#### UNIVERSIDADE FEDERAL DE VIÇOSA

#### CENTRO DE CIÊNCIAS EXATAS E TECNOLÓGICAS

#### DEPARTAMENTO DE INFORMÁTICA

CARLOS HENRIQUE TAVARES BRUMATTI

##### UM ARCABOUÇO COMPUTACIONAL PARA SUPORTAR A TOMADA DE DECISÕES DE ESPECIALISTAS EM CENÁRIOS DE DESASTRES AMBIENTAIS UTILIZANDO DADOS PÚBLICOS

VIÇOSA

2022

CARLOS HENRIQUE TAVARES BRUMATTI

##### UM ARCABOUÇO COMPUTACIONAL PARA SUPORTAR A TOMADA DE DECISÕES DE ESPECIALISTAS EM CENÁRIOS DE DESASTRES AMBIENTAIS UTILIZANDO DADOS PÚBLICOS

Projeto de Pesquisa para o curso de Mestrado *stricto sensu* em Ciência da Computação da Universidade Federal de Viçosa

Orientador: Julio Cesar Soares dos Reis

Co-orientador: Jugurta Lisboa Filho

Viçosa

2022

**Sumário**

[1. INTRODUÇÃO 4](#_Toc117410814)

[2. OBJETIVOS 4](#_Toc117410815)

[2.1. OBJETIVO GERAL 4](#_Toc117410818)

[2.2. OBJETIVOS ESPECÍFICOS 4](#_Toc117410819)

[3. REFERENCIAL TEÓRICO 4](#_Toc117410820)

[3.1 SISTEMAS DE INTELIGÊNCIA GEOGRÁFICA 5](#_Toc117410822)

[3.2 MINERAÇÃO DE DADOS 5](#_Toc117410825)

[3.3 ALGORITMOS DE AGRUPAMENTO 5](#_Toc117410826)

[4. METODOLOGIA 6](#_Toc117410831)

[5. CRONOGRAMA 7](#_Toc117410836)

[6. ORÇAMENTO 8](#_Toc117410837)

[7. REFERÊNCIAS 8](#_Toc117410838)

[8. RESUMO 10](#_Toc117410839)

# INTRODUÇÃO

O desastre ambiental causado pelo rompimento da barragem de Fundão, em Mariana (MG), no dia 05 de novembro de 2015, evidenciou a íntima relação existente entre o meio biótico e o meio antrópico. Segundo Godoy e Dias (2021), o impacto causado pelo desastre ambiental provocou sérias consequências não somente na região de ocorrência, mas em todo o percurso dos rejeitos até a sua chegada ao mar. Calcula-se que 19 pessoas foram vitimadas e que destruiu ou afetou diversas comunidades ao longo do curso do Rio Doce, além de ter sido o maior desastre ambiental do Brasil, até então (Michel Teixeira et al, 2019), com as consequências não se restringindo somente às questões ambientais, mas sim socioeconômicas, políticas e humanas.

Nesse contexto, um dos problemas identificados é a alocação de recursos financeiros, de modo adequado e eficiente, para a recuperação das áreas afetadas. Segundo Barbosa et al (2015), após o desastre, estima-se que as prefeituras das áreas envolvidas terão que gastar cerca de R$150 milhões, além de que há uma proposta da criação de um fundo de US$20 bilhões, ao longo de 10 anos, pelas empresas envolvidas.

Diante deste cenário, soluções computacionais podem ser úteis para apoiar especialistas no processo de tomada de decisão relacionado a essa alocação eficinte dos recursos financeiros. Uma das maneiras existentes é através do processo de Descoberta de Conhecimento em Bases de Dados (do inglês, *Knowledge Discovery in Database –* KDD), principalmente com a etapa de Mineração de Dados desse processo.

# OBJETIVOS



## OBJETIVO GERAL

O objetivo geral deste trabalho é propor uma metodologia voltada para o agrupamento de cidades atingidas por desastres ambientais, buscando assim melhorar a alocação de recursos financeiros utilizados na recuperação dessas áreas afetadas, fornecendo assim suporte ao processo de tomada de decisão pelos responsáveis. Além disso, espera-se a construção de uma ferramenta que possa utilizar dados públicos e privados para a realização das etapas da metodologia proposta via uma interface interativa.

## OBJETIVOS ESPECÍFICOS

Para o objetivo geral ser alcançado, os seguintes objetivos específicos devem ser realizados:

* Identificar os principais algoritmos de agrupamento e as etapas necessárias para suas aplicações.
* Aplicar os algoritmos de agrupamento identificados no conjunto de dados de teste.
* Analisar os agrupamentos formados resultantes da aplicação dos algoritmos de agrupamento, buscando aplicar técnicas para identificar o padrão nos dados de um mesmo agrupamento.
* Projetar e implementar um Sistema de Inteligência Geográfica, assim como toda a arquitetura necessária para sua execução, buscando reproduzir a metodologia validada.

# REFERENCIAL TEÓRICO

Esta seção apresenta os principais conceitos relacionados ao projeto. Dessa forma, foi estruturado em quatro tópicos, sendo eles: Sistemas de Inteligência Geográfica, Mineração de Dados, Algoritmos de Agrupamento e Bacias Hidrográficas.



## SISTEMAS DE INTELIGÊNCIA GEOGRÁFICA

A partir da coleta de dados sobre uma determinada região, é possível utilizar técnicas para a visualização desses dados de forma que qualquer usuário consiga obter algum tipo de conhecimento sobre o cenário local apenas analisando eles. Esse processo de conhecimento de uma área baseado na interpretação dos dados pode ser chamado de Inteligência Geográfica, e os sistemas que possibilitam obter esse tipo de informação são denominados Sistemas de Inteligência Geográfica. Esses sistemas possibilitam a união de vários conceitos já existentes na área da Computação, como Mineração de Dados e Visualização de Dados, juntamente com a área da Geoinformação, por meio dos chamados Sistemas de Informação Geográfica (SIG).



## MINERAÇÃO DE DADOS

O Processo de Descoberta de Conhecimento em Bases de Dados (*Knowledge Discovery in Database – KDD*) utilizado para extrair informações, verificar hipóteses e descobrir novos padrões nos dados, como proposto por Fayyad, Piatetsky-Shapiro, Smyth (1996), pode ser dividido em 04 etapas não lineares, segundo Miller e Han (2009). A primeira etapa é a Seleção, que é responsável por determinar os conjuntos de dados que serão utilizados e definir quais as varáveis de interesse para verificação. A segunda etapa é o Pré-Processamento, onde ocorre a “limpeza” dos dados. Nela, os ruídos e os dados duplicados são eliminados e os dados faltantes são determinados, além de serem definidas também outras bases de dados para obter dados complementares de interesse. Na terceira etapa, que é a de Redução e Projeção, o volume total de dados é reduzido, buscando assim trabalhar somente com um conjunto representativo dos dados gerais, o que facilita a manipulação posterior deles pelos algoritmos de mineração de dados em si. Na última etapa, é obtido o resultado das manipulações e esses valores são interpretados em relação ao contexto geral do problema.

A mineração de dados ocorre na etapa final do processo. De acordo com De Amo (2004), as técnicas existentes na mineração de dados permitem a resolução de problemas de descrição, classificação, regressão, predição, associação e clusterização, sendo que essa última tem o objetivo principal de buscar agrupar possíveis valores com base em algum padrão identificado pelo algoritmo. A Figura 1 ilustra o processo de KDD.

Diagrama

Descrição gerada automaticamente com confiança média

Figura 1. Representação do processo de KDD

## ALGORITMOS DE AGRUPAMENTO

Das várias tarefas que a Mineração de Dados se propõe a resolver, destaca-se o Agrupamento. De acordo com Jain et al. (1999), considerando um conjunto de dados, as técnicas voltadas para tal tarefa buscam gerar agrupamentos ou *clusters* baseados na similaridade dos elementos contidos em um mesmo grupo. Essa similaridade referida é um critério que define o quanto dois ou mais elementos são semelhantes e, consequentemente, devem pertencer a um mesmo conjunto gerado.

Utilizando o aprendizado não-supervisionado, isto é, não sendo necessário fornecer um conjunto prévio de dados para treinamento, fazendo com que o algoritmo seja executado direto sobre o conjunto de dados de interesse, considerando os algoritmos existentes, os mais tradicionais, segundo CASSIANO (2014), podem ser classificados como na Figura 2 a seguir. No primeiro nível da classificação, os algoritmos se dividem em relação a abordagem: hierárquica ou particional. A abordagem hierárquica, segundo ZHANG et al. (1996, apud CASSIANO, 2014), se caracteriza por manter o par de dados mais próximo juntos. Já a abordagem particional divide a base de dados em k grupos, sendo o número k um valor informado pelo usuário, de acordo com ESTER et al. (1996, apud CASSIANO, 2014). No segundo nível, tem-se a subdivisão da abordagem de acordo com a forma de medição da similaridade.

![Diagrama

Descrição gerada automaticamente]()

Figura 2. Classificação dos algoritmos de clusterização. Fonte: Jain et al. (1999), com tradução do autor.



# METODOLOGIA

A metodologia adotada nesse trabalho pode ser vista na Figura 2, a qual ilustra as principais etapas adotadas e agrupadas, assim como os seus produtos gerados.

Diagrama

Descrição gerada automaticamente

Figura 3. Visão geral da metodologia do trabalho

Na etapa 1 é realizada a coleta de dados. Visando a utilização de dados de diversas fontes, é realizada uma busca por dados de características socioeconômicas e ambientais, buscando assim a pluralidade da informação levantada. Para o exemplo proposto, utilizou-se dados do Instituto Brasileiro de Geografia e Estatística (IBGE) e do Cadastro Ambiental Rural (CAR). Em seguida, os principais atributos de interesse são escolhidos pelo usuário. Essa etapa pode contar ainda com a ajuda de um usuário especialista, o qual saberá definir quais são as informações mais relevantes a serem consideradas dado o problema. Caso haja dados privados, esses podem ser também utilizados pelo usuário.

Em seguida, na etapa 2, ocorre o tratamento dos dados. Os dados categóricos são convertidos para dados numéricos e os dados numéricos são adequados conforme o algoritmo de clusterização a ser utilizado. Dessa etapa, é obtido como resultado a matriz de dados a ser utilizada como fonte de dados para o algoritmo.

Na etapa 3, uma técnica de análise de componentes multivariadas é utilizada. Essa etapa é importante para a metodologia adotada pois possibilita que a matriz multidimensional de dados seja convertida para uma matriz bidimensional, possibilitando assim a plotagem dos dados em um gráfico de dispersão bidimensional, por exemplo. Nesse caso, escolheu-se o *Principal Component Analysis* (PCA) como algoritmo para redução da dimensionalidade. Após isso, é realizada a escolha e aplicação de um algoritmo de clusterização, além da determinação do centroide de cada agrupamento formado. Para o exemplo, foi escolhido o algoritmo *Kmeans* por ser um algoritmo clássico e bem-visto na literatura. Por escolher essa técnica, foi necessário a adoção também de um método para a escolha ideal do número de agrupamentos K. Para isso, usou-se o *Silhouette Score*.

Por fim, na etapa 4, logo após a geração de todos os agrupamentos e a determinação de seus centroides, ocorre a plotagem dos dados em um gráfico. Dessa forma, os *clusters* formados podem ser vistos, assim como a distância *inter-clusters* e *intra-clusters*. Nesta etapa é definido o ponto *intra-cluster* mais próximo do centroide determinado. Assim, esse ponto é o elemento mais indicado para ser um representante do agrupamento, ou seja, um elemento que possuirá as propriedades mais características do conjunto a que pertence. Por fim, é realizada a análise dos resultados obtidos. Dessa forma, para cada agrupamento formado, ocorre a análisepor meio do cálculo da média e mediana, e a plotagem desses valores em um *heatmap* para visualização geral dos dados. Assim, é possível identificar os principais atributos que levaram os dados àquele agrupamento e a consequente sugestão de priorização da alocação dos recursos disponíveis nesses mesmos atributos, buscando assim potencializar a recuperação da área atingida pelos desastres.



# CRONOGRAMA

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Atividades | Mês 01 | Mês 02 | Mês 03 | Mês 04 | Mês 05 | Mês 06 | Mês 07 | Mês 08 | Mês 09 | Mês 10 | Mês 11 | Mês 12 |
| Revisão Bibliográfica | X | X | X | X |  |  |  |  |  |  |  |  |
| Propor Metodologia |  |  | X | X | X | X |  |  |  |  |  |  |
| Validação da Metodologia |  |  |  |  |  | X | X |  |  |  |  |  |
| Propor Sistema |  |  |  |  |  |  | X | X | X |  |  |  |
| Escrita e Submissão do Artigo 01 |  |  |  | X | X | X |  |  |  |  |  |  |
| Desenvolvimento Sistema |  |  |  |  |  |  |  | X | X | X |  |  |
| Escrita e Submissão do Artigo 02 |  |  |  |  |  |  | X | X | X |  |  |  |
| Escrita e Submissão do Artigo 03 |  |  |  |  |  |  |  | X | X | X |  |  |
| Escrita da Dissertação |  |  |  |  |  |  |  |  | X | X | X |  |
| Efetuar Correções Sugeridas pela Banca |  |  |  |  |  |  |  |  |  |  | X | X |

# ORÇAMENTO

|  |  |  |
| --- | --- | --- |
| Especificação da Despesa | Valor (R$) | Fonte de Recursos |
| 1. PESSOAL |  |  |
| * 1. Ajuda de custo para realização da pesquisa | 18.000,00 | Recursos próprios |
| 1. SERVIÇOS DE TERCEIROS |  |  |
| 2.1- Tradução | 1.000,00 | Recursos próprios |
| 2.2- Gastos com viagens | 3.600,00 | Recursos próprios |
|  |  |  |
| TOTAL | 22.600,00 |  |

# REFERÊNCIAS

PEREIRA, Michel Teixeira et al. Acidente ambiental em Mariana-MG: um estudo dos impactos econômicos e financeiros nas empresas responsabilizadas. Ágora: revista de divulgação científica, v. 24, p. 44-66, 2019.

GODOY, Sandro Marcos; DIAS, Mateus Benites. O desastre Ambiental de Mariana e o papel da Fundação Renova na reparação dos danos. Direito e Desenvolvimento, v. 12, n.1, p. 37-48, 2021

BARBOSA, Francisco Antônio Rodrigues et al. O desastre de Mariana e suas consequências sociais, econômicas, políticas e ambientais: porque evoluir da abordagem de Gestão dos recursos naturais para Governança dos recursos naturais?. Arquivos do Museu de História Natural e Jardim Botânico da UFMG, v. 24, n. 1-2, 2015

CÂMARA, Gilberto.; MEDEIROS, José S. de. Princípios Básicos em Geoprocessamento. In: ASSAD, E. D., SANO, E. E. Sistema de informações geográficas: aplicações na agricultura. 2. ed. Brasília, Embrapa–SPI; Embrapa–CPAC, 2003. p. 3-11

MILLER, Harvey J.; HAN, Jiawei. Geographic data mining and knowledge discovery. CRC press, 2009.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. From data mining to knowledge discovery in databases. AI magazine, v. 17, n. 3, p. 37-37, 1996.

DE AMO, Sandra. Técnicas de mineração de dados. Jornada de Atualizaçao em Informatica, 2004.

JAIN, Anil K.; MURTY, M. Narasimha; FLYNN, Patrick J. Data clustering: a review. ACM computing surveys (CSUR), v. 31, n. 3, p. 264-323, 1999.

CASSIANO, Keila Mara. Análise de Séries Temporais Usando Análise Espectral Singular (SSA) e Clusterização de Suas Componentes Baseada em Densidade. Pontifícia Universidade Católica do Rio de Janeiro, 2014.

Entenda sobre inteligência geográfica para a indústria e como a Geofusion atua. Disponível em <https://www.geofusion.com.br/blog/inteligencia-geografica-industia/>. Acesso em 25/09/2022.

Inteligência Geográfica: saiba como expandir o seu negócio. Disponível em <https://blog.neoway.com.br/inteligencia-geografica/>. Acesso em 25/09/2022

# RESUMO

**TÍTULO DO PROJETO**

Um arcabouço computacional para suportar a tomada de decisões de especialistas em cenários de desastres ambientais utilizando dados públicos

**EQUIPE**

Julio Cesar Soares dos Reis

Jugurta Lisboa Filho

Carlos Henrique Tavares Brumatti

**INTRODUÇÃO**

Soluções computacionais podem ser úteis para apoiar especialistas no processo de tomada de decisão relacionado a essa alocação eficinte dos recursos financeiros. Uma das maneiras existentes é através do processo de Descoberta de Conhecimento em Bases de Dados (do inglês, *Knowledge Discovery in Database –* KDD), principalmente com a etapa de Mineração de Dados desse processo.

**OBJETIVOS**

O objetivo geral deste trabalho é propor uma metodologia voltada para o agrupamento de cidades atingidas por desastres ambientais, buscando assim melhorar a alocação de recursos financeiros utilizados na recuperação dessas áreas afetadas, fornecendo assim suporte ao processo de tomada de decisão pelos responsáveis. Além disso, espera-se a construção de uma ferramenta que possa utilizar dados públicos e privados para a realização das etapas da metodologia proposta via uma interface interativa

**METODOLOGIA**

Aplicar algoritmos de Mineração de Dados em dados ambientais georreferenciados a fim de agrupar cidades afetadas por desastres ambientais, estabelecendo assim uma metodologia que pode ser adotada no suporte a tomada de decisão por usuários especialistas no problema. Essa interação do usuário com a metodologia ocorrerá por meio de um sistema desenvolvido para essa finalidade.

**PALAVRAS-CHAVE**

Sistema de Inteligência Geográfica; Mineração de Dados; Agrupamento; Dados georreferenciados

**ÁREA DE CONHECIMENTO**

1.03.00.00-7 – Ciência da Computação

1.03.03.04-9 – Sistemas de Informação

**LINHA DE PESQUISA**

DPI-016 - Sistemas de Informação

DPI-046 - Mineração de Dados