UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE TECNOLOGIA E GEOCIÊNCIAS
DEPARTAMENTO DE ENGENHARIA QUÍMICA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA QUÍMICA

LUCIANA PIMENTEL MARQUES

MODELAGEM MATEMÁTICA PARA PREVISÃO DE PARÂMETROS DE QUALIDADE DE ÁGUA EM CORPOS HÍDRICOS

Recife
2018

LUCIANA PIMENTEL MARQUES

MODELAGEM MATEMÁTICA PARA PREVISÃO DE PARÂMETROS DE QUALIDADE DE ÁGUA EM CORPOS HÍDRICOS

Tese apresentada ao Programa de Pós-Graduação em Engenharia Química da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Doutor em Engenharia Química.

Área de concentração: Engenharia Ambiental.

Orientador: Profº. Dr. José Geraldo de Andrade Pacheco Filho
Coorientadora: Profª. Drª Valdinete Lins da Silva
Coorientador: Profº. Dr. Frede de Oliveira Carvalho

Recife
2018

Catalogação na fonte
Bibliotecária: Rosineide Mesquita Gonçalves Luz / CRB4-1361 (BCTG)

M357m Marques, Luciana Pimentel.
Modelagem matemática para previsão de parâmetros de qualidade de água em corpos hídricos / Luciana Pimentel Marque – Recife, 2018.
219 f., il., fig. e tabs.

Orientador: Profº. Dr. José Geraldo de Andrade Pacheco Filho.
Coorientadora: Profª. Drª. Valdinete Lins da Silva.
Coorientador: Profº. Dr. Frede de Oliveira Carvalho.

Tese (Doutorado) – Universidade Federal de Pernambuco. CTG.
Programa de Pós-Graduação em Engenharia Química, 2018.
Inclui Referências e Apêndices.

1. Engenharia Química. 2. Análise de Componentes Principais. 3. Inteligência Artificial. 4. Modelagem Matemática. 5. Qualidade da água. Recursos Hídricos. I. Pacheco Filho, José Geraldo de Andrade (Orientador). II. Silva, Valdinete Lins da (Coorientadora). III. Carvalho, Frede de Oliveira (Coorientador). IV.Título.

660.2 CDD (22. Ed.) UFPE/BCTG-2018/444

LUCIANA PIMENTEL MARQUES

MODELAGEM MATEMÁTICA PARA PREVISÃO DE PARÂMETROS DE QUALIDADE DE ÁGUA EM CORPOS HÍDRICOS

Tese apresentada ao Programa de Pós-Graduação em Engenharia Química da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Doutor em Engenharia Química.

Aprovada em: 31/08/2018.

BANCA EXAMINADORA

_____

Profº. Dr. José Geraldo de Andrade Pacheco Filho (Orientador)
Universidade Federal de Pernambuco

_____

Profª. Drª. Valdinete Lins da Silva (Coorientadora)
Universidade Federal de Pernambuco

_____

Profº. Dr. Frede de Oliveira Carvalho (Coorientador)
Universidade Federal de Alagoas

_____

Profª. Drª. Daniella Carla Napoleão (Examinadora Interna)
Universidade Federal de Pernambuco

_____

Profº. Dr. Gilson Lima (Examinador Externo)
Universidade Federal Rural de Pernambuco

_____

Profº. Dr. José Marcos Francisco da Silva (Examinador Interno)
Universidade Federal de Pernambuco

_____

Profª. Drª. Léa Zaidan (Examinadora Externa)
Pesquisadora

_____

Profº. Dr. Mohand Benachour (Examinador Interno)
Universidade Federal de Pernambuco

AGRADECIMENTOS

Sempre achei esta a parte mais difícil da tese para escrever, talvez porque a vida não se coloca em análise matemática e não é pelo valor que descobrimos a significância das pessoas na nossa caminhada.

Primeiro de tudo, gostaria de agradecer a Deus por me guiar, iluminar e me dar tranquilidade para seguir em frente com os meus objetivos e não desanimar com as dificuldades. Agradeço a Ele também por manter os meus pais (João e Adelaide) ao meu lado, com a saúde que eles estão hoje, por sempre me motivarem, entendendo as minhas faltas e momentos de afastamento e reclusão e me mostrarem o quanto era importante estudar.

Agradeço ao meu irmão (Felipe) e minha cunhada (Deize), por me darem o meu maior presente, o amor da minha vida, minha sobrinha e afilhada, Valentina. Hoje com quase 3 anos, fez parte da fase mais difícil da preparação desta tese e que com o seu olhar e seu abraço me fizeram ter força todos os dias! Sua chegada em nossas vidas nos encheu de fé e motivação... Grata sempre pela sua vida.

Agradeço ao meu namorado, Pedro, com quem eu sei que passarei por muitos e muitos momentos de felicidade como esta e que é a pessoa que a vida escolheu para ser meu companheiro nas horas boas e ruins, e que me tranquiliza com suas palavras de motivação e carinho.

Agradeço às minhas amigas Magda, Pulkra, Mariana Souza, Mariana Galindo, Renata, Wivian, Raquel, Sibéria e Léa, por todos os estímulos, que por vezes foram combustíveis nesta caminhada. Cada uma de vocês foi e é especial na minha vida! Os momentos de lazer e risadas sem dúvida me deram força para chegar até aqui!

Agradeço aos meus colegas de trabalho da Gerência de Planos e Sistemas de Informação, da APAC: Augusto, Alexandro, Elhanie, Erik, Everton e Marcelo Possas. Com eles aprendi o verdadeiro espírito de trabalho em equipe e amadureci muito como profissional.

Gostaria de agradecer profundamente ao professor Frede de Oliveira Carvalho. Considero imensurável o aprendizado por mim obtido nestes quase 10 anos de convívio e amizade. A sua incessante busca pelo conhecimento é contagiante e suas atitudes pessoais e profissionais são dignas de serem seguidas por todos. Resumi-lo a meu coorientador é muito pouco e tenho certeza de que ele sente a importância que teve e tem para mim não só na

condução do trabalho, mas também como conselheiro e até nas horas em que parece que nada está dando certo e que preciso não só de um consolo, mas de um ombro amigo.

Um parágrafo especial dedico a minha sempre orientadora, Valdinete Lins, a quem tenho muito a agradecer desde minha graduação, com suas palavras de apoio e carinho. A senhora é o maior exemplo de caráter e determinação. Grata pela confiança e ser para mim a mãe que todos os alunos do Departamento de Engenharia Química conhecem.

Agradeço ao meu orientador, Prof. Dr. José Geraldo Pacheco, pela orientação, dedicação, paciência e, principalmente pelo seu incentivo, disponibilidade e apoio que sempre demonstrou. Aqui lhe exprimo a minha gratidão.

Agradeço aos professores do PPEQ, Programa de Pós Graduação em Engenharia Química, pelos ensinamentos que passaram desde o mestrado, os quais foram, são e serão muito importantes para mim e para a minha vida profissional, assim como agradeço aos funcionários, Priscila Macêdo e Flávio Garrett, que fazem com que tudo funcione da melhor maneira possível.

Agradeço a Ana Maria Ribeiro, a Helenice Garcia e a Equipe do Professor Frede Carvalho da UFAL, sem eles não teria conseguido concluir este trabalho com toda a complexidade que o envolve.

Por último, agradeço ao PRH28 pela bolsa concedida durante os dois primeiros anos do doutorado, em especial a Celmy Barbosa e Vilckma Oliveira de Santana, pela disponibilidade em ajudar e todo o compromisso colocado em nosso aprendizado.

“Todos têm direito ao meio ambiente ecologicamente equilibrado, bem de uso comum do povo e essencial à sadia qualidade de vida, impondo-se ao Poder Público e à coletividade o dever de defendê-lo e preservá-lo para as presentes e futuras gerações.” (BRASIL, 1988)

RESUMO

O monitoramento da qualidade da água representa uma etapa fundamental, porém consome muito tempo e recursos e requer análise dos dados complexos com múltiplas variáveis. Neste trabalho foram investigadas técnicas de modelagem empírica baseadas em técnicas estatísticas e de inteligência artificial para inferência e previsão de parâmetros de qualidade de água. Nesse estudo foram utilizados dados de três países diferentes, provenientes de um manancial brasileiro, espanhol e chinês. Os dados temporais formam uma série com dados horários de 179 dias que foram utilizados na forma horária e de médias diárias. Inicialmente, foi realizada análise exploratória, baseada em técnicas de estatísticas multivariadas de Análise de Componentes Principais (ACP) e de Mapas Auto-organizáveis de Kohonen (SOM). Os parâmetros físico-químicos da água foram modelados para inferenciar parâmetros de qualidade, tais como clorofila, Demanda Bioquímica de Oxigênio (DBO) e cianobactérias, com base em modelos de inteligência artificial de Redes Neurais Artificiais (RNA) do tipo *Perceptron* Multicamadas (MLP) e *Neuro-Fuzzy* (NF), e modelos estatísticos do tipo Máquina de Vetores de Suporte (SVM). Em seguida os dados na forma de série temporal do manancial brasileiro foram utilizados numa modelagem para previsão futura em curto período de tempo da concentração de clorofila, com base nos modelos citados assim como as suas conjunções com a Transformada *Wavelet* (WT). Observou-se através das ACPs que a degradação da qualidade da água do Brasil, China e Espanha foi atribuída à poluição por matéria orgânica e compostos nitrogenados, e consequentemente encontravam-se eutrofizadas, provavelmente devido aos lançamentos de efluentes domésticos nos corpos de água. Na Análise de agrupamento para os três conjuntos de dados, realizada pelas técnicas SOM e ACP, a matriz de distâncias e o mapa de características mostram que os dados de operação normal formam um agrupamento uniforme. O esquema de cores indica que existem pequenos valores de distâncias entre algumas amostras, indicando a possibilidade de uma modelagem preditiva utilizando estes dados. Na estimativa da qualidade de água através dos dados estacionários, a técnica mais eficiente para predição de clorofila foi o modelo NF (r=0,9396 e RMSE = 3,998), para a DBO foram as RNA (r = 0,9349 e RMSE = 0,6720) e para a Cianotoxina foi o SVM (r = 0,99 e RMSE =2,65). Em relação à estratégia híbrida adotada, na qual a WT foi utilizada no pré-tratamento dos dados de entrada dos modelos, foi percebida a influência positiva na previsão da série temporal do Brasil nos três modelos testados para a previsão de teor de clorofila para série diária e horária. Isoladamente, cada modelo apresentou um desempenho diferente, contudo o melhor modelo foi aquele que

utilizava RNA com WT, através da *Wavelet*-mãe Db20, obtendo-se valores maiores de desempenho para treinamento (R^2=0,99320 e RMSE=0,85026) e teste (R^2=0,99135 e RMSE=2,28134). Portanto, o presente trabalho mostrou que as modelagens em técnicas estatísticas e de inteligência artificial foram eficazes na análise de dados ambientais e podem ser utilizadas para melhoria dos modelos de gestão em caracterização e monitoramento da qualidade da água.

Palavras-chave: Análise de Componentes Principais. Inteligência Artificial. Modelagem Matemática. Qualidade da água. Recursos Hídricos.

ABSTRACT

Monitoring water quality is a critical step, but it consumes a lot of time and resources and requires complex data analysis with multiple variables. This work investigated empirical modeling techniques based on statistical and artificial intelligence techniques for inference and prediction of water quality parameters. In this study we used data from three different countries, coming from a Brazilian, Spanish and Chinese source. The temporal data form a series with hourly data of 179 days that were used in hourly form and daily averages. Initially, an exploratory analysis was performed, based on multivariate statistical techniques of Principal Component Analysis (PCA) and Kohonen's Self-Organizing Maps (SOM). The physico-chemical parameters of the water were modeled for inference of quality parameters, such as chlorophyll, Biochemical Oxygen Demand (BOD) and cyanobacteria, based on Artificial Neural Networks (RNA) models of Multi-layered Perceptron (MLP) and Neuro-Fuzzy (NF), and statistical models of support vector machine (SVM). Then the data in the time series form of the Brazilian source were used in a modeling for a short time future prediction of the chlorophyll concentration, based on the cited models as well as their conjunctions with the Wavelet Transform (WT). It was observed through the ACPs that the degradation of water quality in Brazil, China and Spain was attributed to pollution by organic matter and nitrogenous compounds, and consequently they were eutrophic, probably due to the releases of domestic effluents in the bodies of water. In cluster analysis for the three datasets, performed by the SOM and ACP techniques, the distance matrix and the characteristic map show that the normal operation data form a uniform clustering. The color scheme indicates that there are small distance values between some samples, indicating the possibility of a predictive modeling using this data. The most efficient technique for predicting chlorophyll was the NF model (r = 0.9396 and RMSE = 3,998), for BOD were RNA (r = 0.9349 and RMSE = 0.6720) and for Cyanotoxin was SVM (r = 0.99 and RMSE = 2.65). In relation to the hybrid strategy adopted, in which the WT was used in the pre-treatment of the input data of the models, it was perceived the positive influence in the prediction of the Brazilian time series in the three models tested for the prediction of chlorophyll content for daily series and hourly. Each model presented a different performance, but the best model was the one that used RNA with WT through the Wavelet Db20, obtaining higher performance values for training (R^2 = 0.99320 and RMSE = 0.85026) and test (R 2 = 0.99135 and RMSE = 2.28134). Therefore, the present work showed that the modeling in statistical techniques and artificial intelligence were effective in the analysis of

environmental data and can be used to improve management models in characterization and monitoring of water quality.

Keywords: Principal Component Analysis. Artificial Intelligence. Mathematical Modeling. Water Quality. Water Resources

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ABNT	Associação Brasileira de Normas Técnicas
ACP	Análise de Componentes Principais
ANFIS	Modelo Adaptativo de Inferência Neuro-difusa *(Adaptive Neuro-fuzzy Inference System)*
ANA	Agência Nacional de Águas
ARIMA	Auto Regressivo Integrado Média Móvel *(Auto Regressive Integrated Mobile Average)*
ARIMAX	ARIMA com entrada exógena
BMU	Neurônio Mais Adequado *(Best Matching Unit)*
BPNN	Rede Neural *Backpropagation*
CART	*Classification and Regression Tree*
CA	Análise de Cluster
CETESB-SP	Companhia Ambiental do Estado de São Paulo
CE	Condutividade Elétrica
CF	Constituição Federal
CIES	*Centre for Intelligent Environmental Studies*
CONAMA	Conselho Nacional de Meio Ambiente
CPRH	Agência Estadual de Meio Ambiente de Pernambuco
CT	Coliformes Totais
CWT	Transformadas de Onda Contínua
DBO	Demanda Bioquímica de Oxigênio
DQO	Demanda Química de Oxigênio
DWT	Transformadas de Onda Discreta
ELM	Máquina de Aprendizagem Extrema *(Extreme Learning Machine)*
ETA	Estação de Tratamento de Água
FFBP	Rede *Feedforward Backpropagation*

GMDH	Método de Grupo de Manipulação de Dados *(Group Method of Data Handling)*
GRNN	Rede Neural de Regressão Geral
IA	Inteligência Artificial
IET	Índice de Estado Trófico de Carlson
IQA	Índice de Qualidade de Água
LR	Regressão Linear *(Linear Regression)*
MAE	Erro Absoluto Médio
MERIS	Espectrômetro de Imagem de Resolução Média *(Medium Resolution Imaging Spectrometer)*
MLP	*Perceptron* Multicamadas *(Multi Layer Perceptron)*
MLPNN	Rede Neural *Perceptron* Multicamadas
MPE	Erro de Previsão Médio
MSE	Erro Quadrático Médio
NF	*Neuro fuzzy*
NSE	Coeficiente de Eficiência do Modelo Nash-Sutcliffe
OD	Oxigênio Dissolvido
ORP	Potencial de Oxi-redução
pH	Potencial Hidrogeniônico
PNN	Rede Neural Probabilística
RBMS	*River Biology Monitoring System*
RBF	Funções de Bases Radiais
RBFNN	Redes Neurais do tipo RBF
RNA	Redes Neurais Artificiais
RMSE	Raiz do Erro Médio Quadrado *(Root Mean Square Error)*
r	Coeficiente de Correlação de Pearson
R^2	Coeficiente de Determinação
RLM	Regressão Linear Múltipla

Sabesp	Companhia de Saneamento Básico do Estado de São Paulo
SEMA-SP	Secretaria do Meio Ambiente do Estado de São Paulo
SOM	Mapas Auto-organizáveis de Kohonen *(Self-Organizing Maps)*
SVM	Máquina de Vetores de Suporte *(Support Vector Machine)*
SVR	Máquina de Vetores de Suporte para Regressão *(Support Vector Machine for Regression)*
TSK	Modelo de Takagi-Sugeno-Kang
TW	Transformada *Wavelet*

LISTA DE SÍMBOLOS

C_{sat}	Concentração de Saturação do Oxigênio
K_{O2}	Constante de Solubilidade de Oxigênio na Água
pO_2	Pressão parcial do oxigênio no ar
a_{ij}	Loadings
A	Vetores de Entrada
Wj	Pesos
Th_{ij}	Ativação Residual Interna (threshold)
I	Matriz Identidade
e(x)	Erro
J	Matriz Jacobiana
b_{ij},	Baias
X_j	Variáveis Originais
Z_i	Componente Principal
D_E	Distância Euclideana
i(x)	Representação do Neurônio da Entrada X
$D_{Minkowski}$	Distância Métrica de Minkowski
$D_{Manhattan}$	Distância de Manhattan
L	Lagrange

SUMÁRIO

1 INTRODUÇÃO

As maiores fontes de poluição hídrica se constituem nos lançamentos de efluentes sanitários, industriais e os lançamentos de resíduos de atividades agrosilvipastoris em corpos hídricos superficiais, que por sua vez são em grande parte responsáveis pelo abastecimento humano. Essas fontes poluidoras devem ser identificadas e quantificadas para que um tratamento adequado seja implementado antes do lançamento nos corpos hídricos. Nas indústrias, os processos que utilizam a água como matéria-prima ou como utilidades para manutenção e limpeza de equipamentos, devem ter sistemas de tratamento de efluentes para lançamentos e reuso (SPERLING, 1996; MOTA, 2003; SEMA/SP, 2000). Em relação aos efluentes sanitários, as companhias de saneamento básico são necessárias em todos os municípios do Brasil para que a coleta e o tratamento destes efluentes sejam efetivados (BRASIL, 2000; BRASIL, 2011; CANHAMERO, 2006). Quanto às atividades agrícolas, o uso de agrotóxicos desempenha um papel crítico na qualidade da água, pois estes são usados sem controle sendo carreados por águas de chuvas ou infiltram-se no solo, contaminando assim, as fontes de água (CANHAMERO, 2006).

Apesar de o Brasil possuir, aproximadamente, 12% da água doce superficial do planeta, graves problemas quanto à distribuição e utilização dos recursos hídricos disponíveis são evidentes e estão relacionados aos lançamentos supracitados. Esses problemas são ainda mais graves na região Nordeste do país, que conta com apenas 3% da água do Brasil. Agregado a este fato de forma a acelerar os problemas ambientais, o crescimento e os problemas de abastecimento tornam o Nordeste a região mais carente do país quanto à quantidade e à qualidade da água, conforme consta nos relatórios da Agência Nacional de Águas – ANA (ANA, 2017).

Neste contexto, identificar os problemas relacionados à qualidade da água é essencial, assim como as ações para mitigação desses problemas através do monitoramento e da classificação dos corpos hídricos, bem como estabelecer seu uso mais adequado. Uma das formas de estabelecer um sistema de gestão para este monitoramento da qualidade da água é a definição dos parâmetros ambientais que devem ser mensurados, como forma de analisar as influências e fatores que estão provocando efeitos de degradação da qualidade da água (BRASIL, 2000; MARQUES, 2013; GARCIA, 2012; GARCIA *et al.*, 2012; SILVA, 2015).

É importante ressaltar que em relação ao monitoramento da qualidade de um corpo hídrico, a estratégia de análise experimental constitui um dos aspectos mais impactantes no

sucesso desse monitoramento. A qualidade e a quantidade de dados experimentais necessários para classificar a qualidade de um recurso hídrico são influenciadas por variáveis que muitas vezes independem do executor do projeto, a exemplo, as condições climáticas, econômicas e até mesmo políticas institucionais. É necessário um esforço conjunto dos órgãos oficiais e centros de pesquisas em construir e manter um banco de dados atualizado sobre variáveis ambientais e mais especificamente qualidade da água e podemos afirmar que este esforço pode ser considerado ainda insuficiente e feito de forma precária muitas vezes por falta de recursos.

Os parâmetros ambientais que devem ser incluídos no monitoramento da qualidade dos corpos hídricos formam um conjunto de parâmetros físicos, químicos e biológicos que devem obedecer a limites estabelecidos por normas para cada tipo de uso a que se destina. Neste contexto, a previsão da qualidade de corpos hídricos por meio da análise de parâmetros físico-químicos e biológicos representa, assim, uma importante ferramenta para o monitoramento ambiental, que permite caracterizar ou até mesmo prever com antecedência problemas ambientais graves, como o processo de eutrofização antrópica. Contudo, a previsão da qualidade não é uma tarefa fácil devido à complexidade dos ecossistemas naturais e dos diferentes níveis de interferência do homem sobre o meio ambiente, principalmente na forma de lançamento de efluentes contaminados e modificações físicas causadas por estes (KUO *et al*., 2007; KIM ; SEO, 2015; CETESB, 2018; ANA, 2017; ALAGHA *et al*., 2017).

O monitoramento da qualidade da água consome muito tempo e recursos, sendo a análise dos dados uma tarefa difícil, porque os mesmos são multidimensionais, complexos e não lineares. Como já observado por Gillman e Hails (1997): “Um modelo ecológico, deve descrever as variações ambientais, os graus de exatidão e generalidade, bem como as relações entre os componentes bióticos e abióticos do sistema”. Sendo assim, é indispensável o uso da linguagem matemática para a construção de modelos que sirvam de base para análise dos fenômenos ambientais. Dessa forma, a modelagem permite prever o comportamento de um determinado sistema, de modo que é possível estimar sua evolução.

Estratégias para obtenção de modelos para qualidade de águas superficiais e subterrâneas têm assumido grande importância devido ao crescimento dos problemas de contaminação de corpos hídricos. Há uma grande variedade de tipos de modelos para prever a qualidade da água, contudo os modelos matemáticos empíricos têm sido mais usados devido à complexidade das relações entre os parâmetros monitorados (ZHANG e ZHANG, 2016; CHRISTIANO, 2007).

Nesta abordagem, as técnicas de Inteligência Artificial (IA), têm sido desenvolvidas como alternativas consubstanciadas para avaliação da qualidade da água.

Chou *et al*. (2018) avaliaram a qualidade da água utilizando a aprendizagem de máquina para a previsão do "Índice de Estado Trófico de Carlson", um índice amplamente utilizado para descrever a qualidade da água em reservatórios. Quatro técnicas de inteligência artificial foram usadas para auxiliar nos modelos de gestão da qualidade da água: (i) *Artificial Neural Networks* - RNA, (ii) *Support Vector Machines for regression* - SVR, (iii) *Classification And Regression Tree* - CART e (iv) *Linear Regression* - LR.

Barzegar *et al*. (2018) analisaram a utilidade do uso de Condutividade Elétrica (CE) como indicador de qualidade da água para estimar a mineralização e a salinidade da água. Os objetivos deste estudo foram explorar, pela primeira vez, o modelo de Máquina de Aprendizagem Extrema (ELM) e o modelo híbrido de máquina *wavelet*-extremo (TW-ELM) para prever CE com múltiplos degraus e empregar um método integrado para combinar vantagens dos modelos TW-ELM. Para fins comparativos, também foi desenvolvido um modelo adaptativo de inferência neuro-difusa (NF) e um modelo TW-NF. A área de estudo foi o rio Aji-Chay, na estação hidrométrica de Akhula, no noroeste do Irã.

Alagha *et al.* (2017) empregaram duas técnicas de Inteligência Artificial (IA): a) Redes Neurais Artificiais (RNAs) e b) Máquina de Vetores de Suporte (SVM); para obter uma compreensão mais profunda do processo de salinização (representado pela concentração de cloreto) em aquíferos costeiros complexos influenciados por várias fontes de salinidade. Ambos os modelos foram treinados usando 11 anos de dados de qualidade de água subterrânea de 22 poços municipais na província de Khan Younis, Gaza, Palestina.

O presente trabalho teve como objetivo geral aplicar diferentes técnicas de modelagem matemática empírica para previsão de parâmetros de qualidade de água em corpos hídricos. Tendo ainda como objetivos específicos:

- Estratificar dados ambientais disponíveis na literatura de acordo com a legislação brasileira;
- Realizar análise exploratória de dados, baseado em técnicas estatísticas multivariadas de Análise de Componentes Principais (ACP) e de Mapas Auto-Organizáveis de Kohonen (SOM);
- Aplicar, com base nos parâmetros físico-químicos da água (pH, condutividade, nitrogênio, sólidos suspensos, temperatura da água, fósforo, turbidez, cálcio, entre outros)

ao longo do ano, modelos de inteligência artificial de Redes Neurais e do tipo *Perceptron* Multicamadas (MLP) e *Neuro-Fuzzy*, e modelos estatísticos do tipo Máquina de Vetores de Suporte (SVM) para prever parâmetros de qualidade global, tais como clorofila, Demanda Bioquímica de Oxigênio (DBO) e cianobactérias;

- Modelar uma série de dados temporais para previsão futura em curto período de tempo da concentração de clorofila como parâmetro global de qualidade, com base na transformada *Wavelet* aplicada aos modelos de redes neurais RNA, NF e ao modelo estatístico de SVM.

2 REVISÃO DA LITERATURA

A difusão de modernas tecnologias como microprocessadores e microcomputadores em laboratórios, tem impulsionado a sofisticação das técnicas de instrumentação de análises dos dados experimentais na química e permitiram que a modelagem matemática passasse a desempenhar um importante papel de forma simplificada e prática, ainda que esses processos sejam complexos.

Uma variedade de contaminantes, além de um grande número de práticas destrutivas e a má gestão da água são, atualmente, ameaças aos recursos hídricos no mundo todo. Além disso, sabe-se que água de boa qualidade é um elemento fundamental para a sustentabilidade e desenvolvimento socioeconômico de um país.

2.1 QUALIDADE DA ÁGUA

O Brasil é um país rico em disponibilidade de água com aproximadamente 35 mil metros cúbicos per capita por ano, chegando a 17 vezes o que tem a Alemanha e quase 10 vezes o que a França tem. Contudo, o uso da água no Brasil é feito de forma desequilibrada, gerando carência de abastecimento em várias regiões do País. Esta escassez da água é decorrente da diminuição da sua qualidade, comprometida por desmatamentos, poluição e ocupação irregular. Um exemplo disso é que a maior parte dos esgotos sanitários não tratados é lançada nos rios, os quais degeneram a qualidade de suas águas (REBOUÇAS, 2001, *apud* PEREIRA, 2010; ANA, 2017).

> O Brasil é um dos países que possuem a maior disponibilidade de água doce do mundo. Isso traz um aparente conforto, porém os recursos hídricos estão distribuídos de forma desigual no território, espacial e temporalmente. Esses fatores, somados aos usos da água pelas diferentes atividades econômicas nas bacias hidrográficas brasileiras e os problemas de qualidade de água, geram áreas de conflito (ANA, 2017).

A poluição de cursos d'água em centros urbanos tornou-se um grande problema, pois os rios poluídos não podem mais ser usados como fonte de água. Além da crescente degradação dos ambientes aquáticos naturais, a poluição cria condições favoráveis ao aumento acelerado de doenças como amebíase, cólera, dengue, esquistossomose, febre amarela, febre tifoide, hepatite, leptospirose, malária e outras que apresentam como veículo transmissor a água. Todas oriundas das péssimas condições sanitárias que o crescimento populacional desordenado provoca em regiões carentes como o Nordeste brasileiro (CHRISTIANO, 2007).

A água, recurso natural limitado, constitui-se um bem de domínio público, conforme dispõe a Constituição Federal de 1988 (CF/88) em seus Artigos 20 e 21 e a Lei das Águas (Nº 9.433/97). Como tal, necessita de instrumentos de gestão para monitoramento de sua qualidade visando assegurar às atuais e futuras gerações água disponível em qualidade e quantidade adequadas, mediante seu uso racional e prevenindo situações hidrológicas críticas, com vistas ao desenvolvimento sustentável (WCED, 1987).

O uso das águas dentro de padrões de qualidade apropriados é uma questão importante de saúde pública no Brasil e no mundo, além de constituir uma ação eficaz na prevenção de doenças veiculadas pela água, como algumas epidemias de doenças gastrointestinais, que têm como fonte de infecção a água contaminada (PEREIRA, 2010; ANA, 2017).

Diante de tantos tipos de agressões aos mananciais aquáticos, não é fácil definir a qualidade da água. Para caracterizar uma água, são determinados diversos parâmetros, os quais representam as suas características físicas, químicas e biológicas. Esses parâmetros são utilizados como indicadores da qualidade da água e devem ter valores dentro dos limites estabelecidos para por normas de acordo com seu uso (PEREIRA, 2010; ANA, 2017).

A qualidade da água, no sentido mais amplo de seu conceito, pode ser entendida como o conjunto das características físicas, químicas e biológicas que esse recurso natural deve possuir para atender aos diferentes usos a que se destina (ARAÚJO; SANTAELLA, 2001; LIMA, 2001).

Assim, de acordo com Cunha *et al.* (2001), Secretaria do Meio Ambiente do Estado de São Paulo - SEMA/SP (2000) e Companhia Ambiental do Estado de São Paulo (CETESB, 2018), o conceito de qualidade da água depende do seu uso ou fim, possuindo valor relativo. Tucci *et al.* (2001) e a SEMA/SP (2000) acrescentam, ainda, que a qualidade, enquanto condição natural varia de um corpo hídrico para outro, uma vez que esta é diretamente influenciada pelas condições geológicas, geomorfológicas e de cobertura vegetal particular a cada bacia de drenagem.

Sendo assim, o monitoramento ambiental de um corpo hídrico tem como objetivo acompanhar, sistemática e periodicamente, a resposta do sistema de informações disponíveis, detectando-se em tempo hábil anomalias nos parâmetros de qualidade da água, de forma a evitar ou minimizar consequências indesejadas (PACHECO, 2005). No caso do monitoramento de bacias hidrográficas, o trabalho de monitoramento é muito amplo, não pode se limitar apenas a analisar a qualidade da água com base em um conjunto de parâmetros

limitados, mas deve ser considerada uma série de fatores ambientais e sociais, como o uso desse recurso e das atividades dependentes dele. O monitoramento deve refletir todos os possíveis impactos decorrentes de qualquer atividade na bacia sendo capaz de identificar alterações da qualidade da água e apontar os fatores responsáveis pelas alterações, sendo também capaz de servir de apoio para correções em programas propostos para o gerenciamento da bacia (CHRISTIANO, 2007).

Dessa maneira, na tentativa de enumerar mecanismos de monitoramento da qualidade da água de um corpo hídrico, os órgãos ambientais pré-definiram alguns indicadores físicos, químicos e biológicos que, analisados em conjunto, possibilitam verificar os níveis de poluição de um determinado manancial. Esses indicadores são denominados de parâmetros de qualidade da água, cujos mais importantes para este trabalho, são apresentados a seguir.

2.1.1 Parâmetros de qualidade da água

O consumo de águas dentro dos padrões de potabilidade adequados é uma questão relevante de saúde pública no Brasil e no mundo, além de constituir-se uma ação eficaz na prevenção de doenças veiculadas pela água, como algumas epidemias de doenças gastrointestinais, que têm como fonte de infecção a água contaminada. Para caracterizar uma água, são determinados diversos parâmetros, os quais representam as suas características físicas, químicas e biológicas. Esses parâmetros são indicadores da qualidade da água e constituem impurezas quando alcançam valores superiores aos estabelecidos para determinado uso. Cientistas de todo o planeta vêm incrementando novas técnicas para tentar avaliar de forma cada vez mais eficaz a qualidade da água dos mananciais para auxiliar em sua preservação (CETESB, 2018; VIEIRA, 2010).

A seguir, serão apresentados os principais parâmetros físicos, químicos e biológicos para o monitoramento da qualidade da água.

2.1.1.1 Parâmetros físicos

a. Temperatura

A temperatura representa de forma simplificada a energia cinética das moléculas de um corpo. A diferença de temperatura ao longo de um comprimento (gradiente de temperatura) é o fenômeno responsável pela transferência de calor em um meio. (BRASIL, 2014). A unidade usual da temperatura para fins de acompanhamento da qualidade da água é

o grau Celsius (^{o}C). Este parâmetro afeta praticamente todos os processos físicos, químicos e biológicos que ocorrem na água. Ela pode influenciar no retardamento ou aceleração da atividade biológica, na absorção de oxigênio e precipitação de compostos. A temperatura quando está um pouco elevada, resulta na diminuição de gases pela água, gerando odores e desequilíbrio ecológico (SPERLING, 1996; LIMA, 2001; VIEIRA, 2010).

A temperatura é um importante parâmetro nos corpos de água, conservando as influências de uma série de variáveis físico-químicas, por conseguinte alguns parâmetros (condutividade elétrica, DBO, oxigênio dissolvido e pH) devem ser medidos simultaneamente com a temperatura da água devido a forte relação existente entre eles (CETESB, 2018; VIEIRA, 2010).

Todos os organismos que vivem em ambientes aquáticos são adequados para uma determinada faixa de temperatura e possuem uma temperatura preferencial. Desequilíbrios são suportados, especialmente elevação na temperatura, somente até determinados limites, acima dos quais eles sofrem a morte térmica (organismos superiores) ou a inativação (micro-organismos) (CETESB, 2018; VIEIRA, 2010).

Fatores antropogênicos (usinas termelétricas e despejos industriais) ou naturais podem acarretar variações na temperatura da água. Uma informação importante, é que a temperatura desempenha papéis fundamentais nas atividades metabólicas dos organismos, na velocidade das reações químicas, e na solubilidade de substâncias (SPERLING, 1996; CETESB, 2018; BRASIL, 2014).

No Brasil, as temperaturas das águas apresentam-se, em geral, entre 20ºC a 30ºC. Porém, no sul do país, a temperatura da água no inverno pode baixar a valores entre 5ºC e 15ºC. No entanto, para consumo humano, temperaturas elevadas nas águas aumentam os cenários de reprovação ao uso, por este motivo, águas subterrâneas captadas a grandes profundidades frequentemente necessitam de unidades de resfriamento, a fim de adequá-las ao abastecimento (BRASIL, 2014).

Vale salientar que os fatores que influenciam na temperatura superficial são especialmente: profundidade, latitude, período do dia, altitude, estação do ano e taxa de fluxo (BRASIL, 2014; CETESB, 2018).

b. Condutividade Elétrica

O parâmetro responsável pela determinação da capacidade da água em transmitir corrente elétrica é a condutividade elétrica. Quanto maior for o valor da condutividade

elétrica, maior será a concentração de espécies iônicas dissolvidas e, consequentemente, representa uma forma indireta de medir a concentração de poluentes, principalmente os inorgânicos (LIMA, 2001; CETESB, 2018; BRASIL, 2014). Além disso, altos valores de condutividade elétrica podem indicar características corrosivas da água (CETESB, 2018).

Mesmo que não seja regra existe uma relação direta entre concentração de sólidos totais dissolvidos e condutividade elétrica, já que as águas naturais não são soluções simples, tal correlação é possível para águas de determinadas regiões onde exista a preponderância de um determinado íon em solução. As águas naturais apresentam teores de condutividade na faixa de 10 a 100 $\mu S.cm^{-1}$, já em ambientes poluídos por esgotos sanitários ou industriais os valores podem chegar a 1.000 $\mu S.cm^{-1}$(BRASIL, 2014).

Esteves (1988, *apud* LIMA, 2001) enfatiza que a condutividade elétrica é de grande relevância, visto que pode trazer informações tanto sobre o metabolismo do ecossistema aquático, como da síntese de matéria orgânica (redução dos valores) e decomposição (aumento dos valores), como sobre outros fenômenos que ocorram na sua bacia de drenagem. Isso permite detectar os íons mais diretamente encarregados pelo aumento da condutividade nas águas. A geologia da bacia e o regime das chuvas podem influenciar na composição iônica dos corpos d'água. A condutividade elétrica revela, ainda, as fontes poluidoras nos ecossistemas aquáticos e as diferenças geoquímicas do rio principal e seus afluentes.

c. Turbidez

O parâmetro que representa a propriedade óptica de absorção e reflexão da luz, ou seja, é responsável pelo grau de diminuição de intensidade que um feixe de luz sofre ao percorrer uma amostra de água chama-se de turbidez. Isto ocorre devido ao encontro de partículas sólidas em suspensão, como partículas inorgânicas (areia, silte, argila) e detritos orgânicos (algas e bactérias), que formam coloides e interferem na propagação da luz pela água (LIMA, 2001; SILVA, 2007; VIEIRA, 2010; BRASIL, 2014; CETESB, 2018).

Alguns exemplos de fenômenos que resultam em aumento da turbidez das águas (CETESB, 2018):

- Em épocas de chuva, a erosão das margens dos rios, que é intensificada pelo uso inadequado do solo;
- Os inúmeros efluentes industriais e esgotos sanitários. Um exemplo característico deste fato é devido às atividades de mineração, onde os elevados aumentos na turbidez vêm

levando a formação de grandes bancos de lodo em rios e modificações no ecossistema aquático.

Portanto, o aumento na turbidez diminui o processo da fotossíntese de vegetação enraizada submersa e de algas. Com isso, a redução de plantas pode, por sua vez, eliminar a produtividade de peixes, influenciando nas comunidades biológicas aquáticas. Além disso, afeta adversamente os usos doméstico, industrial e recreacional de uma água (CETESB, 2018).

d. Sólidos

A presença de sólidos de qualquer natureza na água provoca a elevação da cor e da turbidez e a diminuição da transparência, podendo afetar a biota aeróbia e facultativa devido à diminuição da produção fotossintética e, consequentemente, do oxigênio dissolvido no meio hídrico. É o material particulado não dissolvido, encontrado suspenso no corpo d'água, composto por substâncias inorgânicas e orgânicas, incluindo-se aí os organismos planctônicos (fito e zooplâncton) (CETESB, 2018; SILVA, 2007; BRASIL, 2014).

Os sólidos presentes na água existem nas seguintes formas (SILVA, 2007; BRASIL, 2014):

- Em suspensão (sedimentáveis e não sedimentáveis) são as partículas passíveis de retenção por processos de filtração; e
- Dissolvidos (voláteis e fixos). Esta parcela reflete a influência de lançamento de esgotos, além de afetar a qualidade organoléptica da água e por isso é o único referenciado para o padrão de potabilidade, com o limite de 1000mg.L^{-1} de sólidos totais dissolvidos.

O ingresso de sólidos na água pode ocorrer de duas formas, sendo elas naturais (processos erosivos, organismos e detritos orgânicos) ou antropogênicas (lançamento de lixo e esgotos) (SILVA, 2007; BRASIL, 2014).

Mesmo que os parâmetros turbidez e sólidos totais estejam ligados, eles não são sempre correspondentes. Um exemplo disto é que colocando uma pedra em uma bacia com água limpa, confere àquele ambiente elevado valor de sólidos totais, sendo que a sua turbidez pode ser quase nula (BRASIL, 2014).

Os sólidos podem causar prejuízos à vida aquática, principalmente aos peixes, depositando no leito dos rios e devastando organismos que fornecem alimentos ou, também, deteriorando os leitos de desova de peixes. Os sólidos podem manter bactérias e resíduos orgânicos no fundo dos rios, promovendo decomposição anaeróbia. Valores altos de sais

minerais, principalmente sulfato e cloreto, estão ligados à tendência de corrosão em sistemas de distribuição, além de atribuir sabor às águas (CETESB, 2018).

e. Cor

A cor da água é consequência de vários fatores, dentre eles: a) presença de substâncias em solução oriundas, de preferência, dos processos de decomposição que ocorrem no meio ambiente; b) presença de alguns íons metálicos como ferro e manganês; c) plâncton, macrófitas; d) despejos coloridos contidos em esgotos industriais; e e) material em suspensão presente na água. Essa coloração é dita aparente e por isso as águas superficiais estão mais sujeitas a aparentarem cor do que as águas subterrâneas (LIMA, 2001; CETESB, 2018; BRASIL, 2014).

As águas naturais possuem cor que varia entre zero e 200UTNs (Unidades de Turbidez) (LIMA, 2001; BRASIL, 2014). No Brasil, a Resolução CONAMA 20, de 18/06/86, aceita para água bruta até 75 unidades de cor para receber tratamento convencional e depois ser distribuída em sistemas urbanos, não devendo mostrar cor superior a cinco unidades, conforme estabelece a Portaria 36, do Ministério da Saúde.

Fazendo-se uma comparação da amostra da água com um padrão de cobalto-platina tem-se a intensidade da cor, dado em unidades de cor. Valores menores que dez unidades são dificilmente visíveis. Para a caracterização de águas para abastecimento, distingue-se a cor aparente, na qual se consideram as partículas suspensas, da cor verdadeira. Esta última tem-se após centrifugação da amostra. Deve-se salientar que para atender o padrão de potabilidade, a água deve apresentar intensidade de cor aparente inferior a cinco unidades (BRASIL, 2014).

É importante ressaltar que a coloração, realizada na rede de monitoramento, consiste basicamente na observação visual do técnico de coleta no instante da amostragem (CETESB, 2018).

f. Transparência

A transparência está associada à atividade de fotossíntese das algas, que é a capacidade que a água possui em permitir o acesso da luz solar, sendo esse parâmetro, portanto, de grande importância para o meio ambiente, em especial os recursos hídricos (MOTA, 2003; CETESB, 2018).

De acordo com os relatórios da CETESB (2018), a transparência da água pode ser medida com facilidade no campo utilizando-se um disco de Secchi, disco circular branco ou

com setores branco e preto e um cabo graduado, que é mergulhado na água até a profundidade de transparência, em que não seja mais possível visualizar o disco. A partir desta medida, é possível estimar a profundidade da zona fótica, ou seja, a profundidade de penetração vertical da luz solar na coluna d'água, que indica o nível da atividade fotossintética de lagos ou reservatórios.

Quanto maior for a quantidade de sólidos no ambiente, menor será a transparência e, consequentemente, menores serão as taxas de oxigênio dissolvido advindo do metabolismo das algas presentes no meio hídrico (em virtude da baixa penetração de luz no perfil hídrico). Portanto, ações antrópicas (desmatamento e o lançamento de efluentes in natura nos mananciais) interferem diretamente nos níveis de transparência da água, o que pode afetar a biota local (BRAGA *et al*., 2007).

2.1.1.2 Parâmetros químicos

a. Oxigênio Dissolvido

O Oxigênio Dissolvido (OD) é o parâmetro essencial para a caracterização da poluição oriunda de despejos orgânicos. O OD é indispensável para a preservação dos organismos aeróbios (CHRISTIANO, 2007; SILVA, 2007; BRASIL, 2014).

Segundo Vieira (2010), o oxigênio dissolvido na água, pode sair de duas fontes: a) endógena, que corresponde ao oxigênio produzido através da fotossíntese dos organismos aquáticos fotossintetizantes; e b) exógena, que se refere ao oxigênio atmosférico, transferido para água através da difusão.

Os valores de oxigênio dissolvido, em lagos, variam com a profundidade, sendo maior na superfície e menor no fundo, devido à estratificação térmica e outros fatores. A quebra da estratificação, em ambientes de água parada, pode ocasionar na junção do corpo d'água e a recomposição de matéria oxidável sedimentada. Nessas situações a concentração de OD pode diminuir chegando a valores críticos para muitos seres aquáticos, o que não tem relação com processos antrópicos de poluição (VIEIRA, 2010).

Além disso, é importante ressaltar algumas fontes de oxigênio nas águas: a) contribuição de incorporação de oxigênio dissolvido em corpos d'água através da superfície é proporcional à velocidade e depende das particularidades hidráulicas; b) a fotossíntese de algas. A turbidez e a cor elevadas dificultam a penetração dos raios solares e apenas poucas espécies resistentes às condições severas de poluição conseguem sobreviver. A contribuição

fotossintética de oxigênio só é expressiva após grande parte da atividade bacteriana na decomposição de matéria orgânica ter ocorrido, bem como após terem se desenvolvido também os protozoários que, além de decompositores, consomem bactérias clarificando as águas e permitindo a penetração de luz (CETESB, 2018).

A Demanda Bioquímica de Oxigênio (DBO) é tida como parâmetro de fundamental importância na caracterização do grau de poluição dos corpos hídricos. É o parâmetro que representa a quantidade de oxigênio indispensável para haver a oxidação da matéria orgânica biodegradável em condições aeróbicas. É convencionalmente usada a DBO, pois considera a medida a cinco dias, incubada a 20^{o}C, associada à fração biodegradável dos componentes orgânicos carbonáceos. O valor fixado pelo CONAMA para um rio classe II é de 5,0 mg.L-1 (CHAPRA, 1997; SEMA/SP, 2000; LIMA, 2001; MOTA, 2003; CHRISTIANO, 2007; CETESB, 2018).

Em um corpo hídrico, despejos de origem prevalentemente orgânica ocasionam os maiores crescimentos em termos de DBO. A existência de grande quantidade de matéria orgânica pode impulsionar ao completo esgotamento do oxigênio na água, provocando o desaparecimento de peixes e outras formas de vida aquática. Um elevado valor da DBO pode indicar um incremento da microflora presente e interferir no equilíbrio da vida aquática, além de produzir sabores e odores desagradáveis e, ainda, pode obstruir os filtros de areia utilizados nas estações de tratamento de água (CETESB, 2018).

A quantidade de oxigênio empregado pela oxidação química de substâncias orgânicas presentes nas águas é denominada Demanda Química de Oxigênio (DQO) e está associada com a matéria orgânica e seu potencial poluidor (SPERLING, 1996; LIMA, 2001; SILVA, 2007; CETESB, 2018; BRASIL, 2014).

O diferencial existente entre os parâmetros químicos DBO e DQO está no tipo de matéria orgânica estabilizada. À medida que a DBO se refere exclusivamente à matéria orgânica mineralizada por atividade dos micro-organismos, a DQO atinge, também, a estabilização da matéria orgânica ocorrida por processos químicos. A relação entre os valores de DQO e DBO indica a parcela de matéria orgânica que pode ser estabilizada por via biológica, ou seja, a razão DQO/DBO indica o grau de toxicidade presente na água, de tal forma que quanto maior for esta razão, maior será a toxicidade. Tanto a DBO quanto a DQO são expressas em mg.L^{-1}. A concentração média da DBO em esgotos sanitários é da ordem de 300 mg.L^{-1}, o que indica que são necessários 300 miligramas de oxigênio para estabilizar, em

um período de cinco dias e a 20ºC, a quantidade de matéria orgânica biodegradável contida em 1 litro da amostra (BRASIL, 2014).

b. pH

Segundo Christiano (2007), o pH é um índice que indica o grau de acidez, neutralidade ou alcalinidade de uma substância líquida.

De acordo com a Portaria 518/04 do Ministério da Saúde, o pH é padrão de potabilidade, devendo as águas para abastecimento público apresentar valores entre 6,0 a 9,5.

A alcalinidade da água não apresenta problemas para a saúde pública, mas traz incômodo ao paladar. Esse parâmetro representa a habilidade que um sistema aquoso tem de neutralizar ácidos. A maior fração de compostos que possa contribuir para o incremento da alcalinidade na água deve-se principalmente aos bicarbonatos. (LIMA, 2001; BRASIL, 2014).

Corpos hídricos com alta alcalinidade podem, assim, conservar por volta dos mesmos valores de pH, mesmo recebendo contribuições ácidas ou alcalinas. Os principais constituintes da alcalinidade são os bicarbonatos (HCO_3^-), carbonatos (CO_3^{-2}) e hidróxidos (OH^-). Outros ânions, como cloretos, nitratos e sulfatos, não contribuem para a alcalinidade (BRASIL, 2014).

O teor de pH é que diferencia as três formas de alcalinidade na água (bicarbonatos, carbonatos e hidróxidos), como é listado a seguir (BRASIL, 2014):

- pH> 9,4 (hidróxidos e carbonatos);
- pH entre 8,3 e 9,4 (carbonatos e bicarbonatos);
- pH entre 4,4 e 8,3 (apenas bicarbonatos).

c. Série de Nitrogênio

O nitrogênio é um dos elementos essenciais ao metabolismo de ecossistemas aquáticos. Dessa maneira, a presença de nitrogênio no meio aquático pode originar-se das fontes naturais de nitrogênio, tais como: chuva, material orgânico e inorgânico de origem alóctane, de esgotos sanitários e industriais e da drenagem de áreas fertilizadas. As formas em que o nitrogênio apresenta-se nos ambientes aquáticos podem ser: nitrato (NO_3), nitrito (NO_2), amônia (NH_3), íon amônio (NH_4), óxido nitroso (N_2O), nitrogênio molecular (N_2), nitrogênio orgânico dissolvido (aminas, aminoácidos) e nitrogênio orgânico particulado (bactérias, fitoplâncton, zooplâncton e detritos) (LIMA, 2001; SILVA, 2007; CETESB, 2018; BRASIL, 2014).

As diferentes formas dos compostos de nitrogênio encontradas em corpos d'água podem ser utilizadas como indicadores da qualidade sanitária das águas. Mota (2003) salienta que nitrogênio orgânico e amônia estão associados a efluentes e águas recém- poluídas. Com o passar do tempo, o nitrogênio orgânico é convertido em nitrogênio amoniacal e, posteriormente, se condições aeróbias estão presentes, a oxidação da amônia acontece transformando-se em nitrito e nitrato. Conforme ressalta Sperling (1996), em um corpo d'água, a determinação da parcela predominante de nitrogênio pode fornecer informações sobre o estágio da poluição. Os compostos de nitrogênio, na forma orgânica ou de amônia, referem-se à poluição recente, enquanto que nitrito e nitrato à poluição mais remota. A Resolução CONAMA n 357 estabelece limites para amônia não ionizável (NH_3), de 0,02 $mg.L^{-1}$, Nitrato, 10 $mg.L^{-1}$, e Nitrito, 1,0 $mg.L^{-1}$.

Existe uma intensa participação de bactérias no ciclo do nitrogênio, tanto no processo de nitrificação (oxidação bacteriana do amônio a nitrito, e deste a nitrato), quanto no de desnitrificação (redução bacteriana do nitrato ao gás nitrogênio). O nitrogênio é um dos mais importantes nutrientes para o crescimento de algas e macrófitas (plantas aquáticas superiores), sendo facilmente assimilável nas formas de amônio e nitrato. Em condições fortemente alcalinas, ocorre o predomínio da amônia livre (ou não ionizável), que é bastante tóxica a vários organismos aquáticos. Já o nitrato, em concentrações elevadas, está associado à doença da metahemoglobinemia, que dificulta o transporte de oxigênio na corrente sanguínea de bebês. Em adultos, a atividade metabólica interna impede a conversão do nitrato em nitrito, que é o agente responsável por esta enfermidade (CETESB, 2018; BRASIL, 2014).

A concentração de nitrogênio amoniacal é um parâmetro essencial para classificação das águas naturais e é normalmente utilizado na constituição de índices de qualidade das águas. Pois a amônia acima de 5 $mg.L^{-1}$ é um tóxico bastante restritivo à vida dos peixes e, além disso, ela provoca consumo de oxigênio dissolvido das águas naturais ao ser oxidada biologicamente (CETESB, 2018; BRASIL, 2014).

d. Fósforo

Segundo Esteves (1988), o fósforo encontra-se nas águas naturais e residuais, quase exclusivamente na forma de fosfato.

O fósforo pode ser encontrado nas seguintes formas, em corpos hídricos: orgânica e inorgânica, na forma natural (tais como dissolução de compostos do solo e decomposição da matéria orgânica) e derivadas de atividades humanas (despejos sanitários, esgotos,

detergentes, inseticidas fertilizantes) (SILVA, 2007; CETESB, 2018; BRASIL, 2014). FEITOSA *et al.* (1997) dizem que devido a ação de microrganismos, a concentração de fósforo pode ser baixa em águas com menor índice de poluição (< que 0,5mg.L^{-1}) e em águas poluídas valores acima de 1,0 mg.L^{-1}. Os compostos de fósforo são um dos fatores limitantes à vida dos organismos aquáticos e o seu controle, em um corpo d'água, é de importância fundamental no controle ecológico das algas.

O componente inorgânico solúvel é o mais importante no estudo do fósforo por funcionar como alimento para o crescimento de algas e macrófitas. O fósforo é o nutriente mais importante para o crescimento de plantas aquáticas, em razão de ser pouco encontrado em regiões de clima tropical. Quando este crescimento ocorre em excesso, prejudicando os usos da água, caracteriza-se o fenômeno conhecido como eutrofização. Além de fertilizantes, as fontes de poluição por fósforo podem ser provenientes de sabões em pó contendo este composto (BRASIL, 2005; CETESB, 2018; BRASIL, 2014).

2.1.1.3 Parâmetros biológicos

a. Coliformes

O grupo de bactérias determinado coliforme geralmente habita o intestino de homens e animais tornando-se, assim indicadora da contaminação de uma amostra de água por fezes. Como a maioria das doenças associadas com a água é transmitida por via fecal, ou seja, os organismos patogênicos, ao serem eliminados pelas fezes, atingem o ambiente aquático, podendo vir a contaminar as pessoas que utilizam de forma inadequada esta água, conclui-se que as bactérias coliformes podem ser usadas como indicadoras desta contaminação. Quanto maior a população de coliformes em uma amostra de água, maior é a chance de que haja contaminação por organismos patogênicos (LIMA, 2001; CHRISTIANO, 2007; BRASIL, 2014).

Conforme Canhamero (2006), a qualidade de uma água de abastecimento é avaliada usando organismos indicadores, ou seja, um grupo de bactérias denominados de Coliformes, e dividido em três subgrupos: coliformes totais, coliformes fecais e estreptococos fecais.

De acordo Canhamero (2006), os Coliformes Totais (CT) reúnem um grande número de bactérias, entre elas a *Eschrichia Coli*, de origem exclusivamente fecal e que dificilmente se multiplica fora do trato intestinal. O problema é que outras bactérias dos gêneros *Citrobacter*, *Eriterobacter* e *Klebsiella*, igualmente identificadas pelas técnicas laboratoriais

como coliformes totais, podem existir no solo e nos vegetais. Desta forma, não é possível afirmar categoricamente que uma amostra de água com resultado positivo para coliformes totais tenha entrado em contato com fezes. Os Coliformes Fecais pertencem a esse subgrupo dos micro-organismos que aparecem exclusivamente no trato intestinal. Em laboratório, a diferença entre coliformes totais e fecais é feita através da temperatura (os coliformes fecais continuam vivos mesmo a 44ºC, enquanto os coliformes totais têm crescimento a 35ºC). A sua identificação na água permite afirmar que houve presença de matéria fecal, embora não exclusivamente humana.

b. Comunidade Fitoplanctônica

A comunidade de vegetais microscópicos que vivem em suspensão nos corpos hídricos é denominada de fitoplâncton. Eles são constituídos principalmente por algas: clorofíceas, diatomáceas, euglenofíceas, crisofíceas, dinofíceas e xantofíceas e cianobactérias (CETESB, 2018).

O fitoplâncton pode ser utilizado como indicador da qualidade da água, principalmente em reservatórios, e, além disso, a análise da sua estrutura funciona como um meio para avaliar alguns efeitos decorrentes de variações ambientais. Esta comunidade é a base da cadeia alimentar e, portanto, a produtividade dos elos seguintes depende da sua biomassa (CETESB, 2018).

Os seres fitoplanctônicos respondem rapidamente às variações ambientais em virtude da interferência natural ou de atividade humana. Estes são indicadores de estado trófico, podendo ainda ser utilizados como indicadores de poluição por pesticidas ou metais tóxicos (presença de espécies resistentes ao cobre) em reservatórios utilizados para abastecimento. Além disso, a presença de algumas espécies em grande quantidade pode comprometer a qualidade das águas, causando limitações ao seu tratamento e distribuição. Precaução especial é dada às Cianobactérias (Cianofíceas), que possui espécies potencialmente tóxicas. A ocorrência desses organismos tem sido relacionada a eventos de mortandade de animais e de danos à saúde humana (CETESB, 2018; VIEIRA, 2010; BRASIL, 2014).

O Ministério do Meio Ambiente, por meio da Resolução CONAMA 357/2005, exige o monitoramento das células de cianobactérias para o enquadramento e classificação das águas. A Portaria MS 2914/2011, relativa às normas de qualidade para água de consumo humano, estabelece que os responsáveis por estações de tratamento de água para abastecimento público

devem realizar o monitoramento de cianobactérias e o controle das cianotoxinas nos mananciais.

c. Clorofila *a*

De acordo com Vilas *et al*. (2011), a clorofila *a* é um bioindicador de integração dos ecossistemas aquáticos, considerando que é comum a quase todos os organismos fotossintéticos, e sua concentração é amplamente utilizada para estimar a biomassa do fitoplâncton na qualidade da água e estudos ecológicos.

A clorofila *a* é a principal variável indicadora de estado trófico dos ambientes aquáticos, pois ela é a mais universal das clorofilas e representa uma boa parte do material orgânico em todas as algas planctônicas e é, por isso, um indicador da biomassa algal (CETESB, 2018).

Clorofila *a* é um parâmetro fundamental em abundância no fitoplâncton, onde este último junto com a fotossíntese são os principais componentes inter-relacionados da produção primária no ambiente aquático (ÇAMDEVÝREN *et al*., 2005).Portanto, o processo de eutrofização cultural ou antrópica pode ser monitorado através da determinação da concentração de clorofila (VIEIRA, 2010).

2.2 LEGISLAÇÃO AMBIENTAL

No Brasil, os diversos parâmetros da água a serem monitorados e quantificados por técnicas analíticas diferentes para avaliar a qualidade de um determinado corpo d'água e verificar se este apresenta condições satisfatórias para assegurar os seus usos potenciais são indicados na Resolução do Conselho Nacional do Meio Ambiente (CONAMA) na Resolução Nº 357, de 17 de março de 2005 (BRASIL, 2005). Portanto, esta resolução 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.

A Resolução do CONAMA n^0 357 classifica as águas doces (salinidade < 0,05%), salobras (salinidade entre 0,05% e 3%) e salinas (salinidade > 3%), segundo os usos preponderantes atuais ou futuros a que se destinam, em sistema de 13 classes, com os respectivos padrões de qualidade. As águas doces são distribuídas nas classes (especial, 1, 2, 3 e 4), e as águas salinas e salobras, em especial e classes um a três.

As classes para as águas doces segundo seus usos preponderantes são:

> I - classe especial: águas destinadas:
> a) ao abastecimento para consumo humano, com desinfecção;
> b) a preservação do equilíbrio natural das comunidades aquáticas; e,
> c) a preservação dos ambientes aquáticos em unidades de conservação de proteção integral.
> II - classe 1: águas que podem ser destinadas:
> a) ao abastecimento para consumo humano, após tratamento simplificado;
> b) a proteção das comunidades aquáticas;
> c) a recreação de contato primário, tais como natação, esqui aquático e mergulho, conforme Resolução CONAMA no 274, de 2000;
> d) a irrigação de hortaliças que são consumidas cruas e de frutas que se desenvolvam rentes ao solo e que sejam ingeridas cruas sem remoção de película; e
> e) a proteção das comunidades aquáticas em Terras Indígenas.
> III - classe 2: águas que podem ser destinadas:
> a) ao abastecimento para consumo humano, após tratamento convencional;
> b) a proteção das comunidades aquáticas;
> c) a recreação de contato primário, tais como natação, esqui aquático e mergulho, conforme Resolução CONAMA no 274, de 2000;
> d) a irrigação de hortaliças, plantas frutíferas e de parques, jardins, campos de esporte e lazer, com os quais o público possa vir a ter contato direto; e
> e) a aquicultura e a atividade de pesca.
> IV - classe 3: águas que podem ser destinadas:
> a) ao abastecimento para consumo humano, após tratamento convencional ou avançado;
> b) a irrigação de culturas arbóreas, cerealíferas e forrageiras;
> c) a pesca amadora;
> d) a recreação de contato secundário; e
> e) a dessedentação de animais.
> V - classe 4: águas que podem ser destinadas:
> a) a navegação; e
> b) a harmonia paisagística.

Em função dos usos preponderantes das classes relativas à água doce estabelecidos na resolução, a Classe Especial pressupõe os usos mais nobres, e a Classe 4, os menos nobres.

No Art. 42 da Resolução do CONAMA 357/05 cita-se que:

> Enquanto não aprovados os respectivos enquadramentos, as águas doces serão consideradas classe 2, as salinas e salobras pertencem a classe 1, exceto quando as condições atuais de qualidade forem melhores, prevalecendo a aplicação da classe mais rigorosa correspondente (BRASIL, 2005).

Em complemento a esta Resolução, o Art. 5 da Resolução do CONAMA 430/2011, menciona que "os efluentes não poderão conferir ao corpo receptor características de qualidade em desacordo com as metas obrigatórias progressivas, intermediárias e final do seu enquadramento".

Os corpos hídricos do Estado de Pernambuco ainda não foram reenquadrados, e a Agência Estadual de Meio Ambiente de Pernambuco (CPRH) considera os mananciais de água doce pertencentes à classe 2. Este fato é importante na melhoria das condições naturais da qualidade da água no Estado, restringindo aos limites dos parâmetros e aos usos preponderantes atuais ou pretendidos da água doce de classe 2.

Na Tabela 1 está apresentado um resumo do tipo de tratamento requerido para as águas brutas ao abastecimento público.

Tabela 1 - Classificação dos corpos de águas doces e tratamento requerido segundo CONAMA 357/05.

Classificação das águas	Tratamento Requerido
Classe Especial	Desinfecção
Classe 1	Simplificado
Classe 2	Convencional
Classe 3	Convencional ou avançado

Fonte: Adaptado de BRASIL (2005).

De acordo com a resolução CONAMA 357/05, apenas as águas enquadradas na classe especial, classe 1, classe 2 e classe 3 podem ser utilizadas para o abastecimento humano. Deve-se salientar que uma correta classificação do corpo hídrico implica em melhor qualidade da água tratada assim como o menor custo de tratamento na ETA (BRASIL, 2005).

A Agência Estadual de Meio Ambiente (CPRH), de forma complementar, classifica a qualidade da água das bacias hidrográficas com base nos usos preponderantes, de modo a atender ao uso mais restritivo estabelecido no grupo:

1. **Não comprometida** - Enquadram-se, nesta categoria, os corpos de água que apresentam condições de qualidade de água compatíveis com os limites estabelecidos para a classe especial das águas doces, salinas e salobras e classe 1 das águas doces (Resolução CONAMA n° 357/05). Estes corpos d'água apresentam qualidade da água ótima, com níveis desprezíveis de poluição.
2. **Pouco Comprometida -** Enquadram-se, nesta categoria, os corpos de água que apresentam condições de qualidade de água compatíveis com os limites estabelecidos para a classe 2 das águas doces e a classe 1 das águas salinas e salobras (Resolução CONAMA n° 357/05). Estes corpos d'água apresentam qualidade da água boa, com níveis baixos de poluição.
3. **Moderadamente Comprometida** - Enquadram-se, nesta categoria, os corpos de água que apresentam condições de qualidade de água compatíveis com os limites para a classe 3 das águas doces e a classe 2 das águas salinas e salobras (Resolução CONAMA n° 357/05). Estes corpos d'água apresentam qualidade da água regular, com níveis aceitáveis de poluição.
4. **Poluída -** Enquadram-se, nesta categoria, os corpos de água que apresentam condições de qualidade de água compatíveis com os limites estabelecidos para a classe 4 das águas

doces e a classe 3 das águas salinas e salobras (Resolução CONAMA n° 357/05). Estes corpos d'água apresentam qualidade da água ruim, com poluição acima dos limites aceitáveis.

5. **Muito Poluída** - Enquadram-se, nesta categoria, os corpos de água que não se enquadram em nenhuma das classes acima estabelecida. Estes corpos d'água apresentam qualidade da água péssima, com poluição muito elevada.
6. **Não Monitorada**

Na Resolução CONAMA 357/05 para classificação das águas, determinou as condições padrões de qualidade das águas e os limites individuais para cada substância em cada classe. Na Tabela 2 estão dispostos os valores limites dos parâmetros empregados na Resolução desta tese.

Tabela 2 - Limites dos padrões dos parâmetros monitorados segundo a Resolução CONAMA N°357/05 para águas doces (continua).

Parâmetros	Classe 1	Classe 2	Classe 3	Classe 4
Salinidade (0/00)	-	≤0,50	-	-
Clorofila a ($\mu g.L^{-1}$)	≤10	≤30	≤60	-
Densidade de Cianobactérias ($cel.mL^{-1}$)	≤20.000	≤50.000	≤100.000, dessedentação de animais ≤50.000	-
pH	6 a 9	6 a 9	6 a 9	6 a 9
OD ($mg\ O_2.L^{-1}$)	≥6	≥5	≥4	≥2
DBO ($mg\ O_2.L^{-1}$)	≤3	≤5	≤10	-
Nitrogênio Amoniacal Total ($mg\ NH_3.L^{-1}$)	4,5 (pH≤7,5) 2,4 (7,5<pH ≥8,0) 1,2 (8,0<pH ≥8,5) 0,6 (pH>8,5)	4,5 (pH≤7,5) 2,4 (7,5<pH≥8,0) 1,2 (8,0<pH≥8,5) 0,6 (pH>8,5)	16,1 (pH≤7,5) 6,8 (7,5<pH ≥8,0) 2,7 (8,0<pH ≥8,5) 1,2 (pH>8,5)	-
Fósforo Total ($mg\ P.L^{-1}$)	Lêntico ≤0,02 intermediário e tributário de lêntico ≤0,025 lótico e tributário de intermediário≤0,1	Lêntico ≤0,03 intermediário e tributário de lêntico ≤0,05 lótico e tributário de intermediário≤0,1	Lêntico ≤0,05 intermediário e tributário de lêntico ≤0,075 lótico e tributário de intermediário≤0,15	-

Tabela 2 - Limites dos padrões dos parâmetros monitorados segundo a Resolução CONAMAA N°357/05 para águas doces **(**continuação).

Parâmetros	Classe 1	Classe 2	Classe 3	Classe 4
Sólidos Dissolvidos Totais ($mg.L^{-1}$)	500	500	500	-
Cor ($mg\ Pt.L^{-1}$)	-	≤75	≤75	-
Turbidez (UNT)	≤40	≤100	≤100	-
Nitrato ($mg\ N.L^{-1}$)	≤10	≤10	≤10	-
Nitrito ($mg\ N.L^{-1}$)	≤1,0	≤1,0	≤1,0	-
Cloreto Total ($mg\ Cl.L^{-1}$)	250	250	250	-
Sulfato Total ($mg\ SO_4.L^{-1}$)	250	250	250	-
Ferro Dissolvido ($mg\ Fe.L^{-1}$)	0,3	0,3	5,0	-
Zinco total ($mg\ Zn.L^{-1}$)	≤0,18	≤0,18	≤5	-
Coliformes Termotolerantes (NMP/100mL)	≤200 em 80% de 6 amostra/ano	≤1000 em 80% de 6 amostra/ano	≤2.500 contatos secundário ≤1.000 animais confinados ≤4.000 demais usos	-

Fonte: Adaptado de BRASIL (2005)

Para o monitoramento e classificação dos mananciais, faz-se necessária sua caracterização através da quantificação dos diversos parâmetros físicos, químicos e microbiológicos definidos na Tabela 2. É importante salientar que a determinação de alguns desses parâmetros exige um considerável custo com técnicas mais sofisticadas, a exemplo das medidas de concentrações de clorofila-a, cianobactérias e fosforo total, o que torna o desenvolvimento de métodos alternativos de caracterização da qualidade da água ou identificação de parâmetros de menor de custo uma área em constante desenvolvimento e valorada principalmente pelo desenvolvimento de técnicas matemáticas como a estatística multivariada e de inteligência artificial para entendimento, inferência ou previsão de tais parâmetros (CHANG; LIU, 2015, *apud* CHOU *et al*., 2018).

2.3 MODELAGEM DA QUALIDADE DA ÁGUA

A degradação dos recursos hídricos tem aumentado a necessidade de desenvolvimento de projetos relacionados à qualidade da água e ao seu monitoramento. A previsão do comportamento de corpos d'água, através da medida de parâmetros ambientais, funciona como uma importante ferramenta no combate a problemas ambientais, como processo de eutrofização de reservatórios. Entretanto, esta é uma tarefa difícil devido à complexidade dos processos físico-químicos e biológicos causadores desses problemas (KUO *et al*., 2007).

A modelagem permite prever o comportamento de um determinado sistema, de modo que é possível estimar sua evolução. Os ecossistemas são altamente dinâmicos e não lineares, sendo assim, é impossível prever sua evolução sem um modelo flexível capaz de se ajustar aos complexos sistemas ecológicos (ZHANG *et al*., 2007, *apud* CHRISTIANO, 2007).

Quando se refere à modelagem para inferência de parâmetros de qualidade de água pode-se recorrer a modelos fenomenológicos e modelos empíricos sendo estes últimos mais utilizados devido à complexidade e limitações dos primeiros, os quais requerem a especificação de um grande número de parâmetros em função também das diferentes condições de um reservatório, por exemplo. É importante ressaltar que o interesse na análise, inferência e a previsão dos parâmetros da qualidade da água por modelos empíricos têm aumentado substancialmente nos últimos anos, devido à crescente disponibilidade de métodos estatísticos e de sistemas inteligentes para modelagem empírica.

A modelagem através de sistemas inteligentes para ambientes aquáticos vem sendo utilizada uma vez que estes sistemas são altamente complexos, tornando uma modelagem fenomenológica difícil, sendo os modelos empíricos uma alternativa bastante interessante.

Pode-se perceber, diante dos trabalhos obtidos na literatura consultada, que vários modelos vêm e estão sendo desenvolvidos ao longo dos anos para estimativa de parâmetros de qualidade da água. Nesse sentido, técnicas estatísticas multivariadas, regressões, aprendizagem de máquinas baseadas em estatística (SVM) e técnicas de sistemas inteligentes como RNA e *Neuro-fuzzy*, passam a ter uma relevância e importância para a elaboração de modelos que caracterizem a qualidade da água e que englobem variáveis especificas para cada uso da água.

Atualmente pode-se destacar a área de aprendizado de máquinas, com técnicas de sistemas inteligentes como redes neurais artificiais e modelos híbridos (*neuro-fuzzy*) para o ajuste de funções multivariadas não lineares (modelos empíricos), dentre outras, e para

classificação com Mapas Auto-organizáveis de Kohonen, para clusterização. O termo sistemas inteligentes ou inteligência artificial é atribuído aos desenvolvimentos computacionais, pois são sistemas inspirados no comportamento humano ou que tentam reproduzi-lo mesmo estando longe de ser autônomo em inteligência ou raciocínio. Apesar desta ressalva, esses sistemas têm encontrado grande aceitação em diversas áreas do conhecimento, e muitas vezes apresentam desempenho superior quando comparados aos métodos convencionais utilizados para descrever os processos.

As redes neurais são modelos baseados em comportamento de um Sistema físico complexo, que mapeia um conjunto de dados de entrada-saída sem possuir nenhum outro conhecimento prévio do processo, apenas se baseando no histórico dos dados. A Rede Neural Artificial (RNA) é um tipo de técnica de diagnóstico baseado em sistemas inteligentes e é uma poderosa ferramenta para a análise multivariada não linear, além disso, tem se mostrado uma ferramenta essencial para o acompanhamento e modelagem empírica de diferentes tipos de fenômenos.

O sistema *neuro-fuzzy* representa o sistema *fuzzy* - derivado do conceito de conjuntos *fuzzy*, difere dos sistemas lógicos tradicionais em suas características e seus detalhes. Nesta lógica, o raciocínio exato corresponde a um caso limite do raciocínio aproximado, sendo interpretado como um processo de composição de relações nebulosas constitui a base para o desenvolvimento de métodos e algoritmos de modelagem e controle de processos - na forma de redes sujeitas a treinamento, por técnicas análogas às redes neurais. O método de treinamento é o ajuste de parâmetros, com a intenção de minimizar a função erro entre as saídas desejadas e as apresentadas pela rede.

Como máquina de aprendizado estatístico tem-se atualmente se referenciado muito a *Support Vector Machine* (SVM) que se trata de um método computacional de aprendizado de dados baseado na teoria estatística do aprendizado. Este usa o princípio estrutural da minimização de risco com uma boa habilidade de generalização. Ele pode solucionar o problema encontrado pelos métodos convencionais na avaliação da qualidade de água e superar os defeitos de uma lenta velocidade de treinamento, pobre generalização de rede e baixa precisão nas redes neurais artificiais. Deve-se salientar que as técnicas citadas acima são aplicadas tantos para dados estacionários com na forma de séries temporais como modelos regressivos.

Na utilização das técnicas de aprendizado de máquinas para o tratamento de séries temporais verifica-se na literatura que frequentemente são utilizadas a conjunção destas com

as Transformada *wavelet*s como um pré-tratamento dos dados para a modelagem sendo está uma tarefa não muito fácil, pois para isso o usuário desta técnica teria que deter o conhecimento de técnicas de tratamento de sinais (séries temporais) e a de sistemas inteligentes e estatística.

O nome *Wavelet*, também traduzido como "ondaleta" representa uma ferramenta de análise de sinal cuja função é adaptar pequenas ondas a um conjunto de dados com a finalidade de extrair informações tais como padrões, fazer filtragens (em caso de imagem digital), ajustes de curvas, etc.

A Transformada *Wavelet* Discreta utilizada para a finalidade de conjunção é a transformada correspondente à transformada contínua de *Wavelet* para funções discretas (sinais discretos e séries temporais). Esta transformada é utilizada para analisar sinais digitais, e também na compressão de imagens digitais e no tratamento de séries temporais, sendo o resultado de suas aplicações denominado de assinatura de uma série temporal.

Na extração de conhecimento diferente do ajuste de função considerado no modelo empírico de dados empíricos de qualidade dá água, a utilização de sofisticadas ferramentas de diagnóstico pode fornecer a análise e a visualização do conjunto de dados multidimensionais e deve ser considerada. Técnicas de diagnóstico em estatística as quais são muito utilizadas têm suas fragilidades referentes principalmente a não linearidade do fenômeno em questão. De uma forma geral, estas não podem fornecer uma descrição eficiente das relações de causa e efeito, os modelos fenomenológicos requerem a especificação de um grande número parâmetros em função também das diferentes condições dos reservatórios. Por outro lado, as técnicas de diagnóstico baseadas em estatísticas são preferíveis para a implementação de extração de conhecimento em dados de qualidade da água como a técnica de análise multivariada de Análise Componentes Principais (ACP), têm sido amplamente desenvolvidos na análise do sistema hidrológico. No entanto, as limitações dos métodos de análise multivariada clássicos são bem conhecidas. RNA na extração de conhecimento oferece uma alternativa aos métodos estatísticos tradicionais para acompanhamento ideal e determinação do sistema dinâmico, e tem atraído atenção considerável. Neste contexto, o Mapa Auto-organizável de Kohonen é uma técnica de análise de padrões baseado em rede neural com aprendizado não supervisionado, que tem sido amplamente aplicada na análise de dados de qualidade da água e por isso será investigada também neste trabalho.

O modelo de Kohonen é uma rede neural sem supervisão que usa neurônios adaptativos para receber sinais de um evento espacial, sendo em dados ou medidas, como

situações ou frequência. Tendo assim um modelo com neurônios da camada de saída disputando entre si a representabilidade das informações apresentadas aos neurônios de entrada. Segundo, as redes de Kohonen possuem duas habilidades fundamentais: A primeira está em modelar a distribuição das entradas, através da modelagem da função distribuição de probabilidade dos vetores de entrada usada durante o treinamento. A segunda é a de criar mapas que preservam a topologia.

Os modelos desempenham um papel importante na gestão dos recursos hídricos, visto que facilita a tomada de decisão e a comunicação entre os gestores ambientais e a sociedade. Além disso, é importante ressaltar que grande dificuldade para que a análise da qualidade da através de modelos matemáticos está centrada no fato que é necessário um elevado número de amostras para que se tenha um elevado número de dados de entrada para os modelos, e consequentemente, tem-se um elevado custo para obtenção desses dados.

Muitos programas de monitoramento são fomentados pelo poder público ou por projetos de pesquisas que não têm verbas para que os dados sejam obtidos rotineiramente, e algumas vezes, nem todos os parâmetros são medidos por conta do elevado custo, a exemplo da clorofila-a e do fósforo total.

Nos próximos capítulos serão examinadas as principais técnicas utilizadas para avaliação da qualidade da água neste trabalho, chamando a atenção do amplo conhecimento que a manipulação das mesmas exige, assim como de diferentes ferramentas que existem para aplicação das mesmas.

2.3.1 Análise de Componentes Principais (ACP)

Os parâmetros físicos, químicos e biológicos que caracterizam a qualidade das águas sofrem grandes variações no tempo e no espaço, havendo a necessidade de um programa de monitoramento sistemático para obter a real estimativa da variação da qualidade das águas superficiais. Em geral, um programa de monitoramento inclui coletas frequentes nos mesmos pontos de amostragem e análise em laboratório de grande número de parâmetros, resultando em matriz de grandes dimensões e complexa interpretação. Muitas vezes, pequeno número desses parâmetros contém as informações químicas mais relevantes, enquanto a maioria das variáveis adiciona pouco ou nada à interpretação dos resultados em termos de qualidade.

A estatística multivariada consiste em um conjunto de métodos estatísticos utilizados em situações nas quais várias variáveis são medidas simultaneamente, em cada elemento amostral. Em geral, as variáveis são correlacionadas entre si e quanto maior o número de

variáveis, mais complexa torna-se a análise por métodos comuns de estatística univariada. Embora historicamente o uso de métodos multivariados esteja relacionado com trabalhos na Psicologia, Ciências Sociais e Biológicas, há algum tempo eles têm sido aplicados em um grande universo de áreas diferentes, como: Educação, Geologia, Química, Física, Engenharia, Ergonomia, entre outros. Esta expansão na aplicação dessas técnicas somente foi possível graças ao grande avanço da tecnologia computacional e ao grande número de *softwares* estatísticos com módulos de análise multivariada implementados (MINGOTI, 2005).

Portanto, a estatística multivariada é uma alternativa para tratar a complexidade da modelagem e compreensão das relações multivariáveis existentes no tratamento de dados de química analítica ambiental como, por exemplo, dos parâmetros hidrológicos, físicos, químicos e biológicos no estudo da qualidade das águas. Um dos objetivos da utilização da análise multivariada é reduzir a representação dimensional dos dados, organizando-os em uma estrutura que facilita a visualização e interpretação de todo o conjunto de dados. O que pode ser interessante para um pré-processamento de dados para a RNA. Sendo assim, os métodos estatísticos multivariados consideram as amostras e as variáveis em seu conjunto, permitindo extrair informações complementares que a análise univariada não consegue evidenciar.

A estatística multivariada divide-se, basicamente, em dois grupos. O primeiro consiste em técnicas exploratórias de sintetização (ou simplificação) da estrutura de variabilidade dos dados, fazem parte os métodos de análise de componentes principais, análise fatorial, análise de correlações canônicas, análise de agrupamento, análise discriminante e análise de correspondência. O segundo consiste em técnicas de inferência estatística, encontram-se os métodos de estimação de parâmetros, testes de hipóteses, análise de variância, de covariância e de regressão multivariada (MINGOTI, 2005).

Segundo Mingoti (2005), os métodos de estatística multivariada são utilizados com o propósito de simplificar ou facilitar a interpretação do fenômeno que está sendo estudado através da construção de índices ou variáveis alternativas que sintetizem a informação original dos dados; construir grupos de elementos amostrais que apresentam similaridade entre si possibilitando a segmentação do conjunto de dados original; investigar as relações de dependência entre as variáveis respostas associadas ao fenômeno e outros fatores (variáveis explicativas), muitas vezes, com objetivos de previsão, comparar populações ou validar suposições através de testes de hipóteses.

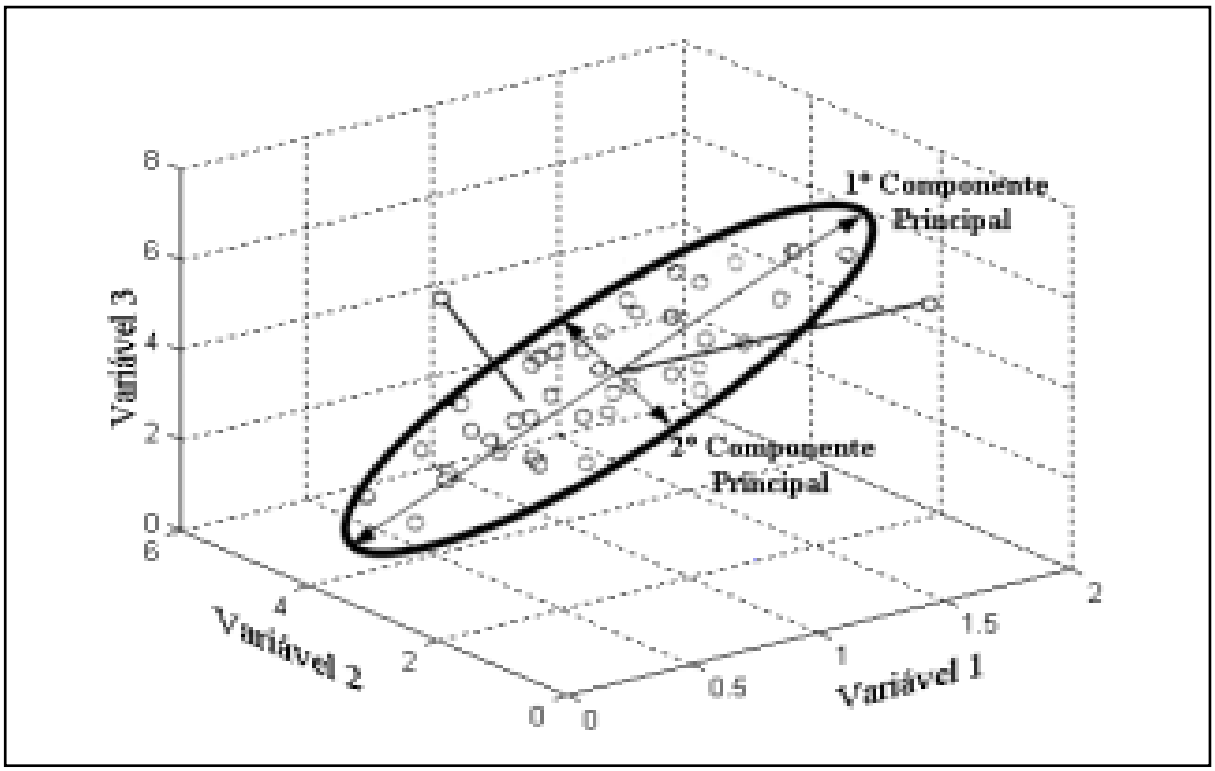
Uma técnica de estatística multivariada que recebe grande interesse e tem se tornado popular em modelagem ecológica ambiental é a Análise de Componentes Principais (ACP). A

Análise de ACP é um procedimento matemático que utiliza uma transformação ortogonal (ortogonalização de vetores) para converter um conjunto de observações de variáveis possivelmente correlacionadas num conjunto de valores de variáveis linearmente não correlacionadas, denominados de componentes principais (Figura 1). O número de componentes principais é menor ou igual ao número de variáveis originais (*CARVALHO et al., 2010).*

O novo conjunto de variáveis são ortogonais entre si e, portanto, não correlacionadas. As primeiras componentes principais (Figura 1) explicam a maior parte da variância total contida no conjunto de dados e podem ser usadas para representá-lo, que contém *m* variáveis, de modo que é possível compactar grande parte da informação linear desse sistema em apenas *k* novas variáveis, onde *k* <*m*. Assim, a i-ésima componente principal de um conjunto de *m* variáveis é definida segundo a Equação 1, na qual Z_i é a componente principal, os a_{ij} são os *loading* e X_j são as variáveis originais.

$$Z_i = a_{i1}X_1 + a_{i2}X_2 + \cdots + a_{im}X_m \qquad (1)$$

Figura 1- Interpretação geométrica das componentes principais.



Fonte: Carvalho *et al.* (2010).

A técnica de Análise de Componente Principal é muito utilizada em estudos de modelagem em ambientes aquáticos e estudos ecológicos, pois oferece um método objetivo para manusear grande quantidade de parâmetros, reduzindo a complexidade de sistemas multidimensionais.

Alguns atores relatam que a maior preocupação quanto à aplicação da análise de componentes principais é o entendimento do modo e ação ou comportamento dos componentes de um sistema ou subsistema, por isso, é importante que se tenha conhecimento fenomenológico do sistema.

Segundo Çamdevýren *et al.* (2005), a análise de ACP, a qual é muito utilizada em estudos de modelagem em ambientes aquáticos e estudos ecológicos, oferece um método objetivo para manusear uma grande parcela de dados abióticos e bióticos e ainda é um redutor de complexidade de sistemas multidimensionais pela maximização dos componentes principais e eliminação de componentes inválidos.

Podemos observar através da literatura consultada que a técnica de ACP vem sendo utilizada para o tratamento de dados de qualidade das águas assim como de dados da química analítica ambiental em geral. Contudo, sua eficácia depende da qualidade e quantidade de dados de campo, que, em geral, tendem a ser esparsos, principalmente quando se trata de sub-bacias inteiras. Apesar disso atualmente ainda não é muito frequente a utilização.

A técnica ACP foi utilizada no presente trabalho buscando um conhecimento maior das relações existentes entre variáveis e amostras no conjunto de dados com informações relevantes, ou seja, dados realmente úteis para a análise da qualidade de água.

2.3.2 Redes Neurais Artificiais

A difusão de modernas tecnologias como microprocessadores e microcomputadores em laboratórios, tem impulsionado a sofisticação das técnicas de instrumentação de análises dos dados experimentais na química e permitiram que a modelagem matemática passasse a desempenhar um importante papel de forma simplificada e prática, ainda que esses processos sejam complexos.

Uma variedade de contaminantes, além de um grande número de práticas destrutivas e a má gestão da água são, atualmente, ameaças aos recursos hídricos no mundo todo. Além disso, sabe-se que água de boa qualidade é um elemento fundamental para a sustentabilidade e desenvolvimento socioeconômico de um país.

A junção de novas tecnologias como a inteligência artificial, computadores mais potentes e a preocupação com os recursos hídricos são os principais motivadores deste trabalho.

O final da década de 1980 ficou marcado com o ressurgimento da área de Redes Neurais Artificiais (RNAs), também conhecida como conexionismo ou sistemas de processamento paralelo e distribuído. Essa forma de computação não-algorítmica é caracterizada por sistemas, que de algum nível, relembram a estrutura do cérebro humano (BRAGA *et al.*, 2007).

As Redes Neurais Artificiais, como modelos simplificados dos neurônios humanos, possuem a capacidade de "aprender" o comportamento de um sistema físico complexo. Neste tipo de abordagem não é necessária a descrição matemática detalhada dos fenômenos envolvidos no processo, apenas o conhecimento das variáveis e amostras envolvidas no mesmo (modelo caixa preta). Uma aplicação dessa capacidade é o monitoramento e tratamento matemático e/ou estatístico de parâmetros ambientais, visto que é um importante instrumento de prevenção e controle de fenômenos naturais.

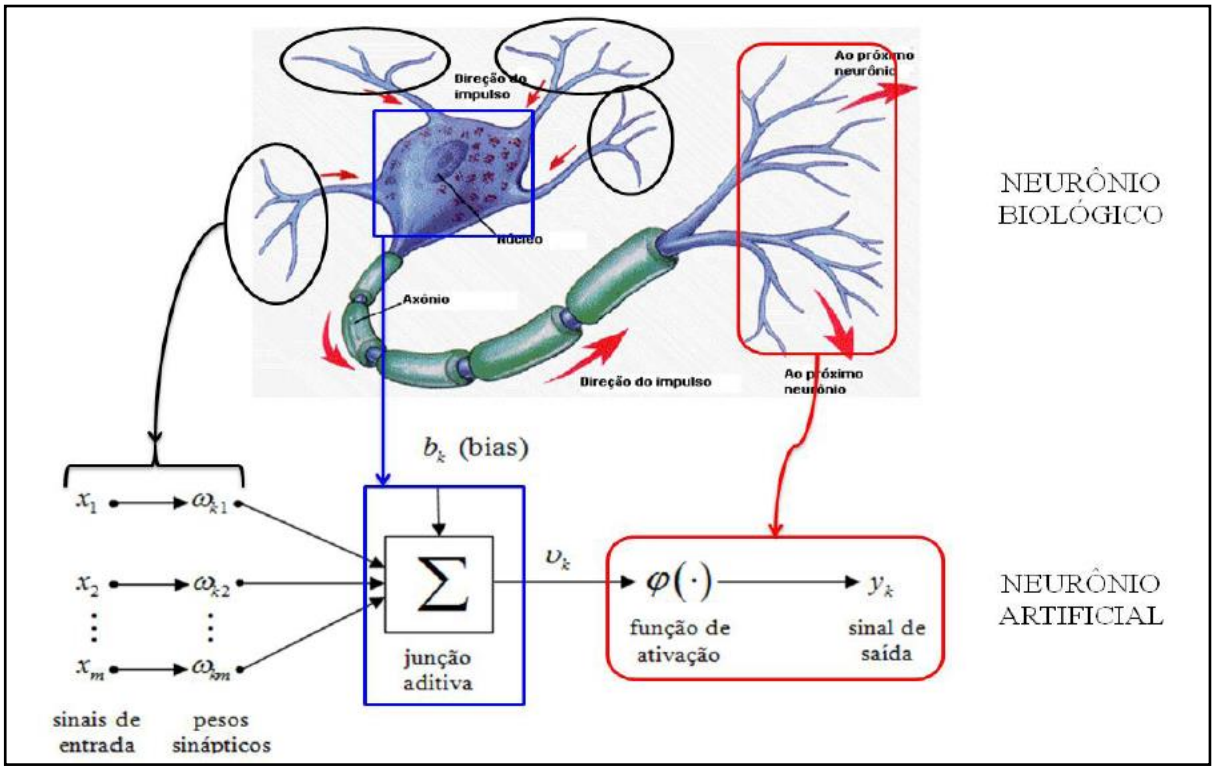
Segundo Braga *et al.* (2007), as RNAs são sistemas paralelos distribuídos compostos por unidades de processamento simples (neurônios artificiais) que calculam determinadas funções matemáticas (normalmente não-lineares). Tais unidades são dispostas em uma ou mais camadas e interligadas por um grande número de conexões, geralmente unidirecionais. Na maioria dos modelos essas conexões estão associadas a pesos, os quais armazenam o conhecimento adquirido pelo modelo e servem para ponderar a entrada recebida por cada neurônio da rede.

As RNAs, como ajuste de função, possuem duas fases de processamento, a de aprendizagem e de utilização, que seria a própria aplicação da rede. A primeira consiste no ajuste dos pesos das conexões em resposta ao estímulo apresentado à rede neural (histórico de dados), e a segunda consiste na maneira pela qual a rede responde a um estímulo de entrada sem que ocorram modificações em sua estrutura de aprendizagem (OLIVEIRA JUNIOR *et al.*, 2007; CARVALHO *et al*., 2009; CARVALHO *et al*., 2010).

Por mais diversas que sejam as aplicações existentes de redes neurais, o aspecto comum a todas é a capacidade de estabelecerem associações entre entradas e saídas conhecidas, através da experimentação de um grande número de situações. As informações de entrada são colocadas em uma rede de nódulos que interagem matematicamente entre si.

Baseado nestas informações surge um mapeamento do modelo entrada-saída, ou seja, as interações entre os nódulos são bem definidas e ajustadas até que as relações entrada-saída desejadas sejam apropriadamente obtidas. A Figura 2 mostra estas relações de entrada-saída através de uma analogia com o neurônio biológico.

Figura 2- Representação esquemática da analogia entre neurônio biológico e neurônio artificial.



Fonte: Souza (2011).

2.3.2.1 Redes Perceptron Multicamadas (Multilayer Perceptron -MLP)

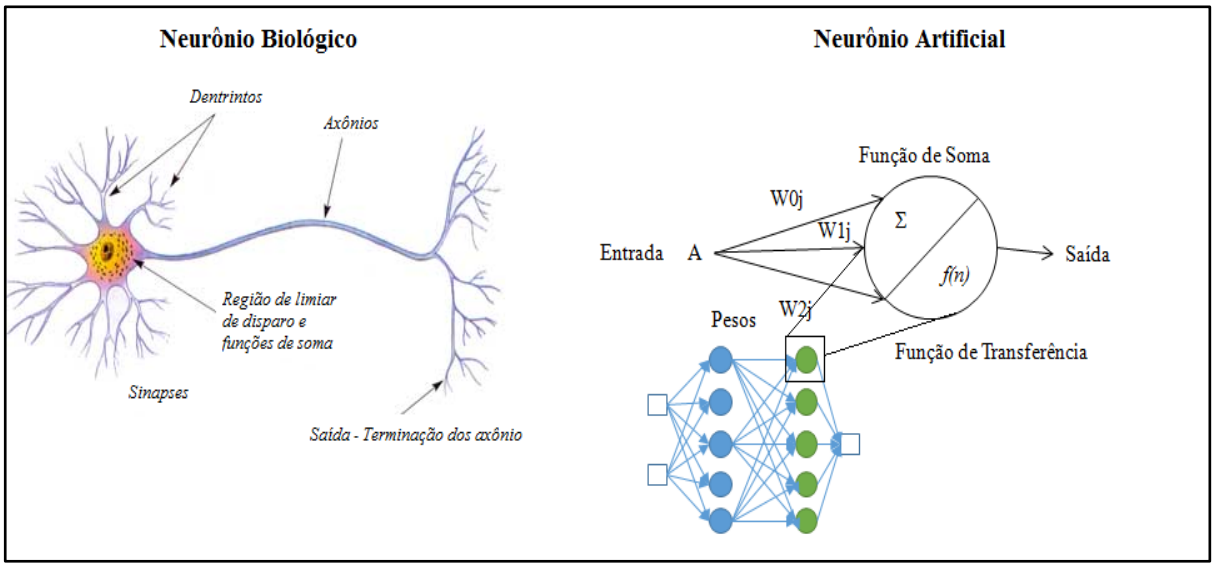
Há um grande número de tipos diferentes de redes neurais que são estudados na atualidade, desde algoritmos mais simples como o *perceptron* até a utilização de programas computacionais extremamente complexos. Porém, o maior número de aplicações práticas conhecidas (cerca de 95%) concentra-se no tipo redes de multicamadas direta MLP (*Perceptron* de Múltiplas Camadas).

Essas RNA são compostas de vários elementos computacionais simples (nodos) que interagem localmente. A arquitetura destes modelos é especificada pelas características do nodo, topologia da rede e algoritmo de treinamento.

Os nodos em RNA são processadores bastante simples inspirados por seus similares biológicos (neurônios cerebrais). Como enfatizado na relação (Figura 3) abaixo os neurônios

artificiais dispõem de entradas (semelhante biologicamente aos dendritos), pesos (semelhante às sinapses), funções de somas, onde são processados os estímulos captados pela entrada, e a funções de transferências (limiar de disparo do neurônio).

Figura 3 - Comparação dos neurônios artificial e biológico.



Verifica-se que o neurônio artificial realiza seus cálculos baseados em suas informações de entrada. Ele faz o somatório do produto entre os vetores de entrada *A* e os pesos *Wj*, subtrai a ativação residual interna, representado como *threshold* Th_{ij} ou Baias b_{ij}, e então passa este resultado para uma forma funcional ou de transferência, *f( )*, mostrada na Equação 2.

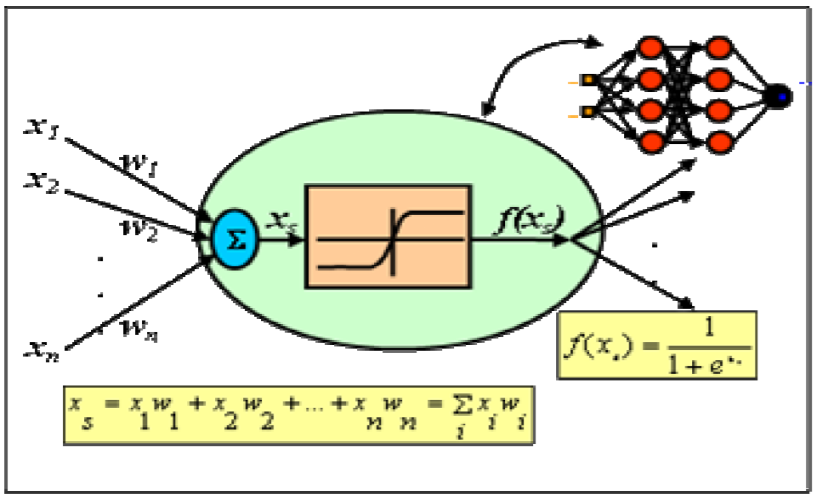
$$f(W_j.A - Th_i) = f(\sum_{i=1}^{n}(W_{ij}.a_i - Th_i)) \quad (2)$$

Neste caso, pode-se utilizar qualquer função, mas os matemáticos e cientistas da computação reportam as funções sigmoides como as mais vantajosas para a maior parte das aplicações.

Redes *Perceptron* de Multicamadas são computacionalmente mais poderosas do que as redes sem camadas escondidas. As MLP (Figura 4) podem tratar dados que não são linearmente separáveis. A precisão obtida e a implementação da função dependem do número de neurônios utilizados nas camadas escondidas. O processamento realizado por cada neurônio é definido pela combinação dos processamentos realizados pelos neurônios da camada anterior que estão conectados a ele. A partir da primeira camada escondida até a

camada de saída, as funções implementadas se tornam cada vez mais complexas. Essas funções definem como é formada a divisão do espaço de decisão (LIU *et al.*, 2008).

Figura 4 - Rede Neural *Feedforward* "típica".

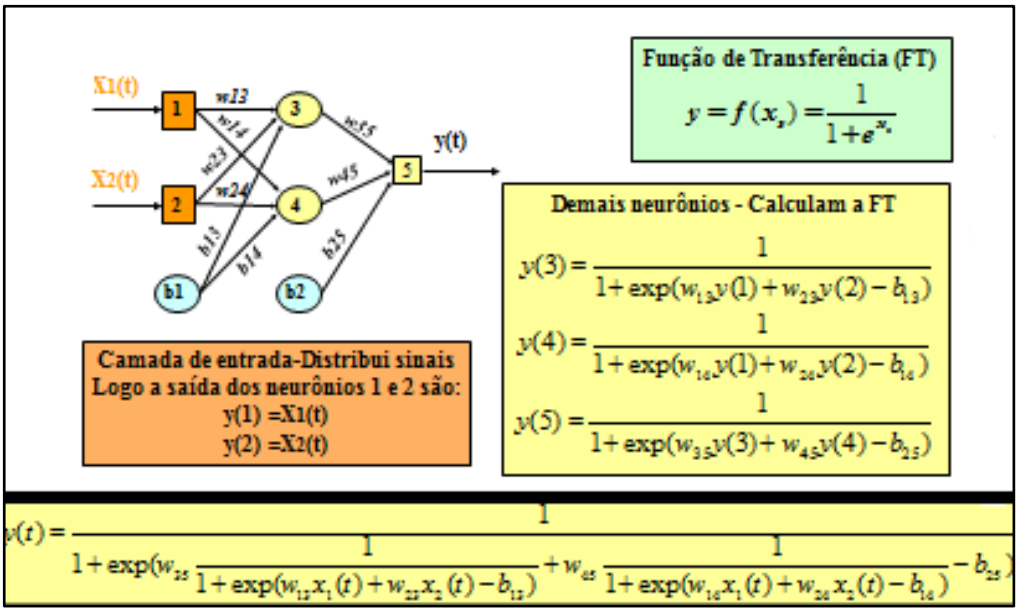


Fonte: Carvalho *et al.* (2009).

As redes *Perceptron* Multicamadas (MLP) são do tipo *feedforward*, ou seja, o processamento da informação se dá no sentido progressivo através das ligações sinápticas entre os neurônios das camadas adjacentes. Estas redes ficaram historicamente populares e tiveram uma ampla utilização a partir do surgimento do algoritmo de aprendizagem conhecido como *Backpropagation* para redes de múltiplas camadas na década de 80.

Na Figura 5 estão representados os principais elementos de uma "típica" Rede Neural *Feedforward*, na qual *x* e *y* representam entrada e saída, respectivamente; *w* representa os pesos e *f(x)*, a função de ativação (logística).

Figura 5 - Topologia de uma rede neural do tipo MLP.



Fonte: Carvalho *et al.* (2007).

A expressão final na parte inferior da figura 5 representa a rede neural projetada que fica representada na forma de uma função de regressão cujos parâmetros ajustáveis (treinamento) são os pesos W_{ij} e threshold Th_{ij}ou Baias b_{ij}. A mais importante propriedade das redes MLP é a capacidade de aproximar qualquer função contínua arbitrária com uma única camada escondida e função de ativação logística.

As redes MLP apresentam no mínimo 3 camadas, com pelo menos um neurônio cada:

- Camada de entrada – na qual os neurônios representam as variáveis de entrada (variáveis independentes) que as distribuem para a(s) camada(s) escondida(s);
- Camada(s) escondida(s) – nas quais os neurônios realizam o processamento, através de regras de propagação e funções de ativação;
- Camada de saída – onde os neurônios representam as variáveis de saída (respostas da rede).

A definição do número de camadas e de neurônios em cada camada intermediária assim como as funções de ativação para cada camada é definido a partir de um método de tentativa e erros.

Em Braga *et al*. (2007) as principais características da rede MLP são:

- Pode ter uma ou mais camadas intermediárias;
- Neurônios das camadas intermediárias e de saída têm funções semelhantes;
- Entrada da função de ativação é o produto interno dos vetores de entrada e de pesos;

- Separa padrões de entrada com hiperplanos;
- Melhor em problemas complexos; e,
- Constrói aproximadores globais para mapeamento entrada-saída.

2.3.2.2 Treinamento da rede

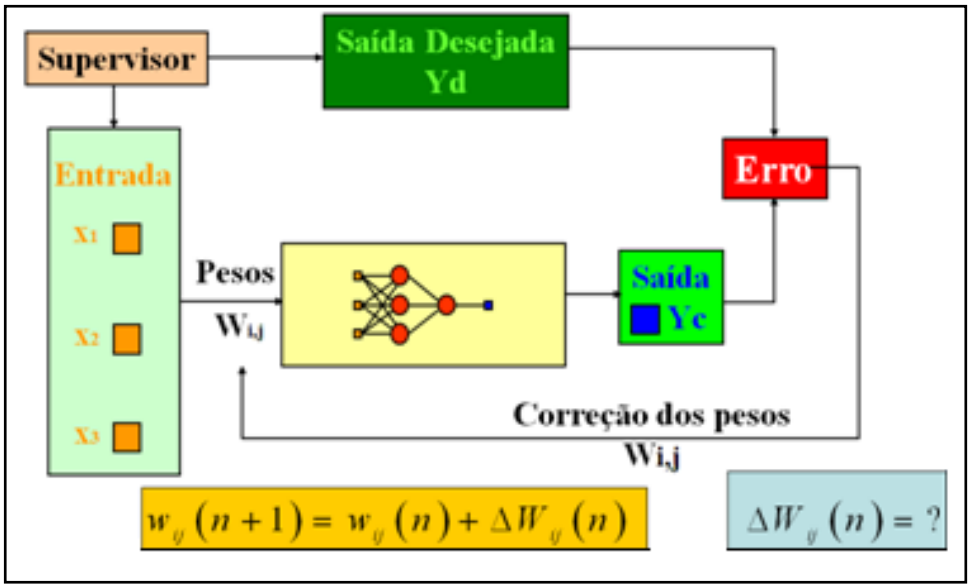
O aprendizado (ou treinamento) de uma rede neural MLP é o processo pelo qual os parâmetros livres da rede neural, ou seja, pesos W_{ij} e *threshold* Th_{ij}ou Baias b_{ij},são ajustados por meio de uma forma continuada de estímulo pelo ambiente externo, sendo o tipo específico de aprendizado definido pela maneira particular como ocorrem os ajustes dos parâmetros livres (Braga, 2007).

Aprendizado supervisionado é utilizado para treinamento da rede neural quando esta é utilizada como ajustadora de funções multivariáveis e não lineares, caso dos inferenciadores ou analisadores virtuais. Neste tipo de aprendizado, necessariamente, pressupõe-se a existência de um supervisor, ou professor externo (que são o conhecimento dos dados de saída), o qual é responsável por estimular as entradas da rede por meio de padrões de entrada e observar a saída calculada pela mesma, comparando-a com a saída desejada. Como a resposta da rede é função dos valores atuais do seu conjunto de pesos, estes são ajustados através de um processo que aproximar a saída da rede da saída desejada.

Para realização deste processo que, na realidade, matematicamente se trata de uma otimização, existem alguns algoritmos já estabelecidos que apresentam características peculiares nas suas utilizações. Dentre os algoritmos citados na literatura, pode se destacar, pela frequência de utilização, dois algoritmos baseados em métodos de otimização diferentes: o de retropropagação do erro (*backpropagation*), com a regra do delta generalizado, que é um método de gradiente descendente bastante utilizado e citado frequentemente na literatura; e o método de Levenberg-Marquardt, baseado no método quasi-Newton. Por se tratarem de algoritmos que envolvem o cálculo de derivadas, supõem-se a utilização de funções de ativação contínuas, principalmente as sigmóides.

A Figura 6 ilustra a fase de treinamento de uma RNA, em que *x* é o vetor de entrada de dados; *yd* e *yc* são os vetores de saída desejado e calculado, respectivamente; *w* é o vetor de pesos e *n* é o número de iterações. Por simplificação não se colocou os *threshold* Th_{ij}ou Baias b_{ij} na figura, porem seguem a mesma lógica.

Figura 6 - Fase de treinamento de uma RNA MLP.



Fonte: Carvalho *et al*. (2009)

A maior parte dos algoritmos de treinamento é baseada nos métodos de gradientes descendentes e de Newton. As abordagens baseadas nos métodos de Newton apresentam, em geral, melhores resultados pelo fato de serem métodos de segunda ordem, apresentando uma convergência quadrática próxima ao mínimo. No entanto, estes métodos são limitados pelo grande espaço de memória requerido e pelo volume de cálculos matriciais envolvidos, o que os torna praticamente inviáveis para redes de grande dimensão (GARCIA *et al*. (2004).

Diversos outros métodos, denominados quasi-Newton, têm sido propostos com o intuito de reduzir a memória requerida e o volume de cálculos processado. Segundo Karul *et al*. (2000), estes métodos se baseiam em simplificações da matriz de Hessian que reduzem o volume e simplificam o cálculo matricial.

Um método simplificado para o treinamento de RNA é o algoritmo de Lavenberg-Marquart (Equação 3), desenvolvido para se obter uma rápida velocidade de treinamento.

$$\Delta x = \left[J^T(x)J(x) + \mu I\right]^{-1} J^T(x)e(x) \qquad (3)$$

Sendo: *I* é a matriz identidade, *e(x)* é o erro e *J* é a matriz Jacobiana a qual contém as primeiras derivadas dos erros obtidos pela rede em relação aos pesos e bias.

O valor μ é sucessivamente diminuído após cada iteração e aumenta quando determinado peso aumenta a performance da função. Desta maneira, o desempenho da função

será reduzido a cada iteração. A aplicação do algoritmo de Levenberg-Marquart para o treinamento de redes neurais é descrito em detalhes por Garcia *et al*. (2004).

De acordo com Garcia *et al*. (2004), uma rede neural *Feedforward Backpropagation* com um número suficiente de neurônios na camada interna pode aproximar qualquer função. Porém, devemos estar cientes que a RNA pode, de fato, memorizar somente os dados disponíveis ao contrário de generalizá-los, fato este chamado de sobre ajuste de dados (*overfitting*). Um típico modelo de RNA sobre ajustada imita o conjunto de dados de treinamento satisfatoriamente. Porém, tem uma má estimativa para os dados que não foram incluídos no treinamento. Para uma boa generalização, o sobre-ajuste deve ser evitado tomando-se precauções apropriadas. Um sobre-ajuste pode ser evitado por um dos dois métodos descritos a seguir:

- Regularização: envolve modificações na função de desempenho por considerar somente um número mínimo de neurônios na camada intermediária, suficiente para treinar o sistema.
- Parada prematura: envolve a parada do treinamento quando o erro para o conjunto de dados de validação/teste começa a aumentar e uma nova inicialização é aconselhável se partindo de diferentes estimativas de pesos W_{ij} e *threshold* Th_{ij}ou Baias b_{ij}.

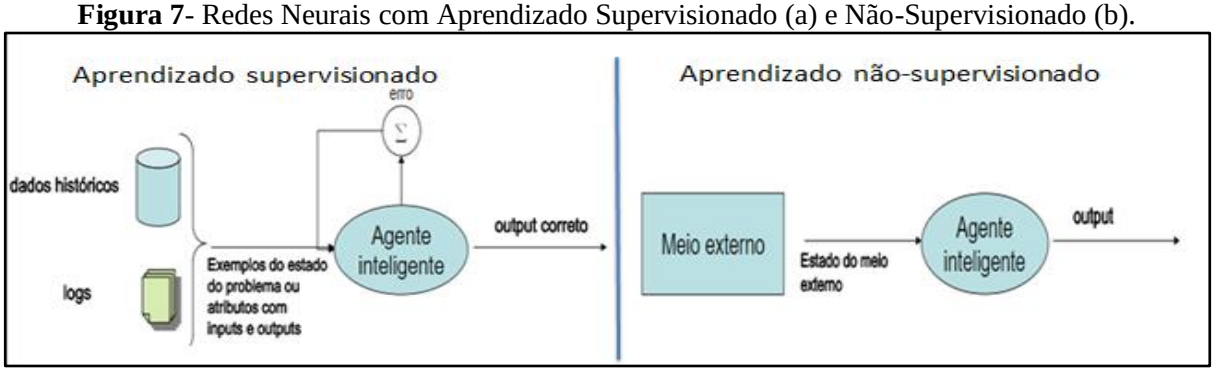
Para decidir quando parar o treinamento e consequentemente avaliar o resultado, o conjunto de dados é dividido aleatoriamente em três subconjuntos. Um subconjunto é usado para o treinamento, um para a validação e outro para o teste.

Para desempenho da parada prematura, os erros (diferenças entre valores medidos e calculados) o conjunto total de dados separados, quando da utilização do *software* MatLab, em treinamento, validação e teste. O erro do conjunto de dados de validação, normalmente, diminui após o início do treinamento, entretanto, quando a rede começa o sobre-ajuste, o erro do conjunto de validação começa a aumentar. Se este aumento continuar, o treinamento é finalizado num número pré-definido de iterações. O conjunto de dados de teste é comparado com o conjunto de validação para verificar se ambos exibem um comportamento similar. Se os erros de validação e os de teste não mostram um comportamento similar isto deve indicar uma má divisão de dados.

2.3.3 Mapas Auto-organizáveis de KOHONEN (SOM)

Inspirados no córtex cerebral humano, os mapas auto-organizáveis consistem em uma rede neural que gera como saída representações bidimensionais (mapas) de banco de dados de alta dimensionalidade. Estes foram desenvolvidos por Teuvo Kohonen (KOHONEN, 1995; KOHONEN, 2001; KOHONEN, 2013), e podem analisar dados por agrupamentos com o intuito de descobrir estruturas e padrões multidimensionais. Além disso, pode ser considerada uma rede neural com aprendizado não supervisionado e competitivo (Figura 07), pois não necessita de um vetor de saída conhecido como vetor alvo (o supervisor) (MESQUITA, 2002).

Figura 7- Redes Neurais com Aprendizado Supervisionado (a) e Não-Supervisionado (b).



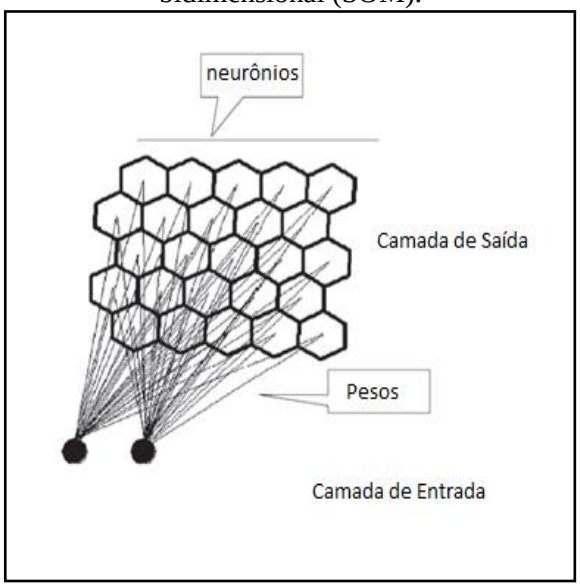
Fonte: http://slideplayer.com.br/slide/45785/

No começo da década de 1980, estes mapas foram consolidados como redes neurais por Kohonen (KOHONEN, 1995). Os mapas auto-organizáveis podem ser definidos como sendo redes neurais competitivas com um alto grau de interconexão entre seus neurônios e que são aptas a formar mapeamentos preservando a topologia entre os espaços de entrada e saída. Podem ser aplicados para problemas não lineares de alta dimensionalidade, tais como: extração de características e classificação de imagens e padrões acústicos, controle adaptativo de robôs, equalização, demodulação e transmissão de sinais assim como em aplicações nas áreas de estatística, processamento de sinais, química e medicina (AFFONSO, 2011).

Apesar da bastante ampla literatura existente sobre os métodos RNA, em especial *Multilayer Perceptron* (MLP), ou Rede de Multicamadas, (MAIER e DANDY, 2000; ASCE, 2000a, b; DAWSON e WILBY, 2001; WU e LO, 2008; SINGH *et al.* 2009; VILAS *et al.*, 2011), é importante ressaltar que há uma carência de ampla revisão da literatura sobre a eficiência de técnicas de aprendizagem não supervisionada.

Em geral, o algoritmo SOM agrupa as amostras ou padrões pré-definidos em classes e também ordena as classes em mapas significativos (topologia de preservação ou propriedade de ordenação). A sua estrutura típica consiste de duas camadas: uma camada de entrada e uma camada de saída ou de Kohonen (Figura 8). A camada de entrada contém um neurônio para cada variável (por exemplo, DBo, DQo, etc.) no conjunto de dados. As camadas de neurônios de Kohonen são conectadas a cada neurônio na camada de entrada através de pesos ajustáveis ou parâmetros de rede. Os vetores de peso na camada de Kohonen dão uma representação da distribuição dos vetores de entrada de uma forma ordenada (NEPOMUCENO, 2006; KALTEH; HJORTH; BERNDTSSON, 2008).

Figura 8 - Exemplificação de uma estrutura de uma 5 x 5 mapa de auto-organização bidimensional (SOM).



Fonte: Adaptada de Kalteh; Hjorth; Berndtsson (2008).

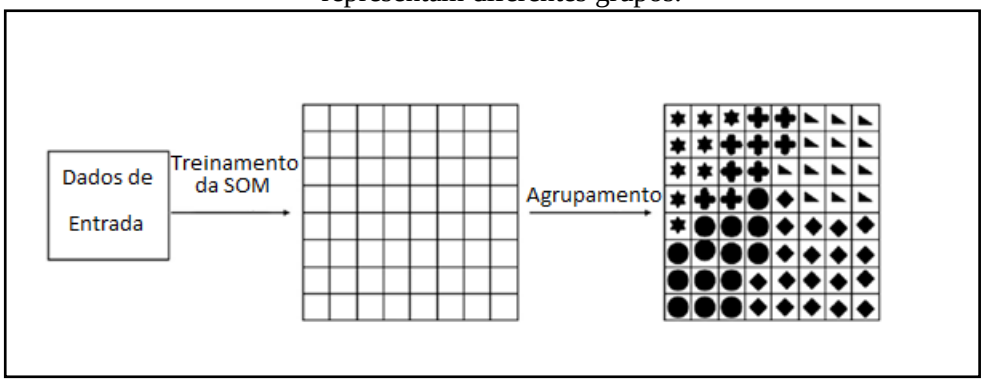
Segundo Kalteh, Hjorth, e Berndtsson (2008), para aplicar sucessivas SOM são necessárias três importantes etapas:

1) A coleta de dados e normalização: a parte mais importante na normalização é para evitar que variáveis tenham maior impacto em relação a outras variáveis. Consequentemente, a normalização, por transformação de todas as variáveis na gama de, por exemplo, 0 e 1, garante que todas as variáveis têm igual importância na formação da SOM.
2) Formação: após a preparação de dados e normalização, um vetor de entrada a partir da matriz de dados é apresentado ao procedimento de treinamento interativo para formar a SOM. Recomenda-se que o número de iterações deve ser, pelo menos, 500 vezes o

número de neurônios na camada de saída (HAYKIN, 1999; KOHONEN, 2001). No início do treinamento, vetores de pesos devem ser inicializados usando qualquer um ao acaso ou um método de inicialização linear. Os vetores de peso são também chamados de referência ou vetores da tabela de codificação. Inicialização aleatória de vetores de peso é o método mais comumente usado. Independentemente do tipo de método de inicialização que é aplicado, o SOM utiliza um tipo de aprendizagem que é chamado de aprendizado competitivo, sem supervisão, ou de auto-organização para combinar com cada vetor de entrada com um neurônio na SOM. Com base no aprendizado competitivo, os neurônios de saída desta rede competem entre si para serem ativados com o resultado de que apenas um neurônio de saída (ou um neurônio por grupo) será ativado em cada iteração. Um neurônio de saída que vence tal competição é chamado neurônio vencedor (*winner-takes-all neuron*). Uma maneira de induzir tal tipo de competição entre os neurônios de saída é usar conexões inibitórias laterais entre eles (ou seja, caminhos de realimentação negativa), ideia originalmente proposta por Rosenblatt em 1958.

3) Colhendo informações da SOM treinada: uma vez que o treinamento da SOM foi realizado o mapa resultante pode ser pós-processado com base em visualização, armazenamento em cluster ou fins de modelagem locais. Esta questão foi exaustivamente investigada por Vesanto (2002). Como mencionado acima, o SOM é um algoritmo de agrupamento supervisionado. Um exemplo de agrupamento SOM é mostrado na Figura 9.

Figura 9 - O diagrama de um grupo de dois níveis do SOM. Diferentes símbolos representam diferentes grupos.



Fonte: Adaptado de WU e CHOW (2004).

2.3.3.1 Treinamento da rede SOM

Os neurônios em uma rede SOM são posteriormente ordenados e apresentados em gráficos gradeados (treliça ou *lattice*), normalmente mono ou bi-dimensionais. Mapas de dimensões maiores são também possíveis, porem mais raros. Os neurônios se tornam

seletivamente "ajustados" a vários estímulos (padrões de entrada) ou classes de padrões de entrada ao longo de um processo competitivo de aprendizado. A localização destes neurônios (que são os neurônios vencedores) se torna ordenada entre si de tal forma que um sistema de coordenadas significativo é criado na treliça, para diferentes características de entrada (AFFONSO, 2011).

O SOM é, portanto, caracterizado pela formação de um mapa topográfico dos padrões de entrada, no qual as localizações espaciais (ou coordenadas) dos neurônios na treliça são indicativas de características estatísticas (implícitas) contidas nos padrões de entrada (AFFONSO, 2011).

O funcionamento de um SOM pode ser compreendido em etapas distintas, a etapa competitiva na qual se define o neurônio mais adequado (*Best Matching Unit*), etapa cooperativa e a etapa adaptativa. A escolha da melhor correspondência entre o vetor de entrada e o vetor peso é feita por meio do critério da menor distância (euclidiana) entre o vetor de pesos por ela armazenado e o vetor de entrada, matematicamente expresso pela Equação 4.

$$i(x) = \arg\min ||x - w_j ||, j= 1, 2, ...n \quad (4)$$

Na qual: i(x) é a representação do neurônio da entrada x, e w_j é o vetor peso.

Entre as funções de distâncias utilizadas para quantificar a "semelhança entre os vetores da rede" e, portanto o quanto eles se aproximam do vetor de dados apresentado, uma das mais empregadas é a Distância Euclideana (D_E), definida pela Equação 5.

$$D_E = \sqrt{(x1 - y1)^2 + (x2 - y2)^2 + \cdots + (xn - yn)^2} \quad (5)$$

Em que: Xn são as coordenadas dos vetores de entrada e yn são as coordenadas dos vetores-protótipo (pesos das redes auto-organizáveis).

Outros tipos de distâncias que podem ser citadas é a similaridade métrica de Minkowski, e distância de Manhattan respectivamente, representadas pelas Equações 6 e 7.

$$D_{\text{Minkowski}} = \sqrt[p]{\sum_{k=0}^{n} |x_k - y_k|^p} \quad (6)$$

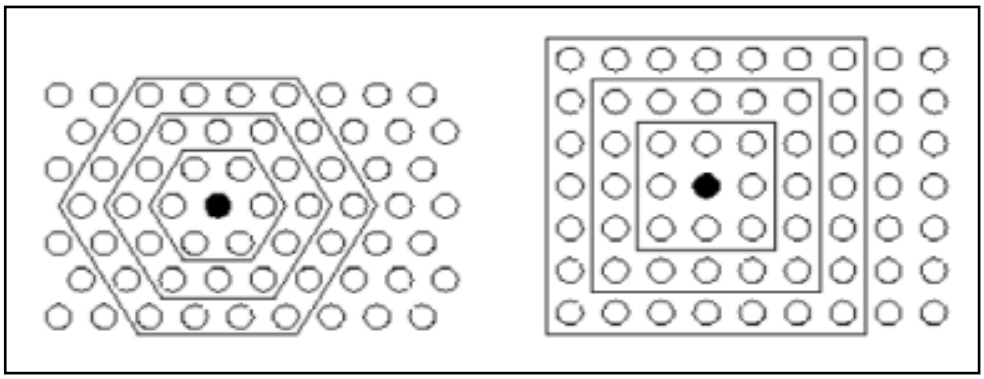
Distância métrica de Minkowski, citada como uma generalização da métrica euclidiana em aplicações na área de psicologia.

$$D_{\text{Manhattan}} = \sum |X - Y| \quad (7)$$

Distância de Manhattan.

Na etapa cooperativa, são definidos os vizinhos dentro de uma distância obtida a partir da BMU (*Best Matching Unit*) obtida na primeira etapa, competitiva. Sumariamente o processo de treinamento da rede, consiste na otimização da distância entre os neurônios. Na minimização das distâncias é definida a vizinhança topológica por meio da interatividade entre os neurônios (um neurônio ativado tende a excitar os neurônios em sua vizinhança imediata). Cada atribuição de novos valores e distâncias abrangendo toda a rede é chamada de época. Pela repetição da adaptação de pesos (vetores-protótipo) é possível determinar o melhor número de épocas de treinamento para cada matriz, o que constitui a etapa adaptativa. Os neurônios nessa vizinhança são atualizados a cada iteração. Na Figura 10, são ilustradas a formação de vizinhança a partir do neurônio vencedor em topologia hexagonal e retangular.

Figura 10 - Representações das Etapas Competitiva e Cooperativa de treinamento da SOM.



Fonte: VESANTO, 2002.

2.3.3.2 Algoritmo de aprendizagem e arquitetura de rede SOM

A rede neural é composta por uma camada de entrada e uma camada de neurônios. Os neurônios ou unidades são organizados em uma grade retangular ou hexagonal e são totalmente interligados. Cada um dos vetores de entrada está também ligado a cada uma das unidades. O algoritmo de aprendizagem aplicado à rede pode ser dividido em seis etapas (KOHONEN, 2001; KOHONEN, 2013; Li *et al*., 2018):

1) Uma matriz *m* x *n* é criada do conjunto de dados com *m* linhas de amostras e *n* colunas de variáveis. A matriz consiste de m vetores de entrada de comprimento n. A classificação dos vetores de entrada é baseado na medida de similaridade, por exemplo, Distância Euclidiana. A fim de evitar distorções na classificação devido a diferenças em unidade ou faixa de medição das variáveis, uma normalização é feita. Isto pode ser feito através da fixação média igual a zero e variância igual a 1 ou por redimensionando a gama de cada variável no intervalo de [0,1].

2) Cada unidade é atribuído aleatoriamente um peso inicial de referência do vetor ou com um comprimento igual ao comprimento dos vectores de entrada (n)
3) Um vetor de entrada é mostrado para a rede e a distância Euclidiana entre o vetor de entrada X e todos os vetores de referencia M_i são calculados de acordo com:

$$\left||X - M|\right| = \sqrt{\sum_{i=1}^{n}(xi - mi)^2} \quad (8)$$

4) A melhor unidade de melhor harmonização Mc é escolhido de acordo com:

$$\left||X - Mc|\right| = min\{||X - Mi||\} \quad (9)$$

5) Os pesos da unidade de melhor harmonização e a unidade na sua vizinhança N (t) são adaptados de modo a que os novos vetores de referência encontram-se mais próximo do vetor de entrada (Figura 11). O fator α(t) controla a taxa de variação dos vetores de referência e é denominada de taxa de aprendizagem.

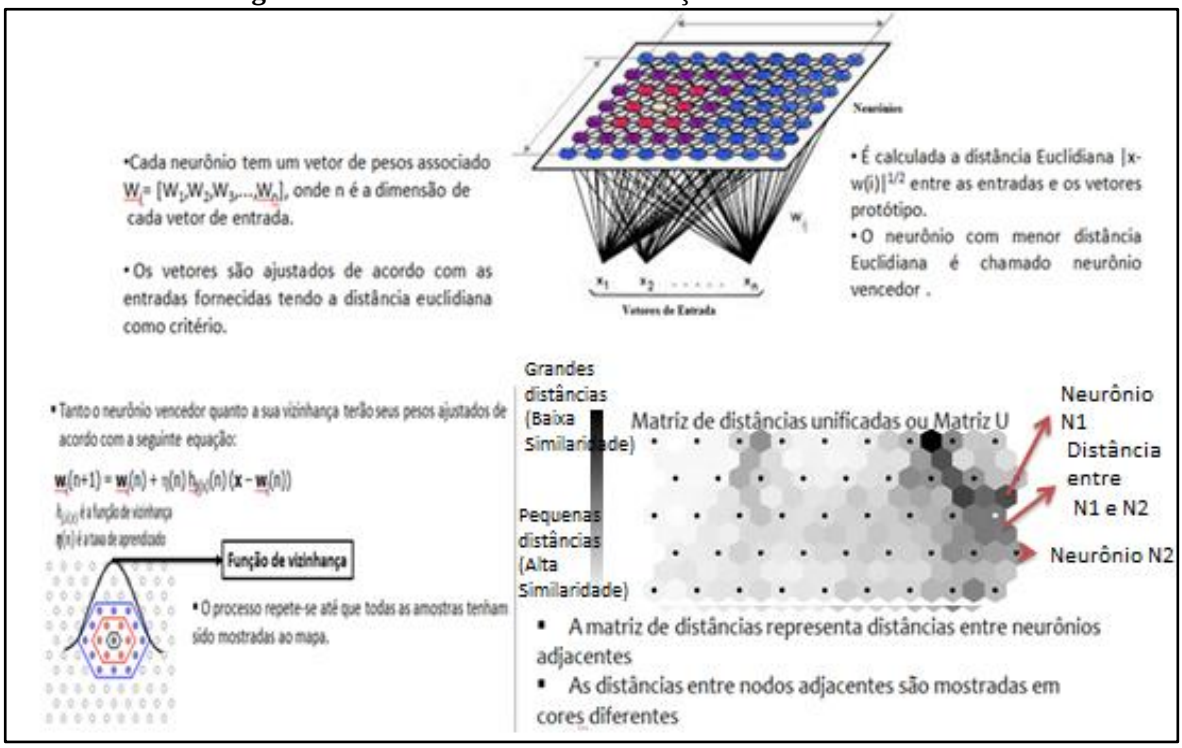
$$Mi(t+1) = \begin{cases} Mi(t) + \alpha(t)[X(t) - Mi(t)] & \nabla \in N(t) \\ Mi(t) & \nabla \in N(t) \end{cases} \quad (10)$$

A taxa de adaptação da unidade é controlada pela função de vizinhança h, o que diminui a partir de um na unidade vencedora a zero em unidades localizadas mais longe do raio r. As funções mais comuns são usados em forma de sino (Gauss) ou quadrado (bolha).

6) O terceiro e quinto passos são repetidos até que um número máximo predefinido de iterações é alcançado. Durante estas iterações ambos α e N (t) diminuem , forçando a rede a convergir .

Portanto, o algoritmo mapa de auto-organização é uma técnica de rede neural não supervisionada que classifica dados de acordo com a sua semelhança, plotando os dados de alta dimensão em uma grade bidimensional em uma topologia de forma de preservação, de modo que a estrutura do conjunto de dados é processada na visualização (KOHONEN, 1995; KOHONEN, 2001; KOHONEN, 2013; Li *et al.*, 2018). Na Figura 11 é resumido este procedimento.

Figura 11- Procedimento de classificação dos dados na SOM.



A não-linearidade e a complexidade das variáveis envolvidas na qualidade da água levaram muitos investigadores a utilizar modelos de RNAs para simular estas variáveis devido à capacidade de tais modelos de lidar com relações complexas, não-lineares.

A questão da divisão dos dados em treinamento, testes e validação em subconjuntos de desenvolvimento de modelos de RNA foi abordada por Bowden *et al*. (2002). Este problema é um dos mais importantes no desenvolvimento de modelos de RNA. Ele tem de ser levado a cabo de uma maneira eficiente, devido ao fato de a divisão dos dados em subconjuntos pode ter uma influência significativa sobre o desempenho do modelo de RNA obtido. Como modelos RNA aprendem com exemplos de dados de treinamento, o modelo terá pouca capacidade de generalização, se os dados de treinamento não são representativos do domínio de modelagem. Em seguida, o modelo estará enfrentando exemplos que estão além do conjunto de treinamento. E também, a inclusão de muitos exemplos repetitivos em um conjunto de treinamento só irá retardar o processo de formação. Por conseguinte, existe uma necessidade de um procedimento de princípio para dividir os dados em subconjuntos, em vez de se dividir os dados arbitrariamente em subconjuntos sem levar em conta as propriedades estatísticas, o que é o processo mais comum em aplicações correntes.

O uso de redes neurais artificiais (RNAs) em problemas relacionados com a qualidade de água tem recebido um crescente interesse na última década. O método relacionado do

mapa de auto-organização (SOM) é um método de aprendizagem não supervisionado para analisar, agrupar e modelar vários tipos de grandes bases de dados. No entanto, há ainda uma notável falta de revisão bibliográfica abrangente para a SOM, juntamente com os procedimentos de treinamento e manipulação de dados e potencial aplicabilidade. Consequentemente, o presente trabalho tem por objetivo, em primeiro lugar, explicar o algoritmo e, em segundo lugar, analisar as aplicações publicadas com ênfase nos problemas de qualidade de água, a fim de avaliar a forma como a SOM pode ser utilizada para resolver este problema específico. Conclui-se que a SOM é uma técnica promissora adequada para investigar, modelar e controlar muitos tipos de processos e sistemas e fenomenos.

Métodos de aprendizagem não supervisionados ainda não foram totalmente testados de forma abrangente dentro da engenharia. No entanto, ao longo dos anos, a SOM tem apresentado um aumento constante no número de aplicações em recursos hídricos devido à robustez do método.

2.3.4 Sistema *Neuro-Fuzzy*

De acordo com Oliveira Júnior *et al.* (2007), os sistemas *neuro-fuzzy* consistem na representação do sistema *fuzzy* na forma de redes passíveis de treinamento, por técnicas semelhantes às usadas em redes neurais. O processo de treinamento na verdade é o ajuste de parâmetros, com o objetivo de minimizar a função erro entre as saídas desejadas e as apresentadas pela rede. Os sistemas *neuro-fuzzy* têm como objetivo, então, conjugar a capacidade de aprendizagem das redes neurais à interpretação característica dos sistemas fuzzy.

Define-se inferência como a passagem, através de regras válidas, do antecedente (SE) ao consequente (ENTÃO) de um objeto de estudo. Na lógica *fuzzy*, essa passagem é realizada mediante a iteração, determinada pelas regras de inferência, entre as variáveis linguísticas de entrada (SE), gerando um conjunto de dados de saída (ENTÃO). Essas regras são aplicadas aos conjuntos fuzzy através das variáveis linguísticas e são construídas mediante a operação entre os conjuntos (OLIVEIRA JUNIOR *et al.*, 2007).

A lógica *fuzzy* foi desenvolvida a partir da teoria de conjuntos nebulosos, elaborada na década de 60 pelo professor Lofti A. Zadeh, para tratar do aspecto vago da informação (SANDRI; CORREA, 1999; CAMPOS FILHO, 2004). Esta teoria é uma extensão da lógica booleana, que incorpora valores parciais de verdade. Em vez das proposições serem

"completamente verdadeiras" ou "completamente falsas", elas recebem um valor que representa seu grau de verdade. Em sistemas fuzzy, os valores são indicados por um número (chamado de valor de verdade) no intervalo de 0 a 1, onde 0,0 representa falso absoluto e 1,0 representa a verdade absoluta, o que possibilita a representação de conceitos imprecisos, sem perder a precisão matemática no tratamento (SANDRI; CORREA, 1999).

A lógica difusa é capaz de lidar com informações aproximadas de forma sistemática e, portanto, é adequada para mapear sistemas não lineares e é usada para modelar sistemas complexos, onde existe um modelo inexato ou sistemas em que a ambiguidade ou a imprecisão são comuns. Os sistemas *fuzzy* são baseados em regras em que um conjunto de regras difusas representa um mecanismo de decisão para ajustar os efeitos de certos estímulos do sistema. Com uma base de regras eficazes, os sistemas de *fuzzy* podem substituir uma decisão de um humano qualificado. A base das regras reflete o conhecimento especializado humano, expresso como variáveis linguísticas, enquanto as funções de associação representam a interpretação especializada dessas variáveis.

A lógica *fuzzy* é baseada na teoria dos conjuntos *fuzzy* e esta é em grande parte uma extensão da teoria dos conjuntos tradicionais. Para exemplificar os dois tipos de conjuntos, tradicional e *fuzzy*, vejamos como exemplo a caracterização da idade de um indivíduo associada a uma determinada faixa etária, por exemplo, indivíduo adolescente. Sendo indivíduo adolescente parte de um grupo onde a faixa etária está no intervalo de números reais entre 12 e 18, conforme a lógica tradicional podemos denotar:

A = {conjunto de indivíduo adolescente}

Como o adolescente é um indivíduo entre 12 e 18 anos de idade, define-se A no intervalo:

A = {12,18}

Conforme a definição formal de um conjunto tradicional tem-se:

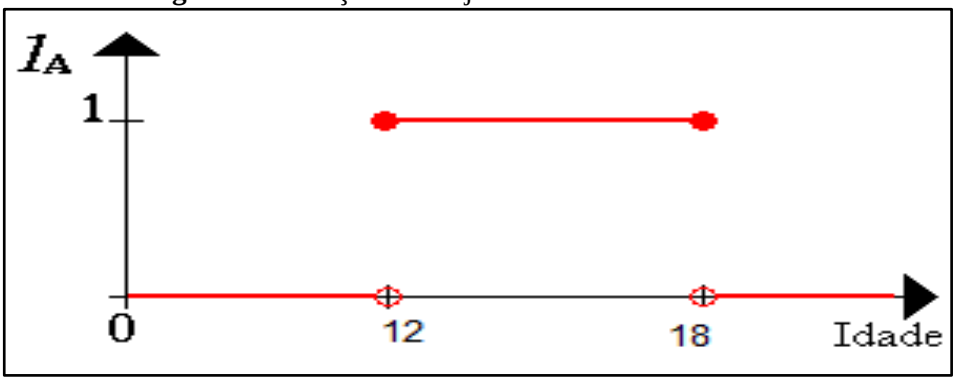
$$f_A\,(x): X \rightarrow 0{,}1 \qquad (11)$$

onde,

$$f_A(x) = \begin{cases} 1, se\ x \in A \\ 0, se\ x \notin A \end{cases}$$

O conjunto A representado pela lógica tradicional tem a sua função característica, onde 1 significa que o individuo pertence ao conjunto de adolescentes e 0 que o individuo não pertence ao conjunto de adolescentes, como mostra a Figura 12.

Figura 12 - Função do conjunto tradicional adolescente.



Fonte: Modificado de SANDRI; CORREA (1999).

Sendo assim, a ideia de pertinência faz parte da caracterização do conjunto *fuzzy* que representa uma generalização da ideia apresentada pelos conjuntos da lógica tradicional. O conjunto *fuzzy* é completamente caracterizado pela definição de um intervalo de pertinência que varia de zero (0) a um (1), sendo 1 pertencendo 100%, 0 pertencendo 0% e o intervalo entre 0 e 1 é pode assumir valores infinitos adotando graus de pertinência individuais. Assim, os valores desse intervalo podem ser considerados como medidas que expressam a possibilidade de um dado elemento ser membro de um conjunto *fuzzy*. Desta forma, uma sentença pode ser "parcialmente verdadeira e parcialmente falsa".

A representação do grau de pertinência é definida por meio de uma função característica generalizada chamada de função de pertinência:

$$\mu\ (x): \qquad [0,1] \qquad (12)$$

Sendo: (x): $[0,1]$

(x): $[0,1]$

$\mu\ (x) < 1\ se\ X\ \rightarrow [0,1]$ ind

Em que *X* é o universo e *A* é o subconjunto *fuzzy* de *X*. Essa função associa a cada elemento *x* de *X* o grau (x), com o qual *x* pertence a *A*.

A representação acima indica o grau com que um elemento *x* pertence ao subconjunto *A*, grau este que pode assumir infinitos valores no intervalo [0,1]. A representação formal como um conjunto é:

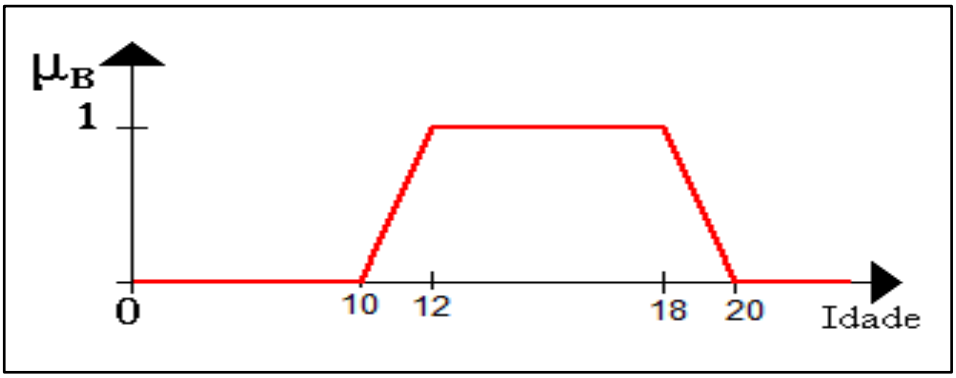
$$\{\ (\mu\ (x)\} \qquad (x): \quad [0,1] \qquad (13)$$

Um conjunto *fuzzy* definido no universo de discurso *A* é caracterizado por uma função de pertinência μ_A, a qual mapeia os elementos de *X* para o intervalo [0,1]. Assim, a função de

pertinência associa a cada elemento *x* pertencente a *X* um número real (x)no intervalo [0,1], que representa o grau de pertinência do elemento *x* ao conjunto *A.*

Para exemplificar um conjunto de individuo adolescente conforme a lógica *fuzzy* utiliza-se a representação dada por sua função característica (na figura 2.3 é representada uma função trapezoidal) onde 1 significa que o individuo pertence ao conjunto de individuo adolescente e 0 que o individuo não pertence ao conjunto de individuo adolescente e no intervalo entre 0 e 1 o conjunto pode assumir valores infinitos adotando graus de pertinência individuais para o conjunto de individuo adolescente como mostra a Figura 13.

Figura 13 - Função trapezoidal do conjunto *fuzzy* adolescente.



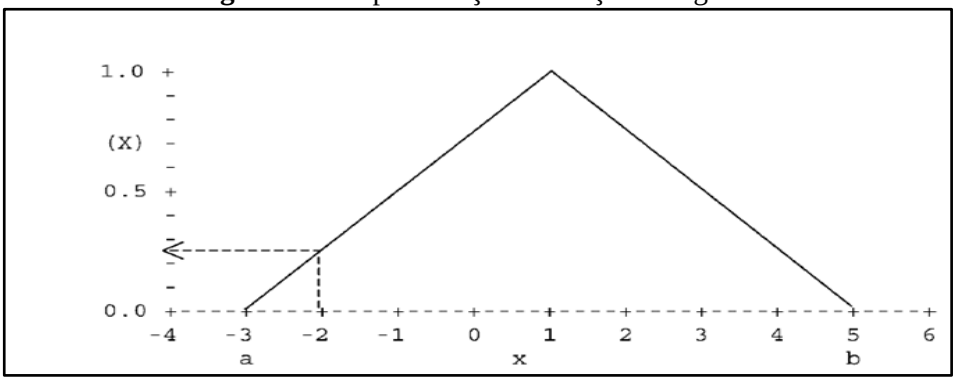
Fonte: Modificado de SANDRI; CORREA (1999).

Considerando o exemplo da Figura 13, quem tem de 12 a 18, teria 100% de pertinência ao conjunto de indivíduo adolescente e quem entre 10 e 12 ou entre 18 e 20 pertence parcialmente ao conjunto de indivíduo adolescente. Sendo assim, existe uma função de pertinência quando se trata de um conjunto *fuzzy*.

2.3.4.1 As funções de pertinência

A função de pertinência que reflete o conhecimento que se tem em relação à intensidade com que o objeto pertence ao conjunto *fuzzy*. Existem várias formas de representar uma função *fuzzy* de pertinência, sendo que, as mais usuais são a triangular, gaussiana, trapezoidal, sigmóide bipolar, S e quadrática, sendo todas definidas no intervalo de pertinência de 0 a 1. A figura 14 mostra a função a triangular e sua representação matemática.

Figura 14 - Representação da função triangular.



Fonte: SILER e BUCKLEY, 2005.

Na função triangular os números *fuzzy* começam a subir a partir de zero x = a; atingir um máximo de 1 em x = b; e declínio para zero em x = c. Em seguida, a função μ (x) de um número *fuzzy* triangular é representada na equação (14) ou (15) e na Figura 14.

$$trimf(x;a,b,c) = \begin{cases} 0, x \le a \\ (x-a)/(b-a), \ a < x \le b \\ (c-x)/(c-b), \ b < x \le c \\ 0, x > c \end{cases} \tag{14}$$

Ou

$$trimf(x;a,b,c) = max\left(\min\left(\frac{x-a}{b-a}, \frac{c-x}{c-b}\right), 0\right) \tag{15}$$

Uma representação matemática semelhante pode ser feita para as funções gaussiana, trapezoidal, sigmóide bipolar, S e quadrática. A função de pertinência irá refletir o grau de pertinência do elemento *x* para o conjunto *fuzzy* Sendo importante escolher qual função representa melhor o objeto de estudo, atualmente essa escolha pode ser baseada em trabalhos da literatura.

Em geral, um sistema *fuzzy* faz corresponder a cada entrada *fuzzy* uma saída *fuzzy*. No entanto, espera-se que a cada entrada *crisp* (um número real, ou par de números reais, ou n-dupla de números reais) faça corresponder uma saída *crisp*. Neste caso, um sistema *fuzzy* é uma função de R^n em R construída de alguma maneira específica. Os módulos que seguem indicam a metodologia para a construção desta função (Amendola, *et al*., 2005):

- **Módulo de *fuzzyficação*:** é o que modela matematicamente a informação das variáveis de entrada por meio de conjuntos *fuzzy*. É neste módulo que se mostra a grande importância do especialista do processo a ser analisado, pois a cada variável de entrada devem ser atribuídos termos linguísticos que representam os estados desta variável e, a

cada termo linguístico, deve ser associado um conjunto *fuzzy* por uma função de pertinência;

- **Módulo da base de regras:** é o que constitui o núcleo do sistema. É neste módulo onde "se guardam" as variáveis e suas classificações linguísticas;
- **Módulo de inferência:** é onde se definem quais são os conectivos lógicos usados para estabelecer a relação *fuzzy* que modela a base de regras. É deste módulo que depende o sucesso do sistema *fuzzy* já que ele fornecerá a saída (decisão) *fuzzy* a ser adotado a partir de cada entrada *fuzzy;*
- **Módulo de *defuzzyficação*:** que traduz o estado da variável de saída *fuzzy* para um valor numérico.

Vale salientar que no procedimento de inferência é analisado o grau de pertinência associado aquele mesmo valor numérico no universo de discurso relacionando-os a uma base de regras conforme a condicional se – então, como foi mostrado.

O tipo de inferência ocorre:

If(Se) < antecedente >*then(Então)* < conseqüente > (16)

ou

SE<situação>*ENTÃO*<ação> (17)

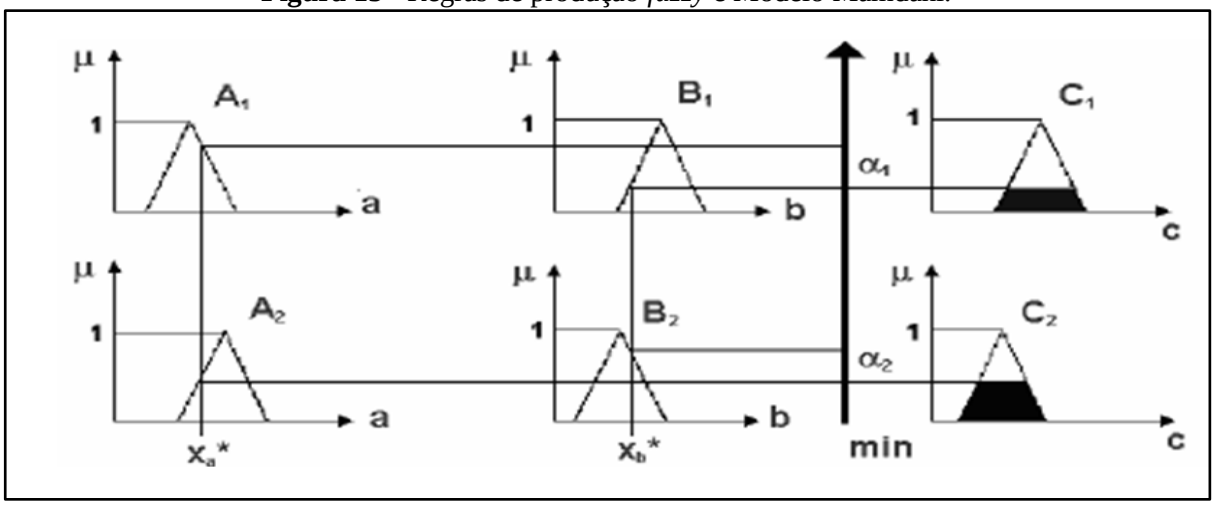
Na lógica clássica a inferência é dada pela comparação, sendo que só é permitida uma "compatibilidade exata", já no raciocínio difuso é possível adotar um valor aproximado dependendo da pertinência ou *fuzzyficação* dessa variável ao conjunto *fuzzy* (WESTPHAL, 2003; SIVANANDAM, SUMATHI e DEEPA, 2007).

A saída ou resposta do sistema são fornecidas diferentemente, a depender do modelo de inferência escolhido (CAMPOS FILHO, 2004). É importante ressaltar que existem na literatura diferentes métodos de inferência *fuzzy* com diferentes propriedades. O *Fuzzy Logic Toolbox* do MatLab oferece duas opções: o Método de Mamdani e o Método de Sugeno. A principal diferença entre esses dois método é que o Método de Sugeno (TSK) utiliza dados para a criação do sistema de inferência, enquanto que no Método de Mamdani, as características do sistema de inferência são definidas pelo usuário.

No modelo de inferência Mamdani a conclusão de cada regra especifica um termo difuso que necessita da utilização de um cálculo, denominado *defuzzyficação* para gerar uma resposta do modelo. Essa *defuzzyficação* pode ser realizada através de diferentes tipos de

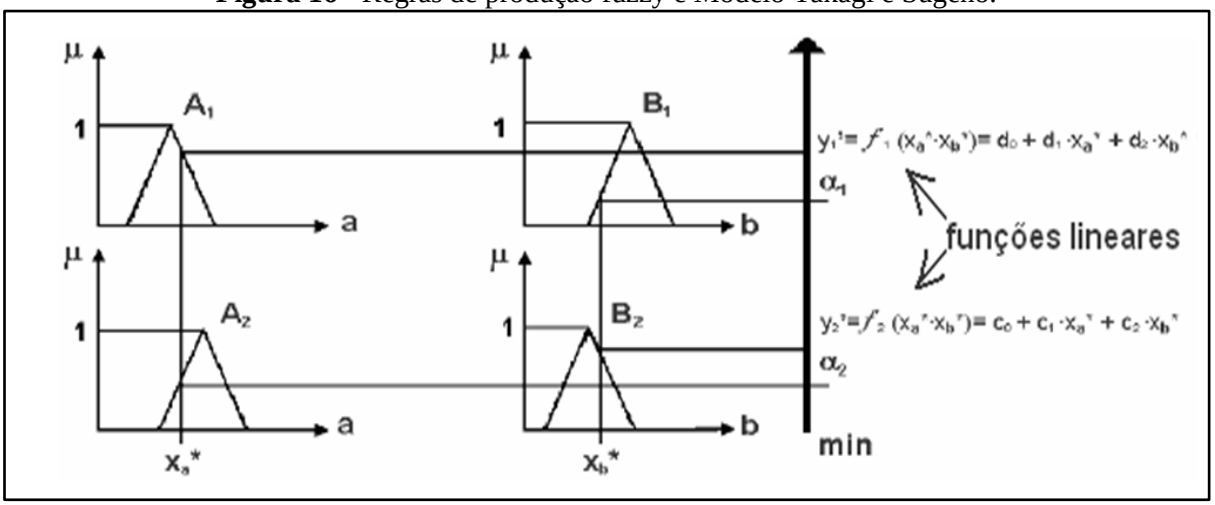
saída. No modelo de interpolação, como os do tipo (TSK), fornecem como saída um indicador, sempre no intervalo entre 0 e 1 que corresponde a média das regras do modelo.

Figura 15 - Regras de produção *fuzzy* e Modelo Mamdani.



Fonte: Modificado de MEDEIROS, 2006.

Figura 16 - Regras de produção fuzzy e Modelo Takagi e Sugeno.



Fonte: Modificado de MEDEIROS, 2006.

Podemos resumir os tipos de inferência nos sistemas fuzzy e enquadra-los em um dos seguintes (CALDEIRA *et al*., 2007):

- Tipo 1: A saída global é a média ponderada das saídas de cada regra por meio dos graus de ativação das respectivas regras. As funções de pertinência dos termos consequentes devem ser monótonas crescentes;
- Tipo 2: A saída é resultante da aplicação da interferência máx-mín seguida do passo de *defuzzyficação*;

- Tipo 3: Sistemas TSK, nos quais a parte consequente de cada regra é uma combinação linear de variáveis de entrada acrescida de um termo constantes. A saída é a média ponderada das saídas de cada regra.

Tal estrutura viabiliza a adaptação dos sistemas *fuzzy* sob treinamento a dado *training set*, de modo que, no final, obtenhamos resultados compreensíveis, ao contrário dos casos típicos das RNA, de difícil interpretação por humanos. Embora possamos utilizar esta técnica em diversos tipos de sistemas fuzzy, os casos práticos são predominantemente do tipo Takagi-*Sugeno-Kang* (TSK). Sua utilidade é evidenciada quando lidamos com sistemas fortemente não lineares, de comportamento variável no tempo etc. (CALDEIRA *et al*., 2007).

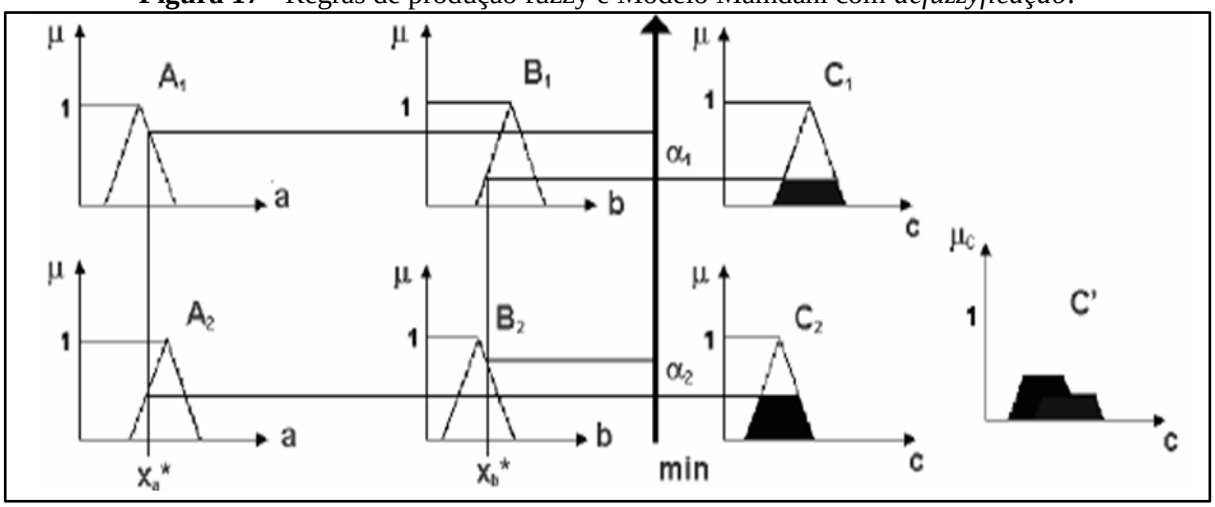
2.3.4.2 Defuzzyficação

No caso Mamdani, num sistema de inferência *fuzzy* a característica de saída é obtida a partir de valores *defuzzyficados* de produção de conjuntos *fuzzy* resultantes da agregação de diferentes resultantes de cada regra (fornecidas após a *fuzzyficação*) da base de regras de inferência distribuídas no universo de discurso.

Na Figura 17 é exemplifica agregação utilizando o conector mínimo do conjunto A1 e B1 é representada por C1, agregação do conjunto A2 e B2 é representada por C2 e a essa agregação total referente às duas regras é representada por C'. A agregação depende do tipo de conector (E, OU) escolhido para o relacionamento entre as variáveis.

Após a agregação, será calculada a forma de resposta do modelo, ou seja, o ponto relacionado ao grau de pertinência da agregação ou distribuição.

Figura 17 - Regras de produção fuzzy e Modelo Mamdani com *defuzzyficação*.

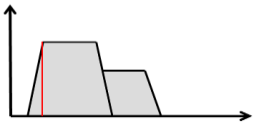
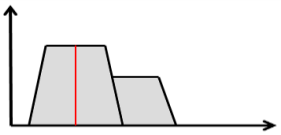
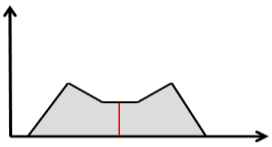


Fonte: MEDEIROS, 2006.

Os métodos de *defuzzyficação* mais utilizados são: o primeiro máximo, média dos máximos e centro da área ou centróide. Esses métodos têm sua forma e gráfico apresentados na Tabela 3.

- Método do primeiro máximo o valor de saída corresponde ao ponto em que o grau de pertinência da distribuição atinge o primeiro valor máximo;
- No método centróide, mais utilizado, o valor de saída corresponde ao centro da gravidade da função de distribuição;
- No método da média do máximos o valor de saída corresponde ao ponto médio entre os valores que tem maior grau de pertinência.

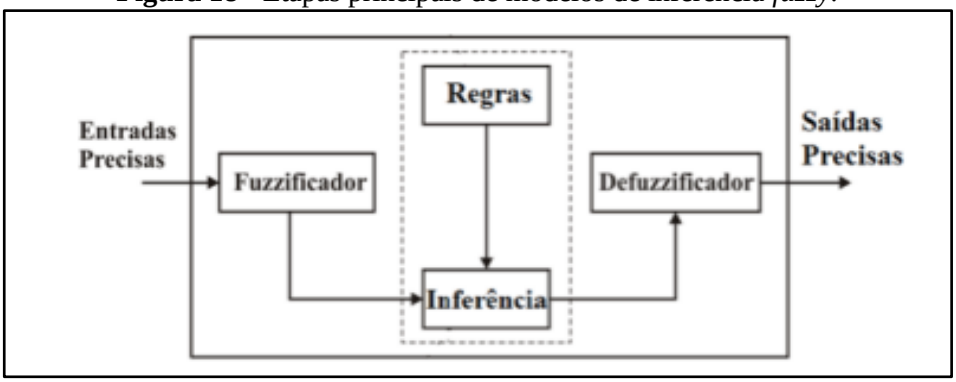
Tabela 3 - Comparativo dos métodos de *defuzzyficação*.

Método	**Fórmula**	**Gráfico**
Primeiro máximo		
Média dos máximos	$\frac{(x \qquad\qquad)}{}$	
Centróide	$\frac{\sum x}{\sum w}$	

Após a entrada das variáveis numéricas precisas, são ativadas as regras (*fuzzyficação*), em seguida o sistema de inferência determina como as regras (determinadas por especialistas) são combinadas, como resultado tem-se uma agregação entre as respostas das regras e após a escolha do tipo de resposta em relação a distribuição dos dados agregados (*defuzzyficação*) tem-se a resposta do modelo no domínio das variáveis de saída num correspondente universo de discurso. As entradas e saídas do sistema são denominadas respectivamente, *fuzzyficação* e

defuzzyficação e correspondem às etapas principais de modelos de inferência *fuzzy* (MALUTTA, 2004).

Figura 18 - Etapas principais de modelos de inferência *fuzzy*.



Conforme Simões e Shaw (2007), a combinação dessas duas tecnologias pode conter uma resposta completa aos problemas de projeto de sistemas inteligentes. A Tabela 4 compara as características entre sistema *fuzzy* e neurais.

Tabela 4 - Comparação entre sistemas *fuzzy* e Rede Neurais Artificiais.

Sistema de inferência *Fuzzy*	**Redes Neurais Artificiais**
Conhecimento a priori pode ser incorporado	Conhecimento a priori não utilizado
Utiliza o conhecimento linguístico	Capacidade de aprendizado
De fácil interpretação (Se-Então)	Caixa preta
Fácil interpretação e implementação	Algoritmos de aprendizado complexos
Conhecimento disponível	Dificuldade para extração de conhecimento

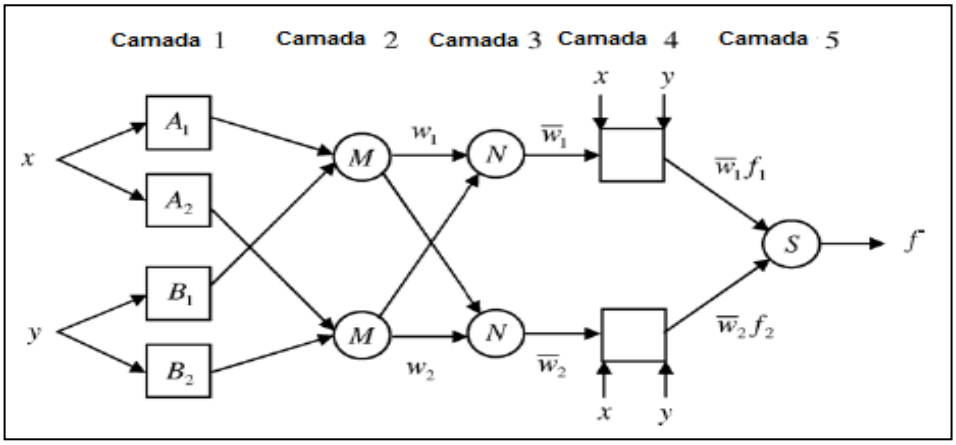
Fonte: adaptado de Silva (2015).

Os sistemas *neuro-fuzzy* objetivam conjugar a capacidade de aprendizado das redes neurais à interpretabilidade característica dos referidos sistemas (Mamdani, TSK etc.), síntese de funções-objetivo e algoritmos de minimização global adequados. À medida que o algoritmo progride rumo ao mínimo global, o respectivo sistema *fuzzy* apresenta comportamento cada vez mais próximo do desejado e traduzido pelos elementos do *training set* (CALDEIRA *et al*., 2007).

2.3.4.3 Arquitetura e treinamento de uma rede neuro-fuzzy

Proposto por Jang (1993) *apud* Caldeira *et al*. (2007), o modelo *Neuro-Fuzzy*(NF) utiliza predominantemente modelo do tipo TSK parametrizados e realiza treinamento por conjugação de técnicas de gradiente e mínimos quadrados. O modelo NF apresenta a seguinte configuração (Figura 19).

Figura 19 - Arquitetura de um modelo NF com duas entradas e uma saída para o modelo de TSK.

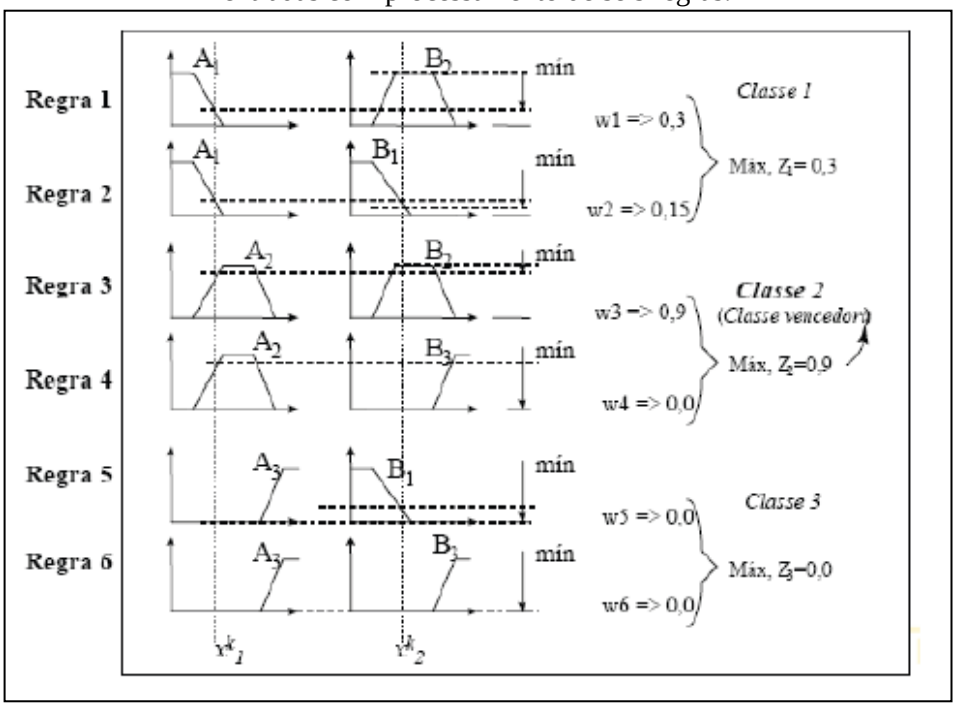


Fonte: Adaptada de Jang (1993) *apud* Caldeira *et al*. (2007).

O método mapeia as várias partes de um sistema de inferência *fuzzy* em uma rede adaptativa *feedforward* de cinco camadas a qual é treinada de modo supervisionado, usando dado *training set.* Ao final do treinamento, os parâmetros devidamente ajustados são mapeados inversamente em um sistema *fuzzy* de estrutura idêntica ao original, que deverá se comportar, aproximadamente, como o sistema gerou o *training set* utilizou para treinamento, ou seja, às mesmas entradas deverão corresponder saídas próximas (CALDEIRA *et al*., 2007).

A Figura 20 mostra algumas regras utilizadas para classificação em um modelo de *fuzzy* TSK.

Figura 20 - Exemplo de modelo Fuzzy Sugeno de primeira ordem de duas entradas com processamento de seis regras.



Fonte: Nunes (2009).

As diversas camadas apresentam as seguintes funcionalidades (CALDEIRA *et al*., 2007):

- Camada 1: Cada neurônio da camada 1 apresenta comportamento adaptativo e é responsável pelo cálculo do valor das funções de pertinência das entradas apresentadas ao sistema. A dinâmica do treinamento determina a evolução dos vários parâmetros.
- Camada 2: Essa camada realiza a aplicação de *t*-normas às saídas da camada anterior. O modelo NF utiliza, tipicamente, o produto algébrico. Trata-se de camada com elementos estáticos.
- Camada 3: Essa camada também é estática, sendo responsável pelo cálculo do grau de ativação normalizado das várias partes consequentes do sistema TKS subjacentes.
- Camada 4: É adaptativa e responsável pelo cálculo de cada uma das partes consequentes do sistema TKS atual. Seus parâmetros correspondem aos coeficientes das expressões afins.
- Camada 5: É fixa e responsável pela soma dos sinais de entrada locais e respectiva propagação para saída.

Nos últimos anos, vários estudos foram publicados utilizando o sistema de inferência adaptativa *neuro-fuzzy* (NF) aplicados nas previsões e classificações da área ambiental e de recursos hídricos, sendo, portanto, mais uma justificativa para o desenvolvimento deste trabalho.

Neste trabalho pretendeu-se investigar a estratégia para a fácil aplicação em que ambas as técnicas pudessem ser aplicadas.

2.3.5 Máquina de Vetores de Suporte (SVM)

A Máquina de Vetores de Suporte (SVM), proposta por Vapnik (1995), é um algoritmo de aprendizado supervisionado utilizado na classificação e regressão de dados para conjuntos linearmente ou não linearmente separáveis, através da escolha de um hiperplano ótimo de forma que a separação dos conjuntos de dados seja máxima.

Um hiperplano é um objeto matemático que pode ser um ponto, caso estejamos interpretando dados em uma dimensão (R), uma reta, para duas dimensões (R^2), um plano para três dimensões (R^3) e assim sucessivamente. A Equação 18 fornece o hiperplano geral (KOWALCZYK, 2017).

(18)

Sendo: x é o vetor dos dados, w é o vetor ortogonal ao hiperplano, que o caracteriza, e b é o vetor distância entre o hiperplano e a origem. A dimensão do hiperplano é dada pelos vetores w e x. A equação geral do hiperplano é obtida a partir da generalização da equação de uma reta, na qual se substitui y por x_2 e x por x_1, conforme mostram as equações 19, 20 e 21 (KOWALCZYK, 2017).

(19)

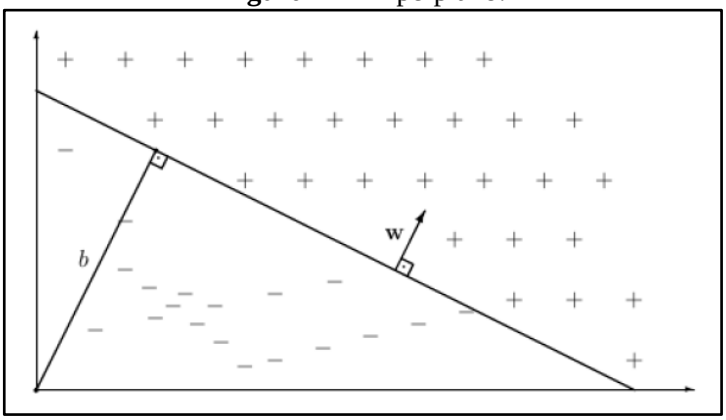
(20)

(21)

Sendo (x) e (a 1), obtem-se a equação geral do hiperplano:

(22)

Figura 21 - Hiperplano.

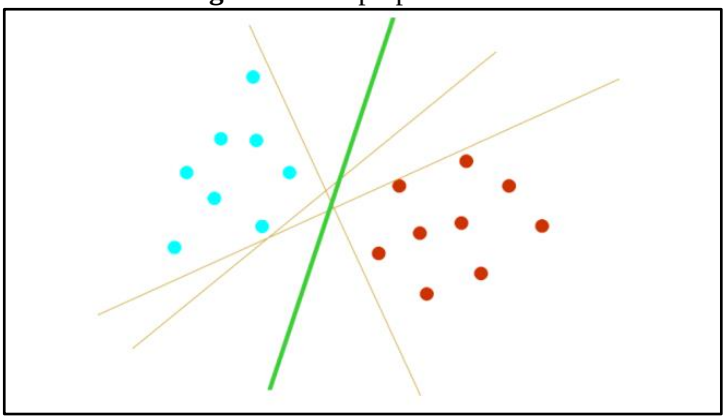


Fonte: (GONÇALVES, 2016)

O objetivo de um algoritmo classificador é obter o menor número de previsões erradas possível na etapa de treinamento. Nessa etapa, a SVM escolhe um plano aleatório que terá sua inclinação (vetor w) ajustada de acordo com o conjunto de dados de treinamento, cuja classificação já é conhecida. Esse procedimento é repetido até que o hiperplano divida corretamente os dados de treinamento.

O problema deste procedimento é que existem diversos hiperplanos que podem separar corretamente o conjunto de dados de treinamento. É possível que a escolha de um hiperplano obtenha bons resultados na etapa de treinamento, mas não seja nem um pouco desejável para a classificação dos outros dados, este é denominado de *Overfitting*. Para contornar esse problema, deve-se utilizar uma regra para decidir qual dos hiperplanos possíveis deve ser escolhido, diminuindo o erro na classificação dos dados (GONÇALVES, 2016).

Figura 22 - Hiperplano ótimo.



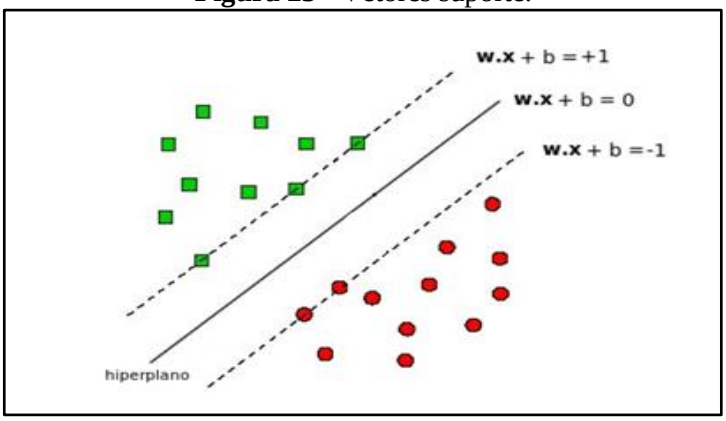
Fonte: (GUNN, 1998)

Para o entendimento dessa regra, é necessária a definição do conceito de margem de classificação. A margem de classificação é uma região no espaço que se encontra entre os conjuntos de dados, ou seja, sua espessura é a menor distância entre os conjuntos. A margem é limitada pelos chamados vetores suporte, que são planos paralelos ao hiperplano escolhido e servem de parâmetro para se determinar o hiperplano que melhor divide os conjuntos de dados, o hiperplano ótimo. Os vetores suporte apresentam as equações 23 e 24 (GIRARDELLO, 2010):

(23)

(24)

Figura 23 - Vetores suporte.



Fonte: Adaptado de Gonçalves, 2016.

A regra de decisão tem o objetivo de escolher o hiperplano que esteja o mais longe possível dos dois conjuntos de dados, simultaneamente, ou seja, o que detenha a maior margem de classificação e por consequência, maior distância entre os vetores suporte. Uma

característica importante que se pode observar na equação dos vetores suporte é que eles são equidistantes do hiperplano ótimo, pois $w \cdot x + b = 1$ e $w \cdot x + b = -1$.

A SVM utiliza uma função hipótese, $f(x) = sgn,$ para a classificação dos dados, levando em consideração as margens estabelecidas pelos vetores suporte (Equação 25) (KOWALCZYK, 2017):

$$f(x) \quad n(w \quad b) \quad \left\{ \right. \tag{25}$$

Sendo o y_i o rótulo do vetor de dados x_i, se x_i recebe o rótulo ele pertence ao conjunto que se encontra acima do hiperplano ótimo, caso x_i receba o rótulo 1, pertencerá ao conjunto que se encontra abaixo do hiperplano. Combinando as equações dos vetores suporte, obtém-se a Equação 26 (LORENA; CARVALHO, 2003).

$$(w \quad + b) \geq 1 \tag{26}$$

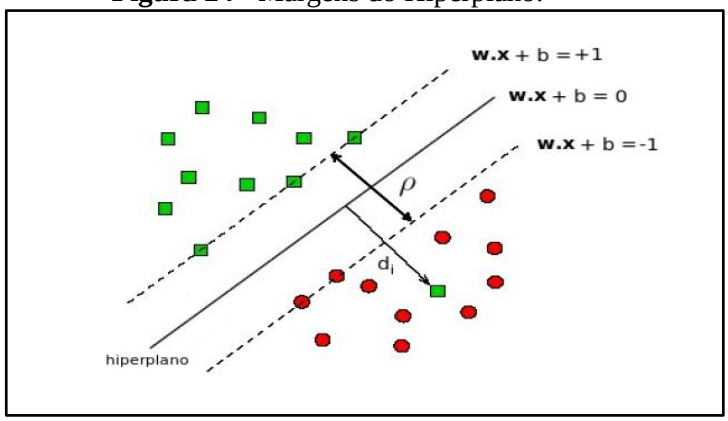
As distâncias entre um dado qualquer e o hiperplano é dada pela seguinte equação derivada da geometria analítica:

$$d(w \quad x) \quad \frac{}{\|w\|} \quad \frac{(w \quad b)}{\|w\|} \tag{27}$$

Como, $y_i(w \quad b) \quad 1$:

$$\overline{\|w\|} \tag{28}$$

Figura 24 - Margens do Hiperplano.



Fonte: (GONLÇALVES, 2016).

Se o vetor suporte está posicionado no limite do conjunto de dados, a distância entre ele e o hiperplano deverá ser o limite inferior da desigualdade, ou seja, a distância entre o

vetor suporte e o hiperplano é $/\|w\|$. Sendo ρ a espessura da margem, dada por $\rho=d_{+}+d_{-}$, obtemos o tamanho da margem em função do vetor ortogonal ao plano:

$$\overline{\|w\|} \tag{29}$$

Desta forma, a escolha do hiperplano ótimo se transforma num problema de otimização onde o objetivo é maximizar ρ, que é a espessura margem de classificação ou a distância entre os vetores suporte, que seria equivalente a minimizar a norma do vetor ortogonal ao hiperplano, w.

Minimizar $\|w\|$ é equivalente a minimizar $\frac{1}{2}\|w\|^2$. Sendo assim, o problema se torna passível de resolução por programação quadrática. Essa otimização é realizada utilizando o método clássico dos multiplicadores de Lagrange, dado pela seguinte equação (BEM-HUR; WESTON, 2010):

$$L(w \quad \alpha) \quad -\|w\| \quad \sum \alpha \ (y \ (w \qquad b) \quad 1) \tag{30}$$

Na qual é denominado de multiplicador de Lagrange e o problema passa a ser a minimização de L, que é uma função de w, b e . Para encontrar os pontos de mínimo da função, o método iguala as derivadas parciais da função de Lagrange a 0:

$$— \tag{31}$$

$$— \tag{32}$$

Derivando em relação às respectivas variáveis, obtemos as seguintes equações:

$$\sum \tag{33}$$

$$\sum \alpha \tag{34}$$

Substituindo as equações (25) e (26) na equação (22), obtém-se um novo problema de otimização, no qual será maximizada a Equação 35 (GIRARDELLO, 2010):

$$\sum \alpha_i - - \sum \sum \alpha \qquad (Xi \quad) \tag{35}$$

Equação esta sujeita às seguintes restrições:

$$\alpha_i \geq 0, \qquad i = 1,2,3,\dots,n \tag{36}$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0 \tag{37}$$

Através da otimização da função acima descrita, obtemos o valor máximo de α_i, que denominaremos α_i^*. Assim, podem-se obter os valores otimizados de w, w^*:

$$w^* = \sum_{i=1}^{n} \alpha_i^* y_i X_i \tag{38}$$

Para a obtenção do parâmetro otimizado b^*, é utilizada a equação derivada da condição Karush-Kuhn-Tucker, que é uma exigência para que a escolha do hiperplano seja ótima. Caso o hiperplano ótimo seja construído de forma a cumprir a condição de Karush-Kuhn-Tucker, essa será a escolha ótima para a situação (KOWALCZYK, 2017).

$$\alpha_i^*(y_i(\langle w^* \cdot x_i \rangle + b^*) - 1) = 0 \tag{39}$$

Após a obtenção dos parâmetros ajustados do hiperplano, w^*e b^*, é possível classificar um novo vetor de dados utilizando a função hipótese e a nova equação do hiperplano:

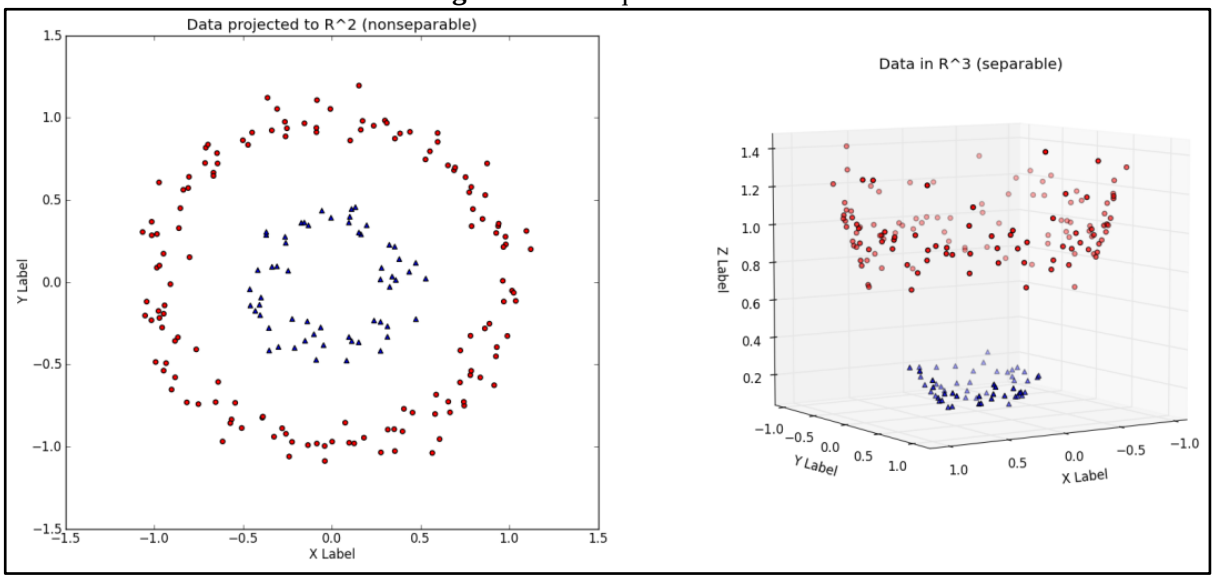
$$sgn(w^* \cdot x_i + b^*) = \begin{cases} Se\ w^* \cdot x_i + b^* \geq +1, & y_i = +1 \\ Se\ w^* \cdot x_i + b^* \leq -1, & y_i = -1 \end{cases} \tag{40}$$

O procedimento descrito acima só pode ser realizado para dados linearmente separáveis, pois existe um hiperplano que pode separá-los. Uma vez que a maioria dos dados relativos a fenômenos reais não são linearmente separáveis, é necessária a utilização de algum método específico para manipular o conjunto de dados de forma que esse passe a ser linearmente separável.

Para resolver o problema da não linearidade do conjunto de dados, devem ser realizadas transformadas não lineares através do uso de funções Kernel, $\phi(x)$. Esta transformação migrará o conjunto de dados de uma dimensão R^n para R^m, sendo $m > n \,\forall\, m, n \in \mathbb{Z}$, pois em uma dimensão maior, será mais provável que o conjunto de dados, antes não linearmente separável, seja linearmente separável, de acordo com o teorema de Cover (HAYKIN, 2009).

No exemplo a seguir, é mostrado um conjunto não linearmente separável em R^2 que é submetido a uma transformação não linear, através da função $\phi(x)$ (Equação 41), e então passa a ser representado em R^3.

Figura 25 - Truque do Kernel.



Fonte: Kim, 2015.

$$\phi(x1, x2) \rightarrow (z_1, z_2, z_3) = \left(x_1^2, \sqrt{2x_1x_2}, x_2^2\right) \tag{41}$$

As principais funções Kernel utilizadas estão listadas na Tabela 5.

Tabela 5 - Tipos de funções Kernel.

Tipo de Kernel	Função ϕ correspondente	Comentários
Linear	$x_i^T x_j$	-
Polinomial	$\left(x_i^T x_j + k\right)^p$	p e k devem ser especificados pelo usuário
Gaussiano (RBF)	$exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)$	A amplitude σ^2 é especificada pelo usuário
Sigmoide	$\tanh(\gamma \cdot x_i^T x_j + k)$	Utilizando somente para alguns valores de γ e k

Para a classificação em conjuntos não linearmente separáveis, realizamos o mesmo procedimento de otimização para encontrar o hiperplano ótimo em dados linearmente separáveis, a diferença é que substituimos o vetor de dados $x$por $\phi(x)$. Para a classificação de um vetor de dados X, a função decisão assumirá, após a otimização dos parâmetros, a seguinte forma (GONÇALVES, 2016):

$$sgn(w^* \cdot \phi(X) + b^*) \tag{42}$$

Apesar de ter sido desenvolvido para classificação, o SVM também pode ser utilizado para a regressão de dado fazendo algumas alterações no mecanismo, mas mantendo a ideia

geral do hiperplano. Seja o seguinte conjunto de dados de treinamento:{( ), ( ), , ( )}

O objetivo é encontrar uma função f, tal que, f: $\mathbb{R}$, ou seja, encontrar uma função que receba os dados de entrada contidos no vetor e retorne o valor de y corretamente. Essa função pode ser a equação do hiperplano:

$$f(w,b) = \hat{y}_m = w \cdot x_m + b \tag{43}$$

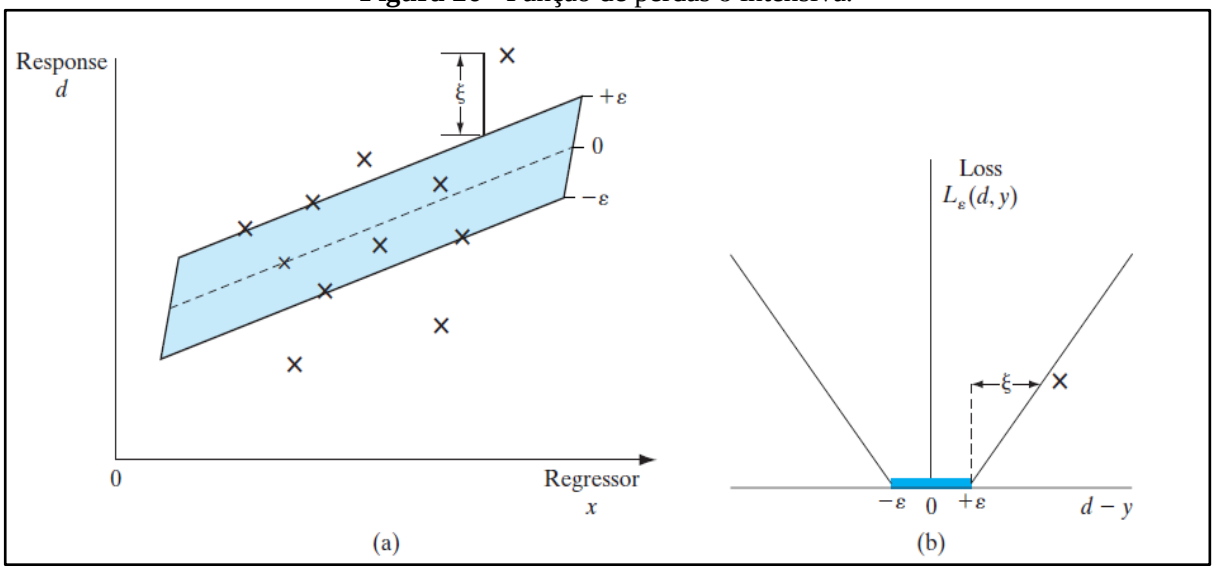
Com o intuito de encontrar a função f, primeiramente, é definido a função de perdas -intensiva (ε-*intensive loss function*). A função de perda convencional seria dada pelo módulo da diferença entre o valor esperado (y) e o valor obtido pela função $\hat{y}$.

$$L(y,\hat{y}) = |y - \hat{y}| \tag{44}$$

A função de perdas ε-intensiva é construída de forma que os pontos que estejam dentro da margem de erro ε, não influenciem no ajuste da equação do hiperplano, ou seja, assumirá a forma da Equação 45 (GUNN, 1998).

$$(y,\hat{y})\begin{cases} \hat{y}| & \\ & \hat{y}| \\ & \hat{y}| \end{cases} \tag{45}$$

Figura 26 - Função de perdas ε-intensiva.



Fonte: HAYKIN, 2009.

Para obter os parâmetros otimizados w^* e , da equação do hiperplano, devemos minimizar a função risco (46), sujeito às seguintes restrições (HAYKIN, 2009):

$$-\|w\| \qquad \sum|\hat{y} \tag{46}$$

$$y_i - \hat{y}_i \leq \varepsilon + \xi_i \quad (47)$$

$$\hat{y}_i - y_i \leq \varepsilon + \xi_i' \quad (48)$$

$$\xi_i, \xi_i' \geq 0 \quad (49)$$

O objetivo das restrições (47) e (48), é fazer com que a distância entre o valor verdadeiro e o valor previsto seja menor a cada iteração. A restrição (49) garante que os valores das margens externas ξ_i *e* ξ_i', sejam sempre positivos.

Esse problema de otimização deve ser resolvido pelo método dos multiplicadores de Lagrange, assim como o SVM para classificação. Levando em consideração as restrições impostas mostradas na Equação 50 (HAYKIN, 2009).

$$\begin{aligned} J(w, \xi_i, \xi_i', \alpha, \alpha', \gamma, \gamma') = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i') - \sum_{i=1}^{n}(\gamma_i\xi_i + \gamma_i'\xi_i') \\ - \sum_{i=1}^{n}\alpha_i(w \cdot x_i + b - y_i + \varepsilon + \xi_i) \\ - \sum_{i=1}^{n}\alpha_i'(y_i - w \cdot x_i - b + \varepsilon + \xi_i') \end{aligned} \quad (50)$$

Sendo: γ e γ' são os novos multiplicadores de Lagrange adicionados à equação para garantir que as restrições sejam atendidas e que os multiplicadores originais, α_i e α_i', assumam a forma de variável e que as margens externas também sejam minimizadas. Derivando a Equação 49 e igualando a zero serão obtidas as seguintes expressões:

$$w^* = \sum_{i=1}^{n}(\alpha_i - \alpha_i')\, x \quad (51)$$

$$\sum_{i=1}^{n}(\alpha_i - \alpha_i') = 0 \quad (52)$$

$$\alpha_i + \xi_i = C \quad (53)$$

$$\alpha_i' + \xi_i' = C \quad (54)$$

Através da equação (51) é possível obter o valor otimizado para o vetor ortogonal, w^*. Para encontrar o vetor *bias* otimizado (b^*), é utilizada a condição de Karush-Kuhn-Tucker, condição também utilizada para encontrar esse parâmetro na SVM para classificação. Aplicando essa condição, são obtidas as seguintes equações (BASAK *et al*., 2007):

$$\alpha_i(\varepsilon + \xi_i + y_i - \hat{y}_i) = 0 \quad (55)$$

$$\alpha_i'(\varepsilon + \xi_i' + \hat{y}_i - y_i) = 0 \quad (56)$$

$$(C - \alpha_i)\xi_i = 0 \quad (57)$$

$$(C - \alpha_i')\xi_i' = 0 \quad (58)$$

Das considerações de Karush-Kuhn-Tucker, são obtidas as seguintes informações:

$$\alpha_i \alpha_i' = 0 \quad (59)$$
$$\xi_i = 0 \; Para \; 0 < \alpha_i < C \quad (60)$$
$$\xi_i' = 0 \; Para \; 0 < \alpha_i' < C \quad (61)$$

Como o intervalo abrange $0 < \alpha_i, \alpha_i' < C$, logo, $\alpha_i, \alpha_i' \neq 0$. Sendo assim:

$$(\varepsilon + \xi_i + y_i - \hat{y}_i) = 0 \; Para \; 0 < \alpha_i < C \quad (62)$$
$$(\varepsilon + \xi_i' + \hat{y}_i - y_i) = 0 \; Para \; 0 < \alpha_i' < C \quad (63)$$

Tendo em vista as equações (62) e (63):

$$(\varepsilon + y_i - \hat{y}_i) = 0 \; Para \; 0 < \alpha_i < C \quad (64)$$
$$(\varepsilon + \hat{y}_i - y_i) = 0 \; Para \; 0 < \alpha_i' < C \quad (65)$$

Considerando a equação do hiperplano com parâmetros otimizados (66) e aplicando as restrições estabelecidas nas equações (62) e (63), é obtida uma relação parab^* (Equações 67, 68 e 69) (HAYKIN, 2009).

$$f(x_i) = \hat{y}_i = w^* \cdot x_i + b^* \quad (66)$$
$$b^* = \hat{y}_i - w^* \cdot x_i \quad (67)$$
$$b^* = y_i - \varepsilon - w^* \cdot x_i \; Para \; 0 < \alpha_i < C \quad (68)$$
$$b^* = y_i + \varepsilon - w^* \cdot x_i \; Para \; 0 < \alpha_i' < C \quad (69)$$

A partir das equações 67, 68 e 69 é possível obter o valor do vetor bias otimizado (b^*), sabendo o vetor ortogonal otimizado (w^*), já calculado, o vetor de dados x_i e o margem de erro ε.

O procedimento acima foi realizado tomando como base um conjunto de dados linearmente separável. Para resolver problemas cujo conjunto de dados seja não linearmente separável basta substituir o vetor de dados x_i, pelo vetor de dados transformado por uma função Kernel, $\phi(x)$, como realizado na SVM para classificação. Neste trabalho a SVM foi utilizada para regressão na inferência de parâmetros de qualidade de água na forma estacionária e de séries temporais.

2.3.6 Transformada *Wavelet e* sua Conjunção

Caracterizada pela alta complexidade, dinamismo e não estacionariedade, a previsão hidrológica e hidro-climatológica sempre representaram como um desafio para os hidrologistas que reconhecem seu papel essencial na gestão ambiental e de recursos hídricos, bem como na mitigação de desastres relacionados à água. Nos últimos anos assistiu-se a um

aumento significativo do número de abordagens científicas aplicadas à modelização e previsão hidrológica, incluindo as abordagens particularmente baseadas em "dados" ou "orientadas por dados". Essas abordagens de modelagem envolvem equações matemáticas extraídas do processo físico na bacia hidrográfica, mas a partir de uma análise de séries temporais de entrada e saída concorrentes (Solomatine e Ostfeld, 2008). Esses modelos podem ser definidos com base em conexões entre as variáveisde estado do sistema (entrada, variáveis internas e de saída) com apenas um número limitado de suposições feitas sobre o comportamento físico do sistema. Exemplos típicos de modelos baseados em dados são as curvas de classificação, o método de hidrograma unitário e vários modelos estatísticos (Regressão Linear, LR, Multi Linear, Auto Regressivo Integrado Média Móvel, ARIMA) e métodos de aprendizagem mecânica. Os modelos de séries temporais de caixa preta convencionais, tais como ARIMA, ARIMA com entrada exógena (ARIMAX) e Regressão Linear Múltipla (RLM) são modelos lineares e assumem a estacionaridade do conjunto de dados. Tais modelos são incapazes de lidar com não-estacionariedade e não-linearidade envolvidos em processos hidrológicos. Como resultado, muitos pesquisadores têm se concentrado no desenvolvimento de modelos capazes de modelar processos não lineares e não estacionários.

Os métodos de Inteligência Artificial (IA), baseados em dados, mostraram-se promissores na modelagem e previsão de processos hidrológicos não-lineares e no manuseio de grandes quantidades de dinâmica e ruído ocultas em conjuntos de dados. Tais propriedades de modelos baseados em IA são bem adequadas para problemas de modelagem hidrológica. Foram utilizadas inúmeras ferramentas ou técnicas de IA, incluindo versões de otimização de pesquisa, otimização matemática, bem como métodos de lógica, classificação, aprendizado estatístico e probabilidade (LUGER, 2005).

Entre as aplicações mais amplas de métodos de IA, *Fuzzy*, RNAs e SVM são amplamente utilizados em diferentes campos da hidrologia. Desde a sua emergência na hidrologia, o desempenho eficiente de técnicas de IA, tais como modelos baseados em dados, tem sido relatado em uma ampla gama de processos hidrológicos (por exemplo, precipitação, fluxo de fluxo, escoamento de chuva, carga de sedimento, água subterrânea, seca, neve derretida, evapotranspiração, qualidade da água, etc.). O número de investigadores ativos nesta área aumentou significativamente ao longo da última década, tal como o número de publicações. Várias dezenas de aplicações bem-sucedidas para modelagem de processos

hidrológicos em qualidade da água usando RNA, *Fuzzy*, e SVM foram relatadas, com alguns exemplos listados, nos capítulos anteriores.

Apesar da flexibilidade e utilidade dos métodos baseados em IA na modelação de processos hidrológicos, eles têm algumas desvantagens com respostas altamente não estacionárias, isto é, que variam numa grande escala de frequências, de hora em hora a multidecadal. Em tais casos de "sazonalidade", a falta de dados de entrada/saída, pré/pós-processamento, pode não permitir que os modelos de IA para lidar adequadamente com dados não-estacionários. Aqui, os modelos híbridos que combinam os esquemas de dados pré e pós-processamento com técnicas de IA podem desempenhar um papel importante.

Modelos hidrológicos híbridos podem tirar proveito dos modelos de caixa preta (IA) e sua capacidade de descrever eficientemente os dados observados em termos estatísticos, bem como outras informações prévias, ocultas nos registros observados. Os modelos híbridos aqui discutidos representam a aplicação conjunta de métodos baseados em IA com a transformada *wavelet* para melhorar o desempenho geral do modelo.

A transformada *wavelet* tem aumentado em uso e popularidade nos últimos anos desde a sua criação no início dos anos 80, mas ainda não é tão amplamente utilizado como a transformada de Fourier. No entanto, a análise de Fourier tem uma desvantagem significativa: a transformada de Fourier de um sinal no domínio da frequência resulta na perda de informações de tempo, de tal forma que se torna impossível dizer quando ocorreu um evento particular. Em contraste, a análise *wavelet* permite o uso de intervalos de tempo longos quando é necessária informação mais precisa de baixa frequência, e regiões mais curtas quando a informação de alta frequência é de interesse.

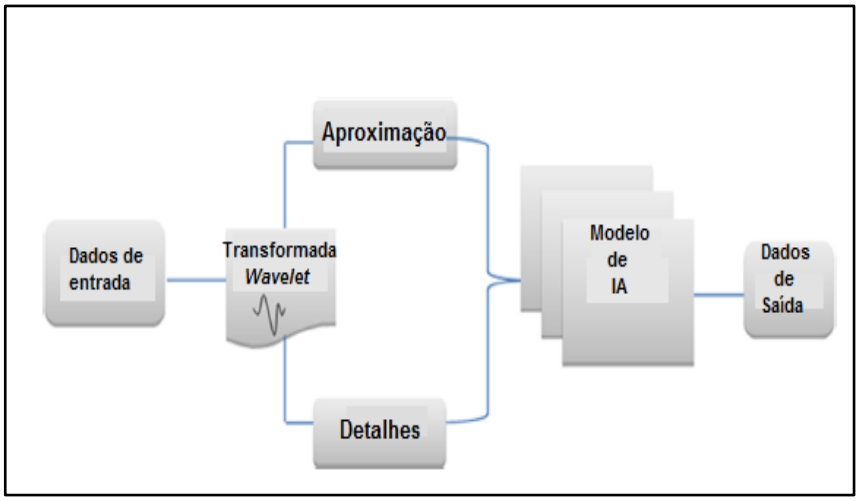
Como um avanço no processamento do sinal, as transformações *wavelet* podem evitar de forma confiável as falhas do modelo IA ao lidar com o comportamento não-estacionário dos sinais. Uma técnica matemática útil na análise numérica e manipulação de conjuntos de sinais multidimensionais, a análise *wavelet* fornece uma representação em escala de tempo do processo e de suas relações. De fato, a propriedade principal da transformada *wavelet* é a sua capacidade de fornecer uma localização em escala de tempo de um processo. A transformada *wavelet* tem atraído atenção significativa desde o seu desenvolvimento teórico em 1984 (Grossmann e Morlet, 1984). Uma série de estudos hidrológicos recentes têm implementado análise *wavelet*, como por exemplo, os trabalhos de Kim e Valdes (2003); Zhou *et al.*, 2008; Kisi (2009a,b, 2010); Nourani *et al.*(2009a,b, 2011); Adamowski e Sun (2010); Maheswaran e Khosa (2012); Partal e Kisi (2007); Sang (2012); Tiwari e Chatterjee (2010).

A transformada *Wavelet* é aplicável na extração de informações não triviais e potencialmente úteis, ou conhecimento, a partir dos grandes conjuntos de dados disponíveis em ciências experimentais (registros históricos, reanálise, simulações de modelos climáticos globais, etc.). Fornecer informações explícitas de forma legível pode ser usado para resolver problemas de diagnóstico, classificação ou previsão. Em uma revisão das aplicações da transformada *wavelet* na modelagem de séries temporais hidrológicas, Sang (2013) destacou as informações multifacetadas que podem ser extraídas de tais análises: caracterização e compreensão das escalas multi-temporais das séries hidrológicas, identificação de sazonalidades e tendências, e dados de ruído. Portanto, a capacidade da transformada *wavelet* para decompor sinais não estacionários em sub-sinais em diferentes escalas temporais (níveis) é útil na interpretação dos processos hidrológicos (NASON e SACHS, 1999; ADAMOWSKI, 2008; ADAMOWSKI *et al*., 2009; KISI, 2010; MIRBAGHERI *et al*., 2010; SANG, 2012).

Dependendo das capacidades individuais dos métodos *wavelet* e IA, pode-se inferir que um modelo híbrido composto por ambos teria simultaneamente as vantagens de ambas as técnicas. A abordagem *wavelet*-IA combinada é uma metodologia útil, fundamentada tanto na transformada *wavelet* como em várias técnicas de modelagem de IA. Permite a construção de modelos articuláveis com aplicações tão amplas em hidrologia como o desembocamento, otimização, remediação de funções RNA ativas, bem como a previsão de processos hidrológicos. Neste último caso, os modelos de *wavelet*-IA foram explorados por hidrólogos, uma vez que a combinação permite uma elucidação detalhada dos sinais, tornando o método híbrido uma ferramenta eficaz para prever fenômenos hidrológicos. Nas tarefas de previsão, o método de IA de *wavelet* híbrido (Figura 27) segue um procedimento de dois passos (Nourani *et al*., 2014):

a) Utilização da transformada *wavelet* para o pré-processamento de dados de entrada. Isto inclui fornecer uma representação de tempo-frequência de um sinal em diferentes períodos no domínio do tempo, assim como uma informação considerável sobre a estrutura física dos dados.
b) Extração de recursos do sinal principal para servir como entradas de IA, e permitindo o modelo completo para processar os dados.

Figura 27 - Diagrama esquemático do modelo de previsão de *wavelet*-IA híbrido.



Fonte: Adaptada de Nourani *et al*. (2014)

A seleção de uma wavelet mãe eficiente e nível de decomposição são duas questões importantes no primeiro passo. A seleção apropriada da wavelet mãe constitui a decisão mais importante associada ao primeiro passo; Tanto no caso de transformadas de onda discreta e contínua (TWD e TWC, respectivamente). TWC e TWD constroem uma representação tempo-frequência na forma de um sinal contínuo ou discreto, respectivamente. Análises detalhadas sobre o desempenho de diferentes wavelets mãe em simulações hidrológicas levaram à conclusão de que para determinar a wavelet mãe ideal para um dado problema uma variedade de wavelets mãe deve ser testada através de um processo de tentativa e erro (MAHESWARAN e KHOSA, 2012a; NALLEY et al., 2012; NOURANI et al., 2011; SANG, 2012).No entanto, semelhança de forma entre a wavelet mãe e a série de temposem tratamento é muitas vezes a melhor orientação na escolha de uma confiável wavelet mãe. Geralmente, as wavelets mãe com uma forma de suporte compacta (por exemplo, Daubechies-1, Haar e Daubechies- 4, db4) são os mais eficazes na geração de características de localização de tempo para séries temporais que têm uma memória curta e características transitórias de curta duração. Em contraste, as wavelets mãe com uma forma de suporte amplo (por exemplo, Daubechies-2, db2) produzem previsões confiáveis para séries temporais com características de longo prazo (MAHESWARAN e KHOSA, 2012a).

Como o TWD começa com um conjunto discreto de dados e considera um conjunto de balanças diádicas, é compatível com a observação discreta de sinais hidrológicos. Para estudar o sinal, a discretização vem em primeiro lugar e, como resultado, os níveis de decomposição seguem. Embora a seleção apropriada da escala máxima seja também importante na TWC, ela

desempenha um papel essencial no TWD devido ao procedimento de decomposição e extração de subséries dominantes que não podem ser descritas tão facilmente quanto com o TWC. Portanto, juntamente com a escolha do tipo *wavelet* mãe, a determinação do nível de decomposição apropriado (escala) é outro subpasso importante no primeiro passo quando se aplica TWD. Em estudos iniciais, o nível ótimo de decomposição foi geralmente determinado através de um processo de tentativa e erro, mas posteriormente foi introduzida na literatura uma fórmula que relaciona o nível mínimo de decomposição, L, com o número de pontos de dados dentro da série temporal Ns (AUSSEM *et al*., 1998, NOURANI *et al*., 2009b, WANG*et al.*, 2013):

$$t[\log \quad] \tag{70}$$

Mais tarde, Nourani *et al*. (2011) criticou o resultado desta fórmula, afirmando que, tendo sido derivado para sinais totalmente autorregressivos (AR), considera apenas o comprimento das séries temporais, sem prestar atenção aos efeitos sazonais. Uma vez que muitas características sazonais podem ser incorporadas em sinais hidrológicos, uma visão precisa do processo em estudo e atenção à periodicidade do processo pode ser útil na seleção de um nível de decomposição apropriado para a análise de TWD diádico. O nível de decomposição l contém l pormenores e, como exemplo, no caso da modelização diária, refere-se ao modo 2n dia em que n = 1, 2, . . . , L (por exemplo, modo de 21 dias, modo de 22 dias, modo de 23 dias que é modo quase semanal, modo de 24 dias, modo de 25 dias que é modo quase mensal, etc.), portanto, Dependendo do tipo de *wavelet*, do nível de decomposição e do tipo de método IA aplicado, várias abordagens podem ser examinadas de acordo com o objetivo no desenvolvimento do modelo de *wavelet*-IA híbrido. Neste contexto, os métodos de IA podem ser considerados em três categorias básicas: otimização, lógica, classificação e aprendizagem estatística; Com base na utilização de IA sobre um destes três campos, podem ser inferidos diferentes propósitos para o modelo *wavelet*-IA híbrido. Geralmente, a aplicação coletiva de métodos de otimização e análise de *wavelet*s leva ao reconhecimento de insumos ótimos para modelos de IA (KUO *et al*., 2010a, b, WANG *et al*., 2011). Caracterização e classificação de insumos dominantes a serem utilizados na previsão (NOURANI *et al*., 2013, 2014), juntamente com a detecção de sazonalidade (NOURANI *et al*., 2009a, b, 2011, 2012), bem como a redução / remoção de ruído das séries temporais hidrológicas (CAMPISI *et al*., 2012, GUO *et al*., 2011) são elementos importantes que contribuem para uma melhor previsão para planejamento futuro através de modelos híbridos *wavelet*-IA.

2.3.7 Estado da Arte sobre Modelagem aplicada à Qualidade da Água

Dentre os trabalhos publicados envolvendo a análise de dados e a modelagem para estimativa da qualidade da água utilizando técnicas de máquinas de aprendizado estatístico e de inteligência artificial, os relacionados a seguir merecem destaque e foram aqui descritos nas Tabelas 06 e 07, com os trabalhos com dados estacionários e dinâmicos, respectivamente:

Tabela 6 - Tipos de modelagem de parâmetros de qualidade da água com dados estacionários reportados na literatura.

Ano do Artigo	Nº de Dados	Parâmetro Global	Técnicas
Li *et al.* (2016)	969 amostras	Oxigênio Dissolvido (OD)	Regressão Linear Múltipla (MLR), Redes Neurais Artificiais (BPNN) e Máquina de Vetores de Suporte (SVM).
Hameed *et al.* (2017)	5233 amostras	Oxigênio dissolvido (OD), demanda bioquímica de oxigênio (DBO), demanda química de oxigênio (DQO), sólido suspenso (SS), nitrogênio amoniacal e valor de pH (pH).	Redes Neurais Função De Base Radial (RBFNN) e o modelo de Rede Neural *back propagation*
Csábrági *et al.* (2017)	409 dados	Oxigênio Dissolvido (OD)	Regressão Multivariável Linear, Rede Neural "Multilayer Perceptron" (MLPNN), Rede Neural de Função de Base Radial (RBFNN) e Rede Neural de Regressão Geral
Khan e Chai (2017)	17.280 dados	Índice de Qualidade da Água (IQA)	Redes Neurais Artificiais (RNA) e Sistema de Inferência Adaptativo Neuro-*Fuzzy* (ANFIS) e ensemble: RNA-NF
Djerioui *et al.* (2018)	774 amostras	Cloro	Análise De Componentes Principais (ACP), Máquina de Vetores de Suporte (SVM), Máquinas de Aprendizado Extremo (ELM)
Chou *et al.* (2018)	1635 dados	Índice de Estado Trófico (IET)	Redes Neurais Artificiais (RNA) e Máquina de Vetores de Suporte (SVM), Regressão Linear (LR), Árvore de Classificação e Regressão (CART) e ensembles: RNAs+CART+SVR, RNAs+CART+LR, CART+SVR+LR, RNAs+SVR+LR e RNAs+CART+SVM+LR

Tabela 7 - Tipos de modelagem de parâmetros de qualidade da água reportados na literatura: Dados Dinâmicos.

Ano do Artigo	Nº de Dados	Parâmetro Global	Técnicas
Faruk (2010)	108 (mensais)	Oxigênio Dissolvido (OD)	ARIMA e RNA
Ravansalar (2015)	274 (mensais)	Condutividade Elétrica (CE)	RNA e RNA-*Wavelet*
Alizadeh *et al.* (2015)	Diários (650) e horários (540)	Temperatura, OD, clorofila, salinidade e turbidez	RNA e RNA-*Wavelet*
Zhang (2016)	60 amostras (mensais)	Nitrogênio total e fósforo total	ARIMA e RBFNN (Redes Neurais do tipo RBF)
Yu *et al.* (2018)	365 amostras	Demanda Química de Oxigênio (DQO) e Demanda Bioquímica de Oxigênio (DBO)	Análise de Componentes Principais (ACP) e Máquina De Aprendizagem Extrema (ELM)
Haghiabi *et al.* (2018)	660 dados	pH, condutividade específica (EC), bicarbonato (HCO3), sulfatos(SO4), cloretos (Cl), sólidos totais dissolvidos (TDS), Sódio (Na), magnésio (Mg), cálcio (Ca)	Redes neurais artificiais (RNA), Método de grupo de manipulação de dados (GMDH, do inglês: Group Method of data Handling) e Máquina de Vetores de Suporte (SVM, do inglês: Support Vector Machine)
Barzegar *et al.* (2018)	315 dados mensais	Condutividade Elétrica (CE)	Máquina De Aprendizagem Extrema (ELM), NF e os modelos híbridos de máquina Wavelet-extremo (WT-ELM) e Wavelet-NF

2.3.7.1 Análise de Dados

No estudo, Peeters e Dassargues (2006), a análise de componentes principais é comparada com o mapa (SOM) algoritmo de auto-organização de Kohonen. Ambas as técnicas tiveram sucesso na distinção entre ambos os aquíferos em estudo e revelaram as relações entre as variáveis. A principal vantagem da ACP é a quantificação matemática de correlação entre variáveis e a expressão dos dados originais no subespaço definido pelos componentes principais. A visualização do SOM-análise, por outro lado permite uma interpretação direta da estrutura de conjunto de dados em que pode ser identificado até mesmo relações não-lineares entre as variáveis. Além disso, o SOM-algoritmo pode lidar com uma quantidade limitada de valores em falta no conjunto de dados, ao contrário da ACP.

Astel *et al*. (2007) aplicaram três técnicas de classificação (projeções de carga e pontuação com base em Análise de Componentes Principais (ACP), Análise de Cluster e Mapas Auto-organizáveis de Kohonen (SOM)) para um grande conjunto de indicadores ambientais de indicadores químicos da qualidade da água do rio. O estudo foi realizado utilizando dados de monitorização da qualidade da água a longo prazo. As vantagens do algoritmo SOM e sua capacidade de classificação e visualização para grandes conjuntos de dados ambientais são enfatizadas. Os resultados obtidos permitiram detectar clusters naturais de locais de monitoramento com similar qualidade de água e identificar importantes variáveis discriminantes responsáveis pelo agrupamento. O clustering SOM permite a observação simultânea de mudanças espaciais e temporais na qualidade da água. A abordagem quimiométrica revelou padrões diferentes de locais de monitoramento denominados condicionalmente "tributário", "urbano", "rural" ou "fundo". Esta separação objetiva poderia conduzir a uma optimização das redes de monitorização dos rios e a uma melhor identificação das alterações naturais e antropogénicas ao longo do rio.

Zimmermann *et al*. (2008), avaliaram as alterações na qualidade da água do Rio Tibagi causados pelas atividades urbanas e industriais na região de Ponta Grossa. O estudo envolveu o monitoramento físico-químico e microbiológico da massa de água, que foram avaliados por uma rotina de análise de componentes principais. A ACP mostrou que o efeito de fontes pontuais associado à atividade industrial contribue para o aumento da concentração total de nitrogênio amoniacal e a redução do oxigênio dissolvido na região estudada. Este trabalho demonstra a importância da estatística multivariada.

Bernardi et. al. (2009) utilizaram a Análise de Componentes Principais (ACP) para avaliar os principais parâmetros físico-químicos do Alto Rio Madeira e seus tributários, Estado de Rondônia, Brasil. As coletas foram realizadas em janeiro de 2003 a dezembro de 2004 (composta por 23 transcectos de amostragem, sendo 11 transcectos no período de estiagem e 12 transcectos no período de cheia). Os principais parâmetros físico-químicos indicam que a Bacia do Alto Rio Madeira engloba três diferentes grupos de águas em relação ao pH (levemente alcalino a levemente ácido), condutividade (alta e baixa), concentração de oxigênio dissolvido e o teor de sólidos em suspensão. A ACP destes parâmetros demonstrou a influência do pH, condutividade e sólidos em suspensão na variabilidade total da hidroquímica. Os escores da ACP revelaram diferenças resultantes da influência sazonal (estação seca e chuvosa). O fluxo e os parâmetros físico-químicos podem ser discriminados de acordo com a estação do ano, que por sua vez são afetados pelo tipo de solos marginais, o

uso do solo (agricultura e pecuária), áreas alagadas e pelas características específicas dos tributários.

Pinto *et al*. (2013) afirmam que a utilização de um grande número de indicadores de qualidade de água torna oneroso o monitoramento dos corpos hídricos a longo tempo. Nesse sentido, a Análise de Componentes Principais (ACP) pode ser considerada como uma técnica promissora de grande auxílio na gestão dos recursos hídricos, pois permite diminuir a dimensionalidade dos dados, facilitando a sua análise. Eles propuseram um índice de qualidade de água (IQA) para a região das nascentes do Rio Grande, em duas situações de uso do solo (uma sub-bacia sob a Mata Atlântica e outra predominantemente sob pastagem) na Serra da Mantiqueira, Minas Gerais. Os indicadores com maiores pesos, segundo a ACP, foram os *Coliformes totais*, nitrato, *Coliformes termotolerante*, demanda química de oxigênio e temperatura. A sub-bacia sob Mata Atlântica apresentou os melhores resultados de IQA, denotando a importância desse tipo de ambiente na manutenção da qualidade da água dos mananciais na região da Serra da Mantiqueira.

An *et al.* (2016) utilizaram a Análise de Componentes Principais (ACP) e um Mapa Auto-Organizado (SOM) para analisar um conjunto de dados complexos obtidos de 2009 a 2011 a partir das estações de monitoramento de água nos rios Tolo Harbor e Channel Water Control Zone (Hong Kong). A ACP foi inicialmente aplicada para identificar os componentes principais (CPs) entre os parâmetros não lineares e complexos da qualidade da água superficial. A SOM seguiu a ACP e foi implementada para analisar as complexas relações e comportamentos dos parâmetros. Os resultados revelam que a ACP reduziu os parâmetros multidimensionais para quatro CPs significativos que são combinações dos originais. As relações positivas e inversas dos parâmetros foram mostradas explicitamente pela análise de padrões nos planos componentes. Verificou-se que ACP e SOM são ferramentas eficientes para capturar e analisar o comportamento de dados de qualidade de água de superfície superficiais multivariáveis, complexos e não lineares.

2.3.7.2 Modelagem

Lu e Lo (2002) retrataram o diagnóstico de reservatório de água utilizando a lógica fuzzy para representar o processo de eutrofização em termos de parâmetros como fósforo total e clorofila-a. Nesta pesquisa, um método de avaliação complementar, mapa auto-organizado (SOM), tem sido usado para diagnosticar a qualidade da água para desenvolver um classificador de estado trófico e é ilustrado com um estudo de caso da avaliação trófica do

reservatório de Fei-Tsui em Taiwan. O banco de dados histórico foi coletado da agência de gerenciamento do Reservatório Fei-Tsui de 1987 a 1995, no total de 110 registros de dados. Os resultados do SOM são comparados com os do índice Carlson e da lógica *fuzzy*, mostrando que os registros inconsistentes podem ser mapeados para a zona de dados conflitantes do mapa de saída do SOM. Além disso, o SOM cria um eixo de diagnóstico no mapa para expressar o status trófico do corpo de água. Enquanto o modelo SOM for bem treinado, novos registros podem ser avaliados e classificados como um dos três níveis tróficos ou casos especiais. Se condições especiais de qualidade da água forem expressas na saída do SOM, esses dados podem revelar que, tanto o fósforo total (PT) quanto a clorofila-a (chl-a) estão acima do normal.

Bowden *et al*. (2002) apresentaram dois métodos com base num algoritmo genético e um SOM que eram capazes de se dividir os dados disponíveis em subconjuntos com propriedades estatísticas comuns. Estes foram usados para desenvolver modelos RNA *back-propagation* para a previsão de salinidade no rio Murray em Murray Bridge, Austrália do Sul, com 14 dias de antecedência. Eles compararam o desempenho dos modelos obtidos com um modelo que foi desenvolvido por meio de divisão de dados arbitrária. Eles encontraram que os modelos desenvolvidos utilizando as técnicas apresentadas superaram o convencional. Verificou-se que o SOM também pode ser usado como uma ferramenta para encontrar saídas, visto que os modelos de RNAs podem dar resultados pobres em algumas partes da série temporal. Para este fim, Bowden *et al*. (2002) utilizaram um SOM para o agrupamento dos dados. Um SOM com 10 x 10 neurônios na camada de Kohonen foi utilizado de modo que os padrões de entrada foram agrupados em grupos 100, onde 49 consistiram em três ou mais padrões. De cada um desses agrupamentos três padrões de dados foram incluídos na amostra, um para cada treinamento, testes e validação. Para os 51 grupos que contenham menos de três padrões, eles usaram o registro de amostra no cluster com apenas um registro como formação e no caso de clusters com dois registros, um deles foi usado para treinamento e outro para testes.

Outra questão que também precisa de uma grande atenção no desenvolvimento de modelos RNA é a determinação de variáveis de entrada significativas. Isto é devido ao fato de que a apresentação de todas as variáveis de entrada potenciais para a RNA e contar com a rede para identificar os mais críticos podem criar problemas. Como Bowden *et al*. (2005a, b) apontou, que há várias desvantagens com esta abordagem, incluindo o aumento da complexidade e requisitos de memória computacional, dificuldade de aprendizagem,

convergência errada e mau desempenho do modelo, o aumento da complexidade do modelo e, consequentemente, uma dificuldade em compreender o modelo, bem como aumento do ruído devido à inclusão de variáveis de entrada espúrias. Por conseguinte, existe uma necessidade de introduzir um método de determinação de entrada eficiente para selecionar um modelo de forma mais parcimoniosa.

Esta questão foi abordada por Bowden *et al*. (2005a, b), que apresentou duas metodologias, uma baseada no algoritmo parcial de informação mútua (PMI) e o outro baseado em um SOM combinado com um algoritmo genético híbrido e uma rede neural regressão geral (SOM-GAGRNN). O primeiro usa uma medida parcial do critério de informação mútua, a fim de determinar as entradas que tenham relação altamente significativa com a variável de saída a ser o sistema modelado e a segunda usa um SOM para agrupar as variáveis de entrada em grupos de entradas semelhantes. Em seguida, eles selecionaram uma variável de entrada de cada cluster para que a entrada com a menor distância euclidiana aos pesos do cluster fosse selecionada a partir de cada cluster. Assim, a dimensionalidade do espaço de entrada foi reduzida e as variáveis obtidas foram introduzidas no GAGRNN para determinar quais entradas que têm relação significativa com a variável do sistema de produção que está sendo modelado. Os autores testaram os métodos propostos em vários conjuntos de dados sintéticos e concluíram que, em termos de desempenho preditivo, ambos os métodos eram bons enquanto que do ponto de vista da obtenção de informações valiosas sobre o sistema sob investigação, o primeiro foi recomendado.

Para verificar as abordagens acima em dados reais, foram aplicados para a determinação de entrada para um modelo de RNA na previsão de salinidade de 14 dias de antecedência no rio Murray em Murray Bridge, South Austrália. As variáveis de entrada para a RNA incluíram salinidade diariamente, fluxo e dados do nível do rio em 16 locais ao longo do rio. Com 60 defasagens, havia, portanto, 960 variáveis de entrada. As técnicas propostas foram comparadas com dois métodos utilizados em estudos anteriores por Maier e Dandy (1996, 1997) para a previsão de salinidade. Os autores descobriram que as novas técnicas levaram a mais parcimoniosa (em termos de número de entradas) modelos RNA e alegaram que os modelos desenvolvidos por meio das novas técnicas tiveram maior capacidade de generalização. No trabalho de Trautmann (1995) afirma-se que “para entradas de poucas dimensões é possível fazer uma avaliação subjetiva dos dados através de mapas gráficos, mas com entradas de muitas dimensões pode haver contraindicações. Essa abordagem não constitui uma abordagem sistemática e objetiva para se montar uma arquitetura ótima ou fazer

a comparação entre vários mapas a partir de diferentes condições de início". Ele também menciona que "pode existir problemas para novos padrões de entrada, já que adaptações em aplicações de monitoramento são necessárias, para quando aparecer um conjunto de treinamento que seja muito dissimilar aos protótipos já existentes". No caso, ele investiu não apenas no modelo SOM clássico, mas também em um modelo incremental que aceitava o aumento do número de nós do mapa para contemplar os novos conjuntos de treinamentos.

Parinet *et al.* (2004) proporcionaram com a utilização do ACP num lago eutrófico, característico de sistemas da Costa do Marfim, a oportunidade para verificar se os valores de todas as variáveis analíticas estão vinculados a causas e efeitos da eutrofização (efeito de validação). Então, nenhum desses valores pode descrever com precisão um estado trófico sozinho. Para resolver esta dificuldade, os autores sugerem que as relações entre as variáveis podem gerar descrições melhores que as próprias variáveis. Onze mil análises foram realizadas durante 23 meses. Os autores mostraram que a Análise de Componentes Principais (ACP) usando coeficientes de regressão linear é uma ferramenta apropriada para essa finalidade.

Çamdevyren *et al*. (2005) utilizaram a ACP para simplificar a complexidade das relações entre variáveis de qualidade da água. Os valores de pontuação obtidos pelascomponentes principais (CP) foram utilizados como variáveis independentes no modelo de regressão. Duas abordagens foram utilizadas na presente análise estatística. Na primeira abordagem, apenas cinco valores de pontuação selecionados obtidos por análise de CP foram utilizados para a previsão dos níveis de clorofila-a. O sucesso preditivo (R) do modelo encontrado foi de 56,3%. Na segunda abordagem, onde todos os valores de pontuação obtidos a partir da análise de CP foram utilizados como variáveis independentes, o poder preditivo foi de 90,8%. Ambas as abordagens poderiam ser utilizadas para prever os níveis de clorofila-a em reservatórios com sucesso.

Diamantopoulou, *et al*. (2005) usaram redes neurais para desenvolver modelos de previsão da qualidade de água com valores mensais com 12 parâmetros (dentre eles: Temperatura, Condutividade Específica, Sódio, Cálcio e oxigênio dissolvido) em um período de 1980-1994 e assim conseguiu preencher os valores faltosos na serie temporal que costumava ser um problema sério nas estações de monitoramento na Grécia. A tabela abaixo mostra o desempenho de cada Rede Neural para a predição dos seis parâmetros de qualidade da água (Nitrato (NO_3^-), Oxigênio Dissolvido (DO), Condutividade Específica (EC), Sódio (Na^+), Cálcio (Ca^{2+}) e magnésio (Mg^{2+})) com os parâmetros de entradas variados. Os

resultados foram satisfatórios, e consequentemente, os modelos de RNA podem ser utilizados para predição de parâmetros de qualidade da água e permitem o preenchimento dos valores faltosos de séries temporais de qualidade da água que é um problema na maioria das estações de monitoramento. Na tabela 08 estão os valores dos Coeficientes de correlação (r), erro absoluto médio (MAE) e raiz do erro quadrático médio (RMSE) para cada modelo de predição.

Tabela 8 - R, MAE, RMSE e as variáveis de entrada para cada RNA para a predição dos seis parâmetros de qualidade da água (Nitrato, Oxigênio Dissolvido, Condutividade Específica, Sódio, Cálcio e magnésio).

	NO_3^- / ANN: 9-19-1/0.9382			ECw / ANN: 10-4-1/0.9744		
Data set	R	MAE	RMSE	R	MAE	RMSE
Total	0.9278	0.962	1.333 (18.67%)	0.9755	18.915	28.377 (5.95%)
Train	0.9382	0.886	1.246 (17.44%)	0.9744	19.119	29.060 (6.09%)
Test	0.8420	1.634	1.941 (27.17%)	0.9863	17.117	21.415 (4.49%)
Inputs	Q, T, ECw, DO, Na^+, Ca^{2+}, SO_4^{2-}, Cl^-, $HCO3^-$			Q, T, pH, Na^+, Mg^{2+},Ca^{2+}, TP, SO_4^{2-}, Cl^-, $HCO3^-$		
	DO / ANN: 9-14-1/0.9188			Na^+ / ANN: 11-24-1/0.9331		
Data set	R	MAE	RMSE	R	MAE	RMSE
Total	0.9103	0.552	0.759 (7.67%)	0.9353	0.035	0.117 (16.65%)
Train	0.9188	0.531	0.735 (7.43%)	0.9331	0.033	0.121 (17.25%)
Test	0.8741	0.736	0.941 (9.52%)	0.9678	0.053	0.069 (9.84%)
Inputs	Q, T, pH, ECw, Na^+, NO_3^-, TP, SO_4^{2-}, Cl^-			Q, T, pH, ECw, DO, Ca^{2+}, NO_3^-, TP, SO_4^{2-}, Cl^-, $HCO3^-$		
	Mg^{2+} / ANN: 8-10-1/0.9640			Ca^{2+} / ANN: 5-10-1/0.9454		
Data set	R	MAE	RMSE	R	MAE	RMSE
Total	0.9660	0.065	0.135 (10.84%)	0.9481	0.223	0.294 (8.90%)
Train	0.9640	0.067	0.145 (11.22%)	0.9454	0.226	0.301 (9.11%)
Test	0.9890	0.046	0.0821 (6.56%)	0.9718	0.193	0.225 (6.81%)
Inputs	pH, ECw, Na^+, Ca^{2+}, NO_3^-, SO_4^{2-}, Cl^-, $HCO3^-$			ECw, Na^+, TP, NO_3^-, SO_4^{2-}		

Fonte: Diamantopoulou, *et al.* (2005)

Outro esforço realizado foi o de Yong (2006), que em seu trabalho avaliou o uso de dois modelos, o perceptron multicamadas (MLP) e o SOM, chegando à conclusão que "ambos os modelos possuíam uma classificação eficiente para o problema proposto", mas nesse trabalho, assim como nos descritos acima, foram usados poucos parâmetros, nesse trabalho foram usados três parâmetros de entrada: o potencial hidrogeniônico (pH), oxigênio dissolvido e temperatura.

O projeto desenvolvido no Centre for Intelligent Environmental Studies (CIES) por Martin (2006) foi realizado com o intuito de analisar a composição biológica das águas. A ferramenta desenvolvida nesse trabalho, o River Biology Monitoring System (RBMS), possui

uma interface amigável e apresenta três tipos de algoritmos, MLP, SOM e Bayes ingênuo, mas, diferentemente dos trabalhos apresentados acima, ela trabalha, apenas, com fatores biológicos.

Strobl *et al*. (2007), que utilizaram diferentes RNAs para classificar o grau de eutrofização de um lago. Quatro arquiteturas de redes neurais artificiais ((i) uma Rede Neural de Regressão (GRNN); (ii) uma Rede Neural Probabilística (PNN); (iii) várias redes neurais de retropropagação; e (iv) uma rede neural polinomial) foram "treinadas" em dados do Eastern Lake Survey, do National Surface Water Survey da Agência de Proteção Ambiental para investigar quais parâmetros físico-químicos são possivelmente de maior importância na determinação do status de eutrofização dos lagos. Dos 110 parâmetros do lago disponíveis, 60 foram escolhidos como entrada para as redes neurais. Os resultados do treinamento com a arquitetura GRNN foram bastante satisfatórios com os coeficientes de determinação múltipla (R^2) para cada classe trófica variando de 0,7614 a 0,9509. Juntamente com os erros quadrados médios (variando de 0,001 a 0,051), indicando que a rede neural "treinada" aprendeu adequadamente o problema dado (ou seja, os fatores individuais de suavização para cada parâmetro de entrada podem ser examinados mais de perto). As várias simulações da rede neural artificial mostraram que, além do fósforo total e do nitrogênio inorgânico, turbidez, condutância específica, elevação do lago e concentração de íon hidrogênio foram identificados como os parâmetros mais significativos que afetam a classificação dos lagos quanto ao estado de eutrofização. Esses resultados sugerem uma associação concebível entre esses parâmetros e a eutrofização do lago, indicando a necessidade de mais estudos sobre essas relações.

No trabalho de Christiano (2007) foi desenvolvido uma rede neural artificial com a finalidade de aprender a dinâmica subjacente à evolução de parâmetros como DBO, a quantidade de oxigênio dissolvido, pH e o nível de coliformes fecais das águas do Rio Jaguaribe, Paraíba. Os resultados mostram que a qualidade das medições influencia decisivamente na capacidade da rede fazer previsões precisas. Quanto maior a qualidade dessas medições, mais precisos podem ser as previsões realizadas por essa técnica. Dados de boa qualidade são necessários, para que se tenha certeza que eles apresentam a realidade da forma mais fiel possível. Sem dados realmente confiáveis, a pesquisa pode ser direcionada para um ponto que não seja realmente o principal problema a ser sanado ou minimizado. Desta forma, as redes neurais construídas provaram poder servir como uma ferramenta, que em conjunto com um monitoramento ambiental cuidadoso, ajude no gerenciamento do

ambiente, permitindo previsões precisas e, consequentemente habilitando o pesquisador a intervir de forma adequada com maior antecipação.

Primpas *et al.* (2007)investigaram o desenvolvimento de um índice multivariável de eutrofização através da análise de componentes principais utilizando nutrientes e valores de clorofila-a, concentrações de fosfato, nitrato, nitrito, amônia do Golfo de Saronikos com uma parcela de dados de ambientes marinhos próximo a Rhodes. A análise dos componentes principais foi aplicada e os coeficientes de nutrientes/clorofila-a do primeiro componente foi utilizado para desenvolver um índice. Um índice multivariado de qualidade da água foi avaliado com uma parcela de dados de nível trófico conhecido e foi considerado eficiente em descrever níveis tróficos.

Wu e Lo (2008), que utilizaram uma RNA do tipo MLP e um modelo adaptativo baseado em rede de sistema de inferência *fuzzy* (NF) para determinação em tempo real da dosagem de coagulante (policloreto de alumínio) a ser utilizada no tratamento da água de superfície do norte de Taiwan. A fim de obter os dados de entrada/saída necessários para desenvolver e validar os modelos RNA e NF foram coletadas amostras de água da ETA no Condado de Taipei, Taiwan. Os parâmetros de qualidade da água desejados foram medidos, incluindo a turbidez, pH, cor. Os 819 pontos de dados foram divididos em dois subconjuntos utilizando o método proposto. Neste trabalho, a raiz quadrada média de erro normalizado (RMSE) é usada como um índice de desempenho para comparar a capacidade de previsão dos modelos treinados por cada conjunto de dados. O coeficiente de correlação de Pearson (r) é usado para comparar a performance relativa dos modelos. O desempenho dos modelos foi considerado suficiente com re RMSE variando respectivamente de 0.8623 e 0.00650 até 0.9256 e 0.00356.

Perendeci *et al.* (2008) propuseram um modelo de sistema adaptativo de inferência neuro-fuzzy (NF) com o uso de variáveis de entrada operacionais disponível on-line e off-line para a estação de tratamento de águas residuais anaeróbia de fábrica de açúcar que opera sob estado instável para estimar a demanda química de oxigênio (DQO) do efluente. Dados de qualidade da água de 192 dias da estação de tratamento foram usados para derivar os modelos. Amostras compostas diárias foram formadas por meio de coleta de amostras da saída da lamela a cada 2 h e dados brutos da análise de DQO e variáveis on-line foram normalizadas entre 0 e 1 estrutura e formatadas no Excel 7.0.A NF foi capaz de estimar o parâmetro de descarga de qualidade da água com sucesso para o caso, quando apenas variáveis on-line limitadas estavam disponíveis sem a necessidade de medição de DQO de

entrada. O coeficiente de correlação de Pearson aceitável (0,8354) e erro quadrático médio baixo (0,1247) foram obtidos para os valores estimados da variável de saída do sistema, no caso o DQO efluente tratado. O modelo pode ser usado no monitoramento e controle do desempenho de plantas de tratamento de águas residuais anaeróbias.

Barbosa (2009) chama a atenção para o fato de o aumento no poder computacional e da crescente quantidade de informações disponibilizada para a linha de pesquisa intiulada de aprendizado de máquinas vem ganhando importância. A mesma está pautada em estudar e desenvolver métodos computacionais para obtenção de sistemas capazes de adquirir conhecimento de forma automática. Segundo os autores, o desafio princípal dos algoritimos de aprendizado é maximizar a capacidade de generalização de seu aprendiz e neste contexto as máquinas de comitê, que são a combinação de mais de uma máquina e aprendizado, pode ser uma alternativa para resolver esse desafio.

Singh *et al.* (2009) estudaram dois modelos baseados em redes neurais artificiais para o cálculo das concentrações do Oxigênio Dissolvido (OD) e da Demanda Bioquímica de Oxigênio (DBO) nas águas do rio Gomti (Índia). Os modelos identificados foram treinados, validados e testados em dados mensais de OD e DBO medidos durante um período de 10 anos. A rede *feedforward* (também chamada MLP) com o algoritmo de aprendizagem *backpropagation* (retropropagação) foi utilizado. As RNAs com diferentes números de neurônios na camada oculta foram construídas, e o desempenho do modelo foi estimado por meio de vários indicadores, incluindo sequência de dados, diagramas de dispersão e medidas quantitativas de RMSE, MAE e R. Os resultados indicaram que a geometria da rede com quatorze neurônios ocultosfoinecessária para um desempenho relativamente melhor de RMSE, MAE e R (30.119 mg.L-1, 16.986 mg.L-1 e 0.841, respectivamente). O estudo mostrou que as redes são ideais para captar as tendências de longo prazo para as variáveis de qualidade da água (OD e DBO), tanto no tempo quanto no espaço. Portanto, para os autores, a rede neural é uma ferramenta eficaz para o cálculo da qualidade da água do rio e também pode ser usado em outras áreas para melhorar a compreensão das tendências da poluição do rio.

Faruk (2010) testou uma aproximação hibrida para séries temporais, formada por um modelo híbrido ARIMA e rede neural que é capaz de explorar os pontos fortes das abordagens tradicionais de séries temporais e redes neurais artificiais. A abordagem proposta consiste em uma metodologia ARIMA e estrutura de rede *feed-forward*, *backpropagation* com um algoritmo de treinamento conjugado otimizado. A abordagem híbrida para previsão de séries

temporais é testada usando observações de 108 meses de dados de qualidade da água, incluindo temperatura da água, boro e oxigênio dissolvido, durante 1996-2004 no rio Büyük Menderes, na Turquia. Os primeiros 72 meses de dados de qualidade da água para boro, oxigênio dissolvido e temperatura da água foram utilizados para o treinamento do modelo. Os 36 restantes meses de dados de qualidade da água foram utilizados para a verificação dos resultados de previsão do modelo. Especificamente, os resultados do modelo híbrido fornecem uma estrutura de modelagem robusta capaz de capturar a natureza não-linear das séries temporais complexas e, assim, produzir previsões mais precisas. Os coeficientes de correlação entre os valores preditos do modelo híbrido e os dados observados para boro, oxigênio dissolvido e temperatura da água são 0,902, 0,893 e 0,909, respectivamente, que são satisfatórios em aplicações de modelos comuns. Os dados de qualidade da água previstos do modelo híbrido são comparados com aqueles da metodologia ARIMA e arquitetura de rede neural usando as medidas de precisão. O resultado do modelo híbrido forneceu um modelo robusto capaz de identificar a natureza não-linear da complexa serie temporal assim produzindo predições mais precisas.

Yan, Zou e Wang (2010) utilizaram um sistema adaptativo de inferência fuzzy baseado em modelo de rede (NF) para classificar o estado da qualidade da água do rio Songhua, rio Liaohe, rio Haihe, rio Huaihe, rio Amarelo, rio Yangtze, rio Pearl, Lago Taihu, lago Chaohu, Lago Dianchi, rio Qiantang e rio Minjiang. Aplicaram vários indicadores físico-químicos e inorgânicos, incluindo oxigênio dissolvido, demanda química de oxigênio, nitrogênio e amônia. Um conjunto de dados (nove semanas com o total de 845 observações) foi coletado de 100 estações de monitoramento em todas as principais bacias hidrográficas da China e utilizados para treinamento e validação do modelo. Comparando o desempenho dos modelos, o modelo NF com função de associação gaussiana teve o melhor desempenho e foi selecionado como o melhor modelo de adaptação. Para todo o conjunto de dados e o conjunto de teste, os maiores valores de coeficiente de correlação de Pearson (0.9665 e 0.9689) e coeficiente de eficiência de Nash-Sutcliffe (0.9340 e 0.9316) e os menores valores da raiz do erro quadrático médio (0.3338 e 0.3704) foram obtidos a partir do modelo NF. Como resultado, o modelo pode prever corretamente 89,59% da qualidade da qualidade do rio, o que demonstrou resultados satisfatórios dessa nova abordagem. Este modelo apresentou melhor desempenho que o modelo RNA e pode gerar valor de saída de forma contínua, o que torna a avaliação da qualidade da água mais compreensível.

Safavi (2010) explorou as capacidades do sistema de inferência adaptativo neuro-fuzzy (NF) para previsões da qualidade da água do rio Zayandehroud da cidade de Isfahan no centro do Irã, com ênfase no oxigênio dissolvido e demanda bioquímica de oxigênio como parâmetros de saída com dados de 16 anos. Pode-se concluir que a velocidade e a profundidade do fluxo desempenham um papel menor como dados de entrada na previsão dos valores de DBO e que o fluxo e a temperatura têm uma sensibilidade mais alta. No estudo de caso realizado no rio Zayandehroud, as previsões de DBO foram obtidas pelo sistema proposto, com um coeficiente de correlação de Pearson de 0,953 no estágio de calibração e 0,931 no estágio de validação, e as previsões de OD foram obtidas com um coeficiente de correlação de Pearson de 0,921 na calibração e 0,904 na etapa de validação. A comparação dos resultados obtidos dos seis diferentes tipos (três modelos para BOD e três para DO) revelou que aumentar o número de parâmetros de entrada não necessariamente aumenta a precisão preditiva do modelo, mas a identificação do efeito de cada parâmetro de entrada e análise de sensibilidade os parâmetros envolvidos desempenham um papel mais importante. A comparação dos resultados fornecidos pelo sistema de inferência adaptativa *neuro-fuzzy* e os valores medidos revelam o alto nível de precisão do modelo proposto.

Noori *et al.* (2010) utilizaram a análise de componentes principais para estratificar as estações de monitoramento e os parâmetros mais importantes em realção a qualidade de águas superficiais do Rio Karoon (no sudoeste do Irã). Parâmetros de qualidade da água incluindo demanda bioquímica de oxigenio (DBO), demanda química de oxigenio (DQO), condutividade elétrica (CE), Nitrato (NO_3^-), Sulfato (SO_4^{-2}), temperatura, Cloreto (Cl^-), oxigênio dissolvido (OD), dureza, Sólidos dissolvidos totais (SDT), pH e turbidez foram medidos em amostras coletadas de 17 estações ao longo do rio Karoon de 1999 a 2002. Neste trabalho, as variações espaciais e as relações entre os parâmetros físicos e químicos foram avaliadas, comprovando que os métodos de ACP podem oferecer uma solução eficaz para a gestão da qualidade da água em casos nos quais se ressalta a complexidade dos dados de qualidade de água.

Primpas *et al.* (2010) desenvolveram um índice simples para classificação do nível trófico do Mar Egeu e comparar com a classificação das Diretrizes da União Europeia para Gestão das Águas, demonstrando que o índice poderia ser utilizado para classificar a qualidade da água em alta, boa, moderada, pobre e ruim. O conjunto de dados, incluindo três subconjuntos, foi coletado das áreas costeiras do Golfo de Saronicos, perto da área metropolitana de Atenas, na Grécia, e do mar alto, perto da ilha de Rodes (648 pontos). A

análise de componentes principais foi aplicada em nitrato, nitrito, amônia, fosfato e clorofila-a em concentrações das águas costeiras do Mar Egeu, Mediterrâneo Oriental e verificou-se que as concentrações de nutrientes inorgânicos que são a causa de eutrofização e concentrações clorofila-a / parâmetros de biomassa fitoplanctônica (o efeito da eutrofização), é a informação adequada para caracterizar os níveis de eutrofização nos corpos d'água. A ACP das cinco variáveis reduziu a dimensionalidade do espaço variável em um componente principal que foi usado para a calibração do índice. O índice de eutrofização resultante foi aplicado com sucesso nas águas costeiras do Mar Egeu.

Talebizadeh e Moridnejad (2011) desenvolveram vários modelos de RNA e NF para prever as flutuações do nível do lago Urmia até seis meses de antecedência, no noroeste do Irã. As variáveis de entrada dos modelos foram a precipitação e evaporação. Os pesquisadores verificaram que os resultados do modelo NF (RMSE=0,08, MAE=0,05 e R^2= 0,99) foram superiores àqueles da RNA (RMSE = 0,14, MAE=0,07 e R^2= 0,99) em que ambos foram mais precisos e com menos incerteza.

Mullai *et al.* (2011) desenvolveram um sistema adaptativo de inferência fuzzy baseado em modelo de rede (NF) para predizer o desempenho de um reator híbrido anaeróbio (AHR) para o tratamento de águas residuais de penicilina-G que foi investigada nas temperaturas ambiente de 30-35ºC para 245 dias em três fases. A qualidade estatística do modelo NF foi significativo devido ao seu alto coeficiente de correlação de Pearson(r) entre experimental e simulado. O r foi encontrado para ser 0,9718, 0,9268 e 0,9796 para as fases I, II e III, respectivamente. Os resultados mostraram que o modelo NF proposto foi bem executado em predizer o desempenho de AHR.

Vilas *et al.* (2011), que utilizaram rede neural artificial do tipo perceptron multicamadas (MLP) para a recuperação de concentração clorofila-a a partir de imagens do Espectrômetro de Imagem de Resolução Média (*Medium Resolution Imaging Spectrometer*-MERIS) sobre a área de estudo. Dois conjuntos de dados de concentração de chla foram utilizados neste estudo. Um conjunto de dados consistiu em medidas analíticas de concentração de clorofila a (chla) fornecidas pelo Instituto Tecnológico para o Controledo Meio Marinho da Galiza (INTECMAR) a partir do programa de monitoramento estabelecido na área de estudo. Contém concentrações chla espectrofluorometricamente determinadas da superfície até uma profundidade de 5 m, para os anos 2002-2006. O segundo conjunto de dados incluiu os resultados de 5 amostragens realizadas em 2007 e 2008 em dias sem nuvens na ria de Vigo e ria de Arousa. Para medições de desempenho, foram utilizados parâmetros

como: coeficiente de correlação de Pearson (r= 0,86), erro médio de previsão (MPE=-0,14), erro médio quadrático (RMSE=0,75), e relativo. Os resultados mostraram que a RNA teve um bom rendimento, com altos R e valores menores de RMSE.

Em Affonso (2011) foi proposto uma metodologia de análise multidimensional dos dados de 22 Unidades de Gerenciamento de Recursos Hídricos (UGRHIs), localizadas no estado de São Paulo, em aproximadamente 136 pontos de coleta, por meio das redes neurais SOM. Estes mapas são usados para estudar e visualizar possíveis correlações entre as diversas variáveis deste banco de dados relativas à análise de compostos inorgânicos e parâmetros físico-químicos referentes à qualidade da água nestas unidades (SOM).

Algumas considerações importantes foram visualizadas no trabalho de Affonso (2011):

- Em estudos aplicados a matrizes ambientais, os mapas auto-organizáveis de Kohonen mostraram-se como uma ferramenta útil, na identificação de correlações conhecidas e não conhecidas, apresentando uma rápida visualização dos resultados;
- Há limitações inerentes ao analista na interpretação de banco de dados multivariados, na compreensão da correlação das variáveis;
- Nos aspectos operacionais, pode-se ressaltar a interatividade da interface do toolbox no estabelecimento da comunicação com outros programas, a linguagem de fácil acesso, e a possibilidade de manipulação da base de dados com agilidade;
- No estudo de classes de dados, o SOM demonstrou a adaptação do resultado aos parâmetros de treinamento definidos, com uma satisfatória representatividade do modelo. Esta metodologia demonstrou sua eficiência e rapidez na análise da base de dados da qualidade da água possibilitando visualizar correlações de forma rápida e dinâmica.

Dastorani e Afkhami (2011) apresenta a aplicação de redes neurais artificiais para prever a seca na estação meteorológica de Yazd. Nesta pesquisa, diferentes arquiteturas de redes neurais artificiais, bem como várias combinações de parâmetros meteorológicos, incluindo média móvel de precipitação de 3 anos, temperaturas máximas, temperaturas médias, umidade relativa, velocidade média do vento, direção do vento predominante e evaporação de 1966 a 2000, foram utilizados como entradas dos modelos. De acordo com os resultados obtidos nesta pesquisa, estruturas dinâmicas de redes neurais artificiais incluindo Rede Recorrente (RN) e Rede Recorrente Temporizada (TLRN) mostraram melhor desempenho para esta aplicação (devido a maior precisão de suas saídas). Finalmente, a rede TLRN com apenas uma camada oculta e a função de transferência de tangente hiperbólica foi

a estrutura de modelo mais apropriada para prever a precipitação média móvel de 3 anos do próximo ano com R de 0,95 e RMSE de aproximadamente 0,05. Em fatos, pela predição da precipitação 12 meses antes de sua ocorrência, é possível avaliar antecipadamente as características da seca. Os resultados indicaram que a combinação de precipitação e temperatura máxima são as entradas mais adequadas dos modelos para obter a maior precisão de saída. Em geral, verificou-se que a RNA é uma ferramenta eficiente para modelar e prever eventos de seca. Os autores afirmam que nas últimas décadas as redes neurais artificiais (RNA) têm se mostrado uma técnica com grande habilidade na modelagem e previsão de séries temporais não lineares e não estacionárias, e, principalmente, na previsão de vários fenômenos onde se pode considerar a qualidade da água estacionária ou como série temporal. Deve-se salientar que a mesma observação pode ser estendida aos sistemas híbridos *Neuro-fuzzy* e a *Support Vector Machine* – SVM.

Garcia (2012) desenvolveu um estudo para que as redes neurais, a lógica *fuzzy* e a análise de componente principal fossem utilizadas como estratégias para avaliação da qualidade da água em corpos hídricos do Estado de Sergipe, com vista à construção de interface fáceis de serem utilizadas em ambiente MatLab. Neste estudo, foram coletados dados ambientais dos reservatórios Jacarecica, da Marcela e da bacia do Rio Poxim, em Sergipe. Para o desenvolvimento da modelagem em termos de redes neurais, foram utilizadas as redes *Multi Layer Perceptron* (MLP) e as redes *Radial Basis Function* (RBF) e um sistema *neuro-fuzzy* para modelar a qualidade da água utilizando como variável de saída a concentração de clorofila-a para caracterizar o fenômeno de eutrofização do sistema. O coeficiente de determinação obteve valores maiores para a rede MLP sem ACP ($R^2 = 0,944$) e uma performance da rede neural no sistema *neuro-fuzzy* mais baixa($R^2= 0,832$), explicada pela qualidade e quantidade de dados utilizados nesse tipo de modelo. Os resultados foram satisfatórios em termos da classificação e descrição do fenômeno de eutrofização entre os níveis de oligotrófico e hipertrófico para corpos hídricos em análise. Dessa forma, as técnicas de inteligência artificial, em particular as redes neurais e a lógica *fuzzy*, foram empregadas com sucesso para um conjunto de dados ambientais, mostrando a viabilidade numérica no que concerne a representação de fenômenos ambientais complexos e importantes para sustentabilidade ambiental dos corpos hídricos.

Wang *et al*. (2013) estudaram um modelo combinado de Transformada *Wavelet* (TW) e Rede Neural Artificial (RNA) para a previsão mensal da série de nitrogênio amoniacal em água de rios para prever a qualidade da água na Estação de Zhushuntun da Região de Harbin,

durante os anos de 2005-2011.Para os modelos de RNA, as séries de dados foram divididas em um conjunto de treinamento (janeiro / 2005-outubro / 2009) e um conjunto de testes (novembro / 2009-outubro / 2011). A WA decompôs séries de tempo originais em subséries diferentes, em que o mais significativo foi escolhido como os dados de treinamento em vez da série original. Em comparação com a RNA tradicional, os modelos WA-RNA foram encontrados mais precisos e confiáveis, com o coeficiente de eficiência do modelo Nash–Sutcliffe (NSE) de 0.7588 e RMSE de 0.0837, do que os únicos modelos RNA na Estação Zhushuntun de Região de Harbin. Os resultados do estudo indicam que WA poderia remover o ruído dos conjuntos de dados originais e o WA-RNA poderia ajudar o tomador de decisão do ambiente a gerir a qualidade da água de forma mais eficaz.

Xu (2013) combinou a transformada *wavelet* com redes neurais para construir um modelo de previsão de qualidade de água e comparou com outros modelos de rede neural e percebeu que ao usar o *wavelet* o erro diminuiu. O modelo foi construído com uma serie temporal com 168 dados horários (entre 21 a 27 de julho no ano de 2010) para prever OD. A precisão da predição foi melhorada grandemente em comparação com a rede neural padrão BP e o Elman neural modelos de rede. O erro percentual absoluto médio caiu de 17,464% e 8,438% para 3,822%.

Marques *et al*. (2013) avaliaram e compararam a aplicabilidade de redes neurais e estatística multivariada na modelagem da qualidade da água utilizando a Clorofila-a como parâmetro de avaliação para os 88 dados de parâmetros ambientais (Profundidade, temperatura da água, temperatura do ar, transparência, condutividade, cor, pH, sólidos totais, sólidos solúveis, oxigênio dissolvido (OD), N-NH4, N-NO2, N-NO3, P-PO4, nitrogênio total e fósforo total) do Reservatório Marcela em Sergipe e 83 dados de 12 parâmetros (temperatura da água, pH, Oxigênio Dissolvido (OD), sólidos totais, salinidade, condutividade, turbidez, %saturação de OD, fósforo total, daphnia e DBO) de reservatórios de seis bacias de Pernambuco. No primeiro teste, com os dados do reservatório de Marcela-SE, o resultado para utilização das redes foi melhor em relação aos dados dos reservatórios em bacias de PE. Este fato é justificado pelo maior número de parâmetros ambientais que influenciam a inferência da clorofila- a que foram disponibilizados para este reservatório. No 1º Teste (Tabela 09), com os dados do reservatório Marcela-SE, a rede RBF com ACP foi a melhor para representar a inferência da clorofila-a, justificadas pelas medidas de desempenho R^2, MPE, MSE e RMSE que são 0,933, 0,054,0,006, 0,076, respectivamente. Porém no 2º Teste (Tabela 10), com os dados de reservatórios em Bacias de PE, foi a rede MLP com ACP

quem melhor determinou, justificadas pelas medidas de desempenho R^2, MPE, MSE e RMSE que são 0,731, 0,092,0,016, 0,126, respectivamente, mostrando a independência de relação entre os conjuntos de dados e o tipo de rede.

Tabela 9 - Medidas de desempenho das redes neurais do 1° teste.

Medidas Desempenho	MLP sem ACP	MLP com ACP	RBF sem ACP	RBF com ACP
R^2	0,912	0,933	0,901	0,933
MPE	0,062	0,057	0,052	0,054
MSE	0,009	0,007	0,007	0,006
RMSE	0,094	0,083	0,084	0,076

Fonte: Marques, *et al*. 2013

Tabela 10 - Medidas de desempenho das redes neurais do 2° teste.

Medidas de Desempenho	MLP sem ACP	MLP com ACP	RBF sem ACP	RBF com ACP
R^2	0,725	0,731	0,693	0,677
MPE	0,093	0,092	0,099	0,103
MSE	0,020	0,016	0,020	0,020
RMSE	0,140	0,126	0,140	0,143

Fonte: Marques, *et al*. 2013

Chang *et al*. (2014) empregaram o sistema adaptativo de inferência baseado no *neuro-fuzzy* na previsão da precipitação de chuva no Reservatório Shihmen da bacia hidrográfica do rio Tahan no norte de Taiwan, antecedentes as advertências de enchentes ou inundações associadas a treze eventos históricos de tufões de 2006-2009, correspondendo a 641 conjuntos de dados horários. No que diz respeito à previsão de chuvas, os resultados demonstraram que o NF alimentou-se com a precipitação assimilada forneceu previsões confiáveis e estáveis, com os coeficientes de correlação de 0,85 e 0,72 para previsão de precipitação com uma e duas horas de antecedência, respectivamente.

Mohammadpour et. al. (2014) utilizaram nesta pesquisa, a Máquina de Vetores de Suporte (SVM) e dois métodos de Redes Neurais Artificiais (RNAs), a saber, *Feedforward Backpropagation* (FFBP) e Função de Base Radial (RBF), foram utilizados para predizer o Índice de Qualidade da Água (IQA) em uma área úmida construída de superfície livre na

Universiti Sains Malaysia (USM). Dezessete pontos da área úmida foram monitorados duas vezes por mês durante um período de 14 meses (de outubro de 2010 a dezembro 2011), e um extenso conjunto de dados (442 dados) foi coletado para 11 variáveis de qualidade da água (temperatura, pH, oxigênio dissolvido, condutividade, sólido suspenso, nitrito, nitrato, nitrogênio amoniacal, demanda bioquímica de oxigênio, demanda química de oxigênio e fosfato). Uma comparação detalhada do desempenho global mostrou que a previsão do modelo de máquina de vetores de suporte (SVM) com coeficiente de correlação de Pearson, r = 0,9984, e erro absoluto médio, MAE = 0,0052, era melhor em relação ao desempenho com redes neurais função de base radial (RBF). Além disso, o resultado fornecido pelo SVM foi comparável com o da rede FFBP (r = 0,9988 e MAE = 0,0044). Esta pesquisa destaca que o SVM e o FFBP podem ser empregados com sucesso para a previsão da qualidade da água em um ambiente de terras úmidas construído em superfície livre. Esses métodos simplificam o cálculo do IQA e reduzem esforços e tempo substanciais, otimizando os cálculos.

Como pode ser observado na literatura consultada os investigadores da área de modelagem, devido ao maior conhecimento e disponibilidade de ferramentas computacionais acessíveis aos trabalhos publicados, começaram a explorar a combinação de diferentes técnicas para a inferência e classificação. Isto é devido ao fato que a eficiência de uma determinada técnica está associada a diversos fatores que dependem da qualidade, quantidade, variabilidade de parâmetros e métricas de divisão de dados para teste e treinamento e a utilização de diferentes técnicas poderia poupar o exaustivo trabalho de tentativa e erro que é exigido quando dados de má qualidade e de reduzida dimensão são utilizados.

BABA *et al.* (2015) foi o primeiro trabalho na literatura consultada que ressalta e define o ensemble em que o principal objetivo é integrar um conjunto de modelos que são usados para resolver diferentes tarefas, de modo a obter um modelo global composto aprimorado, que produz uma precisão e uma estimativa mais confiáveis do que as que podem ser obtidas por meio de um único modelo. Os mesmos autores destacam que os fatores chaves na elaboração de um ensemble são diversidade, estratégias combinadas, número de classificadores ou inferenciadores utilizados, tipos de conjuntos e medidas de desempenho e que deve-se lembrar que esse processo pode resultar em uma grande espaço de armazenamento e tempo computacional. Os autores também reforçam a utilização do R, RMSE e MAE, dentre outros, como medidas favoráveis para a avaliação dos modelos.

No trabalho de Silva (2015) foram aplicadas técnicas de Análises de Componentes Principais (ACP), Redes Neurais Artificiais (RNA), Lógica *Fuzzy* e Sistema *Neuro-fuzzy* para

investigar as alterações da característica da água das barragens de Utinga e do Bita que abastecem de água bruta a ETA Suape é de fundamental importância em função do grande número de variáveis utilizadas para definir a qualidade. Neste trabalho foram realizadas 10 coletas de água em cada área, no período de novembro de 2007 a agosto de 2012, totalizando 120 amostras. Os resultados mostraram uma tendência à degradação da propriedade da água das barragens em decorrência da presença de microrganismos, sais e nutrientes, responsáveis pelo processo de eutrofização, o que se configurou pela maior concentração de fósforo total, *Coliformes termotolerantes*, e diminuição de pH e OD, provavelmente devido à ocorrência de lançamento de efluentes da agroindústria canavieira, industrial e doméstico. A ACP caracterizou mais 76% das amostras permitindo visualizar a existência de mudanças sazonais e uma pequena variação espacial d`água nas barragens. A condição da água das duas barragens foi modelada satisfatoriamente, com razoável precisão e confiabilidade por meio dos modelos estatístico e computacionais; tendo uma quantidade de parâmetros e dados ambientais, que embora limitados, foram suficientes para realização deste trabalho. Ainda assim, fica evidente a eficiência e sucesso da utilização do Sistema *Neuro-fuzzy* (coeficiente de regressão de 0,608 a 0,925) que combina as vantagens das Redes Neurais e da Lógica *Fuzzy* em modelar o conjunto de dados da qualidade da água das barragens de Utinga e Bita.

Mounce (2015) testou o uso de *wavelet* para o monitoramento de vazão, pressão e turbidez para determinar a deterioração da qualidade da água ao longo das redes de distribuição utilizando 288 dados temporais de um mesmo dia. O composto dessas abordagens é aplicado a um extenso conjunto de dados de um sistema de distribuição do Reino Unido, revelando a eficácia da análise pré e pós-lavagem (reduzindo os eventos de descoloração entre 64 e 89%). Com a crescente proliferação de dispositivos de monitoramento e aquisição de dados em tempo real, o potencial para sistemas on-line e o gerenciamento proativo bem informado são aparentes.

Pandhiani e Shabri (2015) combinou transformada *wavelet* discreta, análise dos componentes principais e *Support Vector Machine* (SVM) para estimar valores futuros mensais para a vazão de dois rios no Paquistão. Usando dados mensais obtidos entre 1971 e 2010 para o rio Jhelum e com os obtidos entre 1956 e 2010 para o rio Chenab. O desempenho dos modelos investigados durante este estudo foi medido usando métodos de medição de desempenho bem conhecidos, ou seja, MSE, MAE, R. O modelo combinando as três técnicas (ACP, SVM e *Wavelet*) possui bom desempenho quando comparado ao SVM e a *wavelet* com SVM. O coeficiente de correlação de Pearson (r) para os dados do rio Jhelum e rio Chenab

obtidos pelos modelos SVM é 0,8353 e 0,9457 e pelos modelos *Wavelet* com SVM é 0,9515 e 0,9735 respectivamente, já com as três técnicas em conjunto, o valor r é aumentado para 0,9937 e 0,9991. O MSE obtido pelos modelos SVM é 0,0052 e 0,0021 para os dois conjuntos de dados, respectivamente, e com o modelo em conjunto (*Wavelet* +ACP+SVM), esse valor é reduzido para 0,0002 e 0,0003. Da mesma forma, enquanto o MAE obtido pelo SVM é de 0,0541 e 0,0317, o valor do MAE do modelo *Wavelet* + SVM + ACP é reduzido para 0,0120 e 0,0137. O modelo *Wavelet* + SVM + ACP obteve o melhor valor de MSE e o MAE diminuiu 96% e 86%, respectivamente. O melhor de r aumenta em 19% em comparação com o modelo SVM único para dados Jhelum. Para dados de Chenabo melhor de r aumenta em 5% e o melhor valor obtido para MSE e MAE diminui 78% e 57%, o que mostra uma melhoria substancial na estimativa em comparação com o SVM e *Wavelet*+SVM. Os presentes resultados não apenas retratam o fato de que o modelo *Wavelet* + SVM + ACP é eficiente em termos de tempo, mas também é quase isento de erros, mais confiável e mais estável entre os modelos utilizados.

Alizadeh *et al*. (2015) comparou a utilização das redes neurais artificiais e a conjunção dessas com a transformada *wavelet* na inferência de parâmetros de qualidade de água na Baia de Hilo no Oceano Pacífico. Os autores aplicaram diferentes combinações de parâmetros de qualidade da água assim como diferentes modelos do tipo univariável e multivariável foram aplicados como variáveis de entrada na previsão de valores diários e horários de salinidade da água, temperatura, OD, clorofila, salinidade e turbidez e a combinação destes com a conjunção das *wavelet*s com as redes neurais foi realizada na forma sugerida representando um modelo tradicional bastante apresentado na literatura. Esse modelo consiste basicamente em utilizar as subséries de decomposição *wavelet* do sinal, série temporal original, como dados de entrada nas máquinas de aprendizagem. E como resposta ao treinamento, teremos os dados de sinal original. Os autores reportam que foram utilizados dados diários e horários do período de outubro de 2010 até outubro de 2014. O número de amostras de água selecionadas para os desenvolvimentos de modelos diários e horários é de 650 e 540, respectivamente. Os resultados demonstram que os modelos *Wavelet*+RNA são superiores aos modelos de RNA. Além disso, os modelos horários desenvolvidos para previsão de OD superam os modelos diários de OD. Para os modelos diários, o modelo mais preciso tem coeficiente de correlação de Pearson (r) igual a 0,96, enquanto que para o modelo horário chega a 0,98. No geral, os resultados mostram a capacidade do modelo de monitorar os parâmetros oceânicos, em

condições com dados ausentes ou quando medições e monitoramentos regulares são impossíveis.

Ravansalar (2015) criou modelos utilizando RNA e RNA-*Wavelet* para a previsão de Condutividade Elétrica (CE) no rio Asi na estação de medição de Demirköprü, na Turquia. Ambos os modelos utilizaram 274 conjuntos de dados mensais (1984-2008).Coeficiente de determinação (R^2) foi de 0,949 e 0,381 para os modelos *Wavelet*-RNA e RNA, respectivamente, que resultaram numa performance melhor do modelo *Wavelet*-RNA sobre o RNA assim afirmando que o modelo *Wavelet*-RNA pode ser usado como ferramenta computacional para os parâmetros de qualidade de água.

Li *et al*. (2016) realizaram um estudo comparativo da utilização de diferentes técnicas mais especificamente a Regressão Linear Múltipla (MLR), Redes Neurais Artificiais (BPNN) e *Suporte Vector Machine* (SVM) na inferência do Oxigênio Dissolvido (OD) que é um importante indicador do "estado de saúde" do ecossistema aquático. No estudo foram utilizados variáveis envolvendo fatores físicos, nutriente, substâncias orgânicas e íons sendo eles: Temperatura da Água (TEMP), valores de pH, índice de Permanganato de Potássio ($KMnO_4$), Demanda Bioquímica de Oxigênio (DBO), Nitrogênio Amoniacal (NH_3-N), Nitrogênio Total (Nt), Fósforo Total (TP), Petróleo (PE), Fenol Volátil (VP), Demanda Química De Oxigênio (DQO), Mercúrio (Hg), Chumbo (Pb), Cobre (Cu), Flúor (F), Zinco (Zn), Arsênio (As), Cádmio (Cd), Cromo Hexavalente (Cr), Cianeto (Cyn) e Surfactante Aniônico (LAS), todas que afetam as concentrações de OD. Foram utilizadas 969 amostras coletadas nos rios da China. Para avaliação comparativa da eficiência dos métodos foi utilizado o coeficiente de determinação (R^2), o erro quadrático médio (MSE) e os erros absolutos. Os autores reportam que a SVM foi levemente superior às outras técnicas, com maiores R^2 e menores MSE (Tabela 11).

Tabela 11 - Estatísticas de desempenho de MLR, PSO-BPNN e PSO-SVM durante o treinamento e períodos de teste.

	Treinamento			Teste		
	MLR	**PSO-BPNN**	**PSO-SVM**	**MLR**	**PSO-BPNN**	**PSO-SVM**
R^2	0.20	0.69	0.76	0.22	0.63	0.74
MSE	1.68	1.65	1.64	1.93	1.80	1.62

Fonte: Li *et al*. (2016)

Zhang (2016) desenvolveu um modelo híbrido ARIMA e RBFNN (Redes Neurais do tipo RBF) para previsão de séries temporais de concentração de nitrogênio total e fósforo total para examinar a qualidade da água no lago Chagan, nordeste da China. Os resultados, obtidos através da modelagem com dados mensais entre 2006 e 2011, serviram de base para o monitoramento ecológico no lago Chagan e ajudou a planejar a irrigação. Os resultados revelam o seguinte: (1) as concentrações de Nitrogênio Total (Nt) em Chagan Lake aumentaram ligeiramente de 2006 para 2011, embora as variações anuais de Fósforo Total (Pt) não tenham sido significativas. Os níveis de Nt e Pt foram principalmente classificados como Graus IV e V, (2) Os valores do RMSE híbrido ARIMA e RBFNN para os dados observados e preditos foram 0,139 e 0,036 mg L-1 para Nt e Pt, respectivamente, o que indicou que os modelo híbrido descreve as variações Nt e Pt de forma mais abrangente e precisa do que o modelo ARIMA e RBFNN. Os resultados servem como base teórica para o monitoramento ambiental e ecológico de Chagan Lake e podem ajudar a guiar o planejamento do distrito de irrigação e o projeto de construção de água para a província de Jilin Ocidental.

Hameed *et al*. (2017) chama atenção para que Índices de qualidade da água descrevem as variáveis de qualidade de água num certo ambiente aquático em função do tempo e espaço e que os mesmos são calculados utilizando ainda métodos tradicionais envolvendo demanda de tempo e ainda erros acidentais nos cálculos dos sub índices, assim novas técnicas são requeridas para melhorar a eficiência desses índices. Dentre as técnicas os autores consideram as redes neurais por estas exibirem uma habilidade de capturar os padrões de não linearidades entre predito e preditores. Desta forma os autores compararam duas redes neurais Função de Base Radial (RBF) e o modelo de Rede Neural *back propagation*. Para avaliar a melhor técnica uma avaliação do desempenho e a análise de sensibilidade das variáveis foram realizadas. Segundo os autores, para a Malásia, o Índice de Qualidade da Água é considerado pelo departamento de meio ambiente como função de seis parâmetros sendo eles: Oxigênio Dissolvido (OD), Demanda Bioquímica de Oxigênio (DBO), Demanda Química de Oxigênio (DQO), Sólido Suspenso (SS), Nitrogênio Amoniacal e Valor de pH (pH).O conjunto de dados consistiu em 5233 amostras para período de janeiro de 2001 a outubro de 2010. Esses dados foram divididos em dois conjuntos: treinamento e teste. O trabalho consiste de avaliação de sub índices dessas variáveis e técnicas de inteligência artificial podem ajudar evitando esses cálculos podendo relacionar diretamente as variáveis com os índices. Uma abordagem multicritério foi adotada para avaliar os modelos desenvolvidos, na qual foram utilizados alguns erros de medidas, tais como: Raiz do Erro Quadrático Médio (RMSE),

Coeficiente de Determinação (R^2) e Coeficiente de "Nash–Sutcliffe" (NE). O modelo Rede Neural *back propagation* obteve os melhores resultados de R^2, RMSE e NE (0,7007, 0,0867 e 0,7164, respectivamente) com cinco camadas de entrada de neurônios, seis camadas ocultas de neurônios e um neurônio na camada de saída, enquanto para avaliação de desempenho do modelo RBFNN os melhores resultados de R^2, RMSE e NE são 0,9807, 0,0194 e 0,9803, respectivamente, com valor de spread de 0,6. Portanto, como resultados eles reportam que as redes neurais Função de Base Radial foram melhores.

Csábrági *et al.* (2017) investigaram a utilização das redes neurais no previsão de oxigênio dissolvido em uma seção do rio Danúbio na Hungria. Segundos os autores esse é um dos parâmetros na caracterização das condições de superfície da água. O objetivo do trabalho foi inferenciar este parâmetro em função de parâmetros de fácil medição como pH, temperatura, condutividade elétrica e precipitação, para isso eles testaram modelos lineares e não lineares perfazendo quatro modelos denominados por eles de regressão multivariável linear, Rede Neural *Multilayer Perceptron* (MLPNN), Rede Neural de Função de Base Radial (RBFNN) e rede neural de regressão geral. Os dados utilizados foram relativos a um período de 1998 até 2002, sendo a avaliação do desempenho dos modelos realizada com diferentes medidas estatísticas Erro Quadrático Médio (RMSE), Erro Absoluto Médio (MAE), Coeficiente de Determinação (R^2) e Índice de Concordância de Willmott (IA). Como conclusão os autores ressaltam que de uma forma geral os modelos não lineares foram melhores que os lineares e que em dois casos, a Rede Neural de Regressão Geral forneceu o melhor desempenho, em dois outros casos, a Rede Neural da Função de Base Radial deu os melhores resultados (Tabela 12).

Tabela 12 - Avaliação do desempenho do modelo em conjuntos de dados de treinamento e teste em todas as quatro combinações.

Comb.	Modelo	RMSE ($mg\,L^{-1}$)		MAE ($mg\,L^{-1}$)		R^2		IA	
		Treinamento	Teste	Treinamento	Teste	Treinamento	Teste	Treinamento	Teste
C_A	MLR2	1.14	2.03	0.79	1.38	0.51	0.4	0.82	0.72
	MLPNN	0.65	1.57	0.43	1.28	0.85	0.57	0.93	0.77
	RBFNN	0.84	1.65	0.62	1.25	0.74	0.59	0.92	0.78
	GRNN	0.47	1.42	0.27	1.14	0.93	0.72	0.98	0.87
C_B	MLR2	1.00	1.94	0.75	1.41	0.49	0.44	0.81	0.76
	MLPNN	0.64	1.72	0.46	1.22	0.80	0.58	0.93	0.79
	RBFNN	0.48	1.62	0.37	1.33	0.88	0.54	0.97	0.75
	GRNN	0.35	1.74	0.23	1.27	0.94	0.55	0.98	0.77
C_C	MLR2	1.13	1.57	0.84	1.23	0.43	0.50	0.77	0.72
	MLPNN	0.97	1.46	0.74	1.21	0.58	0.59	0.84	0.72
	RBFNN	0.79	1.43	0.60	1.19	0.72	0.47	0.91	0.75
	GRNN	0.77	1.36	0.56	1.09	0.75	0.60	0.91	0.75
C_D	MLR2	1.31	1.98	0.97	1.41	0.36	0.41	0.72	0.73
	MLPNN	0.96	1.70	0.67	1.21	0.66	0.59	0.88	0.78
	RBFNN	1.11	1.63	0.81	1.17	0.54	0.59	0.83	0.77
	GRNN	0.92	1.70	0.62	1.21	0.70	0.55	0.89	0.77

Fonte: Csábrági *et al.* (2017).

Khan e Chai (2017) investigaram a previsão de qualidade da água representada pelo Indexador de Qualidade da Água (WQI - *Water Quality Index*) construído a partir das variáveis OD, pH, Clorofila, Condutividade, temperatura, alcalinidade, nitrato, turbidez e elevação da superfície da água, utilizando um *ensemble* formado por modelos de Redes Neurais Artificiais (RNA) e adaptativo sistema de inferência *Neuro-Fuzzy* (NF) utilizando a técnica de *ensemble* médio. Os autores reportam que o *ensemble* pode melhorar a precisão da previsão. Os autores utilizaram dados relativos a um ano com amostragem e análises das variáveis realizadas em intervalos de 30 minutos. Na avaliação comparativa, em relação ao desempenho das técnicas, os autores utilizaram o MSE, RMSE e o R^2 (Tabelas 13 e 14). Como conclusão eles reportam que para o conjunto de treinamento e teste dos modelos RNA-NF (RMSE: 0,457 e 0,540) mostram uma melhora significativa no desempenho de previsão em comparação com os modelos individuais RNA (RMSE: 2,709 e 2,682) e o modelo NF (RMSE:1,734 e 1,905). O estudo conclui que o conjunto de modelos de RNA e NF mostra uma melhora significativa no desempenho de previsão em comparação com os modelos individuais. A pesquisa pode revelar-se benéfica para a tomada de decisões em termos de melhoria da qualidade da água.

Tabela 13 - Medidas de desempenho para o treinamento.

	R^2	**MSE**	**RMSE**
RNA	0,972	7,341	2,709
NF	0,988	3,007	1,734
RNA-NF	0,904	0,161	0,457

Fonte: Khan e Chai (2017)

Tabela 14 - Medidas de desempenho para o teste.

	R^2	**MSE**	**RMSE**
RNA	0,973	7,197	2,682
NF	0,986	3,632	1,905
RNA-NF	0,986	0,292	0,540

Fonte: Khan e Chai (2017)

O interessante deste trabalho é que segundo os autores afirmam que, apesar de avanços recentes em modelos de previsão inclusive para evitar o sobre ajuste muito comuns, o desempenho dos modelos pode ser significativamente melhorado se for utilizado um híbrido apropriado de múltiplos modelos do que a utilização de um único. A apredendizagem de *ensembler* refere-se a um proceso de combinação de multiplos preditores para aumentar o desempenho. Ela tem sido usada em várias aplicações, como previsão do consumo de energia (BARAK e SADEGH, 2016) e classificação do câncer (NAGI e BHATTACHARYYA, 2013). No entanto, o uso da apredendizagem de *ensembler* para a previsão da qualidade da água é uma área de pesquisa bastante recente. Um estudo significativo a esse respeito propõe um conjunto homogêneo de modelos de RNA para previsão de parâmetros de qualidade de água (KIM e SEO, 2015), selecionando um modelo ótimo para cada parâmetro de qualidade da água.

Em Li *et al*. (2018) o Mapa de Auto-Organização (SOM) foi usado pelos autores para explorar as características espaciais da qualidade da água na área costeira do sul e centro de Fujian. Dezenove variáveis de qualidade da água (temperatura, salinidade, pH, oxigênio dissolvido, alcalinidade, demanda química de oxigênio, nutrientes NH_4^-N, H_2SiO_3, PO_4^- NO_2^- e NO_3^-, metais pesados / Cu metalóide, Zn, As, Cd, Pb, Hg e Cr^{6+}, e óleo) foram medidos nas camadas de água superficial, média e inferior em 94 locais de amostragem diferentes. Padrões de variáveis de qualidade da água foram visualizados pelos planos SOM, e padrões semelhantes foram observados para aquelas variáveis que se correlacionaram entre si, indicando uma fonte comum. pH, DQO, As, Hg, Pb e Cr^{6+} são provavelmente provenientes de indústrias, enquanto os nutrientes $NH4^-N$, $NO^{2\ -}$ $NO3^-$, e $PO4^{3-}$ foram atribuídos principalmente à agricultura e à aquicultura. O agrupamento *k-means* no SOM agrupou os dados de qualidade da água em nove clusters, que revelou três tipos representativos de água, variando de baixa salinidade a alta salinidade com diferentes níveis de poluição por metais pesados / metalóides e poluição por nutrientes. Mudanças espaciais na qualidade da água refletiram impactos de fatores naturais (vazões ribeirinhas, marés e correntes costeiras), bem como atividades antrópicas (maricultura, descargas industriais e urbanas e efluentes agrícolas). Análise de Componentes Principais (ACP) confirmou os resultados de agrupamento obtidos pela SOM, enquanto o segundo fornece uma classificação mais detalhada e informações sobre as variáveis dominantes que governam os processos de classificação. Os resultados deste estudo sugerem que a SOM é uma ferramenta eficaz para um melhor entendimento dos padrões e processos que conduzem a qualidade da água.

Djerioui *et al*. (2018) investiga motivado por considerar que um das maiores dificuldades em medidas de monitoramento das análises e medidas com aparelhos das variáveis relacionadas à composição da água por várias razões como custo dos sensores, o tempo desprendido para checar, limpar, calibrar e trocar fazem dessa operação muito complicada aliada ainda a questão do monitoramento de qualidade de água exigir um grande número de sensores, assim técnicas de inferência como os Soft sensores podem ser interessante. Dessa forma os autores investigam a utilização das técnicas Máquina de Vetores de Suporte (*SVM*), *Máquinas* de *Aprendizado* Extremo (*ELMs*), técnicas em termos de tempo de aprendizado e alguns parâmetros de erro de regressão e classificação são apresentados como critérios de eficiência. Foram utilizados 774 amostras com as variáveis: temperatura, condutividade, logaritmo negativo da concentração de íon hidrogênio na água, oxigênio dissolvido, sólidos suspensos, demanda bioquímica de oxigênio, cálcio, cloro, magnésio e bicarbonato. Como critério de avaliação foi utilizado o Erro Absoluto Médio (MAE), Erro Quadrático Médio (MSE), Raiz do Erro Quadrático Médio (RMSE) e o coeficiente de correlação de Pearson (r). Como conclusão os autores reportam uma leve vantagem do ELM em relação ao SVM, com valores menores de MAE, MSE, RMSE para treinamento e teste, apresentando coeficiente de correlação de Pearson (r) de 0,86 no teste e 0,99 no treinamento (Tabelas 15 e 16). Na questão de tempo de treinamento, a técnica de ELM também apresenta uma pequena vantagem, porém isso é relativo ao software utilizado, assim como às configurações da máquina utilizada.

Tabela 15 - Valores de MAE, MSE, RMSE e R para as técnicas de SVM e ELM no treinamento dos dados.

	MAE	**MSE**	**RMSE**	**r**
SVM	$9x10^{-4}$	$7x10^{-5}$	$8x10^{-3}$	0,99
ELM	$1,7x10^{-5}$	$4x10^{-8}$	$2x10^{-3}$	0,99

Fonte: Djerioui *et al*. (2018)

Tabela 16 - Valores de MAE, MSE, RMSE e R para as técnicas de SVM e ELM no teste dos dados.

	MAE	**MSE**	**RMSE**	**r**
SVM	1,19	4,25	2,06	0,51
ELM	0,011	7x10-4	0,027	0,86

Fonte: Djerioui *et al*. (2018)

Haghiabi *et al*. (2018) investigaram a eficiência de diferentes técnicas de inteligência artificial incluindo as Redes Neurais Artificiais (RNA), Método de grupo de manipulação de

dados (GMDH) e Máquina de Vetores de Suporte (*SVM*) para predizer parâmetros de qualidade de um rio localizado no sudoeste do irã. A medição das amostras dos componentes de qualidade da água foi realizada mensalmente. Vale ressaltar que na maioria dos meses, várias medidas foram registradas. A linha do tempo da amostragem é de mais de 55 anos. A primeira medida foi relatada em 1960 e até hoje, o monitoramento deste rio ainda está em andamento. Os parâmetros medidos foram: pH, Condutividade Específica (EC), Bicarbonato (HCO_3), Sulfatos(SO_4), Cloretos (Cl), Sólidos Totais Dissolvidos (TDS), Sódio (Na), Magnésio (Mg), Cálcio (Ca) e nove arranjos de modelos foram avaliados um para cada variável inferida. Para a avaliação do desempenho os autores utilizaram o RMSE, R^2 e um índice denominado DDR encontrado unicamente nesse trabalho que é definido como a razão entre o valor predito e o valor observado menos um. Os autores concluíram que em relação à precisão a do GMDH foi levemente menos que a do RNA e SVM sendo o melhor desempenho das SVM, com menor valor de RMSE e maiores valores de coeficiente de determinação (R^2) (Tabela 17).

Tabela 17 - Resumo do desempenho dos modelos aplicados para previsão do componente de qualidade da água.

		GMDH			SVM			RNA		
N°	Saída		R^2	RMSE		R^2	RMSE		R^2	RMSE
1	Ca	Train	0.83	0.333	RBF	0.96	0.160	Tansig	0.92	0.238
		Test	0.85	0.313		0.94	0.193		0.84	0.295
2	Cl	Train	0.87	0.297	RBF	0.97	0.147	Tansig	0.94	0.193
		Test	0.88	0.312		0.95	0.210		0.96	0.178
3	E_C	Train	0.97	29.71	RBF	0.97	26.35	Tansig	0.98	23.88
		Test	0.99	13.84		0.98	28.81		0.96	35.14
4	HCO_3	Train	0.87	0.355	RBF	0.96	0.190	Tansig	0.96	0.192
		Test	0.89	0.334		0.95	0.219		0.92	0.290
5	Mg	Train	0.76	0.345	RBF	0.93	0.184	Tansig	0.92	0.197
		Test	0.79	0.332		0.93	0.199		0.90	0.212
6	Na	Train	0.78	0.317	RBF	0.94	0.157	Tansig	0.85	0.250
		Test	0.77	0.326		0.93	0.186		0.86	0.251
7	SO_4	Train	0.36	0.321	RBF	0.77	0.190	Tansig	0.74	0.185
		Test	0.26	0.297		0.68	0.204		0.68	0.240
8	TDS	Train	0.97	19.34	RBF	0.98	16.40	Tansig	0.98	16.60
		Test	0.99	7.22		0.97	22.31		0.97	19.39
9	pH	Train	0.25	0.358	RBF	0.59	0.270	Tansig	0.29	0.33
		Test	0.22	0.361		0.33	0.340		0.29	0.355

Fonte: Haghiabi *et al*. (2018)

Chou *et al*. (2018) determinaram o Índice de Estado Trófico de Carlson através de parâmetros de qualidade da água (temperatura da água superficial, Demanda Bioquímica de Oxigênio (DBO), Sólido Suspenso (SS), Demanda Química de Oxigênio (DQO) e Amônia (NH_3)) em estações de monitoramento de 20 reservatórios em Taiwan, entre os anos de 1995 e 2016, utilizando quatro técnicas de inteligência artificial, dentre elas as Redes Neurais

Artificiais (RNA) e Máquina de Vetores de Suporte (SVM). Uma interface amigável que integra um modelo de regressão metaheurística foi desenvolvida para avaliar o desempenho preditivo e compará-lo com os dois cenários constituintes. A comparação compreensiva demonstrou que o modelo RNA em conjunto, baseado em um método de escalonamento, obteve RMSE, MAE e MAPE mais baixos (3,941, 3,131 e 6,786%, respectivamente) e SI = 0,250,o que o torna mais preciso do que os outros modelos sozinho, em conjunto e o modelo de regressão metaheurística híbrida. Tanto a precisão da previsão quanto a eficácia da aplicação são consideradas para apoiar os profissionais no planejamento de obras de gerenciamento de água. Assim, este estudo fornece uma nova abordagem para uso potencial na avaliação da qualidade da água.

Ao longo da realização da presente tese observaram-se na literatura diversos trabalhos que tiveram objetivos de aplicar as diferentes técnicas, da inteligência artificial e de análise estatística multivariada para descrever o comportamento dos mais variados ambientes hídricos. Deve-se ressaltar que a aplicabilidade delas esta relacionada também as ferramentas utilizadas, pois os pacotes computacionais que disponibilizam muitas das técnicas aqui citadas são pagos, trazendo inclusive algumas limitações. Neste sentido, as análises comparativas dentre as diferentes técnicas citadas aqui não estão sendo menos exploradas atualmente, pois a eficiência de cada técnica está sujeita a diversos paramentos como qualidade, quantidade, sendo os esforços direcionados para a utilização das mesmas na forma de *enseble* ou comité de máquinas que é a utilização de diversas técnicas em conjunto na modelagem de um determinado sistema utilizando-se de uma métrica como, média aritmética do resultado de cada técnica, por exemplo, para modelagem (Zhang, 2016; Mounce, 2015; Alizadeh, 2015; Ravansalar, 2015). Desta forma, optou-se neste trabalho por explorar a aplicabilidade de diferentes técnicas para análise dos dados (ACP e SOM), As técnicas RNA's, SVM, e *Neuro fuzzy* como modelos de inferência para os dados estacionários e conjunção dessas com a transformada *wavelet* discreta para o prognóstico de séries temporais com número reduzido de amostras. Embora uma análise comparativa esteja sempre relativizada pela natureza dos modelos propostos, a mesma foi realizada em algumas situações.

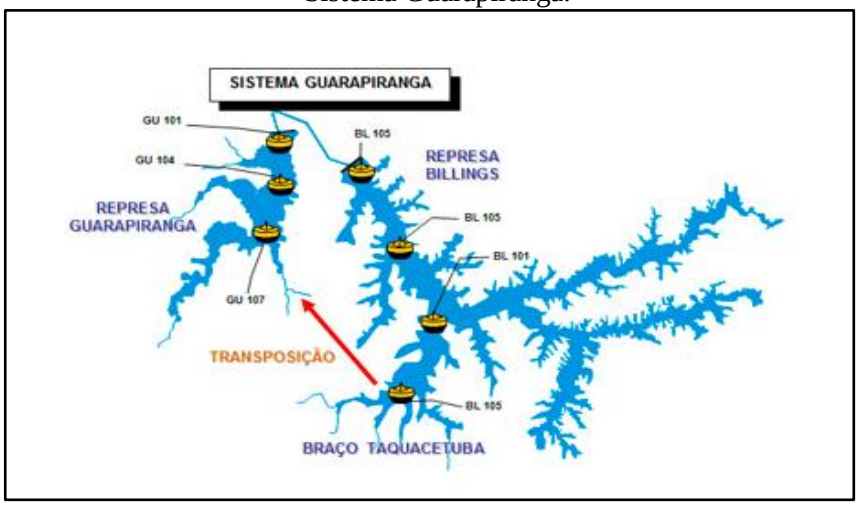
3 MATERIAIS E MÉTODOS

Foram realizadas pesquisas constantes em diversos bancos de dados tanto no Brasil como em outros países a fim de obter dados suficientes para a realização do projeto. Através dessas pesquisas escolheram-se os dados do Brasil, da Espanha e da China para realizá-lo.

3.1 DADOS REAIS DE RESERVATÓRIOS DE ÁGUA UTILIZADOS

Os dados utilizados do Brasil são referentes à represa de Guarapiranga, o qual é o segundo maior sistema de água potável da Região Metropolitana de São Paulo, localizado nas proximidades da Serra do Mar. Seu manancial é formado pelos rios Embu-Mirim, Embu-Guaçu, Santa Rita, Vermelho, Ribeirão Itaim, Capivari e Parelheiros. As represas de Guarapiranga e Billings produzem 15 mil litros de água por segundo e abastece 4,9 milhões de pessoas das Zonas Sul e Sudoeste da Capital. Possui uma área de 26,6 km^2 e a área da bacia é de 631 km^2, tem uma capacidade máxima de 171 bilhões de litros.

Figura 28 - Meio do corpo das Represas Guarapiranga e Billings – Sistema Guarapiranga.



Fonte: Sabesp (Companhia de Saneamento Básico do Estado de São Paulo).

A análise dos mananciais é feita por sondas instaladas em Unidades de Monitoramento Remoto (UMR), a GU 104 foi utilizada para dados estacionários enquanto a GU 107 para dinâmicos. O monitoramento realiza análises a cada 30 min a uma profundidade de 0,3m, diretamente na represa. A Sabesp é a empresa que está responsável por tal monitoramento. Foram disponibilizadas para este trabalho 512 amostras em 179 dias, com séries diárias e horárias, nos quais os parâmetros analisados são: Condutividade, Oxigênio Dissolvido (OD), pH, Potencial de Oxi-redução (ORP), Temperatura e Turbidez.

O Rio LamTsuen é um rio chinês que atravessa o distrito de Tai Po, uma cidade do sul de Hong Kong na China. A nascente do rio localiza-se a 750 m acima do nível do mar, tem uma grande bacia de cerca de 18,5 Km^2, com volume médio anual total de 14,6 milhões de metros cúbicos e comprimento de 12 km.

Figura 29 - Rio LamTsuen, Hong Kong.

Fonte: Google.

Foram disponibilizadas para este trabalho 288 amostras do Rio LamTsuen, em Hong Kong, nos quais os parâmetros analisados são: Temperatura, Sólidos suspensos, Sólidos totais, Fósforo, Nitrito, Nitrato, OD (mg.L-1 e porcentagem de saturação), Condutividade, Demanda Química de Oxigênio (DQO), Amônia e Demanda Bioquímica de Oxigênio (DBO).

Os dados utilizados da Espanha são referentes ao reservatório de Trasona nas redondezas de Astúrias, na Espanha. Seu manancial é composto pelos rios Alvares e Narcea. Volume de 31 mil metros cúbicos e capacidade total de 41 hm^3, sua bacia tem área de capitação de 37 km^2.

Figura 30 - Reservatório de Trasona - Espanha.

Fonte: Fernández, *et al*. (2013).

Foram disponibilizadas 151 amostras do Reservatório de Trasona, Espanha, e entre os parâmetros analisados no manancial do Reservatório de Trasona estão cianotoxinas, clorofila, condutividade, alcalinidade, turbidez, fosfato, amônia, nitrito, cálcio, entre outros.

3.2 ANÁLISE EXPLORATÓRIA DOS DADOS E MODELAGEM

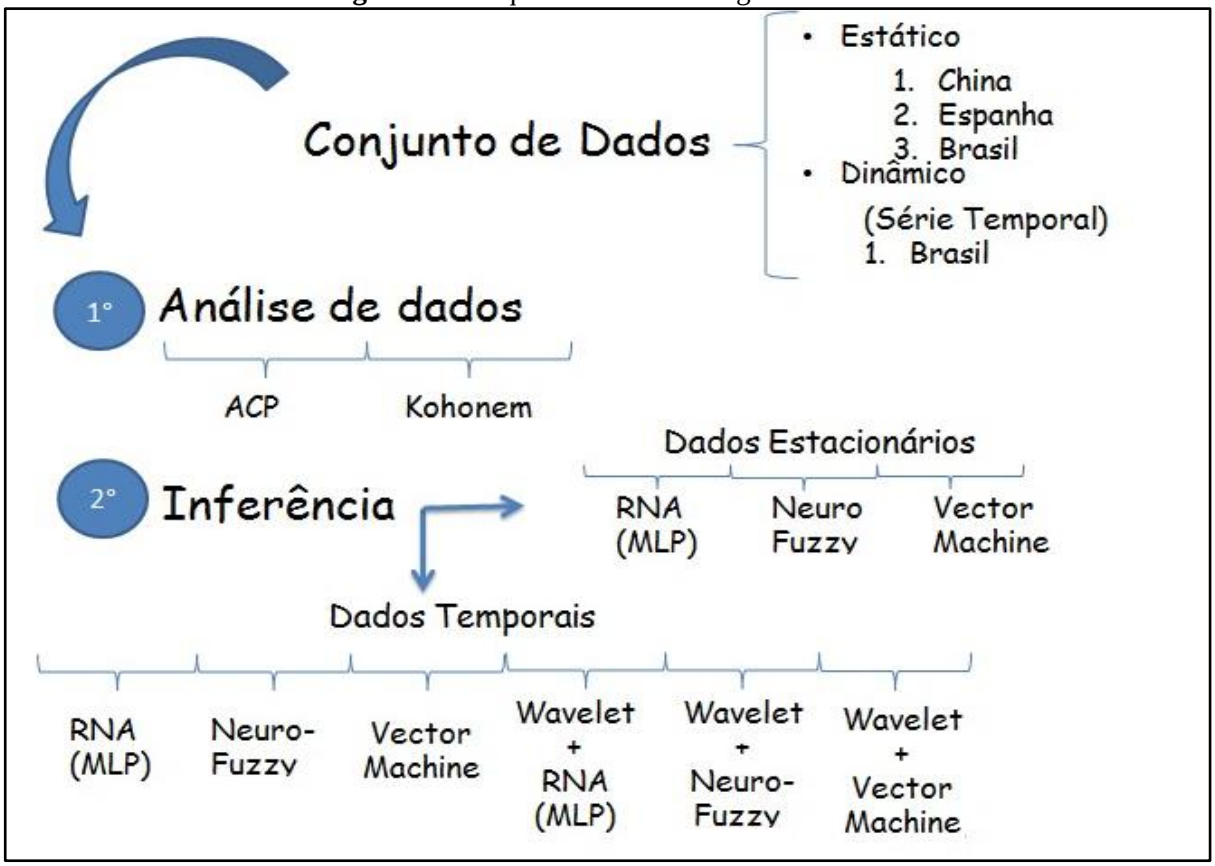
A metodologia utilizada no presente trabalho teve início por uma análise exploratória de dados, baseada em técnicas de estatísticas multivariadas de Análise de Componentes Principais (ACP) e de Redes Neurais Kohonen (SOM). Os parâmetros físico-químicos da água ao longo do ano foram modelados para prever parâmetros de qualidade global, tais como clorofila, Demanda Bioquímica de Oxigênio (DBO) e cianobactérias, com base em modelos de inteligência artificial de Redes Neurais do tipo *Perceptron* Multicamadas (MLP) e *Neuro-Fuzzy*, e modelos estatísticos do tipo Máquina de Vetores de Suporte (SVM). Em seguida os dados do manancial brasileiro foram utilizados numa modelagem de uma série de dados temporais para previsão futura em curto período de tempo da concentração de clorofila como parâmetro global de qualidade, com base na transformada de *wavelet* aplicada aos modelos de redes neurais e estatísticos.

Os dados temporais necessários para a realização do trabalho foram obtidos através de uma revisão bibliográfica que resultou em um conjunto de dados, disponibilizados pela Companhia de Saneamento Básico do Estado de São Paulo (Sabesp), formando uma série temporal com dados horários de 179 dias, utilizados tanto na forma horária quanto na forma de médias diárias, com o intuito de desenvolver estratégias na forma de ferramentas

computacionais construídas em ambiente MatLab que permitam a avaliação e a previsão a curto e longo prazo de variáveis de qualidade de água na forma de séries temporais que estejam direta e indiretamente envolvidas na qualidade da água. As estratégias desenvolvidas utilizarão técnicas de tratamento de sinais, como a transformada *Wavelet* em conjunção com técnicas de sistemas inteligentes tais como Redes Neurais Artificiais.

A metodologia descrita anteriormente pode ser melhor compreendida através do esquema ilustrado na Figura 31.

Figura 31 - Esquema da Metodologia utilizada.



3.2.1 Medidas de Desempenho

A fase de validação, o ajuste do modelo é certificado de acordo com um conjunto de parâmetros (medidas de desempenho), que compara, por exemplo, a concentração de clorofila-a observada e a obtida pelo modelo. Estas medidas de desempenho são obtidas por meio da tentativa e erro e utilizadas em estudos como forma de avaliação dos modelos propostos. A Tabela 18 mostra os parâmetros e como são calculados.

Tabela 18 - Medidas de desempenho para avaliação dos modelos.

Parâmetro	Equação
Coeficiente de Correlação de Pearson(r)	$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{\left[n\sum x^2 - (\sum x)^2\right]\left[n\sum y^2 - (\sum y)^2\right]}}$
Erro de Previsão Médio (MPE)	$MPE = \frac{1}{N}\sum_{i=1}^{N} PE_i$
Erro Quadrático Médio (MSE)	$MSE = \frac{\sum_{i=1}^{N} PE_i^2}{N}$
Raiz do Erro Médio Quadrado (RMSE)	$RMSE = \sqrt{MSE}$

Sendo: N = o número de dados; x = parâmetro de entrada; y = parâmetro de saída;
PE_i (Erro de previsão) = O_P (valor previsto) – O_R (valor real).
Fonte: Vilas *et al.* (2011) e Wu e Lo *et al.* (2008).

O coeficiente de correlação de Pearson (r) compara o desempenho do modelo com a de um modelo de referência, a saída do que é a média de todas as amostras (WU e LO, 2008). Pode, portanto, ser usado para comparar o desempenho relativo dos diferentes modelos de saídas.

Com o Erro de Previsão Médio (MPE) foi possível descobrir se um modelo tende a subestimar (altos valores positivos) ou superestimar (altos valores negativos) as concentrações de clorofila-a observados.

A raiz do Erro Médio Guadrado (RMSE) é usada como um índice de desempenho para comparar a capacidade de previsão de RNA treinada por cada conjunto de dados. O RMSE é conhecido por ser descritivo, quando a capacidade de previsão entre os modelos é comparada (WU e LO, 2008).

4 RESULTADOS E DISCUSSÕES

Com o intuito de prever parâmetros de qualidade de água em corpos hídricos aplicaram-se diferentes técnicas de modelagem matemática empírica.

4.1 ANÁLISE DOS DADOS

As técnicas de redes neurais Kohonen (SOM) e Análise de Componentes Principais (ACP) foram utilizadas neste estudo para uma maior compreensão da qualidade dos dados e relação entre variáveis e amostras. Desta forma para os conjuntos de dados levantados neste trabalho foi aplicada a ACP com o objetivo de maior compreensão das relações entre amostras e variáveis com os objetivos de obter a variável de maior importância no fenômeno investigado e a possibilidade dos dados apresentarem agrupamentos. A técnica SOM foi utilizada com o objetivo de identificar as variáveis e indicar sua influencia no modelo. Desta forma, é possível indicar variáveis que poderão ser dispensadas para a criação do modelo preditivo. Vale salientar que uma análise comparativa do ACP e SOM também foi realizada para cada conjunto de dados na análise de agrupamento na busca de erros grosseiros.

4.1.1 Análise e pré-tratamento estatístico de dados

Todos esses dados foram organizados em diferentes planilhas, devido a tratar de dados distintos, para criar diferentes séries temporais, para então prosseguir com o passo seguinte.

Através do ambiente MatLab foi realizada a retirada de erros grosseiros, *outliers*, e posteriormente feito os tratamentos estatísticos para os parâmetros de cada manancial analisado, sendo eles média ($\overline{X}$), Mediana (Md), Moda (Mo), Valor Mínimo (Min), Valor Máximo (Max), Alcance (Alc), Desvio Padrão (SD), Curtose (B), Simetria (Y) e Coeficiente de Variação (V). Desse modo, produziram-se as Tabelas 19, 20 e 21, mostradas a seguir.

Tabela 19 - Tratamentos estatísticos para as 512 amostras de dados do Brasil.

	$\overline{}$	Md	Mo	Mini	Max	Alc.	SD	B	Y	V
Condutividade	142,653	142,182	137,042	115,229	176,021	60,792	11,953	3,231	0,531	0,084
OD	7,009	7,045	7,694	2,69	10,328	7,637	1,476	2,723	-0,191	0,211
pH	7,572	7,432	9,13	6,397	9,473	3,076	0,664	2,769	0,664	0,088
ORP	169,606	138,386	53,673	53,673	511,696	458,023	90,829	8,247	2,339	0,535
Prof.	0,329	0,3	0,3	0,156	0,568	0,412	0,072	2,883	0,318	0,219
Temp.	22,719	23,253	24	17,217	29,279	12,063	2,745	2,072	-0,076	0,121
Turbidez	7,24	2,942	1,333	0,22	216,494	216,274	26,319	44,362	6,423	3,635
Clorofila	27,328	25,944	2,698	2,698	68,385	65,688	13,787	2,667	0,58	0,504

Tabela 20 - Tratamento estatístico para as 151 amostras de dados da Espanha.

	$\bar{}$	Md	Moda	Min	Max	Alc.	SD	B	Y	V
Cianotoxina	93,404	0	0	0	1871	1871	318,582	21,243	4,243	3,411
Clorofila	706,677	22	5	2	7563	7561	1694,432	10,563	2,858	2,398
MA	1,603	0,276	0,004	0,001	11,05	11,049	2,584	6,02	1,971	1,611
WN	1,280	0,081	0,004	0,001	12,45	12,449	2,636	7,768	2,371	2,06
MAANDWN	7,092	0,023	3x10	2x10	64,35	64,35	16,401	6,999	2,323	2,313
Outras espécies de Cianotoxina	0,676	0,183	0,039	0	6,989	6,989	1,170	12,018	2,827	1,73
Diatos	3,009	2,284	0	0	20,884	20,884	3,003	10,76	2,168	0,998
Crisófitas	0,219	0,183	0	0	0,914	0,914	0,182	3,791	1,018	0,831
Clorofilas	0,072	0,061	0	0	0,21	0,21	0,056	1,907	0,38	0,779
Outras Espécies de fito plâncton	2,904	2,224	0	0	15,297	15,297	2,721	6,456	1,615	0,937
WTEP	13,511	13,1	11	8,3	21,7	13,4	3,144	2,147	0,285	0,233
AMTEM	16,942	16,3	11	7,7	28,6	20,9	4,925	2,11	0,268	0,291
Profundidade Secchdepth	2,311	2,4	2,9	0,5	4,1	3,6	0,916	1,989	-0,03	0,396
Turbidez	6,494	4	2,2	1,4	27,1	25,7	5,836	5,762	1,818	0,899
TP	0,066	0,055	0,044	0,015	0,13	0,115	0,032	1,923	0,411	0,488
Fosfato	0,110	0,101	0,077	0,033	0,221	0,188	0,039	3,063	0,675	0,353
TN	0,610	0,6	0,5	0,5	0,88	0,38	0,11	1,921	0,498	0,18
Nitrato	0,594	0,55	0,5	0,5	1,1	0,6	0,126	4,166	1,353	0,212
Nitrito	0,002	0,002	0,002	0,002	0,01	0,008	0,001	51,141	6,53	0,425
Amônia	0,118	0,11	0,05	0,04	0,26	0,22	0,056	1,916	0,32	0,474
OD	9,781	9,3	8	6,6	14,3	7,7	1,797	2,459	0,499	0,184
Condutividade	293,960	288	277	244	377	133	24,824	3,893	1,05	0,084
Alcalinidade	1,617	1,65	1,8	1,22	1,9	0,68	0,181	2,365	-0,612	0,112
Cálcio	41,533	41,4	50	27,7	51,8	24,1	5,320	2,265	-0,125	0,128
pH	8,519	8,5	9	7,6	9,9	2,3	0,525	2,373	0,275	0,062

Tabela 21 - Tratamentos estatísticos para as 288 amostras de dados da China.

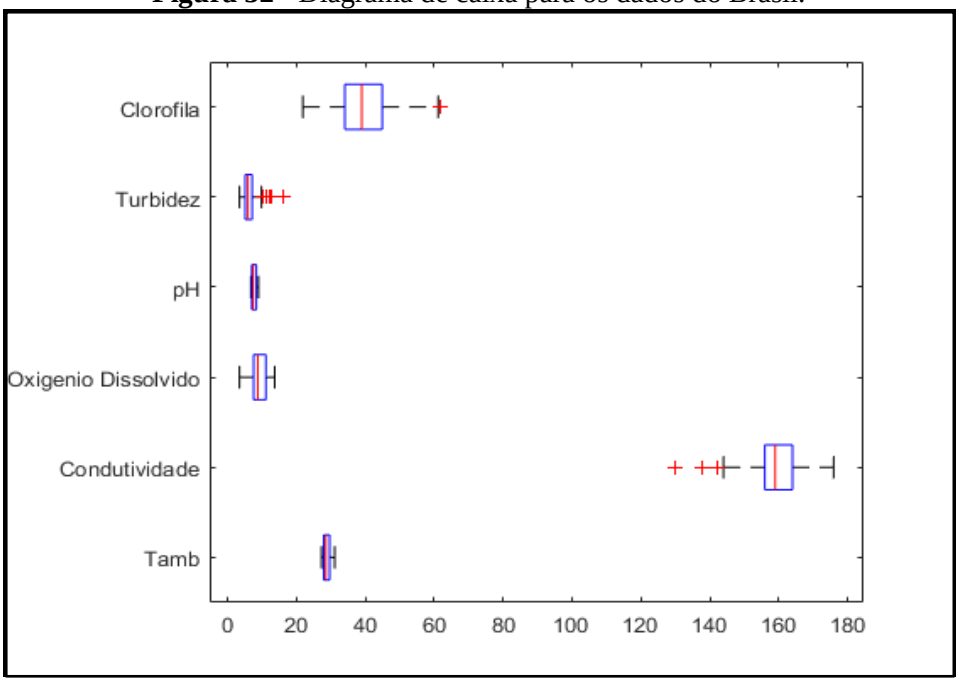
	$\bar{}$	Md	Mo	Min	Max	Alc.	SD	β	Y	V
Temperatura da água (°C)	23,318	23,4	28,2	13,2	30,7	17,5	4,386	1,918	-0,272	0,188
Sólidos suspensos (mg.L^{-1})	5,626	2,1	0,5	0,5	70	69,5	9,305	17,545	3,447	1,654
Sólidos (mg.L^{-1})	95,896	84	110	36	390	354	46,596	16,565	2,968	0,486
Fósforo (mg.L^{-1})	0,0814	0,05	0,04	0,02	0,65	0,63	0,084	15,672	3,098	1,027
Nitrito (mg.L^{-1})	0,024	0,006	0,002	0,002	0,66	0,658	0,064	54,65	6,385	2,641
Nitrato (mg.L^{-1})	0,645	0,64	1,2	0,026	2	1,974	0,404	3,099	0,495	0,626
OD (mg.L^{-1})	8,545	8,5	7,8	6,4	10,4	4	0,838	2,374	0,162	0,098
OD (%saturação)	99,604	100	102	78	121	43	5,446	5,426	-0,746	0,055
Condutividade (µs/cm)	105,783	82	64	29	630	601	74,526	18,087	3,157	0,704
DQO (mg.L^{-1})	4,146	3	2	2	22	20	2,581	12,207	2,249	0,623
Amônia-nitrogênio (mg.L^{-1})	0,184	0,042	0,018	0,009	3,1	3,091	0,417	20,184	3,875	2,265
Demanda bioquímica de oxigênio do (mg.L^{-1})	0,985	0,6	0,1	0,1	8,6	8,5	1,213	12,6	2,815	1,23

4.1.2 Análise de Agrupamentos e Kohonennos dados do Brasil

Inicialmente os dados do Brasil possuíam oito variáveis, mas esse número foi reduzido a seis para que se faça possível a comparação entre estes dados e os dados da Espanha.

Com os dados obtidos foi possível fazer o diagrama de caixa (Figura 32) para permitir a visualização do intervalo entre os dados para cada variável de tal forma que se faz possível identificaras variáveis com maior variação e assim verificar a importância de cada variável no valor final de clorofila.

Figura 32 - Diagrama de caixa para os dados do Brasil.

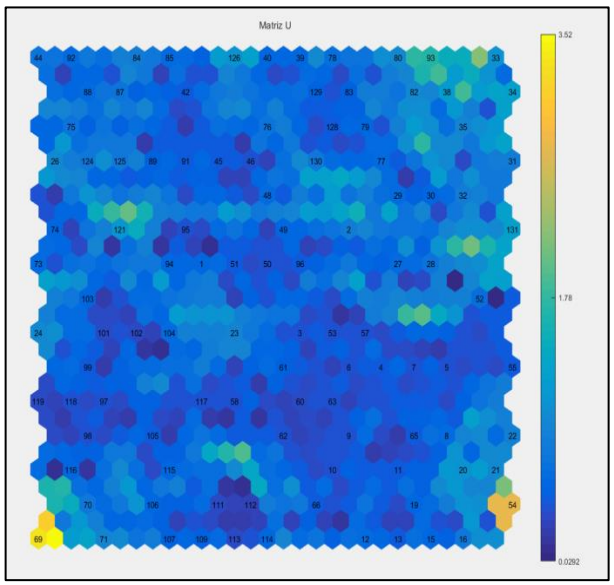


4.1.2.1 Kohonen na análise de agrupamentos do Brasil

A técnica de mapas auto-organizáveis é uma forte ferramenta para agrupar padrões de um espaço multidimensional em um espaço de menor dimensão, geralmente duas dimensões. Também pode ser usado para agrupamento (*clustering*), classificação, estimativa e previsão. Pode até obter performance melhor que métodos de análise atuais pois este sucede em lidar com não-linearidades do sistema, é desenvolvido por dados sem requerer conhecimento profundo do sistema, lida com dados irregulares e com ruídos, e é facilmente atualizado e interpreta informação de múltiplas variáveis e parâmetros. Esse método possui excelente capacidade de visualização que pode ser de grande ajuda nos passos iniciais da determinação da qualidade de água.

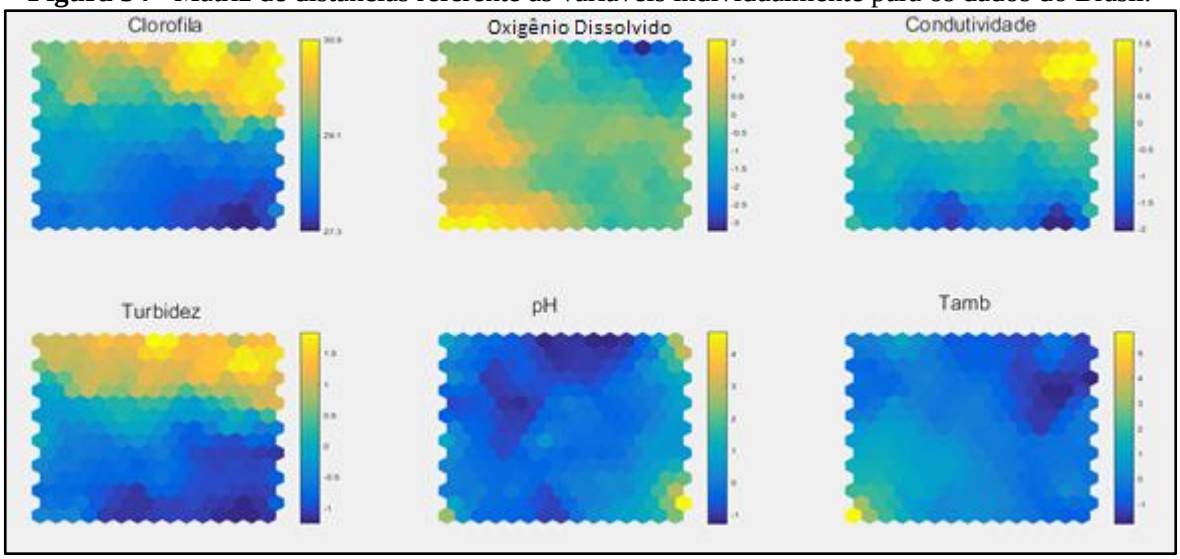
Na Figura 33 (MATRIZ U), a matriz de distâncias e o mapa de características mostram que os dados de operação normal formam um agrupamento uniforme, o esquema de cores indica que existem pequenos valores de distâncias entre algumas amostras.

Figura 33 - Matriz U para os dados do Brasil.



Já na matriz de distâncias (Figura 34) referente às variáveis individualmente, o esquema de cores revela certo grau de dissimilaridade dentro dos grupos normais. No entanto, esta dissimilaridade apresenta caráter uniforme e esta uniformidade indica uma falha devido a um número reduzido de variáveis.

Figura 34 - Matriz de distâncias referente às variáveis individualmente para os dados do Brasil.

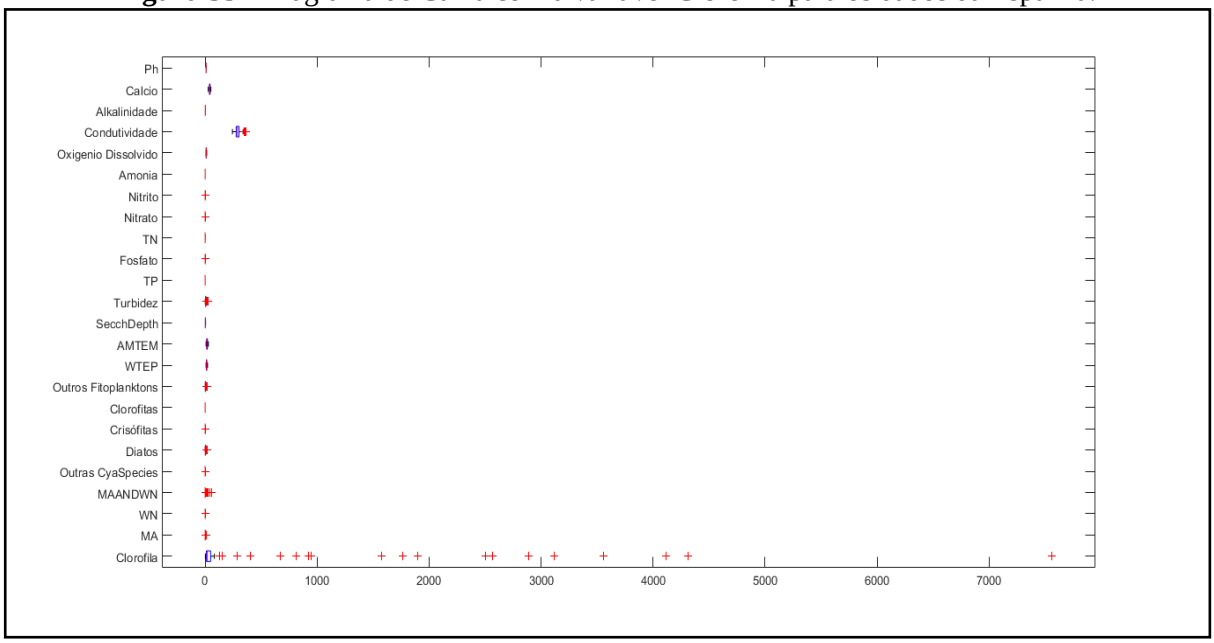


4.1.3 Análise de Agrupamentos aplicado aos dados da Espanha.

Os dados da Espanha possuem 24 variáveis com 135 amostras e estes dados foram escolhidos devido a sua grande amplitude de variáveis e numero de amostras. As variáveis incluem 8 características biológicas para a determinação de poluentes na agua.

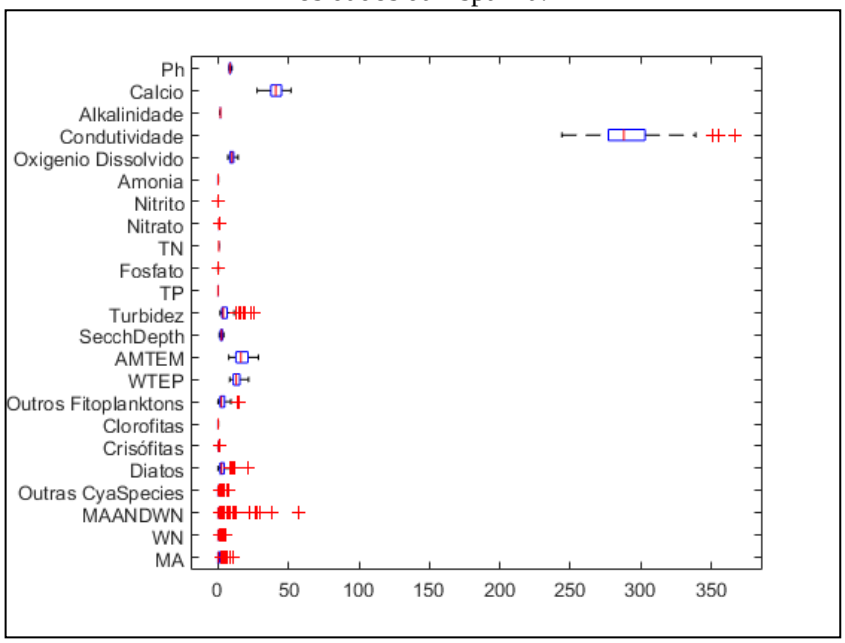
Antes de realizar a análise dos componentes principais, é possível usar o diagrama de caixa (Figura 35) para verificar a amplitude dos dados e se faz claramente visível a distribuição imprecisa da clorofila. Dentre as 135, apenas 20 obtiveram valor de clorofila superior a 100 e entre estes, 12 obtiveram valor superior a 1000. Esses valores que se alteram drasticamente não permitiram a realização do SOM de Kohonen com precisão, então a variável clorofila foi excluída dessa análise.

Figura 35 - Diagrama de Caixa com a variável Clorofila para os dados da Espanha.



Com a variável clorofila excluída, foi refeito o diagrama de caixa (Figura 36).

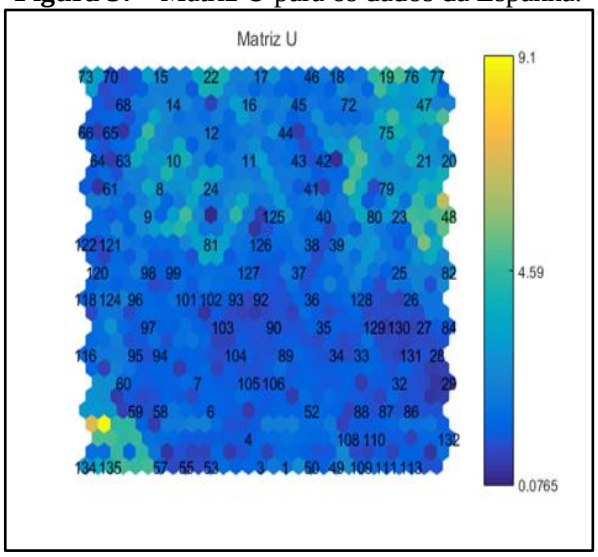
Figura 36 - Diagrama de Caixa com a variável Clorofila excluída para os dados da Espanha.



4.1.3.1 Kohonen na análise de agrupamentos da Espanha

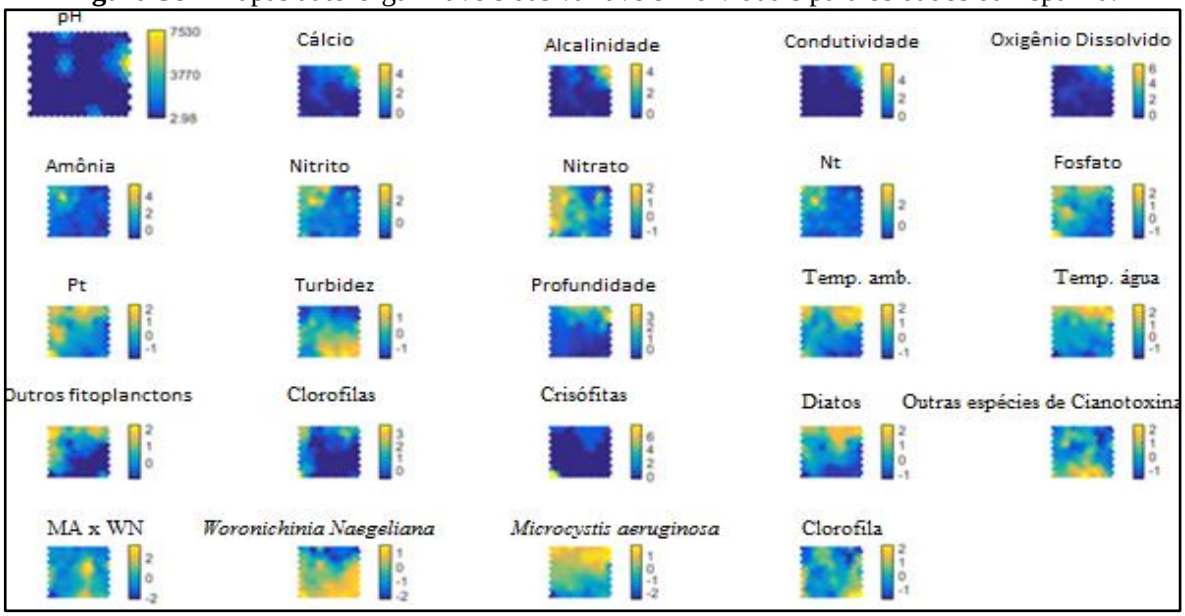
Na Figura37 (MATRIZ U), a matriz de distâncias e o mapa de características mostram que os dados de operação normal formam um agrupamento uniforme, o esquema de cores indica que existem pequenos valores de distâncias entre algumas amostras. O pequeno conjunto de neurônios no lado inferior esquerdo representa um conjunto de dados que apresentam outra dependência além das variáveis utilizadas ou um conjunto de *outliers*.

Figura 37 - Matriz U para os dados da Espanha.



Com os mapas auto-organizáveis das variáveis individuais (Figura 38), é possível identificar as variáveis e indicar sua influência no modelo. Desta forma, é possível indicar variáveis que poderão ser dispensadas para a criação do modelo preditivo.

Figura 38 - Mapas auto-organizáveis das variáveis individuais para os dados da Espanha.

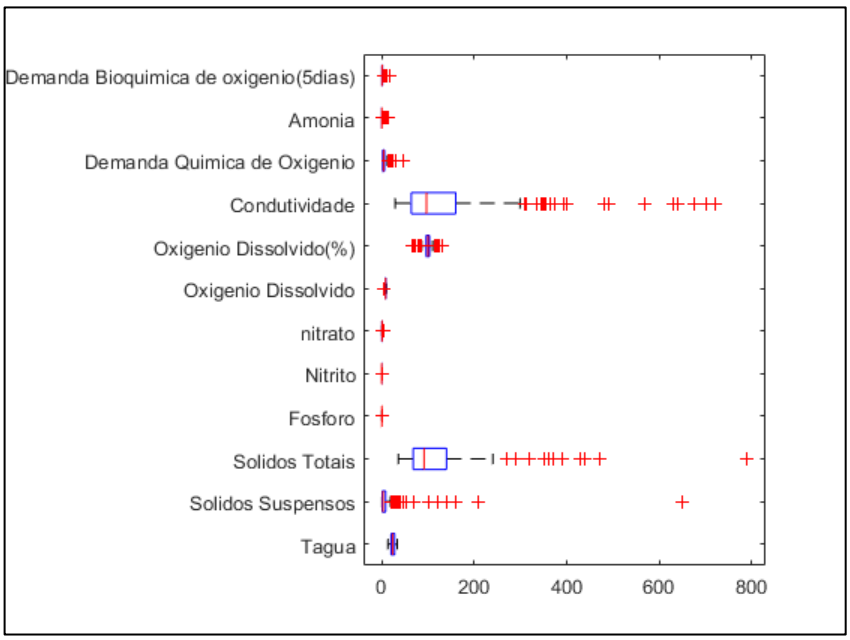


4.1.4 Análise de Agrupamentos aplicado aos dados da China

Os dados da China apresentam doze variáveis onde todas foram utilizadas para o SOM de Kohonen. A comparação entre essas amostras com estas de outros lugares não foi permitida devido á ausência de dados de clorofila, dados estes que são primordiais visto que o interesse desta análise é verificar o caráter preditivo desta variável.

Inicialmente, com os dados obtidos da China, foi possível montar um diagrama de caixas (Figura 39) para analisar as variáveis assim como a amplitude das amostras em relação à média.

Figura 39 - Diagrama de Caixa – Dados da China.

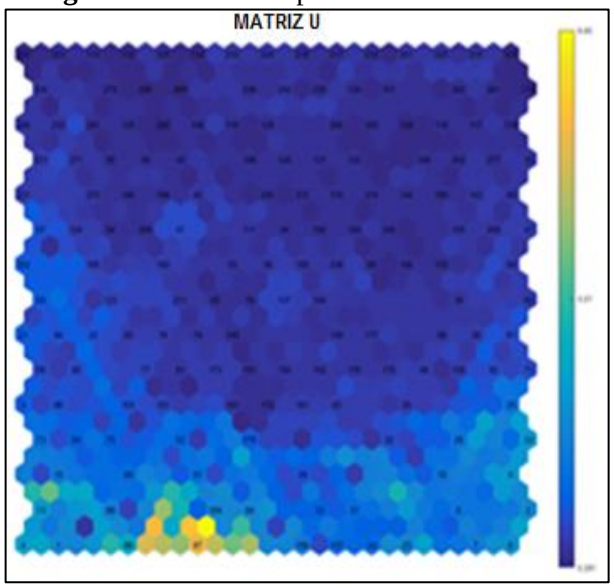


Com o diagrama montado, observa-se que, apesar de alguns dados fora de um intervalo confiável, as variáveis apresentam intervalos bem definidos e a redução dos dados utilizando se apresentou confiável.

4.1.4.1 Kohonen na análise de agrupamentos da China

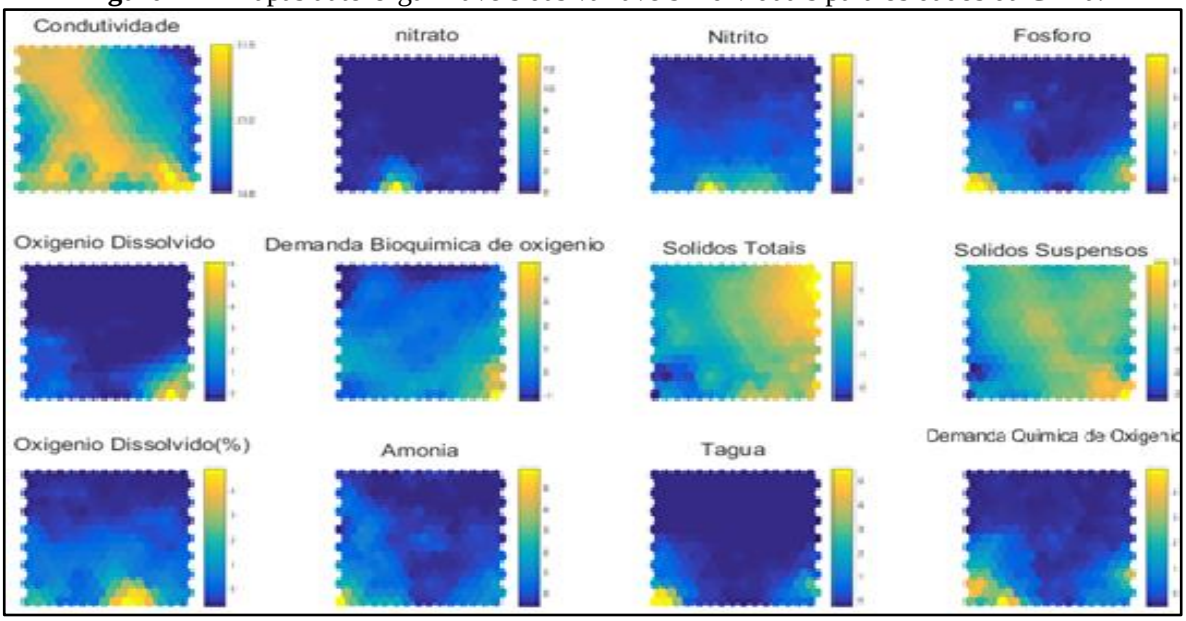
Na Figura 40, a matriz de distâncias e o mapa de características mostram que os dados de operação normal formam um agrupamento uniforme. O esquema de cores indica que existem pequenos valores de distâncias entre algumas amostras.

Figura 40 - Matriz U para os dados da China.



Com os mapas auto-organizáveis das variáveis individuais (Figura 41), é possível identificar as variáveis e indicar sua influencia no modelo. Desta forma, é possível indicar variáveis que poderão ser dispensadas para a criação do modelo preditivo.

Figura 41 - Mapas auto-organizáveis das variáveis individuais para os dados da China.



4.1.5 Análise de Componentes Principais aplicada aos dados das amostras de água do Brasil

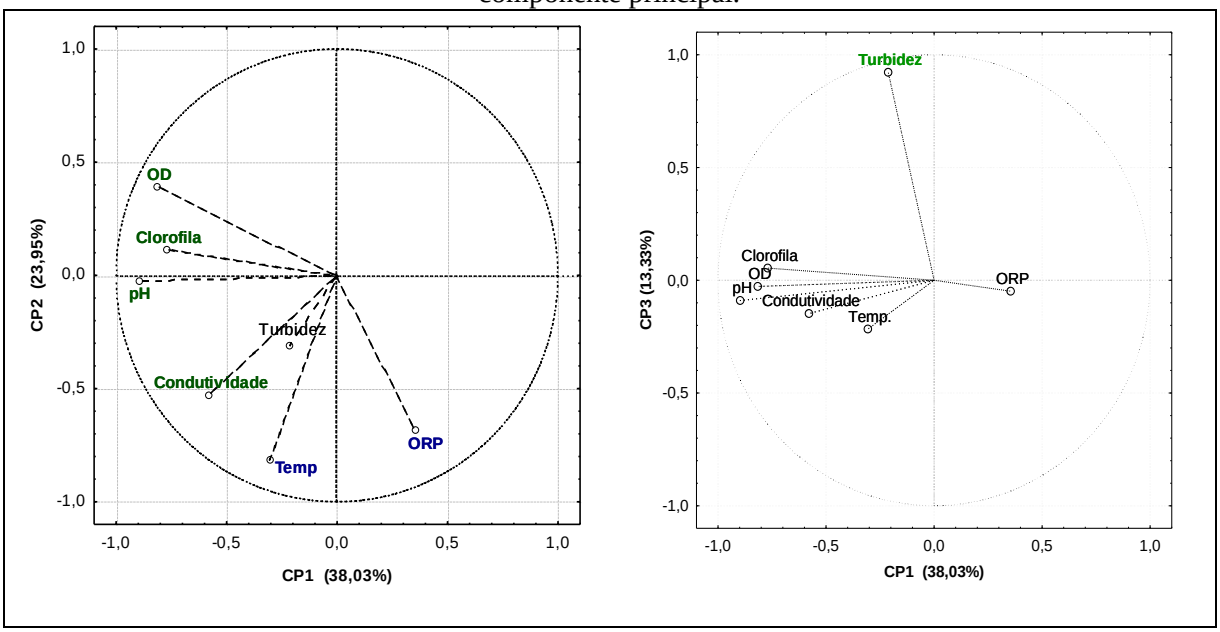
Nos dados do monitoramento da represa Guarapiranga em São Paulo no Brasil, foi aplicada a ACP e a matriz total tem dimensões de 512 (quinhentas e doze) amostras por 7

(sete) parâmetros físico-químicos e biológico da água dos mananciais dos rios Embu-Mirim, Embu-Guaçu, Santa Rita, Vermelho, Ribeirão Itaim, Capivari e Parelheiros.

As três primeiras componentes (Autovalor> 1) explicam 75,3% das variações na qualidade da água do Brasil (Tabela A1 no Apêndice A)e podem ser representadas em dois gráficos dos escores dos objetos (Figuras 45a e 45b).

A Figura 42 apresenta os gráficos dos pesos das variáveis e mostrou três grupos distintos de acordo com suas qualidades de água.

Figura 42 - Gráficos das dispersões e projeções biplot dos pesos dos parâmetros físico-químicos e biológico das amostras da água do Brasil nas (a) duas primeiras CPs e (b) primeira e terceira componente principal.



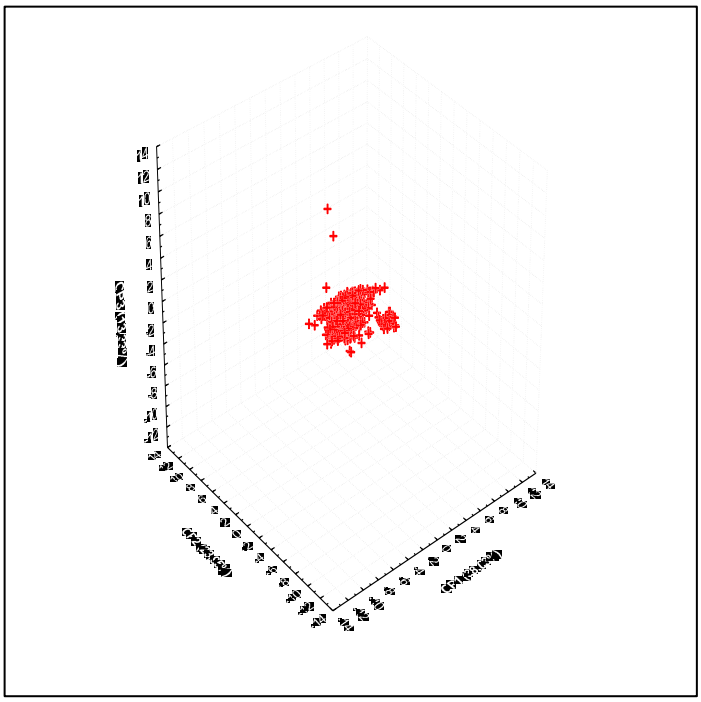
A primeira componente (CP1 - eixo horizontal do gráfico) é a principal mais significativa, que representa 38,02% da variação na qualidade da água, e exibiu altas cargas (à esquerda do gráfico) para pH (-0,897), Oxigênio Dissolvido (-0,815), Clorofila a (-0,770) e Condutividade Elétrica (-0,579), que refletiu o nível de eutrofização. Medeiros *et al*. (2017) aplicaram a ACP na avaliação da qualidade das águas do rio Murucupi no Pará-Brasil. Os pesquisadores relataram que as principais variáveis (pH, a temperatura e o oxigênio dissolvido) influenciaram o nível de qualidade da água do rio Murucupi, provavelmente devido a descarga contínua de esgoto doméstico.

A segunda componente (CP2 - eixo vertical do gráfico) explica 23,95% da variação na qualidade da água, e é altamente influenciada (parte inferior do gráfico) por temperatura da água (-0,816), potencial redox (-0,684) e condutividade (-0,530).

A CP3 (13,33%) apresentou forte associação com turbidez (0,922).

A Figura 43 apresenta o gráfico dos escores das amostras na CP1 (variando de -5,78 a 3,44), CP2 (variando de -3,74 a 2,23) e CP3 (variando de -1,23 a 13,28) que mostra uma maior semelhança entre as amostras plotadas mais próximas, enquanto distingue as amostras de baixa similaridade que estão mais distantes.

Figura 43 - Gráfico das dispersões e projeções3D dos escores das amostras da água do Brasil nas três primeiras CPs.



Na figura 43 observa-se uma nuvem bastante uniforme com as amostras similares, é possível afirmar que os dados se comportam de maneira similar e representam dados semelhantes de composição de água. Verifica-se que os dados não apresentam grupos distintos, de fato apresentam um grupo distribuído uniformemente assim indicando que estes dados colaboram com a modelagem empírica deste conjunto.

4.1.6 Análise de Componentes Principais aplicada aos dados das amostras de água da Espanha

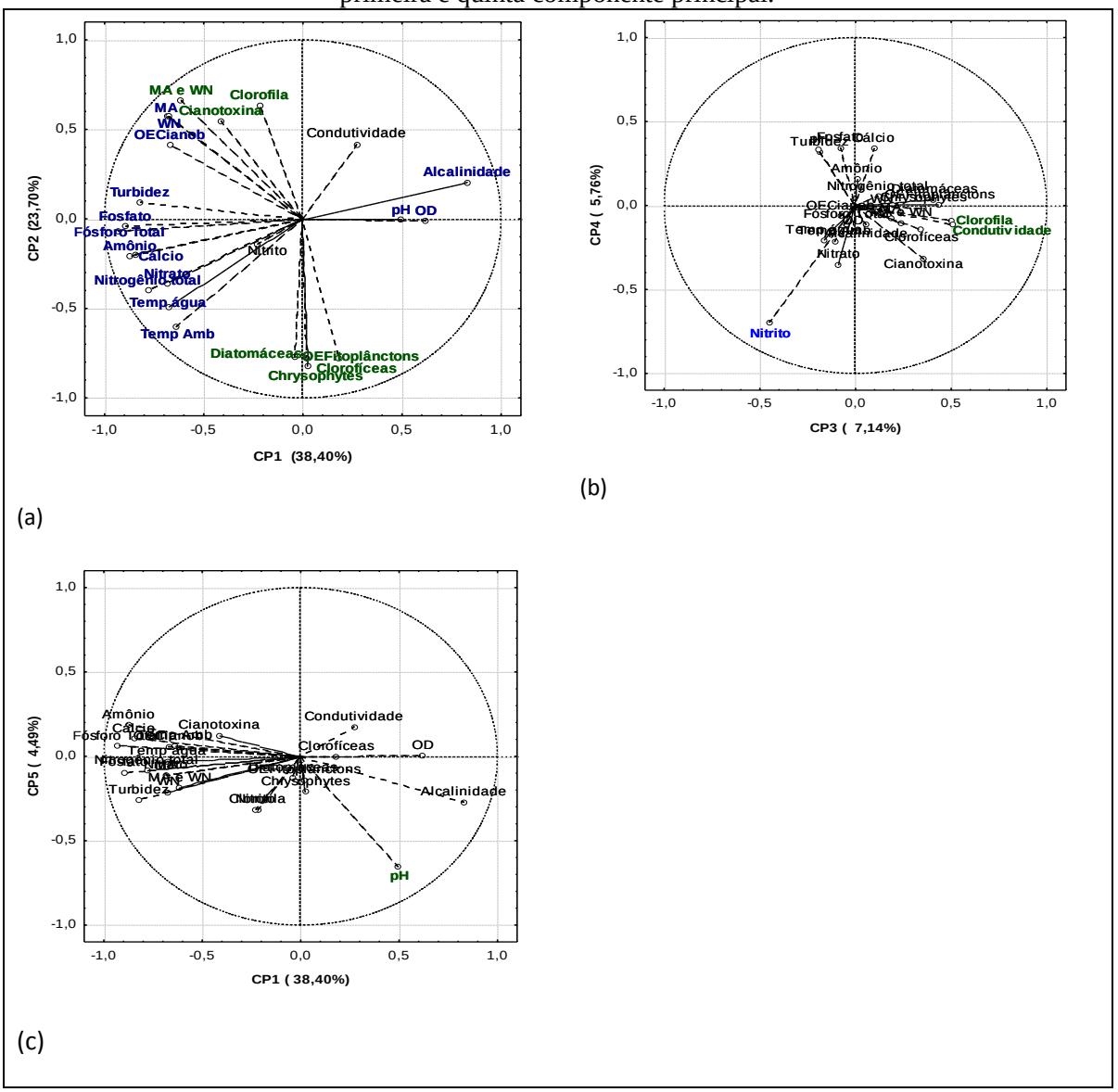
Nos dados do monitoramento do reservatório de Trasona na Espanha, foi aplicada a ACP e a matriz total tem dimensões de 151 (cento e cinquenta e uma) amostras por 24 (vinte e quatro) parâmetros físico-químicos e biológico da água dos mananciais dos rios Alvares e Narcea.

A partir da ACP da matriz dos dados, verifica-se uma grande redução no número de variáveis, uma vez que o melhor ajuste do modelo ocorreu com a inclusão de cinco CPs antes observadas em vinte e quatro dimensões (Tabela A2 no Apêndice A). As cinco CPs (Autovalor> 1) explicam 79,5% das variações na qualidade da água da Espanha e podem ser representadas em três gráficos dos escores dos objetos (Figura 44a, 44b e 44c).

Matiatos *et al.* (2018) aplicaram a ACP para identificar os fatores que controlam a qualidade da água do aquífero da planície do delta do rio Pinios (Tessália) na Grécia. Eles determinaram que cinco componentes (CPs) explicavam 77% da variância total dos parâmetros de qualidade da água; estes foram: (CP1) salinidade; (CP2) interação rochas-silicato-água; (CP3) dureza devido à dissolução de calcita e processos de troca de cátions; (CP4) poluição por nitrogênio; e (CP5) fertilizantes artificiais não relacionados a nitrogênio.

A Figura 44 apresenta os gráficos dos pesos das variáveis e mostrou cinco grupos distintos de acordo com suas qualidades de água.

Figura 44 - Gráficos das dispersões e projeções *biplot* dos pesos dos parâmetros físico-químicos e biológicos das amostras da água da Espanha nas (a) duas primeiras CPs, (b) terceira e quarta CPs e (c) primeira e quinta componente principal.



A primeira componente (CP1 - eixo horizontal do gráfico) representa 38,40% da variação na qualidade da água, e exibiu altas cargas negativa (à esquerda) para Fósforo total (-0,931), Fosfato (-0,896), Amônia (-0,873), Cálcio (-0,847), turbidez (-0,822), Nitrogênio total (-0,781), Nitrato (-0,683), *Microcystis Aeruginosa* (-0,682), Temperatura da água (-0,679), *Woronichinia Naegeliana (*-0,673*),* Outras espécies de Cianobactérias (-0,672), Temperatura Ambiente (-0,638), *Microcystis Aeruginosa*X *Woronichinia Naegeliana*(-0,614),fortes cargas positivas (à direita) para alcalinidade total (0,829), OD (0,621) e pH (0,495).

Zhao *et al.* (2015) aplicaram a ACP para investigar a relação entre fitoplâncton e variáveis ambientais na qualidade da água do Lago Feng-qing. Os pesquisadores relataram

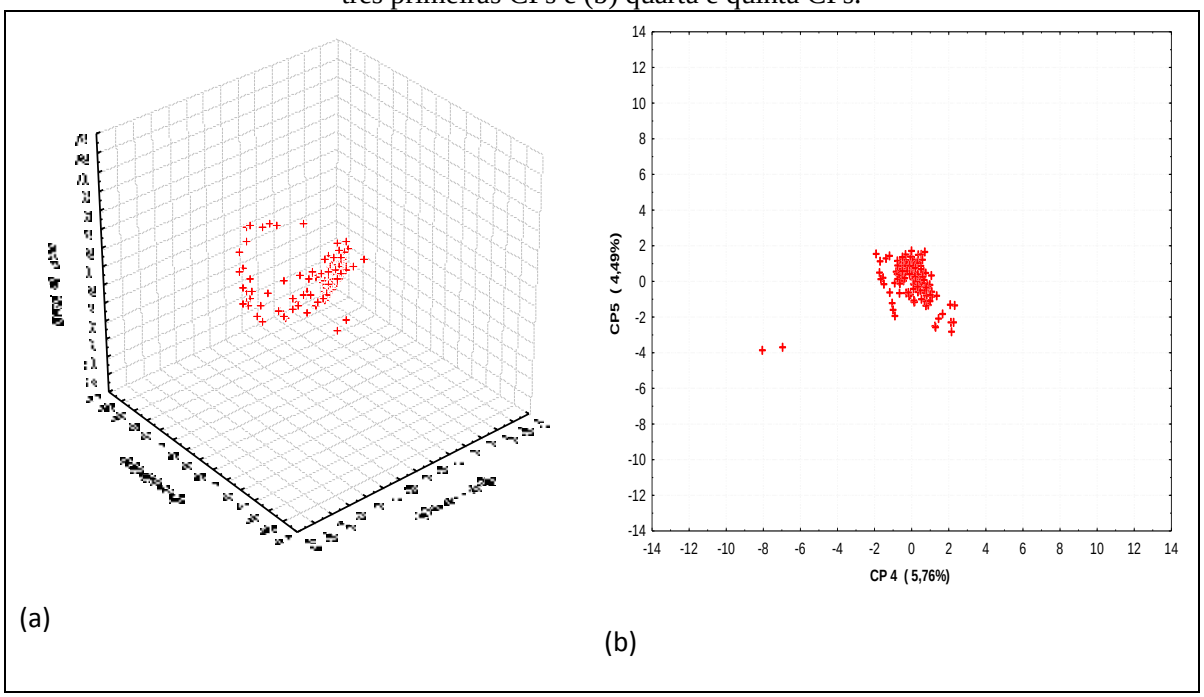
que os principais fatores que afetam o nível de eutrofização do lago foram a temperatura(T), fósforo total (TP), Clorofila-a (Chla), sólidos dissolvidos (SD) e turbidez.

A segunda componente (CP2 - eixo vertical do gráfico) explica 23,70% da variação na qualidade da água, e é altamente influenciada pelas fortes cargas negativas (parte inferior) por *Chrysophytes* (-0,823), Outras espécies de Fitoplânctons (-0,780), Clorofíceas (-0,776), Diatomáceas (-0,773), Temperatura Ambiente (-0,603), Temperatura da água -0,497), e teve fortes cargas positivas (parte superior) *Microcystis AeruginosaX Woronichinia Naegeliana*(0,663), Clorofila-*a* (0,635), *Microcystis Aeruginosa* (0,571), *Woronichinia Naegeliana* (0,571) e Cianotoxina (0,548).

A CP3 (7,14%) apresentou forte associação com condutividade (CE) (0,512) e Clorofila a (0,501). A CP4 (5,76%) incluiu o Nitrito (-0,694). A CP5 (4,49%) conteve o pH (-0,655).

A Figura 45 apresenta os gráficos dos escores das amostras na CP1 (variando de -6,80 a4,74), CP2 (variando de -4,74 a 6,42)e CP3 (variando de -4,71 a 3,67),e na CP4(variando de -8,05 a 2,32) e CP5 (variando de -3,87 a 1,71) que mostram uma maior semelhança entre as amostras, enquanto distingue poucas amostras com baixa similaridade.

Figura 45 - Gráficos das dispersões e projeções dos escores das amostras da água da Espanha nas (a) três primeiras CPs e (b) quarta e quinta CPs.



Pela análise da dispersão dos escores das amostras de água da Espanha nos cinco componentes principais, é possível afirmar que os dados se comportam de maneira semelhante entre si assim facilitando uma modelagem preditiva utilizando estes dados.

4.1.7 Análise de Componentes Principais aplicada aos dados das amostras de água da China

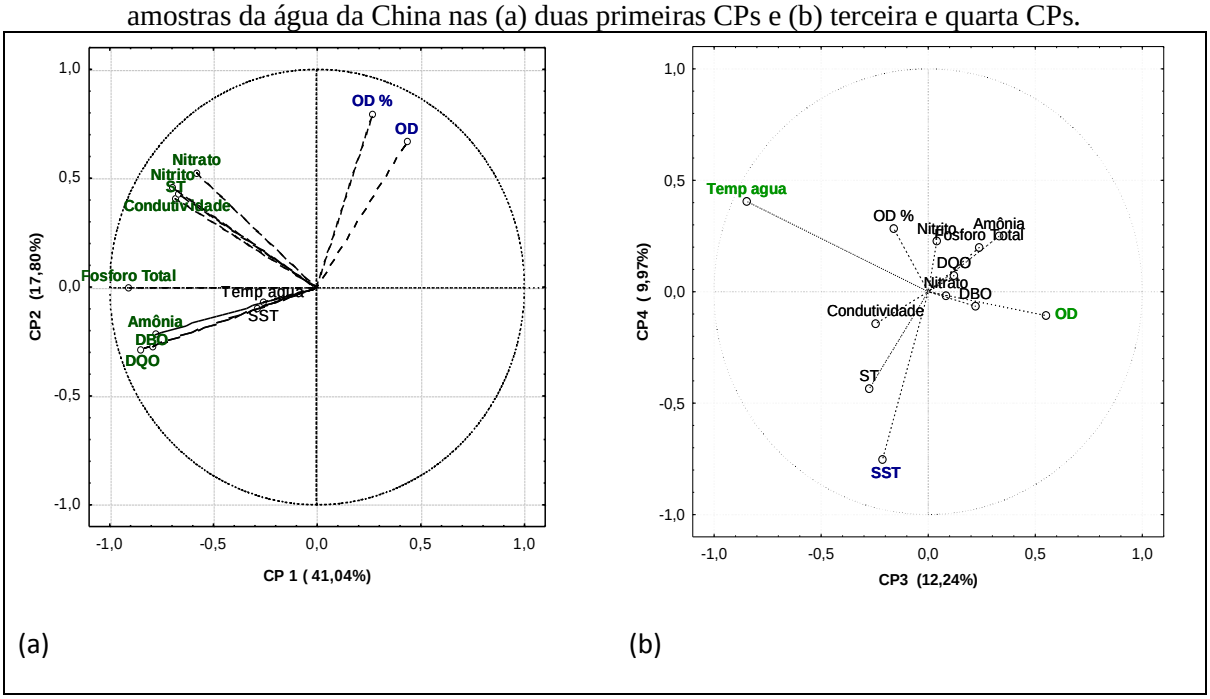
Nos dados do monitoramento do Rio LamTsuen no distrito de Tai Po ao sul de Hong Kong na China, foi aplicada a ACP e a matriz total tem dimensões de 288 (duzentos e oitenta e oito) amostras por 12 (doze) parâmetros físico-químicos da qualidade da água.

A partir da ACP da matriz dos dados, verifica-se uma grande redução no número de variáveis, uma vez que o melhor ajuste do modelo ocorreu com a inclusão de quatro CPs antes observadas em doze dimensões (Tabela A3 no Apêndice A).

As quatro CPs (Autovalor> 1) explicam 81,0% das variações da qualidade da água da China e podem ser representadas em dois gráficos dos escores dos objetos (Figura 49a e 49b).

A Figura 46 apresenta os gráficos dos pesos das variáveis e mostrou quatro grupos distintos de acordo com suas qualidades de água.

Figura 46 - Gráficos das dispersões e projeções biplot dos pesos dos parâmetros físico- químicos das amostras da água da China nas (a) duas primeiras CPs e (b) terceira e quarta CPs.

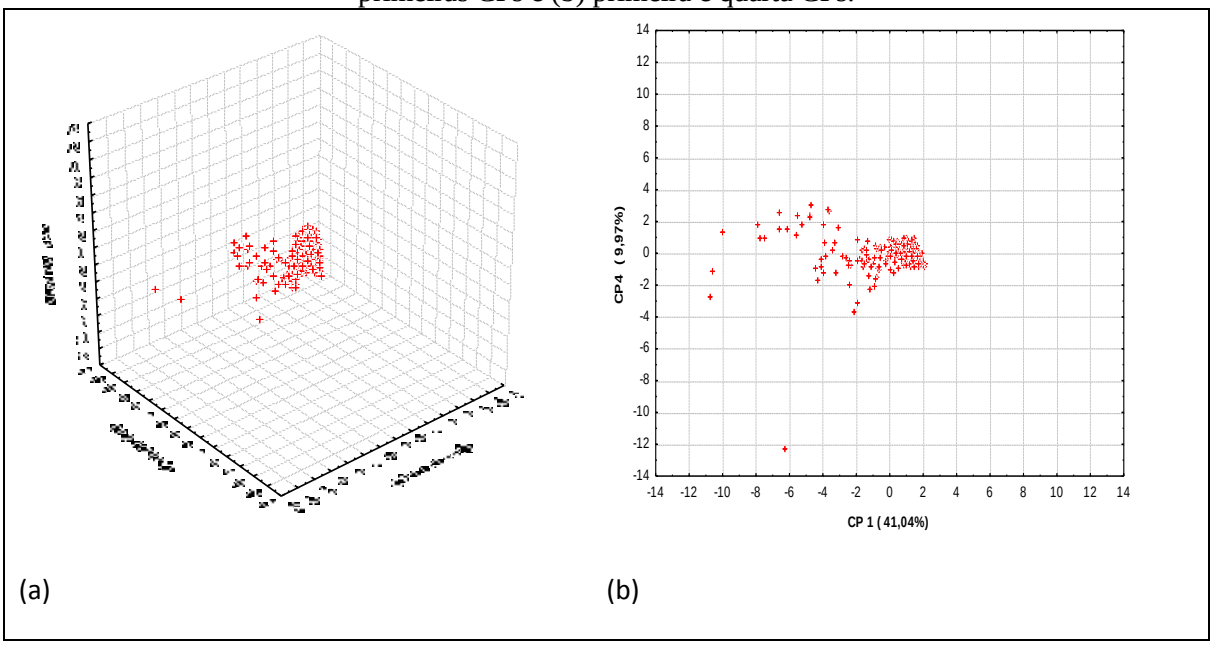


A CP1 representa 41,04% da variação na qualidade da água, e exibiu altas cargas (à esquerda) para Fósforo total (-0,908), DQO (-0,848), DBO (-0,796), Amônia (-0,778), Nitrito (-0,696), sólidos totais (-0,685), condutividade (CE) (-0,666) e Nitrato (-0,580), que são derivados de fontes de poluição orgânica/nutrientes nitrogenados de efluentes sanitários e tratamento de efluentes. Zhou *et al*. (2007) aplicaram ACP no conjuntos de dados de qualidade da água marinha do sul de Hong Kong. Os pesquisadores obtiveram na CP1 (29,44% do total variância) fortes cargas positivas em Nitrato, Nitrogênio total, e DBO, e cargas negativas na Salinidade, que foi atribuído à poluição orgânica das águas residuárias domésticas e do tratamento de esgoto.

A CP2 explica 17,80% da variação na qualidade da água, e é altamente influenciada (parte superior) por OD % (0,797), OD (0,670) e Nitrato (0,526). A CP3 (12,24%) incluiu a temperatura da água (-0,848) e OD (0,551). A CP4 (9,97%) compreendeu os sólidos suspensos totais (-0,753).

A Figuras 47 apresenta os gráficos dos escores das amostras na CP1(variando de -10,75 a 2,14), CP2 (variando de -5,77 a 8,82)e CP3 (variando de -3,72 a 3,55), e na CP1 e CP4 (variando de -12,34 a 2,99).

Figura 47 - Gráficos das dispersões e projeções dos escores das amostras da água da China nas (a) três primeiras CPs e (b) primeira e quarta CPs.



As Figuras 47a e 47b distinguem algumas amostras com baixa similaridade estando mais distantes, mas mostram que em sua grande maioria as amostras tinham uma maior

semelhança na qualidade das águas. Os gráficos apresentam uma maior uniformidade espacial sendo possível modelar estes dados.

4.1.8 Comentários gerais acerca das ACPs

As ACPs analisadas mostram que independentemente das quantidades de parâmetros e local de coleta das amostras de água, a qualidade das águas apresentou uma grande similaridade entres si. Observou-se também que a degradação da qualidade da água do Brasil, China e Espanha foi principalmente atribuída à poluição por matéria orgânica e compostos nitrogenados, e consequentemente encontravam-se eutrofizadas, provavelmente devido aos lançamentos de efluentes sanitários nos corpos d'água.

4.2 REDES NEURAIS ARTIFICIAIS - RNA (MLP) NO TRATAMENTO ESTACIONÁRIO E DINÂMICO DOS DADOS

Uma vez que uma das propostas deste trabalho foi a avaliação da aplicabilidade de diferentes técnicas de sistemas inteligentes e estatística multivariada na modelagem de dados de qualidade de água, deve-se salientar que a avaliação da aplicabilidade da mesmas aos dados utilizados estarão limitadas às opções de variabilidade dos parâmetros para avaliar melhorar a eficiência, de cada técnica, inserida em cada pacote computacional a ser utilizado. Dentre os parâmetros de validade do modelo citados na literatura, o mais importante é a avaliação dos erros para dados de teste e treinamento (Li *et al*., 2016; Zhang, 2016; Hameed *et al*., 2017; Csábrági *et al*., 2017; Khan e Chai, 2017; Djerioui *et al*., 2018; Haghiabi *et al*., 2018; Chou *et al*., 2018). No presente trabalho buscou-se encontrar principalmente um critério de avaliação que fosse disponibilizado nos diferentes pacotes computacionais utilizados neste trabalho.

Dessa forma, para análise da eficiência das diferentes técnicas com os dados os dados de diferentes fontes foi escolhido o coeficiente de determinação (R^2) quadrado, a raiz do erro quadrático médio (REQM) e o coeficiente de correlação de Pearson (r),que foram calculados para os conjuntos de dados de treino e de teste(Clorofila, DBO e Cianotoxina). Esta escolha corrobora com a literatura consultada (Li *et al*., 2016; Zhang, 2016; Hameed *et al*., 2017; Csábrági *et al*., 2017; Khan e Chai, 2017; Djerioui *et al*., 2018; Haghiabi *et al*., 2018; Chou *et al.*, 2018) que demostrou a frequente utilização dos mesmos na análise da eficiência de uma

determinada técnica individualmente, assim como em análise comparativas entre duas ou mais técnicas investigadas em diferentes trabalhos.

Outra questão foi a validação dos resultados quando da utilização de um número reduzido de dados, questão essa abordada na revisão bibliográfica em que se pode perceber a validade da análise aqui realizada, uma vez que foi observado que trabalhos recentes na literatura (Ravansalar, 2015; Zhang, 2016; Barzegar *et al*., 2018) apresentam números de dados semelhantes aos utilizados neste trabalho. Esse fato reforça a dificuldade generalizada em obter-se dados de qualidade de água de uma forma geral.

4.2.1 Redes Neurais em Dados Estacionários

Deve-se nesse momento ressaltar novamente que a eficiências dos modelos de inferência com RNA's são influenciados por diversos parâmetros ligados a sua arquitetura e aprendizado, tais como: estimativas iniciais, divisão de dados para treinamento e teste, combinação de Funções de Ativação, número de camadas intermediárias e número de neurônios nas camadas intermediárias.

Nesta etapa do trabalho foi utilizada uma ferramenta computacional em ambiente MatLab com uma interface para o usuário que limita as possibilidades de variações de alguns parâmetros para rede neural, como número de camadas intermediárias, funções de ativação, algoritmos de treinamento e estimativas iniciais.

Deve-se salientar que a busca dos melhores parâmetros disponibilizados no pacote computacional do MatLab se torna um processo exaustivo porque além de ter que para cada tentativa se fixar determinados parâmetros referentes a arquitetura e aprendizado deve-se executar várias vezes a mesma para variar a estimativa inicial assim como a divisão dos dados de treinamento e teste.

Na ferramenta do MatLab estão fixados o número de camadas (3), as funções de ativação (linear ,tangente hiperbólica e linear) podendo-se variar unicamente o número de neurônios na camada intermediária e os algoritmos de treinamento, sendo eles o Levemberg-Marqguard e a Regularização Bayesiana. Vale salientar que na mesma a inicialização dos pesos e bias e a divisão dos dados 70% para treinamento, 15% para teste e 15% para validação são feitas randomicamente.

Para os algoritmos de aprendizados utilizaram-se os seguintes algoritmos de treinamento através do pacote computacional comercial MatLab7.0:

- Algoritmo Levenberg-Marquardt (Lm) → é o algoritmo de treinamento mais rápido para redes de tamanho moderado. Possui a característica de redução de memória para usar quando o grupo de treinamento é grande.
- Algoritmo de Regularização Bayesiana - Trainbr (Tbr) → é uma modificação do algoritmo de treinamento Lm, para produzir redes que generalizam bem. Reduz a dificuldade de determinar a arquitetura ótima da rede. Estes algoritmos são baseados no método quasi-Newton, e que, a depender do tamanho da rede utilizada, é muito mais eficiente que os do primeiro tipo (Retroalimentação).

Deve-se salientar que no tipo de algoritmo (Tbr) no MatLab a divisão randômica de dados é feita unicamente para treinamento e teste diferente do (Lm) em que parte dos dados vai para a validação na tentativa de evitar o sobre ajuste.

Dessa forma admitindo-se o critério heurístico que os neurônios da camada intermediária devam ser o número de neurônios da camada de entrada mais um, restringiu-se os arranjos para a investigação da eficiência dos três modelos: para Clorofila nos dados do Brasil, DBO nos dados da China e Cianotoxinanos dados da Espanha.

4.2.1.1 Redes Neurais Artificiais aplicadas aos dados das amostras de água do Brasil

Os dados do Brasil tiveram como base 391 amostras de 8 parâmetros distintos como referido anteriormente, lembrando que aqui a rede foi utilizada como ajuste de função e a variável inferida foi a clorofila e os parâmetros para avaliação de eficiência foi a Raiz do Erro Médio Guadrado (RMSE) e o coeficiente de correlação de Pearson (r). Observa-se que os valores obtidos pelo modelo foram adequados para foram bons devido estarem próximos de 1 (Figura 48), ou seja, o modelo criado pela rede mostrou fidelidade aos dados analisados. Além disso, observa-se pelo histograma de erros (Figura 49) que os dados de teste, em sua maioria, segue a mesma linha que o treino, onde quando um cresce, o outro cresce proporcionalmente. Salientando-se que antes da análise foi feita a remoção dos *outliers*, o que gerou resultados melhores para a análise.

A etapa de treinamento se encerra quando o erro atingir o critério de convergência previamente estabelecido. Neste ponto, a rede está treinada e podemos utilizar um conjunto de dados, diferente daquele utilizado no treinamento, para avaliar as propriedades de generalização da rede neural artificial.

As redes neurais geralmente apresentam uma boa capacidade de generalização, ou seja, apresentam resultados satisfatórios quando aplicada em amostras que não participaram do treinamento, especialmente em situações em que os dados apresentem não linearidades, visto que geralmente se utilizam funções de transferência não lineares. Entretanto, em alguns casos, as redes neurais artificiais, mesmo depois de terem sido treinadas, apresentam baixos erros de calibração e elevados erros de previsão, isto é, sobreajuste (*overfitting*), devido ao número excessivo de neurônios utilizados na camada intermediária. Desta forma, neste trabalho variaram-se em duas unidades os números de neurônios da camada intermediária a partir do critério heurístico citado anteriormente pra avaliar a sensibilidade da técnica.

A Tabela 22 apresenta os resultados e as Figuras 48 e 49 apresentam os respectivos gráficos de regressão e histograma de erros para o modelo 6 que representou um melhor resultado. Devemos salientar que foram feitas várias execuções do programa, pois para cada execução a estimativa dos pesos iniciais é feita randomicamente no pacote utilizado e pôde-se perceber durante a execução do trabalho que realmente estas estimativas são determinantes na eficiência da técnica de RNA's e que os r e RMSE estão em consonância com trabalhos encontrados na literatura para este tipo de análise. Deve-se salientar que resultados melhores poderiam ser obtidos a partir de novas tentativas.

Tabela 22 - Resultados do desempenho da RNA nos 6 modelos de clorofila para os dados do Brasil.

Modelo	Combinação de n° de Neurônios	Algoritmo de Treinamento	r Treinamento	RMSE Treinamento	r Teste	RMSE Teste
1	10	LM	0,84761	7,0193	0,7695	10,2572
2	9	LM	0,88924	6,7129	0,8312	8,4621
3	8	LM	0,88616	6,7668	0,8862	8,6888
4	10	LBr	0,90511	6,2333	0,8072	9,0654
5	9	LBr	0,90633	5,9185	0,8356	9,5474
6	**8**	**LBr**	**0,88203**	**6,9004**	**0.8959**	**6,2852**

Figura 48 - Gráficos de regressão para o modelo 6, de clorofila, para dados estacionários do Brasil.

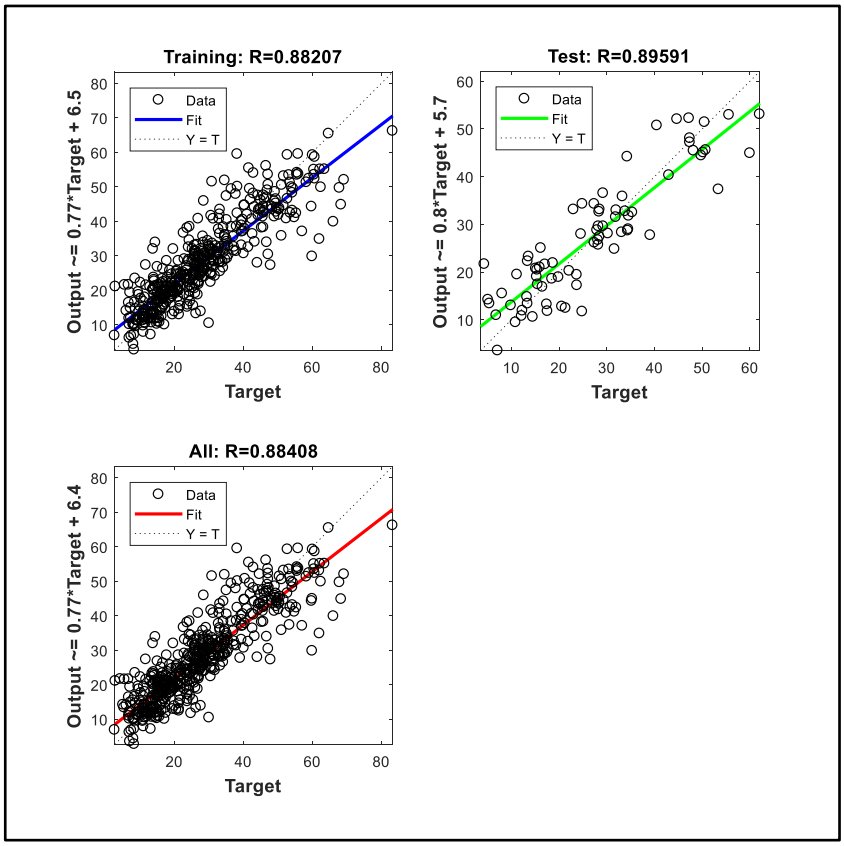
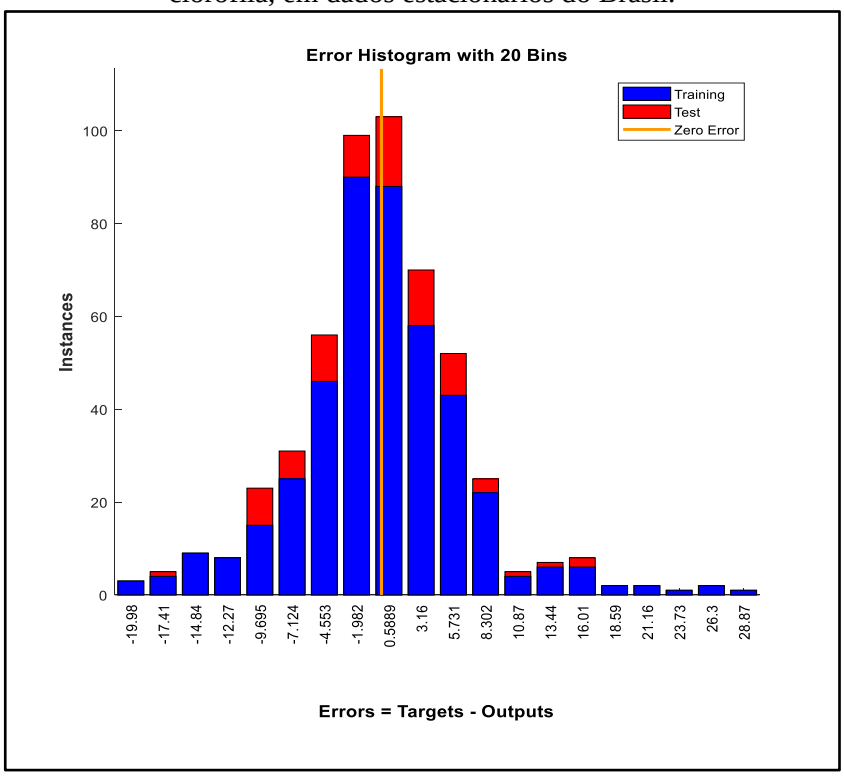


Figura 49 - Gráfico do histograma de erros para o modelo 6, de clorofila, em dados estacionários do Brasil.



4.2.1.2 Redes Neurais Artificiais aplicadas aos dados das amostras de água da China

Os dados da China tiveram como base 240 amostras de 12 parâmetros distintos, como referido anteriormente, lembrando que aqui a rede foi utilizada como ajuste de função e a variável inferida foi a DBO. Observa-se que os valores obtidos para o r foram bons, devido estarem próximos de 1, ou seja, o modelo criado pela rede mostrou fidelidade aos dados analisados. Além disso, observa-se pelo histograma de erros que o teste, em sua maioria, segue a mesma linha que o treino, onde quando um cresce, o outro cresce proporcionalmente. Antes da análise foi feita a remoção dos *outliers* (pontos fora da curva), o que gerou resultados melhores para ela. Vale salientar que quanto mais próximo de 1 for r, maior será a relação entre as variáveis, enquanto que quanto menor for o RMSE , maior será a fidelidade dos dados ao modelo.

A Tabela 23 apresenta os resultados dos modelos, indicando que o modelo 6 apresentou o menor erro. O Gráfico de Regressão e de Histograma para o modelo 6 de DBO dos dados estacionários da amostra de dados da China são apresentados nas Figuras 50 e 51, respectivamente. Os resultados mostram uma boa regressão e que os erro apresentaram uma distribuição normal, indicando que o modelo 6 de RNA é adequado para predição do DBO. Pode-se observar que para os dados e resultados obtidos neste trabalho as RNA's formam melhores para os dados da china que do Brasil. Vale salientar que os dados estão em consonância com trabalhos encontrados na literatura para este tipo de análise. Deve-se salientar que para estes tipos de técnicas sempre existe a possibilidade de obterem-se diferentes a partir de novas tentativas.

Tabela 23 - Resultados do desempenho da RNA nos 6 modelos de DBO para os dados da China.

Modelo	Combinação de n° de Neurônios	Algoritmo de Treinamento	r Treinamento	RMSE Treinamento	r Teste	RMSE Teste
1	14	LM	0,9890	1,0245	0,8072	3,8056
2	13	LM	0,8732	1,3074	0,9349	0,6720
3	12	LM	0,8975	1,2382	0,92382	1,2716
4	14	LBr	0,9222	0,9839	0,8733	2,2062
5	13	LBr	0.9507	0,0643	0,6025	9,8330
6	12	LBr	0,99489	0,0624	0,8152	10,515

Figura 50 - Gráficos de regressão para o modelo de DBO em dados estacionários da China.

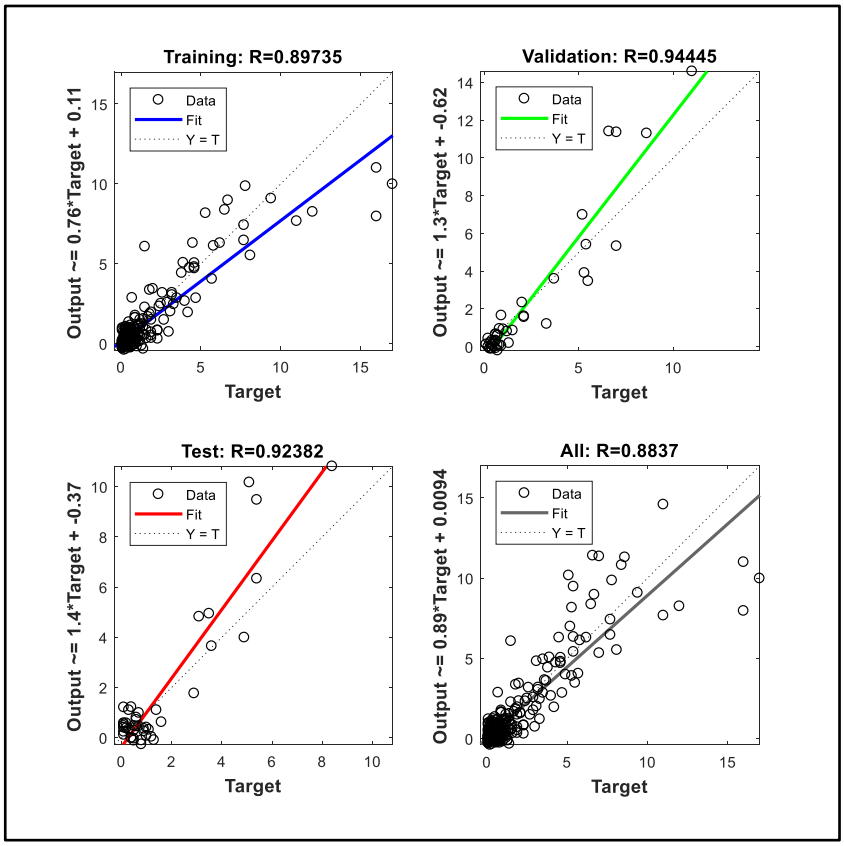
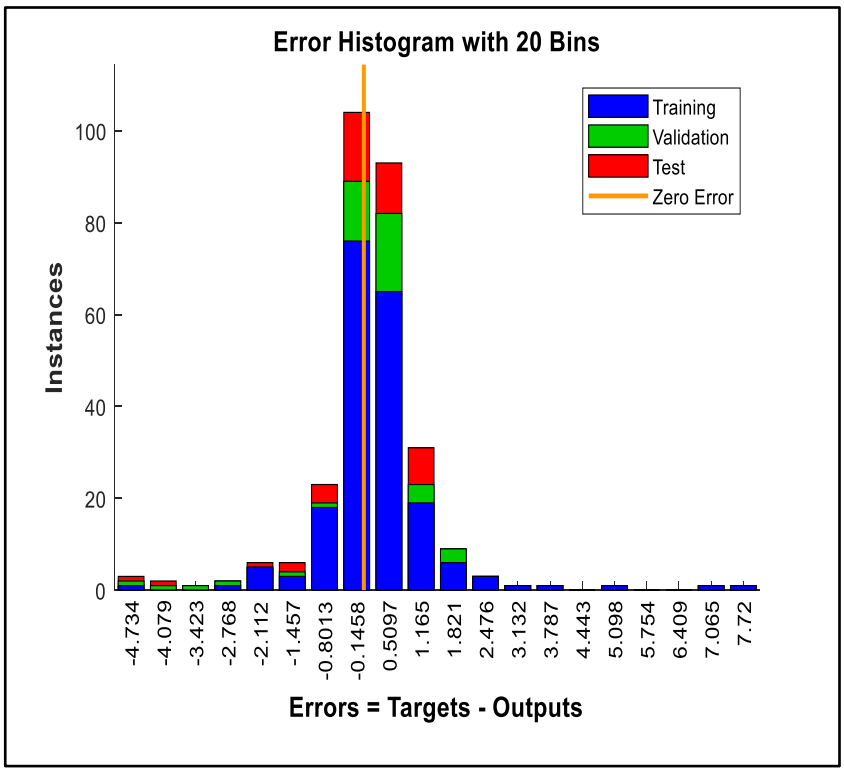


Figura 51 - Gráfico do histograma de erros para o modelo de DBO em dados da China.



4.2.1.3 Redes Neurais Artificiais aplicadas aos dados das amostras de água da Espanha

Os dados da Espanha tiveram como base 151 amostras de 25 parâmetros distintos, como referido anteriormente, lembrando que aqui a rede foi utilizada como ajuste de função e a variável inferida foi a cianobactéria. Vale salientar que estes dados apresentam uma particularidade de apresentarem dados de cianobactérias com valores zeros (ausência de cianobactérias) e muito próximos de zeros, fato que dificultou muito a utilização das RNA's nessa etapa do trabalho. Observa-se que os valores obtidos para o r foram bons devido estarem próximos de 1 (Tabela 24), ou seja, o que caracteriza um sobreajuste. Além disso, observa-se pelo histograma de erros que as barras que representam o teste, em sua maioria, estão bem próximas de zero, coincidindo com os valores nulos de cianotoxina encontrados no rio – vale salientar que não foi feita a detecção de *outliers* para esses dados, pois os mesmos já forma disponibilizados corrigidos. Verifica-se que neste caso tem-se um excesso de parâmetros para um curto número de amostras o que é um problema para a modelagem principalmente por serem fontes de sobreajuste também.

Para este conjunto de dados, com essa problemática, foram feitas diversas tentativas para realizar a modelagem dos dados através das redes neurais com os dois algoritmos Lm e Lbr, disponibilizados na versão do MatLab, porém os resultados apresentavam *over-fitting*, porque o erro RMSE do treinamento foi muito menor do que o RMSE do teste, Desta forma, os modelos não foram adequados.

Abaixo estão apresentados os gráficos de Regressão, para treinamento, validação e teste, e de Histograma para o modelo de cianotoxinas dos dados estacionários da amostra de dados da Espanha nas Figuras 52 e 53, para o LM, e Figuras54 e 55, para Lbr, ambos com a configuração de 25 neurônios na camada intermediária. Deve-se salientar mais uma vez que diversas tentativas foram realizadas na busca de resultados satisfatórios.

Tabela 24 - Resultados do desempenho dos modelos de cianotoxinas via RNA para os dados estacionários da Espanha.

Modelo	Combinação de n° de Neurônios	Algoritmo de Treinamento	r Treinamento	RMSE Treinamento	r Teste	RMSE Teste
1	25	LM	0,9999	0,0005	0,698	86,787
2	25	LBr	0,999	0,00002	0,35934	110,5257

Figura 52 - Gráficos de Regressão do modelo de cianotoxinas em dados estacionários da Espanha (Para LM).

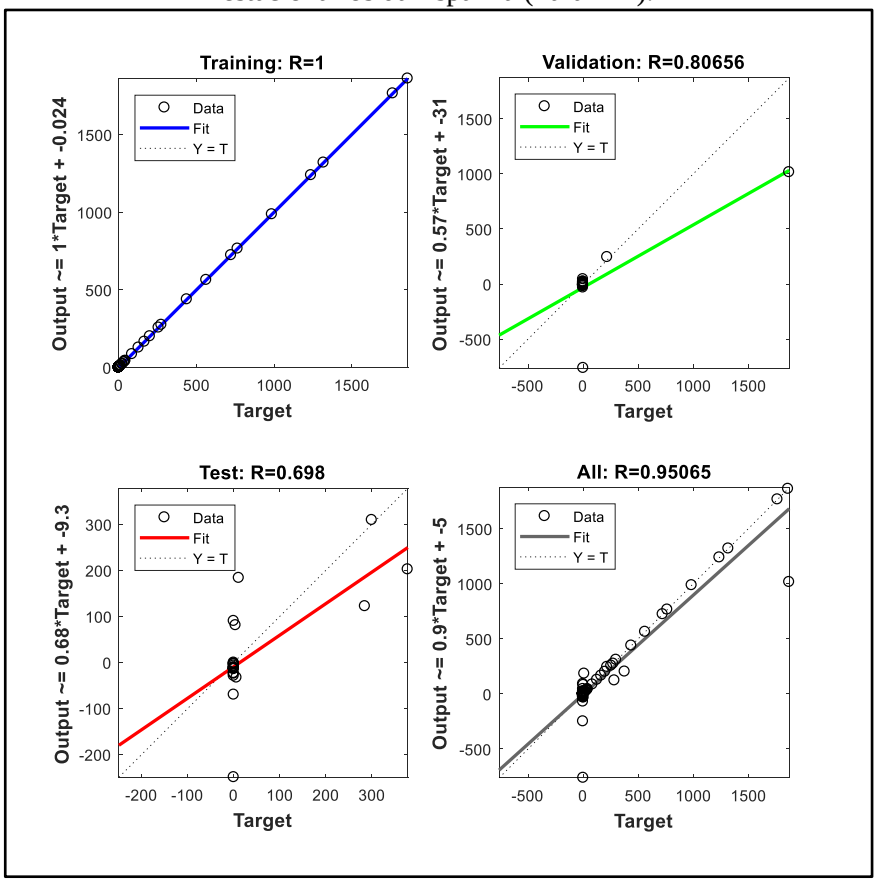


Figura 53 - Gráfico de histograma dos erros do modelo de cianotoxinas da Espanha (para LM).

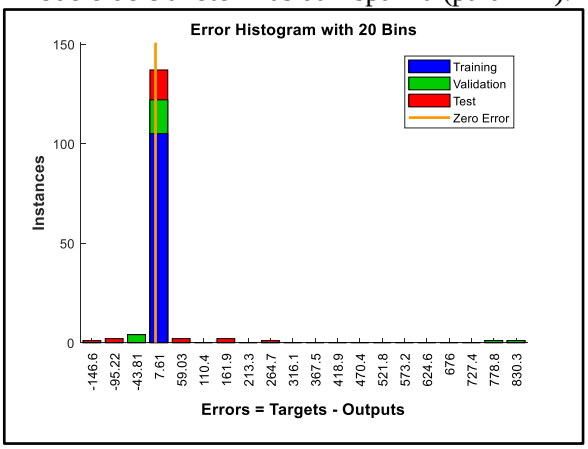


Figura 54 - Gráficos de Regressão do modelo de cianotoxinas da Espanha (Para Lbr).

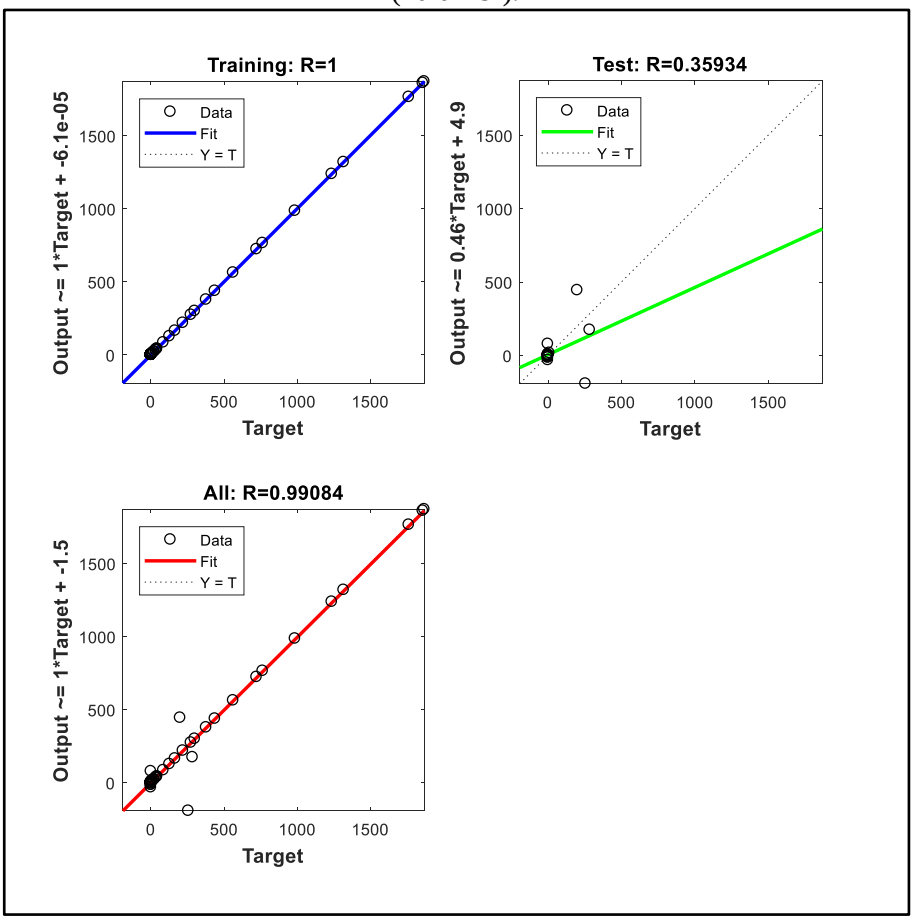
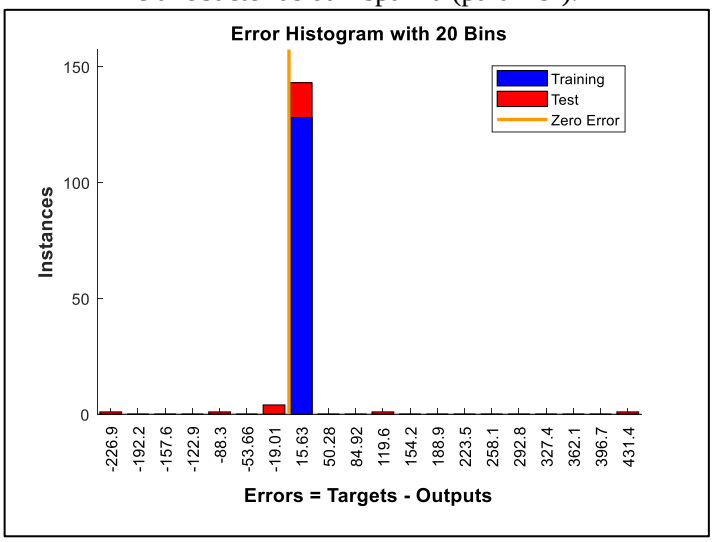


Figura 55 - Gráfico de histograma dos erros do modelo de cianobactérias da Espanha (para Lbr).



Porém verificou-se, na finalização do trabalho, que numa versão mais atualizada do MatLab sua ferramenta para treinamento de RNA'S haviam disponibilizado um algoritmo denominado de *Scaled Conjugate Gradiente* (SCG) em que o algoritmo é finalizado

automaticamente quando a generalização do modelo ou seja o treinamento está caminhando para um sobre ajuste. Este novo algoritmo foi implementado no MatLab, segundo seus construtores, para melhorar a modelagem de dados como os da Espanha, que através dos algoritmos LM e LBr apresentaram novo sobre ajuste. Os resultados para atualização deste algoritmo para os dados da Espanha estão na Tabela 25.

Os Gráficos de Regressão, para treinamento, validação e teste, e de Histograma para o modelo de cianotoxina utilizando o algoritmo SCG dos dados estacionários da amostra de dados da Espanha são apresentados nas Figuras 56 e 57, respectivamente. Apesar do histograma do erro se aproximar de uma distribuição normal, a Tabela 25 mostra que o erro dos três modelos apresenta um valor muito alto, maior do que a média dos dados de cianotoxinas igual a 93,4, mostrados na Tabela 9. Estes resultados mostram que os modelos de RNA de cianotoxinas não foram completamente satisfatórios.

Tabela 25 - Resultados do desempenho modelo de cianotoxinas via RNA para os dados da Espanha, com algoritmo SCG.

Modelo	Combinação de n° de Neurônios	Algoritmo de Treinamento	r Treinamento	RMSE Treinamento	r Teste	RMSE Teste
1	27	SCG	0,95212	109,2501	0,92688	100,03
2	26	SCG	0,92136	121,4084	0,9398	157,1005
3	25	SCG	0,94461	100,0110	0,95645	132,742

Figura 56 - Gráficos de Regressão do modelo de cianotoxinas da Espanha (para SCG).

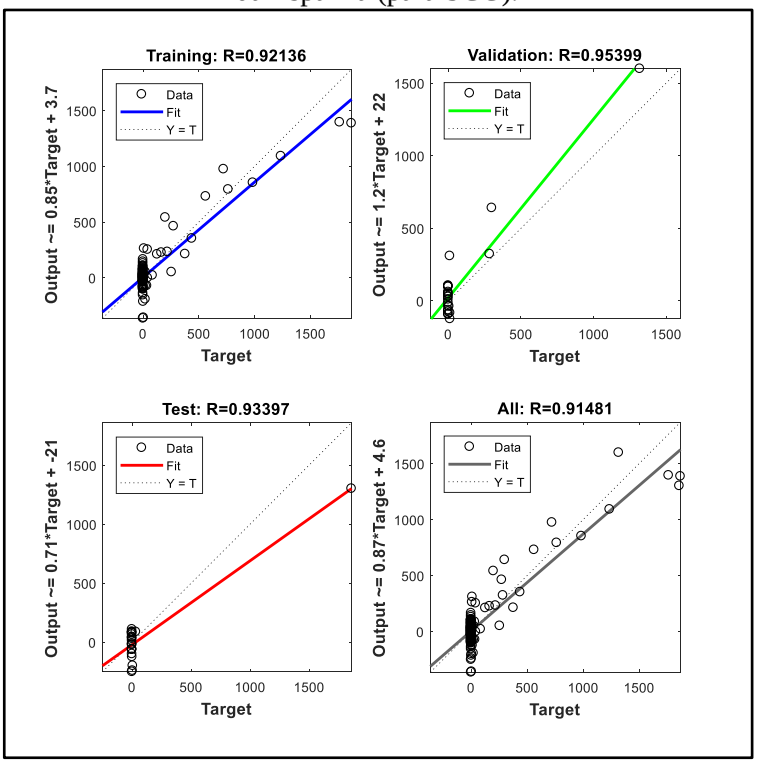
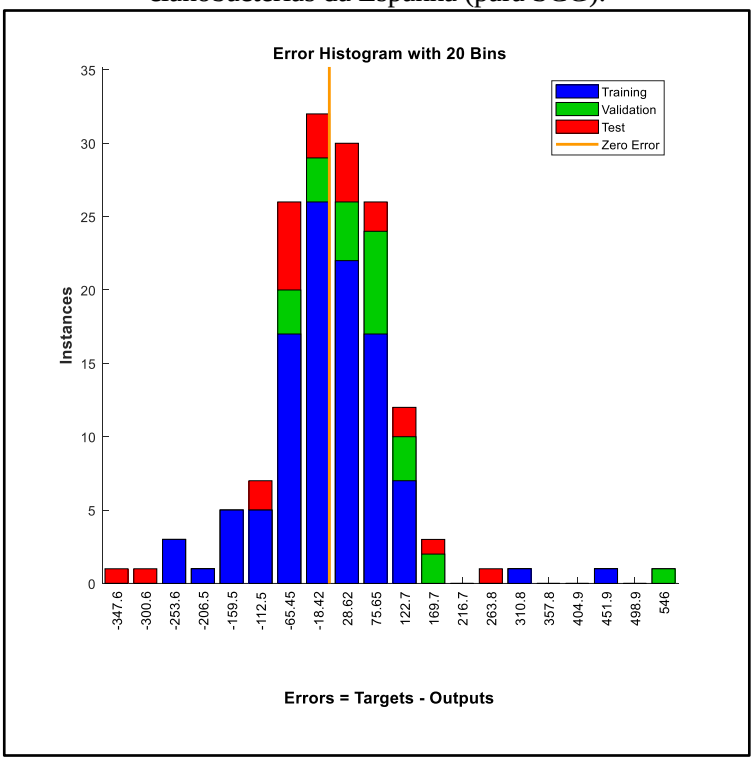


Figura 57 - Gráfico de histograma dos erros do modelo de cianobactérias da Espanha (para SCG).



A Tabela 26 a seguir reúne os melhores resultados para o (r) e o (RMSE) que foram calculados para os conjuntos de dados de treino e de teste utilizados para os três modelos de predição (Clorofila, DBO e Cianotoxina).

Tabela 26 - Desempenho da RNA para os modelos de Clorofila (Brasil), DBO (China) e Cianotoxina (Espanha).

Modelo	**RMSE**		**r**	
	Trein.	**Teste**	**Trein.**	**Teste**
Clorofila (Brasil)	6,9004	6,2852	0,88203	0.8959
DBO (China)	1,3074	0,6720	0,8732	0,9349
Cianotoxina (Espanha)	100,0110	132,742	0,9446	0,95645

Os resultados mostram a aplicabilidade das redes neurais artificiais na qualidade de água ficando clara a extrema sensibilidade dos modelos aos dados, quantidade e qualidade, e parâmetros que buscam maior eficiência. Contudo o modelo para previsão de cianotoxinas no corpo hídrico da Espanha via RNA não apresentou ajuste satisfatório devido ao elevado erro RMSE. Este fato ocorreu em consequência do excesso de zeros ou valores muito próximo assim como a relação número de variáveis, verso número de dados. Deve-se salientar que no artigo que modelou os dados da Espanha, os autores consideram seus resultados satisfatórios (FERNÁNDEZ, *et al*.2013).

4.3 MÁQUINAS DE VETORES DE SUPORTE (SVM)

4.3.1 Maquinas de Vetores de Suporte em dados estacionários

Para avaliação da eficiência das Maquinas de Vetores de Suporte na modelagem e análise exploratória de dados e comparação com as demais técnicas de sistemas inteligentes avaliados neste trabalho, utilizou-se o mesmo conjunto de dados aplicados nas seções anteriores com a Rede Neural Artificial na forma de inferenciadora para a Clorofila, DBO e cianobactérias.

A partir de uma busca na literatura (Li *et al*., 2016; Djerioui *et al*., 2018; Chou *et al*., 2018) foi possível perceber que, diferente das RNA's, o SVM tem a sua fundamentação teórica de aplicação pautada na Teoria de Aprendizagem Estatística e Otimização Matemática. Sendo assim, o conceito básico da utilização da SVM para a regressão consiste

em mapear de forma não linear os dados originais em um espaço de maior dimensão, desta forma o problema de regressão se torna linear.

Vale salientar que para a aplicação da SVM na modelagem de dados utilizou-se uma ferramenta computacional disponível no ambiente Matlab, na qual a inferência do usuário se limitava a variação das funções Kernel disponibilizadas pela ferramenta citada. Desta forma, buscou-se a avaliação das funções: linear, quadrática, cúbica, *fine gaussian, médium gaussian* e *coarse gaussian,* para os mesmos três modelos (Clorofila, DBO e Cianotoxina) utilizados com a RNA. Essas três aplicações diferente das realizadas pelas RNA's foram avaliadas a partir dos parâmetros (r), (MSE), Raiz do Erro Quadrado Médio (*root Mean Square Error* (RMSE)) e o Erro Médio Absoluto (*Mean Absolute Error* (MAE)), uma vez que estes estavam disponibilizados na ferramenta computacional utilizada.

4.3.1.1 SVM aplicada aos dados das amostras de água do Brasil

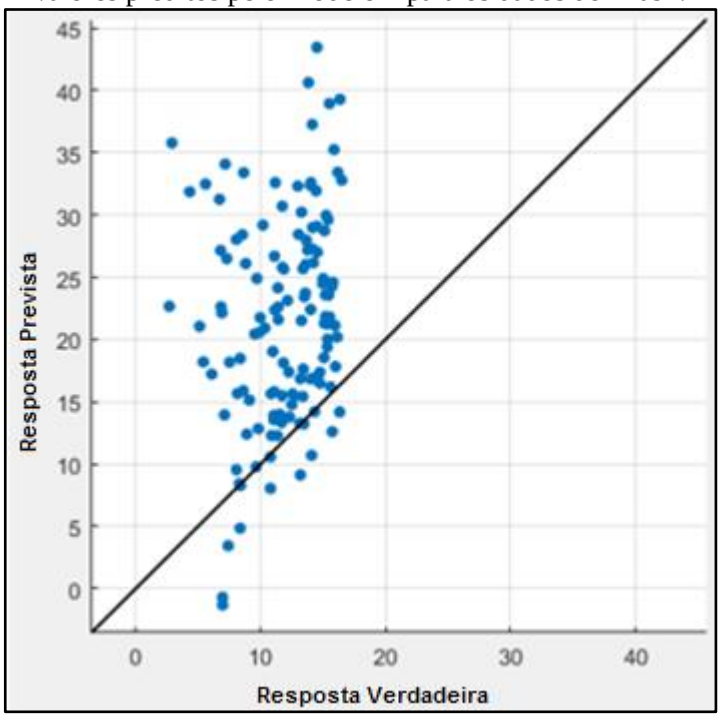
A Tabela 27 mostra os resultados obtidos para as diferentes funções Kernel utilizadas para o treinamento da SVM com os dados das amostras de água do Brasil, nas quais a variável inferenciada é a Clorofila. Podemos perceber de forma geral que o coeficiente de correlação de Pearson apresentou valores entre 0,34 e 0,65, denotando o desempenho inferior da SVM para o conjunto de dados do Brasil, quando comparamos ao desempenho apresentado pela RNA. A função Quadrática foi aquela que apresentou melhor desempenho denotado pelo maior valor de r apresentado (0,65) e os menores valores dos demais erros (MSE, RMSE e MAE).

Tabela 27 - Resultados obtidos para as diferentes funções Kernel utilizadas para o treinamento do modelo de clorofila via SVM com os dados das amostras de água do Brasil.

Modelo	Função Kernel	r	RMSE	MAE
1	Linear	0,55	14,59	13,17
2	**Quadrática**	**0,65**	**12,86**	**10,37**
3	Cúbica	0,58	14,10	11,71
4	Fine Gaussian	0,34	17,65	16,84
5	Medium Gaussian	0,62	13,42	11,98
6	Coarse Gaussian	0,57	14,27	12,92

A Figura 58 mostra um gráfico que compara a correlação entre as medidas observadas (valores de clorofila experimentais) e os valores preditos pelo Modelo 2. A Figura 58 mostra uma baixa correlação entre os valores preditos e os experimentais, comprovando que o modelo SVN não foi satisfatório.

Figura 58 - Gráfico que compara a correlação entre as medidas observadas (valores de clorofila experimentais) e os valores preditos pelo Modelo 2 para os dados do Brasil.



4.3.1.2 SVM aplicada aos dados das amostras da China

Seguindo conforme a aplicação realizada na seção anterior, para os dados do Brasil, realizou-se a análise para o conjunto de dados da China, para os quais a variável a ser inferenciada é a DBO. A Tabela 28 mostra os resultados obtidos para o treinamento das diferentes funções Kernel avaliadas neste trabalho. A função *Kernel* que apresentou o melhor desempenho para esse conjunto de dados foi a *Coarse Gaussian*.

Tabela 28 - Resultados obtidos para as diferentes funções Kernel utilizadas para o treinamento do modelo de DBO via SVM com os dados das amostras de água da China.

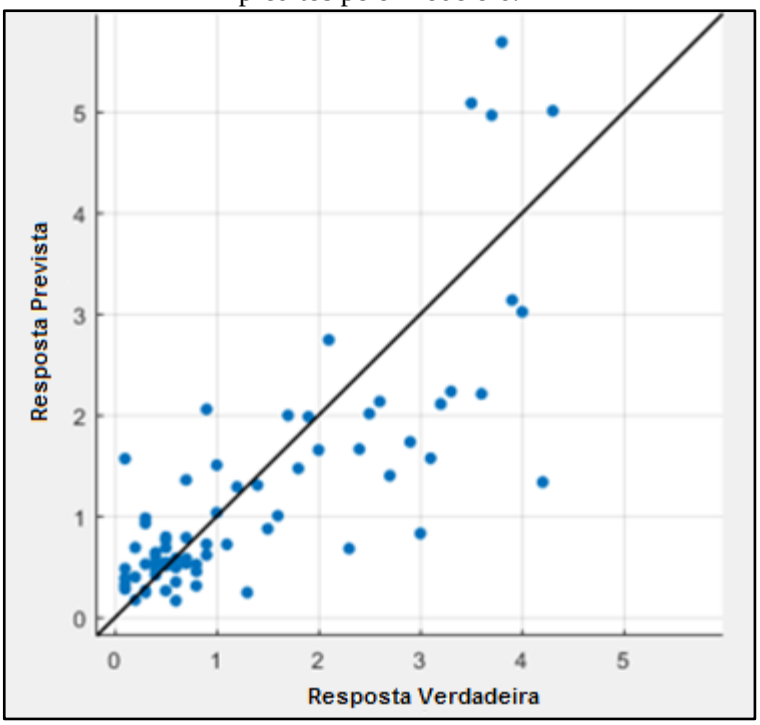
Modelo	Função Kernel	r	RMSE	MAE
1	Linear	0,53	0,92	0,61
2	**Quadrática**	**-20,68**	**6,21**	**1,59**
3	**Cúbica**	**-29,45**	**7,36**	**2,29**
4	Fine Gaussian	0,52	0,92	0,66
5	Medium Gaussian	0,6	0,85	0,56
6	**Coarse Gaussian**	**0,64**	**0,8**	**0,56**

Ao analisarmos a Tabela 28 podemos perceber que assim como observado nos dados do Brasil, os valores de r apresentados para os dados da China também não foram tão bons quanto aqueles apresentados pela RNA. Vale salientar ainda os valores negativos de r para as funções Quadrática (melhor resultado nos dados do Brasil) e Cúbica, esses valores denotam que uma mesma função Kernel pode apresentar resultados de desempenho extremamente diferente para diferentes conjuntos de dados. Além disso, fica evidenciado que a natureza da função Kernel interfere quantitativamente no desempenho da SVM.

É possível perceber também que apesar da SVM ser mencionada na literatura (Li *et al*., 2016; Djerioui *et al*., 2018; Chou *et al*., 2018) como uma ferramenta que pode apresentar bons resultados para poucos dados, esse comportamento não foi observado para os dados do Brasil e China. Sendo assim, fica claro que nesta aplicação o problema não foi a quantidade de dados, mas sim talvez a variedade e qualidade dos dados utilizados, já que esses apresentam baixa amplitude de valores nas suas varáveis. Não podemos deixar de destacar também que a ferramenta se SVM para regressão foi implantada recentemente no pacote MatLab e não tem-se muito relatos da avaliação de sua eficiência. Desta forma, podemos concluir que, neste trabalho, não foi possível obter bons resultados para o conjunto de dados do Brasil e China utilizando essa ferramenta.

A Figura 59 mostra o gráfico com uma comparação entre os valores observados de DBO e os preditos pelo Modelo 6. A Figura 59 mostra uma baixa correlação entre os valores preditos e os experimentais, comprovando que o modelo de DBO via SVM não foi satisfatório.

Figura 59 - Gráfico que compara a correlação entre as medidas observadas (valores de DBO experimentais) e os valores preditos pelo Modelo 6.



4.3.1.3 SVM aplicada aos dados das amostras de água da Espanha

A Tabela 29 apresenta os resultados obtidos para o treinamento de diferentes funções Kernel relativas ao conjunto de dados da Espanha, no qual a variável inferenciada é a Cianotoxina. Deve-se salientar que não se tinha a expectativa de obterem-se resultados satisfatórios para os dados da Espanha pelos problemas citados quando da modelagem com Redes Neurais.

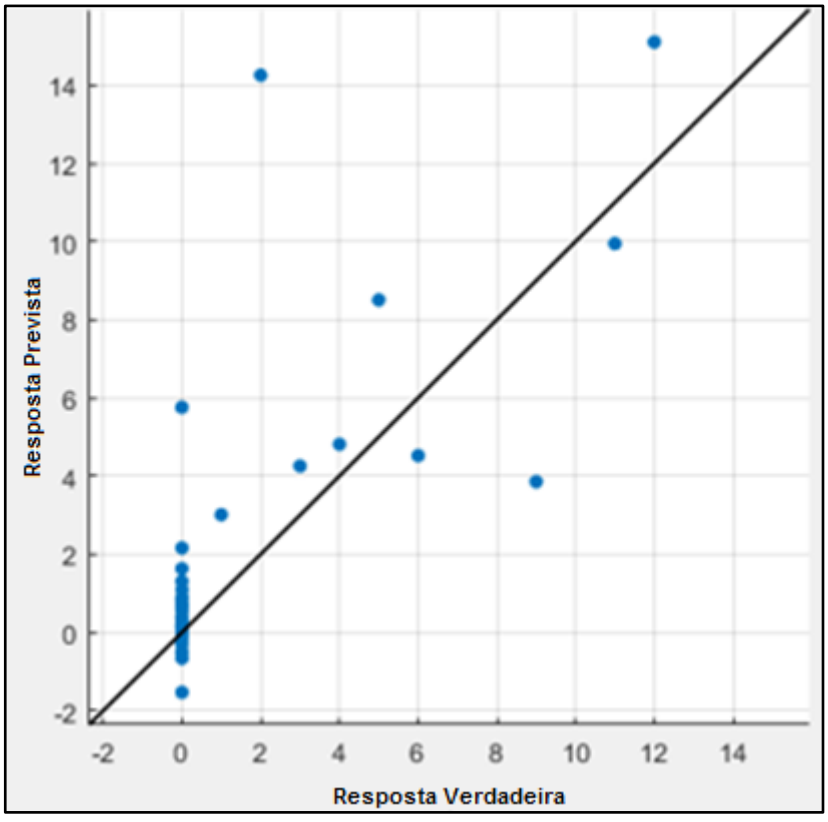
Tabela 29 - Resultados obtidos para as diferentes funções Kernel utilizadas para o treinamento do modelo de ciantoxinas via SVM com os dados das amostras de água da Espanha.

Modelo	Função Kernel	r	RMSE	MAE
1	Linear	0,96	24,54	12,26
2	Quadrática	0,98	18,69	9,16
3	Cúbica	0,48	88,02	31,83
4	Fine Gaussian	0,99	2,85	2,19
5	Medium Gaussian	0,99	6,26	3,44
6	Coarse Gaussian	0,99	2,65	1,44

Analisando os resultados apresentados na Tabela 29, podemos perceber que diferente do que ocorreu para o conjunto de dados do Brasil e China, a SVM apresentou um bom desempenho na modelagem dos dados da Espanha. Esse fato é denotado principalmente pelos valores do coeficiente de correlação de Pearson para as funções Kernel testadas, os quais estiveram bem próximo de 1, com exceção da função cúbica.

Vale salientar ainda que a função Kernel que apresentou o melhor desempenho foi a *Coarse Gaussian*, denotado pelo maior valor do coeficiente de correlação de Pearson e os menores valores dos erros (MSE, RMSE, MAE) apresentados durante o treino. Desta forma, o Modelo 6 (*Coarse Gaussian*) pode ser utilizado para a predição da presença de Cianotoxinas. A Figura 60 mostra a correlação entre os valores observados da concentração de Cianotoxinas e os valores preditos pelo Modelo 6.

Figura 60 - Gráfico da correlação entre as medidas observadas (valores de Cianotoxinas experimentais) e os valores preditos pelo Modelo 6.



O coeficiente de correlação de Pearson (r) e a Raiz do Erro Quadrático Médio (RMSE) que foram calculados para os conjuntos de dados de treino e de teste utilizados para os três modelos (Clorofila, DBO e Cianotoxina) são apresentados na Tabela 30 a seguir.

Tabela 30 - RMSE e R dos modelos do tipo *Suport Vector Machine* (SVM) para os modelos de Clorofila (Brasil), DBO (China) e Cianotoxina (Espanha).

Modelo	**RMSE**	**r**
	Treinamento + Teste	**Treinamento + Teste**
Clorofila (Brasil)	12,86	0,65
DBO (China)	0,8	0,64
Cianotoxina (Espanha)	2,65	0,99

4.4 NEURAL-FUZZY

Devido a dificuldades na utilização da interface *Neuro-Fuzzy*, disponível no MatLab para a utilização, são reportados aqui os dados sem uma representação gráfica destes resultados. Assim na Tabela 31, o (r) e (RMSE), apresentado, foram calculados para os conjuntos de dados de treino e de teste utilizados para os três modelos (Clorofila, DBO e Cianotoxina). E essa ferramenta do NF e a de SVM reportam os critérios de erros para todos os dados conjuntamente.

A Tabela 31 apresenta o resultado geral da aplicação dos modelos de máquinas *neuro-fuzzy* para dados estacionários.

Tabela 31 - Resultado do desempenho dos modelos de máquinas *neuro-fuzzy* para os modelos de Clorofila (Brasil), DBO (China) e Cianotoxina (Espanha).

Modelo	**RMSE**	**r**
Brasil	3,998	0,9396
Espanha	6481,2	0,9715
China	0,13	0,7459

O Resultado dos modelos de máquinas *neuro-fuzzy* para os dados estacionários do Brasil e da Espanha possuíram maiores coeficientes de correlação de Pearson (r). Contudo no modelo de clorofila do Brasil o Erro Quadrático Médio (RMSE) foi de 3,99, sendo considerado pequeno em relação à média dos dados de clorofila que foi 27,3, conforme mostrado na Tabela 7. Por outro lado, o erro RMSE do modelo de cianotoxinas da Espanha foi 6481, que é totalmente insatisfatório quando comparado ao valor médio de cianotoxinas de 93,4, apresentado na Tabela 9.

Embora uma melhor avaliação da aplicabilidade das máquinas de aprendizado tenha ficado um pouco comprometida diante das limitações das ferramentas utilizadas, verifica-se que diante de dados com diferentes características, as redes neurais, por seu ambiente permitir uma melhor avaliação, pareceram mais promissoras.

Nos trabalhos consultados na literatura de uma forma geral percebe-se essa ambiguidade em que não se tem aparentemente uma conclusão de que dentre essas três técnicas aqui testadas qual a mais indicadas para dados de qualidade de água, provavelmente por causa da dependência das características dos dados. Em consonância com a literatura atual, parece ser intuitivo o interesse em utilizar estas diferentes técnicas presentes no MatLab para aplicar a um mesmo conjunto de dados de forma que possa ser escolhido o melhor modelo que se ajusta as dados disponíveis. Observa-se também na literatura a tendência em utilizar para a inferenciação e até também classificação de um conjunto de técnicas na forma das denominadas *enssemble* e comite de máquina que são nada mais que se utilizar de diferentes máquinas de aprendizado tomando-se como resultado diferentes métricas da obtida inferenciação ou classificação

4.5 MODELAGEM DE SÉRIE TEMPORAL COM REDE NEURAL ARTIFICIAL, NEURO-FUZZY, MÁQUINA DE VETORES DE SUPORTE EM CONJUNÇÃO COM A TRANSFORMADA WAVELET.

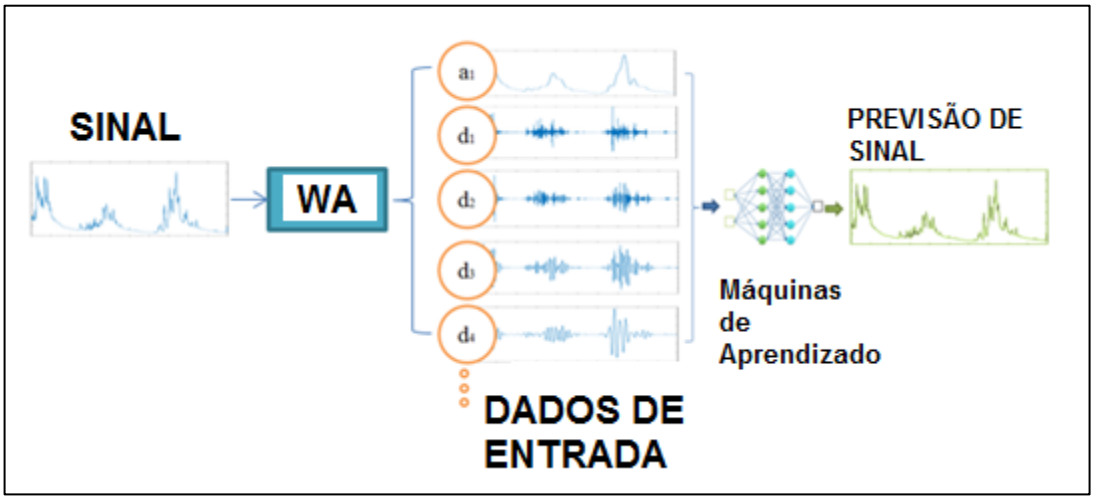
Diferentes tipos de modelos foram utilizados na avaliação da aplicabilidade das máquinas de aprendizado em prognóstico de séries temporais com um número reduzido de amostras.

Primeiramente foi testado o modelo com as máquinas de aprendizado utilizadas como máquinas regressoras, ou seja, os dados de entrada são entradas atrasadas da própria variável que se quer fazer o prognóstico.

Depois duas das arquiteturas de conjunção que envolve o uso da *wavelet* acoplados às máquinas de aprendizado apresentadas na literatura (NOURANI *et al*., 2013, 2014; BARZEGAR *et al*., 2018) são apresentadas nas Figuras 61 e 62. O primeiro, denominado aqui de modelo 2 (Figura 61), representa um modelo tradicional bastante apresentado na literatura (NOURANI *et al*., 2013, 2014; BARZEGAR *et al*., 2018). Esse modelo consiste basicamente em utilizar as subséries de decomposição *wavelet* do sinal, série temporal original, como

dados de entrada nas máquinas de aprendizagem. E como resposta ao treinamento, teremos os dados de sinal original.

Figura 61 - Primeiro modelo (denominado de modelo 2) de acoplagem das máquinas com a transformada wavelet.



O modelo 3 (Figura 62), menos usual que o primeiro, também está presente na literatura (NOURANI et al., 2013, 2014; BARZEGAR et al., 2018). Procede através da combinação das máquinas de aprendizado para a previsão dos componentes de frequência do sinal (série temporal). Combinando posteriormente os valores simulados dos componentes para a reconstrução do sinal original, por meio de uma técnica *wavelet* chamada reconstrução, que funciona basicamente como o processo inverso da decomposição. Em outras palavras, a reconstrução é a soma dos componentes de frequência simulados.

Figura 62 - Segundo modelo (modelo 3) de acoplagem das máquinas com a transformada *wavelet*.

Em relação ao nível de decomposição a ser considerado, pois um método de tentativa e erro para determiná-lo seria demasiadamente exaustivo, foi utilizado o critério de Nourani *et al* (2014) em que N é o número de amostras e F o nível de decomposição a ser utilizado para decomposição da série original fornecendo os coeficientes *wavelet*s a serem considerados como entrada nas máquinas de aprendizado.

$$t[\log N](73) \qquad (71)$$

Em relação aos critérios de avaliação para análise da eficiência das técnicas, foram escolhidos o coeficiente de determinação quadrado (R^2) e o (RMSE) que foram calculados para os conjuntos de dados de treino e de teste utilizados para todos os modelos.

Vale salientar que nesta etapa do trabalho foi necessário a construção de programas em linguagem MatLab que pudessem conectar as rotinas de tratamento *wavelet* de sinais e as máquinas de aprendizado, tarefa essa que foi uma etapa complexa na realização desse trabalho.

A eficiência da conjunção da transformada *wavelet* com as máquinas de aprendizado está diretamente relacionada com a escolha da família *wavelet*, dessa forma se fez necessário testar qual se adequa mais à aplicação na série estudada daquelas disponibilizadas na literatura (NOURANI *et al*., 2013, 2014; BARZEGAR *et al*., 2018). Mais uma vez ficou-se limitada às famílias inseridas no ambiente MatLab construído.

Os modelos referidos serviram para avaliar a conjunção das máquinas de aprendizado, RNA's, NF e SVM, e a transformada *wavelet* observando sua eficiência em relação à mesma atuando como modelos de autorregressivo. Sem a referida conjunção com a transformada estes modelos foram aqui denominados de modelo 1.

Uma questão determinante no trabalho foi a questão dos dados já que essa é muito relevante no quesito dificuldade pois ao longo desta pesquisa percebeu-se a dificuldade de obter os mesmos. Deve-se salientar que na literatura consultada e reportada no item "Estado da Arte", investigações com semelhantes número de amostras foram realizadas por Diamantopoulou, *et al*. (2005) com previsão da qualidade de água com valores mensais com 12 parâmetros em um período de 1980-1994 (168 amostras); Singh *et al.* (2009) na previsão de Oxigênio Dissolvido (OD) e da Demanda Bioquímica de Oxigênio (DBO) em dados mensais medidos durante um período de 10 anos (120 amostras); Faruk (2010) com previsões mensais trabalhando com Oxigênio Dissolvido (OD) em108 meses durante 1996-2004;Ravansalar (2015) na previsão de Condutividade Elétrica (CE) em rios com 274

conjuntos de dados mensais (1984-2008); Zhang (2016) na previsão de séries temporais de concentração de nitrogênio total e fósforo total com dados mensais entre 2006 e 2011 (60 amostras); Haghiabi *et al*. (2018) com o estudo do prognóstico da condutividade específica (EC) em 55 anos (660 dados) e Alizadeh *et al*. (2015) previsão de valores diários (650) e horários (540) de salinidade da água com 4 anos de dados. Também foi observado que se vêm investigando o potencial das máquinas de aprendizado na modelagem com o número de dados restritos. Para fazer esta investigação dados diários foram utilizados, pois representavam o desafio de se tentar modelar com restrição de dados. Porém quando os modelos não se ajustavam foram utilizados dados horários para avaliar a aplicabilidade das técnicas.

Dessa forma os resultados serão apresentados primeiramente por máquinas de aprendizado (RNA'S, NF e SVM) utilizadas como modelos regressores, sem a conjunção *wavelet*. Para essa etapa da análise foram utilizados 85 % dos dados para treinamento e 15% para teste, salientando que estes últimos não participam do processo de treinamento e/ou testes das diferentes técnicas. Em seguida são testados os modelos 2 e 3.

4.5.1 Redes Neurais Artificiais, Neuro-Fuzzy e SVM para o Modelo 1

Neste modelo foram avaliadas a eficiência das máquinas de aprendizado (RNA's, NF e SVM) sem a transformada wavelet.

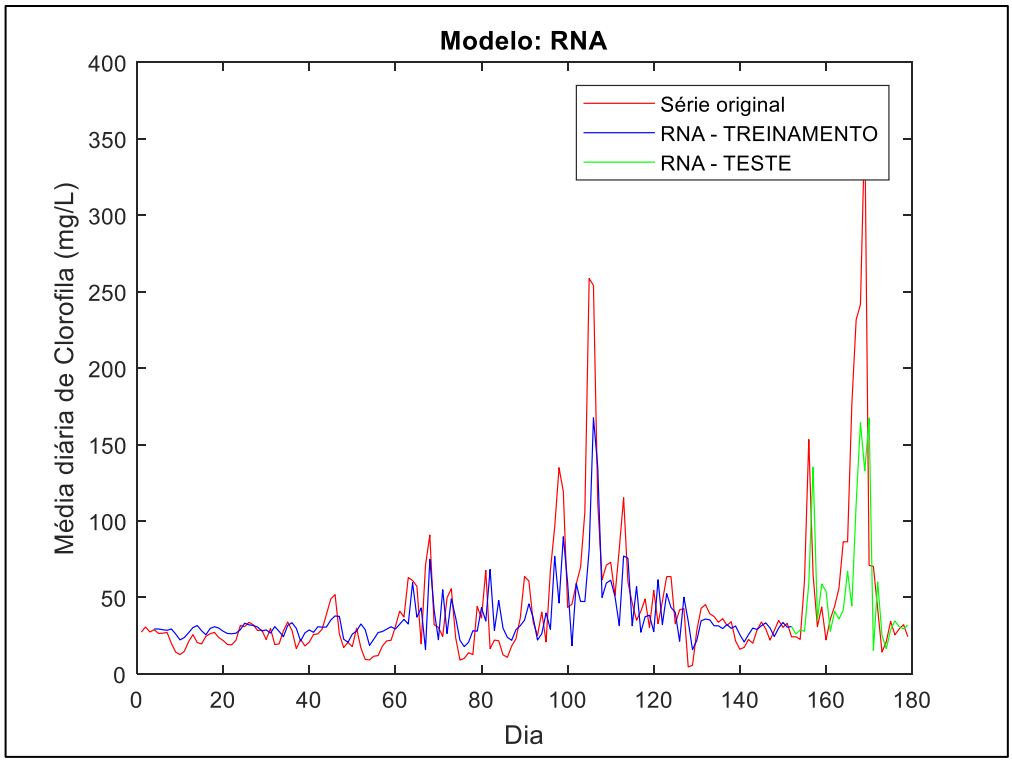
4.5.1.1 Redes Neurais Artificiais

A tabela 32 reporta os resultados para os modelos de máquinas de aprendizado RNA's que foram testados com passos de regressão (atraso) variando de 1 a 5 avaliando os algoritmos de Levenberg-Marquardt e de Regularização Bayesiana. A figura 63 mostra a série original e o resultado da simulação com a rede neural de melhores parâmetros determinada através das ferramentas estatísticas discutidas e presentes na Tabela 29, dividindo em dados de treinamento e teste. Estes parâmetros foram o algoritmo de aprendizagem Levenberg-Marquadt e nº de atraso igual a 3.

Tabela 32 - Resultados para os modelos de máquinas de aprendizado RNA's para a série temporal de clorofila.

Passo de Tempo	**Levenberg - Marquadt**		**Regularização Bayesiana**	
	Treino			
	RMSE	**R²**	**RMSE**	**R²**
1	5,911355175	0,917984	5,570679492	0,919789
2	5,194079322	0,935357	5,089051975	0,936868
3	**4,754350639**	**0,945221**	1,370696903	0,949967
4	5,265952905	0,933098	4,164635638	0,958229
5	4,902530979	0,950154	4,245781436	0,955457
Passo de Tempo	**Levenberg - Marquadt**		**Regularização Bayesiana**	
	Teste			
	RMSE	**R²**	**RMSE**	**R²**
1	4,994448919	0,931786	7,866462356	0,878887
2	6,217149668	0,89261	6,605561293	0,890754
3	**5,501799706**	**0,92833**	6,144157387	0,91915
4	5,800386194	0,948519	7,683456904	0,86263
5	5,559541348	0,917947	8,302305704	0,855587

Figura 63 - Gráfico mostrando a série original e o resultado da simulação com a rede neural de melhores parâmetros para a série temporal de clorofila, dividindo em dados de treinamento e teste.



De uma forma geral pode-se dizer que os modelos se ajustaram bem e a técnica realmente aprendeu sobre a relação temporal do fenômeno, visto que para os dados de teste os resultados foram satisfatórios, obtendo-se valores de r^2 de 0,92833 e RMSE de 5,501799706, na figura 63 pode também ser observada essa eficiência através das linhas verdes.

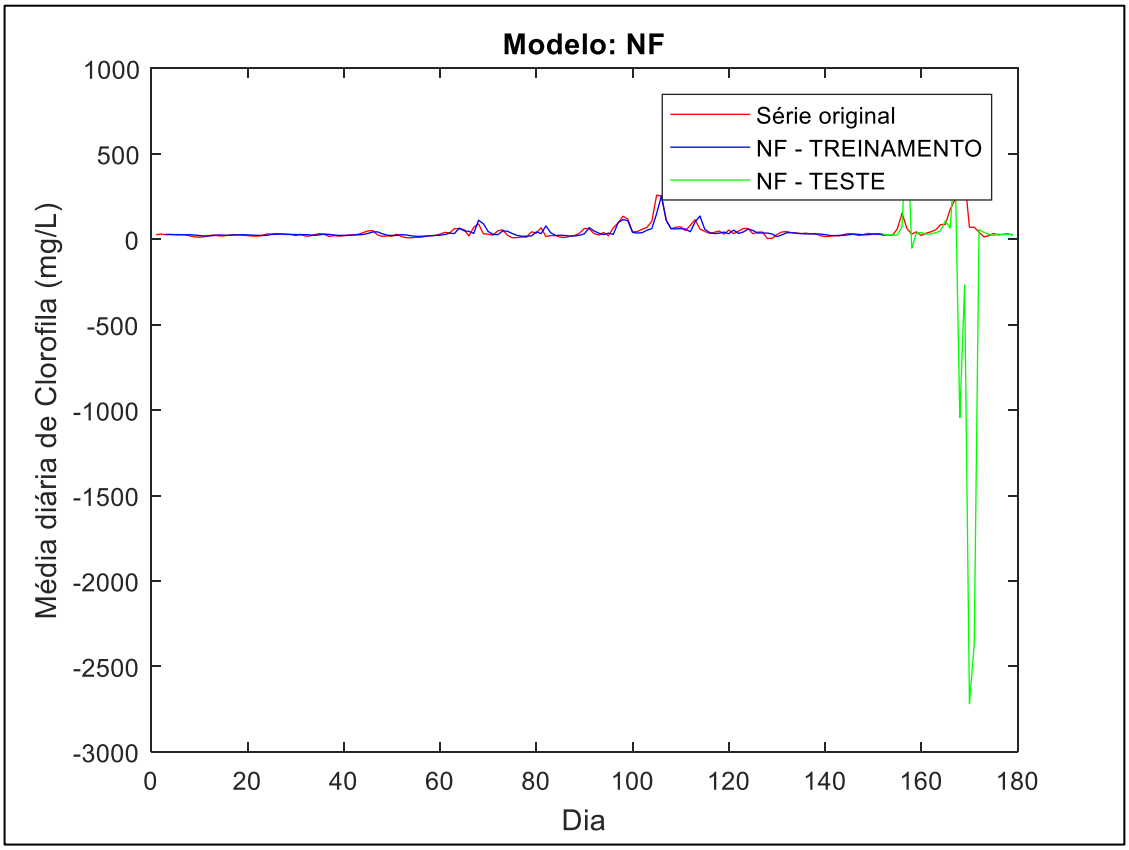
4.5.1.2 Neuro-Fuzzy

A Tabela 33 reporta os resultados para os modelos de máquinas de aprendizado *Neuro-Fuzzy* (NF) que foram testadas com passos de regressão variando de 1 a 5. A Figura 64 mostra a série original e o resultado da simulação através do NF, ferramenta do MatLab para utilização do sistema NF, com os melhores parâmetros determinados através das ferramentas estatísticas discutidas e presentes na Tabela 33, divididos em dados de treinamento e teste. O único parâmetro ajustável e determinado foi o número de atraso igual a 2.

Tabela 33 - Resultados para os modelos de máquinas de aprendizado *Neuro-Fuzzy* testadas com passos de regressão variando de 1 a 5.

Passo de Tempo	RMSE - Treino	R2 - Treino	RMSE - Teste	R2 -Teste
1	4,657681827	0,76795	11,48869009	0,8391
2	**4,043266007**	**0,83201**	**10,21420579**	**0,85909**
3	3,151428248	0,9012	26,83766011	0,66995
4	2,270044052	0,94952	106,2261738	0,70723
5	0,403868791	0,99845	2426,93222	0,55247

Figura 64 - Gráfico mostrando a série original e o resultado da simulação com a NF de melhores parâmetros para a série temporal de clorofila, dividindo em dados de treinamento e teste.



De uma forma geral pode-se dizer que os modelos se ajustaram bem na fase de treinamento, porém na fase de teste não, isso se deve ao fato das limitações impostas ao se utilizar Neuro-Fuzzy através da interface MatLab e não nas linhas de programação.

4.5.1.3 Máquina de Vetores de Suporte - SVM

As Tabelas 34, 35 e 36 reportam os resultados para os modelos de Máquina de Vetores de Suporte (SVM) que foram testadas com passos de regressão variando de 1 a 5 e variando as funções de Kernel. A figura 65 mostra a série original e os resultados da simulação, para os melhores parâmetros encontrados, separados pelas etapas de treinamento e teste, neste caso os parâmetros foram a função de Kernel RBF e número de atraso igual a 2.

Tabela 34 - Resultados para os modelos de máquinas de aprendizado SVM para a série temporal de clorofila testadas com passos de regressão variando de 1 a 5 e função RBF.

RBF				
Passo de tempo	**RMSE - Treinamento**	**R^2 - Treinamento**	**RMSE - Teste**	**R2 - Teste**
1	27.543	0.40632	24.354	0.19565
2	**26.861**	**0.43921**	**20.378**	**0.43686**
3	27.774	0.40436	22.219	0.33049
4	26.201	0.47352	22.218	0.33061
5	26.475	0.46591	22.509	0.31295

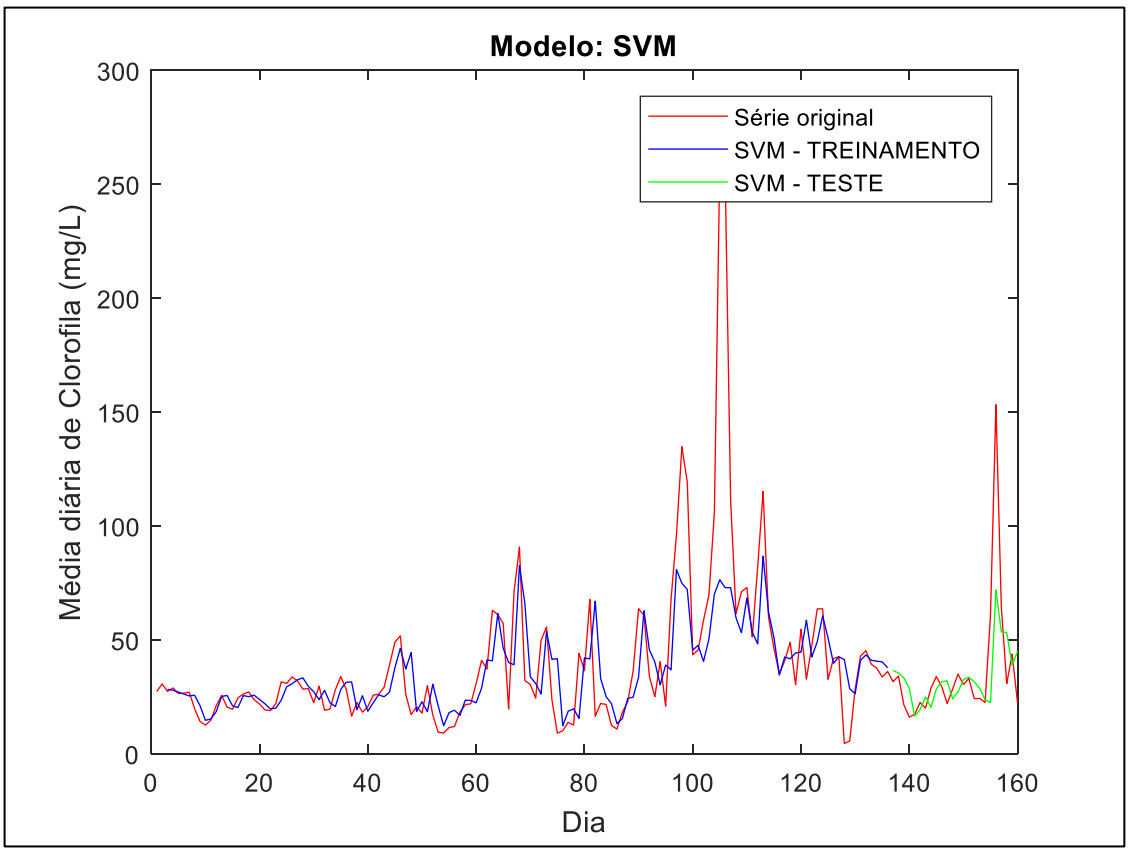
Tabela 35 - Resultados para os modelos de máquinas de aprendizado SVM para a série temporal de clorofila testadas com passos de regressão variando de 1 a 5 e função Polinomial De Grau 3.

POLINOMIAL DE GRAU 3				
Passo de tempo	**RMSE - Treinamento**	**R2 - Treinamento**	**RMSE - Teste**	**R2 - Teste**
1	27.348	0.4147	27.711	-0.041344
2	21.345	0.6459	31.559	-0.35059
3	21.598	0.6398	28.866	-0.12995
4	22.575	0.60915	31.647	-0.35816
5	22.687	0.60781	28.339	-0.089104

Tabela 36 - Resultados para os modelos de máquinas de aprendizado SVM para a série temporal de clorofila testadas com passos de regressão variando de 1 a 5 e função Polinomial De Grau 1 (continua).

POLINOMIAL DE GRAU 1 (LINEAR)				
Passo de tempo	**RMSE - Treinamento**	**R2 - Treinamento**	**RMSE - Teste**	**R2 - Teste**
1	24.255	0.53959	26.825	0.024166
2	23.891	0.55639	25.982	0.084561
3	25.346	0.50395	24.564	0.18174
4	28.42	0.38054	25.666	0.10669
5	28.225	0.39298	25.343	0.12906

Figura 65 - Gráfico mostrando a série original e o resultado da simulação com a SVM de melhores parâmetros para a série temporal de clorofila, dividindo em dados de treinamento e teste.



De uma forma geral pode-se dizer que os modelos com SVM se ajustaram bem na fase de treinamento, porém na fase de teste não tão bem e pode-se dizer que se deve ao fato das limitações impostas ao se utilizar a interface do MatLab.

Observa-se que dentre as três máquinas de aprendizado utilizadas as RNA's foram que apresentaram melhor resultados para o treinamento e principalmente para o teste. Vale salientar que a SVM para Regressão foram implantadas no pacote computacional mais recentemente, sendo sua utilização ainda pouca explorada.

Deve-se observar que os conjuntos de dados utilizados são diferentes, sendo para RNA e NF, 179 dados, e 160 para SVM. Os últimos 19 dados do SVM foram retirados devido à qualidade dos resultados para teste obtidos inicialmente com o conjunto de dados total. O motivo foi porque neste conjunto de dados para teste há um pico muito alto e as técnicas estatísticas e de sistemas inteligentes aqui propostas aprendem e generalizam pela experiência e tal comportamento está presente apenas na fase de teste, tornando-se uma tarefa complexa. A conjunção com a transformada *wavelet* foi testada para melhorar a eficiência de cada técnica.

4.5.2 RNA, NF e SVM com a Transformada *Wavelet* (modelo2)

Neste modelo foram avaliadas a eficiência das máquinas de aprendizado (RNA's, NF e SVM) com a transformada wavelet.

4.5.2.1 RNA com Wavelet

Na utilização das *wavelet*s, um parâmetro determinante é a família *wavelet* a ser utilizada, portanto, nessa análise foras variadas às famílias *wavelet*s mais reportadas na literatura (NOURANI *et al*., 2013, 2014; BARZEGAR *et al*., 2018)e disponíveis na ferramenta computacional utilizada.

Utilizou-se 85% dos dados para treinamento e 15% para teste. Novamente, como no SVM puro, utilizou-se de 160 dados devido ao pequeno número de dados e o comportamento altamente imprevisível na fase de teste.

Deve-se salientar que devido às ferramentas computacionais utilizadas serem diferentes das que foram aplicadas para a conjunção das *wavelet*s com as RNA's e o sistema NF não permitiam o atraso de mais de um passo em detrimento a da SVM que permitia. Dessa forma na aplicação *wavelet* com SVM foi possível variar os atrasos no modelo regressivo (Delay).

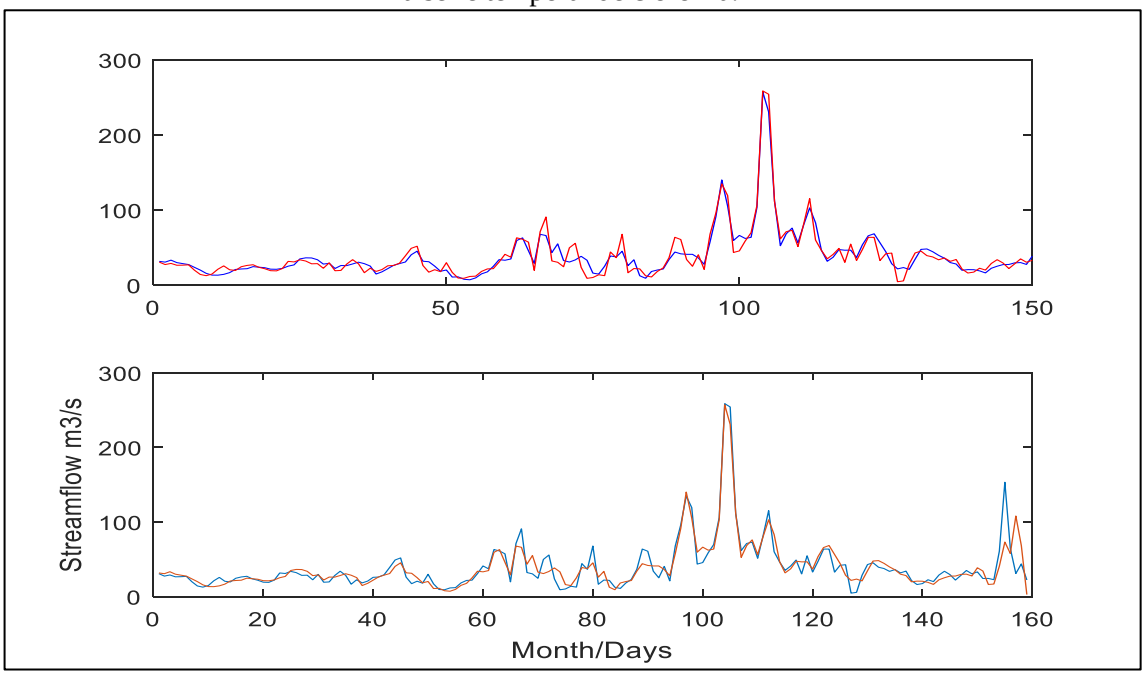
A Tabela 37 e Figura 66 a seguir mostram os resultados de desempenho (R^2 e RMSE) no treinamento e teste da Rede Neural Artificial com a Transformada *Wavelet*, para o modelo

2. Os resultados não foram satisfatórios porque os erros RMSE do conjunto de teste ficaram acima da média dos dados de clorofila que foi 27,3.

Tabela 37 - Resultados do desempenho para RNA com a Transformada *Wavelet* (modelo2) para a série temporal de clorofila.

Modelo	**Algoritmo de Treinamento**	**R^2 Treinamento**	**RMSE Treinamento**	**R^2 Teste**	**RMSE Teste**
	LM	0,9672	12,8659	0,6885	46,6571
Dmey	**LBr**	0,9723	21.5632	0,5985	38,4567
	LM	**0,8481**	**18,9143**	**0,6673**	**24,1032**
Haar	**LBr**	**0,8043**	**23,6782**	**0,76505**	**37.5649**
	LM	0,9589	14,614	0,6345	59,9108
Coif1	LBr	0,9583	27,9996	0,8678	50,2399
	LM	0,9696	15,3887	0,6154	45,8976
Coif3	LBr	0,9081	14,5762	0,8651	43,2367
	LM	0,9308	11,4303	0,9885	46,2254
Coif5	LBr	0,9768	8,8228	0,8543	65,5845
	LM	0,8113	22,1811	0,8171	44,5627
Bior1.1	LBr	0,9581	10,477	0,5277	144,1981
	LM	0,9569	11,0649	0,9467	160,497
Bior3.3	LBr	0,9696	9,05070	0,8105	45,7180
	LM	0,8659	19,1329	0,8441	41,1524
Db2	LBr	0,9765	9,3174	0,8409	54,1237
	LM	0,81502	20,2938	0,9655	56,9871
Db5	LBr	0,9613	19,6752	0,6885	68,564
	LM	0,8990	18,2821	0,8236	31,5632
Db10	LBr	0,96508	9,5131	0,8355	53,4963
	LM	0,9613	10,4969	0,8551	84,4969
Db20	LBr	0,9159	12,8274	0,9711	39,4591

Figura 66 - Resultados do desempenho para RNA com a Transformada *Wavelet* (modelo2) para a série temporal de clorofila.



4.5.2.2 Neuro-Fuzzy e Wavelet

Na análise dos dados temporais utilizando a arquitetura de conjunção *wavelet* com *Neuro-Fuzzy*, os melhores resultados ainda apresentaram valores de R^2 e RMSE muito insatisfatórios provavelmente devido ao número reduzido de dados que foram considerados diários. Nesta Etapa, para avaliar a técnica consideramos os dados horários ao contrário dos dados das médias diárias utilizadas nos outros modelos. Os resultados são apresentados na Tabela 38 e figura 67.

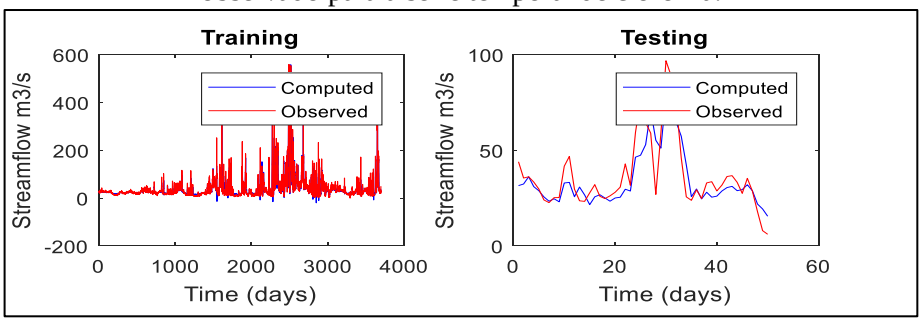
Os resultados não foram melhores que o modelo anterior (*RNA com Wavelet*) porque os erros RMSE do conjunto de teste ficaram abaixo dos resultados da média dos dados de clorofila que foi 27,3. De uma forma geral podemos dizer que o melhor resultado foi utilizando a *wavelt* mãe de Db10 em destaque na tabela 35.

Tabela 38 - Resultados do desempenho para NF com a Transformada *Wavelet* (modelo2) para a série temporal de clorofila.

Modelo1	**R^2 Treinamento**	**RMSE Treinamento**	**R^2 Teste**	**RMSE Teste**
Dmey	0,9266	12,6185	0,6704	10,7986
Haar	0,9011	14,4624	0,20486	11,7007
Coif1	0,9102	13,8288	0,5404	11,6020
Coif3	0,91249	11,8765	0,3554	12,7654
Coif5	0,9258	12,6698	0,5506	10,9690
Bior1.1	0,9011	14,4624	0,2048	11,3604
Bior3.3	O,9165	13,3719	0,4867	11,3604
Db2	0,8989	14,5955	0,6285	9,2836
Db5	0,9284	12,4716	0,5718	11,02466
Db10	**0,9279**	**12,4999**	**0,6888**	**9,5166**
Db20	0,9264	12,6231	0,5925	10,4388

A Figura 67 mostra os gráficos dos resultados da simulação, para treinamento e teste, comparando com o observado.

Figura 67 - Gráficos dos resultados da simulação via modelo NF com a Transformada Wavelet, para treinamento e teste, comparando com o observado para a série temporal de clorofila.



4.5.2.3 SVM e Wavelet

As Tabelas B1, B2 e B3 (Apêndice B) reportam os resultados para os modelos de SVM em que devido à disponibilidade nessa ferramenta pode-se explorar melhor. Desta forma foram testadas com passos de regressão variando de 1 a 5, variando as funções de Kernel e variando ainda as *wavelet*-mãe.

Dentre os parâmetros investigados nas Tabelas B1, B2 e B3, foi observado que os melhores, para o presente conjunto de dados e com o modelo híbrido descrito são função de Kernel polinomial de grau 3, nº de atraso igual a 4 e como *Wavelet*-mãe Coif3. Com estes parâmetros gerou-se a Figura 68 e a Tabela 39 que compara o resultado da simulação. Os resultados mostram que o modelo é pouco satisfatório, porém entre as três combinações foi o que se obteve melhores resultados.

Figura 68 - Gráfico do resultado da simulação e experimental para o modelo de SVM com *wavelet* para a série temporal de clorofila, dividido em teste e treinamento comparando com o sinal original.

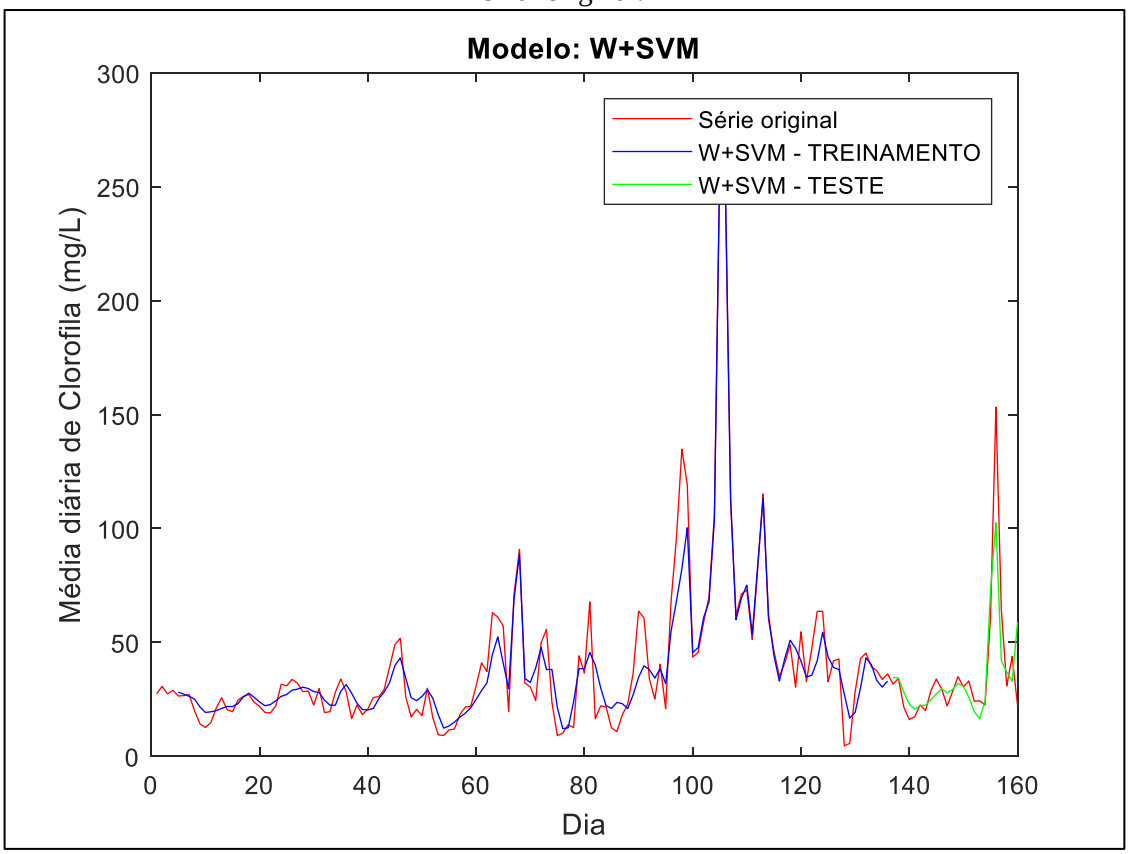


Tabela 39 - Resumo do desempenho para SVM com a Transformada *Wavelet* (modelo2 – Função Polinomial De Grau 3) para a série temporal de clorofila.

POLINOMIAL DE GRAU 3					
Wavelet - mãe	Passo de tempo	RMSE - Treinamento	R^2 - Treinamento	RMSE - Teste	R^2 - Teste
	1	11,7664	0,89165	39,2557	-1,08974
	2	6,92729	0,9627	38,0967	-0,96816
Coif3	3	8,60284	0,94285	24,6318	0,177225
	4	**9.781**	**0.92663**	**14.519**	**0.71413**
	5	8,84994	0,94032	15,2715	0,683737

4.5.3 Modelagem de Série Temporal com RNA, NF e SVM e TWD (modelo3)

Em relação às divisões dos dados, a princípio, os dados foram divididos entre dados para treinamento e teste. Em relação aos critérios de avaliação para a análise da eficiência das técnicas foram escolhido o coeficiente determinação (R^2) e a raiz do erro quadrático médio

(RMSE) que foram calculados para os conjuntos de dados de treino e de teste para os dados do Brasil (Clorofila).

A eficiência da conjunção da transformada *wavelet* com as máquinas de aprendizado está diretamente relacionada com a escolha da família *wavelet*, dessa forma se faz necessário testar no modelo 3 qual se adequa mais à aplicação na série estudada. Mais uma vez ficou-se limitada às famílias inseridas no ambiente MatLab construído.

4.5.3.1 Wavelet com Redes Neurais Artificiais (Modelo 3)

Utilizando o conjunto total de dados (179 dados), dividiu-se novamente em 85% para treinamento e 15% para teste. Para o modelo híbrido de *wavelet* com rede neural, variou-se o atraso e a *Wavelet*-mãe (Tabela B4 no Apêndice B).

Pela Tabela 40 pode-se concluir que os parâmetros que melhoravam e ajustavam o presente modelo e que compreendia melhor a dinâmica estudada foi através dos parâmetros com 5 atrasos e a *Wavelet*-mãe Db20. Tal modelo, com os presentes parâmetros são mostrados graficamente Figura 69. Pode-se dizer que o modelo de RNA com *wavelet* apresentou resultados satisfatórios para predição da série temporal de clorofila.

Figura 69 - Gráfico de representação do modelo RNA com a *wavelet* para a série temporal de clorofila.

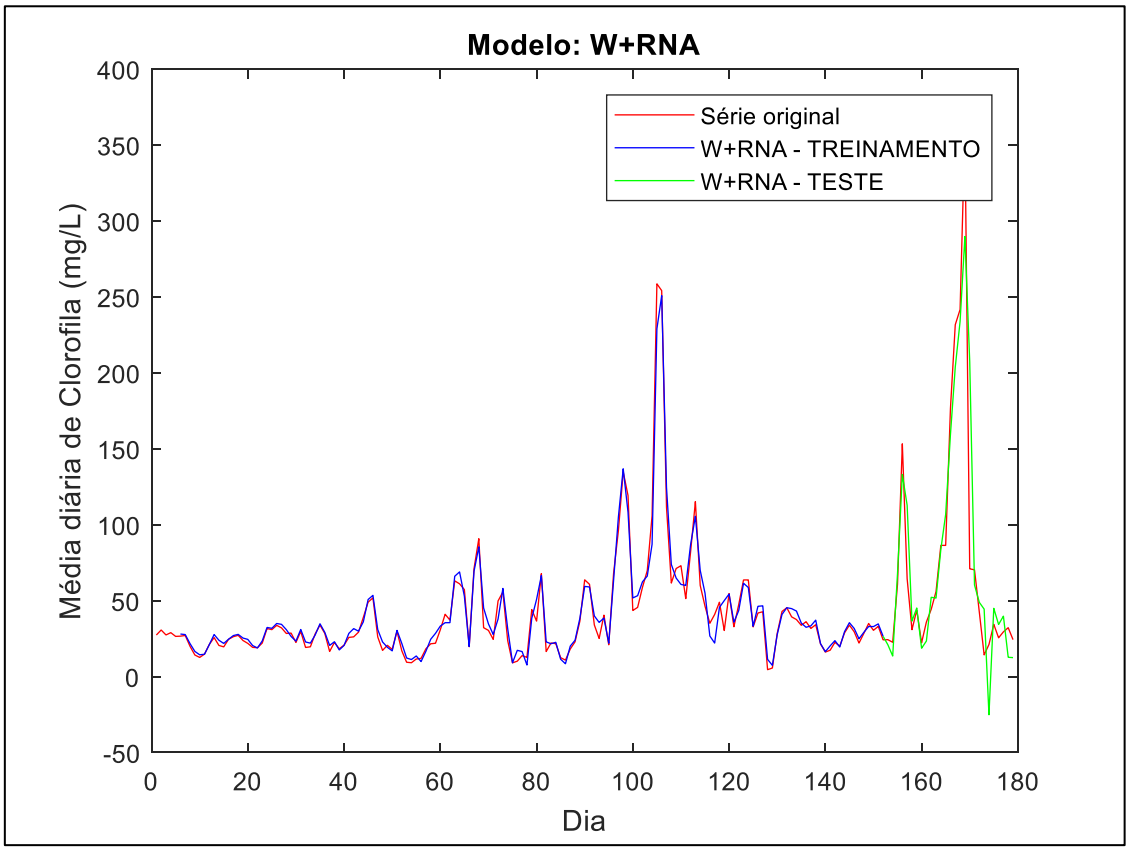


Tabela 40 - Melhor resultado com parâmetros de atraso 5 e a *Wavelet*-mãe Db20.

Wavelet - mãe	Atraso de tempo	RMSE - Treinamento	R^2 - Treinamento	RMSE - Teste	R^2 - Teste
	1	4,68295	0,77695	6,47881	0,94411
	2	4,77399	0,89104	6,65011	0,94264
Db20	3	1,50921	0,97825	6,74996	0,93397
	4	1,03995	0,99019	3,12867	0,98760
	5	**0,85026**	**0,99320**	**2,28134**	**0,99135**

4.5.3.2 Wavelet com Máquina De Vetores de Suporte (Modelo 3)

Para a mesma divisão dos dados, variou-se apenas o passo e a *Wavelet*-mãe (Tabela B5, Apêndice B). Note que, desta vez não foi possível variar a função de Kernel, pois a ferramenta computacional utilizada não permitia, portanto, foi utilizada a RBF padrão da ferramenta que é uma das muito utilizada segundo a literatura (LI *et al*., 2016; HAGHIABI *et*

al., 2018;CHOU *et al*., 2018;DJERIOUI *et al*., 2018).Foram testadas com passos de regressão variando de 1 a 5, variando ainda as *wavelet*-mãe disponibilizadas na ferramenta.

Dentre os parâmetros que foram ajustados, os que melhor entenderam a dinâmica investigada foram passe 2 e *Wavelet*-mãe de Meyer na forma discreta. Este modelo segue na Figura 70. Os resultados mostram que o modelo SVM com *wavelet* para a série temporal de clorofila não apresentou ajuste satisfatório.

Figura 70 - Gráfico de representação do modelo SVM com *wavelet* para a série temporal de clorofila, com parâmetros de passe 2 e *Wavelet*-mãe de Meyer na forma discreta.

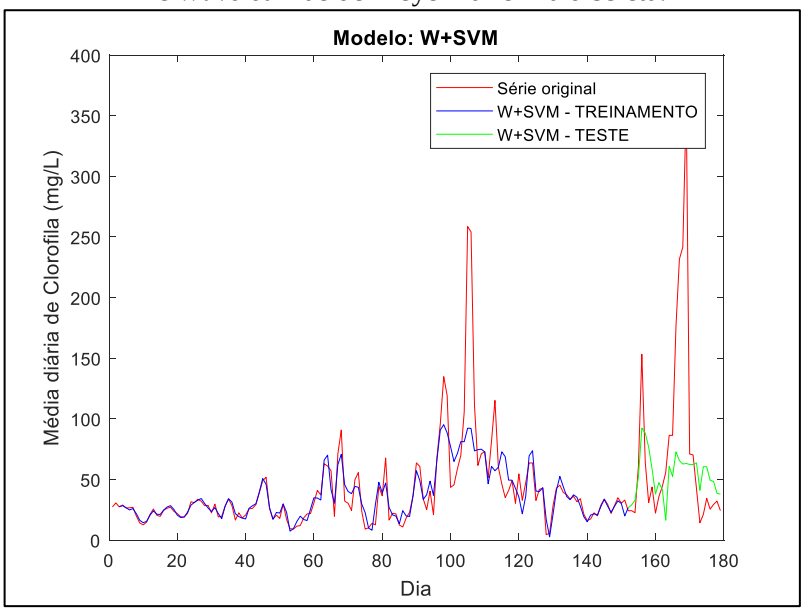


Tabela 41 - Resultados do desempenho da SVM com a *wavelet* para a série temporal de clorofila, variando a *wavelet*-mãe.

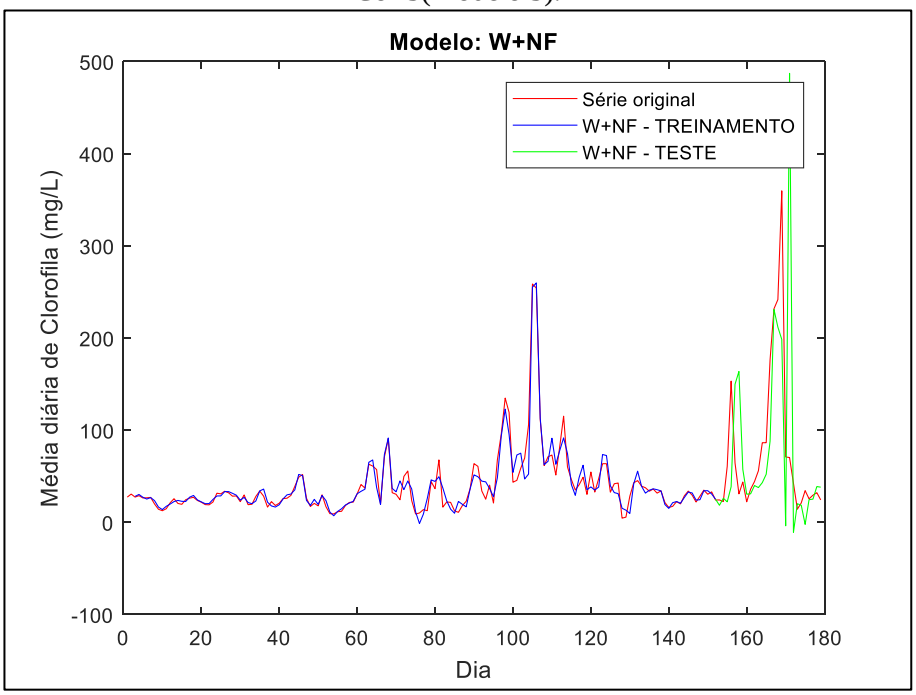
***Wavelet* - mãe**	**Passo de tempo**	**RMSE - Treinamento**	**R2 - Treinamento**	**RMSE - Teste**	**R2 - Teste**
	1	2,10556	0,96324	16,46876	0,38354
	2	**2,33641**	**0,94824**	**15,32123**	**0,64136**
Dmey	3	2,43900	0,94553	16,46269	0,45613
	4	1,94342	0,96831	16,64572	0,41456
	5	2,15260	0,96189	16,49606	0,37199

4.5.3.3 Wavelet com Neuro-Fuzzy (Modelo 3)

Analogamente em relação à divisão dos dados, para os modelos anteriores, variou-se o passo e a *Wavelet*-mãe (Tabela B6, Apêndice B).

Diante de todos os parâmetros ajustados, concluiu-se que os que geraram o melhor modelo preditivo foram passe 2 e Coif3e foram resumidos graficamente na tabela 42. Este resultado segue na Figura 71.

Figura 71 - Gráfico do melhor modelo preditivo com passe 2 e Coif3(Modelo 3).



Diante do que pode ser visto no modelo 3, híbrido, apenas os de RNA e NF conseguiram ter uma boa previsão. Ao longo de todos os modelos investigados, ainda mais, pelos parâmetros estatísticos e análise gráfica, o modelo 3 de redes neurais, para o problema dinâmico, demonstrou superioridade, compreendendo e generalizando de maneira eficaz a dinâmica da concentração da clorofila e, consequentemente, qualidade da água.

Tabela 42 - Resultados do desempenho da NF com a *wavelet*, variando a *Wavelet*-mãe (Modelo 3) para a série temporal de clorofila.

***Wavelet*-mãe**	**Passo de tempo**	**RMSE - Treinamento**	**R2 - Treinamento**	**RMSE - Teste**	**R2 - Teste**
	1	4,14898	0,82289	4,77630	0,96128
	2	1,86188	0,96687	4,65210	0,97311
Coif3	3	1,19227	0,98651	15,93142	0,69128
	4	0,63545	0,99616	28,34061	0,50709
	5	0,37344	0,99868	61,54835	0,32346

4.5.4 Limitações

Cabe, neste momento, ressaltar a dificuldade na obtenção de dados sobre a qualidade da água, principalmente em relação à quantidade de dados exigidos para implementação de ferramentas computacionais. Há necessidade de um esforço conjunto dos órgãos oficiais e centros de pesquisas em construir e manter um banco de dados atualizado sobre variáveis ambientais e mais especificamente qualidade da água. O Brasil e/ou Pernambuco deve investir num melhor monitoramento.

O que fica claro é que a dependência de ferramentas computacionais distintas para cada técnica é prejudicial e uma análise comparativa mais profunda e completa para se estabelecer uma técnica como sendo melhor que a outra é uma tarefa complexa, sem mencionar que pela própria natureza das técnicas utilizadas existem diversos parâmetros para determinar sua eficiência determinados empiricamente.

Vale salientar também, que neste trabalho foi necessário a construção de programas em linguagem MatLab que pudessem conectar as rotinas de tratamento *wavelet* de sinais e as máquinas de aprendizado, tarefa essa que foi uma etapa complexa na realização desse trabalho.

5 CONCLUSÕES

Observou-se a aplicabilidade de Análise de Componentes Principais (ACP) e Mapas Auto-organizáveis de Kohonen para a análise exploratória de dados verificando que foi possível observar as correlações e formações de grupos referentes a variáveis e amostras, assim como na detecção de amostra com pouca semelhança, que podem ser considerados *outliers* do banco de dados.

A degradação da qualidade da água do Brasil, China e Espanha foi principalmente atribuída à poluição por matéria orgânica e compostos nitrogenados, e consequentemente encontravam-se eutrofizadas, provavelmente devido aos lançamentos de efluentes sanitários nos corpos d água.

Os melhores modelos obtidos para os dados estacionários foram: para clorofila, *Neuro-Fuzzy* (R=0,9396 e RMSE = 3,998); para a DBO, RNA (R = 0,9349 e RMSE = 0,6720); e para a Cianotoxina, Máquina de Vetores de Suporte (R = 0,99 e RMSE =2,65). Isto mostra que o melhor tipo de modelo depende do sistema avaliado e dos parâmetros e dados disponíveis.

O uso de modelagem híbrida com uso da transformada *wavelet* na entrada dos modelos resultou na melhoria da previsão da série temporal de clorofila para série diária e horária e o melhor modelo de uma forma geral foi o modelo 3, ou seja Rede Neural Artificial com *Wavelet*-mãe Db20 e atraso 5, com desempenho do teste (R^2=0,99135 e RMSE=2,28134).-O presente trabalho mostrou que as modelagens e técnicas estatísticas e de inteligência artificial foram eficazes na análise de dados ambientais e podem ser utilizadas para melhoria dos modelos de gestão em caracterização e monitoramento da qualidade da água.

Há necessidade de um esforço conjunto dos órgãos oficiais e centros de pesquisas em construir e manter um banco de dados atualizado sobre variáveis ambientais e mais especificamente qualidade da água. O Brasil e/ou Pernambuco deve investir num melhor monitoramento.

6 SUGESTÕES DE TRABALHOS FUTUROS

Os seguintes trabalhos são sugeridos com base nos resultados da tese:

- Fazer uma abrangente pesquisa para obtenção de dados de qualidade de água de melhor qualidade e quantidade através de órgãos governamentais locais, nacionais e internacionais;
- Utilização de uma única plataforma que permita uma maior flexibilização da utilização das máquinas de aprendizado e de preferência que permita uma forma mais automática sem que o usuário tenha que interferir a cada modificação de algum parâmetro na busca de uma maior eficiência;
- Utilizar os Comités de Máquinas ou *ensemble* que são, como referidos anteriormente, um conjunto de máquinas de aprendizado que fazem a inferência ou prognóstico através de alguma métrica dos resultados individuais de cada uma. Vale salientar que na literatura consultada não se observou a existência de uma ferramenta comercial, ou de uso livre, dessa natureza.

REFERÊNCIAS

ADAMOWSKI, J. River flow forecasting using wavelet and cross-wavelet transform models. **Hydrol Process**, n.22, p. 4877–4891, 2008.

ADAMOWSKI, K.; PROKOPH, A.; ADAMOWSKI, J. Development of a new method of wavelet aided trend detection and estimation. **Hydrol Process,** n. 23, p. 2686–2696, 2009.

ADAMOWSKI, J., SUN, K. Development of a coupled wavelet transform and neural network method for flow forecasting of non-perennial rivers in semi-arid watersheds**. J. Hydro**, v. 390, n. 1–2, p. 85–91, 2010.

AFFONSO, G. S. **Mapas auto - organizáveis de Kohonen (SOM) aplicados na avaliação dos parâmetros da qualidade da água.** 2011. Dissertação (Mestrado) - IPEN-Universidade de São Paulo, São Paulo, 2011.

AGÊNCIA NACIONAL DE ÁGUAS (ANA). **Conjuntura dos recursos hídricos no Brasil 2017**: relatório pleno. Brasília: ANA, 2017.

ALAGHA, J.S.; SEYAM, M.; SAID, M.A.M.; MOGHEIR, Y. Integrating an artificial intelligence approach with k-means clustering to model groundwater salinity: the case of Gaza coastal aquifer (Palestine). **Hydrogeol J.**, n. 25, p. 2347–2361, 2017.

ALIZADEH, M. J.; KAVIANPOUR, M. R. Development of wavelet-RNA models to predict water quality parameters in Hilo Bay, Pacific Ocean. **Marine Pollution Bulletin**, n. 98, p. 171–178, 2015.

ALONSO FERNÁNDEZ,J.R.; DÍAZ MUNIZ, C.; GARCIA NIETO, P.J.; DE COS JUEZ, F.J.; SÁNCHEZ LASHERAS, F.; ROQUENÍ, M.N. Forecasting the cyanotoxins presence in fresh waters: a new model based on genetic algorithms combined with the MARS technique. **Ecological Engineering,** n. 53, p. 68–78, apr. 2013. Disponível em: <http://dx.doi.org/10.1016/j.ecoleng.2012.12.015>. Acesso em: 10 nov.2018.

AMENDOLA, M.; SOUZA, A.L.; BARROS, L.C. **Manual do uso da teoria dos conjuntos fuzzy no matlab 6.5**. [s.l.]: [s.n.], 2005. 46 p.

AN. Y; ZOU, Z.; LI, R. Descriptive Characteristics of Surface Water Quality in Hong Kong by a Self-Organising Map. **Int. J. Environ. Res. Public Health,** v.13, n.1, p.115, 2016. Disponível em: <http://doi:10.3390/ijerph13010115. 2016>. Acesso em: 10 nov.2018.

ARAÚJO, J.C.; SANTAELLA, S.T. Gestão da Qualidade. In: CAMPOS, Nilson; STUDART, Ticina (ed.). **Gestão das Águas**. 2. ed. Porto Alegre, RS: ABRH, 2001.242 p.

ASCE Task Committee on application of Artificial Neural Networks in Hydrology Artificial neural networks in hydrology. I: preliminar concepts**. Journal of Hydrologic Engineering,** v.5, n.2, p.115-123, 2000.

ASCE Task Committee on application of Artificial Neural Networks in Hydrology Artificial neural networks in hydrology. II: hydrologic applications. **Journal of Hydrologic Engineering,** v.5, n.2, p.124-137, 2000.

ASTEL, A. *et al*. Comparison of self-organizing maps classification approach with cluster and principal components analysis for large environmental data sets. **Water Research,** v. 41, n. 19, p. 4566-4578, nov.2007.

AUSSEM, A.; CAMPBELL, J.; MURTAGH, F. Wavelet-based feature extraction and decomposition strategies for financial forecasting. **J. Comput. Intel. Fin.**, v.6, n. 2, p. 5– 12, 1998.

BABA, N. M. *et al*. Current issues in ensemble methods and its applications. **Journal of Theoretical and Applied Information Technology,** v.81, n. 2, p.20, nov.2015.

BARAK, S.; SADEGH, S. S. Forecasting energy consumption using ensemble ARIMA–ANFIS hybrid algorithm. **Int. J. Electr. Power Energy Syst**., v. 82, p. 92–104, 2016.

BARBOSA, B. H. G. **Computação Evolucionária e Máquinas de Comitê na Identificação de Sistemas Não-Lineares.** 2009. Tese (Doutorado) - Universidade Federal de Minas Gerais, Belo Horizonte, 2009.

BARZEGAR, R.; MOGHADDAM, A.A.; ADAMOWSKI, J.; OZGA-ZIELINSKI, B. Multi-step water quality forecasting using a boosting ensemble multi-wavelet extreme learning machine model. **Stoch Environ Res Risk Assess,** v. 32, p.799–813, 2018.

BASAK, Debasish; PAL, Srimanta; PATRANABIS, Dipak C. Support Vector Regression. **Neural Information Processing – Letters and Reviews,** v. 11, n. 10, 2007.

BEN-HUR A., WESTON J. A User's Guide to Support Vector Machines: Data Mining Techniques for the Life Sciences. **Methods in Molecular Biology**, v.609, 2010.

BERNARDI, J.V.E.; LACERDA, L.D.; DÓREA, J.G.; LANDIM, P.M.B.; GOMES, J.P.O.; ALMEIDA, R.; MANZATTO A.G. ; BASTOS W.R. Aplicação da análise das componentes principais na ordenação dos parâmetros físico-químicos no alto Rio Madeira e Afluentes, Amazônia Ocidental. **Geochimica Brasiliensis**, v. 23, n. 1, p. 079-090, 2009.

BOWDEN, G.J.; MAIER, H.R.; DANDY, G.C. Optimal division of data for neural network models in water resources applications. **Water Resources Research,** v. 38, n. 2, 2002. Disponível em: <http://doi:10.1029/2001WR000266>. Acesso em: 10 nov.2018.

BOWDEN, G.J.; DANDY, G.C.; MAIER, H.R. Input determination for neural network models in water resources applications: part 1 e background and methodology. **Journal of Hydrology,** v. 301, p. 75-92, 2005**.**

BOWDEN, G.J., MAIER, H.R., DANDY, G.C. Input determination for neural network models in water resources applications: part 2: case study: forecasting salinity in a river. **Journal of Hydrology,** v. 301, p.93-107, 2005.

BRAGA, A.P.; CARVALHO, A.P.L.F.; LUDERMIR, T.B.; **Redes Neurais Artificiais:** Teorias e Aplicações. 2. ed. Rio de Janeiro: LTC, 2007.

BRASIL. Ministério da Saúde. Fundação Nacional de Saúde (FUNASA). **Manual de controle da qualidade da água para técnicos que trabalham em ETAS.** Brasília: Funasa, 2014. 112 p.

BRASIL. Ministério da Saúde. Portaria nº 518, de 25 de março de 2004. Estabelece os procedimentos e responsabilidades relativos ao controle e vigilância da qualidade da água para consumo humano e seu padrão de potabilidade, e dá outras providências. **Diário Oficial da União**, República Federativa do Brasil, Poder Executivo, Brasília, DF, 26 mar. 2004. Seção 1, p. 266-270.

BRASIL. Ministério do Meio Ambiente. Resolução nº 357, de 17 de março de 2005. 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. Conselho Nacional do Meio Ambiente (CONAMA). Brasília–DF, 2005. Disponível em: <http://www.mma.gov.br/port/conama/res/res05/res35705.pdf> Acesso em: 08 maio 2017.

BRASIL. Ministério do Meio Ambiente. Resolução n° 001, de 21 de janeiro de 1986. Conselho Nacional do Meio Ambiente (CONAMA). Brasília–DF, 1986. Disponível em: <http://www.lei.adv.br/001-86.htm>. Acesso em: 20 abr. 2015.

BRASIL. Ministério do Meio Ambiente. Resolução nº 20, de 18 de junho de 1986. Conselho Nacional do Meio Ambiente **(**CONAMA). Brasília–DF, 1986. Disponível em:<http://www.mma.gov.br/port/conama/res/res86/res2086.html>. Acesso em: 20 abr. 2015.

BRASIL. Lei 9.433, de 8 de janeiro de 1997. **Recursos Hídricos: conjunto de normas legais**. 3. ed. Brasília, DF: MMA, 2004. 243 p.

BRASIL. Ministério da Saúde. Portaria nº 36, de 19 de janeiro de 1990. Aprovação de normas e o padrão de Potabilidade da Água destinada ao Consumo Humano, a serem observadas em todo o território nacional. **Diário Oficial da União**, República Federativa do Brasil, Poder Executivo, Brasília, DF.

BRASIL. Ministério da Saúde. Portaria nº 1469, de 29 de dezembro de 2000. Estabelece os procedimentos e responsabilidades relativos ao controle e vigilância da qualidade da água para consumo humano e seu padrão de potabilidade, e dá outras providências. **Diário Oficial da União**, República Federativa do Brasil, Poder Executivo, Brasília, DF.

BRASIL. Ministério da Saúde. Portaria nº 2914.2011. Dispõe sobre os procedimentos de controle e de vigilância da qualidade da água para consumo humano e seu padrão de potabilidade. **Diário Oficial da União**, República Federativa do Brasil, Poder Executivo, Brasília, DF.

BRASIL. **Constituição da República Federativa do Brasil de 1988.** Disponível em: <http://www.planalto.gov.br/ccivil_03/constituicao/constituicaocompilado.htm.>. Acesso em: 15 out. 2016.

CALDEIRA, A. M. *et al*. **Inteligência Computacional aplicada à Administração, Economia e Engenharia em Matlab**®. São Paulo: Thomson Learning, 2007.

CAMPOS FILHO, Pio. **Método para apoio à decisão na verificação da Sustentabilidade de uma Unidade de Conservação, usando Lógica *Fuzzy***. 2004. 210f. Dissertação (Doutorado em Engenharia de Produção) - Universidade Federal de Santa Catarina, Florianópolis, 2004.

CAMPISI, S.; ADAMOWSKI, J.; ORON, G. Forecasting urban water demand via wavelet-denoising and neural network models: case study: City of Syracuse, Italy. **Water Resour. Manage,** v. 26, p.3539–3558, 2012.

CANHAMERO, M. **O saneamento Ambiental através do Conhecimento de Novas Técnicas de Análise para os Recursos Hídricos.** Apostila do curso e capacitação Técnica em Recursos Hídricos com ênfase na sub-bacia hidrográfica Billings Tamanduaté. 2006. Disponível em: <www.agds.org.br/cursos/pdf/apostilapro fmagali.pdf.>. Acesso em: 15 jan. 2017.

CARVALHO F. O.; GARCIA, C. A. B.; GARCIA, H. L; COELHO, F. A.; WANDERLEY, H.S.; SANTOS, R.R.C.; CAVALCANTI, J.R.A. Sistemas Inteligentes (Redes Neurais) e Estatística Multivariada (ACP) aplicados a qualidade da água. In: SIMPÓSIO BRASILEIRO DE RECURSOS HÍDRICOS, XVIII, 2009, Campo Grande, MS.

CARVALHO, F. O.; GARCIA, C. A. B.; GARCIA, H. L.; ALVES, J. P. H.; PIMENTEL, W. R. O.; SILVA, J. L. Modelagem do processo de eutrofização no reservatório Jacarecica II – Sergipe/Brasil utilizando redes neurais artificiais e estatística multivariada. In: CONGRESSO IBERO LATINO-AMERICANO SOBRE MÉTODOS COMPUTACIONAIS EM ENGENHARIA (CILAMCE), 2007, Porto.

CARVALHO F. O.; COELHO, F. A.; SILVA, A. N.; GARCIA, H. L.; SILVA, V. L.; GARCIA, C. A. B. Redes Neurais e Estatística Multivariada na qualidade da água. In: CONGRESSO BRASILEIRO DE ENGENHARIA QUÍMICA, XVIII, 2010, Foz do Iguaçu.

ÇAMDEVÝREN H.; DEMÝR, N.; KANIK, A.; KESKÝN, S. Use of principal component scores in multiple linear regression models for prediction of Chlorophyll – a in reservoirs. **Ecological Modelling**, v. 181, p.581-589, 2005.

COMPANHIA AMBIENTAL DO ESTADO DE SÃO PAULO. **Qualidade das águas interiores no estado de São Paulo:** 2017. SÃO PAULO: CETESB, 2018. Disponível em: <https://cetesb.sp.gov.br/aguasinteriores/wpcontent/uploads/sites/12/2018/06/Relat%C3%B3rio-de-Qualidade-das-%C3%81guas-Interiores-no-Estado-de S%C3%A3o-Paulo-2017.pdf>. Acesso em: 21 abr. 2018.

CHRISTIANO, D. **Uso de Redes Neurais Artificiais, aplicadas no Rio Jaguaribe, João Pessoa, PB, como ferramenta de previsão para o gerenciamento ambiental**. 2007. Mestrado (Dissertação) - Universidade Federal da Paraíba, João Pessoa, 2007.

CHAPRA, S.C. **Surface Water-Quality Modeling**. New York, NY: McGraw-Hill, 1997. 844 p.

CHANG, F.J.; CHIANG, Y.M.; TSAI A, M.J.; SHIEH, M.C.; HSU, K.L.; SOROOSHIAN, S. Watershed rainfall forecasting using neurofuzzy networks with the assimilation of multi-sensor information. **Journal of Hydrology**, v. 508, p. 374–384, 2014.

CHOU, J-S; HO, C.C.; HOANG, H.S**.** Determining quality of water in reservoir using machine learning**. Ecological Informatics,** v .44, mar. 2018, p. 57-75.

CSÁBRÁGI, A.; MOLNÁR, S.; TANOS, P.; KOVÁCS, J. Application of artificial neural networks to the forecasting of dissolvedoxygen content in the Hungarian section of the river Danube. **Ecological Engineering**, v. 100, p. 63-72. 2017.

CUNHA, A.C.; SIQUEIRA, E.Q; CUNHA, H.F.A. Avaliação das Equações de Previsão do Coeficiente de Reaeração no Modelo QUAL2E para Modelagem de Oxigênio Dissolvido: estudo de caso no ribeirão do Feijão (São Carlos - SP). **Revista de Ciência e Tecnologia do Estado do Amapá**, v. 2, n.1. p. 90-111, 2001.

DASTORANI, M. T. & AFKHAMI, H. Application of artificial neural networks on drought prediction in Yazd (Central Iran). **Desert**, v.16, p.39-48, 2011.

DAWSON, C.W.; WILBY, R.L. Hydrological modelling using artificial neural networks. **Progress in Physical Geography,** v.25, n.1, p. 80-108, 2001.

DIAMANTOPOULOU, M. J; ANTONOPOULOS, V. Z.; PAPAMICHAIL, D.M. The use of a neural network technique for the prediction of water quality parameters of axios river in northern Greece. **European Water,** v. 11/12, 55-62, 2005.

DJERIOUI, M; BOUAMAR, M; LADJAL, M;ZERGUINE, A. Chlorine Soft Sensor Based on Extreme Learning Machine for Water Quality Monitoring. **Arabian Journal for Science and Engineering,** 19 abr. 2018. Disponível em: <https://doi.org/10.1007/s13369-018-3253-8>. Acesso em: 23 jul. 2018

EIGER, S. Autodepuração dos Cursos D'água. In: MANOLE *et al.*. **Reuso de Água.** Barueri, SP: [s.n.], 2003. 579 p.

ESTEVES, F.de A. **Fundamentos de Limnologia**. Rio de Janeiro: Interciência, 1988. 575p.

FARUK, D. O. A hybrid neural network and ARIMA model for water quality time series prediction. **Engineering Applications of Artificial Intelligence**, v.23, p. 586–594, 2010.

EITOSA, F. A. C.; FILHO, J. M. **Hidrogeologia: Conceitos e Aplicações.** Recife: LABHID–UFPE; Fortaleza, CE: Editora CPRM, 1997.

FLETCHER, D.; GOSS, E. Forecasting with neural networks: an application using bankruptcy data. ***Information and Management***, v.*24*, *n.*3, p.159-167, 1993. Disponível em: <http://doi:10.1016/0378-7206(93)90064-z>. Acesso em: 10 nov.2018.

GARCIA, H.L. **Desenvolvimento de estratégias para utilização de sistemas inteligentes no monitoramento da qualidade da água**. 2012. Tese (Doutorado). Universidade Federal de Pernambuco, Recife, 2012.

GARCIA, H. L.; SILVA, V. L.; MARQUES, L. P.; GARCIA, C. A. B.; CARVALHO, F. O. Avaliação da qualidade da água utilizando a teoria fuzzy. **Scientia Plena**, v.8, n.7, 2012.

GARCIA H.L.; ALVES J. P. H.; CARVALHO F. O. Modelagem de processos de eutrofização em reservatórios usando a técnica de redes neurais. In: CONGRESSO BRASILEIRO DE ENGENHARIA QUÍMICA, XV, set. 2004, Curitiba. **Anais…** Curitiba: [s.n.], 2004. 1 CD-ROM.

GILLMAN, M.; HAILS, R. An Introduction to Ecological Modelling: Putting Practice into Theory. **Ecological Society Journal of Animal Ecology,** v.66, p. 912-919, 1997.

GIRARDELO, A.D. **Um estudo sobre o uso de Máquinas de Vetores Suporte em problemas de classificação.** 2010. Monografia (Graduação) - Universidade Estadual do Oeste do Paraná – UNIOESTE, Cascavel, 2010.

GONÇALVES, A.R. **Máquina de Vetores Suporte**. 2016. Disponível em: <https://www-users.cs.umn.edu/~agoncalv/arquivos/pdfs/svm.pdf>. Acesso em: 6 de Fevereiro de 2018.

GUNN, Steve R. **Support Vector Machines for Classification and Regression**. University of Southampton. 1998. Disponível em: <http://users.ecs.soton.ac.uk/srg/publications/pdf/SVM.pdf>. Acesso em: 02 fev. 2018.

GUO, J.; ZHOU, J.; QIN, H.; ZOU, Q.; LI, Q. Monthly stream-flow forecasting based on improved support vector machine model. **Expert Syst. Appl**., v.38, p.13073-13081, 2011.

GROSSMANN, A.; Morlet, J. Decomposition of Hardy function into square integrable wavelets of constant shape. **J. Math. Anal**., v.5, p. 723-736, 1984.

HAGHIABI, A.M.; NASROLAHI, A.H.; PARSAIE, A. Water quality prediction using machine learning methods. **Water Quality Research Journal**, v.53, n.1, 2018. Disponível em: <http://doi: 10.2166/wqrj.2018.025>. Acesso em: 10 nov.2018.

HAMEED , M; SHARQI, S.S.; YASEEN, Z.M.; AFAN, H.A.; HUSSAIN, A.; ELSHAFIE, A. Application of artificial intelligence (AI) techniques in water quality index prediction: a case study in tropical region, Malaysia . **Neural Comput. and Applic.**, v.28, sup.1, p. 893-905, dec.2017. Disponível em: <http://doi:10.1007/s00521-016-2404-7>. Acesso em: 10 nov.2018.

HAYKIN, S., **Neural Networks:** a comprehensive foundation. New Jersey: Prentice Hall, 1999.

HAYKIN, S. **Neural Networks and Learning Machines**. 3. ed. [s.l.]: Pearson Prentice Hall, 2009.

KALTEH, A. M.; HORTH, P. BERNDTSSON, R. Review of the self – Organizing Maps (SOM) approach in water resources: Analysis, modeling and application. **Environ. Modell & Softw.**, v.23, p.835-845, 2008.

KARUNANITHI, N.; GRENNEY, W.J.; WHITLEY, D.; BOVEE, K. Neural Networks for River Flow Prediction. **Journal of Computing in Civil Engineering,** v.8, n.2, apr.1994.

KHAN, Y. e CHAI, S.S. Ensemble of RNA and ANFIS for Water Quality Prediction and Analysis: a data driven approach. **Journal of Telecommunication, Electronic and Computer Engineering,** v.9, n. 2-9, 2017.

KIM, T.W.; VALDES, J.B. Non-linear model for drought forecasting based on a conjunction of wavelet transforms and neural networks. **J. Hydrol. Eng**., v.8, n.6, p. 319-328, 2003.

KIM, E. **Everything You Wanted to Know about the Kernel Trick**. 2015. Disponível em: <http://www.eric-kim.net/eric-kim-net/posts/1/kernel_trick.html>. Acesso em: 30 maio 2018.

KIM, S. E.; SEO, I. W. Artificial Neural Network ensemble modeling with conjunctive data clustering for water quality prediction in rivers. **J. Hydro-environment Res**., apr. 2015.

KISI, O. Neural networks and wavelet conjunction model for intermittent stream-flow forecasting. **J. Hydrol. Eng.**, v.14, n.8, p.773-782, 2009.

KISI, O. Neural network and wavelet conjunction model for modelling monthly level fluctuations in Turkey. **Hydrol. Process.,** v.23, n.14, p.2081-2092, 2009.

KISI, O. Daily suspended sediment estimation using neuro-wavelet models. **Int. J. Earth Sci**., v.99, p.1471-1482, 2010.

KOWALCZYK, A. **Support Vector Machine: Succinctly.** Morrisville, NC, USA: Syncfusion, 2017.

KOHONEN, T. **Self-Organizing Maps**. New York: Springer, 1995.

KOHONEN, T. **Self-organizing maps.** 3rd ed. Berlin–Heidelberg, Germany: Springer, 2001.

KOHONEN, T. Essentials of the self-organizing map. **Neural Networks**, v.37, p. 52-65, 2013.

KUO, J. WANG, I. LUNG, W. A hybrid neural–genetic algorithm for reservoir water quality management. **Water Research,** v.40, p.1367-1376, 2007.

KRISHNA, B. Comparison of wavelet based ANN and regression models for reservoir inflow forecasting. **J. Hydrol. Eng**., v.19, n.7, jul. 2014. Disponível em: <http://dx.doi.org/10.1061/(ASCE)HE.1943-5584.0000892>. Acesso em: 10 nov.2018.

KUO, C.C.; GAN, T.Y.; YU, P.S. Wavelet analysis on the variability, teleconnectivity and predictability of the seasonal rainfall of Taiwan. **Mon. Weather Rev**., v.138, n.1, p.162-175, 2010.

KUO, C.C.; GAN, T.Y.; YU, P.S. **Seasonal stream-flow prediction by a combined climate-hydrologic system for river basins of Taiwan**. **J. Hydrol**., v.387, p.292-303, 2010.

LI, X; SHA, J; E WANG, Z. Comparative study of multiple linear regression, artificial neural network and support vector machine for the prediction of dissolved oxygen. **Hydrology Research**, 2016

LI,T.; SUN, G.; YANG, C.; LIANG, K.; MA, S.; HUANG, L. Using self-organizing map for coastal water quality classification: towards a better understanding of patterns and processes. **Science of the Total Environment,** v.628–629, p.1446-1459, jul.2018.

LIMA, E.B.N.R. **Modelagem Integrada para Gestão da Qualidade da Água na Bacia do Rio Cuiabá.** Rio de Janeiro: [s.n], 2001.184 p.

LIU, Y.L.A.; MACAU, E.E.N.; BARROSO, J.J.; SILVA, J.D.S. Uso de Rede Neural Percéptron Multi-Camadas na Classificação de Patologias Cardíacas. **Tend. Mat. Apl. Comput.**, v.9, n.2, p.255-264, 2008.

LORENA, Ana C.; CARVALHO, André C. P. L. F. **Introdução às Máquinas de Vetores Suporte**. São Carlos: Instituto de Ciências Matemáticas e de Computação - USP, 2003.

LUGER, G.F. **Artificial Intelligence: Structures and Strategies for Complex Problem Solving.** 5. ed. [s.l.]: Addison-Wesley, 2005

LU, R.S.; LO, S.L. Diagnosing reservoir water quality using self-organizing maps and fuzzy theory. **Water Research**, v.36, p.2265-2274, 2002.

MAHESWARAN, R.; KHOSA, R. Comparative study of different wavelet hydrologic forecasting. **Comput. Geosci**., v.46, p.284–295, 2012.

MAIER, H.R.; DANDY, G.C. The use of artificial neural networks for the prediction of water quality parameters. **Water Resources Research**, v.32, n.4, p.1013-1022, 1996.

MAIER, H.R.; DANDY, G.C. Determining inputs for neural network models of multivariate time series. **Microcomputers in Civil Engineering**, v.12, n.5, p.353-368, 1997.

MAIER, H.R.; DANDY, G.C. Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications. **Environmental Modelling and Software**, v.15, p.101-124, 2000

MALUTTA C. **Método de apoio à tomada de decisão sobre a adequação de aterros sanitários utilizando a Lógica *Fuzzy*.** 2004. 221f. Dissertação (Doutorado em Engenharia de Produção) - Universidade Federal de Santa Catarina, Florianópolis, 2004.

MARQUES, L. P.; GARCIA, H. L.; COELHO, F. A.; Frede de Oliveira Carvalho; SILVA, V. L. Utilização de redes neurais artificiais e análise de componentes principais no monitoramento da qualidade da água. **Química no Brasil**, v. 7, p. 33-44, 2013.

MARTIN, R. W. e WALLEY W. J. **A River Biology Monitoring System (RBMS) for English and Welsh Rivers**. School of Computing, Staffordshire University, 1995. Disponível em: <http://www.cies.staffs.ac.uk/rbms.htm> Acesso em: Janeiro, 2006.

MATIATOS,I.; PARASKEVOPOULOU,V.; LAZOGIANNIS,K.; BOTSOU,F.; DASSENAK

IS, M.; GHIONIS, G.; ALEXOPOULOS,J. D.; POULOS,S. E. Surface–ground water interactions and hydrogeochemical evolution in a fluvio-deltaic setting: The case study of the Pinios River delta. **Journal of Hydrology,** v.561, p.236-249, 2018. Disponível em: <https://doi.org/10.1016/j.jhydrol.2018.03.067.>. Acesso em:

MEDEIROS, A.V. **Modelagem de sistemas dinâmicos não lineares utilizando Sistemas Fuzzy, Algoritmos Genéticos e Funções de Base Ortonormal.** 2006. Dissertação (Mestrado) - Universidade Estadual de Campinas, Campinas, 2006.

MEDEIROS, A. C.; FAIAL, K. R. F.; FAIAL, K. C. F.; LOPES, I. D. S.; LIMA, M. O.; GUIMARÃES, R. M.; MENDONÇA, N. M. Quality index of the surface water of Amazonian rivers in industrial areas in Pará, Brazil. **Marine Pollution Bulletin**, v. 123, p. 156-164, 2017. Disponível em: <http://doi: 10.1016/j.marpolbul.2017.09.002>. Acesso em: 10 nov.2018.

MESQUITA, N. R. **Classificação de defeitos em tubos de gerador de vapor de plantas nucleares utilizando mapas auto-organizáveis.** 2002. Tese (Doutorado) - Escola Politécnica de Engenharia, Universidade de São Paulo, São Paulo, 2002.

MIRBAGHERI, S.A., NOURANI, V., RAJAEE, T., ALIKHANI, A. Neuro-fuzzy models employing wavelet analysis for suspended sediment concentration prediction in rivers. **Hydrol. Sci. J.**, v.55, n.7, 1175-1189, 2010.

MINGOTI, S. A. **Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada**. Belo Horizonte: Editora UFMG, 2005.

MOHAMMADPOUR, R.; SHAHARUDDIN, S.; CHANG, C.K.; ZAKARIA, N.A.; GHANI, A.; CHAN, N.W. Prediction of water quality index in constructed wetlands using support vector machine. **Environ Sci Pollut Res Int.**, v.22, n.8, p.6208-6219, apr. 2015. Disponível em: <http://doi: 10.1007/s11356-014-3806-7>. Acesso em: 10 nov.2018.

MOUNCE, S. R. e BOXALL, J. B. **Automated data-driven approaches to evaluating and interpreting water quality time series data from water distribution systems read.** University of Sheffield online paper. 2015.

MOTA, S. **Introdução à Engenharia Ambiental**. 3.ed. Rio de Janeiro, RJ: ABES, 2003. 419 p.

MULLAI, P.; ARULSELVI, S.; NGO, HH.; SABARATHINAM, PL. Experiments and ANFIS modelling for the biodegradation of penicillin-G wastewater using anaerobic hybrid reactor. **Bioresource Technol,** v.102, n.9, p.5492-5497, mar.2011. Disponível em: <http://doi:10.1016/j.biortech.2011.01.085>. Acesso em: 10 nov.2018.

NAGI, S.; BHATTACHARYYA, D. K. Classification of microarray cancer data using ensemble approach. **Netw. Model. Anal. Heal. Informatics Bioinforma**, v.2, n.3, p.159-173, 2013.

NALLEY, D.; ADAMOWSKI, J.; KHALIL, B. Using discrete wavelet transforms to analyze trends in stream-flow and precipitation in Quebec and Ontario (1954– 2008). **J. Hydrol**, v.475, p.204-228, 2012.

NASON, G.P.; SACHS, R. Wavelets in time series analysis. **Philos. Trans. Roy. Soc**., v.357, p.2511-2526, 1999.

NEPOMUCENO, V.S. **Classificação de qualidade da água em reservatórios utilizando um algoritmo de aprendizagem competitiva e não supervisionada**. Monografia (Graduação) - Universidade Federal de Pernambuco, Recife, 2006.

NOURANI, V.; ALAMI, M.T.; AMINFAR, M.H. A combined neural-wavelet model for prediction of Ligvanchai watershed precipitation. **Eng. Appl. Artif. Intel**., v.22, n.3, p.466-472, 2009.

NOURANI, V.; KOMASI, M.; MANO, A. A multivariate RNA-wavelet approach for rainfall–runoff modeling. **Water Resour. Manage**., v.23, n.14, p.2877-2894, 2009.

NOURANI, V.; KISI, Ö.; KOMASI, M. I. K.. Two Hybrid Artificial Intelligence Approaches for Modeling Rainfall–Runoff Process. **Journal of Hidrology,** 2011.

NOURANI, V.; KOMASI, M.; ALAMI, M. Hybrid Wavelet-genetic programming approach to optimize ANN modeling of rainfall–runoff process. **J. Hydrol. Eng**., v.16, n.6, p.724-741, 2012.

NOURANI, V.; HOSSEINI BAGHANAM, A.; ADAMOWSKI, J.; GEBREMICHEAL, M., Using self-organizing maps and wavelet transforms for space-time pre-processing of satellite precipitation and runoff data in neural network based rainfall–runoff modeling**. J. Hydrol.,** v.476, p.228-243, 2013.

NOURANI, V.; BAGHANAM, A.H.; ADAMOWSKI, J.; KISI, O. Applications of hybrid wavelet–Artificial Intelligence models in hydrology: a review. **Journal of Hydrology,** v.514, p.358-377, 2014.

NOORI, R. *et al.*; Multivariate statistical analysis of surface water quality based on correlations and variations in the data set. **Desalination**, v.260, n.1-3, p.129-136, 2010.

NUNES, I. S. **Material Didático da Disciplina Sistemas Inteligentes Avançados** (SEL5757). São Carlos: Universidade de São Paulo, Departamento de Engenharia Elétrica – EESC, 2009.

OLIVEIRA JUNIOR, H. A.; CALDEIRA, A. M.; MACHADO, M. A.; SOUZA, R. C. e TANSCHEIT, R. **Inteligência Computacional aplicada à Administração, Economia e Engenharia em Matlab**. São Paulo: Thomson Learning, 2007.

PANDHIANI, S.M.; SHABRI, A. B. A Hybrid Model for Monthly Time Series Forecasting. **Appl. Math. Inf. Sci**., v.9, n.6, p.2943-2953, 2015.

PEREIRA, A.A. **Avaliação da qualidade da água: Proposta de novo índice alicerçado na lógica FUZZY.** 2010. Tese (Doutorado) - Programa de Pós-Graduação em Ciências da Saúde Brasília, Universidade de Brasília, Distrito Federal, 2010.

PACHECO, M.P. **Monitoramento Ambiental: Uma estratégia para minimização de desastres naturais.** [Rio de Janeiro]: Quartas Científicas, UERJ, 2005. 27p.

PARINET, B.; LHOTE, A.; LEGUBE, B. Principal Components analysis: An appropriate tool for water quality evaluation and management application to a tropical lake system. **Ecological Modelling**, v.178, p.295-311, 2004.

PARTAL, T.; KISI, O. Wavelet and neuro-fuzzy conjunction model for precipitation forecasting. **J. Hydrol**., v.342, p.199-212, 2007.

PEETERS, L.; DASSARGUES, A. **Comparison of Kohonen's Self-Organizing Map algorithm and principal component analysis in the exploratory data analysis of a groundwater quality dataset.** 2006. Disponível em: <https://www.researchgate.net/publication/228795917_Comparison_of_Kohonen's_Self-Organizing_Map_algorithm_and_principal_component_analysis_in_the_exploratory_data_analysis_of_a_groundwater_quality_dataset>. Acesso em: 04 ago. 2017

PERENDECI, A.; ARSLAN, S.; CELEBI, S. S.; TANYOLA, A. Prediction of effluent quality of an anaerobic treatment plant under unsteady state through ANFIS modeling with on-line input variables. **Chemical Engineering Journal**, v.145, p.78-85, 2008.

PINTO, L. C.; MELLO, C.R. de; FERREIRA, D.F.; AVILA, L. F. Índice de qualidade de água em duas situações de uso do solo na Serra da Mantiqueira. **Ciênc. Agrotec**., v.37, n.4, p.338-342, 2013. Disponível em: <http://dx.doi.org/10.1590/S1413-70542013000400007.>. Acesso em: 10 nov.2018.

PRIMPAS, I.; TSIRTSIS, G.; KARYDIS, M. Quantitative assessment of eutrophication: a proposed multivariate index. In: INTERNATIONAL CONFERENCE ON ENVIRONMENT SCIENCE AND TECHNOLOGY, 10, sept. 2007, Kos Island, Greece. **Proceedings…**Greece: [s.n.], 2007. p. 5-7.

PRIMPAS, I. *et al*. Principal component analysis: Development of a multivariate index for assessing eutrophication according to the European water framework directive. **Ecological Indicators**, v.10, p.178-183, 2010.

RAVANSALAR, M. e RAJAEE, T. **A hybrid model of Wavelet with Artificial Neural Network formonthly prediction of river water quality**.Conference paper, oct.2014

RODRIGUES, R.B. **Sistema de Suporte à Decisão Proposto para a Gestão Quali-Quantitativa dos Processos de Outorga e Cobrança pelo Uso da Água**. 2005.155 p. Tese **(**Doutorado) **–** Universidade de São Paulo, São Paulo, 2005.

ROSENBLATT, F. The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain. **Psychological Review**, v.65, p.386-408, 1958.

SIMÕES, M.G.; SHAW, I.S. **Controle e Modelagem Fuzzy.** 2.ed. São Paulo: Edgard Blucher, 2007.

SINGH, K. P.; BASANT, A.; MALIK, A.; JAIN, G. Artificial neural network modeling of the river water quality: a case study. **Ecological Modelling**, v.220, p.888-895, 2009.

SINGH K. P.; NIKITA, B.; SHIKHA, G. Support vector machines in water quality management. **Analytica Chimica Acta,** v.703, p.152-162, 2011.

SIVANANDAM, S.N.; SUMATHI, S.; DEEPA, S.N. **Introduction to fuzzy logic using Matlab**. Heidelberg: Springer, 2007.430 p.

STROBL, R.O.; FORTE, F.; PENNETTA, L. Application of artificial neural networks for classifying lake eutrophication status. **Lakes e Reservoirs: Research and Management**, v.12, p.15-25, 2007.

SAFAVI, H. R. Prediction of River Water Quality by Adaptive Neuro Fuzzy Inference System (ANFIS). **Journal of Environmental Studies**, v.36, n.53, 2010.

SANDRI, S.; CORREA, C. Lógica nebulosa. In: ESCOLA DE REDES NEURAIS, 5.,1999, São José dos Campos. **Anais...** São José dos Campos: Conselho Nacional de Redes Neurais, ITA-SP, 1999.

SANG, Y.F., A practical guide to discrete wavelet decomposition of hydrologic time series. **Water Resour**. **Manage**., v.26, n.11, p.3345-3365, 2012.

SANG, Y.F. A review on the applications of wavelet transform in hydrology time series analysis**. Atmos**. **Res**., v.122, p.8-15, 2013.

SEMA/SP - SECRETARIA DE ESTADO DO MEIO AMBIENTE. **A Qualidade das Águas**. 2. ed. São Paulo, SP: CETESB, 2000. 44 p. (Série Manuais Ambientais).

SILER, W.; BUCKLEY, J. J. **Fuzzy Expert Systems and Fuzzy Reasoning**. New Jersey: John Wiley & Sons Inc., 2005.

SILVA, A.M.B. **Avaliação da Qualidade da Água Bruta Superficial das Barragens de Bita e Utinga de Suape aplicando Estatística e Sistemas Inteligentes.** 2015. Tese (Doutorado) - Universidade Federal de Pernambuco, Recife, 2015.

SILVA, I.A.F. **Descoberta de conhecimento em base de dados de monitoramento ambiental para avaliação da qualidade da água.** 2007. Dissertação (Mestrado) - Universidade Federal de Mato Grosso, Mato Grosso, 2007.

SOLOMATINE, D.P.; OSTFELD, A. Data-driven modelling: some past experiences and new approaches. **J. Hydroinform**., v.10, n.1, p.3-22, 2008.

SPERLING, M.V. **Introdução à Qualidade das Águas e ao Tratamento de Esgotos. Série Princípios do Tratamento Biológico de Águas Residuárias** v. 1. 2.ed. Belo Horizonte: UFMG, 1996. 243 p.

SOUZA, A.D. de. **Estudo de perda de carga em escoamento multifásico utilizando técnicas de inteligência artificial com ênfase no escoamento de petróleo**. 2011. Dissertação (Mestrado) - Programa de Pós-Graduação em Engenharia Química, Universidade Federal de Sergipe, São Cristóvão, Sergipe, 2011.

TALEBIZADEH, M.; MORIDNEJAD, A. Uncertainty analysis for the forecast of lake level fluctuations using ensembles of ANN and ANFIS models. **Expert Systems with Applications**, v.*38, n.*4, p.4126-4135, 2011.

TRAUTMANN, T.; DENOEUX, T. Comparison of dynamic feature map models for environmental Monitoring. **Proc. of the International Conference on Artificial Neural Network**, p.73-78, 1995.

TIWARI, M.K.; CHATTERJEE, C. Development of an accurate and reliable hourly flood forecasting model using wavelet–bootstrap–RNA (WBRNA) hybrid approach. **J. Hydrol.** v.394, p.458-470, 2010.

TUCCI, C.E.M.; HESPANHOL, I.; CORDEIRO NETO, O.M. **Gestão da Água no Brasil.** Brasília: UNESCO, 2001.190 p.

VAPNIK, V. N. **The nature of statistical learning theory**. New York: Springer-Verlag New York, Inc., 1995. Disponível em: <http://portal.acm.org/citation-.cfm?id=211359>. Acesso em: jul. 2017.

VESANTO, J. **Data exploration process based on the self-organizing map**. 2002. (Ph.D. thesis). Disponível em: <http://lib.tkk.fi/Diss/2002/isbn9512258978/isbn9512258978.pdf>. Acesso em: 15 ago. 2017

VIEIRA, M.R. **Os principais parâmetros monitorados pelas sondas multiparâmetros são: pH, condutividade, temperatura, turbidez, clorofila ou cianobactérias e oxigênio dissolvido**. 2010. Disponível em: <https://www.agsolve.com.br/news_upload/file/Parametros%20da%20Qualidade%20da%20Agua.pdf>. Acesso em: 15/10/2016.

VILAS, L.G.; SPYRAKOS, E.; PALENZUELA, J.M.T. Neural network estimation of chlorophyll a from MERIS full resolution data for the coastal waters of Galician rias (NW Spain). **Remote Sensing of Environment**, v.115, p.524-535, 2011.

WANG, Y.; WANG, Y.; GUO, L.; ZHAO, Y.; ZHANG, Z.; WANG, P. A Wavelet Neural Network Hybrid Model for Monthly Ammonia Forecasting in River Water. **Research Journal of Applied Sciences, Engineering and Technology,** v.6, n.2, p.345-348, 2013.

WANG, H.; LEI, X.; JIANG, Y.; SONG, X.; WANG, Y. Flood simulation using parallel genetic algorithm integrated wavelet neural networks. **Neurocomputing,** v.74, p.2734–2744, 2011.

WCED. **Our Common Future**. Oxford, UK: Oxford University Press, 1987. 20 p.

WESTPHAL, J. T**. Modelagem Difusa de um Sistema Especialista Médico: Avaliação dos Fatores de Internação em Crianças Queimadas**. 2003. 115f. Dissertação (Mestrado em Ciência da Computação) - Programa de pós-graduação em ciência da Computação, Universidade Federal de Santa Catarina, Florianópolis, 2003.

WU, G; LO, S. Predicting real-time coagulant dosage in water treatment by artificial neural networks and adaptive network-based fuzzy inference system**. Engineering Applications of Artificial Intelligence,** v.21, p.1189-1195, 2008.

WU, S.; CHOW, T.W.S. Clustering of the self-organizing map using a clustering validity index based on inter-cluster and intra-cluster density. **Pattern Recognition,** v.37, p.175-188, 2004.

XU, L. Study of short-term water quality prediction model based on wavelet neural network. **Mathematical and Computer Modelling**, v.58, p.807-813, 2013.

YAN, H.; ZOU, Z.; WANG, H. Adaptive neuro fuzzy inference system for classification of water quality status. **Journal of Environmental Sciences**, v.22, n.12, p.1891-1896, 2010.

YONG, C. K.; LIM, C. M. **An Integrated Water Quality Monitoring System using Artificial Neural Networks.** Singapore: Ngee RNA Polytechnic - Electronic and Computer Engineering Department, 2001. Disponível em: <http://www.np.edu.sg/~yck/ nn_waterquality.pdf>. Acesso em: jan. 2006.

ZHANG, L.; ZHANG, G. X. Water Quality Analysis and Prediction Using Hybrid Time Series and Neural Network Models. **J. Agr. Sci. Tech**., v.18, p.975-983, 2016.

ZHAO, Hai-jiang; WANG, Yi; YANG, Liang-liang; YUAN, Luo-wei; PENG, Dang-cong. Relationship between phytoplankton and environmental factors in landscape water supplemented with reclaimed water. **Ecological Indicators**, v. 58, p. 113-121, 2015. Disponível em: <https://doi.org/10.1016/j.ecolind.2015.03.033.>. Acesso em: 10 nov.2018.

ZHOU, F.; GUO, H.; LIU, Y.; JIANG, Y. Chemometrics data analysis of marine water quality and source identification in Southern Hong Kong. **Marine Pollution Bulletin,** v. 54, p. 745-756, 2007. Disponível em: < https://doi.org/10.1016/j.marpolbul.2007.01.006.>. Acesso em: 10 nov.2018.

ZHOU, H.C.; PENG, Y.; LIANG, G.H. The research of monthly discharge predictor corrector model based on wavelet decomposition. **Water Resour. Manage.**, v.22 , p.217-227, 2008.

ZIMMERMANN, C.M.; GUIMARÃES, O.M.; PERALTA-ZAMORA P.G. Avaliação da qualidade do corpo hídrico do rio Tibagi na região de ponta grossa utilizando análise de componentes principais (ACP). **Química Nova,** v. 31, n. 7, p. 1727-1732, 2008.

APÊNDICES

APÊNDICE A - MATRIZ DE PESO FATORIAL DAS VARIÁVEIS DA QUALIDADE DE ÁGUA

Tabela A1 - Matriz de peso fatorial das sete variáveis da qualidade de água do Brasil nos sete componentes principais selecionados.

Variáveis	**CP1**	**CP2**	**CP3**	**CP4**	**CP5**	**CP6**	**CP7**
Condutividade	**-0,579**	**-0,530**	-0,148	0,516	0,164	0,262	-0,004
OD	**-0,815**	0,390	-0,027	-0,303	-0,110	0,179	-0,215
pH	**-0,897**	-0,027	-0,090	-0,352	0,023	0,039	0,246
Potencial de Oxi-redução (ORP)	0,355	**-0,684**	-0,049	-0,187	-0,591	0,138	0,006
Temp.	-0,306	**-0,816**	-0,216	-0,243	0,248	-0,246	-0,111
Turbidez	-0,212	-0,310	**0,922***	-0,049	0,077	0,017	-0,008
Clorofila	**-0,770**	0,111	0,054	0,367	-0,426	-0,275	-0,007
Autovalores	**2,662**	**1,676**	**0,933**	0,713	0,638	0,257	0,119
Variância explicada (%)	38,03	23,95	13,33	10,19	9,12	3,68	1,71
Variância acumulada (%)	**38,0**	**62,0**	**75,3**	85,5	94,6	98,3	100,0

Tabela A2 - Matriz de peso fatorial das 12 variáveis da qualidade de água da China nos doze componentes principais selecionados.

Variáveis	**CP1**	**CP2**	**CP3**	**CP4**	**CP5**	**CP6**	**CP7**	**CP8**	**CP9**	**CP10**	**CP11**	**CP12**
Temperatura da água	-0,262	-0,071	**-0,848**	0,405	-0,143	0,131	-0,070	-0,019	-0,007	-0,001	-0,011	0,027
Sólidos Suspensos Totais	-0,285	-0,101	-0,213	**-0,753**	-0,410	0,348	0,050	0,001	0,050	-0,018	0,038	0,001
Sólidos Totais	**-0,685**	0,406	-0,275	-0,435	0,295	-0,016	0,019	0,032	0,003	0,024	-0,108	-0,002
Fósforo Total	**-0,908**	-0,005	0,239	0,198	-0,014	0,186	0,034	-0,041	0,143	0,143	0,021	0,001
Nitrito	**-0,696**	0,457	0,040	0,228	-0,269	-0,181	0,243	0,291	0,056	-0,039	0,007	0,000
Nitrato	**-0,580**	**0,526***	0,083	-0,018	-0,380	-0,362	-0,054	-0,317	-0,012	-0,015	-0,002	0,001
OD	0,432	**0,670***	**0,551***	-0,107	0,029	0,204	-0,038	0,033	-0,061	0,001	-0,004	0,033
OD (%)	0,267	**0,797***	-0,161	0,283	-0,119	0,387	-0,125	0,017	-0,082	0,002	0,001	-0,026
Condutividade	**-0,666**	0,426	-0,246	-0,144	0,522	-0,086	-0,067	-0,006	0,013	-0,028	0,095	0,002
DQO	**-0,848**	-0,285	0,120	0,072	0,047	0,147	0,167	-0,059	-0,354	0,000	0,006	0,000
Amônia	**-0,778**	-0,218	0,332	0,250	0,129	0,327	-0,019	-0,108	0,159	-0,109	-0,029	-0,001
DBO (5 dias)	**-0,796**	-0,276	0,221	-0,065	-0,169	-0,108	-0,390	0,199	-0,072	-0,001	-0,004	0,000
Autovalores	**4,92**	**2,14**	**1,47**	**1,20**	0,83	0,68	0,27	0,24	0,19	0,04	0,02	0,00
Variância explicada (%)	41,04	17,80	12,24	9,97	6,89	5,63	2,27	2,04	1,60	0,30	0,20	0,02
Variância acumulada (%)	**41,04**	**58,84**	**71,07**	**81,05**	87,94	93,57	95,84	97,88	99,49	99,78	99,98	100,00

Tabela A3 - Matriz de peso fatorial das vinte e quatro variáveis da qualidade de água da Espanha nos vinte e quatro componentes principais selecionados (continua).

Variáveis	CP1	CP2	CP3	CP4	CP5	CP6	CP7	CP8	CP9	CP10	CP11	CP12
Cianotoxina	-0,413	**0,548***	0,357	-0,319	0,123	-0,181	0,018	-0,082	-0,147	0,388	0,190	0,062
Clorofila	-0,218	**0,635***	**0,501***	-0,094	-0,316	0,051	-0,043	0,079	0,309	-0,191	-0,054	0,028
Microcystis Aeruginosa (MA)	**-0,682**	**0,571***	0,184	-0,078	-0,113	-0,038	-0,204	-0,019	-0,097	-0,101	0,007	-0,103
Woronichinia Nalgiliana (WN)	**-0,673**	**0,571***	0,138	-0,023	-0,212	0,037	0,124	0,188	-0,016	0,097	0,053	0,019
MA e WN	**-0,614**	**0,663***	0,239	-0,103	-0,189	-0,008	0,032	0,109	-0,072	0,082	0,024	-0,055
Outras Espécies de Cianobactérias	**-0,672**	0,415	-0,077	-0,063	0,057	-0,019	-0,139	-0,076	-0,265	-0,424	0,227	0,106
Diatomáceas	-0,042	**-0,773**	0,409	0,033	-0,130	-0,015	-0,172	-0,009	0,003	0,039	-0,030	0,343
Chrysophytes	0,023	**-0,823**	0,352	-0,015	-0,212	-0,016	-0,223	-0,035	0,008	0,031	0,039	0,014
Clorofíceas	0,178	**-0,776**	0,342	-0,140	-0,005	-0,056	0,103	0,133	-0,033	-0,083	0,168	-0,298
Outras Espécies de Fitoplânctons	0,016	**-0,780**	0,438	0,002	-0,140	-0,096	-0,124	-0,052	-0,096	-0,045	-0,028	-0,114
Temperatura da água	**-0,679**	**-0,497**	-0,164	-0,208	-0,030	0,165	0,169	0,206	0,208	-0,003	0,181	0,068
Temperatura Ambiente	**-0,638**	**-0,603**	-0,109	-0,216	0,065	0,240	0,153	0,006	0,088	-0,019	0,174	0,032
Turbidez	**-0,822**	0,093	-0,188	0,315	-0,261	0,152	-0,138	0,137	0,029	0,070	-0,046	-0,041
Fósforo Total	**-0,931**	-0,066	-0,045	-0,116	0,061	-0,082	-0,099	-0,024	0,114	0,018	-0,184	-0,032
Fosfato	**-0,896**	-0,042	-0,078	0,343	-0,096	0,050	0,006	0,013	0,041	0,015	-0,092	-0,023
Nitrogênio total	**-0,781**	-0,400	0,118	0,053	-0,087	0,057	0,342	0,048	-0,193	-0,021	-0,066	0,043
Nitrato	**-0,683**	-0,364	-0,087	-0,353	-0,110	-0,140	0,244	-0,056	-0,228	-0,071	-0,312	0,026
Nitrito	-0,228	-0,141	-0,451	**-0,694**	-0,316	0,081	-0,226	-0,226	0,098	0,065	-0,015	-0,068
Amônio	**-0,873**	-0,207	0,008	0,155	0,186	0,256	0,010	-0,167	0,054	-0,005	0,042	-0,007
Oxigênio Dissolvido	**0,621***	-0,010	-0,014	-0,154	0,001	0,592	-0,214	0,301	-0,263	0,071	-0,117	-0,013
Condutividade	0,271	0,417	**0,512***	-0,110	0,175	0,530	0,192	-0,308	0,068	-0,044	-0,098	-0,027
Alcalinidade	**0,829***	0,204	0,056	-0,114	-0,272	-0,207	0,236	0,060	0,072	-0,075	-0,069	0,022
Cálcio	**-0,847**	-0,199	0,097	0,340	0,104	-0,061	-0,077	-0,133	0,013	0,106	-0,039	-0,110
pH	**0,495***	-0,005	-0,192	0,334	**-0,655**	0,174	0,148	-0,257	-0,125	0,071	0,155	-0,008
Autovalores	**9,215**	**5,688**	**1,714**	**1,382**	**1,077**	0,971	0,659	0,507	0,476	0,448	0,391	0,279
Variância explicada (%)	38,40	23,70	7,14	5,76	4,49	4,05	2,75	2,11	1,98	1,87	1,63	1,16
Variância acumulada (%)	**38,4**	**62,1**	**69,2**	**75,0**	**79,5**	83,5	86,3	88,4	90,4	92,2	93,9	95,0

Tabela A3 - Matriz de peso fatorial das vinte e quatro variáveis da qualidade de água da Espanha nos vinte e quatro componentes principais selecionados (conclusão).

Variáveis	CP13	CP14	CP15	CP16	CP17	CP18	CP19	CP20	CP21	CP22	CP23	CP24
Cianotoxina	0,043	0,013	-0,163	-0,021	-0,039	0,010	0,046	0,018	0,013	0,012	-0,002	-0,009
Clorofila	-0,022	-0,060	-0,119	-0,014	-0,149	-0,028	-0,031	0,003	-0,001	0,021	-0,025	0,002
Microcystis Aeruginosa (MA)	-0,169	-0,187	0,034	-0,035	0,088	0,025	0,024	-0,014	-0,001	0,001	0,059	-0,056
Woronichinia Nalgiliana (WN)	0,130	0,173	0,167	0,050	-0,033	-0,022	-0,087	0,000	-0,025	-0,011	0,051	-0,034
MA e WN	-0,061	-0,074	0,148	0,005	0,092	-0,039	0,011	0,024	0,017	-0,031	-0,059	0,072
Outras Espécies de Cianobactérias	0,081	0,119	-0,063	-0,008	0,024	-0,007	-0,001	-0,003	-0,007	0,003	-0,015	0,014
Diatomáceas	-0,203	0,103	0,082	-0,072	0,019	0,012	0,009	0,012	0,017	0,001	0,005	-0,001
Chrysophytes	0,120	-0,062	-0,080	0,248	0,088	-0,049	-0,023	-0,028	-0,021	-0,017	0,011	0,005
Clorofíceas	-0,201	0,170	-0,024	0,023	-0,035	-0,008	0,032	0,011	0,008	-0,004	0,004	0,001
Outras Espécies de Fitoplânctons	0,288	-0,058	0,099	-0,164	-0,042	0,049	0,026	0,021	0,013	0,003	-0,007	-0,002
Temperatura da água	0,073	-0,042	0,002	-0,091	0,067	-0,067	0,094	-0,093	-0,019	0,019	-0,005	-0,012
Temperatura Ambiente	-0,010	-0,118	-0,045	-0,034	0,020	0,046	-0,099	0,109	0,006	0,006	0,050	0,031
Turbidez	0,023	0,079	-0,033	0,074	0,023	0,113	0,074	0,026	0,021	0,093	0,003	0,018
Fósforo Total	-0,001	0,084	-0,063	-0,044	0,024	0,031	0,043	0,043	-0,147	-0,061	0,002	0,010
Fosfato	0,037	0,057	-0,122	-0,049	0,007	-0,032	0,021	-0,006	0,118	-0,097	0,033	0,005
Nitrogênio total	-0,047	-0,064	-0,045	0,042	-0,022	0,123	-0,052	-0,073	-0,012	-0,029	-0,058	-0,024
Nitrato	-0,009	-0,017	-0,009	0,035	-0,046	-0,097	0,036	0,018	0,014	0,053	0,034	0,013
Nitrito	-0,020	0,080	0,034	-0,006	-0,007	0,034	-0,042	-0,035	0,047	-0,011	-0,030	-0,011
Amônio	-0,002	-0,017	0,082	0,082	-0,034	-0,062	0,046	0,102	0,015	-0,002	-0,063	-0,058
Oxigênio Dissolvido	0,005	0,002	-0,088	-0,061	0,001	-0,043	-0,023	0,020	-0,014	-0,007	-0,017	-0,013
Condutividade	0,034	0,083	0,016	0,009	0,073	0,032	0,031	-0,041	0,005	0,009	0,028	0,016
Alcalinidade	0,058	0,067	-0,088	-0,037	0,189	0,009	-0,009	0,081	0,021	0,009	-0,029	-0,038
Cálcio	-0,039	0,053	-0,060	-0,115	0,077	-0,076	-0,125	-0,042	-0,016	0,065	-0,025	0,001
pH	-0,051	-0,021	-0,026	-0,052	-0,039	-0,040	0,027	0,010	-0,060	-0,011	0,011	0,004
Autovalores	0,256	0,194	0,171	0,149	0,111	0,070	0,066	0,053	0,046	0,032	0,026	0,017
Variância explicada (%)	1,07	0,81	0,71	0,62	0,46	0,29	0,27	0,22	0,19	0,13	0,11	0,07
Variância acumulada (%)	96,1	96,9	97,6	98,2	98,7	99,0	99,3	99,5	99,7	99,8	99,9	100,0

APÊNDICE B – RESULTADOS DA MODELAGEM DE SÉRIE TEMPORAL COM REDE NEURAL ARTIFICIAL, NEURO-FUZZY EMÁQUINA DE VETORES DE SUPORTE EM CONJUNÇÃO COM A TRANSFORMADA WAVELET.

Tabela B1 – Resultados do desempenho para SVM com a Transformada *Wavelet* (modelo 2 – Função RBF) para a série temporal de clorofila (continua).

		RBF			
***Wavelet* - mãe**	**Passo de tempo**	**RMSE - Treinamento**	**R^2 - Treinamento**	**RMSE - Teste**	**R^2 - Teste**
	1	24.882	0.51549	24.333	0.19706
	2	24.696	0.52596	22.868	0.29085
Dmey	3	24.1	0.55153	20.795	0.41357
	4	24.633	0.53464	21.996	0.34387
	5	25.163	0.51754	22.597	0.30757
	1	25.343	0.49737	22.646	0.30453
	2	25.994	0.47484	24.04	0.21627
Haar	3	26.252	0.46785	24.092	0.21291
(db1)	4	24.752	0.53015	22.52	0.31223
	5	25.275	0.51322	22.693	0.30168
	1	25.637	0.48564	22.443	0.31693
	2	25.415	0.49799	22.371	0.32134
Coif1	3	25.667	0.4913	23.259	0.26636
	4	24.556	0.53755	20.301	0.44109
	5	26.552	0.46279	23.983	0.21998
	1	25.403	0.495	22.836	0.29285
	2	24.515	0.53289	21.9	0.34961
Coif3	3	24.742	0.52732	22.166	0.33371
	4	24.562	0.53733	20.822	0.41206
	5	25.077	0.52083	21.856	0.35221

Tabela B1 – Resultados do desempenho para SVM com a Transformada *Wavelet* (modelo 2 – Função RBF) para a série temporal de clorofila (continua).

		RBF			
***Wavelet* - mãe**	**Passo de tempo**	**RMSE - Treinamento**	**R^2 - Treinamento**	**RMSE - Teste**	**R^2 - Teste**
	1	25.156	0.50476	23.293	0.26422
	2	24.593	0.52991	22.751	0.2981
Coif5	3	24.145	0.54985	20.282	0.44215
	4	24.778	0.52914	22.011	0.34299
	5	24.56	0.54038	20.546	0.42755
	1	25.092	0.50728	22.104	0.33742
	2	25.638	0.48912	23.676	0.23984
Bior 1.1	3	24.38	0.54106	21.963	0.34586
	4	26.009	0.48118	23.956	0.22177
	5	25.234	0.5148	22.599	0.30741
	1	23.89	0.55336	24.23	0.20385
	2	25.063	0.51178	22.563	0.30964
Bior 3.3	3	24.709	0.52855	21.345	0.38218
	4	25.095	0.51703	21.577	0.36862
	5	24.63	0.53774	22.686	0.30208
	1	24.625	0.52544	22.314	0.32481
	2	25.003	0.51413	23.015	0.28168
Db2	3	24.714	0.52839	22.876	0.29037
	4	24.844	0.52662	22.508	0.31301
	5	25.231	0.51491	22.462	0.31577
	1	25.099	0.507	21.157	0.39296
	2	24.199	0.54486	19.667	0.4755
Db5	3	24.425	0.53934	20.904	0.40741
	4	24.325	0.54622	21.067	0.39814
	5	24.595	0.53905	20.847	0.41062

Tabela B1 – Resultados do desempenho para SVM com a Transformada *Wavelet* (modelo 2 – Função RBF) para a série temporal de clorofila (conclusão).

		RBF			
***Wavelet* - mãe**	**Passo de tempo**	**RMSE - Treinamento**	**R^2 - Treinamento**	**RMSE - Teste**	**R^2 - Teste**
	1	24,7784	0,51951	24,1978	0,20597
	2	24,3973	0,53738	22,2486	0,32873
Db10	3	25,3348	0,50439	23,1391	0,27393
	4	25,6378	0,4959	23,0001	0,28263
	5	25,0959	0,5201	20,6928	0,41933
	1	25,2253	0,50202	24,0806	0,21364
	2	24,5237	0,53257	22,5189	0,31232
Db20	3	25,1172	0,51287	23,5114	0,25037
	4	25,199	0,51301	23,142	0,27374
	5	25,0692	0,52112	22,0454	0,34094

Tabela B2 – Resultados do desempenho para SVM com a Transformada *Wavelet* (modelo 2 – Função Polinomial De Grau 3) para a série temporal de clorofila (continua).

POLINOMIAL DE GRAU 3					
***Wavelet* - mãe**	**Passo de tempo**	**RMSE - Treinamento**	**R^2 - Treinamento**	**RMSE - Teste**	**R^2 - Teste**
	1	11.348	0.89922	37.43	-0.89992
	2	7.6135	0.95495	17.666	0.57676
Dmey	3	5.484	0.97678	27.975	-0.061278
	4	7.2769	0.95939	15.627	0.66883
	5	7.4019	0.95825	17.676	0.5763
	1	22.221	0.61356	85.981	-9.0253
	2	17.844	0.75254	78.618	-7.3817
Haar	3	14.996	0.82635	84.869	-8.7674
	4	15.039	0.82654	67.282	-5.1389
	5	17.442	0.76817	24.816	0.16487
	1	12,1322	0,88481	46,2918	-1,906
	2	9,68678	0,92707	32,1841	-0,40466
Coif1	3	11,2649	0,90202	33,7913	-0,54845
	4	5,00291	0,9808	90,5678	-10,1233
	5	11,9296	0,89156	22,0206	0,342426
	1	11,7664	0,89165	39,2557	-1,08974
	2	6,92729	0,9627	38,0967	-0,96816
Coif3	3	8,60284	0,94285	24,6318	0,177225
	4	**9.781**	**0.92663**	**14.519**	**0.71413**
	5	8,84994	0,94032	15,2715	0,683737
	1	11,2412	0,90111	38,0261	-0,96088
	2	7,05642	0,9613	26,9141	0,01769
Coif5	3	6,22173	0,97011	31,9153	-0,3813
	4	8,21131	0,94829	15,1598	0,688344
	5	6,93932	0,96331	17,5786	0,580957

Tabela B2 – Resultados do desempenho para SVM com a Transformada *Wavelet* (modelo 2 – Função Polinomial De Grau 3) para a série temporal de clorofila (continua).

POLINOMIAL DE GRAU 3					
***Wavelet* - mãe**	**Passo de tempo**	**RMSE - Treinamento**	**R^2 - Treinamento**	**RMSE - Teste**	**R^2 - Teste**
	1	22,293	0,61106	92,745	-10,6646
	2	18,6282	0,7303	71,3159	-5,89701
Bio 1.1	3	12,4237	0,88082	110,878	-15,6717
	4	17,5726	0,76317	41,0455	-1,28464
	5	13,342	0,86436	31,1474	-0,31562
	1	15,2478	0,81805	79,0225	-7,46817
	2	8,76362	0,94031	40,9416	-1,27309
Bio 3.3	3	9,13779	0,93553	32,4915	-0,43162
	4	9,51917	0,9305	36,2093	-0,77798
	5	9,06183	0,93743	29,2443	-0,15977
Db2	1	11,1439	0,90281	81,6392	-8,03827
	2	7,27165	0,9589	26,3784	0,056406
	3	7,96218	0,95105	19,0736	0,506651
	4	7,99654	0,95096	21,7708	0,357257
	5	5,95433	0,97298	18,0541	0,557984
Db5	1	13,41	0,85927	77,6702	-7,18081
	2	8,8278	0,93943	46,0547	-1,87631
	3	7,57736	0,95567	49,7924	-2,36212
	4	10,3133	0,91843	33,8119	-0,55034
	5	10,9369	0,90885	23,9407	0,222751
Db10	1	13,8014	0,85093	21,0614	0,398464
	2	6,64535	0,96568	22,6477	0,304435
	3	5,72175	0,97472	38,226	-0,98155
	4	5,26698	0,97872	33,4213	-0,51473
	5	5,27447	0,9788	26,6596	0,036184

Tabela B2 – Resultados do desempenho para SVM com a Transformada *Wavelet* (modelo 2 – Função Polinomial De Grau 3) para a série temporal de clorofila (conclusão).

POLINOMIAL DE GRAU 3					
***Wavelet* - mãe**	**Passo de tempo**	**RMSE - Treinamento**	**R^2 - Treinamento**	**RMSE - Teste**	**R^2 - Teste**
Db20	1	13,2533	0,86254	32,825	-0,46116
	2	8,89	0,93858	20,6302	0,422843
	3	8,56181	0,9434	24,8335	0,163699
	4	9,70801	0,92772	21,653	0,364194
	5	7,64667	0,95545	22,643	0,30473

Tabela B3 – Resultados do desempenho para SVM com a Transformada *Wavelet* (modelo 2 Polinomial de Grau 1) para a série temporal de clorofila (continua).

POLINOMIAL DE GRAU 1 (LINEAR)					
***Wavelet* - mãe**	**Passo de tempo**	**RMSE - Treinamento**	**R^2 - Treinamento**	**RMSE - Teste**	**R^2 - Teste**
Dmey	1	21.252	0.64654	19.955	0.46002
	2	22.087	0.62085	19.128	0.50383
	3	21.346	0.64816	18.587	0.53151
	4	24.266	0.5484	20.921	0.40648
	5	22.416	0.61712	20.317	0.44022
Haar	1	24,2315	0,54048	12,3449	0,79334
	2	24,1883	0,54527	17,8665	0,56712
	3	27,3581	0,42207	19,2209	0,499
	4	26,4402	0,46385	18,8084	0,52027
	5	27,0307	0,44325	20,6688	0,42068
Coif1	1	21,104	0,65145	18,5864	0,53153
	2	20,3175	0,67917	16,9802	0,609
	3	24,7521	0,52692	20,6018	0,42443
	4	22,2404	0,62065	18,9325	0,51393
	5	26,0661	0,48227	22,5321	0,31152

Tabela B3 – Resultados do desempenho para SVM com a Transformada *Wavelet* (modelo 2 Polinomial de Grau 1) para a série temporal de clorofila (continuação).

POLINOMIAL DE GRAU 1 (LINEAR)					
	1	21,6343	0,63371	18,6918	0,52621
	2	18,8816	0,72291	15,8427	0,65963
Coif3	3	17,3731	0,76694	14,0354	0,73286
	4	23,3147	0,58311	19,9707	0,45915
	5	21,9107	0,63419	19,3826	0,49054
	1	19,517	0,7019	18,1307	0,55422
	2	19,242	0,71223	16,3007	0,63967
Coif5	3	20,663	0,67032	17,3016	0,59406
	4	21,8578	0,63359	18,8676	0,51725
	5	23,477	0,58002	20,3886	0,43628
	1	24,6671	0,52381	11,3959	0,82389
	2	24,6354	0,5283	17,9843	0,56139
Bio 1.1	3	27,7815	0,40404	19,8888	0,46358
	4	27,6479	0,41375	20,4028	0,4355
	5	27,3616	0,42953	21,3159	0,38384
	1	22,577	0,60109	25,9301	0,08821
	2	23,0845	0,58583	24,1178	0,21121
Bio 3.3	3	24,4128	0,53981	24,7734	0,16774
	4	28,6303	0,37135	26,2783	0,06355
	5	28,3638	0,38698	26,1301	0,07409
	1	23,0371	0,58466	22,6613	0,3036
	2	21,6505	0,63569	21,3743	0,38046
Db2	3	27,2738	0,42562	24,3048	0,19893
	4	27,3365	0,42688	24,4641	0,1884
	5	27,7115	0,41485	24,5013	0,18592

Tabela B3 – Resultados do desempenho para SVM com a Transformada *Wavelet* (modelo 2 Polinomial de Grau 1) para a série temporal de clorofila (conclusão).

POLINOMIAL DE GRAU 1 (LINEAR)					
***Wavelet* - mãe**	**Passo de tempo**	**RMSE - Treinamento**	**R^2 - Treinamento**	**RMSE - Teste**	**R^2 - Teste**
	1	21,4186	0,64098	22,491	0,31403
	2	22,4107	0,60965	20,122	0,45093
Db5	3	26,0795	0,47483	21,9475	0,34678
	4	24,7774	0,52916	21,2163	0,38958
	5	25,9305	0,48764	22,4429	0,31696
	1	22,5975	0,60037	21,3778	0,38026
	2	18,8236	0,72461	17,5355	0,58301
Db10	3	19,0856	0,71873	18,911	0,51503
	4	21,161	0,65658	21,1275	0,39468
	5	22,3514	0,61932	22,0034	0,34345
	1	20,9078	0,65789	20,8492	0,41053
	2	23,7348	0,56216	20,2721	0,44271
Db20	3	23,5523	0,57168	20,1842	0,44753
	4	22,9108	0,59743	20,0683	0,45385
	5	23,8805	0,56545	20,9575	0,40438

Tabela B4 – Resultados do desempenho para RNA com a Transformada *Wavelet* (modelo 3) para a série temporal de clorofila (continua).

Wavelet - mãe	Passo de tempo	RMSE - Treinamento	R2 - Treinamento	RMSE - Teste	R2 - Teste
	1	0,68374	0,99563	1,90825	0,99378
	2	2,25761	0,95093	6,19322	0,96424
Dmey	3	1,70417	0,97216	5,31432	0,97032
	4	0,97026	0,99114	2,11979	0,99355
	5	2,02771	0,98509	0,61809	0,99314
	1	4,09023	0,83568	10,86600	0,86280
	2	4,28019	0,81098	8,56113	0,90693
Haar	3	3,57309	0,87239	13,16511	0,77519
	4	3,16465	0,89997	9,89414	0,87123
	5	3,65636	0,86761	9,73078	0,84076
	1	4,16389	0,81986	8,77143	0,91960
	2	3,18842	0,90129	8,87547	0,93039
Coif1	3	2,46134	0,94108	7,26932	0,94897
	4	2,85354	0,92106	11,39342	0,73681
	5	2,74131	0,93142	6,14386	0,93277
	1	4,55203	0,79321	6,72094	0,93755
	2	4,33486	0,84106	7,93089	0,91436
Coif3	3	2,46431	0,94229	4,29302	0,96972
	4	1,65276	0,97622	4,06227	0,98047
	5	1,17733	0,98685	4,46833	0,98395
	1	4,17540	0,81910	8,16327	0,93565
	2	2,23300	0,95427	5,66966	0,96507
Coif5	3	1,80158	0,96925	4,95358	0,95557
	4	1,20433	0,98610	3,19593	0,98751
	5	0,97952	0,99093	4,28054	0,98636
	1	6,15232	0,78561	12,48519	0,88528
	2	3,57687	0,87599	8,22946	0,90951
Bio 1.1	3	4,97263	0,75815	10,38894	0,87732
	4	3,44064	0,88188	10,49285	0,88197
	5	4,03918	0,84989	9,92406	0,88484

Tabela B4 – Resultados do desempenho para RNA com a Transformada *Wavelet* (modelo 3) para a série temporal de clorofila (conclusão).

***Wavelet* - mãe**	**Passo de tempo**	**RMSE - Treinamento**	**R2 - Treinamento**	**RMSE - Teste**	**R2 - Teste**
	1	4,64661	0,78539	7,11133	0,92822
	2	3,09307	0,90650	6,21474	0,96650
Bio 3.3	3	2,60063	0,94094	3,75899	0,97549
	4	2,70254	0,93225	6,52511	0,95904
	5	1,86384	0,96678	15,72768	0,66826
	1	4,08154	0,82893	8,64332	0,91960
	2	4,01074	0,87008	5,53516	0,95792
Db2	3	2,96511	0,91418	9,84215	0,89181
	4	2,62177	0,93378	16,04120	0,54752
	5	2,46712	0,95249	18,23102	0,27419
	1	4,33820	0,80502	7,76763	0,94302
	2	2,45988	0,94165	6,23522	0,95051
Db5	3	2,21156	0,95369	6,27344	0,96822
	4	1,64222	0,97466	5,65809	0,97180
	5	1,44838	0,98007	6,48491	0,95342
	1	3,98460	0,83741	7,77895	0,92183
	2	1,89618	0,96681	4,85036	0,97215
Db10	3	1,86132	0,96688	4,94277	0,96669
	4	1,21955	0,98760	3,63758	0,98433
	5	1,11969	0,98912	2,58107	0,98973
Db20	1	4,68295	0,77695	6,47881	0,94411
	2	4,77399	0,89104	6,65011	0,94264
	3	1,50921	0,97825	6,74996	0,93397
	4	1,03995	0,99019	3,12867	0,98760
	5	**0,85026**	**0,99320**	**2,28134**	**0,99135**

Tabela B5 – Resultados do desempenho para SVM com a Transformada *Wavelet* (modelo 3) para a série temporal de clorofila (continua).

Wavelet - mãe	Passo de tempo	RMSE - Treinamento	R2 - Treinamento	RMSE - Teste	R2 - Teste
	1	2,10556	0,96324	16,46876	0,38354
	2	**2,33641**	**0,94824**	**15,32123**	**0,64136**
Dmey	3	2,43900	0,94553	16,46269	0,45613
	4	1,94342	0,96831	16,64572	0,41456
	5	2,15260	0,96189	16,49606	0,37199
	1	4,25876	0,81597	15,67705	0,61428
	2	3,48339	0,87934	16,20586	0,53513
Haar	3	3,59986	0,86969	16,23607	0,46901
	4	3,17427	0,90396	16,66343	0,32627
	5	3,46121	0,89270	16,56774	0,28252
	1	4,26884	0,83237	15,87167	0,54292
	2	3,35097	0,90323	16,56653	0,47693
Coif1	3	3,25822	0,90629	16,82052	0,40353
	4	2,92985	0,93203	16,94491	0,27657
	5	2,81240	0,93576	16,56533	0,34318
	1	3,74473	0,87017	15,32188	0,58445
	2	2,54468	0,94255	15,65184	0,57065
Coif3	3	2,58909	0,94013	16,23453	0,47457
	4	2,53641	0,94776	16,50394	0,34772
	5	2,59154	0,94558	16,32238	0,35954
	1	3,78523	0,85862	15,38798	0,57670
	2	2,41386	0,94687	15,68343	0,57244
Coif5	3	2,51494	0,94174	16,36521	0,43865
	4	2,32708	0,95474	16,45479	0,37687
	5	2,37198	0,95324	16,27114	0,39538

Tabela B5 – Resultados do desempenho para SVM com a Transformada *Wavelet* (modelo 3) para a série temporal de clorofila (continuação).

Wavelet - mãe	Passo de tempo	RMSE - Treinamento	R2 - Treinamento	RMSE - Teste	R2 - Teste
	1	4,21307	0,82032	15,41493	0,63129
	2	3,52534	0,87684	16,07638	0,54639
Bio 1.1	3	3,58790	0,87015	16,33738	0,44662
	4	3,21388	0,90178	16,62769	0,32179
	5	3,48669	0,89047	16,56291	0,28529
	1	4,41905	0,79696	16,00375	0,50783
	2	3,62533	0,86920	15,85812	0,56464
Bio 3.3	3	3,49843	0,88127	15,80601	0,55393
	4	3,40470	0,88994	16,16199	0,47328
	5	3,36912	0,89165	16,25546	0,44839
	1	4,60391	0,77450	15,16740	0,53595
	2	3,91024	0,84544	15,68821	0,51033
Db2	3	3,84513	0,85170	16,04431	0,44885
	4	3,66756	0,86336	16,50515	0,36426
	5	3,73135	0,86048	16,62167	0,35895
	1	4,06878	0,83374	15,86096	0,50496
	2	2,97331	0,91898	15,92011	0,49985
Db5	3	2,95998	0,92517	16,04899	0,43004
	4	2,83530	0,93492	16,08136	0,39849
	5	2,96314	0,92861	16,24223	0,36379
	1	3,84005	0,84964	15,72832	0,56474
	2	2,34968	0,94757	15,57979	0,63635
Db10	3	2,29852	0,95134	15,91257	0,55439
	4	1,96494	0,96671	15,92011	0,53050
	5	2,12259	0,96532	15,88679	0,48832

Tabela B5 – Resultados do desempenho para SVM com a Transformada *Wavelet* (modelo 3) para a série temporal de clorofila (conclusão).

Wavelet - mãe	Passo de tempo	RMSE - Treinamento	R2 - Treinamento	RMSE - Teste	R2 - Teste
	1	3,81563	0,85663	15,58236	0,57445
	2	2,54584	0,93992	15,47966	0,60689
Db20	3	2,55572	0,94163	16,18889	0,44461
	4	2,24062	0,95898	16,18456	0,41070
	5	2,27607	0,95855	16,00625	0,44619

Tabela B6 – Resultados do desempenho para NF com a Transformada *Wavelet* (modelo 3) para a série temporal de clorofila (continua).

Wavelet - mãe	Passo de tempo	RMSE - Treinamento	R2 - Treinamento	RMSE - Teste	R2 - Teste
	1	4,16185	0,82202	5,00919	0,96223
	2	1,90835	0,96512	3,67913	0,97730
Dmey	3	1,15853	0,98725	10,10891	0,85298
	4	0,47174	0,99788	28,90588	0,54713
	5	0,24498	0,99943	42,97790	0,28092
	1	4,16713	0,82657	6,01357	0,94090
	2	3,47433	0,88146	7,54142	0,90213
Haar	3	3,05550	0,90828	71,87837	-0,23212
	4	2,72153	0,92674	245,66848	0,34759
	5	2,57630	0,93493	1414,21356	0,24000
	1	4,21248	0,81537	5,37476	0,95550
	2	2,56747	0,93712	6,62103	0,93333
Coif1	3	1,94546	0,96359	25,55230	0,33515
	4	1,09229	0,98854	90,63057	0,14580
	5	0,75486	0,99458	143,36666	0,36773
	1	4,14898	0,82289	4,77630	0,96128
	2	1,86188	0,96687	4,65210	0,97311
Coif3	3	1,19227	0,98651	15,93142	0,69128
	4	0,63545	0,99616	28,34061	0,50709
	5	0,37344	0,99868	61,54835	0,32346

Tabela B6 – Resultados do desempenho para NF com a Transformada *Wavelet* (modelo 3) para a série temporal de clorofila (conclusão).

Wavelet - mãe	Passo de tempo	RMSE - Treinamento	R2 - Treinamento	RMSE - Teste	R2 - Teste
	1	4,16485294	0,82166	4,77063937	0,96248
	2	1,76360	0,97031	4,25829	0,97301
Coif5	3	1,09895	0,98853	13,31503	0,78082
	4	0,54194	0,99720	28,91833	0,42056
	5	0,31402	0,99906	56,71155	0,36410
	1	4,16713	0,82657	6,01357	0,94090
	2	3,47433	0,88147	7,54142	0,90213
Bio 1.1	3	3,05550	0,90828	71,87837	-0,23212
	4	2,72153	0,92674	245,66848	0,34759
	5	2,57630	0,93493	1705,87221	0,24784
	1	4,04228	0,83130	5,56911	0,94342
	2	2,48707	0,93999	9,04743	0,86999
Bio 3.3	3	1,69623	0,97167	56,76531	0,31727
	4	0,96666	0,99106	103,74006	0,12582
	5	0,54197	0,99727	141,50618	0,09632
	1	4,09329	0,82654	6,61400	0,95053
	2	2,85238	0,92089	15,24795	0,60340
Db2	3	1,83510	0,96763	21,30751	0,45482
	4	1,14258	0,98755	1029,56301	0,05627
	5	0,67645	0,99565	224,62636	-0,59245
	1	4,11193	0,82560	4,90785	0,95665
	2	2,13237	0,95717	8,90736	0,85096
Db5	3	1,48856	0,97903	19,04600	0,43120
	4	0,83946	0,99343	102,10779	0,11854
	5	0,45913	0,99801	93,16866	0,05195