Mineração de Dados de Qualidade de Água para Agrupamento de Pontos de Amostragem Usados no Monitoramento de Recursos Hídricos

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

[1]Faculdade de Tecnologia – Universidade Estadual de Campinas (Unicamp)
Limeira – São Paulo – Brazil

leo.btd@gmail.com, {camolesi, giselau, celmar}@ft.unicamp.br

Abstract. *The application of advanced computational resources in support of environmental management systems is becoming increasingly common. In this work is used a technique of cluster analysis, whose goal is to discover hydrographic regions homogeneous in relation to their physical, chemical and ecotoxicological properties. For this, the clustering algorithm adopted seeks water sampling sites where measurements of its parameters of quality are similar. We used data from analyzes of water quality of some of the main rivers of the state of São Paulo, from 2005 to 2011. The methodology contributes to a better knowledge of water bodies, allowing reduction of the number of sites to be analyzed in monitoring programs.*

Resumo. *A aplicação de recursos computacionais avançados no suporte aos sistemas de gestão ambiental vem se tornando cada vez mais frequente. Neste trabalho é empregada uma técnica de análise de grupos cujo objetivo é descobrir regiões hidrográficas homogêneas quanto às suas características físicas, químicas e ecotoxicológicas. Para isso, o algoritmo de clusterização adotado busca grupos de pontos de amostragem de água onde as medições de seus parâmetros de qualidade são similares. Foram utilizados dados de análises de qualidade de água de alguns dos principais rios do estado de São Paulo, realizadas entre 2005 e 2011. A metodologia desenvolvida contribui para um melhor conhecimento dos corpos d'água, permitindo a redução da quantidade de pontos a serem analisados em programas de monitoramento.*

1. Introdução

Os corpos hídricos sempre representaram um recurso indispensável para a existência e a manutenção da vida em nosso planeta. No entanto, a crescente urbanização, a expansão demográfica e o desenvolvimento industrial das últimas décadas vêm ocasionando o comprometimento de muitas bacias hidrográficas. Diante deste panorama, o controle de qualidade da água dos corpos hídricos, bem como a compreensão dos fenômenos que interferem em suas características, são essenciais para preservação deste recurso natural.

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 350 pontos de coleta de amostras de água, localizados

ao longo dos rios e reservatórios rastreados. Cada amostra é analisada sob aspectos físicos, químicos e biológicos, formando um rico conjunto de dados (CETESB, 2011).

Este trabalho é parte de um projeto mais amplo, que tem como meta 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 de clusterização apresentada neste trabalho, este projeto abrange outras duas frentes de pesquisa: a descoberta de regras para classificação de ecotoxicidade em amostras de água, conforme apresentado em Bertholdo et. al (2012), e a investigação de associações significativas entre parâmetros de qualidade de água, tais como Cádmio Total, Chumbo Total, Cobre Dissolvido, Níquel Total, Nitrato, Nitrito, Nitrogênio Amoniacal, Oxigênio Dissolvido, Substância Tensoativa, Zinco Total, entre outros.

O enfoque específico deste trabalho está na procura por grupos de pontos de amostragem de água com alto grau de similaridade entre as medições de seus parâmetros de qualidade. Este agrupamento pode gerar inferências úteis a respeito da condições dos corpos d'água, revelando distinções existentes entre os corpos hídricos ou ainda entre trechos de um mesmo corpo hídrico. Outro benefício decorrente do agrupamento dos pontos de amostragem é a redução da quantidade de locais de coleta de água necessários para o monitoramento dos corpos hídricos. Descobertas como esta podem ser úteis na gestão das bacias hidrográficas e podem ser consideradas como base para definir diretrizes estratégicas específicas para cada região gerada pelo agrupamento.

Neste artigo são apresentados os resultados iniciais desta pesquisa, começando pela Seção 2, que apresenta a metodologia adotada para implementar o agrupamento dos pontos de amostragem. A Seção 3 explica como é realizado o monitoramento de qualidade de água no estado de São Paulo. Em seguida, a Seção 4 descreve brevemente o processo de descoberta de conhecimento destacando sua etapa central, a mineração de dados. A Seção 5 detalha a aplicação da técnica de mineração na regionalização dos pontos de amostragem. Os resultados obtidos até o momento são expostos na Seção 6. Por fim, a Seção 7 apresenta as considerações finais referentes a este trabalho.

2. Metodologia

A metodologia escolhida foi guiada pelo processo de descoberta de conhecimento denominado *Knowledge Discovery in Databases* (KDD). Conforme Fayyad et al. (1996), KDD é um processo não trivial de identificar padrões válidos, novos (antes desconhecidos), potencialmente úteis e, essencialmente, compreensíveis em bancos de dados. Nesta pesquisa, as etapas iniciais deste processo, demandou a intensa participação de uma especialista da área de saneamento ambiental, visando auxiliar na escolha e na preparação dos dados. Da mesma forma, na etapa final, atual estágio em que se encontra o estudo, a cooperação desta especialista tem sido fundamental na interpretação e avaliação dos resultados obtidos.

Para agrupar os pontos de amostragem em áreas uniformes, foi utilizado um método de regionalização baseado em grafos. No contexto desta pesquisa, os vértices do grafo representam os pontos de amostragem e as arestas que conectam estes vértices representam os corpos d'água que interligam os pontos. As arestas mais custosas expõem os relacionamentos mais fracos entre os pontos de amostragem, revelando assim as dissimilaridades existentes entre os corpos hídricos e seus diferentes trechos.

3. Monitoramento de Qualidade de Água

A gestão de bacias hidrográficas passou a assumir crescente importância no Brasil à medida que os efeitos da degradação ambiental sobre a disponibilidade de recursos hídricos foram aumentando (Jacobi et al., 2007). No estado de São Paulo, a CETESB analisa e acompanha a qualidade da água dos rios, lagos e reservatórios desde 1974. Para isso, dispõe de uma ampla rede de monitoramento distribuída por 22 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 em laboratório (CETESB, 2011). A Figura 1 mostra esta divisão, classificando as UGRHIs em grupos conforme suas respectivas vocações.

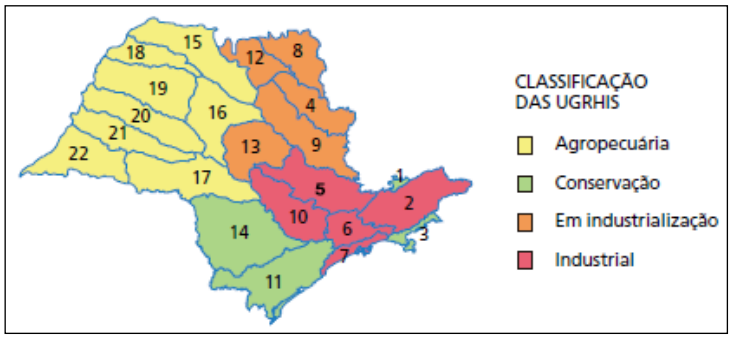

Figura 1. Classificação das 22 UGRHIs por vocação (CETESB, 2011).

A análise laboratorial contempla dezenas de parâmetros de qualidade, os quais podem estar relacionados a aspectos físicos, químicos, microbiológicos, hidrobiológicos e toxicológicos. 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, gera um volume anual estimado de 65.000 análises (CETESB, 2012), considerando que cada análise corresponde a uma medição de um parâmetro em um ponto de amostragem, realizada em uma data específica.

Estas análises são realizadas com base nas normas da Resolução CONAMA 357/2005, legislação ambiental regulamentada pelo Conselho Nacional de Meio Ambiente (CONAMA, 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 et al., 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).

4. Descoberta de Conhecimento em Bases de Dados

Conforme Silva (2007), a capacidade de uma organização de 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. Nas últimas décadas, foram desenvolvidos processos que podem auxiliar na descoberta de informações não triviais

em grandes repositórios de dados e, assim, dar um significado mais representativo e abrangente aos dados existentes nestes repositórios.

Dentre os processos já desenvolvidos para extração de informações ocultas e relevantes em conjuntos de dados, talvez o KDD seja um dos mais difundidos no meio computacional. Este processo é formado por uma série de etapas, que compreende desde a preparação do conjunto de dados a ser estudado – 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 estudado, 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.

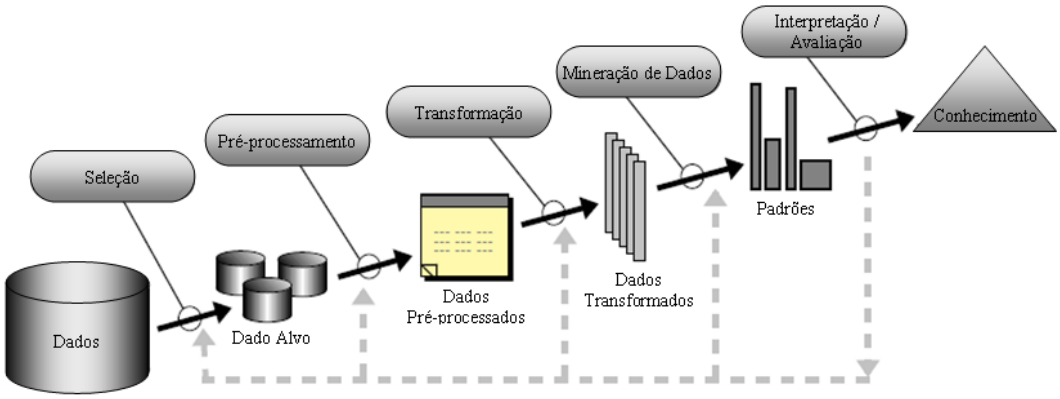

Figura 2. Etapas que compõem o processo de KDD. Adaptado de Fayyad et al. 1996.

Dentre as cinco etapas do KDD, a mineração de dados pode ser considerada a principal, pois é nessa fase em que são extraídas as informações implícitas presentes no conjunto de dados. Segundo Berry (2004), a mineração de dados consiste na exploração e análise de grandes quantidades de dados, visando a descoberta de padrões e regras significativas. Para atingir seu objetivo, a mineração de dados utiliza-se de técnicas de diferentes áreas do conhecimento como: estatística, banco de dados, reconhecimento de padrões, inteligência artificial, visualização de informação, aprendizagem de máquina, entre outras. Atualmente, a mineração de dados vem sendo aplicada nos mais diversos cenários, tais como: área acadêmica, finanças, comércio, marketing, medicina, genética, telecomunicações e meio ambiente.

Entre as técnicas de mineração de dados está a *Análise de Cluster*, que procura organizar objetos em grupos homogêneos. Uma das técnicas de clusterização utilizadas para o agrupamento de objetos distribuídos em espaços geográficos é a regionalização. Conforme Neves et al. (2002), regionalização é um processo de agrupamento que busca uma nova repartição do espaço de estudo em um número menor de objetos, resultando em novas regiões com dimensões geográficas mais abrangentes. Alguns motivos para se realizar este agrupamento são: aumento da representatividade dos valores dos atributos, redução dos efeitos da imprecisão nos valores das variáveis, redução de erros associados ao posicionamento geográfico de eventos, e redução no custo de análise dos dados.

4.1. Trabalhos Relacionados

Há uma série de trabalhos científicos que aplicam conceitos de mineração de dados na descoberta de conhecimento em bancos de dados de monitoramento hidrográfico. Diniz et al. (2012) utiliza de técnicas de regionalização hidrológica, para possibilitar a transferência de dados e informações entre bacias com características similares. O trabalho visa identificar regiões hidrologicamente homogêneas no Estado da Paraíba, utilizando mineração de dados, através da técnica de clusterização, possibilitando assim a identificação de padrões que permitam a transposição de dados de uma região para outra. Foram utilizados algoritmos com métodos baseados em partição, métodos hierárquicos, e métodos baseados em redes neurais, e aplicados índices de validação estatística nos agrupamentos gerados.

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

Seixas et al. (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 podem ser utilizados 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.

Ramachandra Rao e Srinivas (2006) propõem um processo de *clustering* que mescla algoritmos aglomerativos hierárquicos e um algoritmo de agrupamento particional para identificar grupos de bacias hidrográficas semelhantes. A eficácia da análise de cluster híbrido na regionalização é investigado com o uso de dados de bacias hidrográficas do estado de Indiana nos EUA.

Karimipour et al. (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 o quantidade e a localização da poluição industrial e os indicadores de qualidade da água.

Comparativamente a estas pesquisas, este trabalho distingue-se por aplicar um método de agrupamento hierárquico divisivo para regionalização do espaço, o qual é baseado na obtenção de uma Árvore Geradora Mínima (AGM) seguida da poda desta árvore, proporcionando a separação dos pontos de amostragem de água em regiões homogêneas e equilibradas em termos de número de objetos.

5. Processo para Regionalização de Pontos de Amostragem

Esta Seção apresenta todos os passos percorridos durante este processo, desde a seleção e preparação dos dados brutos até a etapa de mineração dos dados pré-processados, bem como a ferramenta desenvolvida para agrupamento dos pontos de amostragem de água.

5.1. Pré-processamento dos Dados

Alguns autores, como Tan et al. (2009), tratam as atividades de KDD, anteriores à mineração de dados, como uma atividade única de "pré-processamento". Nesse trabalho, esta etapa compreendeu as atividades apresentadas a seguir.

Seleção dos Dados

Esta pesquisa utilizou como base análises de água realizadas entre os anos de 2005 a 2011, nos 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: Alumínio Dissolvido, Cádmio Total, Chumbo Total, Chuva 24h, Cloreto Total, Cobre Dissolvido, Condutividade, Ferro Dissolvido, Manganês Total, Níquel Total, Nitrato, Nitrito, Nitrogênio Amoniacal, Oxigênio Dissolvido, pH, Sólidos Totais, Substância Tensoativa, Temperatura Água, Toxicidade, Turbidez e Zinco Total.

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 (CONAMA, 2005), 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 CONAMA, 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

5.2. Identificação de Grupos Homogêneos de Pontos de Amostragem de Água

Neste trabalho, a tarefa de agrupamento dos pontos de amostragem foi realizada a partir do método apresentado em Assunção et. al. (2002), o qual se baseia na obtenção de uma Árvore Geradora Mínima (AGM), seguida da poda desta árvore para propiciar a divisão

dos objetos em regiões uniformes. Algumas referências na área de mineração de dados, como Tan et al. (2009), mencionam técnicas interessantes de agrupamento hierárquico divisivo baseadas na geração de AGMs.

Este método é dividido em duas fases: na primeira, é gerada uma árvore (AGM) a partir do grafo correspondente ao conjunto de dados. Esta árvore é escolhida de forma a garantir que a soma dos custos associados às arestas seja a menor possível. A AGM é obtida a partir do grafo por meio do **algoritmo de Prim**. Nesta abordagem, a cada estágio, uma nova aresta é adicionada à árvore e o algoritmo para somente quando todos os vértices forem visitados. Vale destacar que o custo de cada aresta é inversamente proporcional à similaridade entre os objetos.

No cenário desta pesquisa, os vértices do grafo equivalem aos pontos de amostragem de água e as arestas que conectam estes vértices correspondem aos corpos hídricos que conectam os pontos de amostragem. As arestas mais custosas representam os relacionamentos mais fracos entre os pontos de amostragem. Os custos das arestas são determinados por meio da distância euclidiana entre os atributos *i* e *k* dos dois vértices da aresta, conforme sugere Assunção et al. (2002), e cujo cálculo é dado por:

$$Custo(i,k) = \sqrt{\sum_{j=1}^{m}(x_{ij} - x_{kj})^2}$$

Sendo: *m*, número de atributos;
x, valores dos atributos;

A segunda etapa do método consiste em retirar as arestas mais caras, que no cenário desta pesquisa equivale a eliminação dos relacionamentos mais custosos entre os pontos de amostragem, remanescentes após a geração da AGM. Cada aresta retirada provoca uma divisão na árvore, resultando em duas subárvores desconectadas. São escolhidas *k-1* arestas, para obter *k* regiões. Para produzir regiões mais homogêneas e equilibradas em termos de objetos por região, na fase de poda, a forma de atribuir custos às arestas é modificada e o novo custo é dado pela soma dos quadrados dos desvios, associada a árvore *T*, subtraída pela soma das duas parcelas obtidas da soma dos quadrados dos desvios das duas subárvores geradas pela retirada da aresta da árvore *T*.

5.3. Ferramenta para Agrupamento de Pontos de Amostragem de Água

Para viabilizar o agrupamento dos pontos de amostragem de água e visualizar os resultados gerados por este agrupamento, foi implementada uma ferramenta em linguagem de programação Java, cuja interface principal é apresentada na Figura 3. A ferramenta permite combinar até dez parâmetros simultaneamente, dentre os 21 citados na Seção 5.1. Esta interface pode ser dividida em duas partes:

- **Painel de controle:** Área para configurações dos períodos considerados, informação dos parâmetros configurados, carga dos dados dos pontos de amostragem, geração da AGM e dos grupos de pontos, e opções de visualização das conexões fluviais (relacionamentos) e seus custos (pesos).
- **Painel de visualização (abaixo):** Área para visualização dos pontos de amostragem e suas interligações fluviais.

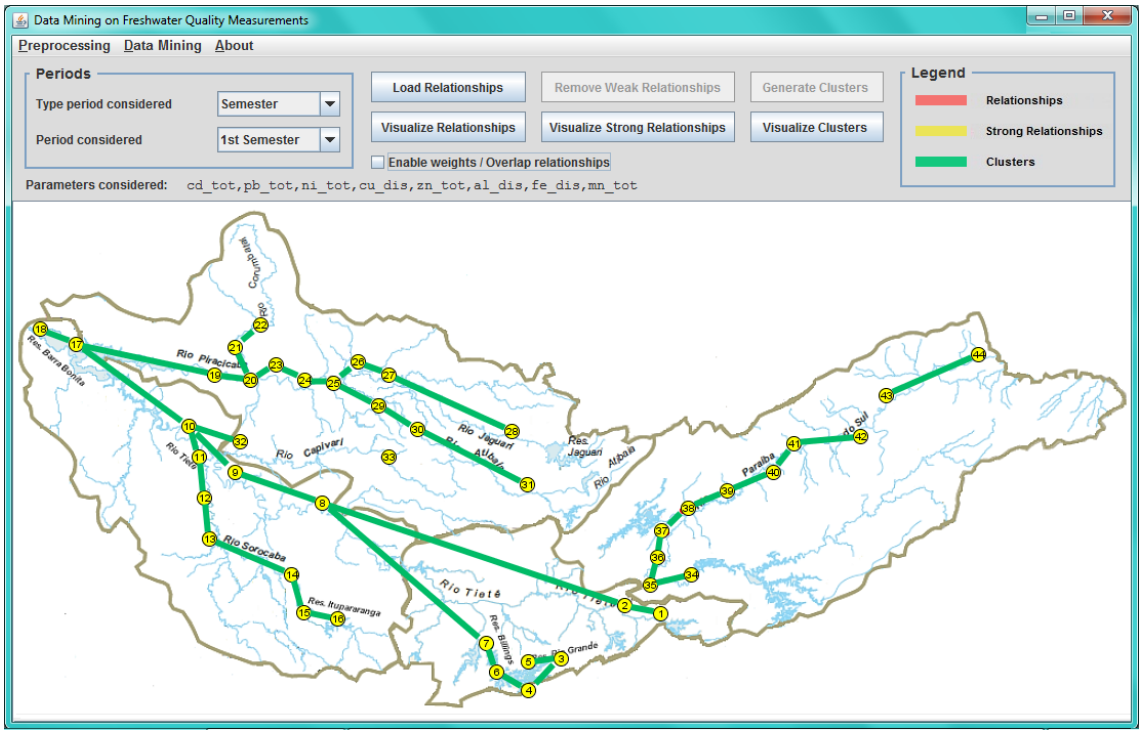

Figura 3. Ferramenta para Agrupamento de Pontos de Amostragem de Água[1]

6. Resultados

O atual estágio da pesquisa encontra-se na fase de experimentação e análise dos resultados obtidos pela mineração de dados. A ferramenta desenvolvida permite configurar diversas variáveis de modo que, ao longo desta etapa, deverá ser realizada uma grande diversidade de experimentos. Por este motivo, nesta Seção são apresentados apenas alguns resultados parciais gerados pela ferramenta.

Os experimentos iniciais consideraram somente a UGRHI 2, que compreende a bacia do Rio Paraíba do Sul, na qual foram considerados 11 pontos de amostragem de água. Os parâmetros de qualidade utilizados nos experimentos foram divididos em quatro categorias: parâmetros relacionados à saúde humana, à vida aquática, a fatores organolépticos e indicadores genéricos, conforme apresentado na Tabela 1.

Tabela 1. Categorias de Parâmetros de Qualidade de Água

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	

[1] cd_tot, pb_tot, ni_tot, cu_dis, zn_tot, al_dis, fe_dis e mn_tot referem-se respectivamente aos metais: Cádmio Total, Chumbo Total, Níquel Total, Cobre Dissolvido, Zinco Total, Alumínio Dissolvido, Ferro Dissolvido e Manganês Total.

Para restringir os experimentos às situações mais interessantes e com maior possibilidade de trazer à tona informações significativas, foram considerados os seguintes períodos:

- **Estações do ano –** As variações sazonais influenciam nas concentrações dos parâmetros de qualidade na água, propiciando a geração de informações específicas para cada época do ano.
- **Anos completos –** Ao considerar os 12 meses do ano, tem-se uma visão geral do comportamento dos parâmetros em cada ponto de amostragem.

Como padrão de regionalização, foi estabelecida a geração de dois grupos de pontos de amostragem em todos os experimentos.

Os experimentos foram organizados em quatro conjuntos. Para cada um deles, foram feitos cinco experimentos, quatro contemplando apenas os três meses de cada estação do ano e um considerando os 12 meses do ano. A seguir, a descrição de cada um destes conjuntos:

- **Experimentos 1 a 5 –** Agrupamento de pontos de amostragem similares em relação às medições de parâmetros relacionados à **saúde humana**.
- **Experimentos 6 a 10 –** Agrupamento de pontos de amostragem similares em relação às medições de parâmetros referentes à **vida aquática**.
- **Experimentos 11 a 15 –** Agrupamento de pontos de amostragem similares em relação às medições de parâmetros considerados **indicadores genéricos**.
- **Experimentos 16 a 20 –** Agrupamento de pontos de amostragem similares em relação às medições de parâmetros associados a **fatores organolépticos**.

A partir da observação dos agrupamentos obtidos nos experimentos, foi possível identificar os seguintes subgrupos de pontos de amostragem em comum:

- **Experimentos 1 a 5 –** 34 a 38; 40 a 43.
- **Experimentos 6 a 10 –** 34 e 35; 37 a 43.
- **Experimentos 11 a 15 –** 34 a 38; 39 a 41; 43 e 44.
- **Experimentos 16 a 20 –** 34 a 39; 42 e 43.

Estes subgrupos existentes dentro dos grupos de pontos de amostragem podem ser melhor visualizados por meio da Tabela 2, onde cada cor representa um subgrupo de pontos que se repetem nos cinco experimentos de cada conjunto de experimentos.

A última linha desta Tabela também exibe os subgrupos de pontos identificados, porém considerando todos os 20 experimentos realizados. Em outras palavras, esta última linha mostra os pontos que aparecem sempre em um mesmo grupo em todos os 20 experimentos. É possível constatar então que os pontos de amostragem **34** e **35**, e **37** e **38** possuem medições similares em todos os experimentos, considerando os 21 parâmetros de qualidade contemplados.

A partir desta constatação, considerando as medições destes parâmetros, pode-se questionar a necessidade de dois destes quatro pontos de amostragem, visto que as medições dos pontos de cada dupla são similares. Com isso, demonstra-se a possibilidade de diminuição do número de pontos de amostragem necessários para o monitoramento dos corpos hídricos, uma das contribuições esperadas deste trabalho.

Tabela 2: Subgrupos de pontos de amostragem identificados nos conjuntos de experimentos.

Conjuntos de Experimentos	Pontos de Amostragem – UGRHI 2										
	34	35	36	37	38	39	40	41	42	43	44
1-5											
6-10											
11-15											
16-20											
1-20											

7. Considerações Finais

Neste artigo, foi apresentada a aplicação de um método de mineração de dados para agrupamento de pontos de amostragem em regiões homogêneas em termos de qualidade de água. Observa-se um expressivo volume de trabalhos relacionados à aplicação da computação na área ambiental, especialmente na gestão de recursos hídricos, o que denota a importância do tema abordado para a comunidade científica.

Quanto aos dados utilizados no estudo, procurou-se contemplar uma amostra significativa das medições de qualidade de água do estado de São Paulo, porém notou-se que o conjunto de dados inicialmente disponível precisou ser drasticamente reduzido. Um dos motivos para esta ocorrência é a grande quantidade de medições incompletas, onde parâmetros relevantes para esta pesquisa não possuíam valor medido. Nesse sentido, a estratégia de redução dos dados adotada visou preservar a etapa de mineração, cujos resultados dependem fortemente da qualidade do conjunto de dados.

Este trabalho enfocou essencialmente a técnica para agrupamento dos pontos de amostragem, sem analisar o grau de homogeneidade das regiões formadas. Fica portanto, como oportunidade de trabalho futuro, a análise das regiões geradas com relação aos seus respectivos níveis de uniformidade, bem como a avaliação de sua utilidade prática na gestão da qualidade das águas.

Referências

Assunção, R.M.; Lage, J.P.; Reis, E. A. (2002) Análise de conglomerados espaciais via árvore geradora mínima. *Revista Brasileira de Estatística*. Rio de Janeiro, Brasil, v. 63, n. 220, p. 7-22.

Berry, M. J. A.; Linoff, G. S. (2004) Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management. Indianapolis: Wiley Publishing, 672 p.

Bertholdo, L.; Silva, C. G.; Umbuzeiro, G. A.; Camolesi Jr., L. (2012) Técnicas de Mineração de Dados na Classificação de Ecotoxicidade de Água para Aplicação na Gestão de Corpos Hídricos. In: *VIII Congresso Nacional de Excelência em Gestão*, Niterói, Brasil, 20 p.

CETESB (2011) Relatório de Qualidade das Águas Superficiais do Estado de São Paulo – 2010. São Paulo: CETESB.

CETESB (2012) Institucional – Companhia Ambiental do Estado de São Paulo – Histórico. http://www.cetesb.sp.gov.br/institucional/institucional/52-Histórico.

CONAMA (2005) Conselho Nacional do Meio Ambiente. Resolução n. 357, de 17 de março de 2005. Brasília: CONAMA.

Diniz, R. B. N.; Soares, V. G.; Cabral, L. A. F. (2012) Uso de Técnicas de Mineração de Dados na Identificação de Áreas Hidrologicamente Homogêneas no Estado da Paraíba. *Revista Brasileira de Recursos Hídricos*. Porto Alegre, Brasil, v. 17, n. 1, p. 65-75.

Duarte, A. A. A.; Bertholdo, L.; Umbuzeiro, G. A.; Camolesi Jr., L.; Silva, C. G. (2011) Processamento e Visualização de Dados para a Descoberta de Conhecimento em Sistemas de Monitoramento de Qualidade de Água. In: *III Workshop de Computação Aplicada à Gestão do Meio Ambiente e Recursos Naturais*, Natal, p. 1409-1418.

Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P. (1996) From data mining to knowledge discovery: An overview. In: *Advances in Knowledge Discovery and Data Mining*, AAAI Press/The MIT Press, England, p. 37-54.

Jacobi, P. R.; Barbi, F. (2007) Democracia e participação na gestão dos recursos hídricos no Brasil. *Revista Katálysis*, Florianópolis, Brasil, v. 10, n. 2, p.237-244.

Karimipour, F.; Delavar, M. R.; Kinaie, M. (2005) Water Quality Management Using GIS Data Mining. *Journal of Environmental Informatics*. Canadá, v. 5, n. 2, p. 61-71.

Neves, M. C.; Câmara, G.; Assunção, R. M.; Freitas, C. C. (2002) Procedimentos Automáticos e Semi-automáticos de Regionalização por Árvore Geradora Mínima. In: *Brazilian Symposium on GeoInformatics*, GeoInfo2002, Caxambu-MG, Brasil, p. 109-116.

Ramachandra Rao, A.; Srinivas, V. V.. (2006) Regionalization of watersheds by hybrid-cluster analysis. In: *Journal Of Hydrology*, v. 318, n.1-4, p. 37 -56.

Seixas, A. J.; Nelson, F. F. E.; Beatriz, S. L. P. L. (2008) Mining spatial and temporal data to classify water quality: a case study. In: *Data Mining IX: Data Mining, Protection, Detection and Other Security Technologies*. Reino Unido, v. 40, p. 83-94.

Shyue, S.; Chen, C.; Chang, C. (2010) Association rule mining for evaluation of regional environments: Case study of Dapeng Bay, Taiwan. *International Journal of Innovative Computing, Information and Control*. v. 6, n. 8, p. 3425-3436.

Silva, I. A. F. (2007) Descoberta de Conhecimento em Base de Dados de Monitoramento Ambiental para Avaliação da Qualidade da Água. 2007. 134 p. Dissertação (Mestrado) – Universidade Federal de Mato Grosso, Cuiabá.

Tan, P.; Steinbach, M.; Kumar, V. (2009) Introdução ao Data Mining – Mineração de Dados. Rio de Janeiro: Editora Ciência Moderna. 900 p.

Umbuzeiro, G. A.; Lorenzetti, M. L. (2009) Fundamentos da Gestão da Qualidade das Águas: Resolução CONAMA 357/2005. Limeira-SP: Biblioteca da Unicamp/CPEA.

Von Sperling, M. (2007) Estudos e modelagem da qualidade da água de rios. Belo Horizonte: Departamento de Engenharia Sanitária e Ambiental – Universidade Federal de Minas Gerais. 588 p. v.7.