Boletim do Observatório Ambiental
Alberto Ribeiro Lamego

Essentia EDITORA IFFLUMINENSE

Artigo Original
e-ISSN 2177-4560
DOI: 10.19180/2177-4560.v14n12020p152-162

Submetido em: 20 fev. 2020
Aceito em: 09 mar. 2020

Avaliação da qualidade de água do Baixo Paraíba do Sul com técnicas de inteligência computacional

Maria Alice Manhães dos Santos https://orcid.org/0000-0003-0502-9980
Mestranda em Sistemas Aplicados à Engenharia e Gestão pelo Instituto Federal de Educação Ciência e Tecnologia Fluminense *Campus* Campos - Centro – Campos dos Goytacazes/RJ – Brasil. E-mail: mrlcmanhaes@gmail.com
José Luiz Lodi Júnior https://orcid.org/0000-0002-4326-2356
Mestrando em Sistemas Aplicados à Engenharia e Gestão pelo Instituto Federal de Educação Ciência e Tecnologia Fluminense - *Campus* Campos - Centro – Campos dos Goytacazes/RJ – Brasil. E-mail: jrlodi@hotmail.com
Milton Erthal Júnior https://orcid.org/0000-0002-9959-3568
Doutor em Produção Vegetal pela Universidade Estadual do Norte Fluminense Darcy Ribeiro. Professor do Instituto Federal de Educação, Ciência e Tecnologia Fluminense, *Campus* Campos-Guarus. Campos dos Goytacazes/RJ – Brasil. E-mail: miltonerthal@hotmail.com
Henrique Rego Monteiro da Hora https://orcid.org/0000-0001-7192-9245
Doutor em Engenharia de Produção pela Universidade Federal Fluminense. Coordenador Adjunto do Mestrado em Sistemas Aplicados à Engenharia e Gestão do Instituto Federal de Educação, Ciência e Tecnologia Fluminense -- *Campus* Campos - Centro – Campos dos Goytacazes/RJ – Brasil. E-mail: dahora@gmail.com

A grande demanda do ser humano pelos recursos naturais, com destaque para a água, gera um desequilíbrio na biodiversidade. As técnicas computacionais vêm sendo ferramentas aliadas nos diversos setores da gestão ambiental. O objetivo deste trabalho é investigar os componentes do índice de qualidade de água e suas interações, por meio das técnicas de mineração de dados. A base de dados utilizada dispõe de 172 registros, entre 2014 e 2018, obtidos nos diferentes pontos de amostragens do Rio Paraíba do Sul. Para aplicação da mineração de dados, foram utilizados os métodos de agrupamento e classificação. Os resultados indicam que os dados se dividem em dois grupos, e o atributo de maior influência é o coliforme termotolerante, indicando problemas relacionados à ausência de infraestrutura de saneamento no entorno da bacia hidrográfica.

Palavras-chave: Qualidade de água. Árvore de decisão. Processo de Descoberta do Conhecimento em Base de Dados.

Quality assessment of Baixo Paraíba do Sul water using computational intelligence techniques

The great demand of human being for natural resources, especially water, generates an imbalance in biodiversity. Computational techniques have been allied tools in different sectors of environmental management. The objective of this work is to investigate the components of the water quality index and their interactions, through data mining techniques. The database used has 172 records, between 2014 and 2018, obtained at the different sampling points of the Paraíba do Sul River. For the application of data mining, the methods of grouping and classification were used. The results indicate that the data are divided into two groups and the attribute with the greatest influence is the thermotolerant coliform, indicating problems related to the lack of sanitation infrastructure around the watershed.

Keywords: Water quality. Decision Tree. Process of Knowledge Discovery in Database.

Evaluación de la calidad del agua del Baixo Paraíba do Sul, con técnicas de inteligencia computacional

La gran demanda de los seres humanos por los recursos naturales, especialmente el agua, genera un desequilibrio en la biodiversidad. Las técnicas computacionales han sido herramientas aliadas en diferentes sectores de la gestión ambiental. El objetivo de este trabajo es investigar los componentes del índice de calidad del agua y sus interacciones, a través de técnicas de minería de datos. La base de datos utilizada tiene 172 registros, entre 2014 y 2018, obtenidos en los diferentes puntos de muestreo del Rio Paraíba do Sul. Para la aplicación de la minería de datos, se utilizaron los métodos de agrupación y clasificación. Los resultados indican que los datos se dividen en dos grupos y el atributo con mayor influencia es el coliforme termotolerante, lo que indica problemas relacionados con la falta de infraestructura de saneamiento alrededor de la cuenca hidrográfica.

Palabras clave: Calidade del agua. Árbol de decisiones. Proceso de descubrimiento de conocimiento en la base de datos.

1 Introdução

A água fornecida pelos corpos hídricos é extremamente importante para satisfazer diversas necessidades, tanto do próprio ecossistema, pela manutenção da biodiversidade, como também das atividades antrópicas, como irrigação, geração de energia, consumo, lazer, entre tantas outras. Todavia, com o aumento populacional e consequente crescimento da urbanização, muitas vezes sem planejamento, aliado ainda ao desenvolvimento das indústrias, ocorre a deterioração dos recursos hídricos nas diferentes bacias hidrográficas (CARPENTER, 1998; BILGEN, 2014; POORE; NEMECEK, 2018).

Nesse sentido, é importante que haja o monitoramento contínuo e uma compreensão adequada do comportamento dos parâmetros físico-químicos e microbiológicos que influenciam a qualidade de água (LOVETT *et al.*, 2007; KACHURIN; KOMASHCHENKO; MORKUN, 2015). A forma mais utilizada para classificar a qualidade de água de determinado curso d'água é o Índice de Qualidade de Água (IQA), em que os valores dos parâmetros de qualidade de água são convertidos em um único valor numérico (JABER; MOHSEN, 2001; EFENDI, 2016; NAUBI *et al.*, 2016). Tal aplicação facilita a expressão dos resultados, uma vez que o IQA classifica a qualidade da água em termos simples: muito ruim, ruim, média, boa ou excelente.

O Norte e Noroeste Fluminense abrigam importantes rios de domínio federal e estadual; em virtude disso, o órgão ambiental do estado, Instituto Estadual do Ambiente – INEA, vem realizando monitoramentos desses cursos d'água de forma contínua, desde 2014. Tais monitoramentos geraram uma ampla base de dados, da qual podem ser extraídas informações a fim de identificar padrões a serem interpretados que são úteis na gestão hídrica.

O avanço tecnológico e o acesso às fontes de dados vêm auxiliando na gestão pública e privada, reduzindo custos por meio do uso de ferramentas fundamentais. O *Knowledge Discovery in Database* (KDD), também chamado de processo de descoberta do conhecimento em base de dados, é um processo difundido composto de diversas etapas, sendo a mineração de dados uma etapa do KDD (FAYYAD; PIATETSKY-SHAPIRO; SMITH, 1996).

Dessa forma, é possível realizar análise de dados por meio de algoritmos que efetuam processamento de um grande volume de dados, sendo tal técnica denominada Mineração de dados. Sua utilidade está presente na automação aplicada para a extração de informações, possuindo métodos próprios para exploração dos dados (CORTES; PORCARO; LIFSCHIZ, 2002).

Uma funcionalidade existente na mineração de dados é a técnica de agrupamento, cuja função é unir um conjunto de dados em subgrupos, denominados *clusters*, os quais possuem semelhanças entre si, de modo a apresentar dados com mais homogeneidade em suas características (LINDEN, 2009). Vale ressaltar que, no agrupamento, os dados se agrupam em função de suas similaridades, não existindo fatores pré-definidos, diferentemente da classificação, em que já existe uma definição de suas classes, ou seja, denomina-se agrupamento de técnica não supervisionada, e classificação de técnica supervisionada.

O principal objetivo da classificação é avaliar as especificidades dos dados e atribuir classes a ele, sendo possível predizer um valor determinado para uma variável que seja categórica (SIMOES, 2008).

Outro importante método muito utilizado é o de Seleção de Atributos, que visa identificar atributos que possuem destaque em um dado conjunto de dados, como por exemplo o *Information*

Gain Attribute, que seleciona os atributos pelo método de ranqueamento, com base na entropia deles, a fim de estabelecer um atributo mais relevante em relação à classe (PARMEZAN *et al.*, 2012).

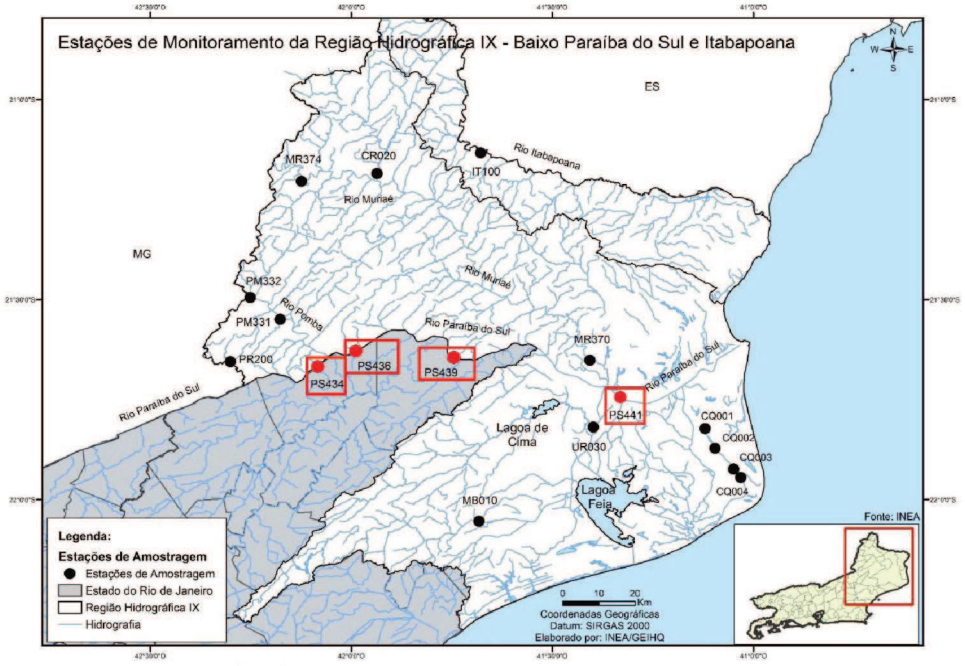
O objetivo deste trabalho é investigar os componentes do índice de qualidade de água dessa região e suas interações, por meio das técnicas de mineração de dados, a partir dos dados de monitoramento do Instituto Estadual do Ambiente.

2 Método

2.1 Área objeto de estudo

A área objeto de estudo compreende uma parte do território da Região Hidrográfica do Baixo Paraíba do Sul e Itabapoana, também denominada de Região Hidrográfica IX, situada na região norte e noroeste fluminense, com destaque para o Rio Paraíba do Sul. Os pontos considerados situam-se nos munícipios de Itaocara (PS0434 e PS 0436), São Fidélis (PS0439) e Campos dos Goytacazes (PS0441), conforme Figura 1.

Figura 1 – Mapa das estações de monitoramento da Região Hidrográfica IX, com destaque em vermelho para as estações consideradas no presente estudo



Fonte: Adaptado de INEA (2018)

2.2 Aquisição de dados

Os dados utilizados na presente pesquisa foram obtidos na página de monitoramento do Instituto Estadual do Ambiente, sendo considerado o período de 2014 a 2018, visto que corresponde a todos os dados disponíveis nessa região. Trata-se de dados referentes ao Índice de Qualidade de Água (IQA) do Rio Paraíba do Sul. Para que o IQA fosse calculado, foi necessário monitorar os seguintes parâmetros: Demanda Bioquímica de Oxigênio (DBO), Fósforo, Nitrato, Oxigênio Dissolvido (OD), Potencial Hidrogeniônico (pH), Turbidez, Coliformes Termotolerantes, Sólidos Totais Dissolvidos (STD), Temperatura da água e Temperatura do ar. Foram utilizadas 172 amostragens de qualidade de água, as quais são avaliadas no laboratório do Instituto Estadual do Ambiente.

2.3 Pré-processamento de dados

Após a coleta dos dados, foi necessário realizar uma análise sobre eles, identificando suas características, significados, e quais informações devem ser consideradas para a técnica de mineração de dados.

2.3.1 Transformação de dados

A transformação dos dados compreendeu a padronização dos parâmetros de qualidade de água em códigos no formato apropriado para o algoritmo a ser aplicado na etapa de mineração de dados, conforme codificação no Quadro 1:

Quadro 1 – Codificação dos parâmetros de qualidade de água

Parâmetro	Codificação	Parâmetro	Codificação
Índice Qualidade de Água	ind_agua	Fósforo	fos
Demanda Bioquímica de Oxigênio	dem_bio	Nitrato	nit
Oxigênio Dissolvido	ox_dis	Temperatura ar	temp_ar
Sólidos Totais Dissolvidos	sol_tot	Temperatura água	temp_agua
Potencial Hidrogeniônico	pot_hid	Turbidez	turb
Coliforme Termotolerante	col_term		

Fonte: Os autores (2019)

2.3.2 Transformação de dados

Para higienização dos dados, foi realizado um tratamento nos dados para que fosse possível uma melhor qualidade na etapa de mineração dos dados, sendo verificada a presença de dados incompletos, os quais foram excluídos.

2.4 Número de clusters

Para definição do número de *clusters* a serem utilizados no algoritmo de agrupamento a ser aplicado, foi utilizado o *software* R, com a aplicação de um *script*.

Após a aplicação do *script*, foi gerado o gráfico com o número de clusters sugerido pelo *software*. Vale ressaltar que foi adotado o método da silhueta, o qual determina o número ideal de grupos em determinado conjunto de dados (ROUSSEEUW, 1987).

2.5 Mineração de dados

Para a presente etapa, foram adotadas 2 (duas) técnicas de mineração de dados: agrupamento e classificação, realizadas no *software* WEKA - Waikato Environment for Knowledge Analysis v. 3.8.3.

2.5.1 Agrupamento

Para realizar a divisão dos dados de acordo com suas características específicas, foi adotada uma técnica não supervisionada de agrupamento. Para isso, foram utilizadas apenas as variáveis independentes dos dados, ou seja, excluído o IQA, e posteriormente aplicado o algoritmo de agrupamento kmeans, o qual consiste na minimização da distância entre as amostras através da distância euclidiana (LINDEN, 2009).

2.5.2 Classificação

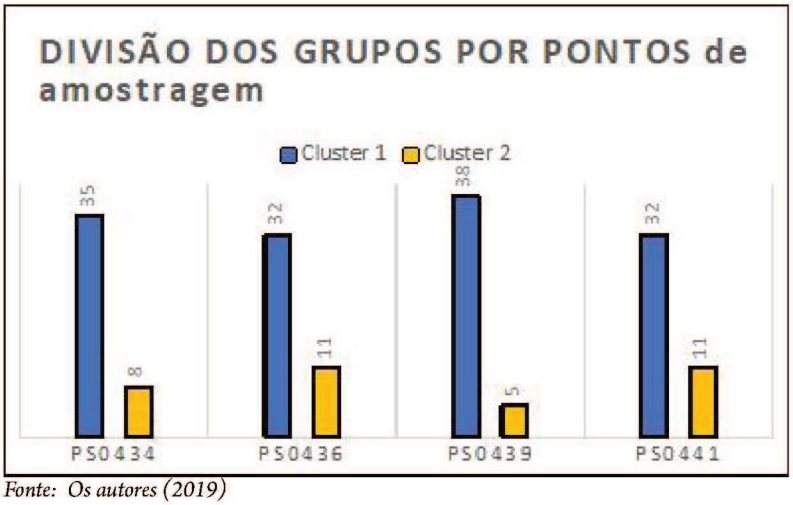
Na etapa de classificação, foi utilizado o algoritmo J48, com a finalidade de gerar a árvore de decisão. A referida árvore é composta de folhas, nós e ramos, sendo as folhas correspondentes às classes, os nós aos atributos, e os ramos aos valores dos atributos.

3 Resultados

3.1 Qualidade de água do Rio Paraíba do Sul

Os dados de qualidade de água utilizados na presente pesquisa demonstram que a água do Rio Paraíba do Sul, entre Itaocara e Campos dos Goytacazes, encontra-se, em 2,91% das amostras enquadradas, na categoria ruim, em 59,30% na categoria médio e em 37,79% na categoria bom, conforme Figura 2.

Figura 2 - Representação do enquadramento das amostras nas categorias de índice de qualidade de água

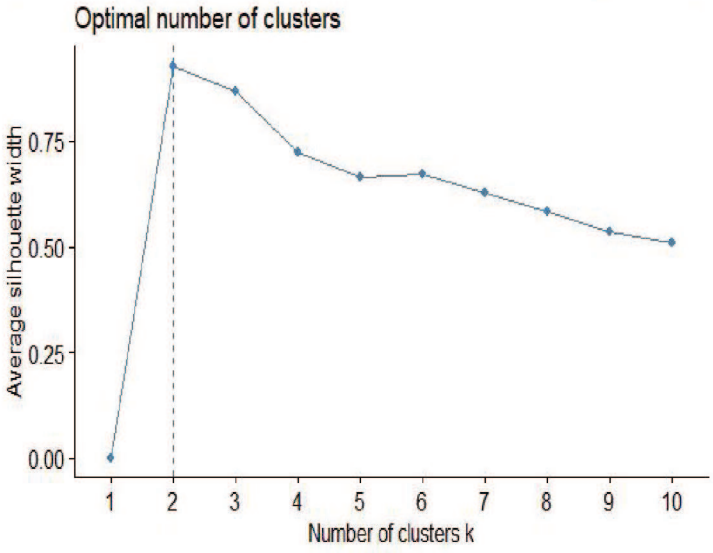


Fonte: Os autores (2019)

3.2 Determinação de clusters para agrupamento

Como produto do *script* extraído do *software* Rstudio, encontrou-se o número 2 para a quantidade de *clusters* a serem aplicados na técnica de agrupamento, conforme Figura 3.

Figura 3 – Gráfico gerado pelo Rstudio, indicando o número ideal de clusters a serem aplicados na técnica de agrupamento



Fonte: Os autores (2019)

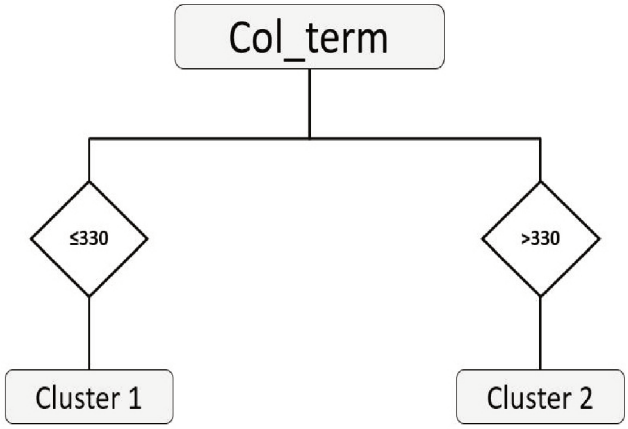
3.3 Mineração de dados

Após determinado o número de *clusters,* foi aplicada a técnica de agrupamento no WEKA, em que os dados foram divididos em dois subgrupos denominados *Cluster* 1 e *Cluster* 2. EXPLICAR A DIVISÃO DE CLUSTER - período chuvoso e período seco, turbidez e sdt.

Considerando que os dados foram divididos em dois grupos, fez-se necessário compreender qual foi o parâmetro mais importante que determinou essa divisão. Para isso, foi adotada a técnica de seleção de atributos por meio do InfoGain Attribute Eval pelo método de ranqueamento, onde foi possível constatar que o parâmetro col_term, ou seja, Coliformes Termotolerantes, possui destaque com o valor correspondente a 0.729.

Além disso, foi realizada a classificação pelo método J48, sendo possível obter a Árvore de Decisão, conforme Figura 4. Cabe frisar que esse método se assemelhou com o *Infogain* no que se refere ao atributo principal.

Figura 4 – Árvore de decisão com desfecho clusters



Fonte: Os autores (2019)

Dessa forma, conforme pode ser observado na Figura 4, o atributo principal para os dados em questão no método de classificação foi o Coliforme Termotolerante, com desfecho para os clusters. É possível observar que, se o valor do coliforme for menor ou igual a 330 NMP/100mL, enquadra-se como Cluster 1; caso seja maior, o desfecho torna-se o Cluster 2.

4 Discussão

De acordo com os resultados obtidos, foi possível observar que o parâmetro coliforme termotolerante destacou-se como atributo principal, tanto na árvore decisória quanto na seleção de atributos. Tal afirmação corrobora com a ausência de tratamento de efluentes sanitários, ou em alguns casos devido à ineficiência de tratamento, e posterior despejo no Rio Paraíba do Sul,

o que é constatado em pesquisas nessa região (GONÇALVES, 2016; LOURENÇO; PRADO, 2019; OLIVEIRA, 2006). Von Sperling (2005) afirma que os coliformes termotolerantes estão diretamente ligados à ausência do tratamento do esgotamento sanitário ou à sua ineficiência.

Uma pesquisa realizada por Gonçalves (2016), com o intuito de avaliar a qualidade de água dos estados de Minas Gerais, Rio de Janeiro e São Paulo, identificou que no Ponto PS0441, situado na calha principal do Rio Paraíba do Sul em Campos dos Goytacazes, ocorre contribuição de lançamento de esgoto, confirmando os resultados obtidos na presente pesquisa ao ser apresentado destaque para o *Cluster* 2 nesse ponto, conforme Figura 4.

Vale ressaltar que a definição de Cluster 2, no presente estudo, corresponde aos pontos cujo valor de coliforme termotolerantes foi superior a 330 NMP/100 mL. Rocha (2014) realizou uma análise por meio da metodologia de agrupamento em reservatórios do semiárido e encontrou 4 grupos de clusters, sendo estes relacionados respectivamente a: metais pesados (Pb, Cr e Hg), poluição por matéria orgânica, metais pesados (Zn e Ni) e eutrofização.

Outra forma de realizar agrupamento é pela metodologia de análise de componentes ambientais (ACP), ou *Principal Component Analysis (PCA)*. Carreira *et al.* (2010) apontaram, em estudo realizado em águas subterrâneas na ilha de Santiago (Cabo Verde), pelo método do PCA, a separação dos pontos em dois grupos em função de parâmetros relacionados a Cálcio, Magnésio e Cloro.

Dota (2014) identificou que os algoritmos FT, J48graft e J48 são mais indicados para compor o método de classificação, sendo J48 adotado no presente estudo. Tal afirmação se deu pela realização de testes com os algoritmos através da taxa de classificação incorreta (ICC).

Técnicas de associação foram utilizadas por Babbar & Babbar (2017) para prever o índice de qualidade de água, por meio das técnicas de Naive Bayes, árvore de decisão, K-Nearest Neighbor, Support Vector Machine, Redes Neurais Artificiais e Rule-based Classifiers (Classificadores baseados em regras), e foi possível concluir que Support Vector Machine e Árvores de Decisão foram os melhores classificadores.

5 Considerações finais

Por meio das consultas às bases acadêmicas, foi possível observar um bom quantitativo de pesquisas relacionadas à gestão de recursos hídricos envolvendo métodos computacionais, o que torna perceptível uma relevância desse tema para a área.

A aplicação de mineração de dados apresentou resultados que condizem com outros estudos ambientais de qualidade de água no mesmo local, provando que essa metodologia é eficaz. Todavia, cabe frisar que a técnica de mineração de dados é muito ampla, sendo possível realizar diversos ajustes no padrão dos dados.

Uma importante contribuição do presente estudo foi que o parâmetro coliforme termotolerante possuiu destaque no índice de qualidade de água da região do Baixo Paraíba do Sul, o que pode estar relacionado a despejos de efluente sanitário sem tratamento ou com ineficiência de tratamento, sendo necessária a atenção do poder público a essa problemática.

Outra questão de grande importância é que a aplicação de técnicas de mineração de dados pode otimizar o monitoramento ambiental, uma vez que é possível identificar padrões de pontos de amostragem, podendo reduzir ou aumentar esses pontos.

Referências

BILGEN, S. Structure and environmental impact of global energy consumption. *Renewable and Sustainable Energy Reviews*, v. 38, p. 890–902, Oct. 2014.

CARPENTER, S. R. *et al.* Nonpoint pollution of surface waters with phosphorus and nitrogen. *Ecological Applications*, v. 8, n. 3, p. 559–568, Aug. 1998.

CORTES, S. C.; PORCARO, R.M.; LIFSCHIZ, S. Mineração de Dados: Funcionalidades Técnicas e Abordagens. *PUC - Rio Inf. MCC 10112,* maio 2002.

EFFENDI, H. River Water Quality Preliminary Rapid Assessment Using Pollution Index. *Procedia Environmental Sciences*, v. 33, p. 562–567, 2016.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM*, v. 39, n. 11, p. 27–34, 1 Nov. 1996.

GONÇALVES, F. M. *Bacia hidrográfica do rio Paraíba do Sul*: avaliação integrada da qualidade das águas dos estados de Minas Gerais, Rio de Janeiro e São Paulo. 2016. Dissertação (Mestrado em Saneamento, Meio Ambiente e Recursos Hídricos) – UFMG, Escola de Engenharia, 2016. Disponível em: https://repositorio.ufmg.br/bitstream/1843/BUOS-AAKR5Q/1/dissertacao_fabricia_2016_1.pdf. Acesso em: 2019.

JABER, J. O.; S. MOHSEN, M. Evaluation of non-conventional water resources supply in Jordan. *Desalination,* v. 136, n. 1–3, p. 83–92, May 2001.

KACHURIN, N.; KOMASHCHENKO, V.; MORKUN, V. Mining production Environmental monitoring atmosphere of mining territories. *Metallurgical and Mining Industry,* n. 6, p. 4, 2015.

LINDEN, R. Técnicas de agrupamento. *Revista de Sistemas de Informação da FSMA*, v. 4, n. 4, p. 18-36, 2009.

LOURENÇO, T.; PRADO, R. Índices de saneamento ambiental em regiões hidrográficas do estado do Rio de Janeiro. *Revista de Gestão de Água da América Latina*, v. 16, n. 1, p. 7–7, 3 jul. 2019.

LOVETT, G. M. et al. Who needs environmental monitoring? *Frontiers in Ecology and the Environment*, v. 5, n. 5, p. 253–260, jun. 2007.

NAUBI, I. *et al.* Effectiveness of Water Quality Index for Monitoring Malaysian River Water Quality. *Polish Journal of Environmental Studies*, v. 25, n. 1, p. 231–239, 2016.

OLIVEIRA, V. S. D. *Percepção social acerca da degradação ambiental e medidas de qualidade de água do rio Paraíba do Sul no trecho entre Itaocara e São João da Barra, RJ*. 2006. Monografia (Graduação em Ciências Biológicas) – Universidade Estadual do Norte Fluminense, 2006.

PARMEZAN, A. R. S. *et al. Avaliação de Métodos para Seleção de Atributos Importantes para Aprendizado de Máquina Supervisionado no Processo de Mineração de Dados*. 2012. Relatório Técnico do Laboratório de Biotecnologia – Universidade Estadual do Oeste do Paraná, 2012. Disponível em: http://sites.labic.icmc.usp.br/aparmezan/publications/pdf/BIBLIOTECA_000_RT_002.pdf. Acesso em: 2019.

POORE, J.; NEMECEK, T. Reducing food's environmental impacts through producers and consumers. *Science*, v. 360, n. 6392, p. 987–992, 1 Jun. 2018.

ROUSSEEUW, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, v. 20, p. 53–65, Nov. 1987.

SIMÕES, A. C. A. *Mineração de Dados baseada em Árvores de Decisão para Análise do Perfil de Contribuintes*. 2008. Dissertação (Mestrado) - Universidade Federal de Pernambuco, UFPE, 2008. Disponível em: https://repositorio.ufpe.br/bitstream/123456789/1476/2/acas.pdf. Acesso em: 2019.

VON SPERLING, M. *Introdução à qualidade das águas e ao tratamento de esgotos*: princípios do tratamento biológico de águas residuárias. Belo Horizonte: Ed. UFMG, 2005. v. 1.