Universidade Federal de Juiz de Fora
Instituto de Ciências Exatas
Bacharelado em Sistemas de Informação

Descoberta de Padrões e Análise de Correlação entre Eventos Climáticos Extremos e Arboviroses no Brasil utilizando Técnicas de Mineração de Dados

Ítalo de Almeida Ribeiro

JUIZ DE FORA
JULHO, 2026

Descoberta de Padrões e Análise de Correlação entre Eventos Climáticos Extremos e Arboviroses no Brasil utilizando Técnicas de Mineração de Dados

Ítalo de Almeida Ribeiro

Universidade Federal de Juiz de Fora

Instituto de Ciências Exatas

Departamento de Ciência da Computação

Bacharelado em Sistemas de Informação

Orientador: Fabricio Martins Mendonça

JUIZ DE FORA

JULHO, 2026

Descoberta de Padrões e Análise de Correlação entre Eventos Climáticos Extremos e Arboviroses no Brasil utilizando Técnicas de Mineração de Dados

Ítalo de Almeida Ribeiro

MONOGRAFIA SUBMETIDA AO CORPO DOCENTE DO INSTITUTO DE CIÊNCIAS EXATAS DA UNIVERSIDADE FEDERAL DE JUIZ DE FORA, COMO PARTE INTEGRANTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE BACHAREL EM SISTEMAS DE INFORMAÇÃO.

Aprovada por:

Fabricio Martins Mendonça
Doutor em Ciência da Informação pela UFMG

Gleiph Ghiotto Lima de Menezes
Doutor em Computação pela UFF

Wagner Antonio Arbex
Doutor em Engenharia de Sistemas e Computação pela UFRJ

JUIZ DE FORA

14 DE JULHO, 2026

Resumo

O aumento de eventos climáticos extremos exige análises integradas para a gestão de riscos. A Ciência da Computação contribui significativamente nesse contexto por meio da integração de dados e extração de conhecimento. No Brasil, bases governamentais como o Atlas Digital de Desastres e a plataforma AdaptaBrasil MCTI (Ministério da Ciência, Tecnologia e Inovação) são fundamentais nessa área, porém heterogêneas e desintegradas. Esta pesquisa propõe a aplicação do processo de Descoberta de Conhecimento em Bases de Dados (KDD) para integrar e analisar essas fontes. Utilizando mineração de dados, processos de ETL (Extração, Transformação e carga), análise de correlação estatística e algoritmos de agrupamento não supervisionado, o objetivo principal é descobrir padrões que relacionem o histórico de extremos hídricos, como secas e inundações, com a vulnerabilidade climática, focando especificamente no risco de incidência de arboviroses. Como resultado, a pesquisa produziu análises gráficas integradas, destacando-se diagramas de dispersão, mapas de calor (heatmaps) de cruzamento de riscos e a segmentação dos municípios brasileiros em cinco perfis de risco composto. Os padrões quantitativos evidenciados nessas visualizações, a exemplo da correlação positiva entre o histórico de secas e a suscetibilidade a dengue, zika e chikungunya, validam os modelos de vulnerabilidade atuais e condizem com a literatura científica recente sobre a influência direta dos extremos climáticos nas crises de saúde pública.

Palavras-chave: Mineração de Dados, Risco Climático, Desastres Naturais, Arboviroses, KDD.

Abstract

The increase in extreme weather events requires integrated analyses for risk management. Computer Science contributes significantly to this context through data integration and knowledge extraction. In Brazil, governmental databases such as the Digital Atlas of Disasters and the AdaptaBrasil MCTI (Ministry of Science, Technology and Innovation) platform are fundamental in this area, yet heterogeneous and unintegrated. This research proposes the application of the Knowledge Discovery in Databases (KDD) process to integrate and analyze these sources. Using data mining, ETL (Extract, Transform, and Load) processes, statistical correlation analysis, and unsupervised clustering algorithms, the main objective is to discover patterns that relate the history of hydrological extremes, such as droughts and floods, to climate vulnerability, specifically focusing on the risk of arbovirus incidence. As a result, the research produced integrated graphical analyses, highlighting scatter plots, risk-crossing heatmaps, and the segmentation of Brazilian municipalities into five composite risk profiles. The quantitative patterns evidenced in these visualizations, such as the positive correlation between the history of droughts and the susceptibility to dengue, zika, and chikungunya, validate the current vulnerability models and align with recent scientific literature on the direct influence of climate extremes on public health crises.

Keywords: Data Mining, Climate Risk, Natural Disasters, Arboviruses, KDD.

Agradecimentos

Gostaria de agradecer minha família por todo apoio dado nesse tempo, além do auxílio nas decisões tomadas e ajudas técnicas. Gostaria de agradecer meus amigos por todos os dias em que reclamei das dificuldades e me ajudaram a seguir em frente. Gostaria de agradecer meus colegas de faculdade por todo o suporte frente as dificuldades enfrentadas no curso. Gostaria de agradecer minha namorada pelo apoio no período de escrita e ter me ajudado em todo o processo. Por fim, gostaria de agradecer o professor orientador pelos ensinamentos e apoio na escrita do trabalho. Obrigado a todos.

Conteúdo

Lista de Figuras

Lista de Tabelas

Lista de Abreviações

ETL	Extração, Transformação e Carga (*Extract, Transform, Load*)
IBGE	Instituto Brasileiro de Geografia e Estatística
IPCC	Painel Intergovernamental sobre Mudanças Climáticas
KDD	Descoberta de Conhecimento em Dados (*Knowledge Discovery in Databases*)
MCTI	Ministério da Ciência, Tecnologia e Inovações
ODS	Objetivos de Desenvolvimento Sustentável
ONU	Organização das Nações Unidas
PCA	Análise de Componentes Principais (*Principal Component Analysis*)

1 Introdução

1.1 Contextualização

A evolução nas análises meteorológicas tem permitido avanços significativos na predição de eventos climáticos em menor escala e em curtos períodos de tempo. Contudo, o cenário atual impõe desafios crescentes e complexos. Segundo Dias (2014), o interesse por eventos climáticos extremos tem se intensificado justamente devido ao aumento observável na sua frequência e escala, bem como pelos severos impactos socioambientais a eles associados. Apesar dos avanços na modelagem climática, ainda existem questões em aberto quanto a uma análise mais integrada dos dados, a fim de auxiliar na representação dos fenômenos. O mesmo autor ressalta que em regiões onde ocorre grande variabilidade no dia a dia, as condições médias e os extremos não são bem representados, permanecendo como um desafio de pesquisa.

Essa dificuldade dos modelos tradicionais em capturar especificidades locais e vulnerabilidades sobrepostas reforça a necessidade de abordagens alternativas, que utilizem o vasto histórico de dados reais para auxiliar nas projeções de risco. No contexto brasileiro, existem bases de dados sólidas e abrangentes, organizadas por órgãos federais, mas que operam de forma isolada. De um lado, tem-se o Atlas Digital de Desastres (BRASIL, 2022), que fornece o registro histórico consolidado das ocorrências e danos em todo o território nacional. Do outro, a plataforma AdaptaBrasil MCTI (INPE, 2025), que oferece índices de risco futuro baseados em projeções climáticas e indicadores socioecológicos.

1.2 Justificativa

A falta de uma integração direta entre o dano materializado registrado no Atlas e o risco projetado pela plataforma AdaptaBrasil cria uma lacuna significativa na validação dos modelos e na formulação de estratégias de mitigação focadas em áreas críticas. O problema de pesquisa abordado neste trabalho reside justamente no desafio técnico e

metodológico de realizar a fusão e a modelagem computacional dessas duas fontes de dados heterogêneas.

O preenchimento dessa lacuna se justifica pela necessidade de gerar novos conhecimentos estruturados sobre o impacto dos extremos climáticos. Ao aplicar técnicas de mineração de dados para cruzar os registros de desastres com as projeções de risco do AdaptaBrasil, torna-se possível descobrir padrões que só emergem da sobreposição de múltiplas dimensões, como a vulnerabilidade hídrica e a suscetibilidade sanitária. Para viabilizar uma análise aprofundada e estatisticamente consistente, a pesquisa delimita seu escopo a um recorte temático específico com foco na intersecção entre desastres hídricos (secas e inundações) e o índice de risco para a ocorrência de arboviroses (dengue, zika e chikungunya). O desenvolvimento deste pipeline de dados viabiliza a extração de inteligência acionável, oferecendo um suporte informacional sólido.

1.3 Questões de Pesquisa

A partir da problematização e do contexto apresentados, o presente estudo é norteado pelas seguintes questões de pesquisa que buscará responder:

1. Como os dados históricos de desastres naturais hídricos (Atlas) se correlacionam com as projeções futuras de risco à saúde, especificamente para arboviroses (AdaptaBrasil)?

2. Quais padrões multivariados de vulnerabilidade climática emergem quando integramos e clusterizamos diferentes dimensões de risco (hídrico, saúde, energético e biodiversidade) em conjunto com o histórico de ocorrências dos municípios?

1.4 Objetivos

Esta seção estabelece as metas a serem alcançadas com a realização desta pesquisa. Inicialmente, é apresentado o objetivo geral, que sintetiza o propósito central da investigação. Na sequência, são detalhados os objetivos específicos, que desdobram a meta principal em etapas práticas para a execução do estudo.

1.4.1 Objetivo Geral

O objetivo geral desta pesquisa é descobrir padrões multivariados e analisar as correlações entre o histórico de eventos climáticos extremos (secas e inundações) e o índice de risco para a ocorrência de arboviroses (dengue, zika e chikungunya) nos municípios brasileiros.

- Gerar uma base de dados integrada a partir da consolidação e compatibilização de fontes governamentais heterogêneas, unificando indicadores climáticos e registros históricos de ocorrências e saúde pública.
- Identificar padrões que relacionem os riscos climáticos futuros e o histórico de desastres naturais com a vulnerabilidade sanitária.
- Desenvolver visualizações gráficas e cartográficas que evidenciem as interseções de vulnerabilidade e o agravamento do risco de arboviroses em municípios com histórico crítico de extremos hídricos.
- Mapear e categorizar o risco composto dos municípios analisados, traduzindo a sobreposição de ameaças climáticas e epidemiológicas em conhecimento aplicável à gestão de desastres.

1.5 Estrutura do Trabalho

Para uma melhor compreensão e encadeamento lógico, o texto deste trabalho está organizado em seis capítulos principais. O Capítulo 1 inclui esta Introdução, responsável por apresentar a contextualização do tema, a justificativa da pesquisa, as questões que norteiam o estudo e os objetivos propostos. O Capítulo 2 apresenta a Fundamentação Teórica, detalhando os conceitos de Ciência de Dados, o processo de Descoberta de Conhecimento em Dados (KDD) e as bases matemáticas dos algoritmos utilizados. O Capítulo 3 aborda os Trabalhos Relacionados, trazendo um panorama do estado da arte sobre a integração e análise de bases de dados heterogêneas de clima e saúde. O Capítulo 4 descreve a Metodologia e o Desenvolvimento da pesquisa, delineando a natureza do estudo, a revisão sistemática da literatura e os procedimentos práticos de extração, transformação e carga (ETL). O Capítulo 5 expõe os Resultados obtidos, discutindo as correlações encontradas,

a formação dos perfis de vulnerabilidade por agrupamento e o mapeamento nacional dos riscos compostos. Por fim, o Capítulo 6 apresenta as Conclusões, destacando as limitações inerentes aos dados e propondo direcionamentos para trabalhos futuros.

2 Fundamentação Teórica

2.1 Desastres Naturais e Mudanças Climáticas

Desastres naturais podem ser definidos como o resultado do impacto de fenômenos naturais extremos ou intensos sobre um sistema social, causando sérios danos e prejuízos que excedem a capacidade da comunidade ou da sociedade atingida em conviver com o impacto (MARCELINO, 2008). Segundo o *Special Report on Extremes* (SREX) do IPCC (FIELD et al., 2012), a frequência desses eventos tem aumentado devido às mudanças climáticas antropogênicas. A compreensão desses fenômenos requer uma análise que vá além do evento físico, incorporando as dimensões de exposição e vulnerabilidade das comunidades afetadas. Metas internacionais, como os Objetivos de Desenvolvimento Sustentável (ODS) da ONU, especificamente os ODS 11 (Cidades e Comunidades Sustentáveis) (Nações Unidas Brasil, 2015a) e ODS 13 (Ação Contra a Mudança Global do Clima) (Nações Unidas Brasil, 2015b), reforçam a necessidade de dados integrados para a redução de riscos de desastres.

2.1.1 Precipitação e Chuvas Intensas

A chuva é formalmente definida como a precipitação de gotas de água com diâmetro superior a 0,5 mm. Quando de grande intensidade e curta duração, são denominadas aguaceiros (TOMINAGA; SANTORO; AMARAL, 2009). No Brasil, a chuva constitui o principal tipo de precipitação e atua como o principal elemento climático deflagrador de desastres naturais, funcionando como "gatilho" para ocorrências como inundações e escorregamentos (TOMINAGA; SANTORO; AMARAL, 2009).

2.1.2 Estiagem e Seca

A estiagem é caracterizada como um período prolongado de baixa pluviosidade ou de ausência total de chuvas, no qual as perdas de umidade do solo por evaporação e trans-

piração superam a reposição hídrica pelas precipitações (TOMINAGA; SANTORO; AMARAL, 2009). Trata-se de um desastre de natureza gradual, podendo um único evento estender-se por mais de um ano. Por representar um estágio inicial e menos severo de deficiência hídrica, a estiagem é considerada o precursor da seca (TOMINAGA; SANTORO; AMARAL, 2009).

A seca, por sua vez, constitui a forma crônica da estiagem (TOMINAGA; SANTORO; AMARAL, 2009): manifesta-se quando o período de escassez pluviométrica se prolonga de forma intensa, afetando grandes extensões territoriais. Atualmente, é reconhecida como um dos desastres naturais de maior ocorrência e impacto no mundo (TOMINAGA; SANTORO; AMARAL, 2009).

No contexto do Atlas Digital de Desastres (BRASIL, 2022), os registros de "Estiagem e Seca" abrangem ambos os fenômenos sob uma mesma tipologia, o que foi considerado na construção do perfil hídrico histórico dos municípios neste trabalho.

2.1.3 Inundação e Eventos Correlatos

A inundação é definida estritamente como o transbordamento das águas de um curso d'água — rio, córrego ou canal — para além de suas margens, atingindo a planície de inundação ou área de várzea (TOMINAGA; SANTORO; AMARAL, 2009). Trata-se de um problema geoambiental de origem hidrometeorológica, diretamente relacionado à quantidade, intensidade e frequência das chuvas, e cuja gravidade é amplificada tanto por condicionantes naturais (grau de saturação do solo e morfologia do relevo) quanto por fatores antrópicos, como a impermeabilização do solo urbano, o desmatamento e a ocupação irregular de áreas de várzea (TOMINAGA; SANTORO; AMARAL, 2009).

É importante distinguir a inundação de fenômenos correlatos frequentemente confundidos (TOMINAGA; SANTORO; AMARAL, 2009):

- **Enchente (ou cheia):** corresponde apenas à elevação do nível da água dentro do leito do rio, atingindo a cota máxima sem necessariamente transbordar para as margens. Toda inundação pressupõe uma enchente, mas nem toda enchente resulta em inundação.

- **Alagamento:** acúmulo momentâneo de água em vias ou áreas urbanas decorrente de deficiência no sistema de drenagem pluvial, podendo ocorrer independentemente do transbordamento de cursos d'água.

- **Enxurrada:** escoamento superficial concentrado e de alta energia cinética, comum em terrenos de elevada declividade ou superfícies impermeabilizadas durante episódios de chuva intensa.

No Atlas Digital de Desastres (BRASIL, 2022), os registros englobam as tipologias "Inundações", "Enxurradas" e "Alagamentos" como categorias distintas, todas consideradas na classificação de eventos de excesso hídrico adotada neste trabalho.

2.2 Ciência de Dados e Mineração de dados

Para analisar e integrar bases de dados tão complexas e heterogêneas como o Atlas de Desastres e o AdaptaBrasil, é fundamental definir os conceitos de Ciência de Dados e Mineração de Dados, que fornecem o ferramental teórico e prático para este trabalho.

2.2.1 Ciência de Dados

A Ciência de Dados é um campo multidisciplinar e abrangente, formalmente definido como a ciência de aprender com os dados (DONOHO, 2017). Com a explosão na capacidade computacional e na tecnologia da informação nas últimas décadas, houve uma geração maciça de dados em diversas áreas do conhecimento. O desafio de compreender, processar e extrair valor dessa imensa quantidade de informação impulsionou o desenvolvimento de novas ferramentas estatísticas e o surgimento de áreas correlatas, como a mineração de dados (*data mining*) e o aprendizado de máquina (*machine learning*) (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). Dessa forma, a Ciência de Dados é frequentemente descrita como um superconjunto que engloba a estatística, a computação e o conhecimento de domínio específico, acrescido da tecnologia necessária para operar em escala com grandes volumes de dados (Big Data).

O escopo da Ciência de Dados envolve o ciclo de vida completo da informação: coleta, gerenciamento, processamento, análise, visualização e interpretação de vastas quan-

tidades de dados heterogêneos (DONOHO, 2017). Para formalizar esse ciclo, Donoho (2017) propõe o conceito de *Greater Data Science* (GDS), que divide a área em seis divisões principais. Para o contexto deste trabalho, destacam-se a Preparação e Exploração de Dados (GDS1) e a Representação e Transformação de Dados (GDS2). Estas fases, que incluem a limpeza de artefatos (*data cleaning*) e a estruturação de variáveis (*data wrangling*), são etapas críticas para lidar com anomalias e garantir a integridade da análise. Na prática da engenharia de dados e sistemas de informação, esse esforço equivale diretamente ao processo de ETL (*Extract, Transform, Load*), aplicado neste projeto para padronizar e unificar os formatos espaciais e temporais distintos das bases do Atlas Digital de Desastres e do AdaptaBrasil.

Além da preparação técnica, a Ciência de Dados engloba a Modelagem de Dados (GDS4) e a Visualização e Apresentação (GDS5) (DONOHO, 2017). A modelagem é o estágio onde algoritmos de aprendizado estatístico e de máquina são aplicados para identificar padrões estruturais implícitos e não triviais (WITTEN; FRANK; HALL, 2011). No paradigma não supervisionado, os algoritmos buscam entender o comportamento multivariado das características e agrupar instâncias semelhantes sem a necessidade de um rótulo prévio (HASTIE; TIBSHIRANI; FRIEDMAN, 2009) — técnica que fundamenta a clusterização (K-Means) utilizada nesta pesquisa para definir o Risco Composto dos municípios.

Por fim, a etapa de visualização (GDS5) cumpre o papel essencial de traduzir a complexidade matemática e algorítmica em informação gerencial e visualmente acessível, consolidando a ponte entre o dado processado e o conhecimento acionável. No escopo deste projeto, isso se materializa na construção do painel analítico.

No contexto deste TCC, a etapa de Preparação e Exploração de Dados (referida por Donoho (2017) como Greater Data Science 1 (GDS1)) é crucial. Esta fase, que inclui *data cleaning* (limpeza) e *data wrangling* (transformação), é essencial para lidar com anomalias e artefatos. Isso equivale diretamente ao processo de ETL (*Extract, Transform, Load*) aplicado para padronizar e unificar os formatos espaciais e temporais distintos das bases do Atlas e AdaptaBrasil.

2.2.2 Mineração de Dados

A Mineração de Dados, frequentemente referida como Descoberta de Conhecimento em Dados (KDD, do inglês *Knowledge Discovery in Databases*), é o processo de extração automatizada de padrões interessantes e não triviais que representam conhecimento, os quais estão implicitamente armazenados em grandes repositórios de dados (HAN; PEI; KAMBER, 2011). Trata-se de um campo multidisciplinar que converge princípios estatísticos, tecnologias de banco de dados e algoritmos de aprendizado de máquina (*machine learning*) para revelar estruturas ocultas em conjuntos de dados massivos (WITTEN; FRANK; HALL, 2011). O processo de KDD não se limita apenas à aplicação de algoritmos, mas engloba etapas cruciais de limpeza, integração, seleção, transformação, mineração propriamente dita, avaliação de padrões e, por fim, a apresentação visual do conhecimento.

No escopo deste projeto de pesquisa, a mineração de dados atua como o motor analítico central para investigar e validar as relações entre os indicadores de risco projetados pela plataforma AdaptaBrasil e o dano histórico mapeado pelo Atlas Digital de Desastres. O painel de visualização gerado a partir dessa modelagem tem o objetivo de consolidar as informações extraídas para facilitar a compreensão técnica do cenário de risco, fornecendo uma base sólida de dados integrados para o planejamento estrutural, a avaliação de vulnerabilidades espaciais e a gestão estratégica de desastres naturais.

Para alcançar os objetivos propostos na integração das bases heterogêneas, as seguintes técnicas de mineração de dados e aprendizado estatístico são exploradas e aplicadas para a extração de conhecimento (HAN; PEI; KAMBER, 2011; HASTIE; TIBSHIRANI; FRIEDMAN, 2009):

- **Análise de Associação e Correlação:** Busca descobrir regras, tendências e relacionamentos frequentes entre conjuntos de variáveis qualitativas ou quantitativas. Neste trabalho, essa técnica é traduzida na investigação estatística (como o coeficiente de Pearson) da correlação cruzada entre um índice de risco climático específico, como a suscetibilidade a arboviroses, e a frequência acumulada de eventos hídricos, permitindo verificar a dependência linear entre a vulnerabilidade projetada e a ocorrência do dano materializado.

- **Clusterização (Agrupamento):** Consiste no aprendizado não supervisionado focado em dividir os dados em subgrupos, de forma que objetos de um mesmo *cluster* sejam altamente similares entre si e consideravelmente diferentes dos objetos alocados em outros grupos (HAN; PEI; KAMBER, 2011). Na presente pesquisa, utiliza-se essa abordagem (por meio do algoritmo K-Means) para agrupar de forma multivariada os municípios brasileiros com base na sobreposição de múltiplas dimensões, como riscos hídrico, de saúde, energético e de biodiversidade, revelando perfis latentes de vulnerabilidade composta no território.

- **Classificação e Regressão:** Compreendem os métodos de aprendizado supervisionado utilizados para prever rótulos categóricos ou estimar tendências numéricas contínuas (WITTEN; FRANK; HALL, 2011). No contexto do processamento realizado, a modelagem por regressão linear simples auxilia na verificação do comportamento e da direção estrutural entre a quantidade de desastres históricos registrados por município e as notas normalizadas dos índices de risco, validando matematicamente os padrões visuais encontrados nos dados.

2.2.3 O Algoritmo K-Means

O algoritmo K-Means é um dos métodos de clusterização por partição mais difundidos na literatura de mineração de dados e aprendizado de máquina (HAN; PEI; KAMBER, 2011). O objetivo primordial desse algoritmo consiste em particionar um conjunto de n objetos multidimensionais em k grupos distintos, definidos a priori pelo pesquisador, de modo que a similaridade entre os elementos pertencentes ao mesmo grupo seja maximizada, enquanto a similaridade entre grupos diferentes seja minimizada (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

O funcionamento do K-Means é estruturado de forma iterativa e baseia-se na minimização da soma dos erros quadráticos (SSE, do inglês *Sum of Squared Errors*), que mede a distância dos pontos em relação ao centro do seu agrupamento. O procedimento algorítmico segue os passos descritos por Han, Pei e Kamber (2011):

1. **Inicialização:** O algoritmo seleciona aleatoriamente k objetos do conjunto de da-

dos para atuarem como os centros iniciais dos grupos, os quais são denominados centroides.

2. **Atribuição:** Cada objeto do conjunto de dados é associado ao centroide mais próximo. Essa proximidade é quantificada por meio de uma métrica de distância no espaço multidimensional, sendo a distância euclidiana a mais comumente empregada (WITTEN; FRANK; HALL, 2011).

3. **Atualização:** Após a alocação de todos os objetos, os novos centroides são recalculados obtendo-se a média aritmética das coordenadas de todos os pontos atribuídos a cada respectivo grupo.

4. **Convergência:** Os passos de atribuição e atualização são repetidos ciclicamente até que nenhum objeto altere sua alocação de grupo ou os centroides permaneçam estáveis, atingindo o critério de parada.

Matematicamente, para um conjunto de dados com atributos normalizados, a distância euclidiana entre um objeto municipal x e o centroide c é calculada de acordo com a Equação (2.1):

$$d(x, c) = \sqrt{\sum_{i=1}^{m} (x_i - c_i)^2} \tag{2.1}$$

Em que m representa o número de dimensões avaliadas, correspondendo, nesta pesquisa, aos quatro índices de risco analisados. Apesar de sua eficiência computacional, o K-Means apresenta sensibilidade à escolha inicial dos centroides e à presença de valores discrepantes (*outliers*), o que requer o tratamento prévio e a normalização das escalas numéricas das variáveis antes de sua execução (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

2.2.4 Análise de Componentes Principais (PCA)

A Análise de Componentes Principais (PCA, do inglês *Principal Component Analysis*) é uma técnica estatística e de mineração de dados amplamente utilizada para a redução de dimensionalidade de conjuntos de dados complexos (HAN; PEI; KAMBER, 2011). Em

bases de dados com múltiplas variáveis, como as diversas dimensões de risco climático analisadas neste estudo, a visualização direta e a interpretação tornam-se inviáveis em um espaço tridimensional ou superior.

O objetivo do PCA é transformar os atributos originais em um novo conjunto de variáveis independentes, denominadas componentes principais, que são combinações lineares das variáveis originais (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). Esse processo matemático garante que a primeira componente principal (PC1) capture a maior proporção possível da variância total dos dados. A segunda componente (PC2), ortogonal à primeira, captura a máxima variância residual, e assim sucessivamente.

Ao projetar os dados multivariados em um plano bidimensional formado apenas por PC1 e PC2, é possível preservar a maior parte das informações estruturais e da variabilidade original, facilitando a identificação gráfica da separação espacial dos agrupamentos gerados por algoritmos como o K-Means (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

2.2.5 Correlação de Pearson e Regressão Linear Simples

Para a descoberta de padrões entre variáveis contínuas, a estatística fornece ferramentas essenciais que foram incorporadas ao processo de mineração de dados, com destaque para a análise de correlação e os modelos de regressão (WITTEN; FRANK; HALL, 2011).

A Correlação de Pearson é uma métrica paramétrica que quantifica o grau de dependência linear entre duas variáveis numéricas. O coeficiente resultante, denotado por r, varia em um intervalo contínuo de -1 a $+1$ (HAN; PEI; KAMBER, 2011). Valores positivos indicam uma relação diretamente proporcional (quando uma variável cresce, a outra também cresce), enquanto valores negativos indicam uma relação inversamente proporcional. Um coeficiente igual ou muito próximo a zero sugere a inexistência de correlação linear, indicando independência estatística entre os atributos. É fundamental ressaltar, contudo, que a existência de correlação estatística não implica causalidade. Conforme explicam Han, Pei e Kamber (2011), atestar que duas variáveis estão correlacionadas não comprova uma relação de causa e efeito direta entre elas, visto que ambas podem ser influenciadas por fatores subjacentes ou terceiras variáveis. Inferir causalidade estritamente a partir de correlações isoladas é considerado inadequado no processo de mineração de

dados (WITTEN; FRANK; HALL, 2011).

Para modelar matematicamente a estrutura dessa relação linear, emprega-se a Regressão Linear Simples. Esse método preditivo busca ajustar uma reta ótima que melhor descreva a dispersão dos dados bidimensionais (HAN; PEI; KAMBER, 2011). A equação da reta é definida por $y = wx + b$, em que y é a variável dependente (o valor a ser estimado), x é a variável independente (o dado de entrada), w representa a inclinação ou peso da reta, e b define a intersecção com o eixo das ordenadas. Na prática computacional, os coeficientes w e b são otimizados iterativamente por meio do método dos mínimos quadrados, cujo objetivo é minimizar o erro acumulado entre os valores reais observados e os valores projetados pela reta teórica (WITTEN; FRANK; HALL, 2011).

3 Trabalhos Relacionados

A literatura científica recente tem explorado de forma intensa o cruzamento de bases de dados heterogêneas, com especial atenção aos repositórios governamentais, para auxiliar a compreensão técnica dos impactos climáticos na gestão em saúde. Esta seção apresenta detalhadamente seis trabalhos que fundamentam a presente pesquisa, abordando tanto as semelhanças temáticas quanto as metodologias aplicáveis à mineração de dados em larga escala.

3.1 Relações Espaçotemporais entre Extremos Climáticos e Arboviroses

O estudo conduzido por Cox et al. (2025) apresenta uma investigação aprofundada sobre a relação espaçotemporal entre eventos climáticos extremos e a transmissão de arboviroses (dengue, zika e chikungunya) em todos os municípios do Brasil, abrangendo o período de 2013 a 2020. O objetivo central dos pesquisadores foi compreender como as anomalias climáticas de longo prazo (a exemplo do El Niño) e os choques de temperatura e precipitação atuam como catalisadores para a proliferação vetorial, gerando cenários de crise aguda na saúde.

Como principal contribuição, os autores utilizaram modelos de regressão avançados para integrar dados de incidência de doenças com variáveis meteorológicas e indicadores socioeconômicos. O trabalho demonstrou matematicamente que a incidência de arboviroses está fortemente associada a períodos extremos de seca ou umidade, sendo severamente agravada por condições de vulnerabilidade social.

Apesar dos resultados robustos, o estudo apresenta limitações inerentes à modelagem de dados climáticos e epidemiológicos. Os próprios autores destacam que a dependência de notificações oficiais de saúde pode mascarar o real cenário devido à subnotificação em municípios menores ou durante períodos interepidêmicos. Além disso, a

aplicação de modelos puramente matemáticos de regressão impõe dificuldades para a extrapolação de resultados em microclimas muito específicos, indicando a necessidade de abordagens complementares que agrupem áreas de risco semelhante.

3.2 Integração de Dados Multidisciplinares para Doenças Sensíveis ao Clima

Na busca por ferramentas operacionais, Dasgupta et al. (2025) propuseram o desenvolvimento do DART (Dengue Advanced Readiness Tools), um *pipeline* de integração de dados escalável e de código aberto voltado especificamente para o monitoramento de doenças infecciosas sensíveis às variações climáticas. O estudo parte da premissa de que a falta de infraestrutura tecnológica unificada é o maior gargalo para a análise preditiva em saúde.

A pesquisa oferece uma contribuição metodológica de alto valor ao abordar o desafio de unificar dados epidemiológicos, socioeconômicos e ambientais que possuem resoluções espaciais e temporais muito distintas. A ferramenta criada automatiza o processo de extração, correção de viés estatístico e agregação temporal, provando que a integração padronizada de múltiplas fontes é o alicerce para qualquer análise robusta de risco.

Como limitações, o estudo aponta os altos custos computacionais exigidos para o processamento de dados globais em tempo real, além da barreira de interoperabilidade quando o modelo tenta ingerir dados de países com infraestruturas governamentais precárias. O trabalho sugere que o *pipeline* necessita de constante calibração manual para lidar com formatos de dados não estruturados ou inconsistentes, um desafio clássico das etapas de preparação de dados.

3.3 Integração e Análise de Dados Médicos e Ambientais usando ETL e BI

Focando na arquitetura de sistemas, o trabalho de Villar et al. (2018) investiga a superação das barreiras técnicas impostas pela heterogeneidade de fontes de dados na pesquisa biomédica e climática. Os pesquisadores estabeleceram como meta projetar um

sistema capaz de ingerir bases díspares e entregá-las de forma visualmente compreensível para usuários finais na Espanha.

A grande contribuição do trabalho é a implementação validada de uma solução tecnológica baseada no uso conjunto de processos de ETL (Extração, Transformação e Carga) integrados a ferramentas de *Business Intelligence* (BI). O estudo relata o processamento de dezenas de milhões de registros governamentais sobre qualidade do ar e admissões hospitalares, comprovando que um *pipeline* bem estruturado de ETL permite normalizar formatos distintos de arquivos, facilitando a agregação de dados e a descoberta visual de padrões sazonais através de painéis dinâmicos.

Por outro lado, o trabalho reconhece como limitação a forte dependência de Data Warehouses rigidamente estruturados, o que dificulta a agilidade em caso de mudança brusca na formatação dos dados fornecidos pelas agências governamentais. Os autores observam que a etapa de transformação (a "Limpeza"no KDD) consumiu a maior parte dos recursos do projeto, demandando a criação de regras rigorosas que não são facilmente escaláveis para outros domínios sem extenso esforço de reprogramação.

3.4 Mineração em Múltiplas Bases de Dados Relacionais Heterogêneas

Em uma abordagem estritamente computacional, Mehenni (2020) propôs o desenvolvimento de um *framework* voltado exclusivamente para a aplicação de mineração de dados sobre múltiplas bases de dados relacionais e fisicamente separadas. O autor analisa a insuficiência técnica dos algoritmos clássicos quando aplicados em repositórios descentralizados.

A pesquisa contribui ao demonstrar que os métodos tradicionais de mineração geralmente assumem um repositório único e padronizado, o que falha ao lidar com a realidade de sistemas governamentais distribuídos, cujas chaves de ligação são incompatíveis ou inexistentes. O trabalho apresenta o desenvolvimento de técnicas de correspondência de esquema (*schema matching*) para resolver a heterogeneidade estrutural de forma algorítmica, operando a harmonização antes mesmo da aplicação de técnicas de classificação

ou agrupamento.

A principal limitação discutida por Mehenni (2020) reside na sobrecarga computacional exigida pelas operações de junção (*join*) em bases de grande escala, especialmente quando não há chaves primárias evidentes. Além disso, o *framework* proposto atua fortemente sobre bancos de dados puramente relacionais, não prevendo rotinas otimizadas para ingestão de dados semiestruturados ou arquivos de texto plano, exigindo a pré-formatação dos repositórios originais.

3.5 Mineração de Dados em Big Data Clínico

Buscando consolidar as técnicas da área, Wu et al. (2021) publicaram uma revisão metodológica abrangente detalhando os modelos, as etapas obrigatórias e os bancos de dados mais frequentemente utilizados em processos de mineração de *Big Data* clínico e epidemiológico.

O trabalho mapeia e explica minuciosamente o fluxo completo de descoberta de conhecimento em saúde, enfatizando a transição desde a limpeza de ruídos e tratamento de valores ausentes até a aplicação de algoritmos avançados de aprendizado de máquina. Os autores dão especial destaque às técnicas não supervisionadas, como o agrupamento hierárquico, o algoritmo K-Means e a Análise de Componentes Principais (PCA). Eles argumentam que a mineração possui vantagens inigualáveis na pesquisa médica em larga escala, pois é capaz de identificar subgrupos de risco latentes sem exigir premissas rígidas sobre a distribuição teórica dos dados.

Embora possua alto valor didático, por tratar-se de um estudo de revisão, a pesquisa carece da apresentação de um estudo de caso aplicado que integre todas as etapas em um *pipeline* unificado. Os autores relatam que o maior gargalo identificado na literatura revisada continua sendo a gestão da qualidade dos dados governamentais de origem, visto que algoritmos como o K-Means são notórios por sua alta sensibilidade a valores extremos e registros ausentes.

3.6 Procedimentos para Vinculação de Dados da Saúde

Por fim, Garcia, Miranda e Sousa (2022) desenvolveram e apresentaram uma metodologia padronizada voltada para a vinculação (*linkage*) determinística de diferentes bancos de dados estruturados em saúde pública. O objetivo prático do estudo foi criar rotinas operacionais de baixo custo para superar o isolamento dos sistemas de informação (SIS) no Brasil.

O estudo detalha os procedimentos computacionais exatos necessários para tratar inconsistências nominais e unificar registros através de operações matemáticas de junção (especialmente o *inner_join* e o *left_join*), utilizando chaves identificadoras comuns (como números de identificação ou códigos regionais). A pesquisa conclui que a técnica de vinculação determinística é uma estratégia altamente eficaz para recuperar registros incompletos, consolidando a confiabilidade das bases de dados antes das análises de inteligência.

Como apontamento crítico, os autores evidenciam que o relacionamento puramente determinístico apresenta baixos índices de sensibilidade se os dados originais sofrerem com erros de digitação recorrentes. Nessas situações, o modelo perde pareamentos válidos devido à falta de tolerância fonética ou aproximação estatística, exigindo que a fase de limpeza (remoção de caracteres especiais e padronização de nomenclatura) seja executada de forma impecável antes de qualquer operação de fusão.

3.7 Síntese e Relação com a Presente Pesquisa

A literatura revisada compõe o alicerce teórico e metodológico que embasa as decisões arquiteturais da presente pesquisa. Longe de atuarem de forma isolada, os trabalhos analisados conectam-se em uma cadeia lógica que justifica todas as etapas do *pipeline* desenvolvido neste Trabalho de Conclusão de Curso, do tratamento da base de dados à apresentação visual dos algoritmos. O tema encontra paralelo no estudo de Cox et al. (2025), que estabelece a pertinência da relação entre extremos hídricos e arboviroses. O modelo proposto neste TCC estende essa temática ao comprovar, por meio da correlação de Pearson, que o agravamento da suscetibilidade sanitária brasileira reflete diretamente o histórico de secas e inundações registradas, trazendo uma validação matemática às

conclusões epidemiológicas da referida obra.

O processo de transição da teoria de tratamento de dados para a prática em bases heterogeneas, baseia-se nas pesquisas feitas por Dasgupta et al. (2025) e Villar et al. (2018). Assim como evidenciado por esses autores, o presente estudo reconheceu a inviabilidade de analisar as projeções do AdaptaBrasil e o histórico do Atlas Digital em seus formatos originais, o que motivou o desenvolvimento rigoroso do processo de ETL. O gargalo computacional de alinhar escalas heterogêneas e a necessidade de projetar um painel de visualização dinâmico estão em perfeita sintonia com a integração multidisciplinar e as ferramentas de *Business Intelligence* validadas em seus achados.

O sucesso da etapa de fusão de dados governamentais é justificado pelas técnicas operacionais discutidas por Mehenni (2020) e Garcia, Miranda e Sousa (2022). A estratégia adotada neste projeto, que utilizou o Código IBGE municipal como chave de por meio da operação de *Left Join*, reflete a dificuldade prática para a vinculação determinística de bases de fontes distintas descritas nestas literaturas. Por fim, a chancela algorítmica para a exploração da base resultante advém de Wu et al. (2021), que apontam o K-Means e a redução de dimensionalidade por PCA como métodos ideais para o manuseio de *Big Data* com características socioambientais. Baseado nessa fundamentação, este TCC aplicou os referidos algoritmos para segmentar mais de cinco mil municípios, provando que a sobreposição de vulnerabilidades cria perfis de risco composto matematicamente consistentes e altamente interpretáveis.

4 Metodologia

Este capítulo apresenta a metodologia adotada para o desenvolvimento desta pesquisa, detalhando as etapas, técnicas e ferramentas computacionais utilizadas na investigação das correlações entre eventos climáticos extremos e o risco de ocorrência de arboviroses no cenário brasileiro.

Para estruturar a apresentação desses procedimentos, o capítulo divide-se em duas partes principais. A Seção 3.1 descreve a metodologia de pesquisa em si, caracterizando sua natureza, sua abordagem e as etapas conceituais planejadas para o estudo. Na sequência, a Seção 3.2 detalha o desenvolvimento da pesquisa, explicitando o arcabouço tecnológico, as bases de dados escolhidas, a implementação do processo de Extração, Transformação e Carga (ETL) e a aplicação prática dos algoritmos de mineração de dados.

4.1 Metodologia de Pesquisa

A metodologia deste trabalho classifica-se quanto à sua natureza, abordagem e objetivos da seguinte maneira (MARCONI; LAKATOS, 2019):

- **Natureza Aplicada:** Objetiva gerar conhecimentos para aplicação prática, voltados à solução de problemas específicos que envolvem a gestão de desastres e o planejamento de políticas públicas de adaptação climática.
- **Abordagem Quantitativa:** Fundamenta-se no processamento de um grande volume de dados numéricos provenientes de bases governamentais, utilizando estatística descritiva, análise de correlações e métricas matemáticas de algoritmos de agrupamento para validar as hipóteses.
- **Caráter Exploratório:** Busca proporcionar maior familiaridade com o problema, visto que busca a integração direta entre o risco futuro (AdaptaBrasil) e o dano histórico (Atlas) para descobrir padrões multivariados.

- **Caráter Qualitativo:** Complementa a etapa quantitativa, uma vez que a interpretação dos padrões encontrados, a definição dos rótulos dos *clusters* e a categorização do "Risco Composto"exigem uma análise profunda do contexto socioambiental para que os dados numéricos gerem inteligência acionável.

4.1.1 Revisão e Seleção da Literatura de Base

Para compor o embasamento teórico e selecionar os trabalhos correlatos discutidos no Capítulo 3, foi conduzido um processo de busca e seleção bibliográfica sistematizado. O objetivo desta etapa foi mapear a literatura científica recente acerca da integração de bases de dados de saúde e ambiente, bem como a aplicação de técnicas de mineração de dados nesse contexto multidisciplinar.

A estratégia de busca fundamentou-se na composição de descritores relacionados aos três eixos principais da pesquisa: os repositórios de dados sanitários, as técnicas de integração de informações heterogêneas e as ferramentas de mineração de dados. Para garantir o alcance internacional e a aderência técnica, formulou-se a seguinte *string* de busca em inglês:

```
("health databases" OR "public health databases" OR "epidemiology")
AND ("database integration" OR "data fusion")
AND ("data mining" OR "machine learning" OR "clustering" OR "k-means")
```

Esta chave de busca foi aplicada em bases acadêmicas consolidadas na área de Computação e Saúde, incluindo repositórios como Scopus, PubMed, IEEE Xplore, ACM Digital Library e Google Scholar. O processo de triagem e seleção do material recuperado ocorreu mediante etapas sucessivas de filtragem, estruturadas da seguinte maneira:

1. **Busca Inicial e Critérios de Inclusão:** A busca retornou um conjunto inicial abrangente de publicações. Como critérios de inclusão, foram considerados artigos completos, publicados em periódicos ou anais de conferências revisados por pares, disponíveis integralmente na língua portuguesa ou inglesa, e que apresentassem aderência técnica direta ao tema do cruzamento de dados de saúde com o uso

de algoritmos. Foram excluídas publicações incompletas, resumos simples, blogs e materiais puramente opinativos.

2. **Leitura de Títulos e Resumos:** O volume bruto de resultados passou por uma triagem primária. Foram avaliados os títulos e os resumos (*abstracts*) para identificar publicações que efetivamente reportassem operações de *linkage*, processos de ETL ou a aplicação de algoritmos não supervisionados em bases governamentais.

3. **Leitura Integral e Seleção Final:** Os trabalhos pré-selecionados foram submetidos à leitura técnica completa. Desta fase, foram extraídos os 6 (seis) artigos de maior relevância metodológica, que compuseram o conjunto principal detalhado no escopo de Trabalhos Relacionados deste TCC, assegurando que o estudo estivesse amparado por casos de sucesso práticos e desafios consolidados na comunidade científica.

O escopo metodológico foi estruturado em quatro etapas principais:

1. **Revisão Bibliográfica:** Levantamento sistemático de trabalhos que aplicaram mineração de dados em desastres naturais no Brasil, além do estudo aprofundado dos manuais do AdaptaBrasil e do Atlas para compreensão das variáveis (BRASIL, 2022; INPE, 2025).

2. **Coleta e Tratamento de Dados (ETL):** Extração, limpeza e junção dos dados governamentais heterogêneos.

3. **Análise e Mineração de Dados:** Execução de análises estatísticas e aplicação de algoritmos de agrupamento.

4. **Cruzamento e Visualização:** Geração de um painel e gráficos analíticos que explicitem as relações cruzadas para auxílio à tomada de decisão.

4.2 Desenvolvimento

A etapa de desenvolvimento do trabalho compreende o processo prático de integração das bases de dados, a implementação das rotinas de ETL (*Extract, Transform, Load*) e a

aplicação das técnicas de Mineração de Dados. Para a execução dessas tarefas, utilizou-se a linguagem de programação Python, na sua versão 3, devido ao seu robusto ecossistema de bibliotecas voltadas à Ciência de Dados.

O fluxo de trabalho foi projetado como um pipeline sequencial e incremental, fundamentado no processo clássico de Descoberta de Conhecimento em Dados (KDD) proposto por Han, Pei e Kamber (2011). Para fins de clareza metodológica, a arquitetura geral do sistema e o caminho percorrido pelo dado, desde a sua captação bruta até a camada de apresentação final, estão detalhados na Figura 4.1.

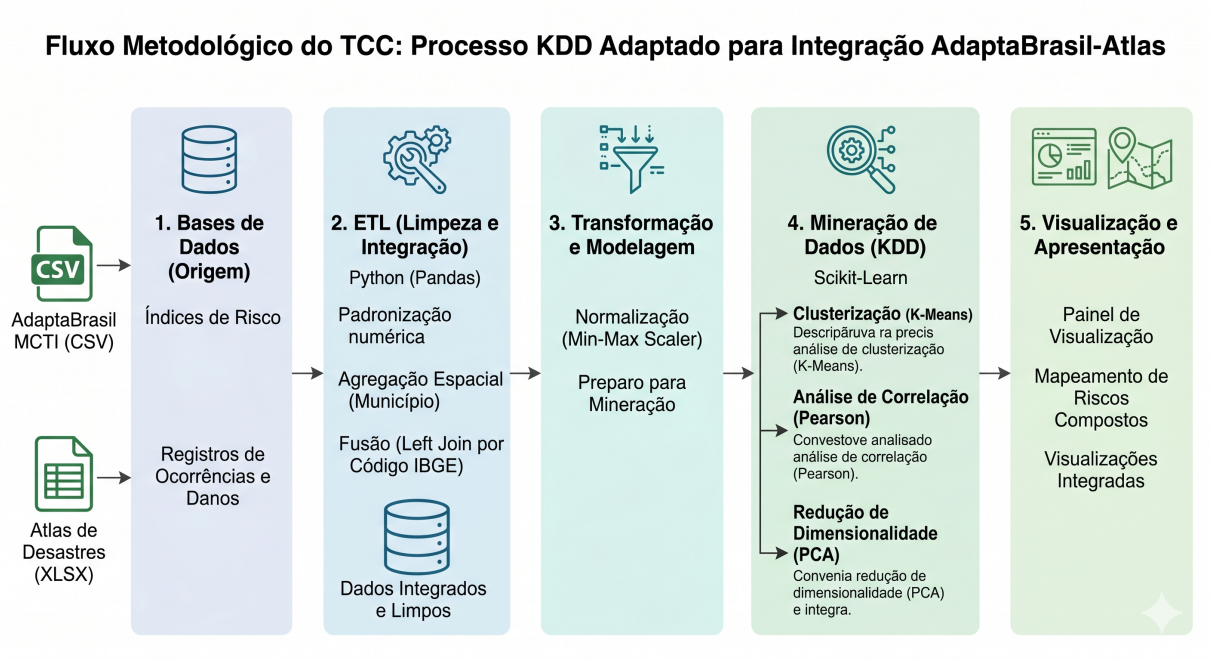

Figura 4.1: Fluxo metodológico do TCC representando o pipeline de dados customizado, estendendo-se desde as fontes de dados de origem até a camada final de visualização e apresentação do conhecimento. Fonte: Elaborado pelo autor com auxílio de ferramentas de IA (2026).

Conforme ilustrado no diagrama estrutural, o pipeline é composto por cinco fases interdependentes que transformam os registros governamentais isolados em uma base de dados unificada e inteligível, dividindo-se em:

1. **Bases de Dados (Origem):** Camada responsável pela ingestão dos arquivos brutos extraídos diretamente dos órgãos federais, englobando as projeções do AdaptaBrasil MCTI e a planilha histórica do Atlas de Desastres.

2. **ETL (Limpeza e Integração):** Fase em que os scripts desenvolvidos em Python

utilizam a biblioteca `pandas` para padronizar os formatos numéricos regionais, agregar os eventos históricos no nível de granularidade municipal e executar a junção das tabelas por meio de uma operação de *Left Join* utilizando o Código IBGE como chave primária.

3. **Transformação e Modelagem:** Etapa focada na preparação matemática dos atributos consolidados, aplicando a normalização *Min-Max Scaler* para unificar as escalas das variáveis e tratar a influência de valores discrepantes.

4. **Mineração de Dados (KDD):** Núcleo algorítmico suportado pela biblioteca `scikit-learn`, onde ocorrem a clusterização multivariada via K-Means, o cálculo estatístico de correlação de Pearson e a redução de dimensionalidade por meio do algoritmo PCA.

5. **Visualização e Apresentação:** Camada de saída que converte os clusters e correlações identificadas em artefatos visuais gráficos e cartográficos, estruturando o painel analítico e a métrica de risco composto.

4.2.1 Bases de Dados Utilizadas

Os dados processados originam-se de duas fontes federais distintas:

- **AdaptaBrasil MCTI:** Foram extraídas quatro bases de dados em formato CSV, representando os índices de risco climático nos setores estratégicos de Recursos Hídricos (2020), Saúde/Arboviroses (2020), Segurança Energética (2019) e Biodiversidade (2017). Dessa plataforma, os dados específicos utilizados na pesquisa englobam os valores numéricos contínuos de risco projetado para cada município, que variam em escalas de 0 a 1 para as dimensões hídrica e energética e de 0 a 100 para saúde e biodiversidade. Adicionalmente, foram utilizadas as suas respectivas classificações categóricas de gravidade, as quais dividem os municípios nas classes Muito baixo, Baixo, Médio, Alto e Muito alto.

- **Atlas Digital de Desastres no Brasil:** Base gerada e gerenciada pelo Ministério do Desenvolvimento Regional (MDR) por meio da Secretaria Nacional de Proteção e Defesa Civil (SEDEC) (BRASIL, 2022). Desta fonte foi extraída a base histórica

consolidada contemplando os registros de ocorrências de desastres, danos humanos e prejuízos materiais entre os anos de 1991 e 2024. Os dados específicos explorados e agregados desta base incluem o número absoluto e acumulado de eventos hidrológicos extremos por município, categorizados em secas, estiagens, inundações, enxurradas e chuvas intensas. Além da frequência dos eventos, foram utilizados os dados quantitativos de impacto, englobando o somatório de óbitos, desabrigados e desalojados, bem como a consolidação monetária dos prejuízos econômicos públicos e privados decorrentes dessas ocorrências.

4.2.2 Processo de ETL (Extração, Transformação e Carga)

A fase de ETL foi implementada utilizando as bibliotecas `pandas` e `numpy`, mapeando a transição da segunda para a terceira camada da arquitetura proposta. As principais tarefas envolveram:

- **Extração e Limpeza Dimensional:** Leitura dos arquivos originais e filtragem das colunas de interesse. Como os dados possuíam o padrão brasileiro de formatação numérica, desenvolveu-se uma rotina para remoção de pontos separadores de milhar e substituição de vírgulas por pontos, permitindo a conversão correta de cadeias de caracteres para tipos numéricos de ponto flutuante (*float*).

- **Agregação Espacial:** A base do Atlas, originalmente orientada a eventos individuais cadastrados via Protocolo S2iD, foi agregada por município utilizando a função de agrupamento `groupby`. Contabilizou-se o total de eventos, o somatório de óbitos, o quantitativo de desabrigados e a consolidação dos prejuízos públicos e privados para cada município, convertendo os tipos de desastres em uma variável categórica combinada por localidade.

- **Junção (*Merge*):** As quatro bases do AdaptaBrasil foram unificadas progressivamente e, em seguida, unidas à base agregada do Atlas por meio de uma operação de junção à esquerda (*Left Join*), tendo como chave primária o Código IBGE do município. Municípios presentes no AdaptaBrasil, mas sem registros no Atlas, tiveram seus contadores de desastres preenchidos com zero através do comando `fillna(0)`,

sinalizando a ausência de eventos documentados no período histórico. Após a integração das bases, os 8 municípios para os quais não havia valores disponíveis nos quatro índices do AdaptaBrasil foram excluídos da etapa de clusterização por K-Means, resultando em um universo de 5.614 municípios para essa análise (frente aos 5.622 municípios com registro histórico de desastres no Atlas).

4.2.3 Modelagem e Mineração de Dados

A biblioteca `scikit-learn` foi empregada para a etapa de mineração e descoberta de conhecimento:

- **Normalização:** Devido à diferença de escala entre os índices (risco hídrico e energético variando de 0 a 1; saúde e biodiversidade de 0 a 100), aplicou-se a técnica de normalização *Min-Max Scaler*, convertendo todos os atributos para o intervalo unificado de [1]. A utilização dessa técnica justifica-se pela necessidade algorítmica de impedir que as variáveis com maiores amplitudes numéricas dominem de forma espúria os cálculos geométricos de distância na etapa de agrupamento.

- **Análise Estatística e Correlação:** Para investigar se a quantidade acumulada de eventos hídricos históricos se relaciona linearmente com o índice de risco de arboviroses, aplicou-se a Correlação de Pearson e a Regressão Linear Simples (HAN; PEI; KAMBER, 2011) para cada tipo de evento. A escolha pela Correlação de Pearson, em detrimento de métricas alternativas como o coeficiente de correlação de Spearman, justifica-se pela natureza contínua dos dados e pelo objetivo analítico da pesquisa. Enquanto o algoritmo de Spearman atua sobre o ranqueamento ou a ordem das variáveis para identificar relações monótonas, a Correlação de Pearson é a métrica paramétrica ideal para atestar a dependência estritamente linear, fornecendo a direção e a magnitude exata dessa proporcionalidade. Segundo Han, Pei e Kamber (2011), o coeficiente de Pearson mede a força da relação em uma escala de -1 a $+1$: valores positivos indicam que ambas crescem juntas, valores negativos indicam relação inversa, e valores próximos de zero indicam independência estatística. É fundamental ressaltar, contudo, que a existência de correlação estatística não implica

causalidade direta, visto que as variáveis podem ser influenciadas por fatores subjacentes (WITTEN; FRANK; HALL, 2011). Paralelamente, a escolha da Regressão Linear Simples frente a modelos de regressão não lineares complexos justifica-se para evitar o sobreajuste aos dados (*overfitting*) e permitir uma interpretação clara do impacto. A Regressão Linear modela a relação pela equação $y = wx + b$, em que w representa a inclinação da reta e b a intersecção no eixo y, com os coeficientes obtidos pelo método dos mínimos quadrados.

- **Clusterização (K-Means):** Para o agrupamento multivariado dos municípios, utilizou-se o algoritmo K-Means. O uso de um algoritmo de agrupamento justifica-se pela insuficiência das análises univariadas frente a um problema socioambiental multifacetado, sendo necessário cruzar os quatro índices de risco de forma simultânea. O K-Means foi eleito especificamente devido à sua alta eficiência computacional e capacidade de gerar partições mutuamente exclusivas, características que o tornam superior a alternativas como o K-Medoids ou agrupamentos hierárquicos para mapear o risco em milhares de municípios de forma escalável. A definição do número ótimo de *clusters* (k) foi avaliada tecnicamente pela combinação do Método do Cotovelo (Análise da Inércia) e do Coeficiente de *Silhouette*. Adotou-se $k = 5$ para manter a aderência analítica às cinco classes de risco da metodologia original do AdaptaBrasil.

- **Redução de Dimensionalidade (PCA):** Para a visualização bidimensional da separação dos clusters, aplicou-se o algoritmo PCA (*Principal Component Analysis*). A escolha deste algoritmo justifica-se pela sua função como ferramenta de auditoria visual e geométrica da clusterização. Como o agrupamento ocorreu em um plano de quatro dimensões, tornando-se impossível de ser plotado em gráficos convencionais, o PCA resolveu o problema ao comprimir o espaço vetorial em duas novas coordenadas bidimensionais ortogonais, garantindo a preservação da máxima variância original dos dados para atestar a qualidade matemática da separação dos grupos.

- **Cálculo do Risco Composto:** Foi desenvolvida uma lógica de pontuação algorítmica para mapear a sobreposição de vulnerabilidades. Atribuiu-se pesos ba-

seados nas classes de risco, e municípios com três ou quatro dimensões ranqueadas simultaneamente como "Alto"ou "Muito alto"foram isolados em uma categoria crítica.

As bibliotecas `matplotlib` e `seaborn` foram utilizadas para a renderização estática dos resultados gerados pelo *pipeline* de dados, incluindo matrizes de correlação, *boxplots*, *heatmaps* de centroides e gráficos de barras empilhadas que compõem as visualizações integradas do estudo.

4.2.4 Classificação do Perfil Hídrico Histórico

Para estruturar a análise, foi desenvolvida uma classificação própria de perfil hídrico histórico para cada município, elaborada a partir dos registros de tipologia de desastres do Atlas Digital (BRASIL, 2022) entre 1991 e 2024. Os municípios foram classificados com base na presença ou ausência de registros das categorias de estiagem/seca e de chuvas intensas/inundações/enxurradas, resultando em quatro perfis mutuamente exclusivos:

- **Sem Evento Hídrico:** municípios sem registro de nenhum evento de seca, estiagem, inundação, enxurrada ou chuva intensa, totalizando 480 municípios (8,5%).
- **Apenas Seca/Estiagem:** com histórico exclusivo de escassez hídrica, correspondendo a 438 municípios (7,8%).
- **Apenas Chuva/Inundação:** com histórico exclusivo de excesso hídrico, englobando 1.435 municípios (25,5%).
- **Seca + Chuva/Inundação:** com histórico combinado de ambos os tipos, representando 3.269 municípios (58,1%).

A predominância do perfil combinado (58,1% dos municípios) evidencia que a maioria dos municípios brasileiros convive historicamente com ambas as formas de extremo hídrico, o que reforça a complexidade da vulnerabilidade climática no território nacional e justifica a necessidade de análises multivariadas.

5 Resultados

Este capítulo apresenta os resultados obtidos na etapa de Análise Exploratória e de Mineração de Dados. As análises foram realizadas sobre um conjunto de 5.622 municípios brasileiros, originados do cruzamento entre os índices de risco do AdaptaBrasil e o histórico de desastres registrado no Atlas Digital (1991 a 2024). O foco desta seção recai sobre a investigação da hipótese de que municípios com maior histórico de eventos hídricos extremos apresentam maior índice de risco para a ocorrência de arboviroses (dengue, zika e chikungunya).

5.1 Análise de Correlação entre Eventos Hídricos e Risco de Arboviroses

Com base nas métricas definidas na metodologia, os resultados da aplicação da Correlação de Pearson e da Regressão Linear Simples para avaliar a relação entre a quantidade acumulada de eventos hídricos e o índice de risco de arboviroses estão demonstrados na Tabela 5.1 e na Figura 5.1.

Tabela 5.1: Correlação de Pearson entre número de eventos hídricos históricos e índice de risco de arboviroses.

Tipo de Evento	r **de Pearson**	p**-valor**	**Inclinação** (w)	n
Seca / Estiagem	+0,264	$< 0{,}001$	+0,952	3.707
Inundação / Enxurrada	−0,134	$< 0{,}001$	−0,840	4.378
Chuvas Intensas	−0,101	$< 0{,}001$	−0,756	2.891

Os resultados revelam padrões distintos segundo o tipo de evento. Para eventos de **seca e estiagem**, observa-se uma correlação positiva ($r = +0{,}264$, $p < 0{,}001$): municípios com maior histórico acumulado desse tipo de evento tendem a apresentar índices mais elevados de risco de arboviroses. A inclinação da reta de regressão ($w = +0{,}952$) indica que, a cada evento adicional de seca registrado, o índice de risco de arboviroses aumenta, em média, aproximadamente 0,95 ponto. Embora de magnitude moderada, essa correlação

é estatisticamente robusta, calculada sobre uma amostra de 3.707 municípios. Em sentido oposto, eventos de **inundação e enxurrada** apresentam correlação negativa ($r = -0{,}134$, $p < 0{,}001$), com inclinação $w = -0{,}840$, indicando que municípios com mais registros desse tipo de evento tendem a ter índices de risco de arboviroses ligeiramente menores. Resultado similar é observado para **chuvas intensas** ($r = -0{,}101$, $p < 0{,}001$, $w = -0{,}756$). Embora as correlações negativas sejam de magnitude fraca, sua significância estatística sobre amostras de larga escala (4.378 e 2.891 municípios, respectivamente) as torna relevantes para a análise.

Esses padrões matemáticos podem ser visualizados de forma mais clara nos diagramas de dispersão a seguir (Figura 5.1).

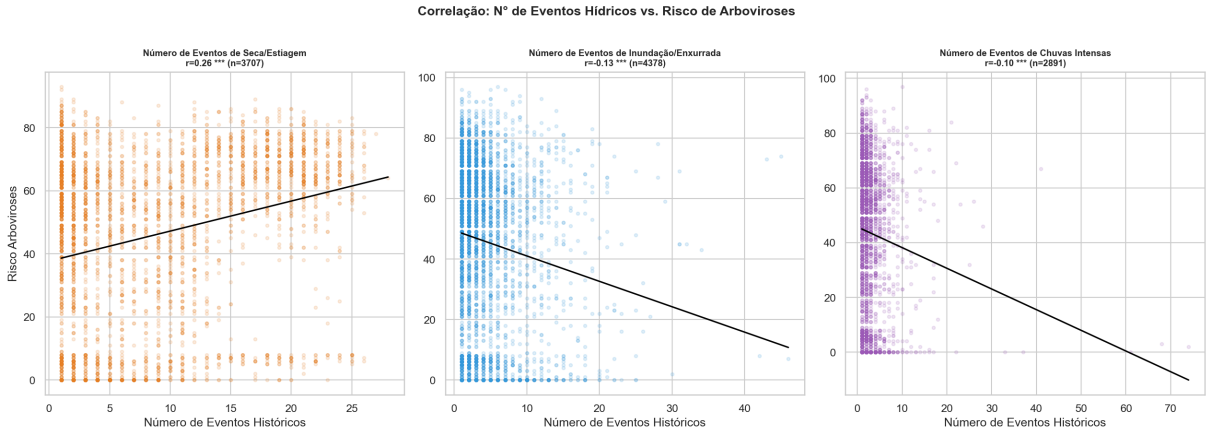

Figura 5.1: Diagramas de dispersão e retas de regressão entre o número histórico de eventos hídricos e o índice de risco de arboviroses, por tipo de evento (1991–2024). Cada ponto representa um município.

No primeiro gráfico, referente aos eventos de **Seca e Estiagem**, o eixo horizontal (X) quantifica o número acumulado de desastres de escassez hídrica por município no período histórico, variando de 0 a 25 ocorrências, enquanto o eixo vertical (Y) expressa o valor do índice contínuo de risco para a ocorrência de arboviroses do AdaptaBrasil, em uma escala de 0 a 100. A dispersão dos pontos revela que, à medida que nos deslocamos para a direita no eixo X (aumento do número de secas registradas), há uma tendência visível de elevação nas notas do índice de risco no eixo Y. É possível observar que, em municípios com registro de cerca de 20 eventos históricos de seca, o modelo projeta valores de risco de arboviroses concentrados em patamares elevados, situando-se frequentemente acima de 60

pontos. A reta de regressão ascendente, com inclinação positiva (w = +0,952), traduz esse comportamento, demonstrando que cada evento histórico adicional de seca está associado a um incremento médio de quase um ponto no índice de risco epidemiológico projetado.

Em contrapartida, o segundo gráfico detalha a relação para os eventos de **Inundação e Enxurrada**, no qual o eixo horizontal representa a frequência de ocorrências (de 0 a cerca de 45 registros) e o vertical exibe o risco de arboviroses. A análise da dispersão revela que o maior adensamento de municípios com altos índices de risco sanitário concentra-se na faixa de baixa frequência de inundações (canto esquerdo do gráfico, com menos de 10 eventos). À medida que avançamos para a direita no eixo X, em direção a municípios com histórico severo de inundações (acima de 30 ocorrências), os pontos de dispersão tendem a se concentrar em faixas de menor vulnerabilidade, situando-se abaixo de 20 pontos de risco no eixo Y. A reta de regressão descendente, com inclinação negativa (w = -0,840), corrobora visualmente essa tendência inversa entre a frequência histórica do desastre hídrico e o risco epidemiológico futuro.

O terceiro gráfico, que analisa as ocorrências de **Chuvas Intensas**, apresenta comportamento análogo às inundações. O eixo horizontal exibe o volume de eventos severos acumulados por município, estendendo-se por uma escala maior, que ultrapassa 70 registros históricos, enquanto o eixo vertical mantém o índice de risco de arboviroses. A dispersão exibe que municípios com frequência extrema de chuvas severas (acima de 50 ocorrências no eixo X) apresentam-se concentrados estritamente na base inferior do eixo vertical Y, com risco sanitário muito próximo a zero. A reta de regressão exibe um declínio suave (w = -0,756), demonstrando de forma coesa que a recorrência sistemática de eventos pluviais severos está associada a uma redução progressiva no índice de risco de arboviroses de longo prazo.

Do ponto de vista socioambiental e epidemiológico, esses comportamentos geométricos observados nos eixos possuem forte respaldo na literatura científica recente. A tendência ascendente observada no gráfico de seca e estiagem corrobora as conclusões de Cox et al. (2025) e Lowe et al. (2021), os quais apontam que períodos prolongados de seca induzem a práticas de armazenamento inseguro e improvisado de água por parte da população. Esse acúmulo de água em recipientes domésticos artificiais atua como criadouro ideal para a

oviposição e reprodução do vetor *Aedes aegypti*, elevando o risco estrutural de transmissão. Por outro lado, o declínio do risco associado a altos índices de inundações e chuvas intensas nos gráficos B e C é consistente com o efeito de lavagem (*flushing*), descrito por Benedum et al. (2018). Episódios de precipitações volumosas e enxurradas contínuas provocam a lavagem mecânica dos criadouros de água parada, eliminando larvas e pupas antes que atinjam a fase adulta, o que reduz temporariamente a população do vetor.

5.2 Distribuição do Risco de Arboviroses por Perfil Hídrico

A Tabela 5.2 apresenta as estatísticas descritivas do índice de risco de arboviroses por perfil hídrico histórico.

Tabela 5.2: Estatísticas descritivas do índice de risco de arboviroses por perfil hídrico histórico.

Perfil Hídrico	***n***	**Média**	**Mediana**	**Desvio Padrão**
Sem Evento Hídrico	480	55,35	62,0	25,12
Apenas Seca/Estiagem	438	52,82	61,0	24,71
Apenas Chuva/Inundação	1.435	47,77	53,0	26,05
Seca + Chuva/Inundação	3.269	45,09	54,0	26,73

Como observado na Tabela 5.2, um resultado notável emerge: os municípios **sem nenhum evento hídrico registrado** apresentam a maior média do índice de risco de arboviroses (55,35) e também a maior proporção de municípios classificados nas categorias "Alto" e "Muito alto" (60,2%). Em sequência, o perfil de seca exclusiva apresenta média de 52,82 e 58,2% dos municípios nessas categorias mais críticas. Os perfis com histórico de chuva ou inundação apresentam médias inferiores (47,77 e 45,09, respectivamente) e proporções menores nas categorias de maior risco (43,3% e 44,1%).

A Figura 5.2 complementa essa leitura ao detalhar a proporção exata de municípios em cada classe de gravidade.

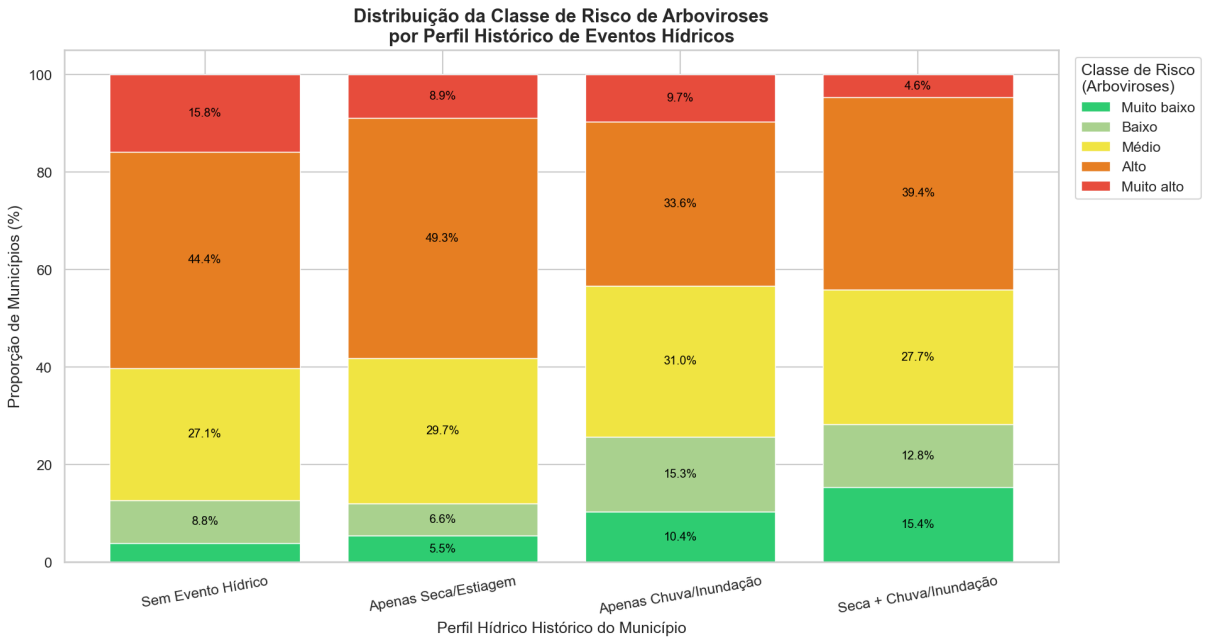

Figura 5.2: Distribuição proporcional da classe de risco de arboviroses (dengue, zika e chikungunya) por perfil hídrico histórico do município. Cada barra representa 100% dos municípios daquele perfil, segmentados por classe de risco do AdaptaBrasil (2020).

Esse padrão, inicialmente contraintuitivo, é consistente com os resultados da correlação de Pearson, que apontou relação positiva para secas e negativa para inundações com o risco de arboviroses.

Uma interpretação plausível reside na natureza do índice de risco do AdaptaBrasil, que incorpora variáveis estruturais de longo prazo, como densidade populacional, cobertura de saneamento e condições climáticas médias, e não apenas a exposição episódica a eventos extremos. Municípios de perfil urbano e costeiro, concentrados em regiões quentes e com maior densidade demográfica, tendem a apresentar alto risco de arboviroses independentemente de seu histórico hídrico.

Diante dessa característica metodológica, cabe ressaltar que o índice de risco do AdaptaBrasil difere fundamentalmente de indicadores operacionais de vigilância entomológica, como o LIRAa (Levantamento Rápido de Índices para *Aedes aegypti*). Enquanto o LIRAa constitui um método de campo, observacional e pontual, voltado à rápida obtenção de indicadores de infestação (como o percentual de imóveis com larvas) para direcionar o controle imediato do vetor (BRASIL, 2024), o índice do AdaptaBrasil avalia a suscetibilidade estrutural e climática a longo prazo. Dessa forma, as análises realizadas

neste estudo não possuem relação direta de composição com os levantamentos de campo do LIRAa, concentrando-se exclusivamente nas projeções de risco climático e vulnerabilidade socioambiental.

5.3 Casos Ilustrativos: Municípios com Maior Histórico Hídrico

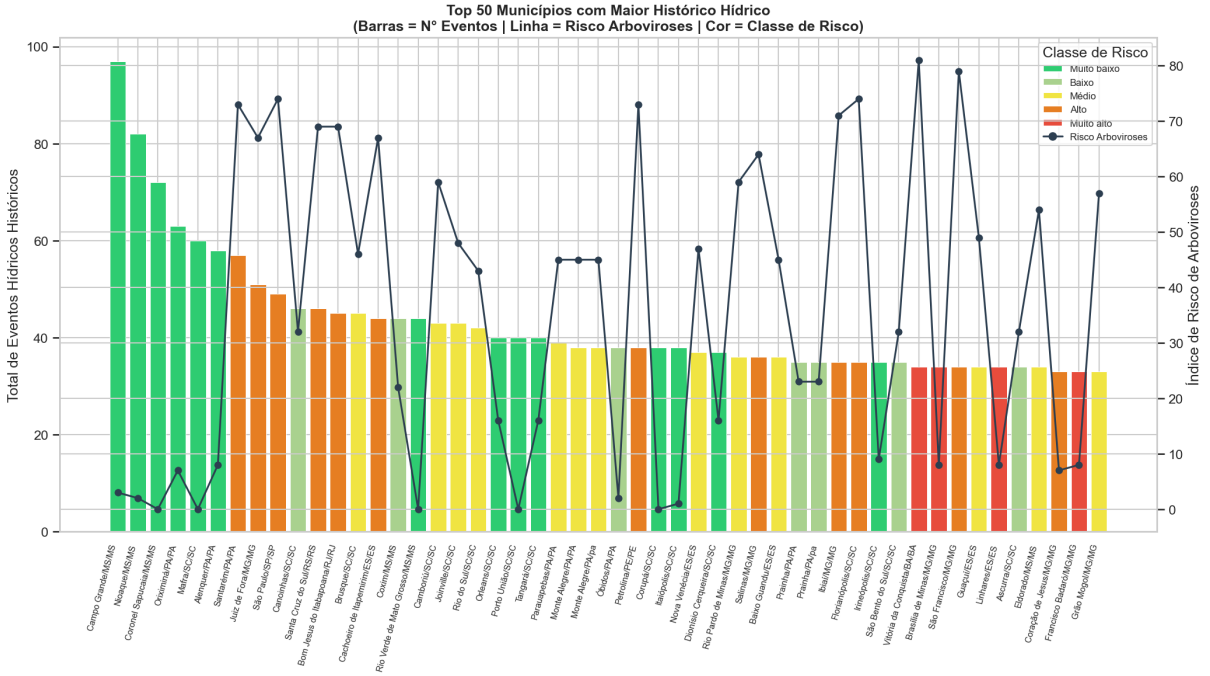

Figura 5.3: Top 50 municípios brasileiros com maior número acumulado de eventos hídricos históricos (1991–2024). As barras indicam o total de eventos e são coloridas pela classe de risco de arboviroses (AdaptaBrasil, 2020). A linha representa o valor contínuo do índice de risco de arboviroses.

A análise do *ranking* dos 50 municípios com maior volume acumulado de eventos hídricos históricos, apresentada na Figura 5.3, reforça a alta heterogeneidade territorial e a complexidade dos padrões encontrados. Para facilitar a compreensão visual desse comportamento pelo leitor, o gráfico foi estruturado sob um sistema de dois eixos ordenados. O eixo horizontal (X) organiza de forma decrescente os municípios segundo o total de desastres hídricos (estiagem, seca, chuvas intensas, enxurradas e inundações) registrados na base consolidada do Atlas Digital (BRASIL, 2022) no período de 1991 a 2024. O eixo vertical esquerdo expressa a quantidade absoluta desses desastres por meio da altura das barras, as quais são coloridas de acordo com a classe de risco categórica para arboviroses

do AdaptaBrasil (Muito Baixo, Baixo, Médio, Alto, Muito Alto) (INPE, 2025). Por fim, o eixo vertical direito mensura o valor contínuo do índice de risco de arboviroses (escala de 0 a 100), representado graficamente pela linha escura com marcadores circulares sobreposta às barras.

A leitura detalhada dessa dispersão revela um forte desacoplamento entre o volume de eventos severos acumulados no passado e a projeção de vulnerabilidade sanitária futura. O município de **Campo Grande, MS**, por exemplo, lidera o ranking nacional com 97 ocorrências hídricas históricas registradas, mas apresenta um índice contínuo de risco de arboviroses extremamente baixo, situando-se próximo a zero na escala do eixo vertical direito, o que o classifica na categoria Muito Baixo. Em contrapartida, ao percorrermos o gráfico em direção à direita, identificam-se localidades que, mesmo com menor volume de desastres hídricos acumulados, exibem uma suscetibilidade epidemiológica substancialmente maior. É o caso de **Santarém, PA**, que ocupa a sexta posição do ranking com 57 ocorrências hídricas históricas e apresenta índice contínuo de risco de arboviroses superior a 60 pontos (classe Alto). Comportamento análogo é observado em **Juiz de Fora, MG**, situado na nona posição com 51 eventos e caracterizado por um risco sanitário igualmente severo (classe Alto).

Essa oscilação da linha contínua de risco ao longo do gráfico de barras demonstra que a quantidade de desastres hídricos não atua, isoladamente, como um preditor univariado para a ocorrência de arboviroses. Essa dissociação decorre do fato de que a vulnerabilidade a doenças transmitidas por vetores, como o mosquito *Aedes aegypti*, é um fenômeno eminentemente multifacetado e dinâmico, condicionado pela intersecção de múltiplas dimensões socioambientais (COX et al., 2025; LOWE et al., 2021). Fatores como a densidade populacional, as deficiências estruturais de saneamento básico, a coleta inadequada de resíduos sólidos e as variações microclimáticas de temperatura atuam de forma combinada no território, sobrepondo-se à mera frequência episódica de inundações ou secas.

A constatação de que modelos lineares simples são insuficientes para decifrar essa complexidade é o que motiva diretamente a aplicação de técnicas de mineração de dados mais robustas. De acordo com Han, Pei e Kamber (2011), quando o objeto de

estudo apresenta múltiplos atributos independentes que interagem de maneira não-linear, as análises univariadas ou de correlação bidimensional tornam-se limitadas, gerando visões parciais ou distorcidas da realidade. Nesses cenários, torna-se indispensável o emprego de métodos de aprendizado de máquina não supervisionado, os quais possuem a capacidade de analisar simultaneamente um espaço de alta dimensionalidade para descobrir estruturas de dados implicitamente armazenadas (DONOHO, 2017).

Conforme validado por Wu et al. (2021), a aplicação de algoritmos de partição baseados em centroides é a estratégia metodológica ideal para reduzir a heterogeneidade de grandes bases de dados, agrupando os objetos em subgrupos (clusters) que compartilham perfis multivariados coesos. No contexto desta pesquisa, a transição para a análise de agrupamento multivariado, detalhada na seção a seguir, faz-se necessária para unificar os quatro índices de risco disponíveis (hídrico, de saúde, energético e de biodiversidade) ao histórico de desastres dos municípios.

5.4 Clusterização de Municípios por Perfil de Risco

Esta seção apresenta a análise de agrupamento multivariado realizada com o objetivo de segmentar os municípios brasileiros de acordo com seus perfis integrados de vulnerabilidade climática e de saúde. Para isso, aplicou-se o algoritmo de aprendizado não supervisionado K-Means sobre as quatro dimensões de risco disponíveis na plataforma AdaptaBrasil MCTI, integrando-as ao histórico de desastres registrado no Atlas Digital.

5.4.1 Pré-processamento e Seleção do Número de Clusters

Conforme estabelecido na metodologia, os quatro índices de risco utilizados como variáveis de entrada passaram pelo processo de normalização para evitar distorções de escala. A seleção do número de agrupamentos foi orientada pela análise conjunta da inércia e do Coeficiente de Silhouette, conforme apresentado na Figura 5.4.

Embora o Coeficiente de Silhouette aponte a partição com dois agrupamentos ($k = 2$) como a ideal sob a perspectiva estritamente matemática, essa divisão revela-se excessivamente genérica para a compreensão da vulnerabilidade socioambiental. A escolha

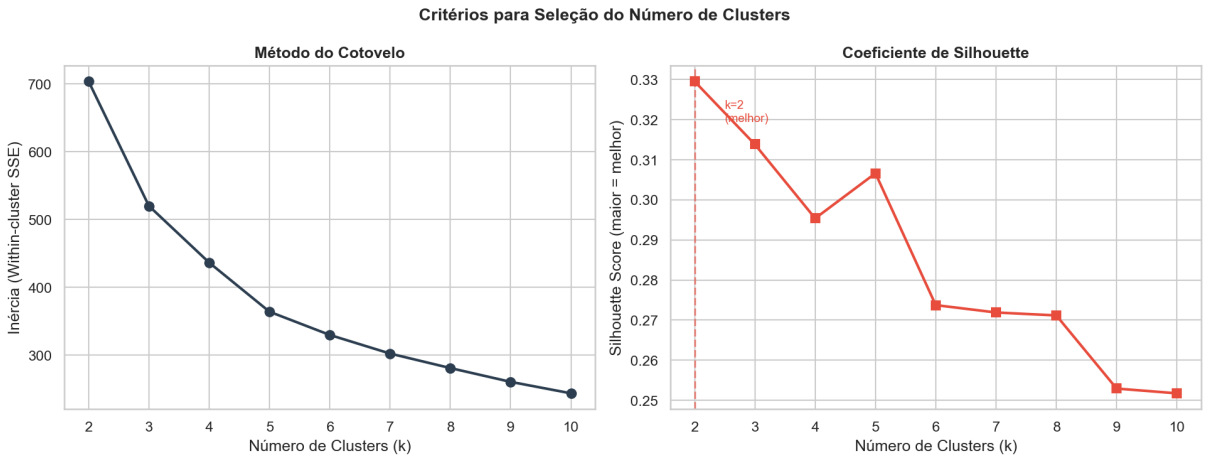

Figura 5.4: Método do cotovelo (inércia) e Coeficiente de Silhouette para valores de k entre 2 e 10. A linha tracejada indica o k ótimo estatístico ($k = 2$); o valor adotado foi $k = 5$ por razões interpretativas.

de cinco clusters ($k = 5$) justifica-se pela necessidade de capturar nuances regionais e perfis específicos de risco composto, permitindo uma separação mais rica e acionável dos municípios sem comprometer a estabilidade matemática do modelo, visto que a curva do cotovelo exibe uma inflexão clara (suavização do decréscimo da inércia) a partir desse ponto.

5.4.2 Perfil dos Clusters

A Tabela 5.3 detalha os valores médios originais de cada cluster, mas, de forma isolada, ela apresenta limitações para a comparação direta entre as dimensões. Como os índices originais possuem escalas distintas, em que o risco hídrico e o energético variam de 0 a 1, enquanto o risco de saúde e o de biodiversidade variam de 0 a 100, os valores numéricos brutos tornam-se matematicamente incomparáveis em termos de magnitude absoluta.

Para superar essa barreira e revelar a real contribuição de cada atributo na composição dos grupos, optou-se pela aplicação da técnica de normalização Min-Max, convertendo todos os índices para o intervalo unificado de para a renderização do *heatmap* dos centroides apresentado na Figura 5.5 (HAN; PEI; KAMBER, 2011). De acordo com a literatura de mineração de dados, a normalização é uma etapa indispensável em algoritmos baseados em distâncias, como o K-Means, pois garante que variáveis com maiores intervalos numéricos originais não dominem de forma espúria os cálculos de similaridade e a visualização dos resultados (WU et al., 2021; DONOHO, 2017).

O *heatmap* traz contribuições analíticas essenciais que não podem ser extraídas diretamente da tabela de médias absolutas. Em primeiro lugar, ele ativa a percepção rápida do leitor por meio de um sistema de cores quentes e frias, permitindo apreender instantaneamente o gradiente de gravidade de cada agrupamento e a interposição de riscos (MARCONI; LAKATOS, 2019). Em segundo lugar, e mais importante, a escala normalizada expressa a posição relativa de cada cluster em relação aos limites de pior situação (valor próximo a 1, em vermelho) e melhor situação (valor próximo a 0, em amarelo-claro) observados no território nacional.

Na Tabela 5.3, por exemplo, a média original de risco de saúde do Cluster E (65,9) e do Cluster D (64,3) parece muito próxima. O *heatmap*, contudo, evidencia que na escala de variação real do país, o Cluster E atinge um patamar de risco hídrico relativo muito mais acentuado (0,75) do que o seu risco à saúde (0,68). Isso revela ao leitor que a pressão sobre os recursos hídricos é a força motriz mais severa dentro do grupo de alta vulnerabilidade, superando a vulnerabilidade sanitária isolada. Essa decodificação visual facilita a identificação imediata de padrões de sobreposição e dos fatores influenciadores do risco composto, reduzindo o esforço cognitivo exaustivo necessário para analisar tabelas numéricas multidimensionais.

Tabela 5.3: Perfil médio dos índices de risco por cluster (K-Means, $k = 5$). Valores dos índices em escala original.

Cluster	***n***	**R. Hídrico**	**R. Saúde**	**R. Energia**	**R. Bio**	**Med. Eventos**
A – Risco Baixo	965	0,394	12,6	0,400	11,8	14
B – Risco Moderado-Baixo	735	0,463	11,4	0,382	53,5	8
C – Risco Moderado	1.166	0,336	56,0	0,337	41,7	6
D – Risco Moderado-Alto	1.130	0,560	64,3	0,396	9,7	16
E – Risco Alto	1.618	0,620	65,9	0,412	59,3	10

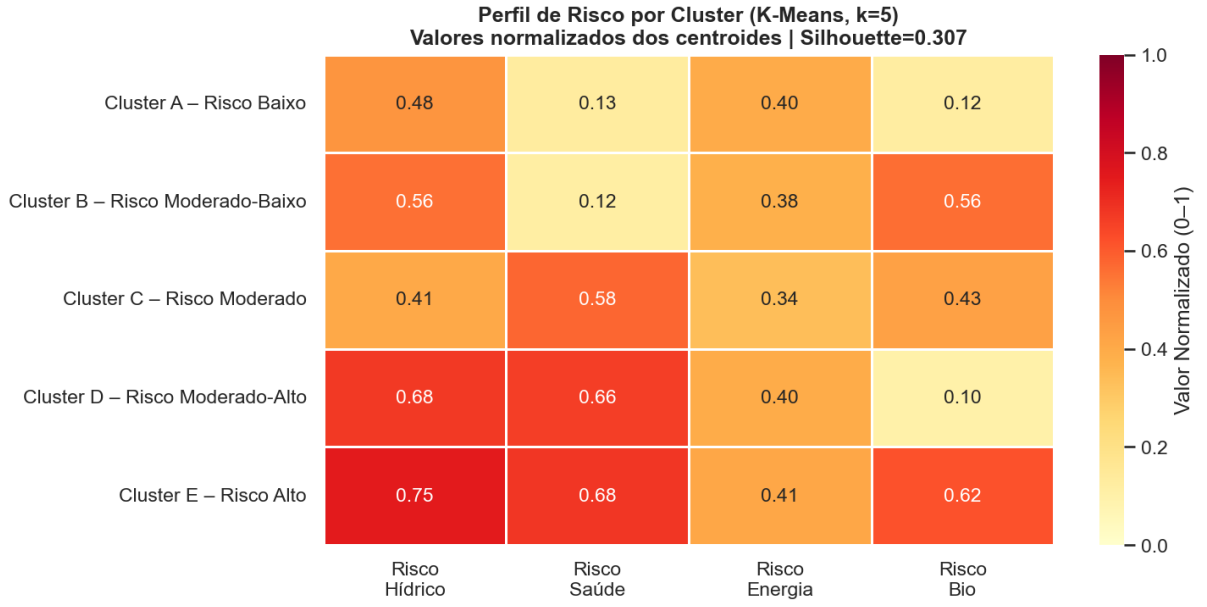

Figura 5.5: *Heatmap* dos centroides normalizados por cluster. Valores mais próximos de 1 (vermelho) indicam maior intensidade do índice de risco na dimensão correspondente.

A análise dos centroides e da distribuição de classes de risco por cluster revela perfis de vulnerabilidade qualitativamente distintos, descritos a seguir:

O **Cluster A – Risco Baixo** (965 municípios, 17,2%) apresenta os menores valores médios em todas as dimensões, com risco de saúde médio de apenas 12,6 e risco de biodiversidade de 11,8. Na dimensão hídrica, 89,8% dos municípios estão nas classes "Muito baixo", "Baixo" ou "Médio", e na dimensão de saúde, 77,5% enquadram-se nas classes "Muito baixo" ou "Baixo". Trata-se do grupo de menor vulnerabilidade climática integrada, embora concentre a maior mediana de eventos históricos de desastre (14 eventos por município), sugerindo que a ocorrência passada de desastres não determina, por si só, o nível de risco projetado.

O **Cluster B – Risco Moderado-Baixo** (735 municípios, 13,1%) se distingue pelo elevado risco de biodiversidade (média 53,5), com 82,2% dos municípios nas classes "Médio", "Alto" ou "Muito alto" nessa dimensão. Em contrapartida, o risco de saúde permanece baixo (média 11,4), com 65,5% dos municípios nas classes "Muito baixo" ou "Baixo". Esse perfil é compatível com municípios localizados em biomas ameaçados, como o Cerrado e a Caatinga, mas com menor densidade urbana e, portanto, menor vulnerabilidade às arboviroses.

O **Cluster C – Risco Moderado** (1.166 municípios, 20,8%) apresenta risco de saúde de magnitude intermediária (média 56,0), com 52,0% dos municípios na classe "Médio" e 37,8% nas classes "Alto" ou "Muito alto". O risco hídrico é o mais baixo entre todos os clusters (média 0,336), com 63,5% dos municípios nas classes "Muito baixo" ou "Baixo", e o risco energético é igualmente reduzido (82,7% na classe "Baixo" ou "Muito baixo"). Esse cluster agrupa predominantemente municípios de porte médio em regiões tropicais, com vulnerabilidade moderada às arboviroses mas sem pressões hídricas ou energéticas críticas.

O **Cluster D – Risco Moderado-Alto** (1.130 municípios, 20,1%) destaca-se pela combinação de alto risco hídrico (média 0,560, com 42,5% dos municípios nas classes "Alto" ou "Muito alto") e elevado risco de saúde (média 64,3, com 67,0% nas classes "Alto" ou "Muito alto"). Em contrapartida, apresenta o menor risco de biodiversidade entre todos os clusters (média 9,7), com 80,4% dos municípios nas classes "Muito baixo" ou "Baixo". Esse perfil sugere municípios em regiões áridas ou semiáridas, onde o estresse hídrico coexiste com alta vulnerabilidade a arboviroses, possivelmente associada ao armazenamento inadequado de água durante os períodos de escassez, condição conhecida como fator de proliferação do *Aedes aegypti*.

O **Cluster E – Risco Alto** (1.618 municípios, 28,8%) representa o grupo de maior vulnerabilidade composta, com os maiores valores médios simultâneos de risco hídrico (0,620), de saúde (65,9), energético (0,412) e de biodiversidade (59,3) em escala original. Na Figura 5.5, esses valores são apresentados na escala normalizada Min-Max [0, 1], correspondendo a 0,75 (hídrico), 0,68 (saúde), 0,41 (energético) e 0,62 (biodiversidade), os maiores em todas as dimensões entre os cinco clusters.

A separação espacial e consistência desses perfis de vulnerabilidade podem ser validadas através da Análise de Componentes Principais (PCA), apresentada na Figura 5.6. Em problemas que envolvem dados de alta dimensionalidade, como as quatro dimensões de risco deste estudo, a visualização geométrica direta das partições geradas pelo algoritmo K-Means é impossível no espaço original. O PCA soluciona essa limitação ao projetar o espaço quadridimensional original em um plano bidimensional ótimo, preservando a maior parte da variância dos dados (HAN; PEI; KAMBER, 2011; WU et al., 2021).

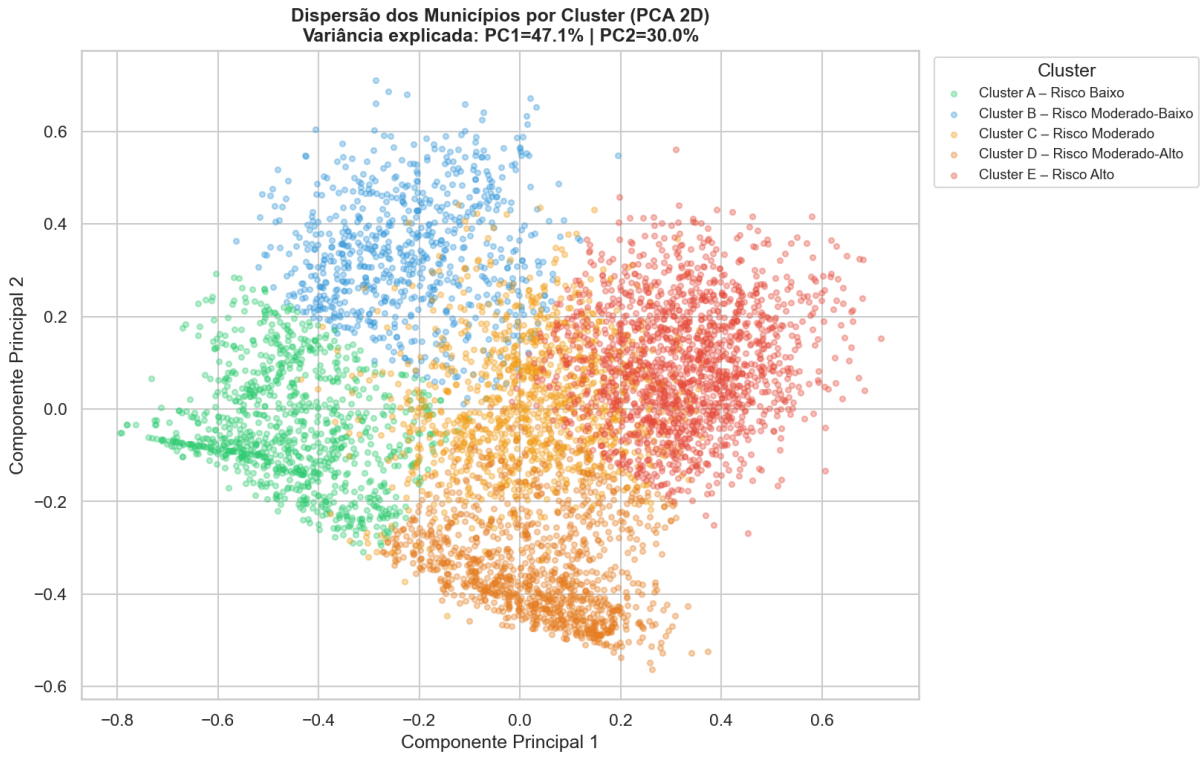

Figura 5.6: Projeção dos municípios no espaço bidimensional das duas primeiras componentes principais (PCA), coloridos por cluster. As componentes PC1 e PC2 explicam conjuntamente a maior parte da variância dos quatro índices de risco normalizados.

As duas primeiras componentes principais explicam conjuntamente 77,1% da variância total dos dados, em que a primeira componente (PC1) captura 47,1% e a segunda (PC2) retém 30,0% da variabilidade original. Essa alta taxa de explicação garante que a simplificação geométrica não resulte em perda significativa de informação estrutural. Ao analisarmos a dispersão dos municípios projetados no plano da Figura 5.6, onde cada ponto colorido representa uma localidade brasileira, é possível extrair conclusões profundas sobre a topologia e a coesão dos clusters:

O eixo horizontal, regido pela PC1, funciona como um indicador sintético do nível geral de vulnerabilidade climática acumulada no território. À medida que nos deslocamos da esquerda para a direita, do domínio negativo para o positivo no eixo X, ocorre uma transição nítida de cenários de baixo risco para patamares de ameaça crítica. O Cluster A (Risco Baixo, em verde) posiciona-se inteiramente no extremo esquerdo da projeção, concentrando-se entre -0,8 e -0,3 no eixo horizontal. Em contraposição absoluta, o Cluster E (Risco Alto, em vermelho) ocupa o quadrante direito da projeção, dispersando-se entre +0,1 e +0,7. A ausência de sobreposição espacial direta entre estes dois grupos em

vermelho e verde confirma que as suas estruturas de dados são mutuamente exclusivas e representam realidades socioambientais polares.

O eixo vertical, governado pela PC2, captura as variações e compensações existentes entre as dimensões de risco que não foram explicadas pela vulnerabilidade geral da PC1. Esse eixo é fundamental para separar agrupamentos que possuem níveis semelhantes de risco médio, mas com focos de suscetibilidade distinots. Isso fica evidente ao compararmos o Cluster B (Risco Moderado-Baixo, em azul) e o Cluster C (Risco Moderado, em laranja). Ambos compartilham faixas semelhantes de transição no eixo horizontal, mas são divididos de forma categórica pelo eixo vertical. O Cluster B projeta-se quase em sua totalidade no semiplano superior positivo da PC2 (entre 0,0 e 0,5), demonstrando que seu peso reside primariamente na dimensão ecológica. Por outro lado, o Cluster C projeta-se de forma inversa, consolidando-se em áreas onde as pressões epidemiológicas e de infraestrutura urbana são preponderantes.

Por fim, a análise espacial confirma a eficácia do K-Means na delimitação de fronteiras lineares ótimas. É possível observar uma transição contínua e geometricamente delimitada entre as classes, onde o Cluster D (Risco Moderado-Alto, em ciano) funciona como uma ponte de ligação entre as regiões de baixa e alta vulnerabilidade, concentrando os municípios com alto risco urbano mas de baixa integridade ecológica florestal. A separação coesa entre as nuvens de pontos de cada cor valida a robustez metodológica do pipeline analítico, assegurando que as classificações subsequentes de risco composto sejam matematicamente fundamentadas.

5.5 Mapeamento de Riscos Compostos

As análises anteriores trataram cada índice de risco de forma independente. Contudo, buscando analisar a interposição de riscos distintos, foi realizada uma nova camada de investigação focada na sobreposição de vulnerabilidades. Municípios que acumulam simultaneamente altos índices de risco hídrico, de saúde, energético e de biodiversidade representam os casos de maior complexidade e, portanto, de maior prioridade para a gestão técnica de desastres. Para capturar essa sobreposição de forma estruturada, foi desenvolvida, pelo autor, uma métrica e uma categorização própria de **risco composto**.

Essa nova dimensão analítica foi formulada pelo autor a partir do cruzamento das classes originais fornecidas pela plataforma AdaptaBrasil (INPE, 2025) e dos resultados quantitativos obtidos na etapa de clusterização.

O risco composto proposto nesta pesquisa é operacionalizado em duas etapas. Na primeira, cada uma das quatro dimensões de risco de um município é verificada quanto à sua classe originária: uma dimensão é considerada **crítica** quando sua classificação no AdaptaBrasil é "Alto" ou "Muito alto". Na segunda etapa, o número de dimensões críticas do município, variando de 0 a 4, determina a nova categoria de risco composto atribuída pelo modelo. Adicionalmente, para diferenciar municípios sem nenhuma dimensão crítica, foi estabelecido o cálculo de um **score total**. Esse *score* é obtido pela soma dos valores numéricos atribuídos às classes de risco originais ("Muito baixo" = 0, "Baixo" = 1, "Médio" = 2, "Alto" = 3, "Muito alto" = 4), resultando em uma escala que varia de 0 a 16. No conjunto analisado, o *score* médio foi de 7,13 (desvio padrão 2,33), com mínimo de 1 e máximo de 13, o que indica ampla dispersão na vulnerabilidade entre os municípios brasileiros. As cinco categorias finais estabelecidas por este estudo e seus respectivos critérios de classificação são apresentadas na Tabela 5.4:

- **Baixa Vulnerabilidade:** nenhuma dimensão em "Alto" ou "Muito alto" e *score* total inferior a 6. Municípios com perfil de risco predominantemente baixo em todas as dimensões.

- **Moderado:** nenhuma dimensão crítica, mas *score* total igual ou superior a 6, indicando acúmulo de riscos intermediários em múltiplas dimensões sem que nenhuma delas atinja nível crítico isoladamente.

- **Vulnerável (1 dimensão):** exatamente uma dimensão em "Alto" ou "Muito alto". Municípios com vulnerabilidade setorial específica, que requerem atenção direcionada em um eixo de risco.

- **Muito Vulnerável (2 dimensões):** duas dimensões simultaneamente em "Alto" ou "Muito alto", configurando sobreposição de vulnerabilidades que amplifica o risco total do território.

- **Crítico (3 a 4 dimensões):** três ou quatro dimensões simultaneamente em "Alto" ou "Muito alto". Trata-se dos municípios de maior prioridade, onde a convergência de múltiplas vulnerabilidades climáticas representa o cenário de maior complexidade para a gestão de riscos. Dos 456 municípios nessa categoria, 98,9% apresentam risco hídrico crítico, 98,5% risco de saúde crítico e 95,4% risco de biodiversidade crítico, enquanto o risco energético crítico está presente em apenas 8,8% dos casos, indicando que a criticidade composta é determinada predominantemente pela tríade hídrico-saúde-biodiversidade.

Tabela 5.4: Critérios de classificação do risco composto por município.

Categoria	**Critério**	***n* (%)**
Baixa Vulnerabilidade	Nenhuma dimensão em Alto/Muito alto e score total < 6	1.243 (22,1%)
Moderado	Nenhuma dimensão crítica e score total ≥ 6	824 (14,7%)
Vulnerável (1 dimensão)	Exatamente 1 dimensão em Alto ou Muito alto	1.844 (32,8%)
Muito Vulnerável (2 dimensões)	Exatamente 2 dimensões em Alto ou Muito alto	1.247 (22,2%)
Crítico (3–4 dimensões)	3 ou 4 dimensões em Alto ou Muito alto	456 (8,1%)

5.5.1 Distribuição Geral e Regional

A Figura 5.7 mostra a distribuição dos municípios pelas categorias de risco composto. Embora a maioria dos municípios (32,8%) se enquadre na categoria "Vulnerável", com ao menos uma dimensão crítica, chama atenção o conjunto de **456 municípios (8,1%) classificados como Críticos**, com três ou quatro dimensões simultaneamente em "Alto" ou "Muito alto".

A distribuição regional dos riscos compostos, apresentada na Figura 5.8 e detalhada na Tabela 5.5, evidencia disparidades marcantes entre as regiões brasileiras. O **Nordeste** concentra a maior proporção de municípios críticos (18,8%) e muito vulneráveis (44,9%), totalizando 63,7% dos seus municípios nas duas categorias mais graves. Em contraste, as regiões **Sul** (0,9% de críticos) e **Norte** (4,4%) apresentam perfis de vulnerabi-

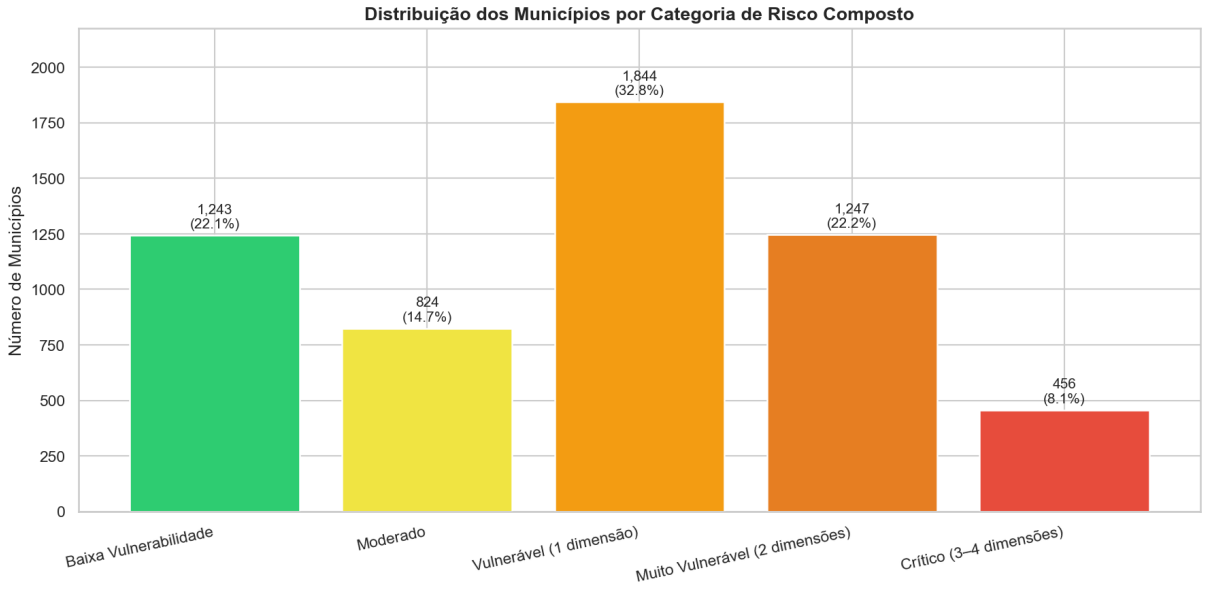

Figura 5.7: Distribuição dos municípios brasileiros por categoria de risco composto, considerando a sobreposição de classes "Alto" e "Muito alto" nos quatro índices do Adapta-Brasil.

lidade composta significativamente menores, com o Sul tendo 51,9% dos seus municípios na categoria de baixa vulnerabilidade.

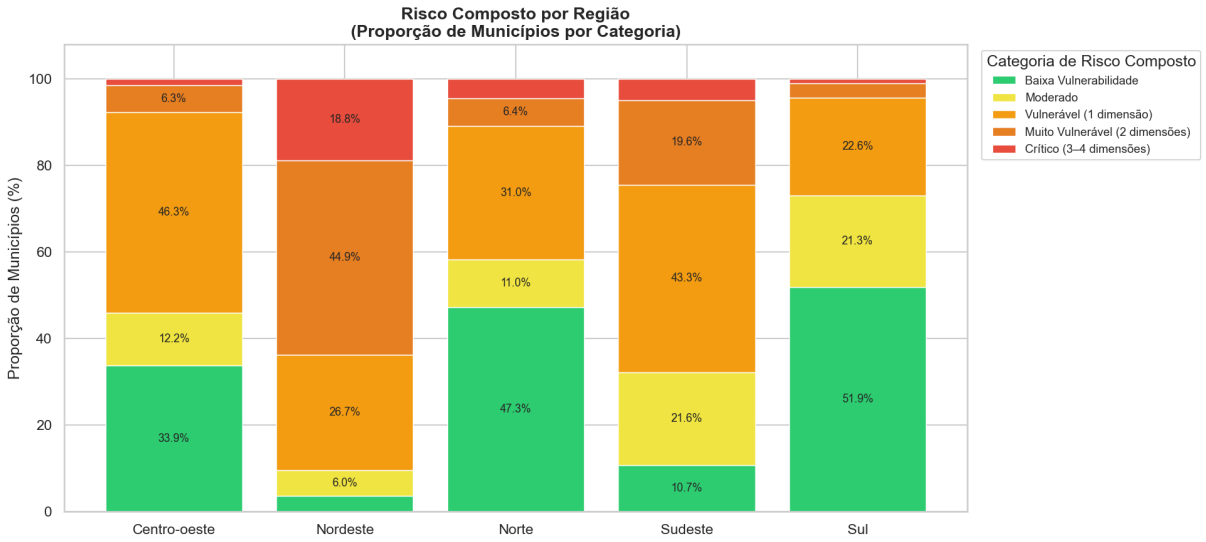

Figura 5.8: Distribuição proporcional das categorias de risco composto por região brasileira. Cada barra representa 100% dos municípios da respectiva região.

5.5.2 Municípios Críticos e Relação com os Clusters

Dos 456 municípios críticos, os cinco estados com maior concentração são **Bahia** (71 municípios), **Pernambuco** (62), **Sergipe** (39), **Ceará** (38) e **Alagoas** (36), todos per-

Tabela 5.5: Proporção de municípios por categoria de risco composto e região (%).

Região	Baixa Vuln.	Moderado	Vulnerável	Muito Vuln.	Crítico
Centro-Oeste	33,9%	12,2%	46,3%	6,3%	1,3%
Nordeste	3,7%	6,0%	26,7%	44,9%	18,8%
Norte	47,3%	11,0%	31,0%	6,4%	4,4%
Sudeste	10,7%	21,6%	43,3%	19,6%	4,8%
Sul	51,9%	21,3%	22,6%	3,3%	0,9%

tencentes ao Nordeste. São Paulo (34) e Paraíba (34) também figuram entre os estados com maior número de municípios críticos.

A Figura 5.9 cruza as categorias de risco composto com os agrupamentos identificados pelo algoritmo K-Means, revelando uma correspondência entre as duas abordagens analíticas. Esse cruzamento é fundamental para validar a coerência interna do modelo, pois demonstra que a clusterização não supervisionada conseguiu capturar organicamente o gradiente de gravidade formulado pela métrica categórica de risco composto.

A leitura detalhada do gráfico evidencia uma progressão clara da sobreposição de vulnerabilidades. No **Cluster A (Risco Baixo)**, a esmagadora maioria dos municípios (70,8%) enquadra-se em baixa vulnerabilidade composta, sendo a presença de municípios críticos estatisticamente residual. À medida que a análise avança para os perfis de transição, a composição de risco se altera. O **Cluster C (Risco Moderado)**, por exemplo, concentra sua maior fatia na categoria Vulnerável (42,2%), refletindo o seu perfil de possuir desafios setoriais específicos (como o risco epidemiológico moderado-alto), mas sem a sobreposição generalizada de ameaças severas em outras dimensões hídricas ou de infraestrutura.

A situação agrava-se progressivamente no **Cluster D (Risco Moderado-Alto)**, onde o acúmulo das pressões hídrica e de saúde empurra a distribuição em direção ao topo da escala. Nesse grupo, 40,2% dos municípios concentram-se na categoria Vulnerável e 32,9% na categoria Muito Vulnerável. A contenção da classe Crítica (apenas 6,0%) neste agrupamento explica-se diretamente pela preservação do índice de integridade da biodiversidade, que impede que os municípios atinjam a sobreposição máxima exigida pela métrica.

Por fim, o ápice do risco sistêmico consolida-se no **Cluster E (Risco Alto)**. Neste agrupamento, a categoria de Baixa Vulnerabilidade inexiste, e a distribuição é amplamente dominada pelas fatias superiores da escala: 42,9% dos municípios são classificados como Muito Vulneráveis e 20,8% como Críticos. Juntas, essas duas categorias de gravidade extrema englobam 63,7% de todo o agrupamento. Essa convergência matemática comprova que as altas médias do Cluster E não resultam de anomalias isoladas, mas da colisão consistente de vulnerabilidades hídricas, sanitárias e ecológicas que ocorrem simultaneamente no mesmo espaço geográfico, sinalizando os territórios de mais alta prioridade técnica para o direcionamento de medidas operacionais de contenção de desastres.

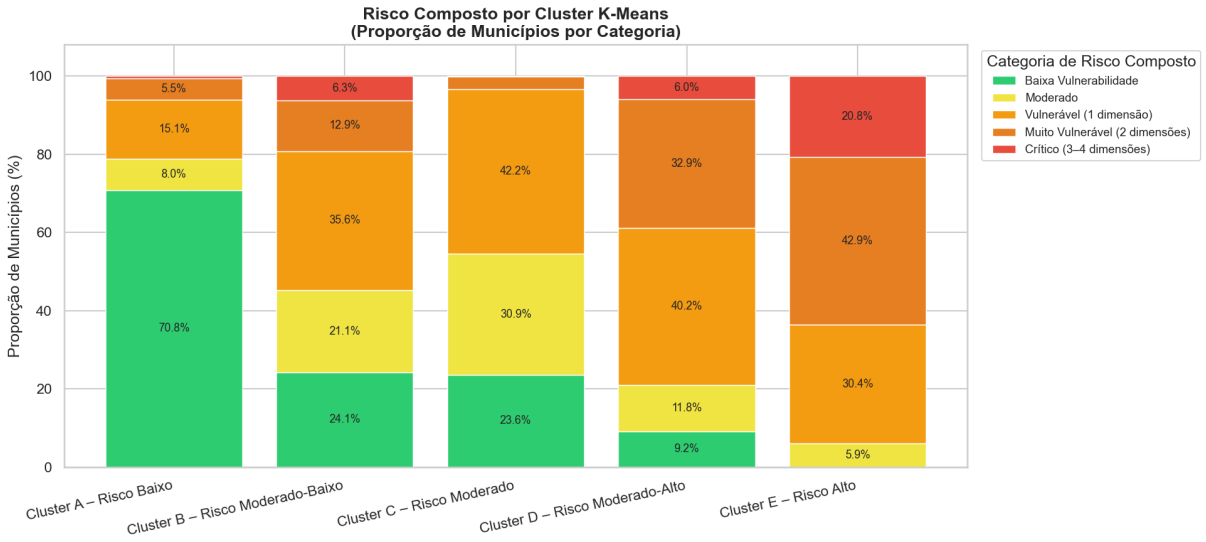

Figura 5.9: Distribuição proporcional das categorias de risco composto por cluster K-Means. A correspondência entre Cluster E e as categorias de maior gravidade valida a coerência interna do agrupamento.

6 Conclusões

Este trabalho apresentou o desenvolvimento e a aplicação de uma modelagem baseada em mineração de dados para investigar as relações entre eventos climáticos extremos e o risco de incidência de arboviroses no cenário brasileiro. Fundamentada nas etapas do processo de Descoberta de Conhecimento em Bases de Dados (KDD), a pesquisa viabilizou a extração, integração e harmonização de bases governamentais de naturezas distintas, especificamente o histórico de ocorrências do Atlas Digital de Desastres e as projeções de vulnerabilidade da plataforma AdaptaBrasil MCTI. A partir dessa consolidação, técnicas de análise de correlação e algoritmos de agrupamento não supervisionado foram aplicados sobre os dados em escala municipal.

Diante dos resultados obtidos, conclui-se que a pesquisa atingiu seu objetivo geral ao descobrir padrões de risco consistentes e estabelecer correlações estatísticas precisas através de técnicas de mineração de dados, extraindo conhecimento prático e integrado do cruzamento entre o histórico do Atlas Digital de Desastres e as projeções da plataforma AdaptaBrasil MCTI. Essa abordagem permitiu responder de forma direta e explícita às duas questões de pesquisa levantadas na introdução. Em resposta à primeira questão (RQ1), a análise revelou como os dados históricos se correlacionam com as projeções futuras de risco à saúde: comprovou-se estatisticamente que municípios com maior histórico de secas e estiagens apresentam uma correlação positiva com o risco projetado para arboviroses, enquanto eventos de inundações e chuvas intensas apresentam correlação negativa. Esses achados validam o modelo numérico ao encontrarem respaldo na literatura científica recente, corroborando as conclusões de Cox et al. (2025) e Lowe et al. (2021), que associam secas severas ao aumento de criadouros artificiais decorrente do armazenamento urbano de água, e os estudos de Benedum et al. (2018), que explicam a redução vetorial temporária promovida pela lavagem mecânica (*flushing*) provocada por chuvas extremas. Em resposta à segunda questão (RQ2), a integração e clusterização de diferentes dimensões de risco em conjunto com o histórico de ocorrências revelaram padrões multivariados claros de vulnerabilidade climática. A aplicação do algoritmo K-Means,

validada metodologicamente por Wu et al. (2021) para bases epidemiológicas de grande escala, segmentou os municípios em cinco perfis distintos, destacando a emergência de grupos com sobreposição crítica de vulnerabilidades (como o Cluster E), onde altos riscos hídricos, de saúde, energéticos e de biodiversidade coexistem no território.

A relevância desta pesquisa consolida-se, sobretudo, na materialização desses padrões por meio dos produtos analíticos e visuais gerados. Os diagramas de dispersão construídos traduziram matematicamente o acoplamento entre a frequência de desastres históricos e a projeção do risco à saúde, enquanto os mapas de calor (*heatmaps*) desvendaram o peso relativo de cada vulnerabilidade nos agrupamentos. A projeção espacial gerada pela Análise de Componentes Principais (PCA) comprovou geometricamente a separação entre as regiões de Risco Baixo e Risco Alto. Esse conjunto de gráficos e indicadores instrumentaliza o diagnóstico de vulnerabilidade climática, provando que a extração de conhecimento do cruzamento de bases agrega valor informacional à gestão técnica de desastres.

Apesar da consistência técnica demonstrada, o estudo apresenta limitações inerentes ao escopo dos dados e às abordagens escolhidas. A principal limitação técnica reside na qualidade e no preenchimento das bases de dados governamentais. Isso fica evidenciado, por exemplo, na exclusão de 8 municípios da etapa de clusterização devido à falta de valores nos índices do AdaptaBrasil, bem como na base do Atlas de Desastres, onde os registros mais antigos tendem a apresentar um menor volume de dados e maior chance de subnotificação em comparação aos registros mais recentes. Essa ausência de registros ou a existência de dados faltantes em municípios de menor porte pode gerar ruídos ou enviesamento no agrupamento espacial. Além disso, no âmbito algorítmico, o K-Means apresenta sensibilidade à presença de outliers e exige uma definição prévia do número de clusters, o que demanda rigorosa validação interna. Por fim, ressalta-se que as técnicas utilizadas identificam padrões de correlação e associação de atributos, não atestando, obrigatoriamente, relações de causalidade direta entre as variáveis ambientais e epidemiológicas.

Como propostas para trabalhos futuros e continuidade desta linha de pesquisa, sugere-se a incorporação de algoritmos de aprendizado de máquina supervisionado (como

Random Forest ou Redes Neurais Artificiais) visando não apenas o agrupamento, mas a predição da evolução do risco de doenças a partir de dados meteorológicos de curto prazo, tais como o volume de precipitação acumulada, a umidade relativa, o número de dias consecutivos sem chuva e as variações de temperaturas máximas e mínimas nas semanas antecedentes. Recomenda-se, também, a expansão do modelo para englobar as demais dimensões do AdaptaBrasil, focando em outras doenças transmissíveis, como, por exemplo, a malária. Por fim, sugere-se a evolução desta modelagem analítica para a construção de um painel interativo (dashboard). A ideia central para esse painel é a criação de uma espécie de mapa interativo do Brasil, onde os dados consolidados e os clusters resultantes desta pesquisa apareçam de forma dinâmica e espacializada, permitindo que gestores e pesquisadores explorem visualmente as áreas de risco composto e utilizem os achados da mineração de dados como ferramenta direta de suporte à tomada de decisão.

Bibliografia

BENEDUM, C. M. et al. Statistical modeling of the effect of rainfall flushing on dengue transmission in singapore. *PLoS Neglected Tropical Diseases*, v. 12, n. 6, p. e0006535, 2018.

BRASIL. *Ministério do Desenvolvimento Regional. Base de dados do Atlas Digital de Desastres no Brasil: manual de aplicação*. Brasília, 2022.

BRASIL. *Levantamento Rápido de Índices para Aedes aegypti (LIRAa)*. 2024. Disponível em: ⟨https://www.gov.br/saude/pt-br/assuntos/saude-de-a-a-z/a/arboviroses/liraa⟩. Acesso em: 30 maio 2026.

COX, V. M. et al. Spatiotemporal relationships between extreme weather events and arbovirus transmission across brazil. *medRxiv*, 2025. Disponível em: ⟨https://doi.org/10.1101/2025.08.01.25332773⟩.

DASGUPTA, A. et al. Scalable, open-access and multidisciplinary data integration pipeline for climate-sensitive diseases. *Wellcome Open Research*, v. 10, n. 467, 2025. Disponível em: ⟨https://doi.org/10.12688/wellcomeopenres.24774.1⟩.

DIAS, M. A. F. d. S. Eventos climáticos extremos. *Revista USP*, São Paulo, n. 103, p. 33–40, nov. 2014.

DONOHO, D. 50 years of data science. *Journal of Computational and Graphical Statistics*, v. 26, n. 4, p. 745–766, 2017.

FIELD, C. B. et al. *Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation*. Cambridge, UK, and New York, NY, USA, 2012.

GARCIA, K. K. S.; MIRANDA, C. B. d.; SOUSA, F. N. e. F. d. Procedimentos para vinculação de dados da saúde: aplicações na vigilância em saúde. *Epidemiologia e Serviços de Saúde*, v. 31, n. 3, p. e20211272, 2022. Disponível em: ⟨https://doi.org/10.1590/S2237-96222022000300004⟩.

HAN, J.; PEI, J.; KAMBER, M. *Data Mining: Concepts and Techniques*. 3rd. ed. Waltham: Morgan Kaufmann, 2011.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2. ed.. ed. New York: Springer, 2009.

INPE. *Teórico-metodológico para avaliação do risco de impacto das mudanças climáticas em Setores Estratégicos: Recursos Hídricos*. São José dos Campos, 2025. Data de revisão: 18 ago. 2025.

LOWE, R. et al. Combined effects of hydrometeorological hazards and urbanisation on dengue risk in brazil: a spatiotemporal modelling study. *The Lancet Planetary Health*, v. 5, n. 4, p. e209–e219, 2021.

MARCELINO, E. V. *Desastres Naturais e Geotecnologias: Conceitos Básicos*. Santa Maria, 2008.

MARCONI, M. d. A.; LAKATOS, E. M. *Metodologia Científica*. 7. ed.. ed. São Paulo: Atlas, 2019.

MEHENNI, T. Multi-database mining: A framework for heterogeneous relational databases. In: *Proceedings of the International Conference on Intelligent Systems and Pattern Recognition (ISPR '20)*. Association for Computing Machinery, 2020. Disponível em: ⟨https://doi.org/10.1145/3432867.3432897⟩.

Nações Unidas Brasil. *Objetivo 11: Tornar as cidades e os assentamentos humanos inclusivos, seguros, resilientes e sustentáveis*. 2015. Disponível em: ⟨https://brasil.un.org/pt-br/sdgs/11⟩. Acesso em: 25 fev. 2025.

Nações Unidas Brasil. *Objetivo 13: Tomar medidas urgentes para combater a mudança climática e seus impactos*. 2015. Disponível em: ⟨https://brasil.un.org/pt-br/sdgs/13⟩. Acesso em: 25 fev. 2025.

TOMINAGA, L. K.; SANTORO, J.; AMARAL, R. d. A. (Ed.). *Desastres naturais: conhecer para prevenir*. São Paulo: Instituto Geológico, 2009.

VILLAR, A. et al. Integrating and analyzing medical and environmental data using etl and business intelligence tools. *International Journal of Biometeorology*, v. 62, p. 1085–1095, 2018. Disponível em: ⟨https://doi.org/10.1007/s00484-018-1511-9⟩.

WITTEN, I. H.; FRANK, E.; HALL, M. A. *Data Mining: Practical Machine Learning Tools and Techniques*. 3. ed.. ed. Burlington: Morgan Kaufmann, 2011.

WU, W.-T. et al. Data mining in clinical big data: the frequently used databases, steps, and methodological models. *Military Medical Research*, v. 8, n. 44, 2021. Disponível em: ⟨https://doi.org/10.1186/s40779-021-00338-z⟩.