#### UNIVERSIDADE FEDERAL DE VIÇOSA

#### CENTRO DE CIÊNCIAS EXATAS E TECNOLÓGICAS

#### DEPARTAMENTO DE INFORMÁTICA

CARLOS HENRIQUE TAVARES BRUMATTI

##### MINERAÇÃO DE DADOS GEOGRÁFICOS: ANÁLISE DO CADASTRO AMBIENTAL RURAL E OUTRAS BASES DE DADOS PÚBLICOS COMO FERRAMENTAS DE APOIO À TOMADA DE DECISÃO EM PROBLEMAS AMBIENTAIS LIGADOS À BACIAS HIDROGRÁFICAS

VIÇOSA

2021

CARLOS HENRIQUE TAVARES BRUMATTI

##### MINERAÇÃO DE DADOS GEOGRÁFICOS: ANÁLISE DO CADASTRO AMBIENTAL RURAL E OUTRAS BASES DE DADOS PÚBLICOS COMO FERRAMENTAS DE APOIO À TOMADA DE DECISÃO EM PROBLEMAS AMBIENTAIS LIGADOS À BACIAS HIDROGRÁFICAS

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

2021

**RESUMO**

**Palavras-chave:** Sistemas de Informação Geográfica; Mineração de Dados; Clusterização; Bacias Hidrográficas

**Sumário**

[1. INTRODUÇÃO 5](#_Toc106976311)

[2. OBJETIVOS 5](#_Toc106976312)

[2.1. OBJETIVO GERAL 5](#_Toc106976315)

[2.2. OBJETIVOS ESPECÍFICOS 5](#_Toc106976316)

[3. REFERENCIAL TEÓRICO 5](#_Toc106976317)

[3.1 SISTEMAS DE INFORMAÇÃO GEOGRÁFICA (SIG) 5](#_Toc106976319)

[3.2 ALGORITMOS DE AGRUPAMENTO 5](#_Toc106976322)

[3.3 PADRÕES DE DADOS ESPACIAIS 5](#_Toc106976327)

[4. METODOLOGIA 5](#_Toc106976328)

[5. RESULTADOS ESPERADOS 7](#_Toc106976329)

[5.1. REQUISITOS FUNCIONAIS E NÃO-FUNCIONAIS 7](#_Toc106976334)

[6. CRONOGRAMA 7](#_Toc106976335)

[7. ORÇAMENTO 7](#_Toc106976336)

[8. REFERÊNCIAS 7](#_Toc106976337)

# INTRODUÇÃO

# OBJETIVOS



## OBJETIVO GERAL

O objetivo geral deste trabalho é propor uma metodologia voltada para o agrupamento de cidades próximas às bacias hidrográficas 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:

* Identificação dos principais algoritmos de clusterização e das etapas necessárias para suas aplicações.
* Criação de um conjunto de dados para teste.
* Aplicação dos algoritmos de clusterização identificados no ocnjunto de dados de teste.
* Avaliação dos agrupamentos formados resultantes da aplicação dos algoritmos de clusterização, buscando identificar um padrão nos dados de um mesmo agrupamento.
* Projetar e implementar um Sistema de Informação Geográfica (SIG), 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 três tópicos, sendo eles: Sistemas de Informação Geográfica (SIG), Algoritmos de Agrupamento e Padrões de Dados Espaciais.



## SISTEMAS DE INFORMAÇÃO GEOGRÁFICA (SIG)



## ALGORITMOS DE AGRUPAMENTO



## PADRÕES DE DADOS ESPACIAIS

# METODOLOGIA

A metodologia adotada nesse trabalho pode ser vista na Figura a seguir. Nela, ilustra-se as principais etapas adotadas, assim como os seus produtos gerados.

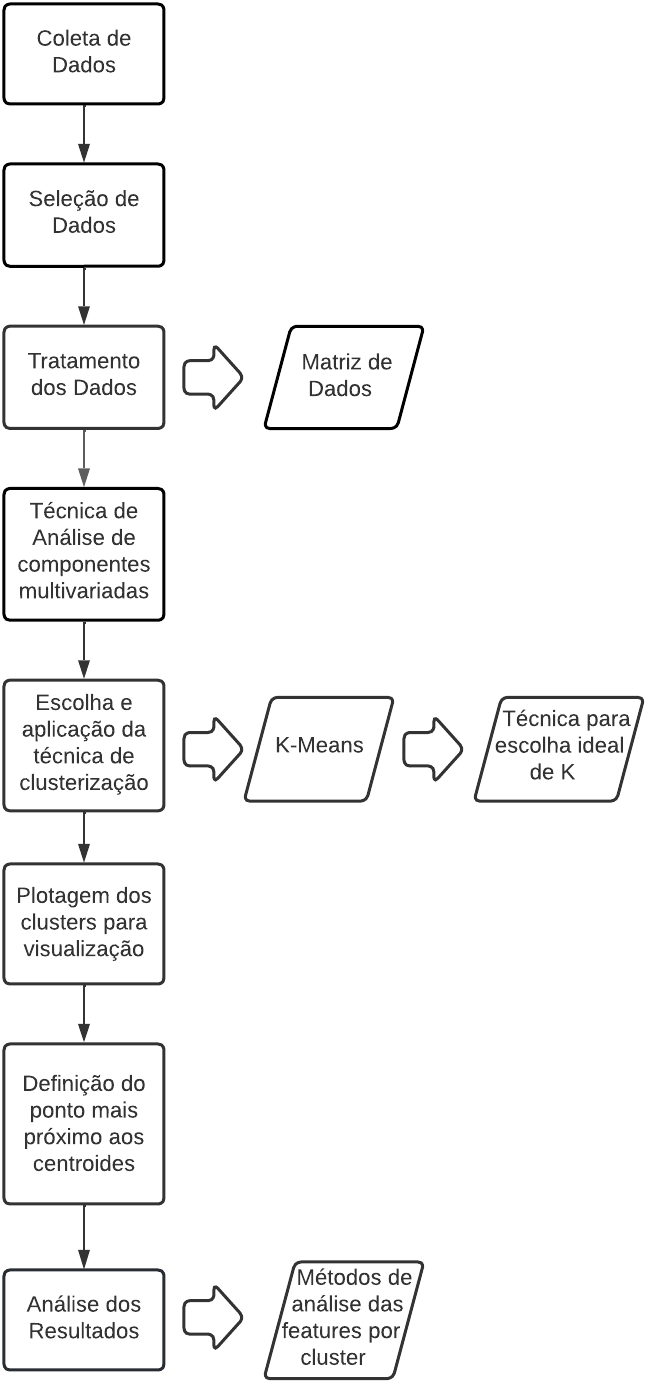


Figura 1. Visão geral da metodologia do trabalho

Na primeira etapa, é realizado a coleta de dados. A fim de serem utilizados dados de diversas fontes, é realizado uma busca por dados de caracteristicas socioeconômicas e ambientais, buscando assim a pluralidade da informação levantada. Para o exemplo inicial proposto, utilizou-se dados do Instituto Brasileiro de Geografia e Estatística (IBGE) e do Cadastro Ambiental Rural (CAR).

Já na segunda etapa, as principais *features* de interesse devem ser escolhidas 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. Para a terceira etapa, ocorrerá o tratamento dos dados. Os dados categóricos devem ser convertidos em dados numéricos e os dados numéricos devem ser adequados conforme o algoritmo de clusterização a ser utilizado. Dessa etapa, é obtido como resultado a matriz de dados a ser utilizado como a fonte de dados para o algoritmo. Na quarta etapa, irá ser utilizado uma técnica de análise de componentes multivariadas. Essa etapa é importante para a metodologia adotada pois possiblita 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 para um exemplo inicial o *Principal Component Analysis* (PCA) como algoritmo para redução da dimensionalidade.

Na quinta etapa é realizada a escolha e aplicação de um algoritmo de clusterização, além da determinação do centroide de cada agrupamento formado. Para o exemplo inicial proposto, 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*.

Na sexta etapa, 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*. Na etapa seguinte, é definido o ponto *intra-cluster* mais próximo do centroide determinado. Dessa forma, esse ponto é o elemento mais indicado para ser um representante do agrupamento, ou seja, um elemento que possuirá as propriedades mais caracteristicas do conjunto a que pertence. Por fim é realizado a análise dos resultados obtidos. Dessa forma, para cada agrupamento formado, ocorre a análise das *features* por 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.

# RESULTADOS ESPERADOS



## REQUISITOS FUNCIONAIS E NÃO-FUNCIONAIS

# CRONOGRAMA

# ORÇAMENTO

# REFERÊNCIAS