Padrões Territoriais do Risco de Inundação no Brasil por Meio de Aprendizado de Máquina

Tereza Margarida Xavier de Melo Lopes [1]; Yasmin Teixeira Bernardo Castelo Branco [2] & Samiria Maria Oliveira da Silva [3]

RESUMO – As inundações urbanas representam um dos principais desafios para a gestão territorial no Brasil, sendo influenciadas por fatores relacionados à urbanização, infraestrutura e ocupação do solo. Este estudo teve como objetivo mapear e categorizar padrões municipais associados ao risco de inundação por meio da clusterização K-means. Foram utilizados três indicadores: parcela de domicílios em situação de risco de inundação (IN040), parcela da área urbana em relação à área total do município (IN042) e taxa de cobertura de pavimentação e meio-fio na área urbana (IN020). Os resultados identificaram diferentes combinações entre risco informado de inundação, urbanização e infraestrutura urbana, evidenciando que a parcela de domicílios em situação de risco não está associada exclusivamente ao grau de urbanização ou à cobertura de pavimentação e meio-fio. Observou-se maior concentração de municípios com elevados valores de IN040 na Região Norte e em áreas específicas do Nordeste e Sudeste. Já os agrupamentos com menores valores de IN040 apresentaram distribuição expressiva em diferentes regiões, sem indicar ausência de risco hidrológico. A clusterização mostrou-se útil para apoiar a análise territorial e orientar estratégias diferenciadas de planejamento, drenagem urbana e mitigação de inundações.

ABSTRACT– Urban flooding represents one of the main challenges for territorial management in Brazil, being influenced by factors related to urbanization, infrastructure, and land use. This study aimed to map and categorize municipal patterns associated with flood risk through K-means clustering. Three indicators were used: the proportion of households located in flood-risk areas (IN040), the proportion of urban area relative to the total municipal area (IN042), and the rate of pavement and curb coverage in urban areas (IN020). The results identified different combinations of reported flood risk, urbanization, and urban infrastructure, showing that the proportion of households in risk areas is not exclusively associated with the degree of urbanization or with

[1]) Doutoranda em Engenharia Civil (Recursos Hídricos), Universidade Federal do Ceará – Campus do Pici, lopes.melo.tereza@gmail.com.
[2]) Graduanda em Engenharia Ambiental, Universidade Federal do Ceará – Campus do Pici, yasmin.tbcb@gmail.com.
[3]) Professora do Departamento de Engenharia Hidráulica e Ambiental, Universidade Federal do Ceará – Campus do Pici, samiriamaria@gmail.com..

pavement and curb coverage. A greater concentration of municipalities with high IN040 values was observed in the North Region and in specific areas of the Northeast and Southeast. In contrast, clusters with lower IN040 values showed a significant distribution across different regions, without indicating the absence of hydrological risk. Clustering proved useful for supporting territorial analysis and guiding differentiated strategies for urban planning, urban drainage, and flood impact mitigation.

Palavras-Chave – Risco de inundação; Clusterização K-means; Análise espacial.

INTRODUÇÃO

De acordo com o Relatório Mundial das Nações Unidas sobre o Desenvolvimento dos Recursos Hídricos de 2024, a mudança climática está projetada para intensificar o ciclo global da água e aumentar ainda mais a frequência e severidade de inundações. No Brasil, o início de 2022 foi marcado por várias cidades brasileiras sendo impactadas por fortes enchentes Mendonça (2023). Esses eventos extremos evidenciam não apenas a vulnerabilidade das infraestruturas urbanas, mas também a ausência de planejamento adequado para lidar com o escoamento superficial, a ocupação desordenada do solo.

Desse modo, compreender a distribuição espacial do risco de inundação é fundamental para subsidiar políticas públicas voltadas à mitigação de impactos e à adaptação. No entanto, apesar do crescente número de estudos sobre inundações urbanas no Brasil, observa-se uma escassez de análises quantitativas que revelem padrões espaciais estruturados de risco em escala nacional, especialmente com base em indicadores técnicos de urbanização e infraestrutura. A maioria das abordagens permanece limitada a estudos de caso ou recortes regionais, o que dificulta a construção de uma agenda nacional de mitigação.

Nesse cenário, a clusterização tem emergido como uma ferramentas robusta na detecção de padrões e estruturas subjacentes em grandes conjuntos de dados. Tal método permite aprimorar a precisão das previsões, uma vez que se ajustam aos comportamentos específicos de cada cluster, garantindo análises espacialmente refinadas e mais assertivas Shi e Wang (2022); Lundell, Bean e Symanzik, (2023).

Estudos recentes reforçam esse potencial: Liu e Mostafavi (2024) utilizaram K-means para identificar fatores relacionados à exposição de imóveis ao risco de enchentes; Tate *et al.* (2025) combinaram análise de componentes principais com clusterização hierárquica para avaliar vulnerabilidades sociais. Essa diversidade metodológica evidencia a flexibilidade da clusterização diante de distintos contextos e variáveis.

Diante desse cenário, este estudo tem como objetivo mapear e categorizar os riscos de inundação por meio da aplicação da técnica de clusterização K-means. A partir da análise conjunta de indicadores de urbanização, infraestrutura urbana e exposição ao risco, busca-se identificar padrões espaciais de vulnerabilidade entre os municípios brasileiros, contribuindo para o direcionamento de estratégias de gestão e mitigação dos impactos associados às inundações.

METODOLOGIA

O processo metodológico foi composto por três etapas principais: i) coleta, organização e pré-processamento dos dados; ii) aplicação da Análise de Componentes Principais (PCA) para transformação multivariada dos indicadores; e iii) clusterização dos municípios brasileiros quanto aos padrões de risco de inundação, utilizando o algoritmo K-means.

Os dados utilizados neste estudo foram obtidos do Sistema Nacional de Informações sobre Saneamento (SNIS, 2022), abrangendo todos os municípios brasileiros e referindo-se ao ano de 2022. As variáveis selecionadas estão descritas na tabela 1:

Tabela 1 – Descrição dos dados de entrada

Código	Nome	Unidade de Medida	Equação	Informações Envolvidas
IN040	Parcela de domicílios em situação de risco de inundação	%	(RI013 / GE008) × 100	GE008: Quantidade total de domicílios urbanos existentes no município RI013: Quantidade de domicílios sujeitos a risco de inundação
IN042	Parcela da área urbana em relação à área total	%	(GE002 / GE001) × 100	GE001: Área territorial total do município (IBGE) GE002: Área urbana total, incluindo áreas urbanas isoladas
IN020	Taxa de Cobertura de Pavimentação e Meio-Fio na Área Urbana	%	(IE019 / IE017) × 100	IE017: Extensão total de vias públicas urbanas do município IE019: Extensão total de vias públicas urbanas com pavimento e meio-fio (ou semelhante)

Os dados foram tratados e analisados utilizando a linguagem R, aplicando procedimentos para remoção de valores ausentes e normalização das variáveis utilizando a técnica de Min-Max Scaling, em que os indicadores são são normalizados para uma faixa idêntica [0,1], a partir da seguinte equação (OCDE, 2008):

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (1)$$

Em que, x representa o valor original do indicador, x_{max} e x_{min} são, respectivamente, o valor mínimo e máximo observador para esse indicador, e x' é o valor normalizado.

Após a normalização dos indicadores, foi aplicada a Análise de Componentes Principais (Principal Component Analysis — PCA), considerando as variáveis IN040, IN042 e IN020. Essa etapa teve como finalidade sintetizar a estrutura multivariada do conjunto de dados e transformar as

variáveis originais em um novo conjunto de componentes ortogonais, reduzindo redundâncias e preservando os principais padrões de variação dos indicadores analisados. Os scores dos componentes principais foram posteriormente utilizados como dados de entrada para a clusterização K-means. Matematicamente, um componente principal pode ser expresso conforme a Equação (2):

$$Z_{ij} = a_{i1}x_{1j} + a_{i2}x_{2j} + a_{i3}x_{3j} + \ldots a_{im}x_{mj}, \quad (2)$$

em que Z_{ij} representa o score do componente principal (i) para a amostra (j); a_{im} corresponde ao peso ou loading da variável (m) no componente (i); x_{mj} representa o valor observado da variável (m) na amostra (j); (i) indica o número do componente principal; (j) representa a unidade amostral, neste caso, o município; e (m) corresponde ao número total de variáveis analisadas Juahir et al. (2011).

Neste estudo, os componentes principais foram obtidos a partir dos indicadores normalizados de exposição ao risco de inundação, urbanização e infraestrutura urbana. Posteriormente, os scores gerados pela PCA foram normalizados e utilizados na etapa de agrupamento, permitindo que a clusterização considerasse a estrutura conjunta dos indicadores, em vez de analisar cada variável isoladamente.

Clusterização k-means

Os métodos de clusterização agrupam n observações em k clusters, de modo que a similaridade intra-cluster seja maximizada e a similaridade entre clusters seja minimizada. O k-means é um método de clusterização por particionamento, onde para cada observação (i), é atribuída a um codificador C, de forma que C(i) = h, onde h, o indicador de cluster, é um elemento do conjunto {1, ..., k}. Os rótulos dos clusters (h) não têm significado e poderiam ser distinguidos por letras, cores ou outras simbologias Hastie *et al*. (2009).

O objetivo geral é obter um agrupamento que minimize a dissimilaridade dentro de cada cluster, ou seja, encontrar um agrupamento das observações i em clusters h que minimize essa função de perda, ou, alternativamente, maximize a similaridade.

Em um qualquer cluster h, podemos separar as distâncias de cada um de seus membros ($i \in h$) para todas as outras observações entre aqueles que pertencem ao cluster ($j \in h$) e aqueles que não pertencem ($j \notin h$):

$$T = \left(\frac{1}{2}\right)\left[\sum_{h=1}^{k}\left(\sum_{i\epsilon h}\sum_{j\epsilon h} d_{ij} + \sum_{i\epsilon h}\sum_{j\notin h} d_{ij}\right)\right] = W + \mathrm{B} \quad (2)$$

onde dij é uma medida de dissimilaridade entre os valores nas observações i e j; k é o elemento do conjunto, h é o indicador de cluster e T é a medida de similaridade. Sendo possível segregar a equação em dois termos principais: W e B, onde W (1º termo) é dissimilaridade interna, e o B (2º termo) refere-se à dissimilaridade entre clusters. Dessa forma, é possível encontrar um valor ótimo minimizando a dissimilaridade interna.

O algoritmo K-means segue o modelo de clusterização por particionamento, utilizando a distância euclidiana ao quadrado como medida de dissimilaridade. Nesse contexto, o objetivo global é encontrar a alocação C(i) de cada observação "i" a um cluster "h" a partir dos "k" clusters, de forma a minimizar a similaridade dentro do cluster sobre todos os k clusters:

$$(W) = (1/2)\sum_{h=1}^{k}\sum_{i\epsilon h}\sum_{j\epsilon h} d_{ij}\left|\left|x_i - x_j\right|\right|^2 \quad (3)$$

onde, em geral, xi e xj são vetores de p dimensões.

Em síntese, minimizar a soma das distâncias entre os pontos é equivalente a minimizar a soma dos desvios ao quadrado entre cada ponto de dados e a média (ou centro) do seu cluster.

Determinação no número ideal de clusters

O número ideal de cluster será definido com a combinação de dois métodos principais: i) o método do cotovelo; ii) o índice de silhueta. A aplicação de cada um é explicada nos tópicos subsequentes.

A determinação do número ideal de clusters foi realizada pelo método do cotovelo. Esse método analisa a soma dos erros quadráticos dentro dos clusters (Within-Cluster Sum of Squares - WSS), conforme a equação:

$$WSS = \sum_{h=1}^{k}\sum_{i\epsilon h} ||x_i - \mu_h||^2 \quad (3)$$

Onde h representa o centróide do cluster h. O gráfico do método do cotovelo é gerado ao plotar os valores de WSS para diferentes números de clusters. O ponto onde a curva apresenta uma redução menos acentuada é identificado como o número ideal de clusters "k" Kodinariya & Makwana (2013).

O índice de silhueta será utilizado como medida de validação para escolher o melhor número de clusters e avaliar a qualidade dos resultados. Em síntese, o índice de silhueta é uma medida que ajuda a avaliar o quão bem os dados foram agrupados em clusters, considerando a proximidade intra-cluster e a distância inter-cluster. Conforme Rousseeuw (1987), o índice de silhueta é definido como:

$$S_i = \frac{b_i - a_i}{max\{a_i, b_i\}} \quad (4)$$

onde "a" é a média da distância euclidiana entre uma observação e todos os outros pontos de dados no mesmo cluster, e "b" é a média da distância euclidiana entre uma observação e todos os outros pontos no próximo cluster. Quanto mais próximo de 1, melhor é a qualidade do agrupamento; quanto mais próximo de -1, pior é a qualidade do agrupamento.

RESULTADOS E DISCUSSÕES

A partir dos resultados obtidos da análise conjunta do método do cotovelo e do índice de silhueta indicou um número de cluster (k) igual a 4 (Figura 1). Além disso, k = 4, proporciona um equilíbrio entre a qualidade estatística do agrupamento e a representação espacial dos diferentes padrões de risco de inundação.

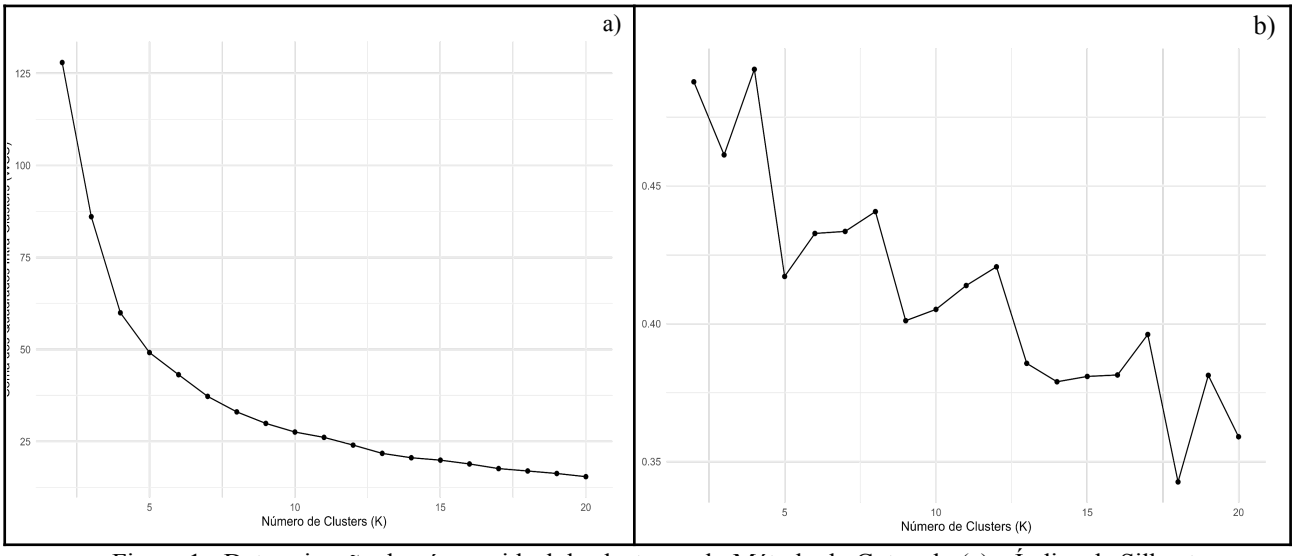


Figura 1 - Determinação do número ideal de clusters pelo Método do Cotovelo (a) e Índice de Silhueta para diferentes valores de k (b).

Após a definição do número ideal de clusters, foi realizada a análise dos agrupamentos obtidos a fim de compreender a relação entre os municípios brasileiros e os indicadores utilizados no estudo: parcela de domicílios em situação de risco de inundação (IN040), parcela da área urbana em relação à área total (IN042) e taxa de cobertura de pavimentação e meio-fio na área urbana (IN020). Os resultados podem ser observados na Figura 2.

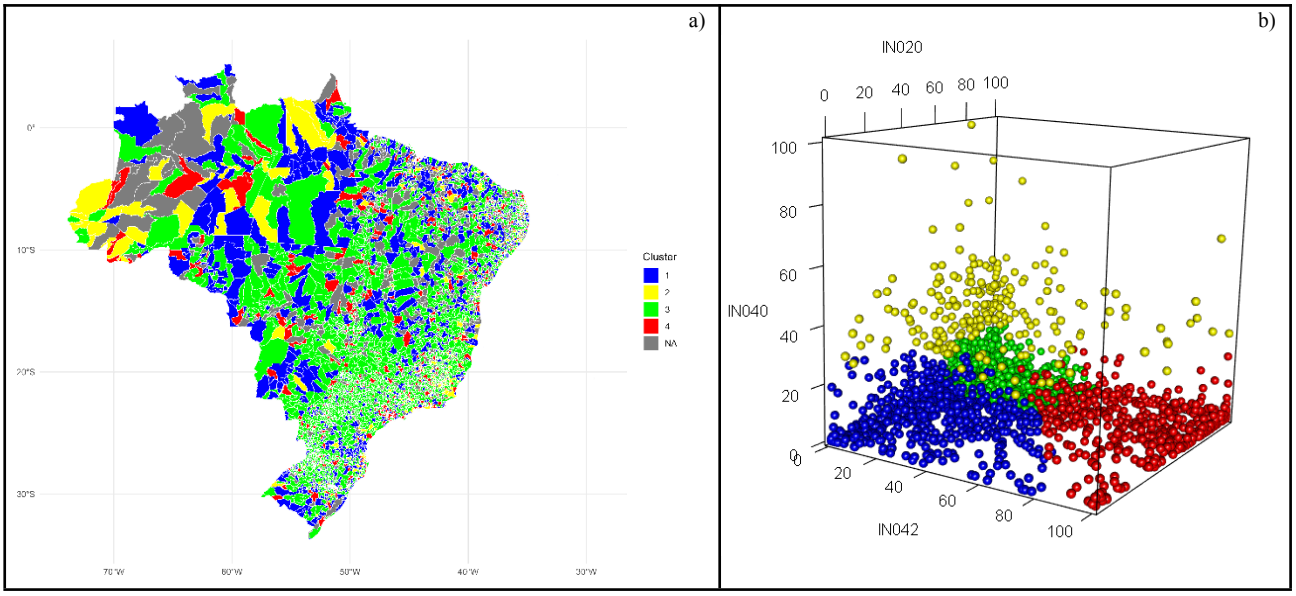


Figura 2 - Mapeamento da clusterização do risco de inundação para os Municípios Brasileiros (a) Visualização tridimensional das variáveis IN040, IN042 e IN020 por cluster (b)

Os municípios pertencentes ao Cluster 1, representado em azul, caracterizam-se por baixos valores de IN040, IN042 e IN020, indicando menor parcela de domicílios urbanos em situação de risco de inundação, menor proporção de área urbana e menor cobertura de pavimentação e meio-fio. Esse padrão sugere que o menor risco informado pode estar associado ao baixo grau de urbanização e à menor artificialização da superfície, condições que tendem a favorecer a infiltração da água e reduzir a concentração de domicílios em áreas urbanas sujeitas à inundação. No entanto, o baixo IN020 também indica menor infraestrutura urbana viária, não devendo ser interpretado como maior capacidade estrutural de mitigação. Espacialmente, esse cluster concentra-se principalmente nas regiões Norte e Nordeste, refletindo municípios menos urbanizados e com menor cobertura de infraestrutura urbana formal.

O Cluster 2, representado em amarelo, caracteriza-se por apresentar valores mais elevados de IN040, indicando maior parcela de domicílios urbanos em situação de risco de inundação. Na visualização tridimensional, esse grupo aparece mais disperso, com municípios que apresentam diferentes combinações de área urbana e cobertura de pavimentação e meio-fio. Esse resultado sugere que a ocorrência de domicílios em situação de risco não depende exclusivamente do grau de urbanização ou da pavimentação, podendo estar relacionada à forma de ocupação do território, à localização das áreas urbanas em relação aos cursos d'água e às limitações dos sistemas de drenagem. Espacialmente, observa-se presença expressiva desse cluster na Região Norte e em áreas específicas do Nordeste e Sudeste.

O Cluster 3, representado em verde, reúne municípios com baixos valores de IN040 e, em geral, maiores valores de IN020, indicando baixa parcela de domicílios em situação de risco de inundação associada a maior cobertura de pavimentação e meio-fio. Esse padrão sugere municípios com infraestrutura urbana viária mais consolidada, embora o IN020 não represente diretamente a eficiência da drenagem urbana. A ampla presença desse cluster no território nacional, especialmente nas regiões Centro-Oeste, Sul e parte do Sudeste, indica municípios nos quais a maior cobertura de infraestrutura urbana formal coincide com menor proporção de domicílios classificados em situação de risco.

O Cluster 4, representado em vermelho, é marcado por valores elevados de IN042, indicando maior proporção de área urbana em relação à área total do município, combinados a diferentes níveis de IN040 e IN020. Esse padrão demonstra que a urbanização, isoladamente, não determina a maior ocorrência de domicílios em situação de risco de inundação. Embora municípios mais urbanizados possam apresentar maior impermeabilização do solo, fatores como planejamento territorial, eficiência dos sistemas de drenagem e ocupação adequada do espaço urbano desempenham papel fundamental na redução dos impactos associados às inundações, conforme destacado por Mendonça, Farias e Buffon (2023). Assim, esse cluster representa municípios mais urbanizados, nos quais o risco de inundação depende da interação entre expansão urbana, infraestrutura e ordenamento territorial.

CONCLUSÃO

De modo geral, a aplicação da PCA associada à clusterização K-means permitiu identificar diferentes padrões municipais relacionados à parcela de domicílios em situação de risco de inundação, ao grau relativo de urbanização e à cobertura de pavimentação e meio-fio. Os resultados evidenciaram que o risco informado de inundação não apresenta relação linear com a urbanização ou com a infraestrutura urbana viária, uma vez que municípios com diferentes níveis de área urbanizada e pavimentação apresentaram condições distintas quanto ao IN040.

O Cluster 1 indicou municípios com baixo risco informado em contexto de baixa urbanização e menor impermeabilização da superfície, enquanto o Cluster 2 reuniu municípios com maior proporção de domicílios em situação de risco, mesmo sem apresentar necessariamente os maiores níveis de urbanização. Já os Clusters 3 e 4 demonstraram que a maior cobertura de infraestrutura urbana e o maior grau de urbanização podem estar associados a diferentes condições de risco, reforçando que a ocorrência de domicílios em áreas sujeitas à inundação depende da interação entre ocupação urbana, infraestrutura e características territoriais.

Assim, os resultados reforçam a importância de abordagens multivariadas para compreender o ordenamento territorial do risco de inundação nos municípios brasileiros. Trabalhos dessa natureza podem contribuir para a identificação de perfis municipais prioritários, auxiliando gestores públicos e tomadores de decisão na formulação de estratégias diferenciadas de planejamento urbano, drenagem, prevenção de desastres e mitigação dos impactos associados às inundações.

REFERÊNCIAS

ALMEIDA, R. et al. Local-Scale Disaster Risk Indicators: Applying the Disaster Risk Indicators in Brazil (Drib) Method to Understand Flood Risks in Brazil. In: MENDONÇA, F.; FARIAS, A.; BUFFON, E. (Eds.). **Urban Flooding in Brazil**. Springer Nature, 2023.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statistical Learning**. Second Edition. Springer New York, NY, v. 27, n. 2, 2009.

KODINARIYA, T. M.; MAKWANA, P. R. Review on Determining Number of Cluster in K-Means Clustering. **International Journal of Advance Research in Computer Science and Management Studies**, v. 1, n. 6, p. 90-95, 2013.

LIU, Chenyue; MOSTAFAVI, Ali. FloodGenome: Interpretable machine learning for decoding features shaping property flood risk predisposition in cites. **Environmental Research: Infrastructure and Sustainability**, v. 5, n. 1, p. 015018, 2025.

LUNDELL, J. F.; BEAN, B.; SYMANZIK, J. Let's talk about the weather: a cluster-based approach to weather forecast accuracy. **Computational Statistics**, v. 38, n. 3, 2023.

MENDONÇA, Francisco; FARIAS, Ariadne; BUFFON, Elaiz (Orgs.). **Urban Flooding in Brazil**. Cham: Springer International Publishing, 2023.

OCDE – ORGANIZAÇÃO PARA COOPERAÇÃO E DESENVOLVIMENTO ECONÔMICO. **Handbook on Constructing Composite Indicators**: Methodology and User Guide. Paris: OECD/European Commission, 2008.

ROUSSEEUW, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. **Journal of Computational and Applied Mathematics**, v. 20, n. C, 1987.

SHI, J.; WANG, Z. A Hybrid Forecast Model for Household Electric Power by Fusing Landmark-Based Spectral Clustering and Deep Learning. **Sustainability**, v. 14, n. 15, 2022.

TATE, Eric; RUFAT, Samuel; ASIF, Rahman; HOOVER, Shelly. Social vulnerability correlates of flood risk to crops and buildings. **Natural Hazards**, 2025.

JUAHIR, H.; ZAIN, S. M.; YUSOFF, M. K.; HANIDZA, T. I. T.; ARMI, A. S. M.; TORIMAN, M. E.; MOKHTAR, M. Spatial water quality assessment of Langat River Basin (Malaysia) using environmetric techniques. **Environmental Monitoring and Assessment**, v. 173, n. 1–4, p. 625–641, 2011.