AUTARQUIA ASSOCIADA À UNIVERSIDADE DE SÃO PAULO

Mapas Auto - organizáveis de Kohonen (SOM) aplicados na avaliação dos parâmetros da qualidade da água

Gustavo Sousa Affonso

Dissertação apresentada como parte dos requisitos para obtenção do Grau de Mestre em Ciências na Área de Tecnologia Nuclear - Reatores

Orientador:
Prof. Dr. Roberto Navarro de Mesquita

São Paulo
2011

Autarquia associada à universidade de São Paulo

MAPAS AUTO - ORGANIZÁVEIS DE KOHONEN (SOM) APLICADOS NA AVALIAÇÃO DOS PARÂMETROS DA QUALIDADE DA ÁGUA.

GUSTAVO SOUSA AFFONSO

Dissertação apresentada como parte dos requisitos para obtenção do Grau de Mestre em Ciências na Área de Tecnologia Nuclear – Reatores

Orientador:
Dr. Roberto Navarro de Mesquita

São Paulo

2011

INSTITUTO DE PESQUISAS ENERGÉTICAS E NUCLEARES
Autarquia associada à universidade de São Paulo

MAPAS AUTO – ORGANIZÁVEIS DE KOHONEN (SOM) APLICADOS NA AVALIAÇÃO DOS PARÂMETROS DA QUALIDADE DA ÁGUA.

GUSTAVO SOUSA AFFONSO

Dissertação apresentada como parte dos requisitos para obtenção do Grau de Mestre em Ciências na Área de Tecnologia Nuclear – Reatores

Orientador:
Dr. Roberto Navarro de Mesquita

São Paulo
2011

“Aos meus pais e minha família.”

AGRADECIMENTOS

Meus sinceros agradecimentos a todos que contribuíram direta e indiretamente para a realização deste trabalho.

Ao Professor Doutor Roberto Navarro de Mesquita por sua orientação e especialmente pela compreensão e paciência.

Ao Professor Doutor Hélio Akira Furusawa por sua cooperação nos diversos aspectos que contemplam este trabalho, por sua disponibilidade e auxilio nos momentos difíceis e por sua amizade.

E ao Centro de Química e Meio Ambiente, pela colaboração na realização deste projeto.

MAPAS AUTO - ORGANIZÁVEIS DE KOHONEN (SOM) APLICADOS NA AVALIAÇÃO DOS PARÂMETROS DA QUALIDADE DA ÁGUA.

Gustavo Sousa Affonso

RESUMO

A atual crescente necessidade de análise de coleções de dados cada vez mais complexas e extensas, nas diversas áreas da investigação científica, tem permitido o desenvolvimento de novas ferramentas para a melhoria da percepção de informações que nem sempre são explícitas e visíveis. Estudos de ferramentas matemáticas que propiciem o destaque de algumas destas informações, ou que inteligentemente reconheçam padrões associados aos diferentes conjuntos de dados, têm demonstrado resultados promissores. No entanto, o sucesso da escolha da metodologia apropriada para a análise dos dados, está vinculado a vários fatores como: a tecnologia disponível para a prospecção destes dados, a adequada coleta e seleção das amostras, e principalmente, a capacidade do pesquisador em interagir com a nova tecnologia de exploração. No presente projeto, é proposta uma metodologia de análise multidimensional dos dados de unidades de gerenciamento de recursos hídricos UGRHIs, localizadas no estado de São Paulo, por meio das redes neurais SOM (Mapas Auto-Organizáveis). Estes mapas são utilizados 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.

SELF - ORGANIZING MAPS OF KOHONEN (SOM) APPLIED IN THE EVALUATION OF PARAMETERS OF WATER QUALITY

Gustavo Sousa Affonso

ABSTRACT

The current increasingly need for data analysis on larger and more complex data collections, in many different areas of scientific research, has induced the development of new tools for the perception improvement of information that not always is explicit and visible at first. Studies of mathematical tools which could enable the highlight of some of this information, or should intelligently recognize patterns associated with these different data collection, have been showing promising results. However, the success of the choice of the appropriate analysis method is associated with several factors: the available technology for this data exploration, the correct gathering and selection of samples, and mainly, the researcher ability to interact with the new exploration technology. In this project we propose a methodology for analyzing multidimensional data from Water Resources Management Units (WRMU's), which are located in São Paulo state, through Self - Organizing Maps (SOM) neural networks. These maps are used to study and visualize possible correlations between the different variables existent in this database, which are derived from analysis of inorganic and physical - chemical parameters related to WRMU's water quality.

SUMÁRIO

Página

LISTA DE FIGURAS

Página

LISTA DE ABREVIATURAS

ART	Adaptive Ressonance Theory
BMU	Best Matching Unit
CESTESB	Companhia Ambiental do Estado de São Paulo
CONAMA	Conselho Nacional do Meio Ambiente
CQMA	Centro de Química e Meio Ambiente
DBO	Demanda Bioquímica de oxigênio
DQO	Demanda Química de oxigênio
ETA	Estação de Tratamento de água
IA	Inteligência Artificial
IAP	Índice de Qualidade das Águas Brutas para Fins de Abastecimento Público
IPEN	Instituto de Pesquisas Energéticas e Nucleares
KSOM	Kohonen Self Organizing Maps
LMS	Least Mean Square Algorithm
MPL	Multilayer Perceptron
MS	Ministério da Saúde
NKT	Nitrogênio Kjeldahl total
NSF	National Sanitation Foundation
OD	Oxigênio Dissolvido
PCA	Principal Components Analysis
PFTHM	Potencial de formação de trihalometanos
pH	Potencial hidrogeniônico
RNA	Rede Neural Artificial
SABESP	Companhia de Saneamento Básico do Estado de São Paulo
SOM	Self Organizing Maps
UGRH	Unidade de Gerenciamento de Recursos Hídricos

1. INTRODUÇÃO

A importância do tratamento da informação é atualmente reconhecida nos mais diversos campos das pesquisas cientificas e sociais, e tem proporcionado o desenvolvimento de novas ferramentas interdisciplinares No entanto, em uma coleção de dados quer seja exígua ou numerosa, a percepção do que representam estas informações nem sempre é direta. Assim, é de grande valia o conhecimento sobre as técnicas disponíveis para manipulação desses dados que permitam o destaque de algumas das informações ou que inteligentemente reconheçam padrões existentes e potencialmente relevantes. O sucesso da escolha da estratégia de prospecção dos dados está vinculado a vários fatores, como à tecnologia disponível para esta prospecção, à coleta e seleção apropriada das amostras, mas principalmente ao conhecimento sobre a informação de interesse e à capacidade que o pesquisador tem de interagir com a tecnologia de exploração. Esta interação, comum a qualquer pesquisa cientifica sistemática, inclui a modificação da metodologia tanto em seus parâmetros fundamentais quanto na utilização apropriada das técnicas de seleção do espaço amostral. O desenvolvimento de técnicas de inteligência artificial nas últimas décadas tem disponibilizado novos recursos ao pesquisador no sentido de automatizar etapas dessas interações. Assim, a interatividade no uso das técnicas que extraiam relações e parâmetros de interesse da base de dados pode ser melhorada e aperfeiçoada na medida em que a própria estratégia e conhecimento do pesquisador sobre o problema são explicitados e sistematizados. Muitos dos trabalhos envolvendo técnicas de inteligência artificial resultaram na possibilidade de comparação objetiva entre diferentes metodologias e estratégias de tratamento dos dados, levando até mesmo ao estabelecimento de padrões de referência (‘*benchmarks*’).

Entre algumas das vantagens na metodologia há a capacidade que as técnicas de inteligência artificial apresentam em repetir exaustivamente tarefas associadas à pequenas modificações dos parâmetros de exploração. Com a possibilidade de se utilizar “diferentes ajustes” operacionais experimentais da ferramenta exploratória, sem que necessariamente estas modificações sejam parte implícita da ferramenta sendo utilizada. Em geral, a inteligência artificial possibilita a automatização de etapas e procedimentos envolvidos no processo de exploração ou mineração de dados.

Entre as diversas técnicas de inteligência artificial, está a técnica dos mapas auto-organizáveis de Kohonen (*Self Organizing Maps*, KSOM) para exploração de bases de dados multidimensionais. Técnica que inicialmente foi estabelecida por Teuvo Kohonen, em 1981, e consiste em uma rede neural artificial interconectada e não supervisionada que permite um mapeamento auto – ajustável do espaço de estados multidimensionais estudado. O SOM pode ser utilizado para um estudo mais amplo da correlação entre as múltiplas variáveis existentes em um fenômeno sem previamente restringir o número de variáveis a serem analisadas (COSTA e NETO, 2007, HONKELA, 2007). Esses mapas permitem uma visualização rápida e ampla de determinadas correlações existentes neste banco de dados, e têm sido empregados nas mais diversas áreas de pesquisa.

No campo da Inteligência Artificial (IA), esta técnica pode ser utilizada em conjunto com outras que possibilitem a automatização de procedimentos de busca e mineração dos dados. Esta integração de diferentes técnicas de IA e a implementação de uma estratégia de prospecção que represente o conhecimento do especialista constitui a chamada implementação de heurística.

O desenvolvimento de metodologias de IA que incluam ou se comparem às análises tradicionais são justificadas pela necessidade da manutenção de características do banco de dados original (caso contrário, pode ocorrer alteração do comportamento do processo, indução à geração de modelos corrompidos, por exemplo). Com o objetivo de obter padrões até então desconhecidos, relativos ao comportamento da qualidade da água, assim como da variância relativa ao tempo (COSTA e NETTO, 2007). Em alguns trabalhos recentes tem sido realizada uma comparação sistemática entre diferentes técnicas de análise multidimensional, incluindo o SOM, envolvendo grandes conjuntos de dados de indicadores químicos da qualidade da água (ASTEL et,al., 2007). Esses trabalhos tem apresentado aplicações na avaliação da qualidade da água, quer seja em sistemas naturais como rios ou associados a algum processo de alteração das características da água. Em, 2008 Kalteh et al., apresentaram um trabalho em que fazem uma revisão da aplicação do SOM em águas. Os autores reforçam a idéia de que o SOM, visto como uma rede neural, pode ser aplicado para a obtenção de agrupamentos (*clustering*), classificação, estimação, predição e mineração de dados (*data mining*) visando o reconhecimento de sinais organização de grande quantidade de dados, monitoramento e análise de processos, modelamento assim como o tratamento das variáveis ambientais. Eles ressaltam a

indicação que os vários estudos sugerem que o SOM pode superar muitos outros métodos aplicados em hidrologia.

Este trabalho se insere neste contexto de desenvolvimento de tecnologias de prospecção de dados e busca contribuir no desenvolvimento de metodologias de análise de qualidade da água através da aplicação do SOM a um conjunto de matrizes ambientais com o auxílio de algumas ferramentas estatísticas.

1.2 OBJETIVOS

Propor uma metodologia utilizando mapas auto organizáveis de Kohonen (SOM), para a análise multidimensional de uma base de dados composta por valores de parâmetros físico-químicos da qualidade da água destinada ao consumo humano (doméstico, industrial e rural) oriunda de pontos de coleta das unidades de gerenciamento de recursos hídricos de diversas regiões do estado de São Paulo.

Objetivos Específicos

Demonstrar o potencial da aplicação dos mapas auto-organizáveis de Kohonen em base de dados de qualidade da água;

Analisar as correlações entre os parâmetros oriundos de análises ambientais visando identificar potenciais correlações.

Identificar e avaliar as limitações desta tecnologia;

Estabelecer um procedimento básico para futuras aplicações.

2. REVISÃO DE LITERUTURA

2.1 Conceitos sobre análise estatística multivariada.

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 dos métodos multivariados esteja em trabalhos na psicologia, ciências sociais e biológicas, mais recentemente eles têm sido aplicados em um grande universo de áreas diferentes como: educação, geologia, química, física, engenharia, etc. Esta expansão na aplicação dessas técnicas somente foi possível graças ao grande avanço da tecnologia e ao grande numero de softwares estatísticos com módulos de análise multivariada. Trabalhos realizados no Centro de Química e Meio Ambiente (CQMA/IPEN), (COTRIM, 2006; REIS, 2006; MARQUES, 2005; LEMES, 2001) utilizando a estatística multivariada demonstraram a aplicabilidade da ferramenta na análise de dados ambientais gerados de naturezas diversas (água bruta e final, sedimento, iodo, entre outros).

Há duas principais técnicas citados na literatura: - um grupo de técnicas exploratórias de sintetização (ou simplificação) da estrutura de variabilidade dos dados e um grupo de técnicas de inferência estatística. Dentro do grupo de técnicas exploratórias podem-se destacar a análise de componentes principais, a análise de correlações canônicas, a análise de agrupamentos, a análise discriminante de correspondência (MINGOTI, 2005).

Do grupo de técnicas de inferência estatística se destacam os métodos de estimação de parâmetros, testes de hipóteses, análise de variância, covariância e de regressão multivariada.

De acordo com Echalar, 1991, uma base de dados multivariada pode ser interpretada como uma descrição das variabilidades em um sistema por meio das séries temporais das variáveis medidas.

A análise de componentes principais tem como principal aplicação a mensuração do grau de inter–relações existentes entre as variáveis envolvidas no processo, e isto pode ser observado na repetitividade de certa característica em uma série temporal denotando que talvez essa informação derive de fatores subjacentes que causem a variabilidade.
Na análise de fatores principais, busca-se substituir a descrição dessa variabilidade de variáveis medidas por outra com um menor número de variáveis, independentes, combinações lineares que representem esses fatores causais, não explícitos da variabilidade medida.

2.2 métodos multivariados e SOM.

Um conjunto importante de aplicações do SOM em análise de qualidade da água tem sido publicadas em anos recentes, demonstrando o interesse internacional na qualificação de novos métodos de análise que utilizam, em primeira instância, os novos recursos de cálculo e de interatividade disponibilizados pelo crescente avanço dos computadores nas últimas décadas.

Algumas revisões importantes foram publicadas bem recentemente analisando os diferentes métodos de análise multidimensional da qualidade da água tanto em seus aspectos espaciais como temporais (Bierman et. al., 2011, Céréghino e Park, 2009, Kalteh et. al, 2008). Os trabalhos analisados e citados por estas revisões recentes ressaltam o potencial apresentado pelo SOM no estudo da qualidade da água na visualização e exploração de relações lineares e não-lineares de dados multidimensionais, consideradas como ferramentas comprovadamente úteis na avaliação da qualidade da água (Céréghino e Park, 2009).

Entre os diversos métodos multivariados disponíveis para análise de tendências e padrões espaciais e temporais em dados de qualidade da água foram comparadas técnicas como Análise de Cluster, Análise Discriminante, Análise Fatorial e Análise de Componentes Principais (Bierman et. al., 2011). Muitos destes métodos estatísticos são utilizados em dados de amostras pontuais e são estendidos para dados adquiridos remotamente por sensoriamento e imageamento. Bierman et. al (2011) afirmam que o SOM, o Semivariograma e a Regressão Pesada Geograficamente (*Geographically Weghted*

Regression (GWR)) são mais apropriados para a análise e representação espacial dos dados relativos à qualidade da água.

Originalmente criada por Kohonen (1981a,b), o SOM foi inicialmente aplicado para reconhecimento de fala. Em 1996, foi aplicado pela primeira vez no estudo da água (Chon et al., 1996) por meio do estudo de comunidades-padrão de bentos em correntes de água e desde então tem sido aplicado com freqüência em diversos estudos relacionados dados ambientais multidimensionais.

Tison et al., 2004 classificou dados biológicos e ambientais baseado na aplicação do SOM em diatomáceas (algas biológicas).

Num trabalho mais direcionado, Mustonen et al., 2008, apresentaram uma avaliação da qualidade da água em uma estação piloto de tratamento de água utilizando uma abordagem com métodos multivariados de exploração de dados com o SOM. Os 7 parâmetros que os autores monitoram foram pH, alcalinidade, dureza, DQO, cloreto, sulfato e ferro.

A utilização concomitante da análise multivariada com a análise de componentes principais (*Principal Components Analysis*, *PCA*) e da análise com as redes neurais de Kohonen pode possibilitar a comprovação e o desenvolvimento de modelos e metodologias que possibilitem uma melhora significativa deste tipo de estudo. Tobiszewski et. al., 2010 se mostram otimistas quanto aos resultados obtidos pela utilização da associação do SOM com um algoritmo especialista. Astel et. al., 2007 já havia feito uma comparação entre a aplicação de SOM para classificação de conjuntos de dados muito grandes com as análise tradicionais como Análise de Agrupamentos (*cluster analysis*) e PCA.

Alguns estudos relacionados ao enfoque deste projeto foram publicados recentemente. A análise de uma planta de tratamento de águas residuais municipais usando o SOM, em um estudo das complexas relações existentes entre as variáveis do processo da planta de tratamento de águas residuais, foi publicado por Hong et. al., em 2002.

Os SOM também foram utilizados recentemente por Garcia e Gozalez, 2004, para estudo do tratamento e monitoramento de águas residuais, em que se propõem o desenvolvimento de técnicas de supervisão para uma planta de tratamento de águas residuais.

O modelamento de uma planta de tratamento de água residual municipal foi feito com algoritmos evolucionários auto–organizáveis por Hong e Bhamidimarri em 2003.
A avaliação do desempenho da remoção dos metais pesados em um experimento construído em zonas úmidas foi feita com a aplicação de mapas auto-organizáveis para elucidar os mecanismos da remoção do metal pesado e para predizer as concentrações, desenvolvido por Lee e Scholz em 2006.

A avaliação da saída de uma Estação de Tratamento de Água (ETA), em um rio no Mediterrâneo usando uma rede neural KSOM e a modelagem de balanço de massa foi publicado por Llorens et.al., em 2008, na qual propõem a avaliação da ferramenta KSOM para auxiliar no controle de quantidade da água e como ferramenta de supervisão.

Algumas outras aplicações recentes do SOM na análise da qualidade da água, utilizando comunidades macro invertebradas tem sido objeto de publicações recentes em biologia (SONG et. al., 2007; LEK e GIRAUDEL, 2001).

Nota–se em um conjunto crescente de pesquisadores de dados ambientais e da qualidade da água o uso de ferramentas mais sofisticadas e iterativas para a obtenção de quadros de correlação entre as variáveis, determinação das variáveis mais importantes e determinantes e a conseqüente possibilidade de obtenção de novos e melhores índices de avaliação da qualidade ambiental. A visão da necessidade do desenvolvimento deste campo de pesquisa é enfatizada por Céréghino e Park, 2009, reconhecendo um certo atraso da aplicação de redes neurais artificiais no campo da pesquisa em pesquisa da água e seus aspectos biológicos e químicos, estimulando que mais estudos de caso e trabalhos sejam coletados.

3. FUNDAMENTAÇÃO TEÓRICA

3.1 Redes Neurais.

Estudos na área de neurofisiologia no último século estimularam cientistas de diversas áreas na compreensão dos processos da mente humana e até mesmo, mimetizá-los no formato de algoritmos. Este campo de estudo e desenvolvimento, muitas vezes descrito como inteligência artificial, gerou versões do neurônio biológico que serviram como base de desenvolvimento de diversas técnicas conhecidas como redes artificiais (BUENO, 2006).

A rede neural artificial (RNA) é um sistema computacional constituído, por um conjunto de unidades de processamento individuais (chamadas de neurônio artificial) interconectados entre si com pesos que podem ser modificados de acordo com os parâmetros de qualidade que avaliam a proximidade entre a resposta desejada e a obtida.
Resumidamente o neurônio artificial, possui entradas semelhantes aos dendritos do neurônio biológico, captando as informações provenientes do meio externo (figura 1), e saídas resultantes semelhantes ao axônio.

A função interna também semelhante ao soma do neurônio biológico, desempenha a função de transformar a informação de entrada em uma nova informação.

A RNA tem a capacidade de aprender à medida que os pesos das interconexões entre os neurônios são ajustados conforme a saída desejada. Assim, uma RNA pode ser utilizada na simulação e obtenção de desempenhos e funções semelhantes as do cérebro humano em relação à cognição e aprendizado.

3.1.1 Modelo biológico

Na constituição do sistema nervoso biológico, o neurônio pode ser definido como célula nervosa altamente especializada ou como as unidades de vias de condução de estímulo nervoso, estimando-se que os seres humanos possuam a quantidade de bilhões de células nervosas interconectadas entre si.

Dos neurônios existentes, propõem–se uma divisão em seus chamados constituintes básicos: o corpo celular, os dendritos e o axônio.

Na captação dos sinais externos transmitidos na forma de impulsos por meio dos dentritos (canais de interconexão) há o processamento interno no corpo celular ou também chamado "soma" para a geração de novas informações. A interface interneural ocorre por meio de reações químicas e em regiões específicas que realizam a comunicação, denominadas de sinapses.

Em uma cadeia neural, a propagação dos estímulos nervosos percebidos pelos dentritos é realizada por meio do filamento central como observado na figura 1 denominado de axônio, o qual os conduz até os dentritos na terminação.

A comunicação entre neurônios é realizada por meio dos canais localizados no terminal axônico ou terminal de transmissão por filamentos sensíveis que desempenham a função de canais transmissores dos estímulos nervosos a outros neurônios, como mostrado na figura 1.

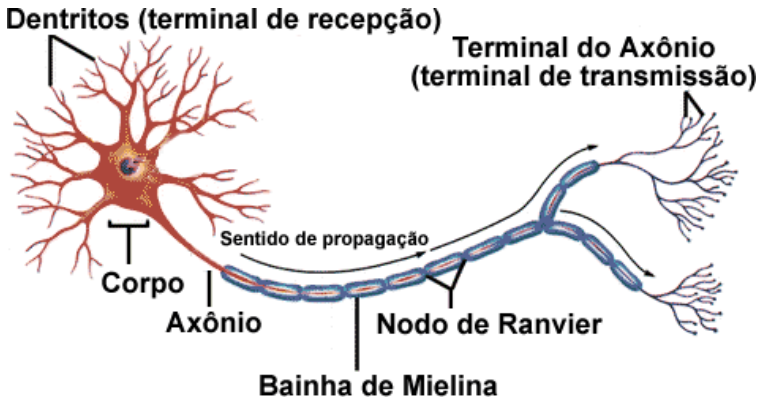

FIGURA 1 – Modelo esquemático do neurônio biológico. (BADIN, 2011).

Entre as terminações axônicas dos neurônios e os dendritos há as regiões de contato denominadas sinapses. A transmissão sináptica pode ser explicada por meio do principio da propagação do influxo nervoso como um processo excitatório por meio da liberação de

substâncias que estimulariam outros neurônios quando o impulso é percebido pelo neurônio (RANSON, 1945).

3.1.2 Breve Histórico das RNA

Entre os métodos de redes neurais artificiais desenvolvidos inicialmente, um dos mais citados e que pode ser considerado originário de vários outros modelos posteriores, é o modelo típico proposto por McMulloch e Pitts em 1943. Este modelo é constituído por um dispositivo com uma saída binária e entradas com ganhos arbitrários, podendo ser excitatórios ou inibitórios. Na Figura 2, é mostrado um esquema de um neurônio artificial típico.

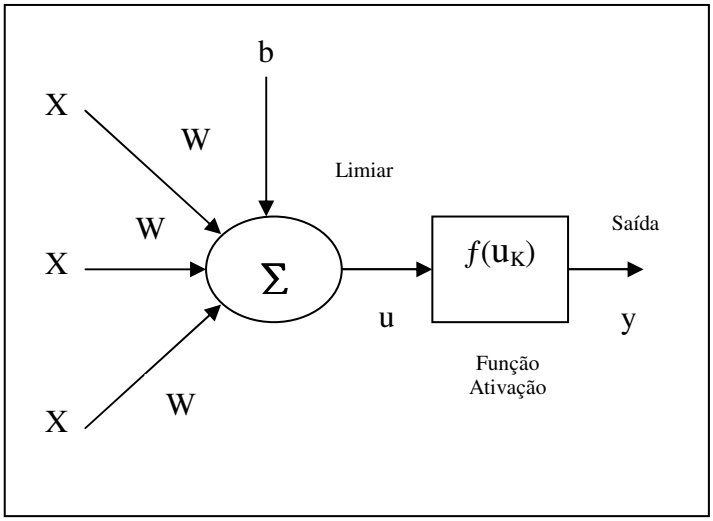

FIGURA 2 - Modelo esquemático do neurônio artificial (BUENO, 2006)

Onde: X_1, X_2,...X_n são os sinais de entrada.

Wk são os pesos sinápticos do neurônio k.

u_K é o integrador linear de saída devido aos sinais de entrada.

b é o bias e y é o sinal de saída do neurônio.

O neurônio k pode ser descrito por meio das equações 1 e 2.

$$U_k = \sum_{j=1}^{p} w_{kj} x_j + b_k \qquad (1)$$

$$y_k = \varphi(u_k) \tag{2}$$

Onde $x_1,x_2,...x_p$ são os sinais de entrada; $w_{k1},w_{k2},...w_{kp}$ são pesos sinápticos do neurônio k; u_k é o integrador linear de saída devido aos sinais de entrada, b_k é o bias; φ(.) é a função de ativação; e y_k é o sinal de saída do neurônio. O "bias" têm a função acrescer uma tendência à saída u_k do combinador linear do neurônio.

No final da década de 1950, o projeto do Perceptron foi desenvolvido por Frank Rosenblatt na Universidade de Cornell, a partir dos estudos de McCulloch. Utilizando a proposta do algoritmo de treinamento da rede baseado no estudo do biólogo Donald Hebb de 1949, que usava o ajuste gradual dos pesos de um discriminador linear. Este projeto utilizava neurônios com pesos ajustáveis para a classificação de padrões linearmente separáveis, inicialmente com 400 células fotoelétricas, e uma arquitetura que consistia de uma camada de neurônios de entrada. A rede era treinada para fornecer saídas de acordo com os dados do conjunto de treinamento, para padrões vetoriais linearmente separáveis. Esquematicamente o perceptron de uma única camada pode ser representado conforme a figura 3.

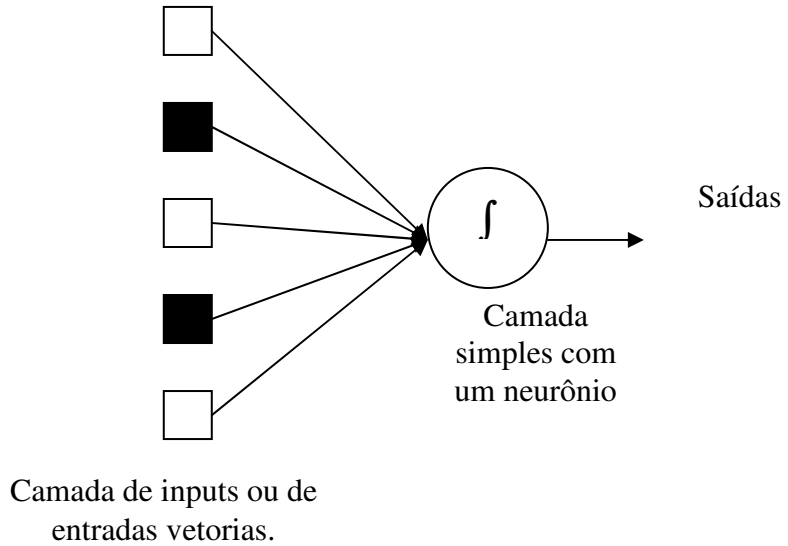

FIGURA 3 - Exemplo ilustrativo do Perceptron de Rosenblatt.

Em 1960 Bernard Widrow e Ted Hoff, desenvolveram um modelo neural linear denominado Adaline (*Adaptive Linear Element*) ou elemento linear adaptativo, e em 1962 a composição de uma rede com múltiplos elementos adaptativos, denominada Madaline (*Multiple – Adaline*) é constituída. Eles também desenvolveram um algoritmo de aprendizado baseado no conceito de minimização do desvio quadrático médio (*Least Mean Square Algorithm – LMS*), também conhecido como regra delta ou método do gradiente decrescente para a minimização do erro. (ROSSI, 2001)

Em 1969, Minsky e Papert publicam o livro intitulado Perceptrons no qual há a restrição a problemas elementares como o ou – Exclusivo e o seu complemento, limitando o teorema desenvolvido por Rosenblatt a classe de problemas linearmente separáveis, no modelo da figura 4 há exemplos ilustram as classes de limitação citadas.

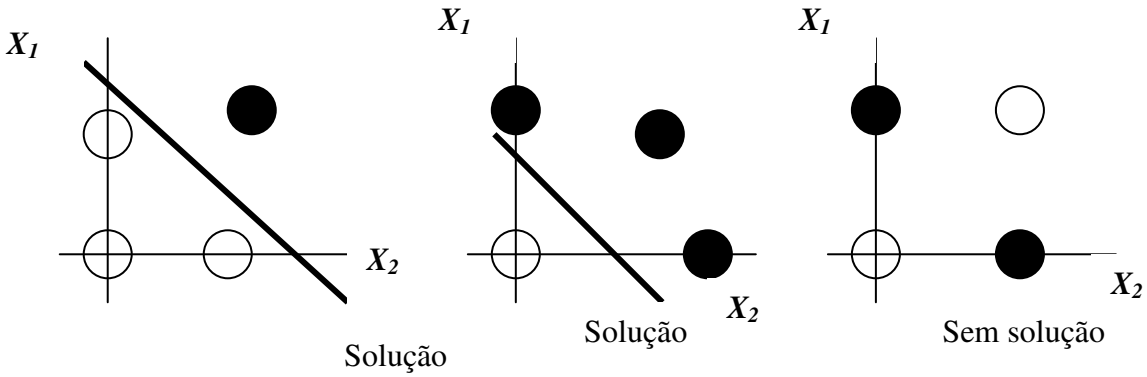

FIGURA 4 – Exemplos de classes linearmente separáveis e inseparáveis do algoritmo discriminante. (FILHO, 1998)

Na década de 70 e inicio dos anos 80 houve uma diminuição na pesquisa e produção científica sobre redes neurais, entretanto apesar de pouca atividade de pesquisas, neste período alguns trabalhos tiveram um considerável destaque. Como os estudos apresentados por Stephen Grossberg, que baseado em trabalhos sobre o aprendizado competitivo em 1987 junto com Carpenter, estabeleceram os princípios para uma nova classe de redes neurais denominadas de ART ou *Adaptive Ressonance Theory*.

Antes do modelo proposto por Stephen Grossberg haviam publicações de modelos como o propostos por James Anderson baseados em modelos biológicos da memória e de reconhecimento em 1968.

Em 1982 Hopfield utilizou a idéia de uma função de energia para um novo modo de funcionamento das redes recorrentes com conexões sinápticas simétricas, onde os elementos são ligados buscando o aprendizado com "um mínimo de energia", tendo como dados de origem as Redes de Hopfield. No mesmo ano, Teuvo Kohonen desenvolve o conceito das "redes auto-organizáveis" na qual utiliza algoritmos competitivos.

Com o desenvolvimento de modelos de memórias associativas, utilizando o conceito de aprendizado competitivo, nos quais as unidades competem para responder a determinada entrada e o elemento vencedor tem os pesos de sua entrada modificados, convergindo para responder com mais força a valores próximos do desejado.

Somente em 1986 houve reinicio, das atividades de desenvolvimento das redes neurais artificiais, com o desenvolvimento do algoritmo de retropropagação (*backpropagation*) por Rumelhart, Hinton e Williams, embora este algoritmo já tivesse sido proposto anteriormente em 1974 por Werbos em sua tese de doutorado, por Parker e LeCun em 1985. Com a publicação do livro intitulado "*Parallel Distributed Processing Explorations in the Microstructures of Cognition"*, editado por Rumelhart e McClelland, o qual apresentava o progresso das redes neurais ressurgiu o grande interesse pela técnica.

3.1.3 O Multi-Layer Perceptron

O Multilayer Perceptron (MLP), ou rede de multicamadas MLP, é uma sofisticação do modelo original do Perceptron com a ampliação do número de camadas interconectadas, e ampliou o espectro de problemas de classificação que podem ser resolvidas pela rede.

Uma MPL pode ser definida como é uma rede interconectada (conexões sinápticas) de neurônios disposta em neurônios de entrada (receptores do meio externo), neurônios da camada interna ou unidades de processamento ocultas (hidden) e neurônios de saída. (JUNIOR, 2005)

Na camada de neurônios de entrada, os vetores (dados) são recebidos e armazenados, dependendo da arquitetura da rede. Na camada mais interna entre a camada de neurônios de entrada e de saída pode haver uma camada intermediária, também chamada de oculta. A camada de saída tem a função de armazenar as respostas obtidas pela rede. O número de neurônios nessa camada corresponde ao tamanho do vetor de saída. (BUENO, 2006, FILHO, 1998). Na figura 5 é apresentada uma ilustração de um exemplo de arquitetura de rede MPL.

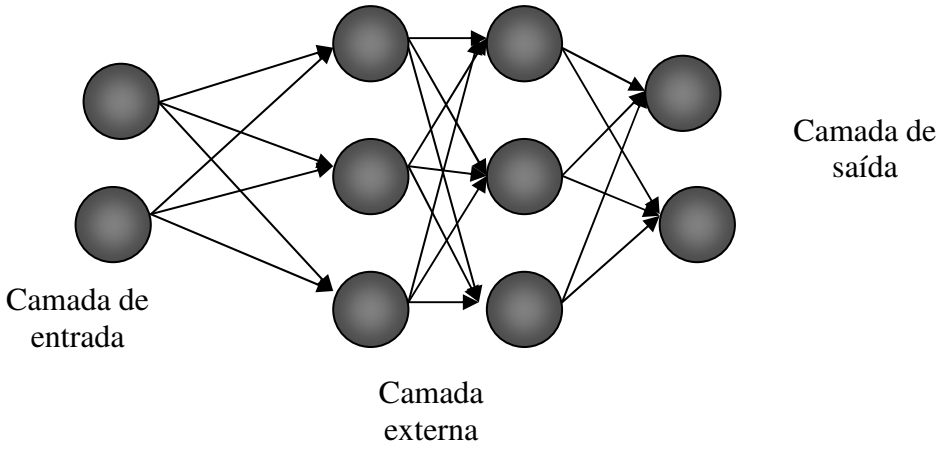

FIGURA 5 - Rede neural MPL.

Este tipo de rede neural exemplifica como as RNA's procuram explorar os princípios adotados pelo cérebro humano, apresentando um processamento altamente paralelo em sua estrutura, além de uma capacidade de generalizar o aprendizado, obtendo respostas mais abrangentes do que os dados apresentados durante o treinamento. Estes dois aspectos fazem com que as redes neurais sejam capazes de solucionar problemas altamente complexos e não-lineares.

3.1.4 Algoritmos de Aprendizagem e Treinamento

Conceitualmente, a aprendizagem da rede neural pode ser definida como um processo adaptativo mediante a resposta aos estímulos externos à rede apresentados durante a fase de treinamento. Quando as respostas desejadas aos dados de entrada são conhecidas, o processo de aprendizagem é chamado de supervisionado, pois são apresentadas à rede simultaneamente as entradas e as saídas desejadas para que ela se auto-configure através da atualização de seus pesos. Esta atualização é feita principalmente por um algoritmo

chamado de retro-propagação (*back-propagation*), que propaga para toda a rede o erro medido entre a resposta obtida e a resposta desejada (meta) da rede em questão. Devido à variedade de algoritmos existentes para o treinamento de redes, são citados alguns dos algoritmos mais comuns encontrados na literatura em uma breve apresentação dos algoritmos de aprendizagem há a definição de classes conforme o emprego de cada regra.

3.1.4.1 Regra de correção de erro

Consistem no processo de modificação dos pesos em função direta das saídas. É estimado por meio do cálculo da diferença entre a saída real gerada e a saída desejada, fornecida em um ensino supervisionado, matematicamente o princípio (LNCC, 2011) pode ser expresso como na equação 3:

$$e_k = d_k - y_k \tag{3}$$

Onde para um estímulo k,

e = sinal de erro*;*

d = saída desejada apresentada durante o treinamento*;*

y = saída real da rede após a apresentação do estímulo de entrada.

3.1.4.2 Regra de gradiente descendente

Esta regra constitui-se de um processo de alteração dos pesos (w_i), onde ocorre a minimização do erro pelo método do mínimo erro médio quadrático, e pode ser expressa pela equação 4:

$$E(w_i) = \frac{1}{2}\sum (x_o - x_p)^2 \tag{4}$$

Onde x_o é o valor observado e x_p é o valor previsto.

3.1.5 Mapas Auto Organizáveis

Os mapas auto–organizáveis inicialmente inspirados no córtex cerebral humano, consistem em uma rede neural que gera como saída representações bidimensionais (mapas) de banco de dados de alta dimensionalidade.

Desenvolvidos por Teuvo Kohonen (KOHONEN, 1981a,b), estes algoritmos podem analisar dados por agrupamentos com o objetivo de descobrir estruturas e padrões multidimensionais. Também pode ser considerada uma rede neural com aprendizado não supervisionado e competitiva, pois não necessita de um vetor de saída conhecido como vetor alvo (MESQUITA, 2002).

Estes mapas foram consolidados como redes neurais por Kohonen em conferência e artigos no começo da década de 1980. 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 de 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.

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), idéia originalmente proposta por Rosenblat em 1958.

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, porém 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.

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.

Como modelo neural, as redes SOM, conceitualmente, podem ser definidas como uma conexão entre a adaptação dos neurônios e padrões de seletividade de características. Sendo consideradas também como uma generalização não linear da heurística para análise de componentes principais (MESQUITA, 2002).

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*). 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 5.

$$i(x) = \arg\min \left\| x - w_j \right\| \quad j = 1,2,...n \tag{5}$$

Onde 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 6.

$$D_E = \sqrt{(x_1 - y_1)^2 + (x_{2-} y_2)^2 + ... + (x_n - y_{1n})^2} \tag{6}$$

Onde x_n são as coordenadas dos vetores de entrada e y_n são as coordenas 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 7 e 8.

$$D_{Minkowski} = \sqrt[p]{\sum_{k=0}^{n} \left| x_k - y_k \right|^p} \tag{7}$$

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

$$D_{Manhat\,\tan} = \sum \left| X - Y \right| \tag{8}$$

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.

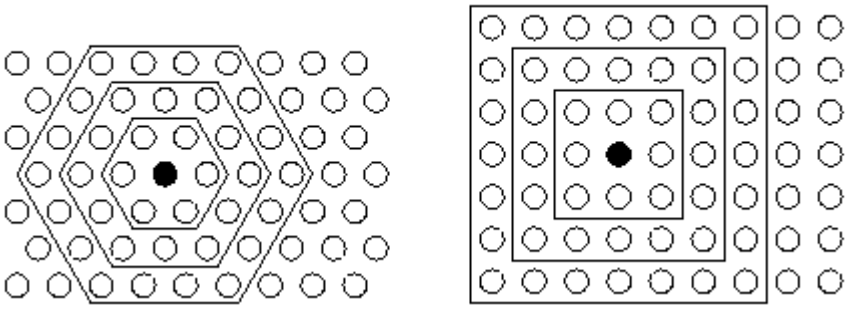

FIGURA 6 - Representações das Etapas Competitiva e Cooperativa de treinamento da SOM (VESANTO, 2009).

Na figura 6 são ilustradas a formação de vizinhança a partir do neurônio vencedor em topologia hexagonal e retangular. Algumas opções de topologia podem ser vistas nas figura 7 (triangular), figura 8 (quadrática) e figura 9 (randômica).

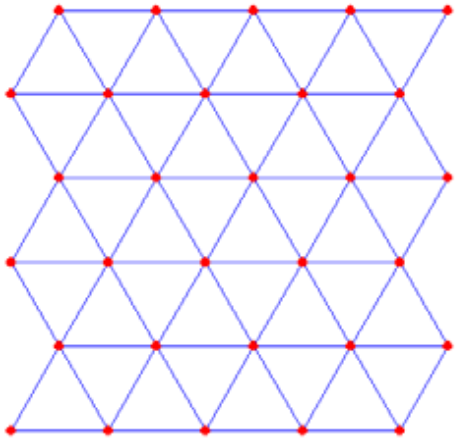

FIGURA 7 – Estrutura de SOM com topologia triangular (LCIS, 2011).

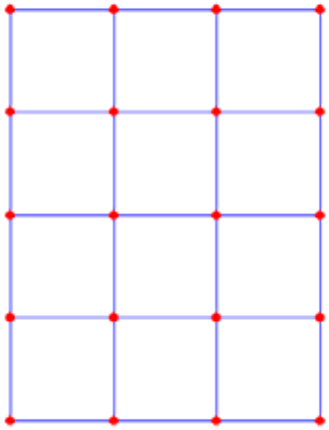

FIGURA 8 – Estrutura de SOM com topologia quadrática (LCIS, 2011).

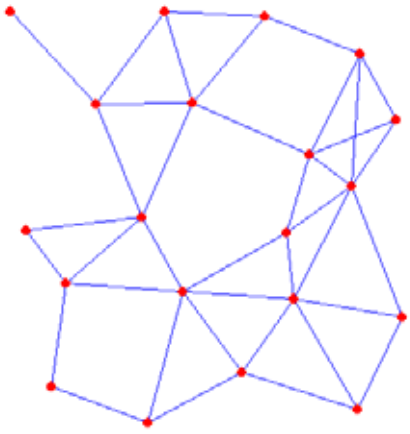

FIGURA 9 – Estrutura de SOM com topologia randômica (LCIS, 2011).

4. METODOLOGIA

No presente trabalho são aplicados os mapas auto-organizáveis na análise de dados físico-químicos de águas que abastecem estações de tratamento de água da Sabesp relativas a algumas unidades de gerenciamento hídrico do Estado de São Paulo.

Uma avaliação inicial das características desse conjunto de dados mostrou interessante potencialidade na aplicação da ferramenta para a busca de padrões de comportamento e correlações. Como estratégia, o banco de dados foi analisado visualmente em busca de eventuais falhas ou "defeitos" na seqüência dos dados que pudessem dificultar a aplicação da ferramenta. Essa avaliação é discutida mais adiante.

4.1 Características dos parâmetros físico-químicos

A despeito de existirem parâmetros definidos em legislação (CONAMA,MS, 2005) para a qualidade de água distribuída para consumo humano, essa qualidade por si só e independente de qualquer referência legal, deve garantir a propriedade para o consumo. Parâmetros como concentração de metais, de substâncias orgânicas, características organolépticas (sabor, odor e cor), acidez/basicidade, presença de coliformes termotolerantes, constituem a maior parte do conjunto das referências legais a serem atendidas. Com a finalidade de tornar mais prático e ágil a avaliação da qualidade da água, alguns organismos de regulação e/ou controle adotam índices de qualidade, considerando-se somente uma fração desses parâmetros. A Companhia Ambiental do Estado de São Paulo, CETESB, (CETESB, 2008), por exemplo, adota os seguintes parâmetros para compor o índice de qualidade de água bruta para fins de abastecimento público, IAP: temperatura da água, pH, oxigênio dissolvido, demanda bioquímica de oxigênio, coliformes termotolerantes, nitrogênio total, fósforo total, resíduo total, turbidez, teste de Ames – genotoxicidade, potencial de formação de trihalometanos – PFTHM, número de células de cianobactérias, cádmio, chumbo, cromo total, mercúrio, níquel, ferro, manganês, alumínio, cobre e zinco. Esses parâmetros apresentam itens comuns com os adotados pela *National Sanitation Foundation* (NSF, 2008) e pela Comunidade Européia (Comunidade Européia, 1998).

Embora a relação disponibilidade/demanda seja positiva em muitas regiões do Estado de São Paulo, não há como distribuir água para consumo sem algum tipo de tratamento. As estações de tratamento de água, ETA, realizam esse tratamento configurando as condições das diversas etapas do processo em função entre outras, das características da água captada (água bruta) e da projeção da qualidade da água final que deverá atender á legislação. Em função da forte correlação entre as características (rio, poço subterrâneo, geomorfologia, clima, entre outras) dos locais de captação e as características físico–químicas da água (COTRIM, 2006), cada ETA, em maior ou menor grau, apresenta uma identidade própria que define, então, o melhor processo para tratamento da água. A aquisição dos dados inicialmente é realizada por meio de relatórios públicos disponibilizados pela CETESB, (CETESB, 2001) os resultados são dispostos em formato de planilhas e expressos em valores ou em estimativas, conforme tabela no anexo 1.

4.1.2 Organização do Banco de Dados

Conceitualmente matrizes ambientais são coleções de resultados multidimensionais dinâmicos de perfis comportamentais de constituintes de interesse legal. (CARVALHO, 2003).

Em análises temporais, são comumente denotadas as dificuldades quanto à variabilidade dos processos, e mencionadas como um limitante e até gerador de incertezas para as interpretações.

No intuito de se suprir estas dificuldades atualmente é empregada uma gama variada de ferramentas matemáticas e estatísticas propostas para manipulações de dados com a preservação das informações relevantes. No entanto há restrições quanto á interpretação dos resultados devido à natureza dos dados originais e incompatibilidades das ferramentas.

Os mapas auto – organizáveis consistem em técnica RNA de reconhecimento de correlações e inter-relações em bancos multidimensionais e mesmo sendo uma técnica de RNA também possui a restrição aos dados originais. No presente trabalho, há a imposição de uma limitação lógica restringindo sobremaneira a manipulação e transformação dos dados de interesse. Originalmente a obtenção dos dados foi realizada para o atendimento a um requisito legal (CONAMA 357 e atualizações e Decreto Estadual 8468/76). São

resultados de análises químicas que, por conseguinte, são limitados tecnológica, financeira e operacionalmente. No processo de exclusão dos valores do banco de dados original proposto para o presente trabalho são apresentados os critérios adotados, admitindo-se que por razões do comprometimento final da análise, haja condições que garantam a qualidade da integridade:

Descontinuidade da análise ou alteração temporal do parâmetro, gerando truncamento na seqüência;

Grande quantidade de valores faltantes, originando em vazios de dados;

Grande quantidade de valores inferiores aos limites impostos pelas técnicas analíticas.

Resumidamente o modelo de controle da qualidade da água adotado para o abastecimento público, pode ser compreendido como a adaptação de processos para o atendimento de critérios de interesse legal visando o estabelecimento de novos critérios, a análise crítica e desenvolvimento de novas metodologias. (COTRIM, 2006)

A CETESB realiza a monitoração do índice de qualidade de água atualmente em 22 Unidades de Gerenciamento de Recursos Hídricos (UGRH), em aproximadamente 136 pontos de coleta subdivididos conforme tabela 1.

TABELA 1 - Pontos de coleta com sua respectiva numeração.

UGRHI 01	MANTIQUEIRA
UGRHI 02	PARAÍBA DO SUL
UGRHI 03	LITORAL NORTE
UGRHI 04	PARDO
UGRHI 05	PIRACICABA, CAPIVARI E JUNDIAÍ
	-Bacia do Rio Capivari. -Bacia do Rio Jundiaí. -Bacia do Rio Piracicaba.
UGRHI 06	ALTO TIETÊ
	-Bacia do Rio Tietê Alto – Cabeceiras. -Bacia do Reservatório Billings. -Bacia do Reservatório Guarapiranga. -Bacia do Rio Cotia. -Bacia do Rio Tietê Alto - Zona Metropolitana.
UGRHI 07	BAIXADA SANTISTA
UGRHI 08	SAPUCAÍ/GRANDE
UGRHI 09	MOGI-GUAÇU
UGRHI 010	SOROCABA/MÉDIO TIETÊ
	Bacia do Rio Tietê Médio-Superior Bacia do Rio Sorocaba
UGRHI 011	RIBEIRA DE IGUAPE/LITORAL SUL
UGRHI 012	BAIXO PARDO/GRANDE
UGRHI 013	TIETÊ/JACARÉ
UGRHI 014	ALTO PARANAPANEMA
UGRHI 015	TURVO/GRANDE
UGRHI 016	TIETÊ/BATALHA
UGRHI 017	MÉDIO PARANAPANEMA
UGRHI 018	SÃO JOSÉ DOS DOURADOS
UGRHI 019	BAIXO TIETÊ
UGRHI 020	AGUAPEÍ
UGRHI 021	PEIXE
UGRHI 022	PONTAL DO PARANAPANEMA

Fonte: Relatório CETESB 2001.

4.1.3 Características das Unidades de gerenciamento de Recursos hídricos

Para o presente trabalho foram utilizados dados públicos de cinco unidades de gerenciamento de recursos hídricos das regiões de Mantiqueira, Paraíba do Sul, Pardo, Capivari e de Biritiba Mirim, no estado de São Paulo. Estes pontos foram selecionados inicialmente sem a utilização de critérios pré - estabelecidos.

A unidade de gerenciamento de recursos hídricos 01 (UGRH), da região de Mantiqueira está localizada a leste do estado de São Paulo, é composta por 3 municípios e tem como principais atividades: agricultura e industrial com proximidade a UGRH 02 da região de Paraíba do Sul.

Na figura 10 são mostradas as localizações da UGRH 01, e UGRH 02 de acordo com as delimitações da CETESB.

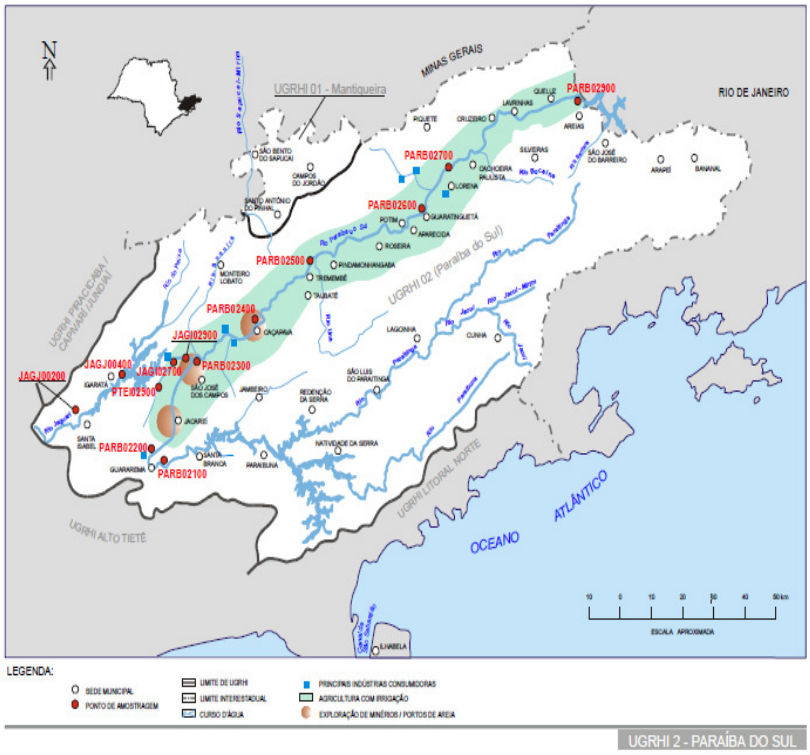

FIGURA 10 – Mapa da Localização geográfica da UGRH 01, região de Mantiqueira e UGRH 02, região de Paraíba do Sul (CETESB, 2001).

A UGRH 04 está situada ao Norte do estado de São Paulo, é composta por 23 municípios, e tem como principais atividades o desenvolvimento agrícola e de segmentos industriais, apresentada na figura 11.

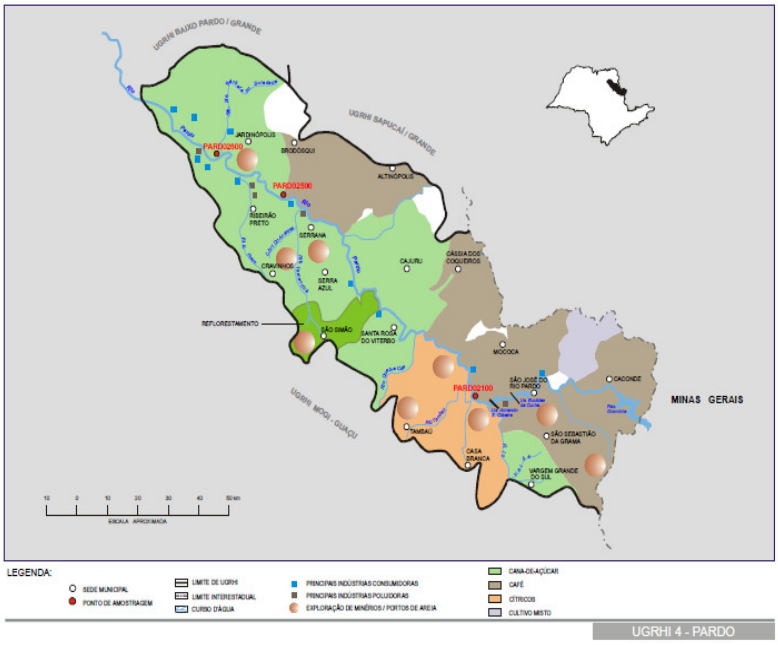

FIGURA 11 – Mapa da localização geográfica da UGRH 4 Rio Pardo (CETESB, 2001).

A UGRH 05 está localizada na região metropolitana do estado São Paulo é composta por 57 municípios e há nessa região intensa atividade industrial, conforme os relatórios da CETESB.

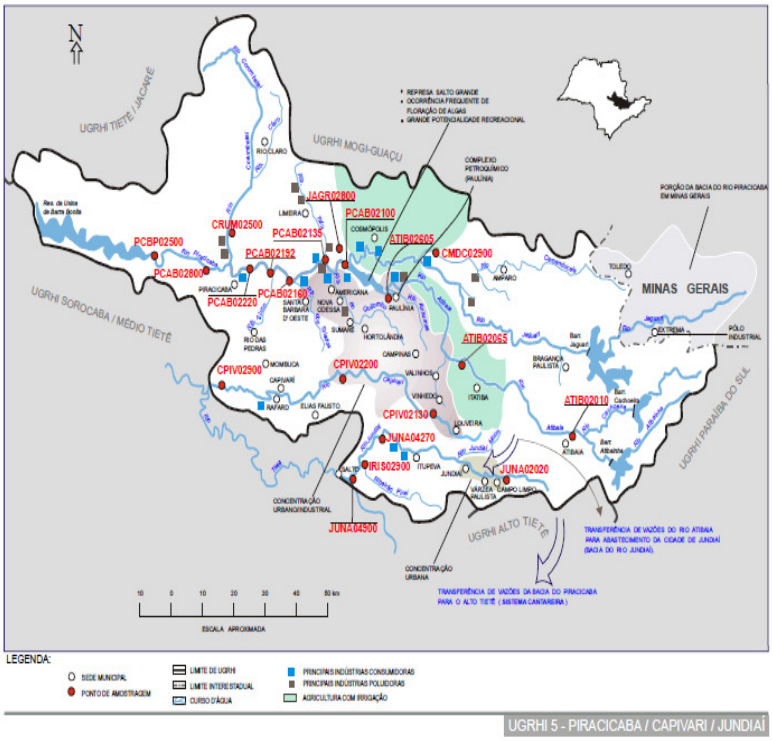

FIGURA 12 – Mapa da localização geográfica da UGRH 05, região de Piracicaba, Capivari e Jundiaí (CETESB,2001).

E a UGRH 06 de acordo com a subdivisão adotada pela CETESB das bacias hídricas, para avaliação da qualidade da água, a distribuição da UGRH está localizada na região metropolitana, composta por 34 municípios e intensa atividade industrial, como atividade primária desenvolvida, como observado na figura 13.

FIGURA 13 – Mapa da localização geográfica da UGRH 06, região do Alto Tietê (CETESB, 2001).

Com a adoção de uma coleta de amostra bimensal em um período de 9 anos ao todo (2000 a 2008), portanto com um tamanho amostral 54 dias. Na análise dos resultados das coletas há informações de 43 indicadores (parâmetros físicos, químicos, hidrobiológicos, microbiológicos e ecotoxicológicos) utilizados segundo relatórios (CETESB, 2000).

A restrição do número de parâmetros avaliados no presente trabalho está na capacidade de correlação e de interpretação final dos resultados pelo analista e é imposta como medida de controle no desempenho da ferramenta.

Na tabela 2 está uma descrição sumaria dos aspetos de relevância para o presente trabalho dos parâmetros, dispostos em grupos conforme a terminologia adotada pela CETESB.

TABELA 2 – Descrição de parâmetros.

I. Parâmetros físicos (campo)	Descrição
Potencial hidrogeniônico (pH) Temperatura do ar Temperatura da água Turbidez Condutividade	Grupo de parâmetros de importância para sustentabilidade do meio aquático e também utilizados em correlações para identificação de possíveis fontes de contaminações antropogenicas (fontes de poluentes de origem humana). * A condutividade em específico é uma expressão numérica da capacidade de condução de corrente elétrica na água, e pode sofrer influências das concentrações iônicas e da temperatura.

II. Parâmetros químicos	Descrição
OD DBO DQO	Para controle de processos em sistemas aquáticos naturais e de uso em estações de tratamento para o estabelecimento de condições mínimas para a manutenção do meio aquático.

III. Compostos inorgânicos	Descrição
Cloreto Fenóis Fósforo Manganês NKT Nitrogênio Kjeldahl total	Indicadores de toxidez possuem relação com outros parâmetros tais como: pH, temperatura da água e turbidez são de interesse legal (tanto para avaliação como o cumprimento da legislação vigente), e em alguns casos possuem associação com produtos oriundos de atividades humanas (contaminante antropogenico). * Obs: o nitrogênio Kjeldahl total é um parâmetro obtido por meio de cáculo, pela soma das formas de nitrogênio orgânico e amoniacal, oriundos de atividades biológicas naturais, e utilizado na avaliação do nitrogênio disponível para as atividades biológicas

Fonte - COTRIM, 2006

A coleta em diferentes períodos é realizada para a inclusão de dados com diferenças sazonais, (períodos de chuva e seca) com o objetivo da incorporação de um modelo real o qual demonstra as alterações dos dados por influencias externa.

A denominação períodos de seca e de chuva são classificações adotadas conforme o índice de precipitação pluviométrica (parâmetro não utilizado) e das variações de temperaturas e são correlacionadas com as estações do ano.

A variação dos resultados amostrais dos parâmetros físico-químicos referentes às UGRHs correspondentes é apresentada no anexo B conforme os resultados das análises físico-químicas mostradas em tabela no anexo A.

4.1.4 Implementação da Metodologia SOM.

O processo de implementação do sistema de análise dos dados ambientais utilizando o SOM é realizado seguindo etapas de formatação e inserção dos dados nas bases de treinamento, conforme o diagrama apresentado na figura 14.

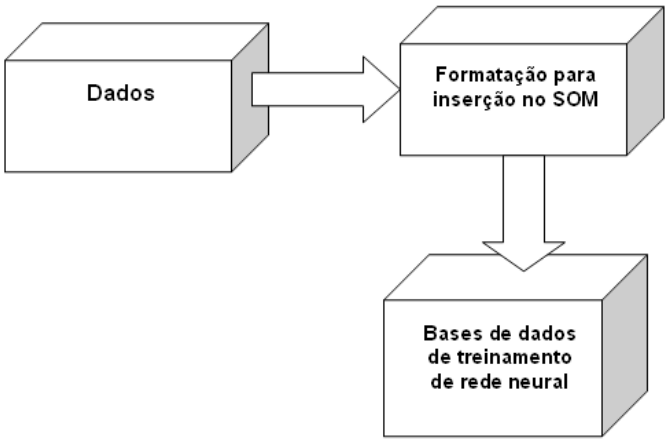

FIGURA 14 - Diagrama da formatação da base de dados.

Inicialmente a escolha das variáveis que seriam utilizadas no estudo seguiu o critério do uso do maior número de variáveis aptas a serem inseridas numericamente na rede SOM, de forma a possibilitar a investigação do mais amplo espectro de possíveis correlações. A

formatação dos dados é uma etapa fundamental e necessária para a correta utilização da ferramenta de análise a ser utilizada: MATLAB (MATHWORKS, 2004),

Na organização dos dados em formato de planilhas eletrônicas consideram-se o ponto de coleta, data da análise, e os parâmetros físico-químicos que serão descritos adiante. Depois de formatada a planilha é transportada para o espaço de trabalho do MATLAB por meio de um "plug-in" (programa adicional ao Matlab) chamado de Excel Link que possibilita a transformação da planilha em uma matriz multidimensional que pode então ser manipulada e trabalhada pelas funções do Matlab e pelo "toolbox" SomToolbox (SOM TOOLBOX, 2005). A partir deste conjunto de funções é possível escolher uma seqüência de comandos que efetuarão as etapas necessárias ao treinamento da rede neural Som que gerará como saída o chamado Mapa de Protótipos ou Mapa de Kohonen. Este aplicativo adicional (Excel Link) possibilita uma conexão ágil e interativa entre os dois programas (Excel e Matlab) de forma que se possa atualizar e fazer a análise pretendida com matrizes multidimensionais.

O SOM Toolbox possui uma interface visual que possibilita a escolha dos parâmetros de treinamento, incluindo o erro almejado. A rede neural é então treinada e após a verificação dos parâmetros de qualidade do treinamento, é possível a visualização dos resultados iniciais que podem ser avaliados de acordo com gráficos gerados pelo próprio aplicativo, onde se pode avaliar com grande agilidade o grau de inter–relação entre as variáveis utilizadas.

A figura 15 apresenta um diagrama de blocos das etapas do procedimento executadas no programa MATLAB com o recurso do SOM Toolbox.

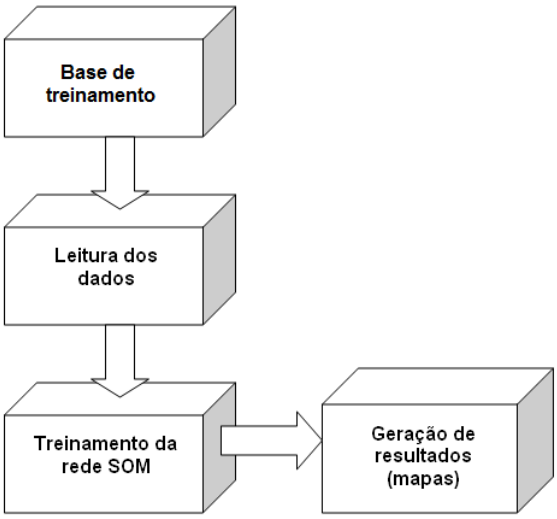

FIGURA 15 - Procedimento realizado do transporte das variáveis para a geração de resultados no SOM Toolbox.

4.1.5 Descrição do procedimento utilizado para o treinamento do SOM.

O procedimento do treinamento do SOM é inicialmente realizado com o transporte e transformação da planilha eletrônica previamente organizada em uma matriz bidimensional no espaço de trabalho do programa Matlab pelo Excel Link, como comentado na inserção dos dados. A partir de uma variável gerada pertencente a uma classe do Matlab denominada "data struct" que tem como função o armazenamento da estrutura de informações presentes na planilha original. A variável pode armazenar outras variáveis como, por exemplo: "double", "inteiro", "string" e outras, concatenadas por campos que definem diferentes tipos de informação de interesse. O "toolbox" Somtoolbox tem funções de treinamento que atuam diretamente sobre este tipo de variável. Por meio do comando:

>>Smatrix = som_data_struct(matrix);

A matriz *matrix* importada pelo Excel Link para o espaço de trabalho do Matlab é transformada em (*Smatrix*), variável da classe estrutura e abriga em si campos de informação apropriados (mesmo que inicialmente vazios), para a manipulação pela função de treinamento da rede presente no Somtoolbox. Na etapa de normalização da matriz de dados, etapa necessária à otimização do treinamento da rede, possibilitando com que o algoritmo de treinamento convirja mais rapidamente e ao mesmo tempo, que a saída possa

ser apresentada visualmente se comparando a variabilidade proporcional da variável, e não seus valores absolutos. Assim, o comportamento da variância dos parâmetros que compõe a base de dados pretende ser graficamente estabelecido, possibilitando uma rápida observação de correlações importantes. O algoritmo de treinamento da rede neural minimiza a distância entre os vetores-protótipo existentes em cada célula de rede SOM. Experimentalmente definido, o processo de normalização de dados para a presente matriz é denominado *"logistic"* e adotado pelos critérios de menor erro de quantificação e de menor erro topográfico. Esta normalização é obtida com o seguinte comando:

>>Smatrix = som_normalize(Smatrix,'logistic');

Onde a variável Smatrix é normalizada escalando todos os vetores para valores no intervalo entre zero e um de acordo com a função logistic definida por duas operações algorítmicas (iterativas) definidas nas equações 9 e 10.

$$x_{escalado} = (x_{antigo} - média(x_{antigo})) \Big/ \delta\,(x_{antigo}) \tag{9}$$

$$x_{novo} = 1 \Big/ (1 + \exp(-x_{escalado}) \tag{10}$$

Após a etapa de normalização, inicia-se a identificação dos dados da matriz conforme os parâmetros fisico – químico e de abreviaturas que indiquem as regiões e as informações de interesse como: região e data de coleta, realizada pela modificação do campo "label" e "comp_names" presentes na variável estrutural Smatrix, por meio de comandos do tipo:

>>Smatrix.comp_names{1,1}='pH';

>>Smatrix.labels{1,1}='MANT1A';

Na compreensão das abreviaturas utilizadas na rotulagem dos vetores protótipos tomando como exemplo, o elemento da matriz localizado na primeira linha e primeira coluna com o rótulo "MANT", referente ao ponto de coleta de Mantiqueira, "1", referente ao ano de 2001 e a letra "A" referente ao primeiro mês de coleta. Sistema de nomenclatura o qual

adotado para a rápida visualização dos agrupamentos por pontos de coleta, anos e meses, que são os vetores de entrada com 13 ordenadas identificados pelos parâmetros físico – químicos.

Nas opções disponíveis de métodos de inicialização dos protótipos no Somtoolbox, há a inicialização *"randômica"* (aleatória) ou a inicialização *"linear"*. A opção entre os métodos pode ser usada para definição de parâmetros tais como de qualidade do treinamento como a velocidade. No treinamento da rede há duas opções de algoritmos diferentes disponíveis: o treinamento seqüencial ou o treinamento em batelada. Sendo que a opção de treinamento de rede seqüencial é recomendada para grandes matrizes ou que ofereçam alguma dificuldade de processamento pelo custo computacional necessário. No presente projeto a escolha da inicialização foi randômica e com o treinamento dos dados em batelada, devido às características da matriz de dados, que apesar de ampla não exigia do algoritmo um tempo de treinamento muito prolongado que justificasse um treinamento seqüencial.

O treinamento tradicional de um SOM passa por duas etapas, uma primeira mais grosseira denominada originalmente de *"rough"* onde um número inicial grande de raio de vizinhança (neurônios vizinhos ao neurônio vencedor – "best match unit" (BMU)) é utilizado, modificando de uma só vez uma quantidade proporcionalmente alta dos neurônios que compõe a rede. Após esta primeira etapa, segue-se a fase mais refinada (*"finetuning")* que utiliza um raio menor de vizinhança, modificando menos neurônios por iteração. O treinamento da rede é um processo contínuo de comparação entre os vetores-protótipos de cada neurônio e os vetores-amostra que compõe a base de dados. Esta comparação utiliza diferentes definições de distância entre os vetores, e a mais utilizada ("*default*") é a que utiliza a distância euclidiana. Assim em iterações sucessivas se encontra o BMU e se modifica esta unidade e seus vizinhos de forma proporcional à distância "medida" entre a a amostra e o protótipo. O comando básico utilizado para iniciar o treinamento é:

>>Smatrixmap = som_make(Smatrix);

Utilizado com mais opções para alterar os modos e parâmetros de treinamento na verificação da qualidade do erro final de quantificação e do erro topográfico, parâmetros de

controle de qualidade da rede entre os parâmetros existentes como definição do tamanho da rede, especificação do número de épocas de treinamento e do treinamento das variáveis em batelada ou randômica, etc.

No relatório do treinamento, são reportados os seguintes parâmetros: tamanho do mapa utilizado (dimensões da matriz de protótipos otimizada pelo algoritmo de acordo com o tamanho da matriz de amostras e da variabilidade apresentada por ela), o número de épocas de treinamento e o tempo de treinamento para cada fase de treinamento, "*rough*" e "*finetune*", o erro final de quantificação e o erro final topográfico.

Ao final da etapa de treinamento da rede, na visualização do mapa são apresentados os mapas resultantes por meio do comando:

>> *som_show (Smatrixmap);*

Com o qual são gerados 13 mapas dos chamados componentes planos e um mapa, auxiliar da matriz de distância dos protótipos presentes em cada neurônio. Este comando tem muitas possibilidades de utilização, ativando diferentes formas de apresentação dos mapas já treinados e presentes na variável Smatrixmap. Pode-se visualizar desde as matrizes de distância vetorial representadas por "*umat*", como também a seleção de variáveis que se deseje observar.

Entre as opções de visualização e apresentação de resultados além do comando descrito acima pode – se também, utilizar o comando para visualização dos rótulos por distribuição de freqüência:

>>*som_show_add('label",Smatrixmap_freq):*

Obtém-se a distribuição por freqüência, uma importante ferramenta para a rápida visualização do número de vetores presentes no mesmo grupo com seus referentes rótulos e qual a freqüência por repetição no mapa, conforme exemplo na figura 34-a nos resultados de mapa por distribuição por freqüência.

A matriz de distâncias proporciona uma visualização bidimensional das distâncias médias entre cada variável e um protótipo de vetor correspondente a cada neurônio.

A figura 16 mostra a ordem na qual os protótipos são arranjados no mapa SOM. O tamanho do mapa é otimizado pelo algoritmo e, no exemplo, as dimensões do mapa gerado são 12 linhas por 7 colunas. A matriz de dados utilizada como entrada da rede tem 257 linhas por 13 colunas (parâmetros).

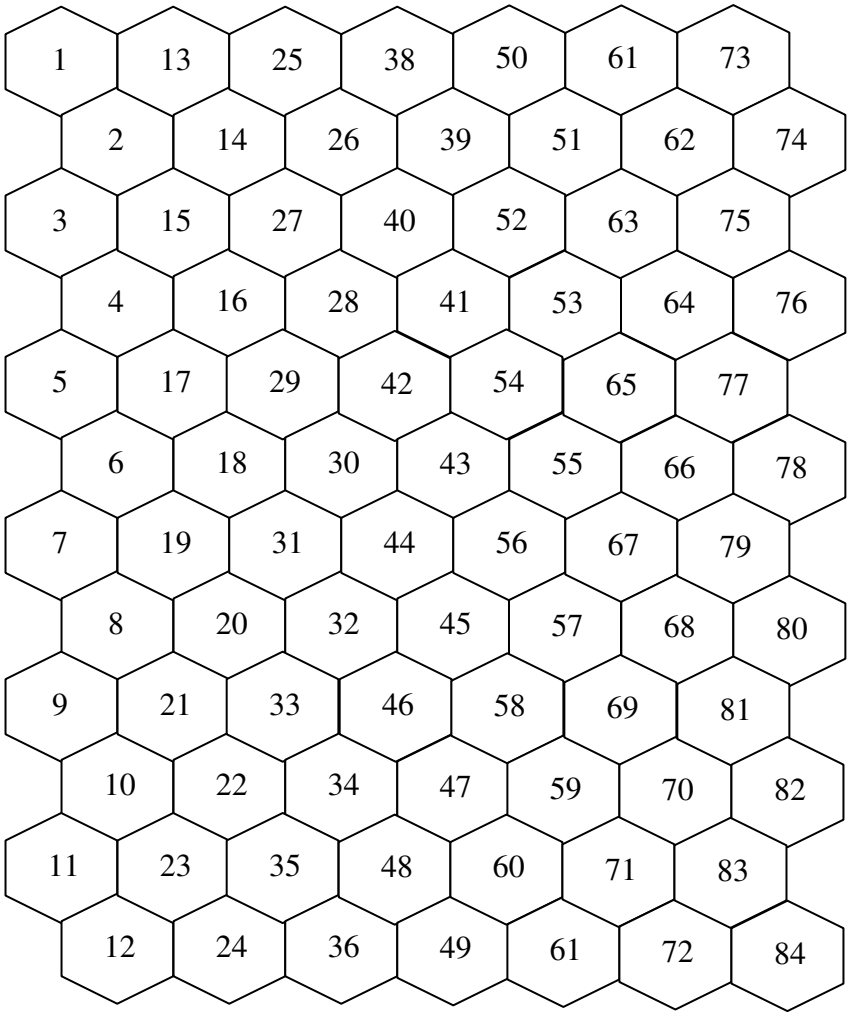

FIGURA 16 – Exemplo do ordenamento dos protótipos de vetores.

5.0 Resultados e Discussão

5.1 Apresentação dos resultados SOM

Na análise dos mapas é proposto um estudo comparativo do conjunto de mapas obtidos aplicando ao mapa um algoritmo de agrupamento, para o reconhecimento do número de "clusters" mais representativos presentes no mapa.

Na análise dos dados, são empregadas diferentes estruturas de dados, em duas abordagens distintas, inicialmente com uma matriz de 257 linhas agrupados em 13 parâmetros (ordenadas), com um total de 3341 elementos. E em uma abordagem posterior com uma matriz de 257 linhas em 10 parâmetros

Devido ao número de resultados para apresentação, no presente trabalho é proposta uma divisão em: Matrizes para estudo de similaridades entre pontos de coleta, Matrizes para estudo de similaridades entre parâmetros fisico-químicos, e Gráficos dos protótipos de vetores no intuito de se proporcionar nas conclusões um ordenamento.

5.2 Estudo de similaridades entre pontos de coleta.

Os protótipos de vetores de 13 ordenadas, (parâmetros), é inicializado randomicamente, apesar de não se empregar pré-tratamento, os dados são normalizados, e treinados em 200 épocas. A escala de cores para visualização dos clusters no mapa pode ser definida entre opções disponibilizadas no *toolbox* (LCIS, 2011).

O tamanho do mapa auto-organizável é definido por critérios pré-estabelecidos no algoritmo, com relação com o tamanho do banco de dados de treinamento. Alguns parâmetros para treinamento como o estudo de variação dos mesmos para otimização do erro final de quantificação e erro final de topográfico, (parâmetros utilizados para controle de qualidade da rede e dos resultados), são experimentais.

A rotulagem dos protótipos é realizada conforme a base de treinamento, no presente trabalho adotam – se as iniciais de acordo com as regiões de coleta, ano da coleta e meses. Além da matriz de dados uma matriz de rótulos e treinada.

Após o treinamento, na figura 17, é apresentado o mapa da matriz de distância entre vetores como um dos resultados principais com tabelas auxiliares de legendas 3 a 5, para as conclusões finais.

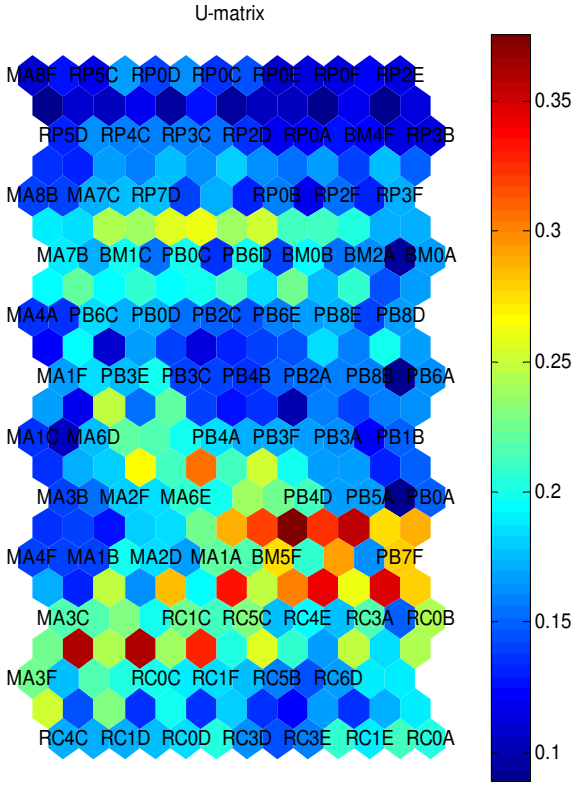

FIGURA 17 - Mapa da matriz de distância entre vetores com os rótulos.

TABELA 3 - Legenda para as regiões de coleta:

Região de coleta	Nome abreviado
Mantiqueira	MA
Biritiba Mirim	BM
Rio Capivari	RC
Rio Paraíba	PB
Rio Pardo	RP

TABELA 4 - Legenda para os meses de coleta

Meses *(período compreendido entre meses)	Letra
Janeiro - Fevereiro	A
Março - Abril	B
Maio - Junho	C
Julho - Agosto	D
Setembro - Outubro	E
Novembro - Dezembro	F

* Obs: Aos meses são atribuídos períodos conforme a data de amostragem diferenciada.

TABELA 5 - Legenda para os anos

Anos	Inicial abreviada
2000	0
2001	1
2002	2
2003	3
2004	4
2005	5
2006	6
2007	7
2008	8

Do mapa da matriz da figura 17 são gerados como saídas 84 protótipos de vetores em uma topologia hexagonal de 7 colunas por 12 linhas, com o perfil médio dos pontos de coleta (protótipo de vetor gerado a partir dos dados de entrada).

Os erros de quantificação e topológico obtidos experimentalmente são de 0.330 e 0.012 respectivamente (menores índices apontados com o uso da função de normalização "*logistic"*.)

Entre as opções de visualização de resultado há a opção da apresentação dos componentes planos juntamente com a matriz de distância vetorial, nessa opção, os protótipos de vetores podem ser visualizados individualmente conforme os parâmetros, como apresentado na figura 18.

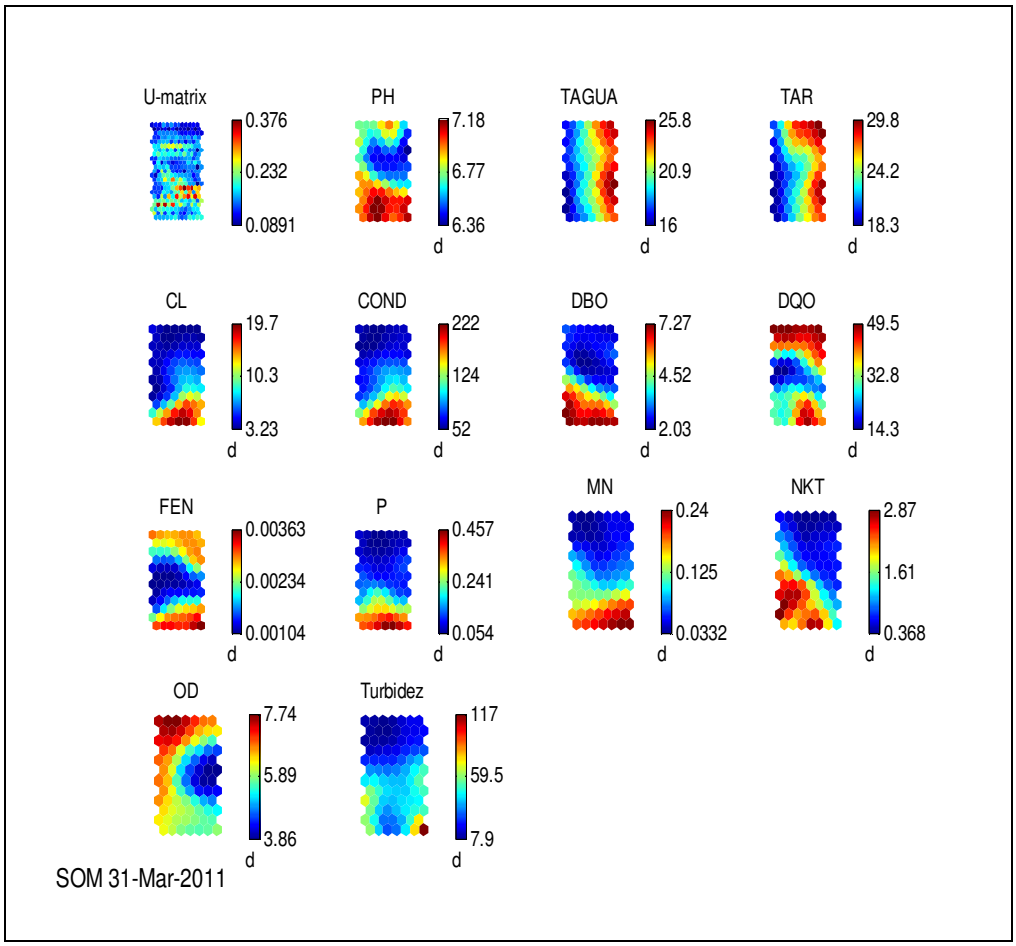

FIGURA 18 - Componentes planos gerados a partir da grande matriz.

Como pode ser observado na figura 18, no campo dos títulos os nomes dos parâmetros físico-químicos são substituídos por iniciais abreviadas de acordo com a (tabela 6 Legenda dos parâmetros físico-químicos).

TABELA 6 – Legenda de parâmetros físico – químicos

Lista de Parâmetros físico – químicos.
PH = Potencial hidrogenionico,
TAGUA = Temperatura da água,
TAR = Temperatura do ar,
CL = Cloreto,
COND = Condutividade,
DBO = Demanda bioquímica de oxigênio
DQO = Demanda química de oxigênio,
FEN = Fenóis,
P = Fósforo,
MN = Manganês,
NKT = Nitrogênio Kjeldahl Total,
OD = Oxigênio dissolvido

A escala de gradação (barra lateral) na figura 18, de cada mapa mostra a variação de cada parâmetro (não normalizado) de acordo com a base de treinamento.

Além da matriz de distância vetorial em estudo, é gerada a matriz universal (U-matriz) para complementação da análise dos agrupamentos. Neste mapa são escolhidos os rótulos relativos às melhores correspondências (BMUs) entre o vetor e o protótipo da célula (neurônio) da matriz. Assim, o mapa das BMUs é apresentado na figura 19.

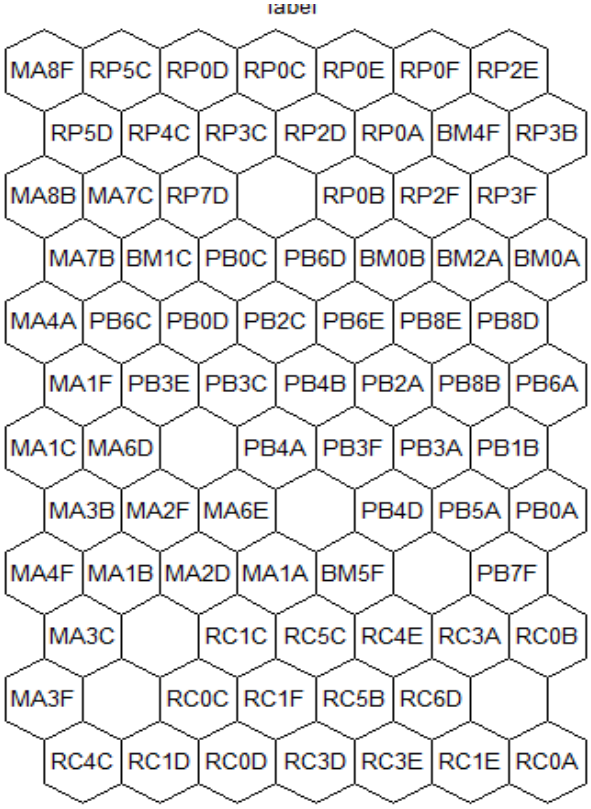

FIGURA 19 - Mapa indicativo dos rótulos característicos (BMUs) dos pontos de coleta na matriz principal.

No presente trabalho a definição do melhor número de clusters distintos para o agrupamento é obtido pela função métrica Davies-Bouldin, com a aplicação de um algoritmo de clusterização (k-means) ao mapa e utilizando o número otimizado. Pode-se visualizar clusters delimitados conforme a figura 20 do mapa, com rotulagem sobreposta para destaque dos agrupamentos. (DAVIES, 1979)

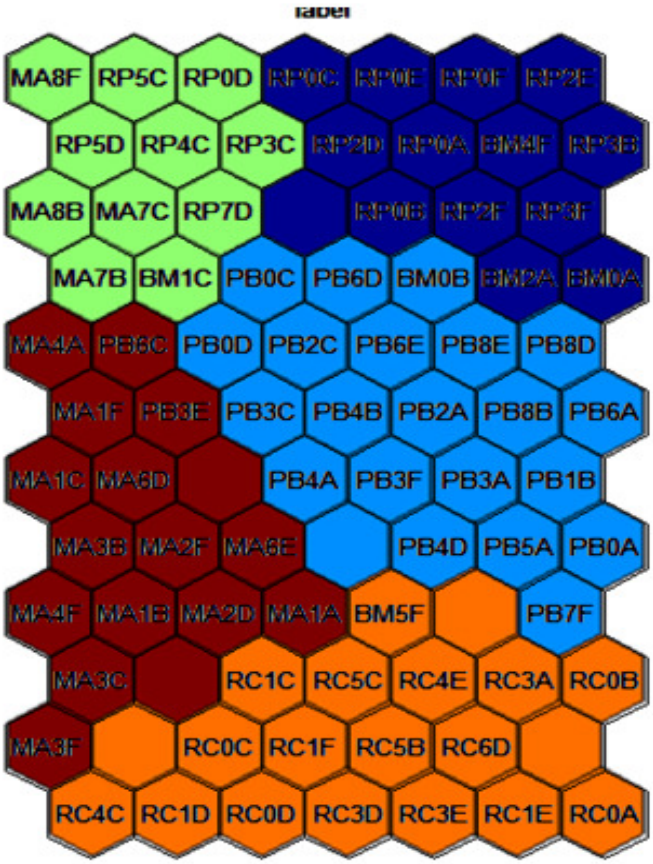

FIGURA 20 - Mapa com rotulagem sobreposta para destaque dos grupos

Na figura 20 pode-se observar os cinco grupos em destaque em diferentes cores, apenas para diferenciação dos clusters e para determinação de suas respectivas delimitações.

Na matriz alterada são excluidos três parâmetros da matriz original de dados (Fenóis, DQO e DBO) resultando em uma nova matriz com as seguintes dimensões: 257 linhas em 10 colunas (total de 2570 elementos), a nova base de dados é treinada conservando-se os mesmos parâmetros do primeiro experimento.

Nas saídas são gerados 80 protótipos de vetores (ver figura 21-c) e após treinada a rede o erro final de quantificação é de 0.254, e o erro final topográfico é de 0.012.

Os novos resultados da matriz de distância vetorial são apresentados com uso de diferentes recursos como na figura 21-a por distribuição por frequencia e por distribuição por votação na figura 21-b.

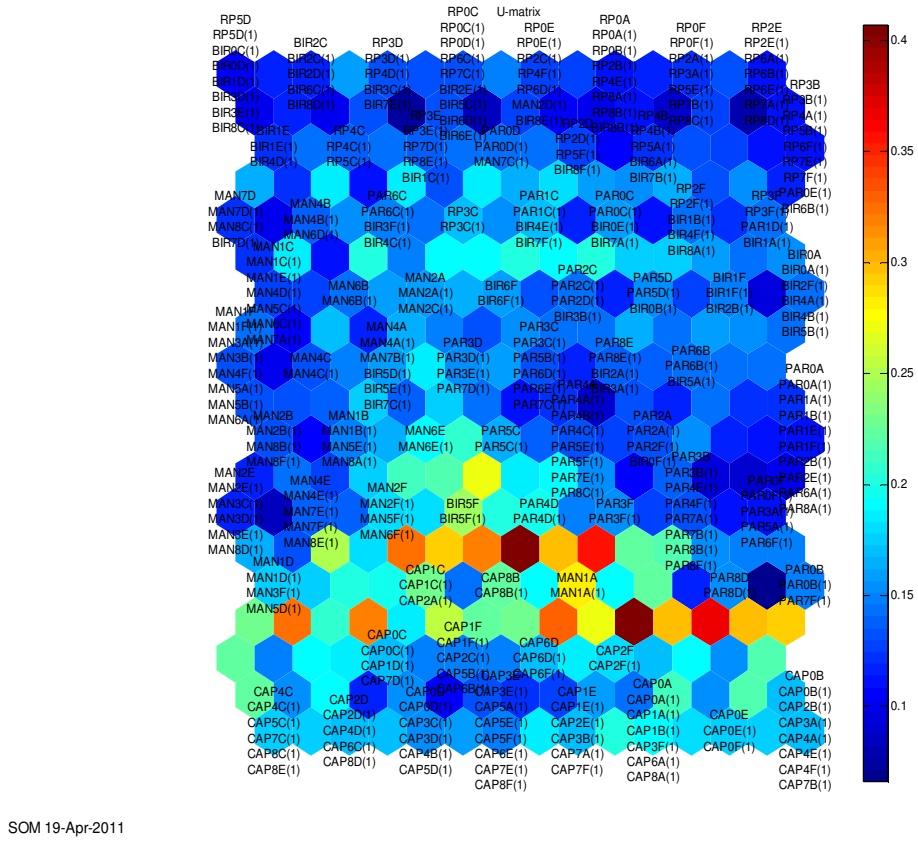

FIGURA 21-a - Apresentação do mapa das distâncias vetoriais, por distribuição de frequencia de rótulos, da matriz modificada (257 linhas por 10 parâmetros)

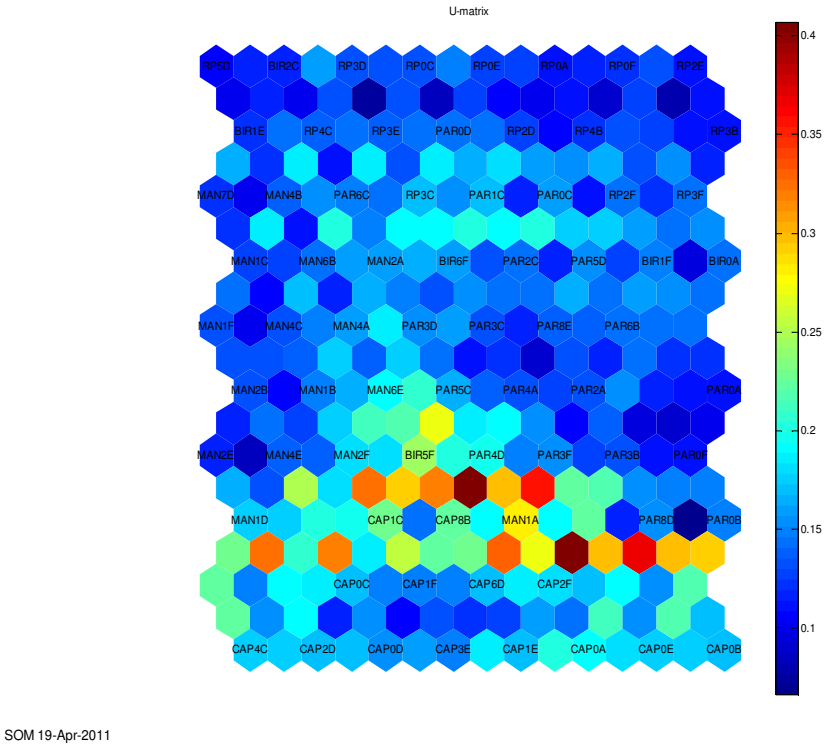

FIGURA 21-b - Apresentação por votação do mapa das distâncias vetoriais com os rótulos da matriz modificada (257 linhas por 10 parâmetros).

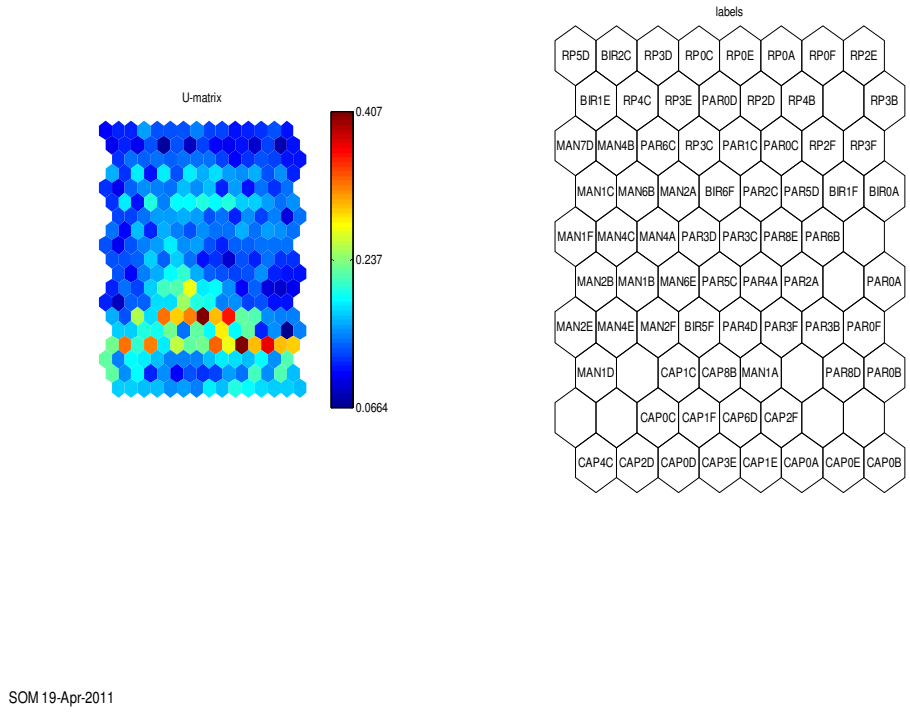

FIGURA 21-c - Apresentação do mapa das distâncias vetoriais com o mapa geral rotulado obtido da matriz modificada (257 linhas por 10 parâmetros)

5.3 Estudo de similaridades entre parâmetros fisico-químicos.

Na reorganização da disposição inicial da matriz em uma disposição transposta, a ordem de inserção é alterada, resultando em uma matriz de 13 linhas por 257 parâmetros, para análise de correlações entre os parâmetros físicos químicos,

Priorizando a análise dos parâmetros, não são gerados os mapas dos componentes planos.

O reordenamento proposto é realizado no próprio Matlab por meio do comando:

>> *Smatrixg' = Smatrixg;*

Onde a variável *"Smatrixg' "* é a transposta de *"Smatrixg"*.

Na análise da matriz inversa são estudadas as correlações dos parâmetros físico – químicos por meio do comportamento amostral, conforme figuras 22 e 23, e obtidos conservando os parâmetros iniciais de treinamento .

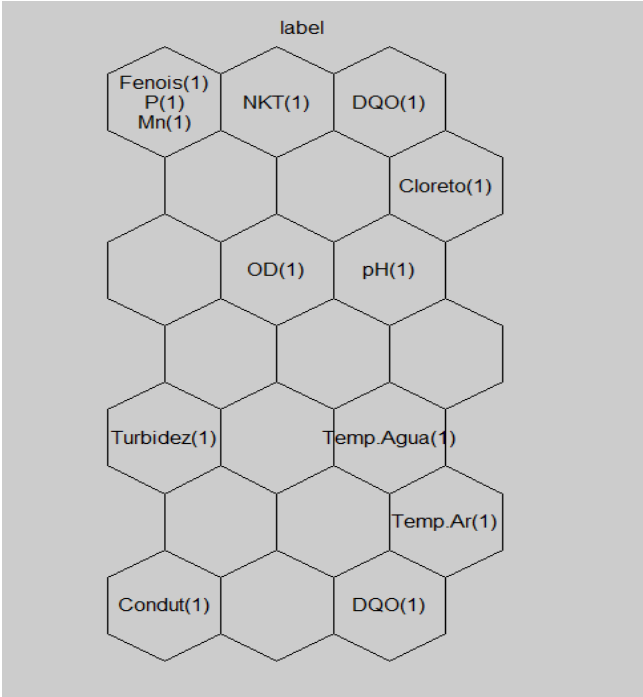

FIGURA 22 - Mapa geral rotulado obtido da matriz inversa.

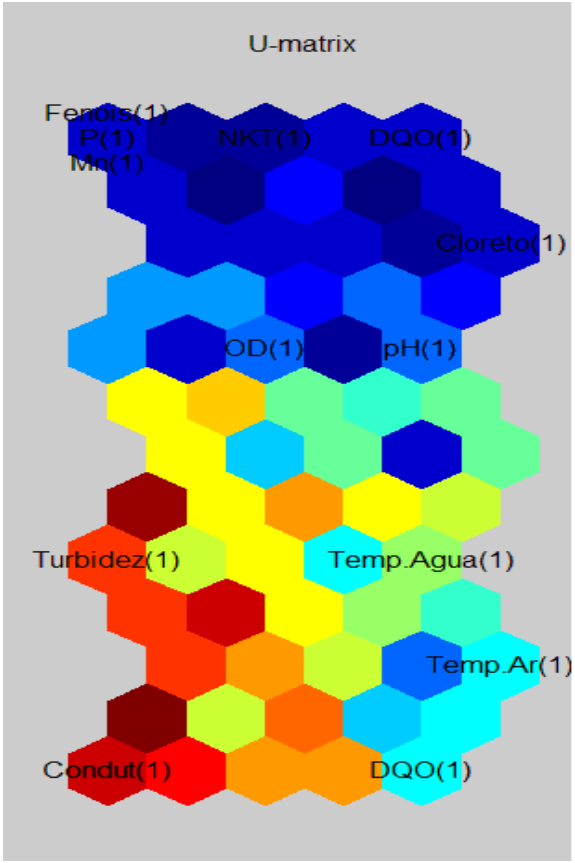

FIGURA 23 - Mapa de distância entre vetores da matriz transposta com os rótulos dos parâmetros físico – químicos.

Na descrição figuras 22 e 23 os mapas gerados têm as seguintes características: dimensões de 3 x 7 células, (figura 35) e 5 x 13(figura 36) com as condições iniciais mantidas o erro final de quantificação é de 0.530, e o erro topográfico de 0.000

Na transposição da matriz alterada, é gerada uma matriz de 10 linhas por 257 colunas, após o treinamento da base de dados, obteve-se como saída 14 protótipos de vetores em uma topologia hexagonal com um erro final de quantificação de 0.530, e um erro final topográfico de 0.000.

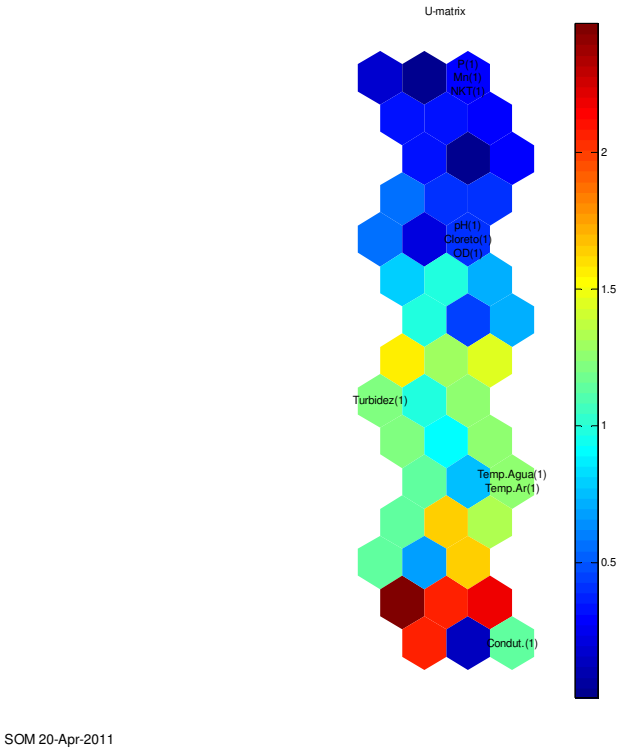

FIGURA 24 - Mapa de distância entre vetores da matriz transposta da matriz modificada de 257 linhas por 10 parâmetros.

Nos resultados do mapa da matriz modificada da figura 24, pode-se observar a identificação dos clusters por meio dos rótulos. Para análise e demonstração das similaridades por reconhecimento visual são extraidos do mapa os chamados "*codebooks*".

5.4 Gráficos dos protótipos de vetores

Os *codebooks* consistem no conjunto de prototípos de vetores gerados e treinados no algoritmo. No presente trabalho, eles são obtidos nas células rotuladas, utilizando-se como critério de escolha as informações do próprio mapa.

O estudo dos *codebooks* por meio de gráficos pode proporcionar, em uma rápida visualização, a relação de similaridade entre os dados de entrada e as saídas geradas no algorítmo para a definição do perfil médio dos parâmetro por região.

No programa Matlab a extração do *codebook* e a plotagem do gráfico é realizada por meio dos comandos:

>> figure;plot(PMnNKT);
>> pHCLOD = Sgminvmap_freq.codebook(11,:);

Os gráficos são gerados conforme o tamanho da matriz (número de amostras) e a normalização, como mostrados nas figuras 25 a 29.

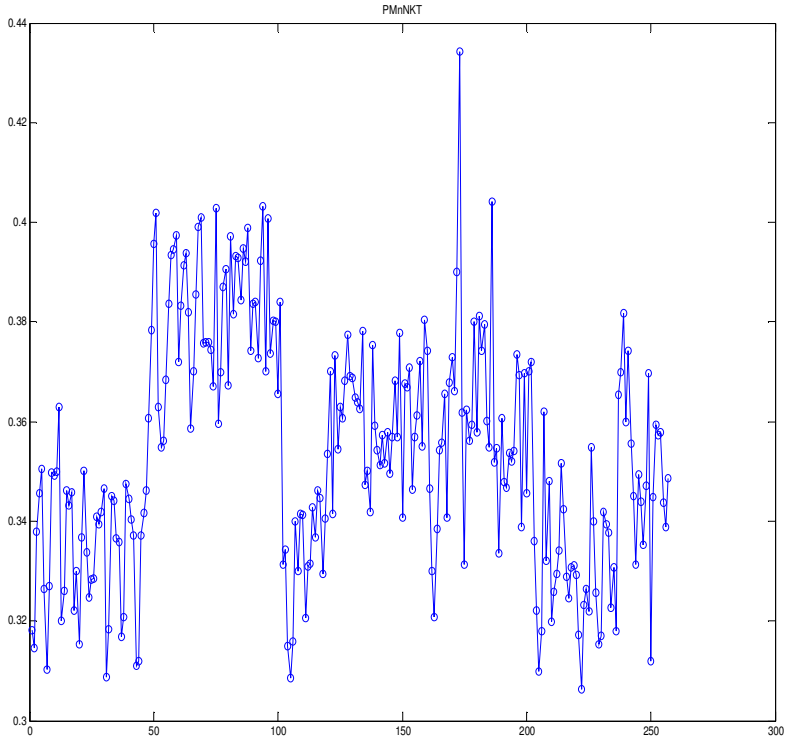

FIGURA 25 - Gráfico do protótipo de vetor mais característico do cluster nomeado "PMnNKT".

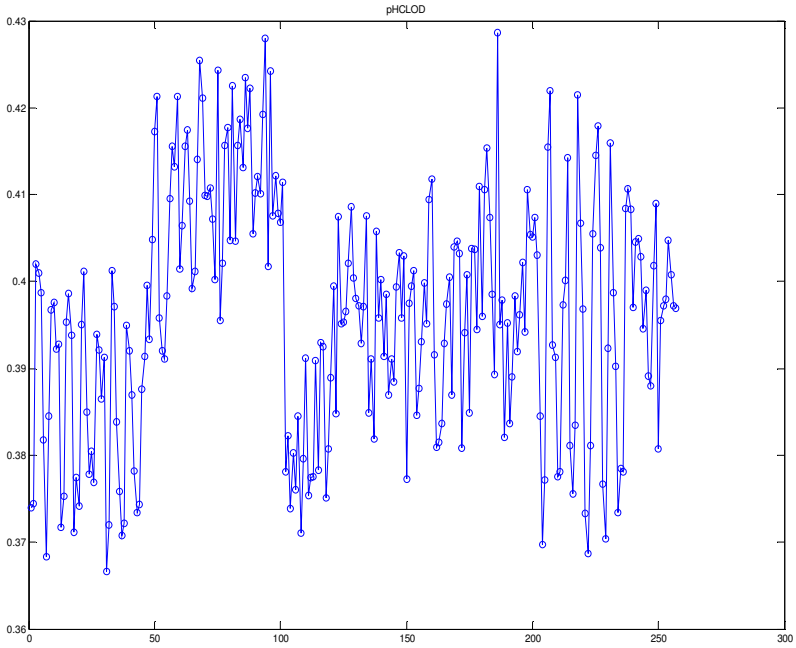

FIGURA 26 - Gráfico do protótipo de vetor mais característico do cluster nomeado “pHCLOD”.

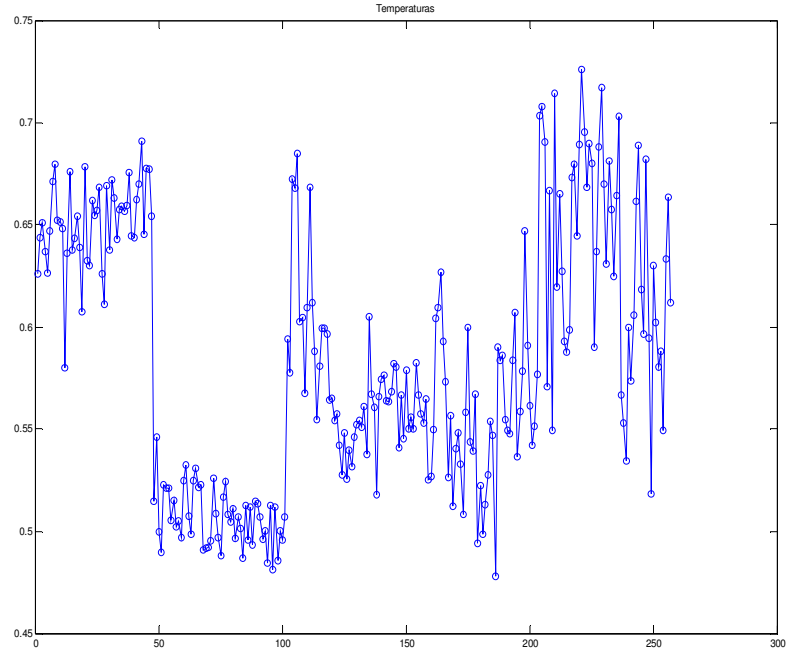

FIGURA 27 - Gráfico do protótipo de vetor mais característico do cluster nomeado “Temperaturas”.

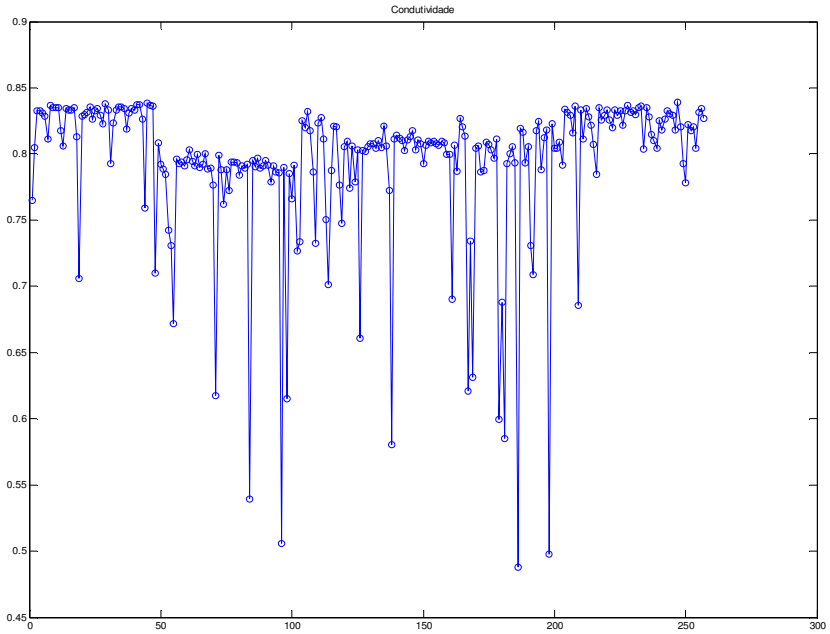

FIGURA 28 – Gráfico do protótipo de vetor mais característico do cluster nomeado “Condutividade”.

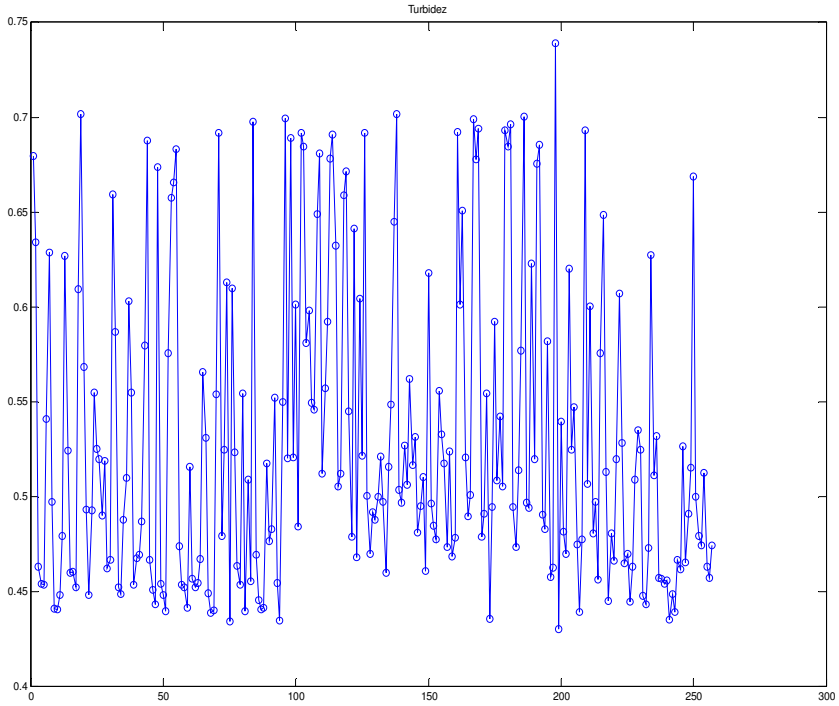

FIGURA 29 - Gráfico do protótipo de vetor mais característico do cluster nomeado “Turbidez”.

Cada gráfico de protótipo de vetor é nomeado de acordo com rótulo correspondente separadamente, como pode ser observado há uma aglutinação de até três parâmetros.

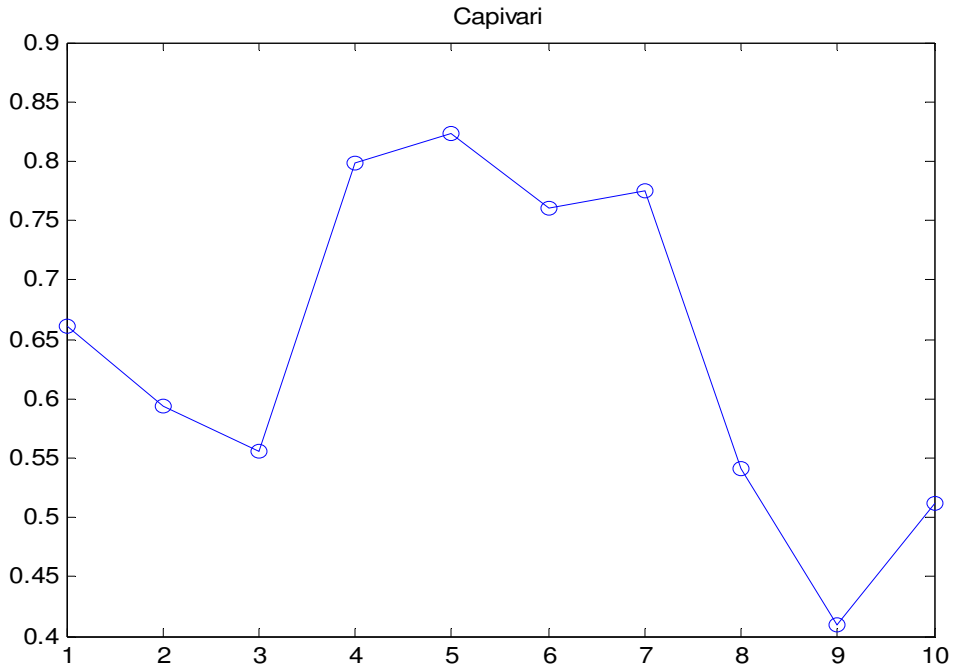

FIGURA 30 - Protótipo de vetor obtido a partir da matriz modificada referente aos dados da região do Rio Capivari.

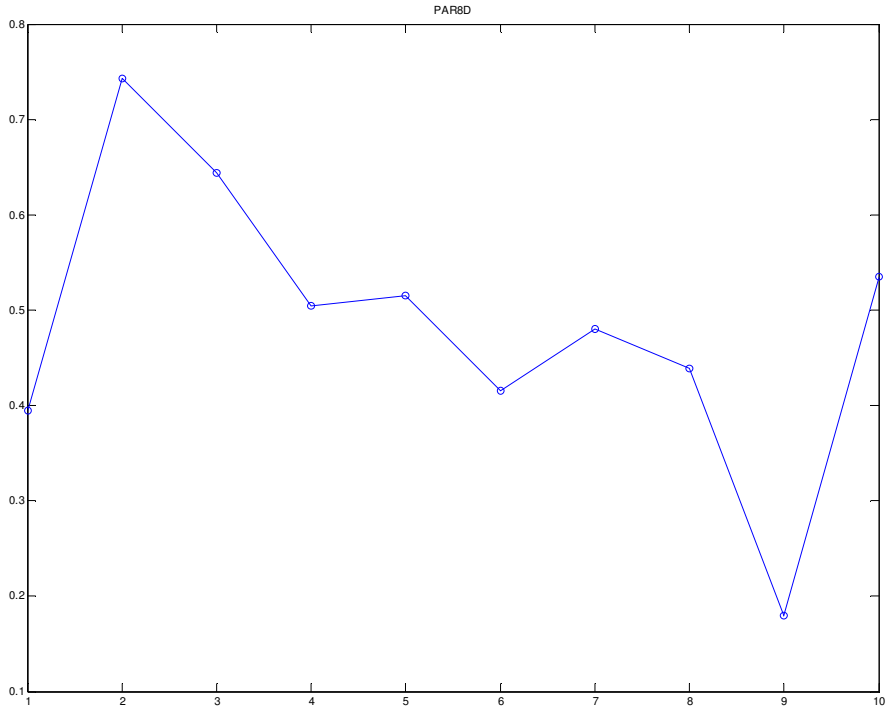

FIGURA 31 - Protótipo de vetor obtido a partir da matriz modificada referente aos dados da região do Rio Paraíba da coleta do dia 19/08/2008.

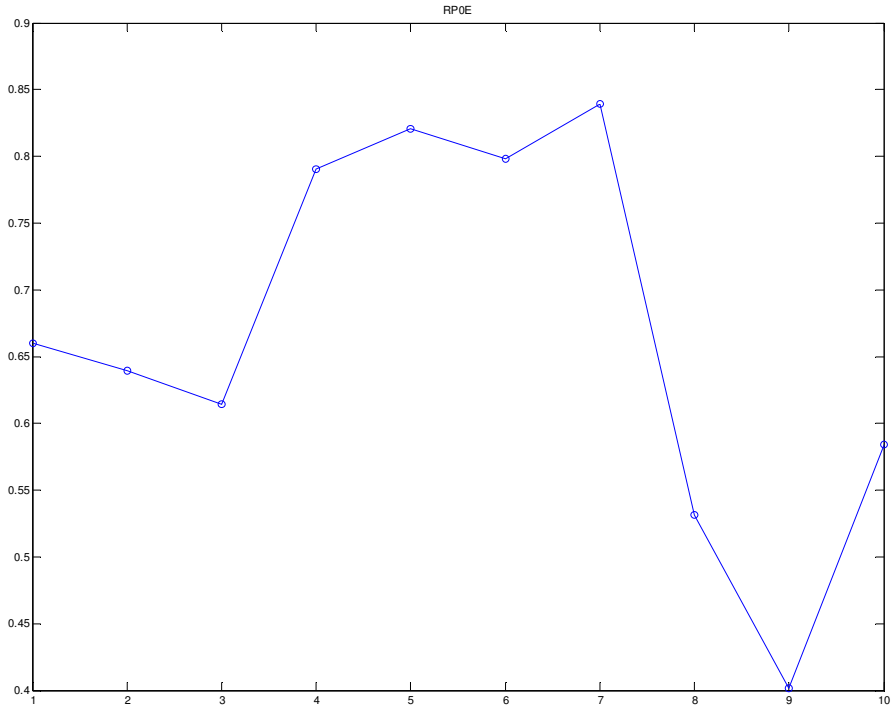

FIGURA 32 - Protótipo de vetor obtido a partir da matriz modificada referente aos dados da região do Rio Pardo da coleta do dia 03/10/2000.

6.0 CONCLUSÕES.

6.1 Matrizes para estudo de similaridades entre pontos de coleta.

Nos mapas de distância vetorial nas figuras 21-a 21-b e 21-c, pode-se observar clusters com protótipos mais próximos como observados na escala lateral (0,1 a 0,20) associados aos pontos de coleta das regiões do Rio Pardo, Rio Paraíba do sul como BMUs sugerindo a proximidade de comportamento destes protótipos.

Nos protótipos do cluster denominado Rio Paraíba (PB), há a proximidade dos protótipos associados aos períodos referentes aos últimos meses de coleta, e com distribuição proporcional correspondente aos 8 anos de coleta.

No cluster há 22 células (ver figura 20) com uma área de aproximadamente 26% da área do mapa. Na interpretação da proximidade, observa-se um padrão dos protótipos referentes aos últimos meses, que podem indicar a relação de sazonalidade nos períodos de coleta e a distribuição proporcional no período compreendido.

Também é observado no mesmo grupo a presença de dois protótipos rotulados "*BM2A"* e *"BM0A* referentes à região de Biritiba Mirim respectivamente dos anos de 2002 e 2000 do mês de Janeiro que podem indicar correlação entre pontos distintos de coleta.

O comportamento dos protótipos do cluster denominado Rio Pardo (RP), sugere associações com referência aos períodos dos últimos meses de coleta e aos anos iniciais. Com 14 células, esse grupo possui aproximadamente 16,67% da área do mapa, e também apresenta uma correlação com a região de Biritiba Mirim no mês de Março, como observado na célula *"BM0B"*.

Na interpretação da proximidade de protótipos do cluster Rio Pardo, há características de sazonalidade e de similaridade no período inicial de coleta.

No cluster de Mantiqueira (MA), há uma distribuição proporcional dos protótipos de vetores referentes ao período integral respectivo de coleta, sem constatação de

predominâncias, e com a repetição dos últimos meses de coleta (sazonalidade). Ocupa aproximadamente 20,4% da área do mapa.

No cluster de Rio Capivari (RC), também há uma distribuição proporcional dos protótipos de vetores referentes ao período integral de coleta sem predominâncias, apresentando repetitividade referente aos últimos meses de coleta (sazonalidade). Apresenta uma área de aproximadamente 23% do mapa, com correlação com a região de Biritiba Mirim do mês de dezembro do ano 2005.

Há em um cluster em específico com correlações entre os pontos de coleta referentes às regiões de: Biritiba Mirim (BM), Rio Pardo (RP), e Mantiqueira (MA). Na distribuição dos protótipos deste *cluster*, pode-se observar a proporcionalidade referente aos meses de coleta, e a predominância referente aos últimos anos de coleta. Observa-se a dessemelhança observada em comparação aos quatro outros *clusters*.

Para a compreensão da similaridade e das possíveis contribuições entre os pontos de coleta nos corpos hídricos, é realizado um estudo comparativo com o mapa do estado de São Paulo e das delimitações das UGRHs obtido nos relatórios da CETESB.

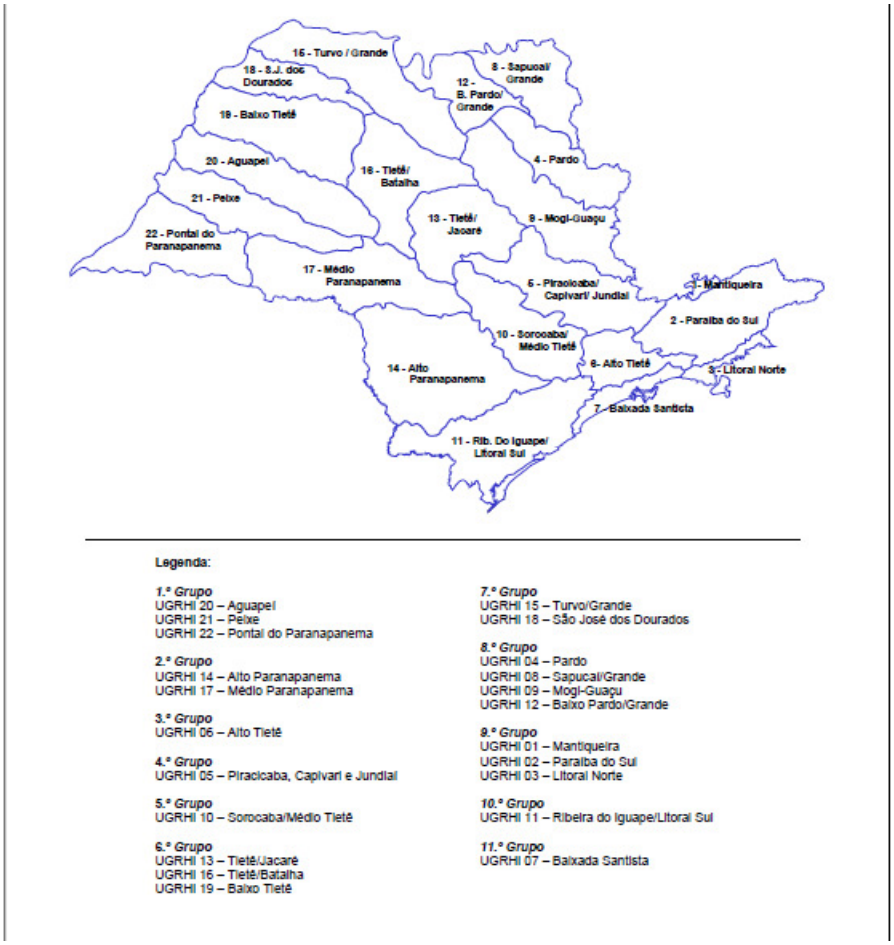

FIGURA 33 - Mapa do estado de São Paulo com as 22 UGRHIs organizadas em 11 grupos. (CETESB, 2001)

Conforme o mapa da figura 33, da região de São Paulo, as 22 UGRHIs encontram-se organizadas em 11 grupos delimitados, nos quais pode – se observar os pontos de coleta nos seguintes grupos:

UGRHI 01 – Mantiqueira, 9°grupo.

UGRHI 02 – Paraíba do sul, 9°grupo.

UGRHI 04 – Rio Pardo, 8°grupo.

UGRHI 05 – Capivari, 4° grupo.

UGRHI 19 – Baixo Tietê (Biritiba Mirim), 6°grupo.

Com base no mapa do estado de São Paulo, na figura 33, há proximidades entre as regiões de Mantiqueira, Paraíba do Sul e Capivari. Devido às características semelhantes, há a

proximidade entre os clusters de forma semelhante á apresentada figura 20. Assim como o distanciamento entre a região do Rio Pardo e a região de Biritiba mirim (Baixo Tietê).

Na comparação entre os mapas da figura 33 e 20 é possível identificar o reconhecimento das características geográficas nos agrupamentos dos mapas SOM.

Esta propriedade denota uma vantagem do emprego da ferramenta tanto em classificação de dados por grau de semelhança entre variáveis como na classificação por semelhança comparativa entre grupos demonstrados no mesmo resultado.

6.2 Matrizes para estudo de similaridades entre parâmetros fisico-químicos.

Na visualização do mapa da matriz transposta nas figuras 22 e 23 pode-se notar um cluster com protótipos mais próximos associados aos parâmetros pH, Cloreto e OD como BMUs, sugerindo a proximidade de comportamento destes protótipos, com um número total de 11 protótipos gerados.

Há proximidades entre os protótipos associados aos parâmetos Fósforo, Manganês e Nitrogênio Kjeldahl Total (NKT), localizados no canto superior do mapa. As temperatura da água e temperatura do ar estão localizados na parte inferior, e a Turbidez e Condutividade, em áreas isoladas.

Devido às características especificas dos parâmetros Fenóis DQO e DBO, decidiu-se fazer um estudo com uma matriz de dados sem a utilização destes parâmetros. Eles apresentavam valores constantes e erro indefinido e poderiam estar alterando treinamento da rede. Este novo critério aplicado para exclusão dos três parâmetros foi descrito no tópico 4.1.2 (Organização do banco de dados).

No mapa da matriz alterada na figura 24 pode-se notar um cluster de protótipos próximos como observados na escala lateral (0-0,8) associados aos parâmetros pH, Cloreto e OD.

Há também um comportamento semelhante observado no cluster de protótipos associados aos parâmetros Fósforo (P), Manganês (Mn) e NKT. A proximidade dos protótipos associados aos parâmetros Temperatura de água e do ar está na escala lateral de variação

de 1-1,5. A Turbidez e a Condutividade encontram-se em áreas isoladas distintas, um comportamento comum dos dois parâmetros também notado em outros mapas.

6.3 Gráficos dos protótipos de vetores

Os gráficos (Anexo B) são os comportamento amostrais dos parâmetros não normalizados e representam os dados de entrada do banco de dados. As figuras 25 a 29 dos gráficos dos protótipos de vetores, consistem no conjunto das saídas. Por meio da comparação entre ambos, pode-se reconhecer áreas similares.

A extração dos protótipos de vetores e o reconhecimento por meio dos gráficos dos mesmos além de ser um recurso adicional, demonstra os comportamento amostrais nas características dos perfis médios gerados.

6.4 Considerações finais.

Em estudos aplicados a matrizes ambientais, os mapas auto – organizáveis de Kohonen se mostraram como uma ferramenta útil, na identificação de correlações conhecidas e não conhecidas, apresentando uma rápida visualização dos resultados.

No estudo de classes de dados, o SOM demonstrou a adaptação do resultado aos parâmetros de treinamento definidos, com uma satisfatória representatividade de 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

Cabe ressaltar que há limitações inerentes ao analista na interpretação de banco de dados multivariados, na compreensão da correlação das variáveis.

Nos aspetos 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.

Trabalhos Futuros

Devido ao potencial da ferramenta, a continuidade de trabalhos na avaliação de matrizes ambientais com SOM está na certificação da ferramenta como instrumento indicativo para determinação e estabelecimento de critérios e parâmetros.

Tanto para satisfazer a necessidade de geração de informação como base de dados como para melhoria na qualidade do tratamento da água por meio de obtenção de novos índices.

ANEXO - A Tabela dos valores dos parâmetros físico químicos (período de 2000 a 2008)

Referente aos resultados dos 13 parâmetros coletados nas UGRHs das cinco regiões, organizados no formato de planilha

Valores dos parâmetros físico – químicos no período de 2000 à 2008.														
		pH	Temp. Água	Temp. Ar	Cloreto Total	Condutividade	DBO (5,20)	DQO	Fenóis	Fósforo Total	Manganês	NKT	OD	Turbidez
		U.pH	ºC	ºC	mg/L	µS/cm	mg/L	mg/L	mg/L	mgP/L	mg/L	mgN/L	mgO_2/L	UNT
Rio Pardo	04/12/2006	6,2	27	32	2,3	66,4	2	50	0,003	0,029	0,04	0,25	5,5	15
Rio Pardo	28/02/2007	6,6	25,5	28,5	2	57,1	2	50	0,001	0,046	0,06	0,33	7,1	30
Rio Pardo	11/04/2007	6,8	26	30	2	55,1	2	50	0,001	0,01	0,06	0,41	6,5	20
Rio Pardo	14/06/2007	6,7	21,4	28	2,5	57,8	2	50	0,001	0,034	0,03	0,6	7,8	3,82
Rio Pardo	08/08/2007	5,9	19,4	26	1,5	53,8	2	50	0,001	0,037	0,04	0,15	7,6	5,75
Rio Pardo	22/10/2007	6,3	25,5	28	2,5	57	4	50	0,002	0,01	0,03	0,2	7	6,12
Rio Pardo	03/12/2007	6,1	28	34	3	63,4	2	50	0,002	0,14	0,04	0,81	6,3	10,2
Rio Pardo	18/02/2008	7,7	25,9	29	1,5	48,1	2	50	0,002	0,039	0,0568	0,26	7,8	21,6
Rio Pardo	09/04/2008	7,2	24,4	29	2	48	2	50	0,002	0,043	0,0488	0,36	8,7	52,3
Rio Pardo	12/06/2008	6,9	25	27,5	1,5	50,2	2	50	0,002	0,016	0,0233	0,27	7,3	4,83
Rio Pardo	11/08/2008	6,7	24	33	2,5	53,1	2	50	0,003	0,007	0,0111	0,15	7,8	2,2
Rio Pardo	21/10/2008	6,3	22,8	25	2,5	52,2	2	50	0,003	0,009	0,0502	0,18	8,7	1,48
Rio Capivari	11/01/2000	7,5	25	29	9	157	4	42	0,004	0,332	0,28	1,3	5,4	200
Rio Capivari	21/03/2000	7	25	31	7,8	150	3	18	0,005	0,18	0,21	0,98	7,6	13
Rio Capivari	09/05/2000	7,3	21	21	16,3	232	6	18	0,003	0,261	0,16	2	6,1	18
Rio Capivari	11/07/2000	7,2	17	21	21,8	278	11	15	0,003	0,41	0,27	2,5	4,5	15
Rio Capivari	19/09/2000	7,3	22	28,7	11	176	11	34	0,007	0,201	0,19	1,6	7	80
Rio Capivari	20/11/2000	7,4	22	27	9	145	8	37	0,005	0,195	0,15	0,93	7	140
Rio Capivari	09/01/2001	7	26	26	7,9	147	9	38	0,003	0,357	0,38	0,49	6,6	160
Rio Capivari	13/03/2001	7,4	24	28	7,9	155	9	31	0,003	0,272	0,42	1,3	4,9	255

Rio Capivari	01/02/2006	7	23	28	9	124	6	50	0,003	0,7	0,29	0,5	6	506
Rio Capivari	24/04/2006	7,3	20	26	17	190	6	50	0,003	0,4	0,18	1,44	6,1	26
Rio Capivari	05/06/2006	7,4	16,5	19,5	22	200	6	50	0,003	0,5	0,23	2,36	6,2	15
Rio Capivari	08/08/2006	7	27	24	21	215	4	50	0,003	0,5	0,19	2	5,5	12
Rio Capivari	02/10/2006	7,3	21	28	33	302	7	50	0,003	0,6	0,22	5	5,2	19
Rio Capivari	11/12/2006	7,2	23	26	17	196	5	50	**0,002**	0,3	0,21	1	5,6	53
Rio Capivari	05/02/2007	7,1	25	28	19	221	6	50	0,001	0,3	0,24	1	4,6	34
Rio Capivari	23/04/2007	6,7	23,5	28	24	232	8	50	0,005	0,05	0,21	4	5,7	41
Rio Capivari	04/06/2007	6,7	16,5	12,5	16	157	6	50	0,005	0,6	0,13	1	6	60
Rio Capivari	07/08/2007	7,1	17	26	23	220	8	50	0,005	0,2	0,19	4	6,5	22
Rio Capivari	01/10/2007	4,5	20	27	50	369	10	50	0,005	0,7	0,3	5	5,2	16
Rio Capivari	10/12/2007	7,1	25	27	18	218	10	50	0,005	0,4	0,21	0,05	5	80
Rio Capivari	11/02/2008	7	23,5	26	8	115	6	50	0,005	0,9	0,35	2	6,4	664
Rio Capivari	01/04/2008	7,2	22	26	20	196	7	50	0,002	0,3	0,09	2	6,3	55
Rio Capivari	02/06/2008	6,6	16	18	13	134	8	50	0,002	0,5	0,2	1	7,3	306
Rio Capivari	05/08/2008	6,6	18	23	18	215	7	50	0,002	0,4	0,3	3	5,6	61
Rio Capivari	06/10/2008	6,4	17	17	23	177	8	50	0,002	0,3	0,2	2	5,7	102
Rio Capivari	16/12/2008	6,6	22	26	22	215	6	50	0,002	0,4	0,3	4	4,8	39
R. Paraiba	16/02/2000	6,5	28	31	6,8	68	3	14	0,002	0,09	0,14	0,57	4,2	89
R. Paraiba	05/04/2000	6,7	24	28	6,6	71	2	18	0,001	0,08	0,14	0,84	5,4	86
R. Paraiba	28/06/2000	6,7	22	28	2,7	49	1	8	0,001	0,07	0,04	0,21	6,3	22
R. Paraiba	15/08/2000	6,6	20	26	5	45	1	4	0,001	0,03	0,05	0,36	6,8	23
R. Paraiba	18/10/2000	6,7	28	34	6,9	57	1	26	0,001	0,07	0,07	0,19	6,0	20
R. Paraiba	14/12/2000	6,6	28	28	8,4	89	3	8	0,001	0,08	0,11	0,38	4,5	30
R. Paraiba	20/02/2001	6,5	29	34	6,1	88	3	16	0,001	0,05	0,1	1,3	3,5	68
R. Paraiba	04/04/2001	6,4	27	30	4,9	88	2	22	0,001	0,06	0,08	0,3	3,3	106
R. Paraiba	26/06/2001	6,9	22	22	4,8	66	2	6	0,001	0,06	0,03	0,33	5,9	16
R. Paraiba	21/08/2001	6,4	22	34	4,5	56	2	10	0,001	0,08	0,03	0,26	6	21

R. Paraiba	21/06/2006	6,5	19	21	5,2	78	2	4	0,001	0,08	0,05	0,44	7,6	16
R. Paraiba	15/08/2006	6,3	19	24	5,1	84	2	50	0,001	0,08	0,05	0,44	5,3	24
R. Paraiba	24/10/2006	6,1	20	20	7,5	86	2	50	0,001	0,09	0,05	1	5,2	20
R. Paraiba	06/12/2006	6,6	24	27	7,2	112	2,6	50	0,001	0,13	0,09	0,58	3,6	44
R. Paraiba	22/02/2007	6,5	25	26	7	109	1,4	50	0,001	0,15	0,09	0,49	3,9	28
R. Paraiba	19/04/2007	7	23	27	6,4	94	0,4	50	0,001	0,06	0,07	0,37	4,2	28
R. Paraiba	20/06/2007	6,6	20	20	5,7	75	2	**36,22**	0,001	0,08	0,03	0,33	5	12
R. Paraiba	28/08/2007	6,5	17	17	5,8	94	2	**36,22**	0,001	0,14	0,04	0,37	4,7	19
R. Paraiba	19/10/2007	6,6	21	20	6,6	87	2	**36,22**	0,002	0,1	0,06	0,58	4,1	21
R. Paraiba	05/12/2007	6,7	27	29	11	153	2	**36,22**	0,002	0,13	0,14	0,77	3	16
R. Paraiba	21/02/2008	6,5	25	28	4,8	97	2	**36,22**	0,002	0,05	0,04	0,67	2,9	59
R. Paraiba	03/04/2008	6,6	24	24	8,1	123	2	**36,22**	0,002	0,1	0,09	0,75	3,5	25
R. Paraiba	18/06/2008	6,7	21	23	6,5	105	2	**36,22**	0,002	0,11	0,07	0,7	4,7	18
R. Paraiba	19/08/2008	6,7	23	26	9,4	125	2	**36,22**	0,003	0,13	0,12	0,63	3,8	19
R. Paraiba	02/10/2008	6,5	22	23	4,7	84	2	**36,22**	0,003	0,11	0,05	0,43	3,6	31
R. Paraiba	03/12/2008	6,6	25	25	5,8	109	2	**36,22**	0,003	0,05	0,08	0,77	2,7	33
Mantiqueira	16/01/2001	7	25	22	5,9	110	8	30	0,001	0,38	0,15	**1,41**	4,8	29
Mantiqueira	01/03/2001	6,9	19	18	2	91	9	22	0,001	0,26	0,14	**1,41**	5,4	13
Mantiqueira	23/05/2001	6,7	15	20	2,1	74	5	16	0,001	0,14	0,13	**1,41**	7	21
Mantiqueira	17/07/2001	7,2	13	18	4,7	104	9	24	0,009	0,34	0,24	**1,41**	5,7	14
Mantiqueira	18/09/2001	6,6	13	16	5,6	91	6	18	0,001	0,25	0,14	**1,41**	7,5	15
Mantiqueira	13/11/2001	7	18	20	2,8	56	1	4	0,001	0,15	0,08	**1,41**	6,4	87
Mantiqueira	22/01/2002	5,9	21	20	2,9	61	3	4	0,001	0,13	0,12	1	7	32
Mantiqueira	03/05/2002	7,4	19	21	2,1	52	2	4	0,001	0,01	0,23	0,78	7,6	40
Mantiqueira	05/02/2002	5,5	22	20	2	56	2	4	0,001	0,13	0,09	1,6	7	15
Mantiqueira	07/10/2002	7,2	21	26	4,7	78	7	17	0,001	0,34	0,06	3,7	7,5	15
Mantiqueira	18/09/2002	7,5	17	20	4,5	71	5	10	0,001	0,22	0,13	2,9	6,1	16
R. Paraiba	27/04/2006	6,5	23	23	7,4	104	4	8	0,001	0,11	0,07	0,26	3,2	23

Mantiqueira	10/07/2007	6,4	12	20	4,5	86	9	50	0,001	0,2	0,08	2,26	7,7	41
Mantiqueira	12/09/2007	6,9	17	21	4,3	85	5	**36,22**	0,001	0,29	0,09	3,52	5,5	9
Mantiqueira	28/11/2007	7,5	21	23	1,8	83	3	**36,22**	0,002	0,32	0,12	5	6,3	9,9
Mantiqueira	07/01/2008	7,1	19	22	3,3	6,8	4	**36,22**	0,002	0,15	0,15	1,28	5,9	47
Mantiqueira	12/03/2008	7,5	18	20	1,2	65	2	**36,22**	0,002	0,17	0,09	1,25	6,4	0,7
Mantiqueira	13/05/2008	6,7	12	13	2,7	51	3	**36,22**	0,002	0,13	0,06	1	7,3	17
Mantiqueira	15/07/2008	7	9,9	19	3,3	74	5	**36,22**	0,002	0,23	0,11	2,72	7	13
Mantiqueira	11/09/2008	7,1	16	23	5,1	94	5	50	0,003	0,34	0,07	2,9	6,2	13
Mantiqueira	26/11/2008	7,3	17,4	21	2,3	68	5	50	0,003	0,19	0,07	1,22	6,8	42
Biritiba.Mirim	13/01/2000	6,1	25	33	3,6	47	**3,99**	39	0,003	0,03	0,07	0,6	3,8	13
Biritiba.Mirim	27/03/2000	6,1	22	25	4,6	37	3	25	0,003	0,11	0,04	0,31	4,8	13
Biritiba.Mirim	24/05/2000	7,5	15	20	3,7	28	4	25	0,003	0,04	0,002	0,76	7,8	4
Biritiba.Mirim	26/07/2000	6,7	14	17	7,1	59	3	25	0,003	0,11	0,002	0,27	7,4	3
Biritiba.Mirim	21/09/2000	6,6	21	28	6,2	49	3	25	0,003	0,13	0,02	0,87	6,1	7
Biritiba.Mirim	30/11/2000	6,2	22	23	8,1	66	4	52	0,003	0,3	0,05	1,12	3,5	106
Biritiba.Mirim	11/01/2001	6,6	25	30	3,41	41	3	26	0,004	0,07	0,03	0,67	5,7	9,24
Biritiba.Mirim	27/03/2001	6,7	24	27	5,83	69	3	29	0,003	0,11	0,03	0,86	5,7	35,7
Biritiba.Mirim	30/05/2001	5,9	18	25	4,06	43	1	12	0,003	0,07	0,002	0,1	8	6,48
Biritiba.Mirim	25/07/2001	6,9	16	18	4,18	43	3	25	0,003	0,03	0,02	1,88	6,7	9,2
Biritiba.Mirim	18/09/2001	6,4	15	17	4,67	51	3	50	0,003	0,03	0,02	1,03	8,4	5,08
Biritiba.Mirim	29/11/2001	6,3	21	29	5,39	87	9	50	0,003	0,39	0,06	1,92	5	38
Biritiba.Mirim	15/01/2002	6,4	21	24	3,9	65	3	50	0,003	0,22	0,03	0,66	4,1	50
Biritiba.Mirim	26/03/2002	6,3	22	28	6,5	49	3	50	0,003	0,15	0,06	0,61	4,6	12
Biritiba.Mirim	28/05/2002	6,5	15	24	3,9	32	3	50	0,003	0,03	0,03	0,03	10	1
Biritiba.Mirim	25/07/2002	6,8	15	25	4,8	44	3	50	0,003	0,05	0,01	0,34	8,9	7
Biritiba.Mirim	17/09/2002	6,6	20	29	4	41	3	50	0,003	0,04	0,04	0,03	7,8	4
Biritiba.Mirim	25/11/2002	6	23	31	3,8	36	3	50	0,003	0,08	0,11	1,31	4,8	10
Mantiqueira	15/05/2007	6,9	18	25	3,2	64	2	50	0,001	0,01	0,08	3,56	6,9	11

Biritiba.Mirim	11/09/2007	7,1	18,4	24,5	3,97	45,8	3	50	0,003	0,03	0,02	0,81	7,5	**40,79**
Biritiba.Mirim	27/11/2007	6,8	21,3	22,5	5,93	68,6	3	50	0,003	0,07	0,05	0,64	6,5	14,48
Biritiba.Mirim	08/01/2008	6,7	23,5	29,5	7,71	100,5	3	50	0,003	0,09	0,08	0,5	6,2	15,43
Biritiba.Mirim	27/03/2008	6,9	22,3	29,4	6,79	91,4	3	50	0,003	0,05	0,07	0,65	7	12,77
Biritiba.Mirim	08/05/2008	6,9	16,5	19,3	6,62	86,8	3	50	0,009	0,08	0,04	0,5	7,8	22
Biritiba.Mirim	24/07/2008	7	17,9	23	2,89	50,7	3	50	0,003	0,15	0,02	0,72	7,9	5,28
Biritiba.Mirim	04/09/2008	7,2	20	31	4,49	50,9	3	**36,22**	0,003	0,05	0,03	0,62	7,6	3,95
Biritiba.Mirim	25/11/2008	7	21,4	26	5,58	69,3	4	**36,22**	0,003	0,04	0,06	0,92	7	9,62
Biritiba.Mirim	13/05/2004	6,5	18,5	21	4,16	39	3	50	0,003	0,05	0,02	0,34	6,9	11
Biritiba.Mirim	29/07/2004	6,4	14,5	19	4,35	40,7	3	50	0,003	0,06	0,01	0,18	8	2,5
Biritiba.Mirim	16/09/2004	6,7	19,2	27	4,33	40,8	3	50	0,003	0,03	0,03	0,62	5,7	0,9
Biritiba.Mirim	03/11/2004	6,8	24	30,5	6,48	58,8	3	50	0,003	0,06	0,07	0,76	5,7	7,45
Biritiba.Mirim	11/01/2005	6,72	24,5	24	4,31	61,7	3	50	0,005	0,19	0,08	0,39	4,27	39
Biritiba.Mirim	01/03/2005	6,3	24,1	28	4,78	55,1	3	50	0,003	0,12	0,07	0,91	4,2	13
Biritiba.Mirim	12/05/2005	6,52	19,6	28,5	0,5	37,7	3	50	0,003	0,04	0,02	0,28	7,04	11,51
Biritiba.Mirim	28/07/2005	6,8	17,2	20	5,99	76,9	3	78	0,003	0,1	0,1	0,84	6,7	7,6
Biritiba.Mirim	13/09/2005	6,5	17,7	18	7,12	84,1	3	50	0,003	0,13	0,12	0,8	6,4	8,4
Biritiba.Mirim	03/11/2005	7	20	25,5	8,68	138,2	3	50	0,003	0,24	0,007	1,79	5,9	12,5
Biritiba.Mirim	10/01/2006	7,1	24	28	5,74	84,5	3	50	0,003	0,14	0,06	1	5,7	7,2
Biritiba.Mirim	08/03/2006	6,1	23,6	29	6,78	106,5	3	50	0,003	0,1	0,07	0,87	6,7	2,8
Biritiba.Mirim	11/05/2006	6,8	18,6	21	3,75	59,9	3	50	0,003	0,02	0,03	0,52	7,2	3,9
Biritiba.Mirim	27/07/2006	6,8	17,2	28,5	3,32	44,9	3	50	0,003	0,11	0,02	0,92	7,8	0,8
Biritiba.Mirim	12/09/2006	6,7	19,2	32	4,23	41,6	3	50	0,003	0,11	0,04	0,74	7,5	4,3
Biritiba.Mirim	07/11/2006	6,2	21,2	19,5	4,25	56,9	3	50	0,003	0,09	0,04	0,38	6,2	5,7
Biritiba.Mirim	09/01/2007	6,7	23,3	24,5	6,09	78,8	3	50	0,003	0,07	0,09	0,77	5,8	22,4
Biritiba.Mirim	09/03/2007	7	25,9	28,5	4,09	50,3	3	50	0,003	0,07	0,08	0,57	6	4,72
Biritiba.Mirim	09/05/2005	6,8	19,4	14	4,54	52,9	3	50	0,003	0,18	0,04	2,13	6,6	10,5
Biritiba.Mirim	26/07/2007	6,5	15,3	16,5	8,59	114,4	3	50	0,003	0,16	0,04	0,59	7,8	30,27

ANEXO - B Perfis amostrais dos treze parâmetros .

Nesta seção são mostrados os comportamentos amostrais dos resultados referentes a Tabela dos valores dos parâmetros físico - químicos no período de 2000 à 2008 (Anexo A).

Os gráficos correspondem a variação das concentrações dos parâmetros em 257 dias de coleta.

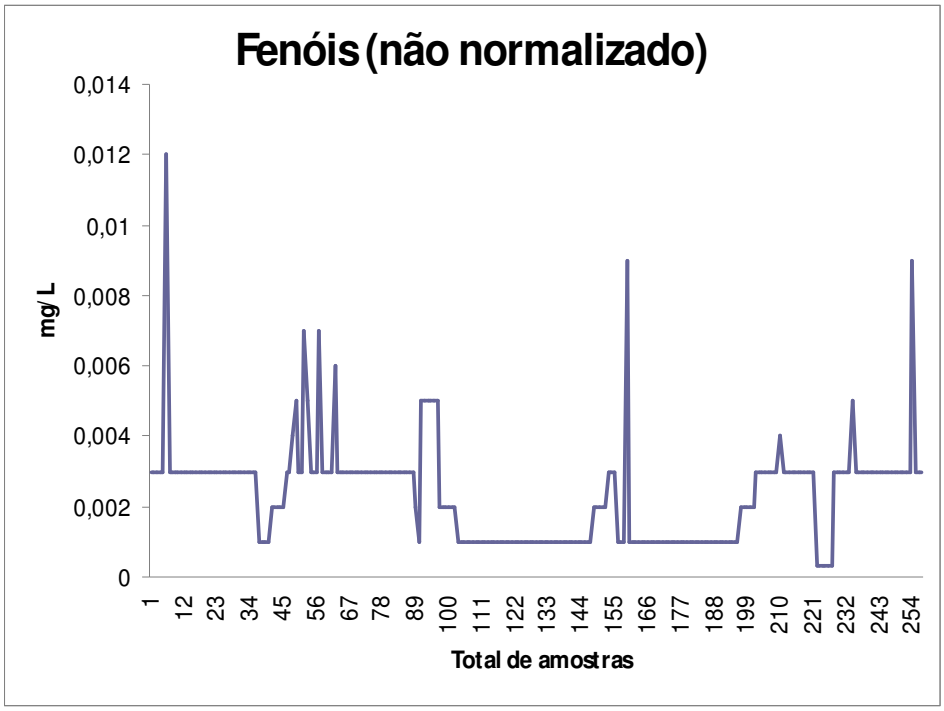

FIGURA B.1 - Gráfico do comportamento não normalizado dos valores das coletas referentes ao periodo de 2000 a 2008 da variável Fenóis.

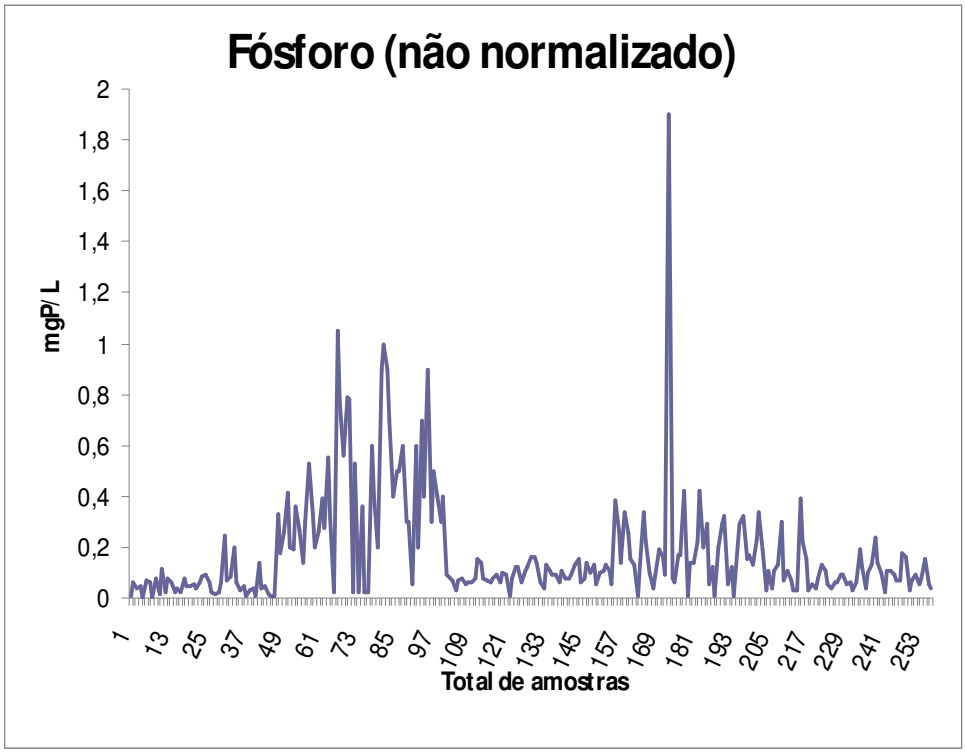

FIGURA B.2 – Gráfico do comportamento não normalizado dos valores das coletas referentes ao período de 2000 a 20008 da variável Fósforo.

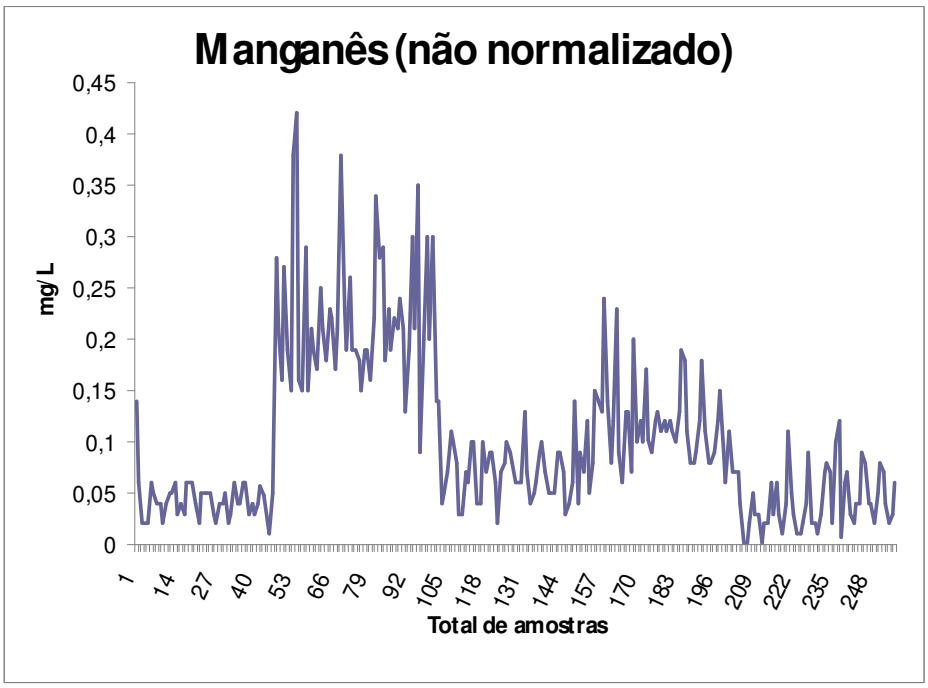

FIGURA B.3 – Gráfico do comportamento não normalizado dos valores das coletas referentes ao período de 2000 a 2008 da variável Manganês.

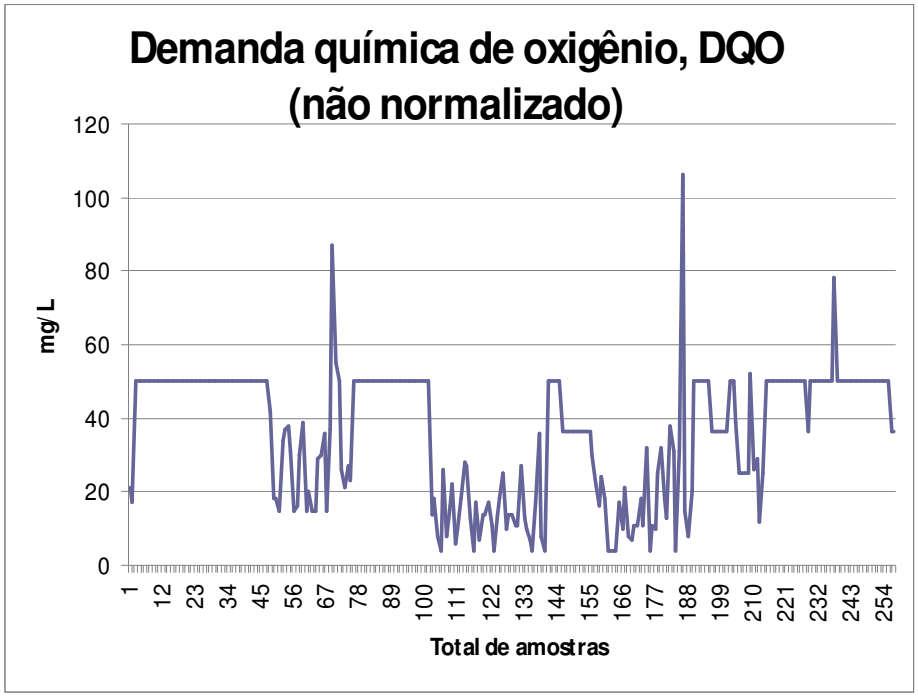

FIGURA B.4 – Gráfico do comportamento não normalizado dos valores das coletas referentes ao período de 2000 a 2008 da variável Demanda Química de Oxigênio.

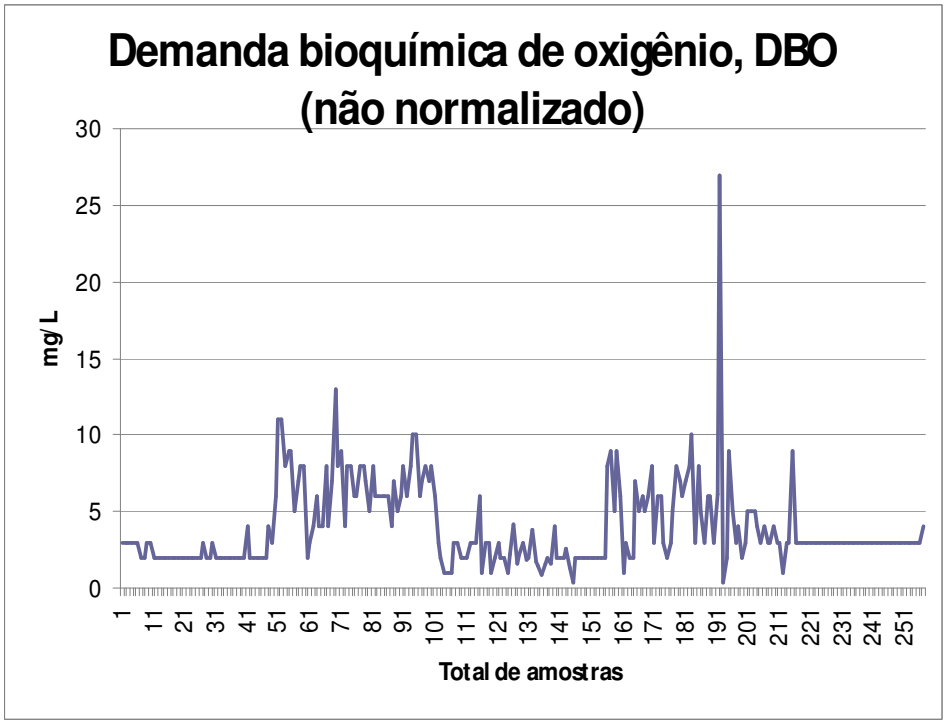

FIGURA B.5 – Gráfico do comportamento não normalizado dos valores das coletas referentes ao período de 2000 a 2008 da variável Demanda Bioquímica de oxigênio.

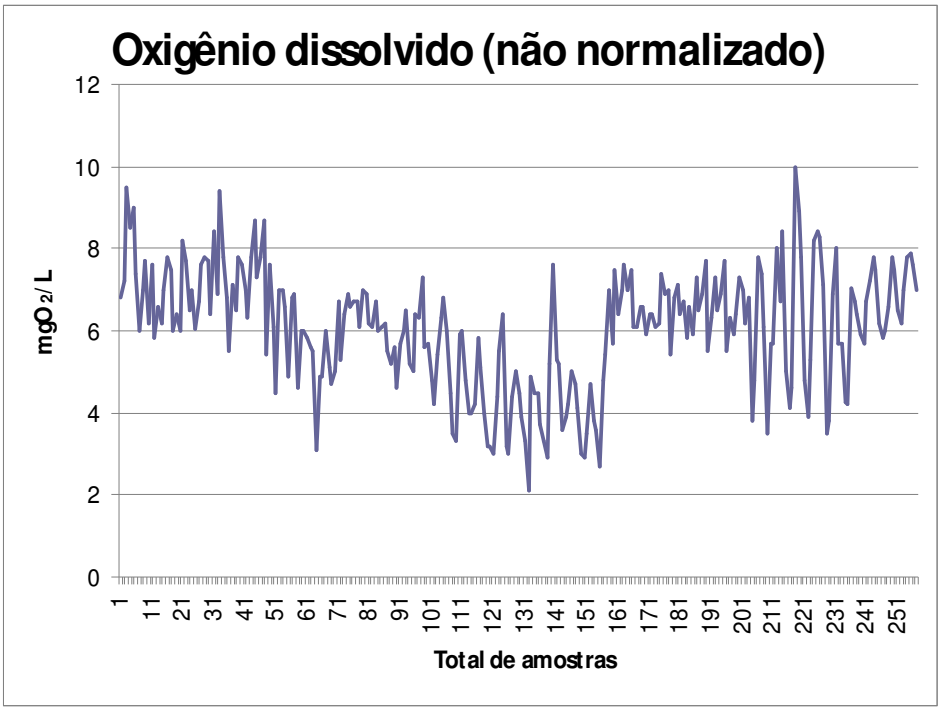

FIGURA B.6 – Gráfico do comportamento não normalizado dos valores referentes ao período de 2000 a 2008 da variável Oxigênio Dissolvido.

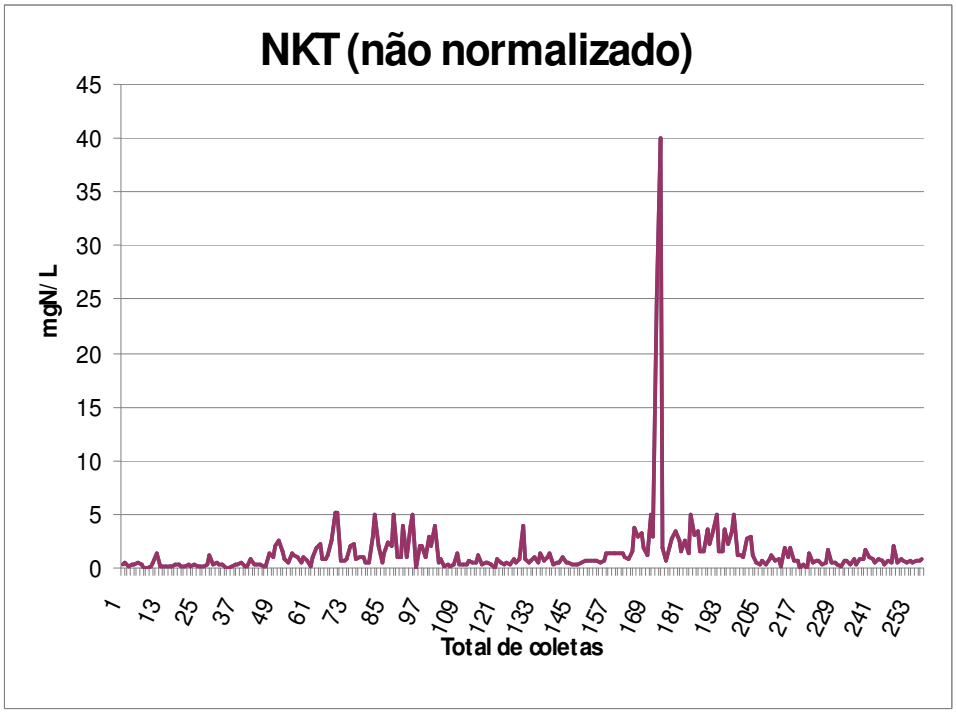

FIGURA B.7 – Gráfico do comportamento não normalizado dos valores referentes ao período de 2000 a 2008 da variável Nitrogênio Kjeldahl Total.

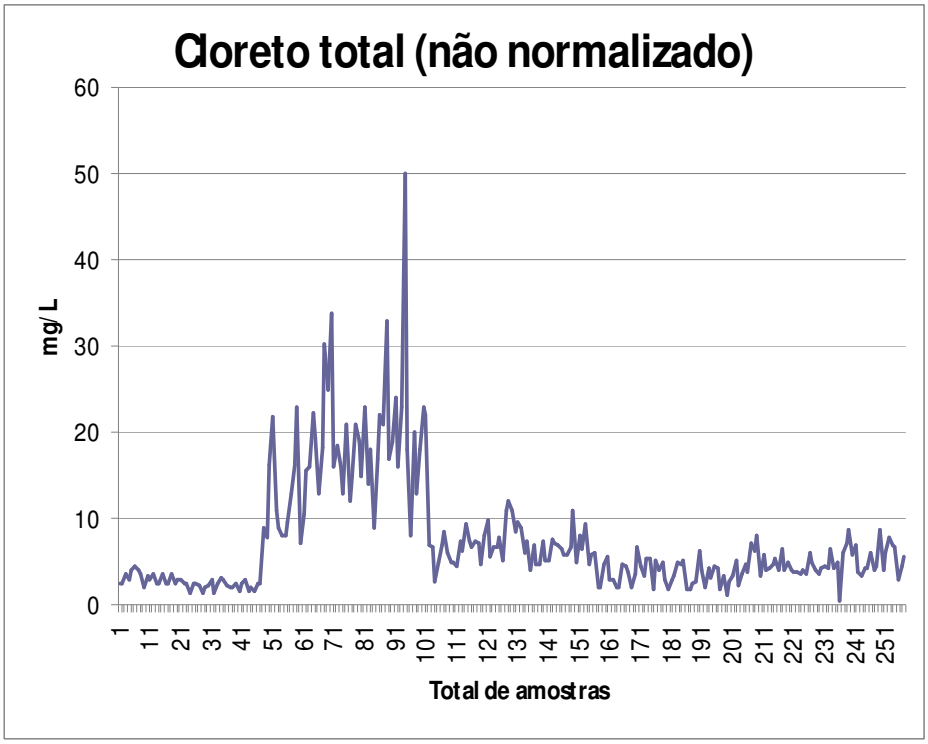

FIGURA B.8 – Gráfico do comportamento não normalizado dos valores de coleta referentes ao período de 2000 a 2008 da variável Cloreto Total.

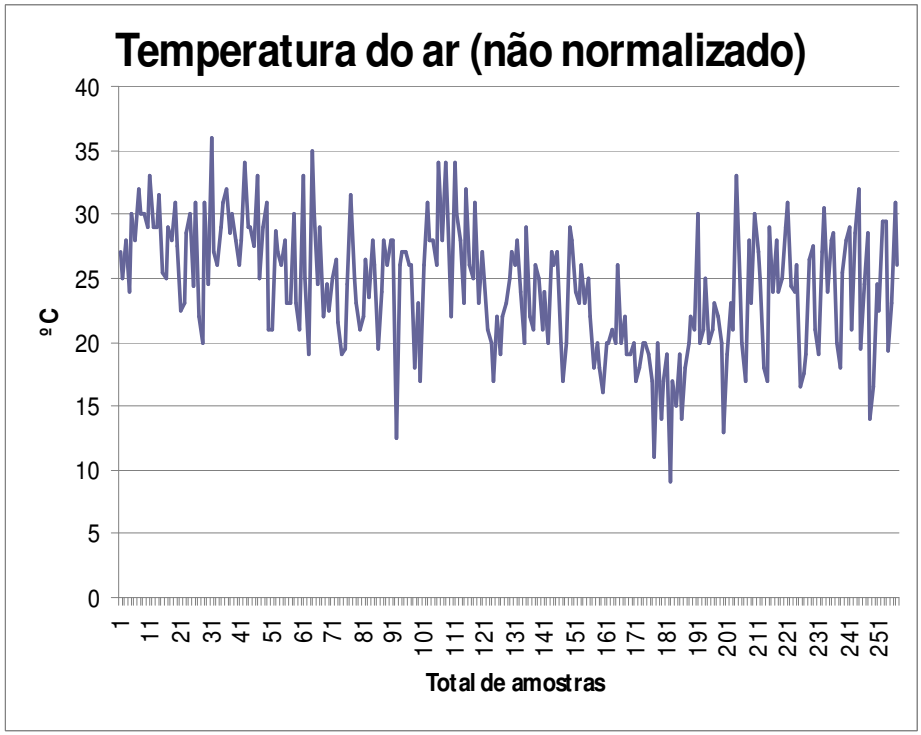

FIGURA B.9 – Gráfico do comportamento não normalizado dos valores de coletas referentes ao período de 2000 a 2008 da variável Temperatura do ar.

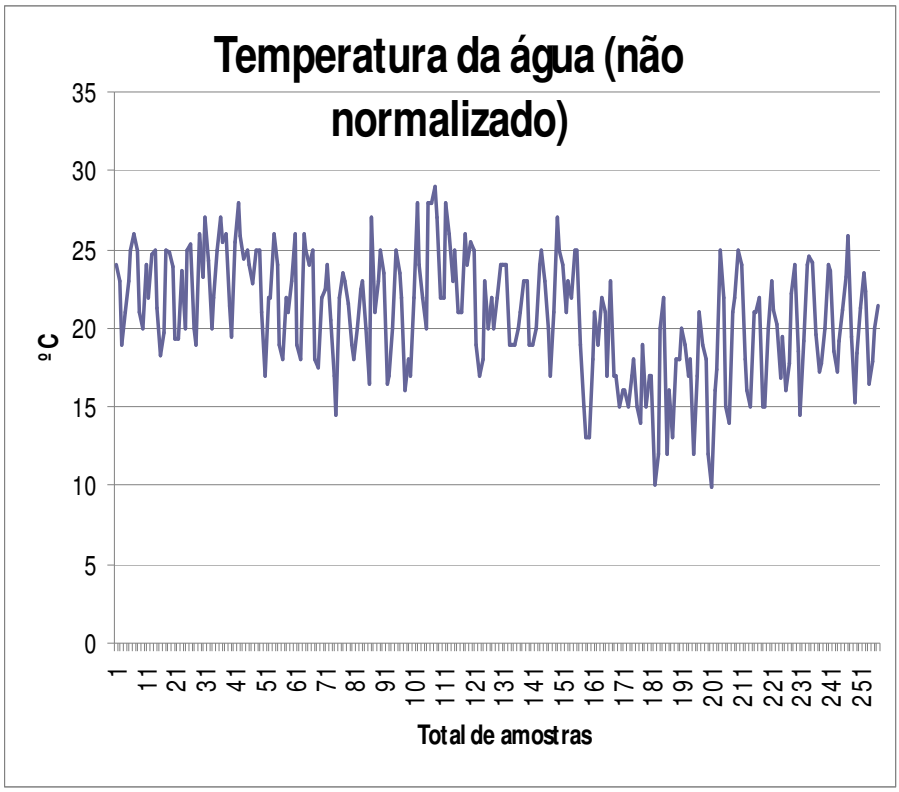

FIGURA B.10 – Gráfico do comportamento não normalizado dos valores das coletas referentes ao periodo de 2000 a 2008 da variávelTemperatura da água.

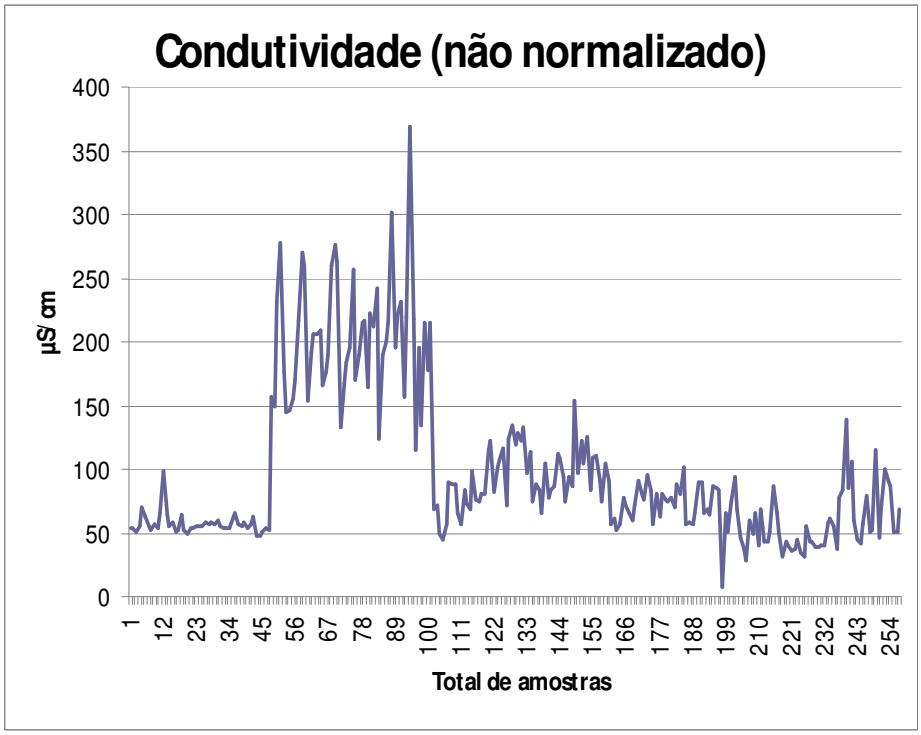

FIGURA B.11 – Gráfico do comportamento não normalizado dos valores de coleta referentes ao período de 2000 a 2008 da variavel Condutividade.

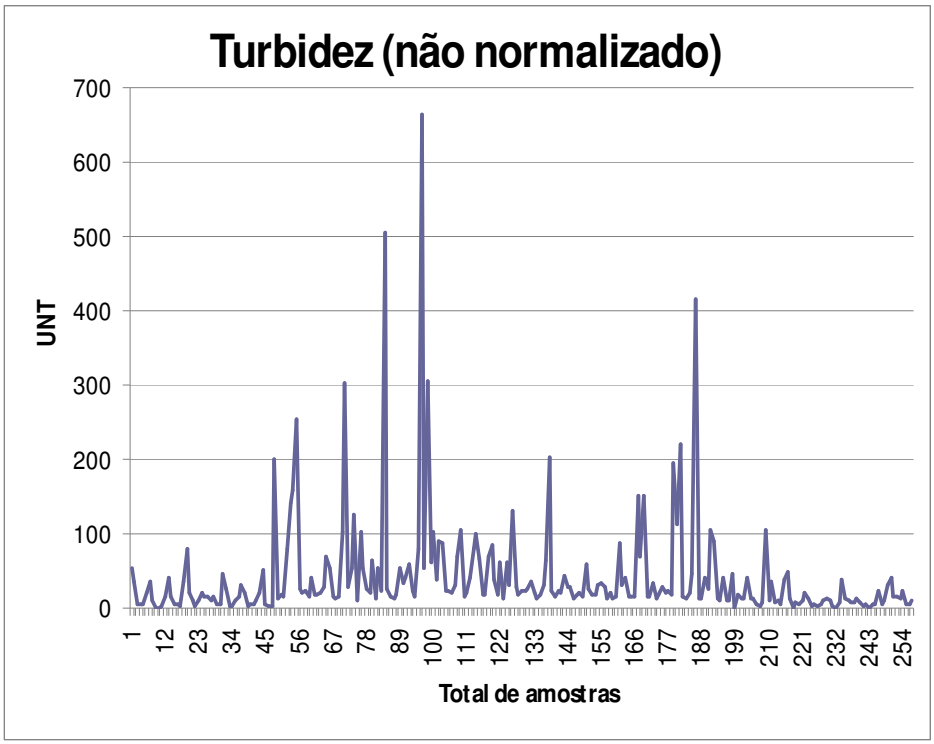

FIGURA B.12 - Gráfico do comportamento não normalizado dos valores de coleta referentes ao período de 2000 a 2008 da variável Turbidez.

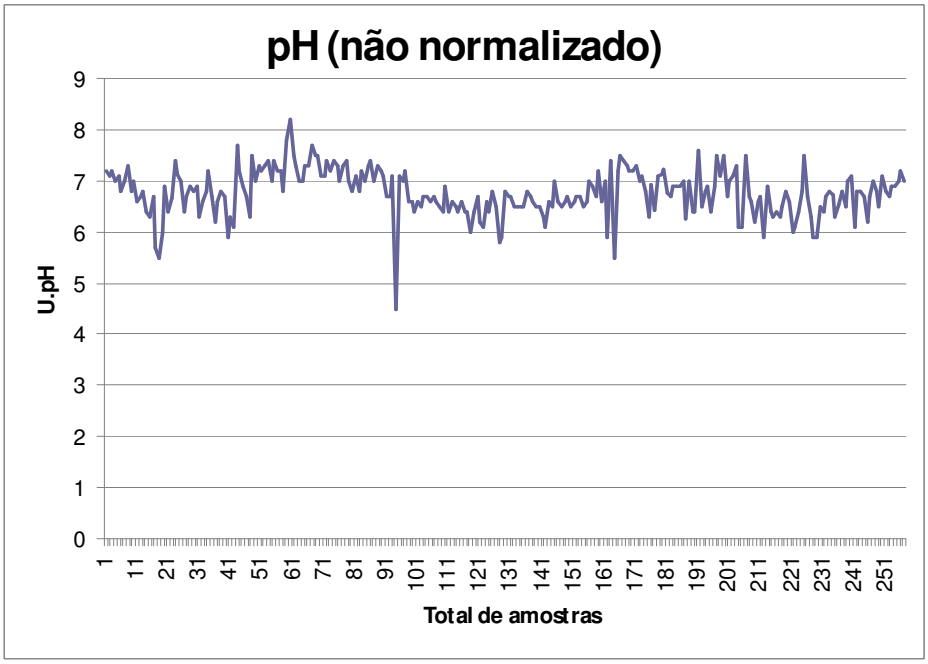

FIGURA B.13 - Gráfico do comportamento não normalizado dos valores de coletas referentes ao período de 2000 a 2008 da variável Potencial hidrogeniônico.

REFERÊNCIAS BIBLIOGRÁFICAS

ASTEL A, S. TSAKOVSKI, P. BARBIERI, V. SIMEONOV, Comparison of self – organizing maps classification approach with cluster and principal components abalysis for large environmental data sets. **Water Research**, 41(19), p. 4566 – 4578, 2007.

BADIN JR., H., **Redes Neurais Artificiais - Parte 2**, **-** Neurônio Biológico, Mecatrônica Atual, disponível em: http://www.mecatronicaatual.com.br/secoes/leitura/553, acessado em 20/01/2011.

BIERMAN P.H., LEWIS M. OSTENDORF B., TANNER J., A review of methods for analysing spatial and temporal patterns in coastal water quality. **Ecological Indicators**, 11, p.103-114, 2011.

BRASIL. MINISTÉRIO DA SAÚDE, **Portaria n. 518, 2004.** Ministério da saúde do Brasil, D.O.U., de 25 de março de 2004, Brasília.

BRASIL. MINISTÉRIO DO DESENVOLVIMENTO URBANO E MEIO AMBIENTE CONSELHO NACIONAL DO MEIO AMBIENTE (CONAMA). **Resolução n. 357,** D.O.U., de 17 de Março de 2005, Brasília.

BUENO E.I., **Utilização de redes neurais artificiais na monitoração e detecção de falhas em sensores do reator IEA – R1,** Dissertação (Mestrado), Instituto de Pesquisas Energéticas e Nucleares – IPEN, São Paulo, 2006.

CARVALHO M.A.G, **Métodos estatísticos para análise de dados de monitoração ambiental.** Tese (Doutorado), Instituto de Pesquisas Energéticas e Nucleares – IPEN, São Paulo, 2003.

CÉRÉGHINO R., PARK Y.-S., Review of Self-Organizing Map (SOM) approach in water resources: Commentary. **Environmental Modelling & Software**, 24, 945-947, 2009.

CETESB – Companhia de Tecnologia de Saneamento de São Paulo. **Relatório de qualidade das águas interiores do estado de São Paulo 2000.** CD, 2v, il, série de relatórios CETESB, São Paulo, 2001.

CETESB – Companhia de Tecnologia de Saneamento de São Paulo. **Variáveis de qualidade das águas.** Disponível em: http://www.cetesb.sp.gov.br/Agua/rios/variaveis.asp , Acesso em: 19/07/2008.

CONAMA - MINISTÉRIO DO MEIO AMBIENTE. CONSELHO NACIONAL DO MEIO AMBIENTE, "Resolução n.º 357 de 17/03/2005, D.OU. n.º 53, Brasília, Brasil.

COSTA F.A.J, NETTO A..L.M, Segmentação de mapas auto – organizáveis com espaço de saída 3 – D, **Revista Controle &Automação,** v.18 n.2., p. 150 – 162, Abr./Mai/Jun. 2007.

COTRIM M.E.B, **Avaliação da qualidade da água na bacia hidrográfica do Ribeira de Iguape com vistas ao abastecimento público,** Tese (Doutorado), Instituto de Pesquisas Energéticas e Nucleares – IPEN, São Paulo, 2006.

DAVIES D.L, BOULDIN D.W. A Cluster Separation Measure. **IEEE - Transactions on Pattern Analysis and Machine Intelligence**, New York, abr. 1979, vol PAM1-1. n.o 2.

DECRETO ESTADUAL n.º 8468 de 08/09/76 obtido em: http://www.cetesb.sp.gov.br/Institucional/documentos/Dec8468.pdf, acessado em 21/01/11.

ECHALAR M.A.F., **Estudo da estrutura de fontes de aerossóis em Cubatão com uso de PIXE e modelos receptores,** Dissertação (Mestrado), Instituto de Física da Universidade de São Paulo, São Paulo, 1991.

FILHO B.D. B, **Redes neurais para controle de sistemas de reatores nucleares,** Tese (Doutorado), Instituto de Pesquisas Energéticas e Nucleares – IPEN, São Paulo, 1998.

GARCIA L.H, GONZÁLES M.I, Self – organizing map and clustering for wastewater treatment monitoring. **Engineering Applications of Artificial Intelligence,**17, p.215 – 225, 2004.

HONG T.S.Y, BHAMIDIMARRI R, Evolutionary self – organizing modeling of a municipal wastewater treatment plant, **Water Research,** 37 (6), p. 1199 – 1212, 2003.

HONG T.S.Y, ROSEN R.M, BHAMIDIMARRI R, Analysis of a municipal wastewater treatment plant using a neural network based pattern analysis. **Water Research,** 37, p. 1608 – 1618, 2003.

HONKELA T, **References on self – organizing map,** disponível em:http://mlab.uiah.fi/~timo/som/references.html , Acesso em novembro de 2007.

IGAMI, M.P.Z.; ZARPELON, L.M.C. (Org). **Guia para a elaboração de dissertações e teses**: preparado para orientação dos alunos de Pós-graduação do IPEN. São Paulo: IPEN, Divisão de Informação e Documentação Científicas, 2002. Disponível em: https://www.ipen.br/conteudo/upload/200609111605540.guia_teses.pdf. Acesso em: 25/11/2010.

JUNIOR D.W, **Identificação de padrões em sistemas supervisórios de instalações de reatores nucleares e em sistemas de gasodutos utilizando mapas auto – organizáveis.** Dissertação (Mestrado), Instituto de Pesquisas Energéticas e Nucleares – IPEN, São Paulo, 2005.

KALTEH A.M, HORTH P, BERNDTSSON R, Review of the self – organizing map (SOM) approach in water resources: Analysis, modeling and application. **Environ. Modell & Softw.** 23, p, 835 – 845, 2008.

KOHONEN T, Automatic formation of topological maps of patterns in a self – organizing system. In: Oja, E. e Simula, O., Eds. **Proc. 2SCIA, Scand. Conf. on Image Analysis,** p 214 – 220, Helsinki, Finland, 1981a.

KOHONEN T, Construction of similarity diagrams for phonemes by a self – organizing algorithm. Report TKK – F, A463, Helsinki Uni. Technol., Finland, 1981b.

LCIS, Laboratoty of Computer and Information Science, Som Toolbox 2.0, obtido em: http://www.cis.hut.fi/somtoolbox/, acesado em: 22/01/2011.

LEE H.B, SCHOLZ M, Application of the self – organizing map (SOM) to assess the heavy metal removal performance in experimental wetlands. **Water Research,** 40 (18), p. 3367 – 3374, 2006.

LEK S, GIRAUDEL J.L, Acomparison of self – organizing map algorithm and some conventional statistical for ecological community ordination.**Ecological modeling,** 146 (1-3), p. 329 – 339, 2001.

LEMES M.J.L, **Avaliação de metais e elementos traço em água e sedimentos das bacias hidrográficas dos rios Mogi – guaçu e Pardo,** Dissertação (Mestrado), Instituto de Pesquisas Energéticas e Nucleares – IPEN, São Paulo, 2001.

LLORENS E, THIERY F, GRIEU S, POLIT M, Evaluation of WWTP discharges into a Mediterranena river using KSOM neural networks and mass balance modeling. **Chemical Engineering Journal,** 142 (2), p 135 – 146, 2008.

LNCC, Laboratório Nacional de Computação Científica, Tutorial em Redes Neurais, obtido em: http://www.lncc.br/~labinfo/tutorialRN/frm1_aprendizado.htm, acessado em 21/01/2011.

McCULLOCH W.S, PITTS W.H, Alogical calculus of the ideas immanent in nervous activity. **Bulletin of Mathematical Biophysics,** 5, p. 115 – 133, 1943.

MARQUES N.M, **Avaliação do impacto de agrotóxicos em áreas de proteção ambiental, pertencentes à bacia hidrográfica do Rio Ribeira de Iguape, São Paulo. Uma contribuição à análise crítica da legislação sobre o padrão de potabilidade.** Tese (Doutorado), Instituto de Pesquisas Energéticas e Nucleares – IPEN, São Paulo, 2005.

MATHWORKS: Matlab versão 6.5, [S.I], Mathworks Inc., 2004, Conjunto de programas, 2 CD-ROM.

MESQUITA N.R, **Classificação de defeitos em tubos de gerador de vapor de plantas nucleares utilizando mapas auto – organizáveis,** Tese (Doutorado), Escola Politécnica de engenharia, USP, São Paulo, 2002.

MICROSOFT Project for Windows, version XP: project planning software. [S.I]: Microsoft Corporation, 2001, Conjunto de programas 1 CD - ROM.

MINGOTI A.S, **Análise de dados através de métodos de estatística multivariada: Uma abordagem aplicada,** Ed UFMG, Belo Horizonte, 2005.

MUSTONEN S.M, TISSARI S, HUIKKO L, KOLEHMAINEN M, LEHTOLA M.J, HIRVONEN A, Evaluating online data of water quality charges in a pilot drinking water distribution system with multivariate data exploration methods.**Water Research,** 42(10 – 11), p. 2421 – 2430, 2008.

NETO A.E.P, **Modelos receptores aplicados à determinação da estrutura de fontes de aerossóis remotos.** Tese (Doutorado), Instituto de Física, USP, São Paulo, 1985.

RANSON. S.W, **Anatomia do Sistema Nervoso – sob o ponto de vista de desenvolvimento e função,** ed. 7, cap 4, p. 31 – 44, Renascença s.a, São Paulo, 1945.

REIS T.L.E, **Abordagem sistêmica do sistema de tratamento de água de registro, São Paulo, com ênfase na avaliação de impacto do descarte dos resíduos na bacia hidrográfica do rio Ribeira de Iguape.** Tese (Doutorado), Instituto de Pesquisas Energéticas e Nucleares – IPEN, São Paulo, 2006.

ROSEMBLATT F, The Perceptron: a probabilistic model for information storage & organization in the brain. **Psychological Review,** 65, p. 386 – 408, 1958.

ROSSI. S.P.H.R, **Utilização de redes neurais na monitoração da potência do reator IEA – R1,** (Tese) Doutorado, Instituto de Pesquisas Energéticas e Nucleares – IPEN, São Paulo, 2001.

SOM TOOLBOX versão 2.0, obtida no Laboratory of Computer and Information Science, Finland, Mar. 17 2005, [S.I]. Disponível em: http://www.cis.hut.fi/projects/somtoolbox/ .

TISON J., PARK Y.S., COSTE, M. DELMAS F., GIRAUDEL, J.L.,Use of unsupervised neural networks for eco-regional zonation of hydrosystems through diatom communities: case study of Adour-Garonne watershed**. Archiv für Hydrobiologie,** 159, 409-422, 2004.

TOBISZEWSKI M., TSAKOVSKI S., SIMEONOV V., NAMIÉSNIK J., Surface water quality assessment by the use of combination of multivariate statistical classification and expert information. **Chemosphere**, 80, p.740-746, 2010.

VESANTO, J., SOM implementation in SOM Toolbox, obtido em: VESANTO, J., SOM implementation in SOM Toolbox, obtido em: http://www.cis.hut.fi/somtoolbox/documentation/somalg.shtml, acessado em: 21/01/2009.