

UNIVERSIDADE ESTADUAL DE CAMPINAS
FACULDADE DE TECNOLOGIA

Leonardo Bertholdo

**CLASSIFICAÇÃO, ASSOCIAÇÃO E REGIONALIZAÇÃO DE
DADOS DE CORPOS HÍDRICOS: APLICAÇÃO NO
MONITORAMENTO DA ÁGUA NO ESTADO DE SÃO PAULO**

Limeira, 2013.

Leonardo Bertholdo

**CLASSIFICAÇÃO, ASSOCIAÇÃO E REGIONALIZAÇÃO DE
DADOS DE CORPOS HÍDRICOS: APLICAÇÃO NO
MONITORAMENTO DA ÁGUA NO ESTADO DE SÃO PAULO**

Dissertação apresentada ao Curso de
Mestrado da Faculdade de Tecnologia da
Universidade Estadual de Campinas, como
requisito para a obtenção do título de Mestre
em Tecnologia.

Área de Concentração: Tecnologia e Inovação

Orientador: Prof. Dr. Luiz Camolesi Júnior

Limeira, 2013.

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca da Faculdade de Tecnologia
Vanessa Evelyn Costa - CRB 8/8295

B461c Bertholdo, Leonardo, 1975-
Classificação, associação e regionalização de dados de corpos hídricos :
aplicação no monitoramento da água no estado de São Paulo / Leonardo
Bertholdo. – Limeira, SP : [s.n.], 2013.

Orientador: Luiz Camolesi Júnior.
Dissertação (mestrado) – Universidade Estadual de Campinas, Faculdade de
Tecnologia.

1. Monitoramento ambiental. 2. Mineração de dados. 3. Gestão de recursos
hídricos. I. Camolesi Júnior, Luiz. II. Universidade Estadual de Campinas.
Faculdade de Tecnologia. III. Título.

Informações para Biblioteca Digital

Título em outro idioma: Classification, association and regionalization of data of water bodies :
application in the monitoring of the water in the state of São Paulo

Palavras-chave em inglês:

Environmental monitoring

Data mining

Water resources management

Área de concentração: Tecnologia e Inovação

Titulação: Mestre em Tecnologia

Banca examinadora:

Luiz Camolesi Júnior [Orientador]

João Eduardo Ferreira

Marco Antonio Garcia de Carvalho

Data de defesa: 10-07-2013

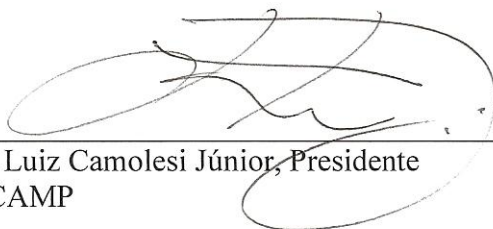
Programa de Pós-Graduação: Tecnologia

DISSERTAÇÃO DE MESTRADO EM TECNOLOGIA
ÁREA DE CONCENTRAÇÃO: TECNOLOGIA E INOVAÇÃO

CLASSIFICAÇÃO, ASSOCIAÇÃO E REGIONALIZAÇÃO DE DADOS DE CORPOS
HÍDRICOS: APLICAÇÃO NO MONITORAMENTO DA ÁGUA NO ESTADO DE SÃO
PAULO.

Autor: Leonardo Bertholdo

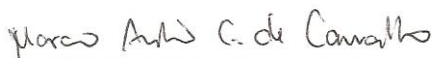
A Banca Examinadora composta pelos membros abaixo aprovou esta Dissertação:



Prof. Dr. Luiz Camolesi Júnior, Presidente
FT/UNICAMP



Prof. Dr. João Eduardo Ferreira
IME/USP



Prof. Dr. Marco Antonio Garcia de Carvalho
FT/UNICAMP

Agradecimentos

Antes de mais nada agradeço a Deus por ter me dado saúde física e mental para enfrentar os desafios pelos quais passei nos últimos três anos.

À minha querida mãe especialmente pelo apoio sempre incondicional aos meus estudos, a despeito de todas as dificuldades inerentes às circunstâncias em que vivíamos. Ao meu pai por compartilhar comigo tantas experiências de vida, positivas e negativas, que em muitos momentos me foram e são tão úteis, certamente permanecerão eternas em minha memória.

À minha esposa Ana Paula pelo amor e companheirismo em todos os momentos e por me ensinar que um pouco de otimismo e uma certa dose de ousadia são ingredientes essenciais para se realizar grandes conquistas.

Ao meu orientador, o professor Luiz Camolesi, pelo acompanhamento e compreensão durante todo o mestrado e pela confiança depositada em mim. À professora Gisela, pelas valiosíssimas contribuições na área de saneamento ambiental, e também a sua orientanda Francine Vacchi pelo apoio dado. Ao professor Celmar pelos questionamentos levantados, pela colaboração na área de visualização da informação, e também ao seu ex-bolsista Bruno Oliveira pelo apoio na fase inicial desse trabalho. A todos esses professores, agradeço especialmente pelas sempre importantes contribuições realizadas durante as revisões dos artigos que juntos publicamos.

Por fim, agradeço a todas as outras pessoas que direta ou indiretamente contribuíram de alguma forma para que este trabalho se concretizasse.

Resumo

A aplicação de recursos computacionais avançados no suporte aos sistemas de gestão ambiental vem se tornando cada vez mais frequente nas últimas décadas. A capacidade de processar e explorar grandes volumes de dados de forma sistemática, inerente a tais recursos, possibilita a extração de informações abrangentes e sintéticas, as quais podem servir como um importante insumo para o processo de controle ambiental. Nesse trabalho são empregadas técnicas de mineração de dados para a descoberta de conhecimento implícito no domínio de monitoramento de qualidade de água em corpos hídricos. A pesquisa está dividida em três frentes: a primeira busca descobrir regras de classificação de ecotoxicidade em amostras de água por meio de uma técnica de modelagem previsiva. Na segunda parte do estudo emprega-se uma técnica de análise associativa para investigar a presença de relacionamentos fortes entre os parâmetros que medem a qualidade de água. Por fim, a última frente utiliza uma abordagem de regionalização para encontrar pontos de amostragem de água similares com relação às medições de seus parâmetros de qualidade. Os resultados obtidos proporcionaram algumas descobertas, entre elas: a associação de determinados parâmetros de qualidade à toxicidade crônica da água, a existência de correlações entre alguns dos parâmetros de qualidade de água e a presença de grupos homogêneos entre os pontos de amostragem de água.

Palavras chave: Monitoramento ambiental, Mineração de dados, Gestão de recursos hídricos.

Abstract

The application of advanced computational resources at the support to the environmental management systems is becoming increasingly frequent in recent decades. The ability to process and explore large volumes of data in systematic way, inherent in these resources, makes it possible to extract information comprehensive and synthetic, which can serve as an important input to the environment control process. This work used data mining techniques to discover implicit knowledge in the field of monitoring water quality in water bodies. The research is divided into three fronts: the first seeks to discover classification rules of ecotoxicity in water samples using a predictive modeling technique. In the second part of the study is used an associative analysis technique to investigate the presence of strong relationships between the parameters that measure the quality of water. Finally, the last front uses a approach of regionalization to find water sampling sites similar in relation to the measurements of their quality parameters. The results provided some discoveries, including: the association of certain quality parameters to the chronic toxicity of the water, the existence of correlations between some of the parameters of water quality and presence of homogeneous groups between the water sampling sites.

Key words: Environmental monitoring, Data mining, Water resources management.

Lista de Ilustrações

Figura 1: Classificação das 22 UGRHs por vocação (CETESB, 2012)	7
Figura 2: Medições do ponto de amostragem JUNA02020 no ano de 2008 (CETESB, 2008)	10
Figura 3: Etapas que compõem o processo de KDD. Traduzido de Fayyad et al. (1996)	14
Figura 4: Exemplo de classificação. Adaptado de Tan et al. (2009)	22
Figura 5: Modelo de classificação baseado em regras	23
Figura 6: Modelo para geração de regras de associação	25
Figura 7: Processo de KDD adaptado	35
Figura 8: UGRHs consideradas após aplicação dos critérios gerais	37
Figura 9: Seleção dos pontos de amostragem em função dos critérios aplicados	38
Figura 10: Fluxo de tarefas do pré-processamento de dados brutos	41
Figura 11: Esquema de conversão dos dados brutos	42
Figura 12: Diagrama Entidade-Relacionamento do banco de dados de medições	43
Figura 13: Estrutura das tabelas de pré-processamento	44
Figura 14: Estrutura das tabelas para mineração dos dados	44
Figura 15: Exemplo de conjunto de medição eliminado	45
Figura 16: Algoritmo de Cobertura Sequencial	53
Figura 17: Algoritmo Apriori	55
Figura 18: Algoritmo de Prim e Poda da AGM	58
Figura 19: Interface principal da ferramenta desenvolvida	59
Figura 20: Interface para imputação dos dados	60
Figura 21: Interface para discretização dos dados	61
Figura 22: Interface para eliminação de categorias	62
Figura 23: Interface para seleção de parâmetros	63
Figura 24. Ferramenta para classificação de toxicidade em amostras de água	64
Figura 25. Visualização da cobertura e precisão das regras geradas	65
Figura 26. Visualização do desempenho do conjunto de regras aplicado	66
Figura 27. Ferramenta para geração de regras de associação entre os parâmetros	

de qualidade de água	68
Figura 28. Visualização da quantidade de regras geradas em função do suporte e da confiança configurados	69
Figura 29. Filtro para visualização de regras específicas	70
Figura 30: Ferramenta para agrupamento de pontos de amostragem de água	71
Figura 31: Formação de ciclo entre pontos de amostragem de água	72
Figura 32: Janelas para informação da quantidade de grupos a serem formados	73
Figura 33: Geração de grupos de pontos de amostragem no Rio Paraíba do Sul	74

Lista de Tabelas

Tabela 1: Tipos de rede de monitoramento de água (CETESB, 2012)	6
Tabela 2: Classificação das águas doces conforme seus usos (Von Sperling, 2007)	8
Tabela 3: Principais parâmetros de qualidade de água (CETESB, 2012)	9
Tabela 4: Exemplo de análise associativa	24
Tabela 5: Exemplo de análise de grupos. Adaptado de Tan et al. (2009)	27
Tabela 6: Pontos de amostragem considerados após aplicação dos critérios específicos	38
Tabela 7: Parâmetros de qualidade considerados após aplicação dos critérios específicos	41
Tabela 8: Transformação dos identificadores dos parâmetros de qualidade	47
Tabela 9: Categorização dos parâmetros contínuos e discretos	48
Tabela 10: Medições médias normalizadas dos parâmetros Sólidos Totais e Turbidez no Rio Paraíba do Sul	74
Tabela 11: Ocorrências dos parâmetros nos experimentos específicos de classificação	100
Tabela 12: Ocorrências dos parâmetros nos antecedentes das regras geradas	103
Tabela 13: Ocorrências dos parâmetros nos experimentos específicos de associação.....	112
Tabela 14: Comparativo das combinações entre as categorias de parâmetros	114
Tabela 15: Subgrupos de pontos de amostragem identificados nos conjuntos de experimentos (UGRHIs 5, 6 e 10)	137
Tabela 16: Subgrupos de pontos de amostragem identificados nos conjuntos de experimentos (UGRHI 2)	137

Lista de Abreviaturas e Siglas

Abreviaturas

cm – Centímetros
mg/dL – Miligramas por decilitro
mg/L – Miligramas por litro
UV – Ultravioleta
°C – Grau Celsius
µS/cm – Micro Siemens por centímetro

Siglas

CETESB – Companhia Ambiental do Estado de São Paulo
CONAMA – Conselho Nacional do Meio Ambiente
IAP – Índice de qualidade das águas para fins de abastecimento público
IB – Índice de balneabilidade
IET – Índice do estado trófico
IQA – Índice de qualidade das águas
IVA – Índice de qualidade das águas para proteção da vida aquática
KDD – Knowledge Discovery in Databases
PDF – Portable Document Format
SGBDR – Sistema Gerenciador de Banco de Dados Relacional
SQL – Structured Query Language
UGRHI – Unidade de Gerenciamento de Recursos Hídricos
XML – eXtensible Markup Language

Sumário

1. Introdução	1
1.1 Contexto	1
1.2 Objetivos	2
1.3 Organização do Trabalho	3
2. Gestão de Recursos Hídricos	4
2.1 Breve Histórico	4
2.2 Rede de Monitoramento de Qualidade de Água no Estado de São Paulo	5
2.3 Parâmetros de Qualidade de Água	8
2.4 Índices de Qualidade de Água	11
3. Descoberta de Conhecimento e Aplicações	13
3.1 Processo para extração de informações	13
3.1.1 Seleção dos Dados	15
3.1.2 Pré-processamento dos Dados	15
3.1.3 Transformação dos Dados	16
3.1.4 Mineração dos Dados	17
3.1.5 Interpretação e Avaliação dos Padrões	17
3.2 Mineração de Dados	18
3.2.1 Classificação	21
3.2.2 Análise Associativa	24
3.2.3 Análise de Cluster	27
3.2.4 Outras Tarefas da Mineração de Dados	29
3.3 Visualização de Dados	30
3.4 Mineração de Dados Aplicada ao Monitoramento de Recursos Hídricos	31
4. Metodologia e Ferramenta	34
4.1 Processo de Descoberta de Conhecimento	34
4.1.1 Seleção dos Dados	36
4.1.2 Pré-processamento dos Dados	41
4.1.3 Mineração dos Dados	49
4.1.4 Visualização dos Dados	50

4.1.5	Interpretação e Avaliação dos Padrões	51
4.2	Método para Classificação de Toxicidade em Amostras de Água	51
4.3	Método para Identificação de Associações entre os Parâmetros de Qualidade	54
4.4	Método para Regionalização de Pontos de Amostragem de Água	55
4.5	Ferramenta para Descoberta de Conhecimento em Dados de Monitoramento de Água	59
4.5.1	Funcionalidades de Pré-processamento	59
4.5.2	Funcionalidades para Classificação de Toxicidade	63
4.5.3	Funcionalidades para Identificação de Associações entre Parâmetros	67
4.5.4	Funcionalidades para Regionalização de Pontos de Amostragem	70
5.	Experimentação	75
5.1	Classificação de Toxicidade	75
5.1.1	Configurações de Pré-processamento.....	75
5.1.2	Configurações de Mineração de Dados.....	76
5.1.3	Experimentos	76
5.2	Identificação de Associações entre Parâmetros	77
5.2.1	Configurações de Pré-processamento.....	78
5.2.2	Configurações de Mineração de Dados.....	78
5.2.3	Experimentos	79
5.3	Regionalização de Pontos de Amostragem	81
5.3.1	Configurações de Pré-processamento.....	81
5.3.2	Configurações de Mineração de Dados.....	81
5.3.3	Experimentos	82
6.	Conclusões	83
6.1	Contribuições	84
6.2	Dificuldades Encontradas	85
6.3	Trabalhos Futuros	85
6.4	Artigos Publicados e Aceitos	87
	Referências Bibliográficas	89
	Apêndice A – Classificação de Toxicidade em Amostras de Água – Resultados e Avaliações	93

Apêndice B – Identificação de Associações entre os Parâmetros de Qualidade de Água – Resultados e Avaliações	105
Apêndice C – Regionalização de Pontos de Amostragem de Água – Resultados e Avaliações	115
Apêndice D – Ferramentas e Bibliotecas Utilizadas	138

1 Introdução

Nesse capítulo são apresentados o contexto em que se insere esta pesquisa, os objetivos perseguidos, além de uma breve descrição da estrutura desta dissertação.

1.1 Contexto

O surgimento da vida em nosso planeta está intrinsecamente relacionado à existência da água. Ela é um dos elementos primordiais que compõem a biosfera, tornando possível a perpetuação dos organismos vivos. Nesse cenário, os corpos hídricos possuem um papel fundamental, pois transportam água para as mais remotas regiões terrestres, sendo responsáveis pelo equilíbrio de muitos ecossistemas, os quais dependem diretamente da água provida por esses cursos naturais. Na esfera humana, além de possibilitar nossa sobrevivência, a água fornecida pelos corpos d'água satisfaz as mais diversas necessidades, tais como abastecimento público e industrial, irrigação agrícola, produção de energia elétrica, transportes fluviais, atividades de lazer, entre outras.

Não obstante a essa realidade, a expansão demográfica e industrial das últimas décadas vem ocasionando o comprometimento de muitos corpos hídricos, como rios, lagos e reservatórios. A interferência do homem, por meio de ações como o despejo doméstico ou industrial e a aplicação de defensivos agrícolas no solo, contribui para a introdução de compostos tanto orgânicos como inorgânicos na água, afetando a sua qualidade. Vale destacar ainda que a água doce é um recurso natural limitado pelo alto custo da sua obtenção a partir de formas menos convencionais, como as águas marinhas e subterrâneas (Alves et al., 2008). Diante desse 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 desse bem.

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), a qual mantém atualmente quase 370 pontos fixos de coleta de amostras de água, localizados ao longo dos rios e reservatórios monitorados. Cada amostra é analisada sob aspectos físicos, químicos,

biológicos e toxicológicos, formando um valioso conjunto de dados contendo informações específicas referentes às condições ambientais dos corpos hídricos contemplados (CETESB, 2012).

Entretanto, devido ao enorme volume de dados produzidos, a análise deste conjunto por meio de métodos convencionais mostra-se ineficiente quando o objetivo é a descoberta de informações não triviais. Esse cenário torna indispensável o uso de técnicas especiais que, por meio de cruzamentos entre as informações, conseguem trazer à tona novos conhecimentos que podem auxiliar estrategicamente na gestão e nas tomadas de decisão referentes a um determinado domínio. Entre as técnicas utilizadas para este fim, está a mineração de dados, que, por meio da associação de diferentes áreas de conhecimento, propõe métodos para descoberta de informações implícitas e relevantes em bancos de dados. A mineração de dados pode ser aplicada aos mais diversos problemas, alguns exemplos são: tarefas preditivas para classificação de elementos em categorias, busca de associações entre variáveis do conjunto de dados, extração de grupos de dados intimamente relacionados, detecção de anomalias, entre outros.

1.2 Objetivos

Essa pesquisa tem como meta a descoberta de conhecimento útil e expressivo em meio a dados de monitoramento de água levantados pela CETESB entre os anos de 2005 e 2011, em uma região delimitada do estado de São Paulo. O estudo pode ser dividido em três partes, cada qual representando um dado tipo de análise, onde o intuito básico consiste em revelar informações interessantes ocultas que possam agregar significado aos dados considerados no estudo.

A primeira questão diz respeito à toxicidade da água, atualmente mensurada por meio de testes que determinam os efeitos tóxicos em organismos aquáticos, causados por um ou mais agentes químicos (CETESB, 2011). Nesse contexto, o objetivo é descobrir regras de classificação que permitam prever a ecotoxicidade de amostras de água somente com base nos valores medidos em outros parâmetros de qualidade. Uma vez descobertos, esses padrões poderiam ser utilizados na predição da toxicidade de futuras amostras de água.

A questão seguinte busca a descoberta de correlações entre os parâmetros de qualidade de água, visto que muitas das interações entre estes parâmetros ainda não foram esclarecidas de

forma conclusiva. Nesse trabalho, as regras de associação extraídas a partir dos dados de monitoramento de qualidade de água são representadas por meio de regras de associação (ou implicação), as quais revelam as correlações existentes entre os diversos parâmetros contemplados. Os padrões descobertos podem desvendar relacionamentos desconhecidos entre os parâmetros de qualidade de água.

Finalmente, a terceira questão tem como propósito encontrar grupos de pontos de amostragem de água com alto grau de similaridade entre as medições de seus parâmetros de qualidade, levando em conta especialmente a ecotoxicidade. Esse agrupamento pode gerar inferências interessantes a respeito das condições dos corpos d'água, além de gerar insumos úteis para a organização e planejamento das redes de monitoramento de qualidade de água.

Para viabilizar a busca destes três objetivos, foi desenvolvida uma ferramenta gráfica específica, cujo papel foi auxiliar no pré-processamento dos dados de monitoramento de qualidade de água mencionados e realizar, por meio de algoritmos, as tarefas de mineração de dados necessárias para a descoberta de regras e padrões.

1.3 Organização do Trabalho

Esta dissertação está organizada da seguinte forma: o Capítulo 2 apresenta um breve histórico da gestão de recursos hídricos no Brasil, destacando a atual rede de monitoramento de qualidade de água existente no estado de São Paulo. Em seguida, o Capítulo 3, descreve o processo de descoberta de conhecimento em base de dados detalhando sua etapa central, a mineração de dados, além de apresentar alguns trabalhos relacionados. No Capítulo 4 é explanada a metodologia empregada, bem como a ferramenta desenvolvida para aplicação das técnicas de mineração nos dados de monitoramento de água. Os experimentos realizados são apresentados no Capítulo 5. Por fim, o Capítulo 6 apresenta as considerações finais referentes a este trabalho, suas contribuições, as dificuldades encontradas durante a pesquisa e os possíveis trabalhos futuros.

2 Gestão de Recursos Hídricos

Este Capítulo apresenta um histórico resumido da gestão dos recursos hídricos no Brasil, destacando a rede de monitoramento de qualidade de água mantida pela CETESB no estado de São Paulo. Também são relacionados os principais parâmetros físicos, químicos, biológicos e toxicológicos, considerados nas análises laboratoriais das amostras de água, bem como os índices de qualidade utilizados na avaliação da qualidade da água.

2.1 Breve Histórico

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). Com a Constituição de 1988 a participação da sociedade civil na gestão dos recursos naturais e, especialmente na gestão das águas, passou a ser um preceito fundamental para nortear todas as políticas públicas do setor. No estado de São Paulo, a Constituição Estadual de 1989 já havia incorporado novos conceitos à questão dos recursos hídricos: a gestão descentralizada, participativa e integrada; a divisão por bacia hidrográfica; e o aproveitamento múltiplo dos recursos hídricos. Em 1991, ano marcado pela enorme mobilização em torno da despoluição do rio Tietê, o governo federal encaminhou ao Congresso Nacional o primeiro projeto de lei que tratava da Política Nacional de Recursos Hídricos. A sociedade manifestou através das organizações civis a necessidade de integração entre os sistemas de recursos hídricos e meio ambiente e o estado de São Paulo, instituiu, por meio da Lei 7.663, o Sistema Estadual de Recursos Hídricos (SOS Fundação Mata Atlântica, 2012).

A partir deste sistema, o território paulista foi dividido em 22 regiões hidrográficas, institui-se a gestão por bacia, com participação efetiva da sociedade civil no processo decisório. A Lei paulista reforçou preceitos do Código de Águas e da Constituição ao contemplar instrumentos de gestão, como o Plano de Bacias, a cobrança pelo uso da água e o Fundo Estadual de Recursos Hídricos, para utilização direta nos Comitês de Bacias – colegiados, com poder deliberativo, que reúnem representantes dos municípios, dos órgãos de Estado e da sociedade

civil organizada para gestão integrada, descentralizada e participativa das águas. Em 1997, foi sancionada a Lei 9.433 que define a Política Nacional de Recursos Hídricos, cujos objetivos são listados a seguir (SOS Fundação Mata Atlântica, 2012):

- Assegurar à atual e às futuras gerações a necessária disponibilidade de água, em padrões de qualidade adequados aos respectivos usos.
- A utilização racional e integrada dos recursos hídricos, incluindo o transporte aquaviário, com vistas ao desenvolvimento sustentável.
- A prevenção e a defesa contra eventos críticos, de origem natural ou decorrente do uso integrado dos recursos hídricos.

2.2 Rede de Monitoramento de Qualidade de Água no Estado de São Paulo

No estado de São Paulo, a implantação dos comitês de bacia hidrográfica e de outras agências ambientais descentralizadas sucedeu a criação de uma instituição que se tornou centro de referência para questões ambientais. A CETESB, criada em 1968, é responsável pelo controle, fiscalização, monitoramento e licenciamento de atividades geradoras de poluição, com a preocupação fundamental de preservar e recuperar a qualidade das águas, do ar e do solo (CETESB, 2013). Desde 1974, a CETESB analisa e acompanha a qualidade da água dos rios, lagos e reservatórios do estado de São Paulo por meio de uma ampla rede de monitoramento. Conforme CETESB (2012), dentre os principais objetivos desta rede estão:

- Fazer um diagnóstico da qualidade das águas superficiais do Estado avaliando sua conformidade com a legislação ambiental.
- Avaliar a evolução temporal da qualidade das águas superficiais do Estado.
- Identificar áreas prioritárias para o controle da poluição das águas, como trechos de rios e estuários onde a qualidade de água possa estar mais degradada, possibilitando ações preventivas e corretivas da CETESB e de outros órgãos.

- Subsidiar o diagnóstico e controle da qualidade das águas doces utilizadas para o abastecimento público, verificando se suas características são compatíveis com o tratamento existente, bem como para os múltiplos usos.
- Dar subsídio técnico para a execução dos Planos de Bacia e Relatórios de Situação dos Recursos Hídricos, para a cobrança do uso da água e para o estudo do enquadramento dos corpos hídricos.
- Fornecer subsídios para a implementação da Política Nacional de Saneamento Básico (Lei 11.445/2007).

O programa de monitoramento de qualidade de água da CETESB é formado por 4 tipos de rede, conforme apresentado na Tabela 1, os quais permitem um melhor diagnóstico da qualidade das águas visando seus múltiplos usos.

Tabela 1: Tipos de rede de monitoramento de água (CETESB, 2012).

Monitoramento CETESB	Objetivos	Início de Operação	Pontos	Frequência	Variáveis
Rede Básica	Fornecer um diagnóstico geral dos recursos hídricos no Estado de São Paulo.	1974	369	Semestral/ Bimestral	Físicas Químicas Biológicas
Rede de Sedimento	Complementar o diagnóstico da coluna d'água.	2002	25	Anual	Físicas Químicas Biológicas
Balneabilidade de Rios e Reservatórios	Informar as condições da água para recreação de contato primário/banho à população.	1994	28	Semanal/Mensal	Biológicas
Monitoramento Automático	Controle de fontes poluidoras domésticas e industriais, bem como controle da qualidade da água destinada ao abastecimento público.	1998	16	Horária	Físicas Químicas

A rede de monitoramento da CETESB cobre as 22 Unidades de Gerenciamento de Recursos Hídricos (UGRHs) do estado de São Paulo. Cada uma destas unidades possui vários pontos de amostragem, de onde são colhidas as amostras de água que, posteriormente, são analisadas em laboratório (CETESB, 2012). A Figura 1 mostra essa divisão, classificando as UGRHs em grupos conforme suas respectivas vocações.

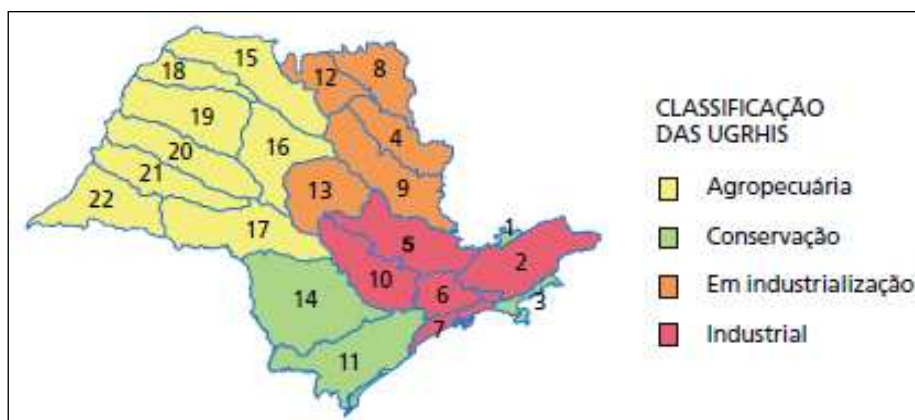


Figura 1: Classificação das 22 UGRHs por vocação (CETESB, 2012).

A análise laboratorial contempla dezenas de parâmetros, os quais podem estar associados a aspectos físicos, químicos, microbiológicos, hidrobiológicos e ecotoxicológicos. Em cada ponto de amostragem é medido um determinado conjunto de parâmetros, que muitas vezes difere de um ponto para outro. Estes dados são publicados anualmente pela CETESB por meio de arquivos em formato PDF disponíveis em seu portal na Internet. Cada arquivo contém as medições realizadas nas amostras coletadas em um dado ponto de amostragem. Somente a rede básica gera um volume anual aproximado 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.

Essas 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). A Tabela 2 mostra a classificação das águas doces em função de seus usos preponderantes, segundo a Resolução CONAMA 357/2005.

Tabela 2: Classificação das águas doces conforme seus usos (Von Sperling, 2007).

Uso	Especial	1	2	3	4
Abastecimento para consumo humano	X (a)	X (b)	X (c)	X (d)	
Preservação do equilíbrio natural das comunidades aquáticas	X				
Preservação de ambientes aquáticos em unid. de conserv. de proteção integral	X				
Proteção das comunidades aquáticas		X (h)	X		
Recreação de contato primário (*)		X	X		
Irrigação		X (e)	X (f)	X (g)	
Aqüicultura e atividade de pesca			X		
Pesca amadora				X	
Dessedentação de animais				X	
Recreação de contato secundário				X	
Navegação					X
Harmonia paisagística					X
Notas:					
a) com desinfecção					
b) após tratamento simplificado					
c) após tratamento convencional					
d) após tratamento convencional ou avançado					
e) hortaliças consumidas cruas e de frutas que se desenvolvam rentes ao solo e que sejam ingeridas cruas sem remoção de película					
f) hortaliças, plantas frutíferas e de parques, jardins, campos de esporte e lazer, com os quais o público possa vir a ter contato direto					
g) culturas arbóreas, cerealíferas e forrageiras					
h) de forma geral, e em comunidades indígenas					
(*) conforme Resolução CONAMA 274/2000 (balneabilidade)					

2.3 Parâmetros de Qualidade de Água

A análise das amostras de água contempla dezenas de variáveis, as quais podem estar relacionadas a aspectos físicos, químicos, microbiológicos, hidrobiológicos e toxicológicos. Apesar da grande diversidade de parâmetros de qualidade existentes, apenas parte destes são consideradas em cada ponto de amostragem, sendo que o conjunto de parâmetros analisados pode variar consideravelmente de um ponto para outro. A Tabela 3 apresenta os parâmetros ou variáveis de qualidade mais comuns contemplados nas análises.

Tabela 3: Principais parâmetros de qualidade de água (CETESB, 2012).

Grupo	Variáveis
Físicos	Condutividade, Cor Verdadeira, Série de Sólidos (Dissolvidos e Totais), Salinidade, Temperatura da Água e do Ar, Transparência e Turbidez
Químicos	Alumínio Total e Dissolvido, Ametrina, Arsênio, Atrazina, Bário, Boro, Cádmio, Carbono Orgânico Dissolvido, Carbono Orgânico Total, Chumbo, Cloreto, Clorpirifos, Cobre Total e Dissolvido, Compostos Orgânicos Voláteis, Cromo, Demanda Bioquímica de Oxigênio (DBO5,20), Demanda Química de Oxigênio (DQO), Demeton-O, Demeton-S, Dureza, Fenóis Totais, Ferro Total e Dissolvido, Fluoreto, Fósforo, Gution, Hidrocarbonetos Aromáticos Policíclicos (HPAs), Manganês, Malation, Mercúrio, Metolacolor, Molinato, Níquel, Óleos e Graxas, Oxigênio Dissolvido, Paration, Pendimetalina, pH, Potássio, Potencial de Formação de Trihalometanos, Propanil, Série Nitrogênio (Kjeldahl, Amoniacal, Nitrato e Nitrito), Simazina, Sódio, Surfactantes e Zinco.
Microbiológicos	<i>Escherichia coli</i> e Coliformes Termotolerantes, número de Células de Cianobactérias
Hidrobiológicos	Clorofila <i>a</i> e Feofitina <i>a</i> e Comunidades Fitoplancônica e Zooplancônica
Ecotoxicológicos	Microcistinas, Ensaio de Toxicidade Aguda com a bactéria luminescente – <i>V. fischeri</i> (Sistema Microtox); ensaio de Toxicidade Crônica com o microcrustáceo <i>Ceriodaphnia dubia</i> e ensaio de Mutação Reversa (teste de Ames).

Todos os anos, a CETESB disponibiliza publicamente em seu portal na Internet relatórios correspondentes a cada ponto de amostragem. Cada relatório contém as medições realizadas em todos os parâmetros de qualidade considerados em um determinado ponto. A Figura 2 apresenta um exemplo com as medições bimestrais realizadas no ponto JUNA02020, localizado no Rio Jundiáí, no ano de 2008.

A coluna “Padrão CONAMA” apresenta os limites de aceitação máximos e mínimos referentes aos parâmetros de qualidade, conforme os valores estabelecidos pela Resolução 357/2005. Estes limites podem variar conforme a classe à qual o ponto de amostragem está associado. Por exemplo: Para pontos da Classe 2, o valor máximo de Zinco Total é de 0,18 mg/L. Já para a Classe 3, o valor máximo é de 5 mg/L.

Nas seis colunas referentes às medições bimestrais de fevereiro à dezembro de 2008, além dos valores normais registrados, também existem alguns que merecem especial atenção:

- **Valores com “*”** – Indica que o valor medido está acima do Padrão CONAMA.
- **Valores com “<”** – Indica que o valor medido está abaixo do Padrão CONAMA, porém não é possível detectar o valor exato.
- **Valores com “i <”** – Indica que o valor medido pode estar abaixo ou acima do Padrão CONAMA, não sendo possível detectar o valor exato.

Os valores com “<” e “i <” ocorrem porque os níveis mínimos de detecção de um dado parâmetro podem variar dependendo da amostra de água. Por exemplo: amostras muito poluídas precisam ser diluídas para serem analisadas e, com isso, pode não ser mais possível detectar o nível de presença de um determinado parâmetro.

Resultados dos parâmetros e indicadores de qualidade das águas								
Código do Ponto : 00SP05245JUNA02020			Classe : 02			Ano : 2008		
UGRHI: PIRACICABA/CAPMARI/JUNDIAI								
Local : Rio Jundiaí - UGRHI 05 - Ponte na Av. Aderbal da Costa Madeira, 50m a jusante do lançamento da Krupp,(Ind. Siderúrgica).								
Descrição do Parâmetro	Unidade	Padrão CONAMA	18/02/2008	08/04/2008	11/06/2008	12/08/2008	15/10/2008	09/12/2008
			11h35	14h15	15h00	15h00	14h15	15h00
Parâmetro : Campo								
Chuva 24h	-		Sim	Não	Não	Não	Não	Não
Coloração	-		Marrom	Amarela	Amarela	Amarela	Amarela	Amarela
pH	U.pH	entre 6 e 9	6,6	7	6,3	6	6,6	6,5
Temp. Água	°C		23,5	24	18,5	19,5	26	24
Temp. Ar	°C		34,5	28	26	31	35	30
Parâmetro : Físico-Químicos								
Alumínio Dissolvid	mg/L	máximo 0,1	* 0,5	0,08	0,06	0,1	0,04	0,04
Cádmio Total	mg/L	máximo 0,001	< 0,001	* 0,002	< 0,001	< 0,001	< 0,001	< 0,001
Chumbo Total	mg/L	máximo 0,01	< 0,01	< 0,01	< 0,01	< 0,01	< 0,01	< 0,01
Cloreto Total	mg/L	máximo 250	8	8	8	9	16	13
Cobre Dissolvido	mg/L	máximo 0,009	* 0,08	* 0,01	0,001	0,001	0,002	0,002
Condutividade	µS/cm		103	109	120	118	179	164
DBO (5, 20)	mg/L	máximo 5	2	* 12	* 11	* 9	* 20	* 13
DQO	mg/L		< 50	< 50	< 50	< 50	< 50	< 50
Fenóis Totais	mg/L	máximo 0,003	i < 0,005	* 0,009	0,002	* 0,005	< 0,002	0,003
Ferro Dissolvido	mg/L	máximo 0,3	* 1	0,2	0,2	0,3	0,3	0,3
Fósforo Total	mg/L	máximo 0,1	* 0,2	* 0,3	* 0,3	* 0,4	* 0,6	* 0,3
Manganês Total	mg/L	máximo 0,1	0,1	* 0,2	0,1	* 0,2	* 0,2	* 0,3
Mercurio Total	mg/L	máximo 0,0002	< 0,0002	* 0,0003	< 0,0002	< 0,0002	< 0,0002	< 0,0002
N. Amoniacal	mg/L	máximo 3,7	1	0,7	2	0,8	3	3
Níquel Total	mg/L	máximo 0,025	< 0,01	< 0,01	0,01	< 0,01	i < 0,1	< 0,01
Nitrato	mg/L	máximo 10	1,1	4,7	2,5	1,1	1,3	1
Nitrito	mg/L	máximo 1	0,05	0,04	0,03	0,03	0,09	0,1
NKT	mg/L		1	0,7	4	0,9	3	3
OD	mg/L	mínimo 5	* 3,8	6,1	6,9	7,4	* 3,2	* 2,3
Sól. Dissolv. Total	mg/L	máximo 500	200	190	163	188	114	114
Sol. Total	mg/L		227	273	205	261	144	150
Subst. Tensoat.	mg/L	máximo 0,5	0,06	0,4	0,4	0,4	* 0,6	* 0,9
Sulfato Total	mg/L	máximo 250	< 2	< 2	< 2	6	9	6
Turbidez	UNT	máximo 100	36	72	24	72	23	25
Zinco Total	mg/L	máximo 0,18	0,01	0,1	0,1	0,03	0,02	0,02
Parâmetro : Microbiológicos								
Coli Termo	UFC/100mL	máximo 1000	* 79000	* 330000	* 130000	* 170000	* 230000	* > 200000
Parâmetro : Ecotoxicológicos								
Toxicidade	-	Não Tóxico	Não Tóxico			Não Tóxico		

Figura 2: Medições do ponto de amostragem JUNA02020 no ano de 2008 (CETESB, 2008).

2.4 Índices de Qualidade de Água

As análises dos parâmetros individuais são utilizadas na composição de diversos indicadores de sustentabilidade ambiental. Estes índices fornecem uma visão geral do estado em que se encontra um determinado recurso natural, pois integram os resultados de diversas variáveis por meio de um único indicador.

Os indicadores ambientais são utilizados na gestão do desenvolvimento sustentável, que pode ser definido como o desenvolvimento que atende às necessidades do presente sem comprometer a capacidade das futuras gerações atenderem às suas necessidades (Brundtland, 1987). Os indicadores ambientais são de grande valia, sobretudo por servirem de insumo para composição dos chamados “indicadores de sustentabilidade” que, segundo Maranhão (2007), representam um aprofundamento dos indicadores ambientais no sentido de integrar os territórios dos indicadores econômicos, sociais e ambientais, visto que o desenvolvimento sustentável requer um tipo de visão integrada do mundo.

Para avaliar a qualidade da água, a CETESB utiliza índices que integram as medições de diversos parâmetros físicos, químicos, biológicos e toxicológicos, por meio de um determinado número (CETESB, 2012). São eles:

- **Índice de qualidade das águas (IQA)** – Considera os parâmetros de qualidade que indicam o lançamento de efluentes sanitários para o corpo d’água, fornecendo uma visão geral sobre as condições de qualidade das águas superficiais. Este índice é calculado para todos os pontos da rede básica. Parâmetros de qualidade avaliados: Temperatura, pH, Oxigênio Dissolvido, Demanda Bioquímica de Oxigênio, Escherichia coli ou Coliformes Termotolerantes, Nitrogênio Total, Fósforo Total, Sólidos Totais e Turbidez.
- **Índice de qualidade das águas para fins de abastecimento público (IAP)** – Contempla, além dos parâmetros considerados no IQA, as substâncias tóxicas e os parâmetros que afetam a qualidade organoléptica da água, advindas principalmente de fontes difusas. Este índice é calculado apenas nos pontos que são coincidentes com captações utilizadas para abastecimento público. Parâmetros de qualidade avaliados: Temperatura, pH, Oxigênio Dissolvido, Demanda Bioquímica de Oxigênio, Escherichia

coli ou Coliformes Termotolerantes, Nitrogênio Total, Fósforo Total, Sólidos Totais, Turbidez, Ferro, Manganês, Alumínio, Cobre, Zinco, Potencial de Formação de Trihalometanos, Número de Células de Cianobactérias (Ambiente Lântico), Cádmio, Chumbo, Cromo Total, Mercúrio e Níquel.

- **Índice do estado trófico (IET)** – Avalia a qualidade da água quanto ao enriquecimento por nutrientes e seu efeito relacionado ao crescimento excessivo das algas. Assim como o IQA, este índice é calculado para todos os pontos da rede básica. Parâmetros de qualidade avaliados: Clorofila a e Fósforo Total.
- **Índice de qualidade das águas para proteção da vida aquática (IVA)** – Considera, além dos parâmetros do IET, os parâmetros essenciais para a vida aquática. Parâmetros de qualidade avaliados: Oxigênio Dissolvido, pH, Ensaio Ecotoxicológico com Ceriodaphnia dubia, Cobre Dissolvido, Zinco, Chumbo, Cromo, Mercúrio, Níquel, Cádmio, Surfactantes, Clorofila a e Fósforo Total.
- **Índice de balneabilidade (IB)** – Indica as condições de contato primário das praias de água doce. Os reservatórios impactados por lançamentos domésticos são avaliados semanalmente, enquanto que aqueles em melhores condições, mensalmente. Parâmetros de qualidade avaliados: Escherichia coli, Coliformes Termotolerantes e Enterococos.

3 Descoberta de Conhecimento e Aplicações

Este Capítulo apresenta o processo de descoberta de conhecimento em base de dados, fazendo um breve resumo de cada uma de suas fases. É dado um enfoque especial para a etapa central deste processo, a mineração de dados. Por fim, são apresentados alguns trabalhos onde são empregadas técnicas de mineração de dados na área ambiental.

3.1 Processo para extração de informações

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 (Silva, 2007). Em contrapartida, institutos científicos, indústrias, corporações e governos vêm acumulando volumes cada vez maiores de dados, impulsionados também pela versatilidade e alcance proporcionados pela Internet (Silva, 2004).

As análises realizadas pela CETESB originam um grande e valioso conjunto de informações referentes à qualidade da água dos corpos hídricos. No entanto, se analisadas por meio de técnicas convencionais, a descoberta de informações que possam contribuir para a gestão da qualidade de água torna-se bastante improvável.

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 (*Knowledge Discovery in Databases*) seja um dos mais difundidos no meio computacional. 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.

O processo de KDD é interativo, iterativo, cognitivo e exploratório (Silva, 2004). Interativo porque o usuário pode intervir no processo e controlar o curso das atividades. Iterativo porque cada etapa depende dos resultados da etapa anterior, além disso, as etapas podem se repetir caso o resultado esperado não seja alcançado. O KDD também é cognitivo, pois envolve capacidade de percepção e interpretação dos dados por parte do usuário. Por fim, é um processo exploratório, visto que busca padrões e hipóteses em meio aos dados.

Este processo é formado por uma série de etapas, que compreende desde a seleção do conjunto de dados a ser estudado até a interpretação dos padrões e regras geradas por técnicas de mineração de dados. A Figura 3 apresenta as cinco fases que compõem o KDD.

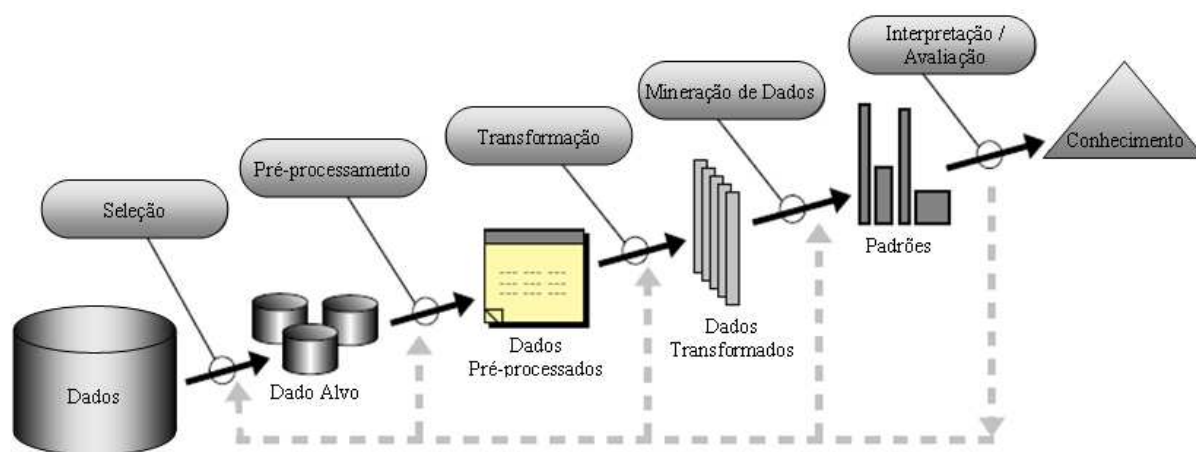


Figura 3: Etapas que compõem o processo de KDD. Traduzido de Fayyad et al. (1996).

Na etapa de seleção é escolhido o conjunto de dados a ser estudado, contendo todas as variáveis que possuem chance de serem utilizadas. No pré-processamento são realizados ajustes nos dados selecionados, como por exemplo: eliminação de dados redundantes, recuperação de dados incompletos e tratamento de dados discrepantes (*outliers*). A fase de transformação contempla a normalização e a centralização dos dados, de modo a reduzir o tempo de processamento da mineração. A etapa de mineração de dados é onde são implementados os algoritmos responsáveis pelo levantamento de padrões e regras implícitos em meio ao conjunto de dados. Por fim, a interpretação e avaliação analisam os resultados obtidos na etapa anterior, visando entender o significado e identificar a utilidade das informações descobertas (Prass, 2004). Na maior parte deste processo, é fundamental a cooperação de um especialista no domínio tratado, cujos conhecimentos e 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). As etapas deste processo são detalhadas nas próximas Seções.

3.1.1 Seleção dos Dados

O primeiro passo no processo de descoberta de conhecimento é a seleção do conjunto de dados a ser estudado. Essa etapa pode se tornar bastante complexa, caso os dados sejam oriundos de diferentes fontes como *data warehouses*, planilhas, sistemas legados (Prass, 2004). Além disso, este passo tem grande impacto sobre o resultado final do processo, uma vez que um conjunto de dados inadequado pode não proporcionar as descobertas esperadas, ou pior ainda, propiciar falsas descobertas.

Nesta etapa devem ser selecionadas todas as variáveis que possuem chance de serem utilizadas durante o processo. Tomando como exemplo um banco de dados relacional, essa tarefa seria equivalente à definição de quais tabelas e atributos serão utilizados na mineração dos dados. Uma vez selecionadas as variáveis, deve-se escolher um conjunto de registros que satisfaçam uma determinada condição. Por exemplo: todos os registros de venda dos três principais modelos de carro fabricados entre os anos de 2007 a 2012.

3.1.2 Pré-processamento dos Dados

A etapa de pré-processamento tem importante papel no processo de KDD, uma vez que a qualidade dos dados determina a eficiência dos algoritmos de mineração. Normalmente, o conjunto de dados definido na fase de seleção necessita de alguns ajustes, como por exemplo: eliminação de dados redundantes e inconsistentes, recuperação de dados incompletos e tratamento de dados discrepantes com relação ao conjunto (*outliers*). Assim como na etapa de seleção, nessa fase o auxílio de um especialista no domínio é essencial. Segundo Prass (2004), entre os problemas mais comuns encontrados em conjuntos de dados estão:

- **Dados ausentes (*missing values*):** Ausência de valores para determinadas variáveis, ou seja, registros com dados incompletos. Por exemplo: um banco de dados com medições de qualidade do ar no qual alguns dias possuem informação do nível de dióxido de enxofre, porém outros dias não a possuem. Algumas das alternativas para solucionar esse problema são: imputação (preenchimento manual dos dados ausentes por um especialista no domínio ou preenchimento automático via software); substituição do valor faltante pela média aritmética da variável; exclusão do registro inteiro (técnica radical, porém elimina o risco de impacto na confiabilidade dos resultados).
- **Dados inexistentes:** Necessidade de informações que não existem no conjunto de dados. Por exemplo: um cadastro de clientes onde não há informação referente à idade destes. A solução nesse caso consiste em fazer a transformação ou a combinação de outros dados relacionados a fim de se obter o dado inexistente. No exemplo mencionado, um possível recurso seria obter a idade dos clientes a partir de suas datas de nascimento, considerando que essa informação existe no cadastro.
- **Dados discrepantes (*outliers*):** Dados que possuem valores extremos, atípicos ou bastante distintos dos demais registros. Por exemplo: um paciente diabético, com histórico semanal de glicemia em torno de 100 mg/dL, que em uma dada semana tem um pico de 1000 mg/dL. Para esse problema, a exclusão do dado pode ser a solução, desde que exista a certeza de que o dado é consequência de um erro, por exemplo, a inserção equivocada de um zero a mais. Os *outliers* devem ser cuidadosamente analisados, pois podem representar, por exemplo, comportamentos não usuais, tendências ou até mesmo transações ilegais.

3.1.3 Transformação dos Dados

A terceira etapa do processo de KDD tem como intuito uniformizar os dados selecionados e pré-processados nas fases anteriores, de modo a reduzir o tempo de processamento dos algoritmos de mineração. Os dados são formatados para que obedeçam a um padrão, permitindo futuras comparações sem que haja a necessidade de executar conversões durante a realização das

consultas. Também é nesta etapa que os dados dispersos entre as diferentes fontes de dados costumam ser centralizados em um repositório único.

Um exemplo dessa transformação poderia ser aplicado no seguinte caso: os dados de sexo provenientes da fonte de dados A estão no formato “M/F”, já os dados de sexo provenientes da fonte de dados B estão no formato “Masculino/Feminino”. Nesse caso, é essencial que os valores do atributo “sexo” sejam padronizados, de modo que seus formatos sejam uniformes independentemente da fonte de dados que os proveu.

3.1.4 Mineração dos Dados

Dentre as cinco etapas do processo de KDD, a mineração de dados pode ser considerada a principal, pois é nessa fase em que são extraídas de fato 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 et al., 2004). Para isso, utiliza-se 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.

Apesar de a mineração de dados ser considerada a etapa central do processo de descoberta de conhecimento, as três etapas anteriores são as mais custosas e demoradas devido à série de atividades inerentes à preparação dos dados que serão minerados. Esta etapa do processo de KDD é explicada com mais detalhes na Seção 3.2.

3.1.5 Interpretação e Avaliação dos Padrões

O estágio final do processo de KDD tem como função interpretar e avaliar as informações obtidas na etapa de mineração de dados e, com isso, alcançar o objetivo de descobrir informações relevantes e antes desconhecidas na base de dados. Assim como ocorre nos dois primeiros passos do processo de KDD, nessa etapa a participação de um ou mais especialistas no domínio é essencial.

É bastante comum que o resultado obtido seja inconclusivo. Nesse caso, o processo pode retornar a qualquer uma das etapas anteriores, inclusive à etapa de seleção dos dados. De acordo com Prass (2004), as ações tomadas mais frequentemente para essa situação são: modificar o conjunto de dados inicial, trocar o algoritmo de mineração ou ambas.

3.2 Mineração de Dados

O enorme volume de informações armazenadas nos repositórios de dados atuais podem guardar conhecimentos valiosos que, sem a aplicação de métodos específicos, podem permanecer ignorados ou inacessíveis para o ser humano. Nos anos 90, surgiu uma linha de pesquisa na área da computação denominada *data mining*, ou mineração de dados. Essa abordagem estabeleceu técnicas especiais para extração de informações implícitas em bases de dados, ou seja, informações ocultas ou presentes nas “entrelinhas” de um determinado conjunto de dados. Esse conhecimento é obtido por meio da busca de padrões e relacionamentos entre as variáveis e seus dados, sendo normalmente expresso na forma de regras de classificação, associação e agrupamento.

A mineração de dados difere da recuperação de dados, visto que esta última se restringe à extração de informações explícitas de um conjunto de dados. Tal abordagem vem se mostrando bastante útil no sentido de proporcionar diretrizes para a transformação de dados brutos em informações de valor estratégico, sendo aplicada nos mais diferentes domínios.

Na área acadêmica, as aplicações da mineração de dados variam conforme o tipo de dado contemplado. Conforme Silva (2004), alguns exemplos de dados são:

- Dados armazenados em *data warehouses*: repositórios com dados de boa qualidade, integrados, estratégicos, históricos e com infraestrutura de processamento.
- Dados espaciais com elementos geográficos, imagens de sensoriamento remoto, imagens médicas, estudos ambientais, vigilância territorial, planejamento urbano, entre outros.
- Dados multimídia: extração de padrões a partir de animações, áudio, vídeo, imagens e textos.

- Dados de séries temporais: mercado de ações, processos de produção, experimentos científicos, tratamentos médicos, análises de tendências, de históricos e de similaridades.
- Dados textuais: informações não estruturadas disponíveis em artigos científicos, livros, e-mails, páginas Web.
- Dados da Web: imensos e complexos repositórios de dados de conteúdo, uso e estrutura, distribuídos de forma global por meio de uma ampla e rica coleção de hiperlinks.

Para tirar proveito dos imensos volumes de informação que coletam a todo instante, as empresas investem continuamente em novas tecnologias que agreguem valor ao negócio. As aplicações mais comuns ocorrem nos seguintes empreendimentos (Silva, 2004):

- Área financeira: detecção de fraudes e lavagem de dinheiro, análise de mercados, tendências de especulação, análise de crédito de consumidores, classificação de clientes para estratégias de marketing.
- Comércio: análise de vendas, comportamento dos clientes, giro de produtos, fenômenos sazonais, preferências regionais, avaliação de campanhas publicitárias, análise do grau de fidelidade dos clientes.
- Telecomunicações: detecção de invasões e comportamentos anômalos em sistemas, avaliação de uso e tráfego, análise de padrões de consumo.
- Genética: atividades de mapeamento de sequências, busca por similaridades e comparações, e identificação de sequências genética co-ocorrentes.
- Medicina: aperfeiçoamento de diagnósticos e tratamentos, mineração de imagens tomográficas, precisão na prescrição de exames e procedimentos.

Na mineração de dados, os problemas de descoberta de conhecimento são chamados de “tarefas”. Segundo Tan et al. (2009), normalmente, as tarefas de mineração são divididas em duas categorias principais, as tarefas previsivas e as tarefas descritivas.

Tarefas previsivas:

Buscam prever o valor de um dado atributo baseado nos valores de outros atributos. O atributo a ser previsto é chamado de variável dependente ou alvo, e os atributos usados para

previsão são chamados de variáveis independentes ou explicativas. Os algoritmos que implementam esse tipo de tarefa são conhecidos como algoritmos de **aprendizagem supervisionada**, pois as classes, ou possíveis valores, do atributo alvo de cada registro da base de treinamento são previamente conhecidas. Entre as tarefas previsivas estão a classificação – usada para variáveis alvo discretas – e a regressão – usada para variáveis alvo contínuas.

Exemplo: Supondo que uma loja de cosméticos deseje descobrir o perfil das mulheres que comprem cosméticos importados, pode-se extrair uma regra preditiva que diga, por exemplo, que mulheres com renda superior a US\$ 5.000,00 e idade entre 30 e 50 anos tendem a comprar esse tipo de cosmético. Nesse caso, os atributos Sexo, Renda e Idade são as variáveis explicativas, enquanto que o atributo Produtos Comprados é a variável alvo.

Tarefas descritivas

Buscam extrair padrões, como correlações, agrupamentos e anomalias, que resumam os relacionamentos implícitos entre os dados. Normalmente são tarefas exploratórias e requerem métodos de pós-processamento para validar e explicar os resultados da mineração, como por exemplo técnicas de visualização de informação. Os algoritmos que implementam esse tipo de tarefa são denominados algoritmos de **aprendizagem não supervisionada**, pois além de não existir necessariamente um atributo alvo específico, as classes, ou possíveis valores, dos atributos de cada registro da base de treinamento não são conhecidas a priori. Exemplos de tarefas descritivas são a análise associativa, a análise de grupos (ou *clustering*) e a detecção de anomalias.

Exemplo: Supondo que uma montadora de automóveis deseje descobrir associações entre os dados de venda dos veículos, pode-se extrair uma regra associativa que diga, por exemplo, que no verão tendem a ocorrer mais vendas de carros esportivos vermelhos. Nesse caso, os atributos Estação do Ano, Categoria, Cor e Vendas encontram-se associados entre si.

Nas próximas Seções são detalhadas as três tarefas da mineração de dados contempladas nesta pesquisa: a classificação, a análise associativa e a análise de grupos.

3.2.1 Classificação

O principal objetivo da classificação é organizar elementos em categorias pré-definidas, gerando um modelo que reduza a chance de erro entre os valores previsto e real da variável alvo (Tan et al., 2009). Entre os algoritmos usados para este fim estão as Árvores de Decisão, os Classificadores Baseados em Regras, de Vizinheiro mais Próximo, bayesianos e as redes neurais artificiais.

A Figura 4 mostra um exemplo clássico da área de mineração de dados, que é a descoberta da espécie de uma flor com base no tamanho de suas pétalas. O conjunto de dados possui 150 flores Íris, as quais são divididas em três espécies: Setosa, Versicolour e Virginica. Essas flores foram categorizadas conforme o tamanho de suas pétalas da seguinte forma:

- **Largura das pétalas (cm):** Pequena (0–0,75), Média (0,75–1,75) e Grande (1,75– ∞).
- **Comprimento das pétalas (cm):** Pequena (0–2,5), Média (2,5–5) e Grande (5– ∞).

A partir dessas categorias foram criadas as seguintes regras de classificação:

- Largura e comprimento pequenos = Setosa.
- Largura e comprimento médios = Versicolour.
- Largura e comprimento grandes = Virginica.

Conforme se pode notar, o resultado da regra de classificação foi perfeito para a espécie Setosa, pois todos os seus pontos ficaram dentro de seu respectivo quadrante. Para as espécies Versicolour e Virginica, a grande maioria dos pontos também ficaram dentro de seus respectivos quadrantes, porém ocorreram algumas classificações indevidas, além de flores que ficaram sem classificação.

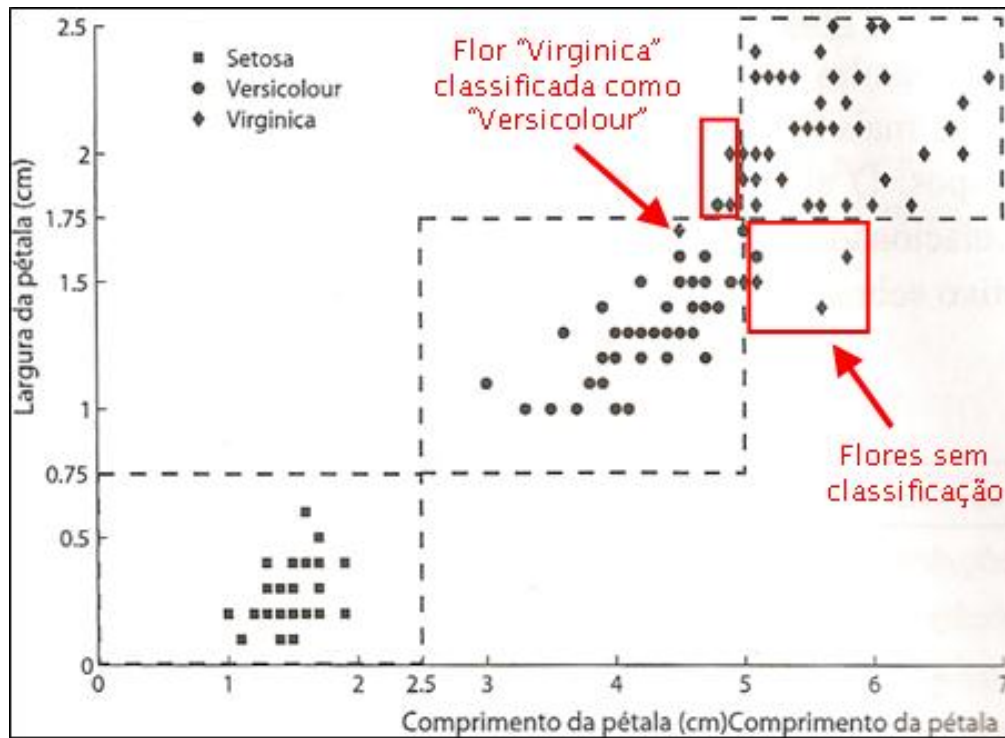


Figura 4: Exemplo de classificação. Adaptado de Tan et al. (2009).

Na maioria dos casos, os modelos previsivos gerados pelos algoritmos de classificação não conseguem gerar regras que se apliquem com total acurácia a todos os registros cobertos. No entanto, é possível adotar filtros para se definir níveis de cobertura e precisão aceitáveis para as regras extraídas, de modo a descartar aquelas menos confiáveis. Os conceitos de cobertura e precisão são explanados a seguir.

A construção de um modelo de classificação baseado em regras genérico é apresentado na Figura 5. Na fase inicial, um conjunto de treinamento, contendo registros cujas classes são conhecidas, é selecionado. Este conjunto é utilizado como insumo para construção do modelo de classificação, que nada mais é que o conjunto de regras de classificação encontrado. No momento seguinte, este modelo é aplicado a um conjunto de testes, contendo registros cujas classes devem ser previstas. Por fim, o desempenho do modelo é avaliado com base na taxa de erros ao classificar os registros da base de testes.

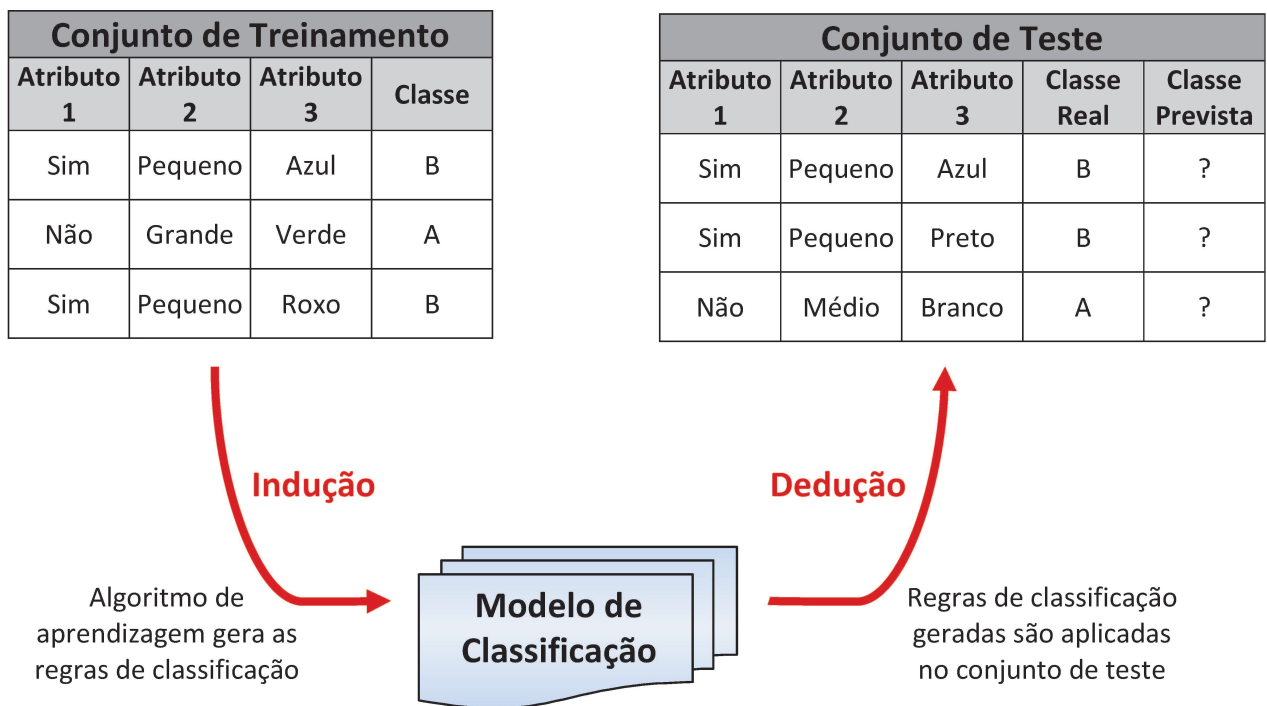


Figura 5: Modelo de classificação baseado em regras.

Durante a construção do modelo, o algoritmo gera um conjunto de regras condicionais a partir da base de dados de treinamento, sendo cada regra composta por um **antecedente**, também chamado de pré-condição, que contém os valores das variáveis explicativas, e um **consequente**, que contém o valor da classe prevista. Em seguida, estas regras aprendidas pelo algoritmo são aplicadas à base de testes, de modo a atribuir uma classe à cada registro desta base. Um exemplo de regra gerada a partir dos dados apresentados na Figura 5 seria:

***Se** Atributo1=Sim e Atributo2=Pequeno **Então** Classe=B*

Para avaliar a qualidade de uma regra de classificação existem medidas básicas como a **cobertura** e a **precisão**. A primeira visa determinar a taxa de registros que se enquadram no antecedente da regra e, portanto, disparam esta regra. A segunda define a taxa de registros que se enquadram tanto no antecedente quanto no consequente da regra e, portanto, além de disparar esta regra, também pertencem à classe prevista pela regra. Os cálculos destas medidas podem ser expressos da seguinte forma:

$$\text{Cobertura} = \frac{\text{Registros que satisfazem ao antecedente da regra}}{\text{Quantidade total de registros}}$$

$$\text{Precisão} = \frac{\text{Registros que satisfazem ao antecedente e ao consequente da regra}}{\text{Registros que satisfazem ao antecedente da regra}}$$

3.2.2 Análise Associativa

A análise associativa é usada para descobrir padrões que representem características altamente associadas em um conjunto de dados. Normalmente, os padrões descobertos são descritos na forma de regras de implicação ou subconjuntos de características (Tan et al., 2009). Ao contrário do que ocorre na classificação, aqui a classe a que cada amostra pertence não é conhecida a priori, tampouco se sabe o número de classes possíveis. Os algoritmos de análise associativa mais comuns são o Apriori, a Árvore de Padrão Freqüente, o Algoritmo por Amostragem e o Algoritmo de Partição (Elmasri et al., 2005).

A Tabela 4 apresenta um exemplo clássico que demonstra o conceito de análise associativa, onde são listadas dez compras realizadas em um armazém. Ao observar as compras, pode-se notar que aquelas que contêm o produto “Leite” também contêm o produto “Fraldas”. Tal fato indica uma clara associação entre a compra dos dois produtos, em outras palavras, a compra de leite implica na compra de fraldas, ou vice-versa.

Tabela 4: Exemplo de análise associativa.

Compras	Produtos
1	Arroz, Feijão, Leite, Fraldas
2	Café, Sal, Queijo, Peixe
3	Arroz, Feijão, Café, Leite, Fraldas, Ovos
4	Arroz, Feijão, Peixe, Carne
5	Ovos, Arroz, Feijão
6	Peixe, Leite, Fraldas
7	Arroz, iogurte, Sal, Ovos
8	Café, Sal, Carne, Ovos
9	Arroz, Leite, Fraldas, Cebola
10	iogurte, Ovos, Queijo, Leite, Fraldas

Assim como ocorre na modelagem previsiva, na maioria dos casos, os modelos associativos gerados pelos algoritmos de análise associativa não geram regras 100% confiáveis. Da mesma forma, pode-se adotar filtros para se definir níveis mínimos de suporte e confiança, conceitos que são apresentados a seguir.

A construção de um modelo para geração de regras de associação genérico é apresentado na Figura 6. 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**.

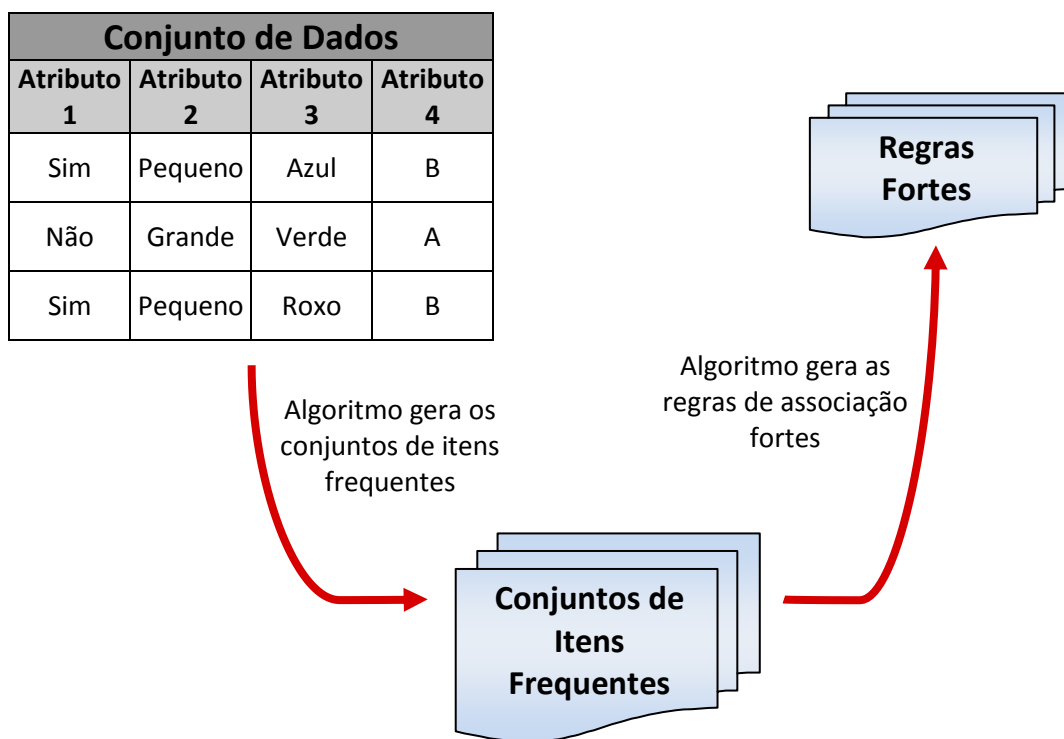


Figura 6: Modelo para geração de regras de associação.

Assim como na geração do modelo previsivo, apresentado na Seção anterior, o resultado final deste algoritmo são regras condicionais compostas por um **antecedente** e um **consequente**. A diferença é que, ao contrário da abordagem anterior, não há necessariamente uma variável alvo específica para a qual se deseja descobrir o valor. No enfoque da análise associativa, tanto o antecedente quanto o consequente da regra podem possuir de uma a n variáveis, desde que não se

repetam em ambos os lados da regra. Baseando-se nos dados apresentado na Figura 6, um exemplo de um conjunto de itens frequentes e as possíveis regras geradas a partir deste seria:

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*

*Se Atributo2=Pequeno e Atributo4=B **Então** Atributo1=Sim*

*Se Atributo2=Pequeno **Então** Atributo1=Sim e Atributo4=B*

*Se Atributo4=B **Então** Atributo1=Sim e Atributo2=Pequeno*

Durante o processo de geração das regras de associação, a qualidade da possível regra é avaliada por meio de medidas como o **suporte** e a **confiança**. A primeira é utilizada como base para a geração dos conjuntos de itens frequentes, e determina a taxa de registros que contém 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 contém o consequente da regra candidata dentre aqueles que possuem o antecedente desta regra. Estas medidas podem ser expressas da seguinte forma:

$$\text{Suporte} = \frac{\text{Quantidade de ocorrências do conjunto}}{\text{Quantidade total de registros}}$$

$$\text{Confiança} = \frac{\text{Quantidade de ocorrências do conjunto}}{\text{Quantidade de registros que possuem o antecedente da regra}}$$

3.2.3 Análise de Cluster

A análise de grupos, também conhecida por *clustering*, procura encontrar grupos de dados intimamente relacionados de modo que dados que pertençam ao mesmo grupo sejam mais semelhantes entre si do que com os que pertencem a outros grupos. Segundo Tan et al. (2009), quanto maior a semelhança dentro de um grupo e maior a diferença entre os grupos, melhor ou mais distinto será o agrupamento. Assim como ocorre na análise associativa, aqui a classe a que cada amostra pertence também não é conhecida a priori, tampouco se sabe a quantidade de classes possíveis. Alguns dos algoritmos de análise de grupos mais usados são o K-means, o BIRCH, o DBSCAN e o Agrupamento Hierárquico.

A Tabela 5 apresenta um exemplo simples de *clustering*, onde oito notícias são agrupadas conforme a categoria a que pertencem. Nesse contexto, o algoritmo de mineração busca grupos de palavras chave que tenham relação semântica entre si. O resultado é o agrupamento das notícias em duas categorias: Economia, destacada em vermelho, e Saúde, destacada em azul.

Tabela 5: Exemplo de análise de grupos. Adaptado de Tan et al. (2009).

Notícias	Palavras Chave
1	Dólar:1, Indústria:4, País:2, Empréstimo:3, Negócio:2, Governo:2
2	Maquinário:2, Trabalho:3, Mercado:4, Indústria:2, Emprego:3, País:1
3	Paciente:4, Sintoma:2, Remédio:3, Saúde:2, Clínica:2, Doutor:2
4	Emprego:5, Inflação:3, Aumento:2, Desemprego:2, Mercado:3, País:2
5	Farmacêutico:2, Empresa:3, Remédio:2, Vacina:1, Gripe:3
6	Morte:2, Câncer:4, Remédio:3, Pública:4, Saúde:3, Diretor:2
7	Doméstico:3, Previsão:2, Ganho:1, Mercado:2, Venda:3, Preço:2
8	Médico:2, Custo:3, Aumento:2, Paciente:2, Saúde:3, Cuidado:1

A análise de *cluster* pode ter diferentes enfoques, dependendo do tipo de agrupamento que se necessita gerar. Quando se deseja organizar objetos em grupos espaciais, um tipo de abordagem bastante utilizado é a **regionalização**. Conforme Neves et al. (2002), regionalização é um procedimento de agrupamento de objetos em regiões homogêneas e contíguas do espaço. Ela 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 e taxas associadas às unidades de área; redução dos efeitos da imprecisão nos valores das variáveis; redução dos erros associados ao posicionamento geográfico de eventos; e redução no custo de análise dos dados (Wise et al., 1997; Openshaw et al., 1995 apud Neves et al., 2002). Conforme Neves et al. (2002), existem três abordagens utilizadas para a condução da regionalização:

- **Primeira abordagem:** O processo se dá em dois estágios independentes. No primeiro estágio é executado um procedimento convencional de *clustering*, utilizando somente os atributos não-espaciais. No segundo estágio, os *clusters* são reavaliados, observando as relações de vizinhança dos objetos. Assim, objetos similares agrupados na fase inicial, mas sem contiguidade espacial, serão separados no segundo estágio formando regiões distintas. Segundo Wise et al. (1997) apud Neves et al. (2002), o inconveniente desta abordagem está na falta de controle sobre o número de regiões resultantes. Os casos com pequena dependência espacial entre os objetos, por exemplo, tenderão a produzir um número elevado de regiões.
- **Segunda abordagem:** A similaridade entre os objetos é avaliada considerando simultaneamente a posição geográfica dos objetos e seus atributos não-espaciais. Nessa implementação, a avaliação da similaridade contém duas componentes ponderadas, uma para os atributos não-espaciais e a outra para a distância geográfica. Assim, se o peso dado para a componente de distância for forte o suficiente, objetos com dados não-espaciais similares ficarão em grupos distintos após o processo de classificação, valendo o mesmo para a situação inversa. O ponto negativo aqui é que cada componente é uma função isolada, ou seja, as duas componentes utilizam em seus cálculos variáveis que medem fenômenos distintos e em unidades diferentes.
- **Terceira abordagem:** O relacionamento de vizinhança entre os objetos é explicitado por meio de dispositivos auxiliares, como matrizes ou grafos. No caso do uso de uma matriz, ela é chamada de matriz de contiguidade (C), onde cada elemento c_{ij} , indica se os objetos i e j são contíguos ou não. Dessa forma, $c_{ij} = 0$ para objetos não contíguos, e $c_{ij} = 1$ para objetos contíguos. De forma equivalente, quando é utilizado grafo, cada objeto é representado por um vértice. Quando os objetos são vizinhos, existe uma aresta ligando os dois vértices correspondentes no grafo (Maravalle et al., 1997); (Gordon, 1996).

3.2.4 Outras Tarefas da Mineração de Dados

Além das tarefas de classificação, análise associativa e análise de grupos existem outros tipos de problemas de descoberta de conhecimento, nos quais podem ser aplicadas técnicas de mineração de dados. A seguir, uma breve descrição de outras três tarefas bastante empregadas e estudadas no meio computacional.

- **Regressão ou Estimativa** – Tarefa preditiva semelhante à classificação, porém é utilizada para indução de variáveis alvo contínuas, ao invés de discretas. Entre os exemplos de aplicação da regressão estão: a previsão de demanda dos consumidores por um novo produto, a projeção de índices de investimentos financeiros, previsão da quantidade de precipitação em uma determinada região.
- **Deteção de Anomalias** – Também chamada de mineração de exceções, essa tarefa visa identificar dados cujas características são fortemente discrepantes do restante dos dados. Nesse caso, o desafio do algoritmo de mineração é diferenciar as anomalias verdadeiras das falsas, evitando assim rotular indevidamente dados normais como anômalos. Algumas das aplicações mais comuns desta tarefa são: deteção de fraudes, invasão de sistemas computacionais, perturbações no meio ambiente e diagnóstico de doenças.
- **Sumarização** – Tarefa cujo objetivo é encontrar descrições sintéticas que facilitem a compreensão das diferentes partes existentes em um conjunto de dados. A sumarização ganhou importância sobretudo após o advento da Internet, que demandou novos métodos para sumarizar o enorme volume de informações gerado. Exemplos de aplicação da sumarização envolvem: resenhas de notícias, resumos dos índices da bolsa de valores, sinopses de textos novelísticos, extratos de livros científicos, resumos de previsões meteorológicas.

3.3 Visualização de dados

A integração de técnicas de mineração de dados e visualização de informação é referenciada na literatura como Mineração Visual de Dados ou *Visual Data Mining*. Essa abordagem híbrida aproxima o usuário e o processo de descoberta de conhecimento em termos de técnicas de visualização eficientes, capacidade de interação e transferência de conhecimento (Rabello, 2007).

A união destas duas técnicas pode potencializar enormemente a exploração de informação, observando que a utilização intercalada pode causar penalidades relativas às deficiências e limitações de cada uma (Wong, 1999). Este mesmo autor define duas formas de integração das técnicas de visualização:

- Acoplamento forte, onde a visualização e o processo analítico são integrados em uma única ferramenta, aproveitando os pontos fortes de cada uma das áreas.
- Acoplamento fraco, onde as áreas são simplesmente intercaladas, possibilitando um aproveitamento parcial do potencial de cada uma delas no uso em conjunto.

De acordo com Han e Kamber (2000 apud Barioni, 2002), a visualização de informação e a mineração de dados podem ser integrados das seguintes formas:

- Os resultados da mineração de dados podem ser representados por meio de formas visuais.
- Dados armazenados em banco de dados podem ser visualizados sob diferentes níveis de abstração, podendo ser utilizadas diferentes combinações de atributos.
- Visualizar as etapas do KDD de forma que o usuário possa acompanhar o processo desde a extração dos dados até a apresentação do resultado.
- Mineração de dados visual: Ferramentas de visualização de informação podem ser utilizadas tanto para extrair conhecimentos quanto para a análise dos resultados obtidos por uma técnica de mineração de dados.

Ademais, em casos que envolvem grande volume de dados, o usuário pode selecionar porções da base de dados de interesse utilizando técnicas de visualização de informação, diminuindo assim a árdua tarefa exercida no entendimento dos resultados da mineração para grande volume de dados (Rabello, 2007).

3.4 Mineração de Dados Aplicada ao Monitoramento de Recursos Hídricos

As atividades de monitoramento ambiental exigem uma abordagem multidisciplinar envolvendo profissionais das áreas de química, física, biologia, matemática, estatística e ciência da computação (Artiola et al., 2004). Nesse contexto, a mineração de dados tem se mostrado bastante útil no sentido de proporcionar diretrizes para a transformação de dados brutos em informações de valor técnico e estratégico.

Conforme Silva (2007), a descoberta de conhecimento em bases de dados de monitoramento ambiental, utilizando técnicas de mineração de dados, pode ser uma ferramenta importante para o processo de tomada de decisão, realizado por órgãos e gestores de recursos hídricos. Há uma série de trabalhos científicos que aplicam conceitos de mineração de dados na descoberta de conhecimento em 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 para mineração de regras de associação e análise da árvore de decisão, com apoio das ferramentas de mineração de dados Weka e Clementine.

Fernandes et al. (2009) apresenta um sistema de *data warehousing* para armazenamento dos dados de qualidade da água de uma determinada região de Portugal. Além de organizar e uniformizar as informações em uma base de dados, a ferramenta procura auxiliar na descoberta do conhecimento através da aplicação das técnicas de mineração de dados, como a classificação e a regressão linear.

Magaia (2009) aborda o papel dos sistemas de suporte à decisão na análise da qualidade da água. O autor propõe o desenvolvimento de um sistema para este fim específico, o qual é empregado em uma estação de tratamento de água. A ferramenta tem como objetivo coletar e fornecer estruturas e meios para a exploração multidimensional dos dados, bem como a sua classificação e geração de modelos através de mecanismos de *data mining*.

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.

Belo et al. (2006) apresentam o sistema AQUA para análise e validação de parâmetros de qualidade da água. Primeiramente, são coletados os dados de leituras efetuadas pelas estações de coleta automática ou por técnicos de laboratório. Em seguida, esses dados coletados são analisados e validados e, posteriormente, integrados e consolidados numa base de dados específica. O sistema oferece mecanismos de detecção de anomalias e permite a exploração dos dados, tendo o apoio de algumas funcionalidades de georreferenciamento, geração de relatórios e gráficos. Apesar de não implementar técnicas de mineração de dados, o *data warehousing* do sistema está preparado de forma a assegurar futuras tarefas de mineração de 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.

Guimarães (2005) apresenta o desenvolvimento de um sistema de mineração de dados baseado em algoritmos genéticos denominado MinAG. A ferramenta realiza a tarefa de classificação de dados contínuos e destina-se a minerar propriedades do solo e da água. O estudo de caso referente à água utilizou-se de uma base de dados contendo medições de amostras de água no Estado da Flórida nos Estados Unidos, às quais contemplavam a análise de 47 diferentes parâmetros físicos e químicos.

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.

Ailamaki et al. (2003) descreve uma pesquisa interdisciplinar para descoberta de conhecimento em grandes bases de dados ambientais, com redes de sensores biológicos e químicos, a fim de melhorar a qualidade da água potável e a segurança na tomada de decisões a respeito desse tema. Para isso, são combinados algoritmos para mineração de dados espacial-temporal, métodos para modelar a qualidade da água e um sofisticado *framework* de análise de decisão.

Apesar de não tratar diretamente o tema da mineração de dados para monitoramento de qualidade de águas, Gibert et al. (2008) discute o papel do pré e pós-processamento no processo de descoberta de conhecimento em sistemas ambientais. Segundo o texto, estes sistemas são particularmente complexos e seus usuários exigem bastante clareza em seus resultados. O trabalho mostra possíveis caminhos para alcançar esse objetivo.

Comparativamente a essas pesquisas, este trabalho distingue-se por aplicar juntamente três importantes tarefas da mineração de dados, modelagem previsiva, análise associativa e análise de grupos, sobre um mesmo conjunto de dados, proporcionando assim diferentes perspectivas deste conjunto para uma análise mais abrangente dos especialistas na área de monitoramento ambiental.

4 Metodologia e Ferramenta

A metodologia desta pesquisa foi guiada pelo processo de descoberta de conhecimento denominado *Knowledge Discovery in Databases* (KDD). Conforme Fayyad et al. (1996), este processo é dividido em cinco etapas: Seleção, Pré-processamento, Transformação, Mineração de dados e Interpretação e Avaliação. Essa divisão pode sofrer variações dependendo do enfoque e das especificidades de cada trabalho. Alguns autores, por exemplo, tratam todos os procedimentos anteriores à mineração de dados como uma etapa única de “pré-processamento”, visto que são atividades fortemente relacionadas. Da mesma maneira, nesta pesquisa o processo de descoberta de conhecimento também foi adaptado para contemplar de forma mais satisfatória as características deste trabalho.

Quanto às técnicas de mineração de dados, para o problema da classificação de amostras de água, foi aplicada a abordagem de **modelagem previsiva**. Esta foi realizada por meio da técnica de classificação baseada em regras, onde os registros de uma base de dados são classificados a partir de regras obtidas por meio de um algoritmo de aprendizagem. A questão das correlações entre os parâmetros de qualidade de água foi tratada por meio de técnicas de **análise associativa**, a qual varre o conjunto de dados em busca de relacionamentos fortes entre as variáveis pesquisadas. Finalmente, para o problema do agrupamento dos pontos de amostragem de água, aplicou-se uma técnica para **análise de grupos** baseada em regionalização, que procura descobrir subconjuntos de objetos com características altamente semelhantes entre si. O processo de descoberta de conhecimento adaptado, as técnicas de mineração de dados empregadas, bem como a ferramenta desenvolvida para este trabalho são descritos nas próximas Seções deste Capítulo.

4.1 Processo de Descoberta de Conhecimento

Esta Seção apresenta a metodologia aplicada no processo de descoberta de conhecimento em meio aos dados de monitoramento de qualidade de água, abrangendo desde a seleção e preparação dos dados brutos até a etapa de interpretação e avaliação dos resultados obtidos pela mineração dos dados.

O processo de descoberta adaptado para este trabalho foi dividido nas seguintes etapas: seleção dos dados brutos, pré-processamento dos dados selecionados, mineração dos dados pré-processados, visualização dos padrões e regras obtidas e interpretação e avaliação dos resultados. A Figura 7 apresenta as cinco fases que compõem este processo.

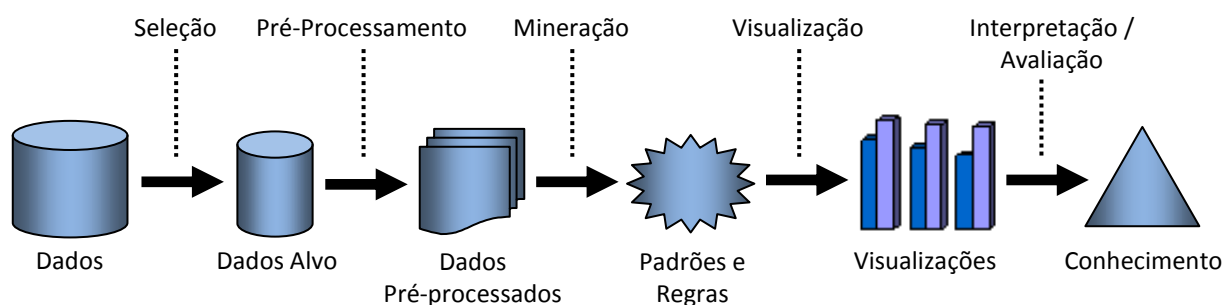


Figura 7: Processo de KDD adaptado.

As duas principais diferenças deste processo com relação ao proposto por Fayyad et al. (1996) são:

- Incorporação da etapa de Transformação à fase de Pré-processamento, que passou a contemplar as seguintes atividades: conversão dos dados brutos, centralização dos dados convertidos, imputação de dados faltantes, transformação, discretização e normalização dos dados.
- Inclusão da etapa de Visualização, que consiste no emprego de representações visuais para sintetizar os resultados obtidos pela mineração de dados, auxiliando assim na etapa seguinte de interpretação e avaliação.

Deve-se ressaltar que as etapas iniciais deste processo, seleção e pré-processamento, demandou a forte participação de uma especialista da área de saneamento ambiental, visando auxiliar na escolha e preparação dos dados. Da mesma forma, na etapa final, a cooperação desta especialista foi fundamental na interpretação e avaliação dos resultados obtidos.

4.1.1 Seleção dos Dados

Neste trabalho, o conjunto de dados a ser analisado foi selecionado com base em critérios gerais, relacionados a aspectos mais abrangentes dos dados, e critérios específicos, associados a características mais peculiares dos dados. Estes critérios foram aplicados por meio da verificação dos dados em estado bruto, no caso os arquivos de medições em formato PDF disponibilizados pela CETESB. Foi adotada a diretriz de se eliminar o máximo de dados possíveis por meio da simples observação visual, haja vista a dificuldade em se converter os dados brutos para um formato adequado a um banco de dados relacional. Nesse sentido, a estratégia empregada proporcionou uma redução e um maior controle dos casos de exceção durante o processo de conversão dos dados. A seguir, os critérios gerais e as respectivas descrições de como foram aplicados:

Critérios gerais para seleção dos dados:

- **Tipo de rede de monitoramento** – Foram selecionados somente pontos da Rede Básica, na qual as amostras de água são coletadas e analisadas manualmente. Nesta rede não são contempladas análises de sedimentos e balneabilidade dos rios, tampouco análises oriundas de sistemas de monitoramento automático. Esta rede abrange quase 85% dos pontos da rede de monitoramento da CETESB.
- **Aspecto temporal** – Foram contempladas as análises realizadas entre os anos de 2005 a 2011, nos quais os dados se mostraram com um maior grau de completude.
- **Aspecto espacial** – Das 22 UGRHIs existentes no estado de São Paulo, foram consideradas as UGRHIs: 2 (Paraíba do Sul), 5 (Piracicaba/Capivari/Jundiaí), 6 (Alto Tietê) e 10 (Sorocaba/Médio Tietê), conforme apresentado na Figura 8. O propósito foi selecionar as UGRHIs mais populosas, com aproximadamente 70% dos habitantes do estado, e fortemente industrializadas, uma vez que os rios de regiões com esse perfil normalmente são bastante impactados pela atividade industrial.

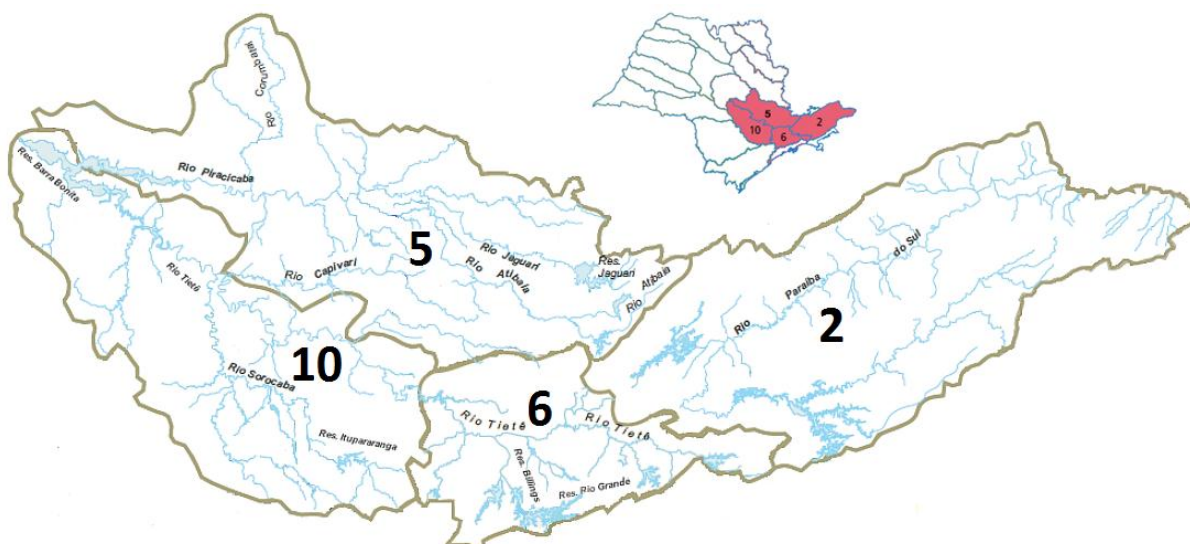


Figura 8: UGRHIs consideradas após aplicação dos critérios gerais.

Após a aplicação dos critérios gerais, dos 322 pontos de amostragem, existentes em média nos sete anos, permaneceram 167, todos localizados nas quatro UGRHIs selecionadas e integrantes da Rede Básica da CETESB.

Os critérios específicos para seleção dos dados levaram em conta especialmente a questão da completude, uma das premissas básicas para que a etapa de mineração de dados seja bem sucedida. A seguir, são apresentados cada um dos critérios específicos empregados na seleção dos dados, bem como a ordem em que foram aplicados:

Cr terios espec ficos para sele  o dos pontos de amostragem:

1. Somente pontos dos corpos hídricos que possuem dois ou mais pontos de amostragem.
2. Somente pontos que estão presentes em todos os anos.
3. Somente pontos que possuem análise de Toxicidade, parâmetro essencial neste estudo.
4. Somente pontos pertencentes à Classe 2. Para manter a uniformidade dos dados, foram descartados quatro pontos, dois pertencentes à Classe 0 (Especial) e dois pertencentes à Classe 3.

Após a aplicação destes critérios, resultou na permanência de 44 pontos de amostragem de água, considerados com maior riqueza e uniformidade nos dados. A Figura 9 mostra como os pontos de amostragem foram selecionados em função dos critérios aplicados:

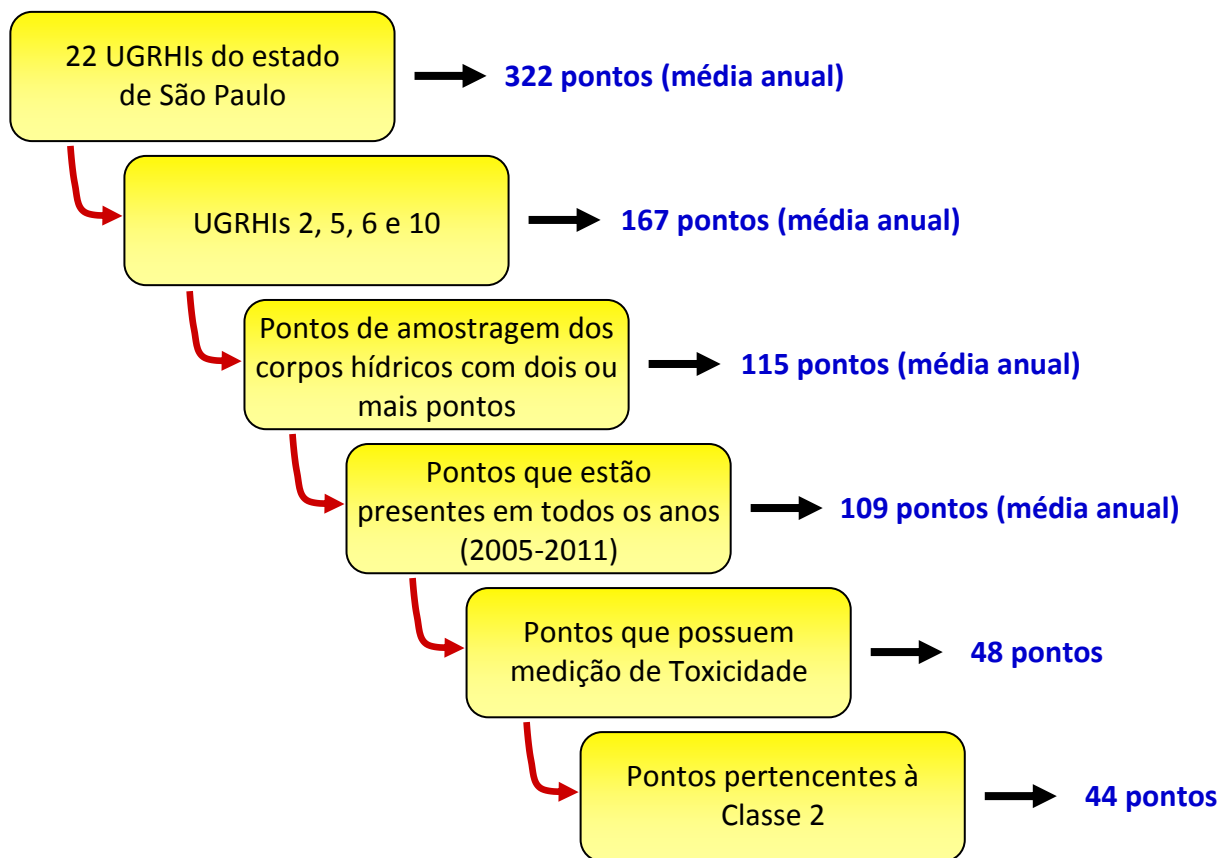


Figura 9: Seleção dos pontos de amostragem em função dos critérios aplicados.

A Tabela 6 apresenta a descrição de cada um dos 44 pontos de amostragem selecionados.

Tabela 6: Pontos de amostragem considerados após aplicação dos critérios.

UGRHI	Ponto	Corpo Hídrico	Localização	Município
2	PARB02050	Rio Paraíba do Sul	Captação de Santa Branca, no bairro Angola de Cima.	Santa Branca
2	PARB02100	Rio Paraíba do Sul	Ponte na rodovia SP-77, no trecho que liga Jacareí a Santa Branca.	Santa Branca
2	PARB02200	Rio Paraíba do Sul	Junto à captação do município de Jacareí	Jacareí
2	PARB02300	Rio Paraíba do Sul	Ponte de acesso ao loteamento Urbanova, em São José dos Campos.	São José Dos Campos
2	PARB02310	Rio Paraíba do Sul	Na captação de São José dos Campos, no canal de adução com extensão de 750m.	São José Dos Campos
2	PARB02400	Rio Paraíba do Sul	Ponte na rua do Porto, no trecho que liga Caçapava ao bairro Menino Jesus.	Caçapava
2	PARB02490	Rio Paraíba do Sul	Na captação da SABESP em Taubaté que abastece Tremembé	Tremembé

UGRHI	Ponto	Corpo Hídrico	Localização	Município
2	PARB02530	Rio Paraíba do Sul	Na captação da SABESP de Pindamonhangaba.	Pindamonhangaba
2	PARB02600	Rio Paraíba do Sul	Na captação de Aparecida.	Aparecida
2	PARB02700	Rio Paraíba do Sul	Ponte na rodovia BR-459, no trecho que liga Lorena a Piquete.	Lorena
2	PARB02900	Rio Paraíba do Sul	Ponte na cidade de Queluz.	Queluz
5	ATIB02010	Rio Atibaia	Junto à captação do município de Atibaia.	Atibaia
5	ATIB02065	Rio Atibaia	Na captação de Campinas, na divisa entre os municípios de Campinas e Valinhos.	Campinas
5	ATIB02605	Rio Atibaia	Ponte da Rodovia SP 332 que liga Campinas a Cosmópolis.	Paulínia
5	CPIV02130	Rio Capivari	Na captação de Campinas-ETA Capivari na Rodovia dos Bandeirantes.	Campinas
5	CPIV02900	Rio Capivari	Ponte no canavial, próximo à foz do Rio Tietê.	Tietê
5	CRUM02200	Rio Corumbataí	Ponte na Estr. Assistência/Paraisolândia.	Rio Claro
5	CRUM02500	Rio Corumbataí	Na captação de Piracicaba.	Piracicaba
5	JAGR02100	Rio Jaguari	Ponte na rodovia SP 95 no trecho que liga Bragança Paulista/Amparo (Km 9).	Bragança Paulista
5	JAGR02500	Rio Jaguari	Na ponte da rodovia SP-332, próximo às captações de Paulínia e Hortolândia.	Paulínia
5	JAGR02800	Rio Jaguari	Na captação de Limeira.	Americana
5	PCAB02100	Rio Piracicaba	Junto à captação de água de Americana, na localidade de Carioba.	Americana
5	PCAB02135	Rio Piracicaba	Na ponte de concreto da estrada Americana-Limeira, divisa de Limeira-Sta. Bárbara d'Oeste.	Limeira
5	PCAB02192	Rio Piracicaba	Ponte a 50 m do Km 135,3 da estrada Piracicaba-Limeira, próx. à Usina Monte Alegre.	Piracicaba
5	PCAB02220	Rio Piracicaba	Margem esquerda, 2,5 Km a jusante da foz do Rib. Piracica-Mirim, na captação de Piracicaba.	Piracicaba
5	PCAB02800	Rio Piracicaba	Em frente à fonte sulfurosa, junto ao posto 4D-07 do DAEE, na localidade de Artemis.	Piracicaba
6	BILL02100	Reserv. Billings	No meio do corpo central, na direção do braço do Bororé.	São Paulo
6	BILL02500	Reserv. Billings	No meio do corpo central, sob a ponte da rodovia dos Imigrantes.	São Bernardo do Campo
6	BILL02900	Reserv. Billings	Próximo à barragem reguladora Billings-Pedras (Summit Control).	São Bernardo do Campo
6	RGDE02200	Reserv. do Rio Grande	No Clube Prainha Tahiti Camping Náutica, na altura do Km 42 da rodovia SP-31.	Ribeirão Pires
6	RGDE02900	Reserv. do Rio Grande	Próximo à rodovia Anchieta, junto à captação da SABESP.	São Bernardo do Campo

UGRHI	Ponto	Corpo Hídrico	Localização	Município
6	TIET02050	Rio Tietê	Ponte na rodovia que liga Mogi das Cruzes a Salesópolis (SP-88).	Biritiba Mirim
6	TIET02090	Rio Tietê	Na captação principal do município de Mogi das Cruzes.	Mogi Das Cruzes
10	SOIT02100	Reserv. Itupararanga	No meio do corpo central, lado esquerdo da Praia do Escritório, em frente a uma ilha.	Ibiúna
10	SOIT02900	Reserv. Itupararanga	Próximo a barragem, na estrada que liga Ibiúna a Votorantim.	Votorantim
10	SORO02100	Rio Sorocaba	Ponte Pinga-Pinga, na Av. Marginal, na cidade de Sorocaba.	Sorocaba
10	SORO02500	Rio Sorocaba	Ponte no Bairro de Americana Velha, em Tatuí	Tatuí
10	SORO02700	Rio Sorocaba	Na ponte à montante da captação do Município de Cerquilha.	Cerquilha
10	SORO02900	Rio Sorocaba	Ponte na estrada que liga Laranjal Paulista à localidade de Entre Rios.	Laranjal Paulista
10	TIBB02100	Reserv. de Barra Bonita	No meio do corpo central, a jusante da confluência Braços Tietê e Piracicaba.	Botucatu
10	TIBB02700	Reserv. de Barra Bonita	No meio do corpo central, na direção do Córrego Araquazinho.	São Manuel
10	TIET02350	Rio Tietê	A cerca de 300 m da ponte da Rodovia do Açúcar (SP-308), na Fazenda Santa Isabel.	Salto
10	TIET02400	Rio Tietê	Ponte na rodovia SP-113, que liga Tietê a Capivari, em Tietê.	Tietê
10	TIET02450	Rio Tietê	Ponte na estrada para a fazenda Santo Olegário, em Laranjal Paulista.	Laranjal Paulista

Critérios específicos para seleção dos parâmetros de qualidade:

1. Parâmetros considerados com maior potencial de trazer à tona informações relevantes.

Foram divididos em quatro categorias: relacionados à saúde humana, à vida aquática, a fatores organolépticos e indicadores genéricos.

2. Parâmetros que constam em pelo menos 80% dos pontos de amostragem.

A aplicação destes critérios específicos resultaram na seleção de 25 parâmetros de qualidade. Contudo, devido a escassez de dados, os parâmetros Cromo Total, Sódio, Potássio e Absorbância no UV foram descartados, resultando em um conjunto final de 21 parâmetros, conforme apresentado na Tabela 7.

Tabela 7: Parâmetros de qualidade considerados após aplicação dos critérios específicos.

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	

4.1.2 Pré-processamento dos Dados

Nesse estudo, a etapa de pré-processamento compreendeu atividades para conversão, centralização, imputação, transformação, discretização e normalização dos dados. A Figura 10 mostra o fluxo de tarefas realizadas durante o pré-processamento dos dados brutos e, em seguida, é apresentado o detalhamento de cada tarefa.

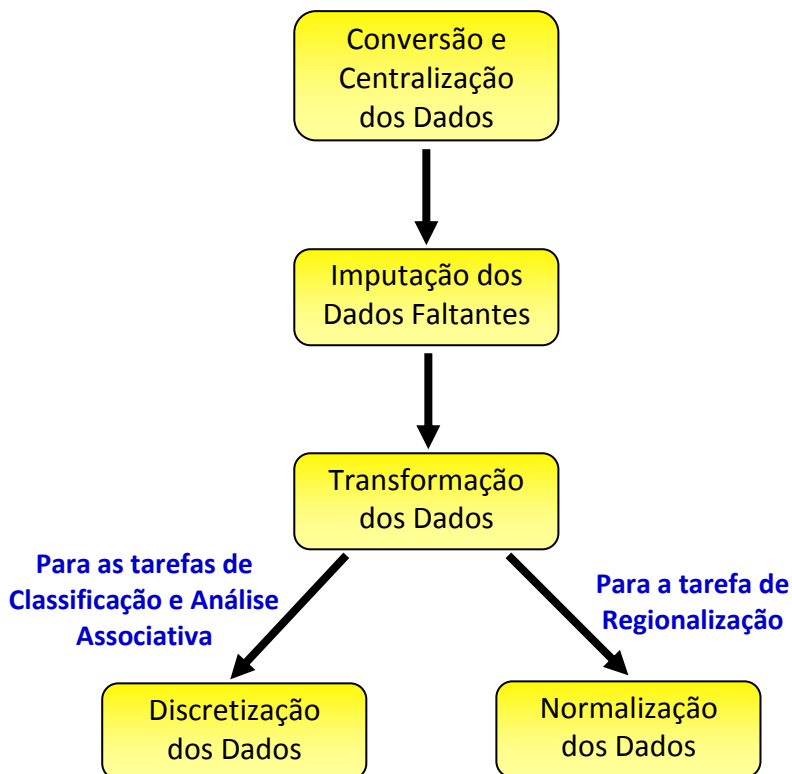


Figura 10: Fluxo de tarefas do pré-processamento de dados brutos.

Conversão e Centralização dos Dados

Depois de selecionados, os dados brutos foram centralizados em um repositório criado por meio do 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 foi realizada em várias etapas e consumiu a maior parte do tempo da etapa de pré-processamento, uma vez que os arquivos originais tinham pequenas diferenças entre si, demandando tratamentos específicos para que estas não impactassem na exatidão e na confiabilidade dos dados recuperados. A Figura 11 ilustra o processo de conversão dos dados originais até o armazenamento no banco de dados.

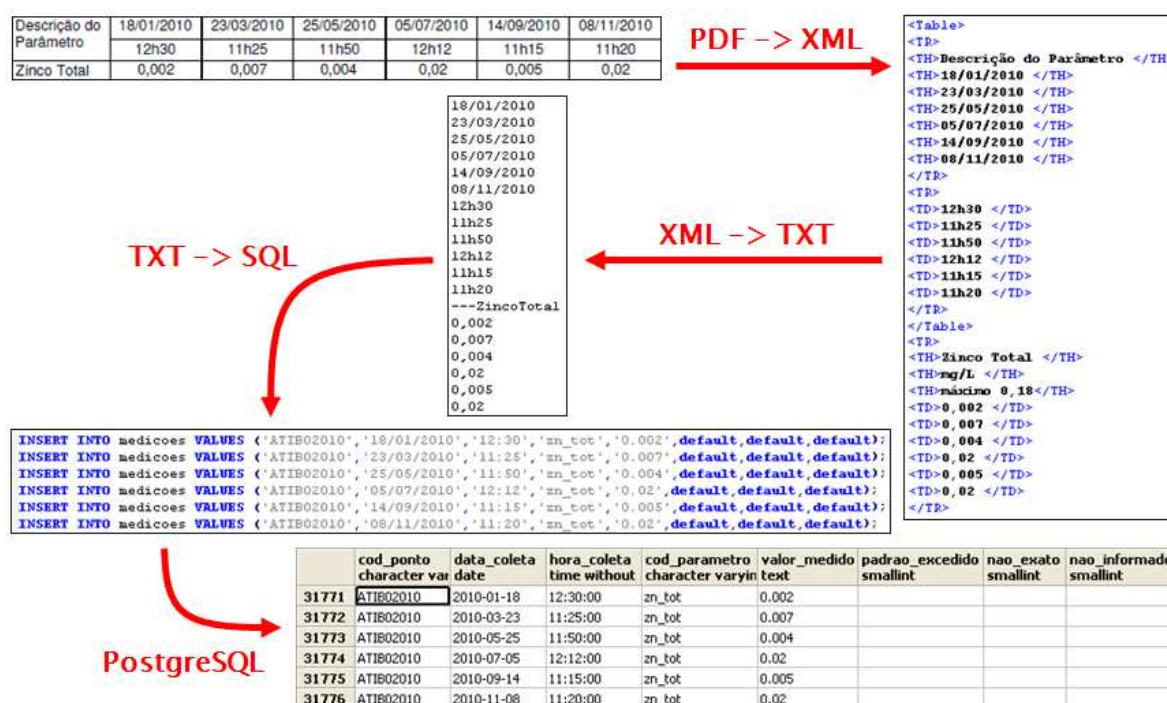


Figura 11: Esquema de conversão dos dados brutos.

Primeiramente, os arquivos PDF foram convertidos para o formato XML (*eXtensible Markup Language*) com o auxílio da própria ferramenta Adobe© Acrobat. Em seguida, por meio de dois conversores, implementados durante a pesquisa especificamente para esta finalidade, foram efetuadas as conversões de XML para o formato texto (TXT), e deste para o formato SQL (*Structured Query Language*). Por fim, os comandos SQL gerados foram executados, permitindo a inserção dos dados no gerenciador PostgreSQL.

A Figura 12 apresenta o projeto lógico da base de dados criada para armazenamento dos dados pré-processados. Nesse diagrama são representados os relacionamentos entre as cinco entidades básicas envolvidas: UGRHIs, pontos de amostragem das UGRHIs, tipos de parâmetro de qualidade, parâmetros de qualidade e, por fim, as medições realizadas em cada parâmetro de cada ponto de amostragem em uma determinada data.

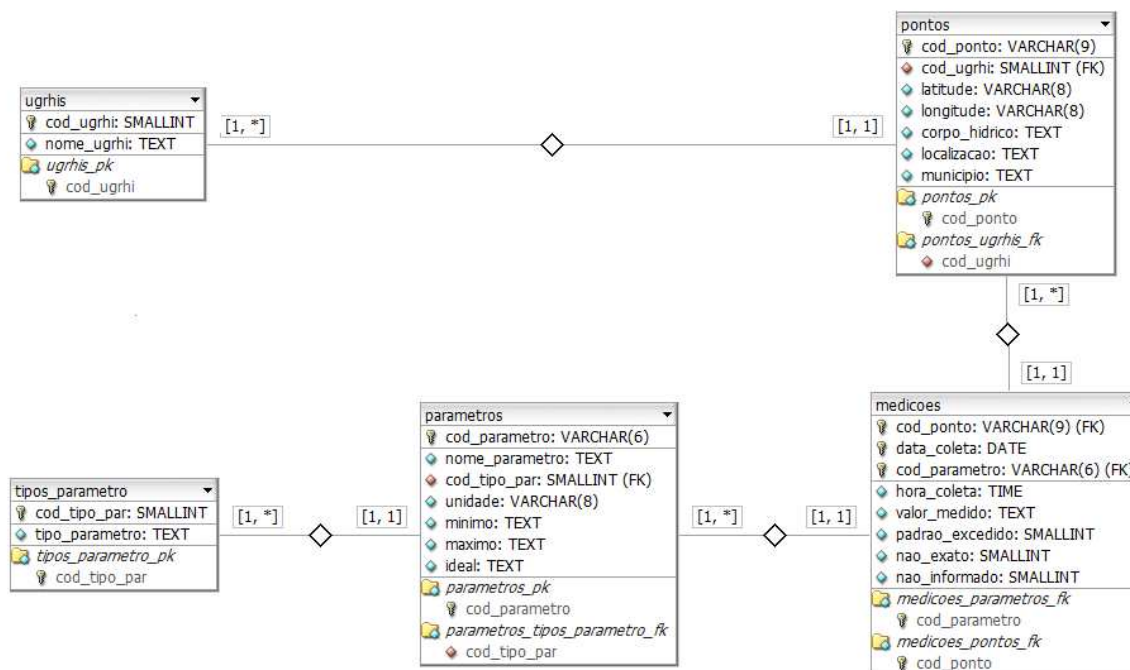


Figura 12: Diagrama Entidade-Relacionamento do banco de dados de medições.

Além destas entidades principais, também foram criadas tabelas auxiliares para armazenamento das configurações das tarefas de pré-processamento disponíveis na ferramenta desenvolvida: imputação, discretização, seleção de valores dos parâmetros (categorias) e seleção dos parâmetros a serem considerados durante a mineração dos dados. A Figura 13 mostra a estrutura destas tabelas.

imputacao parametro: VARCHAR(6) janeiro: VARCHAR(6) fevereiro: VARCHAR(6) marco: VARCHAR(6) abril: VARCHAR(6) maio: VARCHAR(6) junho: VARCHAR(6) julho: VARCHAR(6) agosto: VARCHAR(6) setembro: VARCHAR(6) outubro: VARCHAR(6) novembro: VARCHAR(6) dezembro: VARCHAR(6)	discretizacao parametro: VARCHAR(6) limite1: VARCHAR(6) valor_disc1: VARCHAR(2) limite2a: VARCHAR(6) limite2b: VARCHAR(6) valor_disc2: VARCHAR(2) limite3: VARCHAR(6) valor_disc3: VARCHAR(2)	selecao_categorias parametro: VARCHAR(6) categoria: VARCHAR(2) selecao: VARCHAR(3)	selecao_parametros parametro: VARCHAR(6) selecao: VARCHAR(3)
--	--	--	---

Figura 13: Estrutura das tabelas de pré-processamento.

A tabela Medições foi concebida visando facilitar a entrada dos dados oriundos dos arquivos PDF disponibilizados pela CETESB, porém sua estrutura não era adequada para a fase de mineração dos dados. Por este motivo, também foi necessário criar tabelas específicas para cada uma das três tarefas de mineração desenvolvidas nessa pesquisa: classificação (tabelas de treinamento e teste para as UGRHIs 2, 5, 6 e 10), análise de associação e análise de grupos. A Figura 14 apresenta a estrutura destas tabelas.

<div>dataset_training_2_5</div> <div><div></div>cod_ponto: VARCHAR(9)</div> <div><div></div>data_coleta: DATE</div> <div><div></div>cd_tot: TEXT</div> <div><div></div>pb_tot: TEXT</div> <div><div></div>ni_tot: TEXT</div> <div><div></div>ox_dis: TEXT</div> <div><div></div>toxidd: TEXT</div>	<div>dataset_training_6_10</div> <div><div></div>cod_ponto: VARCHAR(9)</div> <div><div></div>data_coleta: DATE</div> <div><div></div>cd_tot: TEXT</div> <div><div></div>pb_tot: TEXT</div> <div><div></div>ni_tot: TEXT</div> <div><div></div>ox_dis: TEXT</div> <div><div></div>toxidd: TEXT</div>	<div>dataset_test_2_5</div> <div><div></div>cod_ponto: VARCHAR(9)</div> <div><div></div>data_coleta: DATE</div> <div><div></div>cd_tot: TEXT</div> <div><div></div>pb_tot: TEXT</div> <div><div></div>ni_tot: TEXT</div> <div><div></div>ox_dis: TEXT</div> <div><div></div>toxidd: TEXT</div> <div><div></div>toxidd_prev: TEXT</div>	<div>dataset_test_6_10</div> <div><div></div>cod_ponto: VARCHAR(9)</div> <div><div></div>data_coleta: DATE</div> <div><div></div>cd_tot: TEXT</div> <div><div></div>pb_tot: TEXT</div> <div><div></div>ni_tot: TEXT</div> <div><div></div>ox_dis: TEXT</div> <div><div></div>toxidd: TEXT</div> <div><div></div>toxidd_prev: TEXT</div>
<div>dataset_association</div> <div><div></div>cod_ponto: VARCHAR(9)</div> <div><div></div>data_coleta: DATE</div> <div><div></div>cd_tot: TEXT</div> <div><div></div>pb_tot: TEXT</div> <div><div></div>ni_tot: TEXT</div> <div><div></div>ox_dis: TEXT</div> <div><div></div>toxidd: TEXT</div>	<div>dataset_clusterization</div> <div><div></div>cod_ponto: VARCHAR(9)</div> <div><div></div>data_coleta: DATE</div> <div><div></div>cd_tot: TEXT</div> <div><div></div>pb_tot: TEXT</div> <div><div></div>ni_tot: TEXT</div> <div><div></div>ox_dis: TEXT</div> <div><div></div>toxidd: TEXT</div>		

Figura 14: Estrutura das tabelas para mineração dos dados.


Estas entidades são criadas dinamicamente, a partir da tabela básica Medições, durante as configurações de pré-processamento disponíveis na ferramenta desenvolvida. Basicamente, o que se faz nesse processo é transformar os códigos dos parâmetros, cadastrados na coluna

“cod_parametro” da tabela Medições, em colunas nas novas tabelas a serem utilizadas para mineração. Como uma das configurações é a seleção dos parâmetros que serão considerados pela mineração de dados, as estruturas destas tabelas variam de acordo com os parâmetros escolhidos. No exemplo da Figura 12, pode-se observar que durante o pré-processamento foram selecionados apenas cinco parâmetros de qualidade: Cádmio Total (cd_tot), Chumbo Total (pb_tot), Níquel Total (ni_tot), Oxigênio Dissolvido (ox_dis) e Toxicidade (toxidd).

Após a conversão e o armazenamento dos dados em uma base de dados centralizada, foi aplicado um critério específico para seleção dos conjuntos de medição. Vale destacar que o termo “conjunto de medição” refere-se a cada grupo medições de n parâmetros, o qual está associado a um ponto de amostragem e uma data de coleta específicos.

Por se tratar de um parâmetro essencial no contexto deste trabalho, todos os conjuntos de medição que não continham medição de Toxicidade foram descartados, conforme mostra o exemplo da Figura 15. Este último critério eliminou 12.054 das 38.913 medições selecionadas, uma redução de aproximadamente 30% dos conjuntos de medição selecionados até então, restando ao final 26.859 medições.

Conjunto de medição eliminado



Descrição do Parâmetro	Unidade	Padrão CONAMA	07/01/2008	04/03/2008	12/05/2008	22/07/2008	22/09/2008	10/11/2008
			12h50	13h10	13h45	13h20	13h10	13h40
Chuva 24h	-		Sim	Sim	Não	Não	Não	Sim
pH	U.pH	entre 6 e 9	6,3	6,3	6	* 5,9	6,8	6,6
Temp. Água	°C		25	23	18	17	19	25
Alumínio Dissolvid	mg/L	máximo 0,1	* 0,36	* 0,5	* 0,2	0,06	0,1	0,1
Cádmio Total	mg/L	máximo 0,001	< 0,0001	< 0,001	< 0,001	< 0,001	< 0,001	< 0,001
Chumbo Total	mg/L	máximo 0,01	0,004	< 0,01	< 0,01	< 0,01	< 0,01	< 0,01
Cloreto Total	mg/L	máximo 250	4	7	6	7	5	8
Cobre Dissolvido	mg/L	máximo 0,009	< 0,009	0,002	0,001	0,002	< 0,001	0,002
Condutividade	µS/cm		81	91	89	104	75	89
Ferro Dissolvido	mg/L	máximo 0,3	* 0,58	* 0,4	0,3	* 0,5	* 0,4	* 0,7
Manganês Total	mg/L	máximo 0,1	* 0,18	0,1	0,09	0,06	0,1	* 0,2
N. Amoniacal	mg/L	máximo 3,7	0,2	0,7	1	2	1	1
Níquel Total	mg/L	máximo 0,025	< 0,02	< 0,01	< 0,01	< 0,01	< 0,01	< 0,01
Nitrato	mg/L	máximo 10	0,3	6	7,8	0,8	0,7	0,6
Nitrito	mg/L	máximo 1	0,03	0,02	0,02	0,01	0,03	0,02
OD	mg/L	mínimo 5	* 1,8	* 2,1	* 3,1	* 1,1	* 2,6	* 1,2
Sol. Total	mg/L		233	202	175	147	153	800
Subst. Tensoat.	mg/L	máximo 0,5	0,2	0,06	0,09	0,2	0,1	0,2
Turbidez	UNT	máximo 100	* 135	30	17	12	29	16
Zinco Total	mg/L	máximo 0,18	0,03	0,01	0,02	0,03	0,02	0,01
Toxicidade	-	Não Tóxico	Não Tóxico	Não Tóxico		Não Tóxico	Não Tóxico	* Crônico

Figura 15: Exemplo de conjunto de medição eliminado.

Imputação de Dados Faltantes

A ausência de valores para determinados parâmetros, ou a inexatidão destes, pode causar interferências na mineração de dados e, conseqüentemente, gerar resultados distorcidos. A solução mais radical para estes casos é a remoção do registro completo, mesmo que este possua somente um dos atributos com valor faltante. Como se pôde observar no exemplo da Figura 13, esse método foi empregado em conjuntos de medição onde não havia valor de Toxicidade, dada a importância deste parâmetro e a dificuldade em se estimar um valor confiável a este – nos dados da CETESB, a Toxicidade é um parâmetro discreto com apenas três categorias: “Não Tóxico”, “Crônico” e “Agudo”, tornando a atribuição de valores bastante vulnerável a erros. Vale lembrar que, além da Toxicidade, o parâmetro “Chuva 24 horas” também é um parâmetro discreto, porém este não possuía dados faltantes.

Como os 19 parâmetros restantes são todos contínuos, para não reduzir ainda mais a quantidade de conjuntos válidos, foi empregado um método para atribuição de valores, exceto para o parâmetro Cloreto Total que também não possuía dados faltantes. Os critérios adotados nesse método foram estabelecidos de forma empírica, visando o mínimo impacto sobre o conjunto de dados. Para isso, foram considerados dois critérios:

Critério 1 – Em medições abaixo do padrão da resolução CONAMA 357/2005 (CONAMA, 2005), porém sem valor exato conhecido, foi imputado o próprio valor medido. Constatou-se que aproximadamente 21% das medições encontravam-se nessa situação. Exemplo:

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

Critério 2 – 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). Aproximadamente 8% das medições estavam em uma dessas situações. Exemplos:

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

Deve-se destacar que a utilização do valor médio mensal nas situações do Critério 2 é o padrão adotado pela ferramenta. Porém, esta foi implementada de modo a permitir a customização dos valores a serem imputados para essas situações, conforme será descrito na Seção 4.5.1.

Transformação dos Dados

Visando evitar conversões durante o processo de mineração de dados 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 6 caracteres, conforme apresentado na Tabela 8.

Tabela 8: Transformação dos identificadores dos parâmetros de qualidade.

Parâmetro	Código	Parâmetro	Código	Parâmetro	Código
Alumínio Dissolvido	al_dis	Ferro Dissolvido	fe_dis	pH	pot_hi
Cádmio Total	cd_tot	Manganês Total	mn_tot	Sólidos Totais	soltot
Chumbo Total	pb_tot	Níquel Total	ni_tot	Substância Tensoativa	sub_te
Chuva 24h	chuvas	Nitrato	nitrat	Temperatura Água	tmp_ag
Cloreto Total	cloret	Nitrito	nitrit	Toxicidade	toxidd
Cobre Dissolvido	cu_dis	Nitrogênio Amoniacal	n_amon	Turbidez	turbid
Condutividade	condut	Oxigênio Dissolvido	ox_dis	Zinco Total	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. Essa padronização é mais bem detalhada na descrição da discretização a seguir.

Discretização dos Dados

Normalmente, os algoritmos de classificação e análise associativa requerem que os atributos contínuos sejam categorizados por meio de valores discretos, processo denominado **discretização**. A conversão de um atributo contínuo em discreto envolve duas tarefas: definir quantas categorias devem existir e como será feito o mapeamento dos valores contínuos para os valores discretos.

Nessa pesquisa, a discretização padrão dos dados de monitoramento de qualidade água foi realizada de forma empírica, por meio da inspeção visual dos dados. Essa abordagem segundo

Tan et al. (2009) pode ser eficaz em determinadas situações. A Tabela 9 mostra como os parâmetros contínuos foram discretizados, bem como os mnemônicos utilizados para identificação dos valores na base de dados. Esta Tabela também mostra os mnemônicos adotados para os parâmetros que já são discretos nos próprios dados brutos, Chuva 24h e Toxicidade.

Vale destacar que essa forma de discretização é o padrão adotado pela ferramenta desenvolvida para a pesquisa, porém esta foi implementada de modo a permitir a customização das faixas a serem categorizadas e de seus respectivos mnemônicos, conforme descrito na Seção 4.5.1.

Tabela 9: Categorização dos parâmetros contínuos e discretos.

Parâmetros Contínuos	Mnemônico (Valor Discretizado)	Descrição
pH	AB	Abaixo – Abaixo do limite inferior do Padrão CONAMA.
	PC	Padrão CONAMA – Dentro do Padrão CONAMA.
	AC	Acima – Acima do limite superior do Padrão CONAMA.
Temperatura Água ¹ , Condutividade ¹ , Sólidos Totais ¹	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 CONAMA.
	AB	Abaixo – Abaixo do Padrão CONAMA em até 60%.
	MA	Muito Abaixo – Abaixo do Padrão CONAMA 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 CONAMA.
	AC	Acima – Acima do Padrão CONAMA em até 3x.
	MA	Muito Acima – Acima do Padrão CONAMA mais que 3x.
Parâmetros Discretos	Mnemônico	Descrição
Chuva 24h ¹	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 <i>Ceriodaphnia Dubia</i> .
	CR	Crônico – Resposta fisiológica do microcrustáceo <i>Ceriodaphnia Dubia</i> .
	AG	Agudo – Forte resposta fisiológica do microcrustáceo <i>Ceriodaphnia Dubia</i> .

¹ Parâmetros de acompanhamento, sem valores de Padrão CONAMA.

Normalização dos Dados

Ao contrário dos algoritmos utilizados nas tarefas de classificação e análise associativa, o algoritmo de regionalização empregado neste estudo não possibilita o uso de atributos discretos, devido às diversas operações matemáticas necessárias, como o cálculo da distância euclidiana e cálculos de dispersão. Por este motivo, a discretização mencionada anteriormente não foi aplicada aos dados a serem processados pelo algoritmo de regionalização. Por outro lado, foi necessário efetuar a normalização dos dados, visto que os valores dos parâmetros considerados possuem unidades de medida diferentes e, conseqüentemente, amplitudes extremamente distintas. Para este fim, foi utilizada uma técnica de normalização segundo a amplitude, onde todos os parâmetros assumiram valores entre 0 e 1. Tais valores foram obtidos por meio da seguinte equação:

$$y = \frac{x - \min}{\max - \min}$$

Sendo: y , valor normalizado do parâmetro;
 x , valor não normalizado do parâmetro;
 \min , valor mínimo do parâmetro;
 \max , valor máximo do parâmetro;

4.1.3 Mineração dos Dados

As abordagens de mineração de dados utilizadas nesta pesquisa foram escolhidas a partir de uma pesquisa bibliográfica visando levantar métodos já utilizados na área ambiental. Conforme apresentado em Bertholdo et al. (2012), para o problema da predição de ecotoxicidade em amostras de água, foi aplicada uma das abordagens centrais da mineração de dados, a **modelagem previsiva**, que busca construir um modelo para prever o valor de um dado atributo com base nos valores de outros atributos do conjunto de dados. Esta modelagem foi realizada por meio da técnica de classificação baseada em regras, na qual os registros de da base de dados de teste são classificados a partir das regras obtidas a partir da mineração dos registros de uma base de dados de treinamento. No âmbito dos dados de qualidade de água, cada registro da base de dados de monitoramento é representado pela análise de uma amostra de água coletada de um dado ponto de um corpo hídrico, em uma data específica, na qual são medidos diversos

parâmetros de qualidade. Nesse contexto, o objetivo da técnica é descobrir regras que possam, com base nos valores destes parâmetros, definir o nível de toxicidade de cada amostra de água.

O problema dos relacionamentos entre os parâmetros de qualidade de água foi tratado por meio de uma metodologia conhecida como **análise de associação**, a qual varre o conjunto de dados em busca de correlações fortes entre as variáveis pesquisadas. Para encontrar essas conexões, o algoritmo utilizado considerou a frequência na qual cada possível regra de associação se aplica ao conjunto de dados, e também a confiabilidade denotada por estas regras. No contexto desta pesquisa, esta técnica foi empregada com o objetivo de descobrir regras que apontem conexões interessantes entre os parâmetros de qualidade de água medidos.

Finalmente, para a questão do agrupamento dos pontos de amostragem de água, aplicou-se um método para **análise de grupos**. Por envolver o mapeamento de regiões geográficas, foi utilizado um algoritmo que promove a regionalização do espaço por meio da eliminação das arestas mais custosas entre os vértices de um grafo. No domínio tratado nesta pesquisa, os vértices deste grafo são representados pelos pontos de amostragem e as arestas que conectam estes vértices representam os corpos d'água que interligam os pontos de amostragem. Nesse cenário, as arestas mais custosas expõem os relacionamentos mais fracos entre os pontos de amostragem. O objetivo dessa abordagem foi revelar as dissimilaridades mais expressivas entre os corpos hídricos e seus diferentes trechos.

4.1.4 Visualização dos Dados

Muitas vezes o conhecimento resultante da aplicação da mineração de dados pode não ser facilmente interpretados pela simples observação das regras e padrões encontrados. Segundo Tan et al. (2009), em alguns casos, a análise dos dados pode ser executada usando-se ferramentas não visuais e depois os resultados serem apresentados visualmente para avaliação pelo especialista do domínio.

Por este motivo, neste trabalho foram utilizadas algumas técnicas visuais, para que os resultados obtidos pela mineração de dados pudessem ser apresentados de forma sintética, auxiliando na sua interpretação e avaliação. Para o problema de classificação de amostras de

água, foram utilizados gráficos de linhas e de pizza; para a análise associativa dos parâmetros de qualidade, um gráfico de barras e uma tela de filtro para visualização das regras mais interessantes; por fim, para a regionalização dos pontos de amostragem, usou-se um grafo que conecta estes pontos sobre o mapa das quatro UGRHs contempladas.

4.1.5 Interpretação e Avaliação dos Padrões

A partir da visualização das informações retornadas pela mineração de dados, pode-se avaliar e interpretar a significância do conhecimento descoberto em meio aos dados de monitoramento de qualidade de água. Nesse estágio, foram analisadas: as **regras de classificação** geradas, explicitando quais combinações “parâmetro-valor” estão mais associadas à toxicidade da água; as **regras de associação**, demonstrando quais e como os parâmetros estão associados entre si; e os **agrupamentos de pontos de amostragem**, evidenciando as regiões mais homogêneas quanto aos parâmetros considerados. Essas análises foram realizadas com o apoio de uma especialista da área de saneamento ambiental, a qual também avaliou a importância do conhecimento obtido.

4.2 Método para Classificação de Toxicidade em Amostras de Água

Atualmente, a toxicidade de uma amostra de água é mensurada por meio de testes ecotoxicológicos, que consistem na determinação de efeitos tóxicos em organismos aquáticos causados por um ou mais agentes químicos. Os efeitos tóxicos agudos caracterizam-se por serem mais drásticos, causados por elevadas concentrações de agentes químicos e, em geral, manifestam-se em um curto período de exposição dos organismos. Os efeitos tóxicos crônicos são mais sutis, causados por baixas concentrações de agentes químicos dissolvidos e são detectados em prolongados períodos de exposição ou por respostas fisiológicas adversas na reprodução e crescimento dos organismos vivos (CETESB, 2011).

A primeira frente desta pesquisa teve como objetivo descobrir padrões de classificação de ecotoxicidade a partir das medições de outros parâmetros de qualidade. Uma vez descobertos,

esses padrões podem ser utilizados na predição da toxicidade de futuras amostras de água, minimizando a utilização de organismos vivos nas análises ecotoxicológicas, tornando estas análises mais rápidas e eficazes, ou então descobrir que o conjunto de parâmetros/valores adotados são insuficientes para efetuar essa predição, indicando a necessidade de análises adicionais ou alteração dos padrões vigentes.

Para atingir esse objetivo, foi utilizada a técnica de classificação baseada em regras, uma abordagem da mineração de dados que busca construir um modelo, a partir de um conjunto de registros previamente rotulados, capaz de classificar os registros de outros conjuntos ainda não rotulados. No contexto desta pesquisa, o algoritmo aprende um conjunto de regras condicionais a partir da base de dados de treinamento, onde o atributo alvo a ser previsto é a toxicidade. Para avaliar a qualidade de cada possível regra de classificação são utilizadas as medidas de **cobertura** e **precisão**, já explicadas na Seção 3.2.1. Em seguida, as regras aprendidas pelo algoritmo são aplicadas à base de testes, de modo a atribuir um valor de toxicidade à cada conjunto de medição desta base. Um exemplo de regra gerada seria:

Se Níquel=AC e Chumbo=AC e Zinco=MA Então Toxicidade=CR

Nessa pesquisa, o modelo de classificação foi gerado pelo algoritmo de cobertura sequencial apresentado em Tan et al. (2009). Este algoritmo faz uma busca pelas melhores regras para prever cada classe, no caso os valores de Toxicidade: Não Tóxico, Crônico e Agudo. Dentre as diversas abordagens aplicadas nesse domínio, esta técnica foi considerada uma das mais apropriadas para a tarefa de classificação, pois permite extrair regras diretamente dos dados, ao contrário de outros métodos que extraem regras indiretamente, a partir de outros modelos como árvores de decisão e redes neurais. Durante a busca das regras, todos os conjuntos de medição com classe igual a que está sendo pesquisada são considerados **positivos**, e todos os outros conjuntos são considerados **negativos**. Uma regra é considerada satisfatória se cobrir a maioria dos conjuntos positivos e poucos negativos. A Figura 16 mostra de forma resumida o funcionamento deste algoritmo.

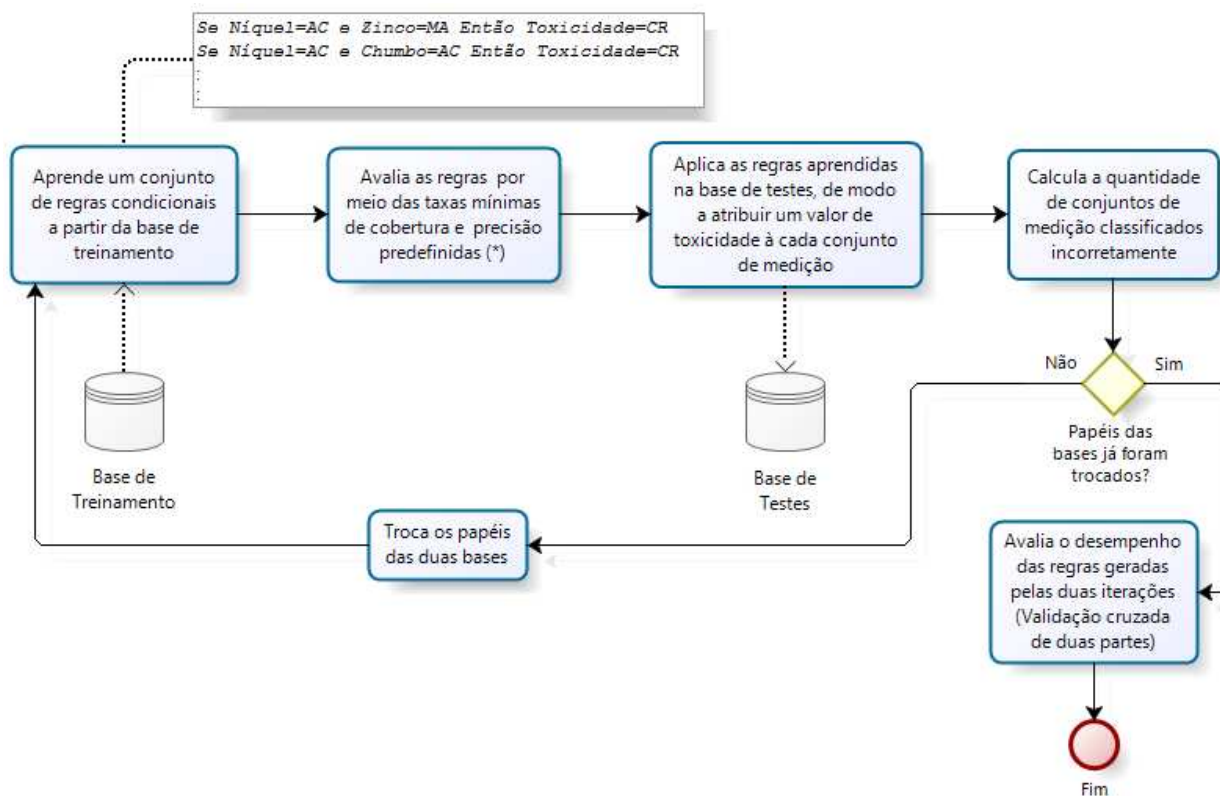


Figura 16: Algoritmo de Cobertura Sequencial.

Por fim, o desempenho geral das regras de classificação geradas foi avaliado pelo método da validação cruzada de duas partes. Nessa abordagem, a base de dados é dividida em dois subconjuntos com quantidades de registros semelhantes. Em um primeiro momento, um dos subconjuntos é utilizado como base de treinamento, ou seja, as regras de classificação são extraídas a partir deste subconjunto. Em seguida, as regras extraídas são aplicadas ao outro subconjunto, que faz o papel de base de teste. Por fim, é calculada a taxa de erro das regras aplicadas nesta base de teste. No segundo momento, os papéis são invertidos, de modo que o subconjunto de treinamento passa a ser de teste e vice-versa. A taxa de erro total é então calculada pela média das duas execuções. Neste trabalho, a divisão da base de dados foi realizada de modo a gerar um subconjunto contendo à medições das UGRHIs 2 e 5 e outro subconjunto contendo as medições das UGRHIs 6 e 10, visto que esta divisão proporcionou uma quantidade semelhante de registros nos dois subconjuntos gerados.

4.3 Método para Identificação de Associações entre os Parâmetros de Qualidade

De forma geral, os relacionamentos entre os parâmetros de qualidade de água são pouco conhecidos. Existem várias questões intrigantes e obscuras, difíceis de serem esclarecidas, devido à complexidade inerente ao enorme volume de dados gerados pelas medições dos parâmetros nas amostras de água. Alguns exemplos: Qual o impacto dos metais na condutividade? Como as chuvas interferem no oxigênio dissolvido? Os sólidos totais influenciam na turbidez? O pH tem alguma relação com a condutividade? Em que medida os metais afetam a toxicidade?

A segunda questão tratada neste trabalho teve como meta encontrar possíveis relacionamentos entre os parâmetros de qualidade de água. Uma vez conhecidas, essas correlações podem proporcionar uma compreensão mais profunda de como os parâmetros de qualidade interagem entre si, validar associações preconcebidas empiricamente, além de gerar subsídios interessantes para a tomada de decisões estratégicas com relação ao gerenciamento dos corpos hídricos.

No contexto deste trabalho, esses relacionamentos são expressos por meio de regras de associação (ou implicação). Para viabilizar a extração destas regras do conjunto de dados de monitoramento de qualidade de água, foi utilizado o algoritmo Apriori apresentado em Tan et al. (2009), um dos mais difundidos para a geração de regras de associação. Assim como no mecanismo de cobertura sequencial, apresentado na Seção anterior, o resultado final deste algoritmo são regras condicionais compostas por um **antecedente** e um **consequente**. A diferença é que, ao contrário da abordagem anterior, não há um parâmetro alvo específico para o qual se deseja descobrir o valor, no caso a toxicidade. No enfoque da análise associativa, tanto o antecedente quanto o consequente da regra podem possuir de um a n parâmetros, desde que evidentemente não se repitam em ambos os lados da regra. Um exemplo de regra gerada seria:

Se Cádmio=AC e Ferro=MA Então Condutividade=AT e pH=AB

Durante o processo de geração das regras de associação, a qualidade de cada possível regra foi avaliada por meio das medidas de **suporte** e **confiança**, explicadas na Seção 3.2.2.

No algoritmo Apriori, 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 um limite de suporte mínimo considerado. No cenário desta pesquisa, os itens são representados por combinações “parâmetro-valor”, por exemplo: “Nitrato=PC”. 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 as chamadas **regras fortes**, que representam os relacionamentos mais significativos entre os **itens frequentes**. A Figura 17 apresenta de maneira sucinta o funcionamento deste algoritmo.

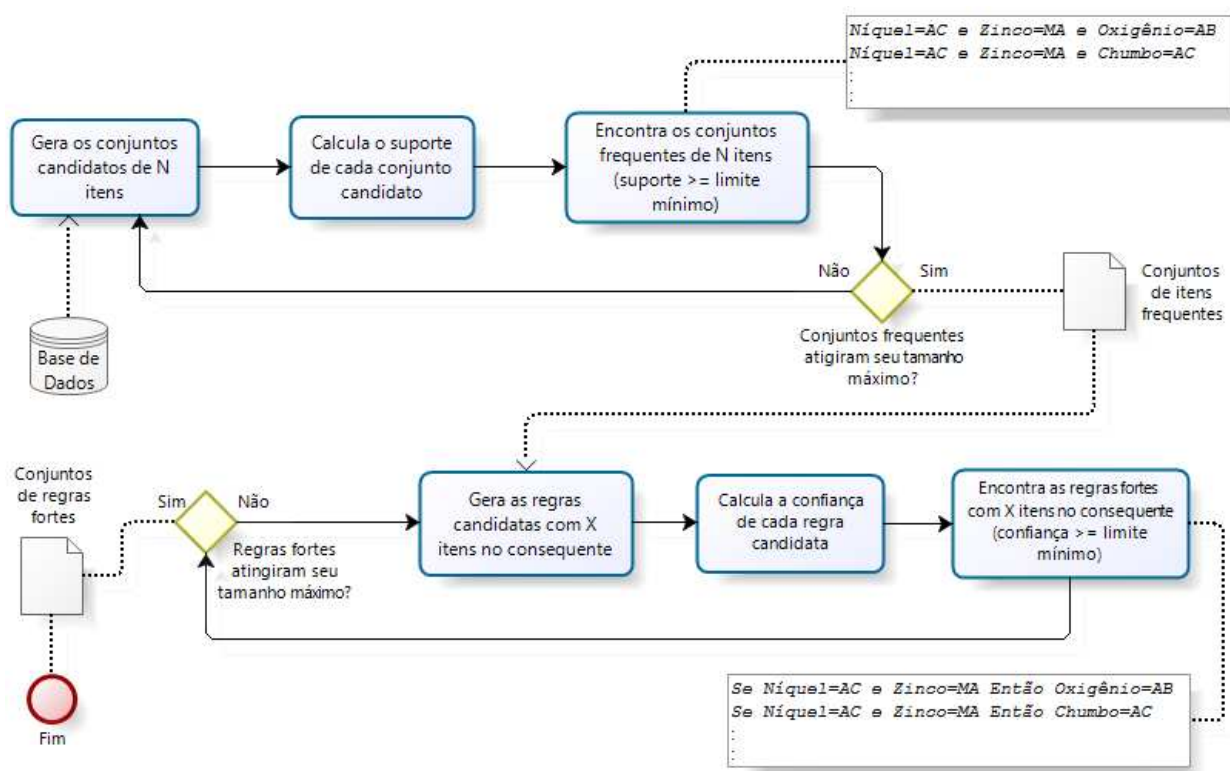


Figura 17: Algoritmo Apriori.

4.4 Método para Regionalização de Pontos de Amostragem de Água

O terceiro tema estudado neste trabalho visou a descoberta de grupos de pontos de amostragem similares em relação às medições de seus parâmetros de qualidade. A regionalização, ou separação, destes pontos em áreas homogêneas pode gerar inferências úteis a

respeito das 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 possibilitar uma melhor distribuição dos 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 trabalho, a tarefa de agrupamento foi realizada a partir do método apresentado em Assunção et. al. (2002), o qual se baseia na geração de uma árvore geradora mínima (AGM), seguida da poda desta árvore para propiciar a divisão dos objetos em regiões homogêneas. A geração da AGM tem como propósito preparar o grafo para a fase de poda, eliminando as arestas de maior custo, porém sem gerar grupos separados de arestas. Uma vez gerada a AGM, qualquer aresta removida provoca a criação de dois grupos. Por este motivo, a fase seguinte, de poda, consiste na remoção de arestas da AGM para formação dos grupos. Algumas referências na área de mineração de dados, como Tan et al. (2009), mencionam a utilização de técnicas de agrupamento hierárquico divisivo baseadas na geração de AGMs.

Na primeira fase deste método, é gerada uma árvore a partir do grafo correspondente ao conjunto de dados. Esta árvore, denominada árvore geradora mínima (AGM), 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, que faz a árvore crescer naturalmente a partir de um dado vértice. A cada estágio, uma nova aresta é adicionada à árvore e o algoritmo para somente quando todos os vértices forem visitados. O resultado é um grafo conexo de custo mínimo, onde não existem ciclos. 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 do cálculo da distância euclidiana entre os atributos i e k dos dois vértices da aresta, conforme sugere Assunção (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;

Na segunda etapa do método é realizada a **poda** da AGM, que assim como na primeira etapa também consiste em remover as arestas mais caras. Como a AGM é um grafo sem ciclos, qualquer aresta retirada provoca uma divisão na árvore, resultando em duas subárvores desconectadas. São escolhidas $k-1$ arestas, para obter k regiões.

Antes de prosseguir com a explanação desta etapa, é preciso ressaltar que o método de cálculo do custo das arestas utilizado na primeira etapa não é adequado para a poda da AGM, pois considera apenas a dissimilaridade local entre cada dupla de vértices adjacentes, não levando em conta a árvore como um todo. Como consequência, o algoritmo tenderá a remover arestas que gerarão regiões com baixo grau de homogeneidade interna.

Por este motivo, na fase de poda a forma de atribuir custos às arestas é alterada, de modo que o custo de cada aresta passa a variar em função dos desvios internos das duas subárvores geradas pela sua possível remoção – quanto menores forem estes desvios maior será o custo da aresta. Como o algoritmo remove sempre as arestas mais custosas, as regiões obtidas em cada poda possuirão sempre os menores desvios internos possíveis, propiciando assim a geração de regiões homogêneas com relação aos atributos considerados. O novo custo das arestas é dado por:

$$Custo da aresta l = SSD_T - SSD_l ,$$

Sendo:

i) SSD_T é soma dos quadrados dos desvios, associada à árvore T , dada por:

$$SSD_T = \sum_{j=1}^m \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2 ,$$

Sendo:

n , o número total de objetos (nós) em T ;

x_{ij} , o atributo j do objeto i ;

m , o número de atributos considerados na análise;

\bar{x}_j , o valor médio do atributo j .

ii) SSD_l é a soma das duas parcelas obtidas da soma dos quadrados dos desvios das duas subárvores, T_a e T_b , geradas pela retirada da aresta l da árvore T :

$$SSD_l = SSD_{T_a} + SSD_{T_b}$$

Para obter a soma dos quadrados dos desvios para as duas subárvores, são calculados os valores médios dos m atributos, tal como feito para o cálculo de SSD_T , porém, considerando-se apenas os atributos referentes aos objetos pertencentes a cada subárvore de T , T_a e T_b . Vale salientar que, a cada aresta podada, os custos das arestas remanescentes nas duas subárvores geradas são recalculados, e a aresta de maior custo entre as duas subárvores é removida. Esse processo é repetido até que um critério de parada seja atendido. No caso da ferramenta desenvolvida para esta pesquisa, tal critério é definido pelo número de grupos a serem gerados, o qual é informado pelo usuário. A Figura 18 mostra simplificada o funcionamento deste algoritmo.

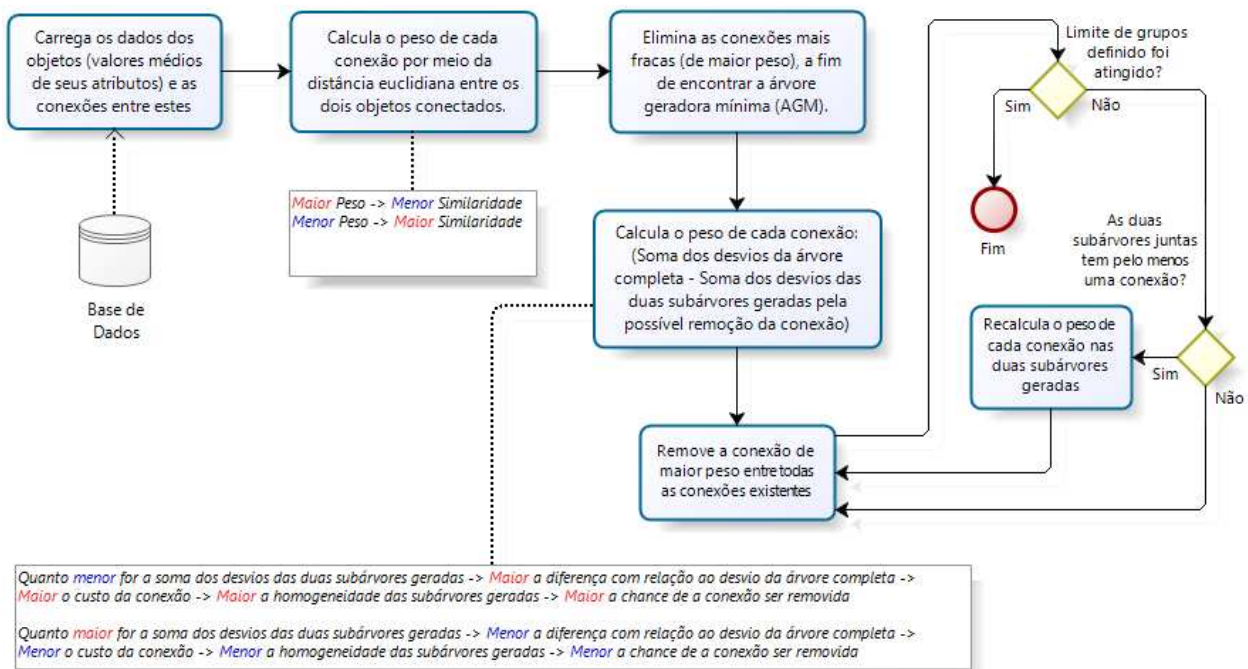


Figura 18: Algoritmo de Prim e Poda da AGM.

4.5 Ferramenta para Descoberta de Conhecimento em Dados de Monitoramento de Água

Esta Seção apresenta a ferramenta gráfica desenvolvida para a descoberta de padrões e regras em meio aos dados de monitoramento de qualidade de água. Sua interface principal é exibida na Figura 19. O menu **Preprocessing** traz todas as configurações de pré-processamento oferecidas pela ferramenta, e o menu **Knowledge Discovery** mostra as funcionalidades disponíveis para classificação, análise associativa e análise de grupos.

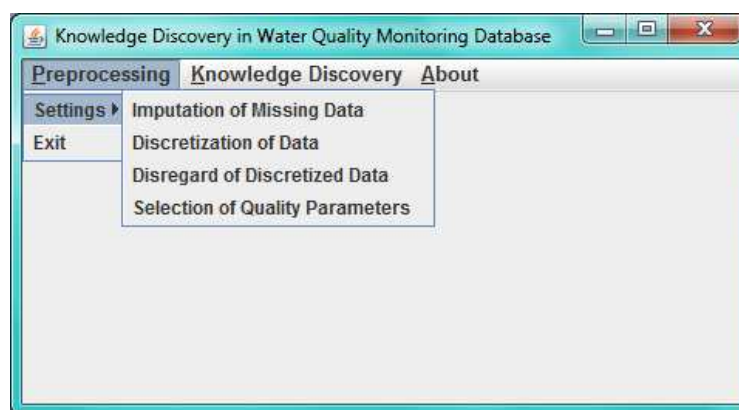
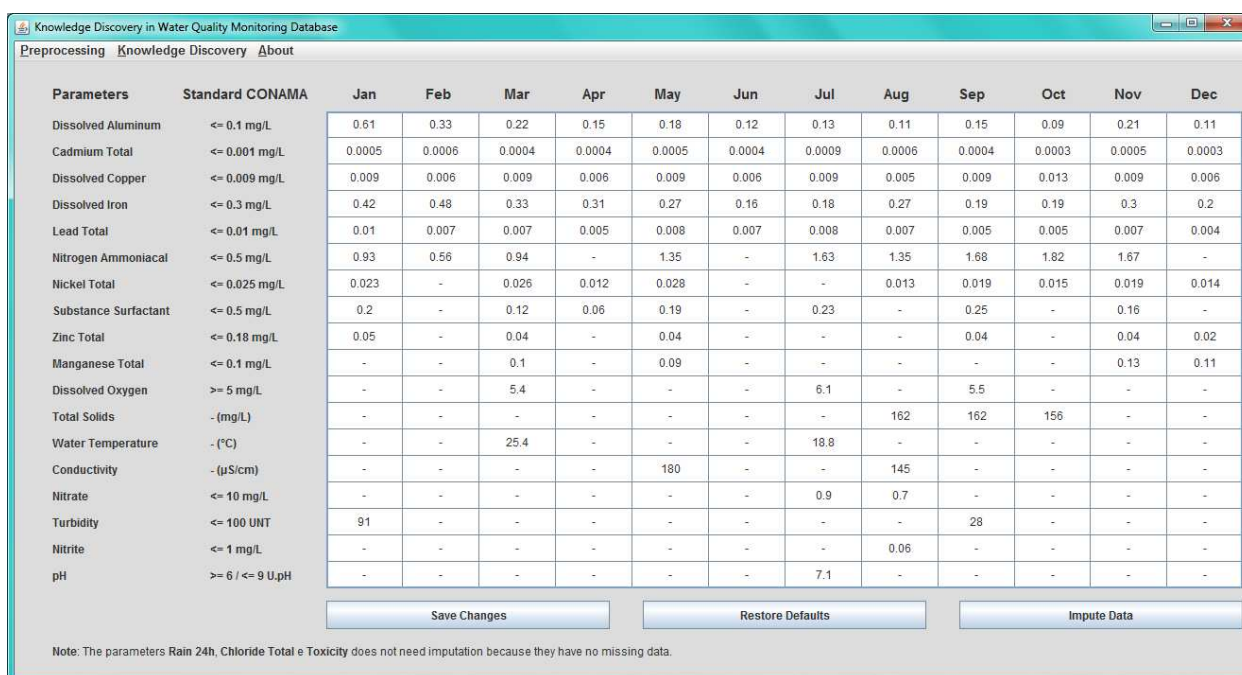


Figura 19: Interface principal da ferramenta desenvolvida.

4.5.1 Funcionalidades de Pré-processamento

As decisões a respeito das questões da imputação e da discretização impactam diretamente nos resultados obtidos pela mineração de dados. Existem inúmeros métodos, diferentes daqueles utilizados como padrão neste trabalho, que também podem ser empregados para esta mesma finalidade. Para a discretização, por exemplo, Tan et al. (2009) afirma que a melhor abordagem é aquela que produz o melhor resultado para a técnica de mineração de dados a ser utilizada. Por este motivo, notou-se a necessidade de flexibilizar a ferramenta, de modo a permitir que o usuário possa utilizar seus próprios métodos para definir quais valores devem ser imputados e de que forma os dados devem ser categorizados.

Para a questão da imputação em medições com valores faltantes, ou em medições onde não foi possível detectar se o valor estava abaixo ou acima do Padrão CONAMA, a ferramenta disponibiliza uma interface de configuração. Nela o usuário pode imputar os valores desejados, ao invés de assumir os valores padrão adotados pela ferramenta, os quais são baseados nos valores médios mensais dos parâmetros. Esta interface é apresentada na Figura 20. Como se pode observar, há vários campos preenchidos com traço. Tais campos são desabilitados para edição, pois dispensam a imputação de valores já que não existem medições faltantes para o parâmetro e mês relacionados. Todos os demais campos, sem traço, são editáveis. Além disso, os parâmetros Chuva 24h, Cloreto Total e Toxicidade não aparecem na interface pois não possuem medições faltantes em nenhum mês.



Parameters	Standard CONAMA	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Dissolved Aluminum	<= 0.1 mg/L	0.61	0.33	0.22	0.15	0.18	0.12	0.13	0.11	0.15	0.09	0.21	0.11
Cadmium Total	<= 0.001 mg/L	0.0005	0.0006	0.0004	0.0004	0.0005	0.0004	0.0009	0.0006	0.0004	0.0003	0.0005	0.0003
Dissolved Copper	<= 0.009 mg/L	0.009	0.006	0.009	0.006	0.009	0.006	0.009	0.005	0.009	0.013	0.009	0.006
Dissolved Iron	<= 0.3 mg/L	0.42	0.48	0.33	0.31	0.27	0.16	0.18	0.27	0.19	0.19	0.3	0.2
Lead Total	<= 0.01 mg/L	0.01	0.007	0.007	0.005	0.008	0.007	0.008	0.007	0.005	0.005	0.007	0.004
Nitrogen Ammoniacal	<= 0.5 mg/L	0.93	0.56	0.94	-	1.35	-	1.63	1.35	1.68	1.82	1.67	-
Nickel Total	<= 0.025 mg/L	0.023	-	0.026	0.012	0.028	-	-	0.013	0.019	0.015	0.019	0.014
Substance Surfactant	<= 0.5 mg/L	0.2	-	0.12	0.06	0.19	-	0.23	-	0.25	-	0.16	-
Zinc Total	<= 0.18 mg/L	0.05	-	0.04	-	0.04	-	-	-	0.04	-	0.04	0.02
Manganese Total	<= 0.1 mg/L	-	-	0.1	-	0.09	-	-	-	-	-	0.13	0.11
Dissolved Oxygen	>= 5 mg/L	-	-	5.4	-	-	-	6.1	-	5.5	-	-	-
Total Solids	-(mg/L)	-	-	-	-	-	-	-	162	162	156	-	-
Water Temperature	-(°C)	-	-	25.4	-	-	-	18.8	-	-	-	-	-
Conductivity	-(µS/cm)	-	-	-	-	180	-	-	145	-	-	-	-
Nitrate	<= 10 mg/L	-	-	-	-	-	-	0.9	0.7	-	-	-	-
Turbidity	<= 100 UNT	91	-	-	-	-	-	-	-	28	-	-	-
Nitrite	<= 1 mg/L	-	-	-	-	-	-	-	0.06	-	-	-	-
pH	>= 6 / <= 9 U.pH	-	-	-	-	-	-	7.1	-	-	-	-	-

Note: The parameters Rain 24h, Chloride Total e Toxicity does not need imputation because they have no missing data.

Figura 20: Interface para imputação dos dados.

A ferramenta também disponibiliza uma interface para customização da discretização padrão apresentada na Tabela 9. Esta interface pode visualizada na Figura 21. É possível categorizar os parâmetros de qualidade obedecendo as seguintes regras:

- **Padrão CONAMA** – Somente os mnemônicos podem ser configurados.

- **Padrão Baixo** – É possível configurar os valores dos campos “<=” e seus respectivos mnemônicos, pois se referem a parâmetros de acompanhamento, que não possuem Padrão CONAMA associado.
- **Padrão Intermediário** – Os valores dos campos “>” assumem o valor do campo Padrão CONAMA ou Padrão Baixo, e não podem ser alterados. É possível configurar os valores dos campos “<=” e seus respectivos mnemônicos. A diferença entre estes dois campos pode ser de até 1000%.
- **Padrão Alto** – Os valores dos campos “>” assumem o valor do campo ‘<=’ do Padrão Intermediário, e não podem ser alterados.
- **Exceções** – Para todas as faixas do pH, somente os mnemônicos podem ser configurados. Os parâmetros Chuva 24h e Toxicidade já são discretos nos próprios dados brutos e, por este motivo, não aparecem na interface.

The interface is titled "Knowledge Discovery in Water Quality Monitoring Database" and includes tabs for "Preprocessing", "Knowledge Discovery", and "About". It displays a table of parameters with their respective standard values and mnemonics for three categories: Standard CONAMA, Intermediate Standard, and Standard High.

Parameters	Standard CONAMA		Intermediate Standard			Standard High	
	<=	Value Discr.	>	<=	Value Discr.	>	Value Discr.
Dissolved Aluminum (mg/L)	0.1	PC	0.1	0.3	AC	0.3	MA
Cadmium Total (mg/L)	0.001	PC	0.001	0.003	AC	0.003	MA
Chloride Total (mg/L)	250	PC	250	750	AC	750	MA
Dissolved Copper (mg/L)	0.009	PC	0.009	0.027	AC	0.027	MA
Dissolved Iron (mg/L)	0.3	PC	0.3	0.9	AC	0.9	MA
Manganese Total (mg/L)	0.1	PC	0.1	0.3	AC	0.3	MA
Nitrogen Ammoniacal (mg/L)	0.5	PC	0.5	1.5	AC	1.5	MA
Nickel Total (mg/L)	0.025	PC	0.025	0.075	AC	0.075	MA
Nitrate (mg/L)	10	PC	10	15	AC	15	MA
Nitrite (mg/L)	1	PC	1	3	AC	3	MA
Lead Total (mg/L)	0.01	PC	0.01	0.03	AC	0.03	MA
Substance Surfactant (mg/L)	0.5	PC	0.5	1.5	AC	1.5	MA
Turbidity (UNT)	100	PC	100	300	AC	300	MA
Zinc Total (mg/L)	0.18	PC	0.18	0.54	AC	0.54	MA
Dissolved Oxygen (mg/L)	>= 5	Value Discr. PC	< 5	>= 2	Value Discr. AB	< 2	Value Discr. MA
pH (U.pH)	Below Standard CONAMA < 6 Value Discr. AB		Standard CONAMA >= 6 <= 9 Value Discr. PC			Above Standard CONAMA > 9 Value Discr. AC	
Conductivity (µS/cm)	Standard Low <= 200 Value Discr. BX		Intermediate Standard > 200 <= 400 Value Discr. MD			Standard High > 400 Value Discr. AT	
Total Solids (mg/L)	200 Value Discr. BX		200 Value Discr. MD			400 Value Discr. AT	
Water Temperature (°C)	21 Value Discr. BX		21 Value Discr. MD			27 Value Discr. AT	

Buttons at the bottom: Save Changes, Restore Defaults, Discretize Data, Restore Database.

Note: Parameters already discretized: Rain 24h -> SI (Yes) and NO (No) / Toxicity -> NT (Non Toxic), CR (Chronic) and AG (Acute)

Figura 21: Interface para discretização dos dados.

Em consultas realizadas na base de dados, percebeu-se que a distribuição das medições entre as categorias se mostrou bastante diferente para alguns parâmetros. Por exemplo, para o Nitrato, 99.5% das medições estão dentro do Padrão CONAMA, 0.3% acima e 0.2% muito acima. Essa disparidade entre as categorias, abre espaço para geração de muitas regras sem importância ou mesmo falsas, visto que em praticamente todas as medições este parâmetro aparecerá dentro do Padrão CONAMA. Em outras palavras, essa situação propicia a geração de regras equivocadas como “**Se** Nitrato=PC **Então** Toxicidade=CR”.

Por este motivo, percebeu-se a necessidade de se descartar categorias muito frequentes antes da mineração dos dados. Para isso, foi criada a interface mostrada na Figura 22, onde é possível marcar e eliminar da base de dados estas categorias. No canto inferior direito, o usuário tem a opção de selecionar uma determinada faixa de frequência (por exemplo: entre 70 e 100%) e marcar simultaneamente todas as categorias que se encontram dentro desta faixa.

Figura 22: Interface para eliminação de categorias.

Por fim, considerando os 21 parâmetros selecionados e disponíveis na tabela Medições, a ferramenta permite a configuração de até 11 parâmetros (10 parâmetros + Toxicidade) para as tarefas de classificação de amostras de água e análise associativa de parâmetros de qualidade, e

até 10 parâmetros para a tarefa de regionalização de pontos de amostragem. Essa restrição foi implementada para manter um nível satisfatório de controle com relação ao desempenho dos algoritmos de mineração. Outro argumento para essa limitação é que a seleção de uma grande variedade de parâmetros poderia aumentar a complexidade dos padrões encontrados, dificultando a compreensão dos resultados.

A Figura 23 apresenta a interface dessa funcionalidade. Após a seleção dos parâmetros de qualidade, pode-se gerar as bases utilizadas na mineração de dados para cada uma das tarefas contempladas neste trabalho: classificação, associação e agrupamento.

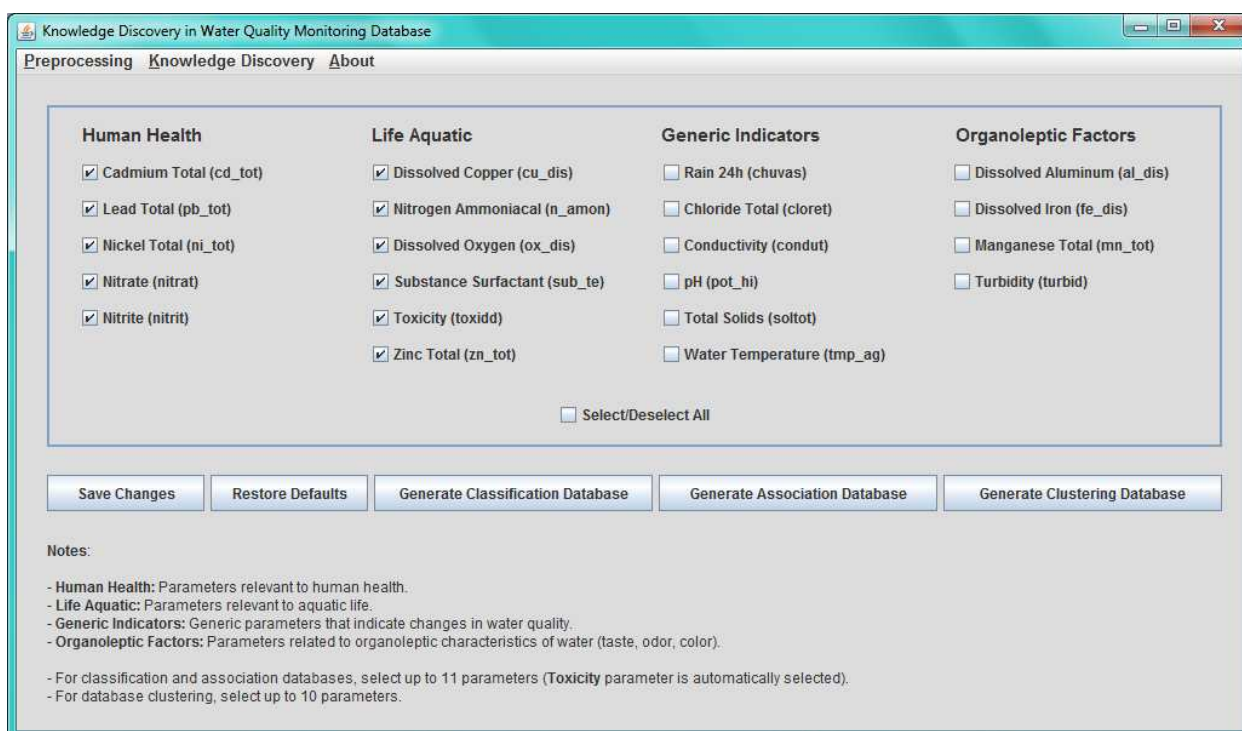


Figura 23: Interface para seleção de parâmetros.

4.5.2 Funcionalidades para Classificação de Toxicidade

A classificação de toxicidade em amostras de água demandou a implementação de funcionalidades para: configuração de taxas mínima de cobertura e precisão para as regras geradas; configuração de informações apresentadas durante a extração das regras; geração das regras a partir da base de treinamento; visualização da cobertura e da precisão de cada regra;

aplicação das regras obtidas na base de teste; visualização do desempenho das regras aplicadas; e apresentação do resultado da validação cruzada. A interface principal desta funcionalidade, apresentada na Figura 24, pode ser dividida em duas partes:

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

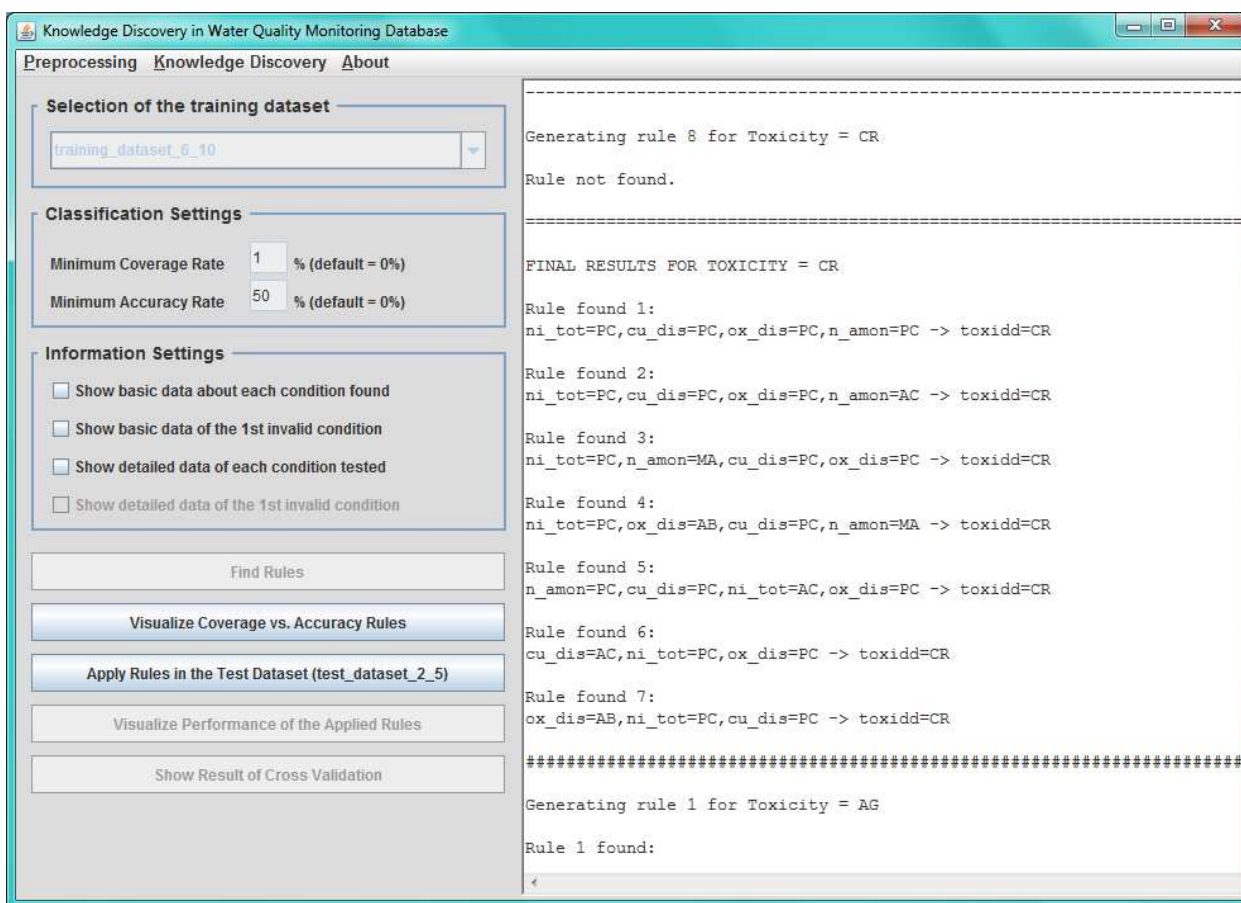


Figura 24. Ferramenta para classificação de toxicidade em amostras de água.

Antes de iniciar a classificação, primeiramente é necessário selecionar a base de dados de treinamento, que será utilizada para o aprendizado do algoritmo. Em seguida, é possível configurar as taxas de cobertura e precisão mínimas que devem ser consideradas na busca de regras, caso não sejam configuradas, são procuradas todas as regras possíveis, independentemente de suas taxas de cobertura e precisão. Também é possível definir algumas

opções de visualização, que permitem configurar até quatro níveis de detalhamento das informações de processamento. Por fim, o botão **Find Rules** inicia o processo de busca de regras de classificação para toxicidade de água.

Após gerar as regras, pode-se visualizar o comportamento das taxas de cobertura e precisão durante a formação de cada regra encontrada por meio do botão **Visualize Coverage vs. Accuracy Rules**. O gráfico de linhas obtido permite visualizar de forma rápida o processo de formação de cada regra gerada, bem como a cobertura e a precisão final alcançados. Essa visão pode auxiliar na análise das regras que devem ser consideradas ou descartadas para a classificação da toxicidade da água. Por exemplo, regras com cobertura mais baixa ou precisão menos satisfatória podem ser desconsideradas pelo analista. Na Figura 25, pode-se notar como as taxas de cobertura e precisão tendem a seguir direções opostas conforme a regra vai sendo aumentada com novas condições (ou parâmetros). Esse fenômeno indica que, em geral, quanto maior a precisão de uma regra, menor será sua cobertura, e vice-versa.

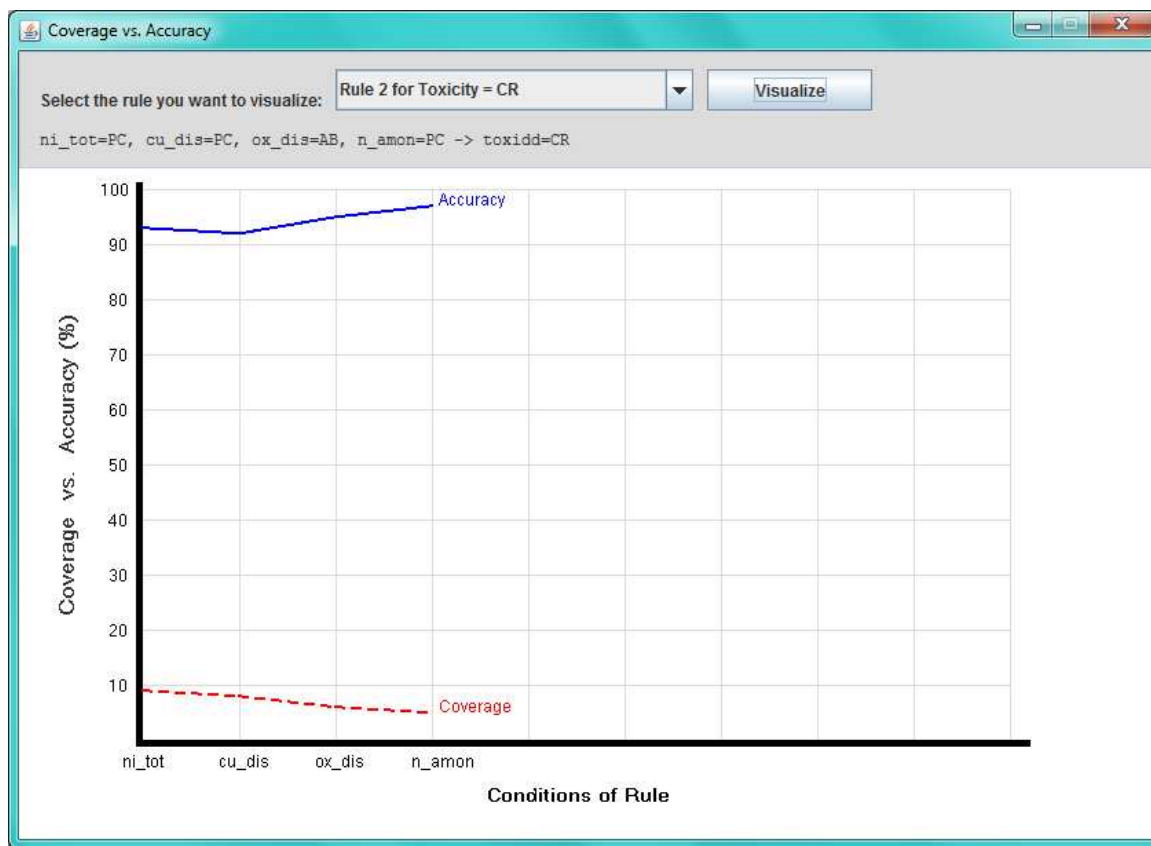


Figura 25. Visualização da cobertura e precisão das regras geradas.

Uma vez geradas as regras, estas podem ser aplicadas na base de teste ao acionar o botão *Apply Rules in the Test Dataset*. Nesse momento, é calculada e apresentada a quantidade de conjuntos de medição classificados incorretamente, bem como a taxa de precisão das regras aplicadas. Vale lembrar que, essa verificação é possível porque as classes dos conjuntos de medição são conhecidas tanto na base de treinamento quanto na base de teste, característica indispensável para se aplicar o método de validação cruzada de duas partes.

Após aplicar as regras geradas na base de teste, pode-se utilizar o botão *Visualize Performance of the Applied Rules*, para visualizar o desempenho do conjunto de regras gerado ao ser aplicado na base de teste. Tal visualização é realizada por meio de um gráfico de pizza, conforme exibido na Figura 26. Cada fatia da pizza representa uma das regras do conjunto gerado. O número inteiro e a porcentagem associados a cada fatia, informam respectivamente a quantidade e a porcentagem de conjuntos de medição cobertos pela regra, considerando todos os conjuntos da base de teste. A fatia referente ao último item da legenda representa sempre a quantidade e a porcentagem de conjuntos de medição classificados incorretamente.

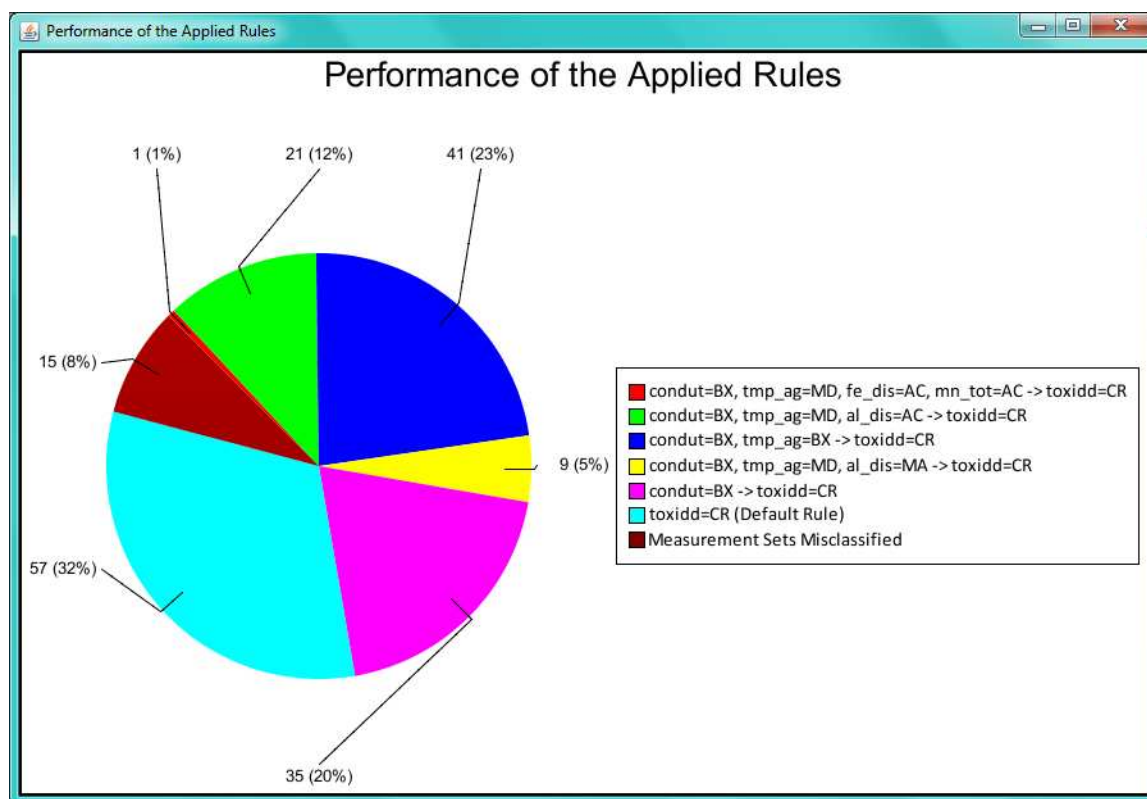


Figura 26. Visualização do desempenho do conjunto de regras aplicado.

O passo seguinte consiste em trocar os papéis das duas bases utilizadas e repetir o mesmo procedimento, de modo que a base que era de treinamento passe a ser a base de teste e vice-versa. Por fim, o desempenho das regras geradas pelas duas iterações pode ser avaliado por meio do botão *Show Result of Cross Validation*.

4.5.3 Funcionalidades para Identificação de Associações entre Parâmetros

A ferramenta para identificação de associações entre os parâmetros de qualidade de água é composta das seguintes funcionalidades: configuração de taxa mínima de suporte, para geração dos conjuntos de parâmetros frequentes; configuração de taxa mínima de confiança, para geração das regras fortes; configuração das informações apresentadas durante a geração dos conjuntos de parâmetros frequentes; geração dos conjuntos de parâmetros frequentes; configuração das informações apresentadas durante a geração das regras fortes; geração das regras fortes; visualização da quantidade de regras geradas em cada execução; filtro para visualização de regras específicas.

A tela principal desta funcionalidade é semelhante à interface para classificação de toxicidade, conforme apresentado na Figura 27. 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.

A primeira tarefa a ser feita é configurar a taxa de suporte mínima, a ser considerada na busca dos conjuntos de parâmetros frequentes, e a taxa de confiança mínima, a ser considerada na busca das regras fortes. Pode-se também definir os dados a serem mostrados na área de mensagens durante o processamento: conjuntos de parâmetros candidatos, conjuntos de parâmetros frequentes, regras de associação candidatas e regras de associação fortes. O botão *Generate Parameters Sets* gera primeiramente os conjuntos de parâmetros candidatos e, em seguida, os conjuntos de parâmetros frequentes. Já o botão *Generate Association Rules* extrai as regras candidatas a partir dos conjuntos de parâmetros frequentes gerados na etapa anterior e, na sequência, seleciona as regras fortes dentre as regras candidatas geradas.

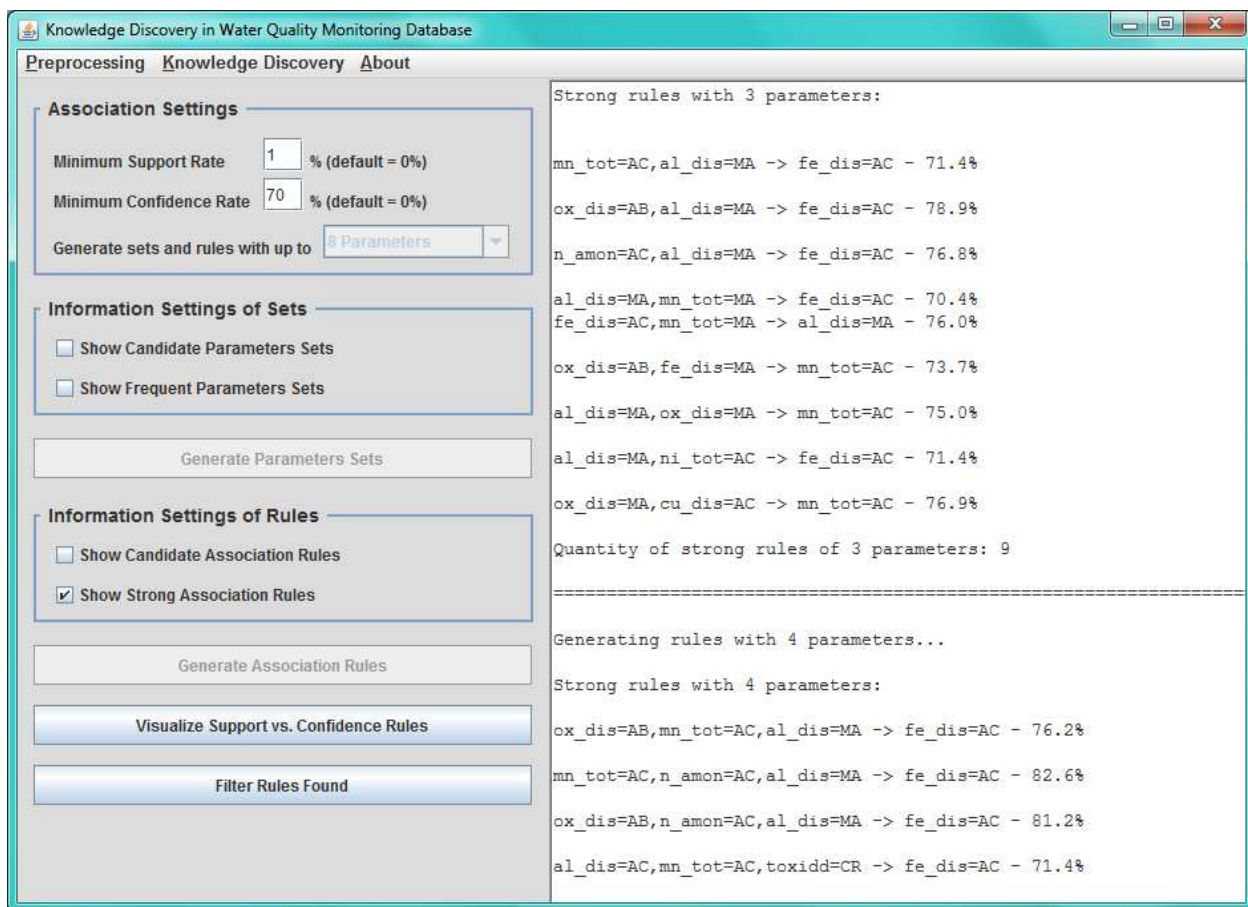


Figura 27. Ferramenta para geração de regras de associação entre os parâmetros de qualidade de água.

Após a extração das regras fortes, pode-se visualizar a quantidade de regras candidatas e fortes geradas em função do suporte e da confiança configurados em cada execução. Isso é feito por meio do botão *Visualize Support vs. Confidence Rules*. O gráfico de barras gerado permite visualizar como o suporte e a confiança, configurados a cada execução, influenciam na quantidade de regras candidatas e fortes geradas. Os números exibidos neste gráfico podem auxiliar o analista na definição dos níveis de suporte e confiança mais adequados para a extração das regras. Na Figura 28, pode-se perceber que quanto menor é a confiança considerada, maior é a quantidade de regras geradas.

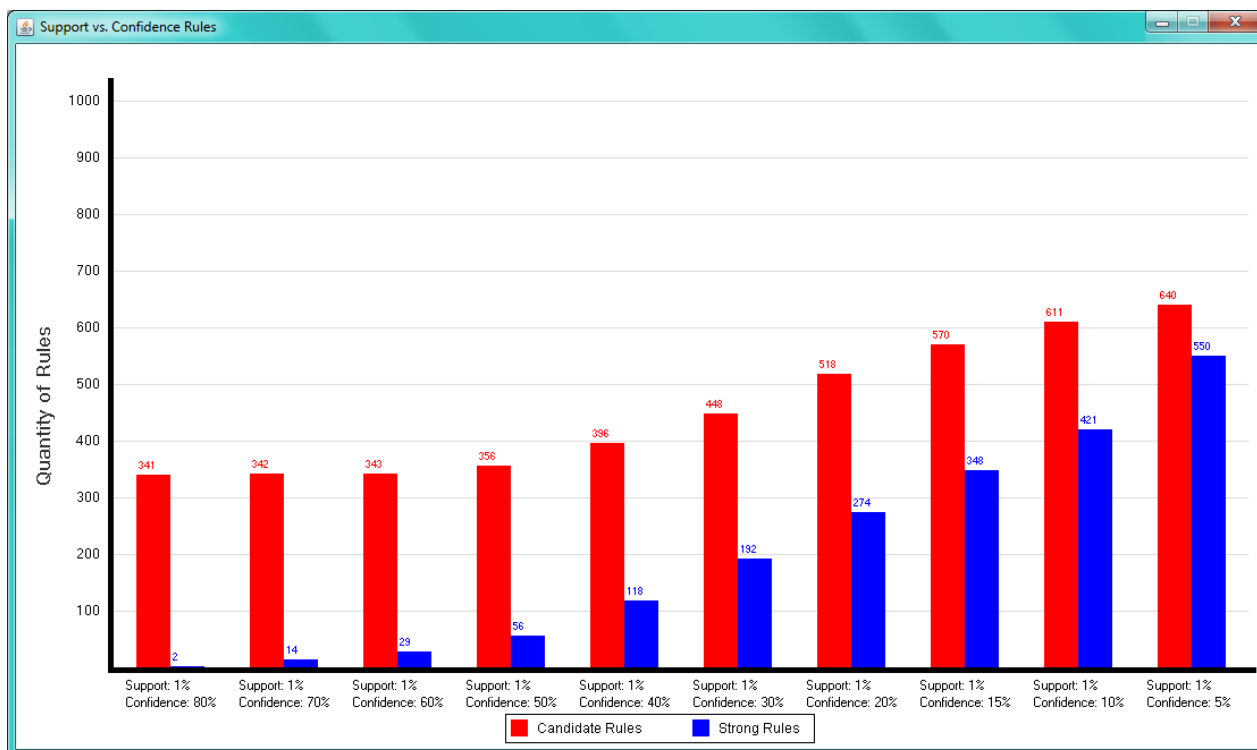


Figura 28. Visualização da quantidade de regras geradas em função do suporte e da confiança configurados.

Outra visualização disponível, ainda que não gráfica, pode ser acessada pelo botão **Filter Rules Found**. Essa funcionalidade, cuja interface é exibida na Figura 29, permite filtrar regras específicas, dentre as regras fortes geradas, com base em determinadas condições informadas pelo usuário. Por exemplo, é possível selecionar todas as regras fortes cujos antecedentes contenham os parâmetros-valores Oxigênio Dissolvido = AB e Nitrogênio Amoniacal = AC, e cujos consequentes contenham o parâmetro-valor Toxicidade = CR.

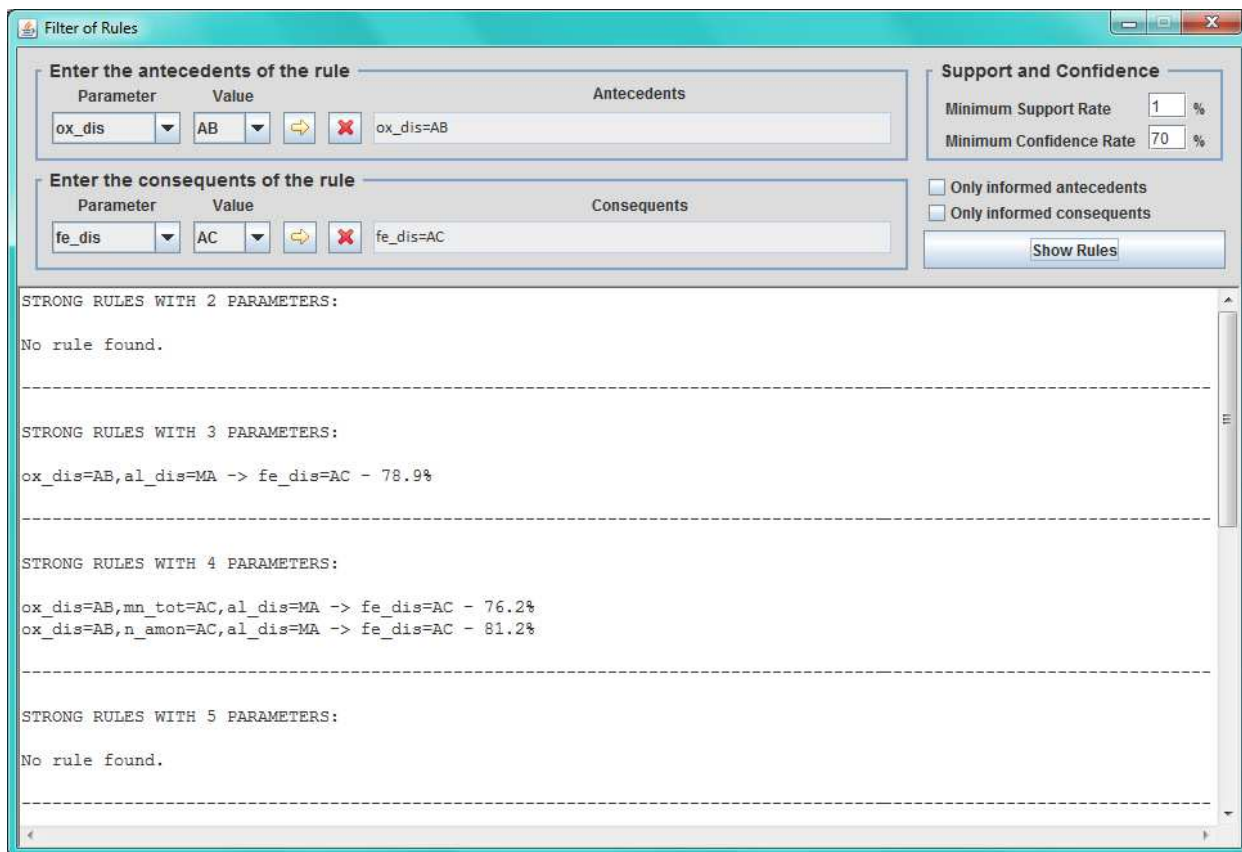


Figura 29. Filtro para visualização de regras específicas.

4.5.4 Funcionalidades para Regionalização de Pontos de Amostragem

O aplicativo desenvolvido para agrupar os pontos de amostragem de água compreende funcionalidades para: seleção do período contemplado na pesquisa, carga e visualização dos pontos de amostragem e de suas conexões fluviais, remoção dos ciclos existentes, visualização das conexões sem ciclos (árvore geradora mínima), geração e visualização dos grupos de pontos de amostragem. Além disso, são disponibilizados recursos para visualização de informações específicas, como os pesos das conexões fluviais e o sentido dos corpos hídricos, e operações para *zoom* e deslocamento do mapa das UGRHIs.

A interface principal, apresentada na Figura 30, pode ser dividida em duas partes:

- **Painel de controle (acima)** – Destina-se às configurações dos períodos considerados (bimestres, estações do ano, semestres e anos), informação dos parâmetros

contemplados, botões de comando e opções de visualização das conexões fluviais, seus pesos (custos) e seus sentidos.

- **Painel de visualização (abaixo)** – Área na qual se pode visualizar os resultados do processamento e os pontos de amostragem e suas conexões sobre o mapa.

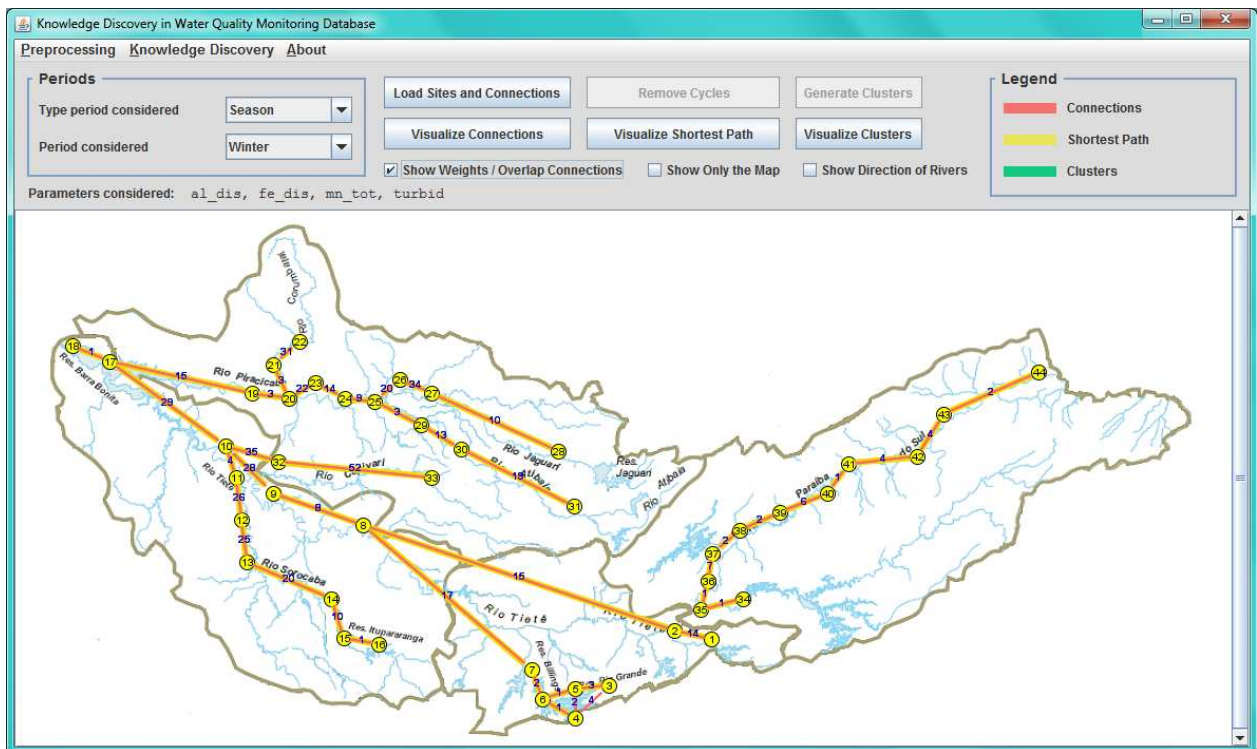


Figura 30: Ferramenta para agrupamento de pontos de amostragem de água.

O processo de agrupamento é iniciado por meio do botão ***Load Sites and Connections***. Ao selecioná-lo, a ferramenta efetua a carga dos pontos de amostragem de água, que serão os vértices do grafo, e das conexões entre os pontos, que serão as arestas que ligarão os vértices do grafo. Estas arestas representam as conexões fluviais entre os pontos de amostragem.

Vale ressaltar que este processo de carga forma uma grafo desconexo, visto que não existe conexão fluvial entre a UGRHI 2 e as UGRHIs 5, 6 e 10. Por este motivo, todo o processo de agrupamento é feito separadamente para cada um dos dois grupos, visando considerar essa divisão hidrográfica natural entre as UGRHIs. O resultado da carga é uma matriz de adjacência, que mostra as conexões entre os pontos de amostragem, bem como os custos de cada uma dessas conexões. Estes custos são obtidos por meio do cálculo da distância euclidiana entre os

parâmetros considerados dos dois pontos ligados pela conexão, conforme sugere Assunção (2002).

Após carregar os pontos de amostragem e suas respectivas conexões, o usuário pode utilizar o botão **Visualize Connections** para visualizar no mapa os pontos de amostragem e suas conexões fluviais, as quais são representadas por meio de linhas vermelhas. Se desejado, pode-se também habilitar a visualização dos pesos das conexões marcando a caixa de seleção **Show Weights/Overlap Connections**. Os pesos exibidos denotam as dissimilaridades entre os parâmetros medidos nos pontos de amostragem. Quanto maior o valor, maiores são as diferenças entre as medições dos dois pontos conectados.

No passo seguinte, o usuário remove os ciclos – conexões que, juntas, criam um caminho fechado no grafo – por meio do botão ***Remove Cycles***. O sistema mostra então as conexões remanescentes após a eliminação dos ciclos. É importante ressaltar que, nesse momento, o objetivo não é formar divisões no grafo, mas apenas eliminar tais “ciclos” existentes. No contexto desta pesquisa, os ciclos representam as conexões entre os pontos de amostragem situados em reservatórios de água. Como o fluxo desses corpos hídricos se move em velocidade muito baixa, suas águas acabam interligando naturalmente seus pontos de amostragem, daí a existência dos ciclos nesse tipo de corpo d’água. A Figura 31 mostra ciclos formados na região das represas Billings e Rio Grande.

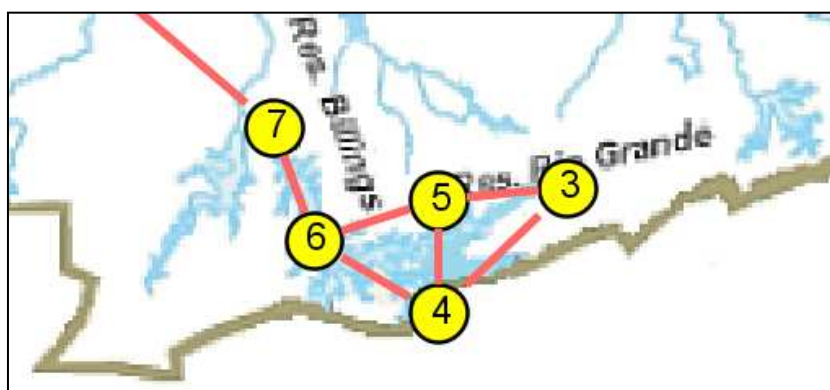


Figura 31: Formação de ciclo entre pontos de amostragem de água.

Na sequência, o usuário pode visualizar a árvore geradora mínima criada por meio do botão **Visualize Shortest Path**. Suas conexões são representadas por meio de linhas amarelas. Da mesma forma descrita anteriormente, pode-se habilitar a visualização dos pesos das conexões marcando a caixa de seleção **Show Weights/Overlap Connections**.

A partir desse ponto, qualquer conexão que for eliminada gerará dois grupos de pontos de amostragem. Dessa forma, o passo seguinte consiste na geração dos grupos a partir da poda da árvore gerada. Para isso, o usuário utiliza o botão **Generate Clusters**. Ao acioná-lo, a ferramenta abre duas janelas, conforme mostrado na Figura 32, onde se deve informar a quantidade de grupos a serem gerados no conjunto de UGRHIs 5, 6 e 10 e na UGRHI 2. Nesse instante são removidas as conexões de maior custo, sendo apresentados: quantidade de conexões removidas, quais foram estas conexões e seus respectivos pesos.

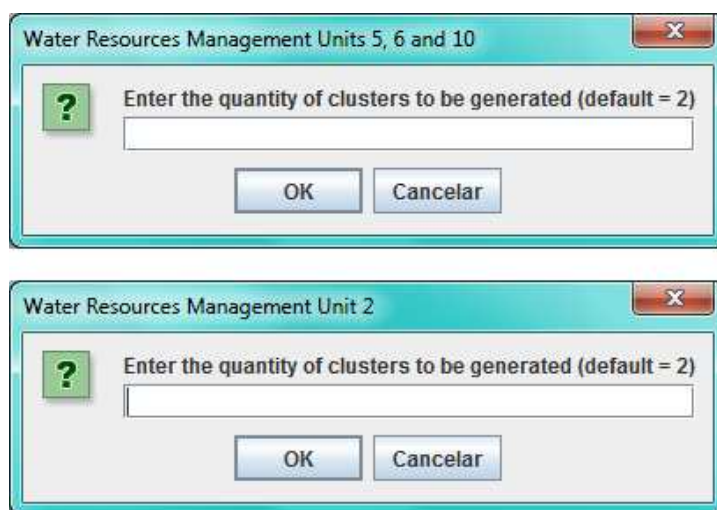


Figura 32: Janelas para informação da quantidade de grupos a serem formados.

Por fim, o usuário pode visualizar os grupos gerados pela poda por meio do botão **Visualize Clusters**. Tais grupos são representados por linhas verdes que conectam seus respectivos pontos de amostragem.

Vale lembrar que a forma de atribuir custos às conexões é alterada na fase de poda, pois passa a variar em função da soma dos quadrados dos desvios associada às subárvores geradas. O intuito dessa mudança é obter regiões mais homogêneas com relação às medições dos parâmetros de qualidade considerados.

O custo de cada aresta do grafo é dado pela diferença entre os desvios da árvore completa e a soma dos desvios das duas subárvores geradas pela remoção da aresta na qual se está calculando o custo. Quanto maior for esta diferença, mais alto será o custo da aresta. Consequentemente, as chances de sua remoção originar regiões mais homogêneas também serão maiores.

Esse benefício pode ser observado no exemplo da Figura 33, onde são considerados os parâmetros Sólidos Totais e Turbidez nos meses de inverno. O algoritmo detectou que a remoção da aresta entre os pontos de amostragem 35 e 36 proporcionaria a melhor dupla de grupos em termos de homogeneidade.

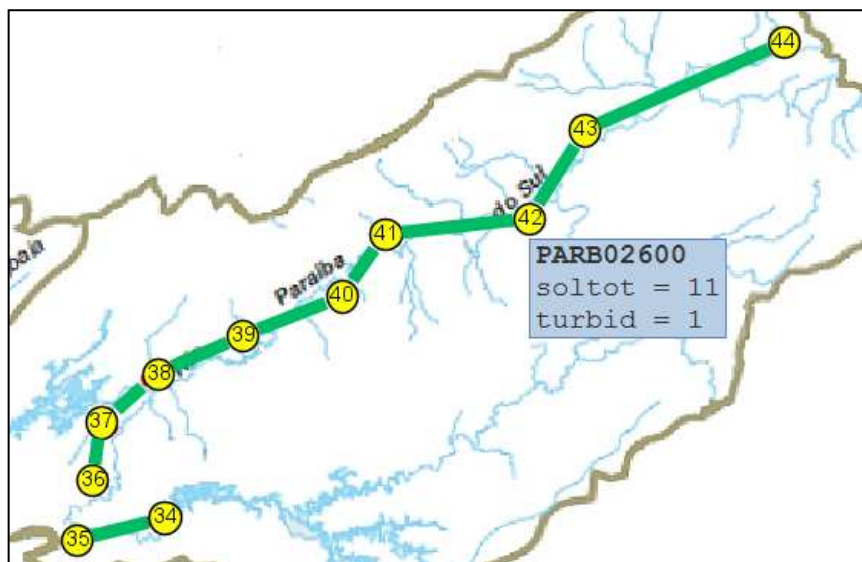


Figura 33: Geração de grupos de pontos de amostragem no Rio Paraíba do Sul.

Conforme a Tabela 10, de fato, observa-se que os pontos 34 e 35 (em azul) possuem medições similares, e os pontos 36 a 44 (em amarelo) são bastante homogêneos. Por outro lado, nota-se também que esses dois grupos de medições diferem consideravelmente entre si.

Tabela 10: Medições médias normalizadas dos parâmetros Sólidos Totais e Turbidez no Rio Paraíba do Sul.

Pontos de Amostragem	34	35	36	37	38	39	40	41	42	43	44
Sólidos Totais	6	6	9	9	10	9	10	10	11	11	12
Turbidez	0	0	0	0	1	1	1	1	1	1	1

5 Experimentação

Este Capítulo apresenta os experimentos realizados na ferramenta para descoberta de conhecimento em dados de monitoramento de qualidade de água, bem como os resultados obtidos pela mineração dos dados. A primeira Seção aborda a questão da classificação de toxicidade em amostras de água. A segunda tem como foco a análise associativa entre os parâmetros de qualidade de água. Por fim, a última Seção trata o agrupamento dos pontos de amostragem de água. Os resultados alcançados são apresentados e avaliados nos Apêndices A, B e C deste trabalho.

5.1 Classificação de Toxicidade

Esta Seção apresenta os experimentos para classificação de toxicidade em amostras de água realizadas por meio da ferramenta desenvolvida neste trabalho. São descritas desde as configurações de pré-processamento e mineração de dados considerados, até os resultados alcançados e suas respectivas análises.

5.1.1 Configurações de Pré-processamento

Foram consideradas as seguintes configurações de pré-processamento:

- **Imputação dos Dados** – Foi utilizado o padrão de atribuição de valores pré-configurado na ferramenta, o qual considera o valor médio mensal de cada parâmetro de qualidade nos sete anos contemplados (2005-2011).
- **Discretização dos Dados** – Foi utilizado o padrão de categorização pré-configurado na ferramenta, o qual é apresentado na Tabela 9.
- **Seleção de Categorias** – Para obter informações mais significativas, descartando regras sem importância ou mesmo pouco confiáveis, foram descartadas todas as medições que

se encontravam dentro do Padrão CONAMA, muitas das quais apresentavam ocorrência acima de 90% considerando o total de registros. Pelo mesmo motivo, as medições com valor “Não Tóxico” também foram desprezadas. Para os parâmetros de acompanhamento Condutividade, Sólidos Totais e Temperatura da Água, além do parâmetro Chuva 24h, foram mantidas todas as categorias, pois estes não possuem Padrão CONAMA associado.

- **Seleção de Parâmetros** – Os parâmetros foram divididos em cinco grupos, cada qual representando um experimento diferente, conforme descrito na Seção 5.1.3.

5.1.2 Configurações de Mineração de Dados

Foram consideradas as seguintes configurações de mineração de dados:

- **Cobertura** – Em função do descarte das medições dentro do Padrão CONAMA e das medições de Toxicidade com valor “Não Tóxico”, as quais contemplavam a maior parte da base de dados, foi necessário trabalhar com uma taxa de cobertura mínima bastante baixa. Por este motivo, para tornar viável a descoberta de regras significativas baseadas em medições fora do Padrão COMANA e com Toxicidade Crônica ou Aguda, a taxa de cobertura mínima das regras de classificação a serem encontradas foi estabelecida em 1%.
- **Precisão** – Visando evitar a geração de regras com baixo nível de confiabilidade, a taxa mínima de precisão das regras de classificação a serem encontradas foi definida em 80%.

5.1.3 Experimentos

Foram realizados cinco experimentos referentes à classificação de toxicidade em amostras de água: quatro específicos, um para cada categoria de parâmetros de qualidade, e um generalista, o qual foi baseado nos resultados obtidos nos quatro anteriores. A seguir, a descrição de cada um dos experimentos:

- **Experimento 1** – Classificação de toxicidade em amostras de água considerando apenas os parâmetros relacionados à **saúde humana**. Este experimento envolveu os parâmetros: Cádmio Total, Chumbo Total, Níquel Total, Nitrato, Nitrito e Toxicidade.
- **Experimento 2** – Classificação de toxicidade em amostras de água considerando apenas os parâmetros relacionados à **vida aquática**. Este experimento envolveu os parâmetros: Cobre Dissolvido, Nitrogênio Amoniacal, Oxigênio Dissolvido, Substância Tensoativa, Zinco Total e Toxicidade.
- **Experimento 3** – Classificação de toxicidade em amostras de água considerando apenas os parâmetros relacionados a **indicadores genéricos**. Este experimento envolveu os parâmetros: Chuva 24h, Condutividade, pH, Sólidos Totais, Temperatura Água e Toxicidade. O parâmetro Cloreto Total foi descartado, pois 100% de suas medições se apresentavam dentro do Padrão CONAMA.
- **Experimento 4** – Classificação de toxicidade em amostras de água considerando apenas os parâmetros relacionados a **fatores organolépticos**. Este experimento envolveu os parâmetros: Alumínio Dissolvido, Ferro Dissolvido, Manganês Total, Turbidez e Toxicidade.
- **Experimento Generalista** – Classificação de toxicidade em amostras de água contemplando apenas os parâmetros considerados mais significativos nos quatro experimentos anteriores, além da Toxicidade que é atributo alvo. A significância foi mensurada com base nas ocorrências dos parâmetros nos experimentos.

Os resultados alcançados nos cinco experimentos realizados, bem como suas respectivas avaliações e interpretações são apresentados no Apêndice A deste trabalho.

5.2 Identificação de Associações entre Parâmetros

Esta Seção apresenta os ensaios realizados para análise associativa de parâmetros de qualidade de água, baseados na ferramenta desenvolvida para este trabalho. São descritas desde as configurações de pré-processamento e mineração de dados contempladas, até os resultados atingidos e suas respectivas avaliações.

5.2.1 Configurações de Pré-processamento

Foram consideradas as seguintes configurações de pré-processamento:

- **Imputação dos Dados** – Foi utilizado o padrão de atribuição de valores pré-configurado na ferramenta, o qual considera o valor médio mensal de cada parâmetro de qualidade nos sete anos contemplados (2005-2011).
- **Discretização dos Dados** – Foi utilizado o padrão de categorização pré-configurado na ferramenta, o qual é apresentado na Tabela 9.
- **Seleção de Categorias** – Para obter informações mais significativas, descartando regras sem importância ou mesmo pouco confiáveis, foram descartadas todas as medições que se encontravam dentro do Padrão CONAMA, muitas das quais apresentavam ocorrência acima de 90% considerando o total de registros. Pelo mesmo motivo, as medições com valor “Não Tóxico” também foram desprezadas. Para os parâmetros de acompanhamento, Condutividade, Sólidos Totais e Temperatura da Água, além do parâmetro Chuva 24h, foram mantidas todas as categorias, pois estes não possuem Padrão CONAMA associado.
- **Seleção de Parâmetros** – Os parâmetros foram divididos em sete grupos, cada qual representando um experimento diferente, conforme descrito na Seção 5.2.3.

5.2.2 Configurações de Mineração de Dados

Foram consideradas as seguintes configurações de mineração de dados:

- **Suporte** – Em função do descarte das medições dentro do Padrão CONAMA e das medições de Toxicidade com valor “Não Tóxico”, as quais contemplavam a maior parte da base de dados, foi necessário trabalhar com uma taxa de suporte mínimo bastante baixa. Por este motivo, para tornar viável a descoberta de regras significativas baseadas em medições fora do Padrão COMANA e com Toxicidade Crônica ou Aguda, a taxa de suporte mínimo das regras de associação a serem encontradas foi estabelecida em 1%.

- **Confiança** – A fim de evitar a geração de regras com baixo nível de confiabilidade, a taxa mínima de confiança das regras de associação a serem encontradas foi definida a partir de 80%. Como alguns experimentos geraram um número muito grande de regras, foi necessário aumentar a taxa de confiança, a fim de limitar esse número. Desse modo, foi possível se concentrar somente nas regras mais confiáveis, facilitando a análise dos resultados. A descrição de cada experimento é apresentada na próxima Seção. As taxas mínimas de confiança estabelecidas para cada um deles foram:
 - **Experimento 1** – Taxa Mínima de Confiança = 80%
 - **Experimento 2** – Taxa Mínima de Confiança = 90%
 - **Experimento 3** – Taxa Mínima de Confiança = 80%
 - **Experimento 4** – Taxa Mínima de Confiança = 95%
 - **Experimento 5** – Taxa Mínima de Confiança = 80%
 - **Experimento 6** – Taxa Mínima de Confiança = 95%
 - **Experimento Generalista** – Taxa Mínima de Confiança = 80%

5.2.3 Experimentos

Foram realizados sete experimentos para identificação de associações entre parâmetros de qualidade de água: seis específicos, um para cada possível dupla de categorias de parâmetros de qualidade, e um generalista, o qual foi baseado nos resultados obtidos nos seis anteriores. A seguir, a descrição de cada um dos experimentos:

- **Experimento 1** – Identificação de associações entre os parâmetros relacionados à **saúde humana** e à **vida aquática**. Este experimento envolveu 11 parâmetros: Cádmio Total, Chumbo Total, Níquel Total, Nitrato, Nitrito, Cobre Dissolvido, Nitrogênio Amoniacal, Oxigênio Dissolvido, Substância Tensoativa, Zinco Total e Toxicidade.
- **Experimento 2** – Identificação de associações entre os parâmetros relacionados à **saúde humana** e a **indicadores genéricos**. Este experimento envolveu 10 parâmetros:

Cádmio Total, Chumbo Total, Níquel Total, Nitrato, Nitrito, Chuva 24h, Condutividade, pH, Sólidos Totais e Temperatura Água. O parâmetro Cloreto Total foi descartado, pois 100% de suas medições se apresentavam dentro do Padrão CONAMA.

- **Experimento 3** – Identificação de associações entre os parâmetros relacionados à **saúde humana** e a **fatores organolépticos**. Este experimento envolveu 9 parâmetros: Cádmio Total, Chumbo Total, Níquel Total, Nitrato, Nitrito, Alumínio Dissolvido, Ferro Dissolvido, Manganês Total e Turbidez.
- **Experimento 4** – Identificação de associações entre os parâmetros relacionados à **vida aquática** e a **indicadores genéricos**. Este experimento envolveu 11 parâmetros: Cobre Dissolvido, Nitrogênio Amoniacal, Oxigênio Dissolvido, Substância Tensoativa, Zinco Total, Toxicidade, Chuva 24h, Condutividade, pH, Sólidos Totais e Temperatura Água. O parâmetro Cloreto Total foi descartado pelo motivo já descrito no Experimento 2.
- **Experimento 5** – Identificação de associações entre os parâmetros relacionados à **vida aquática** e a **fatores organolépticos**. Este experimento envolveu 10 parâmetros: Cobre Dissolvido, Nitrogênio Amoniacal, Oxigênio Dissolvido, Substância Tensoativa, Zinco Total, Toxicidade, Alumínio Dissolvido, Ferro Dissolvido, Manganês Total e Turbidez.
- **Experimento 6** – Identificação de associações entre os parâmetros relacionados a **indicadores genéricos** e a **fatores organolépticos**. Este experimento envolveu 9 parâmetros: Chuva 24h, Condutividade, pH, Sólidos Totais, Temperatura Água, Alumínio Dissolvido, Ferro Dissolvido, Manganês Total e Turbidez.
- **Experimento Generalista** – Identificação de associações entre os parâmetros relacionados contemplando apenas os parâmetros considerados mais significativos nos seis experimentos anteriores. A significância foi mensurada com base nas ocorrências dos parâmetros nos experimentos.

Os resultados alcançados nos sete experimentos realizados, bem como suas respectivas avaliações e interpretações são apresentados no Apêndice B deste trabalho.

5.3 Regionalização de Pontos de Amostragem

Esta Seção apresenta os experimentos para agrupamento dos pontos de amostragem de água em regiões homogêneas, realizado por meio da ferramenta desenvolvida neste trabalho. São apresentadas desde as configurações de pré-processamento e mineração de dados aplicadas, até os resultados obtidos e suas respectivas análises.

5.3.1 Configurações de Pré-processamento

Foram consideradas as seguintes configurações de pré-processamento:

- **Imputação dos Dados** – Foi utilizado o padrão de atribuição de valores pré-configurado na ferramenta, o qual considera o valor médio mensal de cada parâmetro de qualidade nos sete anos contemplados (2005-2011).
- **Seleção de Parâmetros** – Os parâmetros foram divididos em quatro grupos, cada qual representando uma categoria de parâmetros diferente. Os experimentos realizados com cada grupo são descritos na Seção 5.3.3.

5.3.2 Configurações de Mineração de Dados

Foram consideradas as seguintes configurações de mineração de dados:

- **Períodos considerados** – Para restringir os experimentos às situações mais interessantes e com maior possibilidade de trazer à tona informações significativas, foram priorizadas as configurações baseadas nos 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** – Ao considerar os 12 meses do ano, tem-se uma visão geral do comportamento dos parâmetros em cada ponto de amostragem.

- Para as UGRHIs 5, 6 e 10, que reunidas possuem 33 pontos de amostragem e são interligadas hidrograficamente, foi estabelecida a geração de seis grupos de pontos em todos os experimentos. Proporcionalmente, para a UGRHI 2 que possui 11 pontos, foi definida a criação de dois grupos nos experimentos.

5.3.3 Experimentos

Foram realizados quatro grupos de experimentos para regionalização dos pontos de amostragem de água. Para cada um destes grupos, 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 dos grupos de experimentos:

- **Experimentos 1 a 5** – Agrupamento de pontos de amostragem de água similares em relação às medições de parâmetros associados à **saúde humana**. Este experimento envolveu os parâmetros: Cádmio Total, Chumbo Total, Níquel Total, Nitrato e Nitrito.
- **Experimentos 6 a 10** – Agrupamento de pontos de amostragem de água similares em relação às medições de parâmetros associados à **vida aquática**. Este experimento envolveu os parâmetros: Cobre Dissolvido, Nitrogênio Amoniacal, Oxigênio Dissolvido, Substância Tensoativa, Zinco Total e Toxicidade.
- **Experimentos 11 a 15** – Agrupamento de pontos de amostragem de água similares em relação às medições de parâmetros associados a **indicadores genéricos**. Este experimento envolveu os parâmetros: Chuva 24h, Cloreto Total, Condutividade, pH, Sólidos Totais e Temperatura Água.
- **Experimentos 16 a 20** – Agrupamento de pontos de amostragem de água similares em relação às medições de parâmetros associados a **fatores organolépticos**. Este experimento envolveu os parâmetros: Alumínio Dissolvido, Ferro Dissolvido, Manganês Total e Turbidez.

Os resultados alcançados nos quatro grupos de experimentos realizados, bem como suas respectivas avaliações e interpretações são apresentados no Apêndice C deste trabalho.

6 Conclusões

Essa pesquisa teve como propósito a descoberta de informações úteis, ocultas em meio a um conjunto circunscrito de dados de monitoramento de qualidade de água, levantados pela CETESB entre os anos de 2005 e 2011. Para isso, foi aplicada uma metodologia baseada no tradicional processo de KDD e em tarefas e técnicas conhecidas da área de mineração de dados. Durante a pesquisa, pôde-se observar o 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.

Ainda que muitos dos resultados obtidos tenham se mostrado interessantes, deve-se ressaltar que a descoberta de conhecimento é um processo intrinsecamente exploratório e iterativo, característica que demanda muitos ajustes e, consequentemente, novas iterações e experimentos em busca de padrões em meio aos dados. Além disso, a quantidade de variações de configuração oferecidas pela ferramenta desenvolvida implica em uma enorme gama de possibilidades, as quais não foram esgotadas nos experimentos realizados. Dessa forma, muitas das decisões tomadas ao longo desta pesquisa são passíveis de adequações e consequentes reavaliações dos padrões encontrados.

Enfim, os resultados obtidos nesta pesquisa demonstraram a efetividade do processo de descoberta de conhecimento e das técnicas de mineração de dados empregadas na extração de informações implícitas nos dados de monitoramento de qualidade de água. As descobertas proporcionadas por essa metodologia podem originar subsídios valiosos para as diversas áreas envolvidas, desde os técnicos que fazem as análises laboratoriais das amostras de água até os tomadores de decisão responsáveis pela definição das futuras políticas públicas para gestão dos recursos hídricos. Nas próximas Seções são apresentadas as contribuições deste trabalho, as principais dificuldades encontradas durante a pesquisa e alguns possíveis trabalhos futuros.

6.1 Contribuições

As contribuições gerais deste trabalho estão relacionadas a questões ambientais, econômicas e sociais, uma vez que a água é um recurso fundamental para o ecossistema de nosso planeta, para a economia mundial e, conseqüentemente, para a manutenção da sociedade como conhecemos hoje. Este estudo também contribui no âmbito da saúde pública, pois a água tem papel crucial na qualidade de vida da população. Pela perspectiva da gestão ambiental, esta pesquisa pode auxiliar no entendimento de como as atividades humanas vem afetando nossos corpos d'água, propiciando métodos e subsídios úteis para o monitoramento de outras bacias hidrográficas suscetíveis aos impactos ambientais causados pelo homem. Por fim, a ferramenta desenvolvida contribui no âmbito da computação, pois permite constatar a efetividade de alguns dos mais tradicionais algoritmos de mineração, quando aplicados ao problema do monitoramento de qualidade de água em corpos hídricos.

De um ponto de vista mais específico, este trabalho proporciona as seguintes contribuições:

Predição da toxicidade em amostras de água:

- Compreensão de como os parâmetros de qualidade de água afetam a toxicidade da água dos corpos hídricos.
- Redução do uso de organismos vivos nas análises ecotoxicológicas.
- Análises de ecotoxicidade mais rápidas e eficazes.

Correlação entre os parâmetros de qualidade de água:

- Extração de correlações ocultas entre os parâmetros de qualidade de água.
- Comprovação da validade de associações entre parâmetros pressupostas de forma empírica.

Regionalização de Pontos de Amostragem:

- Revelação das distinções existentes entre os corpos hídricos ou ainda entre trechos de um mesmo corpo hídrico.

- Levantamento de subsídios para apoiar a distribuição dos locais de coleta de água necessários para o monitoramento dos corpos hídricos.
- Geração de insumos para definir diretrizes estratégicas específicas para cada região gerada pela regionalização.

6.2 Dificuldades Encontradas

Uma das primeiras dificuldades encontradas neste trabalho adveio do formato em que se encontravam disponíveis os dados brutos. Como foi mostrado, a extração destes dados a partir de arquivos em formato PDF precisou ser realizada em várias etapas, sendo necessária inclusive a implementação de duas aplicações específicas para esta finalidade. Esse cenário fez com que a fase inicial deste trabalho fosse uma das mais custosas.

Um dos obstáculos encontrados durante a pesquisa se refere à qualidade dos dados disponíveis. Devido à grande quantidade de medições incompletas, onde 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 das técnicas de mineração de dados. Tal medida fez com que o conjunto de dados inicialmente disponível fosse drasticamente reduzido.

Os comportamentos específicos apresentados por alguns parâmetros de qualidade também dificultaram a descoberta de informações relevantes nas tarefas de classificação de amostras de água e associação entre os parâmetros. O número diminuto de medições fora do Padrão CONAMA, em alguns parâmetros, fez com que pouco pudesse ser descoberto quando estes eram considerados nos experimentos.

6.3 Trabalhos Futuros

Com relação aos aspectos mais gerais desta pesquisa, em trabalhos futuros, outros algoritmos poderiam ser aplicados para as mesmas finalidades apresentadas, utilizando o mesmo

conjunto de dados. Isto poderia proporcionar um comparativo interessante no que tange à significância dos resultados obtidos em ambas as pesquisas.

Na etapa de pré-processamento dos dados, poderiam ser considerados métodos mais elaborados e confiáveis para imputação dos dados. Uma abordagem interessante seria utilizar técnicas que considerem o componente da probabilidade e estimem os valores dos parâmetros baseando-se não somente nos dados, mas também na relação entre os parâmetros observados.

Especificamente sobre as tarefas de classificação e análise associativa, um aprofundamento sobre a questão do desequilíbrio entre as classes dos parâmetros seria bastante útil. Neste trabalho, utilizou-se um método empírico, que tratou esse problema por meio da eliminação das medições muito frequentes, que no caso estavam predominantemente dentro do Padrão CONAMA. A aplicação de métodos mais apropriados para tratar essa questão poderia trazer resultados mais satisfatórios na mineração dos dados.

Outra interessante contribuição para a abordagem da análise associativa, seria a 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.

Com relação à tarefa de clusterização, este trabalho não se preocupou em analisar o grau de homogeneidade dos grupos de pontos de amostragem formados. Uma oportunidade de trabalho futuro, está na análise dos grupos gerados com relação aos seus respectivos níveis de uniformidade. Outra possível melhoria, seria a inclusão de uma heurística para criação de grupos com um número mínimo de pontos de amostragem, visto que a técnica utilizada na fase de poda gerou grupos com apenas um ponto de amostragem em diversas ocasiões.

Apesar de não ser o foco principal deste trabalho, uma evolução possivelmente necessária na ferramenta desenvolvida está relacionada à questão da visualização dos dados. No período de avaliação e interpretação dos resultados, percebeu-se que as representações gráficas implementadas poderiam ser mais adequadas e melhor exploradas. Essa melhoria auxiliaria fortemente na etapa de análise do conhecimento gerado.

6.4 Artigos Publicados e Aceitos

Durante o mestrado foram publicados três artigos referentes a esta pesquisa. Seus títulos, resumos e eventos onde foram publicados são apresentados a seguir.

“Processamento e Visualização de Dados para a Descoberta de Conhecimento em Sistemas de Monitoramento de Qualidade de Água”

III Workshop de Computação Aplicada à Gestão do Meio Ambiente e Recursos Naturais – Congresso da Sociedade Brasileira de Computação (CSBC 2011, Natal-RN)

Este trabalho propõe o uso de técnicas de visualização e mineração de dados na descoberta de conhecimento no domínio de dados de monitoramento de qualidade de água. No atual estágio da pesquisa, são analisadas algumas técnicas de visualização para a representação e identificação de padrões de comportamento nestes dados. Para realização deste trabalho, foi selecionado, pré-processado e transformado um grande conjunto de dados relativos a medições de qualidade de água dos rios do estado de São Paulo. Com isso, espera-se que estudos semelhantes possam ser reproduzidos e empregados em áreas mais amplas, independentemente dos parâmetros de qualidade de água envolvidos.

“Técnicas de Mineração de Dados na Classificação de Ecotoxicidade de Água para Aplicação na Gestão de Corpos Hídricos”

VIII Congresso Nacional de Excelência em Gestão (CNEG 2012, Niterói-RJ)

Dentre as diversas formas de ação que promovem a sustentabilidade, a inovação tecnológica pode ser considerada uma das mais importantes. Neste trabalho são aplicadas técnicas de mineração de dados na descoberta de conhecimento no domínio de dados de monitoramento de qualidade de água, para prover subsídios úteis e relevantes que auxiliem na tomada de decisão em sistemas de gestão ambiental. No estágio atual da pesquisa, está sendo utilizada uma técnica de modelagem previsiva conhecida como classificação baseada em regras, onde o objetivo é descobrir regras que possam, com base nos valores de determinados parâmetros químicos, prever o nível de ecotoxicidade de uma amostra de água. Foram utilizados dados referentes a análises de água dos principais corpos hídricos do estado de São Paulo, realizadas entre os anos de 2005 e 2010. Espera-se obter uma forma confiável, rápida e eficaz para prever os níveis de ecotoxicidade de água em rios, lagos e reservatórios com base em análises de parâmetros químicos, ou indicar a complementaridade dessas medições em busca da otimização das redes de monitoramento e consequente melhoria da gestão dos recursos naturais.

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

IV Workshop de Computação Aplicada à Gestão do Meio Ambiente e Recursos Naturais – Congresso da Sociedade Brasileira de Computação (CSBC 2013, Maceió-AL)

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.

Além destes três trabalhos, um artigo baseado naquele publicado no CNEG 2012 foi aceito, porém ainda não publicado. Neste congresso foi feita uma seleção dos melhores trabalhos, e este artigo foi um dos escolhidos para publicação no *journal IJESD*. A seguir, seu título e resumo.

“Data Mining Techniques for Water Ecotoxicity Classification for Application on Water Resources Management”

International Journal of Environment and Sustainable Development (IJESD 2012, Reino Unido)

Among the various forms of action that promote sustainability, technological innovation can be considered one of the most important. This paper applied data mining techniques to discover knowledge in the field of water quality monitoring data, providing useful and relevant support for decision making in environmental management systems. At the current stage of research, a predictive modeling technique, known as rule-based classification, was used to find rules that can, based on the values of certain chemical parameters, predict the level of ecotoxicity of a water sample. We used data from analyzes of water from main water bodies of São Paulo state, from 2005 to 2010. We expect to get a reliable, fast and effective way to predict the ecotoxicity levels of water in rivers, lakes and reservoirs based on analyzes of chemical parameters, or indicate the complementarity of these measurements for optimization of monitoring networks and the consequent improvement natural resources management.

Referências Bibliográficas

AILAMAKI, A.; FALOUTSOS, C.; FISCHBECK, P. S.; SMALL, M. J.; VANBRIESEN, J. An environmental sensor network to determine drinking water quality and security. **SIGMOD Record**, v. 32, n. 4, p. 47-52, dez. 2003.

ALVES, E. C.; SILVA, C. F.; COSSICH, E. S.; TAVARES, C. R. G.; FILHO, E. E. S.; CARNIEL, A. Avaliação da qualidade da água da bacia do rio Pirapó – Maringá, Estado do Paraná, por meio de parâmetros físicos, químicos e microbiológicos. **Acta Scientiarum. Technology**, v. 30, n. 1, p. 39-48, 2008.

ARTIOLA, J.F.; PEPPER, I.L.; BRUSSEAU, M. L. Monitoring and Characterization of the Environment. **Environmental Monitoring and Characterization**. Elsevier Academic Press, chap. 1, p. 1-9, 2004.

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

BARIONI, M. C. N. **Visualização de operações de junção em sistemas de bases de dados para mineração de dados**. Dissertação (Mestrado) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, Brasil, 2002.

BELO, O.; LOURENÇO, A.; SARMENTO, P.; MAGRIÇO, A.; PINHO, J. L.; LIMA, M.; VIEIRA, J. AQuA: um sistema de informação para análise e a validação de parâmetros de qualidade da água em Alqueva. In: CONGRESSO DA ÁGUA, 8, Figueira da Foz, Portugal: Universidade do Minho, 2006.

BERRY, M. J. A.; LINOFF, G. S. **Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management**. Wiley Publishing, Inc., 672 p., 2004.

BERTHOLDO, L.; SILVA, C. G.; UMBUZEIRO, G. A.; CAMOLESI JR., L. 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 (CNEG/2012), Niterói, Brasil, **Anais...** Niterói: Universidade Federal Fluminense, 2012a.

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**. Reino Unido, 2012b. No prelo.

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: IV WORKSHOP DE COMPUTAÇÃO APLICADA À

GESTÃO DO MEIO AMBIENTE E RECURSOS NATURAIS (WCAMA/2013), Maceió, Brasil, **Anais...** Maceió: Universidade Federal de Alagoas, p. 1036-1046, 2013.

BRUNDTLAND, G.H. (chair.) “Our Common Future” – Report on the World Commission on Environment and Development. New York, United Nations Environmental Programme, 1987.

CETESB. **Relatório de Qualidade das Águas Superficiais do Estado de São Paulo – 2008**. São Paulo: CETESB, 2008. Disponível em: <http://www.cetesb.sp.gov.br/agua/aguas-superficiais/35-publicacoes/-relatorios>. Acesso em: 29 abr. 2013.

CETESB. **Relatório de Qualidade das Águas Superficiais do Estado de São Paulo – 2011**. São Paulo: CETESB, 2011. Disponível em: <http://www.cetesb.sp.gov.br/agua/aguas-superficiais/35-publicacoes/-relatorios>. Acesso em: 29 abr. 2013.

CETESB. **Relatório de Qualidade das Águas Superficiais do Estado de São Paulo – 2012**. São Paulo: CETESB, 2012. Disponível em: <http://www.cetesb.sp.gov.br/agua/aguas-superficiais/35-publicacoes/-relatorios>. Acesso em: 29 abr. 2013.

CETESB. Institucional – CETESB - Companhia Ambiental do Estado de São Paulo – Histórico, 2013. Disponível em: <http://www.cetesb.sp.gov.br/institucional/institucional/52-Histórico>. Acesso em: 29 abr. 2013.

CONAMA. Conselho Nacional do Meio Ambiente. **Resolução n. 357, de 17 de março de 2005**. Brasília, 27 p., 2005.

DINIZ, R. B. N.; SOARES, V. G.; CABRAL, L. A. F. 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, 2012.

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: III WORKSHOP DE COMPUTAÇÃO APLICADA À GESTÃO DO MEIO AMBIENTE E RECURSOS NATURAIS (WCAMA/2011), Natal, Brasil, **Anais...** Natal: Universidade Federal do Rio Grande do Norte, p. 1409-1418, 2011.

ELMASRI, R.; NAVATHE, S. B. Conceitos de Data Mining. In: ELMASRI, Ramez; NAVATHE, Shamkant B. **Sistemas de Banco de Dados**. 4. ed. Pearson, chap. 27, p. 624-645, 2005.

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

FERNANDES, J.; DUARTE, A. S. Um Sistema de Data Warehousing para a Área da Qualidade da Água. Universidade do Minho, Portugal, 2009.

GIBERT, K.; IZQUIERDO, J.; HOLMES, G.; ATHANASIADIS, I.; COMAS, J.; SÀNCHEZ-MARRÈ, M. On the role of pre and post-processing in environmental data mining. In: INTERNATIONAL CONGRESS ON ENVIRONMENTAL MODELLING AND SOFTWARE, Barcelona, Espanha, **Proceedings...** Barcelona: International Environmental Modelling and Software Society, p. 1937-1958, 2008.

GORDON, A. D. A survey of constrained classification. **Computational Statistics & Data Analysis**, v. 21, p. 17-29, 1996.

GUIMARÃES, A. M. **Aplicação de computação evolucionária na mineração de dados físico-químicos da água e do solo**. 145p. Tese (Doutorado) – Faculdade de Ciências Agronômicas, Universidade Estadual Paulista, Botucatu, Brasil, 2005.

HAN, J.; KAMBER, M. **Data Mining – Concepts and Techniques**. 1.ed. New York: Morgan Kaufmann, 550 p., 2000.

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

KARIMIPOUR, F.; DELAVAR, M. R.; KINAIE, M. Water Quality Management Using GIS Data Mining. **Journal of Environmental Informatics**. Canadá, v. 5, n. 2, p. 61-71, 2005.

MAGAIA, L. P. T. **O papel dos sistemas de suporte à decisão na análise da qualidade da água**. Dissertação (Mestrado) –Universidade do Minho, Portugal, 2009.

MARANHÃO, N. **Sistema de Indicadores para Planejamento e Gestão dos Recursos Hídricos de Bacias Hidrográficas**. Tese (Doutorado) – Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brasil, 2007.

MARAVALLE, M.; SIMEONE, B.; NALDINI, R. Clustering on trees. **Computational Statistics & Data Analysis**, v. 24, p. 217-234, 1997.

NEVES, M. C.; CÂMARA, G.; ASSUNÇÃO, R. M.; FREITAS, C. C. Procedimentos Automáticos e Semi-automáticos de Regionalização por Árvore Geradora Mínima. In: BRAZILIAN SYMPOSIUM ON GEOINFORMATICS (GeoInfo/2002), Caxambu, Brasil, **Anais...** p. 109-116, 2002.

OPENSHAW, S.; WYMER, C. Classifying and regionalizing census data. In: OPENSHAW, S., Ed., **Census users' handbook**. Cambridge: GeoInformation International, p. 460, 1995.

PRASS, F. S. KDD: Processo de descoberta de conhecimento em bancos de dados. Grupo de Interesse em Engenharia de Software, Florianópolis, v.1, p. 10-14, 2004.

RABELO, E. **Avaliação de Técnicas de Visualização para Mineração de Dados**. 103p. Dissertação (Mestrado) – Departamento de Informática, Universidade Estadual de Maringá, Maringá, Brasil, 2007.

RAMACHANDRA RAO, A.; SRINIVAS, V. V. Regionalization of watersheds by hybrid-cluster analysis. **Journal Of Hydrology**, v. 318, n.1-4, p. 37 -56, 2006.

SEIXAS, A. J.; NELSON, F. F. E.; BEATRIZ, S. L. P. L. Mining spatial and temporal data to classify water quality: a case study. In: DATA MINING IX: DATA MINING, PROTECTION, DETECTION AND OTHER SECURITY TECHNOLOGIES, Cadiz, Espanha, **Proceedings...** Cadiz: University of Cadiz, v. 40, p. 83-94, 2008.

SHYUE, S.; CHEN, C.; 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, n. 8, p. 3425-3436, 2010.

SILVA, I. A. F. **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, 2007.

SILVA, M. P. S. Mineração de Dados: Conceitos, Aplicações e Experimentos com Weka. In: IV ESCOLA REGIONAL DE INFORMÁTICA - RJ/ES, Vitória - Rio das Ostras, Brasil, nov. 2004.

SOS FUNDAÇÃO MATA ATLÂNTICA. Uma política pública para as águas, 2012. Disponível em: <http://www.rededasaguas.org.br/politicas-publicas/>. Acesso em: 02 mar. 2012.

TAN, P.; STEINBACH, M.; KUMAR, V. **Introdução ao Data Mining – Mineração de Dados**. Rio de Janeiro: Editora Ciência Moderna. 900 p., 2009.

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

VON SPERLING, M. **Estudos e modelagem da qualidade da água de rios**. Belo Horizonte, Brasil: Departamento de Engenharia Sanitária e Ambiental – Universidade Federal de Minas Gerais, v.7, 588 p., 2007.

WISE, S.; HAINING, R.; MA, J. Regionalisation Tools for The Exploratory Spatial Analysis of Health Data. In: FISCHER, M.M.; GETIS, A., Eds., **Recent Developments in Spatial Analysis: Spatial Statistics, Behavioural Modelling, and Computational Intelligence**. Berlin: Springer, p. 83-100, 1997.

WONG, P. C. Visual data mining. **IEEE Computer Graphics and Applications**, Los Alamitos, v.19, no.5, p. 20-21, set./out. 1999.

Apêndice A – Classificação de Toxicidade em Amostras de Água – Resultados e Avaliações

Este Apêndice apresenta os cinco experimentos realizados para classificação de toxicidade em amostras de água. Os quatro primeiros são específicos para cada categoria de parâmetros de qualidade: “Saúde Humana”, “Vida Aquática”, “Indicadores Genéricos” e “Fatores Organolépticos”. O último experimento é genérico e cobre os parâmetros considerados mais significativos nos quatro experimentos anteriores, além da Toxicidade que é atributo alvo. A avaliação e a interpretação dos resultados obtidos é realizada no final do Apêndice.

Experimento 1

A) Regras extraídas a partir da base de treinamento das UGRHIs 2 e 5:

FINAL RESULTS FOR TOXICITY = NT

Rule not found.

FINAL RESULTS FOR TOXICITY = CR

Rule not found.

FINAL RESULTS FOR TOXICITY = AG

Rule not found.

*Default rule found: **toxidd=CR***

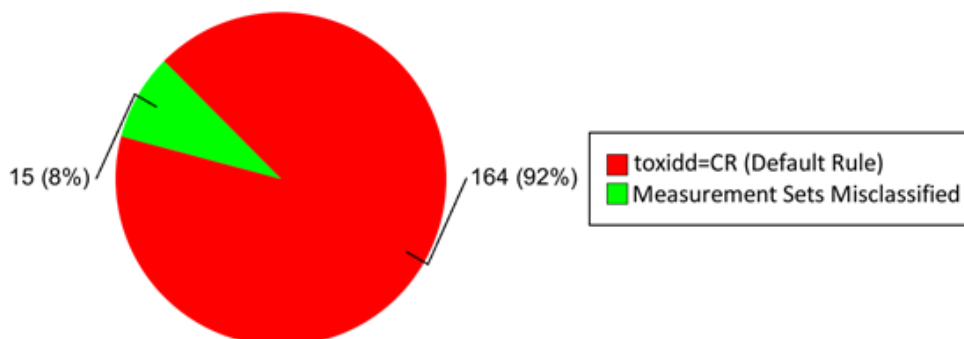
Resultado da aplicação das regras obtidas na base de treinamento das UGRHIs 2 e 5 na base de testes das UGRHIs 6 e 10:

Measurement Sets (Total): 179

Measurement Sets Misclassified: 15

Accuracy: 91,62%

Desempenho das regras aplicadas na base de testes das UGRHIs 6 e 10:



B) Regras extraídas a partir da base de treinamento das UGRHIs 6 e 10:

FINAL RESULTS FOR TOXICITY = NT

Rule not found.

FINAL RESULTS FOR TOXICITY = CR

ni_tot=AC -> toxidd=CR

FINAL RESULTS FOR TOXICITY = AG

Rule not found.

Default rule found: toxidd=CR

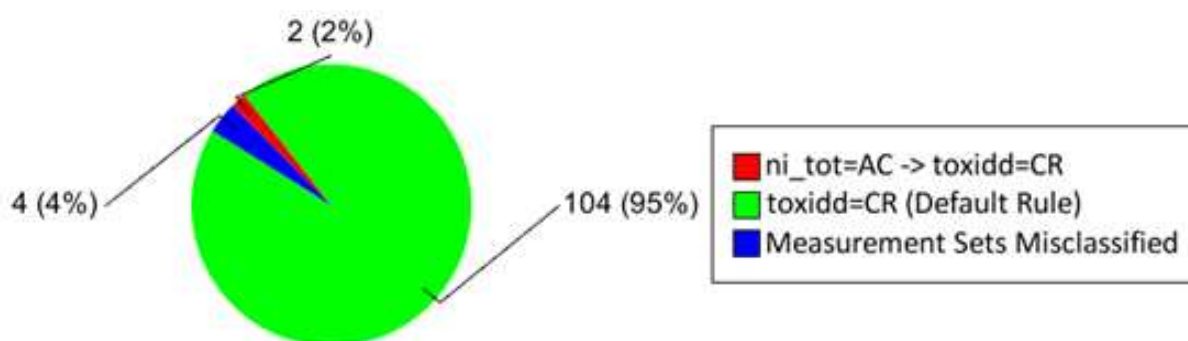
Resultado da aplicação das regras obtidas na base de treinamento das UGRHIs 6 e 10 na base de testes das UGRHIs 2 e 5:

Measurement Sets (Total): 110

Measurement Sets Misclassified: 4

Accuracy: 96,36%

Desempenho das regras aplicadas na base de testes das UGRHIs 2 e 5:



C) Resultado da validação cruzada referente aos dois processamentos:

Measurement Sets (Total): 289

Measurement Sets Misclassified: 19

Accuracy: 93,43%

Experimento 2

A) Regras extraídas a partir da base de treinamento das UGRHIs 2 e 5:

FINAL RESULTS FOR TOXICITY = NT

Rule not found.

FINAL RESULTS FOR TOXICITY = CR

ox_dis=AB -> toxidd=CR

n_amon=AC -> toxidd=CR

FINAL RESULTS FOR TOXICITY = AG

Rule not found.

*Default rule found: **toxidd=CR***

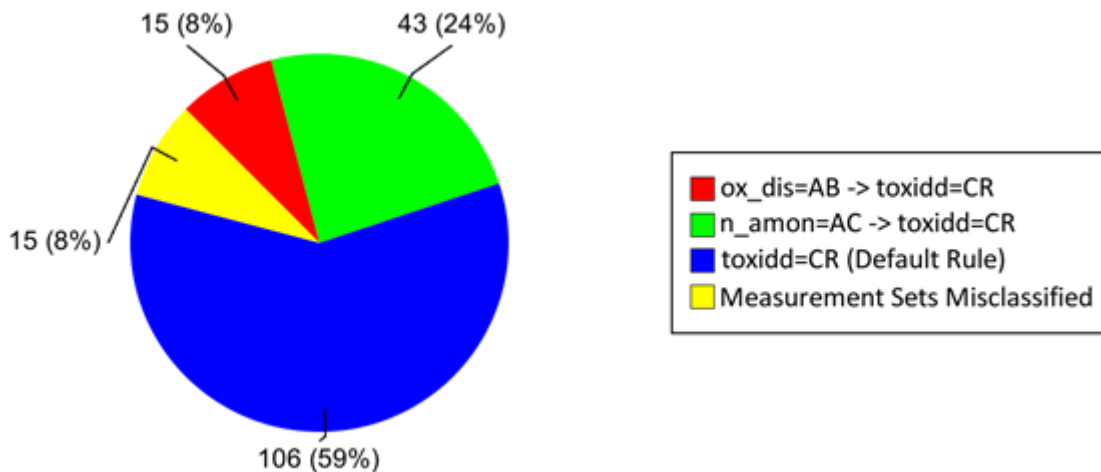
Resultado da aplicação das regras obtidas na base de treinamento das UGRHIs 2 e 5 na base de testes das UGRHIs 6 e 10:

Measurement Sets (Total): 179

Measurement Sets Misclassified: 15

Accuracy: 91,62%

Desempenho das regras aplicadas na base de testes das UGRHIs 6 e 10:



B) Regras extraídas a partir da base de treinamento das UGRHIs 6 e 10:

FINAL RESULTS FOR TOXICITY = NT

Rule not found.

FINAL RESULTS FOR TOXICITY = CR

n_amon=AC -> toxidd=CR

n_amon=MA,ox_dis=AB -> toxidd=CR

n_amon=MA -> toxidd=CR

FINAL RESULTS FOR TOXICITY = AG

Rule not found.

*Default rule found: **toxidd=CR***

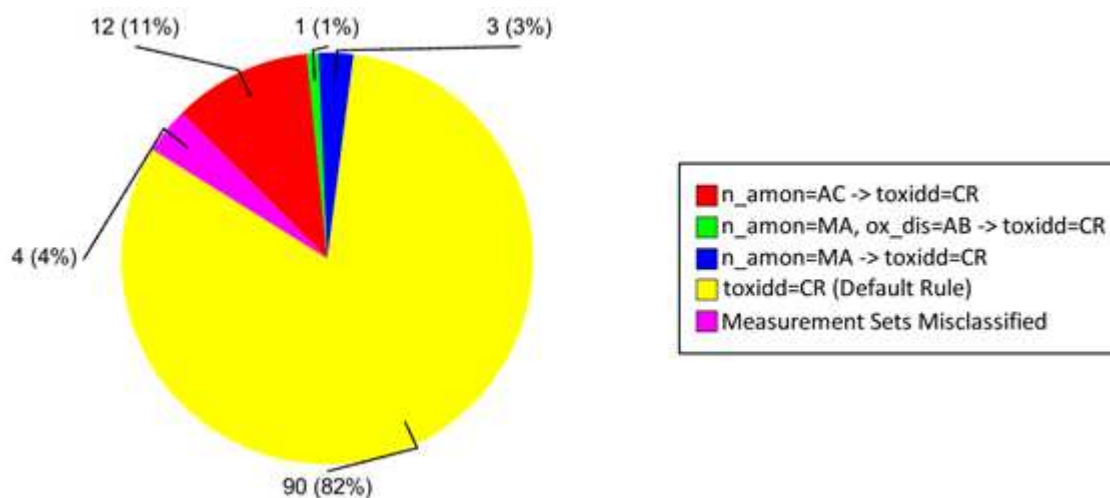
Resultado da aplicação das regras obtidas na base de treinamento das UGRHIs 6 e 10 na base de testes das UGRHIs 2 e 5:

Measurement Sets (Total): 110

Measurement Sets Misclassified: 4

Accuracy: 96,36%

Desempenho das regras aplicadas na base de testes das UGRHIs 2 e 5:



C) Resultado da validação cruzada referente aos dois processamentos:

Measurement Sets (Total): 289

Measurement Sets Misclassified: 19

Accuracy: 93,43%

Experimento 3

A) Regras extraídas a partir da base de treinamento das UGRHIs 2 e 5:

FINAL RESULTS FOR TOXICITY = NT

Rule not found.

FINAL RESULTS FOR TOXICITY = CR

condut=BX,soltot=BX,chuvas=NO,tmp_ag=BX -> toxidd=CR

condut=BX,tmp_ag=MD,soltot=BX,chuvas=NO -> toxidd=CR

condut=BX,chuvas=SI,soltot=BX,tmp_ag=MD -> toxidd=CR

condut=BX,chuvas=SI,soltot=BX,tmp_ag=BX -> toxidd=CR

condut=BX,tmp_ag=MD,soltot=MD,chuvas=SI -> toxidd=CR

condut=BX,tmp_ag=MD -> toxidd=CR

tmp_ag=AT,condut=BX -> toxidd=CR

condut=MD -> toxidd=CR

FINAL RESULTS FOR TOXICITY = AG

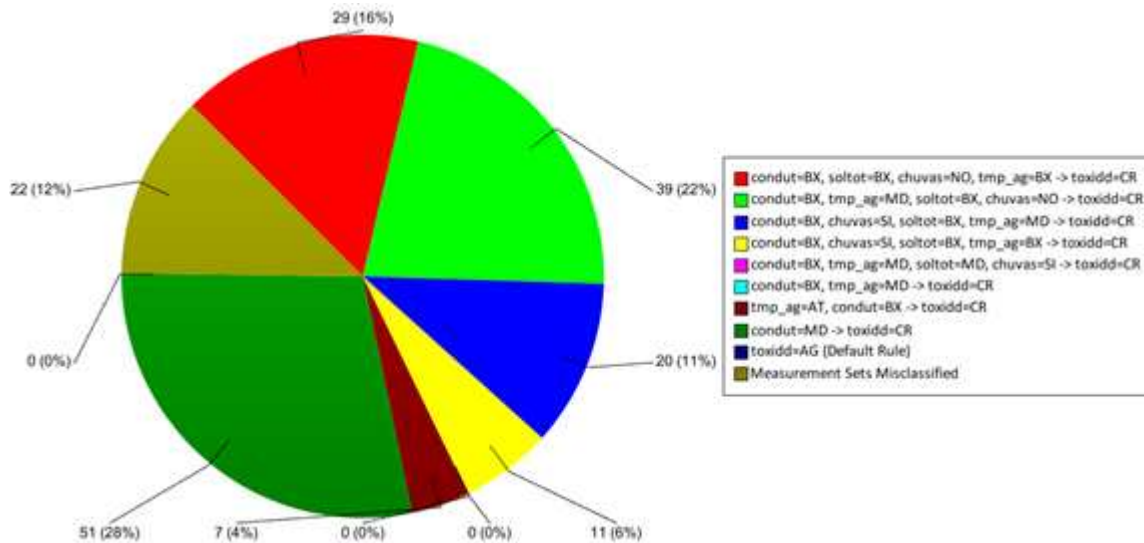
Rule not found.

Default rule found: toxidd=AG

Resultado da aplicação das regras obtidas na base de treinamento das UGRHIs 2 e 5 na base de testes das UGRHIs 6 e 10:

Measurement Sets (Total): 179
 Measurement Sets Misclassified: 22
 Accuracy: 87,71%

Desempenho das regras aplicadas na base de testes das UGRHIs 6 e 10:



B) Regras extraídas a partir da base de treinamento das UGRHIs 6 e 10:

FINAL RESULTS FOR TOXICITY = NT

Rule not found.

FINAL RESULTS FOR TOXICITY = CR

soltot=BX,condut=BX,chuvas=NO,tmp_ag=MD -> toxidd=CR

soltot=BX,condut=BX,tmp_ag=BX,chuvas=NO -> toxidd=CR

soltot=BX,tmp_ag=MD,condut=MD,chuvas=NO -> toxidd=CR

chuvas=SI,soltot=BX,condut=BX,tmp_ag=MD -> toxidd=CR

chuvas=SI,soltot=BX,condut=MD,tmp_ag=MD -> toxidd=CR

FINAL RESULTS FOR TOXICITY = AG

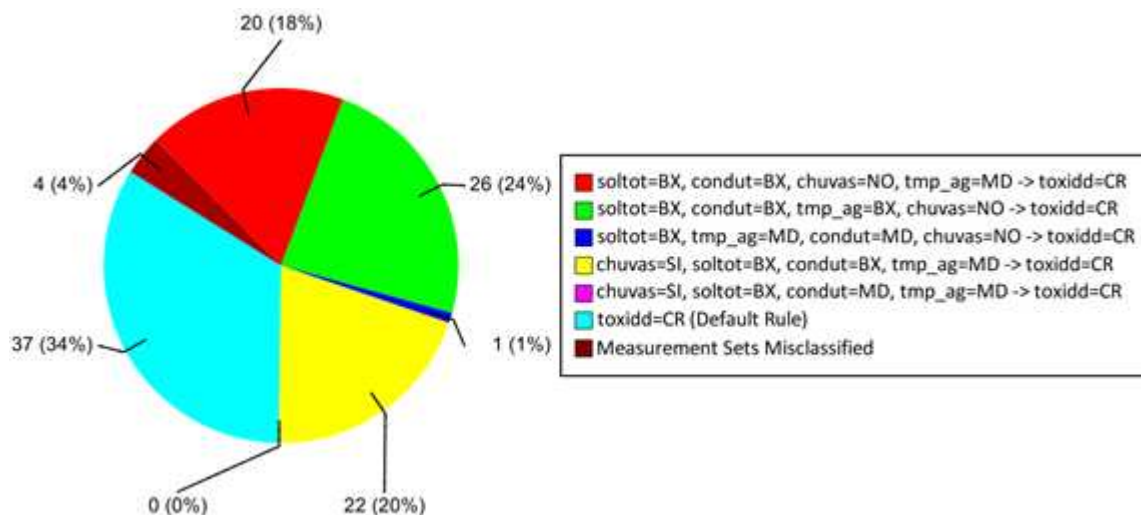
Rule not found.

Default rule found: toxidd=CR

Resultado da aplicação das regras obtidas na base de treinamento das UGRHIs 6 e 10 na base de testes das UGRHIs 2 e 5:

Measurement Sets (Total): 110
 Measurement Sets Misclassified: 4
 Accuracy: 96,36%

Desempenho das regras aplicadas na base de testes das UGRHIs 2 e 5:



C) Resultado da validação cruzada referente aos dois processamentos:

Measurement Sets (Total): 289

Measurement Sets Misclassified: 26

Accuracy: 91,00%

Experimento 4

A) Regras extraídas a partir da base de treinamento das UGRHIs 2 e 5:

FINAL RESULTS FOR TOXICITY = NT

Rule not found.

FINAL RESULTS FOR TOXICITY = CR

fe_dis=AC,mn_tot=AC,al_dis=MA -> toxidd=CR

fe_dis=AC,al_dis=AC,mn_tot=AC -> toxidd=CR

fe_dis=AC,al_dis=AC -> toxidd=CR

fe_dis=AC,al_dis=MA -> toxidd=CR

al_dis=AC -> toxidd=CR

fe_dis=AC -> toxidd=CR

FINAL RESULTS FOR TOXICITY = AG

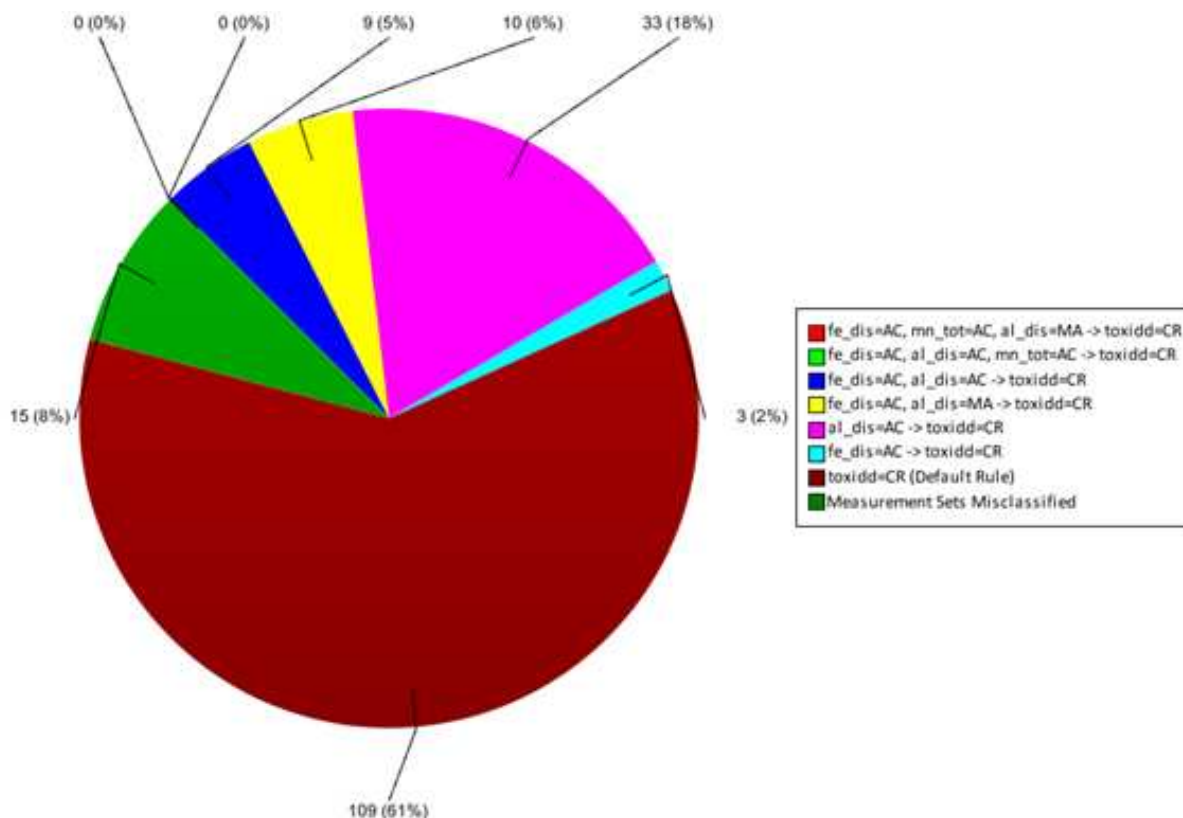
Rule not found.

Default rule found: toxidd=CR

Resultado da aplicação das regras obtidas na base de treinamento das UGRHIs 2 e 5 na base de testes das UGRHIs 6 e 10:

Measurement Sets (Total): 179
 Measurement Sets Misclassified: 15
 Accuracy: 91,62%

Desempenho das regras aplicadas na base de testes das UGRHIs 6 e 10:



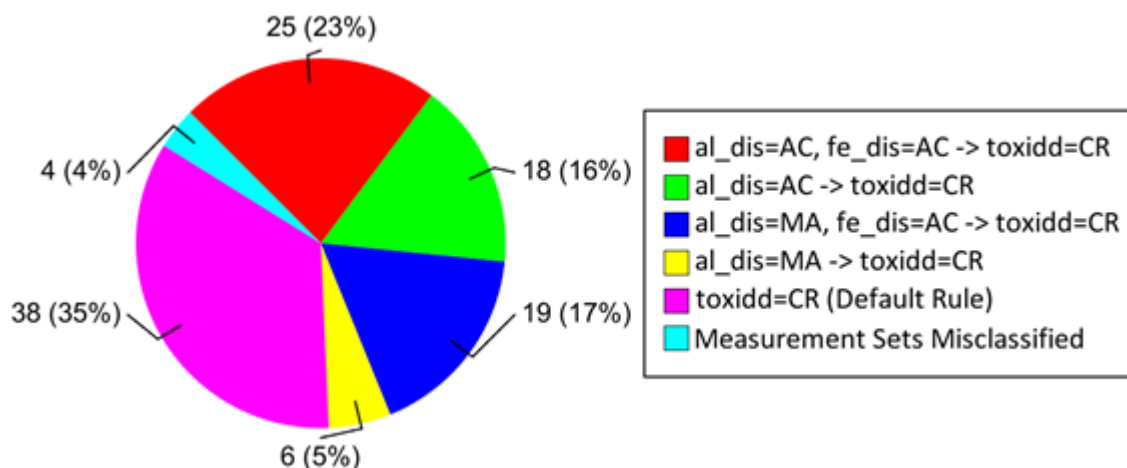
B) Regras extraídas a partir da base de treinamento das UGRHIs 6 e 10:

FINAL RESULTS FOR TOXICITY = NT
 Rule not found.
FINAL RESULTS FOR TOXICITY = CR
al_dis=AC,fe_dis=AC -> toxidd=CR
al_dis=AC -> toxidd=CR
al_dis=MA,fe_dis=AC -> toxidd=CR
al_dis=MA -> toxidd=CR
FINAL RESULTS FOR TOXICITY = AG
 Rule not found.
 Default rule found: **toxidd=CR**

Resultado da aplicação das regras obtidas na base de treinamento das UGRHIs 6 e 10 na base de testes das UGRHIs 2 e 5:

Measurement Sets (Total): 110
 Measurement Sets Misclassified: 4
 Accuracy: 96,36%

Desempenho das regras aplicadas na base de testes das UGRHIs 2 e 5:



C) Resultado da validação cruzada referente aos dois processamentos:

Measurement Sets (Total): 289
 Measurement Sets Misclassified: 19
 Accuracy: 93,43%

Experimento Generalista

Os parâmetros considerados mais significativos nos quatro experimentos anteriores, foram selecionados com base nas ocorrências dos parâmetros nestes experimentos, as quais são apresentadas em ordem decrescente na Tabela 11.

Tabela 11: Ocorrências dos parâmetros nos experimentos específicos de classificação.

Parâmetros	Total de Ocorrências
Condutividade	13
Temperatura Água	12
Chuva 24h	10
Sólidos Totais	10
Alumínio Dissolvido	9
Ferro Dissolvido	7
Nitrogênio Amoniacal	4

Parâmetros	Total de Ocorrências
Oxigênio Dissolvido	2
Manganês Total	2
Níquel Total	1
Cádmio Total	0
Chumbo Total	0
Nitrato	0
Nitrito	0
Cobre Dissolvido	0
Substância Tensoativa	0
Zinco Total	0
pH	0
Turbidez	0

Apesar de mais frequentes, as manifestações dos parâmetros de acompanhamento Condutividade, Sólidos Totais e Temperatura da Água, além do parâmetro Chuva 24h, são pouco conclusivas, pois além de se tratarem de parâmetros sem definição de Padrão CONAMA, apresentam valores de medição preponderantemente baixos ou irregulares nas regras de classificação geradas. Por este motivo, estes parâmetros não foram considerados nesse experimento. Sendo assim, a partir das ocorrências levantadas, inferiu-se que os parâmetros considerados mais significativos para utilização no experimento generalista foram:

- Alumínio Dissolvido
- Ferro Dissolvido
- Nitrogênio Amoniacal
- Oxigênio Dissolvido
- Manganês Total
- Níquel Total

A) Regras extraídas a partir da base de treinamento das UGRHIs 2 e 5:

FINAL RESULTS FOR TOXICITY = NT

Rule not found.

FINAL RESULTS FOR TOXICITY = CR

fe_dis=AC,mn_tot=AC,ox_dis=AB -> toxidd=CR

fe_dis=AC,al_dis=AC -> toxidd=CR

fe_dis=AC,al_dis=MA -> toxidd=CR

al_dis=AC -> toxidd=CR

fe_dis=AC -> toxidd=CR

FINAL RESULTS FOR TOXICITY = AG

Rule not found.

*Default rule found: **toxidd=CR***

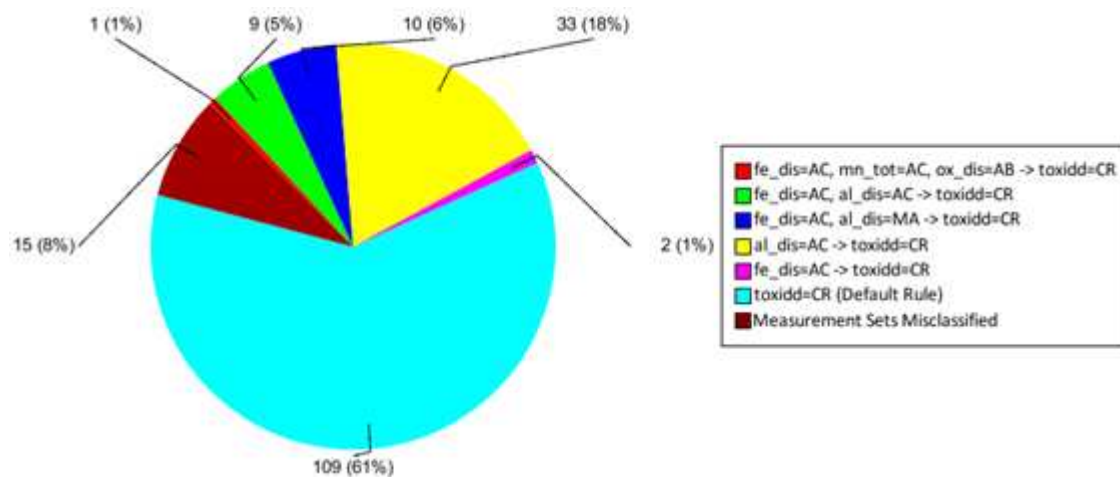
Resultado da aplicação das regras obtidas na base de treinamento das UGRHIs 2 e 5 na base de testes das UGRHIs 6 e 10:

Measurement Sets (Total): 179

Measurement Sets Misclassified: 15

Accuracy: 91,62%

Desempenho das regras aplicadas na base de testes das UGRHIs 6 e 10:



B) Regras extraídas a partir da base de treinamento das UGRHIs 6 e 10:

FINAL RESULTS FOR TOXICITY = NT

Rule not found.

FINAL RESULTS FOR TOXICITY = CR

al_dis=AC, n_amon=MA -> toxidd=CR

n_amon=AC, fe_dis=AC, al_dis=MA -> toxidd=CR

n_amon=AC, al_dis=AC -> toxidd=CR

n_amon=AC -> toxidd=CR

n_amon=MA -> toxidd=CR

FINAL RESULTS FOR TOXICITY = AG

Rule not found.

*Default rule found: **toxidd=CR***

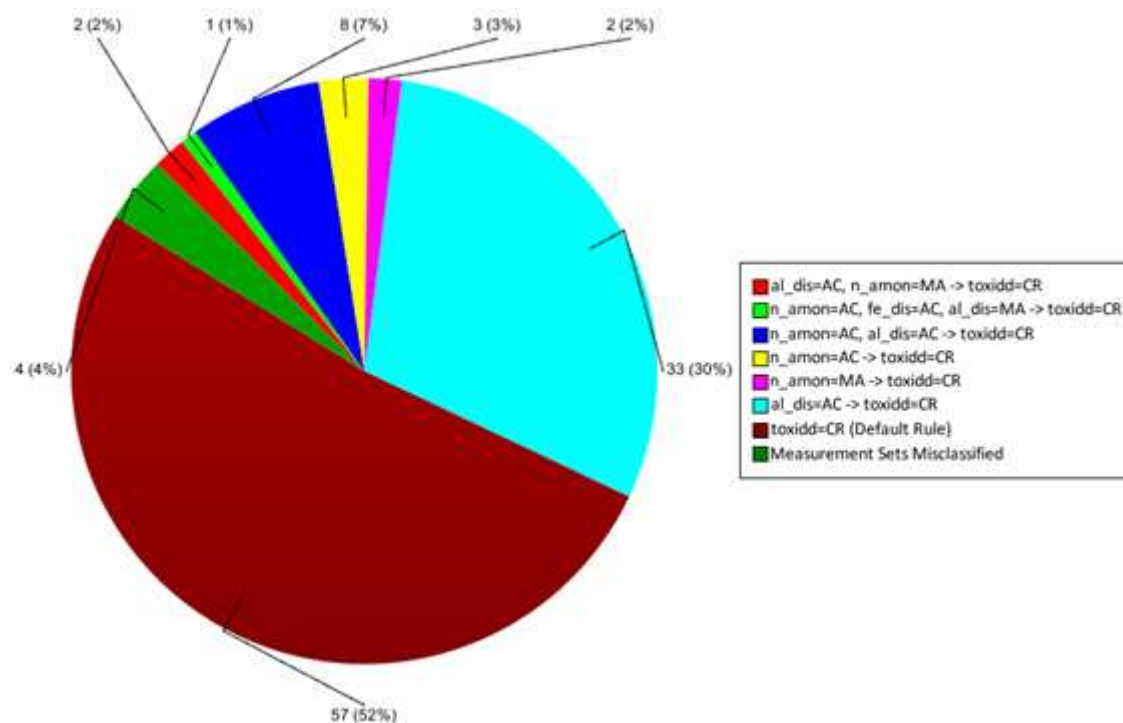
Resultado da aplicação das regras obtidas na base de treinamento das UGRHIs 6 e 10 na base de testes das UGRHIs 2 e 5:

Measurement Sets (Total): 110

Measurement Sets Misclassified: 4

Accuracy: 96,36%

Desempenho das regras aplicadas na base de testes das UGRHIs 2 e 5:



C) Resultado da validação cruzada referente aos dois processamentos:

Measurement Sets (Total): 289

Measurement Sets Misclassified: 19

Accuracy: 93,43%

Avaliação e Interpretação

A Tabela 12 apresenta a quantidade de ocorrências de cada parâmetro nos antecedentes das regras para Toxicidade “Crônica” geradas no experimento generalista.

Tabela 12: Ocorrências dos parâmetros nos antecedentes das regras geradas.

Ocorrências para toxidd=CR	
al_dis	6
fe_dis	5
n_amon	5
ox_dis	1
mn_tot	1

Com base nesses resultados, pode-se inferir que os parâmetros mais associados à Toxicidade “Crônica” são: Alumínio Dissolvido, Ferro Dissolvido, Nitrogênio Amoniacal, e em menor proporção, Oxigênio Dissolvido e Manganês Total. Devido à escassez de medições contendo Toxicidade “Aguda”, a ferramenta não gerou nenhuma regra de classificação para este valor.

Pôde-se verificar que o índice de acerto na aplicação das regras de classificação geradas ficou acima de 90% em todos os experimentos, o que denota um bom desempenho dos conjuntos de regras gerados para a predição de Toxicidade “Crônica” em amostras de água.

Apêndice B – Identificação de Associações entre os Parâmetros de Qualidade de Água – Resultados e Avaliações

Este Apêndice apresenta os sete experimentos para identificação de associações entre parâmetros de qualidade de água. Os seis primeiros são específicos para cada possível dupla de categorias de parâmetros de qualidade, conforme descrito a seguir:

- **Experimento 1** – “Saúde Humana” e “Vida Aquática”.
- **Experimento 2** – “Saúde Humana” e “Indicadores Genéricos”.
- **Experimento 3** – “Saúde Humana” e “Fatores Organolépticos”.
- **Experimento 4** – “Vida Aquática” e “Indicadores Genéricos”.
- **Experimento 5** – “Vida Aquática” e “Fatores Organolépticos”.
- **Experimento 6** – “Indicadores Genéricos” e “Fatores Organolépticos”.

O último experimento é genérico e cobre os parâmetros considerados mais significativos nos seis experimentos anteriores. A avaliação e a interpretação dos resultados apresentados nos experimentos é realizada no final do Apêndice.

Experimento 1

Regras extraídas:

Strong rules with 3 parameters:

ox_dis=MA,sub_te=AC -> n_amon=MA - 90.9%

Quantity of strong rules of 3 parameters: 1

Experimento 2

Regras extraídas:

Strong rules with 2 parameters:

pot_hi=AC -> soltot=BX - 92.0%

cd_tot=MA -> condut=BX - 100.0%

Quantity of strong rules of 2 parameters: 2

Strong rules with 3 parameters:

condut=BX,chuvas=NO -> soltot=BX - 92.6%
condut=BX,tmp_ag=BX -> soltot=BX - 93.3%
condut=BX,tmp_ag=AT -> soltot=BX - 90.0%
soltot=BX,ni_tot=AC -> condut=BX - 96.3%
soltot=BX,ni_tot=AC -> chuvas=NO - 92.6%
condut=BX,soltot=AT -> chuvas=SI - 90.9%
condut=BX,pot_hi=AC -> soltot=BX - 100.0%
tmp_ag=MD,pot_hi=AC -> soltot=BX - 100.0%
soltot=BX,pb_tot=MA -> condut=BX - 90.9%

Quantity of strong rules of 3 parameters: 9

Strong rules with 4 parameters:

condut=BX,chuvas=NO,tmp_ag=MD -> soltot=BX - 90.9%
condut=BX,chuvas=NO,tmp_ag=BX -> soltot=BX - 94.8%
condut=BX,chuvas=NO,ni_tot=AC -> soltot=BX - 96.2%
soltot=BX,chuvas=NO,ni_tot=AC -> condut=BX - 100.0%
soltot=BX,condut=BX,ni_tot=AC -> chuvas=NO - 96.2%
soltot=BX,ni_tot=AC -> condut=BX,chuvas=NO - 92.6%
soltot=BX,tmp_ag=MD,ni_tot=AC -> condut=BX - 95.0%
soltot=BX,tmp_ag=MD,ni_tot=AC -> chuvas=NO - 95.0%
chuvas=NO,tmp_ag=BX,condut=AT -> soltot=MD - 90.5%
condut=BX,chuvas=SI,tmp_ag=AT -> soltot=BX - 100.0%
condut=BX,chuvas=NO,pot_hi=AC -> soltot=BX - 100.0%
chuvas=NO,tmp_ag=MD,pot_hi=AC -> soltot=BX - 100.0%
condut=BX,tmp_ag=MD,pot_hi=AC -> soltot=BX - 100.0%

Quantity of strong rules of 4 parameters: 13

Strong rules with 5 parameters:

condut=BX,chuvas=NO,tmp_ag=MD,ni_tot=AC -> soltot=BX - 100.0%
soltot=BX,chuvas=NO,tmp_ag=MD,ni_tot=AC -> condut=BX - 100.0%
soltot=BX,condut=BX,tmp_ag=MD,ni_tot=AC -> chuvas=NO - 100.0%
soltot=BX,tmp_ag=MD,ni_tot=AC -> condut=BX,chuvas=NO - 95.0%

Quantity of strong rules of 5 parameters: 4

Experimento 3

Regras extraídas:

Strong rules with 3 parameters:

fe_dis=AC,turbid=AC -> al_dis=MA - 84.8%
fe_dis=AC,turbid=MA -> al_dis=MA - 84.6%

Quantity of strong rules of 3 parameters: 2

Strong rules with 4 parameters:
fe_dis=AC,mn_tot=AC,turbid=AC -> al_dis=MA - 85.7%
Quantity of strong rules of 4 parameters: 1

Experimento 4

Regras extraídas:

Strong rules with 2 parameters:
condut=AT -> n_amon=MA - 95.3%
toxidd=AG -> soltot=BX - 100.0%
Quantity of strong rules of 2 parameters: 2

Strong rules with 3 parameters:
n_amon=AC,toxidd=CR -> soltot=BX - 96.1%
soltot=MD,condut=AT -> n_amon=MA - 96.1%
chuvas=NO,condut=AT -> n_amon=MA - 95.7%
tmp_ag=BX,condut=AT -> n_amon=MA - 100.0%
condut=BX,pot_hi=AC -> soltot=BX - 100.0%
tmp_ag=MD,pot_hi=AC -> soltot=BX - 100.0%
condut=BX,toxidd=AG -> soltot=BX - 100.0%
condut=AT,sub_te=AC -> n_amon=MA - 100.0%
tmp_ag=MD,toxidd=AG -> soltot=BX - 100.0%
Quantity of strong rules of 3 parameters: 9

Strong rules with 4 parameters:
condut=BX,tmp_ag=BX,toxidd=CR -> soltot=BX - 98.5%
condut=BX,chuvas=NO,n_amon=MA -> soltot=BX - 96.2%
chuvas=NO,soltot=MD,condut=AT -> n_amon=MA - 97.4%
chuvas=NO,n_amon=AC,toxidd=CR -> soltot=BX - 97.0%
condut=BX,n_amon=AC,toxidd=CR -> soltot=BX - 96.8%
tmp_ag=BX,soltot=MD,condut=AT -> n_amon=MA - 100.0%
chuvas=NO,tmp_ag=MD,condut=AT -> n_amon=MA - 95.8%
chuvas=NO,tmp_ag=BX,condut=AT -> n_amon=MA - 100.0%
condut=BX,chuvas=SI,tmp_ag=AT -> soltot=BX - 100.0%
condut=BX,n_amon=MA,toxidd=CR -> soltot=BX - 100.0%
condut=BX,n_amon=MA,toxidd=CR -> chuvas=NO - 100.0%
chuvas=NO,ox_dis=MA,condut=AT -> n_amon=MA - 100.0%
n_amon=MA,ox_dis=AB,condut=AT -> soltot=MD - 100.0%
condut=BX,chuvas=NO,pot_hi=AC -> soltot=BX - 100.0%
chuvas=NO,n_amon=MA,pot_hi=AC -> soltot=BX - 100.0%
condut=BX,toxidd=CR,pot_hi=AC -> soltot=BX - 100.0%
chuvas=NO,tmp_ag=MD,pot_hi=AC -> soltot=BX - 100.0%
tmp_ag=MD,n_amon=MA,pot_hi=AC -> soltot=BX - 100.0%

condut=BX,tmp_ag=MD,pot_hi=AC -> soltot=BX - 100.0%
tmp_ag=MD,toxidd=CR,pot_hi=AC -> soltot=BX - 100.0%
condut=MD,ox_dis=AB,sub_te=AC -> chuvas=NO - 100.0%
Quantity of strong rules of 4 parameters: 21

Strong rules with 5 parameters:

condut=BX,chuvas=NO,tmp_ag=BX,toxidd=CR -> soltot=BX - 98.0%
soltot=BX,chuvas=NO,tmp_ag=BX,toxidd=CR -> condut=BX - 96.2%
condut=BX,chuvas=NO,n_amon=MA,tmp_ag=BX -> soltot=BX - 95.8%
condut=BX,chuvas=NO,n_amon=AC,toxidd=CR -> soltot=BX - 100.0%
chuvas=NO,tmp_ag=MD,n_amon=AC,toxidd=CR -> soltot=BX - 100.0%
chuvas=NO,tmp_ag=BX,soltot=MD,condut=AT -> n_amon=MA - 100.0%
condut=BX,chuvas=SI,tmp_ag=BX,toxidd=CR -> soltot=BX - 100.0%
condut=BX,chuvas=NO,n_amon=MA,toxidd=CR -> soltot=BX - 100.0%
soltot=BX,condut=BX,n_amon=MA,toxidd=CR -> chuvas=NO - 100.0%
condut=BX,n_amon=MA,toxidd=CR -> soltot=BX,chuvas=NO - 100.0%
chuvas=NO,soltot=MD,ox_dis=MA,condut=AT -> n_amon=MA - 100.0%
condut=BX,tmp_ag=BX,n_amon=AC,toxidd=CR -> soltot=BX - 100.0%
chuvas=NO,tmp_ag=MD,n_amon=MA,pot_hi=AC -> soltot=BX - 100.0%
chuvas=NO,tmp_ag=BX,n_amon=AC,toxidd=CR -> soltot=BX - 100.0%
Quantity of strong rules of 5 parameters: 14

Strong rules with 6 parameters:

condut=BX,chuvas=NO,tmp_ag=MD,n_amon=AC,toxidd=CR -> soltot=BX - 100.0%
Quantity of strong rules of 6 parameters: 1

Experimento 5

Regras extraídas:

Strong rules with 3 parameters:

fe_dis=AC,turbid=AC -> al_dis=MA - 84.8%
al_dis=AC,turbid=AC -> ox_dis=AB - 80.0%
fe_dis=AC,turbid=MA -> al_dis=MA - 84.6%
ox_dis=MA,sub_te=AC -> n_amon=MA - 90.9%
Quantity of strong rules of 3 parameters: 4

Strong rules with 4 parameters:

fe_dis=AC,mn_tot=AC,turbid=AC -> al_dis=MA - 85.7%
Quantity of strong rules of 4 parameters: 1

Experimento 6

Regras extraídas:

Strong rules with 3 parameters:

fe_dis=AC,turbid=AC -> conduat=BX - 97.8%
soltot=BX,fe_dis=MA -> conduat=BX - 95.7%
al_dis=MA,tmp_ag=AT -> fe_dis=AC - 95.0%
al_dis=MA,fe_dis=MA -> conduat=BX - 100.0%
conduat=BX,pot_hi=AC -> soltot=BX - 100.0%
soltot=BX,turbid=AC -> conduat=BX - 100.0%
tmp_ag=MD,pot_hi=AC -> soltot=BX - 100.0%
fe_dis=AC,turbid=MA -> chuvas=SI - 100.0%
Quantity of strong rules of 3 parameters: 8

Strong rules with 4 parameters:

tmp_ag=MD,al_dis=MA,turbid=AC -> conduat=BX - 95.2%
tmp_ag=MD,fe_dis=AC,turbid=AC -> conduat=BX - 100.0%
fe_dis=AC,al_dis=MA,turbid=AC -> conduat=BX - 97.4%
chuvas=SI,fe_dis=AC,turbid=AC -> conduat=BX - 100.0%
fe_dis=AC,soltot=MD,turbid=AC -> conduat=BX - 97.0%
fe_dis=AC,mn_tot=AC,turbid=AC -> conduat=BX - 96.4%
soltot=BX,tmp_ag=MD,fe_dis=MA -> conduat=BX - 95.0%
conduat=BX,chuvas=SI,tmp_ag=AT -> soltot=BX - 100.0%
tmp_ag=MD,al_dis=MA,fe_dis=MA -> conduat=BX - 100.0%
soltot=BX,al_dis=MA,tmp_ag=AT -> fe_dis=AC - 100.0%
conduat=BX,fe_dis=AC,soltot=AT -> chuvas=SI - 100.0%
conduat=BX,soltot=AT,turbid=MA -> chuvas=SI - 100.0%
soltot=MD,al_dis=MA,mn_tot=MA -> tmp_ag=MD - 100.0%
conduat=BX,chuvas=NO,pot_hi=AC -> soltot=BX - 100.0%
soltot=BX,tmp_ag=MD,turbid=AC -> conduat=BX - 100.0%
soltot=BX,chuvas=SI,fe_dis=MA -> conduat=BX - 100.0%
conduat=BX,al_dis=MA,soltot=AT -> chuvas=SI - 100.0%
mn_tot=AC,al_dis=MA,fe_dis=MA -> conduat=BX - 100.0%
chuvas=SI,al_dis=MA,tmp_ag=AT -> fe_dis=AC - 100.0%
soltot=BX,fe_dis=AC,turbid=AC -> conduat=BX - 100.0%
soltot=BX,al_dis=MA,turbid=AC -> conduat=BX - 100.0%
chuvas=NO,tmp_ag=MD,pot_hi=AC -> soltot=BX - 100.0%
conduat=BX,fe_dis=AC,turbid=MA -> chuvas=SI - 100.0%
conduat=BX,mn_tot=MA,soltot=AT -> chuvas=SI - 100.0%
conduat=BX,al_dis=MA,tmp_ag=AT -> fe_dis=AC - 100.0%
fe_dis=AC,al_dis=MA,turbid=MA -> chuvas=SI - 100.0%
fe_dis=AC,soltot=AT,turbid=MA -> chuvas=SI - 100.0%
al_dis=MA,soltot=AT,turbid=MA -> chuvas=SI - 100.0%
conduat=BX,tmp_ag=MD,pot_hi=AC -> soltot=BX - 100.0%
soltot=BX,chuvas=SI,turbid=AC -> conduat=BX - 100.0%
conduat=BX,al_dis=AC,fe_dis=MA -> soltot=BX - 100.0%

soltot=BX,al_dis=AC,fe_dis=MA -> condu=BX - 100.0%
condu=BX,mn_tot=AC,soltot=AT -> tmp_ag=MD - 100.0%
chuv=SI,al_dis=MA,fe_dis=MA -> condu=BX - 100.0%
condu=BX,mn_tot=MA,turbid=MA -> chuv=SI - 100.0%
tmp_ag=MD,fe_dis=AC,turbid=MA -> chuv=SI - 100.0%
mn_tot=AC,al_dis=MA,tmp_ag=AT -> fe_dis=AC - 100.0%
 Quantity of strong rules of 4 parameters: 37

Strong rules with 5 parameters:

condu=BX,tmp_ag=MD,chuv=NO,fe_dis=AC -> soltot=BX - 141.5%
soltot=BX,tmp_ag=MD,chuv=NO,fe_dis=AC -> condu=BX - 141.5%
condu=BX,chuv=NO,fe_dis=AC,tmp_ag=BX -> soltot=BX - 95.2%
tmp_ag=MD,fe_dis=AC,al_dis=MA,turbid=AC -> condu=BX - 100.0%
soltot=BX,fe_dis=AC,mn_tot=AC,al_dis=MA -> condu=BX - 97.0%
tmp_ag=MD,chuv=SI,fe_dis=AC,turbid=AC -> condu=BX - 100.0%
tmp_ag=MD,fe_dis=AC,soltot=MD,turbid=AC -> condu=BX - 100.0%
chuv=SI,fe_dis=AC,al_dis=MA,turbid=AC -> condu=BX - 100.0%
fe_dis=AC,soltot=MD,al_dis=MA,turbid=AC -> condu=BX - 96.3%
condu=BX,chuv=SI,soltot=MD,al_dis=MA -> tmp_ag=MD - 96.0%
tmp_ag=MD,mn_tot=AC,al_dis=MA,turbid=AC -> condu=BX - 95.8%
fe_dis=AC,mn_tot=AC,al_dis=MA,turbid=AC -> condu=BX - 95.8%
tmp_ag=MD,fe_dis=AC,mn_tot=AC,turbid=AC -> condu=BX - 100.0%
chuv=SI,fe_dis=AC,soltot=MD,turbid=AC -> condu=BX - 100.0%
chuv=SI,fe_dis=AC,soltot=MD,al_dis=MA -> tmp_ag=MD - 95.7%
mn_tot=AC,soltot=MD,al_dis=MA,turbid=AC -> condu=BX - 95.5%
chuv=SI,soltot=MD,al_dis=MA,turbid=AC -> tmp_ag=MD - 95.5%
soltot=BX,chuv=NO,mn_tot=AC,al_dis=MA -> condu=BX - 95.0%
chuv=SI,mn_tot=AC,al_dis=MA,turbid=AC -> condu=BX - 95.0%
fe_dis=AC,mn_tot=AC,soltot=MD,turbid=AC -> condu=BX - 95.0%
chuv=SI,fe_dis=AC,mn_tot=AC,turbid=AC -> condu=BX - 100.0%
condu=BX,mn_tot=AC,al_dis=AC,tmp_ag=BX -> soltot=BX - 100.0%
tmp_ag=MD,chuv=SI,fe_dis=AC,mn_tot=MA -> condu=BX - 100.0%
chuv=NO,soltot=MD,al_dis=MA,turbid=AC -> condu=BX - 100.0%
chuv=SI,fe_dis=AC,al_dis=MA,mn_tot=MA -> condu=BX - 100.0%
condu=BX,tmp_ag=MD,soltot=AT,turbid=MA -> chuv=SI - 100.0%
condu=BX,soltot=MD,al_dis=MA,mn_tot=MA -> tmp_ag=MD - 100.0%
soltot=MD,al_dis=MA,mn_tot=MA,turbid=AC -> tmp_ag=MD - 100.0%
soltot=BX,tmp_ag=MD,chuv=SI,fe_dis=MA -> condu=BX - 100.0%
condu=BX,chuv=SI,fe_dis=AC,tmp_ag=AT -> soltot=BX - 100.0%
condu=BX,al_dis=MA,mn_tot=MA,turbid=AC -> tmp_ag=MD - 100.0%
soltot=BX,chuv=SI,al_dis=MA,tmp_ag=AT -> fe_dis=AC - 100.0%
fe_dis=AC,mn_tot=AC,al_dis=AC,tmp_ag=BX -> soltot=BX - 100.0%
condu=BX,tmp_ag=MD,al_dis=MA,soltot=AT -> chuv=SI - 100.0%
tmp_ag=MD,fe_dis=AC,soltot=MD,mn_tot=MA -> condu=BX - 100.0%
condu=BX,fe_dis=AC,soltot=MD,mn_tot=MA -> tmp_ag=MD - 100.0%
tmp_ag=MD,fe_dis=AC,mn_tot=MA,turbid=AC -> condu=BX - 100.0%

condut=BX,fe_dis=AC,mn_tot=MA,turbid=AC -> tmp_ag=MD - 100.0%
tmp_ag=MD,mn_tot=AC,al_dis=MA,fe_dis=MA -> condut=BX - 100.0%
condut=BX,fe_dis=AC,al_dis=MA,soltot=AT -> chuvas=SI - 100.0%
chuvas=SI,soltot=MD,al_dis=MA,mn_tot=MA -> tmp_ag=MD - 100.0%
chuvas=SI,al_dis=MA,mn_tot=MA,turbid=AC -> tmp_ag=MD - 100.0%
 Quantity of strong rules of 5 parameters: 42

Strong rules with 6 parameters:

condut=BX,tmp_ag=MD,fe_dis=AC,chuvas=NO,mn_tot=AC -> soltot=BX - 218.8%
soltot=BX,tmp_ag=MD,fe_dis=AC,chuvas=NO,mn_tot=AC -> condut=BX - 218.8%
soltot=BX,condut=BX,fe_dis=AC,chuvas=NO,mn_tot=AC -> tmp_ag=MD - 218.8%
soltot=BX,tmp_ag=MD,chuvas=SI,fe_dis=AC,al_dis=AC -> condut=BX - 96.0%
soltot=BX,tmp_ag=MD,chuvas=SI,fe_dis=AC,mn_tot=AC -> condut=BX - 95.8%
tmp_ag=MD,chuvas=SI,fe_dis=AC,al_dis=MA,turbid=AC -> condut=BX - 100.0%
tmp_ag=MD,fe_dis=AC,soltot=MD,al_dis=MA,turbid=AC -> condut=BX - 100.0%
soltot=BX,tmp_ag=MD,fe_dis=AC,mn_tot=AC,al_dis=MA -> condut=BX - 95.7%
condut=BX,chuvas=SI,fe_dis=AC,soltot=MD,al_dis=MA -> tmp_ag=MD - 95.2%
tmp_ag=MD,chuvas=SI,fe_dis=AC,soltot=MD,turbid=AC -> condut=BX - 100.0%
condut=BX,chuvas=SI,soltot=MD,al_dis=MA,turbid=AC -> tmp_ag=MD - 95.0%
tmp_ag=MD,fe_dis=AC,mn_tot=AC,al_dis=MA,turbid=AC -> condut=BX - 100.0%
chuvas=SI,fe_dis=AC,soltot=MD,al_dis=MA,turbid=AC -> condut=BX - 100.0%
tmp_ag=MD,chuvas=SI,fe_dis=AC,mn_tot=AC,turbid=AC -> condut=BX - 100.0%
tmp_ag=MD,chuvas=SI,mn_tot=AC,al_dis=MA,turbid=AC -> condut=BX - 100.0%
tmp_ag=MD,mn_tot=AC,soltot=MD,al_dis=MA,turbid=AC -> condut=BX - 100.0%
chuvas=SI,fe_dis=AC,mn_tot=AC,al_dis=MA,turbid=AC -> condut=BX - 100.0%
condut=BX,chuvas=SI,mn_tot=AC,soltot=MD,al_dis=MA -> tmp_ag=MD - 100.0%
tmp_ag=MD,fe_dis=AC,mn_tot=AC,soltot=MD,turbid=AC -> condut=BX - 100.0%
chuvas=SI,fe_dis=AC,mn_tot=AC,soltot=MD,turbid=AC -> condut=BX - 100.0%
chuvas=SI,mn_tot=AC,soltot=MD,al_dis=MA,turbid=AC -> condut=BX - 100.0%
tmp_ag=MD,chuvas=NO,soltot=MD,al_dis=MA,turbid=AC -> condut=BX - 100.0%
condut=BX,soltot=MD,al_dis=MA,mn_tot=MA,turbid=AC -> tmp_ag=MD - 100.0%
 Quantity of strong rules of 6 parameters: 23

Strong rules with 7 parameters:

tmp_ag=MD,chuvas=SI,fe_dis=AC,soltot=MD,al_dis=MA,turbid=AC -> condut=BX - 100.0%
tmp_ag=MD,chuvas=SI,fe_dis=AC,mn_tot=AC,al_dis=MA,turbid=AC -> condut=BX - 100.0%
tmp_ag=MD,fe_dis=AC,mn_tot=AC,soltot=MD,al_dis=MA,turbid=AC -> condut=BX - 100.0%
tmp_ag=MD,chuvas=SI,fe_dis=AC,mn_tot=AC,soltot=MD,turbid=AC -> condut=BX - 100.0%
tmp_ag=MD,chuvas=SI,mn_tot=AC,soltot=MD,al_dis=MA,turbid=AC -> condut=BX - 100.0%
 Quantity of strong rules of 7 parameters: 5

Experimento Generalista

Os parâmetros considerados mais significativos nos seis experimentos anteriores, foram selecionados com base nas ocorrências dos parâmetros nestes experimentos, as quais são apresentadas em ordem decrescente na Tabela 13.

Tabela 13: Ocorrências dos parâmetros nos experimentos específicos de associação.

Parâmetros	Total de Ocorrências
Condutividade	154
Sólidos Totais	140
Chuva 24h	108
Temperatura Água	107
Ferro Dissolvido	80
Alumínio Dissolvido	69
Turbidez	68
Manganês Total	51
Nitrogênio Amoniacal	33
Toxicidade	22
pH	21
Níquel Total	12
Oxigênio Dissolvido	7
Substância Tensoativa	4
Cádmio Total	1
Chumbo Total	1
Nitrato	0
Nitrito	0
Cobre Dissolvido	0
Zinco Total	0

Da mesma forma como ocorreu no experimento de classificação, apesar de mais frequentes, as manifestações dos parâmetros de acompanhamento Condutividade, Sólidos Totais e Temperatura da Água, além do parâmetro Chuva 24h, são pouco conclusivas. Por este motivo, não foram considerados neste experimento. Com isso, a partir das ocorrências levantadas, inferiu-se que os 10 parâmetros considerados mais significativos para utilização no experimento generalista foram:

- Ferro Dissolvido
- Alumínio Dissolvido
- Turbidez

- Manganês Total
- Nitrogênio Amoniacal
- Toxicidade
- pH
- Níquel Total
- Oxigênio Dissolvido
- Substância Tensoativa

Regras extraídas:

Strong rules with 3 parameters:

fe_dis=AC,turbid=AC -> al_dis=MA - 84.8%

al_dis=AC,turbid=AC -> ox_dis=AB - 80.0%

fe_dis=AC,turbid=MA -> al_dis=MA - 84.6%

ox_dis=MA,sub_te=AC -> n_amon=MA - 90.9%

Quantity of strong rules of 3 parameters: 4

Strong rules with 4 parameters:

fe_dis=AC,mn_tot=AC,turbid=AC -> al_dis=MA - 85.7%

Quantity of strong rules of 4 parameters: 1

Avaliação e Interpretação

A partir dos resultados do experimento generalista, pode-se inferir três interessantes relações 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 estas relações parecem caminhar juntos, pois nas regras de associação encontradas tendem a aparecer reunidos e dissonantes quanto ao Padrão CONAMA.

A Tabela 14 mostra as combinações entre as categorias dos parâmetros realizadas nos experimentos, o número de regras geradas a partir destas combinações e a confiança considerada na extração das regras. 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, mesmo com a limitação imposta pela taxa mínima de confiança definida, originaram um número maior de regras.

Por outro lado, as combinações de categorias dos experimentos 1, 3 e 5 parecem representar relações mais fracas, pois mesmo com uma taxa mínima de confiança mais baixa, poucas regras foram encontradas.

Tabela 14: Comparativo das combinações entre as categorias de parâmetros.

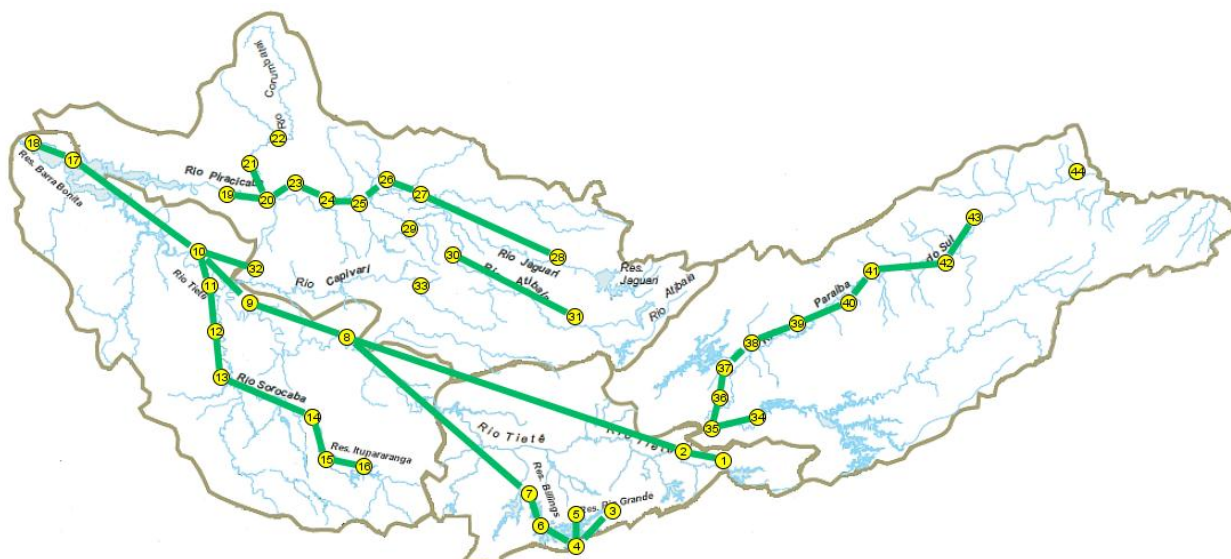
Experimentos	Combinações entre Categorias	Regras geradas	Confiança mínima considerada
1	Saúde Humana e Vida Aquática	1	80%
2	Saúde Humana e Indicadores Genéricos	29	90%
3	Saúde Humana e Fatores Organolépticos	3	80%
4	Vida Aquática e Indicadores Genéricos	47	95%
5	Vida Aquática e Fatores Organolépticos	5	80%
6	Indicadores Genéricos e Fatores Organolépticos	115	95%

Também é interessante observar que, das três relações citadas, duas contém apenas parâmetros de uma mesma categoria, Alumínio Dissolvido, Turbidez e Ferro Dissolvido pertencem todos à categoria “Fatores Organolépticos”, e Nitrogênio Amoniacal, Substância Tensoativa e Oxigênio Dissolvido à categoria “Vida Aquática”. Isto denota que não existem fortes relações entre parâmetros de categorias distintas.

Apêndice C – Regionalização de Pontos de Amostragem de Água – Resultados e Avaliações

Este Apêndice apresenta os quatro grupos de cinco experimentos para regionalização dos pontos de amostragem de água. Cada grupo contemplou uma categoria de parâmetros de qualidade: “Saúde Humana”, “Vida Aquática”, “Indicadores Genéricos” e “Fatores Organolépticos”. Para cada um destes grupos, foram feitos cinco experimentos, quatro contemplando apenas os três meses de cada estação do ano e um considerando os 12 meses do ano. A avaliação e a interpretação dos resultados apresentados nos experimentos é realizada no final do Apêndice.

Experimento 2 – Estação do Ano = Verão



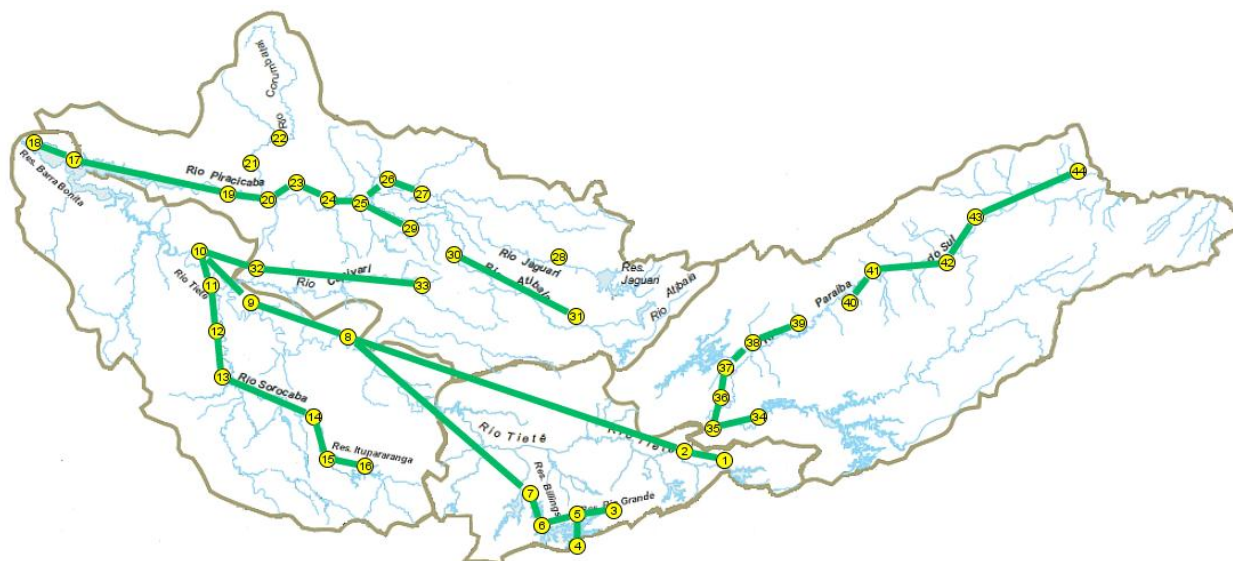
Grupos de pontos de amostragem de água gerados para as UGRHs 5, 6 e 10:

Grupos	Pontos de Amostragem - UGRHs 5, 6 e 10																																
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33
1																																	
2																																	
3																																	
4																																	
5																																	
6																																	

Grupos de pontos de amostragem de água gerados para a UGRHs 2:

Grupos	Pontos de Amostragem – UGRH 2										
	34	35	36	37	38	39	40	41	42	43	44
1											
2											

Experimento 3 – Estação do Ano = Outono



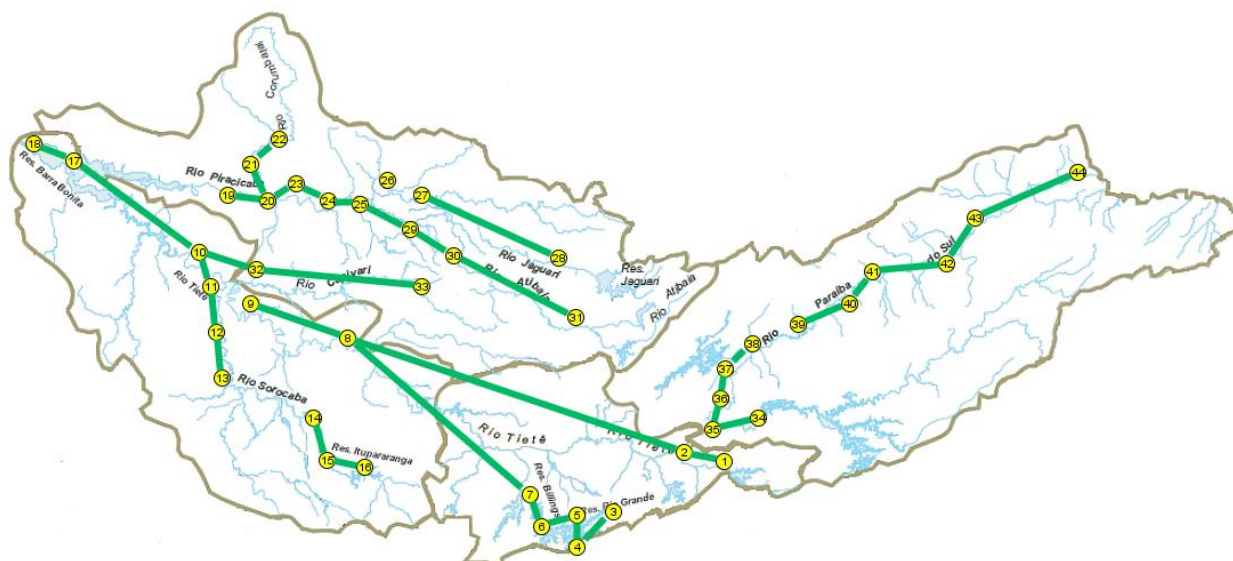
Grupos de pontos de amostragem de água gerados para as UGRHIs 5, 6 e 10:

[illegible]

Grupos de pontos de amostragem de água gerados para a UGRHIs 2:

[illegible]

Experimento 4 – Estação do Ano = Inverno



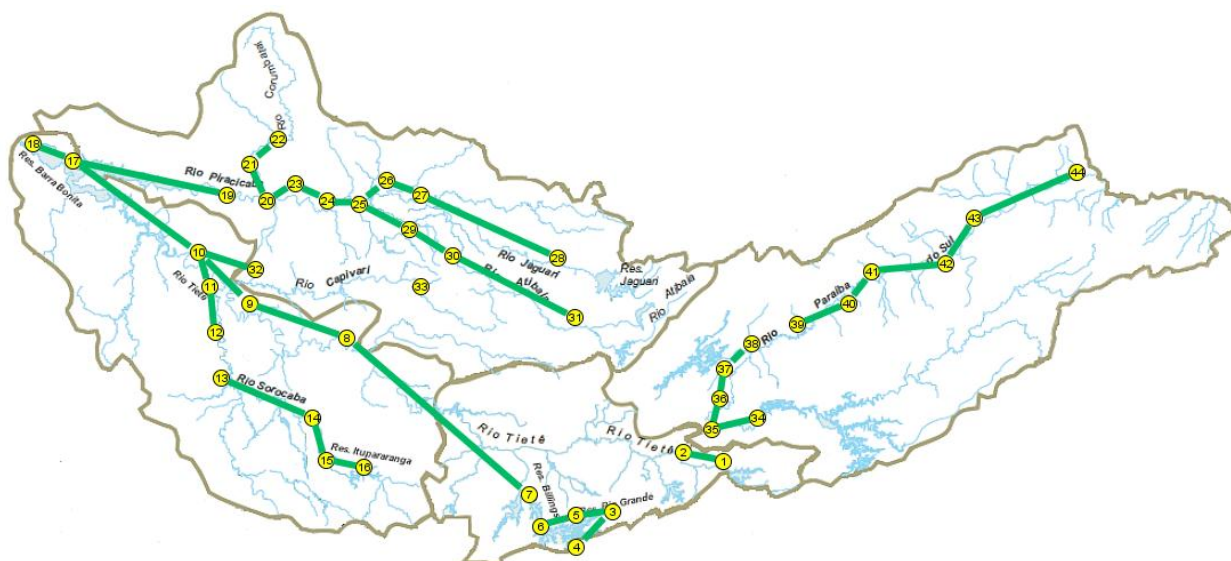
Grupos de pontos de amostragem de água gerados para as UGRHIs 5, 6 e 10:

Grupos	Pontos de Amostragem - UGRHIs 5, 6 e 10																																	
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	
1																																		
2																																		
3																																		
4																																		
5																																		
6																																		

Grupos de pontos de amostragem de água gerados para a UGRHIs 2:

[illegible]

Experimento 5 – Ano Completo



Grupos de pontos de amostragem de água gerados para as UGRHIs 5, 6 e 10:

[illegible]

Grupos de pontos de amostragem de água gerados para a UGRHs 2:

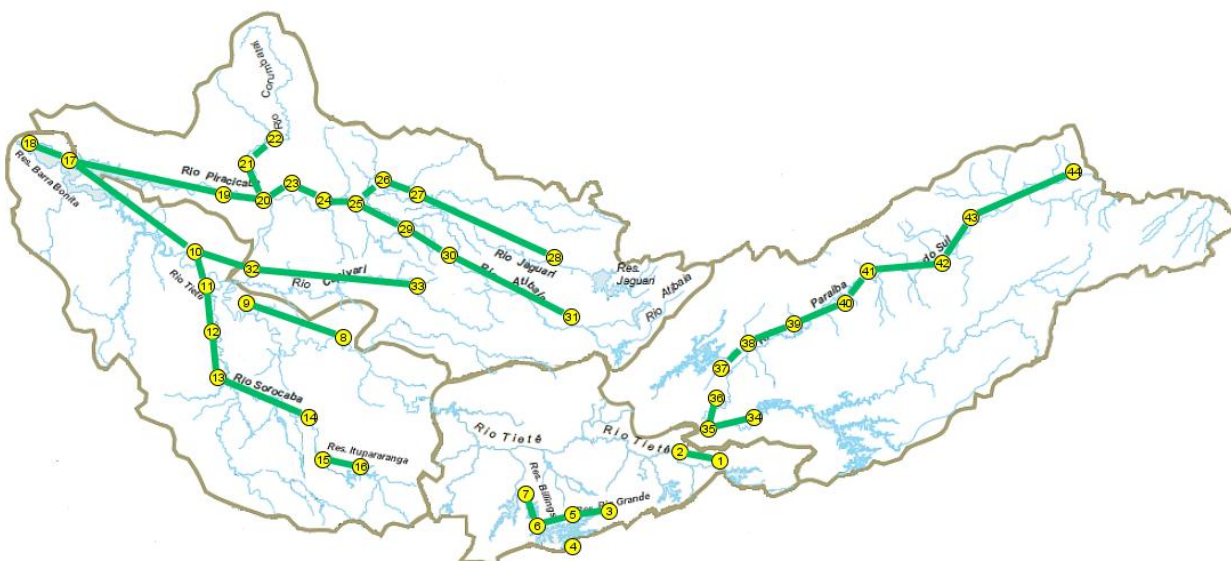
[illegible]

Experimentos 6 a 10

Parâmetros considerados:

Cobre Dissolvido, Nitrogênio Amoniacal, Oxigênio Dissolvido, Substância Tensoativa, Zinco Total, Toxicidade

Experimento 6 – Estação do Ano = Primavera



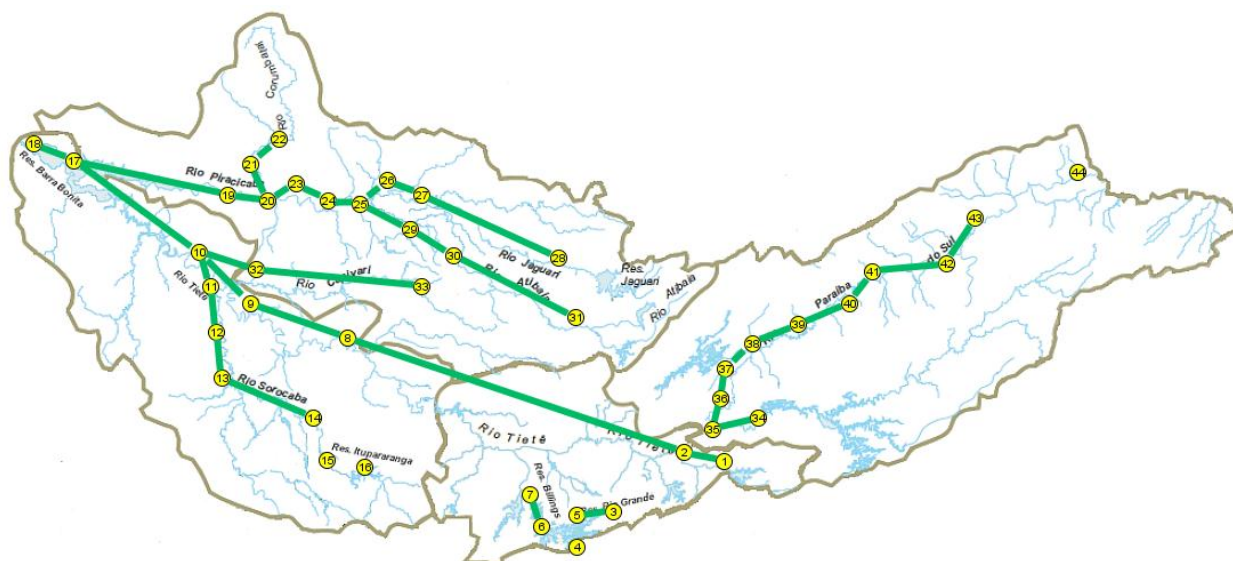
Grupos de pontos de amostragem de água gerados para as UGRHIs 5, 6 e 10:

[illegible]

Grupos de pontos de amostragem de água gerados para a UGRHs 2:

[illegible]

Experimento 7 – Estação do Ano = Verão



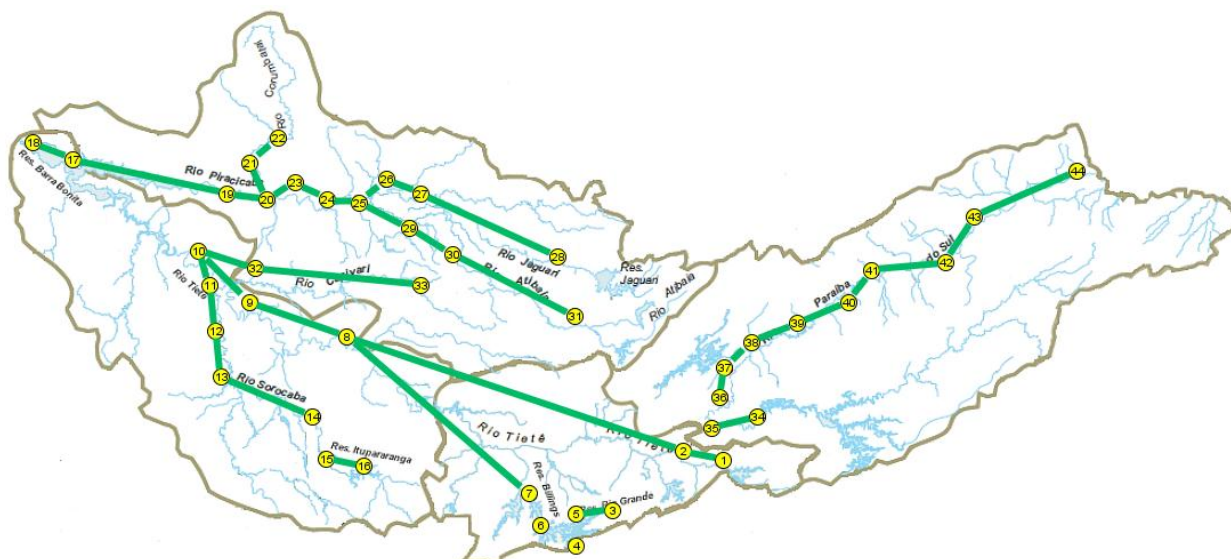
Grupos de pontos de amostragem de água gerados para as UGRHIs 5, 6 e 10:

[illegible]

Grupos de pontos de amostragem de água gerados para a UGRHIs 2:

[illegible]

Experimento 8 – Estação do Ano = Outono



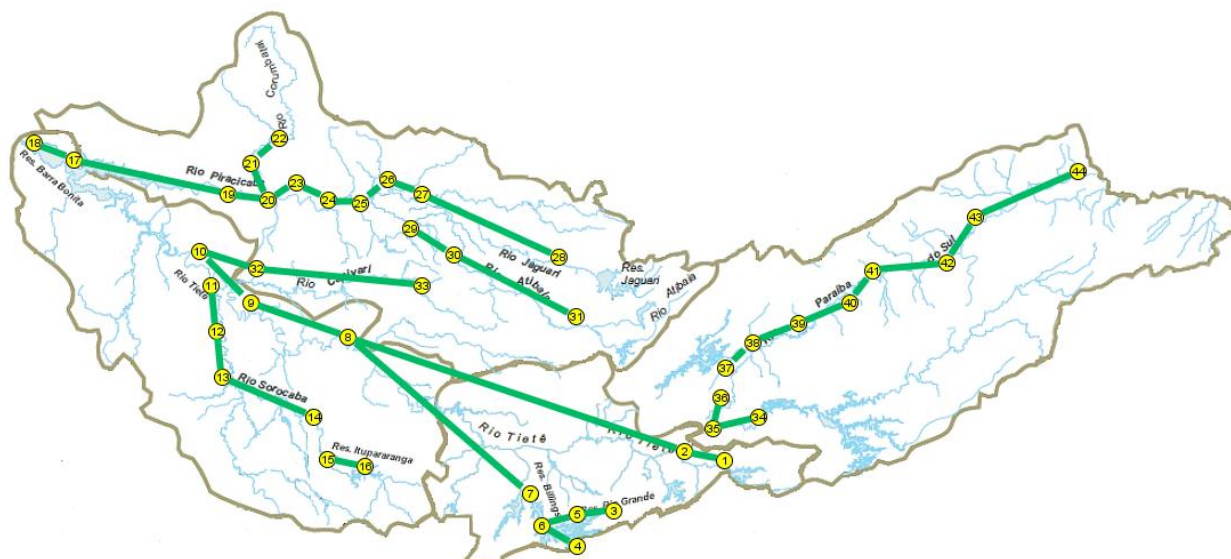
Grupos de pontos de amostragem de água gerados para as UGRHs 5, 6 e 10:

[illegible]

Grupos de pontos de amostragem de água gerados para a UGRHIs 2:

[illegible]

Experimento 9 – Estação do Ano = Inverno



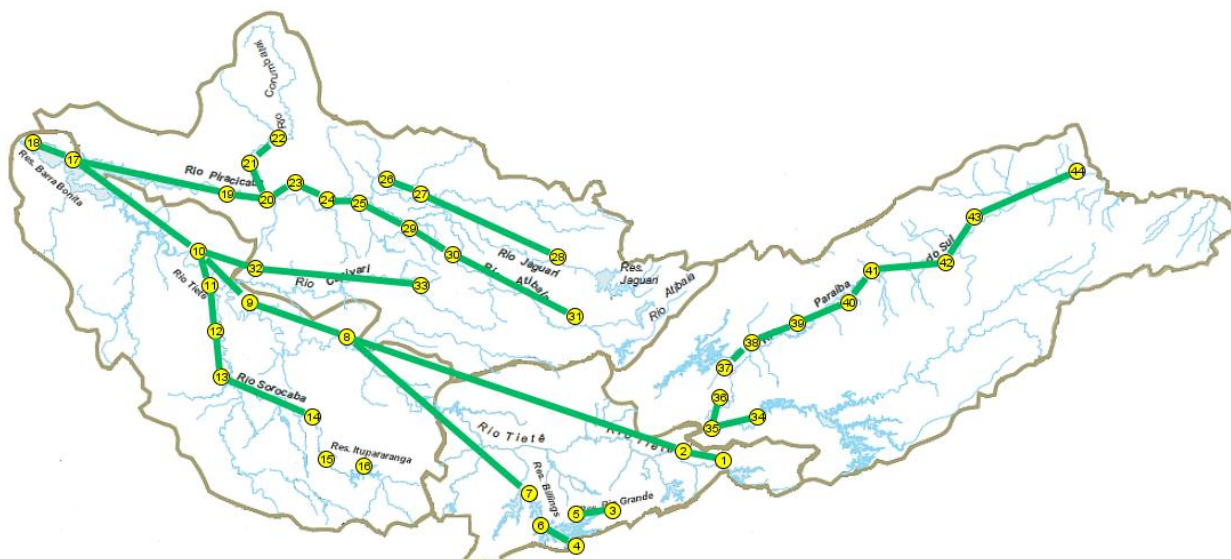
Grupos de pontos de amostragem de água gerados para as UGRHs 5, 6 e 10:

Grupos	Pontos de Amostragem - UGRHs 5, 6 e 10																																
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33
1																																	
2																																	
3																																	
4																																	
5																																	
6																																	

Grupos de pontos de amostragem de água gerados para a UGRHs 2:

Grupos	Pontos de Amostragem – UGRHI 2												
	34	35	36	37	38	39	40	41	42	43	44		
1													
2													

Experimento 10 – Ano Completo



Grupos de pontos de amostragem de água gerados para as UGRHIs 5, 6 e 10:

Grupos	Pontos de Amostragem - UGRHIs 5, 6 e 10																																	
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	
1	■	■				■	■	■	■	■	■	■	■	■		■	■	■	■	■	■	■	■	■	■				■	■	■	■	■	
2			■		■																													
3				■		■																												
4															■																			
5																■																		
6																											■	■	■					

Grupos de pontos de amostragem de água gerados para a UGRHs 2:

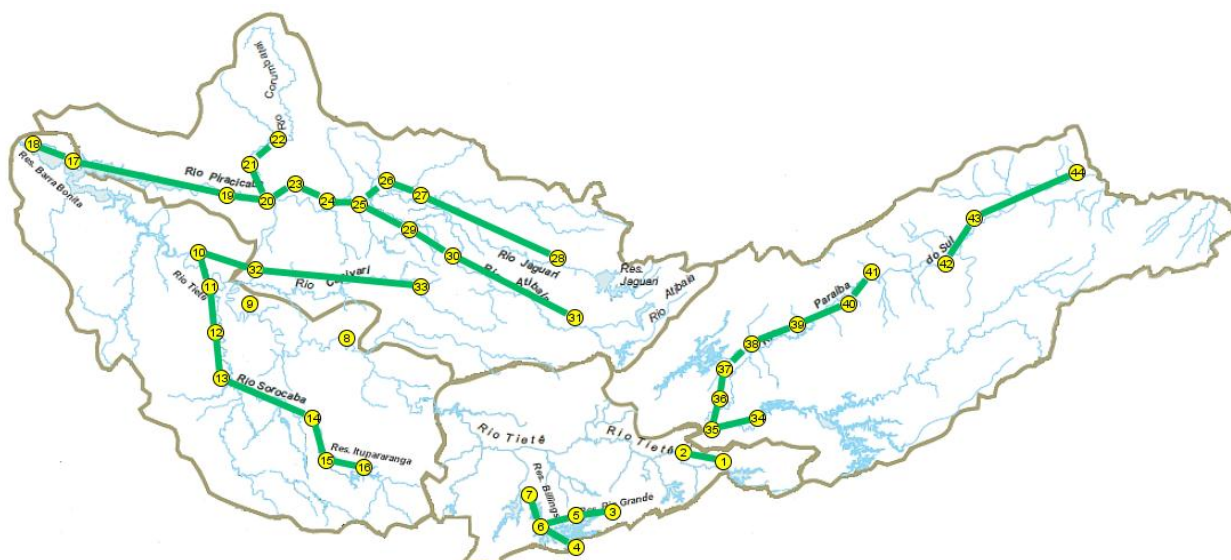
[illegible]

Experimentos 11 a 15

Parâmetros considerados:

Chuva 24h, Cloreto Total, Condutividade, pH, Sólidos Totais, Temperatura Água

Experimento 11 – Estação do Ano = Primavera



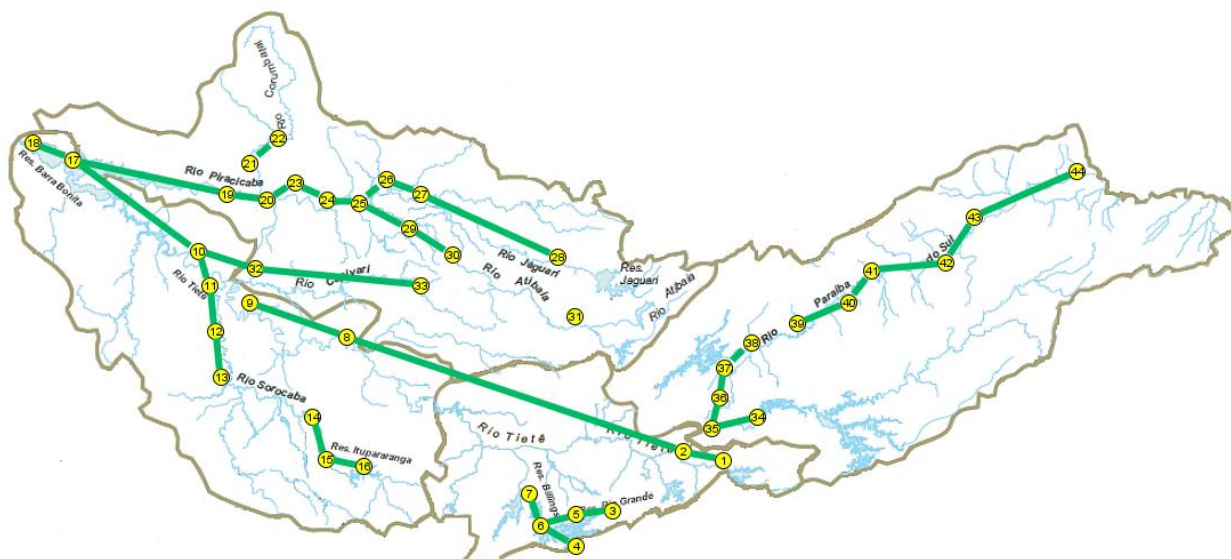
Grupos de pontos de amostragem de água gerados para as UGRHIs 5, 6 e 10:

[illegible]

Grupos de pontos de amostragem de água gerados para a UGRHIs 2:

[illegible]

Experimento 12 – Estação do Ano = Verão



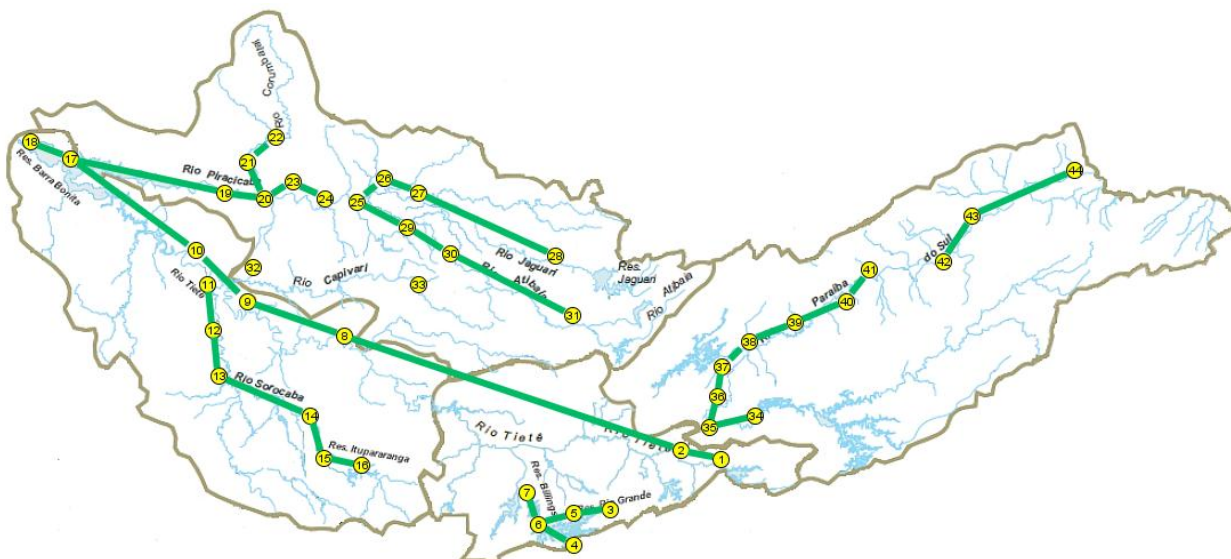
Grupos de pontos de amostragem de água gerados para as UGRHIs 5, 6 e 10:

[illegible]

Grupos de pontos de amostragem de água gerados para a UGRHIs 2:

[illegible]

Experimento 13 – Estação do Ano = Outono



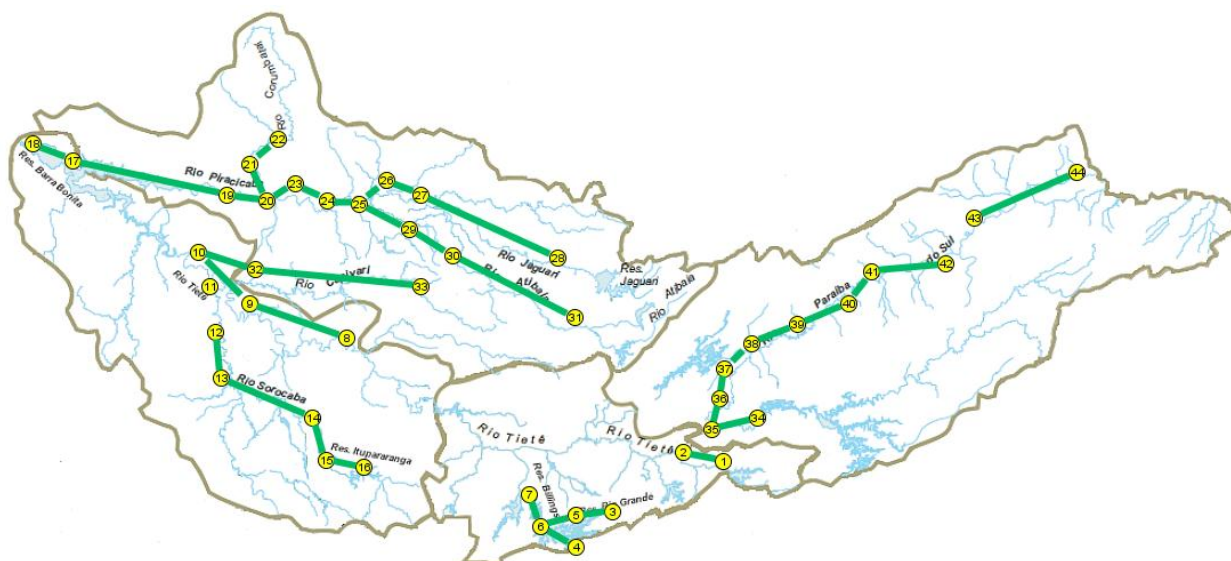
Grupos de pontos de amostragem de água gerados para as UGRHs 5, 6 e 10:

[illegible]

Grupos de pontos de amostragem de água gerados para a UGRHIs 2:

[illegible]

Experimento 14 – Estação do Ano = Inverno



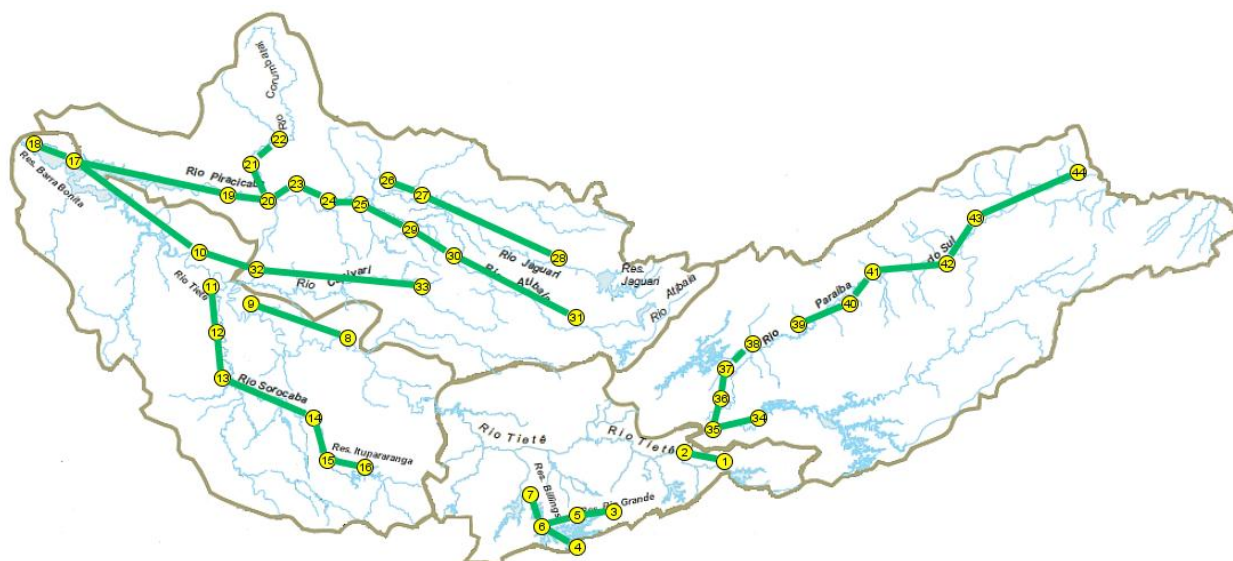
Grupos de pontos de amostragem de água gerados para as UGRHs 5, 6 e 10:

Grupos	Pontos de Amostragem - UGRHs 5, 6 e 10																																
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33
1																																	
2																																	
3																																	
4																																	
5																																	
6																																	

Grupos de pontos de amostragem de água gerados para a UGRHs 2:

Grupos	Pontos de Amostragem – UGRH 2										
	34	35	36	37	38	39	40	41	42	43	44
1											
2											

Experimento 15 – Ano Completo



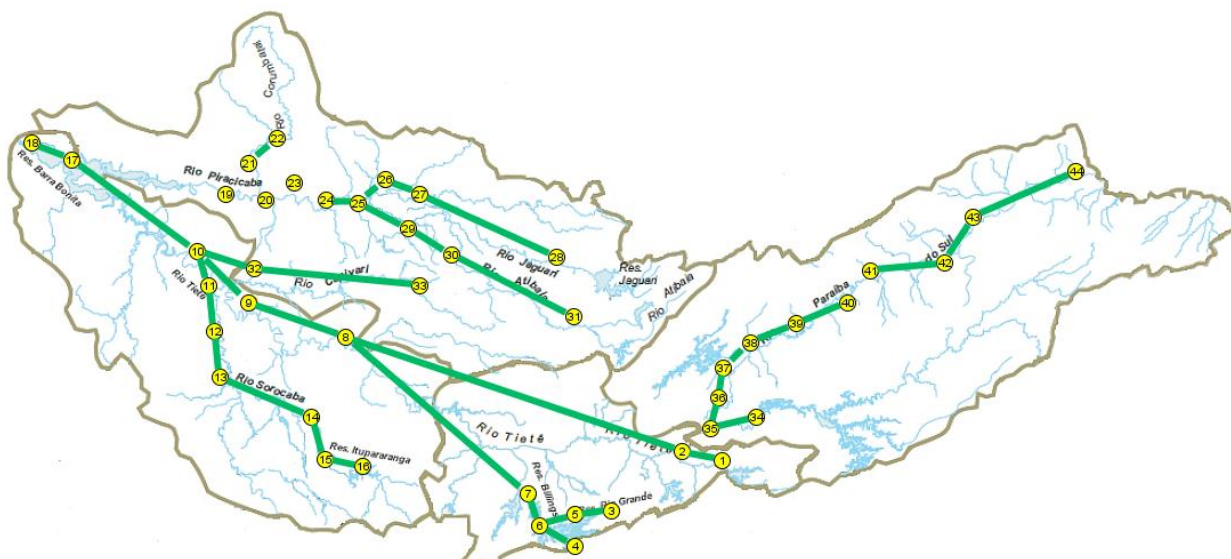
Grupos de pontos de amostragem de água gerados para as UGRHs 5, 6 e 10:

Grupos	Pontos de Amostragem - UGRHs 5, 6 e 10																																	
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	
1	■	■																																
2			■	■	■	■	■	■																										
3									■	■																								
4											■							■	■	■	■	■	■	■	■	■			■	■	■	■	■	
5												■	■	■	■	■																		
6																											■	■	■					

Grupos de pontos de amostragem de água gerados para a UGRHIs 2:

[illegible]

Experimento 17 – Estação do Ano = Verão



Grupos de pontos de amostragem de água gerados para as UGRHIs 5, 6 e 10:

[illegible]

Grupos de pontos de amostragem de água gerados para a UGRHIs 2:

[illegible]

Experimento 18 – Estação do Ano = Outono



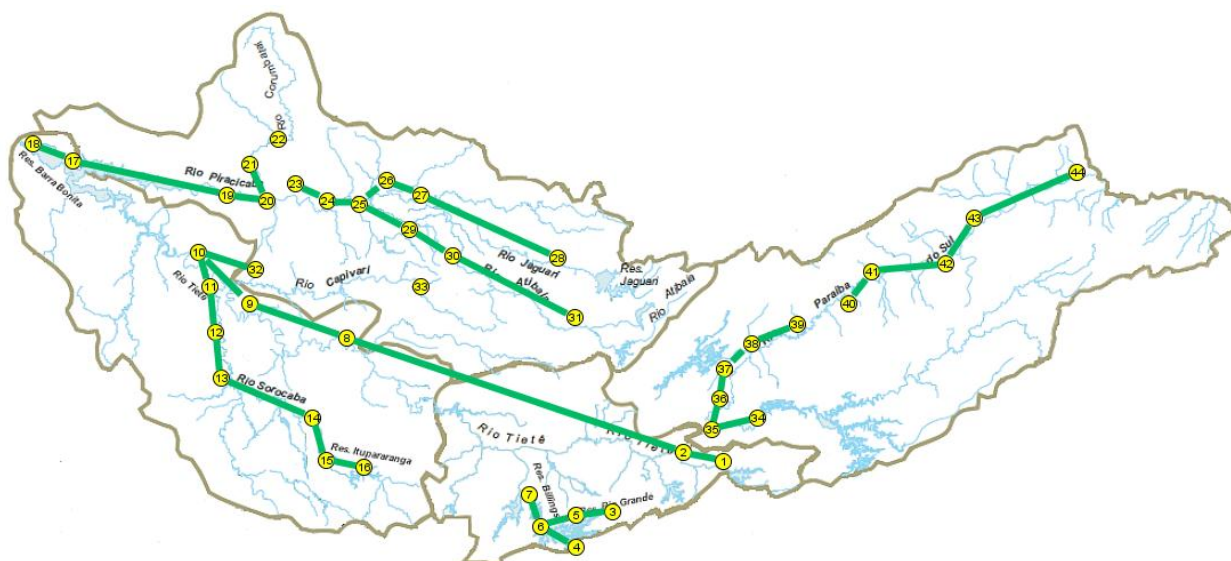
Grupos de pontos de amostragem de água gerados para as UGRHIs 5, 6 e 10:

[illegible]

Grupos de pontos de amostragem de água gerados para a UGRHIs 2:

[illegible]

Experimento 19 – Estação do Ano = Inverno



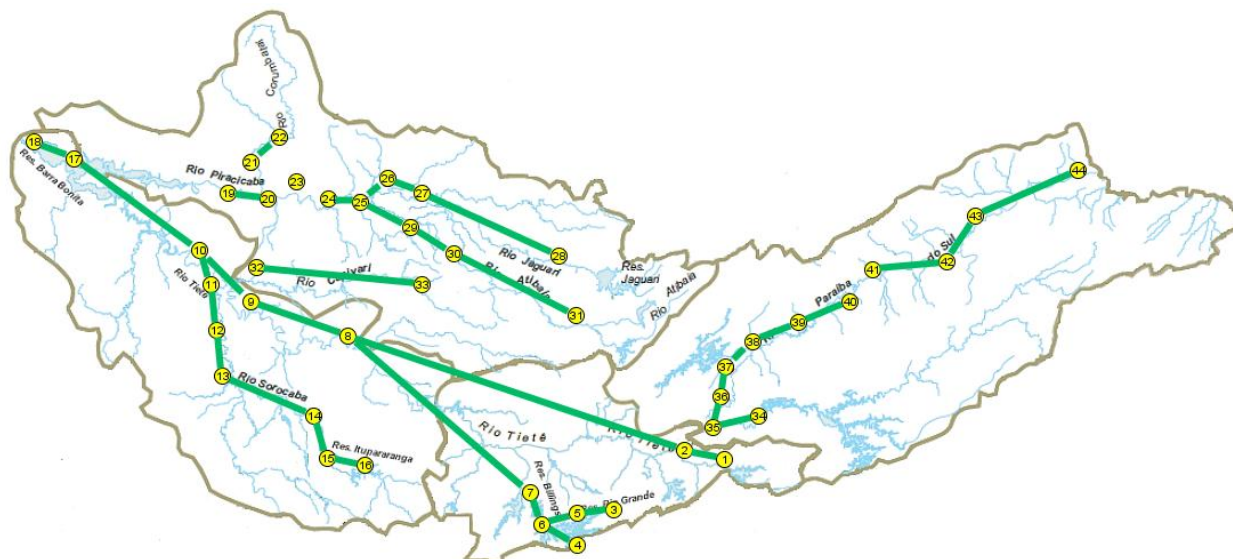
Grupos de pontos de amostragem de água gerados para as UGRHIs 5, 6 e 10:

[illegible]

Grupos de pontos de amostragem de água gerados para a UGRHIs 2:

[illegible]

Experimento 20 – Ano Completo



Grupos de pontos de amostragem de água gerados para as UGRHIs 5, 6 e 10:

[illegible]

Grupos de pontos de amostragem de água gerados para a UGRHs 2:

[illegible]

Avaliação e Interpretação

A partir da observação dos grupos de pontos de amostragem obtidos nos experimentos, é possível chegar às seguintes inferências:

- Nos experimentos 1 a 5, que contemplam os parâmetros Cádmio Total, Chumbo Total, Níquel Total, Nitrato e Nitrito, foram identificados os seguintes subgrupos de pontos de amostragem em comum:
 - **UGRHIs 5, 6 e 10:** 1 e 2; 3 a 6; 7 a 9; 10 a 12; 14 a 16; 20 e 23 a 25.
 - **UGRHIs 2:** 34 a 38; 40 a 43.
- Nos experimentos 6 a 10, que contemplam os parâmetros Cobre Dissolvido, Nitrogênio Amoniacal, Oxigênio Dissolvido, Substância Tensoativa, Zinco Total e Toxicidade, foram identificados os seguintes subgrupos de pontos de amostragem em comum:
 - **UGRHIs 5, 6 e 10:** 1 e 2; 3 e 5; 8 e 9; 10, 31 e 32; 11 a 14; 17 a 25; 26 a 28; 29 a 31.
 - **UGRHIs 2:** 34 e 35; 37 a 43.
- Nos experimentos 11 a 15, que contemplam os parâmetros Chuva 24h, Cloreto Total, Condutividade, pH, Sólidos Totais e Temperatura Água, foram identificados os seguintes subgrupos de pontos de amostragem em comum:
 - **UGRHIs 5, 6 e 10:** 1 e 2; 3 a 7; 12 e 13; 14 a 16; 17 a 20, 23 e 24; 21 e 22; 25, 29 e 30; 26 a 28.
 - **UGRHIs 2:** 34 a 38; 39 a 41; 43 e 44.
- Nos experimentos 16 a 20, que contemplam os parâmetros Alumínio Dissolvido, Ferro Dissolvido, Manganês Total e Turbidez, foram identificados os seguintes subgrupos de pontos de amostragem em comum:
 - **UGRHIs 5, 6 e 10:** 1 e 2, 3 a 7; 8 a 16; 17 e 18; 24 a 26 e 29 a 31; 27 e 28.
 - **UGRHIs 2:** 34 a 39; 42 e 43.

Estes subgrupos existentes dentro dos grupos de pontos de amostragem podem ser melhor visualizados por meio das Tabelas 15 e 16, onde cada cor representa um subgrupo de pontos que se repetem nos cinco experimentos de cada conjunto de experimentos.

A última linha das Tabelas 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. Pode-se constatar que os pontos de amostragem 1 e 2, 3 e 5, 34 e 35, e 37 e 38 possuem sempre medições similares, considerando os 21 parâmetros de qualidade contemplados em todos os experimentos.

Essa constatação contempla uma das contribuições esperadas neste trabalho, pois a descoberta de grupos de pontos de amostragem com medições similares traz subsídios para possíveis melhorias na distribuição destes pontos na rede de monitoramento de qualidade de água. O conhecimento levantado pode sugerir tanto a remoção de pontos de amostragem, em locais onde estes se apresentam contíguos e com medições sempre similares, como a adição de pontos de amostragem em trechos de corpos hídricos que ligam regiões muito díspares com relação às medições dos parâmetros de qualidade.

Tabela 15: Subgrupos de pontos de amostragem identificados nos conjuntos de experimentos (UGRHIs 5, 6 e 10).

Conjuntos de Experimentos	Pontos de Amostragem - UGRHIs 5, 6 e 10																																	
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	
1-5																																		
6-10																																		
11-15																																		
16-20																																		
1-20																																		

Tabela 16: Subgrupos de pontos de amostragem identificados nos conjuntos de experimentos (UGRHI 2).

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											

Apêndice D – Ferramentas e Bibliotecas Utilizadas

Este Apêndice descreve brevemente as principais ferramentas e bibliotecas de software utilizadas no desenvolvimento deste trabalho.

Ferramentas

- **Adobe© Acrobat 9 Pro Extend** – Ferramenta para criação e edição de arquivos em formato PDF. Neste trabalho, foi utilizada no processo de extração dos dados das análises de água, existentes nos arquivos PDFs disponibilizados pela CETESB. Por meio da conversão de tais arquivos para o formato XML, foi possível obter um formato estruturado, o que viabilizou a leitura dos dados pelo programa de extração.
- **DB Designer© (Versão 4.0.5.6 Beta)** – Ferramenta para projeto de banco de dados que integra a modelagem, projeto, implementação e manutenção em um mesmo ambiente. Foi utilizada para modelagem e normalização da base de dados responsável pelo armazenamento dos dados da CETESB.
- **Eclipse© SDK (Versão 4.2.0)** – Ambiente integrado para desenvolvimento de programas, aplicativos e ferramentas, de forma otimizada e padronizada, baseando-se nas iniciativas de software livre. Foi utilizada para implementação da ferramenta de descoberta de conhecimento apresentada neste trabalho.
- **PostgreSQL© (Versão 9.1)** – Sistema de gerenciamento de banco de dados relacional, que suporta quase todas as construções SQL, além de trabalhar com grandes volumes de dados. Nesse trabalho, foi usado na criação e manutenção da base para armazenamento dos dados da CETESB.

Bibliotecas

- **Chart2D© (Versão 1.9.6)** – Biblioteca Java para a visualização de dados quantitativos utilizando gráficos bidimensionais, tais como: gráfico de pizza, de linha, de barras, de dispersão, entre outros. Foi utilizada na geração do gráfico de pizza apresentado na funcionalidade de classificação de toxicidade em amostras de água.
- **JGraphX© (Versão 1.6.1.2)** – Biblioteca para construção de grafos interativos baseados em aplicações Java Swing. É projetada principalmente para o uso em ambiente *desktop*,

embora tenha uma versão para JavaScript que permite seu uso em ambiente Web. Neste trabalho, foi empregada para criação dos grafos gerados na funcionalidade para regionalização de pontos de amostragem de água.