#### 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

**Sumário**

[1. INTRODUÇÃO 4](#_Toc89764860)

[2. OBJEIVOS 7](#_Toc89764861)

[3. OBJETIVOS ESPECÍFICOS 7](#_Toc89764862)

[4. REFERENCIAL TEÓRICO 7](#_Toc89764863)

[5. TRABALHOS RELACIONADOS 11](#_Toc89764864)

[6. METODOLOGIA 11](#_Toc89764865)

[7. CRONOGRAMA 12](#_Toc89764866)

[8. ORÇAMENTO 13](#_Toc89764867)

[9. REFERÊNCIAS 14](#_Toc89764868)

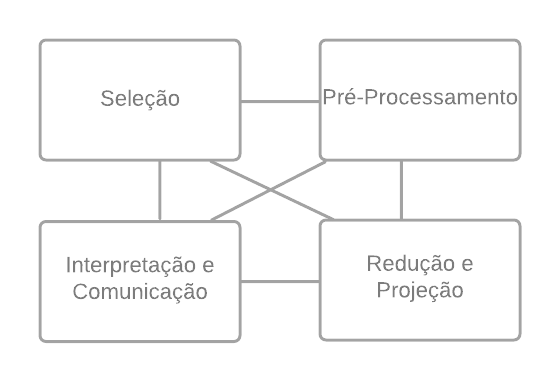
# INTRODUÇÃO

O desastre ambiental causado pelo rompimento da barragem de Fundão, em Mariana (MG), no dia 05 de novembro de 2015[[1]](#footnote-1), 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. Estas consequências não se restringem somente às questões ambientais, mas sim socioeconômicas, políticas e humanas. 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.

Uma das áreas atingidas com o desastre foi a Bacia Hidrográfica do Rio Paraopeba (BHRP) – Minas Gerais, Brasil. Abrangendo uma área de 12.054,25 km² e contando com um total de 1.318.885 milhão de habitantes (IBGE/CENSO 2010) distribuídos ao longo de 48 municípios, a BHRP é um dos principais afluentes da Bacia Hidrográfica Federal do Rio São Francisco, fazendo parte da chamada Circunscrição Hidrográfica SF3, conforme a Deliberação Normativa CERH-MG nº 66, de 17 de novembro de 2020. As principais atividades econômicas desses municípios giram em torno desse recurso hídrico (IGAM, 2013), destacando-se as atividades extrativistas, industriais e agropecuárias.

Com o desastre ocorrido, parte das atividades econômicas foram comprometidas, uma vez que as áreas de ocorrência delas foram atingidas diretamente ou indiretamente. Buscando auxiliar no processo de retomada dessas atividades, dados públicos disponíveis podem ser utilizados no processo de tomada de decisão futuro. Como exemplo desses dados, tem-se os dados ambientais do Cadastro Ambiental Rural (CAR). Os dados referentes às propriedades rurais devem ser informados de maneira eletrônica por todos aqueles que possuem uma propriedade rural ou que são responsáveis legais pelas mesmas. Sendo assim, estes dados possuem a capacidade de descrever quantitativamente e qualitativamente uma dada região rural. Utilizando técnicas adequadas, é possível aplicar o Processo de Descoberta de Conhecimento em Bases de Dados (*Knowledge Discovery in Database – KDD*) a fim de extrair informações, verificar hipóteses e descobrir novos padrões nos dados, como proposto por Fayyad, Piatetsky-Shapiro, Smyth (1996).

Miller e Han (2009) definem o processo de KDD como sendo constituído por quatro etapas não lineares: Seleção, Pré-Processamento, Redução e Projeção, e Interpretação e Comunicação dos Dados, conforme a Figura 01 abaixo:



(Figura 01: Etapas do KDD segundo Miller e Han(2009))

A primeira etapa é a Seleção. Ela 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, e é 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.

Hand (1999), por outro lado, definiu a mineração de dados como uma etapa independente no processo de KDD, consistindo em um conjunto de ideias, ferramentas e métodos de várias outras áreas, como a Estatística e o Aprendizado de Máquina, por exemplo, para a aplicação em um conjunto de dados. Dessa forma, a mineração de dados é um conceito multidisciplinar, voltado para o entendimento do problema e o uso das melhores técnicas disponíveis.

Com base nos conceitos apresentados até agora, esse trabalho busca aplicar as técnicas de mineração de dados sobre os dados disponíveis no CAR em relação à BHRP, buscando assim verificar como ocorre a distribuição das principais coberturas vegetais e hídricas ao longo da bacia e a relação existente entre elas e os aspectos socioeconômicas da região, evidenciando assim a importância da bacia para as cidades locais. Com os resultados, espera-se que possam ser úteis para o processo de tomada de decisão relativos à priorização de áreas para o reflorestamento e recuperação dos recursos hídricos dado o impacto socioeconômico local, além de se ter uma visão mais ampla da região.

# OBJETIVOS

* 1. **OBJETIVO GERAL**
  2. **OBJETIVOS ESPECÍFICOS**
* Construir uma ferramenta com mapa e filtros interativos que permita a visualização dos dados antes e depois da aplicação dos algoritmos de mineração de dados.
* Usar técnicas de *Explainable IA (XAI)* para determinar as variáveis mais importantes consideradas pelos algoritmos.
* Determinar possíveis áreas de interesse para a alocação de recursos visando a recuperação ambiental e socioeconômica dessas localidades.

# REFERENCIAL TEÓRICO

Esta seção apresenta os principais conceitos relacionados ao projeto. Dessa forma, foi estruturado em quatro tópicos, sendo eles: cadastro ambiental rural; sistema nacional de cadastro ambiental rural; mineração de dados e dados geoespaciais. Aqui, discute-se brevemente fonte dos dados explorada e como esses dados são disponibilizados para a sociedade, assim como qual é o tipo de dados tido como objeto de trabalho. Além disso, como se pretende utilizar técnicas de mineração de dados, é apresentado uma visão geral sobre elas.

* 1. CADASTRO AMBIENTAL RURAL (CAR)

Com os objetivos de auxiliar o planejamento ambiental e econômico, o controle e o monitoramento das áreas rurais e a recuperação de áreas degradadas (De Alcântara Laudares, Da Silva e Borges 2014), o Governo Brasileiro definiu por meio da lei nº 12.651, de 25 de maio de 2012, a criação do Cadastro Ambiental Rural (CAR[[2]](#footnote-2)). O registro é obrigatório para todos os imóveis rurais e é composto de um conjunto de informações disponibilizadas pelo próprio proprietário rural.

(Geração do identificador da propriedade)

Os dados disponibilizados no CAR representam um conjunto de informações ambientais sobre a propriedade rural subdivididas em 18 categorias, conforme a Tabela 01 a seguir:

|  |  |
| --- | --- |
| **Categoria** | **Descrição** |
| Área de Proteção Permanente (APP) |  |
| Área de Altitude Superior a 1800 m |  |
| Área Consolidada |  |
| Área com Declividade Superior a 45 graus |  |
| Área Imóvel |  |
| Área Pousio |  |
| Área de Topo de Morro |  |
| Banhado |  |
| Borda de Chapada |  |
| Hidrografia |  |
| Manguezal |  |
| Nascente Olho D’Água |  |
| Reserva Legal |  |
| Restinga |  |
| Servidão Administrativa |  |
| Uso Restrito |  |
| Vegetação Nativa |  |
| Vereda |  |

(Tabela 01: Categorias das informações disponibilizadas pelo CAR)

* 1. SISTEMA NACIONAL DE CADASTRO AMBIENTAL RURAL (SICAR)

O Sistema Nacional de Cadastro Ambiental Rural (SICAR) é o sistema eletrônico para gerenciamento dos dados do CAR. Criado pelo Decreto nº 7.830, de 17 de outubro de 2012, é através dele que o usuário consegue realizar o cadastro dos seus dados por meio de uma interface web. Dessa forma, a aplicação pode ser classificada como sendo um exemplo de Sistema de Informação Geográfica Voluntário (*Volunteered Geographic Information –* VGI), pois, segundo Goodchild (2009), **sistemas VGI são...**

Além disso, é no SICAR onde ficam disponíveis para *download*, em formato *shapefile* (shp[[3]](#footnote-3)), todos os dados informados. Assim, é possível filtrar os dados por estado e município, e realizar uma cópia local dos dados lá presentes.

* 1. MINERAÇÃO DE DADOS

A mineração de dados é uma das etapas finais do processo de Descobrimento de Informação em Bases de Dados (*Knowledge Discovery in Databases – KDD)*. 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 agrupamento (clusterização). Uma breve descrição das técnicas pode ser vista na Tabela 01 a seguir:

|  |  |
| --- | --- |
| **Problema** | **Descrição** |
| Descrição | Descrição dos padrões e tendências observadas. Comumente utilizada na análise exploratória dos dados. |
| Classificação | Identifica um determinado conjunto de dados e determina a qual dos grupos disponíveis ele pertence. |
| Regressão | Com base em um conjunto de dados do tipo numérico, formula uma equação que descreve eles. |
| Predição | Com base em um conjunto de dados, tenta estimar quais os possíveis valores que tendem a aparecer. |
| Associação | Busca identificar dois ou mais valores que apresentam uma relação de dependência entre si. |
| Agrupamento (Clusterização) | Busca agrupar possíveis valores com base em algum padrão identificado ao longo da execução. |

(Tabela 01: Descrição dos possíveis problemas resolvidos com a mineração de dados)

Para cada um dos problemas apresentados acima, há um conjunto de algoritmos específicos que podem solucionar eles. Há a possibilidade também de um mesmo algoritmo resolver mais de um tipo de problema. Como exemplo de técnicas comumente utilizadas, pode-se destacar:

* Mineração de Itens Frequentes
* Árvores de Decisão
* Classificação Bayesiana
* Classificação Baseada em Regras
* Redes Neurais
* SVM (*Support Vector Machine*)
* Classificação por Regras de Associação
* Aprendizado Tardio
* Algoritmo Genético
* Conjuntos Aproximados
* Conjuntos Nebulosos
* Regressão Linear
* Regressão Não-Linear
* Métodos de Particionamento (K-means e K-Medoids)
* Métodos Hierárquicos (Aglomerativos e Divisivos)
* Métodos Baseados na Densidade
* Métodos Baseados em Grade
* Métodos Baseados em Modelos

A mineração de dados aplicado em dados geoespaciais pode ser classificada como sendo uma mineração de estruturas complexas, ou seja, em dados de formato não-estruturado, de acordo com Camilo e Silva (2009). O autor afirma também que essa classificação ocorre porque a concepção da mineração de dados é para atuar em tipos de dados estruturados, como banco de dados ou arquivos em formato csv.

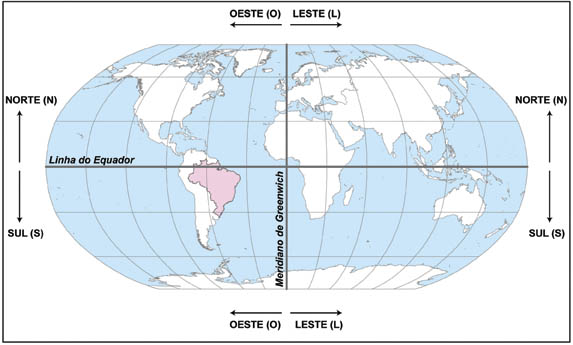
* 1. DADOS GEOESPACIAIS

Dados Geoespaciais (ou dados georreferenciados[[4]](#footnote-4)) são um conjunto específico de dados que contém além de informações sobre um evento ou objeto, a presença de um componente espacial que representa a localização daquele fenômeno em relação à superfície da Terra. Esse componente pode ser um ou mais pares ordenados do tipo latitude/ longitude e, associado a ele, há uma geometria, que pode ser classificada como ponto, linha ou polígono. Além dessas duas características, o dado geoespacial apresenta ainda um atributo temporal, representando o intervalo do tempo de ocorrência.

* + 1. **LOCALIZAÇÃO**

(citar algum autor)

A localização de um dado ou fenômeno no espaço pode ser representada através de um par de valores: latitude e longitude. Esses valores são definidos considerando o ângulo formado entre a posição do objeto e a linha do equador e o meridiano de Greenwich, respectivamente, sendo ambos chamados de linhas imaginárias. A disposição dessas linhas no globo terrestre pode ser vista na Figura 0x seguinte:



(Figura 0x: Representação das linhas imaginárias no globo terrestre. Fonte: https://educa.ibge.gov.br/images/vamoscontar/atividades/mapa\_mundo\_coordenadas.jpg. Acesso em: 10 jan 2022)

* + 1. **SISTEMA DE REFERÊNCIA DE COORDENADAS**

(citar algum autor)

(exemplo: Universal Transverse Mercator (UTM))

(sistema de coordenadas projetadas e não-projetadas)

* + 1. **FORMATO DOS DADOS GEOESPACIAIS**

(citar algum autor)

(dados vetoriais e dados rasterizados)

# TRABALHOS RELACIONADOS

(Mineração de dados espaciais: queimadas e faltas em linhas de transmissão)

(Quaisquer trabalhos que utilizam o CAR)

# METODOLOGIA

(dados = CAR, PIB dos municípios – IBGE, outros dados que buscam descrever os aspectos socioeconômicos e ambientais da região

Clusterização, K-Means, PCA , *silhouette score* e método *elbow*

QGIS para determinar o número de *features* por categoria de dados do CAR

)

Foram obtidos os dados do CAR para os 48 municípios que compõem a BHRP no dia 21 de outubro de 2021. A importância dessa bacia ocorre pelo fato dela ser um dos principais afluentes do rio São Francisco, além dela apresentar extrema relevância para o abastecimento público de água da região. Sendo assim, a primeira etapa de trabalho consiste na sumarização dos dados, buscando identificar e determinar a quantidade de coberturas vegetais e hídricas existentes em cada munícipio, assim como a área total ocupada. Essa etapa é importante para conhecimento inicial dos dados e resumo das informações presentes. Os dados são obtidos de forma manual e inseridos em um arquivo do formato CSV. Em seguida, será feita a análise descritiva para os dados obtidos, buscando-se entender como os dados se manifestam na região de interesse. Essa primeira análise será importante também para servir de base para futuras análises quantitativas e qualitativas acerca dos dados. No final do projeto, após todas as etapas da mineração de dados, será implementado um visualizador para os dados geoespaciais e para os resultados obtidos. Dessa forma, como um dos objetivos desse trabalho é de servir como ferramenta de apoio à tomada de decisão, essa etapa de interação entre os usuários e os dados é de extrema importância para o objetivo proposto.

# CRONOGRAMA

A Tabela 02 apresenta uma proposta de cronograma para o desenvolvimento do projeto. Nele, foi considerado como base a realização de todas as etapas do KDD e foi considerado também o período de 12 meses para a realização da pesquisa.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Atividade** | **Jan** | **Fev** | **Mar** | **Abr** | **Mai** | **Jun** | **Jul** | **Ago** | **Set** | **Out** | **Nov** | **Dez** |
| Revisão de literatura | x | x | x | x | x | x | x | x | x | x |  |  |
| Seleção | x | x | x |  |  |  |  |  |  |  |  |  |
| Pré-processamento |  | x | x | x |  |  |  |  |  |  |  |  |
| Transformação |  |  |  |  | x |  |  |  |  |  |  |  |
| Mineração de Dados |  |  |  |  | x | x | x | x |  |  |  |  |
| Interpretação |  |  |  |  |  |  |  | x | x |  |  |  |
| Interface de Visualização |  |  |  |  |  |  | x | x |  |  |  |  |
| Escrita de Artigos e Submissões |  |  | x | x | x | x | x | x | x | x | x |  |
| Defesa de Dissertação |  |  |  |  |  |  |  |  |  |  |  | x |

(Tabela 02 – Cronograma do Projeto)

# ORÇAMENTO

A Tabela 03 apresenta uma estimativa de orçamento para a realização do projeto. Nela, está incluso os principais gastos que serão necessários no decorrer da pesquisa.

|  |  |  |
| --- | --- | --- |
| **Especificação das Despesas** | **Valor** | **Fonte de Recursos** |
| PESSOAL |  |  |
| Bolsa de Estudos | 18 x 1.500,00 = 27.000,00 | Fonte Própria |
| PUBLICAÇÃO |  |  |
| Traduções | 1.200,00 | Fonte Própria |
| Taxas de Publicações | 1.000,00 | Fonte Própria |
| Viagens para congressos | 1.200,00 x 3 = 3.600,00 | Fonte Própria |
| MATERIAIS |  |  |
| Material de Escritório | 100,00 | Fonte Própria |
| TOTAL | 32.900,00 | Fonte Própria |

(Tabela 03: Orçamento previsto)

# REFERÊNCIAS

CHRISTEN, Martin. **Geospatial Analysis for the Smart City**. In: BIG DATA BBQ, 2021, Mannheim. Disponível em: < https://github.com/martinchristen/bigdatabbq2021 >. Acesso em: 23 out. 2021.

WESTRA, Erik. **Python geospatial development**. Birmingham: Packt Publishing, 2010.

LISBOA F., J. Porto Alegre: CPGCC da UFRGS, 1995. 1995. 69p, (TI-491)

COUTO, Rodrigo de Souza. Estruturação, modelagem e implementação de banco de dados geoespaciais para o cadastro ambiental rural. 2017.

ARAUJO, Marcos Gabriel Pereira. Aplicação de técnicas de mineração de dados para análise climática no estado do Tocantins. 2021.

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.

PIB do setor agropecuário registrou alta de 5,7% no primeiro trimester. Disponível em:< https://www.gov.br/agricultura/pt-br/assuntos/noticias/pib-do-setor-agropecuario-registrou-alta-de-5-7-no-primeiro-trimestre>. Acesso em: 23 out. 2021

BORGES, Sérgio. O desastre da barragem de rejeitos em Mariana, Minas Gerais: aspectos socioambientais e de gestão na exploração de recursos minerais. **Cuadernos de Geografía: Revista Colombiana de Geografía**, v. 27, n. 2, p. 301-312, 2018.

ESPINDOLA, Haruf Salmen; NODARI, Eunice Sueli; SANTOS, Mauro Augusto dos. Rio Doce: riscos e incertezas a partir do desastre de Mariana (MG). **Revista Brasileira de História**, v. 39, p. 141-162, 2019

**CAR**

EMBRAPA TERRITORIAL. **Agricultura e preservação ambiental**: uma análise do cadastro ambiental rural. Campinas, 2020. Disponível em: < www.embrapa.br/car >. Acesso em:  28 out. 2021

**What is geospatial data ?** Disponível em: https://www.ibm.com/topics/geospatial-data. Acesso em: 05 novem. 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

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

**Bacia do Rio Paraopeba.**Disponível em: http://www.feam.br/component/content/article/901-bacia-do-rio-paraopeba. Acesso em: 10 novem. 2021.

**QUALIDADE DAS ÁGUAS SUPERFICIAIS DE MINAS GERAIS EM 2013.** Disponível em: http://www.igam.mg.gov.br/images/stories/qualidade\_aguas/2014/resumo-executivo-2013.pdf. Acesso em: 11 novem. 2021.

GOODCHILD, Michael F. Citizens as sensors: the world of volunteered geography. **GeoJournal**, v. 69, n. 4, p. 211-221, 2007

**DELIBERAÇÃO NORMATIVA CERH-MG Nº 66, DE 17 DE NOVEMBRO DE 2020**. Disponível em: http://www.siam.mg.gov.br/sla/download.pdf?idNorma=52900. Acesso em: 13 novem. 2021

HAND, David J. Statistics and data mining: intersecting disciplines. **Acm Sigkdd Explorations Newsletter**, v. 1, n. 1, p. 16-19, 1999.

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

DE ALCÂNTARA LAUDARES, Sarita Soraia; DA SILVA, Kmila Gomes; BORGES, Luís Antônio Coimbra. Cadastro Ambiental Rural: uma análise da nova ferramenta para regularização ambiental no Brasil. **Desenvolvimento e Meio Ambiente**, v. 31, 2014

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

CAMILO, Cássio Oliveira; SILVA, João Carlos da. Mineração de dados: Conceitos, tarefas, métodos e ferramentas. **Universidade Federal de Goiás (UFC)**, v. 1, n. 1, p. 1-29, 2009.

**OGC Standards.** Disponível em: https://www.ogc.org/docs/is. Acesso em: 17 novem. 2021

**Introdução às linhas imaginárias e coordenadas geográficas.** Disponível em: https://educa.ibge.gov.br/professores/educa-atividades/17650-linhas-imaginarias.html. Acesso em: 10 jan. 2022

1. https://mundoeducacao.uol.com.br/biologia/acidente-mariana-mg-seus-impactos-ambientais.htm [↑](#footnote-ref-1)
2. https://www.florestal.gov.br/inventario-florestal-nacional/?option=com\_content&view=article&id=74&Itemid=94 [↑](#footnote-ref-2)
3. https://docs.fileformat.com/gis/shp/ [↑](#footnote-ref-3)
4. https://www.ibm.com/topics/geospatial-data [↑](#footnote-ref-4)