UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE CIÊNCIAS EXATAS E NATURAIS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

APLICAÇÃO DE TÉCNICAS ESTATÍSTICAS E DE INTELIGÊNCIA COMPUTACIONAL NA CLASSIFICAÇÃO DE CICLOS HIDROLÓGICOS EM RESERVATÓRIOS DE ÁGUA NA REGIÃO AMAZÔNICA: UM ESTUDO DE CASO

JEAN CARLOS AROUCHE FREIRE

UFPA / ICEN / PPGCC
Campus Universitário do Guamá
Belém-Pará-Brasil
05/2014

UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE CIÊNCIAS EXATAS E NATURAIS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

JEAN CARLOS AROUCHE FREIRE

APLICAÇÃO DE TÉCNICAS ESTATÍSTICAS E DE INTELIGÊNCIA COMPUTACIONAL NA CLASSIFICAÇÃO DE CICLOS HIDROLÓGICOS EM RESERVATÓRIOS DE ÁGUA NA REGIÃO AMAZÔNICA: UM ESTUDO DE CASO

Dissertação submetida à Banca Examinadora do Programa de Pós-Graduação em Ciência da Computação da UFPA para a obtenção do Grau de Mestre em Ciência da Computação.

Orientador: Dr. Jefferson Magalhães de morais

UFPA / ICEN / PPGCC
Campus Universitário do Guamá
Belém-Pará-Brasil
05/2014

A minha esposa
e à minha mãe
com amor e carinho

AGRADECIMENTOS

Agradeço primeiro à Lei Divina, por ter me dado a oportunidade de minha existência, pela minha saúde, pela minha família, pela minha esposa e por tudo que conquistei e também pelo conhecimento adquirido.

À minha esposa Ilza Léia Ramos Arouche, pelo apoio, compreensão e compartilhamento dos momentos alegres e difíceis de minha vida. Uma eterna companheira que a Lei Divina me agraciou em colocá-la no meu caminho.

À minha família, especialmente à minha mãe, por ter sempre cuidado de mim em todos os momentos da minha vida com sua paciência, carinho e apreço.

A meu orientador, Prof. Jefferson Magalhães de Morais, por todo seu conhecimento, compreensão, dedicação e contribuição de orientação para o desenvolvimento deste trabalho, e principalmente por ter acreditado em mim.

À minha co-orientadora Profa. Terezinha Ferreira de Oliveira, colegas do grupo de pesquisas *Soft Computing*, professores e todos auxiliares do Instituto de Ciência Exatas e Naturais (ICEN), pelo apoio, orientação, contribuição e dedicação para o desenvolvimento desse trabalho.

Ao Programa de Pós-graduação em Ciência da Computação (PPGCC) da Universidade Federal do Pará, especialmente ao Prof. Francisco Edson Lopes da Rocha, por ter me aceitado como aluno e que tornou possível a realização desse trabalho.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), por contribuir de forma direta e indiretamente para a realização do meu estudo, principalmente na parte financeira.

À Eletrobras-Eletronorte por ter fornecido as amostras de análise de água dos sítios amostrais da Usina Hidrelétrica de Tucuruí que possibilitou a realização da pesquisa.

A todos aqueles que não foram citados, porém que de certa forma contribuíram para realização do meu estudo.

SUMÁRIO

LISTA DE ABREVIATURAS

AD - Análise discriminante

AE - Análise exploratória

AF - Análise fatorial

AM - Aprendizado de máquina

AMU - Anáise multivariada

ARFF - *Attribute-Relation file format*

Ca - Cálcio

Fe – Ferro

IA - Inteligência computacional

K – Potássio

KMO – *Kaiser-Meyer-Olkin*

KNN - *K-nearest neighbors* - (K- vizinhos mais próximos)

Mg - Magnésio

MLP - *Multilayer perceptron* (Perceptron de múltiplas camadas)

Na - Sódio

NH_4 - Amônia

NO_3 – Nitrito

NTU -*Nephelometric Turbidity* Units (Unidades Nefelométricas de Turbidez)

OD - Oxigênio dissolvido

pH - Potencial de hidrogênio

PO_4 – Fosfato

RBF - Radial basis functions (Funções de base radial)

RDS - Reserva de Desenvolvimento Sustentável

RF - Random forest (Floresta aleatória)

RNA - Rede neural artificial

STS - Sólidos totais em suspensão

SVM - *Support vector machine* - Máquinas de vetores de suporte

TSS - *Total suspended solids*

UHE - Usina hidrelétrica

ZPVS - Zona de Preservação da Vida Silvestre

WEKA - *Waikato environment for knowledge analysis*

LISTA DE SÍMBOLOS

μs - Micro segundos

σ - Desvio padrão

∑- Somatório

Φ - Função de estimação

k- Kernel (Núcleo)

λ - Autovalores

ω - vetor peso

x_v - vetor de suporte

χ^2 - Teste de bartlett de esfericidade

r^2_{ij} - Coeficiente de correlação observado entre as variáveis *i* e *j*

a^2_{ij} - Coeficiente de correlaçãoo parcial entre as mesmas variáveis

$d_{euclidiana}$ - Distância euclidiana

err(f) - Taxa de erro

Ac(f) - Acurácia

sens(f) - Sensibilidade

esp(f) - Especificidade

prec(f) - Precisão

df - Degree free (Grau de liberdade)

n - Tamanho da amostra

$\overline{X}_1$ - Média da amostra 1

$\overline{X}_2$ - Média da amostra 2

S - Desvio padrão de uma amostra

Sx_1x_2 - Desvio padrão de duas amostras

H_0 - Hipótese nula

H_1 - Hipótese alternativa

t - t-Student test

V_1 - Fator de variação de valores dos parâmetros de classificadores computacionais

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

RESUMO

Este estudo avalia a qualidade da água do reservatório da Usina Hidrelétrica de Tucuruí de acordo com o ciclo hidrológico da região e da disposição espacial dos diferentes sítios de coleta distribuídos nas zonas à montante da barragem no período de 2009 a 2012 a partir da alteração de 17 parâmetros físico-químicos e de metais da água extraídos de seis fatores que representaram 71,01% de variabilidade total dos dados. Foi observado que as maiores variações do NO_3, NH_4, Total P, PO_4 e STS ocorreram no período de enchentes, podendo ser uma indicação do estado trófico nos sítios amostrais em decorrência da existência de pólos pesqueiros ou da densidade populacional no entorno desses sítios. Para classificação do ciclo hidrológico foram utilizados seis classificadores: análise discriminante, redes neurais artificiais, k-vizinhos mais próximo, máquinas de vetores de suporte com núcleo radial e polinominal, e random forest. Os resultados obtidos indicaram que o classificador random forest foi o que apresentou melhor desempenho com percentual de classificação de 7,80% de predições incorretas. Enquanto que o *t-Student test* indica que random forest e k-vizinhos mais próximo tem em média taxa de predições incorretas iguais com índice de significância fixado em $\alpha = 5\%$.

PALAVRAS-CHAVE: Reservatórios; Hidrológia; Classificação de ciclos hidrológicos; Estatística Matemática-Processamento de dadoss; Inteligência Computacional.

ABSTRACT

This study evaluates the quality of the water reservoir of the Hydroelectric Plant Tucuruí according to the regional hydrological cycle and the spatial arrangement of the different sampling sites distributed in areas upstream of the dam in the period 2009-2012 from the amendment of 17 parameters physico-chemical and metals from water extracted of six factors that accounted for 71.01% of total data variability. It was observed that the greatest variations of NO_3, NH_4, , totalP, PO_4 and STS occurred in the period of floods and may be an indication of trophic status in the sampling sites due to the existence of fishing poles or population density in the vicinity these sites. Discriminant analysis, artificial neural networks, k-nearest neighbors, support vector machine with polynomial and radial core and random forest: classification of the hydrological cycle to six classifiers were used. The results indicate that the random forest classifier showed the best performance with a percentage rating of 7.80% of incorrect predictions. While Student t test indicates that random forest and k-nearest neighbors have an average rate of incorrect predictions with equal significance index set at $\alpha = 5\%$.

KEYWORDS: Reservoirs; Hydrology; Classification of hydrological cycles; Mathematical Statistics-processing data types; Computational Intelligence.

CAPÍTULO 1

Introdução

1.1 Visão Geral

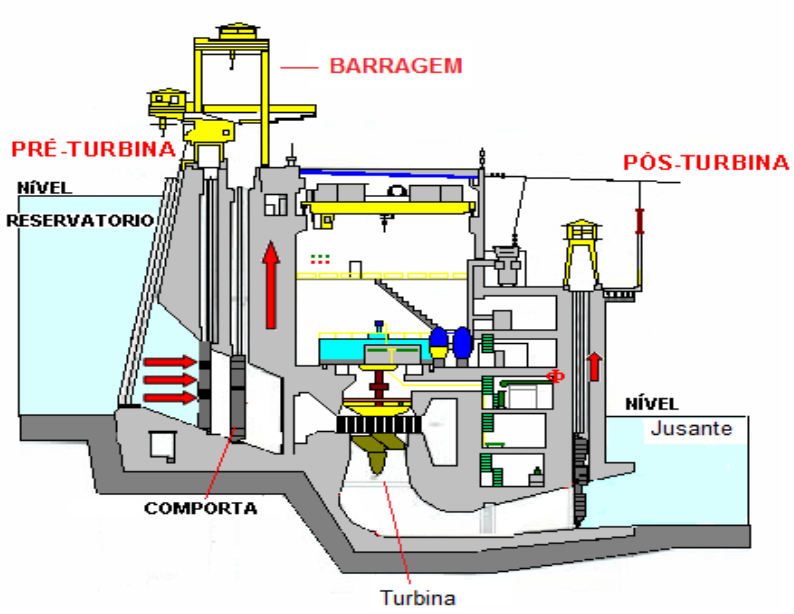
Devido ao crescimento da população mundial, a demanda por energia elétrica é cada vez maior. Para atender esta demanda crescente, muitos países, incluindo Brasil, estão investindo cada vez mais na ampliação da sua capacidade de geração de energia elétrica. Nesse contexto, a maioria dos países tem como base de sua matriz geradora de eletricidade derivados de combustíveis fósseis. Um típico exemplo desta forma de geração de energia elétrica são as termoelétricas.

Contudo, existe uma preocupação à nível mundial quanto aos impactos que este tipo geração de energia provoca no meio ambiente, devido principalmente às emissões de CO_4 produzidas a partir da queima destes combustíveis. Por isso, busca-se, hoje, fontes de energias alternativas que sejam menos agressivas ao meio ambiente, no que diz respeito às emissões de caborno na atmosfera. Dentre essas fontes alternativas, destacam-se as usinas hidrelétricas.

As usinas hidrelétricas aproveitam a diferença de energia potencial entre o nível de água à montante e à jusante. Quando a água cai do nível mais elevado para o menos elevado, dentro de um tubo, essa energia potencial é transformada em energia cinética e de pressão, que por sua vez faz girar a turbina produzindo energia elétrica.

No cenário Brasileiro, as usinas hidrelétricas são as que mais contribuem na matriz energética nacional. Isto porque o Brasil dispõe de um dos maiores potenciais hídricos do mundo. Apesar do país possuir numerosos rios com potencial hidrelétrico, muitos deles são caracterizados por grande vazão e pequena declividade. Por isso, muitas vezes faz-se necessário a criação de barragens e de reservatórios de regularização (vide Figura 1). Estes reservatórios são assim chamados porque em períodos mais chuvosos eles tendem a encher e em períodos menos chuvosos eles tendem a esvaziar evitando inundações e estiagens, respectivamente.

Figura 1: Aproveitamento hidrelétrico do rio de planalto (Rebouças et al., 2006).



Excetuando o reservatório de Três Marias, no Rio São Francisco, que foi concebido para atender a múltiplos usos (navegação, irrigação e produção de energia elétrica), os reservatórios brasileiros foram planejados e construídos visando unicamente à produção de energia elétrica (Rebouças et al., 2006). Embora a atividade pesqueira, que já existia antes do fechamento da barragem, se tornou mais intensa com a formação do reservatório, atraindo multidões em busca de trabalho, emprego e renda (Mérona et al., 2010).

O mecanismo de funcionamento de um reservatório está relacionado principalmente com o ciclo hidrológico e o volume de água, os gradientes verticais e horizontais e os processos de circulação produzidos pelo vento, pelo aquecimento e resfriamento térmico ou pelas descargas relacionadas com o uso (Rebouças et al., 2006).

Com base nesse contexto, fica claro a importância de estudar os ciclos hidrológicos já que a operação do reservatório, isto é, a variação da quantidade de água estocada modifica a disponibilidade para a usina situada rio abaixo, afetando também o abastecimento de água, a diluição de efluentes de cidades e/ou indústria, a irrigação, a navegação, o controle de enchente e a recreação (Bittencourt e Amadio, 2013; Rebouças et al., 2006).

Além disso, a qualidade da água do reservatório é influenciada pelo ciclo hidrológico a partir da variação das características microbiológicas, físicas e químicas de modo que o monitoramento é fundamental para o gerenciamento já que influencia a preservação e o equilíbrio dos ecossistemas (Atobatele e Ugwumba, 2008; Bertholdo et al., 2013; Bittencourt e Amadio, 2013;

Chapman, 1996; Galvão, 1999; Gastaldini et al., 2002; Kazi et al., 2009; Lewis, 2000; Singh et al., 2009; Vrana et al., 2005).

Vários estudos tem sido realizados para melhoria do gerenciamento, a partir do aprimoramento do monitoramento, da aplicação de modelos ecológicos e de técnicas estatísticas e da inteligência computacional para a implantação de sistemas de suporte à decisão, para o controle efetivo da qualidade da água considerando os diversos aspectos da legislação brasileira. Grande parte destes trabalhos foram realizados com o objetivo de identificar as alterações sazonais dos parâmetros físico-químicos e nas concentrações de metais em ambientes aquáticos (Coletti et al., 2010; Hauser-Davis et al., 2010; Kazi et al., 2009; Ouyang, 2005; Song et al., 2011; Zahraie e Hosseni, 2009; Wang et al., 2009; Wenner et al., 2005). Dentre as técnicas utilizadas para esse propósito se destacam contribuições recebidas da inteligência computacional e estatística (Cavalcante et al., 2013; Hauser-Davis et al., 2012; Singh et al., 2009; Zahraie e Hosseni, 2009; Wang et al., 2009).

1.2 Motivação

As represas da Amazônia produzem grandes alterações ambientais resultantes de modificações no nível da água, aumento da oxigenação e perda de matéria orgânica à jusante.

Além desses impactos, verifica-se que a ocupação do solo após o enchimento e funcionamento da represa é outro fator de perda da diversidade, como aconteceu próximo a represa de Tucuruí com os grandes desmatamentos resultantes da ocupação imediatamente posterior à construção e o enchimento do reservatório.

Por essas razões e, sabendo que o parque hidrelétrico brasileiro é um dos maiores do mundo, a motivação para este estudo foi propiciar a melhoria do gerenciamento de reservatórios a partir da aplicação das técnicas aqui propostas. Pois, o estudo das variacões sazonais pode-se fazer estimativas climáticas que contribuam na administração dos recursos hídricos em reservatórios na região amazônica. Essas perspectivas podem ser potencializadas com o auxílio de recursos computacionais como suporte aos sistemas de gestão ambiental

1.3 Objetivos

1.3.1 Objetivo geral

O objetivo geral deste trabalho consiste na aplicação de técnicas estatísticas e da inteligência computacional para classificar os ciclos hidrológicos considerando a alteração nos parâmetros físico-químicos e na concentração dos metais na água do reservatório da UHE de Tucuruí.

1.3.2 Objetivo específico

Especificamente, este trabalho busca:

- Avaliar a qualidade da água do reservatório de acordo com o ciclo hidrológico da região. Para isso deve-se levar em conta a distribuição espacial dos diferentes sítios de coleta dispostos nas zonas à montante da barragem.
- Selecionar os parâmetros (atributos) mais relevantes para o processo de classificação utilizando técnicas estatísticas de análise multivariada: análise fatorial e análise discriminante.
- Avaliar de forma sistemática diferentes técnicas de inteligência computacional aplicadas a classificação de ciclos hidrológicos.
- Aplicar testes de significância para estimar o grau de equivalência entre os classificadores avaliados.

1.4 Metodologia

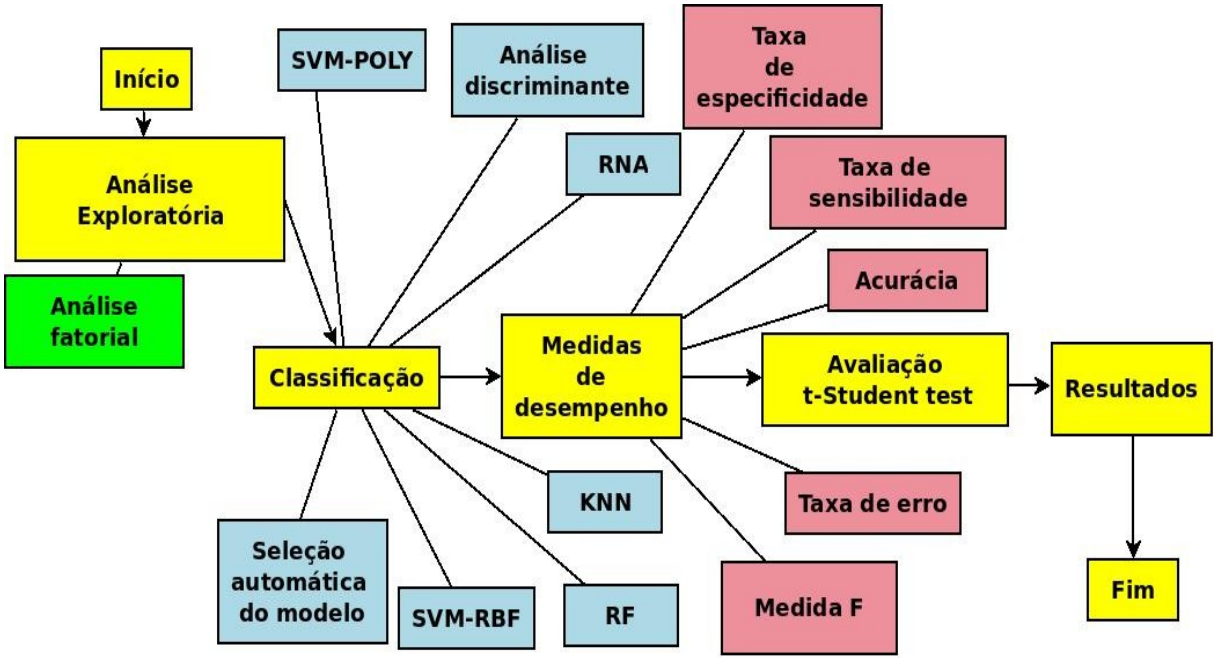
A área escolhida para o estudo de caso foi a região à montante da barragem da Usina Hidrelétrica (UHE) de Tucuruí. Foram coletadas, no período de 2009 a 2012, 423 amostras de água considerando 39 parâmetros físico-químicos e metais de nove sítios amostrais.

Foram utilizados no estudo técnicas estatísticas de análise multivariada. Tais técnicas serviram como mecanismos para exploração dos dados, trazendo assim confiabilidade para o pré-processamento destes e para classificação dos ciclos hidrológicos.

A metodologia adotada no estudo é apresentada na Figura 2, em que pode-se observar que foi feita inicialmente a análise exploratória de dados, incluindo a normalização das variáveis do estudo. A aplicação da Análise Fatorial (AF) serviu para estabelecer a estrutura de variação dos dados e extrair os fatores. Com esta técnica foi possível selecionar as variáveis mais relevantes da

base de dados. Já a Análise Discriminate (AD) foi empregada na classificação dos ciclos hidrológicos. Todos os procedimentos estatísticos foram feitos utilizando o *software* SPSS.

Figura 2: Fluxo da metodologia utilizada no estudo.



No que tange a aplicação das técnicas de inteligência computacional, com o auxílio do *software* WEKA, foi feita a seleção automática do modelo para escolha dos valores dos parâmetros dos classificadores. Além disso, com a base de dados adaptada para arquivos ARFF, padrão de arquivo adotado pelo WEKA, foi possível obter os resultados da classificação dos ciclos hidrológicos utilizando as mesmas variáveis de entrada empregadas na análise estatistica. Tal procedimento foi realizado com os seguintes classificadores: *Random Forest* (RF), Rede Neural Artificial (RNA), *K-Nearest Neighbors* (KNN) e *Support Vector Machine* (SVM) considerando *kernel* de base radial e polinominal.

Os resultados foram processados com validação cruzada com percentuais padrão divididos em treinamento e teste. As medidas de desempenho foram extraídas da matriz de confusão avaliando a taxa global de erro, exatidão, precisão, sensibilidade, especificidade e a medida *F*. Por fim, foi aplicado o *t-Student test* para análisar a significância da diferença da classificação obtida pelos classificadores.

1.5 Estrutura da dissertação

Este trabalho está dividido em cinco capítulos, organizados da seguinte maneira: o primeiro capítulo discute de forma introdutória a motivação, a metodologia, os objetivos e a estrutura da pesquisa.

O Capítulo 2 faz um levantamento dos trabalhos relacionados nos últimos três anos que envolvem a importância dos estudos sobre a qualidade da água, assim como estudos da sazonalidade ambiental que possam impactar os ecossistemas aquáticos em diversos aspectos, contextualizando assim o entendimento da proposta da pesquisa.

O Capítulo 3 descreve detalhadamente a área de estudo, as características dos sítios amostrais no reservatório da UHE de Tucuruí, os parâmetros físico-químico e metais da água, as técnicas estatísticas e computacionais, assim como, os procedimentos e as ferramentas utilizadas para validação dos resultados.

No Capítulo 4 são apresentados os resultados obtidos considerando as técnicas estatísticas, tanto a análise exploratória de dados quanto a Análise Discriminante. Ainda neste capítulo discute-se os resultados obtidos pelos classificadores, e também os resultados da aplicação do *t-Student test*. Finalmente, no Capítulo 5 as conclusões e as propostas de trabalho futuro são apresentadas.

CAPÍTULO 2

Estado da Arte

Existem inúmeros trabalhos relacionados à proposta do presente estudo. A maioria abordam técnicas estatísticas e de inteligência computacional na predição de contextos ambientais, sendo alguns destacados nessa sessão.

Em Hauser-Davis et al. (2010) foram utilizadas as técnicas de análise discriminante e redes neurais artificiais para classificação de três espécies de peixes em diferentes locais no estado do Rio de Janeiro, Brasil. As redes neurais foram as que apresentaram os resultados mais satisfatórios, mesmo com grupos de tamanhos diferentes, indicando esta técnica como uma excelente alternativa para problemas de classificação com dados desbalanceados.

Em Song et al. (2011) foi realizada uma analise da qualidade da água do rio Beijiang, China, para determinar a influência das águas agrícolas e urbanas no reservatório. Para isso, foram utilizados 18 parâmetros a partir de coletas mensais de amostras de treze pontos de coleta durante os anos de 2005 e 2006. Uma analise de correlação bivariada foi utilizada para avaliar as correlações dos parâmetros, enquanto a análise de componentes principais foi aplicada para extrair as variáveis que mais influenciam nas variações regionais. Seis componentes foram extraídos, os quais explicam mais que 78,00% e 84,00% da variância total agrícola e urbana respectivamente. A regressão linear multivariada foi utilizada para estimar a contribuição de todas as fontes de poluição identificadas. Altos coeficientes de determinação das equações de regressão sugerem que a técnica é aplicável para estimativa das fontes da maioria dos parâmetros físico-químicos.

Murtojarvi et al. (2011) fizeram monitoramento da qualidade da água no litoral do arquipélago da Finlândia que é totalmente fragmentado e exige uma rede de observação densa, levando ao custo econômico considerável. Os autores propuseram um método de otimização para determinar um conjunto de sítios amostais que dispõe de formações adequadas com precisão eficiente.

Em Ramin et al. (2012) foi desenvolvido um estudo utilizando modelagem baysiana para predição de qualidade da água concluindo que não existe um único modelo para sistemas ecológicos, mas várias descrições adequadas de diferentes bases e estruturas conceituais.

Liu e Chen (2012) utilizaram modelos de circulação tridimensional e RNA para estudar a estratificação térmica da qualidade da água de lagos. Os autores concluíram que apesar do bom desempenho das RNAs o modelo de circulação tridimensional é melhor, pois consegue simular os processos físicos de acordo com as variações espacial e temporal simultaneamente.

Holguin-Gonzalez et al. (2013) elaboraram uma estrutura genérica para apoiar à decisão no gerenciamento de gestão da qualidade da água, a partir de um modelo hidráulico, físico-mecanicista, com modelos ecológicos aquáticos. Os resultados apresentaram a importância de modelos integrados aos objetivos propostos.

Em Sharma et al. (2013) foi apresentada uma ferramenta automatizada (*Surface Water Quality Assessment*), que considera 14 diferentes variáveis físico-químicas, toxicológicas e bacteriológicas. Os resultados demonstraram que o sistema é útil para analisar a variação do local e do tempo onde a amostra de água é coletada.

Em Ren et al. (2012) foi realizado um estudo sobre a associação entre a biodiversidade da macrozooplâncton e os índices físico-químicos da água em oito locais na Bacia do Rio Wenyuhe na China de abril de 2010 a março de 2011 a partir de uma rede neural do tipo mapa auto-organizável. Os autores concluíram que essa associação pode ser útil na definição de qualidade da água e integridade ecológica na gestão dos ecossistemas aquáticos, especialmente para ambientes complexos.

Em Cavalcante et al. (2013) foi desenvolvido um estudo comparativo utilizando técnicas de Intelegência Computacional e Estatística para classificar a qualidade da água no período chuvoso e seco dos reservatórios da UHE de Tucuruí e Coaracy Nunes,a partir dos parâmetros físico-químicos e dos metais. Para o reservatório de Tucuruí os resultados apresentaram 91,00% de predições corretas com AD e 100,00% com RNA e para o reservatório de Coaracy Nunes obteve-se 96,40% com AD e 100,00% com RNA.

Em Borges Pedro et al. (2014) foi analisada a influência dos ciclos hidrológicos sobre os parâmetros físico-químico da água na parte médio do rio Solimões entre 2004 a 2011 em diversos corpos d'água utilizando estatística descritiva e de agrupamento. Os resultados apresentados mostraram que a variação do nível da água durante os ciclos avaliados ao longo dos sete anos de monitoramento influenciou a qualidade da água nos lagos da Reserva de Desenvolvimento Sustentável Mamirauá e nos rios Solimões e Tefé. As variações mais acentuadas observadas foram nos parâmetros transparência, oxigênio dissolvido e condutividade nos períodos de seca.

Em Lobato (2014) foram utilizadas técnicas estatísticas multivariadas na criação de indicadores e adaptação de índices na construção de um sistema fuzzy que permitiram caracterizar e classificar água em relação ao estado trófico a partir da análise exploratória dos dados referentes aos parâmetros físico-químicos, a clorofila e os metais da água do reservatório da Usina Hidrelétrica de Tucuruí coletados nos anos 2009 a 2012.

Os trabalhos de Cavalcante et al. (2013) e de Lobato (2014) serviram de base para este trabalho, diferenciando nos parâmetros físico-químico e nos metais, nas técnicas utilizadas e no enfoque da abordagem dos ciclos hidrológicos. Apesar de serem pesquisas relacionadas com a qualidade da água, os trabalhos anteriores não abordam a classificação dos ciclos hidrológicos em relação a alteração dos parâmetros físico-químico e metais da água nas UHEs da Amazônia quanto à sazonalidade, fazendo desta uma pesquisa original e relevante.

CAPÍTULO 3

Materiais e Métodos

3.1 Área de estudo

Os estudos para a construção de uma hidrelétrica para o aproveitamento do potencial do Rio Tocantins tiveram início em 1957. Na década de 1960 foram feitos estudos de alternativas para o eixo da barragem, as decisões tomadas na década de 1970 e a inauguração em 1984. Ela foi construída para suprir energia para a produção de alumínio e estimular a indústria regional e articular as ligações e produzir energia para abastecer o país em escala nacional (Mérona et al., 2010).

A primeira etapa da implantação ocorreu entre 1975 e 1989, com doze unidades principais com a capacidade total de 3960 MW e posteriormente duas unidades auxiliares que elevaram a capacidade instalada para 4000 MW. A segunda etapa inaugurada no final de 2008 elevou a capacidade instalada para 7960 MW. A barragem acarretou a formação de um grande lago de cerca 200km de extensão e uma área de aproximadamente 2875 km^2.

A UHE de Tucuruí (Figura 3) está localizada no Rio Tocantins, no estado do Pará, fica cerca de 7 km da cidade de Tucuruí e a 300 km em linha reta da cidade de Belém. O rio Tocantins com seu principal afluente, o Araguaia, constitui uma bacia própria, ora denominada Bacia do Tocantins, ora Bacia do Tocantins-Araguaia. Nascido no planalto central brasileiro, este rio percorre grandes extensões recobertas por cerrados antes de penetrar em áreas de floresta amazônica densa (Eletrobras-Eletronorte, 2008).

Figura 3: Vista aérea da UHE de Tucuruí (Fonte: Eletrobras-Eletronorte, 2008).

As informações de localização geográfica e características principais da UHE de Tucuruí, área de estudo da pesquisa, são apresentadas no Quadro 1.

Quadro 1: Características gerais da área estudada.

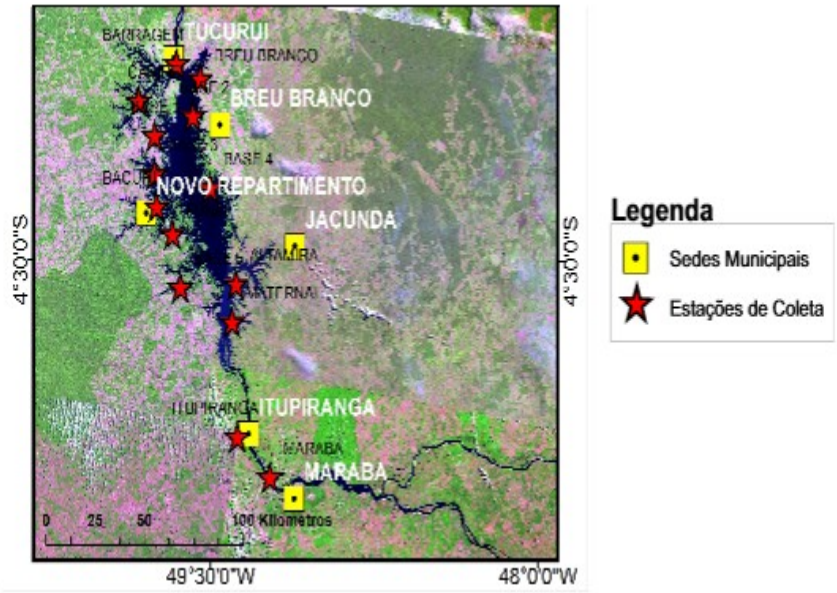
Características	Descrição
País	Brasil
Latitude	03° 43'-- 05° 15'S
Longitude	49° 12'--50° 00' W
Vegetação	Úmida
Rio	Tocantins
Data do fechamento	Set/84
Duração do Enchimento	6 meses
Area de Inundação km^2	2.875 km^2
Área de Bacia de Drenagem	758.000 km^2
Comprimento Total	170 km
Volume Total	50,29 km^2
Profundidade Máxima	77 m
Profundidade Média	19.8 m
Tempo de Residência Aproximado	51 dias
Vazão Média de Longo termo	11.170 m^3

3.1.1 Cacaracterísticas dos Sítios amostrais do estudo

A construção da barragem UHE de Tucuruí provocou um grande aumento na população nas proximidades da obra e deslocamento da população rural da área a ser inundada. Houve um aumento da população na região de Marabá em decorrência do desenvolvimento dos projetos de mineração na região de Serra Pelada e Carajá, bem como a colonização agrícola no entorno da Transamazônica.

A atividade pesqueira, que já existia antes do fechamento da barragem se tornou mais intensa com a formação do reservatório. A grande disponibilidade de recursos pesqueiros gerados pela formação do reservatório atraiu um grande número de pessoas em busca de trabalho, emprego e renda (Mérona et al., 2010). A Figura 4 exibe as estações de coletas de peixes à montante da barragem.

Figura 4: Estações de coletas de peixes à montante da barragem.



As características dos reservatórios são as regiões com influência do rio (à montante), com a influência da represa e das descargas (próximas à barragem) ou as localizadas no meio do reservatório como o funcionamento do lago. Mérona et al. (2010) dividiram à montante da barragem em quatro zonas distintas. Zona: 1 inferior do reservatório, próxima à barragem; Zona 2: mediana do reservatório; Zona 3: superior do reservatório e Zona 4: situada entre Itupiranga e Marabá.

Na Zona 1, que está próxima à barragem, estão os sítios amostrais na margem esquerda, um braço do rio Caraipé, os sítios Caraipé 1 (C1) e Caraipé 2 (C2). Na margem direita está Breu Branco. Na região do Caraipé há, desde 2002, a Reserva de Desenvolvimento Sustentável RDS Alcobaça. Esta região se encontra bastante desmatada. Este sítio amostral (MBB) se situa na cidade de Breu Branco, na margem direita do reservatório, apresentando-se antrópica com poucas áreas de floresta primária. Essa região é caracterizada pelo lançamento de boa parte da drenagem superficial da cidade (Lobato, 2014; Mérona et al., 2010).

Segundo Lobato (2014), a qualidade da água coletadas nos sítios amostrais C1, C2 e MBB, apesar de estarem localizados próximos à região onde residem muitos moradores, foi classificada como Boa para C2 e MBB, enquanto para C1 suas águas ficaram muito próximas dessa classificação.

Na Zona 2, região mediana do reservatório, estão os sítios amostrais Montante Belauto (MBL), Montante Pucuruí (MP), Montante 3 (M3).

O sítio amostral MBL está localizado na margem direita do reservatório, nesta região existe a área da Zona de Preservação da Vida Silvestre ZPVS Base 4, área de proteção integral não permitindo a presença de moradores. O sítio amostral MP está na margem esquerda, na região da Reserva de Desenvolvimento Sustentável RDS Pucuruí Ararão, onde existe desembarque de peixes no Polo Pesqueiro para seguir ao município de Novo Repartimento, sendo assim uma região bastante antrópica com a diminuição da vegetação nativa nos últimos dez anos (Lobato, 2014).

A qualidade da água coletada no sítio amostral M3 dependem da vazão afluente do reservatório, por estar localizado na parte central, aproximadamente 60 km em linha reta da barragem.

Na Zona 3, região superior do reservatório, se encontram os sítios Montante Lontra (ML) e Montante Jacundá Velho (MJV). O sítio amostral ML está localizado onde existe a maior extensão contínua de vegetação nativa, na antiga calha do igarapé Bacuri, próximo da terra indígena dos índios Parakanãs. E o sítio amostral MJV está localizado perto de desembarque de peixe do Porto Novo (município de Jacundá) e Porto da Colônia (município de Goianésia do Pará). Esta região é bastante antrópica com poucas áreas de vegetação nativa.

De acordo com Lobato (2014) a qualidade da água na UHE de Tucuruí, considerando os mesmos anos de estudo desta pesquisa, 2009 a 2012, coletados nos sítios amostrais nas zonas 2 e 3, foram classificadas como aceitáveis.

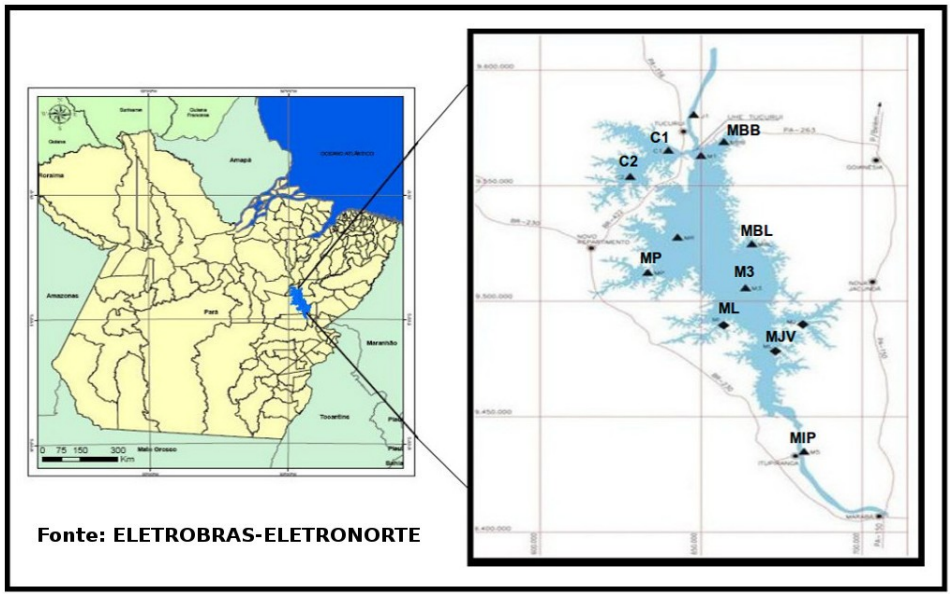
Na Zona 4, situada entre as cidades de Itupiranga e Marabá, se encontra o sítio amostral Montante Ipixuna (MIP) que está distante aproximadamente 130 km em linha reta da barragem. Da mesma forma, que a qualidade da água coletada nos sítios das zonas 2 e 3, a qualidade coletada

neste sítio foi considerada aceitável. Estes resultados já eram esperados, pois MIP está próximo da cidade de Itupiranga, possuindo intensa ocupação desde os fins da década de 80.

Para coleta de amostras de água existem doze sítios nessas zonas. Neste estudo foram investigados nove sítios que não apresentam grandes profundidades. Não foram considerados os seguintes sítios amostrais. Montante 1, Montante 5 e Montante Repartimento.

A Figura 5 exibe os sítios amostrais utilizados no estudo: Caraipé 1(C1), Caraipé 2(C2), Breu Branco(MBB), Ipixuna (MIP), Jacundá Velho (MJV), Lontra (ML), Pucuruí (MP), Montante 3 (M3) e Belauto (BE).

Figura 5: Distribuição dos sítios amostrais do reservatório.



\

3.1.2 Coleta e análises das amostras de água

As coletas e análises foram realizadas de janeiro de 2009 a dezembro de 2012. Neste período foram feitas campanhas anuais à montante da barragem, considerando o ciclo hidrológico da região na seca (setembro à novembro), enchente (dezembro à janeiro), cheia (março à maio) e vazante (meses de junho à agosto), em nove sítios amostrais. A Figura 6 mostra o comportamento do nível de água (em metros) à montante da barragem nos quatro anos mencionados (Cavalcante et al., 2013).

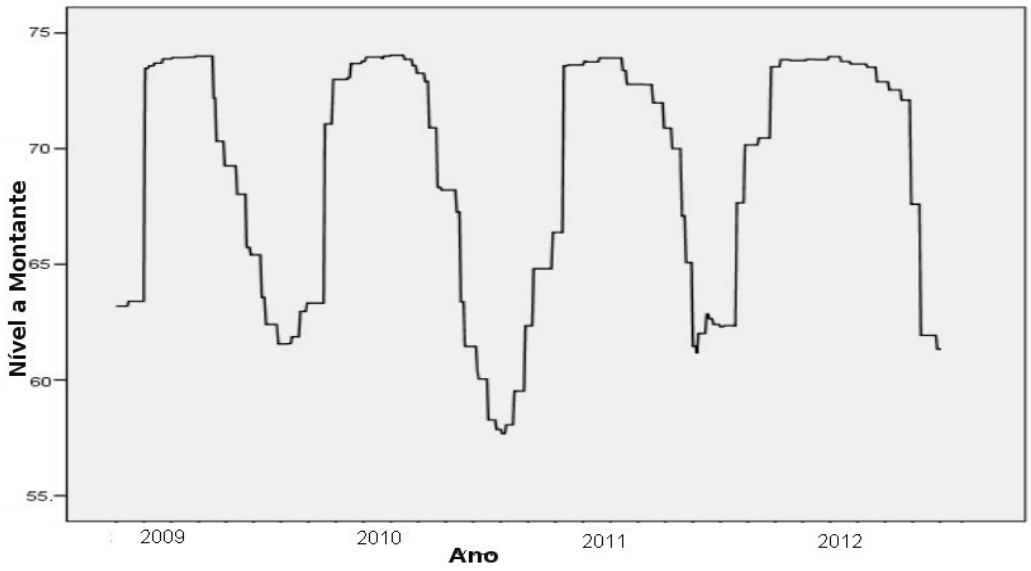
As amostras foram armazenadas em frascos eppendorf de 50 ml para metais e acidificadas com ácido nítrico. Os frascos utilizados para coleta foram inicialmente lavados com água ultra-pura

para remover sujeiras e plásticos remanescentes. Em seguida foram submetidos à solução de lavagem de acido nítrico 10\%, por um período de 24h. A preservação das amostras foi realizada segundo a norma *Standard Methods* (2013).

Os parâmetros pH, condutividade elétrica e temperatura foram determinados no local da coleta, através da utilização de aparelhos portáteis, enquanto que os demais parâmetros foram medidos no Laboratório. E a análise de água para determinação dos metais foi feita pela técnica multielementar com espectrometria de emissão óptica em plasma por acoplamento indutivo (ICP OES).

A classificação foi realizada em base de dados fornecida pela Eletrobras-Eletronorte contendo um total de 423 registros de análise de água de nove sítios amostrais do reservatório da UHE de Tucuruí entre os anos de 2009 a 2012. Foram considerados 17 variáveis entre 39 da base original de dados para os classificadores mapeados a partir dos parâmetros físico-químicos e metais objetivando a classificação do ciclo hidrológico (CH), que afetam o nível de água: seca, enchente, cheia e vazante. A Figura 6 mostra o comportamento à montante do reservatório nos 4 anos mencionados.

Figura 6: Nível de água (m) à montante da barragem nos de estudo.



De acordo com a Figura 6, observa-se que o ano de 2009 foi o período que o reservatório permaneceu menos cheio, e o ano de 2012, mais cheio, diferente da transição de 2010 para 2011 que teve sua maior baixa. Outro fator relevante, é que claramente se observa uma sequência padrão dos quatro ciclo hidrológicos conforme o comportamento do nível à montante, ou seja, em cada ano o reservatório inicia com seu nível mais baixo, depois vai enchendo até chegar ao ponto mais alto, e

logo após, começa a esvaziar. Entretanto, devido a sazonalidade climática ser oscilante e imprecisa, os dados mostram que esse comportamento se mantém diferente para cada ano.

As amostras de água foram coletadas no reservatório da UHE de Tucuruí utilizando a garrafa de van dorn de 2,5 litros de capacidade, em no máximo três profundidades definidas pelo disco de Secchi (superfície, Secchi e dobro do Secchi). Os parâmetros físico-químicos e metais analisados nas águas do reservatório com suas unidades de medidas foram: Secchi (m), Temperatura (ºC), pH, OD (mg O_2/L), Condutividade (µs cm^{-1}), Ferro (mg L^{-1}), Ca (mg L^{-1}), Mg (mg L^{-1}), K (mg L^{-1}), Na (mg L^{-1}), NH_4 (µg L^{-1}), NO_3(mg L^{-1}), totalPO_4 (mg L^{-1}), PO_4 (mg L^{-1}), STS(mg L^{-1}), Turbidez (NTU), Clorofila-a (mg L^{-1}).

3.1.3 Características dos parâmetros físico-químicos e metais da água do estudo

Os parâmetros físico-químicos e metais da água apresentam características específicas que podem ser avaliadas sobre a qualidade hidrológica nos corpos dŽágua. Esses aspectos sofrem uma grande influência do clima, geomorfologia e condições geoquímicas do lençol aquífero e na base de drenagem (Chapman, 1996). Entretanto, com a criação de represas, há uma grande modificação nessas características como foi descrito por Mérona et al. (2010) e Oduwole (1997).

Na época da construção da UHE de Tucuruí, percebeu-se que durante o período de cheia, ocorreu a diminuição no nível da água e uma grande extensão do leito do rio ficou exposta quando tinha o total fechamento da barragem, acarretando grandes mudanças nos parâmetros físico-químicos e metais da água.

Na parte à jusante, a quantidade de sólidos totais em suspensão diminuiu devido à interrupção do fluxo do rio, com o aumento no final de abril com as enchentes voltando a diminuir na época de seca. A transparência aumentou atingindo em março de 1986 quatro metros.

Logo após o fechamento da barragem também houve um rápido decaimento no teor de oxigênio dissolvido, especialmente no trecho perto da barragem, mas a água foi logo oxigenada após a abertura das comportas. Nos anos seguintes, o oxigênio dissolvido apresentou uma grande saturação no período de cheias e baixa saturação no período de seca. O teor de CO_2 seguiu uma tendência inversa à do oxigênio em decorrência da oxidação da amônia durante o período seco. Houve um grande aumento na concentração de nutrientes particularmente do fósforo, devido à decomposição da matéria orgânica no reservatório.

Depois do completo enchimento do reservatório, nos meses de cheia, o teor de oxigênio no fundo chegou a ultrapassar 2mg L^{-1} tornando-se estável com o decorrer dos anos. No primeiro ano,

o teor de CO_2 apresentou menores valores no período seco quando a fotossíntese é alta e maiores no período de cheia com a grande intensidade dos processos de oxigenação.

Na superfície do reservatório, os maiores valores de nutrientes ocorreram na cheia e os menores na seca, quando esses elementos são reciclados pelos produtores primários.

Essas modificações físicas e químicas na qualidade da água, principalmente a disponibilidade de nutrientes, permitiram o desenvolvimento dos vegetais aquáticos, com uma maior concentração de fitoplâncton e macrófitas aquática no lago do que à jusante. Esses acontecimentos acarretaram consequências drásticas nas comunidades de peixes refletindo na produção pesqueira na região abaixo da barragem e no reservatório.

Os 17 parâmetros físico-químicos e metais que foram extraídos a partir das 39 variáveis iniciais pela técnica de Análise Fatorial com as suas unidades de medidas são:

3.1.3.1 Secchi

O secchi, corresponde a um procedimento para medir a transparência da água, cujo ponto de partida é a utilização de um prato chato com peso suficiente para afundar a determinadas profundidades. De côr branca ou em 4 partes intercaladas de branco e preto, preso ao centro por um cordão ou bastão, o secchi serve como referencial para as coletas das amostras de água na superfície, o dobro do secchi, meio e fundo com base de medida em metros (Bezerra, 2012). A Figura 7 apresenta o disco de secchi para medir a transparência da água.

Figura 7: Disco de secchi

3.1.3.2 Temperatura

A temperatura, com sua unidade de medida em graus Celsius ºC, representa um papel relevante nos corpos d'água, pois influência inúmeros parâmetros físico-químicos. Algumas variáveis como viscosidade, tensão superficial, calor específico, constante de ionização, compressibilidade e calor latente de vaporização diminuem com o aumento da temperatura.

Fazendo o caminho inverso, a condutividade térmica e a pressão aumentam. Essas variações interferem nos ecossistemas aquáticos, pois possuem limites de tolerância térmica, propiciando ou não, um ambiente favorável para o crescimento, reprodução, desova e incubação de inúmeros organismos. Assim, variações de temperatura fazem parte do ciclo natural climático e está condicionada a latitude, altitude, estação do ano, período do dia, taxa de fluxo e profundidade. Entretanto, a elevação de temperatura nos corpos d'água é provocada por despejos de resíduos industriais, domésticos e usinas termoelétricas (Atobatele e Ugwumba, 2008; Cetesb, 2009; Chapman, 1996).

3.1.3.3 Potencial de hidrogênio

O potencial de hidrogênio (pH) mede a concentração de íon hidrogênio na água, sendo uma escala que apresenta valores neutros, básicos e altos. São considerados ácidos valores abaixo de sete, acima de sete e menor que oito e meio o ideal para vida aquática, e abaixo de quatro prejudiciais. Tais características vem fortemente assegurar que ecossistemas não toleram uma variação em grande escala do pH em águas naturais por muito tempo que podem acarretar mortes de espécies aquáticas, sendo influenciado diretamente pela concentração de dióxido de carbono (Cavalcante, 2013; Li et al., 2013).

3.1.3.4 Oxigênio dissolvido

O oxigênio dissolvido (OD), tendo sua unidade de medida em miligramas de oxigênio por litro (mg O_2/L), é o agente oxidante mais importante em águas naturais, pois envolve uma reação química de transferência de elétrons, sendo que cada um dos átomos da molécula é reduzido do estado de oxidação zero até o estado de oxidação -2, formando assim H_2O. Esse processo se torna essencial para os ecossistemas aquáticos, cujo ponto determinante é de fornecer uma fonte de oxigênio dentro do habiat natural da vida aquática (Fiorucci e Filho. 2005).

3.1.3.5 Condutividade elétrica

A condutividade elétrica, medida em microsegundo por centímetro cúbico (µs cm^{-1}), é a forma que a água possui de conduzir corrente elétrica. Tal aspecto se desenvolve, devido a presença de íons dissolvidos na água, que são partículas carregadas eletricamente, sendo a potência elétrica, diretamente proporcional a quantidade de íons dissolvidos na água. Em águas continentais, o cálcio, o magnésio, o potássio, o sódio, carbonatos, carbonetos, sulfatos e cloretos são responsáveis diretamente pelo valores da condutividade elétrica. Nessa direção, esses fatores não determinam, especificamente, quais os íons que estão presentes em determinada amostra de água, mas pode contribuir para possíveis reconhecimentos de impactos ambientais que ocorram na bacia de drenagem ocasionados por lançamentos de resíduos industriais, mineração, esgotos entre outros (Cetesb, 2009).

3.1.3.6 Ferro

O ferro (Fe), medido em miligramas por litro (mg L^{-1}), é encontrado geralmente em águas subterrâneas devido à dissolução do minério pelo gás carbônico da água. Em águas superficiais, o nível de ferro se acentua no período chuvoso devido ao carreamento de solos e a ocorrência de processos secundários que levam a erosão das margens. O ferro, apesar de não se apresentar como um agente tóxico, pode trazer inúmeros problemas para o consumo humano em altas concentrações. Águas que contêm a presença de ferro caracterizam-se por apresentar cor acentuada e turbidez baixa e em muitas estações de tratamento de água, este problema só é resolvido mediante a aplicação de cloro, denominada de pré-cloração (Mann et al., 2013; Cetesb, 2009).

3.1.3.7 Cálcio

O cálcio (Ca), medido em miligramas por litro (mg L^{-1}), é um elemento pertencente ao grupo dos alcalinos terrosos, sendo um íon responsável em determinar a dureza da água. Quanto maior sua concentração, maior será a dureza da água. Altas concentrações de cálcio podem ser prejudiciais em organismos vivos, porém em dosagens aceitáveis se torna essencial para a vida dos ecossistemas aquáticos, principalmente em corpos de água doce como reservatório (Rahman et al., 2013; Yiang et al., 2012).

3.1.3.8 Magnésio

Assim como o cálcio, o magnésio (Mg) que é medido em miligramas por litro (mg L^{-1}), é também determinante para dureza da água. Sendo considerado um dos parâmetros físico-químicos e metais da água, o magnésio passa a ser um dos agentes responsáveis na concentração de íons que caracteriza a formação da dureza da água. Como o cálcio, altas taxas de magnésio na água podem ser prejudiciais para os ecossistemas em corpos de água e para saúde humana (Avni et al., 2013).

3.1.3.9 Potássio

O potássio (K), medido em miligramas por litro (mg L^{-1}), sendo um elemento essencial para nutrição animal, é encontrado em maiores concentrações em rochas, resistindo a ação do tempo, e em menor concentração nos corpos de água doce acumulado em estruturas minerais pela biota aquática. Como é largamente utilizado na indústria, pricipalmente em fertilizantes para agricultura, certamente se torna efluente industrial entrando através de resíduos nos corpos de água naturais e altas concentraçõs de potássio podem indicar a ocorrência de fontes quentes e salmouras (Yustseven et al., 2005; Cetesb, 2009).

3.1.3.10 Sódio

O sódio (Na), medido em miligramas por litro (mg L^{-1}), é um parâmetro que é encontrada em todas as águas naturais. De natureza altamente solúvel, é um elemento ativo dentro dos organismos vivos e abundante na natureza. Altas concentrações de sódio em corpos de água podem contribuir diretamente para morte dos ecossistemas aquáticos e para saúde humana, de maneira tal que o aumento das concentrações de sódio na água podem ser oriundos de lançamentos de esgotos domésticos, efluentes industriais e do uso de sais em rodovias para controlar neve e gelo, principalmente, nos países da América do Norte e Europa (Cetesb, 2009; Furtado et al., 2011)

3.1.3.11 Amônia

A amônia (NH_4), medido em microgramas por litro (µg L^{-1}), é o resultado do despejo de resíduos industriais, domésticos e esgotos em corpos de água. Geralmente esse processo influencia a quebra de nitrogênio orgânico e matéria inorgânica no solo e água, excreção pela biota, redução

do gás nitrogênio no corpo de água por microrganismo e da troca de gases com a atmosfera (Cavalcante et al., 2013; Sims et al., 2012).

3.1.3.12 Nitrato

O nitrato (NO_3), medido em miligramas por litro (mg L^{-1}), é uma consequência do nitrogênio orgânico oxidado em corpos de água doce. Isso deve a concentrações de resíduos de efluentes domésticos, industriais e de esgotos. Quando grande concentração de nitrato nas amostras de águas, indica que o foco da poluição da água está próximo do local da coleta (Elmi et al., 2011; Cetesb, 2009).

3.1.3.13 Total de fósforo

O total de fósforo (totalP), medido em miligramas por litro (mg L^{-1}), aparece em águas naturais oriundos principalmente de descargas de efluentes, como os esgotos sanitários. Nestes, os detergentes superfosfatados empregados em larga escala domesticamente constituem a principal fonte. Alguns efluentes industriais, como os de indústrias de fertilizantes, pesticidas, químicas em geral, conservas alimentícias, abatedouros, frigoríficos e laticínios, apresentam fósforo em quantidades excessivas. As águas drenadas em áreas agrícolas e urbanas também podem provocar a presença excessiva de fósforo em águas naturais (Cetesb, 2009; Schoumans et al., 2013).

3.1.3.14 Fosfato

O fosfato (PO_4), medido em miligramas por litro (mg L^{-1}), aparece em corpos de água oriundas principalmente, de resíduos dos esgotos sanitários entre outros fatores. Assim como o nitrogênio, o total de fósforo constitui-se em um dos principais nutrientes para os processos biológicos, ou seja, é um dos chamados macro-nutrientes e de grande relevância para os ecossistemas aquáticos (Neal et al., 2000; Cetesb, 2009).

3.1.3.15 Sólidos totais em suspensão

Os sólidos totais em suspensão (STS), medido em miligramas por litro (mg L^{-1}), é a porção dos sólidos totais que fica retida em um filtro que propicia a retenção de partículas de diâmetro

maior ou igual a 1,2 µm. São caracterizados através de amostras líquidas ou sólidas com a finalidade de se verificar a possibilidade de degradação aeróbia/anaeróbia dos sólidos em suspensão, tendo sido empregados também quando em sólidos totais sedimentados, para estimar o conteúdo orgânico do sedimento, em corpos d'água (Bilotta e Brazier, 2008; Chapman, 1996).

3.1.3.16 Turbidez

A turbidez, medida em unidades nefelométricas de turbidez (NTU em inglês), é o parâmetro da qualidade da água que indica a redução da transparência devido a frequência de resíduos sólidos em suspensão. Essa transparência é diretamente influenciada pela redução da luz, e quanto maior a turbidez, menor a transparência. Um fator relevante dentro desse contexto, é quanto mais partículas sólidas flutuando na água, menor será a capacidade de luz, trazendo assim uma déficit de claridade na água que prejudica a fotossíntese para os ecossistemas subaquáticos (Atobatele e Ugwumba, 2008; Lessels e Bishop, 2012).

3.1.3.17 Clorofila-a

A clorofila-a, medida em miligramas por litro (mg L^{-1}), sendo um dos parâmetros de pigmentação da água, assim como os carotenóides e ficobilinas, são responsáveis pelo processo de fotossíntese. A clorofila-a é a mais universal das clorofilas e representa, aproximadamente, de 1% a 2% do peso seco do material orgânico em todas as algas planctônicas e é, por isso, um indicador da biomassa algal. Assim, a clorofila a é considerada a principal variável indicadora de estado trófico dos ambientes aquáticos (Novoa, 2012; Cetesb, 2009).

3.2 Regulamentação da qualidade da água no Brasil

No Brasil a qualidade da água é regulada pelo Conselho Nacional do Meio Ambiente (CONAMA) através da Resolução nº 357 de 2005 que dispõe sobre a classificação dos corpos de água e diretrizes ambientais para o seu enquadramento, bem como estabelece as condições e padrões de lançamento de efluentes, e dá outras providências. O CONAMA também através das Portarias do Ministério da Saúde nº 1469/2000 e 2914/2011 que estabelece os procedimentos e responsabilidades relativos ao controle e vigilância da qualidade da água para consumo humano e seu padrão de potabilidade, e fornece outras providências. As regulamentações visam estimular os

limites de valores máximos e mínimos aceitáveis dos parâmetros físico-químicos e metais, estabelecendo níveis individuais para cada substância em cada classe de água doce, assim como, águas salinas.

A Tabela 1 apresenta os valores máximos para os 17 parâmetros físico-químicos e metais extraídos pela análise fatorial que serão utilizados na classificação dos ciclos hidrológicos para a água doce de classe 2. A classe 2 das águas doces serve para: o abastecimento do consumo humano, após tratamento convencional; proteção da comunidade aquática; recreação de contato primário (natação, esqui aquático e mergulho); irrigação de hortaliças e plantas frutíferas e de parques, jardins, campos de esporte e lazer, com os quais o público possa a vir a ter contato direto; aquicultura e atividade de pesca.

Tabela 1. Valores máximos para os 17 parâmetros físico-químicos para a água doce de classe 2 de acordo com as três legislações.

Substância	Res 357/2005	Port 1469/2000	Port 2914/2011
Transp(Secchi)	-	-	-
Temp	-	-	-
pH	6 ↔ 9	6,5 ↔ 8,5	6 ↔ 9
OD	⩽ 5 mg O_2/L	⩽ 2 mg O_2/L	-
Cond Elétrica	-	-	-
FeTotal	⩽ 0.3 L^{-1}	⩽ 0.3 L^{-1}	⩽ 0.3 L^{-1}
Ca	-	-	-
Mg	-	-	-
Na	-	⩽ 200 mg L^{-1}	⩽ 200 mg L^{-1}
K	-	-	-
NH_4	⩽ 0.0 mg L^{-1}	⩽ 1.5 mg L^{-1}	⩽ 1.5 mg L^{-1}
NO_3	⩽ 10.0 mg/L	⩽ 10.0 mg/L	⩽ 10.0 mg/L
PO_4	⩽ 0.025 mg/L	-	-
Ptotal	⩽ 10.0 mg/L	-	-
STS	⩽ 500.0 mg/L	-	-
Clorofila-a	⩽ 10.0 mg/L	-	-
Turbidez	⩽ 100 uT	⩽ 5 uT	⩽ 5 uT

3.3 Técnicas estatísticas

3.3.1 Análise multivariada

Na atual sociedade da informação, percebe-se que existe em trânsito um volume acentuado de dados que cobrem inúmeros temas das atividades humanas em diferentes áreas de conhecimento. A complexidade e a extensão de vários fenômenos requer cada vez mais uma análise de muitas variáveis em óticas diferentes, trazendo assim a necessidade de procedimentos mais adequados na interpretação de dados e tomada correta de decisões. Além disso, o advento tecnológico, principalmente na área computacional, tem propiciado avanços e facilidades no que diz respeito à extração de informações e análise de dados em diversas áreas.

A técnica estatística de análise de dados multivariada têm comprovado sua efetividade para lidar com grandes volumes de informações complexas. Tratam-se de métodos multidimensionais que permitem a confrontação entre duas ou mais variáveis. Esse mecanismo extrai as tendências mais sobressalentes e hierarquizada, eliminando efeitos que perturbam a percepção global (Bake et al., 2008).

Para estatística, na resolução de problemas do mundo real, é necessário uma análise de todo contexto sobre o universo do domínio de conhecimento para tomadas de decisão. Para isso deve-se levar em consideração inúmeros fatores extraídos das variáveis iniciais, no qual cada uma tem seu peso e sua relevância na hora de uma escolha. Neste sentido, é aconselhável procedimentos sistemáticos, e não intuitivos, para se identificar os fatores mais significativos na tomada de uma decisão.

Assim, estabelecer relações, mensurar e manipular fatores que são considerados significativos ao entendimento do fenômeno em análise deve ser feito de forma mais precisa e efetiva. Os métodos estatísticos, para analisar variáveis, estão dispostos em dois grupos: estatística univariada que trata de variáveis de maneira isolada, e outro que trata as variáveis de forma conjunta, chamada de estatística multivariada (Vicini e Souza, 2005).

Neste trabalho foram utilizadas técnicas estatísticas multivariada. A análise fatorial foi utilizada para selecionar os parâmetros físico-químicos e metais da água mais relevantes para o processo de classificação. Já a análise discriminante foi utilizado como um classificador dos ciclos hidrológicos da UHE de Tucuruí.

3.3.2 Análise Fatorial

A Análise Fatorial é uma técnica estatística multivariada que objetiva reduzir o número de variáveis observáveis em fatores subjacentes não observáveis. Estes fatores refletem de forma precisa a representação do fenômeno em análise sem grande perda de informação dos dados originais para tornar os dados observados mais claros para a interpretação (Bakke et al., 2008; Hair et al., 1998; Zeller e Carmines, 1980).

Dentre as técnicas de AF o presente estudo utilizou a matriz de correlação, cujos elementos são os valores das correlações entre variáveis, ou seja, o grau de relacionamento entre elas. O teste de esfericidade de Bartlett é utilizado para testar a hipótese de que as variáveis são correlacionadas na população. A estatística teste é apresentada na Equação 3.1.

$$\chi 2 = -[(n-1) - \frac{2p+5}{6} \ln |R|] \quad (3.1)$$

Onde:

n = tamanho da amostra
p = número de variáveis
$|R|$ = determinante da matriz de correlação

A comunalidade é a proporção da variância que uma variável compartilha com todas as outras variáveis consideradas. É também a proporção de variância explicada pelos fatores comuns. Os autovalores (*Eigenvalues*) representam a variância total explicada por cada fator. As cargas fatoriais são as correlações simples entre as variáveis e os fatores. Já a matriz de fatores é composta pelas cargas dos fatores de todas as variáveis em todos os fatores extraídos. E os escores fatoriais são obtidos a partir das estimativas dos fatores.

A medida de adequação de *Kaiser-Meyer-Olkin* (KMO) é um índice usado para avaliar a adequação da análise fatorial. Os valores obtidos através deste teste variam entre 0 e 1, sendo aceitável valores acima de 0,5 (Vicini e Souza, 2005; Hair et al., 1998). A Equação 3.2 apresenta a estatística teste KMO.

$$KMO = \frac{\sum i * \sum_{i}^{r^2 ij} .}{\sum i * \sum_{j}^{r^2 ij} . + \sum i * \sum_{j}^{a^2 ij} .} \quad (3.2)$$

Onde:

$r^2 ij$ = é o coeficiente de correlação observado entre as variáveis i e j.

$a^2 ij$ = é o coeficiente de correlação parcial entre as mesmas variáveis, que é simultaneamente, uma estimativa das correlações entre os fatores. Os a_{ij} deverão estar próximos de zero, pelo fato de os fatores serem ortogonais entre si.

Neste estudo, os pré-requisitos para adequação da analise fatorial utilizados como critérios de escolha dos fatores que melhor representam a estrutura dos dados foram: o KMO maior que 0,50 e o teste de Bartlett para verificar a correlação entre as variáveis.

A quantidade de fatores que foram extraídos foi escolhida com base em critérios pré-determinados. E o critério escolhido e adotado no estudo foi de raiz latente que assume qualquer fator individual e explica a variância de pelo menos uma variável. Cada variável contribui com um valor 1 do auto-valor total. Logo, apenas os fatores que tenham raízes latentes ou autovalores maiores que 1 são considerados significantes, os demais fatores são descartados. Para melhorar a interpretação obtida pela análise foi utilizada a rotação fatorial *varimax* que é a mais comum entre as principais abordagens ortogonais, pois simplifica as colunas da matriz fatorial e apresenta também bons resultados (Jonson e Wichern, 2001; Hair et al., 1998).

Para análise de dados foi utilizado o software estatístico SPSS *Statistical Package for Social Science*, que permite realizar cálculos complexos e visualizar seus resultados de forma simples e autoexplicativas (Bahiense, 2013).

3.3.3 Análise discriminante

A AD é provavelmente uma das técnicas de classificação estatística mais antiga e difundida na comunidade científica. Seu objetivo é discriminar uma variável dependente categórica a partir de variáveis independentes utilizando modelos lineares, como a função de Fisher, e não lineares, como funções quadráticas, cúbicas, entre outras. E também, construindo regras para discriminação de classes de objetos dentro de conjunto de dados (Lachenbruch, 1975).

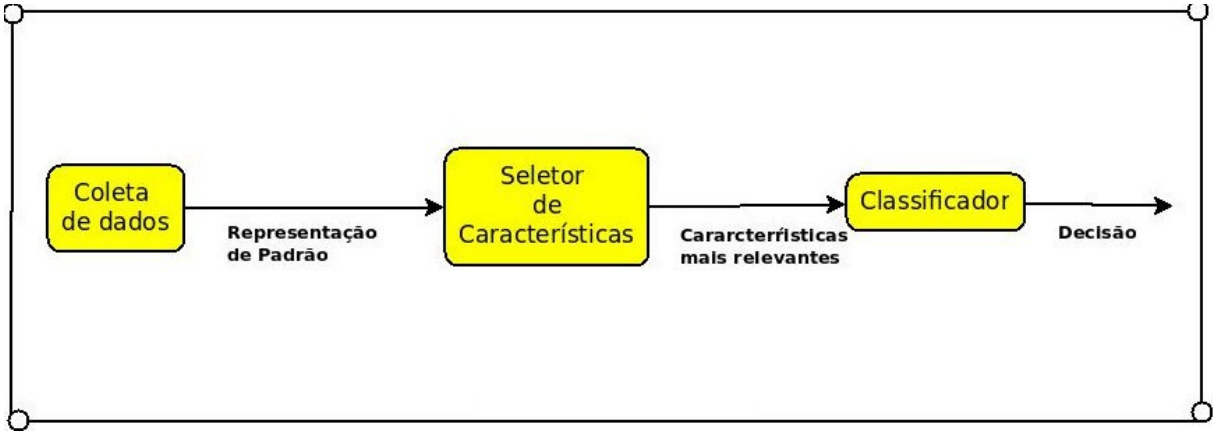
Na AD as amostras envolvidos devem ser divididas em dados de análise e de teste, pois a função discriminante utiliza do último para averiguar sua eficácia na classificação das classes. Para

isso, depois que a função foi obtida é necessário avaliar o nível de significância, sendo o λ de Wilks um dos critérios mais comuns. Esta estatística determina o critério de otimalidade (relacionado com a probabilidade de erro de classificação), a qual avalia a separação entre as classes para cada variável a partir da razão entre a dispersão das médias das classes e a variância total (Cavalcante, 2013; Hair et al., 1998; Johnson e Wichern, 2001). Para classificação dos ciclos hidrológicos a discriminação estabeleceu pesos às variáveis iniciais para maximizar a variância entre as classes através da função discriminante.

3.4 Classificadores computacionais

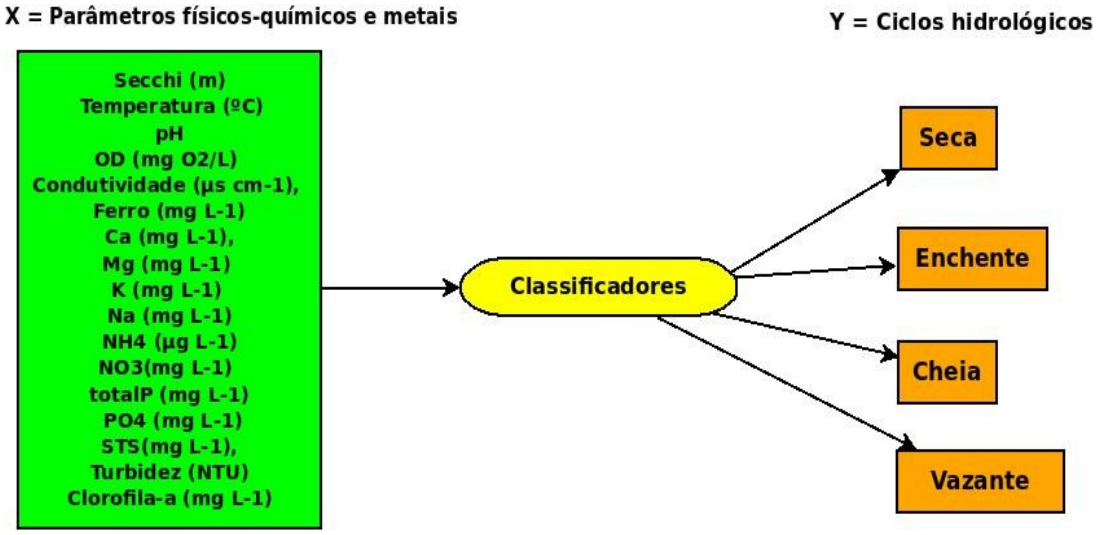
Um *classificador* **F** é um mapeamento $F : \mathbb{R}^k \rightarrow \{1..., Y\}$, onde *K* é a dimensão do vetor de entrada $x \in \mathbb{R}^k$ e o *rótulo* y ∈ {1...., *Y*} é a classe. Quando se treina um classificador computacional usando aprendizado supervisionado, é dado um *conjunto de treinamento* T={(x_1, y_1),..., (x_n, y_n)} contendo *N exemplos* de (x,*y*). A Figura 8 apresenta o fluxo de alguns dos processos principais envolvidos em um sistema de classificação de padrões.

Figura 8: Fluxo dos processos de um classificador (Fonte: Ferreira, 2007).



Neste trabalho foram utilizados classificadores computacionais para predição de 4 ciclos hidrológicos definidos como seca, enchente, cheia e vazante, levando em cosideração como variáveis de entrada, os valores das análises dos parâmetros físico-químicos e metais da água, coletados no reservatório da UHE de Tucuruí. A Figura 9 apresenta o modelo adotado no estudo com suas respectivas entradas e saídas desejadas, juntamenente com as 4 classes associadas.

Figura 9: Modelo de classificação dos ciclos hidrológicos.



Percebe-se que na Figura 9 os parâmetros físico-químicos e metais da água são os atributos em x_i, e os ciclos hidrológicos são representados em y_i com suas classes associadas: seca, enchente, cheia e vazante.

Para classificação este trabalho adota os algoritmos disponíveis no software WEKA (*Waikato Environment for Knowledge Analysis*)[1]. Assim, a próxima seção discutirá brevemente este software e as seções seguintes listarão os principais classificadores utilizados. Como o objetivo deste trabalho não é discutir cada um desses classificadores em profundidade, busca-se prioritariamente ilustrar como os mesmos são usados no WEKA.

3.4.1 WEKA

O WEKA é reconhecido pela comunidade científica como um sistema de referência em aprendizado de máquina e mineração de dados. O sistema é formado por um conjunto de implementações de algoritmos de diversas técnicas de inteligência computacional, e foi implementado na linguagem de programação Java, tornando-o acessível nas principais plataformas computacionais (Witten e Frank, 2005; Hall et al, 2009).

O WEKA inclui algoritmos de regressão, classificação, agrupamento, regras de associação e seleção de parâmetros (atributos). Atualmente está na versão 3.6.10, sendo organizado em três

1 Software de domínio público desenvolvido na Universidade de Waikato na Nova Zelândia. Disponível em http://www.cs.waikato.ac.nz/ml/weka/.

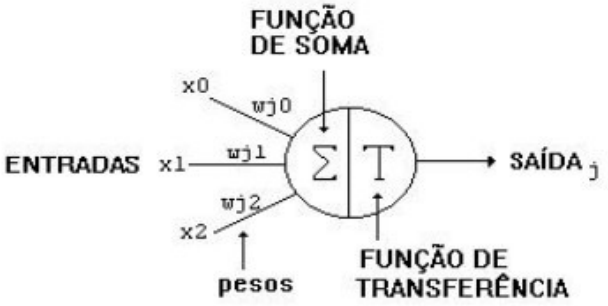
módulos de operação. O primeiro, “*simple Command Line Interface*" (CLI), a interação do usuário com WEKA ocorre através de linhas de comando. O “*Explorer*" é considerado o principal módulo, pois executa a interface gráfica para execução dos algoritmos de aprendizado de máquina suportados pelo WEKA. E o terceiro, é o módulo “*Experimenter*" no qual o usuário, também por meio de interface gráfica, executa testes estatísticos em diferentes algoritmos simultaneamente a fim de avaliar os resultados obtidos.

Antes de utilizar o pacote WEKA, os dados devem ser convertidos para um dos formatos de arquivo suportados pelo WEKA. Neste trabalho o formato adotado é próprio do WEKA denominado de arff (*Attribute-Relation File Format*). O arquivo no formato arff é um arquivo ASCII composto de três partes. A primeira parte, chamada de relação indicada pelo marcador @relation que fica na primeira linha do arquivo, identifica o nome da relação. A segunda parte, iniciada sempre com o marcador @attribute, contém a lista de todos os parâmetros (ou atributos), onde se deve definir o tipo de parâmetro ou os valores que eles podem assumir, ao utilizar os valores estes devem estar entre chaves separados por vírgula. A terceira, encontra-se logo após a linha com o marcador @data e consiste das instances, isto é, os dados a serem minerados com o valor dos parâmetros para cada instância (linha). O formato arff não especifica qual parâmetro é a classe, pois isso permite que o mesmo seja mais flexível. Assim, o que o WEKA entende por instance, reúne o que a notação aqui adotada chama de instance *x* mais o rótulo *y*.

3.4.2 Redes neurais artificiais

Redes Neurais Artificiais (RNAs) são sistemas computacionais distribuídos compostos de unidades de processamento simples que computam funções matemáticas, sendo densamente interconectadas. Tais unidades são conhecidas como neurônios artificiais que ficam dispostas em uma ou mais camadas intermediárias por um grande número de conexões. Geralmente essas conexões possuem pesos associados que regula a entrada recebida por cada neurônio na rede para posteriormente produzir a saída (Kezunovic e Rikalo, 1996). O comportamento inteligente de uma RNA vem das interações entre as unidades de processamento da rede. A Figura 10 exibe a estrutura de um neurônio artificial.

Figura 10: Estrutura de um neurônio artificial.



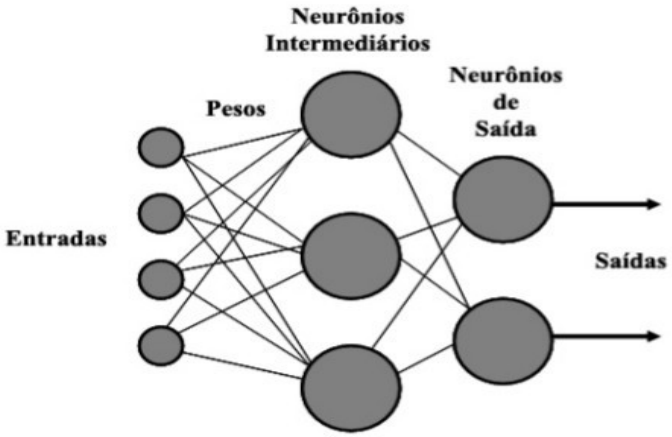
O ajuste de pesos é realizado em função de um cálculo que aponta a quantidade de erro do resultado (saída). Este ajuste procura corrigir os pesos de modo que se produza a saída desejada diante da respectiva entrada. Dentre os diversos tipos de cálculos para este fim, a Regra Delta é a mais utilizada.

Existem muitos tipos de algoritmos de aprendizado para redes neurais artificiais. Estes diferem entre si principalmente pelo modo como os pesos são modificados. Os mais conhecidos são o aprendizado supervisionado e o não supervisionado. No aprendizado supervisionado a rede neural recebe um conjunto de entradas e saídas de dados. Neste tipo de algoritmo, o aprendizado ocorre por meio dos ajustes nos pesos, os quais são modificados até que os erros entre os padrões de saída gerados pela rede tenham um valor desejado ou próximo do desejado. Já no aprendizado não supervisionado, a rede neural trabalha os dados de forma a determinar algumas propriedades do conjunto de dados. A partir destas propriedades é que o aprendizado é constituído.

Denomina-se iteração ou época uma apresentação de todos os pares (entrada e saída) do conjunto de treinamento no processo de aprendizado. A correção dos pesos numa iteração pode ser executada de modo *standard* ou em *batch*. No modo *standard* o erro é estimado a cada apresentação de um conjunto de treino à rede. Enquanto que no modo *batch* estima-se o erro médio após todos os exemplos do conjunto de treinamento serem apresentados à rede.

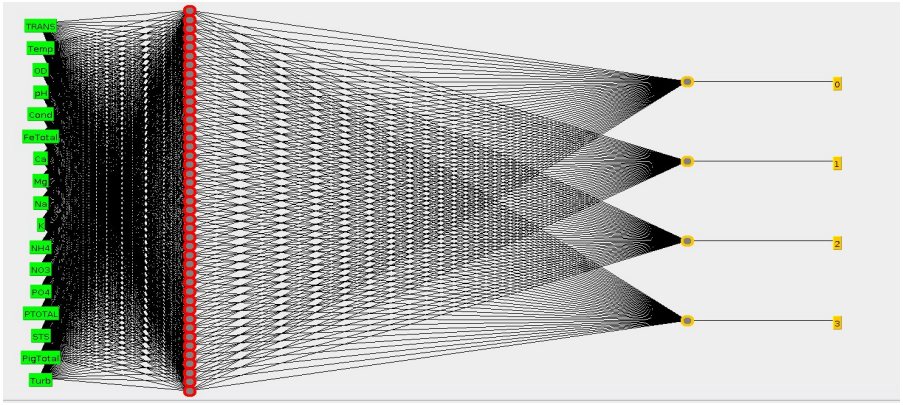
Dentre os paradigmas neural tem-se o perceptron de múltiplas camadas (MLP), que consiste de um conjunto de nós fonte os quais formam a camada de entrada, uma ou mais camadas escondidas e uma camada de saída. Uma MLP é uma generalização do modelo perceptron, forma mais simples de uma rede neural usada para classificação de padrões. A Figura 11 exibe um rede neural do tipo perceptron de múltiplas camadas.

Figura 11: Exemplo de uma rede neural MLP.



O número de neurônios na camada de entrada é determinado pela dimensão do espaço observado. Em contrapartida, a quantidade de neurônios na camada de saída é determinado pela dimensionalidade requerida pela resposta. Por exemplo, na classificação dos ciclos hidrológicos o número de neurônios na camada de entrada depende dos valores dos parâmetros físico-químico e metais da água. Quanto ao número de neurônios na camada de saída, seria dada por 4 neurônios que representam: seca, enchente, cheia e vazante (adotada neste trabalho). Além disso, uma MLP leva em consideração os seguintes aspectos: determinação do número de camadas escondidas, de neurônios em cada uma destas camadas, assim como a especificação do tipo de algoritmo de aprendizado supervisionado utilizado. O algoritmo utilizado neste trabalho para treinamento das redes neurais MLP foi o *backpropagation* também chamado de regra delta generalizada (Duda et al., 2010; Haykin, 1998; Braga et al., 2007). A Figura 12 apresenta o modelo da RNA adotado no estudo extraído do *software* WEKA.

Figura 12: RNA adotada no estudo.



O algoritmo consiste em um processo de aprendizagem supervisionada que utiliza um conjunto pré-determinado de pares de exemplo de entrada e saída para ajustar os pesos da rede através de um esquema de correção de erros realizado em ciclos de propagação. O *backpropagation* é dividido em duas fases: a primeira fase consiste em propagar (*forward*) o vetor de entrada a partir da primeira até a última camada e comparar o valor da saída com o valor desejado. A segunda fase consiste em retropropagar (*backward*) o erro partindo da última camada até chegar a camada de entrada ajustando os pesos dos neurônios das camadas intermediárias. Após ajustar todos os pesos da rede, é apresentado mais um conjunto de exemplos encerrando uma época. Este processo é repetido até que o erro torna-se aceitável para o conjunto de treinamento, momento denominado de convergência da rede.

O comportamento de uma rede neural MLP durante seu treinamento varia mediante a alteração de algumas de suas características (Haykin, 1998):

- Inicialização dos pesos: Os pesos das conexões entre os neurônios podem ser inicializados uniformemente ou de forma aleatória.
- Taxa de aprendizado: A taxa de aprendizado controla a velocidade do aprendizado, aumentando ou diminuindo o ajuste de pesos que é efetuado a cada iteração durante o treinamento. Intuitivamente, seu valor deve ser maior que 0 e menor que 1. Se a taxa de aprendizado for muito pequena, o aprendizado ocorrerá muito lentamente. Caso a taxa seja muito grande (maior que 1), a correção seria maior do que o erro observado, fazendo com que a rede neural ultrapassasse o ponto de aprendizado ótimo, tornando o processo de treinamento instável.
- Parametrização da função de transferência: Também conhecida como limiar lógico, essa função é quem define e envia para fora do neurônio o valor passado pela função de ativação. A função de ativação pode ter muitas formas e métodos. As mais conhecidas são: função linear, função sigmóide e função exponencial.

Neste trabalho foi utilizada a rede neural MLP implementada no WEKA. Sendo os principais parâmetros do modelo descrito abaixo:

- -L: Corresponde à taxa de aprendizado utilizada pelo algoritmo *backpropagation*. Este valor deve ser entre 0 e 1 (Padrão é 0.3).

- -M: Taxa de momento para o algoritmo *backpropagation*. Este valor deve ser entre 0 e 1 (Padrão é 0.2).
- -N: Este parâmetro corresponde ao número de épocas para treinamento da rede. O Padrão é 500.
- -H: Corresponde à quantidade de neurônios em camadas ocultas que podem ser criadas na rede.

Por exemplo o (-H 3, 2) cria duas camadas intermediárias com 3 e 2 neurônios, respectivamente. Outra forma de representar pode ser através do uso de letras: a opção a corresponde a (números de parâmetros + número de classes)/2; as outras opções são: i (número de parâmetros), o (número de classes) e, t (números de parâmetros + número de classes).

Dentre os paradigmas neurais disponíveis, as RNAs multicamada, com o algoritmo de treinamento supervisionado *backpropagation*, é uma das mais utilizadas na práatica e, pelas características do problema abordado neste estudo, foi o paradigma adotado.

3.4.3 Support vector machine

As máquinas de vetores de suporte (SVM - Support Vector Machine) constituem uma técnica de aprendizado de máquina fundamentada na teoria de aprendizado estatístico. As SVMs vem sendo muito utilizada pela comunidade de Aprendizado de Máquina (AM) nos últimos anos, pois os resultados da aplicação dessa técnica são iguais ou superiores a outros técnicas de IC. O objetivo de um classificador SVM consiste em encontrar um hiperplano (supérficie de decisão) que máximize a separação no espaço de classes (Vapnik, 1995; Haykin, 1998; Faceli et al., 2011).

Um ponto importante no algoritmo de aprendizagem por vetor de suporte é a função kernel ou função núcleo. Na literatura, várias possibilidades de kernel SVM são apresentadas em aplicações envolvendo reconhecimento de padrões, tais como: linear, polinomial, sigmóide e funções de base radial (RBF - *Radial Basis Function*). As SVMs, e outros métodos kernel, podem ser caracterizados como uma função de estimação Φ definida pela equação 3.3.

$$\frac{1}{V}\sum_{v=1}^{V} .(\varphi(x_v),x_v)+\lambda\|\not{\sigma}\|^2_{H_k} \qquad (3.3)$$

Onde:

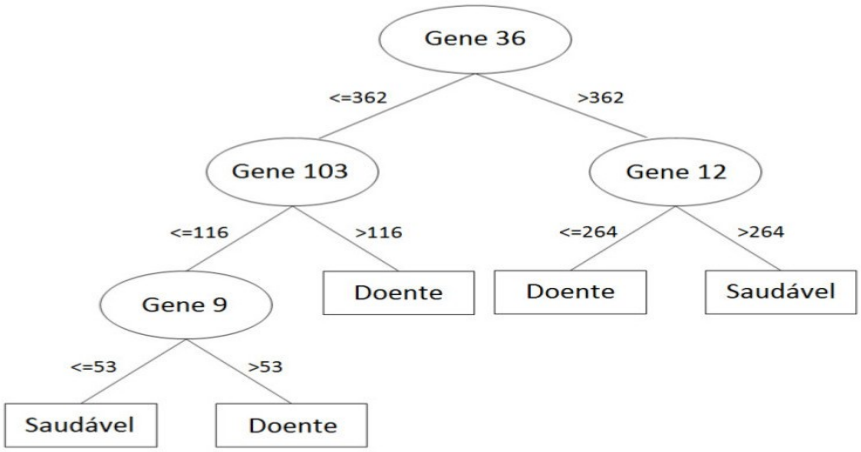
- H_k = corresponde ao espaço euclidiano gerado pelo kernel *K*;
- $\Phi = h + b$, onde *h* corresponde ao produto do vetor peso (ω) pelo vetor de suporte (x_v) com $h \in H_k$;
- *b* = corresponde ao bias, $b \in R$;
- $L(\Phi(x_v), y_v))$ = corresponde a função perda (risco fundamental);
- $\lambda =$ corresponde a autovalores;
- *V* = corresponde ao número de exemplos de treino.

O *default* no WEKA do classificador SVM é o kernel POLY, e um módulo extra para o Kernel RBF. Na RBF o classificador implementa uma versão da LibSVM otimizada para lidar com o problema de multi-classes, as SVMs são organizadas na LibSVM no esquema one-versus-one (ou *all-pairs*)(EL-Manzalawy e Honavar, 2005; Riflin e Klautau, 2012).
No escopo desse trabalho, foram utilizadas as funções kernel RBF e POLY. Assim, os principais parâmetros utilizados foram o -C, parâmetro de penalidade do termo de erro, e o -G, largura da função dos kernels.

3.4.4 Random forest

O classificador Random Forest utiliza mecanismos de dividir para conquistar para resolver problemas de decisão. A idéia é que problemas mais complexos podem ser divididos em problemas mais simples em forma recursiva. Essas soluções podem ser combinadas em forma de árvore para gerar uma solução do problema complexo. A proposta é que dividindo o espaço de *instances* em subespaços que são ajustados em diferentes modelos, sendo formalmente estruturado em um grafo acíclico em que cada nó, ou é um nó de divisão com dois ou mais sucessores dotado de um teste condicional, ou um nó folha rotulado com uma nova classe (Faceli et al., 2011). A Figura 13 apresenta um exemplo de árvore de decisão com seus nós de decisão e nós-folhas associados as classes desejadas pelo modelo.

Figura 13: Exemplo de estrutura de uma árvore de decisão (adaptado de Oshiro, 2013).



Observa-se na Figura 13 que os nós na árvore de decisão estão representando os gens numerados, e para cada decisão é associado uma escala contendo um valor referencial que determinará a classe saudável ou doente de pacientes em estudo.

Nessa direção, o preditor RF constrói diversas árvores de decisão que serão usadas para classificar um novo exemplo por um mecanismo de voto majoritário. Cada árvore de decisão usa um subconjunto de atributos selecionados aleatoriamente a partir do conjunto original, contendo todos os atributos (Oshiro, 2013).

3.4.5 K-nearest neighboors

Os classificadores vistos anteriormente são caracterizados pelo fato de utilizarem os dados de treinamento para construírem um modelo de classificação, o qual, uma vez encontrado e testado, estará pronto para testar qualquer padrão novo. Diferentemente desses classificadores, o classificador K-vizinhos mais próximos (KNN - *K-Nearest Neighbors*) utiliza os próprios dados de treinamento como modelo de classificação, isto é, para cada novo padrão que se quer classificar, utiliza-se os dados do treinamento para verificar quais são os exemplos nessa base de dados que são "mais próximos" do padrão em análise. A cada novo padrão a ser classificado faz-se uma varredura nos dados de treinamento, o que provoca um grande esforço computacional.

Considerando um conjunto de treinamento e seja $z = (x_1 , \ldots , x_n)$ uma nova *instance*, ainda não classificada. A fim de classificá-la, calcula-se as distâncias, através de uma medida de similaridade, entre *z* e todos os exemplos do conjunto de treinamento e considera-se os *K* exemplos mais próximos (com menores distâncias) em relação *z*. Verifica-se então, qual a classe que aparece

com mais frequência, entre os *K* vizinhos encontrados. O padrão *z* será classificado de acordo com a classe *y* mais frequente dentre os *K* exemplos encontrados.

A distância entre duas *instances* é calculada utilizando-se uma medida de similaridade. Uma medida de similaridade bastante popular é a distância euclidiana d_{eucl} (Haykin, 1998; Witten e Frank, 2005). Tal medida calcula a raiz quadrada da norma do vetor diferença entre os vetores z e $\hat{z}$ (Equalção 3.4).

$$d_{eucl} = \sqrt{\sum_{i=1}^{k} .(z_i, \hat{z}_i)^2} \qquad (3.4)$$

O KNN no WEKA é implementado na classe IBK e seus principais parâmetros são:

- -N : número de centros (ou K).
- -S: esta opção gera aleatoriamente os centros.

3.4.6 Seleção automática do modelo

Esta seção aborda o importante problema de seleção automática do modelo, isto é, a escolha de valores dos parâmetros a serem usados em técnicas de classificação em IC. Um exemplo desse mecanismo seria o ajuste do número de neurônios na camada escondida de uma rede neural que frequentemente, é realizada por meio de validação-cruzada (*cross-validation*) (Witten e Frank, 2005).

Esta é uma abordagem computacionalmente custosa, mas é importante para evitar ajustar o modelo usando-se o mesmo conjunto a ser usado no teste, pois isso faria com que o modelo eventualmente se viciasse no conjunto de teste. Se esse fosse o caso, a taxa de erro no conjunto de teste não exprimiria adequadamente a capacidade de generalização do modelo.

Tendo-se o erro de classificação no conjunto de validação como figura de mérito, a melhor combinação dos possíveis valores nos parâmetros dos classificadores foram procuradas com o auxílio de um *grid* (produto cartesiano dentre as opções sugeridas ao programa de busca.

Cada parâmetro compondo o *grid* (cada eixo cartesiano) foi especificado a partir de incremento linear ou geométrico (usando-se adição ou multiplicação, respectivamente). Em ambos os incrementos as opções dos parâmetros dos classificadores são especificadas de acordo com um

valor mínimo V_{min} um valor máximo V_{max} e o número de passos V_p. O fator de variação das opções dos parâmetros de um determinado classificador utilizando o incremento linear V_l é dado por:

$$V_l = \frac{V_{max} - V_{min}}{V_p - 1} \tag{3.5}$$

No caso do incremento geométrico o fator de variação V_g é igual a V_l. A Tabela 2 apresenta o *grid* adotado no estudo para o procedimento de seleção do modelo.

Tabela 2: Grid de seleção do modelo.

Classificador	Parâmetros	Incremento	Min	Max	V_g ou V_l	Valor *Grid*	Pontos Grid
RNA	H	Linear	11	171	32	11, 43, 75, 107, 139 e 171	90
	L	Linear	0.1	0.9	0.2	0.1, 0.3, 0.5, 0.7 e 0.9	
	M	Logarítimica	0.1	0.9	0.2	0.1, 0.3 e 0.9	
RF	I	Linear	100	1000	100	100, 200, 300, … 1000	10
SVM-RBF	G	Logarítimica	0.1	100	10	0.1, 1, 10 e 100	16
	C	Logarítimica	0.1	100	10	0.1, 1, 10 e 100	
SVM-POLY	G	Logarítimica	0.1	100	10	0.1, 1, 10 e 100	16
	C	Logarítimica	0.1	100	10	0.1, 1, 10 e 100	
KNN	C	Linear	1	15	2	1, 3, 4, 5, , 7, 9, 11, 13 e 15	8

3.5 Avaliação das técnicas de classificação

3.5.1 Medidas de desempenho

Alguns classificadores são capazes de prover *scores* de confiança $f_i(x)$ para cada classe *i = 1*,...., Y, tais como a probabilidade de distribuição sobre *y*. Por conveniência, pode-se assumir que todo classificador retorna um vetor ($f_1(x)$,.....,$f_y(x)$) com *Y scores*. Se o classificador naturalmente não retornar *scores* de confiança, o vetor de *scores* é criado com um *score* unitário $f_j(x)=1$ para a classe *j* sugerida pelo classificador, enquanto que os *scores* das outras classes são zero: $f_1(x) = 0$, *i*

≠ *j*. Assim, a decisão final é sempre baseada no valor máximo dos *scores* (chamada regra *max-wins*):

$$F(x)=\arg\max_{i=1,\ldots,J} f_i(x) \tag{3.6}$$

Um *conjunto de teste* {(x_1,y_1,....., x_R,y_R} contendo *R* exemplos e disjunto do conjunto de treino pode ser usado para calcular a taxa de erro de classificação

$$E_f=\frac{1}{R}\sum_{r=1}^{R}.\tau\left(F(x_r)\neq y_r\right) \tag{3.7}$$

onde τ é a função indicador, que é um (1) caso o argumento seja verdadeiro e zero (0) caso contrário. O erro E_f é uma estimativa da capacidade de generalização do classificador (Witten e Frank, 2005).

A taxa de erro varia entre 0 e 1, e valores próximos a 0 são índices melhores nas predições dos classificadores. Faceli et al. (2011) afirmam que a taxa de acerto, apresentada na Equação 3.8, serve como complemento da taxa de erro.

$$ac(f)=1-E_f \tag{3.8}$$

A taxa de acerto também varia entre 0 a 1, entretanto ao contrário da taxa de erro, valores próximos a 1 apresentam melhores predições nas classificações. Um ponto relevante dentro desse contexto é que tanto a taxa de erro ou acerto, entre outras medidas de desempenho de um classificador podem ser obtidas a partir da matriz de confusão, cuja dimensão corresponde ao número de classes existentes em um determinado conjunto de *instances*. Na diagonal principal estão a quantidade de acertos em cada classe e os elementos fora desta, correspondem a quantidade de erros.

Na matriz de confusão (M_c) as linhas representam as classes verdadeiras, e as colunas, as classes preditas pelo classificador. Logo, cada elemento m_{ij} de uma M_c apresenta o número de exemplos da classe *i* classificados como pertencentes a classe *j*. Para *Y* classes, M_c tem a dimensão *Y x Y* (Oshiro, 2013; Faceli et al., 2011). A Tabela 3 apresenta um exemplo de matriz de confusão para um problema de duas classes.

Tabela 3: Matriz de confusão para um problema de duas classes.

	Classe	Predita
	+	-
+	VP	FN
-	FP	VN

Onde:

- VP: Vedadeiros positivos equivalem a todos exemplos classificados corretamente.
- VN:Vedadeiros negaivos equivalem a todos exemplos classificados como verdadeiros incorretamente.
- FP: Falsos positivos equivalem a todos os exemplos classificados corretamente como negativos.
- FN: Falsos negativos equivalem a todos os exemplos classificados incorretamente como negativos .

Assim como pode-se extrair a taxa de erro e taxa de acerto através da M_c, Equações 3.7 e 3.8, respectivamente, também é possível obter outras métricas de desempenho de um classificador como taxa de especificidade (*esp(f)*, Equação 3.12), taxa de sensibiidade (*sens(f)*, Equação 3.11), taxa de precisão (*precisao(f)*, equação 3.13) entre outras.

$$E_f = \frac{FP+FN}{n} \tag{3.9}$$

$$ac(f) = \frac{VP+VN}{n} \tag{3.10}$$

$$sens(f) = \frac{VP}{VP+FN} \tag{3.11}$$

$$esp(f) = \frac{VN}{VN + FP} \tag{3.12}$$

$$precisao(f) = \frac{VP}{VP + FP} \tag{3.13}$$

Outra métrica que também pode mensurar o desempenho de um classificador é a medida *F*, que é uma média harmônica obtida através das taxas de precisão e sensibilidade, Equação 3.14 (Silva et al., 2012).

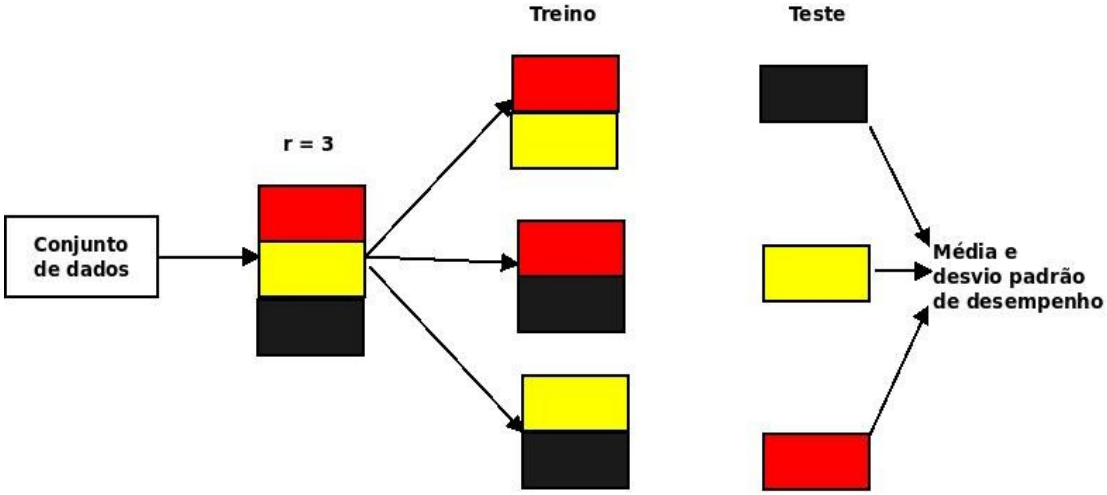
$$medida - F(f) = 2[\frac{precisao(f) * sens(f)}{precisao(f) + sens(f)}] \tag{3.13}$$

3.5.2 Validação cruzada

A validação dos modelos de classificação deste trabalho foi feito através da validação cruzada (*cross-validation*) com a qual os dados são fragmentados em dois subconjuntos, denominados conjunto de treinamento e conjunto de teste.

De modo geral a técnica de validação cruzada divide a base de dados em *r* partes (*r-fold cross-validation*). Destas, *r-1* são utilizadas para o treinamento e uma serve como base de testes. O processo é repetido *r* vezes, de forma que cada parte é utilizada uma vez como conjunto de testes. Ao final, a correção total é calculada pela média aritmética e desvio padrão dos resultados obtidos em cada etapa, obtendo-se assim uma estimativa da qualidade do modelo de conhecimento gerado e permitindo análises estatísticas (Santos et al., 2009). O processo é ilustrado na Figura 14.

Figura 14: Exemplo de validação cruzada (Fonte: Faceli et al., 2011).



Como se pode observar no exemplo da Figura 14, a validação cruzada com o método (*r-fold*), dividiu o conjunto de dados em três subconjuntos *r*, depois que foi feito o treinamento a partir da combinação de conjunto dois a dois, e por fim se realiza avaliação de desempenho.

3.5.3 t-Student test

Neste trabalho, o *t-Student test* foi utilizado para comparar as taxas de acerto dos classificadores utilizados, com o qual se testou as hipóteses abaixo, a partir da Equação 3.15 (Scudino, 2008).

- $H_0: \overline{X}_1 = \overline{X}_2$ não existe diferença significativa entre as médias dos classificadores 1 e 2.
- $H_1: \overline{X}_1 \neq \overline{X}_2$ existe diferença significativa entre as médias dos classificadores 1 e 2.

A estatística de teste será dada por:

$$t = \frac{(\overline{X_1} - \overline{X_2})}{S_{x_1 x_2} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \qquad (3.15)$$

Em que $S_{x_1 x_2}$ é:

$$S_{x_1 x_2} = \sqrt{\frac{(n_1 - 1) S_{x_1} + (n_1 - 1) S_{x_2}}{n_1 + n_2 - 2}} \quad (3.16)$$

Sendo o grau de liberdade para esses casos igual a:

$$df = n_1 + n_2 - 2 \quad (3.17)$$

Onde:

- $\overline{X}_1$ corresponde a média da taxa de acerto do classificador 1.
- $\overline{X}_2$ corresponde a média da taxa de acerto do classificador 2.
- n_i é o número de *folds* para cada classificador.
- Sx_i corresponde ao desvio padrão da taxa de acerto do classificador i:

A idéia básica é que exista uma hipótese nula H_0 e outra alternativa H_1 para depois confrontrar os resultados obtidos em valor tabelado t_{tab} da distribuição *t-Student* de acordo com os graus de liberdade *df*. O valor calculado t_{calc} é extraído na equação 3.14 e o valor de t_{tab} é obtido conforme a Tabela 4, na qual apresenta valores de significância em até 20 unidades.

Na resposta dos testes de hipóteses, um valor é comparado com o nível de significância previamente escolhido, sendo chamado de *p-valor* ou valor *p*. O *p-valor* (nível de significância observado) é o menor nível de significância em que H_0 seria rejeitada, quando um procedimento de teste específico é usado em um determinado conjunto de dados. Assim, quando *p-valor* < α implica na rejeição de H_0 no nível α . Ou se *p-valor* > α implica na não rejeiçãode H_0 no nível α. Então, em vários estudos as respostas poderão vir referenciando o nível de significância ou *p-valor*.

Tabela 4: Valores de *t*, segundo o grau de liberdade e o valor de α (Dancey e Reidy, 2006}

Grau de Liberdade (*df*)	10%	5% (α)	1%
1	6,31	12,71	63,66

2	2,92	4,30	9,92
3	2,35	3,18	5,84
4	2,13	2,78	4,60
5	2,02	2,57	4,03
6	1,94	2,45	3,71
7	1,90	2,36	3,50
8	1,86	2,31	3,36
9	1,83	2,26	3,25
10	1,81	2,23	3,17
11	1,80	2,20	3,11
12	1,78	2,18	3,06
13	1,77	2,16	3,01
14	1,76	2,14	2,98
15	1,75	2,13	2,95
16	1,75	2,12	2,92
17	1,74	2,11	2,90
18	**1,73**	**2,10**	**2,88**
19	1,73	2,09	2,86
20	1,73	2,09	2,84

O presente trabalho utilizou validação cruzada com dez *folds*, tanto para técnicas computacionais como para técnicas estatísticas, e também aplicou a Equação 3.15 para averiguar a significância dos resultados obtidos no desempenho dos classificadores com nível de significância $\alpha = 5\%$.

CAPÍTULO 4

Resultados e discussões

4.1 Resultados da análise fatorial

Conforme mencionado na Seção 3.3.2 a análise fatorial teve como objetivo reduzir o número de variáveis observáveis em fatores subjacentes não observáveis, identificando as variáveis mais representativas para a proposta do trabalho.

A Tabela 5 apresenta as comunalidades, KMO e teste de Bartlett obtidos antes da análise fatorial. Analisando a mesma observa-se que o KMO e as comunalidades resultantes da análise fatorial ficaram acima de 0.5. Além disso o teste de Bartlett mostrou-se significativo, ou seja abaixo de 0.5, indicando a adequação da técnica à proposta do trabalho.

Tabela 5: Comunalidades, KMO e teste de Bartlett da análise fatorial.

KMO	=	0,763
Teste de esfericidade de Bartlett	sig	0,000
=	Comunalidades	=
Variáveis	Inicial	Extração
Transp	1	0,772
Temp	1	0,541
OD	1	0,645
pH	1	0,688
Cond	1	0,660
FeTotal	1	0,725
Ca	1	0,689
Mg	1	0,726
Na	1	0,814
K	1	0,763
NH_4	1	0,728
NO_3	1	0,686
PO_4	1	0,622
PTOTAL	1	0,838

STS	1	0,713
PigTOTAL	1	0,703
Turb	1	0,761

Através do *software* SPSS, inicialmente foram extraídas as combinações de variáveis que explicam o maior percentual de variância. Em seguida, foram obtidas as combinações que apresentavam percentuais cada vez menores de variância. A quantidade de fatores que foram extraídos foi escolhida com base no critério de raiz latente.

A Tabela 6 apresenta o resultado da AF com os fatores extraídos explicando 71,01% dos dados na matriz rotacionada, com 26,39% de variância para o primeiro fator.

Tabela 6: Percentual de variância dos fatores obtidos na Análise Fatorial.

Componente	Somas de extrações de cargas ao quadrado		Somas de rotação de cargas ao quadrado		
	% de Variância	Acumulativo	Total	% de Variância	Acumulativo
1	28.448	28.448	4.486	26.385	26.385
2	12.349	40.797	1.692	9.995	36.341
3	9.007	49.804	1.582	9.308	45.648
4	7.924	57.728	1.468	8.638	54.284
5	7.333	65.061	1.430	8.409	62.693
6	5.955	71.015	1.415	8.322	71.015

A rotação fatorial ortogonal é a mais comum e entre as principais abordagens ortogonais, a utilizada no trabalho foi a *varimax*, pois simplifica as colunas da matriz fatorial e apresenta também bons resultados (Hair et al., 1998). A Tabela 7 apresenta a matriz fatorial com valores rotacionados com *varimax* utilizados no trabalho.

Tabela 7: Matriz fatorial rotacionada usando *varimax*.

Parâmetros	1	2	3	4	5	6
Transp (*)	-0,768	-0,080	0,163	-0,217	0,064	-0,318
Temp	-0,457	-0,260	0,007	-0,212	0,046	0,486
OD	0,112	-0,053	-0,745	0,076	0,235	0,115
pH	0,149	-0,246	-0,475	0,549	0,069	0,272
Cond	0,237	-0,004	0,108	-0,060	0,714	0,281

FeTotal (*)	0,752	0,177	0,339	-0,036	-0,093	-0,057
Ca	-0,034	0,082	-0,073	-0,129	0,766	0,269
Mg	0,061	0,094	0,011	0,833	0,131	0,036
Na	-0,037	0,876	0,067	-0,064	0,095	0,166
K	0,252	0,793	0,116	0,087	-0,219	-0,032
NH_4	0,352	0,162	0,632	0,229	0,152	0,320
NO_3	0,378	-0,088	0,357	0,506	0,350	-0,171
PO_4 (*)	0,740	0,004	0,098	-0,106	-0,181	-0,147
PTotal (*)	0,865	0,199	0,205	-0,037	-0,057	0,062
STS (*)	0,827	-0,028	-0,094	0,077	0,047	0,111
Clorofila-a	0,096	0,186	-0,046	0,040	-0,005	0,809
Turb (*)	0,843	-0,051	-0,188	0,109	-0,020	-0,028

Na Tabela 7 é possível observar que as variáveis Transp, FeTotal, PO_4, PTotal, STS e Turb, marcadas, foram extraídas para o primeiro fator que representou 26.39% da variância total. Observa-se ainda a correlação negativa entre transparência (-0.768) e as demais variáveis deste fator, resultado este esperado na literatura.

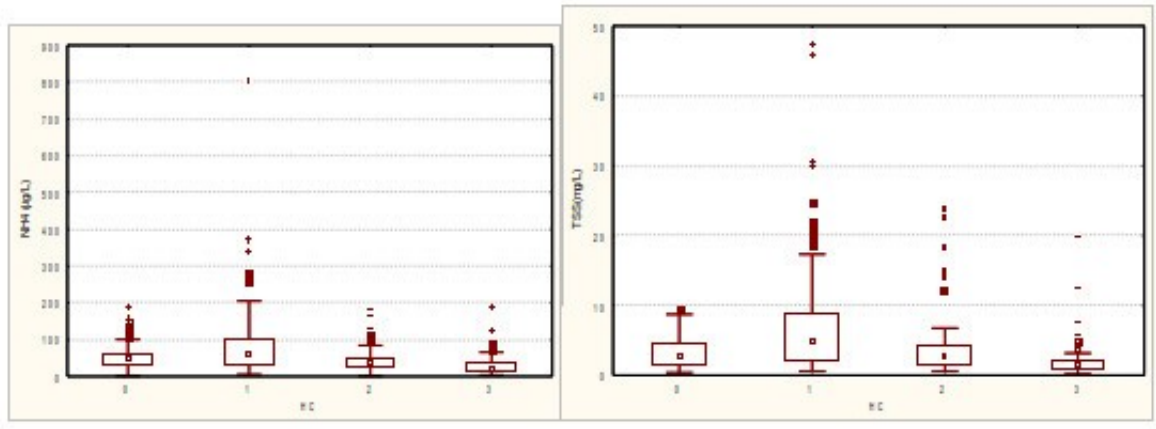
A Tabela 8 mostra o intervalo de confiança de 95% para média dos parâmetros físico-químicos e metais considerando o ciclo hidrológico: 0 - seco; 1- enchendo; 2- cheio e 3 esvaziando. Observa-se que as maiores variações ocorreram quando o nível do reservatório estava enchendo, para o NO_3, NH_4, Total P, PO_4 e STS. A Figura~\ref{fig:boxsplot} mostra os Box-plots dessa diferença considerando o NH_4 e o STS.

Tabela 8: Intervalo de confiança para média de 95% dos parâmetros físico-químicos e metais considerando o ciclo hidrológico: 0-seca; 1-enchente; 2-cheia e 3-vazante.

Componente	Ciclos Hidrológicos			
	seca	enchente	cheia	vazante
Transp (m)	1,90±0,16	1,45±0,17	1,70±0,14	3,00±0,15
PO_4 (mg/L)	20,30±1,71	30,17±4,91	23,90±2,12	16,34±1,75
Temp(ºC)	30,20±0,11	29,60±0,17	29,60±0,16	30,20±0,22
OD (O_3 /L)	5,68±0,18	6,29±0,19	5,55±0,24	5,83±0,24
pH (n)	7,17±0,05	7,24±0,05	7,00±0,06	7,08±0,05
Cond (µs cm^{-1})	49,70±2,71	8,00±1,38	40,35±1,65	44,05±1,43
NH_4 (mg/L)	46,89±7,13	60,71±19,99	35,65±6,05	22,15±4,86
NO_3 (mg/L)	10,07±4,00	39,38±8,82	26,50±6,49	11,64±4,57
PTotal (mg/L)	11,85±1,07	14,41±1,95	17,14±1,38	11,17±1,47

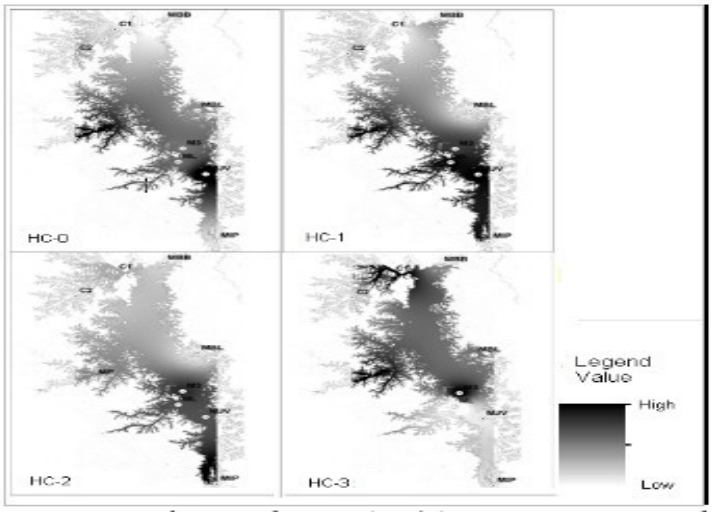
STS (mg L^{-1})	3,00±0,44	4,78±1,63	2,80±0,79	1,40±0,46
Clorofila-a (mg/L)	5,71±0,97	5,95±0,73	4,76±0,43	4,05±0,42
Turbidez (NTU)	3,32±0,61	8,18±2,32	5,28±1,36	1,80±0,46
Ca (mg/L)	4,03±0,10	4,24±0,17	3,84±0,08	4,16±0,13
Mg (mg/L)	42±0,10	1,46±0,08	1,46±0,06	1,36±0,07
Na (mg/L)	2,60±0,32	2,65±0,31	2,30±0,24	2,60±0,15
K (mg/L)	1,60±0,11	1,50±0,12	1,50±0,05	1,40±0,07
FeTotal (mg/L)	0,43±0,06	0,84±0,17	0,82±0,11	0,42±0,09

Figura 15: Boxplots do NH_4 e STS por ciclo hidrológico.



A Figura 15 apresenta a estimação do PO_4 (mg/L) para o ciclo hidrológico, obtida pelo krigagem utilizando inverso da distância ponderada (Deustch e Journel, 1996). Pode-se observar também na Tabela 8 que no período de enchente há a maior média e a maior variação de concentração 30.37±4.91 (mg/L) de PO_4. Isto pode ser uma indicação do estado trófico nos sítios amostrais MP, ML,MJV e MIP, enquanto que, a vazante possui a menor média, 16.34±1.75 mg/L não foram evidenciados altas concentrações nesses sítios. A mesma análise se aplica para o STS e clorofila-a. Isto decorre da existência de um polo pesqueiro no sítio amostral MP que está próximo da sede do Município de Novo Repartimento. Já o ML é próximo da aldeia indígena dos índios Parakanãs. O MJV está próximo de dois portos, o Porto Novo no Município de Jacundá e o Porto da Colônia. O MIP sofre influência da população do Município de Itupiranga.

Figura 16: Estimação do PO_4 (mg/L) por sítio amostral.



4.2 Resultados da classificação

Em relação a classificação baseada na Análise Discriminante foi posível obter uma taxa global de classificação incorreta de 12.3% e 19.1% para a construção do modelo e validação cruzada, respectivamente. O valor de λ de Wilks resultou em 0,239 (Wilks' *Lambda test*) com o teste significativo (p=0.000) indicando adequação da técnica.

Já em relação aos classificadores computacionais, a escolha dos melhores modelos foi realizada a partir de um procedimento automático de seleção implementado no WEKA pela classe CVParameterSelection. A Tabela 9 apresenta o *grid* de parâmetros adotado para o procedimento de seleção automática do modelo.

Tabela 9: Resultado do *grid* de seleção do modelo.

Classificador	Parâmetros	Valor *Grid*	Melhor Valor
RNA	H	11, 43, 75, 107, 139 e 171	43
	L	0.1, 0.3, 0.5, 0.7 e 0.9	0.3
	M	0.1, 0.3 e 0.9	0.3
RF	I	100, 200, 300, … 1000	100
SVM-RBF	G	0.1, 1, 10 e 100	10
	C	0.1, 1, 10 e 100	100
SVM-POLY	G	0.1, 1, 10 e 100	1
	C	0.1, 1, 10 e 100	10
KNN	C	1, 3, 4, 5, , 7, 9, 11,	1

		13 e 15	

As Figuras 17, 18, 19, 20 e 21 apresentam a evolução da taxa de acertos de todos os classificadores computacionais utilizados no estudo.

Figura 17: Gráfico da evolução da taxa de acerto da RNA considerando o número de neurônios na camada escondida.

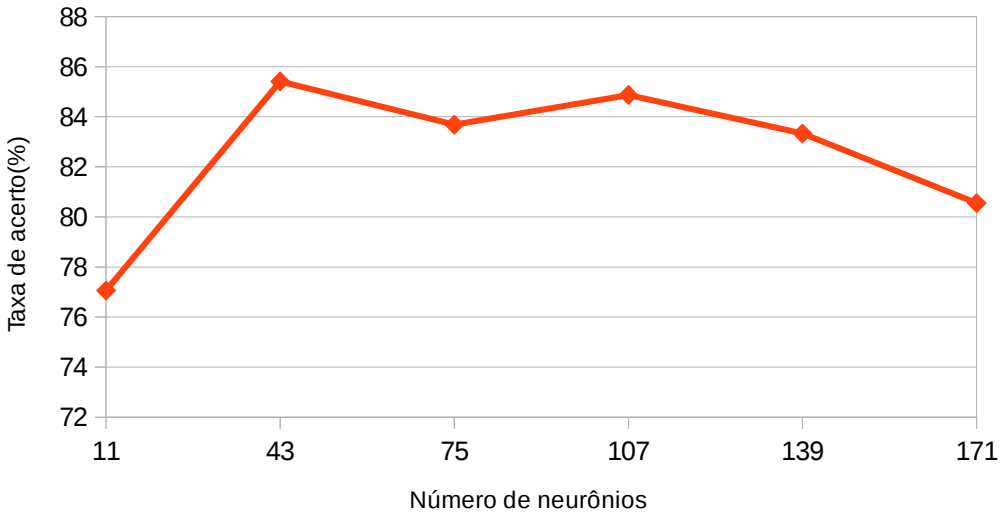


Figura 18: Gráfico da evolução da taxa de acerto do Random Forest considerando o número de árvores.

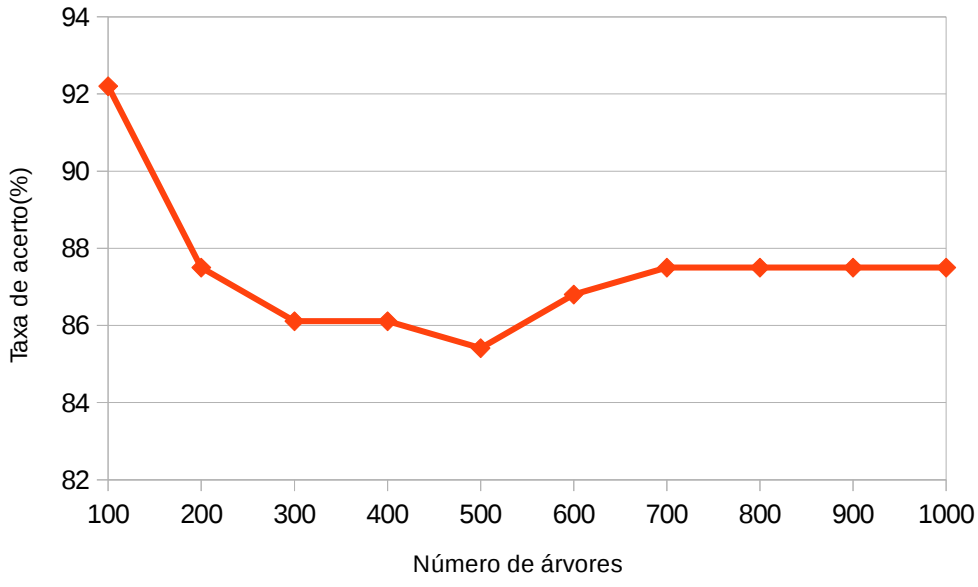


Figura 19: Gráfico da evolução da taxa de acerto do KNN considerando o número de vizinhos mais próximos.

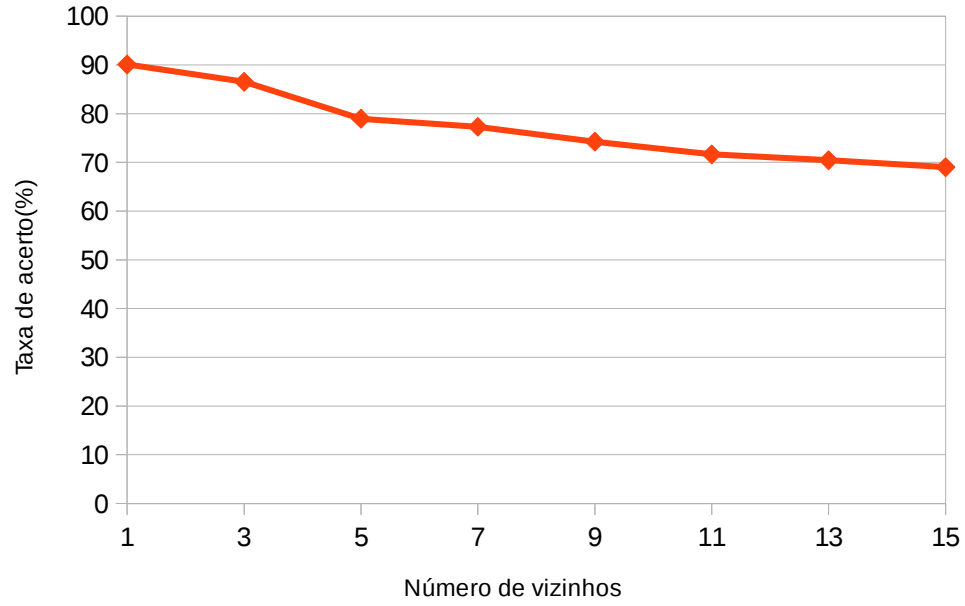


Figura 20: Gráfico da evolução da taxa de acerto da SVM-POLY considerando largura da função dos kernels.

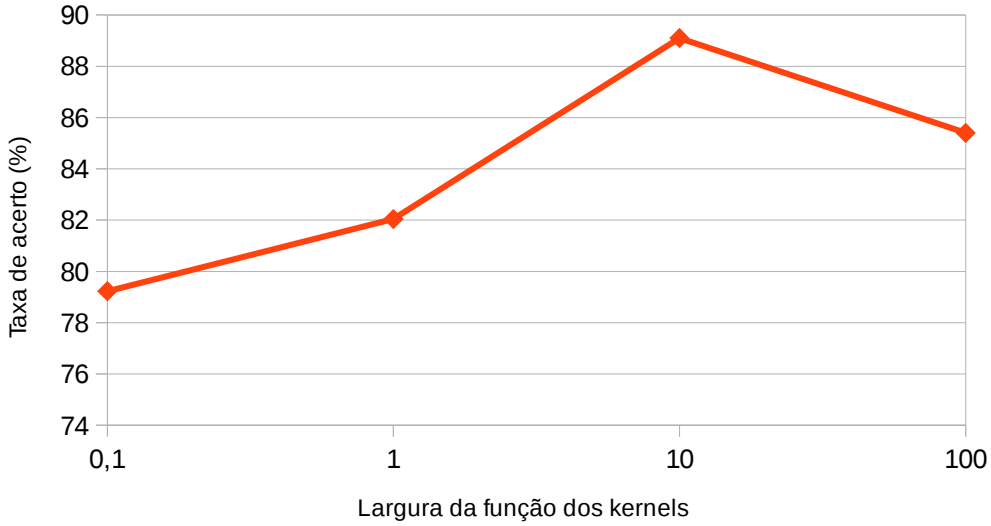
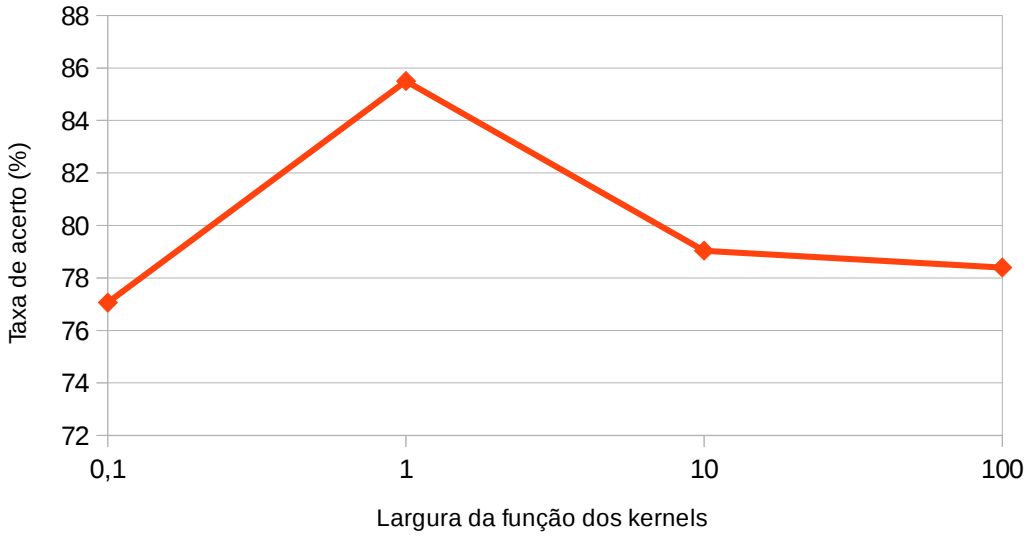


Figura 21: Gráfico da evolução da taxa de acerto da SVM-RBF considerando largura da função dos kernels.



Após vários testes a arquitetura mais robusta para RNA foi com apenas uma camada escondida com 43 neurônios fixados em *H*, taxa de aprendizagem *L* de 0.3 e taxa de momentum *M* de 0.3, sendo que o número de épocas fixado em 2000. Para os classificadores SVMs, os parâmetros otimizados foram gama *G* e penalidade do erro *C*. Depois de vários experimentos o melhor desempenho para SVM-RBF foi 10 em *G* e 100 em *C* e para o classificador SVM-POLY foi 1 em *G* e 10 para *C*. Para o classificador Random Forest a seleção de modelo levou em consideração o principal parâmetro deste classificador, o número de árvores a ser gerada *I*. Para o classificador KNN, *K* é o número de vizinhos com o valor 1.

A Tabela 10 apresenta os desempenhos dos classificadores utilizados levando em consideração a taxa global classificação incorreta E_f e a acurácia (ac(f) = 1 - E_f) como medidas de desempenho.

Tabela 10: Taxa de erro dos classificadores utilizados no estudo.

Classificador	Taxa de erro(E_f) em %	Melhor Valor
AD	19,1	80,9
RNA	15,5	84,5
KNN	9,9	90,1
SVM-RBF	10,9	89,1
SVM-POLY	14,5	85,5
RF	7,8	99,2

De acordo com os resultados apresentados na Tabela 10, pode-se observar que o classificador Random Forest foi o que tendeu a ter o melhor desempenho apresentando uma taxa de erro 7,80% sendo que os classificadores RNA e SVM-POLY foram os que apresentaram um pior desempenho exceto em relação a análise discriminate.

Foram adotadas outras medidas de desempenho na predição dos ciclos hidrológicos utilizadas neste trabalho: a taxa de sensibilidade T_s, taxa de especificidade T_e e a medida F. A Tabela 11 apresenta os resultados obtidos considerando essas métricas. Analisando a tabela pode-se observar que o classificador Random Forest foi o que apresentou as melhores taxas de especificidade, sensibilidade e medida F em quase todas as variações sazonais climáticas confirmando assim o seu melhor desempenho na classificação de ciclos-hidrologicos em relação aos demais classificadores analisados. Além disso, pode-se considerar que este classificador apresentou um bom desempenho, visto que a classificação de ciclo hidrológicos é uma tarefa complexa, pois há uma grande variação dos impactos climáticos em anos distintos.

Tabela 11: Taxa de sensibilidadT_s , Taxa de especificidade T_e e a medida F obtida pelos classificadores.

Classificador	Ciclo Hidrológico	T_s	T_e	F
RNA	Seca	84,9	86,8	85,9
	Enchente	87	87,9	87,4
	Cheia	85,2	82,9	84
	Vazante	82,5	82,5	82,5
SVM-RBF	Seca	90,3	92,3	91,3
	Enchente	95,4	86,6	90,7
	Cheia	88	88	88
	Vazante	83,3	90,5	86,8
SVM-POLY	Seca	81,7	85,4	83,5
	Enchente	85,2	89,3	87,2
	Cheia	85,2	81,4	83,3
	Vazante	80,7	78	79,3
KNN	Seca	95,6	98,3	93,2
	Enchente	92,5	96,2	94,8
	Cheia	85,1	94,4	84,8
	Vazante	87,7	95,5	88,1
RF	Seca	97,6	99,4	95,1
	Enchente	94,4	98,2	96,7
	Cheia	86,8	96,3	86,5
	Vazante	89,5	97,5	90,2

Tabela 12: Resultado do *t-Student test* com %5 de significância.

Classificador	Grau de Liberdade	T_{cal}	T_{tab}	t-Student test	Média
RF->RNA	18	15,52	2,10	$T_{cal} > T_{tab}$	Diferente
RF->KNN	18	1,94	2,10	$T_{cal} < T_{tab}$	Iguais
RF->SVM-POLY	18	14,80	2,10	$T_{cal} > T_{tab}$	Diferente
RF->SVM-RBF	18	8,00	2,10	$T_{cal} > T_{tab}$	Diferente
RF->AD	18	13,50	2,10	$T_{cal} > T_{tab}$	Diferente

De acordo com os resultados apresentados na Tabela 12 em relação ao teste de significância estatística *t-Student test*, utilizando um nível de significância α = 5%, pode-se inferir que os classificadores Random Forest e KNN tem estatisticamente médias iguais mesmo com valores diferentes. Enquanto que os classificadores AD, SVM-RBF e SVM-POLY tiveram taxa média de acertos diferentes em relação ao RF e KKN, todos com valores de p = 0.000.

CAPÍTULO 5

Conclusão

Alguns trabalhos na literatura, que tratam da análise da qualidade de água de reservatórios, consideraram apenas o padrão de sazonalidade sem levar em consideração as características das regiões localizadas à montante da barragem. Este foi o fator preponderante que motivou o desenvolvimento desta dissertação, que procurou preencher esta lacuna através de investigação criteriosa dos diversos aspectos que envolvem essa temática. Assim, partindo-se desse pressuposto, este trabalho avaliou sistematiamente técnicas estatísticas e de inteligência computacional para classificação dos ciclos hidrológicos a partir da análise dos parâmetros físico-químicos e metais da água coletados em 9 sítios amostrais nas zonas à montante da barragem da UHE de Tucuruí.

A partir das 39 variáveis inicialmente analisadas, foi possível extrair seis fatores através da técnica de análise fatorial, reduzindo para 17 parâmetros físico-químicos e metais que explicaram 71,01% de variância total com 26,39% de variância para o primeiro fator.

Ao longo do trabalho foi observado que as maiores variações do NO_3, NH_4, Total P, PO_4 e STS ocorreram no período de enchentes, podendo ser uma indicacação do estado trófico nos sítios amostrais MP, ML, MJV e MIP, em decorrência da existência de pólos pesqueiros ou da maior densidade populacional no entorno do sítio.

A partir da avaliação das medidas de desempenho dos classificadores para o ciclo hidrológico: análise discriminante, redes neurais artificiais, k-vizinhos mais próximo, máquinas de vetores de suporte e random forest, foi possível verificar que o classificador random forest foi o que apresentou melhor desempenho considerando as topologias e arranjos testados com percentual de classificação de 7.80% de predições incorretas. Vale ressaltar que a grande variação sazonal climática da região amazônica pode ter interferido na obtenção de resultados mais precisos. Essa precisão também é muitas vezes afetada, pois as concentraçõess dos parâmetros físico-químicos da água não são homogêneos nas estações de amostragem, tornando a classificação de ciclos hidrológicos um problema complexo.

Adicionalmente, apesar de não constituir o objeto principal do trabalho, a análise do conjunto de dados obtidos pela técnica multivariada Análise Fatorial pode fornecer informações pertinentes a sistemas similares acerca da estrutura do processo de medição.

Portanto, a aplicação de técnicas de inteligência computacional e estatísticas na classificação de ciclos hidrológicos tem uma contribuição construtiva na administração dos recursos hídricos em reservatórios de água em regiões com grande variação sazonal como é o caso da Amazônia.

5.1 Trabalhos Futuros

Como sugestões para trabalhos futuros relacionados com as contribuições apresentadas neste trabalho, é possível citar:

- Como a qualidade da água do reservatório é influenciada pelo ciclo hidrológico a partir da variação das características físicas, químicas e microbiológicas, sugere-se que sejam investigados aspectos relacionados ao tempo de residência da água, a microbiologia nos sítios amostrais, correlacionando também à produção pesqueira e as atividades agrícolas, bem como o tipo de ecossistema associado aos sítios.
- Evidentemente, que estudos como este, têm validade local, isto é, cada represa tem suas especificidades em decorrência da pronfundidade, geomorfologia, e principalmente condições geoquímicas. Por isso, como trabalhos futuros deve-se utilizar dados de outras represas para testar o grau de generalização dos classificadores aqui apresentados.
- Avaliar outras técnicas eficientes e de baixo custo computacional para seleção dos parâmetros físico-químicos mais relevantes ao problema. Adicionalmente, avaliar outros algoritmos de classificação tais como os Bayesianos e Sistemas Fuzzy.

Referências Bibliográficas

Atobatele, O., Ugwumba, O. A. (2008). Seasonal 1 variation in the physicochemistry of a small tropical reservoir (Aiba Reservoir, Iwo, Osun, Nigeria). African Journal of Biotechnology, v.7, June, pp. 1962–1971.

Avni, N., Eben-Chaime, M., Oron, G. (2013). Optimizing desalinated sea water blending with other sources to meet magnesium requirements for potable and irrigation waters. Water Research v. 47, pp. 2164–2176.

Bahiense, J. (2013). Análise estatística utilizando o SPSS. Software available at http://www.prograd.uff.br/estatistica/sites/default/files/Apostila-SPSS.pdf. Bakke, H,A., Leite,

A.S. M., Silva, L. B. (2008). Estatística multivariada: aplicação da análise fatorial na engenharia de produção. Revista Gestão Industrial 4, pp. 17-23.

Bertholdo, L., Júunior, L. C., Umbuzeiro, G. de A., da Silva, C. G. (2013). Mineração de Dados de Qualidade de Agua para Agrupamento de Pontos de Amostragem Usados no Monitoramento de Recursos Hídricos. WCAMA - CSBC 1, pp. 1036-1046.

Bezerra, C. dos S. (2012). Caracterização hidrogeoquímica do reservatório da usina hidrelétrica de Coracy Nunes. Dissertação de Mestrado, UFPA, 2012.

Bilotta, G. S., Brazier, R. F. (2008). Understanding the influence of suspended solids on water quality and aquatic biota. Water Research, v. 42(12), p. 2849-2861.

Bittencourt, M. M., Amadio, S. A. (2007). Proposta para identificação rápida dos períodos hidrológicos em áreas de várzea do rio Solimões-Amazonas nas proximidades de Manaus. Acta Amazonica 37(2), pp. 303-308.

Braga, A. P., Carvalho, A. C. P. L., Neurais Ludermir, T. B. (2000) Redes Artificiais: teoria e aplicações. LTC - Livros Técnicos e Científico ed. 2, pp. 0-260.

Borges Pedro, J. P., Lima, R., Gomes, M. C. R. L., Trindade, M. E. de J., Cavalcante, D. P., Oliveira, J. A. de., Hercos, A. P.; Zucchi, N., Lima, C. B. de., Pereira, S. A., Queiroz, H. L. de. (2014) Influence of the Hydrologcal Cycle on Physical and Chemical Variables of Water Bodies in the Várzea Areas of the Middle Solimões River Region (Amazonas, Brazil). UAKARI 9(2), pp. 75-90.

Cabena, P., Hadjinian, P., Stader, R., Verhees, J., Zanasi. A. (Discovering Data) (1997). Mining: From Concept to Implementation. Prentice Hall.

Cavalcante, Y.L., Hauser-Davis., R.A., Saraiva, A.C.F., Brand˜ao, I.L.S., Oliveira, T.F., Silveira, A.M. (2013). Metal and physico-chemical variations at a hydroelectric reservoir analyzed by Multivariate Analyses and Artificial Neural Networks: Environmental management and policy/decision-making tools. Science os The Total Environment 442, 509-514.

Cavalcante, Y. L. (2013). Aplicação das técnicas análise multivariada e redes neurais artificiais na classificação das águas de reservatórios de hidrelétricas: um estudo de caso na região amaôznica. Dissertação de mestrado, UFPA, 2013.

Cetesb. (2009). Significado ambiental e sanitário das variáveis de qualidade das águas e dos sedimentos e metodologias analíticas e de amostragem. Companhia ambiental do estado de São Paulo, 2009.

Chapman, D. (Ed.). (1996). Water Quality Assessments: A guide to the use of biota, sediments and water in environmental monitoring. Revista Brasileira de Engenharia Agrícola e Ambiental, Campina Grande 14(5), 517-522.

Coletti, C., Testezlaf, R., Ribeiro, T. A. P., Souza, R. T. G., Pereira, D.A. (2010). Water quality index using multivariate factorial analysis. 2. ed. London: UNESCO/WHO/UNEP, p. 651.

Dancey, C. P., Reidy, J.(2006). Estatística sem Matemítica para Psicologia: usando SPSS para Windows. [Tradução VIALI, L.]. 3a ed. Porto Alegre: Artmed.

Deustch, C. V., Journel, A. G. (1996). GSLIB: Geostatistical Software Library and Guide, Beta Version. Oxford University Press, New York, pp.361.

Duda, R.O.; Hart, P.R.; Stork, D.G. (2001). Pattern classification. Wiley.

El-Manzalawy, Y., Honavar, V. (2005). Water quality index using multivariate factorial analysis.

WLSVM: Integrating LibSVM into Weka Environment, <http://www.cs.iastate.edu/ yasser/wlsvm>, (accessado 12 Abril 2005).

Eletrobras-Eletronorte. (2008). Manual do sistema de gestão ambiental - UHE Tucuruí. Eletrobras-Eletronorte, Tocantins.

Elmi, A. A., Mandramoto, C., Hamel, C. (2011). Influence of water table and nitrogen management on residual soil NO_3 and denitrification rate under corn production in sandy loam soil in Quebec. Agriculture, Ecosystem & Environment, v. 79(2-3), p.187-197, 2011.

Faceli, k., Lorena, A. C., Gama, J., Carvalho, A. C. P. L. F. Inteligência artificial: uma abordagem de aprendizado de máquina. Editora LTC, 2011.

Ferreira, M. R. P. (2007). Análise discriminante clássica e de núcleo: avaliações e algumas contribuições relativas aos métodos boosting e bootstrap. Dissertação de Mestrado, UFPE.

Fiorucci, A. R., Filho, E. B. A. (2005). A importância do oxigênio dissolvido em ecossistemas aquáticos. Química Nova na Escola, v. 22, 2005.

Furtado, P. S., Poersch, L. H., Junior, W. W. (2011). Effect of calcium hydroxide, carbonate and sodium bicarbonate on water quality and zootechnical performance of shrimp Litopenaeus vannamei reared in bio-flocs technology (BFT) systems. Aquaculture, v.321(1-2), p. 130-135.

Galvão, C. de O. (1999). Sistemas inteligentes: Aplicações a recursos hídricos e ciências ambientais. UFRGS: ABRH.

Gastaldini, M. C. C., Sefrin, G. F. F., Paz, M. F. (2002). Diagnóstico atual e previsão futura da qualidade das águas do rio Ibicuí utilizando o modelo QUAL2E. Engenharia Sanitária e Ambiental 7(4), 129-138.

Hall, M.; Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I. H. (2009). The weka data mining software: an update. SIGKDD Explor. Newsl. 11, 10-18.

Hair, J. F., Tatham, R. L., Anderson, R. E., Black, W. (1998). Multivariate data analysis. 5. ed. New Jersey: Prentice Hall.

Hauser-Davis, R. A., Oliveira, T.F., Silveira, A. R., Silva, T. B. Ziolli, R. L. (2010). Case study: Comparing the use of nonlinear discriminating analysis and Artificial Neural Networks in the classification of three fish species: acaras (Geophagus brasiliensis), tilapias (Tilapia rendalli) and mullets (Mugil liza). Ecological Informatics (Print) 5, 474-478.

Hauser-Davis, R. A., Oliveira, T. F.; Silveira, A. M., Protázio, J. M. B., Ziolli, R. L. (2012). Logistic regression and fuzzy logic as a classification method for feral fish sampling sites. Environ Ecol Stat, 19, 473-483.

Haykin, S. (1998). Neural Networks: A comprehensive Foundation. Prentice Hall, 2 edition, July 1998. p. 900.

Holguin-Gonzalez, J. E., Boets, P., Alvardo, A., Cisneros, F., Carrasco, M. C., Wyseure, G., Nopens, I., Goethals, P. L. M. (2013). Integrating hydraulic, physicochemical and ecological models to assess the effectiveness of water quality management strategies for the River Cuenca in Ecuador. Ecologycal Modeling v. 254, p.114, 2013.

Kazi, T.G., Arain, M.B., Jamali, M.K., Jalbani, N., Afridi, H.I., Sarfraz, R.A., Baigi, J.A., Shah, A.Q. (2009). Assessment of water quality of polluted lake using multivariate statistical techniques: A case study. Ecotoxicology and Environmental Safety 72, 301- 309.

Kezunovic, M., Rikalo, I. (1996). Detect and classify faults using neural networks. Computer Applications in Power 9, 42-47.

Johnson, R. A., Wichern, D. W. (2001). Applied multivariate statistical analysis. 5. ed. [S.l.]: Prentice Hall, pp.767.

Lachenbruch, P. A. (1975). Discriminant analysis. Hafner Press, New York.

Lessels, J. S., Bishop, T. F. A. (2013). Estimating water quality using linear mixed models with stream discharge and turbidity. Journal of Hydrology v. 498, p. 13-22.

Lewis, W. M. (2000). Basis for the protection and management of tropical lakes. Lakes and Reservoirs: Research and Management 5, 35-48.

Li, M-t. Zhao, L-p, Zhang, J-J. (2013) Effect of Temperature, pH and Salt on Fluorescent Quality of Water Extractable Organic Matter in Black Soil. Journal of Integrative Agriculture 12(7), p.1251-1257.

Liu, W., Chen, W. G. Prediction of water temperature in a subtropical subalpine lake using an artificial neural network and three-dimensional circulation models. (2012). Computers & Geoscience 45, p.13-25.

Lobato, T. C. (2014). Análise estatística e abordagem fuzzy na classificação do estado trófico da água de reservatório na região Amazônica. Dissertação de Mestrado em Matemática e Estatística, ICEN, UFPA.

Mérona, B., Juras, A. A., Santos, G. M, e Cintra, I. H. A. (2010). Os peixes e a pesca no baixo Rio Tocantins: vinte anos depois da UHE Tucuruí. Eletrobras-Eletronorte.

Yiang, L.; Qu, H., Zhang, Y., Li, F. (2012). Effects of partial root-zone irrigation on physiology, fruit yield and quality and water use efficiency of tomato under different calcium levels. Agricultural Water Management 104, p.89-94.

Yustseven, E., Kesmez, Unlukara, A. (2005). The effects of water salinity and potassium levels on yield, fruit quality and water consumption of a native central anatolian tomato species (Lycopersiconesculantum). Agricultural Water Management 78(1-2), p.128-135.

Mann, G.R., Duncan, S.E., Knowlton, K. F., Dietrich, A. D., Okeefe, S. F. (2013). Effects of mineral content of bovine drinking water: Does iron content affect milk quality?. Dairy Science 96(12), p.7478-7489.

Mitchell, T. (1997). Machine learning. (Mcgraw-Hill International Edit) McGraw-Hill Education (ISE Editions).

Murtojarvi, M., Suominen, T., Usipaika, E., Nevalanier, O.S. (2011). Optimising an observational water monitoring network for Archipelago Sea, South West Finland. Com-puters & Geosciences 37, 844-854.

Neal, R., Jarvie, H. P., Sharon, M. H., Whitehead, P. G., Williams, R. J, Neal, M., Harrow, M. Wickham. (2000). The water quality of the River Kennet: initial observations on a lowland chalk stream impacted by sewage inputs and phosphorus remediation. Science of The Total Environment 251-252, p.477-495.

Novoa, S., Chust, G., Sagarminada, Y., Revilla, M., Borja, A., Franco. (2012). Water quality assessment using satellite-derived chlorophyll-a within the European directives, in the southeastern Bay of Biscay. Marine Polluition Bulletin 64(4), p.739-750.

Oduwole, G. A. (1997). Indices of pollution in Ogunpa and Ona Rivers, Ibadan: physico-chemical, trace metal and plankton studies. [S.l.]: PhD degree, University of Ibadan, pp.293.

Oshiro, T. M. (2013). Uma abordagem para construção de uma única árvore a partir do random forest para classificação de bases de expressão gênica. Dissertação de mestrado, USP.

Ouyang, Y. (2005). Evaluation of river water quality monitoring stations by principal component analysis. Water Research 39, 2621-2635.

Rahman, S. M. E., Wang, J., Oh, D. (2013). Synergistic effect of low concentration electrolyzed water and calcium lactate to ensure microbial safety, shelf life and sensory quality of fresh pork. Food Control 30, p.176-183.

Ramin, M., Labenck, T., Boyd, D., Trolle, D., Arhonditsis, G. B. (2012). A Bayesian synthesis of predictions from different models for setting water quality criteria. Ecologycal Modeling 242, 127-145.

Rebouças, A. C., Braga, B. e Tundisi, J. G. (2006). Águas doces do Brasil: capital ecológico, uso e conservação. Escrituras Editora 3. ed., São Paulo.

Ren, Z., Zeng, Y., Fu, X., Zang, G., Chen, L., Chen, J., Chon, T., Wang, Y., Wei, Y. (2013). Modeling macrozooplankton and water quality relationships after wetland construction in the Wenyuhe River Basin, china. Ecologycal Modeling 252, 97-105.

Riflin, R., Klautau, A. (2004). In defense of one-vs-all classification. Machine Learning Research 5, 101-141.

Santos, L. D. M. dos., Mikami, R., Vendramim, A. C. B. K., Kaestner, C. A. A. (2009). Procedimentos de Validação Cruzada em Mineração de Dados para ambiente de Computação Paralela. Escola Regional de Alto Desempenho, 2009.

Scudino, P. A. (2008). A Utilização de alguns testes estatísticos para análise da variabilidade do preço do Mel nos municípios Angra dos Reis e Mangaratiba, Estado do Rio de Janeiro. Monografia de graduação, UFRRJ, 2008.

Sharma, A., Naidu, M., Sargaonkar, A. (2013). Development of computer automated decision support system for surface water quality assessment. Computers & Geosciences 51, p.129-134.

Silva, R. M., Almeida, T. A., Yamakamy, A. (2012). Análise de métodos de aprendizagem de máquina para detecção automática de spams hosts. Simpósio Brasileiro de Segurança da Informação e de Sistemas Computacionais.

Sims, A., Garaj, S., Hu, Z. (2012). Seasonal population changes of ammonia-oxidizing organisms and their relationship to water quality in a constructed wetland. Ecological Engineering 40, p.100-107.

Singh, K. P., Basant, A., Malik, A., Jain, G. (2009). Artificial neural network modeling of the river water quality-A case study. Ecologycal Modeling 220, 888-895.

Schoumans, O.F., Chardon, W. J., Bechmann, M. E., Gascuel-Odoux, C., Hofman, G., Kronvang, B., Rubaek, G. H., Ulen, B., Dorioz, J. M. (2013). Mitigation options to reduce phosphorus losses from the agricultural sector and improve surface water quality: A review. Science of The Total Environment, In Press, September, 2013.

Souza, A. M. (2000). Monitoraçãoo e ajuste de realimentação em processos produtivos multivariados. Tese (Doutorado Engenharia de Produção), Universidade Federal Santa Catarina.

Song, M. W., Huang, P., Huang, F., Zhang, H., Xie, K. Z., Wang, Z. H., He, G. X. (2011). Water quality of a tributary of the Pearl River, the Beijiang, Southern China: implications from multivariate statistical analyses. Environ Monit Assess 172, P.589-603.

Spiegel, M. R. (1993). Estat´ıstica. [Tradução: CONSENTINO, P.] (Coleção Schaum), São Paulo: Makron Books.

Standard Methods. (2013). 1060 Collection and Preservation of Samples. <http://www.standardmethods.org/store/ProductView.cfm?ProductID=123>, (accessado 10 Setembro 2013).

Vrana, B., Allan, I,J., Greewood, R., Mills, G., Dominak, E., Svensson, K., Knutsson, J., Morrison, G. (2005). Passive sampling techniques for monitoring pollutants in water. Trends in Analytical Chemistry 24, 845-868.

Vapnik, V.N. (1995). The nature of statistical learning theory. Springer Verlag. Vicini, L., Souza, A. M. (2005). An´alise multivariada da teoria a pr´atica. UFSM-CCNE, 2005.

Zahraie, B., Hosseni, S. M. (2009). Development of reservoir operation policies considering variable agricultural water demands. Expert System with Applications 36, 4980-4987.

Zhang, B., Qin, Y., Huang, M., Sun, Q., Li, S., Wang, L. SD-GIS-based temporal-spatial simulation of water quality in sudden water pollution accidents. Computers & Geosciences 37, p.874-882.

Zeller, R. A., Carmines, E. G. (1980). Measurement in the social sciences: The link between theory and datas. Cambridge University Press, New York, 198pp.

Wang, X. L., Yin, Z. J., Lv, Y. B., Li, S. F. (2009). Operating rules classification system of water supply reservoir based on Learning Classifier System. Expert Systems with Applications 36, 5654-5659.

Weiss, S. M. Kulikownski, C. A. (1991). Computer system that learn: classification and prediction methods from statistics, neural nets, machine learning, and experts systems. San Francisco: Morgan Kaufmann Publishers Inc.

Wenner, E., Sanger, D., Arendt, M., Holland, A. F., Chen, Y. (2004). Variability in dissolved oxygen and other water-quality variables within the national estuarine research reserve system. Journal of Coastal Research 45, 17-38.

Witten, I. H., Frank, E. (2005). Data mining: practical machine learning tools and techniques with java implementations. Edition Morgan Kaufmann.

APÊNDICE A - Parte do arquivo ARFF contendo a base de dados que foi processado no WEKA

A Figura 22 apresenta parte do arquivo ARFF utilizado no estudo que foi processa no software WEKA. Esse arquivo é dotado de 423 registros com algumas particularidades de instruções codificadas.

Figura 22: Parte do arquivo ARFF utilizado no WEKA

```
@relation QualidadeAgua


@attribute TRANS numeric
@attribute Temp numeric
@attribute OD numeric
@attribute pH numeric
@attribute Cond numeric
@attribute FeTotal numeric
@attribute Ca numeric
@attribute Mg numeric
@attribute Na numeric
@attribute K numeric
@attribute NH4 numeric
@attribute NO3 numeric
@attribute PO4 numeric
@attribute PTOTAL numeric
@attribute STS numeric
@attribute PigTotal numeric
@attribute Turb numeric
@attribute CH {0,1,2,3}
@data

0.400,26.300,4.510,6.500,54.700,0.840,2.240,1.170,3.100,5.900,98.040,37.960,32.930,84.460,17.3
30,4.360,46.900,1
0.400,26.700,4.700,6.500,46.100,2.250,2.080,1.170,3.000,3.300,87.130,37.630,35.570,83.730,17.0
00,2.780,45.600,1
0.400,26.500,4.620,6.440,45.900,2.890,1.760,1.750,3.100,3.600,136.220,48.940,39.730,98.180,16.
670,1.980,45.300,1
2.800,29.600,5.990,7.000,49.400,0.570,2.400,2.140,2.700,3.700,40.770,34.110,2.720,12.280,2.200,
6.660,1.680,1
2.800,29.700,5.800,6.960,49.000,0.450,2.720,1.750,2.600,1.800,65.310,52.720,1.960,11.920,2.000,
5.470,1.550,1
2.800,29.800,5.450,6.930,48.900,0.460,2.720,1.750,2.600,1.700,43.500,14.540,2.340,11.920,2.400,
6.430,1.380,1
1.600,28.900,5.940,6.960,57.500,1.160,3.040,0.970,4.600,2.600,38.040,14.300,6.500,45.840,3.400,
9.280,4.490,1
1.600,28.900,5.770,6.790,57.500,1.660,3.040,0.970,4.200,2.600,62.590,10.640,8.010,41.150,3.400,
8.570,4.280,1
1.600,29.000,5.300,6.850,56.800,0.710,3.200,1.170,4.100,2.500,59.860,11.300,8.010,40.790,3.400,
9.520,4.480,1
1.900,29.000,6.220,6.750,57.100,1.030,3.040,1.940,2.400,2.200,10.770,72.050,9.520,29.240,2.400,
1.900,8.780,1
1.900,29.000,6.040,6.840,57.300,0.900,3.200,1.850,2.400,2.200,10.770,92.920,9.520,30.680,2.500,
1.430,7.900,1
1.500,30.300,6.150,6.720,54.800,0.530,3.680,1.170,2.000,1.500,5.320,43.570,10.650,31.400,3.870,
14.760,8.630,3
1.500,29.900,5.930,6.900,54.500,0.520,3.360,1.650,2.700,1.400,5.320,20.260,25.760,37.900,4.000,
12.610,10.000,3
1.500,29.900,4.860,7.010,54.800,0.500,4.000,1.460,1.300,1.400,13.500,50.170,21.980,25.990,2.80
0,2.140,8.350,3
0.700,29.400,3.190,7.130,52.400,0.660,4.320,1.260,1.500,2.000,8.040,64.470,18.200,29.240,3.400,
```

As instruções apresentadas correspondem a:

- *@Relation* - indica o nome da base de dados;
- *@atribute* - refere-se aos parâmetros da qualidade da água;
- *@atribute* CH - a classe com as 4 saídas que identificam os ciclos hidrológicos;
- *@data* - finaliza as instruções.