliação deve considerar os objetivos específicos do estudo e as características do conjunto de dados. Em situações nas quais os custos associados aos diferentes tipos de erro não são equivalentes, métricas baseadas apenas na acurácia ou no erro médio podem não representar adequadamente o desempenho do modelo. Nesses casos, a utilização conjunta de múltiplos indicadores permite uma análise mais abrangente e consistente, reduzindo o risco de interpretações simplificadas ou enviesadas (Powers, 2011; Fawcett, 2006).

Associadas às métricas de desempenho, as estratégias de validação desempenham papel fundamental na estimativa da capacidade de generalização dos modelos. A divisão dos dados em conjuntos de treinamento e teste, bem como a aplicação de procedimentos de validação cruzada, são práticas amplamente recomendadas para reduzir o risco de sobreajuste (*overfitting*) e para obter estimativas mais realistas do desempenho dos algoritmos em dados não utilizados no ajuste do modelo (James *et al*., 2013). Dessa forma, a avaliação de modelos de aprendizado de máquina deve ser compreendida como um processo integrado, que articula métricas quantitativas e procedimentos de validação, contribuindo para a robustez e a confiabilidade dos resultados obtidos.

4 CONTEXTUALIZAÇÃO DO ARTIGO 1

O primeiro artigo, intitulado "*Dynamics of linear erosion processes in urban areas: an approach based on GIS and CART algorithm*", tem um papel fundamental na composição da presente tese, uma vez que, a partir da pergunta de pesquisa, norteou a elaboração da hipótese: as mudanças na cobertura do solo são os principais desencadeadores dos processos erosivos em áreas urbanas.

Em suma, foi investigada a influência das condições naturais e antrópicas na ocorrência de processos erosivos lineares em áreas urbanas. Para isso, foram analisadas as feições erosivas lineares de ravinas e voçorocas mapeadas pelo Instituto de Pesquisas Tecnológicas (IPT) nas áreas urbanas do estado de São Paulo. Essas feições foram então associadas a atributos de material de origem, uso do solo, erosividade das chuvas e característica das encostas.

A metodologia adotada neste estudo desempenhou um papel crucial ao evidenciar a relevância do pré-processamento adequado dos dados para a implementação eficaz dos algoritmos de aprendizado de máquina. A correta preparação dos dados, incluindo limpeza, transformação e seleção de variáveis, mostrou-se fundamental para garantir a qualidade e a confiabilidade dos resultados obtidos. Além disso, a modelagem foi estruturada com base em um procedimento de validação cruzada em 10 partições (10-fold cross-validation), associado a uma estratégia sistemática de ajuste de parâmetros, que considerou diferentes configurações de complexidade da árvore de decisão, a introdução de custos assimétricos de classificação para lidar com o desbalanceamento entre classes e a avaliação do efeito da poda do modelo.

A opção pelo algoritmo de Árvore de Classificação e Regressão (CART) foi estratégica, levando em consideração as características das variáveis dependentes e independentes de entrada, bem como os objetivos específicos do estudo. Ao adotar um conjunto de critérios que conciliou desempenho preditivo, equilíbrio entre classes e parcimônia do modelo, foi possível realizar uma análise precisa e abrangente da importância dos fatores de influência na previsão dos processos erosivos lineares em áreas urbanas.

Os resultados indicaram, dentre outros aspectos, a importância dos atributos analisados na explicabilidade de ocorrência das feições erosivas de ravinas e voçorocas, sendo possível identificar que as variáveis de natureza antrópica têm maior relevância, especificamente, as relacionadas ao uso do solo. As variáveis contextualizadas neste conjunto - uso a montante, uso a jusante e cobertura do solo - se mostraram as mais relevantes para incitar o fenômeno estudado, uma vez que a soma dos valores de importância explicativa delas representa o maior percentual, seguido por material de origem, características do relevo e erosividade.

As conclusões deste artigo enfatizaram a pertinência de compreender como o uso do solo de fato influencia o desencadeamento de processos erosivos e fomentaram o debate acerca de como esta variável deve ser incorporada às análises. Isso devido a natureza progressiva e dinâmica que caracteriza a formação das erosões lineares: não apenas a configuração estática do uso do solo, mas sim as suas transições ao longo do tempo e no espaço circundante, considerando a área de contribuição que efetivamente interage com o fenômeno. Neste sentido, a pesquisa evidenciou a necessidade de avanço metodológico nestes aspectos.

O texto do artigo, escrito em inglês, foi avaliado e aceito para publicação pelo periódico científico "*Revista Brasileira de Cartografia*", estando em fase de editoração textual. A versão completa do artigo encontra-se em apêndice neste documento (APÊNDICE A – Artigo 1).

5 CONTEXTUALIZAÇÃO DO ARTIGO 2

O segundo artigo, intitulado "*Sampling strategies for machine learning-based linear erosion studies: a review approaching contributing area*", tem papel fundamental no direcionamento do procedimento metodológico principal da tese. A partir das considerações e conclusões do artigo 1, este artigo buscou investigar aspectos metodológicos a serem considerados na obtenção e estruturação dos dados afim de viabilizar a implementação do modelo de aprendizado de máquina.

A partir das premissas apresentadas na introdução deste artigo, se estabelece a hipótese de que amostragem e área de influência são tópicos sensíveis em estudos de natureza ambiental desenvolvidos por meio de metodologias e aplicações que incorporam técnicas de inteligência artificial.

Dessa forma, o propósito deste artigo foi investigar essa suposição inicial através da realização de uma análise bibliométrica e da extração sistemática de informações utilizando inteligência artificial (IA), abrangendo 127 documentos científicos pertinentes tanto à erosão linear quanto ao aprendizado de máquina, datados de 2017 a 2023. A extração de informações utilizando uma ferramenta de IA buscou responder as seguintes questões:

I. Quais métodos de análise multivariada são mencionados?
II. Quais métodos de aprendizado de máquina são mencionados?
III. Que métodos estatísticos são mencionados?
IV. Quais linguagens de programação são mencionadas?
V. Quais ferramentas ou materiais baseados em GIS são mencionados?

Além disso, para enriquecer a discussão, foram minuciosamente examinados documentos que tratavam especificamente sobre amostragem e área de influência, dentro do contexto de um estudo exploratório-descritivo.

Os resultados indicaram que a implementação de técnicas de aprendizado de máquina no contexto da análise de erosão linear depende fortemente de estratégias eficazes de amostragem. A discussão destacou diversos métodos de amostragem adaptados para lidar com a natureza espacial dos dados geoespaciais, enfatizando a importância de considerar fatores como desequilíbrio de classes, autocorrelação espacial e recursos computacionais disponíveis.

A amostragem de não ocorrência (instância negativas do fenômeno) emerge como um aspecto crítico no aprendizado de máquina, especialmente para lidar com o desequilíbrio de classes inerente a muitos cenários do mundo real. Fornecer um conjunto representativo de exemplos negativos ajuda a mitigar vieses (*bias*) e garante a acurácia do modelo para novos dados não vistos. A amostragem adequada e balanceada de amostras de ocorrência e não ocorrência também ajuda a evitar o sobreajuste e a promover um melhor desempenho do modelo ao facilitar a aprendizagem de padrões relevantes e reduzir o ruído presente em instâncias positivas.

A revisão descritiva-exploratória indicou que técnicas de dispersão espacial, como a técnica de amostragem estratificada espacialmente, mostram-se mais adequadas para o cenário investigado (contextos ambientais). Especificamente, a estratégia de posicionar amostras de não ocorrência fora da área de contribuição hidrológica, com um *buffer* que pode variar de 50 a 2.500 metros da borda dessa área, foi eficaz para representar a diversidade ambiental e capturar condições variadas de terreno. Além disso, a proporção de densidade de amostras de 1 para 1,2 entre ocorrência e não ocorrência contribuiu para garantir uma distribuição equilibrada de pontos de amostragem, permitindo uma cobertura abrangente da área de estudo e uma representação adequada de diferentes condições ambientais. Outra possibilidade levantada entre os estudos analisados é a possibilidade de intervenção a partir do conhecimento do especialista, contudo, a maneira como esse conhecimento é formalizado e operacionalizado metodologicamente não é detalhada.

A revisão permitiu concluir que ainda existe possibilidades para melhorias na integração de dinâmicas complexas de paisagem - como características de uso e ocupação do solo - em modelos de aprendizado de máquina para suscetibilidade à erosão linear. Embora atributos topográficos da área de influência sejam comumente incorporados, verifica-se uma lacuna na plena utilização dos fatores acumulativos para a dinâmica da água, como tamanho da área de drenagem e a distribuição de usos da terra dentro dela. Os esforços de pesquisa futuros apontam para a necessidade de desenvolvimento de modelos que considerem a interação de vários fatores ambientais e sua influência na superfície para aprimorar a acurácia e a confiabilidade das avaliações de suscetibilidade à erosão linear.

O papel deste artigo é fornecer uma base teórica sólida para o desenvolvimento do terceiro artigo da presente tese, incorporando os aspectos metodológicos limitadores identificados até então. O texto foi originalmente escrito em inglês e encontra-se no prelo (pre-proof) do periódico “*Natural Hazards Research*”. A versão do artigo encontra-se em apêndice neste documento (APÊNDICE B – Artigo 2).

6 CONTEXTUALIZAÇÃO DO ARTIGO 3

O terceiro artigo da tese, "*Unveiling Land Use Drivers of Urban Gully Erosion: An Interpretable Machine Learning Approach*", teve como objetivo específico a investigação da importância das atividades humanas, representadas por atributos de mudanças de uso e ocupação do solo, na ocorrência de erosão em áreas urbanas. Este artigo foi motivado pelos resultados dos artigos 1 e 2, os quais indicaram a relevância efetiva do uso do solo no desencadeamento de processos erosivos e o potencial da área de contribuição como delimitador na seleção de amostras de instâncias positivas e negativas.

O método empregado se embasou em testes de correlação dos dados, seguido do emprego de um algoritmo interpretável – árvore de decisão CART – para avaliação da importância das variáveis para a explicabilidade do fenômeno.

Foram investigados os padrões de mudança de cobertura do solo e a influência deles na ocorrência de processos erosivos nas áreas urbanas do estado de São Paulo. O estudo utilizou um inventário pontual de 1.398 processos erosivos em áreas urbanas do IPT e propôs uma estratégia inovadora para a incorporação de instâncias negativas, pontos de não ocorrência de processos erosivos, na proporção de 1:1,2 (Liu *et al.*, 2023). Logo, foram estabelecidas 1.678 instâncias negativas para processos erosivos. Também foi conduzida uma análise comparativa quanto à estratégia de incorporação das classes de transições de uso do solo: (i) considerando um *buffer* de 100 m a partir da feição pontual; (ii) considerando a área de contribuição hídrica delimitada a partir dessa feição pontual.

Investigaram-se os padrões de mudança da cobertura do solo em um intervalo temporal de 20 anos (1990-2010). A definição desse período deve-se a dois fatores principais: (i) o ano final (2010) coincide com a data do inventário de dados erosivos que fundamenta a análise, e (ii) o ano inicial (1990) foi selecionado para capturar a maior variabilidade possível de transições, uma vez que a análise dos dados matriciais, com resolução de 30m, indicou uma baixa variabilidade significativa entre os anos subsequentes, tornando 1990 a opção mais robusta para um *baseline* de qualidade.

A construção dos conjuntos de dados envolveu um trabalho minucioso em ambiente SIG e, posteriormente, focou no desenvolvimento do modelo e ajuste dos hiperparâmetros adequados para a análise desses conjuntos de dados. Para tal, a codificação baseada em frequência, na qual cada variável representa a proporção relativa de uma determinada transição de uso do solo em uma área específica, e a normalização por linha possibilitaram comparações diretas entre diferentes regiões e entre os conjuntos de dados.

Em suma, a metodologia adotada demonstrou-se eficaz para a análise da influência das transições de uso do solo na erosão, combinando técnicas de aprendizado de máquina interpretáveis com uma abordagem sistemática de validação.

Em termos de interpretação dos resultados, as transições de uso do solo mais importantes para a classificação de erosão nas árvores de decisão foram, para ambos os conjuntos de dados, àquelas relacionadas a usos agropecuários que se mantiveram inalterados ou que se converteram em áreas não vegetadas, sugerindo que essas características representam situações mais propensas a sofrer erosão.

As áreas de florestas que permaneceram como florestas, apresentaram características estatísticas bastante similares nos dois conjuntos de dados analisados. Apesar dessa semelhança, a importância da variável variou significativamente entre os modelos ajustados, No modelo baseado nas áreas de contribuição hídrica essas áreas de florestas foram a variável mais relevante, enquanto no modelo ajustado com o conjunto baseado no *buffer* de 100 m, essa variável sequer figurou entre as mais importantes.

A comparação entre as duas estruturas dos dados muito além permitir identificar a estrutura mais adequada para incorporação, ofereceu uma base sólida para estudos futuros sobre o tema. Isso porque apesar da área de contribuição hídrica ter se mostrado mais promissora na identificação de casos positivos para erosão, a estratégia de *buffer* apresentou um melhor equilíbrio preditivo entre as classes.

O texto foi originalmente escrito em inglês e encontra-se em fase de avaliação no periódico “*Earth Surface Processes and Landforms*”. A versão do artigo encontra-se em apêndice neste documento (APÊNDICE C – Artigo 3) [1].

---

[1] O artigo encontra-se submetido a periódico científico e em processo de avaliação. Assim, sua versão definitiva poderá sofrer ajustes decorrentes das recomendações dos pareceristas e das diretrizes editoriais, inclusive quanto ao título e ao periódico no qual será publicado.

7 CONSIDERAÇÕES FINAIS

Os objetivos delineados no início desta pesquisa foram abordados de maneira progressiva e integrada, com os Artigos 1 e 2 representando os alicerces teórico e metodológico que sustentaram as investigações do Artigo 3, o qual, por sua vez, respondeu à questão de pesquisa central e comprovou a hipótese da tese.

O Artigo 1 cumpriu o papel fundamental de validar a premissa inicial, demonstrando de forma robusta que variáveis antrópicas, em especial as relacionadas ao uso e cobertura do solo, são os principais fatores no desencadeamento de processos erosivos lineares em áreas urbanas. Mais do que confirmar essa influência, seus resultados sinalizaram a necessidade de superar a análise estática da paisagem, apontando para a importância de considerar as transições temporais e a área de contribuição que interage espacialmente com o fenômeno.

Esta constatação direcionou o escopo do Artigo 2, que se dedicou a investigar os fundamentos metodológicos para tal avanço. A revisão sistemática realizada evidenciou a amostragem de instâncias negativas e a delimitação de áreas de influência como aspectos críticos para a robustez de modelos de aprendizado de máquina. O artigo consolidou assim as diretrizes para a correta estruturação dos dados, recomendando estratégias espaciais específicas e fornecendo a base teórica sólida necessária para a etapa seguinte.

Finalmente, o Artigo 3 operacionalizou todo o conhecimento acumulado, aplicando-o para identificar os padrões de mudança de uso do solo com maior impacto na erosão linear. A pesquisa confirmou que transições relacionadas a usos agropecuários estáveis ou convertidos em áreas não vegetadas são altamente significativas. A contribuição metodológica deste artigo foi dupla e reveladora: a comparação entre as abordagens da área de contribuição hídrica e do *buffer* de 100m demonstrou que elas possuem potenciais preditivos distintos e complementares. Enquanto a área de contribuição hídrica mostrou-se mais eficaz na identificação de casos positivos de erosão (e.g., elegendo "áreas de floresta preservada" como sua variável mais relevante), a estratégia do *buffer* proporcionou um melhor equilíbrio preditivo geral entre as classes. Portanto, os resultados não indicam a superioridade de uma abordagem sobre a outra, mas sim a necessidade de selecioná-las com base no objetivo específico do modelo – seja para detecção máxima de eventos ou para um balanceamento global de precisão –, abrindo um campo fértil para investigações futuras que busquem validar e estabelecer suas potencialidades de forma mais ampla.

Em síntese, esta tese não apenas alcançou seu objetivo de investigar o efeito das transições de cobertura do solo na erosão linear urbana, mas também ofereceu uma contribuição metodológica substancial. Ao longo dos três artigos, foi construído e validado um arcabouço analítico que integra, de forma inovadora, a dimensão espaço-temporal das mudanças antrópicas à modelagem preditiva. Os resultados aqui obtidos possuem aplicabilidade direta, fornecendo subsídios valiosos para a academia, o planejamento urbano e a gestão de riscos, e contribuindo, em última instância, para o desenvolvimento de cidades mais resilientes e sustentáveis.

REFERÊNCIAS

AHMED, A. M.; RIZANER, A.; ULUSOY, A. H. A novel decision tree classification based on post-pruning with Bayes minimum risk. *Plos one*, v. 13, n. 4, p. e0194168, 2018. DOI: 10.1016/j.cor.2005.12.009.

AMIRI, M. *et al.* Assessment of the importance of gully erosion effective factors using Boruta algorithm and its spatial modeling and mapping using three machine learning algorithms. **Geoderma**, [*s. l.*], v. 340, p. 55–69, 2019. DOI: 10.1016/j.geoderma.2018.12.042.

APRIGLIANO, V. *et al.* Built environment effect on metro ridership in metropolitan area of Valparaíso, Chile, under different influence area approaches. **ISPRS International Journal of Geo-Information**, v. 13, p. 266, 2024.

ARABAMERI, A. *et al*. Novel machine learning approaches for modelling the gully erosion susceptibility. **Remote Sensing**, v. 12, art. 2833, 2020. DOI: 10.3390/rs12172833.

ARRIBAS-BEL, D.; GARCÍA-LÓPEZ, M.; VILADECANS-MARSAL, E. Building(s and) cities: delineating urban areas with a machine learning algorithm. **Journal of Urban Economics**, v. 125, 2021.

BAG, R. *et al*. Modelling and mapping of soil erosion susceptibility using machine learning in a tropical hot sub-humid environment. **Journal of Cleaner Production**, v. 364, art. 132428, 2022. DOI: 10.1016/j.jclepro.2022.132428.

BARAKAT, A. *et al*. Mapping of water-induced soil erosion using machine learning models: a case study of Oum Er Rbia Basin (Morocco). **Earth Systems and Environment**, v. 7, p. 151–170, 2023. DOI: 10.1007/s41748-022-00317-x.

BAZZOFFI, P. *et al*. **Soil erosion**. In: Science. Boca Raton: CRC Press, 1993.

BERTONI, J.; LOMBARDI NETO, F. **Conservação do solo**. São Paulo: Ícone, 1990.

BISHOP, C. M. **Pattern Recognition and Machine Learning**. [S.l.: s.n.]: Springer, 2006.

BONZANINI, H. L.; LUPINACCI, C. M.; STEFANUTO, E. B. A Erosão Linear e Sua Relação com a Morfometria do Relevo na Alta Bacia do Rio Capivara – Botucatu (SP).

Revista Brasileira de Geografia Física, [*s. l.*], v. 15, n. 4, p. 1947–1964, 2022. Available at: https://periodicos.ufpe.br/revistas/index.php/rbgfe/article/view/251894.

BORRELLI, P. *et al*. Land use and climate change impacts on global soil erosion by water (2015–2070). **Proc. of the National Academy of Sciences**, v. 117, p. 21994–22001, 2020.

BORSELLI, L., CASSI, P., TORRI, D., Prolegomena to sediment and flow connectivity in the landscape: A GIS and field numerical assessment. **Catena**, 75, 268–277, 2008.

BRAGA, A. d. P.; CARVALHO, A. P. d. L. F.; LUDERMIR, T. B. **Redes Neurais Artificiais**: Teoria e Aplicações. Rio de Janeiro: LTC Editora, 2000. 262 p. ISBN 9788521615644.

BRASIL. **Lei nº 6.766, de 19 de dezembro de 1979**. [*S. l.: s. n.*], 1979.

BREIMAN, L. *et al*. **Classification and regression trees**. Boca Raton: Chapman & Hall/CRC, 1984.

BRODERSEN, K. H. *et al.* The Balanced Accuracy and Its Posterior Distribution. *In*: , 2010. **2010 20th International Conference on Pattern Recognition**. [*S. l.*]: IEEE, 2010. p. 3121–3124. Available at: http://ieeexplore.ieee.org/document/5597285/.

BUITINCK, Lars *et al*. API design for machine learning software: experiences from the scikit-learn project. **arXiv preprint**, 1309.0238, 2013. DOI: 10.48550/arXiv.1309.0238.

CARVALHO JUNIOR, O. *et al.* Urbanization impacts upon catchment hydrology and gully development using mutli-temporal digital elevation data analysis. **Earth Surface Processes and Landforms**, [*s. l.*], v. 35, n. 5, p. 611–617, 2010. DOI: 10.1002/esp.1917.

CASTILLO, C.; GÓMEZ, J.A. A century of gully erosion research: Urgency, complexity and study approaches. **Earth-Science Reviews**, [*s. l.*], v. 160, p. 300–319, 2016. Available at: https://linkinghub.elsevier.com/retrieve/pii/S0012825216301842.

CATI. **Digital Elevation Model of São Paulo State**. São Paulo: Brazilian Farming Information Center, 2016. Available at: http://mapas.cati.sp.gov.br/mdecati.html.

CATTANEO, A.; NELSON, A.; McMENOMY, T. Global mapping of urban–rural catchment areas reveals unequal access to services. **Proceedings of the National**

Academy of Sciences of the United States of America, v. 118, 2021.

CHALISE, D.; KUMAR, L.; KRISTIANSEN, P. Land degradation by soil erosion in Nepal: a review. **Soil Systems**, v. 3, art. 12, 2019. DOI: 10.3390/soilsystems3010012.

CHAPLOT, V. *et al.* Dynamic modelling for linear erosion initiation and development under climate and land-use changes in northern Laos. **CATENA**, [*s. l.*], v. 63, n 2, p. 318–328, 2005.

CHEN, R. *et al.* Determination of contributing area threshold and downscaling of topographic factors for small watersheds in hilly areas of purple soil. **Land**, v. 13, p. 1193, 2024.

CHRISTOFOLETTI, A. **Geomorfologia**. São Paulo: Editora Blucher, 1980.

DJOUKBALA, O. *et al.* Comparison of the erosion prediction models from USLE, MUSLE and RUSLE in a Mediterranean watershed, case of Wadi Gazouana (N-W of Algeria). **Modeling Earth Systems and Environment**, [*s. l.*], v. 5, n. 2, p. 725–743, 2019. Available at: https://doi.org/10.1007/s40808-018-0562-6.

DOS SANTOS, J. C. N. *et al.* Land use impact on soil erosion at different scales in the Brazilian semi-arid. **REVISTA CIÊNCIA AGRONÔMICA**, [*s. l.*], v. 48, n. 2, p. 251–260, 2017. Available at: http://www.gnresearch.org/doi/10.5935/1806-6690.20170029. Accessed at: 13 Oct. 2021.

DOSHI-VELEZ, F.; KIM, B. Towards a rigorous science of interpretable machine learning. **arXiv preprint**,*1702.08608*, 2017. DOI: 10.48550/arXiv.1702.08608.

DUBE, H. B.; MUTEMA, M.; MUCHAONYERWA, P.; POESEN, J.; CHAPLOT, V. A global analysis of the morphology of linear erosion features. **Catena**, v. 190, art. 104542, 2020. DOI: 10.1016/j.catena.2020.104542.

DUVILLIER, C. *et al.* Development and evaluation of a method to identify potential release areas of snow avalanches based on watershed delineation. **Natural Hazards and Earth System Sciences**, v. 23, p. 1383–1398, 2023.

ENDRES, P. F. *et al.* Quantificação das classes de erosão por tipo de uso do solo no município de Franca - SP. **Engenharia Agrícola**, [*s. l.*], v. 26, n. 1, p. 200–207, 2006. Available at: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-

69162006000100022&lng=pt&tlng=pt.

FALCÃO, K. dos S. *et al*. Surface runoff and soil erosion in a natural regeneration area of the Brazilian Cerrado. **International Soil and Water Conservation Research**, v. 8, p. 124–130, 2020. DOI: 10.1016/j.iswcr.2020.04.004.

FANG, C., ZHOU, L., GU, X., LIU, X., & WERNER, M. A data driven approach to urban area delineation using multi source geospatial data. **Scientific Reports**, v. 15, n. 1, p. 8708, 2025.

FAWCETT, Tom. An introduction to ROC analysis. **Pattern recognition letters**, v. 27, n. 8, p. 861-874, 2006.

FERNÁNDEZ-DELGADO, M. *et al*. Do we need hundreds of classifiers to solve real world classification problems? **Journal of Machine Learning Research**, v. 15, n. 1, p. 3133–3181, 2014.

FERREIRA, T. L.; MARTINS, T. D. Análise da vulnerabilidade das edificações em áreas afetadas por movimentos de massa. **Revista do Departamento de Geografia**, [*s. l.*], v. 41, n. 1, p. e176904, 2021. Available at: https://www.revistas.usp.br/rdg/article/view/176904.

FOLTÊTE, J.; VUIDEL, G. Using landscape graphs to delineate ecologically functional areas. **Landscape Ecology**, v. 32, p. 249–263, 2017.

FRANCO, C. R**. Inteligência Artificial**. Londrina: Uniasselvi, 2014. 168 p. ISBN 978-85-68075-77-7.

FRANKL, A. *et al*. Gully prevention and control: techniques, failures and effectiveness. **Earth Surface Processes and Landforms**, v. 46, p. 220–238, 2020. DOI: 10.1002/esp.5033.

GANASRI, B. P.; RAMESH, H. J. G. F. Assessment of soil erosion by RUSLE model using remote sensing and GIS. **Geoscience Frontiers**, v. 7, n. 6, p. 953-961, 2016.

GHOLAMI, H. *et al*. An explainable integrated machine learning model for mapping soil erosion by wind and water in a catchment with three desiccated lakes. **Aeolian Research**, v. 67–69, art. 100924, 2024. DOI: 10.1016/j.aeolia.2024.100924.

GIGLIO, L. C. **Comparação de estimativas de perda de solo pela EUPS com**

produtos de aeronave remotamente pilotada na análise de processos erosivos em áreas de regeneração. 2025. Trabalho de Conclusão de Curso (Graduação em Gestão e Análise Ambiental) – Universidade Federal de São Carlos, São Carlos, 2025.

GREENE, W. H. **Econometric Analysis**. [*S. l.*]: Prentice Hall, 2003.

GUAN, M.; SILLANPÄÄ, N.; KOIVUSALO, H. Modelling and assessment of hydrological changes in a developing urban catchment. **Hydrological Processes**, v. 29, n. 13, p. 2880-2894, 2015.

GUDIÑO-ELIZONDO, N. *et al*. Measuring ephemeral gully erosion rates and topographical thresholds in an urban watershed using unmanned aerial systems and structure from motion photogrammetric techniques. **Land Degradation & Development**, v. 29, p. 1896–1905, 2018. DOI: 10.1002/ldr.2976.

GUERRA, A. J. T. O início do processo erosivo. In: GUERRA, A. J. T. (org.). **Erosão e conservação dos solos**. Rio de Janeiro: Bertrand Brasil, 2012. p. 50.

GUERRA, A.J.T.; JORGE, M.C.O. **Degradação dos solos no Brasil**. Rio de Janeiro: Bertrand Brasil, 2018.

GUO, Y. *et al.* Modelling the impacts of climate and land use changes on soil water erosion: Model applications, limitations and future challenges. **Journal of Environmental Management**, [*s. l.*], v. 250, p. 109403, 2019.

GUYON, I. *et al.* Gene Selection for Cancer Classification using Support Vector Machines. **Machine Learning**, [*s. l.*], v. 46, n. 1, p. 389–422, 2002. Available at: https://doi.org/10.1023/A:1012487302797.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The elements of statistical learning: data mining, inference, and prediction**. New York: Springer, 2009.

IMWANGANA, F. M. *et al*. Gully erosion associated with peri-urbanization: focus on the catchments around Kimwenza in the south of Kinshasa (DR Congo). **East African Journal of Environment and Natural Resources**, v. 8, n. 2, 2025. DOI: 10.37284/eajenr.8.2.3160.

JAIN, A. K.; MURTY, M. N.; FLYNN, P. J. Data clustering: A review. **ACM Computing Surveys**, v. 31, n. 3, p. 264–323, 1999.

JAMES, G. *et al*. **An introduction to statistical learning: with applications in R**. New York: Springer, 2013.

KUCHKAROVA, D.; ACHILOVA, D. Model of rainwater runoff formation on the surface of complex topography. **IOP Conference Series: Materials Science and Engineering**, [*s. l.*], v. 1030, n. 1, p. 012107, 2021. Available at: https://iopscience.iop.org/article/10.1088/1757-899X/1030/1/012107.

KUHN, C. E. S.; REIS, F. A. G. V.; ZARFL, C.; GRATHWOHL, P. Ravines and gullies, a review about impact valuation. **Natural Hazards**, v. 117, p. 597–624, 2023. DOI: 10.1007/s11069-023-05874-6.

KURSA, M. B.; RUDNICKI, Witold R. Feature Selection with the Boruta Package. **Journal of Statistical Software**, [*s. l.*], v. 36, n. 11, 2010. Available at: http://www.jstatsoft.org/v36/i11/.

LIN, H.; ZHAO, Y. Soil erosion assessment of alpine grassland in the source park of the Yellow River on the Qinghai-Tibetan Plateau, China. **Frontiers in Ecology and Evolution**, v. 9, 2022. DOI: 10.3389/fevo.2021.771439.

LIU, X. *et al*. Data-driven machine learning in environmental pollution: gains and problems. **Environmental Science & Technology**, v. 56, n. 4, p. 2124–2133, 2022.

LIU, Y. *et al.* Optimizing the Sample Selection of Machine Learning Models for Landslide Susceptibility Prediction Using Information Value Models in the Dabie Mountain Area of Anhui, China. **Sustainability**, [*s. l.*], v. 15, n. 3, p. 1971, 2023. Available at: https://www.mdpi.com/2071-1050/15/3/1971.

LUCCHESE, L. V.; DE OLIVEIRA, G. G.; PEDROLLO, O. C. Investigation of the influence of nonoccurrence sampling on landslide susceptibility assessment using Artificial Neural Networks. **CATENA**, [*s. l.*], v. 198, p. 105067, 2021. Available at: https://www.sciencedirect.com/science/article/pii/S0341816220306172.

MAATEN, L. van der; HINTON, G. Visualizing data using t-sne. **Journal of Machine Learning Research**, v. 9, p. 2579–2605, 2009

MACKAY, D.; BAND, L. Extraction and representation of nested catchment areas from digital elevation models in lake-dominated topography. **Water Resources Research**, v. 34, p. 897–901, 1998.

MAGALHÃES, W. A. *et al.* Using the GeoWEPP Model to Predict Water Erosion in Micro-Watersheds in the Brazilian Cerrado. **Sustainability**, [*s. l.*], v. 15, n. 6, p. 4711, 2023. Available at: https://www.mdpi.com/2071-1050/15/6/4711.

MAWE, G. *et al*. Mapping urban gullies in the Democratic Republic of the Congo. **Nature**, v. 644, p. 952–959, 2025. DOI: 10.1038/s41586-025-09371-7.

MAWE, G. *et al*. What controls the expansion of urban gullies in tropical environments? Lessons learned from contrasting cities in DR Congo. **Catena**, v. 241, art. 108055, 2024. DOI: 10.1016/j.catena.2024.108055.

MCELWEE, P. *et al.* The impact of interventions in the global land and agri-food sectors on Nature's Contributions to People and the UN Sustainable Development Goals. **Global Change Biology**, [*s. l.*], v. 26, n. 9, p. 4691–4721, 2020. Available at: https://onlinelibrary.wiley.com/doi/10.1111/gcb.15219.

MILLER, T. Explanation in artificial intelligence: Insights from the social sciences. **Artificial Intelligence**, v. 267, p. 1–38, 2019. DOI: 10.1016/j.artint.2018.07.007.

MINGERS, J. An empirical comparison of pruning methods for decision tree induction. **Machine Learning**, v. 4, p. 227–243, 1989. DOI: 10.1023/A:1022604100933.

MOHRI, M.; ROSTAMIZADEH, A.; TALWALKAR, **A. Foundations of Machine Learning**. 2. ed. New York, USA: MIT Press, 2018.

MOLNAR, C. **Interpretable machine learning**. 2. ed. Leanpub, 2022.

MORENO-MONROY, A.; SCHIAVINA, M.; VENERI, P. Metropolitan areas in the world: delineation and population trends. **Journal of Urban Economics**, art. 103242, 2020. DOI: 10.1016/j.jue.2020.103242.

MORGAN, R. P. C. **Soil erosion and conservation**. Oxford: Blackwell Publishing, 2009.

MORTOJA, M.; YIGITCANLAR, T.; MAYERE, S. What is the most suitable methodological approach to demarcate peri-urban areas? A systematic review of the literature. **Land Use Policy**, v. 97, 2020.

MOTA, A. *et al.* Efeito da Escala na Obtenção de Parâmetros Morfométricos em Pequena Bacia Experimental (Map scale effect on morfometric parameters estimation in a small experimental basin). *In*: , 2013. **Anais [...]**. [*S. l.: s. n.*], 2013.

MURDOCH, W. J. *et al*. Definitions, methods, and applications in interpretable machine learning. **Proc. Natl. Acad. Sci. U.S.A**, v. 116, n. 44, p. 22071-22080, 2019. DOI: 10.1073/pnas.1900654116.

NATURAL CAPITAL PROJECT. **InVEST 3.17.0**. Stanford University, University of Minnesota, Chinese Academy of Sciences, The Nature Conservancy, World Wildlife Fund, Stockholm Resilience Centre, Royal Swedish Academy of Sciences. 2025.

NEARING, M. A. *et al*. A Process-Based Soil Erosion Model for USDA-Water Erosion Prediction Project Technology. **Transactions of the ASAE**, [*s. l.*], v. 32, n. 5, p. 1587–1593, 1989.

OLIVATTO, T. F.; DI LOLLO, J. A. Sampling strategies for machine learning-based linear erosion studies: a review approaching contributing area. **Natural Hazards Research** (no prelo), 2025.

ONU. **Objetivos de Desenvolvimento Sustentável**. [*S. l.*], 2015. Available at: https://nacoesunidas.org/pos2015/agenda2030/. Accessed at: 26 Jun. 2023.

PECK, R; OLSEN, C; DEVORE, J L. **Introduction to Statistics and Data Analysis**. [*S. l.*]: Cengage Learning, 2015.

PEDREGOSA, F. *et al*. Scikit-learn: Machine learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011.

PHALAK, P. *et al*. Analysis of decision tree: a survey. ***International Journal of Engineering Research and Technology***, v. 3, 2014. DOI: 10.17577/IJERTV3IS030302.

POESEN, J. *et al.* Gully erosion and environmental change: Importance and research needs. **Catena**, [*s. l.*], v. 50, no. 2–4, p. 91 – 133, 2003.

POESEN, J. Soil erosion in the Anthropocene: research needs. **Earth Surface Processes and Landforms**, v. 43, n. 1, p. 64–84, 2018.

POLOVINA, S. *et al*. Soil erosion assessment and prediction in urban landscapes: a new G2 model approach. **Applied Sciences**, v. 11, 2021. DOI: 10.3390/app11094154.

POWERS, D. M. W. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness & correlation. **Journal of Machine Learning Technologies**, v. 2, n. 1, p. 37–63, 2011. DOI: 10.48550/arXiv.2010.16061.

POURGHASEMI, H. R.; YOUSEFI, S.; KORNEJADY, A.; CERDÀ, A. Performance assessment of individual and ensemble data-mining techniques for gully erosion modeling. **Science of the Total Environment**, v. 609, p. 764–775, 2017. DOI: 10.1016/j.scitotenv.2017.07.198.

POURGHASEMI, H.; SADHASIVAM, N.; KARIMINEJAD, N.; COLLINS, A. Gully erosion spatial modelling: role of machine learning algorithms in selection of the best controlling factors and modelling process. **Geoscience Frontiers**, v. 11, p. 2207–2219, 2020. DOI: 10.1016/j.gsf.2020.03.005.

RAHMATI, O. *et al*. Contribution of physical and anthropogenic factors to gully erosion initiation. **Catena**, v. 201, art. 105925, 2022. DOI: 10.1016/j.catena.2021.105925.

RAHMATI, O. *et al*. Evaluation of different machine learning models for predicting and mapping the susceptibility of gully erosion. **Geomorphology**, v. 298, p. 118–137, 2017. DOI: 10.1016/j.geomorph.2017.09.006.

REECE, D. A. *et al.* Using Google Earth Imagery to Target Assessments of Ephemeral Gully Erosion. *In*: , 2023, St. Joseph, MI. **Soil Erosion Research Under a Changing Climate, January 8-13, 2023, Aguadilla, Puerto Rico, USA**. St. Joseph, MI: American Society of Agricultural and Biological Engineers, 2023. Available at: https://elibrary.asabe.org/abstract.asp?JID=1&AID=53833&CID=soil2023&T=1.

RENARD, K.G *et al*. **Predicting Soil Erosion by Water**: A Guide to Conservation Planning. Washington: USDA. Agricultural Handbook. 1997

RIEGER, W. A phenomenon-based approach to upslope contributing area and depressions in DEMs. **Hydrological Processes**, v. 12, p. 857–872, 1998.

SAHA, A. *et al.* Optimization modelling to establish false measures implemented with ex-situ plant species to control gully erosion in a monsoon-dominated region with novel in-situ measurements. **Journal of Environmental Management**, [*s. l.*], v. 287, p. 112284, 2021.

SANDRI, M.; ZUCCOLOTTO, P. A bias correction algorithm for the Gini variable importance measure in classification trees. **Journal of Computational and Graphical**

Statistics, v. 17, p. 611–628, 2008. DOI:10.1198/106186008X344522.

SEMERARO, T. *et al.* Planning of urban green spaces: an ecological perspective on human benefits. **Land**, v. 10, p. 210, 2021.

TARODIYA, R.; LEVY, A. Surface erosion due to particle-surface interactions - A review. **Powder Technology**, [*s. l.*], v. 387, p. 527–559, 2021.

TELLES, T. S.; GUIMARÃES, M. F.; DECHEN, S. C. F. The costs of soil erosion. **Revista Brasileira de Ciência do Solo**, [*s. l.*], v. 35, no. 2, p. 287–298, 2011.

VANMAERCKE, M. *et al*. How fast do gully headcuts retreat? **Earth-Science Reviews**, v. 154, p. 336–355, 2016. DOI: 10.1016/j.earscirev.2016.01.009.

VELASTEGUI-MONTOYA, A. *et al.* Land Use and Land Cover in Tropical Forest: Global Research. **Forests**, [*s. l.*], v. 13, no. 10, p. 1709, 2022. Available at: https://www.mdpi.com/1999-4907/13/10/1709.

WANG, K. *et al.* A perception and judgement of contributing factors for allocating urban residential land: a systematic review and statistical analysis. **Land**, v. 12, 2023.

WIJESUNDARA, N C; ABEYSINGHA, N S; DISSANAYAKE, D M S L B. GIS-based soil loss estimation using RUSLE model: a case of Kirindi Oya river basin, Sri Lanka. **Modeling Earth Systems and Environment**, [*s. l.*], v. 4, n. 1, p. 251–262, 2018. Available at: https://doi.org/10.1007/s40808-018-0419-z.

WILLIAMS, J. R. SEDIMENT ROUTING FOR AGRICULTURAL WATERSHEDS 1. **JAWRA Journal of the American Water Resources Association**, [*s. l.*], v. 11, n. 5, p. 965–974, 1975. Available at: https://onlinelibrary.wiley.com/doi/10.1111/j.1752-1688.1975.tb01817.x.

WILLIAMS, J. R.; NICKS, A. D.; ARNOLD, J. G. Simulator for Water Resources in Rural Basins. **Journal of Hydraulic Engineering**, [*s. l.*], v. 111, n. 6, p. 970–986, 1985. Available at: https://ascelibrary.org/doi/10.1061/%28ASCE%290733-9429%281985%29111%3A6%28970%29.

WISCHMEIER, W. H.; SMITH, D. D. **Predicting rainfall erosion losses: a guide to conservation planning**. [*S. l.*]: Department of Agriculture, Science and Education Administration, 1978.

WISCHMEIER, W. H.; SMITH, D. D. **Predicting rainfall-erosion losses from cropland east of the Rocky Mountains: Guide for selection of practices for soil and water conservation**. [*S. l.*]: Agricultural Research Service, US Department of Agriculture, 1965.

YOO, C. *et al.* Change of rainfall–runoff processes in urban areas due to high-rise buildings. **Journal of Hydrology**, [*s. l.*], v. 597, p. 126155, 2021. Available at: https://linkinghub.elsevier.com/retrieve/pii/S002216942100202X.

ZHOU, Q. *et al*. Measuring interpretability for different types of machine learning models. In: SPRINGER. **Trends and Applications in Knowledge Discovery and Data Mining**: PAKDD 2018 Workshops, BDASC, BDM, ML4Cyber, PAISI, DaMEMO, Melbourne, VIC, Australia, June 3, 2018, Revised Selected Papers 22. [S.l.: s.n.], 2018. p. 295–308.

ZHOU, Z.-H. **Machine learning**. Singapore: Springer Nature, 2021.DOI: 10.1007/978-981-15-1967-3.

ZOLEZZI, G. *et al*. Urban gully erosion in sub-Saharan Africa: a case study from Uganda. **Land Degradation & Development**, v. 29, p. 849–859, 2018. DOI: 10.1002/ldr.2865.

APÊNDICE A – ARTIGO 1

Rev. Bras. Cartogr, vol. x, aaaa DOI: http://dx.doi.org/10.14393/rbcvxxn0a-xxxx

Revista Brasileira de Cartografia
ISSN 1808-0936 | https://doi.org/10.14393/revbrascartogr
Sociedade Brasileira de Cartografia, Geodésia, Fotogrametria e Sensoriamento Remoto

Dynamics of linear erosion processes in urban areas: an approach based on GIS and CART algorithm

Dinâmica dos processos erosivos lineares em áreas urbanas: uma abordagem apoiada em SIG e algoritmo CART

Tatiane Ferreira Olivatto[1], José Augusto Di Lollo[2] e Denise Balestrero Menezes[3]

1 Postgraduate Program in Urban Engineering, Federal University of São Carlos (UFSCar), São Carlos-SP, Brazil. tatianeolivatto@ufscar.br. ORCID: https://orcid.org/0000-0002-5770-7088
2 Engineering Faculty, São Paulo State University (UNESP), Ilha Solteira-SP, Brazil. jose.lollo@unesp.br. ORCID: https://orcid.org/0000-0002-6703-5377
3 Postgraduate Program in Urban Engineering, Federal University of São Carlos (UFSCar), São Carlos-SP, Brazil. denisebm@ufscar.br. ORCID: https://orcid.org/0000-0003-2962-3028

Recebido: 08.2025 | Aceito: 02.2026

Abstract: Linear erosion is a significant geomorphological process that progresses through distinct stages—rills, gullies, and ravines—shaped by natural and anthropogenic factors. While extensively studied in rural environments, its occurrence and dynamics in urban areas remain understudied, despite their implications for infrastructure stability and urban planning. This study investigates the relative influence of substrate conditions, land use, rainfall erosivity, and geomorphometric variables on the occurrence of gullies and ravines in urban areas, using a case study in São Paulo State, Brazil. An inventory of 1,398 mapped urban erosion features was analyzed using Geographic Information Systems (GIS) and a Classification and Regression Tree (CART) model. Results indicate that anthropogenic factors—particularly triggering actions related to surface runoff and stormwater management—are the dominant controls on urban linear erosion, surpassing traditional natural predictors such as rainfall erosivity. Geological conditions and geomorphometric attributes, especially elevation and plan curvature, play a secondary but meaningful role by conditioning the susceptibility of terrain units. The results also shown a cluster of geological formations with a lower propensity for gully development. The final pruned CART model achieved balanced classification performance (overall accuracy of 72.7%, AUC = 0.78), demonstrating robust generalization and limited overfitting. These findings highlight the central role of urban hydrological alteration in erosion processes and emphasize the need to integrate erosion risk assessment into urban planning and stormwater management strategies, contributing to more resilient and sustainable cities.
Keywords: Gully. Ravine. Land Use. Geographic Information System. Decision Tree.

Resumo: A erosão linear é um importante processo geomorfológico que evolui por estágios distintos — sulcos, ravinas e voçorocas — sendo condicionada por fatores naturais e antrópicos. Embora amplamente estudada em ambientes rurais, sua ocorrência e dinâmica em áreas urbanas permanece pouco investigada, apesar de suas implicações para a estabilidade da infraestrutura e o planejamento urbano. Este estudo analisa a influência relativa das condições do substrato, do uso do solo, da erosividade das chuvas e de variáveis geomorfométricas na ocorrência de ravinas e voçorocas em áreas urbanas, a partir de um estudo de caso no estado de São Paulo, Brasil. Um inventário de 1.398 feições de erosão linear urbana mapeadas foi analisado usando Sistemas de Informação Geográfica (SIG) e um modelo de Árvores de Classificação e Regressão (CART). Os resultados indicam que fatores antrópicos — particularmente ações desencadeadoras associadas ao escoamento superficial e ao manejo das águas pluviais — constituem os principais controles da erosão linear urbana, superando preditores naturais tradicionais, como a erosividade da chuva. As condições geológicas e os atributos geomorfométricos, especialmente a altitude e o plano de curvatura, exercem um papel secundário, porém relevante, ao condicionarem a suscetibilidade das unidades do terreno. Os resultados também evidenciaram um agrupamento de formações geológicas com menor propensão ao desenvolvimento de voçorocas. O modelo CART final podado apresentou desempenho de classificação equilibrado (acurácia global de 72,7%, AUC = 0,78), indicando boa capacidade de generalização e baixa suscetibilidade ao sobreajuste. Esses achados reforçam o papel central das alterações hidrológicas induzidas pela urbanização nos processos erosivos e destacam a necessidade de integrar a avaliação do risco de erosão ao planejamento urbano e às estratégias de gestão das águas pluviais, contribuindo para cidades mais resilientes e sustentáveis.
Palavras-chave : Voçoroca. Ravina. Uso do Solo. Sistema de Informação Geográfica. Árvore de Decisão.

1 INTRODUCTION

Soil erosion is among the most damaging environmental degradation processes on the globe, reaching urban and rural populations. The consequences of erosion processes, besides damaging the environment (Chalise et al., 2019; Issaka & Ashraf, 2017), have economic and social reflexes, such as loss of agricultural areas, compromising food production and material and human losses in urban occupations in risk zones (Pierce & Lal, 2017; Pimentel, 2006; Rodrigo-Comino et al., 2018; Telles et al., 2011; van Zelm et al., 2018; Zdruli et al., 2016).

Estimates show that 50% of global soil erosion is caused by water (Blanco-Canqui & Lal, 2010). The Brazilian Southeastern region holds most of the country's economic activities (55.4%), as well as 42.6% of the Brazilian population (IBGE, 2010). Most erosion processes are observed in the Paraná River basin, where natural and anthropic conditions have intensified these processes (Anache et al., 2017; Costa et al., 2015; Dorici et al., 2016; Sparovek et al., 2007).

Several researchers rely on studies about accelerated soil erosion as core environmental degradation, which is triggered by activities such as changes in land cover, inappropriate agricultural practices, as well as wrong occupation of slope areas, water bodies' neighbourhoods' and land prone to erosion, besides associated socio-environmental issues (Mota, 1995; Zuquette et al., 2013; Carvalho et al., 2019).

Understanding erosion susceptibility as inherent to the environment, it is possible to define different potentials of predisposition to erosion-processes development depending on a combination of several conditions. Cunha and Guerra (2009) categorized the factors triggering and/or controlling erosion processes as follows: geomorphometric features, land use/cover, geological material, and rainfall erosivity.

In this context, the spatial distribution and dynamics of erosion-conditioning factors are typically assessed using cartographic approaches, particularly through susceptibility maps. This can be achieved using different information treatment techniques, as observed in Pons, Pejon and Zuquette (2007), Di Lollo and Sena (2013), Galharte, Villela and Crestana (2014), Durães and Mello (2016), Javidan et al. (2019) and Prieto-Amparán et al. (2019).

Despite the increasing concerns over soil erosion in urban environments, most existing studies focus on rural areas, leaving a significant gap in understanding the specific drivers and mechanisms of linear erosion in cities. Traditional approaches often emphasize natural factors, such as geomorphology and rainfall erosivity, but fail to adequately incorporate anthropogenic influences, which are dominant in urban settings. This study advances the field by integrating both natural and human-induced variables—such as upstream and downstream land use—into an erosion susceptibility analysis, employing a Classification and Regression Tree (CART) model. While decision tree-based methods have been widely applied in environmental modeling, their use in urban erosion studies remains limited. By applying this methodological approach to a densely populated and highly urbanized environment, this research provides new insights into the role of land use dynamics and stormwater management in shaping urban erosion processes. Moreover, the framework developed is not limited to a specific geographic region and can be adapted to assess erosion susceptibility in various urban contexts worldwide.

Acknowledging the relevance of this topic to help better understand erosion processes' triggering and evolution, and also considering the scarcity of studies focusing on urban areas, the aim of the present study was to assess the influence of natural and anthropic conditions on linear erosion processes observed in São Paulo State urban areas.

1.1 Gully erosion in urban areas

Urban linear erosion is widely recognized as a direct outcome of anthropogenic disturbances to local hydrological systems. Rapid and frequently unplanned urban growth modifies natural runoff dynamics through vegetation removal, soil compaction, surface sealing, and the artificial concentration of flow along streets, plots, and drainage structures. These changes increase runoff volume and velocity, allowing surface water to exceed soil resistance and initiate incision processes that may rapidly evolve into larger erosional features

Rev. Bras. Cartogr, vol. x, n. x, aaaa DOI: http://dx.doi.org/10.14393/rbcvxxnx-xxxx

(Júnior et al., 2010; De Albuquerque et al., 2020; Mawe et al., 2025).

A consistent finding in the literature is the central role of roads, unpaved surfaces, and deficient stormwater drainage in triggering urban gullies. Streets, road embankments, drainage ditches, and concentrated discharge points often function as preferential flow paths, promoting headcut initiation and channel deepening, particularly on steep or topographically convergent slopes (Gudiño-Elizondo et al., 2018; Rahmati et al., 2022; Imwangana et al., 2025). Evidence from both Global South cities and Brazilian case studies shows that even areas with moderate natural susceptibility can develop severe erosion once urban hydrology is profoundly altered (Zolezzi et al., 2018; Mawe et al., 2024).

Natural controls such as slope gradient, soil texture, lithology, and groundwater behavior strongly condition the spatial distribution and rate of erosion, but their influence is commonly amplified by human interventions. Sandy or deeply weathered soils, typical of many tropical and subtropical urban environments, are particularly vulnerable when combined with increased surface runoff and reduced infiltration capacity (Mahamba et al., 2023; Chen et al., 2024; IPT, 2012). Several studies indicate that erosion accelerates once channels intersect the groundwater table, intensifying headward retreat and lateral expansion (Frankl et al., 2020; Mawe et al., 2024).

Urban linear erosion produces substantial environmental, social, and economic impacts, including sediment delivery to downstream channels, reservoir siltation, infrastructure damage, housing loss, land devaluation, and population displacement (Kuhn et al., 2024; Mawe et al., 2025). These effects are most acute in urban peripheries, where informal or poorly serviced settlements frequently coincide with steep slopes and inadequate drainage infrastructure, a pattern long documented in Brazilian cities (Almeida Filho et al., 2001).

Overall, urban linear erosion reflects a structural mismatch between land-use practices and the physical capacity of the landscape. Contemporary studies converge in showing that effective mitigation relies less on isolated corrective works and more on integrated strategies combining land-use regulation, stormwater management, vegetation cover maintenance, and planning informed by geomorphological and hydrological constraints (Frankl et al., 2020; Mawe et al., 2025).

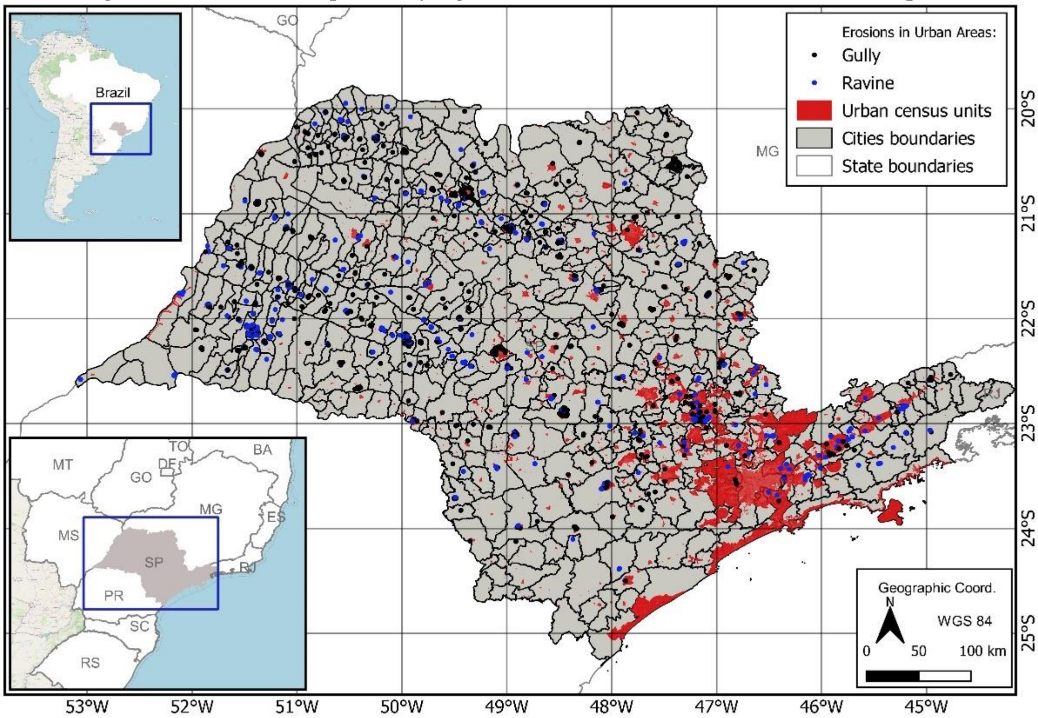
2 MATERIALS AND METHODS

São Paulo State was chosen as the study area, with a focus on urban regions defined by census-based urban-rural classifications (IBGE, 2010). The erosion features analyzed in this study were obtained from the official inventory produced by the Technological Research Institute (IPT). According to the methodological report (IPT, 2012), the identification of urban erosion processes was carried out through visual interpretation of high-resolution satellite images available in Google Earth (SPOT-5 and GeoEye-1), complemented by the analysis of 1:50,000 IBGE topographic maps. The erosion features identified through image interpretation were subsequently evaluated and systematized in standardized inventory forms, followed by field surveys, contributing to the improvement of database quality. Therefore, the present study did not perform the delimitation or validation of erosion features inventoried in 2010, relying instead on this previously established and validated official dataset. The location of São Paulo state, the identification of urban census sectors and the 1,398 analyzed urban erosion features (949 gullies; 449 ravines) are shown in Figure 1.

It is important to emphasize that the classification of linear erosion features adopted in this study follows the conceptual framework used by the IPT, which is consistent with a widely applied geomorphological tradition in Brazilian erosion studies. According to this approach, linear erosion developed from rills - shallow and narrow incisions formed by concentrated surface runoff – to ravines. Ravines are elongated erosive features, generally longer than wide, with variable depths, limited lateral branching, and without interception of the groundwater table. Their lateral expansion results from surface runoff within the channel, promoting undercutting at the base of the slopes and subsequent mass movements. Minimum depths reported in the literature for ravines range from approximately 30 cm (Tricart, 1977) to 50 cm (Imeson & Kwaad, 1980). When surface runoff processes evolve to intercept the groundwater table, a combination of surface and subsurface erosion mechanisms leads to the development of gullies (locally referred to as voçorocas), which are larger and more complex erosive forms associated with processes such as piping, sand liquefaction, and slope failures (Oliveira, 1994; Salomão, 1994).

 DOI: http://dx.doi.org/10.14393/rbcvxxnx-xxxx

Figure 1 – Location map identifying urban census sectors and urban erosion spots.



Source: Authors (2026).

This study therefore adopts the erosion typologies as defined in the IPT inventory, acknowledging that alternative classifications may exist in the international literature. The investigation aimed not only to understand this phenomenon but also to identify the data required to guide the analysis and methodological choices. Accordingly, two main steps were established: (i) data extraction and processing and (ii) data mining.

2.1 Data extraction and processing step

This stage consisted of extracting data of interest from several sources and processing them in the Geographic Information System (GIS) environment of ArcGIS Pro 2.8.3 software (Esri, 2021). A database of ravine and gully linear erosion types mapped by IPT (2012) was elaborated by associating features related to the four categories presented in the introduction section: substrate-related factors (source: IPT, 2012), land use (sources: IPT, 2012; MapBiomas Project, 2010), rainfall erosivity (source: Ricardi & Lima, 2022) and geomorphometric features (source: CATI, 2016).

Each mapped erosion feature was treated as an individual analytical unit and used as the reference element for database construction. Attributes related to substrate characteristics and part of the land use information were already available in the original IPT (2012) erosion inventory and, therefore, did not require spatial overlay procedures, being subjected only to a preprocessing stage for the organization and standardization of categorical variables, as detailed in the subsequent sections. The remaining datasets—land use information derived from the MapBiomas Project, rainfall erosivity, and geomorphometric features obtained from the Digital Elevation Model (DEM)—were spatially intersected with the linear features representing gullies and ravines. As a result, a structured geodatabase was developed in which each erosion feature (row) was associated with a set of attributes (columns) corresponding to the dependent variables of the four analytical categories considered in this study. This procedure enabled the integration of heterogeneous data sources into a single, consistent database for subsequent data mining analyses.

2.1.1 SUBSTRATE-RELATED FACTORS

Data related to substrate characteristics were already available in feature 'assessed database source standard'. They were the only ones not requiring reclassification or processing because they were organized into well-structured categories.

Rev. Bras. Cartogr, vol. x, n. x, aaaa DOI: http://dx.doi.org/10.14393/rbcvxxnx-xxxx

Substrate-related features were observed in features mapped by IPT, namely: geomorphology, geology and pedology. Possible occurrences in each feature are shown in Chart 1.

Chart 1 – Substrate-related features and occurrences linked to erosion processes.

Attributes	Occurrence
Geomorphology[1] (based on Ross & Moroz, 1997)	Lower Ribeira Depression; Middle Paraíba Depression; Paulista Peripheral Depression; Atlantic Plateau; São Paulo Plateau; Western Paulista Plateau; Coastal and Fluvial Plains.
Geology[2] (based on DAEE/UNESP, 1982)	Coastal Complex – Migmatites; Embu Complex – Migmatites; Pinhal Complex – Migmatites; Silvianópolis Complex – Embrechitic gneiss; Silvianópolis Complex – Granulites; Alluvial deposits; Colluvial deposits; Summit deposits; Caçapava Formation; Rio Claro Formation; São Paulo Formation; Tremembé Formation; Açungui Group – Mica schists; Amparo Group – Paragneiss; Bauru Group – Adamantina Formation; Bauru Group – Caiuá Formation; Bauru Group – Marília Formation; Bauru Group – Santo Anastácio Formation; Paraná Group – Furnas Formation; Passa Dois Group – Corumbataí Formation; São Bento Group – Botucatu Formation; São Bento Group – Pirambóia Formation; São Bento Group – Serra Geral Formation; São Roque Group – Phyllites; São Roque Group – Schists; Tubarão Group – Aquidauana Formation; Tubarão Group – Tatuí Formation; Tubarão Group – Itararé Subgroup; Basic suites; Granitoid suites.
Pedology[3] (based on Oliveira et. al., 1999)	Área urbana – solo não classificado (Urban soils / Technosols); Argissolos Vermelho-Amarelos (Acrisols); Argissolos Vermelhos (Acrisols); Cambissolos Háplicos (Cambisols); Gleissolos Melânicos (Gleysols); Latossolos Amarelos (Ferralsols); Latossolos Vermelho-Amarelos (Ferralsols); Latossolos Vermelhos (Ferralsols); Neossolos Litólicos (Leptosols); Neossolos Quartzarênicos (Arenosols); Nitossolos Vermelhos (Nitisols).

[1] Geomorphological units were referenced according to the official Brazilian geomorphological map, with translation limited to generic geomorphological terms while preserving original toponyms.
[2] Geological units follow the official nomenclature of the São Paulo State geological mapping. Formal lithostratigraphic names were preserved, while lithological terms were translated into English.
[3] Soil classes according to the Brazilian Soil Classification System (SiBCS). For international reference, the corresponding classes according to the World Reference Base for Soil Resources (FAO, 2022) are also indicated. The conversion to WRB was performed at the Reference Soil Group level, which does not explicitly account for color attributes used in the SiBCS.

Source: Authors (2026).

2.1.2 LAND USE

Land use data, although already belonged to the original source's features, refer to textual information and were not set into categories, which is a requirement for the CART decision tree algorithm. Therefore, it was necessary to organize the collected information into established categories to make the data mining technique application feasible.

These data's reorganization focused on originally-divergent textual data due to their plural and singular use, to terms' reversed ordering, to detailed featuring in some cases – but generalized in others (for example: surface runoff and rainwater runoff from the urban area discharged at the river head), among others. This process led to reduction ranging from 82 to 10 possible occurrences set for upstream use, from 89 to 8 for downstream use data, and from 134 to 5 for triggering action data. The land use attributes, divided into "upstream use", "downstream use" – both in comparison to the mapped erosion features – and "triggering action", as the established categories are shown in Chart 2.

Considering data descriptive nature, requiring categorization due to its large number of classes, option was made for incorporating into the analyses the land cover data extracted from the MapBiomas Project for the year 2010 (the year of data collection). This incorporation was justified by the greater standardization of classes and by enabling a possible comparison in future applications of this same methodology, as the case may be. This incorporation, included in Chart 2, also allowed the attenuation of any generalizations resulting from the categorization step.

Chart 2 – Land use features and occurrences linked to erosion processes.

Attributes	Occurrence
Upstream use	Urban; Urban in consolidation; Urban and Rail/Highway (or Rail/Highway only); Urban in consolidation and Railway/Highway; Urban and Rural; Urban in consolidation and Rural; Rural; Rural and Railway/Highway (or Railway/Highway only); Drainage; Vegetation
Downstream use	Urban; Urban in consolidation; Urban and Rural; Urban in consolidation and Rural; Rural; Rural and Vegetation; Drainage; Vegetation

Rev. Bras. Cartogr, vol. x, n. x, aaaa DOI: http://dx.doi.org/10.14393/rbcvxxnx-xxxx

Attributes	Occurrence
Triggering actions	Fluvial Erosion; Surface runoff; Stormwater disposal (pipe); Surface runoff and stormwater disposal; Railroad/Highway stormwater Runoff/Disposal
Land cover (Mapbiomas Project, 2010)	Farming; Urbanized Area; Forest; Non-Forest Natural Formation; Agriculture and Pasture Mosaic

Source: Authors (2026).

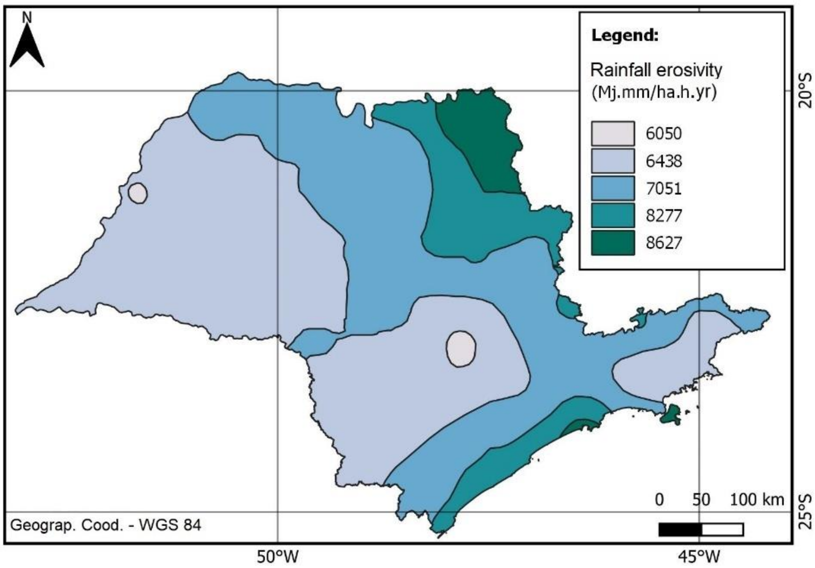
For the upstream and downstream use attributes, occurrences categorized as Rural include agricultural use, pasture (with and without terraces) and exposed soil. Urban features include residential, commercial and industrial uses. It is also important to mention that, although the attribute "triggering action" does not specifically regard soil use, it was added to the analysis due to its association with anthropic activities that influence the environmental soil dynamics.

2.1.3 RAINFALL EROSIVITY

Data on rainfall erosivity were obtained from the study conducted by Ricardi and Lima (2022), who updated rainfall erosivity estimates set for São Paulo State based on yearly and monthly averages between 1997 and 2017. These data resulted from the spatial extrapolation of measurements taken from the state's rain gauge stations during the aforementioned time interval – these datasets were made available by the Water and Electric Power Department (DAEE, 2019).

In order to add these data to the herein proposed analyses, the vectorization of the cartographic product about rainfall erosivity was carried out in GIS environment, in the form of polygons (see Figure 2), followed by the intersection with the erosion records - resulted in values ranging from 6,050 to 8,627 Mj.mm/ha.h.yr. In total, 10 of the 1,398 points were located in regions presenting erosivity of 6,050 Mj.mm/ha.h.yr, 718 were in regions with erosivity 6,438 Mj.mm/ha.h.yr, 520 were in places with erosivity 7,051 Mj.mm/ha.h.yr, 80 were in location with erosivity 8,277 Mj.mm/ha.h.yr, and 70 were in regions with erosivity 8,627 Mj.mm/ha.h.yr.

Figure 2 – São Paulo State Rainfall Erosivity Map for the 1997-2017.



Source: Authors (2026).

2.1.4 GEOMORPHOMETRIC VARIABLES

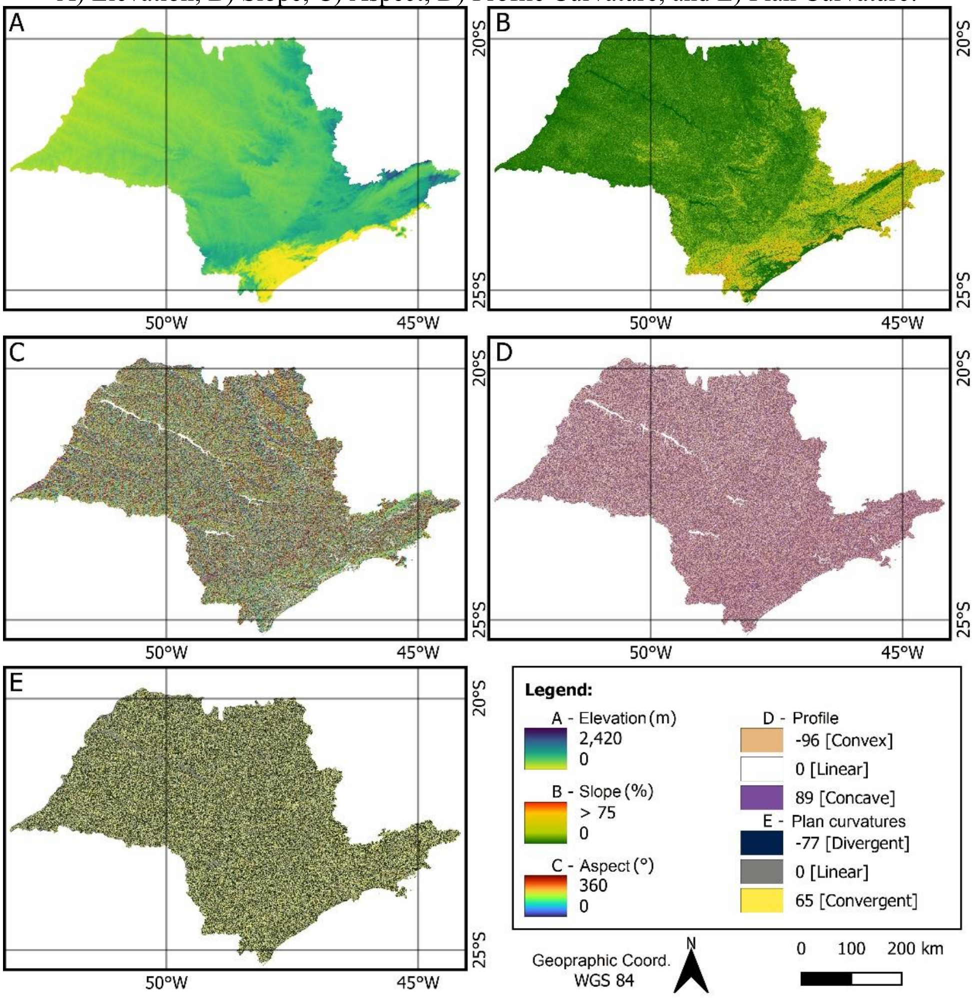
In the relief study the following features were analyzed: elevation, slope, aspect, profile and plan curvature. The Digital Elevation Model (DEM) prepared by the Brazilian Farming Information Center (CATI, 2016) was used for data extraction purposes. It was elaborated based on products from NASA's Shuttle Radar Topography Mission (SRTM), with 30m spatial resolution; corrected with data from NASA and METI's ASTER GDEM V2 sensor, also with 30m spatial resolution, and geoid ripple for São Paulo State.

Rev. Bras. Cartogr, vol. x, n. x, aaaa DOI: http://dx.doi.org/10.14393/rbcvxxnx-xxxx

First of all, based on the aforementioned base, a hypsometric map was elaborated, as well as maps of slope (using the Slope GIS tool), aspect (using the Aspect GIS tool), profile and plan curvature (using the Curvature GIS tool). It was decided not to categorize these data, given their continuous numerical features. Altitudes were recorded in meters, slope in percentage, aspect in degrees (azimuth angle corresponding to the greatest slope of the terrain), profile curvature in negative values for convex, in positive values for concave and in null values for linear; and plan curvature in positive values for divergent, in negative values for convergent and in null values for linear.

Data forming geomorphometric characterization - DEM-based geoprocessing - led to 5 intermediate maps (Figure 3). As from erosive processes records, its locations were crossed with the produced geomorphometric layers – through the Intersect GIS tool. Subsequently, elevation, slope, aspect, profile curvature and plan curvature values recorded for each point were extracted.

Figure 3 – Maps of geomorphometric variables.
A) Elevation; B) Slope; C) Aspect; D) Profile Curvature; and E) Plan Curvature.



Source: Authors (2026).

Altitudes ranged from 0 to 2,420 meters, being the mean altitude 557.36 meters and the most frequent one 487.50 meters. Mean slope was close to 11%, presenting the highest occurrence at the range from 4% to 6% - slopes above 20% remained lower than 10% of all cases. Aspect, which ranged from 0° to 360°, recorded quite varying occurrences, with prevalence in Northeastern (45°), Southeastern (135°), Southwestern (225°) and Northwestern (314°). With respect to the profile curvature, 43.27% of collected data referred to positive values, assigned to concave profile; 49.49% to negative values, assigned to convex profile; and 7.24% to null values, which corresponds to curvature-free regions. Finally, regarding plan curvature, it was possible observing 39.32% positive values, which designate convergent planes; 48.49% negative values, designating divergent planes; and 12.19% null values that underline linear planes.

2.2 Data mining step

The application of data mining methods in studies focused on linear erosion processing have shown overall upward trends, mainly from 2018 onwards (according to search in Scopus database carried out by the present authors in December 2022). Random Forest, Decision Tree and Artificial Neural Networks were the most often applied methods. In this work, the Decision Tree (DT) was chosen because it is a simplified model capable of representing associations between a given dataset (independent variables) and a dependent variable (Medeiros et al., 2014; Quinlan, 1983). IBM SPSS Statistics 20 software (2011) was used for method application purposes.

This technique runs a decision-making procedure (bipartition of occurrences) based on algorithms that classify data according to their homogeneity towards the investigated variable. CART is one of the most often applied algorithms (Breiman et al., 1984), which was also used in the current research. It adopts the Gini Index (GI) to achieve data binary partition – it can range from 0 to 1: the closer to 0, the purer the node, consequently, the more homogenous. Among other words, observations branch from independent variables' attributes - taking into consideration the highest occurrence of a single category of the dependent variable (Breiman et al., 1984).

The decision-making procedure is repeated across multiple levels of the tree. Therefore, several hyperparameter configurations were systematically tested in order to control tree complexity, optimize predictive performance, and address class imbalance between erosion types. Model tuning focused on the main stopping rules of the CART algorithm: maximum tree depth, minimum number of cases in parent and child nodes, and minimum improvement in node impurity based on the Gini Index.

All experiments were conducted using a 10-fold cross-validation procedure, with the maximum tree depth fixed at 10. The tuning strategy was organized into three stages. In the first stage, models without misclassification costs were tested using two alternative minimum node size configurations (49/7 and 70/14 cases in parent and child nodes, respectively), aiming to assess the trade-off between model complexity and classification balance. In the second stage, an asymmetric cost matrix was introduced to mitigate class imbalance, and the same node size configurations were compared. In the third stage, the impact of tree pruning was evaluated by varying the minimum Gini improvement threshold (0.001, 0.002, and 0.003) under the selected cost structure.

Preliminary tests also explored the use of class weights; however, these configurations resulted in reduced predictive performance and were therefore not retained. Instead, an asymmetric misclassification cost matrix was adopted, in which no cost was assigned to correct classifications, while higher penalties were imposed on misclassification errors. Misclassifying a gully as a ravine was assigned a cost of 1.0, whereas misclassifying a ravine as a gully was assigned a higher cost of 1.5. This configuration reflects the greater analytical and practical relevance of correctly identifying gullies, which generally represent more advanced and severe erosion processes.

Model performance was primarily evaluated using cross-validation risk and overall accuracy, with particular emphasis on class-specific accuracies in order to assess balance between ravines and gullies. The final model was selected based on a combination of (i) predictive performance, (ii) balance between classes, and (iii) model parsimony, favoring configurations that reduced overfitting while maintaining interpretability.

In accordance with the implementation available in IBM SPSS Statistics, the average overall accuracy across the folds was used as the primary metric for cross-validation-based model comparison. The Area Under the Receiver Operating Characteristic Curve (AUC–ROC) was computed directly within the software environment. Subsequently, based on the confusion matrix obtained for the selected model, additional performance metrics—including precision, sensitivity (recall), and F1-score—were calculated to provide a more detailed evaluation of classification performance, particularly with respect to class-specific behavior (Powers, 2011). (Powers, 2011).

The attributes used in this investigation, as well as their likely occurrences, were described in item 2.1. The linear erosion type (with two associated categories: ravine and gully) was adopted as dependent variable, and the other variables were the independent variables, as described in Chart 3 (which also indicates the variable's typology).

Rev. Bras. Cartogr, vol. x, n. x, aaaa DOI: http://dx.doi.org/10.14393/rbcvxxnx-xxxx

Chart 3 – Features of the investigated variables.

Type	Nature	Variable	Typology
Dependent	Linear erosion	Ravines	Textual
		Gullies	Textual
Independent	Substrate-related	Geomorphology	Textual
		Geology	Textual
		Pedology	Textual
	Land use	Upstream use	Textual
		Downstream use	Textual
		Triggering action	Textual
		Land cover (MapBiomas)	Textual
	Erosivity	Rainfall erosivity	Numeric
	Geomorphometric	Elevation	Numeric
		Slope	Numeric
		Aspect	Numeric
		Profile curvature	Numeric
		Plan curvature	Numeric

Source: Authors (2026).

3 RESULTS AND DISCUSSION

According to IPT data, 41,262 ravine and gully linear erosion processes spots were identified and characterized in São Paulo State, of which 1,398 were observed in urban areas. If one has in mind that 96,7% of the total population in the state lives in these areas (IBGE, 2022), then, one can see the relevance of learning more about the association between triggered erosion processes in urban areas.

Therefore, the herein suggested data mining methodology aimed to establish these associations by investigating the most important agents for erosion occurrences. The data extraction and processing stage was crucial for structuring the dataset to ensure compatibility with the software and for integrating attributes already recognized in the literature as influential. The data mining step could start once values are extracted from the target features of each of the 1,398 erosion process features.

3.1 Data Mining Step

To evaluate the sensitivity of the CART classifier to hyperparameter choices and to address class imbalance between ravines and gullies, a set of six models was trained and assessed using 10-fold cross-validation. The tested configurations varied the minimum number of cases in parent and child nodes, the minimum improvement in Gini impurity required for splitting, and the use of asymmetric misclassification costs. Table 1 summarizes the structural characteristics of each tree and their respective predictive performances, including class-specific accuracies, overall accuracy, and resubstitution and cross-validation risks.

Table 1 – Summary of CART hyperparameter tests results.

Model	Misclassification Costs		Minimum samples		Min. Gini	Number of		Tree Depth	Accuracy (%)			Risk		
	Rav.	Gully	Parent Node	Child Node		Nodes	Terminal Nodes		Rav.	Gully	Overall	Resubstitution	CV	Diference
1	0.00	0.00	49	7	0.001	49	25	9	51.2	88.7	76.7	0.233	0.336	0.103
2	0.00	0.00	70	14	0.001	31	16	8	57.5	82.6	74.5	0.255	0.300	0.045
3	1.00	1.50	49	7	0.001	49	25	9	74.2	76.0	75.4	0.288	0.418	0.130
4	1.00	1.50	70	14	0.001	31	16	8	75.1	72.4	73.2	0.308	0.407	0.099
5	1.00	1.50	70	14	0.002	29	15	7	79.5	69.0	72.4	0.309	0.413	0.104
6	1.00	1.50	70	14	0.003	19	10	5	80.8	65.0	70.1	0.330	0.399	0.069

Source: Authors (2026).

In the first stage (Models 1 and 2), only the minimum number of cases in parent and child nodes was varied, while no misclassification costs were applied. Both models achieved relatively high overall accuracies

(76.7% and 74.5%, respectively), but exhibited strong class imbalance, with substantially higher accuracy for gullies than for ravines. Model 1, with smaller minimum node sizes (49–7), produced a deeper and more complex tree (depth = 9), whereas Model 2 (70–14) resulted in a simpler structure (depth = 8) with slightly lower overall accuracy but reduced risk difference, suggesting improved generalization.

In the second stage (Models 3 and 4), an asymmetric misclassification cost matrix was introduced to mitigate class imbalance. The inclusion of costs markedly improved the balance between ravine and gully accuracies. Compared to their cost-free counterparts, both models exhibited a substantial increase in ravine accuracy, indicating that the cost structure effectively penalized misclassification of ravines. Model 4 (70–14) achieved a more compact tree while maintaining comparable class-specific accuracies, demonstrating that the cost matrix improved class balance without excessively increasing model complexity.

In the third stage (Models 4, 5, and 6), the minimum Gini improvement required for splitting was increased from 0.001 to 0.002 and 0.003 while maintaining the same node size and cost configuration. As the Gini threshold increased, tree complexity was progressively reduced, with a decrease in both depth and number of terminal nodes. However, this simplification came at the expense of predictive balance: although ravine accuracy continued to increase, gully accuracy declined substantially, leading to lower overall accuracy and reduced class equilibrium in Models 5 and 6. These results indicate that overly restrictive splitting criteria can oversimplify the model and degrade classification performance.

Based on the joint evaluation of predictive performance, class balance, and model complexity, Model 4 was selected as the basis for the final classifier. Although it did not achieve the highest overall accuracy among all tested configurations, it provided the best trade-off between ravine and gully accuracies, avoiding the strong class bias observed in cost-free models and the performance degradation associated with more restrictive Gini thresholds. To further enhance generalization and interpretability, this model was subsequently pruned to a maximum tree depth of five levels.

This pruning strategy substantially reduced structural complexity while preserving virtually the same predictive performance, resulting in a more parsimonious and robust representation of erosion processes. The final pruned model maintains a balanced classification behavior between erosion types and exhibits a moderate difference between resubstitution and cross-validation risks, indicating good generalization capacity and limited susceptibility to overfitting. This configuration therefore represents a robust and interpretable solution for mapping urban linear erosion processes.

Table 2 summarizes the metrics derived from the confusion matrix of the selected pruned model, as well as AUC values and confidence interval (CI) from ROC analysis based on predicted probabilities for each erosion class under 10-fold cross-validation. The final pruned CART model achieved an overall accuracy of 72.7%, with class-specific performance metrics indicate a balanced predictive behavior between erosion types. For ravines, the model obtained a recall of 71.3%, demonstrating good sensitivity in identifying this class, although with moderate precision (55.9%), reflecting residual confusion with gullies. For gullies, precision reached 84.4%, while recall was 73.4%, indicating a high reliability of positive predictions and a satisfactory detection rate.

Table 2 – Performance metrics of the final CART model.

Metric	Ravine	Gully
Accuracy	0.727	
Precision	0.559	0.844
Recall	0.713	0.734
F1-score	0.626	0.785
AUC (ROC)	0.780	
95% CI (AUC)	0.755–0.805	
Resubstitution Risk	0.319	
Cross-Validation Risk	0.397	

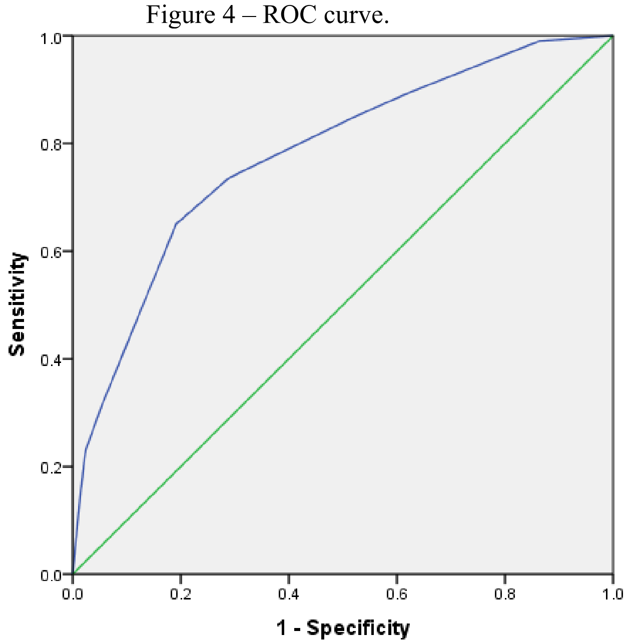
Source: Authors (2026).

The resulting F1-scores (0.626 for ravines and 0.785 for gullies) confirm that the adopted cost-sensitive configuration effectively mitigated the strong class bias observed in preliminary models without misclassification costs. Importantly, the relatively small gap between resubstitution and cross-validation risks

(0.078) further suggests improved generalization and reduced overfitting due to pruning.

ROC analysis further supports the robustness and generalization capacity of the final pruned CART model. The ROC curve yielded an AUC of 0.780, with a 95% confidence interval ranging from 0.755 to 0.805 ($p < 0.001$), indicating good discriminatory ability between erosion types. This result confirms that the classifier performs substantially better than random assignment and maintains consistent predictive power across different probability thresholds. The shape of the ROC curve (see Figure 4) reveals a stable trade-off between sensitivity and specificity over a broad range of cutoff values, suggesting that the model's performance is not overly dependent on a single decision threshold. This behavior is particularly relevant in urban erosion studies, where management priorities may favor either higher sensitivity (to avoid missing severe gully features) or higher specificity (to reduce false alarms), depending on planning and mitigation objectives. Therefore, the ROC analysis reinforces the suitability of the selected model for operational and decision-support contexts, complementing class-specific metrics and supporting the interpretation that the adopted pruning strategy effectively balances predictive performance and model generalization.

Figure 4 – ROC curve.



Note: diagonal segment produced by ties. Source: Produced by the authors at SPSS tool (2026).

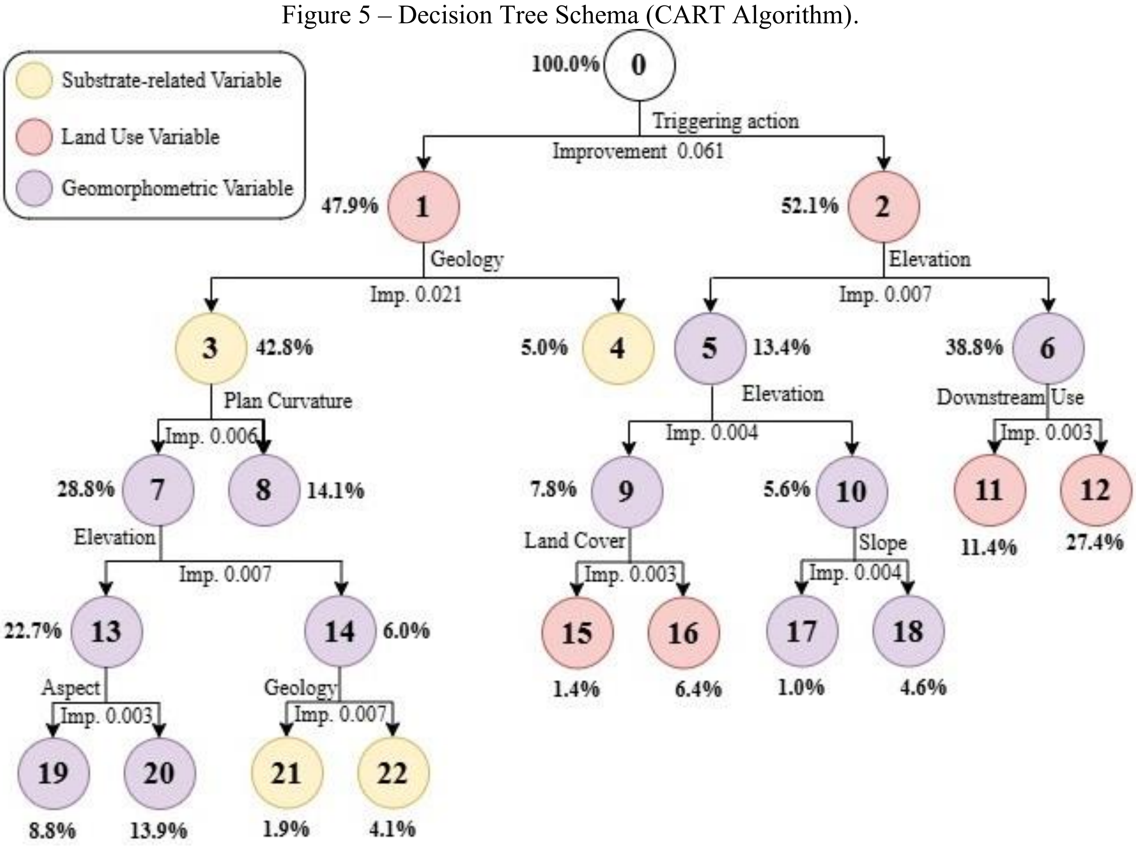
Figure 5 depicts the final CART decision tree obtained after hyperparameter optimization and the incorporation of asymmetric misclassification costs. The tree structure represents the hierarchical decision process based on the explanatory variables, with each node corresponding to a subset of erosion features that is increasingly homogeneous with respect to erosion type. Node purity is measured by the Gini Index, and each split is associated with an improvement value indicating the reduction in impurity. Colored nodes highlight the thematic group of each explanatory variable, distinguishing substrate-related variables, land-use factors, and terrain characteristics.

The model resulted in a tree with moderate depth (5) and 23 nodes - being 12 of them terminal nodes - ensuring both interpretability and generalization. The classification outcomes associated with each node are detailed in Table 3, which reports the relative frequencies of ravines and gullies within each partition. The algorithm calculates the improvement value (Gini Index) of each branching, which ranged from 0.003 to 0.061. These values measure how much the homogeneity related to the dependent variable in the child node was enhanced in comparison to the parent nodes.

The simultaneous analysis of both – Figure 5 and Table 3 - enables identifying: (i) by which values of the explanatory variables (independent) the resulting nodes are characterized, at different levels; and (ii) which nodes presented the highest frequency percentages recorded for erosion types (ravine and gully). Thus, it is possible to establish the factors influencing the prevailing occurrence of a given erosion type in comparison to the others.

Rev. Bras. Cartogr, vol. x, n. x, aaaa DOI: http://dx.doi.org/10.14393/rbcvxxnx-xxxx

Figure 5 – Decision Tree Schema (CART Algorithm).



Source: Authors (2022).

Table 3 – List of Decision Tree nodes and divisions.

Node	Parent Node	Gully		Ravine		Explanatory variable	Resulting Cluster
		N	%	N	%		
0	(root)	449	32.1	949	67.9	Type of Erosion	
1	0	324	48.4	345	51.6	Triggering action	Surface runoff; Railroad/Highway stormwater Runoff/Disposal
2	0	125	17.1	604	82.9	Triggering action	Fluvial Erosion; Stormwater disposal (pipe); Surface runoff and stormwater disposal
3	1	263	43.9	336	56.1	Geology	Silvianópolis Complex – Embrechitic gneiss; Silvianópolis Complex – Granulites; Alluvial deposits; Colluvial deposits; Summit deposits; Caçapava Formation; Rio Claro Formation; Amparo Group – Paragneiss; Bauru Group – Adamantina Formation; Bauru Group – Marília Formation; Bauru Group – Santo Anastácio Formation; Paraná Group – Furnas Formation; São Bento Group – Botucatu Formation; São Bento Group – Pirambóia Formation; São Bento Group – Serra Geral Formation; São Roque Group – Phyllites; Tubarão Group – Aquidauana Formation; Tubarão Group – Tatuí Formation; Tubarão Group – Itararé Subgroup; Basic suites.
4	1	61	87.1	9	12.9	Geology	Coastal Complex – Migmatites; Embu Complex- Migmatites; Pinhal Complex- Migmatites; São Paulo Formation; Tremembé Formation; Açungui Group - Mica schists; Bauru Group – Caiuá Formation; Passa Dois Group – Corumbataí Formation; São Roque Group – Schists; Granitoid Suites.
5	2	56	29.9	131	70.1	Elevation	≤ 453.0
6	2	69	12.7	473	87.3	Elevation	> 453.0
7	3	156	38.8	246	61.2	Plan Curvature	≤ 0.0036
8	3	107	54.3	90	45.7	Plan Curvature	> 0.0036
9	5	22	20.2	87	79.8	Elevation	≤ 416.5
10	5	34	43.6	44	56.4	Elevation	> 416.5
11	6	7	4.4	152	95.6	Downstream Use	Urban in consolidation; Urban and Rural; Urban in consolidation and Rural; Rural and Vegetation
12	6	62	16.2	321	83.8	Downstream Use	Urban; Rural; Drainage; Vegetation
13	7	139	43.7	179	56.3	Elevation	≤ 660.0

Node	Parent Node	Gully		Ravine		Explanatory variable	Resulting Cluster
		N	%	N	%		
14	7	17	20.2	67	79.8	Elevation	> 660.0
15	9	9	47.4	10	52.6	Land Cover	Farming
16	9	13	14.4	77	85.6	Land Cover	Urbanized Area; Forest; Non-Forest Natural Formation; Agriculture and Pasture Mosaic
17	10	1	7.1	13	92.9	Slope	≤ 2.76039
18	10	33	51.6	31	48.4	Slope	> 2.76039
19	13	43	35.0	80	65.0	Aspect	≤ 128.347438
20	13	96	49.2	99	50.8	Aspect	> 128.347438
21	14	14	51.9	13	48.1	Geology	Tubarão Group – Itararé Subgroup; Silvianópolis Complex – Granulites; São Bento Group – Pirambóia Formation; Alluvial deposits; Silvianópolis Complex – Embrechitic gneiss.
22	14	3	5.3	54	94.7	Geology	Colluvial deposits; Summit deposits; Caçapava Formation; Rio Claro Formation; Amparo Group – Paragneiss; Bauru Group – Adamantina Formation; Bauru Group – Marília Formation; Bauru Group – Santo Anastácio Formation; Paraná Group – Furnas Formation; São Bento Group – Botucatu Formation; São Bento Group – Serra Geral Formation; São Roque Group – Phyllites; Tubarão Group – Aquidauana Formation; Tubarão Group – Tatuí Formation; Basic suites.

Source: Authors (2022).

At the root level (Node 0), the algorithm selected “Triggering Action” as the most informative variable, yielding the highest normalized importance (100%). This first bifurcation separates the dataset into two nearly balanced branches (47.9% and 52.1% of the observations), while already producing distinct class compositions: Node 1 presents a near equilibrium between ravines (48.4%) and gullies (51.6%), whereas Node 2 is strongly dominated by gullies (82.9%). This result confirms the central role of runoff concentration, stormwater disposal, and fluvial processes in controlling the spatial differentiation between ravines and gullies in urban environments..

Although this first level presents the highest Gini Index value (0.061), the variable responsible for data partitioning, “Triggering Action”, was identified by the algorithm as the most important variable for the classification, as pointed out in Table 4, which relates the independent variable according to its relevance for the classification process. This finding aligns with studies by Guo et al. (2019), which emphasize the importance of stormwater disposal in accelerating erosion processes in urbanized settings. However, unlike rural contexts, where agricultural practices dominate as erosion triggers, urban areas present unique challenges, such as concentrated runoff, inadequate stormwater infrastructure, and impervious surfaces. This highlights the necessity of urban-specific approaches to manage these processes effectively.

Table 4 – Importance relation of independent variables selected by the CART algorithm.

Independent Variable		Importance	Normalized Importance (%)
1	Triggering action	0.062	100.0
2	Geology	0.049	78.9
3	Elevation	0.036	57.2
4	Geomorphology	0.034	53.7
5	Land cover	0.024	38.8
6	Slope	0.023	37.6
7	Downstream use	0.022	36.0
8	Plan curvature	0.013	20.2
9	Profile curvature	0.009	14.7
10	Erosivity	0.008	13.3
11	Aspect	0.007	10.5
12	Pedology	0.006	9.7
13	Upstream use	0.005	7.7

Note: Colors indicates variable type and are consistent with those used in Figure 5. Source: Authors (2026).

After “Triggering Action”, related to land use, variables “Geology” (related to substrate characteristics) and “Elevation” (related to geomorphometric variables) were the ones better explained data variability. These variables are accountable for data bipartition at the next level (2). Variable “Rainfall

Rev. Bras. Cartogr, vol. x, n. x, aaaa DOI: http://dx.doi.org/10.14393/rbcvxxnx-xxxx

Erosivity” only appears in the 11th position.

By observing node 4, which is explained by variable “Geology”, it is possible to notice that the geological formations cluster in this node holds 87.1% ravine occurrences, indicating less appropriate formations to gullies’ development. The geological formations found in node 3 does not lead to a clear conclusion about erosion types, however, as the analyses proceed to nodes 7 and 8, it is possible to infer that plan curvatures >0.0036 (convergent) holds the highest ravine occurrences and ≤0.0036 (divergent or weakly convergent forms) holds the highest gully occurrences.

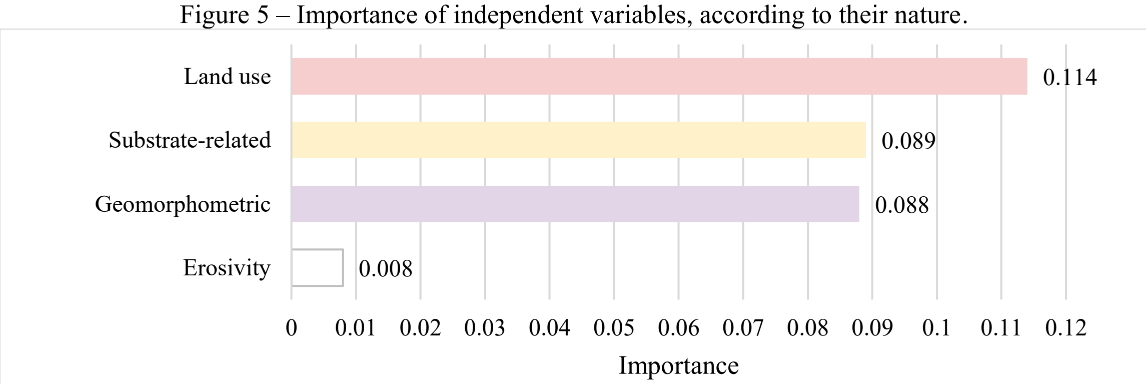
The identification of geologic formations and soil types less prone to gully formation offers valuable insights for land-use planning. These areas could serve as benchmarks for soil conservation practices, guiding interventions in more vulnerable regions. Moreover, understanding these stable conditions can inform urban development policies, ensuring that natural protective factors are preserved or even replicated in erosion-prone areas.

Although “Elevation” ranks among the most important predictors, its discriminatory capacity is limited. The elevation thresholds identified at multiple levels (453 m, 416.5 m, 660 m) generally lead to nodes still dominated by gullies, indicating that altitude alone is insufficient to clearly differentiate erosion types. This reinforces the interpretation that elevation primarily acts as a contextual variable rather than a direct driver of erosion typology.

In contrast, when elevation is examined in combination with slope, a more nuanced pattern emerges. For elevations above 416.5 m, areas characterized by slopes greater than 2.76% exhibit a slightly higher concentration of ravine features (51.6%), as observed in node 18. Although this difference is not pronounced, it suggests a modest tendency toward conditions less favorable to gully development under steeper topography at higher altitudes. Conversely, for elevations below 416.5 m associated with gentler slopes (≤ 2.76%), gully erosion becomes strongly dominant, accounting for 92.9% of the cases (node 17). This contrast indicates that, while elevation alone has limited discriminatory power, its interaction with slope plays a meaningful role in conditioning the susceptibility to the initiation and progression of gully processes.

Figure 5 summarizes the relative importance of variable classes by category (as an interpretative strategy to support overall interpretation and synthesis of results). Land-use-related variables were predominant in explaining erosion processes, highlighting human activities as the primary drivers of linear erosion in urban areas. After that, the main explicative variable is substrate-related, followed by geomorphometric variables and rainfall erosivity, as the less important factor.

Figure 5 – Importance of independent variables, according to their nature.



Note: Colors indicates variable type and are consistent with those used in Figure 5. Source: Authors (2026).

In general, the findings underline the critical role of stormwater management in mitigating urban erosion processes, as widely documented in the literature (Vanmaercke et al., 2021; Rahmati et al., 2022). This pattern does not contradict the existing literature, but rather highlights a shift in the relative importance of controlling factors in urban settings. While rainfall, soil properties, and topography remain essential predisposing conditions for erosion development (Rahmati et al., 2022; Vanmaercke et al., 2021), the modification of hydrological processes by urban land use can be decisive in determining where and how linear erosion features evolve. Similar conclusions have been reported in studies showing that poorly planned urbanization amplifies erosion susceptibility by concentrating runoff and reducing infiltration, even under comparable climatic conditions (Frankl et al., 2020; Zolezzi et al., 2018; Mawe et al., 2025).

This perspective is particularly relevant under climate-change scenarios, in which increases in rainfall intensity are expected to exacerbate erosion hazards where urban drainage systems are inadequate (Kuhn et al., 2024; Chen et al., 2024). These results provide actionable insights for decision-makers, highlighting the need for sustainable practices that integrate urban hydrology with land-use planning

4 CONCLUSIONS

Overall, the procedures applied to process data and the GIS tools incorporated to complement the target attributes investigation were effective in subsidizing the subsequent data mining technique. Based on the recorded results, the used method, decision tree supported by the CART algorithm, was efficient in identifying the variables better explaining the occurrence of each erosion type.

Substrate-related factors and terrain characteristics were shown to act as conditioning elements. The areas least prone to gully development were better explained by substrate-related factors, especially when combined with convergent plan curvature. Although elevation and slope alone showed limited discriminatory capacity, their interaction contributed to differentiating erosion behavior in specific contexts, reinforcing the importance of multivariate analysis.

The results indicate that anthropogenic factors dominate the differentiation between ravines and gullies in urban areas of São Paulo State. Among all explanatory variables, *triggering action* - closely linked to surface runoff (concentrated) and stormwater disposal (diffuse) - emerged as the most influential predictor, highlighting the central role of urban hydrological alteration in erosion processes. Land-use-related variables proved to be the most relevant for inciting the studied phenomenon since the combined explanatory importance values accounted for the highest percentage, followed by substrate-related factors, geomorphometric attributes and rainfall erosivity.

From a methodological perspective, the adoption of asymmetric misclassification costs proved essential for mitigating class imbalance and improving the identification of ravines without compromising gully detection. The subsequent pruning of the decision tree to a maximum depth of five levels resulted in a parsimonious and robust model, with balanced class performance, moderate overall accuracy (72.7%), good discriminatory ability (AUC = 0.78), and limited overfitting, as indicated by the small difference between resubstitution and cross-validation risks.

These outcomes underscore that effective mitigation of urban linear erosion depends less on isolated corrective measures and more on integrated strategies that combine land-use regulation, stormwater management, and planning informed by geomorphological constraints. Equally important is the continuous development and updating of erosion inventories such as the one used in this study, originally compiled in 2010. The maintenance of systematic, up-to-date databases would enable model validation against newly observed erosion features, strengthening its predictive reliability under real-world conditions. Moreover, advances in remote sensing, terrain modeling, and land-use mapping now allow the incorporation of higher-resolution geomorphometric and surface data, which can substantially refine susceptibility assessments. The proposed framework therefore offers a transferable and scalable methodological basis for erosion risk analysis in urban areas, while highlighting the need for sustained data acquisition efforts to support more accurate, dynamic, and resilient urban planning under ongoing urbanization and climate change scenarios.

5 BIBLIOGRAPHY

Almeida Filho, G. S., Gouveia, M., Júnior, J., & Canil, K. (2001). *Prevenção e controle da erosão urbana no Estado de São Paulo*. In Anais do 21º Congresso Brasileiro de Engenharia Sanitária e Ambiental (pp. 1-12).

Anache, J. A. A., Wendland, E. C., Oliveira, P. T. S., Flanagan, D. C. & Nearing, M. A. (2017). Runoff and soil erosion plot-scale studies under natural rainfall: A meta-analysis of the Brazilian experience. *Catena*, 152, 29–39. https://doi.org/10.1016/j.catena.2017.01.003

Arantes, L. T., Carvalho, A. C. P., Carvalho, A. P. P., Lorandi, R., Moschini, L. E. & Di Lollo, J. A. (2021).

Rev. Bras. Cartogr, vol. x, n. x, aaaa DOI: http://dx.doi.org/10.14393/rbcvxxnx-xxxx

Surface runoff associated with climate change and land use and land cover in southeast region of Brazil. *Environmental Challenges*, 3, 100054. https://doi.org/10.1016/j.envc.2021.100054

Blanco-Canqui, H. & Lal, R. (2010). *Principles of Soil Conservation and Management* (1st ed.). Springer Netherlands. https://doi.org/10.1007/978-1-4020-8709-7

Breiman, L., Friedman, J., Olshen, R. & Stone, C. (1984). *Classification and Regression Trees*. Chapman & Hall.

Carvalho, A. P. P., Guerrero, J. V. R., Silva, E. V. E., Pinto, M. J. R., Vaz, R. M. G. F. C., Pereira, C. T., Lorandi, R., Di Lollo, J. A. & Moschini, L. E. (2019). Environmental Fragility to Erosion in an Anthropogenic Watershed in the Northeast of the State of São Paulo, Brazil. *Anuário Do Instituto de Geociências - UFRJ*, 42(3), 7–18. https://doi.org/10.11137/2019_3_07_18

CATI. (2016). *Digital Elevation Model of São Paulo State*. Brazilian Farming Information Center. http://mapas.cati.sp.gov.br/mdecati.html

Chalise, D., Kumar, L. & Kristiansen, P. (2019). Land Degradation by Soil Erosion in Nepal: A Review. *Soil Systems*, 3(1), 12. https://doi.org/10.3390/soilsystems3010012

Chen, Y., Jiao, J., Yan, X., Li, J., Vanmaercke, M., & Wang, N. (2024). Response of gully morphology and density to the spatial and rainy-season monthly variation of rainfall at the regional scale of the Chinese Loess Plateau. *Catena*, 232, 107773. https://doi.org/10.1016/j.catena.2023.107773

Corcovia, L. O. & Alves, R. S. (2019). Machine Learning and Data Mining: evaluation of learning methods. *Revista Interface Tecnológica*, 16(1), 90–101. https://revista.fatectq.edu.br/interfacetecnologica/article/view/562

Costa, C. W., Lorandi, R., de Lollo, J. A., Imani, M. & Dupas, F. A. (2018). Surface runoff and accelerated erosion in a peri-urban wellhead area in southeastern Brazil. *Environmental Earth Sciences*, 77(5), 160. https://doi.org/10.1007/s12665-018-7366-x

Costa, C. W., Piga, F. G., Moraes, M. C. P. de, Dorici, M., Sanguinetto, E. de C., Di Lollo, J. A., Moschini, L. E., Lorandi, R. & Oliveira, L. J. (2015). Environmental Fragility and water scarcity in catchment basins: Headwaters of Araras River – Araras, SP. *Revista Brasileira de Recursos Hídricos*, 20(4), 946–958. http://dx.doi.org/10.21168/rbrh.v20n4.p946-958

Cunha, S. B. & Guerra, A. J. T. (2009). Environmental Degradation. In A. J. T. Guerra & S. B. Cunha (Eds.), *Geomorphology and Environment* (7th ed., pp. 337–379). Bertrand Brasil.

DAEE. (1982). *Geological map of the State of São Paulo* [Map; scale 1:250,000]. Joint project: Department of Water and Electricity and Institute of Geosciences and Exact Sciences - UNESP, Rio Claro.

DAEE. (2019). *Hydrological Database*. Department of Water and Electricity. http://www.hidrologia.daee.sp.gov.br/

De Albuquerque, A. O., de Carvalho Júnior, O. A., Guimarães, R. F., Gomes, R. A. T., & Hermuche, P. M. (2020). Assessment of gully development using geomorphic change detection between pre- and post-urbanization scenarios. *Environmental Earth Sciences*, 79, 1–14. https://doi.org/10.1007/s12665-020-08958-9

Di Lollo, J. A., Guerrero, J. V. R., Abe, A. C. P. & Lorandi, R. (2018). *Land Change, Soil Degradation Processes, and Landscape Management at the Clarinho River Watershed, Brazil*. In San Francisco (Ed.), IAEG/AEG Annual Meeting Proceedings (Vol. 3, pp. 99–106). Springer International Publishing.

Di Lollo, J. A. & Sena, J. N. (2013). Establishing erosion susceptibility: analytical hierarchical process and traditional approaches. *Bulletin of Engineering Geology and the Environment*, 72(3–4), 589–600. https://doi.org/10.1007/s10064-013-0529-9

Dorici, M., Costa, C. W., de Moraes, M. C. P., Piga, F. G., Lorandi, R., de Lollo, J. A. & Moschini, L. E. (2016). Accelerated erosion in a watershed in the southeastern region of Brazil. *Environmental Earth Sciences*, 75(19), 1301. https://doi.org/10.1007/s12665-016-6102-7

Durães, M. F. & Mello, C. R. de. (2016). Spatial distribution of the potential and current soil erosion for the Sapucaí River Basin, MG, Brazil. *Engenharia Sanitaria e Ambiental*, 21(4), 677–685.

https://doi.org/10.1590/s1413-41522016121182

Esri. (2021). *ArcGIS Pro* (2.8.3).

FAO. (2022). *World Reference Base for Soil Resources 2022*: International soil classification system for naming soils and creating legends for soil maps (4th ed.). FAO. https://doi.org/10.4060/cc3166en

Frankl, A., Nyssen, J., Vanmaercke, M., & Poesen, J. (2020). Gully prevention and control: Techniques, failures and effectiveness. *Earth Surface Processes and Landforms*, 46, 220–238. https://doi.org/10.1002/esp.5033

Galharte, C. A., Villela, J. M. & Crestana, S. (2014). Sediment yield estimation due to changes in land use and cover. *Revista Brasileira de Engenharia Agrícola e Ambiental*, 18(2), 199–201. https://doi.org/10.1590/S1415-43662014000200010

Gudiño-Elizondo, N., Biggs, T. W., Castillo, C., Bingner, R. L., Langendoen, E. J., Taniguchi, K., Kretzschmar, T., Yuan, Y., & Liden, D. (2018). Measuring ephemeral gully erosion rates and topographical thresholds in an urban watershed using unmanned aerial systems and structure from motion photogrammetric techniques. *Land Degradation & Development*, 29, 1896–1905. https://doi.org/10.1002/ldr.2976

Guo, Y., Peng, C., Zhu, Q., Wang, M., Wang, H., Peng, S. & He, H. (2019). Modelling the impacts of climate and land use changes on soil water erosion: Model applications, limitations and future challenges. *Journal of Environmental Management*, 250, 109403. https://doi.org/10.1016/J.JENVMAN.2019.109403

IBGE. (2010). *Brazilian Demographic Census*. Brazilian Institute of Geography and Statistics. https://censo2010.ibge.gov.br/

IBGE (2022). *São Paulo / Brasil: Panorama*. https://cidades.ibge.gov.br/brasil/sp/panorama

IBM. (2011). *IBM SPSS Statistics* (20).

Imeson, A. C., & Kwaad, F. J. P. M. (1980). Gully types and gully prediction. *Geografisch Tijdschrift*, 14(5), 430-441.

Imwangana, F. M., Zola, B. J., Kangudia, J. M., & Sanghy, S. (2025). Gully erosion associated with peri-urbanization: Focus on the catchments around Kimwenza in the south of Kinshasa (DR Congo). *East African Journal of Environment and Natural Resources*, 8(2). https://doi.org/10.37284/eajenr.8.2.3160

IPT. (2012). *Registration of Erosion and Flooding Points in the State of São Paulo* (Vol. 2, Issues 131.057-205). Technological Research Institute Technical Report.

Issaka, S. & Ashraf, M. A. (2017). Impact of soil erosion and degradation on water quality: a review. *Geology, Ecology, and Landscapes*, 1(1), 1–11. https://doi.org/10.1080/24749508.2017.1301053

Javidan, N., Kavian, A., Pourghasemi, H. R., Conoscenti, C. & Jafarian, Z. (2019). Gully Erosion Susceptibility Mapping Using Multivariate Adaptive Regression Splines—Replications and Sample Size Scenarios. *Water*, 11(11), 2319. https://doi.org/10.3390/w11112319

Júnior, O. A. C., Guimarães, R. F., Freitas, L. F., Gomes-Loebmann, D., Gomes, R. A. T., Martins, E. S., & Montgomery, D. R. (2010). Urbanization impacts upon catchment hydrology and gully development using multi-temporal digital elevation data analysis. *Earth Surface Processes and Landforms*, 35, 611–629. https://doi.org/10.1002/esp.1917

Kuhn, C., Reis, F., Furegatti, S., Zarfl, C., & Peixoto, A. (2024). Economic impacts of an urban gully are driven by land degradation. *Natural Hazards*, 120, 13995–14026. https://doi.org/10.1007/s11069-024-06727-6

Mahamba, J., Kayitoghera, G., Musubao, M., Chuma, G., & Sahani, W. (2023). Evolution of gully erosion and susceptibility factors in the urban watershed of the Kimemi (Butembo/DR Congo). *Geography and Sustainability*, 4(3), 268-279. https://doi.org/10.1016/j.geosus.2023.07.001

MapBiomas Project. (2010). *Collection 8.0 of the Annual Series of Land Cover and Land Use Maps of Brazil*. https://brasil.mapbiomas.org/

Mawe, G., Landu, E., Imwangana, F., Hubert, A., Dille, A., Bielders, C., Poesen, J., Dewitte, O., &

Vanmaercke, M. (2024). What controls the expansion of urban gullies in tropical environments? Lessons learned from contrasting cities in DR Congo. *Catena*, 241, 108055. https://doi.org/10.1016/j.catena.2024.108055

Mawe, G., Landu, E., Dujardin, E., Imwangana, F., Bielders, C., Hubert, A., Michellier, C., Nzolang, C., Poesen, J., Dewitte, O., & Vanmaercke, M. (2025). Mapping urban gullies in the Democratic Republic of the Congo. *Nature*, 644, 952–959. https://doi.org/10.1038/s41586-025-09371-7

Medeiros, A. R. C., Araújo, Y. B. de, Vianna, R. P. de T. & Moraes, R. M. de. (2014). Decision support model applied to the recognition of non-adherent individuals to antihypertensive therapy. *Saúde em Debate*, 38(100). https://doi.org/10.5935/0103-104.20140016

Mota, S. (1995). *Preservation and Conservation of Water Resources* (2nd ed.). ABES.

Oliveira, A. M. S. (1994). *Technogenic deposits and reservoir silting: The Capivara Reservoir, Paranapanema River* [Doctoral dissertation]. University of São Paulo, São Paulo.

Oliveira, J. B., Camargo, M. N., Rossi, M. & Calderano Filho, B. (1999). *Pedological map of the State of São Paulo* [Map; scale 1:500,000]. Instituto Agronômico de Campinas (IAC); Embrapa Solos.

Pierce, F. J. & Lal, R. (2017). *Monitoring the Impact of Soil Erosion on Crop Productivity*. In Soil Erosion Research Methods (pp. 235–263). Routledge. https://doi.org/10.1201/9780203739358-10

Pimentel, D. (2006). Soil Erosion: A Food and Environmental Threat. *Environment, Development and Sustainability*, 8(1), 119–137. https://doi.org/10.1007/s10668-005-1262-8

Pons, N. A. D., Pejon, O. J. & Zuquette, L. V. (2007). Use of geoprocessing in the study of land degradation in urban environments: the case of the city of São Carlos, state of São Paulo, Brazil. *Environmental Geology*, 53(4), 727–739. https://doi.org/10.1007/s00254-007-0685-y

Powers, D. M. W. (2011). Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation. *International Journal of Machine Learning Technology*, 2(1), 37–63. https://doi.org/10.48550/arXiv.2010.16061

Prieto-Amparán, Pinedo-Alvarez, Vázquez-Quintero, Valles-Aragón, Rascón-Ramos, Martinez-Salvador & Villarreal-Guerrero. (2019). A Multivariate Geomorphometric Approach to Prioritize Erosion-Prone Watersheds. *Sustainability*, 11(18), 5140. https://doi.org/10.3390/su11185140

Quinlan, J. R. (1983). *Learning Efficient Classification Procedures and Their Application to Chess End Games*. In R. S. Michalski, J. G. Carbonell & T. M. Mitchell (Eds.), Machine Learning (pp. 463–482). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-12405-5_15

Rahmati, O., Kalantari, Z., Ferreira, C. S. S., Chen, W., Soleimanpour, S., Kapović-Solomun, M., Seifollahi-Aghmiuni, S., Ghajarnia, N., & Kazemabady, N. (2022). Contribution of physical and anthropogenic factors to gully erosion initiation. *Catena*, 2010, 105925. https://doi.org/10.1016/j.catena.2021.105925

Ricardi, A. de M. & Lima, C. G. da R. (2022). Spatial and temporal variability of rain erosivity (EI30) in the State of São Paulo, Brazil. *Geosciences*, 40(4), 965-985. https://doi.org/10.5016/geociencias.v40i04.15492

Rodrigo-Comino, J., Keesstra, S. & Cerdà, A. (2018). Soil Erosion as an Environmental Concern in Vineyards. The Case Study of Celler del Roure, Eastern Spain, by Means of Rainfall Simulation Experiments. *Beverages*, 4(2), 31. https://doi.org/10.3390/beverages4020031

Ross, J. L. S., & Moroz, I. C. (1997). *Geomorphological map of the State of São Paulo* [Map; scale 1:500,000]. Laboratório de Geomorfologia, Departamento de Geografia, FFLCH/USP; Laboratório de Cartografia Geotécnica, Geologia Aplicada, IPT; Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP).

Salomão, F. X. T. (1994). *Linear erosion processes in Bauru (SP): Cartographic regionalization applied to preventive urban and rural control* [Doctoral dissertation]. University of São Paulo, São Paulo.

Sparovek, G., Correchel, V. & Barretto, A. G. O. P. (2007). The risk of erosion in Brazilian cultivated pastures. *Scientia Agricola*, 64(1), 77–82. https://doi.org/10.1590/S0103-90162007000100012

Telles, T. S., Guimarães, M. de F. & Dechen, S. C. F. (2011). The costs of soil erosion. *Revista Brasileira de Ciência do Solo*, 35(2), 287–298. https://doi.org/10.1590/S0100-06832011000200001

Tricart, J. (1977). *Ecodinâmica*. Secretaria de Planejamento da Presidência da República, Fundação Instituto Brasileiro de Geografia e Estatística, Diretoria Técnica, Superintendência de Recursos Naturais e Meio Ambient. https://books.google.com.br/books?id=41BgAAAAMAAJ

van Zelm, R., van der Velde, M., Balkovic, J., Čengić, M., Elshout, P. M. F., Koellner, T., Núñez, M., Obersteiner, M., Schmid, E. & Huijbregts, M. A. J. (2018). Spatially explicit life cycle impact assessment for soil erosion from global crop production. *Ecosystem Services*, 30, 220–227. https://doi.org/10.1016/j.ecoser.2017.08.015

Vanmaercke, M., Panagos, P., Vanwalleghem, T., Hayas, A., Foerster, S., Borrelli, P., … Poesen, J. (2021). Measuring, modelling and managing gully erosion at large scales: A state of the art. *Earth-Science Reviews*, 218, 103637. https://doi.org/10.1016/j.earscirev.2021.103637

Zdruli, P., Karydas, C. G., Dedaj, K., Salillari, I., Cela, F., Lushaj, S. & Panagos, P. (2016). High resolution spatiotemporal analysis of erosion risk per land cover category in Korçe region, Albania. *Earth Science Informatics*, 9(4), 481–495. https://doi.org/10.1007/s12145-016-0269-z

Zolezzi, G., Bezzi, M., Spada, D., & Bozzarelli, E. (2018). Urban gully erosion in sub-Saharan Africa: A case study from Uganda. *Land Degradation & Development*, 29, 849–859. https://doi.org/10.1002/ldr.2865

Zuquette, L. V., Rodrigues, V. G. S. & Pejon, O. J. (2013). *Recovery of Degraded Areas*. In Environmental Engineering: concepts, technology and management. Elsevier. https://repositorio.usp.br/item/002472425

Funding

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Process 88887.658530/2021-00. The research was also supported by National Council for Scientific and Technological Development (CNPq) - Process #311393/2021-7:

Author Contributions

T. F. Olivatto: Conceptualization, Methodology, Data curation, Formal analysis, Investigation, Visualization Writing- Original draft. J. A. Di Lollo and D. B. Menezes: Supervision, Validation, Writing-Reviewing and Editing.

Conflict of Interest

No conflict of interest to declare.

First Author's Biography

PhD and master degree in Urban Engineering at UFSCar, with a mobility period at the Universidad Nacional del Litoral and a sandwich program at the Universidad Nacional de Mar del Plata, both in Argentina. Graduated in Environmental and Sanitary Engineering from PUC Campinas, including a sandwich program at the University of South Australia. Specialized in Data Science through CeMEAI and ICMC/USP, with technical training in Geomatics from Colégio Técnico de Limeira (UNICAMP), and a specialization in University Teaching. Dedicated to research that integrates geospatial information and remote sensing with artificial intelligence, particularly machine learning.

APÊNDICE B – ARTIGO 2

Journal Pre-proof

Sampling strategies for machine learning-based linear erosion studies: a review approaching contributing area

Tatiane Ferreira Olivatto, José Augusto Di Lollo

PII: S2666-5921(25)00070-8

DOI: https://doi.org/10.1016/j.nhres.2025.08.002

Reference: NHRES 277

To appear in: *Natural Hazards Research*

Received Date: 15 January 2025

Revised Date: 10 August 2025

Accepted Date: 10 August 2025

Please cite this article as: Olivatto, T.F., Di Lollo, J.A., Sampling strategies for machine learning-based linear erosion studies: a review approaching contributing area, *Natural Hazards Research*, https://doi.org/10.1016/j.nhres.2025.08.002.

Sampling strategies for machine learning-based linear erosion studies: a review approaching contributing area

Abstract: Linear erosion is a major socio-environmental challenge, influenced by climatic, geomorphological and anthropogenic factors. This study explores how sampling strategies and topographic contributing areas impact machine learning-based applications in linear erosion research. A bibliometric analysis was conducted to assess historical and emerging trends at the intersection of linear erosion and machine learning. Additionally, a systematic extraction of key methodological insights was performed using artificial intelligence tools, followed by an integrative literature review focusing on sampling techniques and drainage influence areas. Results show that spatially stratified sampling, particularly with a 1:1.2 occurrence-to-non-occurrence ratio and non-occurrence points placed outside the hydrological contributing area, improves model generalization and environmental representativeness. However, its effectiveness depends on an adequate understanding of local landscape dynamics. Advanced oversampling techniques further mitigate class imbalance, while spatial cross-validation addresses spatial autocorrelation and enhances robustness. The use of topographic contributing area as a predictive variable shows significant potential due to its role in controlling runoff concentration and sediment flow. Nevertheless, the lack of standardization in how contributing areas are delineated and integrated into models limits broader applicability. Current machine learning approaches often underexplore these spatial components, reducing their physical consistency. This work consolidates methodological advances in data preparation and emphasizes physically informed models that integrate land use and hydrological processes, enabling more robust and interpretable machine learning applications in erosion studies. Future research should focus on refining sampling frameworks and integrating contributing area metrics to enhance the robustness and interpretability of machine learning predictions in erosion studies.

Keywords: drainage area, influence area, absence data, spatial sampling, gully, ravine.

1. Introduction

Accelerated linear erosion is a cause for socio-environmental concern, especially in susceptible areas due to factors such as climate and anthropogenic interventions (Churu et al., 2024; Poesen et al., 2003). Linear erosion, often manifested as gully erosion, it is a process of detachment and dragging of soil particles caused by water action (Conciani, 2008). Challenges associated with gully erosion include soil degradation, loss of arable land and increased sedimentation in water bodies, contributing to broader ecological imbalances (Junior et al., 2010; Phionah and Extension, 2024; Telles et al., 2011).

Linear erosion extends its influence beyond the natural environment, significantly impacting human societies and often exacerbates existing social inequities. As the landscape transforms, communities may find their homes and agricultural lands compromised, forcing them to relocate. Vulnerable populations may lack the resources and resilience to cope with the impacts of erosion, intensifying social disparities and highlighting the need for inclusive and equitable solutions (Saha et al., 2021). In urban areas, where currently most population lives, linear erosion jeopardizes critical infrastructure and exacerbates urban drainage issues, increasing sedimentation in stormwater systems can that can lead to blockages, flooding and compromised drainage efficiency, posing threats to public safety and the functionality of urban infrastructures (Gudino-Elizondo et al., 2022; Junior et al., 2010; Polovina et al., 2021).

Linear erosion is particularly relevant in urban areas. Urban expansion, combined with vegetation removal and increased impervious surfaces, intensifies surface runoff, leading to the formation of gullies and ravines. These erosion processes compromise critical infrastructures such as roads and drainage systems, exacerbating urban challenges and posing risks to public safety (Polovina et al., 2021). Then, the choice to focus on linear erosion is motivated by the fact that the majority of the population lives in urbanized areas (Moreno-Monroy et al., 2020), requiring a deeper understanding of linear erosion forms and their economic, social and environmental impacts (Kuhn et al., 2023).

Within the context of linear erosion, it is essential to distinguish between its different stages. However, this can be challenging, as the classification of linear erosion features varies globally (Vanmaercke et al., 2016). Dube et al. (2020) and Kuhn et al. (Kuhn et al., 2023) conducted comprehensive worldwide studies on linear erosion morphologies and characterized three primary stages of them: rills, gullies, and ravines.

According to them, rills are the initial stage of concentrated flow erosion, forming shallow channels with incision depths of less than 0.25 meters. They develop on slopes due to surface water flow, detaching and transporting soil particles, and can be easily obliterated by tillage or natural sediment deposition. Gullies represent a more advanced stage of erosion, with depths greater than 0.25 meters. They are often U-shaped in Brazilian classifications and arise when concentrated runoff intensifies, leading to significant soil loss and the expansion of multiple branching channels.

Ravines, in contrast, are deeper and larger than gullies, frequently exhibiting a V-shaped cross-section and reaching depths greater than 0.50 meters. They often connect to aquifers, promoting lateral erosion and exacerbating landscape degradation. Due to their size and permanence, ravines represent a mature stage of linear erosion, significantly altering landforms and ecosystems. In specific global contexts, particularly in arid and semi-arid regions, ravines play a crucial role in the landscape dynamics. In these areas, where vegetation cover is sparse and rainfall is erratic but intense, ravines form as a result of the severe degradation of soil caused by concentrated runoff (Arabameri et al., 2020).

These processes influenced by several key factors, including precipitation, topography, soil properties and land use (Guerra, 2012). Precipitation intensity and frequency play a critical role in shaping erosion patterns, as heavy rainfall events can accelerate surface runoff, leading to the formation of rills and gullies. Topography, including slope steepness and length, directly affects the velocity and concentration of water flow, which in turn determines the erosive power of runoff. Soil properties, such as texture, structure, and organic matter content, influence soil's susceptibility to erosion, with loose or poorly structured soils being more prone to detachment and transport. Land use practices, such as agriculture, deforestation, and urbanization, can further exacerbate erosion by reducing vegetation cover and altering natural drainage patterns. Together, these factors interact in complex ways, creating dynamic and varied erosional landscapes (Chalise et al., 2019; Falcão et al., 2020; Guerra, 2012).

Then, understanding and monitoring these type of erosion is imperative for sustainable land management and environmental preservation. Studying linear erosion is crucial for unravelling its complex dynamics and formulating effective strategies for prevention and control. A comprehensive understanding is essential for informed decision-making and sustainable resource management (Frankl et al., 2021; Telles et al., 2011).

In this sense, the integration of artificial intelligence (AI) have revolutionized the field of gully erosion studies. Recent advancements leverage this concept, mainly through machine learning (ML) applications, to analyse vast datasets, model erosion processes and predict potential gully formation. ML applications offer a nuanced understanding of the intricate relationships between environmental variables, enhancing the accuracy of erosion risk assessments (Amiri et al., 2019; Pourghasemi et al., 2017).

ML techniques offer significant advantages over traditional erosion models, particularly in handling large-scale data and complex nonlinear relationships. Traditional models often rely on predefined equations and assumptions, which can be limiting when dealing with the intricate and variable nature of erosion processes. In contrast, ML algorithms can analyze high-dimensional data and capture subtle patterns without explicit programming. For instance, a study applied neural networks models to identify soil erosion intensity in ecological restoration areas, demonstrating that these models could effectively map nonlinear relationships between erosion factors, leading to more accurate assessments compared to traditional methods (C. Liu et al., 2023).

Furthermore, in coastal engineering for example, ML models have been reported to analyze nonlinear relationships and high-dimensional variables much more rapidly than physics-based models, exhibiting computational efficiency ranging from 22.5 times faster to 4,000 times faster. This acceleration is pivotal for practitioners who require timely insights for effective erosion management (Abouhalima et al., 2024).

By leveraging the capabilities of ML, researchers and practitioners can achieve more precise and efficient erosion assessments, overcoming the limitations of traditional models and enhancing the effectiveness of erosion control strategies. When comparing ML models to the Revised Universal Soil Loss Equation (RUSLE), machine learning models generally provide more accurate and comprehensive results (Lin and Zhao, 2022). Random Forest (RF) and Support Vector Machines (SVM), for example, have overcame traditional bivariate statistical models, achieving predictive accuracy values that exceed 0.9, which indicates their ability to reliably predict gully erosion (Garosi et al., 2019; Pourghasemi et al., 2020; Rahmati et al., 2017).

These machine learning models are more flexible than traditional models such as empirical or statistical regression-based approaches, or physically-based models like the Universal Soil Loss Equation (USLE), which, while useful, lack the predictive power and ability to capture complex interactions between multiple environmental factors. This flexibility and higher accuracy make machine learning models a valuable tool for erosion prediction, allowing for more precise and adaptable erosion management strategies in diverse environments (Bag et al., 2022; Gholami et al., 2024; Pourghasemi et al., 2020).

In addition to that, ML has gained growing prominence in erosion research due to its ability to integrate complex and nonlinear interactions among multiple environmental, geological, and climatic factors. By leveraging large and heterogeneous datasets, ML models enable high-resolution spatial mapping of erosion-prone areas, significantly improving prediction accuracy and identifying zones of heightened risk (Barakat et al., 2023; Pourghasemi et al., 2020). These advancements support more targeted and cost-effective interventions by allowing land managers to anticipate erosion processes and allocate resources more efficiently. Moreover, ML algorithms facilitate factor analysis, helping researchers identify and rank the most influential drivers of erosion, such as slope gradient, rainfall intensity, vegetation cover, and proximity to river networks, which is essential for designing effective mitigation strategies (Bag et al., 2022; Gholami et al., 2024).

Beyond prediction and mapping, the incorporation of explainable machine learning techniques, such as SHAP (Shapley Additive Explanations) and permutation importance, has further expanded the practical utility of these models. These approaches enhance the transparency of ML outputs, enabling stakeholders and decision-makers to understand how different variables contribute to erosion risk (Gholami et al., 2024). As a result, ML not only strengthens scientific understanding of erosion dynamics but also improves communication and trust between researchers, policymakers, and practitioners. Overall, the integration of machine learning into erosion research marks a critical shift toward more precise, data-driven decision-making in environmental and land management contexts.

Recent methodological advancements further underscore the growing importance of machine learning in environmental sciences. For instance, Jemeļjanova et al. (2024) reviewed how supervised machine learning models address spatial autocorrelation—an essential concern in modeling environmental variables—highlighting the lack of standardized approaches and the contextual nature of spatial strategies in the modeling pipeline. Their work underscores the importance of selecting appropriate spatial methods based on study characteristics and shows the potential of explicit spatial covariates and informed data splitting to improve model performance.

Complementing this, Ekeh et al. (2025) examined the transformative potential of machine learning in guiding environmental policy and sustainability. Their interdisciplinary framework integrates geospatial analysis, predictive modeling, and real-time monitoring through IoT to tackle urban and ecological challenges. The study emphasizes the value of machine learning in generating actionable insights for climate resilience, pollution control, and equitable environmental decision-making.

Despite the promising advancements, implementing machine learning in the context of environmental issues presents inherent challenges. Notable limitations are related to insufficient or low-quality data, biased or incomplete datasets, computational resources requirements, limited techniques for effectively incorporating domain-specific knowledge held by experts and scalability challenges in modelling for larger geographic areas or address regional environmental issues (Xian Liu et al., 2022).

Some specific environmental-related limitations include temporal and spatial aspects. Environmental conditions change over time and machine learning models may face difficulties in capturing temporal dynamics (Amato et al., 2020). This limitation is particularly relevant when addressing phenomena with seasonality or long-term trends. In addition, environmental processes often exhibit spatial heterogeneity and machine learning models may struggle to capture these variations accurately (Schmidt et al., 2020). The lack of spatial representativity can lead to limited applicability of models across diverse geographical regions. In addition, Not considering these aspects in machine learning models may not adequately capture uncertainties associated with environmental processes, leading to overly confident predictions (Xian Liu et al., 2022).

Linear erosion processes are triggered by several factors, mainly, relief features, land use/cover, geological material, and rainfall erosivity (Cunha and Guerra, 1996). Then, defining the extent area of the spatial influence of these factors remains a complex task, impacting the accuracy of predictive models. Besides that, scarcity of non-occurrence data poses challenges in training models, as linear erosion occurrences are often imbalanced (Xian Liu et al., 2022).

In fact, the study of linear erosion, particularly gully erosion, is paramount for environmental sustainability. The integration of artificial intelligence and machine learning has enhanced our ability to address these challenges. However, acknowledging and addressing the limitations in implementing machine learning for environmental issues is crucial for advancing accurate and reliable predictive models, ultimately contributing to effective gully erosion management and environmental conservation.

Based on the aforementioned premises, this text establishes the hypothesis that sampling (or instances) and influence area are sensitive topics among linear erosion studies developed by the use of machine learning methodology. Thus, this study aimed to investigate the critical role of sampling strategies and how the topographic contributing area can support predictive modelling among machine learning studies. To achieve this, a bibliometric analysis was conducted to identify historical and emerging trends at the intersection of linear erosion and machine learning. Additionally, a systematic AI-assisted information extraction was performed to synthesize key methodological approaches. A subsequent in-depth literature review focused on sampling strategies and the role of drainage areas in erosion studies. By critically analyzing these aspects, this study aims to highlight gaps in current methodologies and propose improvements for more effective and spatially representative machine learning applications in linear erosion research.

2. Materials and methods

The methodological procedure started with a bibliometric analysis, serving as a foundational exploration of the existing literature framework. Subsequently, a systematic extraction of key information was conducted using artificial intelligence tools, allowing for the comprehensive retrieval and synthesis of relevant data points. Building upon these findings, a thorough literature review was undertaken to delve deeper into topics pertaining to sampling for machine learning applications and the significance of defining the area of influence concerning natural phenomena.

The bibliometric analysis assessed studies on linear erosion developed with or related to machine learning techniques. Bibliometrics is a quantitative method used to strategically analyse patterns and relationships within written scientific publications. It involves the statistical and mathematical analysis of bibliographic data, which includes information about books, articles, or other forms of literature, allowing concepts exploration and providing insights of the main characteristics of a specific knowledge field (Sharifi et al., 2023; van Raan, 2005).

Data was retrieved from Scopus online database, as it is one of the main indexing platforms for international scientific publications besides being besides being widely recognized for its comprehensive coverage and reliable citation metrics (Romanelli et al., 2018). Additionally, Scopus allows for efficient data export and integration with bibliometric analysis tools, which is essential for processing and analyzing citation networks in our study (Falagas et al., 2008).

For these reasons, Scopus was chosen to ensure the reproducibility of our search process and seamless integration with bibliometric tools (Gusenbauer and Haddaway, 2020). The research examined academic publications until December 31, 2023 and a minimum temporal frame was not established, however, the research performed in January 2024, for the purposed search, found documents from 2017.

According to (T. H. F. E. Silva et al., 2021), the development of bibliometric studies and systematic reviews tightly rely on search terms selection and composition. The search was performed considering the titles, keywords, and abstracts, according to the terms and delimiters indicated in Table 1, however, only the Joint Search was adopted to develop the bibliometrics.

Table 1. Terms and strings used in the searches for bibliometrics analysis.

Searched Terms	String Composition	N. Publications	Period
Linear Erosion Gully Ravine	TITLE-ABS-KEY ("linear erosion" OR "linear erosions" OR gully OR gullies OR ravine*)	11,531	1869 - 2023
Machine Learning	TITLE-ABS-KEY ("machine learning")	567,176	1959 - 2023
Joint Search	TITLE-ABS-KEY (("linear erosion" OR "linear erosions" OR gully OR gullies OR ravine*) AND ("machine learning"))	127	2017 - 2023

Source: Prepared by the authors.

Descriptive statistics were applied to summarize basic information regarding retrieve bibliographic data, such as documents typology, their quantitative temporal evolution and the most relevant sources. VOSviewer, a bibliometric software, was employed for network analysis and visualization. This software is widely disseminated for the purpose of bibliometrics and essentially projects the bibliographic information of academic documents onto bibliometric maps (van Eck and Waltman, 2010).

Collaboration patterns among countries and authors were visualized as networks. Authors keyword maps were generated to illustrate relationships between key terms, including the temporal evolution of research trends. In this visualization scheme, data are represented as circles, connections (lines) and colors (clusters or average date). The size of the circles represents the number of documents, the intensity of connection between items is expressed by the thickness of the lines, and the color symbology can represent groups (clusters) with more significant interaction or the average year of occurrence (van Eck and Waltman, 2014). In this study, we chose showing only items with at least three occurrences.

Additionally, an artificial intelligence tool was employed to extract valuable information from indexed bibliographic data, providing a deeper understanding of the content. Abstracts of retrieved publications were processed using Perplexity AI Chat, that is an artificial intelligence-powered language model that uses natural language processing and machine learning to generate human-like text, answer questions, provide information, and assist with various tasks . Perplexity AI Chat is similar to the most known ChatGPT, however, allows users to conduct a human-like conversation based on inputs, such as texts and images (Codina, 2023; A. Pan et al., 2023).

Considering the literature investigated on the introduction section and expert knowledge, Perplexity AI Chat was employed to identify and extract the following key insights:

(i) What multivariate analysis methods are mentioned?

(ii) What machine learning methods are mentioned?

(iii) What statistical methods are mentioned?

(iv) What programming languages are mentioned?

(v) What GIS-based tools or materials are mentioned?

The central topics for extraction were identified based on their inclusion in the set of studies outlined by Vosgerau and Romanowski (2014) and Silva et al. (2021). They approach aims to facilitate the organization and synthesis of the identified publications, encompassing both trends and methodological procedures employed. Furthermore, following their guidelines and adapting their elucidative expressions for the objectives of this paper, the selection of these extraction topics was guided by specific directives: assessing production contributions, understanding the characteristics and methodological qualities of the studies, and gaining insight into the tools employed.

For the literature review stage, in order to refine the selection of articles aligning with the purposed objectives, the search query in Scopus was augmented with the terminology "sampling", which yielded a total of 4 additional documents. Subsequently, a thorough examination of these documents led to the inclusion of 2 more articles, guided by insights gleaned from the discussions and findings presented therein. This systematic and iterative process of literature curation ensured the comprehensive coverage of relevant studies (Paré et al., 2015) addressing the nuances of sampling methodologies for machine learning applications in the context of natural phenomena, specifically, linear erosion.

Given the absence of documents retrieved upon augmenting the search with terms related to drainage or topographic influence area and non-occurrence or absence sampling, the review process concerning these topics transitioned into an integrative literature review approach. Integrative reviews provide a comprehensive synthesis of existing literature by drawing from diverse sources and methodologies, allowing for a nuanced exploration of complex and interdisciplinary topics (Cronin and George, 2023). By adopting an integrative approach, the review aimed to bridge gaps in the literature and offer a holistic understanding of the topics found during the previous methodology steps. Also, through this integrative lens, the review sought to uncover insights from various disciplinary perspectives, thereby enriching the discourse surrounding sampling methodologies and spatial delineation techniques crucial for accurate modelling and prediction of environmental processes.

3. Results and discussions

3.1. Bibliometric analysis

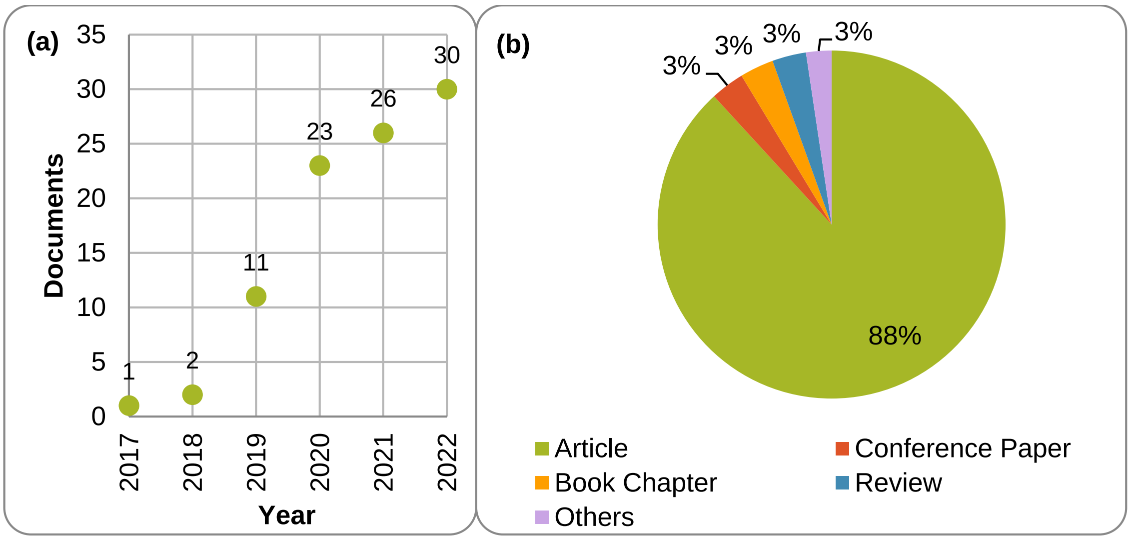
As of the date of the search, there was Scopus has 11,531indexed documents containing the Linear Erosion related terminology, the oldest one is “On the Formation of Ravines by Recent Drift Accumulations” (Kinahan, 1869). The most cited document, named “Gully erosion and environmental change: Importance and research needs” (Poesen et al., 2003), exposed a huge gap as researches dealing with soil water erosion have been concentrated exclusively on sheet (interrill) and rill erosion processes operating at the (runoff) plot scale. Authors conducted a comprehensive review on this, identifying key research questions, and concluded that there is a need for monitoring, experimental and modelling studies of gully erosion as a basis for predicting the effects of environmental change (climatic and land use changes) on gully erosion rates.

Regarding Machine Learning related documents, two documents from 1959 were found, “Some experiments in machine learning” (Campaigne, 1959) and “ Two notes on machine Learning” (Martens, 1959). In this context, the most cited paper is “Random forests” (Breiman, 2001) which introduces the methodology of combining tree predictors, where each tree's reliance on values is based on an independently sampled random vector with a consistent distribution across all trees in the forest.

Finally, the joint investigation returned 127 documents and, differing from the previous searches, the combined one showed itself as a recent and ongoing topic, as the first published paper dates from 2017, being the second most cited, and whereas there is a growth trend over the years (see Figure 1a). In Figure 1b one can see the predominance of article-type documents, 88%, indicating high scientific quality content, once they are usually peer-reviewed.

Corroborating this, most relevant sources also comprise high quality journals, as showed in Table 2. The provided table reveals insightful patterns in the distribution and impact of scholarly publications across various journals. Firstly, the extensive reach of the research is evident as documents are disseminated across a diverse range of 59 journals. This highlights the broad scope and multifaceted nature of the subject matter. Additionally, the concentration of contributions is underscored by the fact that 16 journals have published three or more documents, emphasizing key players in the dissemination of knowledge within this field.

Figure 1. (a) documents by year and **(b)** documents type.



Source: Prepared by the authors.

Table 2. Terms and strings used in the searches for bibliometrics analysis.

Source Title	Documents number	%	CiteScore* (2022)	Number of Citations	Citation Ratio**
Remote Sensing	13	15.9	14.43	395	30.4
Geomorphology	6	7.3	7.8	260	43.3
Science Of The Total Environment	6	7.3	7.3	565	94.2
Geomatics Natural Hazards And Risk	5	6.1	12.9	108	21.6
Geoderma	4	4.9	16.8	350	87.5
Water Switzerland	4	4.9	15.3	32	8.0
Sensors Switzerland	4	4.9	7.2	234	58.5
Land Degradation And Development	4	4.9	6.8	36	9.0
Geoscience Frontiers	4	4.9	5.5	185	46.3
Journal Of Environmental Management	3	3.7	13.4	146	48.7
International Soil And Water Conservation Research	3	3.7	11.9	10	3.3
Catena	3	3.7	9.8	37	12.3
ISPRS International Journal Of Geo Information	3	3.7	6.2	77	25.7
Ecological Informatics	3	3.7	6.1	53	17.7
Geocarto International	3	3.7	4.3	5	1.7
Advances In Science Technology And Innovation	3	3.7	0.6	42	14.0

Note: ***CiteScore** is a bibliometric indicator proposed by Scopus to assess the quality of academic journals. It is widely used in scientometric analyses as a measure of journal impact and relevance (Elsevier, 2021). ****Citation Ratio** represents the number of citations received by a journal divided by the number of documents published in that journal, providing an indication of the average impact per publication. Source: Prepared by the authors.

Notably, the journal "Remote Sensing" emerges as a significant contributor, responsible for 15.9% of the total documents, distinguishing itself with the third-highest CiteScore (14.43). The spectrum of CiteScores, ranging from 0.6 to 16.8, showcases the variability in the scholarly impact of the journals, with an overall average CiteScore of 9.1. The relatively high CiteScore values of the top journals in this field suggest that research on this topic is being published in well-regarded venues, indicating both scientific interest and academic credibility.

In general, this comprehensive analysis of journal distribution and CiteScores provides valuable insights into the landscape of academic contributions in the given subject area. Furthermore, citations number and citation ratio, which is the relation between citations and documents number of a source, evidence journals' visibility and content relevance.

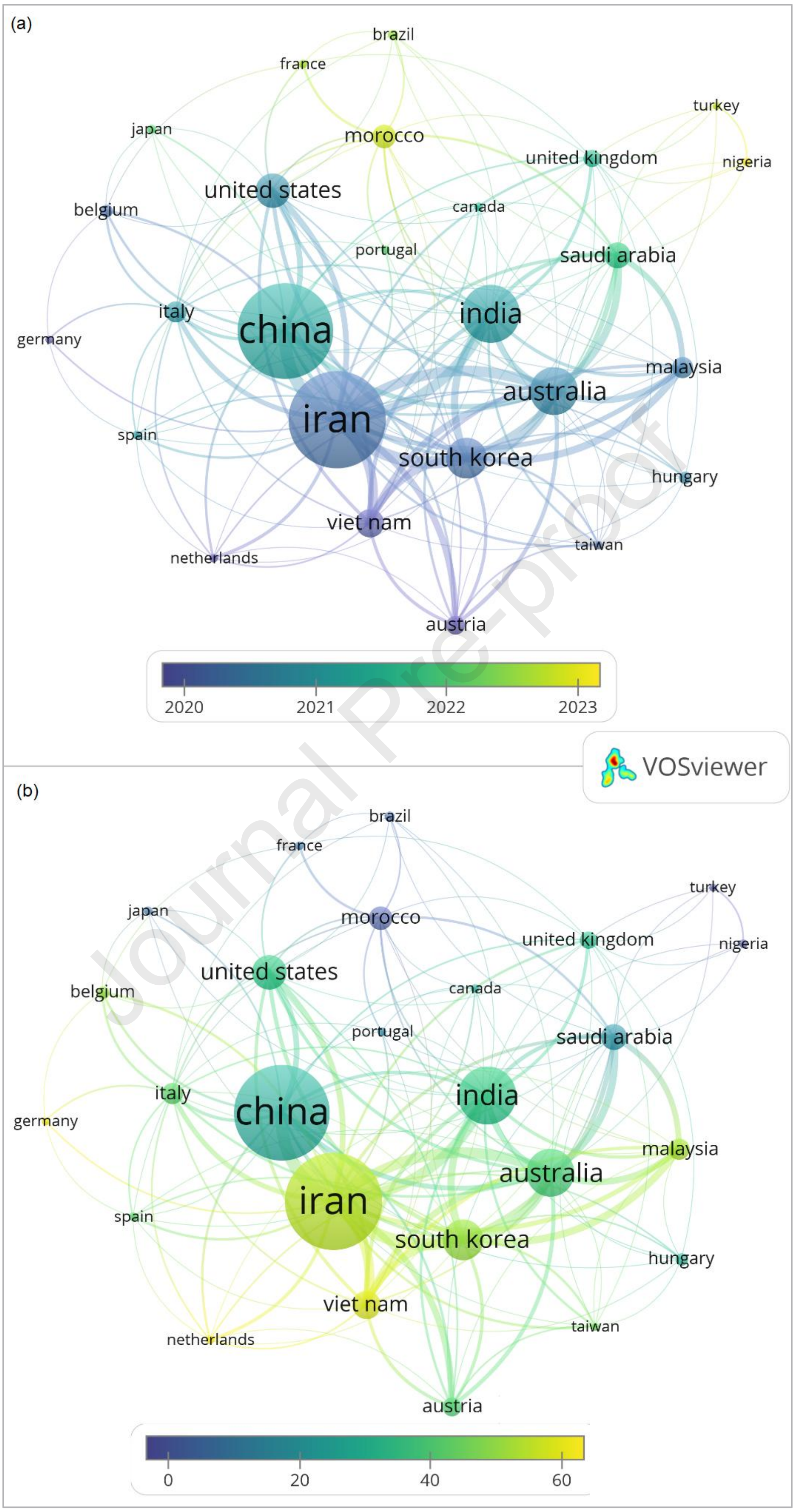
The examination of the data reveals intriguing dynamics within the realm of academic influence among the journals. Despite "Remote Sensing" having the highest number of publications at 13, it falls short in terms of citations, suggesting that quantity does not necessarily correlate with influence in this field of research. On the contrary, "Science Of The Total Environment," with a smaller number of publications at 6, emerges as the most cited journal, signaling its exceptional relevance or heightened visibility within the scholarly community. This is further emphasized by a remarkable citations-to-document ratio of 94.2, indicating a substantial impact per publication. This suggests that while some journals serve as primary outlets for publishing research, others are more influential in shaping the academic discourse, likely due to the perceived credibility and interdisciplinary nature of their readership.

Notably, the "Geoderma" journal stands out with an impressive citation ratio of 87.5, underscoring its significant influence despite a modest number of publications. It's worth noting that while "Remote Sensing" boasts a higher CiteScore than "Science Of The Total Environment," the pinnacle is reached by "Geoderma," suggesting that assessing influence solely based on publication quantity or CiteScore may oversimplify the nuanced landscape of scholarly impact in this particular field. These variations reinforce the importance of considering multiple bibliometric indicators when assessing journal influence, as citation-based metrics may reflect different aspects of academic impact compared to publication frequency or CiteScore alone. These findings prompt a more intricate exploration into the factors contributing to the varying degrees of influence exhibited by these journals.

In Figure 2, we present two bibliometric maps depicting the publication scenario across countries, in which the size of each circle corresponds to the quantity of publications per country. Iran emerges as the frontrunner with 45 publications, followed closely by China with 44 publications. Subsequent positions are occupied by India, Australia, and South Korea, each contributing significantly to the academic discourse in the field. The varying sizes of the circles offer a straightforward yet impactful visualization of the publication contributions by different countries, elucidating the relative prominence and engagement of each nation within the studied research domain.

The network of international contributions is characterized by its heterogeneity, with numerous countries engaging in collaborative publications, as depicted by the varying thickness of connecting lines between circles. Remarkably, countries with a higher volume of publications appear to be more active in international cooperation. The thickness of the connecting lines signifies the extent of collaborative efforts, showcasing the intricate web of scientific exchange.

Figure 2. Documents per country, **(a)** average publication year and **(b)** average citations.



Note: (a) The color scale represents the average publication year, with darker blue tones indicating older publications and yellow tones representing more recent ones. (b) The color scale represents the average number of citations per country, with darker blue indicating fewer citations and yellow highlighting countries with higher citation counts. In both visualizations, the node size corresponds to the number of documents published by each country, while the thickness of the connecting lines represents the strength of collaboration between countries. Source: Prepared by the authors.

Despite the absence of discernible patterns in specific collaboration types, the data underscores the multifaceted nature of global research partnerships. This intricate interplay of collaborative networks accentuates the interconnectedness of the scientific community, emphasizing the importance of international collaboration in advancing knowledge within the studied field.

Figure 2a illustrates a bibliometric network regarding the average publication year of documents by country. Despite the prominence of China, India, and Iran as prolific publishers, this figure unveils a distinct pattern where Morocco, Nigeria, Turkey, France, and Brazil stand out with the most recent average publication years. This highlights not only their active scholarly engagement but also the contemporary nature of their contributions to the field.

In contrast, Figure 2b portrays a bibliometric network showcasing the average citations per document by country. Here, Germany, the Netherlands, Vietnam, Iran, Malaysia, South Korea, and Belgium emerge as notable leaders. These countries exhibit a high average citation per document, emphasizing the impact and recognition of their scholarly output. This dual analysis in Figure 2 provides a nuanced understanding of the global scientific landscape, capturing both the temporal characteristics of publications and the citation impact of different countries in the field.

By visually mapping the connections between keywords based on their occurrence and relationship, the bibliometric network presented in Figure 3 provides a nuanced perspective on the most relevant topics studied by authors. Examining the keywords with the highest frequency of occurrences unveils pivotal themes in the research domain. "Machine learning" and “gully erosion” stand out with the highest frequency and link strength between them, however this was expected as they were searched terms. This prominence underscores the extensive exploration and utilization of machine learning techniques in the context of gully studies.

Additionally, "Random forest", “GIS” (geographic information system), "GES" (gully erosion susceptibility), “soil erosion” and SVM (support vector machine) demonstrate both notable frequency among publications and strong relation with the aforementioned items, emphasizing their importance in the identified publications. The bibliometric representation not only captures the quantitative aspects of keyword occurrences but also offers a visual narrative of the interconnections, providing valuable insights into the focal points and relationships shaping the scientific discussions in the field.

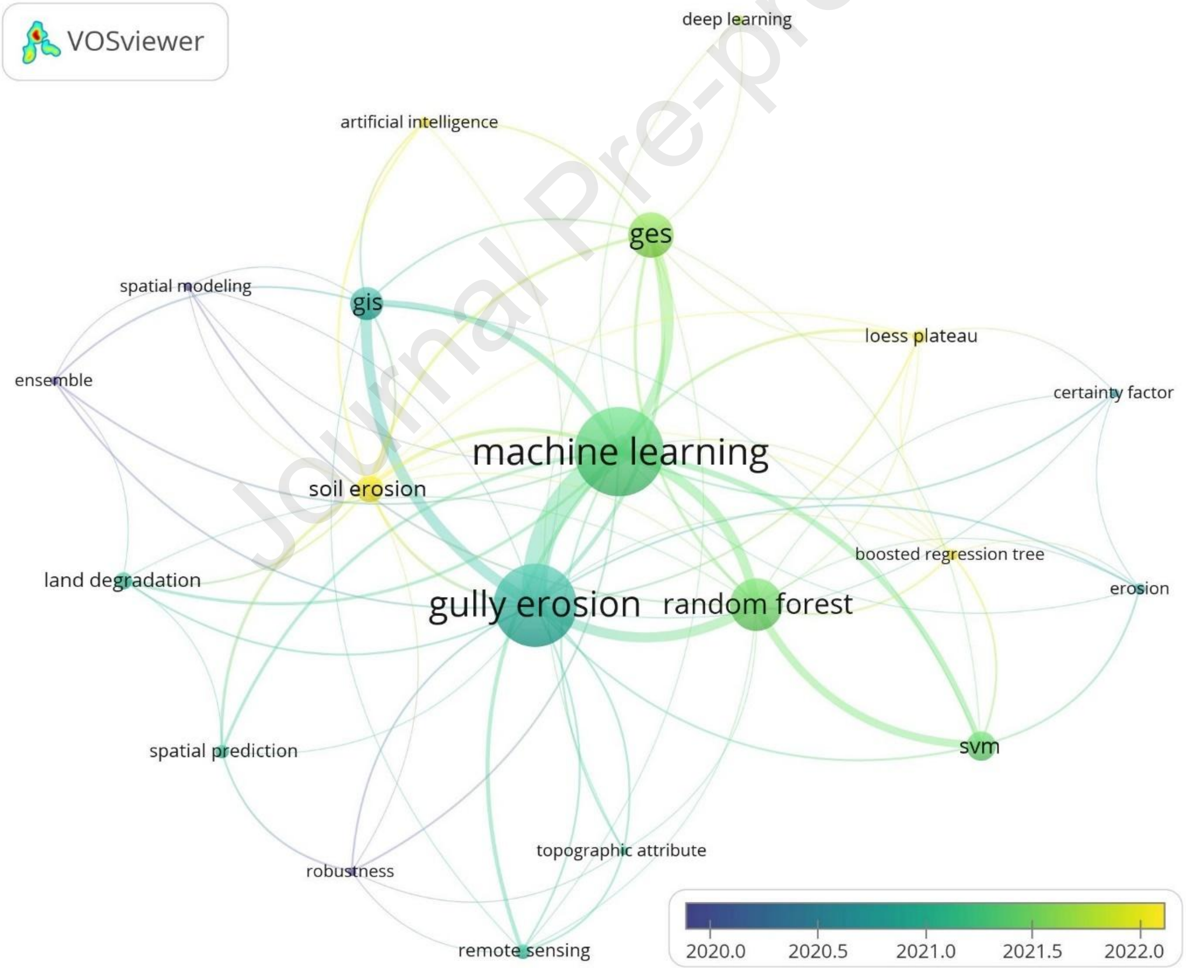
The inclusion of the average year of utilization for each keyword adds a temporal dimension to our analysis, providing insights into the evolving trends within scientific literature. Keywords such as "soil erosion," "boosted regression tree," "loess plateau," and "artificial intelligence" have an average year of occurrence in 2022, signifying their recent prominence in scholarly discussions. The frequent appearance of "loess plateau" and "iran" can be attributed to their association with China and Iran, respectively. The prevalence of these keywords may stem from the substantial volume of publications originating from these regions, contributing to a higher frequency of terms related to specific geographical features.

The keyword "artificial intelligence" maintains a relatively high frequency in most recent years, which can be rationalized by its direct correlation with the most frequent term, "machine learning." Similarly, the recent presence of "soil erosion" aligns with the overarching theme of "gully erosion," emphasizing the interconnectedness of these terms within the scholarly discourse. In the bibliometric network, "soil erosion", most prominent term among the most recent ones, exhibits direct connections with the other recent terms " boosted regression tree", “loess plateau" and “artificial intelligence”. Additionally, its associations with terms from 2021, such as "svm", "random forest", "machine learning", "gully erosion" and "erosion" highlight its continued relevance and interplay with key concepts over time. This temporal analysis enhances our understanding of the dynamic nature of keyword relationships and their evolving roles in scientific research.

The keywords "topographic attribute," "spatial prediction," and "spatial modeling" exhibit limited impact within the bibliometric network, suggesting a relatively lower emphasis on spatial aspects and influence areas in the current discourse. Despite their relevance to environmental phenomena, these terms do not emerge prominently in the network of main keywords. This observation implies that, within the studied scientific domain, there may be a gap in the exploration of spatial dimensions and modeling approaches, despite their potential significance in understanding and predicting erosive processes.

Furthermore, the absence of terms related to the sampling of occurrence and non-occurrence locations in erosion studies is noteworthy. This gap is particularly crucial in the context of machine learning applications, where the inclusion of such terms is essential for robust model development and accurate predictions. This insight into the underrepresentation of certain spatial and methodological aspects underscores potential avenues for the strategic literature review, in which further exploration and investigation may contribute to a more comprehensive understanding of erosion-related phenomena.

Figure 3. Most relevant keywords (average publication year).



Note: The network visualization illustrates the co-occurrence of keywords in the analyzed publications. The node size represents the frequency of occurrence of each keyword, while the thickness of the connecting lines indicates the strength of association between terms. The color scale reflects the average publication year of documents associated with each keyword, with darker blue representing older studies and yellow indicating more recent research. Source: Prepared by the authors.

3.2. Machine learning among gully studies

Following the proposed methodology, this phase scrutinized the content of articles, encompassing titles, abstracts and keywords, through predefined inquiries. These inquiries were systematically addressed with the aid of an artificial intelligence tool, allowing for a comprehensive analysis of the gathered data. This stage is crucial as it will facilitate a more thorough and targeted analysis of the documents retrieved in the initial search, thereby enhancing the understanding of the research landscape within the scope of our investigation.

Tables 3 to 5 present the responses to the three proposed extraction questions: "What multivariate analysis methods mentioned?", "What machine learning methods are mentioned?" and "What statistical methods are mentioned?". The second column of the table provides insights into the methods and tools identified based on the textual content investigated, showing the context of the studies. These tables offer a comprehensive overview of the analytical techniques and methodologies discussed within the literature, aiding in the categorization and synthesis of pertinent information for analysis and discussion conducted in the second stage of this paper.

Table 3. Multivariate analysis methods mentioned

Methods	Study context
Analytical Hierarchy Process (AHP)	Used to weigh the factors and produce a preliminary map per geoenvironmental hazard through Weighted Linear Combination in the Multi-Geoenvironmental Hazards Susceptibility tool.
Weight of Evidence (WoE)	Used for spatiotemporal monitoring of gully erosion sensitivity.
Multicriteria Decision Making (MCDM)	Used for spatiotemporal monitoring of gully erosion sensitivity.
Order of Preference by Similarity to Ideal Solutions (TOPSIS)	Used for mapping gully erosion susceptibility using only topographic related attributes.
Principal Component Analysis (PCA)	Used to identify the processes controlling the variability in gully development.
Genetic Algorithm-Extreme Gradient Boosting (GE-XGBoost)	A hybrid computer education solution used for spatial mapping of the susceptibility of gully erosion.
Machine Learning Classification (MLC)	Used in combination with the statistical weight of evidence model for gully erosion susceptibility mapping.

Source: Prepared by the authors.

Table 4. Machine learning methods mentioned

Methods	Study context

Random Forest (RF)	Used for gully erosion susceptibility prediction in the study of groundwater potential.
Boosted Regression Trees (BRT)	Used for gully erosion susceptibility prediction in the study of groundwater potential.
Reduced Error Pruning Tree (REPTree)	Used for land degradation susceptibility mapping in the study of gravely undulating red and lateritic agro-climatic zones.
Bagging-REPTree	Used for land degradation susceptibility mapping in the study of gravely undulating red and lateritic agro-climatic zones.
Support Vector Machine (SVM)	Used for gully erosion susceptibility prediction.
Multilayer Perceptron Neural Network (MLPNN)	Used for gully erosion susceptibility prediction.
Extreme Gradient Boosting (XGBoost)	Used for gully erosion susceptibility prediction.

Source: Prepared by the authors.

Table 5. Statistical methods mentioned

Methods	**Study context**
Frequency Ratio (FR)	A bivariate statistical method used to evaluate the influence of geo-environmental factors on gully erosion susceptibility.
Modified Information Value (MIV)	Another bivariate statistical method used to assess the influence of geo-environmental factors on gully erosion susceptibility.
Receiver Operating Characteristic (ROC) curve	A graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied.
Area Under the Curve (AUC)	A summary measure of the ROC curve, representing the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one.
True Skill Statistics (TSS)	A measure of the accuracy of a binary classification test.
Root Mean Square Error (RMSE)	A frequently used measure of the differences between values predicted by a model and the values observed.
Mean Absolute Error (MAE)	A measure of errors between paired observations expressing the same phenomenon.
R-squared	A statistical measure of how close the data are to the fitted regression line.

Source: Prepared by the authors.

In sum, the multivariate analysis methods showed in Table 3 were used in combination with the machine learning algorithms showed in Table 4 to develop predictive models for various environmental phenomena, such as gully erosion susceptibility, multi-hazard mapping, soil erosion and landslide susceptibility. Regardin the statistical methods showed in Table 5, they were applied to assess, evaluate and validate the predictive performance of the models developed.

The studies under review predominantly employed a range of programming languages, with Python, R, MATLAB, Java, and C++ emerging as the most commonly utilized ones. Python, renowned for its simplicity and versatility, has gained widespread popularity in scientific computing and data analysis due to its extensive libraries. R, known for its powerful statistical analysis capabilities and comprehensive graphical functionalities, remains a staple in the field of data science and statistical modeling. MATLAB, recognized for its extensive array of mathematical functions and visualization tools, continues to be favored in academic and research settings, particularly in engineering and computational science. Java, celebrated for its platform independence and robustness, finds applications in a diverse range of domains, including enterprise-level software development and mobile applications. Similarly, C++, prized for its efficiency and performance, is often chosen for projects requiring high computational speed and low-level system control. Collectively, these programming languages offer a rich toolkit for researchers to tackle diverse computational challenges across various domains, underscoring their significance in advancing scientific inquiry and innovation (Bogdanchikov et al., 2013; Ozgur et al., 2017).

Although QGIS is widely regarded as the most popular open-source GIS software, mainly due to its robust suite of geospatial analysis tools and a user-friendly and open-source interface (Friedrich, 2014), ArcGIS was the only one mentioned. The absence of QGIS mentions may reflect either a bias towards proprietary software or a lack of emphasis on open-source solutions within the surveyed literature.

In addition to traditional GIS software, the Google Earth Engine (GEE) platform emerged as a notable tool utilized in the analyzed studies. While not strictly a GIS software, GEE provides a powerful cloud-based geospatial analysis platform, enabling researchers to access and analyze large-scale geospatial datasets using Google's computational infrastructure (Zhao et al., 2021). Its ease of use, scalability, and access to a vast repository of remote sensing data make GEE a valuable resource for conducting geospatial analyses, particularly for projects requiring extensive data processing and remote sensing applications.

Within the realm of remote sensing, the studies demonstrated a broad spectrum of applications and techniques. Mentioned prominently were orbital images from Landsat, Sentinel-1, and Sentinel-2 satellites, reflecting their widespread use in environmental monitoring, land cover classification and change detection analyses. Additionally, high-resolution imagery captured by Unmanned Aerial Vehicles (UAVs) emerged as a valuable resource for localized assessments and detailed mapping efforts, including terrain altimetry nformation, particularly in areas where satellite imagery may lack the necessary spatial resolution.

Moreover, the employment of Interferometric Synthetic Aperture RADAR (InSAR) was noted, highlighting its utility in generating precise digital terrain models and monitoring surface deformation with millimeter-level accuracy. InSAR leverages the phase difference between radar signals acquired at different times to infer elevation changes and surface displacements, making it invaluable for studying geohazards, land subsidence, and glacial movements (Osmanoğlu et al., 2016). These diverse remote sensing techniques offer researchers a wealth of information for characterizing Earth's surface dynamics, enhancing our understanding of natural processes and human-induced changes on both local and global scales.

As the discussion about specific characteristics and requirements of studies employing machine learning techniques progresses, it becomes evident that the successful application of predictive models relies heavily on meticulous attention to implementation steps. Particularly in the context of environmental phenomena, such as soil erosion, achieving models that closely mirror real-world scenarios necessitates a thorough consideration of various factors.

Among these, the adequacy and representativeness of the training dataset emerge as critical elements, given the high dependence on sampling in machine learning applications. As such, ensuring that the training dataset encompasses a diverse range of environmental conditions and accurately captures the complexities of soil erosion dynamics becomes paramount. Moreover, addressing issues of data quality, feature selection, and model validation assumes heightened importance in the pursuit of robust and reliable predictive models (Rizeei et al., 2019; Uçar et al., 2020).

By navigating through these challenges, researchers can harness the full potential of machine learning methodologies to unravel intricate environmental processes and inform sustainable management practices effectively. All these mentioned, the second stage of this study, which follows, will address in greater detail aspects related to samping and influence area.

3.3. Machine learning dependence on sampling

Machine learning (ML) encompasses a wide range of computational techniques designed to identify patterns and relationships in data, and, in environmental modeling, it has become a valuable tool for predicting the spatial distribution of diverse phenomena (Jemeļjanova et al., 2024). Most approaches in this field fall under supervised learning, in which the algorithm is trained on labeled datasets—comprising both presence and absence (or occurrence and non-occurrence) samples—alongside a set of explanatory variables derived from environmental, topographic, climatic, or other relevant data sources (Rabby et al., 2023; Stock, 2025; Zhu et al., 2019). The model learns the relationship between these predictors and the known outcomes, enabling it to classify or estimate the likelihood of the phenomenon occurring in unsampled locations. In general, supervised learning workflows involve three main stages: (i) data collection and preprocessing, where datasets are compiled, cleaned, and transformed into suitable input variables; (ii) training and validation, where the model learns from labeled data and is tested to ensure generalizability; and (iii) evaluation, where predictive performance is quantified through statistical and spatial accuracy metrics (L. Breiman et al., 1984; Mitchell, 1997).

This process is inherently dependent on the quality and representativeness of the training data, making the sampling strategy a critical step for ensuring robust and generalizable predictions. In broad terms, sampling involves dealing with class imbalance, reducing dimensionality, balancing data, and using various sampling methods such as systematic, cyclic/sequential, and stratified systematic sampling to improve training time, accuracy, and stability (Uçar et al., 2020).

As it is a spatial phenomenon, which is influenced by the position in space, specific geospatial data management tequiniques are necessary for obtaining representative subsets (Rizeei et al., 2019; Uçar et al., 2020). The most common sampling techniques used in the context of geospatial data are (Eckman and Himelein, 2020; Stein and Ettema, 2003; Stevens and Olsen, 2004; Wang et al., 2012):

- Random Sampling: randomly selecting data points from the entire geospatial dataset. Useful when the distribution of features across the spatial extent is relatively uniform.
- Stratified Sampling: dividing the geospatial area into strata (subsets) based on certain criteria and then sampling randomly from each stratum. Appropriate when

the geospatial area exhibits significant variability, and you want to ensure representation from different strata.

- Cluster Sampling:dividing the entire spatial region into clusters and then randomly selecting entire clusters for analysis. Suitable when spatial autocorrelation is present, and neighboring locations tend to have similar characteristics.
- Systematic Sampling: selecting data points at regular intervals across the geospatial dataset. Useful when there is a regular pattern or grid structure in the geospatial data.
- Spatial Stratified Sampling: similar to stratified sampling but considers the spatial structure of the data. Divides the space into strata based on certain criteria and then samples from each stratum. Effective when there are distinct spatial patterns or clusters in the geospatial data.
- Quasi-random Sampling: using quasi-random sequences like Sobol or Halton sequences to select samples. These sequences have desirable properties for even coverage. Useful when you want a more evenly distributed set of samples compared to completely random sampling.
- Spatially Balanced Sampling: a method that ensures a spatially representative sample by minimizing the imbalance in feature distributions across the spatial extent. Suitable when spatial patterns in the geospatial data need to be captured accurately.
- Adaptive Sampling: Adjusting the sampling strategy dynamically based on the observed data during the sampling process. Useful when the initial understanding of the geospatial data is limited, and the sampling strategy needs to adapt to emerging patterns.

The choice of sampling strategy depends on the characteristics of the target fenomenon, the specific objectives of the analysis and the computational resources available (Stein and Ettema, 2003; Stevens and Olsen, 2004). Considering the use of machine learning in linear erosion analyses, just as important as building an inventory of occurrences of erosion features is building an inventory of non-occurrences (Naghibi et al., 2021).

Non-occurrence sampling, also known as negative or background sampling, is an important aspect of machine learning, especially in binary classification problems. In binary classification, the goal is often to classify instances into one of two classes: positive (presence of a certain feature or event) or negative (absence of that feature or event) (Cucker and Smale, 2001).

The main reason why non-occurrence sampling is crucial in machine learning refers to class imbalance. In many real-world scenarios, as in gully erosion, the classes are imbalanced, meaning the possibility of occurrence has significantly fewer instances than non-occurrence. Non-occurrence sampling helps address this imbalance by providing a representative set of negative examples. Without proper handling of class imbalance, machine learning models may be biased towards the majority class, leading to poor performance on the minority class (Ghori et al., 2021).

In addition, training a machine learning model with only positive instances might result in a model that is overly optimistic and overly sensitive to positive cases. Non-occurrence sampling helps the model learn to distinguish between positive and negative instances, promoting better generalization to new, unseen data. Besides that, non-occurrence sampling helps reduce bias in the model. If the model is trained only on positive instances, it may learn a biased representation of the problem and struggle when faced with negative instances during inference. Incorporating non-occurrence samples ensures a more balanced and unbiased model (Catani et al., 2013).

Preventing overfitting, where a model learns to memorize the training data instead of generalizing to new data, is also achieve by preventing imbalance. This because including negative instances helps the model focus on relevant patterns and avoids memorizing noise present in positive instances. This is especially important when dealing with noisy or imperfect data, where including negative instances aids in learning the true underlying patterns (Y. Liu et al., 2023).

When evaluating model performance, it's important to consider metrics beyond accuracy, such as precision, recall, and F1 score. Also, it is also important consider that the absence of a specific event or condition is equally important as its presence. Then, non-occurrence sampling allows for a more comprehensive assessment of the mentioned metrics, providing insights into the model's ability to correctly classify both positive and negative instances and making it more applicable to various scenarios (Ghori et al., 2021).

3.4. Sampling problem in the context of linear erosion

Considering the dependence of machine learning on sampling and all the premisses described in the previous topic, it become crear that, in the context of environmental and geophysical processes, the selection of a spatially stratified sampling approach is justified by several reasons, among which Terry et al. (2021) and Koldasbayeva et al. (2024) underscore the following:

- Heterogeneous Landscape Conditions: Gully erosion is strongly influenced by topographic, climatic, soil, and land-use conditions, which often vary spatially. A spatially stratified sampling method ensures that all relevant environmental conditions are adequately represented, reducing the risk of underrepresenting critical erosion-prone areas.
- Spatial Autocorrelation Consideration: Since spatial data often exhibit autocorrelation—where neighboring locations share similar characteristics—it is essential to use a sampling method that captures these dependencies. Stratified or cluster sampling approaches help mitigate issues related to redundant information from closely located data points.
- Computational Efficiency: While random sampling may be straightforward, it can lead to redundant data collection in regions with uniform characteristics. By using a spatially balanced or stratified approach, the dataset remains diverse while reducing unnecessary computational complexity.
- Robust Training for Machine Learning Models: Since machine learning models rely on diverse and representative data to learn meaningful patterns, spatially balanced sampling prevents over-representation of any particular land type or erosion condition, thus improving model robustness.
- Alignment with Real-World Decision-Making: In practical applications, land managers and policymakers often rely on spatially explicit models to implement mitigation strategies. By ensuring that the training dataset represents real-world heterogeneity, the model outputs become more applicable for decision-making processes.

In addition to sampling design, several studies emphasize the importance of considering spatial autocorrelation also in the validation stage of machine learning workflows. Jemeļjanova et al. (2024) reviewed how environmental studies using machine learning have addressed spatial autocorrelation across the modeling pipeline—from data exploration to validation. Their findings highlight the lack of a standardized approach, with context-specific strategies proving most effective. In particular, the use of explicit spatial covariates and informed data splitting methods stood out as robust strategies to reduce spatial bias and improve model reliability. Similarly, Liu, Kounadi and Zurita-Milla (2022) demonstrated that the integration of spatial features, such as spatial lag and eigenvector spatial filtering, into random forest models significantly reduced prediction errors and spatial autocorrelation in residuals, improving overall model performance.

Complementary evidence is provided by Filippelli et al. (2024), who warned that random cross-validation schemes can lead to overoptimistic estimates of model performance in spatially structured data. Their work reinforces the need for spatially aware validation strategies—such as spatial blocking or buffering—to ensure independent and representative testing. In the same vein, Stock (2025) demonstrated that spatial block cross-validation, particularly when aligned with natural subregions, leads to more realistic performance metrics. However, block size and configuration must be carefully calibrated to balance bias and variance. These insights underscore the relevance of integrating spatial validation practices alongside sampling strategies to enhance the robustness and transferability of machine learning models in erosion studies.

By carefully selecting an appropriate sampling strategy that includes both erosion-prone and stable areas, and by considering the spatial heterogeneity of the study area, a more accurate and generalizable model for predicting gully erosion susceptibility can be ensured. Then, issues related to heterogeneity and spatial distribution will be addressed in the next section (3.5). Meanwhile, the specific aspects of practical sampling strategies will be discussed considering six main documents, as catalogued in Table 6.

Table 6. Documents considered for review on sampling in erosion context.

Title	Authors	Date	Source	Geographic scope
Assessing the performance of GIS- based machine learning models with different accuracy	Garosi, Y.; Sheklabadi, M.; Conoscenti, C.;	2019	Science of The Total Environment	Hamedan, western Iran

measures for determining susceptibility to gully erosion	Pourghasemi, H. R.; Van Oost, K.			
Machine learning digital soil mapping to inform gully erosion mitigation measures in the Eastern Cape, South Africa	du Plessis, C.; van Zijl, G.; van Tol, J.; Manyevere, A.	2020	Geoderma	Eastern Cape, South Africa
APG: A novel python-based ArcGIS toolbox to generate absence-datasets for geospatial studies	Naghibi, S.A.; Hashemi, H.; Pradhan, B.	2021	Geoscience Frontiers	Chaharmahal and Bakhtiari Province, Iran
Investigation of the influence of nonoccurrence sampling on landslide susceptibility assessment using Artificial Neural Networks	Lucchese, L. V.; de Oliveira, G. G.; Pedrollo, O.C.	2021	Catena	Rio Grande do Sul, Brazil
Stepwise integration of analytical hierarchy process with machine learning algorithms for landslide, gully erosion and flash flood susceptibility mapping over the North-Moungo perimeter, Cameroon	Mfondoum, A.H.N.; Nguet, P.W.; Seuwui, D.T.; Mfondoum, J.V.M.; Ngenyam, H.B.; Diba, I.; Tchindjang, M.; Djiangoue, B.; Mihi, A.; Hakdaoui, S.; Batcha, R.; Tchatchouang, F.C.L; Petcheu, I.C.N.; Beni, L.M.	2023	Geoenvironmental Disasters	North-Moungo perimeter, Cameroon
Optimizing the Sample Selection of Machine Learning Models for Landslide Susceptibility Prediction Using Information Value Models in the Dabie Mountain Area of Anhui, China	Liu, Y.; Meng, Z.; Zhu, L.; Hu, D.; He, H.	2023	Sustainability	Dabie Mountain, Anhui, China
Addressing class imbalance in soil movement predictions	Kumar, P.; Priyanka, P.; Uday; K. V.; Dutt, V.	2024	Nat. Hazards Earth Syst. Sci.	Uttarakhand, northern India
Evaluation of landslide susceptibility based on SMOTE-Tomek sampling and machine learning algorithm	Lv M-z; Li K-l; Cai J-z; Mao J; Gao J-j; Xu H.	2025	PLoS One	Taiping Township, Zhejiang Province, China

Source: Prepared by the authors.

Naghibi et al. (2021) proposed a novel Python-based ArcGIS toolbox that generates absence-datasets for geospatial studies: the Absence Point Generation (APG) toolbox. Authors chose to work with a diversity of environmental phenomena, such as potential, susceptibility to floods, gully and landslides and their main goal was to automate the construction of absence-datasets and improve the quality of your outputs.The methodology began with the manipulation of presence location data, meaning locations where the analyzed features were observed. Then, the APG toolbox computes and sums the frequency ratio (FR) for all driving factors classes, generates a potential or susceptibility map and eliminate 25% of the most areas with the highest susceptibility values.

Additionally, as geospatial phenomena exhibit a continuous nature that goes beyond the representation of individual pixels; where areas in proximity to flood, landslide and gully spots are generally more susceptible and regions closer to springs are likely to have higher potential; zones closer and with higher presence densities are more likely as well. Based on this assumption, authors incorporated two additional criteria, eliminating an user-definined spring buffer of 300m from the area and 10% of the area with the highest spring densities. Subsequently, a set of *n* absence datasets (as defined by the user) is randomly generated. The absence and presence datasets for training and validation are then combined, resulting in the creation of a datasets for both training and validation machine learning purposes.

Unlike the previously mentioned study, Garosi et al. (2019) do not considered the geospatial characteristics of gully features, because the main goal was to compare four machine learning models: Random Forest (RF), Support Vector Machine (SVM), Naive Bayes (NB) and Generalized Additive Model (GAM). Authors founded RF more reliable for susceptibility assessment (Garosi et al., 2019).

Mfondoum et al. (2023) constructed a stepwise integration of analytical hierarchy process (AHP) with machine learning (ML) for landslide, gully erosion and flood susceptibility mapping and risk assessment. Overall, AHP was used to identify and prioritize the factors that contribute to the occurrence of geoenvironmental hazards, helping to focus on the most critical factors and allocate resources more effectively. Then, ML algorithms were used to develop predictive models that can identify areas at high risk, creating susceptibility maps. Classification and Regression Trees (CART), Random Forest (RF) and Gradient Boosting Regression Trees (GBRT) were used as basic learners of the stacked hazard factors, whereas, Support Vector Regression (SVR) was used for a meta-learning.

This study differs from Naghibi et al. (2021) and Garosi et al. (2019) because does not explicitly mention the use of non-occurrence or absence sampling, however, Mfondoum et al. (2023) do mention the use of event records and analyst knowledge. Essentially, this knowledge was used to supplement the event records and identify additional areas from low to high risk of geoenvironmental hazards, based on the knowledge of analysts familiar with the study area. This approach likely allowed the researchers to address the nonoccurence samples problem by leveraging their knowledge of the study area and the available event records to generate representative samples for the predictive models. Liu et al. (2023) did something similar, by creating a disaster zoning map from low to high susceptibility and selecting non-occurrence units among the low-susceptibility area (Y. Liu et al., 2023).

Although being focused on piping, du Plessis et al. (2020) bring valuable insights into the impact of soil erosion near a road network and the measures being taken to mitigate it. Authors describes the development of a Machine learning-based digital soil mapping, created to identify areas where piping could occur. The methodology is based on sampling soil properties at 600 locations within a 500 m buffer around the available road network using the conditioned Latin hypercube sampling (LHS) method.

LHS is a statistical method used for sampling from a multivariate parameter space (Minasny and McBratney, 2006) and, in the context of the study, ensured that the samples were representative of the parameter space by dividing the range of each parameter into equally probable intervals and selecting one sample from each interval. This approach is proven to be efficient given the need to cover the entire range of ancillary variables, ensuring that the samples are evenly distributed across the parameter space, reducing the likelihood of bias and improving the accuracy of the resulting soil database (McBratney et al., 2003; Minasny and McBratney, 2006).

Regarding class imbalance, several studies on landslide indicate that ratio of ccurrence samples and non occurrence should be equal to 1 (Nefeslioglu et al., 2008; Schicker and Moon, 2012; Süzen and Doyuran, 2004). Following this premise, Naghibi et al. (2021) and Lucchese et al. (2021) equally calibrated presence and absence points data, in other words, for each gully sampling, a non-gully sampling was generated (Naghibi et al., 2021). Garosi et al. (2019) did the same, however, in terms of pixels number. In this study, after recognizing and counting pixels occupied by gully erosion, authors randomly selected the same number of non-gully erosion pixels.

Liu et al. (2023) applied a positive/negative ratio of 1:1.2 for the selection of non-landslide points to maintain a balanced approach to sample selection and to ensure that the dataset adequately represented the natural conditions of the study area. This refers to the predominance of non-occurrence cases observed in real world and, by choosing a positive/negative ratio of 1:1.2, authors aimed to mitigate potential biases that could arise from an imbalanced representation, contribute to the robustness and generalization ability of the machine learning model and lead to improved model performance in identifying both landslide-prone and stable areas.

In scenarios where class imbalance is pronounced, as commonly seen in erosion mapping studies, traditional random undersampling or equal ratio strategies may not always provide optimal model performance. To enhance model robustness and sensitivity to minority classes (i.e., erosion occurrences), recent studies suggest employing data augmentation techniques such as the Synthetic Minority Over-sampling Technique (SMOTE) (Lv et al., 2025) or Adaptive Synthetic Sampling (ADASYN) (Kumar et al., 2024). In sum, these techniques generate synthetic samples of the minority class by interpolating between existing data points, improving the learning process of the model by expanding the decision boundaries in underrepresented regions of the feature space.

In their studies, Kumar et al. (2024) and Lv et al.(2025) explored a range of oversampling techniques—such as SMOTE, K-means SMOTE, borderline-SMOTE, SMOTE-Tomek and ADASYN—in the context of soil movement prediction and landslide susceptibility mapping. Their findings revealed that models trained without oversampling consistently underperformed. These results reinforce the promising potential of synthetic oversampling in handling class imbalance, especially in highly skewed environmental datasets where the minority class (e.g., occurrence of erosion or landslides) is underrepresented.

Although not widely applied in gully erosion studies yet, such approaches have shown promising results in other environmental modeling contexts (e.g., landslides, floods), where class imbalance also poses challenges. For example, Liu et al. (2023) emphasize the need to balance datasets when using machine learning models for landslide prediction, highlighting that imbalanced data may hinder the detection of true positives in hazard mapping. The integration of techniques like SMOTE and ADASYN can mitigate these issues by increasing the presence of meaningful minority samples, thereby improving sensitivity and predictive accuracy without requiring extensive new field data.

Furthermore, combining spatially aware sampling strategies with synthetic balancing methods can yield more representative training datasets. This integrated approach not only addresses statistical class imbalance but also retains the spatial heterogeneity crucial to modeling environmental phenomena such as erosion.

Taking into account these studies on linear erosion, machine learning and sampling, they bring relevant contibutions in terms of presence and absense sampling paramters, including location, density and class balance. However, as gully erosion is an hydro-geomorphological phenomenon, driven by several and complex factors, the dynamics of interaction between these factors needs to be considered. (Quoc Bao Pham Kaustuv Mukherjee and Anh, 2020).

3.5. The contributing area role in linear erosion phenomenon

The hydrological contributing area influences linear erosion through geomorphometric factors, climatic factors and human actions, being essential to consider these interactions among the entire topographic area (Bonzanini et al., 2022; Neves, 2022). Several studies suggest that most relevant aspects are slope gradient, drainage area and land cover, with topographic indices being useful for assessing processes' evolution and identifying susceptible areas.

Patton and Schumm (1975) and Begin and Schumm (1979) began modelling gully erosion as a threshold process and assumed that, for a given area of drainage, there is a minimum slope that conditions surface runoff necessary to start the erosion process. From this premise, other approaches were developed to predict how erosive processes are distributed in space, to name a few, the "Composite Topographic Index" (Thorne et al., 1986), the "Topographic Wetness Index" (Moore et al., 1988), the "Sediment Transport Index" (Moore and Burch, 1986) and the "Stream Power Index " (Moore et al., 1988).

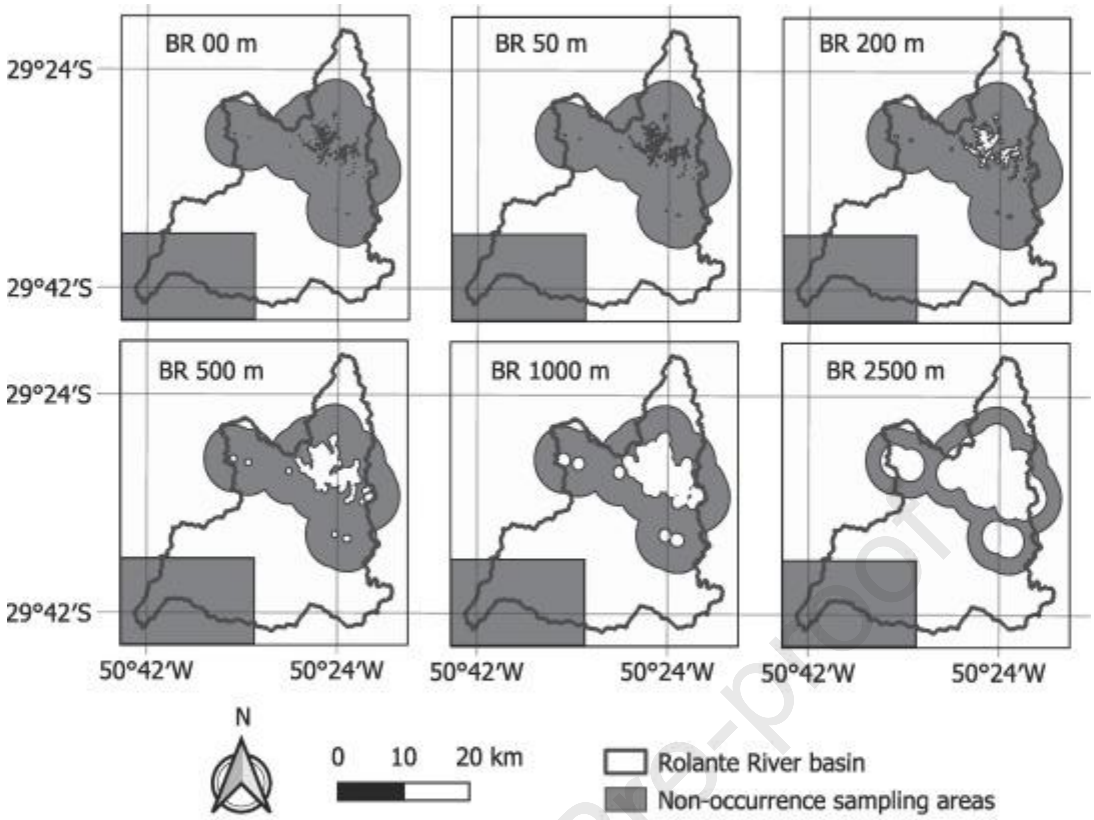
If, on the one hand, each of these indices has a distinct proposal, on the other hand, they all have the specific hydrological contribution area variable. This is an indication of the high correlation in this aspect to erosion development. Similar as a watershed delimited from an outlet, the contribution area (or contributing area or drainage area or catchment area or water/ hydric influence area) can be defined as a basic spatial unit in landscape hydrology and is delineated by the topography of the land, with higher elevations forming natural boundaries that guide the flow of water towards a central drainage point (Jeřábek and Zumr, 2021).

One of the first studies comprising the influence area specifically for linear erosion confirmed a strong topographic control on the extent of the channel network in a watershed, showing that the conditions for gully triggering are controlled strongly by a topographic threshold that has an inverse relationship between upslope catchment area and local gradient (Prosser and Abernethy, 1996). Authors also proposed relating other parameters, such as the presence of piping, rock fragments and land use, however, do not detailed the methodology to do so.

Santisteban et al. (2005) and Muñoz-Robles et al. (2010) explored the relationship between gully erosion and site characteristics (topography, vegetation cover, land use and road infrastructure), focusing on the assessment of the gully volume as a result of a topographic thresholds. Although authors do not consider using the contributing area for attributes incorporation, they concluded that the largest erosional features derive from broader drainage areas.

Lucchese et al. (2021) also do not considered a topography-guided contributing area, however, they proposed three different non-occurrence sampling strategies taking into account the interaction of occurrence samples with the surrounding environment. The first one collected non-occurrence samples from areas within a specified buffer around the landslide scar, as say, a 5 km maximum radius and the minimum distance from the scars varying from 0 to 2.5 km. The rationale is that areas close to where landslides occurred may have had similar triggering conditions, but landslides did not happen. The assumption is that rainfall and other terrain attributes were homogeneous in this proximity. The second strategy extracted non-occurrence sampling from a rectangle located in the valley area, considered a safe region, as landslides are less expected to occur in flat areas. This approach incorporates prior expert knowledge. And the third methodology combines both previous strategies, buffer and rectangle, in which the total number of non-occurrence samples remains constant, resulting in sparser sampling points within these combined areas.

Figure 4. Non-occurrence sampling strategies proposed by Lucchese et al. (2021).



Source: Lucchese et al. (2021).

Instead of concluding which strategy is more efficient, authors conclude that, in many cases, "*expert knowledge intervention should, instead of being applied on the formulation of the sample, (e.g. including a rectangle on the lowlands), be applied to the delimitation of the application areas*" (Lucchese et al., 2021 p.10).

Neves (2022) also based her research on sampling strategies comparison, however, considering the use of 100m buffers and contributing areas - delimitated from head-erosion records - as a separating border to select samples with and without erosions, where gully points were placed within these boundaries and non-gully point outside them. Although the research concluded that buffer was more suitable for variable selection, it is worth mentioning that the author used this limit only for the inside selection of random points, not considering the accumulated influence of the attributes for a registered erosion feature.

In this sense, while traditional non-occurrence sampling strategies may introduce biases due to the inclusion of potentially susceptible areas, integrating topographical thresholds into the sampling process can enhance the representativeness and reliability of non-occurrence samples. This, in turn, leads to more accurate and robust landslide susceptibility models, not only by establishing a reliable methodology for creating non-occurrence samples but also by ensuring that the phenomenon itself - erosive processes - is better represented in terms of its nature as an environmental phenomenon.

Further supporting these premisses, recent studies on absence data sampling in the context of landdslide have demonstrated that considering the spatial distribution of causal factors can significantly improve the quality of non-occurrence samples (Hong et al., 2018; Tsangaratos and Benardos, 2014; Zhu et al., 2019).

For example, the work developed by Rabby, Li and Hilafu (2023) proposed an objective method to determine the critical value for sampling absence data based on the third law of geography: "*if two areas have the same geographical environment, they will experience the same geographical processes*" (Rabby et al., 2023), such as erosion. This study compares two landslide susceptibility maps produced using a random forest model, based on two different absence data sampling methods: (i) a traditional slope-based approach and (ii) a method that incorporates 15 landslide causal factors. While the first approach searches for sample spaces based only on slopes where landslides do not occur, the second identifies areas with the greatest dissimilarities from landslide-prone locations. The susceptibility map generated using the proposed method resulted in higher model fitting and prediction accuracy compared to traditional slope-based sampling methods. Additionally, the authors concluded that maps based on the traditional strategy tended to overestimate high or very high susceptibility zones.

Considering the importance of defining the critical conditions for the generation, development and stabilization of gullies, studies evolving from the ones mentioned above aimed to evaluate the topographical critical limits. Regarding that, one of the most discussed and data explored characterisation of gullies is based on the topographic control of the gully head position, where "*the processes of erosion cannot continue expanding upslope under the given rainstorm intensity and other boundary conditions such as land use, vegetation cover and soil type*" (Torri and Poesen, 2014, p.74).

At this point, it has been showed that the evaluation of the relative importance of the various factors influencing gully formation, which is crucial for a better understanding of gully erosion, relies on a certain influence area. This influence area, frequently associated to the contributing area, can derive from field survey and aerial photographs for identification of erosion spots, followed by drainage area delimitation through Digital Terrain Model (DTM) (Kakembo et al., 2009; Stabile, 2013; Tang et al., 2022; Yu et al., 2023).

Torri and Poesen (2014) conducted a comprehensive literature review on topographic threshold conditions for gully head development in different environments. Authors have highlighted the importance of considering the effects of land use on gully development and the need for more detailed studies to obtain reliable models for prediction and susceptibility.

In terms of land use aspects, models need to cover a complex processes and dynamics. For example, depending of the influence area land uses, there is more or less surface storage of rain water. If the surface use contains vegetation, it is protected from drop impact and delays soil surface sealing. Additionally, soil aggregates are more stable if organic matter is present in the soil, also plant roots increase macroporosity and infiltration capacity. Furthermore, some types of land use can increase friction to overland flow, decreasing runoff velocity and absorbing part of the flow energy (Baets et al., 2007; Torri et al., 2013). There are also impermeable areas which runoff is entire transferred to another land use, superficially and through pipelines, increasing rainwater discharge in a reduce time amount (Torri and Poesen, 2014).

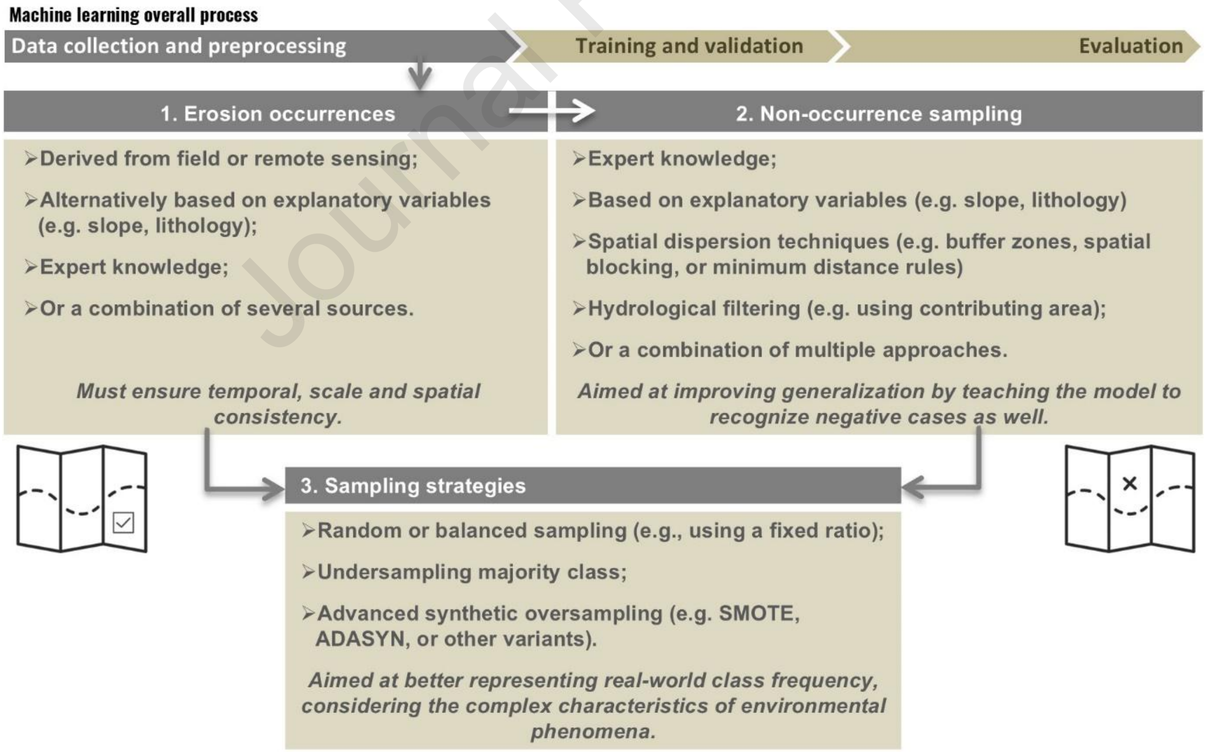
In this context, most of literature review focused on equations' parameters. However, Torri and Poesen (2014) concluded that most studies only provide a qualitative description of the landscape, without specific details on the features of each gully catchment area. Then, they suggested the development of a physically-based model for predicting gully development in various landscapes and under changing environmental conditions. Essentially, assuming natural conditions for rainfall parameters, soil types and slope gradients, land use characterization of the catchment area boundary represents the main interaction leading to an accelerated linear erosion development.

Contrary to this premise, in all the sampling techniques gathered throughout this review, either only topographic attributes of the influence area are incorporated, or this area is used only as a border for the random sampling selection of points or pixels. None of them used the drainage area as an accumulative factor for water dynamics, in other words, the size of each land use surface among the contributing area have not been be integrated into gully erosion susceptibility machine learning models. This means that both, occurrence and non-occurrence samples, could incorporate the land use attributes of the drainage area of each point, proportionally to the surface occupied by different uses.

3.6. Challenges in generalization and applicability

To consolidate the main elements discussed so far, Figure 5 provides a schematic overview of the methodological workflow and key findings of this study. The diagram highlights the data collection and preprocessing stage—central to our analysis—which focuses on sampling strategies for environmental applications. Special attention is given to addressing class imbalance between erosion and non-erosion instances and to the use of contributing area for accurately locating non-erosion points. This synthesis serves as a bridge to the following discussion, which addresses one final and critical challenge: ensuring model generalization and applicability across diverse geographic and environmental contexts.

Figure 5. Summary of the main findings of this study.



Source: Prepared by the authors.

The application of machine learning models in environmental studies, such as erosion prediction, faces significant challenges due to geographic variability and distinct environmental conditions. One of the central questions is whether to develop highly accurate local models for specific areas or broader global models that may be less precise but more widely applicable. Additionally, improving the generalization ability of these models in complex and dynamic geographic environments is essential to ensure their effectiveness.

Local models are trained with data specific to a given region, capturing particular environmental nuances such as soil type, climate, and land use. This specificity results in more accurate predictions within that area, as the model can better represent the spatial and environmental relationships involved. However, this approach has limitations, as the model's applicability remains restricted to the studied area. This means multiple models would need to be developed for different regions, which can be costly and require extensive local datasets that are not always available (Rolf et al., 2024; Shi et al., 2022).

On the other hand, global models are designed to cover multiple regions, allowing for broader applicability. These models use data collected from various geographic areas and can serve as a standardized tool for large-scale environmental analysis. However, by incorporating diverse environments, such models may lose precision concerning specific local conditions, as their approach tends to smooth out regional environmental variations and reduce the influence of particular features. This simplification can result in lower accuracy, especially when dealing with phenomena like erosion, which are highly influenced by regional variables (Rolf et al., 2024; Shi et al., 2022).

Recent studies reinforce this dichotomy between local and global models. Wu et al. (2022) emphasize the importance of capturing local relationships in environmental data through the use of Graph Neural Networks, while Pan et al. (2023) explore generative approaches for spatial networks, highlighting that considering specific spatial interactions improves the modelling of environmental phenomena. Thus, a major challenge in this research field is defining the balance between specificity and applicability.

Beyond this issue, a fundamental aspect of ensuring the robustness of machine learning models applied to environmental studies is their ability to generalize. In complex and dynamic geographic contexts, models must be capable of making accurate predictions even in scenarios different from those used during training. One of the most effective approaches to improving this capability involves adopting spatial cross-validation strategies that account for data spatial autocorrelation and prevent model accuracy from being overestimated (Tziachris et al., 2023). According to the study conducted by Tziachris et al. (2023), "*spatial cross-validation seems to be able to capture the spatial patterns in the data while also reducing the over-optimism bias that is often associated with conventional random cross-validation methods*". This is especially relevant because spatial data possess unique characteristics that can introduce biases into machine learning models due to their inherent spatial autocorrelation. Additionally, optimizing models through regularization techniques and feature selection helps reduce the risk of overfitting and enhances prediction robustness (Koldasbayeva et al., 2024).

Another crucial factor in improving generalization is the model's ability to incorporate new data over time. Since environmental processes are subject to continuous changes, implementing continual learning techniques becomes a relevant strategy to ensure that models remain updated and effective. Bai et al. (2022) propose the concept of temporal generalization, in which dynamic neural networks are adjusted to capture changes in environmental patterns over time. This approach can be particularly promising for dealing with phenomena such as erosion, which may vary according to climatic events and land-use changes.

Given these challenges, the choice between local and global models directly depends on the specific study objectives and data availability. While local models offer higher accuracy in specific areas, global models ensure broader applicability, although they may lose important details. Regardless of the chosen approach, enhancing model generalization should be prioritized through appropriate validation, optimization, and continual learning strategies, ensuring more accurate and applicable predictions across different environmental contexts.

Conclusions

This study reinforces the critical role of sampling strategies and topographic contributing areas in improving machine learning applications for linear erosion studies. By systematically reviewing bibliometric trends and methodological approaches, we identified key challenges and advancements in integrating machine learning into erosion susceptibility assessments.

A central challenge in environmental machine learning lies in the incorporation of absence data, given the inherent class imbalance of many environmental processes. Unlike controlled experimental settings, real-world phenomena such as linear erosion are dynamic, continuously evolving in response to climatic, topographic, and anthropogenic factors. This variability complicates the reliable identification of true absence locations, as areas currently unaffected by erosion may become susceptible over time. Addressing this issue requires the development of adaptive sampling strategies that account for spatial and temporal changes, ensuring more representative training datasets and improving the predictive accuracy of machine learning models.

The findings highlight that spatially stratified sampling techniques, particularly those considering the hydrological contributing area, significantly enhance model robustness. The strategy of placing non-occurrence samples outside the hydrological contributing area—within a buffer range of 300 to 500 meters—proved effective in capturing environmental variability. Additionally, maintaining a 1:1.2 occurrence-to-non-occurrence ratio was instrumental in ensuring a balanced dataset, reducing potential biases, and improving model generalization. Expert knowledge intervention can also be incorporated.

Considering that non-occurrence sampling represents a critical aspect in machine learning, particularly in addressing class imbalance inherent in many real-world scenarios. Providing a representative set of negative examples helps mitigate biases and ensures the model's ability to generalize to new, unseen data. It also aids in preventing overfitting and promoting better model performance by facilitating the learning of relevant patterns and reducing noise present in positive instances. In this context, the contributing area determined by the topographic threshold also showed potential in assisting the placement process of non-occurrence instances. In addition, the adoption of advanced oversampling techniques and the use of spatial cross-validation methods offer complementary strategies to further improve model robustness and applicability in erosion prediction.

Despite these methodological advancements, challenges remain in fully integrating landscape dynamics, particularly land use characteristics, into predictive models. While topographic attributes are widely incorporated, the cumulative effects of land use within contributing areas are often overlooked. Future research should emphasize physically-based models that account for these interactions to refine erosion susceptibility predictions. Furthermore, spatial cross-validation and continual learning techniques must be prioritized to enhance model applicability across diverse landscapes.

Ultimately, this study reinforces the necessity of refining sampling methodologies and integrating multi-factor environmental data to advance machine learning applications in erosion studies. Addressing these gaps will improve the predictive accuracy and practical applicability of models, fostering more effective soil conservation and land management strategies.

Acknowledgments

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001. The research was supported by National Council for Scientific and Technological Development (CNPq) - Process #311393/2021-7.

Bibliography

Abouhalima, M., das Neves, L., Taveira-Pinto, F., Rosa-Santos, P., 2024. Machine Learning in Coastal Engineering: Applications, Challenges, and Perspectives. J. Mar. Sci. Eng. 12. https://doi.org/10.3390/jmse12040638

Amato, F., Guignard, F., Robert, S., Kanevski, M., 2020. A novel framework for spatio-temporal prediction of environmental data using deep learning. Sci. Rep. 10, 22243. https://doi.org/10.1038/s41598-020-79148-7

Amazing Hope Ekeh, Charles Elachi Apeh, Chinekwu Somtochukwu Odionu, Blessing Austin-Gabriel, 2025. Leveraging machine learning for environmental policy innovation: Advances in Data Analytics to address urban and ecological challenges. Gulf J. Adv. Bus. Res. 3, 456–482. https://doi.org/10.51594/gjabr.v3i2.92

Amiri, M., Pourghasemi, H.R., Ghanbarian, G.A., Afzali, S.F., 2019. Assessment of the importance of gully erosion effective factors using Boruta algorithm and its spatial modeling and mapping using three machine learning algorithms. Geoderma 340, 55–69. https://doi.org/https://doi.org/10.1016/j.geoderma.2018.12.042

Arabameri, A., Nalivan, O., Pal, S., Chakrabortty, R., Saha, A., Lee, S., Pradhan, B., Bui, D., 2020. Novel Machine Learning Approaches for Modelling the Gully Erosion Susceptibility. Remote. Sens. 12, 2833. https://doi.org/10.3390/rs12172833

Baets, S. De, Poesen, J., Knapen, A., Galindo, P., 2007. Impact of root architecture on the erosion-reducing potential of roots during concentrated flow. Earth Surf. Process. Landforms J. Br. Geomorphol. Res. Gr. 32, 1323–1345.

Bag, R., Mondal, I., Dehbozorgi, M., Bank, S.P., Das, D.N., Bandyopadhyay, J., Pham, Q.B., Fadhil Al-Quraishi, A.M., Nguyen, X.C., 2022. Modelling and mapping of soil erosion susceptibility using machine learning in a tropical hot sub-humid environment. J. Clean. Prod. 364, 132428. https://doi.org/10.1016/j.jclepro.2022.132428

Bai, G., Chen, L., Zhao, L., 2022. Temporal Domain Generalization with Drift-Aware Dynamic Neural Network. ArXiv abs/2205.1.

Barakat, A., Rafai, M., Mosaid, H., Islam, M.S., Saeed, S., 2023. Mapping of Water-Induced Soil Erosion Using Machine Learning Models: A Case Study of Oum Er Rbia Basin (Morocco). Earth Syst. Environ. 7, 151–170. https://doi.org/10.1007/s41748-022-00317-x

Begin, Z.B., Schumm, S.A., 1979. Instability of alluvial valley floors: a method for its assessment. Trans. ASAE 22, 347–350.

Bogdanchikov, A., Zhaparov, M., Suliyev, R., 2013. Python to learn programming. J. Phys. Conf. Ser. 423, 12027. https://doi.org/10.1088/1742-6596/423/1/012027

Bonzanini, H.L., Lupinacci, C.M., Stefanuto, E.B., 2022. A Erosão Linear e Sua Relação com a Morfometria do Relevo na Alta Bacia do Rio Capivara – Botucatu (SP). Rev. Bras. Geogr. Física 15, 1947–1964. https://doi.org/10.26848/rbgf.v15.4.p1947-1964

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5 – 32. https://doi.org/10.1023/A:1010933404324

Breiman, L., Friedman, J., Olshen, R., Stone, C., 1984. Classification and Regression Trees. Chapman & Hall, Belmont. CA.

Campaigne, H., 1959. Some experiments in machine learning, in: Proceedings of the Western Joint Computer Conference, IRE-AIEE-ACM 1959. Association for Computing Machinery, Inc, pp. 173 – 175. https://doi.org/10.1145/1457838.1457868

Catani, F., Lagomarsino, D., Segoni, S., Tofani, V., 2013. Landslide susceptibility estimation by random forests technique: sensitivity and scaling issues. Nat. Hazards Earth Syst. Sci. 13, 2815–2831. https://doi.org/10.5194/nhess-13-2815-2013

Chalise, D., Kumar, L., Kristiansen, P., 2019. Land Degradation by Soil Erosion in Nepal: A Review. Soil Syst. 3, 12. https://doi.org/10.3390/soilsystems3010012

Churu, H., Kamau, S., Ng'etich, W., Magiroi, K., Bonface, A., Kebeney, S., Wamalwa, F., Mumo, J., 2024. Drivers of Gully Erosion and its Socio-economic and Environmental Effects in a Tropical Semi-arid Environment. Asian J. Soil Sci. Plant Nutr. 10, 13–26. https://doi.org/10.9734/ajsspn/2024/v10i2256

Codina, L., 2023. Alternative search engines to Google with generative AI: analysis of You.com, Perplexity AI and Bing Chat. INFONOMY 1. https://doi.org/10.3145/infonomy.23.002

Conciani, W., 2008. Processos Erosivos: Conceitos e ações de controle. CEFETMT.

Cronin, M.A., George, E., 2023. The Why and How of the Integrative Review. Organ. Res. Methods 26, 168–192. https://doi.org/10.1177/1094428120935507

Cucker, F., Smale, S., 2001. On the mathematical foundations of learning. Bull. Am. Math. Soc. 39, 1–49. https://doi.org/10.1090/S0273-0979-01-00923-5

Cunha, S.B. da, Guerra, A.J.T., 1996. Degradação ambiental. Geomorfol. e meio Ambient. 4, 337–379.

du Plessis, C., van Zijl, G., Van Tol, J., Manyevere, A., 2020. Machine learning digital soil mapping to inform gully erosion mitigation measures in the Eastern Cape, South Africa. Geoderma 368, 114287. https://doi.org/10.1016/j.geoderma.2020.114287

Dube, H.B., Mutema, M., Muchaonyerwa, P., Poesen, J., Chaplot, V., 2020. A global analysis of the morphology of linear erosion features. CATENA 190, 104542. https://doi.org/https://doi.org/10.1016/j.catena.2020.104542

Eckman, S., Himelein, K., 2020. Methods of Geo-Spatial Sampling, in: Hoogeveen, J., Pape, U. (Eds.), Data Collection in Fragile States. Springer International Publishing, Cham, pp. 103–128. https://doi.org/10.1007/978-3-030-25120-8_7

Falagas, M.E., Pitsouni, E.I., Malietzis, G.A., Pappas, G., 2008. Comparison of PubMed, Scopus, Web of Science, and Google Scholar: strengths and weaknesses. FASEB J. 22, 338–342. https://doi.org/https://doi.org/10.1096/fj.07-9492LSF

Falcão, K. dos S., Panachuki, E., Monteiro, F. das N., da Silva Menezes, R., Rodrigues, D.B.B., Sone, J.S., Oliveira, P.T.S., 2020. Surface runoff and soil erosion in a natural regeneration area of the Brazilian Cerrado. Int. Soil Water Conserv. Res. 8, 124–130. https://doi.org/10.1016/j.iswcr.2020.04.004

Filippelli, S.K., Schleeweis, K., Nelson, M.D., Fekety, P.A., Vogeler, J.C., 2024. Testing temporal transferability of remote sensing models for large area monitoring. Sci. Remote Sens. 9, 100119. https://doi.org/10.1016/j.srs.2024.100119

Frankl, A., Nyssen, J., Vanmaercke, M., Poesen, J., 2021. Gully prevention and control: Techniques, failures and effectiveness. Earth Surf. Process. Landforms 46, 220–238. https://doi.org/10.1002/esp.5033

Friedrich, C., 2014. Comparison of ArcGIS and QGIS for applications in sustainable spatial planning. Master’s thesis, Univ. Vienna.

Garosi, Y., Sheklabadi, M., Conoscenti, C., Pourghasemi, H.R., Van Oost, K., 2019. Assessing the performance of GIS- based machine learning models with different accuracy measures for determining susceptibility to gully erosion. Sci. Total Environ. 664, 1117–1132. https://doi.org/10.1016/j.scitotenv.2019.02.093

Gholami, H., Jalali, M., Rezaei, M., Mohamadifar, A., Song, Y., Li, Y., Wang, Y., Niu, B., Omidvar, E., Kaskaoutis, D.G., 2024. An explainable integrated machine learning model for mapping soil erosion by wind and water in a catchment with three desiccated lakes. Aeolian Res. 67–69, 100924. https://doi.org/10.1016/j.aeolia.2024.100924

Ghori, K.M., Awais, M., Khattak, A.S., Imran, M., Fazal-E-Amin, Szathmary, L., 2021. Treating Class Imbalance in Non-Technical Loss Detection: An Exploratory Analysis of a Real Dataset. IEEE Access 9, 98928–98938. https://doi.org/10.1109/ACCESS.2021.3095145

Gudino-Elizondo, N., Brand, M.W., Biggs, T.W., Hinojosa-Corona, A., Gómez-Gutiérrez, Á., Langendoen, E., Bingner, R., Yuan, Y., Sanders, B.F., 2022. Rapid assessment of abrupt urban mega-gully and landslide events with structure-from-motion photogrammetric techniques validates link to water resources infrastructure failures in an urban periphery. Nat. Hazards Earth Syst. Sci. 22, 523–538. https://doi.org/10.5194/nhess-22-523-2022

Guerra, A.J.T., 2012. O início do Processo Erosivo, in: Erosão e Conservação Dos Solos. Berttrand Brasil, Rio de Janeiro, p. 50.

Gusenbauer, M., Haddaway, N.R., 2020. Which academic search systems are suitable for systematic reviews or meta-analyses? Evaluating retrieval qualities of Google Scholar, PubMed, and 26 other resources. Res. Synth. Methods 11, 181–217. https://doi.org/https://doi.org/10.1002/jrsm.1378

Hong, H., Liu, J., Bui, D.T., Pradhan, B., Acharya, T.D., Pham, B.T., Zhu, A.-X., Chen, W., Ahmad, B. Bin, 2018. Landslide susceptibility mapping using J48 Decision Tree with AdaBoost, Bagging and Rotation Forest ensembles in the Guangchang area (China). Catena 163, 399–413. https://doi.org/10.1016/j.catena.2018.01.005

Jemeļjanova, M., Kmoch, A., Uuemaa, E., 2024. Adapting machine learning for environmental spatial data - A review. Ecol. Inform. 81, 102634. https://doi.org/10.1016/j.ecoinf.2024.102634

Jeřábek, J., Zumr, D., 2021. GEOPHYSICAL SURVEY AS A TOOL TO REVEAL SUBSURFACE STRATIFICATION AT A SMALL AGRICULTURAL HEADWATER CATCHMENT: A CASE STUDY. Stavební Obz. - Civ. Eng. J. https://doi.org/10.14311/cej.2021.03.0059

Junior, O.C., Guimaraes, R., Freitas, L., Gomes-Loebmann, D., Gomes, R.A., Martins, E., Montgomery, D.R., 2010. Urbanization impacts upon catchment hydrology and gully development using mutli-temporal digital elevation data analysis. Earth Surf. Process. Landforms 35, 611–617. https://doi.org/10.1002/esp.1917

Kakembo, V., Xanga, W.W., Rowntree, K., 2009. Topographic thresholds in gully development on the hillslopes of communal areas in Ngqushwa Local Municipality, Eastern Cape, South Africa. Geomorphology 110, 188–194. https://doi.org/https://doi.org/10.1016/j.geomorph.2009.04.006

Kinahan, G.H., 1869. V.—On the Formation of Ravines by Recent Drift Accumulations. Geol. Mag. 6, 406 – 408. https://doi.org/10.1017/S0016756800162922

Koldasbayeva, D., Tregubova, P., Gasanov, M., Zaytsev, A., Petrovskaia, A., Burnaev, E., 2024. Challenges in data-driven geospatial modeling for environmental research and practice. Nat. Commun. 15, 10700. https://doi.org/10.1038/s41467-024-55240-8

Kuhn, C.E.S., Reis, F.A.G.V., Zarfl, C., Grathwohl, P., 2023. Ravines and gullies, a review about impact valuation. Nat. Hazards 117, 597–624. https://doi.org/10.1007/s11069-023-05874-6

Kumar, P., Priyanka, P., Uday, K.V., Dutt, V., 2024. Addressing class imbalance in soil movement predictions. Nat. Hazards Earth Syst. Sci. 24, 1913–1928.

https://doi.org/10.5194/nhess-24-1913-2024

Lin, H., Zhao, Y., 2022. Soil Erosion Assessment of Alpine Grassland in the Source Park of the Yellow River on the Qinghai-Tibetan Plateau, China 9. https://doi.org/10.3389/fevo.2021.771439

Liu, C., Li, H., Xu, J., Gao, W., Shen, X., Miao, S., 2023. Applying Convolutional Neural Network to Predict Soil Erosion: A Case Study of Coastal Areas. Int. J. Environ. Res. Public Health 20. https://doi.org/10.3390/ijerph20032513

Liu, Xiaojian, Kounadi, O., Zurita-Milla, R., 2022. Incorporating Spatial Autocorrelation in Machine Learning Models Using Spatial Lag and Eigenvector Spatial Filtering Features. ISPRS Int. J. Geo-Information 11, 242. https://doi.org/10.3390/ijgi11040242

Liu, Xian, Lu, D., Zhang, A., Liu, Q., Jiang, G., 2022. Data-Driven Machine Learning in Environmental Pollution: Gains and Problems. Environ. Sci. Technol. 56, 2124–2133. https://doi.org/10.1021/acs.est.1c06157

Liu, Y., Meng, Z., Zhu, L., Hu, D., He, H., 2023. Optimizing the Sample Selection of Machine Learning Models for Landslide Susceptibility Prediction Using Information Value Models in the Dabie Mountain Area of Anhui, China. Sustainability 15, 1971. https://doi.org/10.3390/su15031971

Lucchese, L.V., de Oliveira, G.G., Pedrollo, O.C., 2021. Investigation of the influence of nonoccurrence sampling on landslide susceptibility assessment using Artificial Neural Networks. CATENA 198, 105067. https://doi.org/https://doi.org/10.1016/j.catena.2020.105067

Lv, M., Li, K., Cai, J., Mao, J., Gao, J., Xu, H., 2025. Evaluation of landslide susceptibility based on SMOTE-Tomek sampling and machine learning algorithm. PLoS One 20, e0323487. https://doi.org/10.1371/journal.pone.0323487

Martens, H.H., 1959. Two notes on machine “Learning.” Inf. Control 2, 364 – 379. https://doi.org/10.1016/S0019-9958(59)80014-0

McBratney, A.B., Santos, M.L.M., Minasny, B., 2003. On digital soil mapping. Geoderma 117, 3–52.

Mfondoum, A.H.N., Nguet, P.W., Seuwui, D.T., Mfondoum, J.V.M., Ngenyam, H.B., Diba, I., Tchindjang, M., Djiangoue, B., Mihi, A., Hakdaoui, S., Batcha, R., Tchatchouang, F.C.L., Petcheu, I.C.N., Beni, L.M., 2023. Stepwise integration of analytical hierarchy process with machine learning algorithms for landslide, gully erosion and flash flood susceptibility mapping over the North-Moungo perimeter, Cameroon. Geoenvironmental Disasters 10, 22. https://doi.org/10.1186/s40677-023-00254-5

Minasny, B., McBratney, A.B., 2006. A conditioned Latin hypercube method for sampling in the presence of ancillary information. Comput. Geosci. 32, 1378–1388. https://doi.org/https://doi.org/10.1016/j.cageo.2005.12.009

Mitchell, T.M., 1997. Machine Learning. McGraw-Hill, Nova Iorque.

Moore, I.D., Burch, G.J., 1986. Physical basis of the length-slope factor in the universal soil loss equation. Soil Sci. Soc. Am. J. 50, 1294–1298.

Moore, I.D., Burch, G.J., Mackenzie, D.H., 1988. Topographic effects on the distribution of surface soil water and the location of ephemeral gullies. Trans. ASAE 31, 1098–1107.

Moreno-Monroy, A., Schiavina, M., Veneri, P., 2020. Metropolitan areas in the world. Delineation and population trends. J. Urban Econ. 103242. https://doi.org/10.1016/j.jue.2020.103242

Muñoz-Robles, C., Reid, N., Frazier, P., Tighe, M., Briggs, S. V., Wilson, B., 2010. Factors related to gully erosion in woody encroachment in south-eastern Australia. CATENA 83, 148–157. https://doi.org/10.1016/j.catena.2010.08.002

Naghibi, S.A., Hashemi, H., Pradhan, B., 2021. APG: A novel python-based ArcGIS toolbox to generate absence-datasets for geospatial studies. Geosci. Front. 12, 101232. https://doi.org/10.1016/j.gsf.2021.101232

Nefeslioglu, H.A., Duman, T.Y., Durmaz, S., 2008. Landslide susceptibility mapping for a part of tectonic Kelkit Valley (Eastern Black Sea region of Turkey). Geomorphology 94, 401–418.

Neves, M. de P., 2022. Suscetibilidade a processos erosivos lineares apoiada em análise estatística multivariada. Universidade Federal de São Carlos.

Osmanoğlu, B., Sunar, F., Wdowinski, S., Cabral-Cano, E., 2016. Time series analysis of InSAR data: Methods and trends. ISPRS J. Photogramm. Remote Sens. 115, 90–102. https://doi.org/https://doi.org/10.1016/j.isprsjprs.2015.10.003

Ozgur, C., Colliau, T., Rogers, G., Hughes, Z., Myer-Tyson, E. ˮBennieˮ, 2017. MatLab vs. Python vs. R. J. Data Sci. 15, 355–371. https://doi.org/10.6339/JDS.201707_15(3).0001

Pan, A., Musheyev, D., Bockelman, D., Loeb, S., Kabarriti, A.E., 2023. Assessment of Artificial Intelligence Chatbot Responses to Top Searched Queries About Cancer.

JAMA Oncol. 9, 1437–1440. https://doi.org/10.1001/jamaoncol.2023.2947

Pan, B., Qin, M., Wang, S., Zhang, Y., Zhao, L., 2023. Controllable Data Generation Via Iterative Data-Property Mutual Mappings. arXiv Prepr. arXiv2310.07683.

Paré, G., Trudel, M.-C., Jaana, M., Kitsiou, S., 2015. Synthesizing information systems knowledge: A typology of literature reviews. Inf. \& Manag. 52, 183–199.

Patton, P.C., Schumm, S.A., 1975. Gully erosion, Northwestern Colorado: a threshold phenomenon. Geology 3, 88–90.

Phionah, B., Extension, K.P., 2024. Effect of Gully Erosion on the Environment 5, 95–100.

Poesen, J., Nachtergaele, J., Verstraeten, G., Valentin, C., 2003. Gully erosion and environmental change: Importance and research needs. Catena 50, 91 – 133. https://doi.org/10.1016/S0341-8162(02)00143-1

Polovina, S., Radić, B., Ristić, R., Kovačević, J., Milčanović, V., Živanović, N., 2021. Soil Erosion Assessment and Prediction in Urban Landscapes: A New G2 Model Approach. Appl. Sci. 11. https://doi.org/10.3390/app11094154

Pourghasemi, H., Sadhasivam, N., Kariminejad, N., Collins, A., 2020. Gully erosion spatial modelling: Role of machine learning algorithms in selection of the best controlling factors and modelling process. Geosci. Front. 11, 2207–2219. https://doi.org/10.1016/j.gsf.2020.03.005

Pourghasemi, H.R., Yousefi, S., Kornejady, A., Cerdà, A., 2017. Performance assessment of individual and ensemble data-mining techniques for gully erosion modeling. Sci. Total Environ. 609, 764 – 775. https://doi.org/10.1016/j.scitotenv.2017.07.198

Prosser, I.P., Abernethy, B., 1996. Predicting the Topographic Limits to a Gully Network Using a Digital Terrain Model and Process Thresholds. Water Resour. Res. 32, 2289–2298. https://doi.org/10.1029/96WR00713

Quoc Bao Pham Kaustuv Mukherjee, A.N.N.T.T.L.S.J.K.A.A.C.T.N.C.D., Anh, D.T., 2020. Head-cut gully erosion susceptibility modelling based on ensemble Random Forest with oblique decision trees in Fareghan watershed, Iran. Geomatics, Nat. Hazards Risk 11, 2385–2410. https://doi.org/10.1080/19475705.2020.1837968

Rabby, Y.W., Li, Y., Hilafu, H., 2023. An objective absence data sampling method for landslide susceptibility mapping. Sci. Rep. 13, 1740. https://doi.org/10.1038/s41598-023-28991-5

Rahmati, O., Tahmasebipour, N., Haghizadeh, A., Pourghasemi, H., 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

Rizeei, H.M., Pradhan, B., Saharkhiz, M.A., Lee, S., 2019. Groundwater aquifer potential modeling using an ensemble multi-adoptive boosting logistic regression technique. J. Hydrol. 579, 124172. https://doi.org/10.1016/j.jhydrol.2019.124172

Rolf, E., Gordon, L., Tambe, M., Davies, A., 2024. Contrasting local and global modeling with machine learning and satellite data: A case study estimating tree canopy height in African savannas. arXiv Prepr. arXiv2411.14354.

Romanelli, J.P., Fujimoto, J.T., Ferreira, M.D., Milanez, D.H., 2018. Assessing ecological restoration as a research topic using bibliometric indicators. Ecol. Eng. 120, 311–320. https://doi.org/10.1016/j.ecoleng.2018.06.015

Saha, A., Pal, S.C., Arabameri, A., Chowdhuri, I., Rezaie, F., Chakrabortty, R., Roy, P., Shit, M., 2021. Optimization modelling to establish false measures implemented with ex-situ plant species to control gully erosion in a monsoon-dominated region with novel in-situ measurements. J. Environ. Manage. 287, 112284. https://doi.org/10.1016/j.jenvman.2021.112284

Santisteban, L.M. De, Casalí, J., López, J.J., Giráldez, J. V., Poesen, J., Nachtergaele, J., 2005. Exploring the role of topography in small channel erosion. Earth Surf. Process. Landforms 30, 591–599. https://doi.org/10.1002/esp.1160

Schicker, R., Moon, V., 2012. Comparison of bivariate and multivariate statistical approaches in landslide susceptibility mapping at a regional scale. Geomorphology 161–162, 40–57. https://doi.org/10.1016/j.geomorph.2012.03.036

Schmidt, L., Heße, F., Attinger, S., Kumar, R., 2020. Challenges in Applying Machine Learning Models for Hydrological Inference: A Case Study for Flooding Events Across Germany. Water Resour. Res. 56. https://doi.org/10.1029/2019WR025924

Sharifi, A., Khavarian-Garmsir, A.R., Allam, Z., Asadzadeh, A., 2023. Progress and prospects in planning: A bibliometric review of literature in Urban Studies and Regional and Urban Planning, 1956–2022. Prog. Plann. 173, 100740. https://doi.org/https://doi.org/10.1016/j.progress.2023.100740

Shi, H., Luo, G., Hellwich, O., Xie, M., Zhang, C., Zhang, Y., Wang, Y., Yuan, X., Ma, X., Zhang, W., Kurban, A., De Maeyer, P., Van de Voorde, T., 2022. Variability and uncertainty in flux-site-scale net ecosystem exchange simulations based on machine learning and remote sensing: a systematic evaluation. Biogeosciences 19, 3739–3756. https://doi.org/10.5194/bg-19-3739-2022

Silva, F.X., Barbosa, E.P., Mendes, V. d. R., Santos, A.L.F. dos, 2021. Three Review Studies Types in Educational Research: Characterization and Analysis. SciELO Prepr. https://doi.org/10.1590/SciELOPreprints.2897

Silva, T.H.F.E., Souza, V.C.O., Paula, M.M. V, 2021. A systematic mapping of literature on the use of geotechnologies in smart cities in the government area, in: ACM International Conference Proceeding Series. https://doi.org/10.1145/3466933.3466942

Stabile, R.A., 2013. Limites críticos topográficos de feições erosivas lineares em Piratininga (SP). Universidade de São Paulo, São Paulo. https://doi.org/10.11606/D.8.2013.tde-16042013-114853

Stein, A., Ettema, C., 2003. An overview of spatial sampling procedures and experimental design of spatial studies for ecosystem comparisons. Agric. Ecosyst. Environ. 94, 31–47. https://doi.org/10.1016/S0167-8809(02)00013-0

Stevens, D.L., Olsen, A.R., 2004. Spatially Balanced Sampling of Natural Resources. J. Am. Stat. Assoc. 99, 262–278.

Stock, A., 2025. Choosing blocks for spatial cross-validation: lessons from a marine remote sensing case study. Front. Remote Sens. 6. https://doi.org/10.3389/frsen.2025.1531097

Süzen, M.L., Doyuran, V., 2004. A comparison of the GIS based landslide susceptibility assessment methods: multivariate versus bivariate. Environ. Geol. 45, 665–679.

Tang, J., Liu, G., Xie, Y., Wu, Y., Wang, D., Gao, Y., Meng, L., 2022. Effect of topographic variations and tillage methods on gully erosion in the black soil region: A case-study from Northeast China. L. Degrad. Dev. 33, 3786–3800. https://doi.org/10.1002/ldr.4423

Telles, T.S., Guimarães, M. de F., Dechen, S.C.F., 2011. The costs of soil erosion. Rev. Bras. Ciência do Solo 35, 287–298. https://doi.org/10.1590/S0100-06832011000200001

Terry, N., Johnson, C.D., Day-Lewis, F.D., Parker, B., Slater, L., 2021. Beware of Spatial Autocorrelation When Applying Machine Learning Algorithms to Borehole

Geophysical Logs. Groundwater 59, 315–319. https://doi.org/https://doi.org/10.1111/gwat.13081

Thorne, C.R., Zevenbergen, L.W., Grissinger, E.H., Murphey, J.B., 1986. Ephemeral gullies as sources of sediment, in: Proceedings of the Fourth Federal Interagency Sedimentation Conference March 24-27, 1986, Las Vegas, Nevada.

Torri, D., Poesen, J., 2014. A review of topographic threshold conditions for gully head development in different environments. Earth-Science Rev. 130, 73–85. https://doi.org/10.1016/j.earscirev.2013.12.006

Torri, D., Santi, E., Marignani, M., Rossi, M., Borselli, L., Maccherini, S., 2013. The recurring cycles of biancana badlands: Erosion, vegetation and human impact. CATENA 106, 22–30. https://doi.org/https://doi.org/10.1016/j.catena.2012.07.001

Tsangaratos, P., Benardos, A., 2014. Estimating landslide susceptibility through a artificial neural network classifier. Nat. Hazards 74, 1489–1516. https://doi.org/10.1007/s11069-014-1245-x

Tziachris, P., Nikou, M., Aschonitis, V., Kallioras, A., Sachsamanoglou, K., Fidelibus, M.D., Tziritis, E., 2023. Spatial or Random Cross-Validation? The Effect of Resampling Methods in Predicting Groundwater Salinity with Machine Learning in Mediterranean Region. Water 15, 2278. https://doi.org/10.3390/w15122278

Uçar, M.K., Nour, M., Sindi, H., Polat, K., 2020. The Effect of Training and Testing Process on Machine Learning in Biomedical Datasets. Math. Probl. Eng. 2020, 1–17. https://doi.org/10.1155/2020/2836236

van Eck, N.J., Waltman, L., 2014. Visualizing Bibliometric Networks, in: Ding, Y., Rousseau, R., Wolfram, D. (Eds.), Measuring Scholarly Impact. Springer International Publishing, Cham, pp. 285–320. https://doi.org/10.1007/978-3-319-10377-8_13

van Eck, N.J., Waltman, L., 2010. Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics 84, 523–538. https://doi.org/10.1007/s11192-009-0146-3

van Raan, A.F.J., 2005. For Your Citations Only? Hot Topics in Bibliometric Analysis. Meas. Interdiscip. Res. Perspect. 3, 50–62. https://doi.org/10.1207/s15366359mea0301_7

Vanmaercke, M., Poesen, J., Van Mele, B., Demuzere, M., Bruynseels, A., Golosov, V., Bezerra, J.F.R., Bolysov, S., Dvinskih, A., Frankl, A., Fuseina, Y., Guerra, A.J.T., Haregeweyn, N., Ionita, I., Makanzu Imwangana, F., Moeyersons, J., Moshe, I., Nazari Samani, A., Niacsu, L., Nyssen, J., Otsuki, Y., Radoane, M., Rysin, I., Ryzhov, Y. V, Yermolaev, O., 2016. How fast do gully headcuts retreat? Earth-Science Rev. 154, 336–355. https://doi.org/https://doi.org/10.1016/j.earscirev.2016.01.009

Vosgerau, D.S.R., Romanowski, J.P., 2014. Estudos de revisão: implicações conceituais e metodológicas. Rev. Diálogo Educ. 14, 165–190.

Wang, J.-F., Stein, A., Gao, B.-B., Ge, Y., 2012. A review of spatial sampling. Spat. Stat. 2, 1–14. https://doi.org/https://doi.org/10.1016/j.spasta.2012.08.001

Wu, L., Cui, P., Pei, J., Zhao, L., Guo, X., 2022. Graph neural networks: foundation, frontiers and applications, in: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. pp. 4840–4841.

Yu, H., Liu, X., Zhan, Y., Gan, S., Deng, X., Wang, X., Zhu, X., 2023. Research on Surface Debris Flow Sensitivity Evaluation Based on RS and GIS, in: Tallón-Ballesteros, A.J., Santana-Morales, P. (Eds.), Digitalization and Management Innovation - Proceedings of DMI 2022, Beijing, China, November 26, 2022, Frontiers in Artificial Intelligence and Applications. IOS Press, pp. 600–612. https://doi.org/10.3233/FAIA230060

Zhao, Q., Yu, L., Li, X., Peng, D., Zhang, Y., Gong, P., 2021. Progress and Trends in the Application of Google Earth and Google Earth Engine. Remote Sens. 13. https://doi.org/10.3390/rs13183778

Zhu, A.-X., Miao, Y., Liu, J., Bai, S., Zeng, C., Ma, T., Hong, H., 2019. A similarity-based approach to sampling absence data for landslide susceptibility mapping using data-driven methods. Catena 183, 104188. https://doi.org/10.1016/j.catena.2019.104188

APÊNDICE C – ARTIGO 3

Received: Added at production | Revised: Added at production | Accepted: Added at production
DOI: xxx/xxxx

RESEARCH ARTICLE

Unveiling Land Use Drivers of Urban Gully Erosion: An Interpretable Machine Learning Approach

Tatiane Ferreira Olivatto[1] | José Augusto Di Lollo[2]

[1]Urban Engineering Graduate Program, Federal University of São Carlos (UFSCar), São Carlos, São Paulo, Brazil

[2]Civil Engineering Department, São Paulo State University (UNESP), Ilha Solteira, São Paulo, Brazil

Correspondence
[1] Rod. Washington Luís, s/n - Monjolinho, São Carlos - SP - Brazil, Zipcode 13565-905.
Email: tatianeolivatto@ufscar.br

Abstract

Surface water erosion has intensified due to human-induced land-use changes. Although linear erosion has been widely studied in rural environments, its dynamics in urban areas—where rapid transformations occur—remain insufficiently explored. Moreover, traditional studies often rely on point-based explanatory variables, disregarding the hydrological nature of the phenomenon. Considering these particularities, this study analyzed linear erosion processes in urban areas across the state of São Paulo, Brazil, examining two decades of land-cover changes and their implications for erosion occurrence. A machine learning model, specifically a Classification and Regression Tree (CART), was employed due to its interpretability, enabling a clear assessment of the relationship between land-use transitions and erosion events. The research also compared data structuring strategies to determine which configuration better captured the influence of land-use dynamics on erosion and non-erosion cases, using hydrological contributing areas as spatial references. Results revealed that land-use changes are strongly related to erosion occurrence, highlighting the role of forest and natural vegetation preservation in preventing the phenomenon. Conversely, transitions toward agricultural and non-vegetated areas—including urbanized zones—were identified as key explanatory factors for erosion occurrence. These findings are expected to support future predictive models and deepen the understanding of urban linear erosion dynamics.

KEYWORDS
Decision Tree, CART, Erosion, Contributing Area.

1 | INTRODUCTION

Soil erosion represents a significant environmental challenge in urban landscapes, with anthropogenic activities playing a crucial role in triggering and accelerating erosive processes (Poesen, Nachtergaele, Verstraeten, & Valentin, 2003; Guo et al., 2019). Among various erosion types, surface water erosion is the most prevalent (Blanco-Canqui & Lal, 2010), manifesting in two primary forms: sheet erosion, characterized by uniform soil particle removal across slopes, and linear erosion, where concentrated surface runoff creates small incisions that may evolve into rills and gullies (Salomão et al., 1999).

In sheet erosion, more frequent in rural areas, the main conditioning factors are precipitation, vegetation cover, topography, and soil characteristics (Salomão et al., 1999; Cunha & Guerra, 2009). These same factors are widely considered in studies on linear erosive processes. However, in urban areas, the influence of anthropogenic landscape modifications stands out, where slope, contributing drainage area, and land use and land cover become determining elements (Chaplot, Giboire, Marchand, & Valentin, 2005; Endres, Pissarra, Borges, & Politano, 2006; Guo et al., 2019).

Pioneering research on linear erosion modeling by Patton and Schumm (1975) and Begin and Schumm (1979) established that for a given drainage area, a minimum slope threshold governs the surface runoff required to initiate erosion. However, these foundational approaches and subsequent traditional methodologies (Moore, Burch, & Mackenzie, 1988; Moore & Burch, 1986; Thorne, Zevenbergen, Grissinger, & Murphey, 1986) treated land cover as a punctual phenomenon rather than integrating it across the entire drainage contribution area. Furthermore, both classical and more recent investigations (Pons, Pejon, & Zuquette, 2007; Galharte, Villela, & Crestana, 2014; Durães & de Mello, 2016; Javidan, Kavian, Pourghasemi, Conoscenti, & Jafarian, 2019; Prieto-Amparán et al., 2019) predominantly focus on rural environments. Conversely, Brazilian urban and peri-urban areas research typically emphasizes understanding surface runoff conditions to evaluate intervention alternatives for process control (Costa et al., 2018; Carvalho et al., 2019; Arantes et al., 2021).

Spatial modeling in urban areas presents particular complexities, as linear erosion is primarily driven by land use dynamics characterized by rapid and continuous changes. Traditional approaches exhibit another limitation by neglecting temporal aspects, considering only current land use as the triggering factor while ignoring the influence of land cover changes over time.

In this context, the main motivation for this study lies in the need to adapt traditional methodologies for modeling erosion susceptibility—widely applied in rural areas—to the specific conditions of urban environments, where land use and land cover undergo rapid and continuous changes. The incorporation of the temporal aspect and anthropogenic dynamics represents an essential advancement for understanding the mechanisms that control the occurrence and evolution of erosive processes in these environments.

The general objective of this work is to investigate the most relevant patterns of land cover changes in triggering linear erosion in urban and peri-urban areas. To this end, we propose to: (i) identify the most suitable structure for incorporating and representing land cover transition classes into the model—considering the buffer and contributing drainage area approaches—used as delimitation strategies for incorporating the land use variable and for allocating the sampling of non-erosion points; and (ii) develop and adjust a classification model capable of identifying land use and land cover change patterns that are more prone to triggering erosive processes in urban areas, evaluating its effectiveness and accuracy.

The main scientific contribution of the study consists of proposing a methodological and operational approach that integrates spatial delimitation strategies (buffer vs. contributing drainage area) with the temporal incorporation of the land use variable, which can be applied to identify and evaluate areas susceptible to linear erosion in urban contexts. This approach aims to support more accurate diagnoses and inform urban planning actions and land management.

2 | LITERATURE REVIEW

A consensus exists among researchers regarding the significant impact of anthropogenic activities on triggering and accelerating erosive processes (Poesen et al., 2003; Guo et al., 2019). In this context, incorporating machine learning (ML) into erosion studies represents a powerful approach for understanding complex relationships among environmental variables, leveraging its capabilities in analysis, modeling, and prediction to support risk assessment (Pourghasemi, Yousefi, Kornejady, & Cerdà, 2017; Amiri, Pourghasemi, Ghanbarian, & Afzali, 2019).

Despite these promising advances, implementing ML in environmental applications presents significant challenges. Notable limitations include insufficient or low-quality data, biased or incomplete datasets, substantial computational resource requirements, and limited techniques for effectively integrating domain-specific expert knowledge. Additionally, scalability challenges arise when modeling extensive geographical areas or addressing regional environmental problems (X. Liu, Lu, Zhang, Liu, & Jiang, 2022).

Environment-specific limitations further complicate ML applications. Temporal dynamics pose difficulties as environmental conditions change over time, and ML models often struggle to capture these evolving patterns (Amato, Guignard, Robert, & Kanevski, 2020). Spatial heterogeneity presents another challenge, as environmental processes frequently exhibit variations that ML models may fail to accurately represent (Schmidt, Heße, Attinger, & Kumar, 2020). Insufficient spatial representativeness can limit model applicability across different geographical regions, while failure to account for these spatiotemporal aspects may lead to inadequate characterization of uncertainties in environmental processes, potentially resulting in overconfident predictions (X. Liu et al., 2022).

Machine learning algorithms can be broadly categorized into supervised and unsupervised approaches. Supervised learning involves training algorithms on labeled datasets to predict outcomes for new inputs, encompassing both classification tasks (assigning categorical labels) and regression tasks (predicting continuous numerical values). Unsupervised learning identifies inherent patterns in unlabeled data through clustering (grouping similar instances) and dimensionality reduction (simplifying data representation while preserving essential characteristics) (?, ?; Alloghani, Al-Jumeily, Mustafina, Hussain, & Aljaaf, 2019).

2.1 | Interpretable ML

Interpretability is crucial in scientific applications where understanding decision processes is as important as prediction accuracy (Miller, 2019; Doshi-Velez & Kim, 2017). Christoph (2020) describes several key situations where interpretability is particularly desirable, including knowledge generation, bias detection, model debugging, and reconciling contradictions between prior knowledge and model outputs.

The evaluation of machine learning model performance is essential in the data mining process, with common metrics including precision, recall, accuracy, and F1-score (Z.-H. Zhou, 2021). However, in many real-world applications (Gorzałczany & Rudziński, 2017), it is important that models demonstrate not only strong performance but also interpretability for users (Piltaver, Luštrek, Gams, & Martinčić-Ipšić, 2016; Q. Zhou, Liao, Mou, & Wang, 2018). Additionally, Doshi-Velez and Kim (2017) emphasizes that even when a model exhibits strong performance based on conventional metrics, it may still provide an incomplete representation of certain real-world tasks.

While performance metrics are well-established, interpretability remains subjective and challenging to quantify (Piltaver et al., 2016; Q. Zhou et al., 2018; Margot & Luta, 2021). Miller (2019) defines interpretability as ”*the degree to which a human can understand the cause of a decision*”, while Murdoch, Singh, Kumbier, Abbasi-Asl, and Yu (2019) provides a more specific definition, describing interpretable machine learning as ”*the use of ML models to extract relevant knowledge about domain relationships contained in data*”.

Among models commonly considered interpretable are linear regression, logistic regression, and decision trees, though this categorization continues to evolve with advances in statistical techniques, visualization tools, and interpretability-focused research (Q. Zhou et al., 2018; Murdoch et al., 2019; Christoph, 2020). The primary limitation of linear regression is its inadequacy for classification problems. Although logistic regression addresses this limitation, its interpretability can be challenging due to the multiplicative nature of weights, unlike the weighted sum formulation of linear regression. Both methods struggle in applications where feature-outcome relationships are nonlinear or involve complex feature interactions. Decision trees, while performing poorly when linear relationships dominate, excel at capturing attribute interactions (Christoph, 2020).

Based on the literature review and the need for interpretability in environmental applications, this work employs Classification and Regression Trees (CART) proposed by (Leo Breiman Jerome Friedman, 1984). This choice was motivated not only by the algorithm's high interpretability through hierarchical structures where each decision can be explicitly explained (Christoph, 2020), but also by its suitability for capturing the nonlinear relationships inherent in the data analyzed in this work.

2.2 | Decision Trees: CART

Decision Trees are supervised learning models applied to both classification and regression tasks. The Classification and Regression Trees (CART) algorithm, introduced by Leo Breiman Jerome Friedman (1984), remains one of the most widely used approaches due to its simplicity, computational efficiency, and high interpretability.

CART constructs trees through recursive binary partitioning of the data. At each step, the algorithm selects the attribute that minimizes impurity within the resulting subsets. For classification, impurity is typically measured using the Gini index, defined for a node t as:

$$Gini(t) = 1 - \sum_{i=1}^{C} p_i^2$$

where p_i denotes the proportion of class i instances in node t, and C is the total number of classes. The split yielding the lowest impurity is chosen as optimal (Leo Breiman Jerome Friedman, 1984).

Tree growth continues until a stopping criterion is met, such as the maximum depth or the minimum number of samples per leaf. These parameters regulate the balance between *overfitting* and *underfitting* (Phalak, Bhandari, & Sharma, 2014). According to Pedregosa et al. (2019) and Buitinck et al. (2013), the most commonly used stopping parameters include:

- **Maximum depth (*max_depth*):** Limits the maximum number of levels in the tree. Deep trees tend to overfit the training data by capturing noise, whereas shallow trees may underfit by failing to model data complexity.
- **Minimum samples required to split a node (*min_samples_split*):** Defines the minimum number of samples required to further divide a node. Larger values help control model complexity by preventing overly specific splits.
- **Minimum samples per leaf (*min_samples_leaf*):** Ensures that each terminal node (leaf) contains a minimum number of samples. This avoids splits that result in leaves with very few observations, which could lead to overfitting.
- **Minimum impurity decrease (*min_impurity_decrease*):** Specifies the minimum impurity reduction required for a split to be performed. This parameter controls split quality and prevents unnecessary branches.

A key advantage of decision trees, and particularly of CART, is their high interpretability. Trees can be visualized hierarchically, and each decision made at a node can be explicitly understood, making the model's decision-making process transparent and accessible (Christoph, 2020).

Furthermore, feature importance can be derived from the contribution of each variable to impurity reduction (e.g., Gini decrease). Variables that appear near the top of the tree generally have greater importance in the model (Sandri & Zuccolotto, 2008). However, despite their simplicity and transparency, decision trees can be sensitive to small variations in the input data, potentially leading to significantly different structures with minor data changes (Hastie, Tibshirani, & Friedman, 2009).

Decision trees also stand out for their computational efficiency. They are generally fast to train and have low memory requirements, especially for small to moderately sized datasets. However, when applied to large datasets or high-dimensional data with many categorical features, their efficiency may decrease, as the algorithm must evaluate all possible splits to identify the optimal partition at each node (Hastie et al., 2009).

Finally, the implementation of decision trees in modern libraries such as *scikit-learn* (Pedregosa et al., 2011) has greatly facilitated their practical use. These libraries provide flexible configuration options for adjusting stopping criteria, pruning methods, and evaluation metrics, allowing researchers and developers to customize models for specific needs while ensuring computational efficiency and reproducibility of experiments.

2.3 | Evaluation Metrics

Model evaluation is a critical step in machine learning to ensure the reliability and validity of results. In binary classification problems—such as the one investigated in this study—several metrics can be used to assess model performance for both positive (1) and negative (0) classes. The main metrics adopted here include accuracy, precision, recall, F1-score, and ROC AUC (*Receiver Operating Characteristic – Area Under the Curve*) (Fawcett, 2006).

Most of these metrics derive from the confusion matrix, which summarizes prediction outcomes into four categories:

TABLE 1 Main evaluation metrics

Metric/Formula	Definition
$Accuracy = \dfrac{TP + TN}{TP + TN + FP + FN}$	Represents the proportion of correctly classified samples among all instances. Although widely used, it can be misleading in the presence of class imbalance.
$Precision = \dfrac{TP}{TP + FP}$	Indicates the proportion of predicted positive cases that are truly positive. It is crucial when minimizing false positives is more relevant than detecting all positives.
$Recall(Sensitivity) = \dfrac{TP}{TP + FN}$	Measures the model's ability to correctly identify positive instances, being essential when minimizing false negatives is a priority.
$F1\text{-}score = 2 \times \dfrac{Precision \times Recall}{Precision + Recall}$	Harmonic mean of precision and recall, balancing both metrics. It is particularly useful for evaluating models trained on imbalanced datasets.
$ROC - AreaUnderCurve(AUC)$	Reflects the model's capacity to discriminate between positive and negative classes at various decision thresholds. The ROC curve plots the true positive rate against the false positive rate, and the AUC represents the classifier's overall performance—values close to 1 denote strong discriminative ability, while 0.5 indicates random prediction.

- True Positives (TP): correctly classified positive instances.
- False Positives (FP): negative instances incorrectly classified as positive.
- True Negatives (TN): correctly classified negative instances.
- False Negatives (FN): positive instances incorrectly classified as negative.

Although accuracy is widely used, it may be misleading for imbalanced datasets. In such cases, macro- and weighted-accuracy variants—computed as the unweighted or weighted mean across classes, respectively—offer fairer performance estimates (Pedregosa et al., 2011).

3 | METHODOLOGY AND METHODS

This study employs an integrated geospatial and machine learning approach to analyze the influence of land use changes on linear erosion occurrence in urban areas. The methodology was applied in São Paulo state, Brazil - the country's most urbanized state with 8,614.62 km² of urban area (IBGE, 2022) and the highest urban population concentration at 96.5% of the state's inhabitants (SEADE, 2020). This context provides an ideal empirical foundation for developing transferable methodologies applicable to rapidly urbanizing regions globally.

The methodological framework comprises five main phases: (1) construction of two distinct datasets using different spatial delimitation strategies; (2) exploratory data analysis to understand data structure and variable relationships; (3) data preparation including train-test-validation splitting with stratification; (4) hyperparameter optimization using cross-validation; and (5) final model validation and interpretation.

Data integration was performed within a Geographic Information System (GIS) environment, combining four primary sources: (i) erosion points inventoried by the Technological Research Institute (IPT, 2012); (ii) land cover maps from 1990 and 2010 provided by the MapBiomas Project (MapBiomas, 2023); (iii) a Digital Elevation Model (DEM) derived from NASA's SRTM mission (CATI, 2016); and (iv) urban census sector polygons from the 2010 Brazilian Census (IBGE, 2010).

3.1 | Sampling strategy

For the proper functioning of supervised machine learning algorithms, it is essential to provide both positive and negative examples of the phenomenon under investigation. In this study, positive examples correspond to erosion occurrences, represented by erosion points inventoried by the Technological Research Institute (IPT, 2012). Since explicit negative erosion samples are not available — given that areas not classified as erosion are implicitly non-erosion — negative points were allocated following a 1:1.2 ratio between positive and negative samples, as recommended by Y. Liu, Meng, Zhu, Hu, and He (2023) for environmental machine learning applications.

The strategy adopted began by delineating the Hydrological Contributing Areas (HCA) associated with IPT-catalogued erosion points through extraction of drainage basins from the DEM. Negative instances (non-erosion areas) were generated by sampling random points within urban sectors (according to census sector polygons) located beyond 2.5 km from the erosion influence zones, thus ensuring spatial separation between classes (Y. Liu et al., 2023; Olivatto & Di Lollo, 2025).

The overall data processing workflow is summarized in the flowchart shown in Figure 1, while the spatial delimitation procedure of the erosion and non-erosion areas is illustrated in Figure 2.

This process resulted in a point vector structure containing a binary response variable `Class`, where value 1 indicates erosion presence (positive class) and value 0 indicates erosion absence (negative class). There are 1,393 positive instances and 1,700 negative instances. These points will be subsequently used to delimit the contributing areas for the next methodological steps.

3.2 | Dataset Construction

Two datasets were systematically constructed to support the development of this study. The overall data construction workflow is summarized in Figure 3.

The difference between them relies on the strategy adopted to delineate the contributing area for the analysis of the phenomenon. In the first dataset, designated as `d_B100`, a circular buffer with a 100-meter

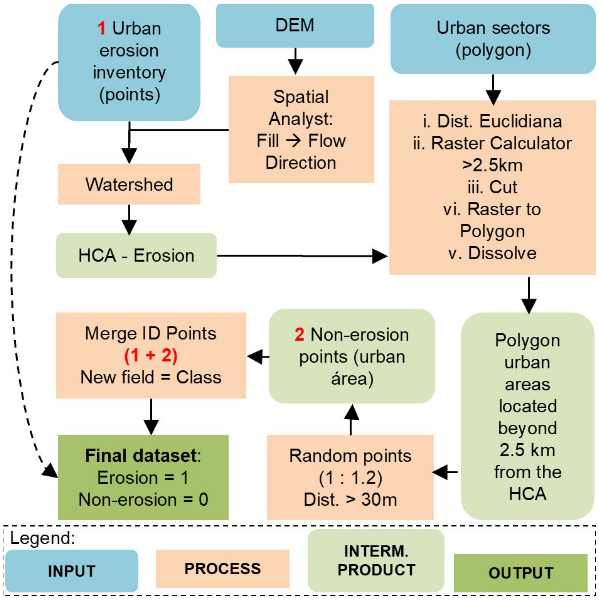


FIGURE 1 Steps for negative sampling and data preparation

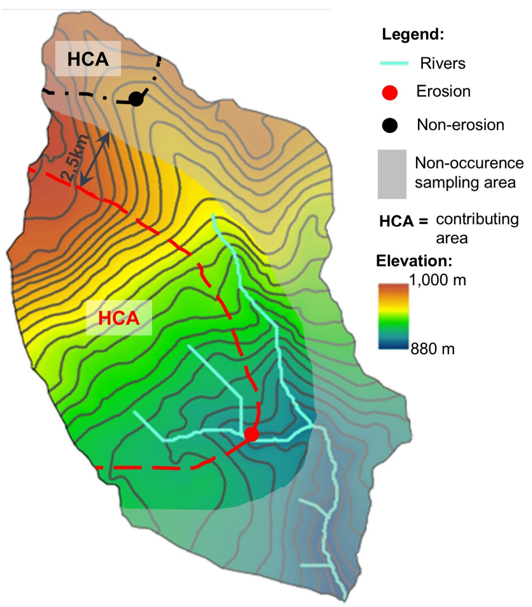


FIGURE 2 Spatial delineation of erosion and non-erosion surfaces

radius was employed, following established protocols (Conoscenti et al., 2014; Dewitte, Daoudi, Bosco, & Eeckhaut, 2015). In the second dataset, designated as `d_HCA`, the delineation criterion was based on the topographic threshold method, defining a HCA for each observation - as suggested in other researches (Neves, 2022; Olivatto & Di Lollo, 2025).

By doing so, each erosion and non-erosion instance corresponds to a specific spatial unit of influence — represented either by a buffer area or by a hydrological contributing area, depending on the dataset. Consequently, both datasets contain the same sample points and binary

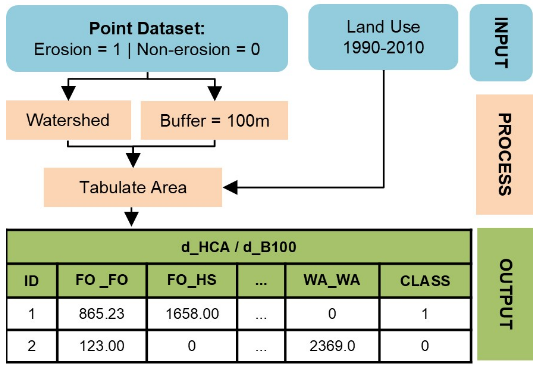


FIGURE 3 Data construction workflow.

response values, differing only in the definition of the spatial unit used for analysis.

Concurrently, the land use transition raster for the 1990-2010 period was developed using GIS mathematical operations (raster calculator), following the methodology proposed by MapBiomas (2023). This enabled identification of land use changes over two decades, categorized into transition classes. The influence areas from both datasets, HCAs and 100-meter buffers, were then intersected with land use transition data, so that each instance represents a distinct area characterized by explanatory variables indicating the proportion of each land-use transition within that area.

These areas were computed as attributes representing the area, in square meters, occupied by land use transition classes. Following GIS processing, data were exported in tabular format for subsequent analysis in Python (Python Software Foundation, 2020) using Google Colab Python Notebook (Google, 2024). In total, both datasets contain the same 25 transition classes, imported according to the official nomenclature from the data source — MapBiomas Project (MapBiomas, 2024) — and renamed to facilitate interpretation (see description in Table 2).

For each observation, the sum of computed values across all attributes equals the total contribution area of that record [`sum(axis=1)`]. In the `d_B100` dataset, this area remains constant since all cases represent a circle with a 100-meter radius. In contrast, for the `d_HCA` dataset, this value varies as each hydrological contributing area has a different size. Therefore, to ensure comparability between observations, frequency-based encoding with row normalization was adopted. This process adjusted each independent variable to represent the percentage occurrence of each transition class, ensuring that values in the same row always sum to 1.00 (Sakai & Goto, 2014; Han, Kamber, & Pei, 2011), as schematically represented in Table 3. To facilitate comparison between datasets, this same strategy was applied to both cases.

An exploratory statistical analysis was conducted to characterize the datasets, including computation of mean, median, standard deviation, and quartiles for all variables. The distribution of land use transition classes was visualized using violin plots, providing insights into data spread and concentration patterns. To investigate relationships

TABLE 2 Land use transition classes coding scheme

Initial Class	Final Class	Transition Code	Renamed Class
Forest	Forest	VALUE_101	FO_FO
Forest	Shrubland	VALUE_110	FO_HS
Forest	Agriculture	VALUE_114	FO_FA
Forest	Non-vegetated	VALUE_122	FO_NV
Forest	Water	VALUE_126	FO_WA
Shrubland	Forest	VALUE_1001	HS_FO
Shrubland	Shrubland	VALUE_1010	HS_HS
Shrubland	Agriculture	VALUE_1014	HS_FA
Shrubland	Non-vegetated	VALUE_1022	HS_NV
Shrubland	Water	VALUE_1026	HS_WA
Agriculture	Forest	VALUE_1401	FA_FO
Agriculture	Shrubland	VALUE_1410	FA_HS
Agriculture	Agriculture	VALUE_1414	FA_FA
Agriculture	Non-vegetated	VALUE_1422	FA_NV
Agriculture	Water	VALUE_1426	FA_WA
Non-vegetated	Forest	VALUE_2201	NV_FO
Non-vegetated	Shrubland	VALUE_2210	NV_HS
Non-vegetated	Agriculture	VALUE_2214	NV_FA
Non-vegetated	Non-vegetated	VALUE_2222	NV_NV
Non-vegetated	Water	VALUE_2226	NV_WA
Water	Forest	VALUE_2601	WA_FO
Water	Shrubland	VALUE_2610	WA_HA
Water	Agriculture	VALUE_2614	WA_FA
Water	Non-vegetated	VALUE_2622	WA_NV
Water	Water	VALUE_2626	WA_WA

Source: Adapted from MapBiomas (2023).

TABLE 3 Data structure schematic representation

ID	Class	Land use transition classes				
		FO_FO	FO_HS	...	A_NV	A_A
0	1	0,0	0,9	...	0,1	0,0
1	0	0,1	0,2	...	0,0	0,7
...	...	...	...	...	...	...
3092	0	0,55	0,05		0,1	0,3

between explanatory variables and erosion occurrence, a comprehensive correlation analysis was performed using three complementary approaches: Pearson correlation to assess linear relationships between continuous variables and binary response; Spearman correlation to evaluate monotonic relationships; and the Mann-Whitney U test for non-parametric comparison of variable distributions between erosion and non-erosion classes (Mann & Whitney, 1947; McKnight & Najab, 2010).

This multi-method approach enabled robust assessment of variable associations with the target phenomenon. Statistical analyses were supported by Matplotlib (Hunter & the Matplotlib development team, 2007) and Seaborn (Waskom & the Seaborn development team, 2021) libraries, while machine learning implementation and evaluation primarily utilized Scikit-Learn (Pedregosa et al., 2011).

3.3 | Experimental Setup and Training

To enable decision tree implementation, each dataset was initially structured into two sets: training-test and validation (in proportions of 80% and 20%, respectively). The training-test set was further divided into training and test subsets, also in an 80%-20% ratio, employing a stratified *K-Fold Cross-Validation* strategy (Kohavi, 1995) with 5 folds. All dataset splits—training, test, and validation—were stratified to maintain the balance between positive and negative erosion instances, as illustrated in Figure 4. The validation set was excluded from training and reserved exclusively for obtaining prediction metrics.

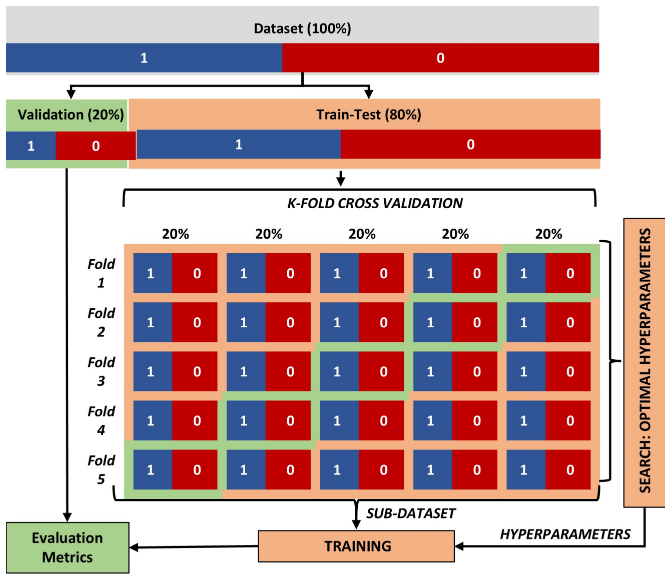


FIGURE 4 Data splitting strategy into subsets

For CART algorithm implementation (Leo Breiman Jerome Friedman, 1984), the `Gini` criterion was used as the impurity measure, along with automatic class weight adjustment (`balanced`). To select optimal model hyperparameters, a parameter combination search was conducted using *RandomizedSearchCV*, with 50 random combinations tested across the following parameter ranges:

- *max_depth*: random integer between 5 and 10 [`randint(5, 10)`];
- *min_samples_split*: random integer between 25 and 50 [`randint(25, 50)`];
- *min_samples_leaf*: random integer between 10 and 30 [`randint(10, 30)`];
- *min_impurity_decrease*: continuous values between 0.001 and 0.009 [`uniform(0.001, 0.009)`].

For each parameter combination, the model was trained using stratified *K-Fold Cross-Validation* with 5 splits based on the training-test set, totaling 250 model fits. The performance metric used to select the most promising hyperparameters was *recall*, also known as sensitivity or true positive rate, which measures the model's ability to correctly identify positive class examples. The choice of recall was motivated by the priority to reduce false negatives, thereby avoiding the loss of positive cases.

3.4 | Model Evaluation and Interpretation

The final models for `d_HCA` and `d_B100` were trained on the training-test sets using the optimized hyperparameters. Subsequently, these models were evaluated on the validation sets, which were previously separated and excluded from the training process. This critical step assesses model performance on completely unseen data, ensuring robust generalization capability. Model evaluation was conducted based on comprehensive metrics including the confusion matrix, classification report (accuracy, precision, recall, and F1-score), and ROC AUC analysis.

Beyond performance metrics, detailed model interpretability was achieved through tree structure visualizations and variable importance analysis, enabling deeper investigation into the underlying erosion processes and key driving factors identified by the models. The analysis extends beyond examining tree structures and feature weights for model explainability, aiming to establish meaningful correlations between the computational results and the physical erosion processes through interdisciplinary collaboration with domain specialists.

4 | RESULTS

4.1 | Statistical Approach

According to Breiman (2001), decision tree–based models — such as the one employed in this study — are not affected by the absolute scale of the variables, as they partition the input space using absolute split thresholds rather than depending on linear relationships among predictors. Although some classes exhibit very low frequencies while others approach unity, no additional transformations were applied, since the adopted frequency-based encoding already meets the purpose of normalizing the proportions.

Table 4 summarizes the descriptive statistics of the explanatory variables. A preliminary inspection of these results indicates that many variables have mean values close to zero, with medians and first quartiles also equal to zero, suggesting that their values are frequently low or null across most of the dataset. However, the classes `FA_FA` and `NV_NV` stand out for presenting higher mean, median, and upper quartile values, suggesting a greater overall representativeness within the dataset.

The comparative analysis of the descriptive statistics of both datasets reveals structural similarities as well as some notable differences in the distribution of attributes. In general, both datasets show relatively low mean values for most classes, suggesting that the presence of land-use transition categories within each sampling unit is often small. However, some differences emerge between the datasets, such as in the case of class `FA_FA`, which has a higher mean value in `d_B100` (0.47) than in `d_HCA` (0.42), along with a lower standard deviation, indicating a more concentrated distribution in this dataset. Moreover, some classes, such as `NV_NV` and `FA_NV`, also display differences in their mean and dispersion values, suggesting regional or methodological variations between datasets.

Regarding variability, the classes exhibit distinct dispersion patterns. Some categories, such as `FO_FO` and `NV_NV`, show high standard deviations in both datasets, indicating greater heterogeneity in the distribution of these transitions. In contrast, several classes, including `HS_HS`, `HS_FA`, and `FO_WA`, present extremely low mean values and standard deviations close to zero, suggesting that these transitions are rare and homogeneous among the samples. The quartile analysis reinforces this trend, as many classes show first and third quartiles equal to zero, indicating that most observations have no positive values for these categories.

Finally, the heterogeneity between the two datasets can also be observed in the range of maximum values. In both datasets, some classes reach maximum values of 1.00, indicating that, in some regions, a given transition class occupies the entire spatial unit. However, differences can be seen in specific classes, such as `FA_FA`, where both the third quartile and the median are higher in `d_B100`, suggesting that this category is more representative in that dataset.

It is important to note that the relationship investigated in this study — the influence of land-use change on erosion occurrence — as well as the data used, refers to a real-world problem. Therefore, considering the spatial scale of data acquisition and the temporal interval adopted, the observed attribute values represent land-use transitions that, in reality, may be sparse and highly variable, as shown in Figure 5. These plots illustrate the behavior of attributes with distributions highly concentrated near zero, confirming the predominance of low values and the absence of significant variability in many of these variables. The combination of these analyses suggests that some variables have limited potential to contribute meaningfully to explaining the response variable.

The CART model implemented in this study performs well with input data exhibiting the characteristics illustrated in Figure 5, where most variables are concentrated around 0. This is because, by splitting the data into segments, the tree selects cutoff values that can effectively adapt to different distributions of explanatory variables, including those highly concentrated around specific values (Leo Breiman Jerome Friedman, 1984).

Regardless of this fact, in a modeling context, it is important to explore the correlations between these continuous variables and the response variable (James, Witten, Hastie, & Tibshirani, 2013). The most traditionally used correlation metrics are those of Pearson (1895) and Spearman (1904). Although the explanatory variables are continuous, the response variable is binary, which requires the use of more appropriate metrics.

Therefore, considering the heterogeneous distribution of the data demonstrated in the violin plot (Figure 5), the Mann–Whitney U test (Mann & Whitney, 1947) was also applied. This test, which is particularly useful when the assumptions of normality or homogeneity of variances

TABLE 4 Descriptive Statistics of the Variables

Class	d_B100							d_HCA						
	Mean	Standard Deviation	Mín.	Q1	Median	Q3	Max.	Mean	Standard Deviation	Mín.	Q1	Median	Q3	Max.
FO_FO	0.13	0.30	0.0	0.00	0.00	0.03	1.00	0.14	0.32	0.0	0.0	0.00	0.00	1.00
FO_HS	0.00	0.01	0.0	0.00	0.00	0.00	0.24	0.00	0.00	0.0	0.0	0.00	0.00	0.03
FO_FA	0.02	0.09	0.0	0.00	0.00	0.00	0.97	0.02	0.12	0.0	0.0	0.00	0.00	1.00
FO_NV	0.00	0.04	0.0	0.00	0.00	0.00	1.00	0.01	0.06	0.0	0.0	0.00	0.00	1.00
FO_WA	0.00	0.00	0.0	0.00	0.00	0.00	0.22	0.00	0.00	0.0	0.0	0.00	0.00	0.05
HS_FO	0.00	0.01	0.0	0.00	0.00	0.00	0.46	0.00	0.02	0.0	0.0	0.00	0.00	1.00
HS_HS	0.00	0.03	0.0	0.00	0.00	0.00	0.64	0.00	0.04	0.0	0.0	0.00	0.00	1.00
HS_FA	0.00	0.01	0.0	0.00	0.00	0.00	0.41	0.00	0.00	0.0	0.0	0.00	0.00	0.25
HS_NV	0.00	0.01	0.0	0.00	0.00	0.00	0.34	0.00	0.03	0.0	0.0	0.00	0.00	1.00
HS_WA	0.00	0.00	0.0	0.00	0.00	0.00	0.00	0.00	0.00	0.0	0.0	0.00	0.00	0.00
FA_FO	0.02	0.07	0.0	0.00	0.00	0.00	0.81	0.02	0.11	0.0	0.0	0.00	0.00	1.00
FA_HS	0.00	0.02	0.0	0.00	0.00	0.00	0.34	0.00	0.03	0.0	0.0	0.00	0.00	1.00
FA_FA	0.47	0.38	0.0	0.03	0.47	0.85	1.00	0.42	0.44	0.0	0.0	0.22	1.00	1.00
FA_NV	0.15	0.25	0.0	0.00	0.00	0.23	1.00	0.17	0.32	0.0	0.0	0.00	0.18	1.00
FA_WA	0.00	0.02	0.0	0.00	0.00	0.00	1.00	0.00	0.03	0.0	0.0	0.00	0.00	1.00
NV_FO	0.00	0.00	0.0	0.00	0.00	0.00	0.16	0.00	0.00	0.0	0.0	0.00	0.00	0.17
NV_HS	0.00	0.00	0.0	0.00	0.00	0.00	0.16	0.00	0.01	0.0	0.0	0.00	0.00	0.29
NV_FA	0.00	0.02	0.0	0.00	0.00	0.00	0.92	0.00	0.02	0.0	0.0	0.00	0.00	1.00
NV_NV	0.16	0.31	0.0	0.00	0.00	0.16	1.00	0.20	0.36	0.0	0.0	0.00	0.22	1.00
NV_WA	0.00	0.00	0.0	0.00	0.00	0.00	0.06	0.00	0.00	0.0	0.0	0.00	0.00	0.05
WA_FO	0.00	0.02	0.0	0.00	0.00	0.00	0.44	0.00	0.02	0.0	0.0	0.00	0.00	0.50
WA_HS	0.00	0.00	0.0	0.00	0.00	0.00	0.19	0.00	0.00	0.0	0.0	0.00	0.00	0.01
WA_FA	0.00	0.02	0.0	0.00	0.00	0.00	0.34	0.00	0.03	0.0	0.0	0.00	0.00	1.00
WA_NV	0.00	0.01	0.0	0.00	0.00	0.00	0.25	0.00	0.01	0.0	0.0	0.00	0.00	0.67
WA_WA	0.01	0.10	0.0	0.00	0.00	0.00	1.00	0.01	0.09	0.0	0.0	0.00	0.00	1.00

Note: Min.= minimum; Q1= 1st Quartile; Q3= 3rd Quartile; Max.= maximum.

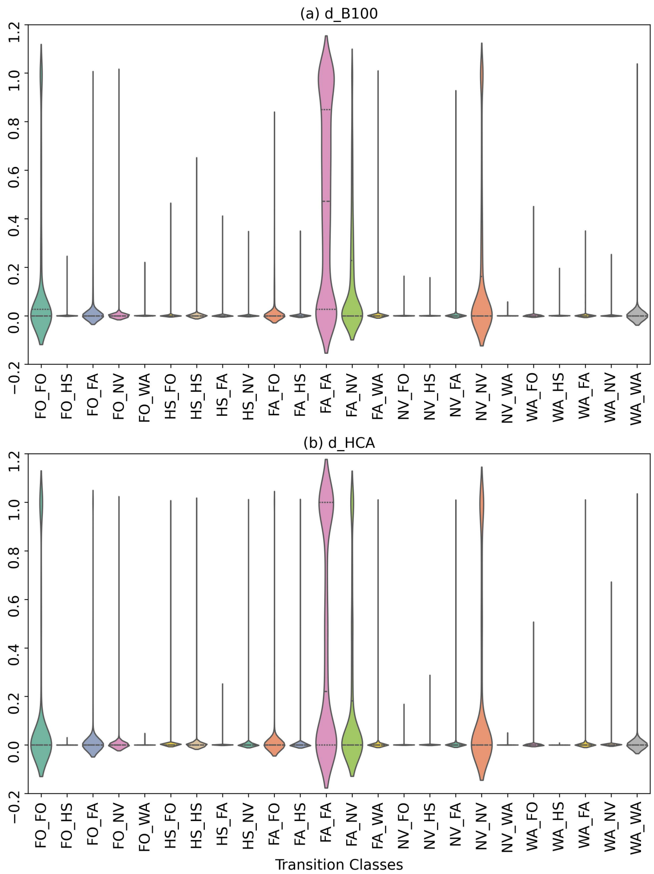


FIGURE 5 ViolinPlot of explanatory variables

are not met since it is a non-parametric test, is widely used to compare the distributions of a continuous variable between two independent groups (McKnight & Najab, 2010)—in this case, the binary values of the response variable. The results are presented in Table **??**.

TABLE 5 Correlation tests results

Class	d_B100			d_HCA		
	Pearson	Spearman	*p-Value*	Pearson	Spearman	*p-Value*
FO_FO	-0.35	-0.34	2.09e-81	-0.35	-0.34	2.49e-79
FO_HS	0.05	0.06	2.18e-03	0.04	0.05	8.07e-03
FO_FA	-0.09	-0.09	2.03e-06	-0.08	-0.05	4.53e-03
FO_NV	-0.01	0.01	6.69e-01	-0.00	0.04	2.52e-02
FO_WA	-0.03	-0.04	2.65e-02	-0.03	-0.06	1.13e-03
HS_FO	0.03	0.04	1.47e-02	-0.01	0.02	2.14e-01
HS_HS	0.04	0.08	2.71e-05	0.03	0.01	6.79e-01
HS_FA	0.02	0.05	4.94e-03	0.03	0.04	4.96e-02
HS_NV	0.04	0.03	7.09e-02	0.03	0.06	1.93e-03
HS_WA	NaN	NaN	1.00e+00	0.02	0.01	5.01e-01
FA_FO	-0.09	-0.07	6.29e-05	-0.08	-0.08	1.17e-05
FA_HS	0.05	0.09	1.50e-06	0.00	0.03	5.92e-02
FA_FA	0.27	0.26	6.68e-48	0.12	0.18	8.23e-23
FA_NV	0.21	0.39	2.24e-104	0.22	0.33	8.05e-76
FA_WA	-0.05	-0.08	1.13e-05	-0.03	-0.06	1.91e-03
NV_FO	-0.01	0.00	8.08e-01	-0.02	0.01	6.70e-01
NV_HS	-0.00	0.00	8.89e-01	-0.02	0.00	8.89e-01
NV_FA	0.02	0.04	1.86e-02	0.01	0.02	3.96e-01
NV_NV	-0.08	0.10	4.70e-08	0.06	0.13	1.80e-13
NV_WA	-0.02	-0.02	2.01e-01	-0.02	-0.03	1.11e-01
WA_FO	-0.07	-0.08	6.05e-06	-0.05	-0.09	1.31e-06
WA_HS	-0.02	-0.00	8.21e-01	0.02	0.01	5.18e-01
WA_FA	-0.06	-0.08	4.35e-06	-0.03	-0.06	1.09e-03
WA_NV	-0.04	-0.05	4.15e-03	-0.02	-0.00	9.19e-01
WA_WA	-0.11	-0.13	2.37e-13	-0.09	-0.11	1.40e-09

The proposed approach, which combines different correlation coefficients, allows for understanding both the linear and monotonic relationships between continuous variables and the categorical response, complemented by a robust hypothesis test for differences between groups.

Pearson's correlation measures the linear relationship between the explanatory variables and the response variable. In both datasets, d_HCA and d_B100, predominantly negative correlations were observed for several variables. For instance, the variable FO_FO showed a correlation of -0.35 in both d_HCA and d_B100, with extremely low p-values (2.49e-79 and 2.09e-81, respectively), indicating a significant inverse relationship between this variable and class 1 of the response variable. This pattern of negative correlation is also observed in other variables, such as FA_FO and WA_FO, suggesting that as the values of these variables increase, the probability of belonging to class 1 decreases. These variables, therefore, play an important role in distinguishing between the classes.

On the other hand, Spearman's correlation, which evaluates monotonic relationships between variables, showed patterns similar to those of Pearson's correlation in many cases. The difference is that, for some variables, the Spearman correlation was slightly higher, which may suggest that the relationship between these variables and the response is not strictly linear but still monotonic. For example, the variable FA_FA exhibited Spearman correlations of 0.18 in d_HCA and 0.26 in d_B100, both statistically significant. This indicates that, although the relationship is not strictly linear, higher values of FA_FA are associated with a higher probability of belonging to class 1, and this association is stronger in d_B100.

Regarding the Mann–Whitney U test, which assesses whether there are significant differences between the distributions of variables for classes 0 and 1, the results support the correlation findings. Many variables, such as FA_FA, FA_NV, FO_FO, and FO_HS, among others, presented extremely low p-values (often below 1e-10), indicating that the distributions of these variables differ significantly between the two classes. This suggests that these variables are indicative of class separability, a decisive aspect for building an effective classification model. The significant Pearson and Spearman correlations for these variables reinforce this conclusion, indicating that they not only exhibit different distributions between classes but also have a consistent association with the response variable.

Conversely, some variables did not show significant differences between classes, such as HS_FO and NV_NV in d_B100. These variables, with higher p-values (above 0.05), suggest that class separation is not well defined by them, which may indicate that they are less useful for distinguishing between classes or that there is greater overlap in their value distributions for classes 0 and 1. In some cases, Pearson and Spearman correlations are also low, supporting this interpretation. In d_HCA, for instance, the variable NV_NV has a p-value of 1.80e-13 and a Spearman correlation of 0.13, indicating that while the correlation is not strong, there is still some statistically significant but less pronounced relationship compared to other variables.

When comparing the two datasets, d_HCA and d_B100, some subtle differences can be observed in correlations and p-values. For several variables, the p-values in d_B100 are slightly lower, as in the case of FO_HS, FA_FA, and FA_NV, suggesting that the explanatory variables have a more pronounced effect on class separation in d_B100. These differences may be attributed to variations in the variable distributions or even structural differences between the two datasets. Similarly, some correlations, such as that of FA_FA, are slightly stronger in d_B100, indicating that this variable has a more robust relationship with the response in that dataset. This behavior can be explored to assess how the characteristics of the two datasets may influence predictive modeling and potentially guide model adjustments.

In terms of class separability, the variables FA_FA, FA_NV, FO_FO, and WA_FO stand out in both datasets, with extremely significant correlations and p-values, suggesting that these variables are the main discriminant factors between classes 0 and 1. Variables with higher p-values, such as HS_WA and NV_NV, indicate lower discriminative power.

Therefore, the analysis of correlations and the Mann–Whitney U test between the datasets d_HCA and d_B100 reveals that, although some subtle differences exist in correlations and p-values, both datasets contain variables with discriminative power for classes 0 and 1. These variables are essential for constructing robust predictive models and for improving the understanding of the relationships between explanatory variables and the response.

An important observation is that a low correlation does not necessarily imply irrelevance for the supervised learning model. In algorithms such as Decision Trees (DT), non-linear interactions between attributes and the response variable can capture patterns not evident from direct correlations (Hastie et al., 2009). Therefore, even variables with weak correlations are considered in the DT model. Another important point is that the CART algorithm can handle complex interactions without requiring explicit data transformations.

In conclusion of the descriptive and exploratory analysis, Pearson and Spearman correlation tests were also conducted to investigate possible correlations among explanatory variables (see Appendix B).

In the d_HCA dataset, Pearson correlations indicate weak relationships among most variables, with some more expressive negative associations, such as between FA_FA and NV_NV ($r \approx -0.45$) and between FA_FA and FA_NV ($r \approx -0.36$). Spearman correlations reinforce this trend, highlighting significant non-linear associations, especially among variables related to the AG component. In the d_B100 dataset, Pearson correlations also suggest predominantly weak relationships, though some variables exhibit more evident structural dependencies. Spearman correlations, which capture monotonic associations, indicate similar patterns, suggesting that while relationships among some variables exist, they do not necessarily follow a strictly linear behavior.

Overall, the analysis suggests that the variables exhibit a low degree of collinearity, which can be advantageous for machine learning models by reducing attribute redundancy.

4.2 | CART training and validation

After organizing the subsets, the training phase was conducted to identify the optimal hyperparameters for the model, as described in the methodological section. The training tests yielded a maximum *recall* of approximately 0.8815 for the d_HCA dataset and 0.0765 for d_B100. For both datasets, the identified hyperparameters were as follows:

- *max_depth* = 7;
- *min_samples_split* = 32;
- *min_samples_leaf* = 13;
- *min_impurity_decrease* ≈ 0.0051.

The final models for `d_HCA` and `d_B100` were trained using the training–testing subsets, with the hyperparameters identified in the optimization stage. Subsequently, these models were evaluated using the validation subsets, which had been previously separated and were not used during training. This step is essential to assess the models' performance on entirely unseen data, ensuring their ability to generalize well. The evaluation was based on classification reports (Table 6), confusion matrices (Figure 6) and ROC–AUC curves (Figure 7).

TABLE 6 Classification reports

	d_HCA			d_B100		
Metric / Class	**Precision**	**Recall**	**F1-Score**	**Precision**	**Recall**	**F1-Score**
0 (Non-erosion)	0.89	0.51	0.65	0.77	0.73	0.75
1 (Erosion)	0.61	0.92	0.73	0.69	0.73	0.71
Accuracy			**0.69**			**0.73**
Macro average	0.75	0.71	0.69	0.73	0.73	0.73
Weighted average	0.76	0.69	0.68	0.73	0.73	0.73

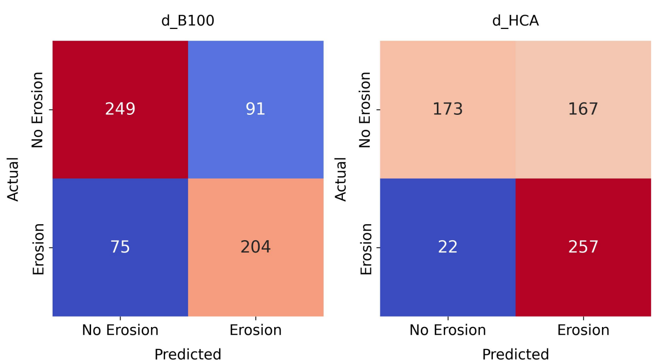


FIGURE 6 Confusion Matrices

The confusion matrices presented in Figure 6 are particularly valuable tools for visualizing the performance of classification models, allowing a detailed examination of the predictions made on the validation set (which was not included in the supervised learning training phase).

For the `d_HCA` dataset, the resulting configuration indicates that, although the model was effective in identifying erosion cases (257 true positives), it also produced a relatively high number of false positives (167), incorrectly classifying many non-erosion cases as positive. For the `d_B100` dataset, the model performed better in avoiding false positives, which is reflected in the higher number of true negatives. However, the increase in false negatives (75) suggests that the model struggled to correctly identify some erosion cases.

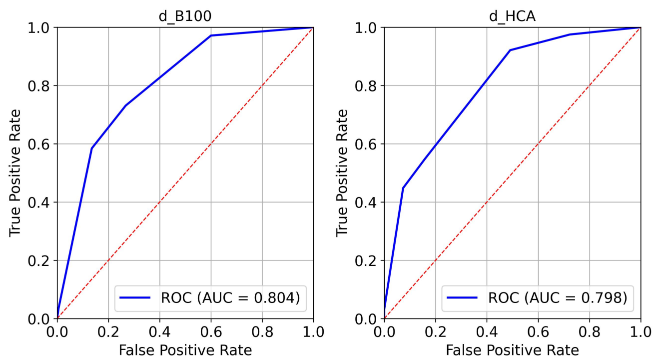


FIGURE 7 ROC–AUC Curves

This difference between the confusion matrices of the two datasets is crucial, as it implies that the `d_HCA` model may be more suitable in situations where the precise identification of positive cases is critical, even if this comes at the cost of more false positives.

As shown in Table **??**, for the `d_HCA` dataset, the model achieved a precision of 0.89 for the "Non-erosion" class and 0.61 for "Erosion," while for `d_B100`, the precision values were 0.77 and 0.69, respectively. This slight decrease in precision for the erosion class in `d_B100` suggests greater difficulty in correctly classifying positive cases within this dataset.

Recall is particularly relevant in this study, as it prioritizes the identification of positive cases. The `d_HCA` model achieved a recall of 0.92 for the "Erosion" class, demonstrating its effectiveness in identifying most true positive cases. In contrast, the `d_B100` model reached a recall of 0.73. This difference indicates that, although both models were optimized for recall, the `d_HCA` model performed clearly better in detecting positive cases.

The *F1-score*, which combines precision and recall, yielded values of 0.73 for `d_HCA` and 0.71 for `d_B100` in the "Erosion" class. This indicates that, although the `d_B100` model achieved higher precision for the "Non-erosion" class, its ability to balance precision and recall for the positive class was not as efficient as that of the `d_HCA` model. The overall accuracy values were 0.69 for `d_HCA` and 0.73 for `d_B100`, reflecting the models' ability to correctly classify both positive and negative cases.

The AUC–ROC metric, which provides a broader view of model performance, showed values of 0.79763 for `d_HCA` and 0.80446 for `d_B100`. Although the AUC–ROC for `d_B100` was slightly higher, this does not necessarily translate into better performance in identifying positive cases, as evidenced by the recall and F1-score results, as well as by the confusion matrix. These results also demonstrate that both models were able to generalize adequately to new data, confirming the validity of the hyperparameter tuning and cross-validation process.

Although both models were trained with the same objective of identifying positive cases, they exhibited significant differences in performance metrics. The `d_HCA` model demonstrated superior capability in identifying positive cases, reflected in its higher recall and better F1-score, while the `d_B100` model achieved higher overall accuracy.

This analysis highlights the importance of hyperparameter selection and the need to balance precision and recall for practical classification applications.

5 | DISCUSSION

This section focuses on model interpretability through the visualization of the resulting decision trees and the assessment of variable importance. Subsection 5.1 discusses the model developed for the `d_B100` dataset, while Subsection 5.2 addresses the model based on the `d_HCA` dataset. Beyond analyzing the structure of the trees and the relative importance of each variable, this section aims to relate the observed patterns to the studied phenomenon, emphasizing an interdisciplinary interpretation supported by domain expertise.

5.1 | Dataset d_B100

The decision tree generated for the `d_B100` dataset follows the same hierarchical binary structure of CART models. Figure 8 illustrates the resulting tree, where each internal node contains a splitting condition and its associated Gini impurity, while the terminal nodes represent the model predictions.

This tree contains nine internal nodes and six leaves. The root node (#0) splits the data based on the `FA_NV` variable with a threshold value of 0.014, dividing instances into those with higher or lower proportions of agricultural-to-non-vegetated transition. Subsequent splits rely on variables such as `FA_FA` and `NV_NV`, until the instances are classified into erosion or non-erosion.

The analysis of variable importance (Figure 9) highlights `FA_NV` as the most influential predictor, with a relative importance of 1.0, accounting for most of the variability explained by the model.

Other variables, such as `FA_FA` and `NV_NV`, also played relevant roles, with relative importance values of 0.512 and 0.073, respectively. These features contributed to the formation of subsequent splits, suggesting that the model depends on them to refine predictions. Interestingly, `FO_FO`, which was the most relevant variable in the `d_HCA` model, had zero importance in this case.

As observed in the following model, most variables exhibited zero importance, implying that they were not informative for tree construction. The presence of many zero-importance features suggests that dimensionality reduction might improve the model by removing irrelevant predictors.

5.2 | Dataset d_HCA

The tree generated by the CART technique for the `d_HCA` dataset also presents a hierarchical structure of binary decisions, where each node represents a question about an explanatory variable. The tree shown in Figure 10 has ten nodes, including the root node, which splits the data according to the `FO_FO` variable (forest-to-forest transition) with a threshold value of 0.048. This initial split separates instances with `FO_FO` values higher or lower than this threshold.

Subsequent nodes further divide the data based on other variables such as `FA_NV` (agricultural-to-non-vegetated transition), `FA_FA` (agricultural-to-agricultural), and `NV_NV` (non-vegetated-to-non-vegetated), until all instances are classified into one of the two final categories: erosion or non-erosion. Some terminal nodes present low Gini impurity values (e.g., node #6), indicating purer splits and higher classification confidence.

The variable importance analysis for this CART model (Figure 11) shows that `FO_FO` is the most influential variable, with a relative importance of 1.0, accounting for most of the variability explained by the model. The variable `FA_NV` follows with a relative importance of 0.626, making it the second most relevant predictor for erosion classification.

Other variables, such as `FA_FA` and `NV_NV`, showed lower relative importance values (0.163 and 0.089, respectively). Although these features contribute to some splits, their influence is minor compared to `FO_FO` and `FA_NV`. From a geomorphological standpoint, forested areas (`FO_FO`) tend to inhibit erosion processes, while transitions from agricultural to non-vegetated areas (`FA_NV`) are more prone to erosion development.

The remaining variables exhibit zero importance, indicating that they did not significantly influence the model's decision process. Given that all land-use types have some conceptual relationship with erosion, this behavior likely reflects insufficient variability in those features to meaningfully contribute to classification.

5.3 | Comparative analysis

The training–testing and validation split strategy was essential to ensure model robustness. During training, a five-fold cross-validation procedure was used, providing a more reliable tuning of hyperparameters. Consequently, the CART decision tree was optimized using *RandomizedSearchCV*, with the following parameters:

- `max_depth` = 7;
- `min_impurity_decrease` ≈ 0.005;
- `min_samples_leaf` = 13;
- `min_samples_split` = 32.

These hyperparameters were consistent across both datasets, suggesting that the underlying land-use transition patterns share similar characteristics, even though the magnitude of their effects differs between datasets. The final evaluation was conducted using an external validation set, unseen during training, ensuring an unbiased estimate of model performance on new data.

The choice of *recall* as the primary evaluation metric was motivated by the aim to correctly identify erosion-prone areas, prioritizing the detection of positive cases. Under this criterion, the `d_HCA` model achieved a recall of 0.885, while the `d_B100` model reached 0.706, indicating

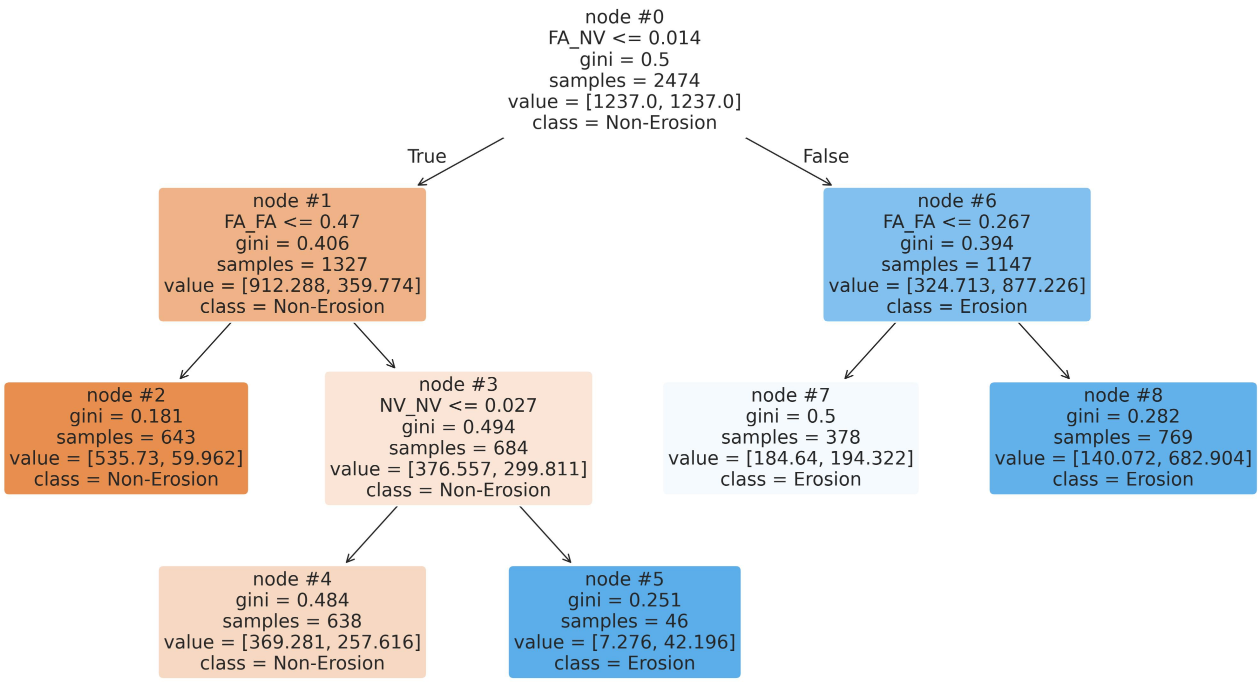


FIGURE 8 Decision tree – d_B100

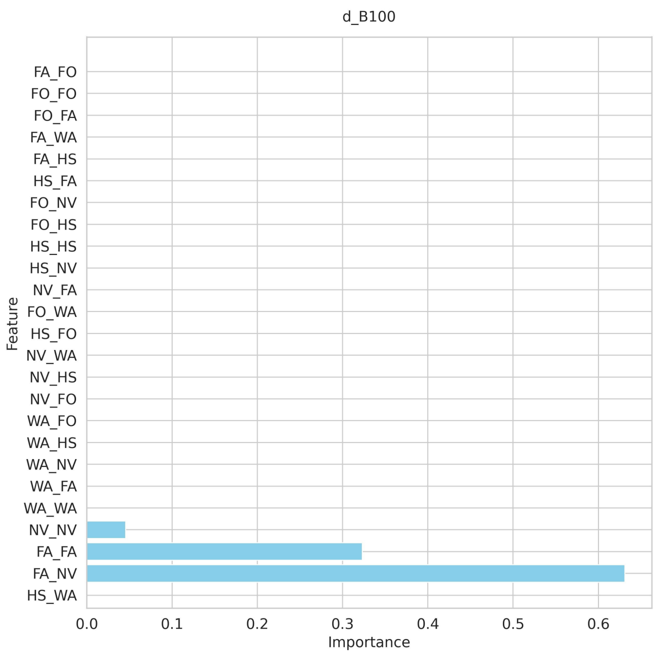


FIGURE 9 Variable importance – d_B100

superior performance of the hydrological contributing area–based approach. This result suggests that using a hydrologically defined spatial unit is more effective for predicting erosion susceptibility.

Although the `d_B100` model achieved slightly higher precision and AUC-ROC values, its F1-score was lower. The weighted precision averages were comparable across models (0.76 for `d_HCA` and 0.73 for `d_B100`), reflecting balanced overall performance.

From an interpretative standpoint, the most influential variables for erosion classification in both models were `FA_NV` and `FA_FA`. These land-use transitions—agricultural-to-non-vegetated and agricultural-to-agricultural—are strongly associated with erosion-prone conditions, highlighting their significance in the predictive process. The `FA_NV` variable, in particular, had the highest relative importance in both datasets, reinforcing its key role in erosion susceptibility.

The `FO_FO` variable (forest-to-forest) exhibited similar descriptive statistics across datasets, with mean and standard deviation values of 0.14 and 0.32 for `d_HCA`, and 0.13 and 0.30 for `d_B100`. Despite this similarity, its importance varied considerably: it was the most relevant predictor in `d_HCA`, but non-influential in `d_B100`. This is noteworthy since the violin plots revealed similar distributions for this variable in both datasets—values concentrated near zero, as expected for urban census sectors where forest persistence is minimal.

While the higher recall of the `d_HCA` model suggests stronger predictive capacity, the `d_B100` decision tree still performed reasonably well, with recall and F1-score values around 0.70. This indicates that the model successfully captured relevant spatial patterns, even under less favorable conditions. Both models demonstrated good discriminative ability, with AUC-ROC values of 0.79763 and 0.80446 for `d_HCA` and

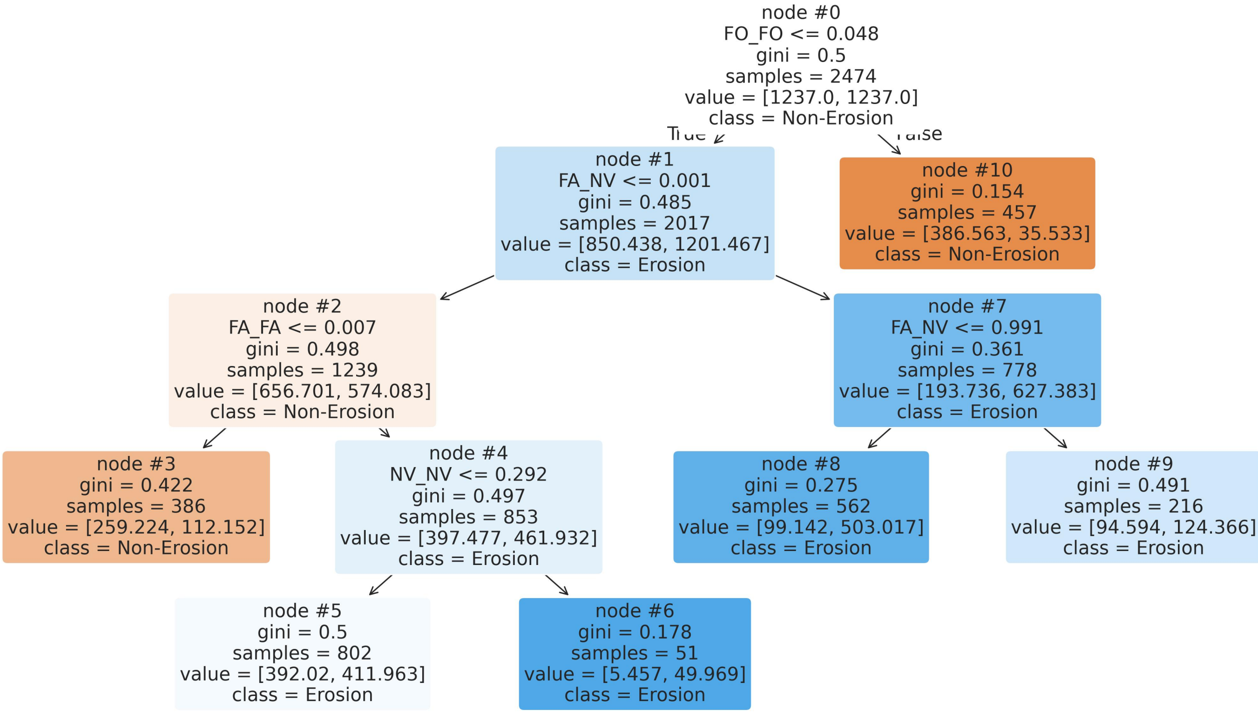


FIGURE 10 Decision tree – d_HCA

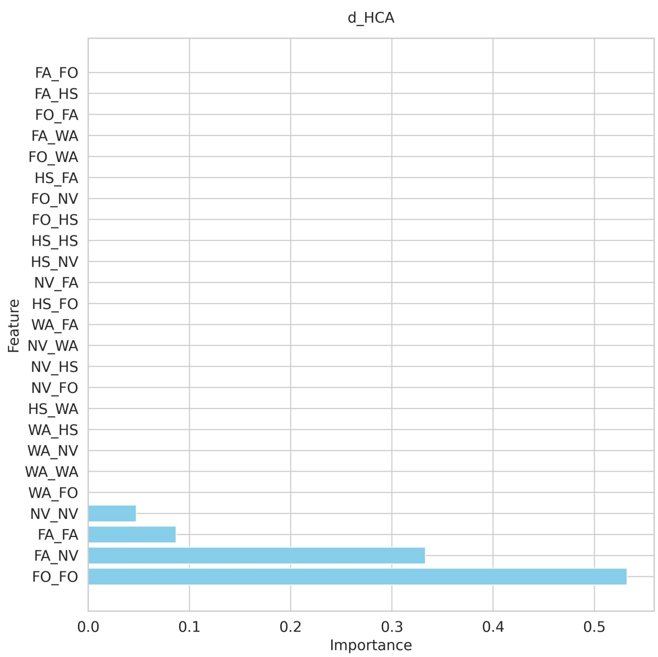


FIGURE 11 Variable importance – d_HCA

`d_B100`, respectively, confirming their capability to distinguish erosion from non-erosion cases.

6 | CONCLUSIONS

This study aimed to analyze the relationship between land-use transitions and erosion occurrence in urban areas. The proposed approach allowed for the evaluation of which data structure provides better performance for erosion prediction while preserving model interpretability. This strategy proved essential for validating the innovative methodology proposed in this work, which introduces the use of the hydrological contributing area as an alternative to the traditional buffer-based approach.

The construction of the datasets required extensive work in a GIS environment. The frequency-based encoding — where each variable represents the relative proportion of a specific land-use transition within a given area — enabled direct comparison between different regions and datasets.

Overall, the decision tree approach proved to be an effective and interpretable tool for analyzing land-use transition data in relation to erosion processes. Land-use transition variables play a central role in determining erosion patterns, and their interpretation enables the identification of key factors driving these processes. The strong recall performance also indicates that the models are capable of correctly detecting erosion-prone areas, a crucial aspect for urban management and planning.

The comparison between the two dataset structures not only enabled the identification of the most suitable configuration for model integration but also provided a solid foundation for future research.

Although the hydrological contributing area (HCA) structure showed superior capability in identifying positive erosion cases, the 100-meter buffer approach exhibited a more balanced predictive performance across classes.

6.1 | Future Work

Building upon the results obtained in this study, several research directions can be pursued to enhance performance while maintaining model interpretability. Given that land-use transition data are often sparse and high-dimensional, refining data preprocessing strategies represents an important direction for future work. Techniques such as feature selection, data augmentation, or dimensionality reduction (e.g., Principal Component Analysis) could be applied to improve model generalization and reduce noise without compromising interpretability.

Another potential avenue involves exploring alternative modeling techniques. While the decision tree method proved effective and interpretable, testing additional algorithms—such as Random Forests, Gradient Boosting, or Explainable Neural Networks—could improve performance and robustness, particularly in handling complex, non-linear relationships between land-use transitions and erosion occurrence.

Finally, the use of hybrid modeling approaches combining supervised learning with rule-based or heuristic systems could offer a balance between predictive power and transparency. For instance, ensemble or stacking frameworks could leverage the strengths of multiple algorithms while maintaining interpretability through an auxiliary decision tree layer.

Overall, these perspectives aim to advance both the methodological framework and the scientific understanding of how land-use transitions influence erosion processes in urban contexts. Continued refinement of these models will be crucial to ensure that predictive outcomes remain both accurate and interpretable, supporting decision-making processes in risk management and territorial planning.

AUTHOR CONTRIBUTIONS

Author 1 contributed to conceptualization, data collection, data curation, formal analysis, investigation, methodology design and implementation, validation, visualization, analysis and interpretation of results, and writing and revision of the original draft. Author 2 was responsible for conceptualization, project supervision, methodological design, analysis and interpretation of results, and critical review and final editing of the manuscript.

ACKNOWLEDGMENTS

The first author was supported by a doctoral scholarship from the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001.

FINANCIAL DISCLOSURE

None reported.

CONFLICT OF INTEREST

The authors declare no potential conflict of interests.

References

Alloghani, M., Al-Jumeily, D., Mustafina, J., Hussain, A., & Aljaaf, A. (2019). A systematic review on supervised and unsupervised machine learning algorithms for data science. *Unsupervised and Semi-Supervised Learning*. doi: 10.1007/978-3-030-22475-2$_1$

Amato, F., Guignard, F., Robert, S., & Kanevski, M. (2020). A novel framework for spatio-temporal prediction of environmental data using deep learning. *Scientific Reports*, *10*(1), 22243. doi: 10.1038/s41598-020-79148-7

Amiri, M., Pourghasemi, H. R., Ghanbarian, G. A., & Afzali, S. F. (2019). Assessment of the importance of gully erosion effective factors using Boruta algorithm and its spatial modeling and mapping using three machine learning algorithms. *Geoderma*, *340*, 55–69. doi: 10.1016/j.geoderma.2018.12.042

Arantes, L. T., Carvalho, A. C. P., Carvalho, A. P. P., Lorandi, R., Moschini, L. E., & Di Lollo, J. A. (2021, apr). Surface runoff associated with climate change and land use and land cover in southeast region of Brazil. *Environmental Challenges*, *3*, 100054. doi: 10.1016/j.envc.2021.100054

Begin, Z. B., & Schumm, S. A. (1979). Instability of alluvial valley floors: a method for its assessment. *Transactions of the ASAE*, *22*(2), 347–350.

Blanco-Canqui, H., & Lal, R. (2010). *Principles of Soil Conservation and Management* (1st ed.). Dordrecht: Springer Netherlands. doi: 10.1007/978-1-4020-8709-7

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Buitinck, L., et al. (2013). Api design for machine learning software: Experiences from the scikit-learn project. *Proceedings of the 21st European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, 327–332. Discusses practical implementations and stopping criteria in decision trees, particularly through the scikit-learn API.

Carvalho, A. P. P., Guerrero, J. V. R., Silva, E. V. E., Pinto, M. J. R., Vaz, R. M. G. F. C., Pereira, C. T., ... Moschini, L. E. (2019). Fragilidade ambiental à erosão em uma bacia hidrográfica antropizada no nordeste do estado de são paulo, brasil. *Anuário do Instituto de Geociências*, *42*(3), 07–18. doi: 10.11137/2019$_{30}$7$_1$8

CATI. (2016). *Modelo Digital de Elevação do Estado de São Paulo*. http://mapas.cati.sp.gov.br/mdecati.html. São Paulo: Centro de Informações Agrícolas de São Paulo. Retrieved from `http://mapas.cati.sp.gov.br/mdecati.html`

Chaplot, V., Giboire, G., Marchand, P., & Valentin, C. (2005, oct). Dynamic modelling for linear erosion initiation and development under climate and land-use changes in northern Laos. *CATENA*, *63*(2-3), 318–328. doi: 10.1016/J.CATENA.2005.06.008

Christoph, M. (2020). *Interpretable machine learning: A guide for making black box models explainable*. Leanpub.

Conoscenti, C., Angileri, S., Cappadonia, C., Rotigliano, E., Agnesi, V., & Märker, M. (2014). Gully erosion susceptibility assessment by means of gis-based logistic regression: A case of sicily (italy). *Geomorphology*, *204*, 399–411. doi: 10.1016/j.geomorph.2013.08.021

Costa, C. W., Lorandi, R., Lollo, J. A., Serikawa, V. S., Ferreira, T. S., Stanganini, F. N., & Neto, P. S. G. (2018, jul). Análise multicritério aplicada à seleção de áreas para implantação de aterros sanitários na Bacia do Ribeirão do Meio (Leme, SP), em escala 1:50.00. *Sociedade & Natureza*, *30*(1 SE - Artigos). doi: https://doi.org/10.14393/SN-v30n1-2018-9

Cunha, S. B., & Guerra, A. J. T. (2009). Degradação ambiental. In A. J. T. Guerra & S. B. Cunha (Eds.), *Geomorfologia e meio ambiente* (7th ed., pp. 337–379). Rio de Janeiro: Bertrand Brasil.

Dewitte, O., Daoudi, M., Bosco, C., & Eeckhaut, M. V. D. (2015). Predicting the susceptibility to gully initiation in data-poor regions. *Geomorphology*, *228*, 101–115. doi: 10.1016/j.geomorph.2014.08.010

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Durães, M. F., & de Mello, C. R. (2016). Distribuição espacial da erosão potencial e atual do solo na Bacia Hidrográfica do Rio Sapucaí, MG. *Engenharia Sanitaria e Ambiental*, *21*(4), 677–685. doi: https://doi.org/10.1590/S1413-41522016121182

Endres, P. F., Pissarra, T. C. T., Borges, M. J., & Politano, W. (2006, apr). Quantificação das classes de erosão por tipo de uso do solo no município de Franca - SP. *Engenharia Agrícola*, *26*(1), 200–207. doi: 10.1590/S0100-69162006000100022

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874.

Galharte, C. A., Villela, J. M., & Crestana, S. (2014, feb). Sediment yield estimation due to changes in land use and cover. *Revista Brasileira de Engenharia Agrícola e Ambiental*, *18*(2), 199–201. doi: 10.1590/S1415-43662014000200010

Google. (2024). *Google Colaboratory*. `https://colab.research.google.com`. Retrieved from `https://colab.google/`

Gorzałczany, M. B., & Rudziński, F. (2017, apr). Interpretable and accurate medical data classification - a multi-objective genetic-fuzzy optimization approach. *Expert Syst. Appl.*, *71*(C), 26–39. doi: 10.1016/j.eswa.2016.11.017

Guo, Y., Peng, C., Zhu, Q., Wang, M., Wang, H., Peng, S., & He, H. (2019, nov). Modelling the impacts of climate and land use changes on soil water erosion: Model applications, limitations and future challenges. *Journal of Environmental Management*, *250*, 109403. doi: 10.1016/J.JENVMAN.2019.109403

Han, J., Kamber, M., & Pei, J. (2011). *Data mining: Concepts and techniques* (3rd ed.). Boston: Elsevier.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). New York: Springer. Discusses the sensitivity of decision trees to variations in input data (Chapter 9).

Hunter, J. D., & the Matplotlib development team. (2007). *Matplotlib: Visualization with python.* Retrieved from `https://matplotlib.org/`

IBGE. (2010). *Censo Demográfico Brasileiro*. Rio de Janeiro: Instituto Brasileiro de Geografia e Estatística. Retrieved from `https://censo2010.ibge.gov.br/`

IBGE. (2022). *Brasil - São Paulo: Panorama.* Retrieved 2024-07-22, from `https://cidades.ibge.gov.br/brasil/sp/panorama`

IPT. (2012). *Cadastro de Pontos de Erosão e Inundação no Estado de São Paulo* (Vol. 2; 131.057-205). São Paulo: Instituto de Pesquisas Tecnológicas.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: With applications in r* (1st ed.). Springer. ISBN: 978-1-4614-7138-7. doi: https://doi.org/10.1007/978-1-4614-7138-7

Javidan, Kavian, Pourghasemi, Conoscenti, & Jafarian. (2019, nov). Gully Erosion Susceptibility Mapping Using Multivariate Adaptive Regression Splines—Replications and Sample Size Scenarios. *Water*, *11*(11), 2319. doi: https://doi.org/10.3390/w11112319

Kohavi, R. (1995). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In *International joint conference on artificial intelligence* (Vol. 14, pp. 1137–1145).

Leo Breiman Jerome Friedman, C. J. S. R. A. O. (1984). *Classification and Regression Trees*. Chapman and Hall/CRC.

Liu, X., Lu, D., Zhang, A., Liu, Q., & Jiang, G. (2022, feb). Data-Driven Machine Learning in Environmental Pollution: Gains and Problems. *Environmental Science & Technology*, *56*(4), 2124–2133. doi: 10.1021/acs.est.1c06157

Liu, Y., Meng, Z., Zhu, L., Hu, D., & He, H. (2023). Optimizing the sample selection of machine learning models for landslide susceptibility prediction using information value models in the dabie mountain area of anhui, china. *Sustainability*, *15*(3). doi: 10.3390/su15031971

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, *18*(1), 50–60. doi: 10.1214/aoms/1177730491

MapBiomas. (2023). *Mapbiomas - método de mapeamento de cobertura e uso do solo.* Retrieved from `https://brasil.mapbiomas.org/metodo_cobertura_e_uso/` Acesso em: 27 dez. 2024.

MapBiomas. (2024). *Mapbiomas - coleção 9 da série anual de mapas de cobertura e uso do solo do brasil.* Retrieved from `https://mapbiomas.org` Acesso em: 27 dez. 2024.

Margot, V., & Luta, G. (2021). A new method to compare the interpretability of rule-based algorithms. *AI*, *2*(4), 621–635. doi: 10.3390/ai2040037

McKnight, P. E., & Najab, J. (2010). Mann-whitney u test. In I. B. Weiner & W. E. Craighead (Eds.), *The corsini encyclopedia of psychology* (pp. 1–1). Wiley. doi: 10.1002/9780470479216.corpsy0524

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*, 1-38. doi: https://doi.org/10.1016/j.artint.2018.07.007

Moore, I. D., & Burch, G. J. (1986). Physical basis of the length-slope factor in the universal soil loss equation. *Soil Science Society of America Journal*, *50*(5), 1294–1298.

Moore, I. D., Burch, G. J., & Mackenzie, D. H. (1988). Topographic effects on the distribution of surface soil water and the location of ephemeral gullies. *Transactions of the ASAE*, *31*(4), 1098–1107.

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Interpretable machine learning: definitions, methods, and applications. *arXiv preprint arXiv:1901.04592*.

Neves, M. (2022). *Análise das mudanças de uso do solo e seus impactos na erosão linear em Áreas urbanas* (Doctoral dissertation, Universidade Federal de São Carlos, São Carlos, SP, Brasil). Retrieved from `https://repositorio.ufscar.br/bitstream/handle/ufscar/16519/Tese_Monique_Neves.pdf?sequence=6&isAllowed=y`

Olivatto, T. F., & Di Lollo, J. A. (2025). Sampling strategies for machine learning-based linear erosion studies: a review approaching contributing area. *Natural Hazards Research*. doi: https://doi.org/10.1016/j.nhres.2025.08.002

Patton, P. C., & Schumm, S. A. (1975). Gully erosion, Northwestern Colorado: a threshold phenomenon. *Geology*, *3*(2), 88–90.

Pearson, K. (1895). Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, *58*, 240–242.

Pedregosa, F., et al. (2019). *Scikit-learn: Machine learning in python*. Online documentation. Available at: `https://scikit-learn.org/stable/`.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, E. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Phalak, M., Bhandari, M., & Sharma, R. (2014). Analysis of decision tree-a survey. *International journal of engineering research and technology*, *3*.

Piltaver, R., Luštrek, M., Gams, M., & Martinčić-Ipšić, S. (2016). What makes classification trees comprehensible? *Expert Systems with Applications*, *62*, 333-346. doi: https://doi.org/10.1016/j.eswa.2016.06.009

Poesen, J., Nachtergaele, J., Verstraeten, G., & Valentin, C. (2003). Gully erosion and environmental change: Importance and research needs [Conference paper]. *Catena*, *50*(2-4), 91 – 133. doi: 10.1016/S0341-8162(02)00143-1

Pons, N. A. D., Pejon, O. J., & Zuquette, L. V. (2007, nov). Use of geoprocessing in the study of land degradation in urban environments: the case of the city of São Carlos, state of São Paulo, Brazil. *Environmental Geology*, *53*(4), 727–739. doi: https://doi.org/10.1007/s00254-007-0685-y

Pourghasemi, H. R., Yousefi, S., Kornejady, A., & Cerdà, A. (2017). Performance assessment of individual and ensemble data-mining techniques for gully erosion modeling [Article]. *Science of the Total Environment*, *609*, 764 – 775. doi: 10.1016/j.scitotenv.2017.07.198

Prieto-Amparán, Pinedo-Alvarez, Vázquez-Quintero, Valles-Aragón, Rascón-Ramos, Martinez-Salvador, & Villarreal-Guerrero. (2019, sep). A Multivariate Geomorphometric Approach to Prioritize Erosion-Prone Watersheds. *Sustainability*, *11*(18), 5140. doi: https://doi.org/10.3390/su11185140

Python Software Foundation. (2020). Python 3.9 Documentação [Computer software manual]. Elk Grove Village. Retrieved from `https://docs.python.org/3.9/`

Sakai, T., & Goto, M. (2014). Categorical feature encoding for classification: A survey. *Journal of Machine Learning*, *1*, 10-25.

Salomão, F. d. T., et al. (1999). Controle e prevenção dos processos erosivos. *Guerra AJT, Silva AS, Botelho RGM, organizadores. Erosão e conservação dos solos: conceitos, temas e aplicações. Rio de Janeiro: Bertrand Brasil*, 229–67.

Sandri, M., & Zuccolotto, P. (2008). A bias correction algorithm for the gini variable importance measure in classification trees. *Journal of Computational and Graphical Statistics*, *17*, 611 - 628. doi: 10.1198/106186008X344522

Schmidt, L., Heße, F., Attinger, S., & Kumar, R. (2020, may). Challenges in Applying Machine Learning Models for Hydrological Inference: A Case Study for Flooding Events Across Germany. *Water Resources Research*, *56*(5). doi: 10.1029/2019WR025924

SEADE. (2020). *População urbana e rural do estado de são paulo*. Retrieved from `https://repositorio.seade.gov.br/dataset/populacao-urbana-e-rural` Acessado em: 22 jul. 2024.

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology*, *15*(1), 72–101.

Thorne, C. R., Zevenbergen, L. W., Grissinger, E. H., & Murphey, J. B. (1986). Ephemeral gullies as sources of sediment. In *Proceedings of the fourth federal interagency sedimentation conference march 24-27, 1986, las vegas, nevada.* (Vol. 1).

Waskom, M., & the Seaborn development team. (2021). Seaborn: statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021. Retrieved from `https://seaborn.pydata.org/` doi: 10.21105/joss.03021

Zhou, Q., Liao, F., Mou, C., & Wang, P. (2018). Measuring interpretability for different types of machine learning models. In *Trends and applications in knowledge discovery and data mining: Pakdd 2018 workshops, bdasc, bdm, ml4cyber, paisi, damemo, melbourne, vic, australia, june 3, 2018, revised selected papers 22* (pp. 295–308).

Zhou, Z.-H. (2021). *Machine learning*. Springer nature.

SUPPORTING INFORMATION

All data, codes, and products generated in this study were organized in a public repository on GitHub, ensuring transparency and reproducibility of the methodological steps. The repository contains the scripts used in the analysis, the processed input files, and the resulting maps from the modeling and validation stages.

The material is available at:
`https://github.com/tatianeolivatto/TCC_USP_TFO`

How to cite this article: Olivatto T. F., & Di Lollo J. A.. Unveiling Land Use Drivers of Urban Gully Erosion: An Interpretable Machine Learning Approach. *J Comput Phys.* 2026; 00(00):1–18.

APPENDIX

A CORRELATIONS MATRICES

Pearson and Spearman correlation matrices of explanatory variables for `d_HCA` and `d_B100`:

Figure A1 - Pearson Correlation Matrix – `d_HCA`;
Figure A2 - Spearman Correlation Matrix – `d_HCA`;
Figure A3 - Pearson Correlation Matrix – `d_B100`;
Figure A4 - Spearman Correlation Matrix – `d_B100`.

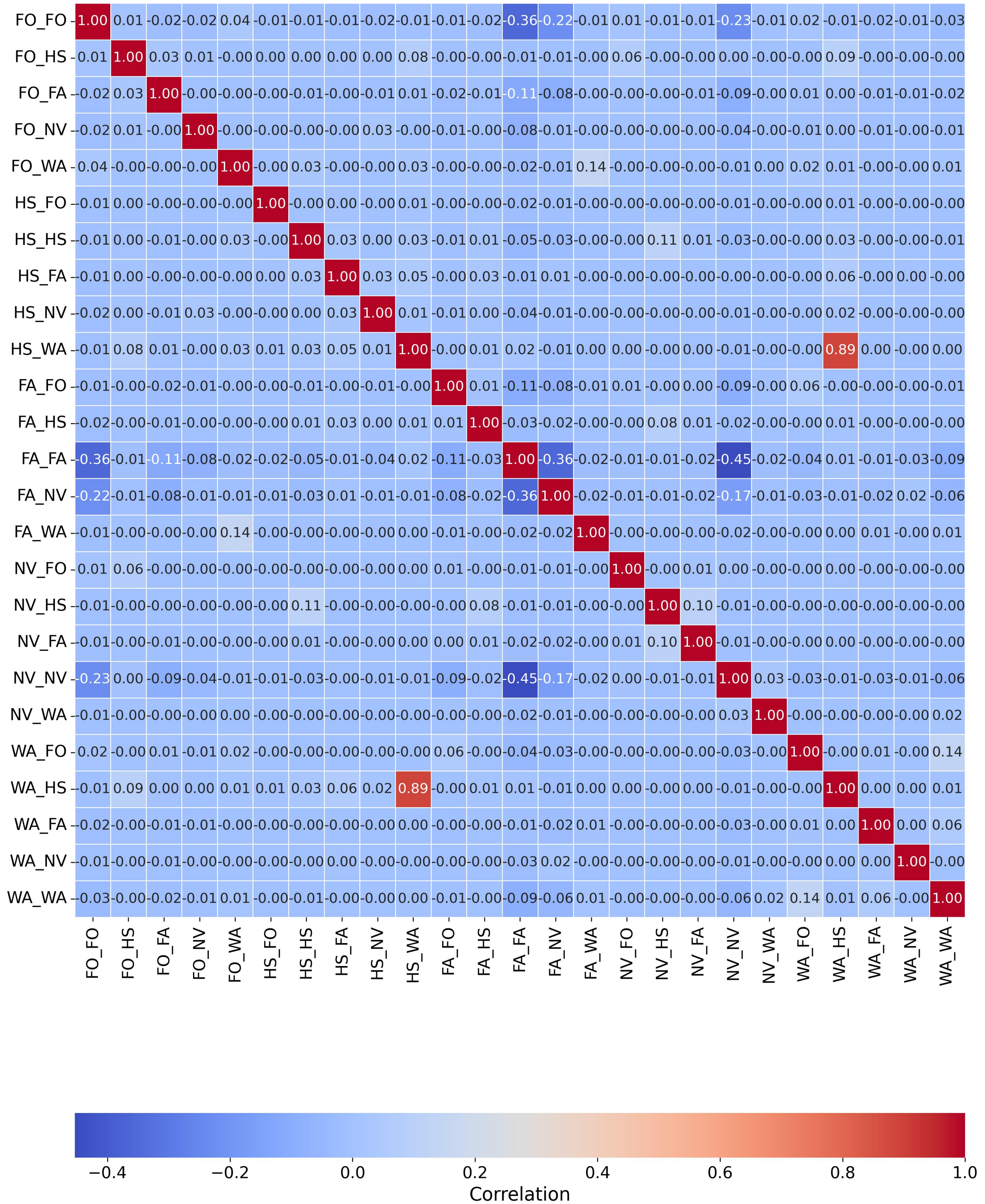


FIGURE A1 Pearson Correlation Matrix – Explanatory Variables (d_HCA).

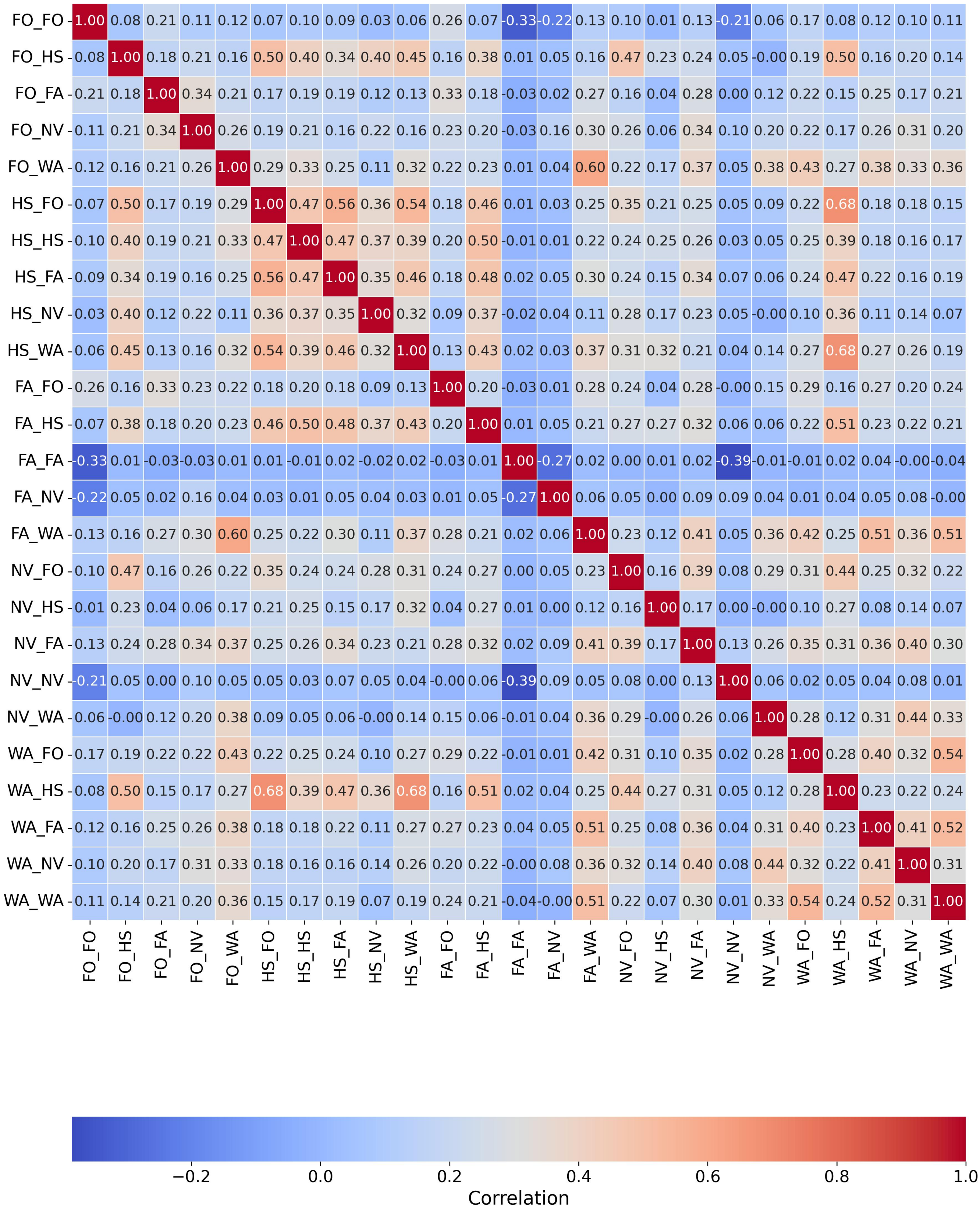


FIGURE A2 Spearman Correlation Matrix – Explanatory Variables (d_HCA).

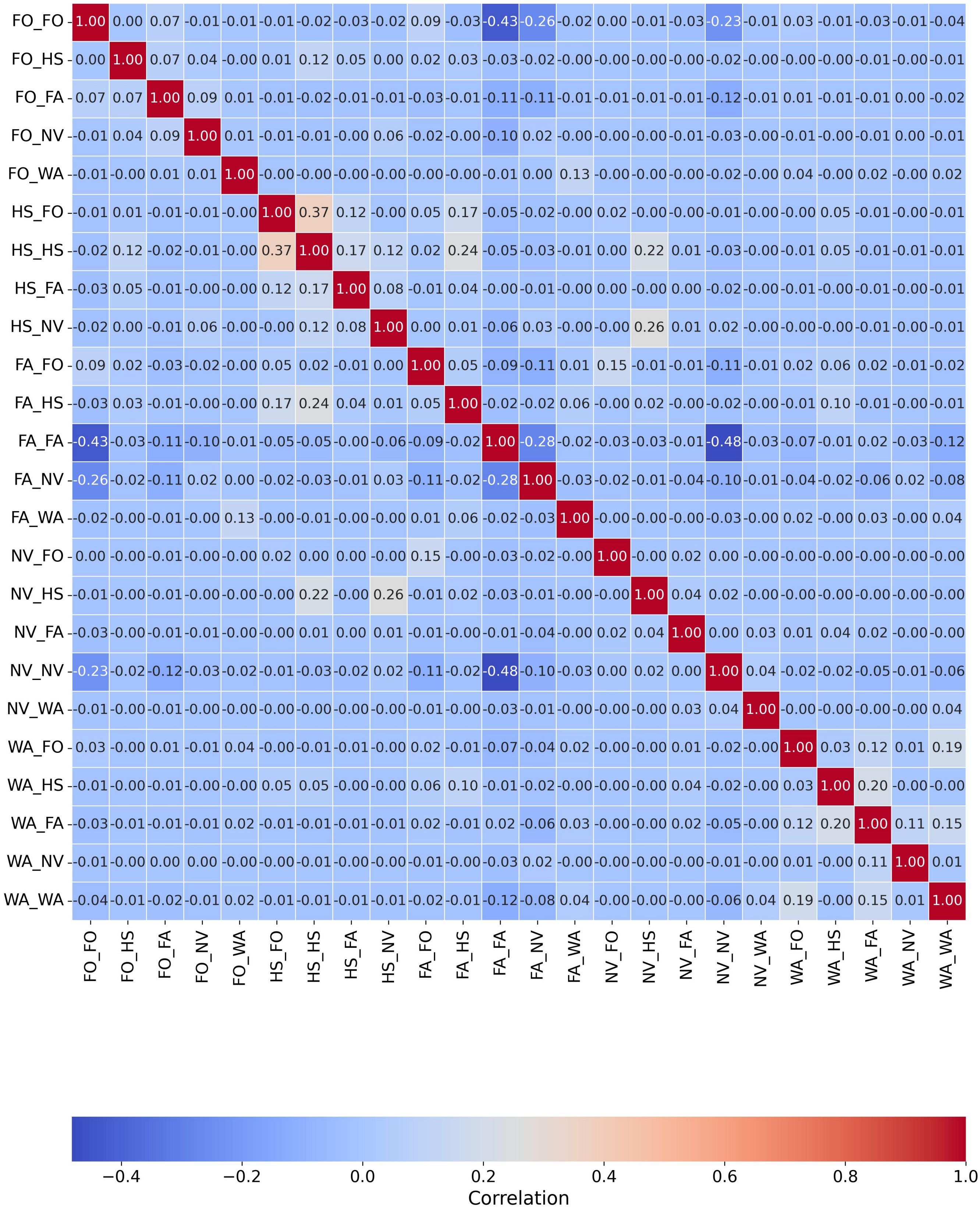


FIGURE A3 Pearson Correlation Matrix – Explanatory Variables (d_B100).

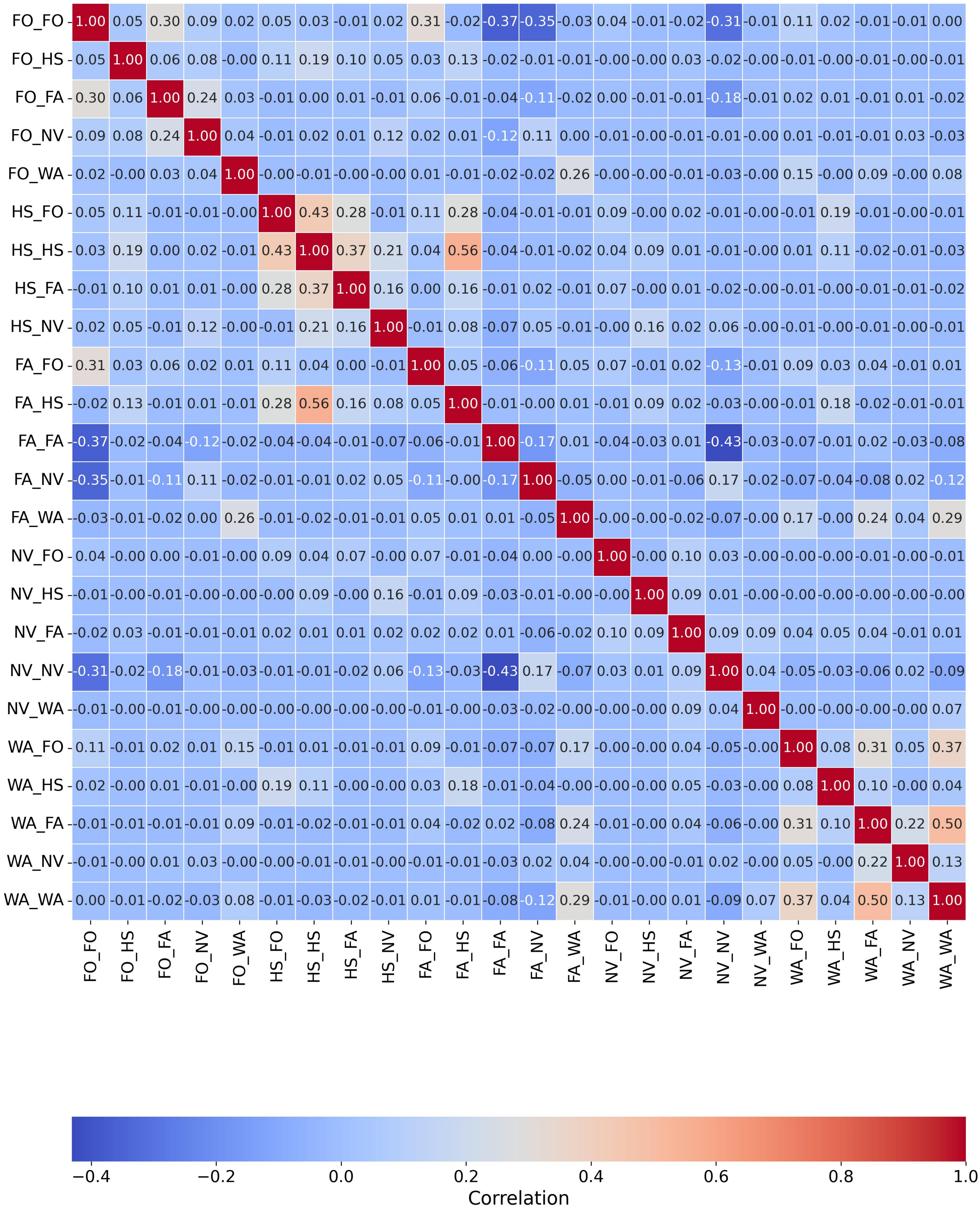


FIGURE A4 Spearman Correlation Matrix – Explanatory Variables (d_B100).