Um Sistema de Inteligência Geográfica para Apoio à Tomada de Decisão em Cenários de Desastres Ambientais Envolvendo Bacias Hidrográficas

Omitido

1

Abstract. *Environmental disasters, like dam failures, cause impacts that go beyond the area of occurrence. From the region of origin to its arrival at sea, the sediments can cause both environmental and economic impacts. Searching for ways to help in the recovery of these degraded areas, this work proposes the development of a computational framework for the decision-making of specialists in this context. Using suitable techniques, it was possible to apply the Knowledge Discovery Process in Database (KDD) in order to group cities close to hydrographic basins, thus presenting the generation of knowledge groups.*

Resumo. *Desastres ambientais, como o rompimento de barragens, causam impactos que vão muito além da área de ocorrência. Da região de origem até a sua chegada ao mar, os resíduos podem causar tanto impactos ambientais quanto econômicos. Buscando formas de auxiliar na recuperação dessas áreas degradadas, neste trabalho é proposto o desenvolvimento de um arcabouço computacional para suportar a tomada de decisões de especialistas neste contexto. Utilizando técnicas adequadas, foi possível aplicar o Processo de Descoberta de Conhecimento em Bases de Dados (Knowledge Discovery in Database – KDD) a fim de agrupar cidades próximas às bacias hidrográficas, possibilitando assim a geração de grupos de cidades.*

1. Introdução

O desastre ambiental causado pelo rompimento da barragem de Fundão, em Mariana (MG), no dia 05 de novembro de 2015[1], evidenciou a íntima relação existente entre o meio biótico e o meio antrópico. Segundo Godoy e Dias (Godoy and 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 também socioeconômicas, políticas e humanas.

Ao longo desse percurso, o impacto ambiental foi mais significativo nas bacias hidrográficas existentes, uma vez que os rios que as compõem foram contaminados. Bacias hidrográficas como as do Rio Paraopeba e do Rio Doce foram as principais atingidas pelos resíduos contaminados.

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 (Barbosa et al. 2015), após o desastre, estima-se que as prefeituras das áreas

[1]http://www.meioambiente.mg.gov.br/component/content/article/13-informativo/2879-desastre-ambiental-em-mariana-e-recuperacao-da-bacia-do-rio-doce

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 à desastres ambientais. Assim, neste trabalho é proposto um arcabouço computacional que inclui, uma metodologia e um Sistema de Inteligência Geográfica que permite agrupar cidades atingidas por desastres ambientais e que são localizadas próximas à bacias hidrográficas.

O texto está organizado conforme detalhado a seguir. A Seção 2 apresenta a fundamentação teórica. Depois, na Seção 3 é descrita a metodologia elaborada. A Seção 4 discute brevemente a arquitetura da ferramenta proposta e, por fim, a Seção 5, apresenta a conclusão e os trabalhos futuros.

2. Fundamentação Teórica

O Processo de Descoberta de Conhecimento em Bases de Dados (do inglês, *Knowledge Discovery in Database* – KDD) pode ser utilizado para extrair informações, verificar hipóteses e descobrir novos padrões nos dados, conforme proposto em (Fayyad et al. 1996). Em suma, o KDD pode ser dividido em 4 etapas não lineares (Miller and Han 2009) principais, conforme apresentado na Figura 1.

A primeira etapa é a (1) `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 (2) `Pré-Processamento`, onde ocorre a "limpeza"e tratamento dos dados. Nela, os ruídos e os dados duplicados são eliminados e os dados faltantes são determinados, além de serem definidas também outras bases para obter dados complementares de interesse. Na terceira etapa, que é a de (3) `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 destes pelos algoritmos de mineração de dados. Na última etapa, (4) `Interpretação e Comunicação` 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 (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.

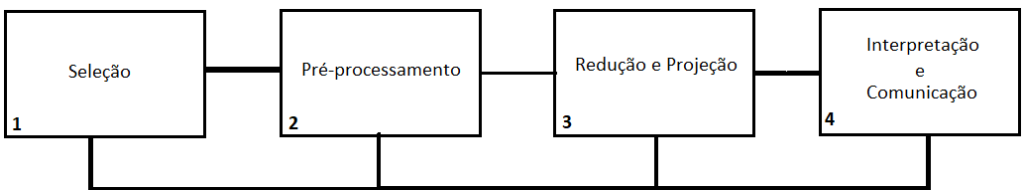

Figura 1. Representação do processo de KDD.

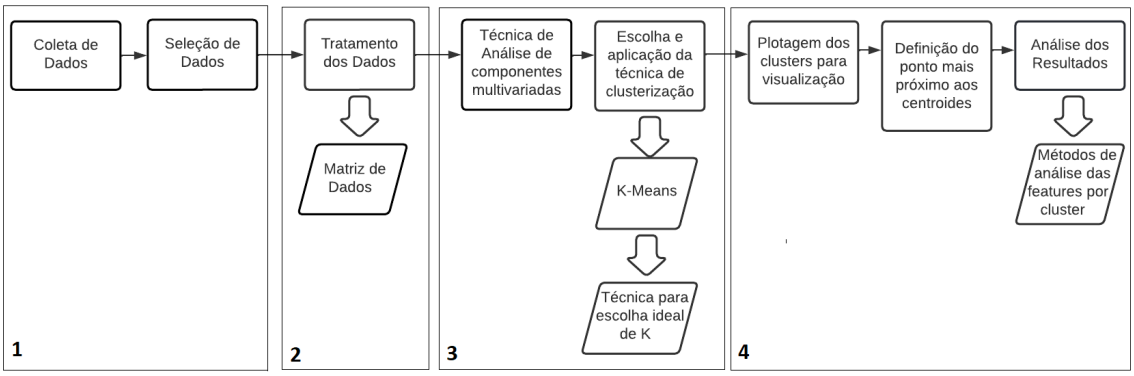

Figura 2. Visão geral da metodologia do trabalho.

3. Metodologia

Uma sumarização da metodologia proposta nesse trabalho é apresentada na Figura 2, a qual ilustra as principais etapas adotadas e agrupadas, assim como os seus produtos gerados.

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)[2] – também explorado em trabalhos anteriores como (Paes et al. 2009) – e do Cadastro Ambiental Rural (CAR)[3], até então pouco explorados. 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 possui o conhecimento necessário para definir quais são as informações mais relevantes a serem consideradas para um determinado problema. Caso haja dados privados, esses podem ser também utilizados pelo usuário. Neste contexto, é válido destacar que a ideia geral é que haja suporte para uma personalização da abordagem proposta conforme necessidade do usuário especialista.

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.

Depois, 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 trabalho, aplicou-se o *Principal Component Analysis* (PCA) (Bro and Smilde 2014) 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 *K-means* por ser um algoritmo clássico e bastante explorado na literatura (Likas et al. 2003). Por escolher essa técnica, foi necessário a adoção também de um método para a es-

[2]https://www.ibge.gov.br/
[3]https://www.car.gov.br/

colha ideal do número de agrupamentos K. Para isso, usou-se o *Silhouette Score* (Rousseeuw 1987), que de maneira geral captura a consistência dentro de um determinado agrupamento de dados. É importante mencionar que, embora tenham sido aplicados algoritmos e/ou técnicas específicas baseadas em critérios pré-definidos, a metodologia é generalizável. Em outras palavras, caso seja de interesse do usuário, os algoritmos e/ou técnicas explorados em alguns componentes podem ser alterados.

Finalmente, 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 análise por meio do cálculo da média e mediana, e a plotagem desses valores em um mapa de calor 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 potencializar a recuperação da área atingida pelos desastres.

4. Arquitetura do Sistema Proposto

A Figura 3 apresenta uma visão geral da arquitetura do sistema proposto, baseando-se na metodologia apresentada na Seção 3. De maneira geral, esta arquitetura pode ser dividida em dois módulos de implementação.

O Módulo 01 (Figura 3) consiste no *upload* dos dados de treinamento (*input*) através do *frontend* do usuário. Uma vez que é carregado no sistema, o usuário fica responsável por selecionar os principais atributos de interesse que serão utilizadas no restante do processo. O *backend* é responsável por realizar todas as etapas do KDD, conforme a metodologia proposta (Seção 3). Uma vez que os agrupamentos (i.e. *clusters*) são formados, os dados voltam para o *frontend* (ou *interface*) no formato *JavaScript Object Notation* (JSON). A seguir, ocorre a plotagem dos resultados em um mapa interativo. A Figura 4 ilustra um exemplo do resultado esperado dessa primeira etapa.

Já o Módulo 02 (Figura 3) da aplicação, permite a escolha e edição de dados geoespaciais por programas de terceiros, incluindo funcionalidades como salvar esses dados em um banco de dados espaciais, como o *Postgres*[4], por exemplo, que por sua vez é conectado a um servidor de mapas, como o Geoserver[5], fornecendo assim os dados nos formatos *Web Map Service* (WMS) e *Web Feature Service* (WFS), conforme especificações da *Open Geospatial Consortium* (OGC)[6], para o programa principal. Esses últimos dados carregados para o sistema permitem que arquivos espaciais estáticos, como *geojson*, *shapefiles* e outros, sejam consumidos pela aplicação, fornecendo mapas base personalizados pelo usuário, conforme a região de interesse.

Por fim, é válido mencionar que todo o sistema é proposto utilizando tecnologias

[4]https://www.postgresql.org/
[5]https://geoserver.org/
[6]https://www.ogc.org/

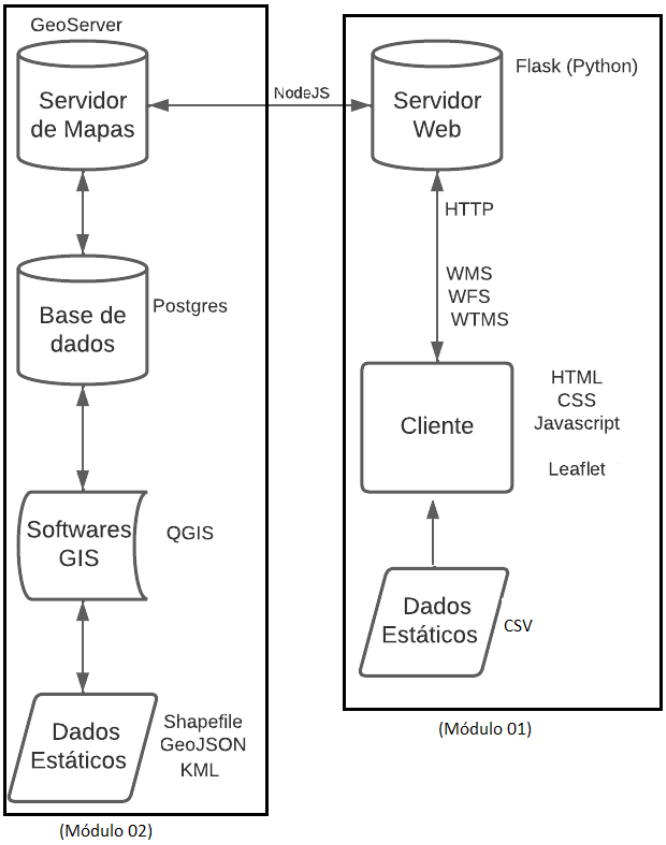

Figura 3. Visão geral da arquitetura do sistema.

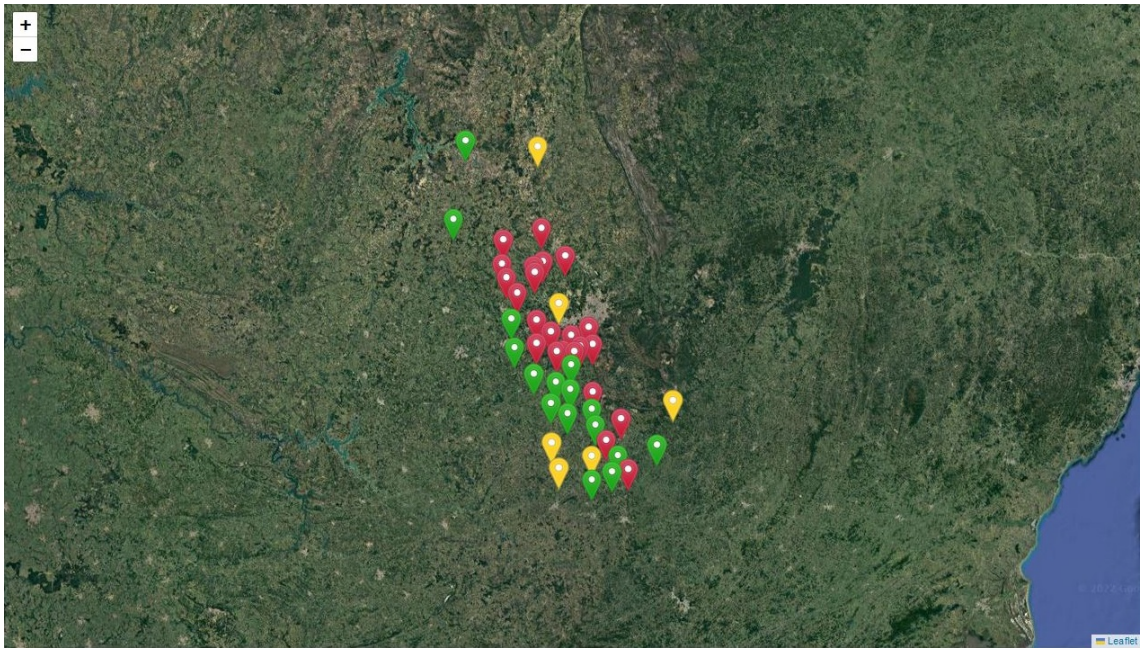

Figura 4. Plotagem dos agrupamentos obtidos a partir de um arquivo de entrada como exemplo.

opensource[7], além de toda a comunicação ocorrer de acordo com os principais padrões

[7]Programas gratuitos e livres construídos e aprimorados por uma comunidade fiel.

de dados existentes, permitindo assim a interoperabilidade entre o sistema aqui proposto outros sistemas existentes.

5. Conclusão

Este artigo apresentou um Sistema de Inteligência Geográfica que busca ser uma ferramenta para auxiliar os usuários responsáveis (i.e., especialistas) pelo processo de tomada de decisão, seja esse usuário pertencente a algum órgão público ou empresa privada, durante a ocorrência de desastres ambientais ligados à bacias hidrográficas e que causam impacto direto nas cidades próximas. Espera-se que que o arcabouço computacional proposto possa ser útil para apoiar decisões neste contexto, provendo aos especialistas informações importantes. Como trabalhos futuros, planeja-se realizar modificações na versão inicial da arquitetura/sistema propostos, principalmente em relação ao Módulo 02, a fim de tornar o funcionamento da arcabouço mais dinâmico, se comportando de maneira geral de acordo com os dados de entrada, além de permitir a seleção dos atributos de interesse pelo usuário e a visualização das principais características de cada agrupamento, de maneira personalizada.

Referências

[Barbosa et al. 2015] Barbosa, F. A. R., Maia-Barbosa, P. M., Nascimento, A. M. A., Rietzler, A. C., Franco, M. W., Paes, T. A., Reis, M., Moura, K. A. F., Dias, M. F., de Paula Ávila, M., et al. (2015). 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*, 24(1-2).

[Bro and Smilde 2014] Bro, R. and Smilde, A. K. (2014). Principal component analysis. *Analytical methods*, 6(9):2812–2831.

[De Amo 2004] De Amo, S. (2004). Técnicas de mineração de dados. *Jornada de Atualizaçao em Informatica*, page 26.

[Fayyad et al. 1996] Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, 17(3):37–37.

[Godoy and Dias 2021] Godoy, S. M. and Dias, M. B. (2021). O desastre ambiental de mariana e o papel da fundação renova na reparação dos danos. *Direito e Desenvolvimento*, 12(1):37–48.

[Likas et al. 2003] Likas, A., Vlassis, N., and Verbeek, J. J. (2003). The global k-means clustering algorithm. *Pattern recognition*, 36(2):451–461.

[Miller and Han 2009] Miller, H. J. and Han, J. (2009). *Geographic data mining and knowledge discovery*. CRC press.

[Paes et al. 2009] Paes, M., Lima, A. A., and Mattoso, M. (2009). Processamento de alto desempenho em consultas sobre bases de dados geoestatísticas usando replicação parcial. In *SBBD*, pages 241–255.

[Rousseeuw 1987] Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.