Metodologia para Avaliação do Impacto de Eventos (Epidemias e Desastres) por *Machine Learning*: Estudo de caso da Barragem de Fundão ⋆

Lucas L. Carneiro * **Ed Wilson R. Vieira** ** **Walmir M. Caminhas** ***

* *Programa de Pós-Graduação em Engenharia Elétrica – Universidade Federal de Minas Gerais – Av. Antônio Carlos 6627, 31270-901, Belo Horizonte, MG, Brasil (e-mail: llcarneiro@ufmg.br).*
** *Departamento de Enfermagem Materno-Infantil e Saúde Pública – Universidade Federal de Minas Gerais – Av. Augusto de Lima, 30190-000, Belo Horizonte, MG, Brasil (e-mail: edwilsonvieira@gmail.com).*
*** *Programa de Pós-Graduação em Engenharia Elétrica – Universidade Federal de Minas Gerais – Av. Antônio Carlos 6627, 31270-901, Belo Horizonte, MG, Brasil (e-mail: caminhas@cpdee.br).*

Abstract: This article proposes a new method for assessing the impact of an event, such as epidemics, natural and technological disasters, in a given set of places. The proposed method uses machine learning techniques and statistical tools to investigate effects of the event, in one or more variables, through different angles. Four distinct analyses are performed, three monovariate (descriptive analysis; Resultant Vectors Graph; and statistical comparison through paired t tests) and one multivariate analysis through Fuzzy c-means clustering algorithm. The proposed method is applied in a case study: the Fundão dam failure in Mariana – MG. This event is investigated with respect to "Number of dengue fever cases" and "GDP" variables. This new method managed to identify the impact, or its absence, in the observed counties after the event. The main results indicated that Mariana and Conselheiro Pena suffered an increase in dengue fever cases level in almost all analysis.

Resumo: Este artigo propõe uma metodologia para a análise do impacto de eventos, como epidemias e desastres naturais e tecnológicos, em um conjunto de localidades. O método proposto utiliza técnicas de *Machine Learning* e ferramentas estatísticas para investigar os efeitos do evento, em uma ou mais variáveis, por diferentes perspectivas. Quatro diferentes análises são realizadas, três monovariadas (análise descritiva; Gráfico de Resultantes Vetoriais; e comparação estatística por testes t pareados) e uma análise multivariada por meio do algoritmo de *clusterização Fuzzy c-means*. O método proposto é aplicado em um estudo caso: o rompimento da barragem de Fundão em Mariana – MG. Este evento é investigado com relação às variáveis "Número de casos de dengue" e "PIB". A metodologia foi capaz de identificar os impactos, ou ausência destes, nos municípios estudados após a ocorrência do evento. Dentre os resultados obtidos, destacam-se os municípios de Mariana e Conselheiro Pena que apresentaram aumento nos níveis de atendimento em quase todas as análises.

Keywords: method; effect; disaster; machine learning; statistics
Palavras-chaves: método; efeito; desastre; aprendizado de maquina; estatística

1. INTRODUÇÃO

Eventos catastróficos como epidemias, desastres naturais e tecnológicos ocorrem frequentemente em todo o mundo, apenas na última década foram registrados mundialmente mais de 5,7 mil desastres (EM-DAT, 2022). Quando eventos dessa magnitude ocorrem, diferentes formas de assistência são providenciadas. Uma resposta imediata serve para controlar os danos e minimizar as perdas, porém, após o evento, um tipo de assistência igualmente importante é a avaliação de impactos para amparar os moradores das localidades atingidas.

Os impactos podem ser de diferentes naturezas, sejam impactos ambientais, na saúde ou até mesmo econômicos. Para investigar esses impactos, diferentes metodologias podem ser utilizadas. Alguns dos métodos empregados são as diferentes formas de regressão, como segmentada (Becquart et al., 2019) e logística (Kishi et al., 2015), além de diferentes tipos de testes estatísticos como t de Student (Becquart et al., 2019) e Difference-in-Difference (Nishi-

⋆ Este trabalho foi assistido pela CAPES, número de processo 88887.682854/2022-00, e pelo CPNq

jima and Rocha, 2020). Vista a variedade de caminhos (métodos) que os especialistas podem seguir para avaliação de impactos, este artigo propõe uma nova metodologia para avaliar impactos de desastres e epidemias a partir do emprego de técnicas de aprendizado de máquina (*machine learning*) em conjunto com análises estatísticas.

Dentre as técnicas utilizadas, destaca-se o algoritmo de Agrupamento Nebuloso *Fuzzy c-means*, que é capaz de particionar um conjunto de dados em grupos compostos por amostras similares entre si e diferentes das demais de forma que estas possuam graus de pertinência a cada grupo (Babuška, 2012; Jain et al., 1999). Esta classe de algoritmos é utilizada em diversos problemas por ser capaz de modelar a incerteza do mundo real, tornando-os mais robustos (Bedregal et al., 2010).

Este artigo é organizado da seguinte forma: a Seção 2 explora como é formulado o problema de pesquisa, ou seja, como que a ocorrência do evento é interpretada; a Seção 3 detalha os algoritmos de aprendizado de máquina utilizados na metodologia proposta; já a Seção 4 descreve, passo a passo, o método proposto e as técnicas que ele utiliza; a Seção 5 apresenta um estudo de caso para exemplificar a aplicação da metodologia em uma situação real; por fim, a Seção 6 apresenta as principais conclusões deste trabalho.

2. FORMULAÇÃO DO PROBLEMA DE PESQUISA

Dada a ocorrência de um evento, sejam N_{LA} o número de localidades atingidas e N_T o número total de localidades observadas, define-se

$$C_T = \{L_1, L_2, \ldots, L_{N_T}\}, \tag{1}$$

como o conjunto total de localidades e

$$C_{LA} = \{L_i \in C_T \mid C_{LA} \subset C_T\} \tag{2}$$

como o conjunto de localidades atingidas, ou seja, as localidades que foram diretamente afetadas pelo evento; onde $|C_{LA}| = N_{LA}$ é a cardinalidade de C_{LA}. Por sua vez, o conjunto C_T compreende todas as localidades imediatamente próximas às atingidas que tenham a mesma granularidade. Alguns exemplos são dados a seguir:

- Selecionar todos os bairros e regiões vizinhas para um evento que afete apenas uma região de um município (ex.: apagões, deslizamento de terra, problemas de mobilidade urbana);
- Selecionar todos os municípios de um estado para um evento que possa afetar uma ou mais cidades (ex.: terremotos, tsunamis, problemas de abastecimento);
- Selecionar todos os estados de um país para um evento que possa afetar um estado ou região (ex.: epidemia, crise hídrica).

Busca-se um conjunto de localidades de controle, estas devem possuir caraterísticas similares às atingidas. Os controles serão utilizados para estabelecer o comportamento padrão [1] que uma localidade atingida deveria obedecer caso o evento não tivesse ocorrido. Assim, com um número N_{LC} de localidades de controle, tem-se

$$C_{LC} = \{L_i \in C_T \mid C_{LC} \subset C_T;\ C_{LC} \cap C_{LA} = \emptyset\} \tag{3}$$

como o conjunto de localidades de controle.

[1] Comportamento padrão se refere ao estado dos atributos antes do evento.

Para avaliar o efeito do evento nas localidades, estuda-se N_a características que possivelmente foram impactadas. [2] Assim, define-se

$$C_A = \{A_1, A_2, \ldots, A_{N_a}\} \tag{4}$$

como o conjunto de atributos que possivelmente foram afetados pelo evento.

Assume-se que os dados estejam estruturados em séries temporais, onde T_e marca o instante da ocorrência do evento. As etapas desta formulação são descritas a seguir:

(1) Definir o conjunto das localidades de controle;
(2) Análise monovariada – Descrever os efeitos do evento observados em cada atributo por localidade atingida;
(3) Análise monovariada – Avaliar se o efeito do evento observado em cada atributo é significativo nas localidades atingidas e nos respectivos controles;
(4) Análise multivariada – Avaliar se o evento provoca mudança no padrão de comportamento das localidades atingidas.

Ao final das etapas, torna-se possível inferir quais localidades mais sofreram impactos decorrentes do evento em estudo.

3. REFERENCIAL TEÓRICO

O método de agrupamento (*clusterização*) Fuzzy c-means (FCM) (Dunn, 1973; Bezdek, 2013) consiste em particionar um conjunto de dados $\mathbf{X}$ em c grupos (*clusters*) por meio da solução do seguinte problema de otimização:

$$\min_{\mathbf{U},\mathbf{K}} \mathcal{J}(\mathbf{X}; \mathbf{U}, \mathbf{K}) = \sum_{i=1}^{c}\sum_{j=1}^{n} \mu_{ij}^m d_{ij}^2 \tag{5a}$$

$$\text{sujeito a} \sum_{i=1}^{c} \mu_{ij} = 1,\ j = 1, \ldots, n \tag{5b}$$

$$0 < \sum_{j=1}^{n} \mu_{ij} < n,\ i = 1, \ldots, c \tag{5c}$$

onde $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\} \in \mathbb{R}^{n\times p}$ é o conjunto de dados com n amostras e p características, $\mathbf{K} = \{\mathbf{k}_1, \mathbf{k}_2, \ldots, \mathbf{k}_c\} \in \mathbb{R}^{c\times p}$ é o conjunto dos c clusters, $\mu_{ij} \in [0, 1] \subset \mathbf{U}$ corresponde à pertinência da amostra j ao cluster i, $m \in (1, \infty)$ é chamado parâmetro de fuzzyficação e $d_{ij}^2 = (\mathbf{x}_j - \mathbf{k}_i)^T \mathbf{I} (\mathbf{x}_j - \mathbf{k}_i)$ é a distância euclidiana entre $\mathbf{x}_j$ e $\mathbf{k}_i$.

Para se obter os pontos estacionários são utilizados multiplicadores de Lagrange $\lambda_i \in \mathbb{R}$ para adicionar a restrição (5b) à função objetivo:

$$\overline{\mathcal{J}}(\mathbf{X}; \mathbf{U}, \mathbf{K}, \lambda) = \sum_{i=1}^{c}\sum_{j=1}^{n} \mu_{ij}^m d_{ij}^2 + \sum_{j=1}^{n} \lambda_i \left(\sum_{i=1}^{c} \mu_{ij} - 1 \right) \tag{6}$$

Ao zerar os gradientes de (6) em relação a $\mathbf{U}$, $\mathbf{K}$ e λ com $m > 1$ e $d_{ij}^2 > 0\ \forall i, j$, obtém-se as equações para atualizar $\mathbf{U}$ e $\mathbf{K}$ dadas por:

[2] Estas são as características que serão avaliadas na metodologia para análise de um provável impacto do evento, elas podem ser diferentes das características do conjunto para seleção dos controles.

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{c}(d_{ij}/d_{kj})^{2/(m-1)}} \tag{7}$$

e

$$\mathbf{k}_i = \frac{\sum_{j=1}^{n}\mu_{ij}^{m}\mathbf{x}_j}{\sum_{j=1}^{n}\mu_{ij}^{m}} \tag{8}$$

Então, algoritmo FCM consiste em atualizar $\mathbf{U}$ e $\mathbf{K}$ alternadamente até que $\|\mathbf{K}^{(t+1)} - \mathbf{K}^{(t)}\| < \epsilon$, onde t é a iteração e $\epsilon > 0$ é a tolerância usada como critério de parada.

A solução ótima da função objetivo (5a) é obtida quando se faz $\mathbf{K} = \mathbf{X}$ com $c = n$. Contudo, este cenário dificilmente é encontrado em problemas do mundo real. Assim, para um problema qualquer de *clusterização*, as soluções obtidas têm caráter local ($c < n$) o que faz a otimização de (5a) ser sensitiva ao valor de inicial de $\mathbf{K}$.

Para contornar o problema de sensitividade, comumente utiliza-se a *clusterização* subtrativa. Este algoritmo inicializa os centros de forma que dado um conjunto de parâmetros, os centros iniciais sempre serão os mesmos, o que garante a replicabilidade dos resultados.

3.1 Clusterização Subtrativa

A *Clusterização* Subtrativa, proposta por Chiu (1994) e baseada no método da montanha (Yager and Filev, 1994), consiste em um método de estimação do número de *clusters* e dos valores iniciais de seus respectivos centros, estes que podem ser utilizados para inicializar algoritmos de *clusterização* baseados na otimização da função custo, como o FCM.

Considera-se cada ponto do conjunto de dados $\mathbf{X}$ como um potencial centro de *cluster* e define-se a medida do potencial de um ponto $\mathbf{x}_i$ como:

$$f_i = \sum_{j=1}^{n} e^{-\alpha\|\mathbf{x}_i - \mathbf{x}_j\|^2}, \tag{9}$$

onde $\alpha = 4/r_a^2$ e r_a é uma constante positiva que define o raio de vizinhança para cada centro de *cluster* (Yang et al., 2010).

Após o potencial de todos os dados serem computados, a amostra com o maior potencial é selecionado como centro de *cluster*. Assim, após o k-ésimo centro for selecionado, o potencial dos dados restantes é ajustado por

$$f_i \leftarrow f_i - f_k^* e^{-\beta\|\mathbf{x}_i - \mathbf{x}_k^*\|^2}, \tag{10}$$

onde $\mathbf{x}_k^*$ é a posição do k-ésimo centro, f_k^* o seu valor de potencial, $\beta = 4/r_b^2$ e r_b é uma constante positiva que representa o raio de vizinhança sobre o qual haverá redução de potencial. Este processo é repetido iterativamente até que o critério de parada seja alcançado. Para isto, são definidas as constantes $\overline{\epsilon}$ e $\underline{\epsilon}$ como o índice de aceitação e o índice de rejeição, respectivamente.

A partir de (9) tem-se que o raio de vizinhança está diretamente relacionado com o número de grupos resultantes. Quanto maior seu valor, maior é a influência dos *clusters* gerados e, portanto, uma menor quantidade de grupos é obtida.

4. METODOLOGIA PROPOSTA

Dada a situação descrita na Seção 2, um diagrama de controle pode ser construído ao calcular a média aritmética e o desvio padrão de todos os dados anteriores ao momento de ocorrência do evento T_e, obedecendo à granularidade dos dados. Por exemplo, supõe-se que seja necessário avaliar a evolução mensal de uma variável e que o intervalo de tempo estudado anterior a T_e seja de três anos. Assim, calcula-se a média e o desvio padrão correspondente a cada mês e é desenhado o limite superior, dado por média + 1,96 desvios padrão, e o limite inferior, dado por média - 1,96 desvios padrão (Sellick, 1993). Estes limites são sobrepostos com os dados atuais (após T_e) para gerar o diagrama, como exemplificado na Figura 1.

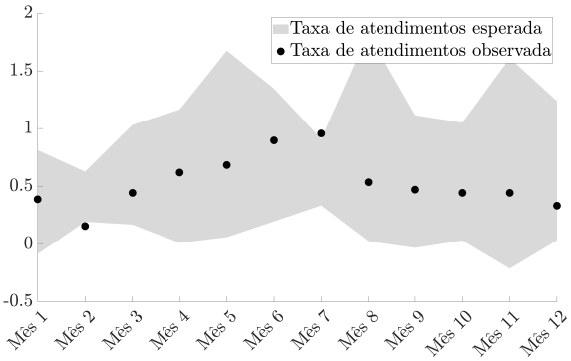

Figura 1. Ilustração de um diagrama de controle.

4.1 Gráfico de Resultantes Vetoriais

O Gráfico de Resultantes Vetoriais (GRV) é uma técnica proposta com a intenção de sintetizar as informações de um ou mais diagramas de controle em apenas um gráfico. Os GRV contemplam, simultaneamente, três informações dos diagramas: atributos acima, dentro ou abaixo dos limites históricos. A partir dos diagramas, cada vez que o atributo ultrapassa o limite superior esperado, é atribuído um vetor unitário na direção de crescimento do eixo das ordenadas (Figura 2a); quando o atributo fica abaixo do limite inferior esperado, um vetor no sentido de decrescimento é atribuído na direção de diminuição do eixo das ordenadas (Figura 2b); e quando o atributo fica dentro dos limites esperados, é atribuído um vetor unitário no sentido de crescimento do eixo das abcissas (Figura 2c). Por fim, após "percorrer" todo o diagrama, os vetores são somados para gerar um resultante (Figura 2d).

Quando mais de um diagrama de controle é utilizado para gerar o GRV, os resultantes vetoriais são exibidos em uma única figura. No GRV, se o resultante estiver no primeiro quadrante, indicará um aumento na variável para o período estudado; se estiver no quarto quadrante, a variável terá estado aquém do esperado; e se estiver próximo do eixo das abcissas, a variável estará dentro dos limites históricos.

4.2 Definição das Localidades de Controle

As variáveis utilizadas para encontrar o conjunto de localidades de controle C_{LC} podem ser as mais diversas, desde

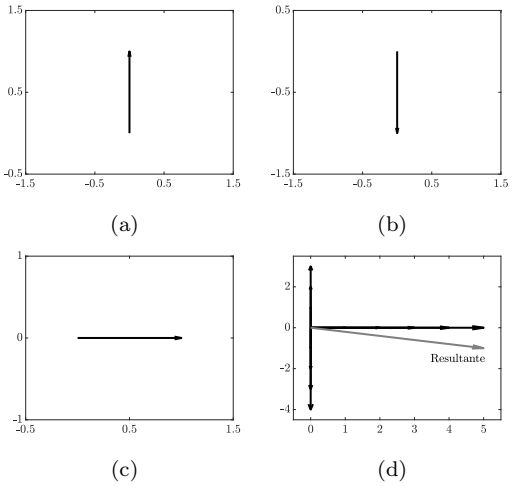

Figura 2. Ilustração da técnica de Gráficos de Resultantes Vetoriais indicando os vetores unitários atribuídos ao gráfico: 2a acima do limite superior esperado; 2b abaixo do limite inferior esperado; 2c dentro dos limites esperados; e 2d vetor resultante.

variáveis socioeconômicas, para obter localidades com semelhantes perfis socioeconômicos, até variáveis pertencentes ao conjunto C_A. Contudo, a escolha destas variáveis dependem do tipo e objetivo do estudo e aconselha-se que sejam escolhidas pelos especialistas que utilizarem esta metodologia. Ressalta-se que nesta etapa utiliza-se apenas dados anteriores ao evento.

A busca por C_{LC} é realizada com base em medida de similaridade *fuzzy*. O algoritmo *Fuzzy c-means* particiona o conjunto de dados escolhido com $m = 2$, por ser um valor comumente utilizado, onde os centros são dados correspondentes às localidades contidas em C_{LA}. Após apenas 1 iteração, as pertinências dos dados correspondentes às demais localidades ($C_T \cap C_{LA}$) são ordenadas de forma decrescente para a escolha dos controles. As localidades com maior pertinência a cada centro são escolhidas como seu respectivo controle. Por fim, deve-se verificar se não há controles repetidos entre as localidades atingidas.

4.3 Análise Monovariada

A análise monovariada é realizada individualmente para cada atributo contido em C_A e é composta por três etapas: análise descritiva, análise por GRV e comparação com controles. A análise descritiva tem o objetivo de apresentar os dados, informar proporções relativas a dados antes e após o evento para cada localidade e identificar possíveis problemas com a base de dados, tal como dados faltantes.

Na segunda etapa, o Gráfico de Resultantes Vetoriais é aplicado nos dados das localidades atingidas. O GRV é uma comparação dos dados do estado atual da localidade atingida (após o evento) com os dados históricos (antes do evento), ou seja, é uma comparação consigo mesmo. Procura-se estabelecer quais foram os principais impactos sofridos nos atributos, se houve aumento, queda, ou permanência dos valores observados na série histórica. Para auxiliar na interpretação, recomenda-se verificar quais localidades apresentam vetores mais próximos do eixo das ordenadas.

Já a comparação com o controle se dá ao avaliar as séries temporais antes e após o evento por meio do teste t pareado (Student, 1908) formulado da seguinte forma:

$$\begin{cases} H_0 : \mu_1 - \mu_2 = 0 \\ H_0 : \mu_1 - \mu_2 \neq 0 \end{cases}, \tag{11}$$

ou seja, a hipótese nula é de que não há diferença entre os períodos anterior e posterior a T_e para o atributo em estudo; já a hipótese alternativa é de que existe uma diferença.

O teste é realizado para cada localidade, atingida e controle, com o nível de significância de 0,05, de forma que a amostra 1 seja composta por dados históricos e a amostra 2 por dados atuais do atributo. Devido à restrição das amostras possuírem o mesmo número de observações, sugere-se calcular a média dos dados anteriores ao evento, analogamente ao processo de construção do diagrama de controle descrito na Seção 4.

O p-valor resultante do teste permite verificar se a diferença observada é significativa para o nível de significância especificado. Dada a realização dos testes, os possíveis resultados são descritos a seguir:

- **Diferença significativa apenas na localidade atingida:** Há evidências de impacto do evento, visto que houve alteração apenas em uma das localidades.
- **Diferença significativa apenas no controle:** Há evidências de impacto do evento, visto que houve alteração apenas em uma das localidades.
- **Diferença significativa na localidade atingida e no controle:** Caso as diferenças tenham mesmo sinal, não há evidências de impacto do evento; se as diferenças têm sinais opostos, há evidências de impacto do evento.
- **Não há diferença significativa:** Não há evidências de impacto do evento.

Contudo, as análises não devem se limitar aos casos mencionados acima. O p-valor pode indicar o quão forte é a evidência contra a hipótese nula. Deste modo, p-valores que excedem mas ainda estão próximos à 0,05 podem indicar um provável impacto do evento, ainda que com um nível de significância diferente (Montgomery, 2001).

4.4 Análise Multivariada

A análise multivariada é utilizada caso $N_a > 1$ e é executada por meio do algoritmo *Fuzzy c-means*. Um conjunto de dados $\mathbf{X}^* = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{N_{2LA}}\} \in \mathbb{R}^{N_{2LA} \times N_a}$ é criado a partir da junção de dados anteriores e posteriores ao evento. Dessa forma, tem-se o termo N_{2LA} que corresponde a N_{LA} observações pré e N_{LA} observações pós-evento; cada uma com N_a características. Por isso, uma restrição desta análise é que os possíveis diferentes conjuntos de dados possam ser agrupados como descrito anteriormente.

O conjunto $\mathbf{X}^*$ é particionado em 3 grupos. Para isso, varia-se o r_a até que este valor de c seja alcançado [3]. A definição da quantidade de grupos teve caráter empírico. Como será melhor explicado na Seção 5, o método proposto utiliza um sistema de defuzificação para transformar

[3] Nenhum valor de r_a é definido nesta metodologia visto que o c resultante depende do conjunto de dados.

os valores de centro de numéricos para linguísticos. Notou-se que caso fosse selecionado $c = 2$, alguns grupos com valores intermediários poderiam ser perdidos, e que caso o c fosse maior que 3, a interpretabilidade dos grupos poderia ser comprometida.

Após o particionamento, verifica-se a pertinência de cada observação aos *clusters*, assim, tem-se uma pertinência para dados pré e outra para dados pós-evento. O objetivo desta análise é verificar se houve mudança de grupo, ou seja, se houve mudança de comportamento para uma dada localidade. Para simplicidade da análise, considera-se que uma observação pertence ao grupo ao qual apresenta maior pertinência. Contudo, ressalta-se que o algoritmo utiliza de lógica *fuzzy* e por isso uma interpretação mais refinada pode ser obtida ao interpretar os graus de pertinência.

4.5 Observações Finais

Esta metodologia propõe a análise do impacto de um avento por três frentes distintas: uma análise introspectiva onde a comparação se dá entre os dados atuais e históricos, por localidade; uma análise comparativa onde uma localidade de controle, que possui comportamento histórico similar à localidade atingida, passa pelas mesmas análises que as atingidas e os dois resultados são avaliados; e uma terceira comparação onde localidades com comportamento similar são agrupadas e a diferença entre o comportamento antes e após o evento é avaliado de uma forma global (localidades atingidas × localidades atingidas).

5. ESTUDO DE CASO – ROMPIMENTO DA BARRAGEM DE FUNDÃO

Esta seção apresenta um estudo de caso para a aplicação da metodologia proposta. Neste estudo de caso, serão avaliados os impactos do rompimento da barragem de Fundão em Mariana no estado de Minas Gerais ocorrido em 5 de novembro de 2015, identificado como o maior desastre socioambiental do país no setor de mineração (IBAMA, 2020). Foram considerados como anteriores ao evento dados de novembro de 2010 a outubro de 2015 e como posteriores ao evento dados de novembro de 2015 a outubro de 2016.

Define-se C_T como composto por todos os $N_T = 853$ municípios de Minas Gerais e C_{LA} como os $N_{LA} = 36$ municípios de Minas Gerais que foram direta ou indiretamente atingidos pela lama decorrente do rompimento da barragem de Fundão que foram contemplados no Termo de Transação e Ajustamento de Conduta (Fundação Renova, 2018). O conjunto C_A é composto por $A_1 =$ Taxa mensal de atendimentos por dengue e por $A_2 =$ Produto Interno Bruto (PIB) municipal anual.

Dados mensais de atendimentos por dengue durante o período de 2010 a 2016 foram extraídos do Tabnet DATASUS (2018) e então convertidos em taxa de atendimentos por mil habitantes utilizando estimativas de população do Instituto Brasileiro de Geografia e Estatística (IBGE) (IBGE, 2022). Por sua vez, os dados de PIB municipal dos anos de 2015 e 2016 para os municípios de Minas Gerais foram extraídos do Cidades@, o sistema agregador de informações do IBGE sobre os municípios e estados do Brasil (IBGE, 2017a). Além disso, uma base de dados secundária contendo o Índice de Desenvolvimento Humano Municipal (IDHm) para os municípios de Minas gerais foi obtida em IBGE (2017b), os valores mais recentes datam de 2010.

5.1 Definição dos Controles

Para a definição dos controles foram definidos como atributos a taxa de atendimento anual por dengue e o valor de IDHm. Essas variáveis foram selecionadas para incluir o efeito da sazonalidade dos casos de dengue na escolha dos controles assim como para considerar o porte socioeconômico do município. A metodologia para determinação dos controles foi aplicada e o resultado simplificado é exibido na Tabela 1. Como não houve localidades repetidas na lista dos municípios mais similares, estes formaram o conjunto das localidades de controle. Dessa forma, o algoritmo de *clusterização* foi capaz de selecionar os controles adequadamente.

Tabela 1. Localidades mais similares às atingidas, ao considerar taxas de atendimento anual por dengue e IDHm, obtidas por meio de *clusterização fuzzy* – são exibidas até as segundas localidades mais similares para simplificação da tabela.

Localidade Atingida	1ª Mais Similar	2ª Mais Similar
Aimorés	Quartel Geral	São Gonçalo do Pará
Alpercata	Morro da Garça	Santa Maria de Itabira
Barra Longa	Caparaó	Rio Pardo de Minas
Belo Oriente	Elói Mendes	Itamarati de Minas
Bom Jesus do Galho	Cônego Marinho	Jacinto
Bugre	Riacho dos Machados	Mendes Pimentel
Caratinga	Santa Juliana	Santa Bárbara
Conselheiro Pena	Lajinha	Januária
Córrego Novo	Lagoa dos Patos	Martins Soares
Dionísio	Guaranésia	Presidente Olegário
Fernandes Tourinho	Capitão Enéas	Machacalis
Galiléia	Guarará	Santo Hipólito
Governador Valadares	Além Paraíba	Santo Antônio do Monte
Iapu	Carmo da Cachoeira	Capelinha
Ipaba	Miradouro	Volta Grande
Ipatinga	Uberaba	Lavras
Itueta	Oratórios	Conceição do Mato Dentro
Mariana	Caxambu	Bicas
Marliéria	Monte Azul	Brasília de Minas
Naque	Centralina	Jaguaraçu
Periquito	Gameleiras	Rio Manso
Pingo-d'Água	Ibiaí	Malacacheta
Ponte Nova	Ibiá	Ijaci
Raul Soares	Pescador	Águas Formosas
Resplendor	Claro dos Poções	Água Comprida
Rio Casca	São João do Oriente	Itanhomi
Rio Doce	São Brás do Suaçuí	Mercês
Santa Cruz do Escalvado	São João do Pacuí	Piedade dos Gerais
Santana do Paraíso	Antônio Prado de Minas	Pimenta
São Domingos do Prata	Manhuaçu	Astolfo Dutra
São José do Goiabal	Uruana de Minas	Chácara
São Pedro dos Ferros	Conceição do Rio Verde	Inimutaba
Sem-Peixe	Abre Campo	Cana Verde
Sobrália	Urucânia	Tarumirim
Timóteo	Sete Lagoas	Coronel Fabriciano
Tumiritinga	Santana de Pirapama	Espinosa

5.2 Análise Monovariada

Os números de atendimento por dengue, antes e após o evento, para dez dos municípios atingidos e seus respectivos controles são exibidos na Tabela 2 [4]. Em um comportamento normal, esperava-se que todos os municípios apresentassem maior porcentagem de atendimentos antes

[4] O número de municípios exibidos na Tabela 2 foi reduzido para uma exibição mais simples e economia de espaço. Os municípios retratados correspondem aos resultados mais relevantes.

do evento (Nov/2010 a Out/2015) do que após (Nov/2015 a Out/2016), devido ao maior número de meses estudados; porém, é observado que alguns municípios como Mariana apresentam alta porcentagem após o evento, o que pode indicar um provável impacto. Além disso, alguns municípios apresentam quantidade de atendimentos extremamente baixas, como Caparaó que apresenta apenas 1 atendimento em 6 anos estudados, o que pode indicar um possível problema no registro de atendimentos daquela cidade.

Tabela 2. Número de atendimentos observados antes e após o evento para dez municípios atingidos selecionados e seus controles.

Localidade	Número de atendimentos		
	Antes do evento (Nov. 2010 - Out. 2015)	Depois do Evento (Nov. 2015 - Out. 2016)	Total
Abre Campo	253 (48,47%)	269 (51,53%)	522
Barra Longa	1 (0,58%)	171 (99,42%)	172
Belo Oriente	204 (41,89%)	283 (58,11%)	487
Bugre	37 (36,63%)	64 (63,37%)	101
Caparaó	1 (100%)	0 (0%)	1
Caratinga	831 (34,61%)	1569 (65,35%)	2401
Caxambu	105 (54,97%)	86 (45,03%)	191
Conselheiro Pena	388 (60,44%)	254 (39,56%)	642
Dionísio	128 (57,40%)	95 (42,60%)	223
Elói Mendes	175 (90,67%)	17 (8,81%)	193
Guaranésia	289 (30,71%)	652 (69,29%)	941
Ipaba	445 (53,23%)	389 (46,53%)	836
Lajinha	291 (53,49%)	253 (46,51%)	544
Mariana	339 (41,14%)	485 (58,86%)	824
Miradouro	275 (63,07%)	161 (36,93%)	436
Riacho dos Machados	97 (95,10%)	5 (4,90%)	102
Rio Doce	3 (27,27%)	8 (72,73%)	11
Santa Juliana	120 (47,62%)	132 (52,38%)	252
São Brás do Suaçuí	5 (33,33%)	10 (66,67%)	15
Sem-Peixe	57 (47,90%)	62 (52,10%)	119

Para a auto-comparação, o GRV foi aplicado aos municípios estudados e os resultados podem ser visualizados nas Figuras 3, municípios com média anual acima do esperado, e 4, demais municípios. Os rótulos entre parênteses contêm a quantidade de meses acima, dentro e abaixo do esperado, respectivamente; e separado por um travessão é exibido valor da média anual da taxa de atendimentos por dengue.

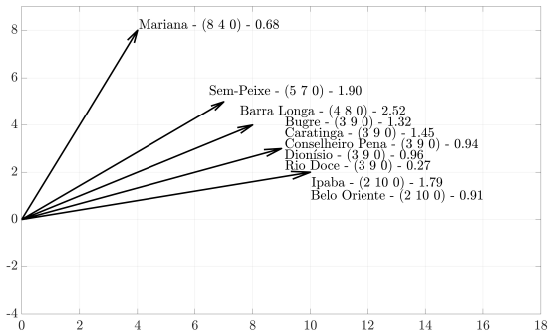

Figura 3. Gráfico de resultantes vetoriais – Municípios com média anual acima do esperado.

Destacam-se os municípios de Barra Longa, Sem-peixe e Mariana por terem apresentado alto número de meses acima do esperado. A maior parte dos demais municípios apresentaram taxas aquém do esperado ou com poucos meses acima da taxa esperada. Aqui o GRV permitiu reduzir o número de imagens necessárias para realizar a análise e identificar os municípios com resultados mais

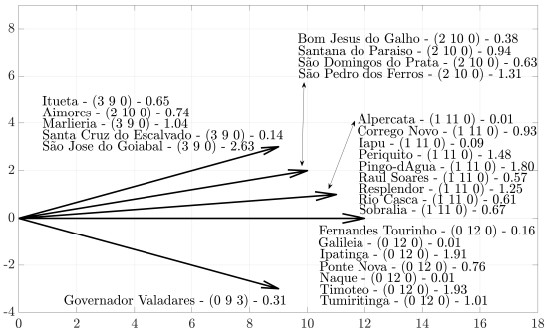

Figura 4. Gráfico de resultantes vetoriais – Demais municípios estudados.

alarmantes. Contudo, ele não permitiu identificar quais meses que ficaram acima do esperado.

A segunda parte da análise compreende os testes estatísticos e comparação com os controles. Os resultados são exibidos na Tabela 3.

Tabela 3. Resultante do teste t pareado para dez municípios atingidos e seus respectivos controles.

Municípios Atingidos					
Localidade	Média Pós-Evento	Média Pré-Evento	Dif. das Médias	Erro Padrão	*p-valor*
Barra Longa	2,5248	0,0027	2,5221	1,3895	0,0968
Belo Oriente	0,9070	0,1373	0,7697	0,6342	0,2503
Bugre	1,3175	0,1520	1,1655	0,8634	0,2042
Caratinga	1,4498	0,1564	1,2934	0,7737	0,1228
Conselheiro Pena	0,9355	0,2854	0,6501	0,3125	0,0617
Dionísio	0,9615	0,2452	0,7163	0,4489	0,1388
Ipaba	1,7882	0,4232	1,3650	0,9002	0,1576
Mariana	0,6791	0,0984	0,5808	0,2786	0,0612
Rio Doce	0,2684	0,0200	0,2483	0,1505	0,1271
Sem-Peixe	1,9020	0,3427	1,5593	1,0412	0,1624
Municípios de Controle					
Caparaó	0,0000	0,0031	-0,0031	0,0031	0,3388
Elói Mendes	0,0519	0,1082	-0,0562	0,0469	0,2552
Riacho dos Machados	0,0435	0,1691	-0,1256	0,0927	0,2027
Santa Juliana	0,8120	0,1589	0,6531	0,3193	0,0655
Lajinha	1,0608	0,2432	0,8176	0,4702	0,1099
Guaranésia	2,8691	0,2540	2,6151	1,5070	0,1106
Miradouro	1,2678	0,4348	0,8330	0,4637	0,0999
Caxambu	0,3338	0,0810	0,2528	0,1311	0,0800
São Brás do Suaçuí	0,2269	0,0232	0,2037	0,1129	0,0987
Abre Campo	1,6773	0,3131	1,3641	0,8493	0,1365

Dos testes exibidos, todos falharam em rejeitar a hipótese nula. Assim, para uma significância de 0,05, não há evidências de impacto do evento para nenhum dos municípios presentes na Tabela 3. Porém, alguns municípios como Conselheiro Pena e Mariana apresentam p-valor próximo ao nível de significância utilizado. Dessa forma, pode-se inferir que o evento impactou Conselheiro Pena e Mariana no sentido de aumento da média de atendimentos por dengue, corroborando com os resultados da Figura 3.

5.3 Análise Multivariada

A análise multivariada foi empregada em um conjunto de dados composto pelos atributos C_A, que foi normalizado para que os valores de cada variável estivessem no intervalo $[0, 1]$. A quantidade $c = 3$ grupos foi obtida com $r_a = 0,1$, e as pertinências estão apresentadas na Tabela 4. A Tabela 5 explica o significado de cada grupo em termos de níveis das variáveis. Os centros dos grupos foram transformados de

valor numérico para valor linguístico por meio da seguinte conversão: foi atribuído "Nível baixo" para valores no intervalo [0, 0,3), "Nível médio" para valores no intervalo [0,3, 0,6) e "Nível alto" para valores em [0,6, 1]. Este processo é chamado de *defuzificação*.

Tabela 4. Pertinências fuzzy antes e após a ocorrência do evento – apenas dez municípios exibidos para simplificação da tabela.

Município Atingido	Período	Pertinências aos Grupos		
		1	2	3
Barra Longa	Pré-Evento	**0,9257**	0,0618	0,0125
	Pós-Evento	0,1183	0,2297	**0,6520**
Belo Oriente	Pré-Evento	**0,8124**	0,1574	0,0302
	Pós-Evento	0,0727	**0,8856**	0,0417
Bugre	Pré-Evento	**0,9774**	0,0194	0,0032
	Pós-Evento	0,1312	**0,6842**	0,1845
Caratinga	Pré-Evento	**0,7966**	0,1710	0,0324
	Pós-Evento	0,0847	0,3577	**0,5576**
Conselheiro Pena	Pré-Evento	**0,9888**	0,0100	0,0013
	Pós-Evento	0,0370	**0,9487**	0,0142
Dionísio	Pré-Evento	**0,9873**	0,0112	0,0016
	Pós-Evento	0,0707	**0,9010**	0,0283
Ipaba	Pré-Evento	**0,8402**	0,1462	0,0136
	Pós-Evento	0,1152	0,3309	**0,5539**
Mariana	Pré-Evento	**0,5094**	0,3528	0,1377
	Pós-Evento	0,2835	**0,6018**	0,1147
Rio Doce	Pré-Evento	**0,9316**	0,0571	0,0113
	Pós-Evento	**0,9784**	0,0191	0,0026
Sem-Peixe	Pré-Evento	**0,9361**	0,0573	0,0066
	Pós-Evento	0,1148	0,2998	**0,5854**

Tabela 5. Significado linguístico dos grupos obtidos na análise multivariada.

Variável	Grupo 1	Grupo 2	Grupo 3
Dengue	Baixo	Médio	Alto
PIB	Baixo	Baixo	Médio

A maior parte dos municípios da Tabela 4 apresentou uma mudança de comportamento no sentido de aumento dos níveis de atendimento por dengue. Ressalta-se que os municípios de Caratinga, Ipaba e Sem-peixe apresentam mudança de grupo para um de nível superior no PIB; contudo, com níveis de pertinência baixos ($\simeq 0,55$) o que pode indicar que o aumento não muito expressivo. Por fim, o município Rio Doce não apresenta mudança de comportamento quanto à análise multivariada.

A análise multivariada permitiu observar as mudanças de comportamento dos municípios com relação aos níveis das variáveis, em contraposição às análises anteriores que foram formas de auto-comparação. O FCM foi capaz de separar grupos bem definidos quanto aos atendimentos por dengue. Contudo, a interpretabilidade com relação aos níveis de PIB foi prejudicada. Ainda assim, foi nítida a mudança de comportamento dos municípios observados.

6. CONCLUSÕES

Este artigo propôs uma metodologia para avaliação de impactos de eventos como epidemias e desastres naturais. O problema de pesquisa foi apresentado e com ele quatro diferentes formas de analisar as variáveis, entre elas técnicas mono e multivariadas. A nova metodologia para avaliação de impactos de eventos foi aplicada ao "Rompimento da barragem de Fundão em Mariana–MG", que configura um desastre tecnológico. As análises mono e multivariáveis foram aplicadas para avaliar os atributos "Taxa mensal de atendimentos por dengue" e "PIB municipal anual". O algoritmo *Fuzzy c-means* foi capaz de auxiliar a seleção de localidades de controle baseando-se em similaridade *fuzzy* e de realizar a análise multivariada dos atributos. Dentre os resultados obtidos, destacam-se os municípios de Mariana e Conselheiro Pena que apresentaram aumento nos níveis de atendimento em quase todas as análises. Desse modo, a metodologia conseguiu identificar impactos, ou ausência de impactos, nos municípios estudados após a ocorrência do evento.

AGRADECIMENTOS

Este trabalho foi assistido pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – CAPES, número de processo 88887.682854/2022-00, e pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico – CNPq

REFERÊNCIAS

Babuška, R. (2012). *Fuzzy modeling for control*, volume 12. Springer Science & Business Media.

Becquart, N.A., Naumova, E.N., Singh, G., and Chui, K.K. (2019). Cardiovascular disease hospitalizations in louisiana parishes' elderly before, during and after hurricane katrina. *International journal of environmental research and public health*, 16(1), 74.

Bedregal, B.R. et al. (2010). A comparative study between fuzzy c-means and ckmeans algorithms. In *2010 Annual Meeting of the North American Fuzzy Information Processing Society*, 1–6. IEEE.

Bezdek, J.C. (2013). *Pattern recognition with fuzzy objective function algorithms*. Springer Science & Business Media.

Chiu, S.L. (1994). Fuzzy model identification based on cluster estimation. *Journal of Intelligent & fuzzy systems*, 2(3), 267–278.

DATASUS (2018). Tabnet. URL `https://datasus.saude.gov.br/informacoes-de-saude-tabnet/`. Acesso em: 08/05/2022.

Dunn, J.C. (1973). A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters.

EM-DAT (2022). Em-dat, the international disaster database. URL `https://public.emdat.be/data`. Acesso em: 10/05/2022.

Fundação Renova (2018). Sobre o termo. URL `https://www.fundacaorenova.org/sobre-o-termo/`. Acesso em: 08/05/2022.

IBAMA (2020). Rompimento da barragem de fundão: Documentos relacionados ao desastre da samarco em mariana/mg. URL `http://www.ibama.gov.br/cites-e-comercio-exterior/cites?id=117`. Acesso em: 08/05/2022.

IBGE (2017a). Produto interno bruto dos municípios. URL https://cidades.ibge.gov.br/brasil/mg/uniao-de-minas/pesquisa/38/47001. Acesso em: 08/05/2022.

IBGE (2017b). Índice de desenvolvimento humano. URL https://cidades.ibge.gov.br/brasil/mg/uniao-de-minas/pesquisa/37/30255. Acesso em: 08/05/2022.

IBGE (2022). Estimativas da população. URL https://www.ibge.gov.br/estatisticas/sociais/populacao/9103-estimativas-de-populacao.html?=&t=downloads. Acesso em: 08/05/2022.

Jain, A.K., Murty, M.N., and Flynn, P.J. (1999). Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3), 264–323.

Kishi, M., Aizawa, F., Matsui, M., Yokoyama, Y., Abe, A., Minami, K., Suzuki, R., Miura, H., Sakata, K., and Ogawa, A. (2015). Oral health-related quality of life and related factors among residents in a disaster area of the great east japan earthquake and giant tsunami. *Health and quality of life outcomes*, 13(1), 1–11.

Montgomery, D.C. (2001). Design and analysis of experiments. john wiley & sons. *Inc., New York*, 1997, 200–1.

Nishijima, M. and Rocha, F.F. (2020). An economic investigation of the dengue incidence as a result of a tailings dam accident in brazil. *Journal of environmental management*, 253, 109748.

Sellick, J.A. (1993). The use of statistical process control charts in hospital epidemiology. *Infection Control and Hospital Epidemiology*, 14. doi:10.2307/30149749.

Student (1908). The probable error of a mean. *Biometrika*, 1–25.

Yager, R.R. and Filev, D.P. (1994). Approximate clustering via the mountain method. *IEEE Transactions on Systems, Man, and Cybernetics*, 24(8), 1279–1284.

Yang, Q., Zhang, D., and Tian, F. (2010). An initialization method for fuzzy c-means algorithm using subtractive clustering. In *2010 third international conference on intelligent networks and intelligent systems*, 393–396. IEEE.