RECONHECIMENTO DE ASSOCIAÇÕES ENTRE PARÂMETROS DE QUALIDADE DE ÁGUA EM CORPOS HÍDRICOS POR MEIO DE MINERAÇÃO DE DADOS

RECOGNITION OF ASSOCIATIONS AMONG PARAMETERS OF WATER QUALITY IN WATER BODIES USING DATA MINING

Leonardo Bertholdo [1]
Luiz Camolesi Júnior [2]
Gisela de Aragão Umbuzeiro [3]
Celmar Guimarães da Silva [4]

Data de entrega dos originais à redação em: 20/06/2015
e recebido para diagramação em: 29/09/2016

A expansão demográfica, a crescente urbanização e o desenvolvimento industrial das últimas décadas vêm comprometendo a qualidade da água de diversos corpos hídricos. Nesse cenário, torna-se indispensável a implementação de soluções tecnológicas que auxiliem no processo de monitoramento ambiental. Neste trabalho, é empregada uma técnica de mineração de dados para investigar a presença de relações fortes entre parâmetros de qualidade de água. Como insumo para a pesquisa, foram utilizados dados de análises da água de alguns dos principais rios do estado de São Paulo. Com isso, espera-se contribuir para uma melhor compreensão dos resultados obtidos em programas de monitoramento de corpos hídricos.

Palavras-chave: Monitoramento Ambiental. Gestão de Recursos Hídricos. Mineração de Dados. Regras de Associação.

Demographic expansion, increasing urbanization and industrial development of recent decades have impaired water quality of various water bodies. In this scenario, it becomes essential the implementation of technological solutions that assist the environmental monitoring process. In this work, a technique of data mining is used to investigate the presence of strong relationships among water quality parameters. For this research input, we used data from water quality analysis of some of the main rivers in the state of Sao Paulo, Brazil. Thus, we aim to contribute to a better understanding of the results obtained in monitoring programs of water bodies.

Keywords: Environmental Monitoring. Water Resources Management. Data Mining. Association Rules.

1 INTRODUÇÃO

A água doce é um dos elementos vitais que compõem a biosfera. Sua degradação e escassez põem em risco a existência e a perpetuação dos organismos vivos. Nesse contexto, os corpos hídricos desempenham um papel fundamental, pois transportam água para as mais remotas regiões, sendo responsáveis pelo equilíbrio de muitos ecossistemas, além de viabilizar as mais diversas atividades humanas. Por outro lado, é preocupante o impacto que diversas bacias hidrográficas vêm sofrendo em consequência da urbanização e da industrialização aceleradas, além do rápido crescimento populacional das últimas décadas. Diante desse cenário, a descoberta de conhecimento útil pode ser extremamente valiosa para uma melhor compreensão dos fenômenos físicos, químicos e ecotoxicológicos observados nos corpos hídricos.

No estado de São Paulo, o monitoramento dos dados sobre a qualidade das águas dos corpos hídricos é realizado pela Companhia Ambiental do estado de São Paulo (CETESB), que mantém mais de 400 pontos de coleta de amostras de água, localizados ao longo dos rios e reservatórios monitorados. Cada amostra é analisada sob aspectos físicos, químicos, biológicos, ecotoxicológicos e bioanalíticos (CETESB, 2015), formando um amplo e rico conjunto de dados.

Este trabalho é parte de um projeto mais amplo, que tem como objetivo a descoberta de conhecimento útil em meio a dados de monitoramento de qualidade de água por meio da aplicação de diferentes técnicas de mineração de dados. Além da abordagem da análise associativa apresentada neste trabalho, este projeto abarca outras duas frentes de pesquisa: a análise de grupos para descoberta de regiões hidrográficas homogêneas quanto às suas características físicas, químicas e ecotoxicológicas, conforme Bertholdo et al. (2013) e a implementação de um modelo previsivo para descoberta de regras para classificação de ecotoxicidade em amostras de água, conforme Bertholdo et al. (2014).

O propósito específico deste trabalho é a descoberta de associações ocultas entre dados de monitoramento de qualidade de água em corpos

1 Mestre em Tecnologia pela Universidade Estadual de Campinas (UNICAMP) - Professor do Instituto Federal de Educação, Ciência e Tecnologia de São Paulo (IFSP) – Campus São Paulo. < l.bertholdo@ifsp.edu.br >.
2 Doutor em Física Computacional pela Universidade de São Paulo (USP) - Professor e pesquisador da Universidade Estadual de Campinas (UNICAMP). < camolesi@ft.unicamp.br >.
3 Doutora em Biologia Molecular pela Universidade de Campinas (UNICAMP) - Professora e pesquisadora da Universidade Estadual de Campinas (UNICAMP). < giselau@ft.unicamp.br >.
4 Doutor em Ciência da Computação pela Universidade Estadual de Campinas (UNICAMP) - Professor e pesquisador da Universidade Estadual de Campinas (UNICAMP). < celmar@ft.unicamp.br >.

hídricos, de modo a gerar novos conhecimentos acerca das interações entre os agentes que interferem nas características de suas águas. Os resultados obtidos proporcionaram a descoberta de correlações entre alguns dos parâmetros de qualidade de água medidos atualmente. Informações como estas podem ser úteis na gestão dos recursos hídricos, uma vez que o conhecimento levantado pode ser aplicado em outras bacias hidrográficas.

Neste artigo são apresentados os resultados desta pesquisa, começando pela Seção 2, que apresenta como é realizado o monitoramento de qualidade de água no estado de São Paulo. A Seção 3 descreve brevemente o processo de descoberta de conhecimento destacando sua etapa central, a mineração de dados, e a técnica de análise associativa. Em seguida, a Seção 4 apresenta a metodologia adotada para implementar a análise associativa com base nas medições dos parâmetros de qualidade de água. A Seção 5 detalha o processo e o algoritmo para análise associativa aplicados, bem como a ferramenta desenvolvida para este fim. Os resultados obtidos são mostrados na Seção 6. Por fim, a Seção 7 apresenta as considerações finais referentes a este trabalho.

2 MONITORAMENTO DE QUALIDADE DE ÁGUA

No estado de São Paulo, o monitoramento da qualidade da água dos rios, lagos e reservatórios é realizado pela Companhia Ambiental do estado de São Paulo (CETESB) desde 1974. Para isso, a CETESB dispõe de uma ampla rede de monitoramento que abrange as 22 regiões hidrográficas do estado, denominadas Unidades de Gerenciamento de Recursos Hídricos (UGRHIs).

Cada uma destas unidades possui vários pontos de amostragem, de onde são colhidas as amostras de água que, posteriormente, são analisadas com base em cerca de 60 parâmetros de qualidade de água, considerados os mais representativos (CETESB, 2015). A Figura 1 mostra esta divisão hidrográfica, classificando as UGRHIs em grupos conforme suas respectivas vocações.

Os parâmetros de qualidade de água medidos podem estar relacionados a aspectos físicos, químicos, microbiológicos, hidrobiológicos, ecotoxicológicos e bioanalíticos. Em cada ponto de amostragem é analisado um determinado conjunto de parâmetros, cujas medições são disponibilizadas anualmente pela CETESB em seu portal na Internet. Somente a rede básica, que visa o monitoramento da água dos rios do estado, gerou em 2014 aproximadamente 98.000 análises (CETESB, 2015), considerando que cada análise corresponde a uma medição de um parâmetro em um ponto de amostragem, realizada em uma data específica.

A Tabela 1 apresenta os parâmetros ou variáveis de qualidade contemplados nas análises. As variáveis principais são monitoradas em mais de 70% dos pontos de amostragem, e as variáveis adicionais são monitoradas em menos de 70% dos pontos (CETESB, 2015).

Monitoramento	Grupo	Principais Variáveis*	Variáveis Adicionais**
Rede Básica	Físicos	Condutividade, Sólido Dissolvido Total, Sólido Total, Temperatura da Água, Temperatura do Ar, Turbidez	Cor Verdadeira, Nível d'água, Salinidade, Transparência, Vazão
	Químicos	Alumínio Dissolvido, Alumínio Total, Bário Total, Cádmio Total, Carbono Orgânico Total, Chumbo Total, Cloreto Total, Cobre Dissolvido, Cobre Total, Cromo Total, DBO (5, 20), Ferro Dissolvido, Ferro Total, Fósforo Total, Manganês Total, Mercúrio Total, Níquel Total, Nitrogênio Amoniacal, Nitrogênio Kjeldahl, Nitrogênio-Nitrato, Nitrogênio-Nitrito, Oxigênio Dissolvido, pH, Potássio, Sódio, Subst. Tensoat. reagem c/ Azul Metileno, Zinco Total	Alcalinidade Total, Arsênio Total, Boro Total, Cafeína, Carbono Orgânico Dissolvido, Compostos Orgânicos Voláteis (COVs)[a], Compostos Orgânicos Semi-Voláteis (Semi-COVs)[a], DQO, Dureza, Fenóis Totais, Fluoreto Total, Herbicidas[b], Hidrocarbonetos Policíclicos Aromáticos (HPAs)[c], Microcistinas, Óleos e Graxas, Pesticidas Organofosforados[d], Potencial de Formação de THM,
	Hidrobiológicos	Clorofila *a* e Feofitina *a*	Comunidades Fitoplanctônica e Zooplanctônica
	Microbiologicos	*Escherichia coli*	*Giardia* e *Cryptosporidium*
	Ecotoxicológicos	Ensaio de Toxicidade Crônica com o microcrustáceo *Ceriodaphnia dubia*	Ensaio de Toxicidade Aguda com a bactéria luminescente - *Vibrio fisheri* (Sistema Microtox®), Ensaio de Mutação Reversa (Teste de Ames)[e],
	Bioanalíticos		Atividade Estrogênica por BLYES

Tabela 1 – Principais parâmetros de qualidade de água.
Fonte: CETESB (2015)

Estas análises são realizadas com base nas normas da Resolução CONAMA 357/2005, legislação ambiental regulamentada pelo Conselho Nacional de Meio Ambiente (BRASIL, 2005), que dispõe sobre a classificação dos corpos hídricos, dá diretrizes ambientais para o seu enquadramento, bem como estabelece condições e padrões de lançamento de efluentes (Umbuzeiro e Lorenzetti, 2010). Esta Resolução também define cinco classes para as águas doces, Especial, 1, 2, 3 e 4, sendo que a Classe Especial pressupõe usos mais nobres e a Classe 4 menos nobres. Estas classes representam um conjunto de condições e padrões de água necessários ao atendimento dos usos preponderantes, atuais ou futuros (VON SPERLING, 2007).

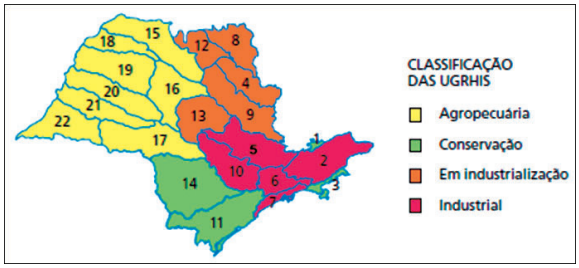

Figura 1 – Classificação das 22 UGRHIs por vocação
Fonte: CETESB (2015)

3 DESCOBERTA DE CONHECIMENTO EM BASES DE DADOS

A capacidade de uma organização tomar decisões é frequentemente associada ao conhecimento que esta possui sobre seu domínio de dados. Um dos problemas dos analistas de informação é a transformação de dados em informação relevante para a tomada de decisão (Silva, 2007). Conforme observado, as análises realizadas pela CETESB originam anualmente um valioso conjunto de informações referentes à qualidade da água dos corpos hídricos. No entanto, se analisadas por meio de técnicas tradicionais, a descoberta de conhecimento útil para a gestão da qualidade de água torna-se bastante limitado.

Nas últimas décadas, foram desenvolvidas técnicas que podem auxiliar na descoberta de informações implícitas em grandes repositórios de dados e, assim,

propiciar uma visão mais profunda do conjunto de dados analisado. Dentre os processos já desenvolvidos para extração de informações ocultas e relevantes em conjuntos de dados, talvez o KDD (*Knowledge Discovery in Databases*) seja um dos mais difundidos no meio computacional. Conforme Fayyad, Piatetsky-Shapiro e Smyth (1996), KDD é um processo não trivial de identificar padrões válidos, novos (antes desconhecidos), potencialmente úteis e, essencialmente, compreensíveis em bancos de dados. Este processo é formado por uma série de etapas, que compreende desde a preparação do conjunto de dados a ser analisado - seleção, pré-processamento e transformação - passando pela mineração dos dados, até a interpretação dos padrões e regras gerados para obtenção do conhecimento.

Na maior parte deste processo, é fundamental a cooperação de um especialista no domínio tratado, cujas habilidades podem contribuir decisivamente para o sucesso na escolha do conjunto de dados a ser analisado, além de auxiliar na definição do tipo de conhecimento a ser descoberto e como tal conhecimento pode contribuir no suporte a decisões (DUARTE et al., 2011). A Figura 2 apresenta as cinco fases que compõem este processo.

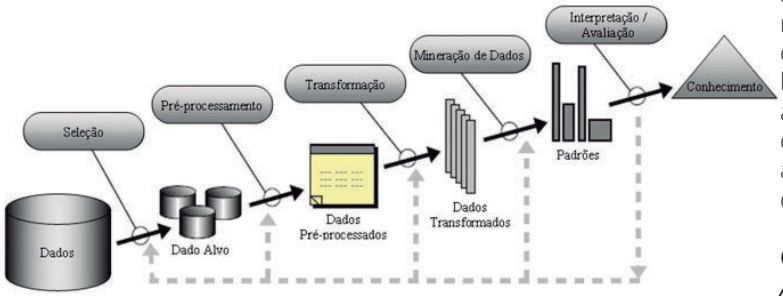

Figura 2 – Etapas que compõem o processo de KDD
Fonte: Adaptado de Fayyad, Piatetsky-Shapiro e Smyth (1996).

A etapa central deste processo é a mineração de dados, quando são extraídas efetivamente as informações implícitas presentes no conjunto de dados. A mineração de dados consiste na exploração e análise de grandes quantidades de dados, visando a descoberta de padrões e regras significativas (BERRY E LINOFF, 2004). Para isso, são utilizados algoritmos e técnicas de diferentes áreas do conhecimento como: estatística, banco de dados, reconhecimento de padrões, inteligência artificial, visualização de informação, aprendizagem de máquina, computação distribuída, entre outras. Atualmente, a mineração de dados vem sendo aplicada nos mais diversos cenários, como: área acadêmica, finanças, comércio, marketing, medicina, genética, telecomunicações e meio ambiente.

3.1 Análise Associativa

Segundo Tan, Steinbach e Kumar (2009), a Análise Associativa, técnica de mineração de dados aplicada neste trabalho, é usada para descobrir padrões que representem características altamente associadas dentro dos dados. A construção de um modelo para geração de regras de associação genérico é apresentada na Figura 3. Ele pode ser dividido em duas etapas: primeiramente, são procurados todos os conjuntos de **itens frequentes** da base de dados. Já a segunda etapa tem como objetivo encontrar regras a partir dos conjuntos de itens frequentes gerados na etapa anterior. Estas são as chamadas **regras fortes**, que representam os relacionamentos mais significativos entre os itens frequentes.

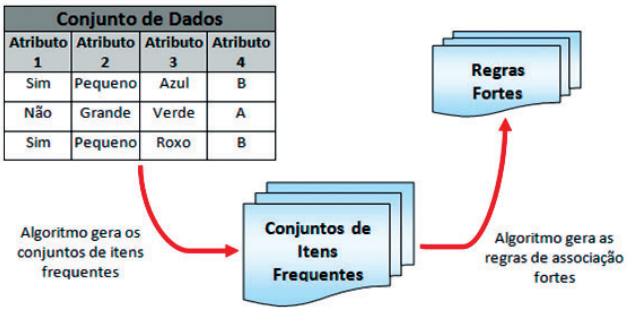

Figura 3 – Modelo para geração de regras de associação

O resultado final deste algoritmo são regras condicionais compostas por um **antecedente** e um **consequente**. Na análise associativa, tanto o antecedente quanto o consequente da regra podem possuir de *1* a *n* variáveis, desde que não se repitam em ambos os lados da regra. Baseando-se nos dados apresentado na Figura 3, um exemplo de um conjunto de itens frequentes e algumas possíveis regras geradas a partir deste seriam:

Conjunto de Itens Frequentes:
Atributo1=Sim e Atributo2=Pequeno e Atributo4=B

Regras Fortes:
Se *Atributo1=Sim e Atributo2=Pequeno* ***Então*** *Atributo4=B*
Se *Atributo1=Sim e Atributo4=B* ***Então*** *Atributo2=Pequeno*
Se *Atributo1=Sim* ***Então*** *Atributo2=Pequeno e Atributo4=B*

Durante o processo de geração das regras de associação, a qualidade da possível regra é avaliada por meio de medidas de **suporte** e **confiança**. O suporte é utilizado como base para a geração dos conjuntos de itens frequentes, e determina a taxa de registros que contêm todos os parâmetros da regra candidata. Já a confiança é verificada durante a geração das regras, as quais são obtidas a partir dos conjuntos de itens frequentes. Ela define a taxa de registros que contêm o consequente da regra candidata dentre aqueles que possuem o antecedente desta regra.

3.2 Trabalhos Relacionados

Assim como ocorre em outros domínios, a aplicação da mineração de dados em bases de monitoramento ambiental pode auxiliar fortemente nas tomadas de decisão, dando assim suporte à gestão de recursos hídricos. Existem diversos trabalhos que aplicam conceitos de mineração de

dados para identificação de correlações em dados de monitoramento de recursos hídricos.

Silva et al. (2013) apresenta o uso de técnicas de mineração de dados para análise exploratória em uma base de dados de ictioplâncton de um reservatório de água doce da Amazônia Legal. A aplicação do algoritmo Apriori permitiu a geração de regras de associação que proporcionaram a descoberta de conhecimento singular para o entendimento do processo de desova dos peixes na bacia do rio Tocantins.

Chen, Shyue e Chang (2010) apresenta um estudo de caso cujo objetivo é determinar os vários padrões que caracterizam os ambientes marinhos da baía de Dapeng, ao sul de Taiwan. Para isso, utilizam técnicas para mineração de regras de associação e análise da árvore de decisão, com apoio das ferramentas de mineração de dados Weka e Clementine.

Seixas, Nelson e Beatriz (2008) investiga a correlação dos dados espaciais e temporais que compõem o conjunto de poluentes da Lagoa Rodrigo de Freitas no Rio de Janeiro. O objetivo principal é obter uma metodologia para a classificação da qualidade da água, que pode ser utilizada em outros corpos hídricos. O trabalho inclui várias etapas de descoberta de conhecimento que são implementadas para atingir as metas, bem como a utilização de técnicas de mineração de dados para agrupar e classificar os dados.

Karimipour, Delavar e Kinaie (2005) estuda a mineração de dados geoespaciais para gestão de dados ambientais e, especialmente, para gestão de qualidade de água. Um estudo de caso realizado na região entre o Azerbaijão e o Irã apresenta a correlação entre a poluição de centros industriais e indicadores de qualidade de água através de mineração de dados geoespaciais. Segundo o estudo, ficam visíveis a relação entre a quantidade e a localização da poluição industrial e os indicadores de qualidade da água.

Comparativamente a estas pesquisas, este trabalho distingue-se por aplicar uma técnica de mineração de dados, baseada na extração de regras de associação, para encontrar relacionamentos fortes entre parâmetros específicos – físicos, químicos e ecotoxicológicos – utilizados no monitoramento de qualidade de água de corpos hídricos.

4 METODOLOGIA

A metodologia deste trabalho foi guiada pelo processo de descoberta de conhecimento *Knowledge Discovery in Databases* (KDD). Conforme já explanado, este processo é formado por uma série de etapas, que compreende desde a seleção do conjunto de dados a ser analisado até a interpretação dos padrões e regras geradas por técnicas de mineração de dados. Nesta pesquisa, as etapas iniciais deste processo, envolveu fortemente conhecimentos da área de saneamento ambiental, para a escolha e preparação dos dados. Da mesma forma, na etapa final, este envolvimento foi vital para a interpretação e avaliação dos resultados obtidos.

Para viabilizar a extração das regras de associação do conjunto de dados de monitoramento de qualidade de água, foi utilizado o **algoritmo Apriori**, apresentado em Tan, Steinbach e Kumar (2009), um dos mais difundidos para a geração de regras de associação. O resultado final deste algoritmo são regras condicionais compostas por um antecedente e um consequente. No contexto desta pesquisa, um exemplo de regra fictícia seria: ***Se** Cádmio e Ferro estão acima do padrão **Então** Condutividade é alta*. Este algoritmo será melhor detalhado na Seção 5.3.

5 PROCESSO DE RECONHECIMENTO DE ASSOCIAÇÕES ENTRE PARÂMETROS

Esta seção apresenta todas as etapas cobertas durante o processo de descoberta de conhecimento, bem como a ferramenta desenvolvida para identificação das correlações entre os parâmetros de qualidade de água.

Alguns autores, como Tan, Steinbach e Kumar (2009), tratam todos os procedimentos anteriores à mineração de dados como uma etapa única de "pré-processamento", visto que são atividades fortemente relacionadas. Nesta pesquisa, estes procedimentos foram divididos duas etapas: Seleção dos Dados e Pré-processamento dos Dados, sendo esta última dividida em cinco sub-etapas.

5.1 Seleção dos Dados

Esta pesquisa utilizou como base análises de água realizadas entre os anos de 2005 a 2011, nas quais os dados se mostraram com um maior grau de completude. Com relação ao aspecto geográfico, foram contempladas as UGRHIs: 2 (Paraíba do Sul), 5 (Piracicaba/Capivari/Jundiaí), 6 (Alto Tietê) e 10 (Sorocaba/Médio Tietê), as quais comportam aproximadamente 70% dos habitantes do estado de São Paulo, além de serem fortemente industrializadas. Nestas quatro UGRHIs foram selecionados 44 pontos de amostragem, considerados os pontos com maior riqueza e uniformidade de dados.

Quanto aos parâmetros de qualidade, foram considerados aqueles com maior possibilidade de trazer à tona informações relevantes e que constavam em pelo menos 80% dos pontos de amostragem. A aplicação destes critérios resultou em um conjunto de 21 parâmetros, divididos em quatro categorias: parâmetros relacionados à saúde humana, à vida aquática, a fatores organolépticos e indicadores genéricos, conforme apresentado na Tabela 2.

Tabela 2 – Parâmetros de qualidade de água considerados separados por categoria

Saúde Humana	Vida Aquática	Indicadores Genéricos	Fatores Organolépticos
Cádmio Total	Cobre Dissolvido	Chuva 24h	Alumínio Dissolvido
Chumbo Total	Nitrogênio Amoniacal	Cloreto Total	Ferro Dissolvido
Níquel Total	Oxigênio Dissolvido	Condutividade	Manganês Total
Nitrato	Substância Tensoativa	pH	Turbidez
Nitrito	Toxicidade	Sólidos Totais	
	Zinco Total	Temperatura Água	

5.2 Pré-processamento dos Dados

Nesse trabalho, a etapa de pré-processamento compreendeu as atividades de conversão, centralização, imputação, transformação e discretização dos dados.

Conversão e Centralização dos Dados

Depois de selecionados, os dados brutos foram centralizados em um repositório criado no Sistema Gerenciador de Banco de Dados PostgreSQL. Contudo, para tornar isto possível, foi necessário converter os dados, que se encontravam em arquivos PDF, para um formato adequado à estrutura de um banco de dados relacional. Essa atividade consumiu a maior parte do esforço da etapa de pré-processamento, uma vez que parte dos arquivos apresentava pequenas diferenças entre si, demandando tratamentos específicos em várias situações.

Imputação de Dados Faltantes

Para reduzir possíveis distorções nos resultados da mineração de dados, foi empregado um método para atribuição de valores aos parâmetros de qualidade com dados faltantes. Os critérios adotados neste método foram estabelecidos de forma empírica, visando o mínimo impacto sobre o conjunto de dados.

Em medições abaixo do padrão da Resolução CONAMA 357/2005[1], porém sem valor exato conhecido, foi imputado o valor medido. Exemplo:

Zinco Total	mg/L	máximo	0,18	< 0,02	Valor imputado = 0,02

Em medições com valores faltantes ou onde não foi possível detectar se o valor estava abaixo ou acima do padrão da Resolução CONAMA 357/2005, o valor foi ignorado sendo imputado um valor médio mensal do parâmetro nos sete anos (2005-2011). Exemplos:

Níquel Total	mg/L	máximo	0,025		Valor imputado = Média
Cádmio Total	mg/L	máximo	0,001	i < 0,005	Valor imputado = Média

Transformação dos Dados

Visando evitar conversões de dados durante o processo de mineração e assim reduzir o tempo de processamento dos algoritmos, os dados referentes aos identificadores e aos valores dos parâmetros de qualidade foram uniformizados. Os parâmetros foram padronizados na base de dados por meio de códigos contendo seis caracteres. Por exemplo, o parâmetro Zinco Total foi transformado em "zn_tot". Com a mesma finalidade, os valores discretizados dos parâmetros na base de dados foram uniformizados para serem representados por apenas duas letras maiúsculas, conforme apresentado na Tabela 3.

Discretização dos Dados

Muitas vezes, a análise associativa requer que os atributos contínuos sejam categorizados por meio de valores discretos. Nessa pesquisa, a discretização dos dados de monitoramento de qualidade água foi realizada por meio da inspeção visual dos dados. Essa abordagem segundo Tan, Steinbach e Kumar (2009) pode ser eficaz em determinadas situações. A Tabela 3 mostra como os parâmetros contínuos foram discretizados na base de dados.

5.3 Identificação de Associações entre Parâmetros de Qualidade de Água

Os relacionamentos entre os parâmetros de qualidade de água ainda não são totalmente conhecidos. Existem várias questões ainda não esclarecidas, devido à complexidade inerente ao enorme volume de dados gerados pelas medições dos parâmetros nas amostras de água.

Neste trabalho, a extração das regras de associação do conjunto de dados de monitoramento de qualidade de água foi realizada por meio do algoritmo Apriori apresentado em Tan, Steinbach e Kumar (2009), um dos algoritmos mais difundidos para a geração de regras de associação. Neste algoritmo, o processo de geração das regras de associação é dividido em duas etapas. A primeira é responsável por encontrar todos os conjuntos de itens frequentes que atendam a um limite de **suporte** mínimo considerado. A segunda etapa tem como objetivo encontrar todas as regras que satisfaçam um limite de **confiança** mínima considerado, a partir dos conjuntos de itens frequentes gerados na etapa anterior. Estas são denominadas **regras fortes**, que representam os relacionamentos mais significativos entre os **itens frequentes**. A Figura 4 apresenta de maneira resumida o funcionamento deste algoritmo.

5.4 Ferramenta para Identificação de Associações entre Parâmetros de Qualidade de Água

Para gerar as regras de associação e possibilitar a visualização dos resultados, foi implementada uma ferramenta em linguagem de programação Java, a qual é composta das seguintes funcionalidades: configuração de taxa mínima de suporte; configuração de taxa mínima de confiança; geração dos conjuntos de parâmetros frequentes; geração das regras fortes; configuração das informações apresentadas durante a execução; visualização da quantidade de regras geradas em cada execução; filtro para visualização de regras específicas. A interface principal, apresentada na Figura 5, pode ser dividida em duas partes:

- **Painel de controle (à esquerda) –** Destina-se às configurações de associação e visualização de informações e aos botões de comando.
- **Área de mensagens (à direita) –** Mostra os resultados do processamento.

1 Este padrão apresenta os limites de aceitação máximos e mínimos referentes aos parâmetros de qualidade, conforme os valores estabelecidos pela Resolução CONAMA 357/2005.

Tabela 3 – Categorização dos parâmetros contínuos e discretos

Parâmetros Contínuos	Valor Discretizado	Descrição
pH	AB	**Abaixo** – Abaixo do limite inferior do padrão da Resolução CONAMA 357/2005.
	PC	**Padrão CONAMA** – Dentro do padrão da Resolução CONAMA 357/2005.
	AC	**Acima** – Acima do limite superior do padrão da Resolução CONAMA 357/2005.
Temperatura Água[1], Condutividade[2], Sólidos Totais[2]	BX	**Baixo** – Dentro da faixa inferior (21 °C, 200 µS/cm, 200 mg/L respectivamente).
	MD	**Médio** – Entre as faixas inferior e superior.
	AT	**Alto** – Dentro da faixa superior (27 °C, 400 µS/cm, 400 mg/L respectivamente).
Oxigênio Dissolvido	PC	**Padrão CONAMA** – Dentro do padrão da Resolução CONAMA 357/2005.
	AB	**Abaixo** – Abaixo do padrão da Resolução CONAMA 357/2005 em até 60%.
	MA	**Muito Abaixo** – Abaixo do padrão da Resolução CONAMA 357/2005 mais que 60%.
Alumínio Dissolvido, Cádmio Total, Cloreto Total, Cobre Dissolvido, Ferro Dissolvido, Manganês Total, Nitrogênio Amoniacal, Níquel Total, Nitrato, Nitrito, Chumbo Total, Substância Tensoativa, Turbidez, Zinco Total	PC	**Padrão CONAMA** – Dentro do padrão da Resolução CONAMA 357/2005.
	AC	**Acima** – Acima do padrão da Resolução CONAMA 357/2005 em até 3x.
	MA	**Muito Acima** – Acima do padrão da Resolução CONAMA 357/2005 mais que 3x.
Parâmetros Discretos	**Mnemônico**	**Descrição**
Chuva 24h[2]	SI	**Sim** – Indica que choveu nas 24 horas anteriores à coleta da amostra.
	NO	**Não** – Indica que não choveu nas 24 horas anteriores à coleta da amostra.
Toxicidade	NT	**Não Tóxico** – Sem resposta fisiológica do microcrustáceo *Ceriodaphnia dubia*.
	CR	**Crônico** – Resposta fisiológica do microcrustáceo *Ceriodaphnia dubia*.
	AG	**Agudo** – Forte resposta fisiológica do microcrustáceo *Ceriodaphnia dubia*.

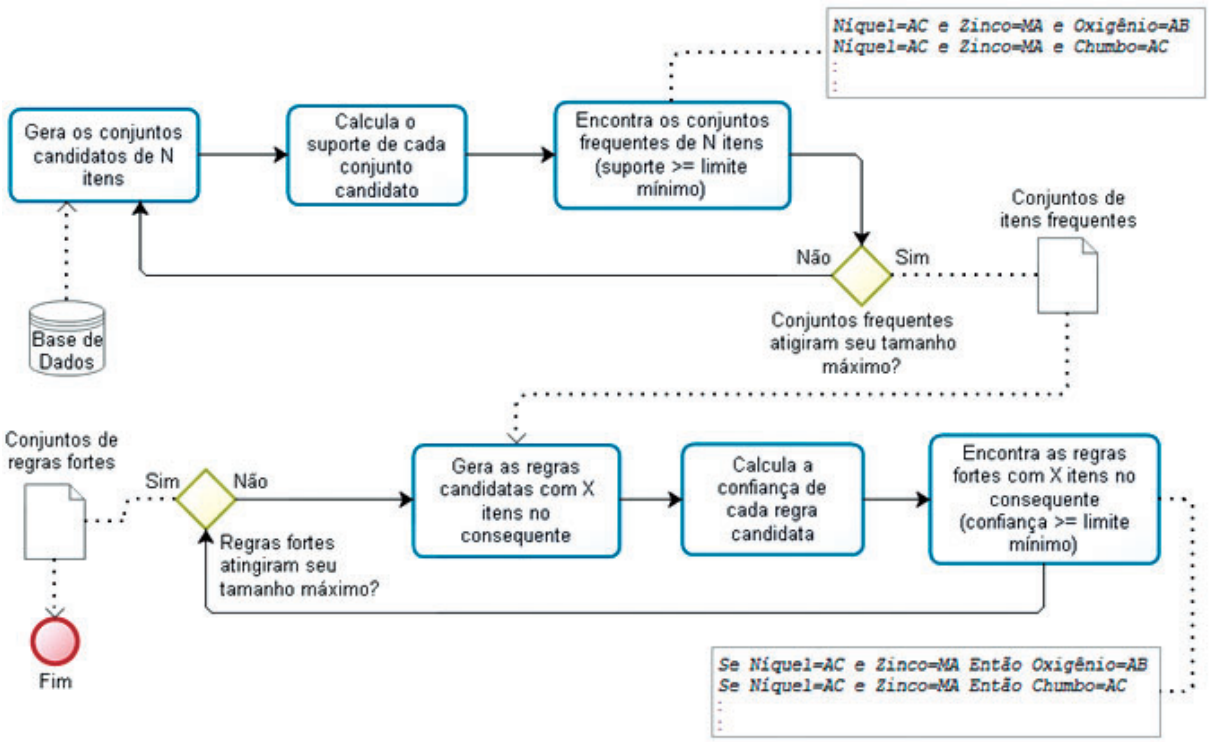

Figura 4 – Esquematização do algoritmo Apriori

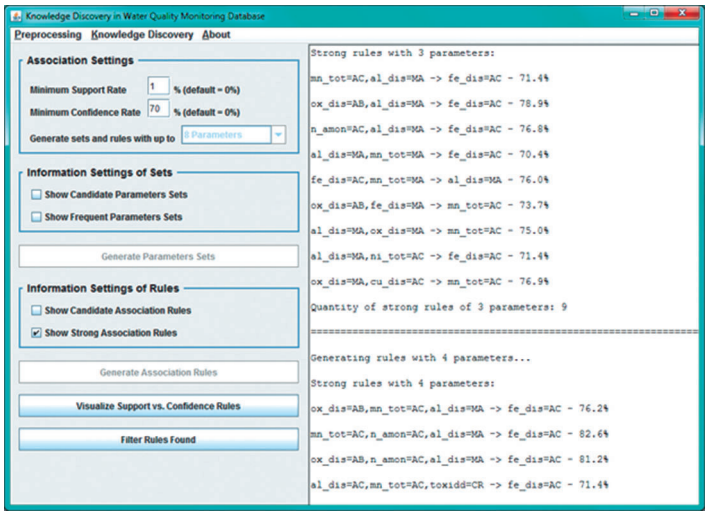

Figura 5 – Ferramenta para geração de regras de associação entre os parâmetros de qualidade de água

6 RESULTADOS

Foram realizados sete experimentos para identificação de associações entre parâmetros de qualidade de água: seis experimentos específicos, um para cada possível dupla de categorias de parâmetros de qualidade, conforme categorias apresentadas na Tabela 2, e um experimento generalista, o qual foi baseado nos resultados obtidos nos seis experimentos anteriores. A seguir, os parâmetros considerados em cada um dos experimentos:

- **Experimento 1** – Parâmetros relacionados à saúde humana e à vida aquática.
- **Experimento 2** – Parâmetros relacionados à saúde humana e a indicadores genéricos.
- **Experimento 3** – Parâmetros relacionados à saúde humana e a fatores organolépticos.
- **Experimento 4** – Parâmetros relacionados à vida aquática e a indicadores genéricos.
- **Experimento 5** – Parâmetros relacionados à vida aquática e a fatores organolépticos.
- **Experimento 6** – Parâmetros relacionados a indicadores genéricos e a fatores organolépticos.
- **Experimento Generalista** – Parâmetros considerados mais significativos nos seis experimentos anteriores. A significância foi mensurada com base nas ocorrências dos parâmetros nos experimentos.

As seguintes regras de associação foram extraídas a partir do experimento generalista[2]:

fe_dis=AC,turbid=AC -> al_dis=MA
al_dis=AC,turbid=AC -> ox_dis=AB
fe_dis=AC,turbid=MA -> al_dis=MA
ox_dis=MA,sub_te=AC -> n_amon=MA
fe_dis=AC,mn_tot=AC,turbid=AC -> al_dis=MA

A partir dos resultados do experimento generalista, pode-se inferir três relações principais entre os parâmetros de qualidade:

- Alumínio Dissolvido, Turbidez e Ferro Dissolvido
- Alumínio Dissolvido, Turbidez e Oxigênio Dissolvido
- Nitrogênio Amoniacal, Substância Tensoativa e Oxigênio Dissolvido

Percebe-se que os três parâmetros que compõem cada uma destas relações parecem caminhar juntos, pois nas regras de associação encontradas tendem a aparecer reunidos quando estão divergentes em relação ao padrão da Resolução CONAMA 357/2005. Observa-se também que, das três relações citadas, duas contêm apenas parâmetros de uma mesma categoria: Alumínio Dissolvido, Turbidez e Ferro Dissolvido pertencem todos à categoria "Fatores Organolépticos", e Nitrogênio Amoniacal, Substâncias Tensoativas e Oxigênio Dissolvido à categoria "Vida Aquática". Essa constatação indica que as relações fortes tendem a ocorrer entre parâmetros de mesma categoria.

Outra inferência também pode ser obtida a partir da Tabela 4, que mostra o número de regras geradas nos experimentos de 1 a 6. Com base nos números apresentados, pode-se notar que as combinações de categorias dos experimentos 2, 4 e 6 se mostraram como relações mais fortes, visto que originaram um número maior de regras.

7 CONSIDERAÇÕES FINAIS

Durante este trabalho, observou-se um grande volume de pesquisas relacionadas à aplicação da mineração de dados na área ambiental, especialmente na gestão de recursos hídricos, o que denota a importância do tema abordado para a comunidade científica. Como contribuições específicas deste trabalho, destacam-se: a geração de subsídios para auxiliar na compreensão dos resultados obtidos no monitoramento

2 al_dis: Alumínio Dissolvido; turbid: Turbidez; fe_dis: Ferro Dissolvido; n_amon: Nitrogênio Amoniacal; sub_te: Substância Tensoativa; ox_dis: Oxigênio Dissolvido; mn_tot: Manganês Total.

Tabela 4 – Comparativo das combinações entre as categorias de parâmetros

Experimentos	Combinações entre Categorias	Regras geradas
1	Saúde Humana e Vida Aquática	1
2	Saúde Humana e Indicadores Genéricos	29
3	Saúde Humana e Fatores Organolépticos	3
4	Vida Aquática e Indicadores Genéricos	47
5	Vida Aquática e Fatores Organolépticos	5
6	Indicadores Genéricos e Fatores Organolépticos	115

dos corpos hídricos e a possibilidade de comprovação da validade de correlações entre parâmetros que venham a ser pressupostas de forma empírica.

Embora tenha-se procurado contemplar uma amostra significativa das medições de qualidade de água do estado de São Paulo, devido à grande quantidade de medições incompletas, em que parâmetros essenciais para as análises não possuíam valor medido, foi necessário utilizar diversos critérios para se chegar a um conjunto de dados satisfatório para a aplicação da mineração de dados. Essa medida fez com que o conjunto de dados inicialmente disponível fosse reduzido.

Em continuidade a este trabalho, deve-se realizar um aprofundamento sobre a questão do desequilíbrio entre as classes dos parâmetros. Nesta pesquisa, esse problema foi tratado por meio da eliminação das medições que se encontravam dentro do padrão da Resolução CONAMA 357/2005, muitas das quais apresentavam ocorrência acima de 90%. A aplicação de métodos mais apropriados para tratar essa questão poderia trazer resultados mais satisfatórios na mineração dos dados. Outra futura abordagem está na utilização de identificadores para indicar as regras que englobam outras mais simples. Essa medida reduziria drasticamente o número de regras geradas, restringindo os resultados às regras mais abrangentes e significativas. Por fim, a representação dessas regras utilizando técnicas de visualização de grafos permitiria analisar de forma mais abrangente todas as regras e possíveis influências entre elas.

REFERÊNCIAS

BERRY, M. J. A.; LINOFF, G. S. **Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management.** Indianapolis: Wiley Publishing, 2004.

BERTHOLDO, L.; SILVA, C. G.; UMBUZEIRO, G. A.; CAMOLESI JR., L. Mineração de Dados de Qualidade de Água para Agrupamento de Pontos de Amostragem Usados no Monitoramento de Recursos Hídricos. In: **Congresso da Sociedade Brasileira de Computação, IV Workshop de Computação Aplicada à Gestão do Meio Ambiente e Recursos Naturais**, Maceió, p. 1036-1046, 2013.

BERTHOLDO, L.; SILVA, C. G.; UMBUZEIRO, G. A.; CAMOLESI JR., L. Data Mining Techniques for Water Ecotoxicity Classification for Application on Water Resources Management. **International Journal of Environment and Sustainable Development**. v. 13, nº 4, p. 408-424, 2014.

BRASIL. Conselho Nacional do Meio Ambiente. Resolução CONAMA n. 357, de 17 de março de 2005. **Diário Oficial da União,** Brasília, 2005.

CETESB. **Relatório de Qualidade das Águas Superficiais do Estado de São Paulo – 2014.** São Paulo: CETESB, 2015.

CHEN, C.; SHYUE, S.; CHANG, C. Association rule mining for evaluation of regional environments: Case study of Dapeng Bay, Taiwan. **International Journal of Innovative Computing, Information and Control.** v. 6, nº 8, p. 3425-3436, 2010.

DUARTE, A. A. A.; BERTHOLDO, L.; UMBUZEIRO, G. A.; CAMOLESI JR., L.; SILVA, C. G. Processamento e Visualização de Dados para a Descoberta de Conhecimento em Sistemas de Monitoramento de Qualidade de Água. In: **Congresso da Sociedade Brasileira de Computação, III Workshop de Computação Aplicada à Gestão do Meio Ambiente e Recursos Naturais**, Natal, p. 1409-1418, 2011.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery: An overview. **Advances in Knowledge Discovery and Data Mining**, AAAI Press/The MIT Press. p. 37-54, 1996.

KARIMIPOUR, F.; DELAVAR, M. R.; KINAIE, M. Water Quality Management Using GIS Data Mining. **Journal of Environmental Informatics.** v. 5, nº 2, p. 61-71, 2005.

SEIXAS, A. J.; NELSON, F. F. E.; BEATRIZ, S. L. P. L. Mining spatial and temporal data to classify water quality: a case study. **Data Mining IX: Data Mining, Protection, Detection and Other Security Technologies**. v. 40, p. 83-94, 2008.

SILVA, I. A. F. **Descoberta de Conhecimento em Base de Dados de Monitoramento Ambiental para Avaliação da Qualidade da Água**. 2007. 133 f. Dissertação (Mestrado). Instituto de Ciências Exatas e da Terra, Universidade Federal de Mato Grosso, Cuiabá, 2007.

SILVA, M. A., TREVISAN, D. Q., PRATA, D. N., MARQUES, E. E. Aplicação do algoritmo Apriori para uma base de dados de ictioplâncton em um reservatório de água doce da Amazônia Legal. In: **X Encontro Nacional de Inteligência Artificial e Computacional.** Fortaleza, 2013.

TAN, P.; STEINBACH, M.; KUMAR, V. **Introdução ao Data Mining – Mineração de Dados**. Rio de Janeiro: Editora Ciência Moderna, 2009.

UMBUZEIRO, G. A.; LORENZETTI, M. L. **Fundamentos da Gestão da Qualidade das Águas: Resolução CONAMA 357/2005.** Limeira: Biblioteca da Unicamp/CPEA, 2009.

VON SPERLING, M. **Estudos e modelagem da qualidade da água de rios**. Belo Horizonte: Departamento de Engenharia Sanitária e Ambiental – Universidade Federal de Minas Gerais, 2007.

(Footnotes)

1 Parâmetros de acompanhamento, sem valores de padrão da Resolução CONAMA 357/2005.