



**COPPE/UFRJ**

INVESTIGAÇÃO DE TÉCNICAS EXPLORATÓRIAS DE DADOS APLICADAS A  
QUALIDADE AMBIENTAL EM SISTEMAS DE INFORMAÇÃO GEOGRÁFICA.

Alexandre Tadeu Politano

Dissertação de Mestrado apresentada ao  
Programa de Pós-graduação em Engenharia  
Civil, COPPE, da Universidade Federal do  
Rio de Janeiro, como parte dos requisitos  
necessários à obtenção do título de Mestre em  
Engenharia Civil.

Orientador: Nelson Francisco Favilla Ebecken

Rio de Janeiro

Junho de 2009

INVESTIGAÇÃO DE TÉCNICAS EXPLORATÓRIAS DE DADOS APLICADAS A  
QUALIDADE AMBIENTAL EM SISTEMAS DE INFORMAÇÃO GEOGRÁFICA.

Alexandre Tadeu Politano

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO  
LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA  
(COPPE) DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE  
DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE  
EM CIÊNCIAS EM ENGENHARIA CIVIL.

Aprovada por:

---

Prof. Nelson Francisco Favilla Ebecken, D.Sc.

---

Prof. Alexandre Gonçalves Evsukoff, D.Sc.

---

Cristina Maria Bentz, D.Sc.

---

Prof. Rodolfo Paranhos, D.Sc.

RIO DE JANEIRO, RJ – BRASIL

JUNHO DE 2009

Politano, Alexandre Tadeu

Investigação de Técnicas Exploratórias de Dados Aplicadas a Qualidade Ambiental em Sistemas de Informação Geográfica / Alexandre Tadeu Politano – Rio de Janeiro: UFRJ/COPPE, 2009.

IX, 137 p.: il.; 29,7 cm.

Orientador: Nelson Francisco Favilla Ebecken

Dissertação (mestrado) – UFRJ/ COPPE/ Programa de Engenharia Civil, 2009.

Referencias Bibliográficas: p. 106-112.

1. Sistema de informação geográfica. 2. Classificação Supervisionada. 3. Classificação não-supervisionada. I. Ebecken, Nelson Francisco Favilla. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia Civil. III. Título.

***“Uma vez descobertas, as verdades são  
fáceis de compreender; o importante é  
descobri-las”***

GALILEU GALILEI (1564 – 1642), ITÁLIA

Dedico este trabalho aos  
meus pais, Jose e Sonia.

## *Agradecimentos*

Gostaria de agradecer a todos que de alguma maneira favoreceram a conclusão desse trabalho, especialmente meus Pais, irmãs, sobrinhos e agregados por todo o apoio de vida, a minha esposa por todo o apoio e ajuda, a minha filha pela compreensão nos momentos de privação dessa fase e aos meus amigos pelo entendimento do meu sumiço temporário.

Aos gerentes da AMA Pedro Penido e Viviana Coelho, que viabilizaram administrativamente o trabalho, a todos os colegas da Petrobrás pelo apoio, especialmente Teresinha e Fátima Guadalupe, e em particular a Cristina Bentz, com quem muito aprendi profissionalmente e foi a responsável pela minha aproximação da área de mineração de dados.

Ao Professor Nelson pela paciência e por permitir uma livre condução do tema, a partir dos pontos norteadores. Ao Professor Alexandre com suas aulas esclarecedoras e de rico conteúdo, fundamentais para meu entendimento e motivação profissional.

Ao colega Fabio Moreira pelas inúmeras idéias trocadas no *line-up*, entre uma série e outra na terapia semanal (surf) e pela sua atenção e habilidade no desenvolvimento dos *scripts* em Python, que facilitou muito minha vida na integração do banco aos resultado do classificados *fuzzy*.

Resumo da Dissertação apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

## INVESTIGAÇÃO DE TÉCNICAS EXPLORATÓRIAS DE DADOS APLICADAS A QUALIDADE AMBIENTAL EM SISTEMAS DE INFORMAÇÃO GEOGRÁFICA.

Alexandre Tadeu Politano

Junho/2009

Orientador: Nelson Francisco Favilla Ebecken

Programa: Engenharia Civil

A Petrobras desenvolveu um projeto de pesquisa para caracterizar ambientalmente a Baía de Guanabara. Nesse sentido, com o objetivo de aumentar o potencial de conhecimento e uso das informações do projeto realizou-se uma exploração multivariada e classificatória em parte dos dados coletados, considerando os parâmetros físico-químicos e biológicos da água adquiridos quinzenalmente durante o período de 2 anos (2005 e 2007), em 10 estações amostrais distribuídas na baía. Avaliou-se uma organização para os dados em modelo de base de dados espacial (geodatabase) e realizou-se uma investigação exploratória, espaço-temporal, dos dados e suas possibilidades de agrupamentos a partir de técnicas de classificação supervisionada (árvore de decisão) e não supervisionada (fuzzy). Explorou-se ainda o potencial de contribuição de ferramentas de geoprocessamento como interface de acesso à base de dados e nos processos de análise, integração e apresentação espacial das informações.

Abstract of Dissertation presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

RESEARCH TECHNIQUES IN EXPLORATORY DATA APPLIED  
ENVIRONMENTAL QUALITY IN GEOGRAPHICAL INFORMATION SYSTEMS

Alexandre Tadeu Politano

June/2009

Advisor: Nelson Francisco Favilla Ebecken

Department: Engenharia Civil

Petrobras has developed a research Project to characterize environmentally the Guanabara Bay. With the aim to increase development potential and the use of the project information, there has been a multivariate exploration and qualifying in parts of the collected data, considering physics-chemistry parameters water biological gotten every fifteen days for the period of 2 years (2005 to 2007), in 10 sample stations distributed in the bay. An organization has been evaluated to the data in the geodatabase, and developed an exploratory investigation, temporal space, of the data and their grouping possibilities taken from classification from supervised techniques (decision-tree) and non supervised. We've also explored the tool contribution potential of geoprocessing as the interface access to the database and in the analyses processes integration and information spatial presentation.

## SUMÁRIO

RESUMO .....	vii
ABSTRACT .....	viii
ÍNDICE DE FIGURAS.....	XII
ÍNDICE DE TABELAS.....	XV
ÍNDICE DE FÓRMULAS .....	XVII
ÍNDICE DE GRÁFICOS .....	XVIII
CAPÍTULO 1 - INTRODUÇÃO .....	1
1.1 JUSTIFICATIVA.....	4
1.2. OBJETIVO .....	5
1.2.1. OBJETIVOS ESPECÍFICOS.....	5
CAPÍTULO 2 - FUNDAMENTAÇÃO TEÓRICA .....	6
2.2 GEOPROCESSAMENTO.....	6
2.3 SISTEMA DE INFORMAÇÃO GEOGRÁFICA .....	7
2.4 MODELO DE DADOS DO SIG .....	9
2.5 ESCALA CARTOGRÁFICA .....	11
2.6 SISTEMA DE COORDENADAS.....	12
2.7 DATUM.....	14
2.8 METADADOS.....	14
2.9 REPRESENTAÇÃO TEMÁTICA .....	15
2.10 ANÁLISE ESPACIAL.....	16
2.11 MINERAÇÃO DE DADOS ( <i>DATA MINING</i> ) .....	18
2.12 PROCESSO DE DATA MINING .....	19
2.13 ANÁLISE MULTIVARIADA.....	20
2.14 ANALISE DE COMPONENTES PRINCIPAIS - APC .....	22
2.15 CLASSIFICADORES .....	23
2.16 ÁRVORE DE DECISÃO .....	24
2.17 <i>FUZZY C</i> -MÉDIAS .....	25
2.18 MÉTRICAS DE DESEMPENHO.....	27
2.18.1 CLASSIFICAÇÃO SUPERVISIONADA .....	27
2.18.2 CLASSIFICAÇÃO NÃO-SUPERVISIONADA .....	28
2.18.2.1 PBM.....	28
2.18.2.2 ÍNDICE XIE E BENI .....	29
2.18.2.3 DISCRIMINANTE DE FISHER .....	29

2.18.2.4	ÍNDICE DE CALINSKI-HARABASZ .....	29
2.19	CONSIDERAÇÕES AMBIENTAIS: .....	30
2.19.1	PARÂMETROS FÍSICO-QUÍMICOS.....	30
2.19.1.1	MATERIAL EM SUSPENSÃO .....	30
2.19.1.2	AMÔNIA.....	30
2.19.1.3	OXIGÊNIO DISSOLVIDO.....	31
2.19.1.4	SALINIDADE.....	31
2.19.1.5	NITROGÊNIO TOTAL .....	31
2.19.1.6	FÓSFORO TOTAL.....	32
2.19.1.7	pH.....	32
2.19.1.8	TEMPERATURA DA ÁGUA .....	33
2.19.1.9	PROFUNDIDADE DA ESTAÇÃO.....	33
2.19.2	PARÂMETROS MICROBIOLÓGICOS .....	33
2.19.2.1	CLOROFILA-A .....	33
2.19.2.2	ABUNDÂNCIA BACTERIANA.....	33
CAPÍTULO 3 - ESTUDO DE CASO .....		35
3.1	RESUMO METODOLÓGICO .....	35
3.2	ÁREA DE ESTUDO .....	36
3.3	COLETA DOS DADOS.....	38
3.4	BANCO DE DADOS .....	39
3.5	ANÁLISE EXPLORATÓRIA DOS DADOS .....	43
3.6	MATRIZ DE CORRELAÇÃO .....	56
3.7	ANÁLISE DOS COMPONENTE PRINCIPAIS – ACP .....	63
3.8	CLASSIFICAÇÃO SUPERVISIONADA .....	67
3.9	CLASSIFICAÇÃO NÃO-SUPERVISIONADA .....	82
3.10	MODELOS DOS CLASSIFICADORES (NÃO-SUPERVISIONADA) .....	95
CAPÍTULO 4 - DISCUSSÃO .....		96
CAPÍTULO 5 - CONCLUSÃO .....		103
CAPÍTULO 6 - RECOMENDAÇÕES .....		105
REFERÊNCIAS BIBLIOGRÁFICAS .....		106
ANEXO I.....		113
ANEXO II.....		115
ANEXO III .....		118
ANEXO IV .....		123

ANEXO V .....	126
ANEXO VI.....	129

## ÍNDICE DE FIGURAS

Figura 1: Componentes básicos de um SIG. ....	7
Figura 2: Esquema de representação em camadas do SIG, adaptado de Zeiler (1999)....	8
Figura 3: Simplificação do modelo de dados do SIG, onde cada elemento representado no mapa tem seu registro correspondente na tabela de dados. ....	9
Figura 4: Figura (a) indica que 1 cm do mapa corresponde a 250.000 cm no terreno. Na figura (b) o segmento de medida corresponde a 50.000 cm. ....	11
Figura 5: O mapa da esquerda apresenta o município pela localização da sede (representação de ponto) e o mapa da direita representa o município pelo polígono do limite de sua área, para a mesma escala de visualização. ....	12
Figura 6: Exemplos de projeções cartográficas e alguns de seus efeitos de distorção em uma ou mais propriedades espaciais: forma, área, distância, e direção (ESRI, 1992). ..	13
Figura 7: Exemplo do modelo de análise de sobreposição ponderada, o produto é um mapa índice que considera todas as variáveis de entrada. ....	17
Figura 8: Apresenta o esquema da interpolação espacial. Estão representados: (A) dados pontuais; (B) ação do algoritmo; e (C) modelo de dados contínuo (matricial). ....	17
Figura 9: Apresenta o processo do KDD, incluindo o ciclo virtuoso de evolução do processo (MOTTA, 2005). ....	20
Figura 10: Representa o esquema da arquitetura da árvore de decisão. ....	24
Figura 11: Representação gráfica do entendimento das possíveis respostas geradas pela teoria clássica dos conjuntos (falso ou verdade) “crisp”, comparando com o modelo “fuzzy”, que representa o valor de resposta como sendo a pertinência, ou melhor, uma graduação entre 0 e 1 informando a possibilidade da informação ser mais falsa ou mais verdadeira. ....	25
Figura 12: Esquema dos processos envolvidos até a geração do conhecimento. ....	36
Figura 13: Apresentação geográfica da área de estudo. ....	38
Figura 14: Malha amostral. ....	38
Figura 15: Aba de exibição das descrições dos metadados. ....	41
Figura 16: Aba de exibição dos metadados para as informações referentes ao sistema de coordenada do layer. ....	42
Figura 17: A aba do metadados associada aos atributos da tabela exhibe todas a estrutura do campo, assim como seu alias e sua definição. ....	42
Figura 18: Mapa temático qualitativo com valores médios para TEMP_AG. ....	51

Figura 19: Mapa temático qualitativo com valores médios para SAL_AG. ....	52
Figura 20: Mapa temático qualitativo com valores médios para OD_AG. ....	52
Figura 21: Mapa temático qualitativo com valores médios para FOS_TOT_AG. ....	53
Figura 22: Mapa temático qualitativo com valores médios para AMON_AG. ....	53
Figura 23: Mapa temático qualitativo com valores médios para NIT_TOT_AG. ....	54
Figura 24: Mapa temático qualitativo com valores médios para MPS. ....	55
Figura 25: Mapa temático qualitativo com valores médios para CLOR_A. ....	55
Figura 26: Mapa temático qualitativo com valores médios para AB_BAC. ....	56
Figura 27: Representação das cinco áreas de qualidade ambiental proposta por Mayer <i>et al.</i> (1989) e adotada no presente estudo como conhecimento <i>a priori</i> . ....	68
Figura 28: Posicionamento das estações de coleta em relação ao mapa de classes de qualidade da água (Mayr <i>et al.</i> , 1989), e a distância de cada estação a borda mais próxima da classe do mapa. ....	69
Figura 29: Mapa apresentando o resultado da análise espacial com a distribuição .....	69
Figura 30: Representação das estações de coleta com os valores obtidos na matriz de confiança de classificação; para comparação visual estão sobrepostos as áreas definidas por Mayr <i>et al.</i> , 1989. ....	79
Figura 31: Modelo de árvore de decisão criado pelo classificador. ....	80
Figura 32: Modelo de classificação ponderado com definição das classes por intervalos iguais; as barras representam a soma das ocorrências para cada estação. ....	85
Figura 33: Modelo de classificação ponderado com definição das classes por quebras naturais. ....	85
Figura 34: Modelo de classificação ponderado com definição das classes por intervalos iguais, considerando somente seis variáveis. ....	86
Figura 35: Modelo de classificação ponderado com definição das classes por quebras naturais, considerando somente seis variáveis. ....	87
Figura 36: Visualização de parte da tabela do banco obtida pela análise. ....	92
Figura 37: Representação da variabilidade espaço-temporal do classificador fuzzy. ....	93
Figura 38: Mapa com a representação dos resultados do modelo fuzzy e sobreposição ponderada em um mesmo tempo (campanha 05). As barras gráficas do mapa do modelo <i>fuzzy</i> estão representando o valor de pertinência da estação para cada classe e os pontos representam a classificação do modelo de superposição; as cores são correspondentes as classes sinalizando o comprometimento ambiental. ....	94
Figura 39: Representação esquemática do modelo de classificação ponderado. ....	95

Figura 40: Representação esquemática do modelo de classificação fuzzy. ....	95
Figura 41: Representação da variabilidade de classificação não supervisionada das estações para o método de sobreposição ponderada.....	100
Figura 42: Consolidação do resultado do classificador não supervisionado fuzzy ( dados somente de superfície considerando as 6 parâmetros definidos como prioritários e para as coletas de superfície. ....	101

## ÍNDICE DE TABELAS

Tabela 1: Apresenta um critério para a validação do modelo .....	28
Tabela 2: Apresenta as áreas de atuação do projeto .....	37
Tabela 3: Relação das estações com a profundidade média.....	39
Tabela 4: Valores da estatística básica das variáveis. ....	46
Tabela 5: Apresenta a estatística básica já considerando a substituição .....	47
Tabela 6: Relação campanha, época do ano, mês e ano da coleta.....	49
Tabela 7: Matriz de correlação considerando todas as variáveis da base de dados. Correlação significativa para $p < ,05000$ .....	58
Tabela 8: Apresenta a matriz de correlação considerando somente os dados de coleta na superfície, no período de seca. Correlação significativa para $p < ,05000$ .....	59
Tabela 9: Apresenta a matriz de correlação considerando somente os dados de coleta na superfície, no período de chuva. Correlação significativa para $p < ,05000$ . ....	60
Tabela 10: Apresenta a matriz de correlação considerando somente os dados de coleta de fundo para o período de seca. Correlação significativa para $p < ,05000$ .....	61
Tabela 11: Apresenta a matriz de correlação considerando somente os dados de coleta de fundo para o período de chuva. Correlação significativa para $p < ,05000$ .....	62
Tabela 12: Interpretação para as faixas de valores obtidas em $p$ pela matriz de correlação.....	63
Tabela 13: Apresenta os valores para as componentes principais.....	66
Tabela 14: Apresenta a sumarização considerando o número de estações e a quantidade de medições para cada classe.....	70
Tabela 15: Variáveis e sua codificação no ambiente da classificação supervisionada. .	76
Tabela 16: Relação entre as classes do conhecimento a priori e adotadas pelo classificador (associação estabelecida por análise espacial). ....	76
Tabela 17: Matriz de confusão criada a partir do resultado do classificador. ....	79
Tabela 18: Apresenta os valores das métricas de validação utilizada, em verde os índices que destacam a eficiência do agrupamento. ....	82
Tabela 19: As faixas de valores estabelecidas pelo método de quebras naturais, consideradas como base das regras do modelo classificador não-supervisionado;.....	87
Tabela 20: Apresenta as classes utilizadas para definir as regras de partição fuzzy, em negrito as variáveis consideradas pelo classificador fuzzy. ....	88
Tabela 21: Valor dos parâmetros da estação BG-34, campanha 05 (superfície).....	93

Tabela 22: Resultado do valor de pertinência de cada classe do modelo fuzzy.....	93
Tabela 23: Consolidação espaço-temporal da avaliação dos classificadores utilizados; as cores da tabela fazem relação com a classe que ocorreu com maior frequência considerando todos os classificadores não supervisionado avaliados. ....	96

## ÍNDICE DE FÓRMULAS

Fórmula 1: Cálculos dos centros dos agrupamentos (classes).....	26
Fórmula 2: Cálculo do valor da pertinência .....	26
Fórmula 3: Índice PBM. ....	28
Fórmula 4: Índice Xie e Beni. ....	29
Fórmula 5: Índice de Calinski-Harabasz .....	30
Fórmula 6: Estrutura da fórmula de cálculo da sobreposição ponderada.....	84

## ÍNDICE DE GRÁFICOS

Gráfico 1: Gráfico de Pareto, que tem por finalidade obter melhor visualização quando se necessita priorizar diversos itens, descrevendo a contribuição de cada fator de correlação entre as variáveis.....	64
Gráfico 2: Correlação entre as variáveis. Observa-se a grande alternância nas barras sugerindo baixa correlação entre as variáveis estudadas.....	65
Gráfico 3: Descreve as informações coletadas para parâmetro de temperatura da água na superfície agrupado pelas classes estabelecidas pelo projeto referência.....	71
Gráfico 4: Histograma gerado a partir dos dados de OD_AG.....	71
Gráfico 5: Histograma gerado a partir dos dados de SAL_AG.....	72
Gráfico 6: Histograma apresentado os dados de FOS_TOT.....	72
Gráfico 7: Gráfico de histograma gerado para os dados do parâmetro pH agrupados pelas classes de água.....	73
Gráfico 8: Histograma da frequência de ocorrência do parâmetro de AMON_AG.....	73
Gráfico 9: Gráfico da distribuição dos valores para NIT_TOT_AG por frequência de ocorrência.....	74
Gráfico 10: Histograma de frequência dos valores considerados para o parâmetro MPS.....	74
Gráfico 11: Valores do parâmetro de CLOR_A apresentados por gráfico de histograma gerado para as coletas de superfície.....	75
Gráfico 12: Histograma gerado a partir dos dados de AB_BAC para as coletas de superfície.....	75
Gráfico 13: Distribuição das classes para a seleção de dados utilizada.....	77
Gráfico 14: Resultado obtido a partir do teste do algoritmo de classificação.....	77
Gráfico 15: Apresenta a área da curva ROC – Critério de validação do modelo.....	78
Gráfico 16: AUC calculada para cada classe.....	78
Gráfico 17: Partição <i>fuzzy</i> referente às classes de importância.....	89
Gráfico 18: Partição <i>fuzzy</i> para a variável de salinidade da água.....	89
Gráfico 19: Classe de partição <i>fuzzy</i> utilizada no modelo de.....	90
Gráfico 20: Partições <i>fuzzy</i> para a variável de clorofila-a para as coletas de água superficial.....	90
Gráfico 21: Partição <i>fuzzy</i> para a variável de nitrogênio total.....	90

Gráfico 22: Pertinência associadas as classes de importância ambiental, partição *fuzzy* para a variável de abundancia bacteriana. .... 91

# CAPÍTULO 1

## INTRODUÇÃO

Em muitos casos, o desenvolvimento das grandes cidades ocorre em regiões onde o meio ambiente torna-se fragilizado frente à magnitude das ações antrópicas. A Baía de Guanabara é um desses ambientes complexos, encurralado pelo avanço da capital do Estado do Rio de Janeiro e de seus municípios adjacentes, sendo este um cenário de grande importância para o desenvolvimento da indústria do petróleo nacional.

Atualmente esta indústria aporta grande quantidade de recursos financeiros em projetos ambientais.

A Baía de Guanabara figura na história do país desde à época do Brasil Colônia, sendo adjacente ao Município do Rio de Janeiro, capital da república por três séculos (1889 a 1960); assim, tornou-se um grande pólo de desenvolvimento tendo em seu entorno a presença de outros municípios, como Niterói, São Gonçalo, Itaboraí, Guapimirim, Magé e Duque de Caxias. É a segunda maior baía do litoral brasileiro com uma área aproximada de 380 km<sup>2</sup> de extensão. Sua geografia abrigada favorece o desenvolvimento econômico, suportando uma infra-estrutura de terminais portuários com intenso tráfego de embarcações. Somado a importância para o desenvolvimento nacional, a baía é um ambiente sensível por natureza com uma ampla bacia de drenagem e a presença de algumas áreas de remanescentes manguezais (BARROS, 2002, GARRIDO *et al.*, 1978, MAYR *et al.*, 1989).

A importância ambiental da baía é reforçada pela presença de uma APA (Área de Proteção Ambiental), e segundo o critério de avaliação do uso das águas adotado pela FEEMA (Fundação Estadual de Engenharia de Meio Ambiente) a baía é classificada como área de águas salinas, que deve ter seu uso destinado à recreação, lazer, turismo, navegação, atividades portuárias, abastecimento industrial, amenização ambiental, estética, conservação da biodiversidade e pesca artesanal e comercial (FEEMA, 1990, FEEMA, 1998, MAYR *et al.*, 1989).

Contudo, a baía vem sendo desgastada ao longo do tempo pelo seu uso insustentável. As principais causas apontadas pela sua degradação ambiental são: aterros e assoreamento; destruição de manguezais; poluição industrial, esgoto e

acidentes ambientais. Esse cenário, somado à importância socioeconômica da região, estimula o desenvolvimento de muitos projetos de pesquisa com financiamentos públicos e/ou privados, abordando os mais diversos temas de relevância para a região tais como: biológicos, físico-químicos, geológicos, socioambiental, caracterização ambiental, recuperação ambiental, entres outros. Todos os estudos geram conhecimento e dados para um melhor entendimento deste recurso natural (BARROS, 2002, VILLAC *et al.*, 1991).

Frente aos desafios encontrados no campo da conservação e no manejo de ecossistemas, diversos segmentos do mercado que atuam na área ambiental vêm fazendo uso dos avanços tecnológicos para organizar e armazenar as informações do meio ambiente. O SIG – Sistema de Informação Geográfica (ou GIS – Geography Information System) consiste em um ambiente tecnológico com a capacidade de armazenar e processar informações com características espacial, representadas por elementos gráficos (localização) e alfanuméricos (atributos) integrados. Assim, um SIG pode ser visto como a combinação de hardware, software, dados, metodologias e recursos humanos que operam de forma harmônica para armazenar, organizar, tratar, analisar e publicar informação geográfica associada a bancos de dados geográficos (NEIA, 1997, FIGUEIRA, 1999). A habilidade do SIG aplicada em dados de meio ambiente constitui um enorme potencial para a comunidade de uma forma geral. A ciência ambiental necessita aprimorar seus mecanismos de gestão de dados para permitir análises nas dimensões espaços-temporais de forma rápida e acessível. A investigação entre pontos amostrais de variáveis abióticas e bióticas pode auxiliar a compreensão das variações espaço-temporais. Esta compreensão possibilita um melhor planejamento e direcionamento na execução das atividades de gestão ambiental (FONSECA, 2003).

Assim essa ferramenta pode prestar fundamental apoio aos processos vinculados aos estudos ambientais, nas tarefas de organização dos dados, integração, análises espaciais e na apresentação dos resultados.

A utilização de algoritmos matemáticos relacionados à mineração de dados pode revelar informações estratégicas com relevante valor científico, podendo descrever características do passado, assim como prever tendências para o futuro (GIMENEZ, 2000). As técnicas de mineração de dados podem ser definidas como o processo de

análise de conjuntos de dados cujo objetivo é a descoberta de padrões de interesse (SOUZA, 2003) auxiliando o entendimento dos desequilíbrios espaços-temporais.

Neste contexto, o trabalho apresenta uma abordagem investigativa do comportamento espaço-temporal, aplicada em ambiente multivariado sobre os valores de parâmetros físico-químicos da água. Este trabalho avaliou duas abordagens classificatórias: uma supervisionada, baseada em um conhecimento a priori, e outra não-supervisionada, através de métodos de particionamento/agrupamento utilizando-se critérios multivariados aplicados sobre os parâmetros estudados.

## 1.1 Justificativa

É importante citar o grande volume de recursos que são investidos nos processos de coleta e análise de parâmetros ambientais, com custos associados ao planejamento e realização das campanhas de coleta (embarcações e equipamento), às análises laboratoriais das amostras e a toda mão de obra especializada, envolvida em todas as fases do processo, para a geração de um conhecimento único para aquele estado de tempo. Porém, muitas vezes essas informações podem armazenar “segredos” revelados somente com o suporte de uma investigação matemática mais apurada ao longo do tempo e do espaço geográfico. Nessa linha, este trabalho pretende contribuir buscando extrair outros conhecimentos da base de dados ambiental, potencializando seu uso na gestão ambiental. Também propõe validar o uso de boas práticas de armazenamento dos dados em base de dados espacial.

No Brasil, a determinação de padrões de qualidade ambiental é ainda baseada no conceito de níveis de concentração máxima admissíveis de poluentes de acordo com o Conselho Nacional de Meio Ambiente (CONAMA). Entretanto, atualmente verifica-se que muitos órgãos ambientais internacionais (USEPA, EEA, JEMAI, entre outras) adotam critérios biológicos como indicadores de qualidade ambiental, estabelecendo faixas de valores a limites graduados de tolerância. Este conceito é bem conhecido no âmbito da ecologia, encontrando-se mais próximo da realidade no que se refere à detecção dos diversos distúrbios aos quais a biota é submetida. Na determinação da qualidade ambiental, é importante distinguir a variação natural daquela variabilidade causada por impactos provocados pelo homem. Desta maneira, faz-se necessário uma profunda inspeção em dados de monitoramento ambiental de longo prazo, com esforços no sentido de relacionar as variáveis ambientais às biológicas, estabelecendo critérios naturais para sua variação e permitindo a melhor compreensão da interferência do homem no ciclo do meio ambiente.

## **1.2. Objetivo**

O objetivo deste trabalho é investigar o potencial de técnicas computacionais (mineração de dados) associadas a problemas de classificação para aumentar a capacidade de gerar conhecimento em base de dados ambiental com características espaço-temporais.

### **1.2.1. Objetivos Específicos**

- Validar o modelo espacial utilizado para armazenamento dos dados;
- Avaliar o potencial de integração da base de dados espacial entre aplicativos de tratamento/análises dos dados;
- Avaliar o comportamento dos dados em um modelo de classificação supervisionada gerado a partir de informações de um conhecimento *a priori*;
- Testar as possibilidades de agrupamento implícitas (classificação não-supervisionada) aos dados;
- Avaliar o potencial de comunicação do conhecimento gerado a partir de mapas temáticos.

# CAPÍTULO 2

## FUNDAMENTAÇÃO TEÓRICA

### 2.2 Geoprocessamento

A integração entre tecnologia espacial e dados ambientais permite agregar valor aos processos de geração de conhecimento em estudos de ecossistemas, principalmente porque, em sua maioria, as questões de meio-ambiente envolvem situações nas quais o “ONDE” é fundamental para o entendimento dos processos. Este entendimento abre uma oportunidade para considerar a adoção de um SIG – Sistema de Informação Geográfica, e o geoprocessamento como ferramenta de trabalho (CÂMARA *et al.*, 2004).

A sociedade organizada gera informação a partir de seus recursos minerais, propriedades, animais, plantas entre outros organizados no espaço geográfico. No entanto, em um passado recente, estas informações eram representadas apenas em documentos e mapas em papel, dificultando os processos de análise que combinassem diversos mapas e dados. Com a evolução tecnológica, tornou-se possível armazenar e representar tais informações em ambiente computacional, possibilitando o crescimento do geoprocessamento (CÂMARA *et al.*, 1998).

De acordo com Rodrigues (1993), geoprocessamento é um conjunto de tecnologias de coleta, tratamento, manipulação e apresentação de informações espaciais voltada para um objetivo específico. Esse conceito é a base para a estrutura do SIG – Sistema de Informação Geográfica.

Em estudos ambientais, os ecossistemas são apresentados como sistemas dinâmicos não lineares com variantes ao longo do tempo e do espaço, sendo, portanto, de extrema complexidade. Assim, o armazenamento estruturado dos dados facilita o entendimento das variações espaço-temporais, essenciais para a o monitoramento ambiental (FIGUEIREDO *et al* 1998).

Segundo Maidment (1992), a área de meio ambiente e recursos hídricos tem intensificado o uso do geoprocessamento, pois esta tecnologia pode auxiliar na identificação da variabilidade espacial das características de corpos hídricos.

### 2.3 Sistema de Informação Geográfica

Um sistema de informação geográfica é uma coleção organizada de recursos de hardware, software, dados geográficos, processos e pessoas (Figura 1), desenvolvida para capturar, armazenar, atualizar, analisar e exibir todas as formas de informação geograficamente referenciada, auxiliando na tomada de decisão dos problemas do mundo real. Por se tratar de um sistema computadorizado associado a coordenadas no espaço, é capaz de suportar e utilizar dados descrevendo regiões da superfície terrestre (ESRI, 1992). O SIG é um sistema desenvolvido a partir de tecnologias de geoprocessamento, promovendo o uso de informação na forma de dados geográficos.

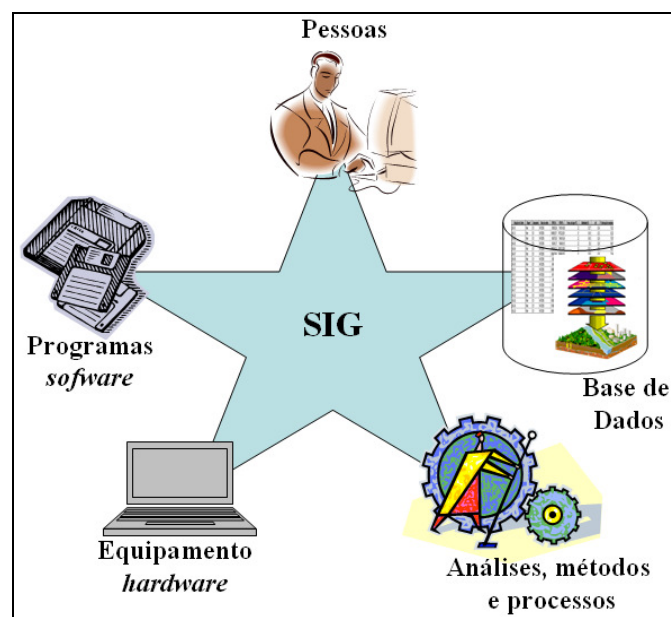


Figura 1: Componentes básicos de um SIG.

A estrutura de armazenamento de dados do SIG descreve o mundo real em níveis de informação (camadas de informação ou temas) com representações geométricas em um espaço geográfico (Figura 2). Cada nível de informação é

representado por um tipo de forma geométrica (ponto, linha, polígono ou pixel) georeferenciando um conjunto de “objetos” do mundo real (ZEILER, 1999).

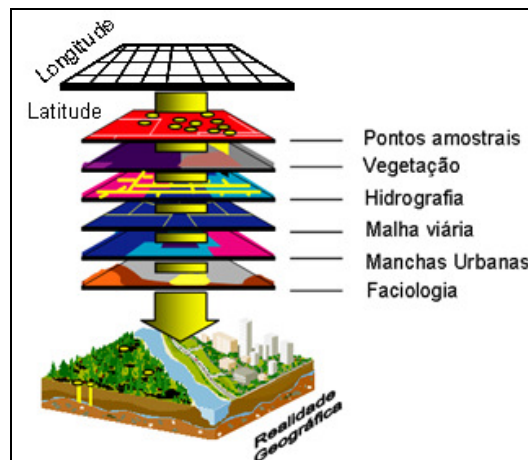


Figura 2: Esquema de representação em camadas do SIG, adaptado de Zeiler (1999).

O SIG entende os dados como relacionamento entre objetos ordenados por camadas, onde cada camada possui a sua característica própria de representação espacial (topologia) e tabela associada (Figura 3). Cada feição no mapa representa um elemento do mundo real descrito por um conjunto de atributos organizado em tabela.

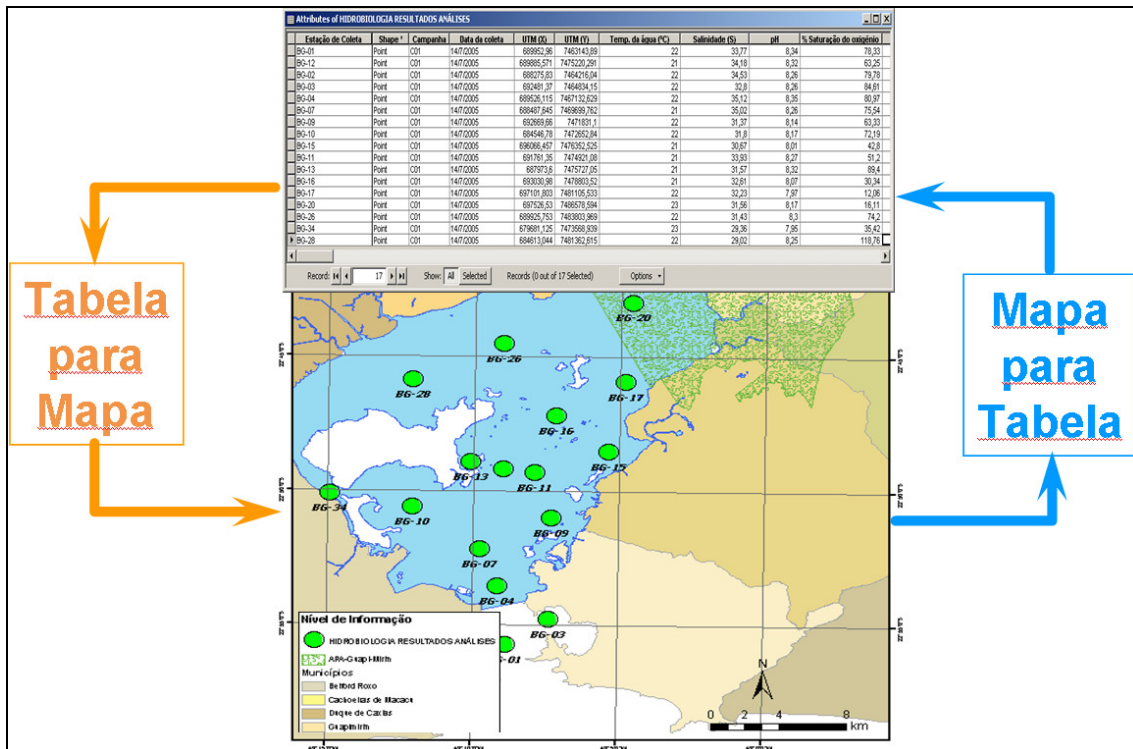


Figura 3: Simplificação do modelo de dados do SIG, onde cada elemento representado no mapa tem seu registro correspondente na tabela de dados.

A ferramenta básica do SIG, baseada no seu conceito, é a interação do usuário com a representação espacial do mapa. Desta forma, uma pesquisa na base tabular retorna, destacando no mapa, a feição espacial correspondente e a operação inversa, ou seja, a seleção de feições no mapa retorna o respectivo registro na tabela de dados.

## 2.4 Modelo de Dados do SIG

Os Modelos de dados do SIG estão relacionados à representação de objetos geográficos e suas inter-relações digitais. Um modelo de dados é um retrato lógico de partes selecionadas do mundo real. A modelagem dos dados espaciais configura-se como o processo de abstração e representação do mundo real em um sistema de computador (ZEILER, 1999).

Para o modelo de dados do SIG, os objetos representados nos mapas, tanto os naturais quanto os produzidos pelo homem, são chamados de feições de mapa, ou simplesmente feições. Cada feição tem uma localização geográfica, uma topologia, um

símbolo que descreve sua característica e um registro que corresponde a uma linha em sua tabela de atributos.

O modelo de representação espacial está dividido basicamente em formas geométricas, como ponto, linha, polígono e pixel:

- O modelo topológico de pontos representa qualquer elemento geográfico associado a um par de coordenadas X e Y; exemplo: ponto de coleta, poste de energia, entre outros. Esta representação permite que o banco de dados armazene as coordenadas de cada feição.
- A representação de segmentos é dada pela feição topológica de linhas armazenando no banco o conjunto de pares de coordenadas X e Y de todos os vértices da reta; exemplo: rios, rodovias, etc. Um atributo padrão do banco para esse formato é o valor do comprimento da linha.
- A topologia de polígono está associada à representação de área e armazena as coordenadas de cada vértice de seus segmentos, de tal forma que o vértice inicial deve obrigatoriamente estar posicionado sobre o vértice final, “fechando” a geometria; exemplo: divisão política administrativa, lotes, etc. As informações de área e perímetro são atributos padrão para esse formato de representação espacial.
- A representação por pixel está relacionada à informação do tipo *raster* ou matricial (imagens ou grandes numéricas), onde a informação é representada/armazenada por um pixel, que é a menor unidade de representação da matriz. Cada pixel pode ter seu valor digital associado a um valor único ou a uma combinação; exemplo: imagem de satélite de sensor ótico (combinação de bandas para a representação em cores “reais”) e modelo digital do terreno (nesse caso o pixel corresponde ao valor de elevação do terreno sob aquela área representada).

O modelo de dados do SIG permite desenvolver regras de negócio associadas a regras topológicas, onde o relacionamento espacial entre as feições de mapa, tais como conectividade, estar contido, adjacência, é levado em conta no modelo.

O detalhe da representação espacial do modelo de dados do SIG está associado à escala de levantamento do dado, que deve considerar o objetivo da análise, considerando que a representação está diretamente relacionada as regras de negocio entre as feições de mapa

## 2.5 Escala Cartográfica

A escala é a relação entre as dimensões dos elementos representados em um mapa e a sua grandeza correspondente medida sobre a superfície da Terra. Toda representação tem uma relação de tamanho (proporção) com o objeto representado (Figura 4). Assim, a representação da superfície terrestre sob a forma de carta/mapa deve ser bastante reduzida, dentro de certa proporção (ESRI, 1992).

Um mapa pode ser milhares ou até milhões de vezes menor que o lugar representado. Com um simples olhar, não há como sabermos a proporção com que o mapa foi desenhado. Por isso a necessidade de adotar um padrão de uso da escala. Escalas maiores são usadas para trabalhos com maior detalhe, como mapeamento, estudos geotécnicos e demais estudos de precisão. Quanto maior o valor da escala cartográfica, menor a quantidade de detalhes apresentado pelas feições do mapa (ESRI, 1992).

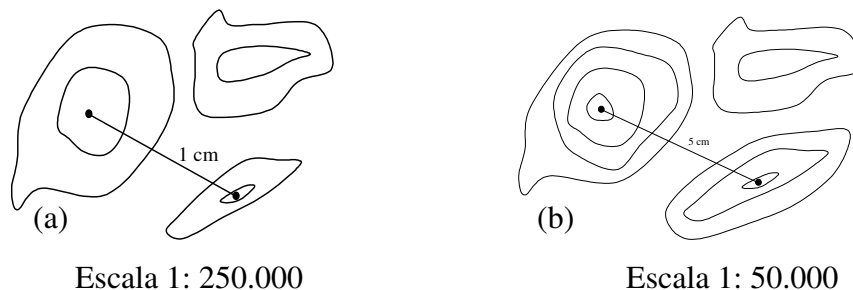


Figura 4: Figura (a) indica que 1 cm do mapa corresponde a 250.000 cm no terreno. Na figura (b) o segmento de medida corresponde a 50.000 cm.

Um município pode ser representado no mapa por uma feição de ponto a partir da localização da sua sede municipal para estudos de pequena escala (poucos detalhes); já em estudos de maior detalhe, esse mesmo município pode ser representado por uma feição de polígono com a representação do limite da sua área (Figura 5). A escala de levantamento do dado é definida pelo propósito do estudo. Em estudos de integração

espacial, a compatibilização entre as escalas das feições deve receber atenção especial para manter a qualidade do produto apresentado.

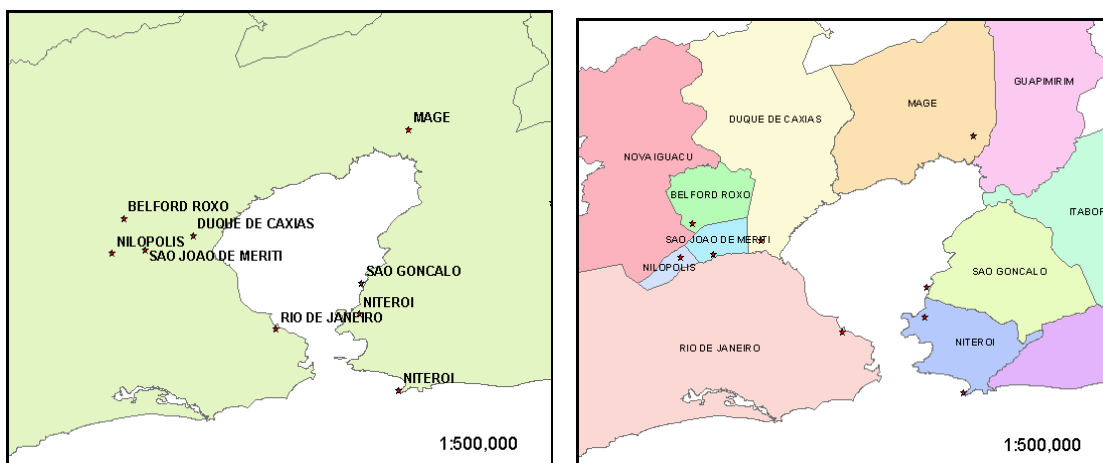


Figura 5: O mapa da esquerda apresenta o município pela localização da sede (representação de ponto) e o mapa da direita representa o município pelo polígono do limite de sua área, para a mesma escala de visualização.

As ferramentas de navegação espacial disponível no SIG devem ser utilizadas com critérios para compatibilizar a escala da base do dado com a escala de visualização.

## 2.6 Sistema de Coordenadas

Considerando que todo SIG é georeferenciado, a posição dos objetos na superfície da terra é representada por valores conhecidos, considerando a unidade geográfica ou projetada.

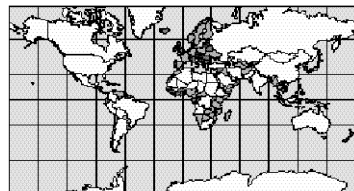
O sistema geográfico representa a superfície em forma de “esfera”, onde os valores são apresentados em graus de latitude e longitude, também conhecidas como coordenadas geográficas. A latitude tem o valor zero no Equador e aumenta até 90 graus para o norte e 90 graus para o sul até os pólos. A longitude tem seu valor de zero grau no Meridiano Principal (Greenwich) e aumenta até 180 graus a leste e 180 graus a oeste até a Linha Internacional de Mudança de Data. Os graus são subdivididos em graus, minutos, e segundos.

O sistema projetado localiza as feições no mapa com medidas de coordenadas planas bidimensional. Um sistema de coordenadas planas descreve a distância de uma

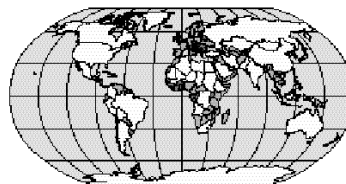
origem (0,0) ao longo de dois eixos separados, um eixo horizontal x, que representa a direção leste-oeste, e um eixo vertical y, representando a direção norte-sul. Pelo fato da terra ser “esférica” e mapas serem apresentados em planos, a conversão de locais da superfície curva para a plana requer uma fórmula matemática chamada de projeção cartográfica. Este processo de aplinar a superfície curva gera distorção na forma, área, distância, e direção.

Felizmente, há muitas projeções cartográficas diferentes. Elas são distinguidos pela sua adequabilidade em representar uma porção particular da superfície da terra, e pela habilidade de preservar distância, área, forma, ou direção (IBGE, 2009). Algumas projeções cartográficas minimizam a distorção em uma propriedade à custa de outra, enquanto outras se esforçam para minimizar a distorção global (figura 6).

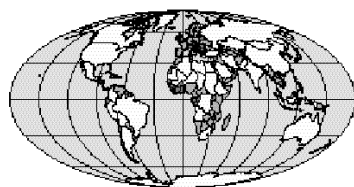
Na atualidade, os aplicativos de geoprocessamento tratam a complexa matemática das projeções cartográficas de forma bastante amigável para o usuário, o que não dispensa o conhecimento do assunto no processo de modelagem espacial da base de dados.



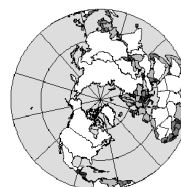
Mercator



Robinson



Mollweide



Azimuthal

Figura 6: Exemplos de projeções cartográficas e alguns de seus efeitos de distorção em uma ou mais propriedades espaciais: forma, área, distância, e direção (ESRI, 1992).

## 2.7 Datum

Outro parâmetro importante no georeferenciamento é o DATUM, que se refere ao modelo matemático teórico da representação da superfície da Terra ao nível do mar. Esse parâmetro corrige ou ajusta o algoritmo que representa melhor a superfície da terra. O DATUM tem como “objetivo” definir um quadro de referência para medir localizações sobre a superfície da terra no plano horizontal.

## 2.8 Metadados

Parte de grande importância do modelo de dados, os metadados, ou metainformação, são dados sobre outros dados, um conjunto de informações organizadas que descrevem o dado de forma descritiva (contextual). Os metadados armazenam outras informações sobre os dados além das já disponíveis em sua tabela de atributos, com o objetivo de facilitar o entendimento dos relacionamentos e a utilidade das informações dos dados.

O papel dos metadados é de suporte no acesso aos dados garantindo a:

- Rastreabilidade: dados necessários para identificar a origem e os processos envolvidos na geração das informações/dados;
- Compatibilidade de uso: dados necessários para determinar se um conjunto de dados se enquadra em determinado fim;
- Acesso: dados necessários para que se adquira um conjunto de dados identificados;
- Transferência: dados necessários para processar e usar um conjunto de dados (ISA, 2001).

Existem alguns padrões e modelos de metadados adotados mundialmente como ISO (IEC 11179) e FGDC - Federal Geographic Data Committee.

## 2.9 Representação Temática

A representação temática está associada à apresentação cartográfica, constituindo um processo móvel que se completa na visualização espacial dos elementos que compõem a realidade (IBGE, 2009).

O SIG, por intermédio de inúmeros aplicativos disponíveis no mercado, permite uma ampla possibilidade de criação temática a partir dos métodos de apresentação e biblioteca de símbolos variados (ESRI, 1992). As representações temáticas devem respeitar o usuário final do mapa, fornecendo um claro entendimento da informação representada. O mapa deve “falar” por si próprio, ou seja, todas as informações necessárias para seu entendimento devem estar representadas.

A utilização de convenções ou padrões de símbolos cartográficos bem definidos auxilia o entendimento da mensagem a ser transmitida pelo mapa, condicionando o usuário ao rápido entendimento da informação.

A criação temática está associada a um nível de informação, as características da simbologia e apresentação do dado são diretamente ligadas ao modelo de dados (topologia + tabela) podendo apresentar variações para as diferentes topologias.

A informação temática pode ser dividida em duas classes, conforme a propriedade de armazenamento da informação associada: classificação temática quantitativa, quando a informação representada é associada a um campo numérico da tabela, e qualitativa (descritiva), quando a representação for associada a um campo da tabela do tipo texto (*string*).

Alguns métodos de representação temática estão listados abaixo:

- Intervalos (Ranges) – o mapa é apresentado a partir da criação de intervalos de valores de um campo numérico;
- Gráfico de Barras (Bar Charts) – criação de gráficos de barras em cada elemento gráfico de acordo com o(s) valor(s) de um ou mais campos da tabela ou mapa;
- Gráfico Graduado (Pie Charts) – criação de gráfico tipo pizza em cada elemento gráfico de acordo com o(s) valor(s) de um ou mais campos da tabela ou mapa;
- Símbolos Graduados (Graduated) – representação dos símbolos graduados (tamanho) de acordo com o valor de cada campo escolhido. Quanto maior o símbolo, maior o valor representado;

- Densidade de Pontos (Dot Density) – pontos distribuídos aleatoriamente, representando valores dos dados;
- Valores Individuais (Individual) – uso de representações diferentes para cada dado escolhido. São muito utilizados para dados qualitativos (ESRI, 1992).

## **2.10 Análise Espacial**

Quando o posicionamento da feição e a sua relação com outra(s) feição(s) são de interesse do estudo, o SIG disponibiliza um conjunto de ferramentas para auxiliar essas questões. Em sua maioria, os aplicativos desenvolvidos em SIG trazem consigo um pacote mínimo de recursos que permite explorar o banco de dados espacial. Entre as possibilidades mais convencionais dessas ferramentas destacam-se: pesquisa direta ao dado espacial; pesquisa lógica na tabela; cálculo de medidas (áreas, perímetros, comprimento); interações a partir de condições topológicas associadas a distâncias, conectividades, superposição e sobreposição entre feições.

Em sua maioria, as análises geradas pelos processos descritos anteriormente já são bastante conclusivas. As ferramentas desenvolvidas no SIG têm sua finalidade na aplicação atendendo a necessidade do sistema. O uso combinado e orientado de operações de análise espacial pode contribuir e gerar resposta para problemas complexos, como no caso de estudos de recursos naturais.

O modelo de dados do SIG suporta operação espacial entre os níveis de informação; essa integração entre camadas considera o posicionamento de cada feição e aceita critério de classe multivariado aplicado a cada feição, conhecida como “sobreposição ponderada” (ESRI, 2002). O termo ponderado deve ser entendido como equilíbrio.

Essa operação de integração espacial consiste em sobrepor classes de informação, apresentada segundo um critério de importância entre as camadas temáticas (Figura 7). Cada camada pode ter um fator de importância de contexto, sendo o peso assimilado pelo modelo de análise.

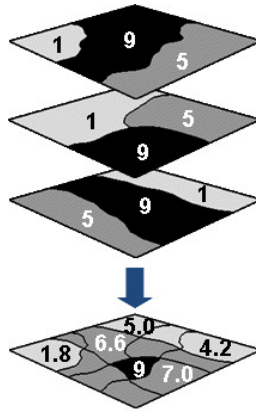


Figura 7: Exemplo do modelo de análise de sobreposição ponderada, o produto é um mapa índice que considera todas as variáveis de entrada.

Outra ciência associada ao SIG é a Geoestatística, que pode ser definida como a associação de uma classe de técnicas usadas para analisar e inferir valores de uma variável distribuída no espaço e/ou no tempo. Tais valores são implicitamente assumidos por correlação dos dados de entrada (CAMARGO, 1999). A geoestatística tem por objetivo a caracterização da dispersão espaço-temporal das grandezas que definem a quantidade e a qualidade de recursos naturais, ou outros fenômenos espaciais em que os atributos apresentem certa estrutura no espaço ou no tempo (CAMARGO, 2007).

Um dos produtos da geoestatística é a geração de superfície contínua de dados. Esse desdobramento se dá a partir de algoritmo de interpolação espacial. Considerando informações pontuais (dados discretos) distribuídas sobre uma determinada área geográfica, o algoritmo relaciona o ponto com sua vizinhança e estima/prediz/associa valor à área onde inicialmente não existia dado espacial (Figura 8).

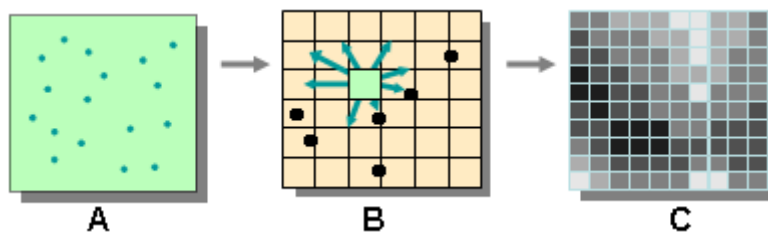


Figura 8: Apresenta o esquema da interpolação espacial. Estão representados: (A) dados pontuais; (B) ação do algoritmo; e (C) modelo de dados contínuo (matricial).

Independente da técnica utilizada para analisar o dado espacialmente, o potencial do SIG está na facilidade em apresentar um problema local e entender o cenário global da área de estudo (“*pense global e aja local*”), orientando às tomadas de decisão.

## **2.11 Mineração de Dados (*Data Mining*)**

A mineração de dados (*Data Mining*) constitui um campo de pesquisa recente em inteligência artificial, cujo objetivo é extrair conhecimento de grandes bases de dados. Fayyad et al (1996) definiu o termo como sendo "...o processo não-trivial de identificar, em dados, padrões válidos, novos, potencialmente úteis e ultimamente compreensíveis" (NAVEGA, 2002).

A mineração de dados vem sendo uma das novidades da ciência da computação. Com a geração de um volume cada vez maior de informação, é essencial tentar aproveitar o máximo possível desse investimento. Esse é um processo analítico projetado para explorar grandes quantidades de dados (tipicamente relacionados a negócios, mercado ou pesquisas científicas), na busca de padrões consistentes e/ou relacionamentos sistemáticos entre variáveis e, então, validá-los aplicando os padrões detectados a novos subconjuntos de dados (PUC-RIO, 2009).

A tarefa de minerar dados pode ser separada em duas grandes linhas de atuação: problemas associados a atividade preditiva ou problemas associados a atividade descritiva (MOTTA, 2005). As atividades preditivas (ou supervisionadas) buscam identificar a classe de uma nova amostra de dados (tendências futuras), a partir do conhecimento adquirido de um conjunto de amostras com classes conhecidas. Estas ações suportam problemas de classificação ou regressão de dados (MOTTA, 2005). Já as atividades descritivas (ou não-supervisionadas) trabalham com um conjunto de dados que não possuem uma classe determinada, buscando identificar padrões de comportamento comuns nestes dados, suportando problemas de sumarização, regras de associação e agrupamentos.

Qualquer que seja a técnica de mineração de dados por classificação supervisionada, ela se utiliza de dados sobre o passado (conjunto de treinamento) para classificar dados futuros (conjunto de execução de um modelo de classificação, ou simplesmente conjunto de execução), ou seja, dados que não pertencem ao conjunto de

treinamento — em geral coletados cronologicamente após os dados de treinamento (BENITEZ, 2001).

Dentre os algoritmos de mineração, pode-se destacar alguns de uso mais comum, como: Algoritmos Estatísticos, Algoritmos Genéticos, Árvores de Decisão, Regras de Decisão, Redes Neurais Artificiais, Algoritmos de Agrupamento e Lógica Fuzzy.

## **2.12 Processo de Data Mining**

Mineração de dados é a parte de um processo maior de descoberta do conhecimento em base de dados (Knowledge Discovery in Database - KDD). KDD consiste, fundamentalmente, na estruturação do banco de dados; na seleção, preparação e pré-processamento dos dados; na transformação, adequação e redução da sua dimensionalidade; e nas análises, assimilações, interpretações e uso do conhecimento extraído do banco de dados (CAZZELA, 2007).

O termo KDD refere-se ao processo global de descobrimento de conhecimento útil em bases de dados, e a atividade de mineração de dados é um passo particular neste processo / aplicação de algoritmos específicos para extrair padrões de dados. Os passos adicionais no processo KDD, como preparação, seleção e limpeza de dados, incorporação de conhecimento anterior apropriado e interpretação formal dos resultados de mineração, asseguram aquele conhecimento útil que é derivado dos dados.

O KDD evoluiu e continua evoluindo da interseção de pesquisas em campos como bancos de dados, aprendizado de máquinas, reconhecimento de padrões, estatísticas, inteligência artificial, aquisição de conhecimento para sistemas especialistas, visualização de dados, descoberta científica, recuperação de informação e computação de alto-desempenho. Como indica a Figura 9, sistemas de software KDD incorporam teorias, algoritmos e métodos de todos estes campos (MOTTA, 2005).

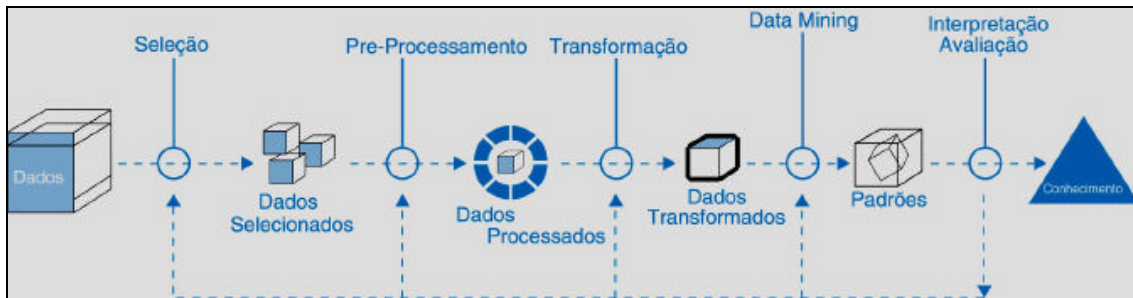


Figura 9: Apresenta o processo do KDD, incluindo o ciclo virtuoso de evolução do processo (MOTTA, 2005).

Os projetos de mineração de dados normalmente se beneficiam da criação de um ciclo, o chamado “ciclo virtuoso de data mining”, em que os algoritmos e a informação obtida utilizam-se de processamentos anteriores do mesmo modelo, incorporando as respostas obtidas em futuros processamentos “aprendizado computacionais” (MOTTA, 2005).

Os diversos algoritmos associados às técnicas computacionais de *mineração* apresentam-se em constante evolução, com o desenvolvimento de processo de dados que buscam encontrar padrões de comportamento associados à informação. A rápida taxa de inovação nas tecnologias de informática está exigindo que, cada vez mais, os profissionais estejam preparados e atualizados para conhecer e enfrentar os desafios da Tecnologia da Informação.

O processo de extração de conhecimento não explícito a partir de bancos de dados é fase mais importante desse processo (mineração de dados ou Data Mining), responsável por extrair, interpretar e relacionar informações provenientes de bancos de dados. Entre as diversas técnicas sendo estudadas e aplicadas, as relacionadas à linha de pesquisa denominada inteligência computacional vem obtendo excelentes resultados. Algoritmos genéticos, redes neurais artificiais e sistemas fuzzy vêm sendo intensamente investigados, sendo que a combinação dessas técnicas em sistemas híbridos tem chamado mais atenção recentemente.

### 2.13 Análise Multivariada

A denominação “Análise Multivariada” corresponde a um grande número de métodos e técnicas que utilizam simultaneamente todas as variáveis na interpretação

teórica do conjunto de dados obtidos. Distingue-se da estatística tradicional, que analisa cada variável ou cada amostra.

A estatística desenvolveu muito um ramo que olha as variáveis de maneira isolada, a “Estatística Univariada”, considerando assim uma análise isolada para determinar um padrão. Este uso simplificado apresenta sua vantagem na simplicidade de cálculo, mas também desvantagens, pois quando um fenômeno depende de muitas variáveis, geralmente este tipo de análise univariada falha, pois não basta conhecer informações estatísticas isoladas, é necessário também conhecer a totalidade destas informações fornecida pelo conjunto das variáveis (SANTOS, 2006).

A utilização de índices compostos permite capturar, simultaneamente, múltiplas dimensões em uma única medida. Entretanto, faz-se necessário um conhecimento detalhado das variáveis disponíveis e dos possíveis efeitos decorrentes da interação existente entre as mesmas, evitando a produção de resultados enganosos (BARILIS, 1997).

Em muitos casos, as relações existentes entre as variáveis não são percebidas, sendo desconsideradas. Porém, pode existir um caso restrito onde as variáveis são independentes entre si, que com razoável segurança, podem explicar um fenômeno complexo.

Ao observar o mundo que nos cerca e sua complexidade, o fator multivariado vai estar presente, necessitando de uma abordagem estatística adequada. A estatística multivariada permite uma visão global das variáveis sendo um instrumento valioso numa pesquisa complexa.

Segundo Stpanian e Garner (1989), os dados ambientais são, na sua maioria, multivariados, não se adequando a análises tradicionais. Muitos são autocorrelacionados no espaço e no tempo e, freqüentemente, são vários os determinantes de um fenômeno ambiental; suas interações tendem a serem complexas, dificultando o desenvolvimento de um modelo eficaz (ROVERE, 2005).

O desenvolvimento estatístico evolui na forma de perceber fenômenos estudados através de mecanismo sintetizador das variáveis envolvidas, procurando reduzir o problema a poucas variáveis. São muitas as técnicas de análise multivariada, assim como seus propósitos de aplicação, que variam de acordo com os objetivos da pesquisa. Se o objetivo é entender o comportamento de como as variáveis se relacionam entre si pode-se destacar o método de análise por componentes principais (APC).

## 2.14 Análise de Componentes Principais - APC

A análise de componentes principais (APC), ou em inglês PCA – Principal Component Analysis, é uma técnica estatística poderosa que pode ser utilizada para redução linear do número de variáveis de um conjunto de dados, consistindo essencialmente em reescrever as coordenadas das amostras em outro sistema de eixo mais conveniente para a análise dos dados. A utilização desta técnica mantém de forma intrínseca a informação original do dado (ZHANG, 2000).

Segundo Johnson e Wichern (2002) a APC explica a estrutura da variância e covariância de um vetor aleatório através de poucas combinações lineares, a partir das variáveis originais. O objetivo geral consiste tanto em reduzir os dados como em facilitar a interpretação, pois consiste transformação dos eixos das variáveis, tornando as novas variáveis (combinações lineares) não correlacionadas (BAPTISTELLA, 2005).

Como resultado, a ACP apresenta novas variáveis (componentes) geradas através de uma transformação matemática espacial realizada sobre as variáveis originais, onde cada componente principal é uma combinação linear de todas as variáveis originais (JAIN, 2000).

Segundo Prado *et al.*, (2002) as características mais importantes das componentes principais, que as tornam mais efetivas que as variáveis originais, são:

- As variáveis podem guardar entre si correlações que são suprimidas nas componentes principais. Ou seja, as componentes principais são ortogonais entre si. Deste modo, cada componente principal traz uma informação estatística diferente das outras.
- Como são decorrentes de processo matemático-estatístico para geração de cada componente que maximiza a informação estatística para cada uma das coordenadas que estão sendo criadas. As variáveis originais têm a mesma importância estatística, enquanto que as componentes principais têm importância estatística decrescente. Ou seja, as primeiras componentes principais são tão mais importantes que podemos até desprezar as demais.
- As componentes geradas podem ser analisadas separadamente devido à ortogonalidade, servindo para interpretar o peso das variáveis originais na combinação das componentes principais mais importantes.

- Podem servir para visualizar o conjunto da amostra apenas pelo gráfico das duas primeiras componentes principais, as quais detêm maior parte da informação estatística (Prado, 2002).

## 2.15 Classificadores

A tarefa de agrupar ou classificar objetos em categorias é uma das atividades mais comuns e primitivas do homem e vem sendo intensificada em função do grande volume de informações disponíveis atualmente, sobre as mais diversas áreas.

Para cada tipo de técnica de mineração de dados tem-se como base um conjunto de algoritmos que são usados na extração de relações relevantes dentro de uma massa de dados, sendo categorizados em classificação supervisionada e não-supervisionada.

A classificação supervisionada se dá a partir de amostras rotuladas onde o algoritmo recebe o conhecimento “*a priori*” e deriva regras de conhecimento que permite rotular novas amostras desconhecidas. Exemplos de algoritmos de classificação supervisionada:

- Paralelepípedo, Mínima distância Euclidiana, Máxima verossimilhança;
- K-vizinhos mais próximos;
- Redes neurais (Back - Propagation , Learning Vector Quantization);
- Árvores de decisão;
- Sistemas Especialistas.

A classificação não-supervisionada gera seu modelo a partir de amostras não-rotuladas e as regras de classificação são definidas pela associação das amostras similares em grupos ou clusters. Algoritmos de clusterização têm por objetivo particionar/separar um conjunto de dados em clusters/grupos de tal forma que indivíduos dentro de um mesmo cluster tenham um alto grau de similaridade, enquanto indivíduos pertencentes a diferentes clusters tenham alto grau de dissimilaridade. O uso desse processo é apropriado quando se conhece pouco ou nada sobre a estrutura de um conjunto de dados. As seguintes técnicas podem ser citadas:

- K-Médias, Isodata, agrupamento em geral;
- Fuzzy C-Médias e variantes, Fuzzy Maximum Like lihood Estimation;
- Mapas Auto-Organizáveis (SOMs).

## 2.16 Árvore de Decisão

Árvore de Decisão é um método de classificação supervisionado estruturado sobre um sistema de aprendizado orientado ao conhecimento, onde o interesse principal consiste em obter descrições simbólicas de fácil compreensão (VASCONCELOS, 2002).

Sendo um classificador simbólico, a árvore de decisão é representada como uma estrutura de dados em árvore, onde cada nó interno indica o teste em um atributo, cada ramo representa um resultado do teste, e os nós terminais (folhas) representam classes ou distribuições de classe. O topo da árvore é a sua raiz. O algoritmo de treinamento constrói a árvore de decisão recursivamente, de cima para baixo (top-down), identificando o atributo mais importante (atributo divisor ou de teste), isto é, aquele que faz a maior diferença para a classificação das amostras disponíveis. Geralmente a árvore apresenta-se como um conjunto de dados ligados ao nó inicial (ou nó raiz, que também é um nó interno) da árvore; dependendo do resultado do teste lógico usado pelo nó, a árvore ramifica-se para um dos nós filhos e este procedimento é repetido até que um nó terminal é alcançado, como indica a Figura 10 (MOTTA, 2005).

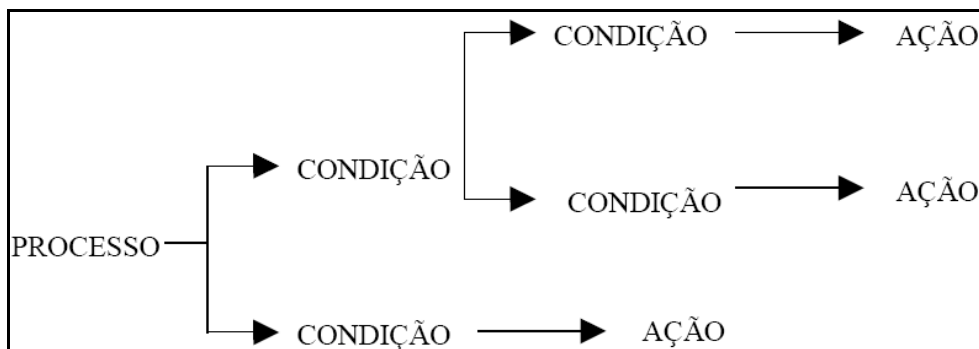


Figura 10: Representa o esquema da arquitetura da árvore de decisão.

Na árvore de decisão, cada nó corresponde a um atributo não-categorico, e cada arco a um possível valor daquele atributo. Uma folha da árvore especifica o valor esperado do atributo categorico para os registros descritos pelo caminho entre a raiz e esta folha. Na árvore de decisão, em cada nó deve ser associado o atributo não-

categorico, que é o mais informativo entre os atributos ainda não considerados no caminho a partir da raiz (SILVA, 2006).

### 2.17 FUZZY C-Médias

*Fuzzy C-Médias* é um método de classificação não-supervisionada que pode realizar a partição de um conjunto de dados utilizando o algoritmo de aproximação *C-means* para estabelecer seus *cluster*. Esse algoritmo particiona um conjunto de dados em grupos de elementos semelhantes, onde um parâmetro “k” define o número de partições inicialmente geradas. No algoritmo *fuzzy c-médias*, cada elemento está associado a uma função de pertinência que indica o grau de pertinência do elemento com relação a um determinado grupo (DRUMMONT, 2003).

Diferentemente da teoria clássica de conjuntos, onde a pertinência é definida a partir dos valores falso ou verdadeiro, ou seja, 0 ou 1 “crisp”, a pertinência FUZZY é definida como sendo um universo de discurso de uma variável definida como A, onde sua pertinência é representada pela função  $\mu_A(u): U \rightarrow [0,1]$ . A função de pertinência de A (u) representa a o grau de compatibilidade entre a x e o conceito expresso por (SÂNDI e CORREA, 1999).

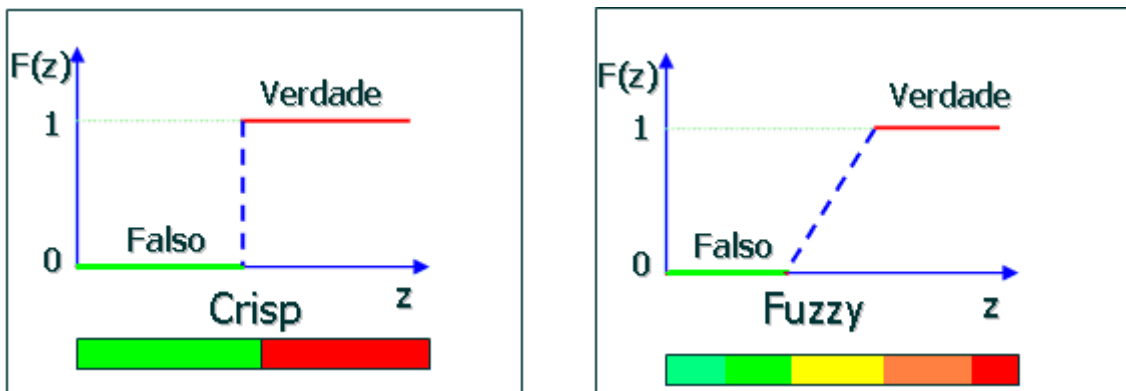


Figura 11: Representação gráfica do entendimento das possíveis respostas geradas pela teoria clássica dos conjuntos (falso ou verdade) “crisp”, comparando com o modelo “fuzzy”, que representa o valor de resposta como sendo a pertinência, ou melhor, uma graduação entre 0 e 1 informando a possibilidade da informação ser mais falsa ou mais verdadeira.

A lógica de operação do algoritmo c-médias está associada ao processo inicial de determinar as posições iniciais de k (centróides dos clusters - numero de classes), e a primeira interação do modelo determina a locação para cada elemento ao centróide mais próximo. Na seqüência do processo os centros são recalculados a partir da locação anterior, esta rotina é realizada até que algum critério de convergência seja estipulado. O valor de k é informado por algum conhecimento *a priori*.

O algoritmo *fuzzy C-medias* tem seu processamento iterativo onde os objetos, inicialmente em posição aleatória, são classificados, em grupos com o número de classes determinada no modelo, a partir desse valor é calculado os centros de cada classe utilizando como base o valor médio dos atributos dos objetos. Na seqüência do processo, os objetos são realocados entre as classes conforme a similaridade entre eles. Considerando um conjunto de pontos  $X = \{ x_1, x_2, \dots, x_n, \}$  é o conjunto de pertinência onde,  $U_{ik}(x) \rightarrow [0,1]$ ,  $i = 1, \dots, k$ , define o grau de pertinência de cada  $x_k$  comparando com cada uma das K classes. Os centros de cada classes são calculados pela formula 1.

$$V_i = \frac{\sum_{k=1}^n (U_{ik})^m * x_k}{\sum_{k=1}^n (U_{ik})^m}$$

Fórmula 1: Cálculos dos centros dos agrupamentos (classes).

Onde  $m$ ,  $1 < m \leq \infty$ , é um coeficiente que vai ponderar o quanto o grau de pertinência e a influência métrica de distância definida, quanto maior o valor maior a sobreposição das classes.

A cada interação o grau de pertinência é recalculado utilizando a formula 2, onde  $d$  é a métrica de distância empregada.

$$U_{ik} = \frac{1}{\sum_{j=1}^c \left( \frac{d_{ik}}{d_{jk}} \right)^{2/(m-1)}}$$

Fórmula 2: Calculo do valor da pertinência.

O algoritmo finaliza quando um determinado número de iterações é alcançado ou quando a matriz  $U = \langle A_{ik} \rangle$  é menor que um certo limiar  $\delta$  de convergência.

No algoritmo c-médias, cada registro classificado pode pertencer a mais de um agrupamento, expresso pelo valor de pertinência.

A função do algoritmo *fuzzy* (FCM) é utilizar o menor valor da função de aproximação c-means para obter um resultado onde a distância de cada registro para o centro do agrupamento mais próximo seja ponderada por um valor de pertinência. Assim, o registro pode pertencer a mais de uma classe, seguindo seu valor de pertinência. Devemos considerar que a soma dos valores de pertinência para cada registro deve ser igual a um (1), incluindo todas as classes.

## 2.18 Métricas de Desempenho

As métricas de desempenho são utilizadas para avaliar o potencial de solução dos algoritmos de classificação.

### 2.18.1 Classificação Supervisionada

A avaliação sobre o resultado dos classificadores não é uma tarefa fácil. Existem várias formas, testes e medidas para estimar o seu desempenho; ROC (Receiver Operating Characteristic) é uma técnica para visualizar, avaliar, organizar e selecionar classificadores baseada em seu desempenho (SILVA, 2006).

Para avaliar o desempenho das análises, o gráfico ROC pode apresentar o limiar entre taxas de acertos e alarmes falsos (taxas de erros) dos classificadores. A área sob a curva do gráfico ROC, definida como AUC (Area Under ROC Curve) apresenta seus valores entre 0 e 1, e é usada como uma medida de habilidade do modelo em discriminar quem obteve uma boa solução e quem não obteve.

A curva do gráfico apresenta a probabilidade para diferentes pontos de corte, onde os valores podem ser verdadeiros positivos (sensibilidade) e verdadeiros negativos (1-especificidade). Elevados valores de sensibilidade representam a indicação correta da presença da variável desfecho, já elevados valores de especificidade representam a indicação correta da ausência da variável desfecho, e AUC é o coeficiente geral do

diagnóstico, derivado da interação entre sensibilidade e especificidade (MACHADO e LADEIRA, 2007, ANDREOSI, 2008).

De acordo com Hosmer e Lemeshow o desempenho do algoritmo pode ser avaliado segundo o critério apresentado na Tabela 1 (ANDREOSI, 2008):

Tabela 1: Apresenta um critério para a validação do modelo de classificação supervisionada.

Valores de AUC	Diagnóstico
AUC = 0.5	Modelo sem poder discriminatório
$0.7 \leq \text{AUC} < 0.8$	Discriminação aceitável
$0.8 \leq \text{AUC} < 0.9$	Discriminação excelente
$\text{AUC} \geq 0.9$	Discriminação extraordinária

## 2.18.2 Classificação Não-Supervisionada

Promover medidas para avaliar o resultado dos classificadores não supervisionados é fundamental, ainda mais quando não se conhece o número de grupos. Os algoritmos de agrupamento sempre buscam o melhor resultado baseando-se no conhecimento “a priori” de um número de grupos, entretanto isto não significa que este é o melhor resultado, o número de grupos pode influenciar no resultado do agrupamento de acordo com certa medida de validação.

Existe uma grande variedade de métricas de validação de cluster aplicadas em problemas de classificação não-supervisionada, a seguir apresentamos algumas dessas métricas:

### 2.18.2.1 PBM

É um método originalmente proposto por Pakhira, Bandyopadhyay e Maulik, é definido como:

$$PBM(K) = \left( \frac{1}{k} \cdot \frac{E_1}{E_K} \cdot D_K \right)^2$$

Fórmula 3: Índice PBM.

Sendo que  $K$  é o número de cluster;  $E1$  é a soma das distâncias até o centro correspondente;  $EK$  é a soma das distâncias até o centro correspondente ponderado pelo valor da função de pertinência; e  $DK$  é o valor máximo da separação entre dois grupos (MARTINS, 2005).

### 2.18.2.2 Índice Xie e Beni (Xie e Beni's Index XB)

Esse método tem como objetivo quantificar a relação da variação total dentro de agrupamentos e a separação de agrupamentos. O número de cluster ótimo seria o resultado que minimize este valor (MARTINS, 2005).

$$XB(c) = \frac{\sum_{i=1}^c \sum_{j=1}^N (\mu_{ij})^m \|x_j - v_i\|^2}{N \min i, j \|x_j - v_i\|^2}$$

Fórmula 4: Índice Xie e Beni.

### 2.18.2.3 Discriminante de Fisher

Método que utiliza informações das categorias associadas a cada padrão para extrair linearmente as características mais discriminantes. Em análise de discriminantes, a separação inter-classes é enfatizada através da substituição da matriz de covariância total do PCA por uma medida de separabilidade, como o critério Fisher. O critério Fisher trata da determinação dos auto-vetores de  $S_w^{-1} \cdot S_b$ , sendo que  $S_w$  é a matriz de espalhamento intra-classes e  $S_b$  é a matriz de espalhamento inter-classes (WELLING, 2005), ou seja, esse método extrai coeficientes que minimizam o espalhamento de padrões de mesma classe entre classes e maximizam o espalhamento de padrões de classes diferentes.

### 2.18.2.4 Índice de Calinski-Harabasz

Esse índice, ilustrado pela Equação abaixo, tem B(K) e W(K) como sendo a somas dos quadrados das distâncias ao centróide, respectivamente entre os clusters e dentro dos clusters, com K clusters.

$$CH(K) = \frac{B(K)/(K-1)}{W(K)/(n-1)}$$

Fórmula 5: Índice de Calinski-Harabasz

O índice não está definido para K = 1. O melhor número de clusters é aquele que maximiza o índice.

## **2.19 Considerações Ambientais:**

Esse tópico do trabalho apresenta algumas considerações necessárias para o entendimento das análises realizadas sobre as variáveis ambientais selecionadas. São informações sobre os parâmetros físico-químicos e microbiológicos investigados.

### **2.19.1 Parâmetros Físico-Químicos**

#### **2.19.1.1 Material em Suspensão**

Qualquer tipo de partícula, seja ela mineral ou orgânica, disponível na coluna d'água, podendo indicar poluição em caso de grandes concentrações (APHA, AWWA, WEF, 1995). Sua principal influência é na diminuição na transparência da água, impedindo a penetração da luz.

#### **2.19.1.2 Amônia**

A amônia é facilmente biodegradável. As plantas o absorvem com muita facilidade, sendo um nutriente muito importante como fornecedor de nitrogênio para a

produção de compostos orgânicos. Em concentrações muito altas, por exemplo, na água de consumo, pode causar danos graves, já que o amoníaco interfere no transporte do oxigênio pela hemoglobina, entre outros efeitos nefastos. Os organismos necessitam, nesse caso, de manter uma baixa concentração de amoníaco que, caso contrário torna-se particularmente tóxico (WIKIPEDIA, 2009) indicando o nível de eutrofização. Este nutriente está presente no esgoto doméstico.

### **2.19.1.3 Oxigênio Dissolvido**

Fundamental para a sobrevivência de organismos aeróbicos, esse gás dissolvido na água é importante na avaliação de qualidade de qualquer corpo d'água. Do ponto de vista ecológico, o oxigênio dissolvido é uma variável extremamente importante, pois é necessário para a respiração da maioria dos organismos que habitam o meio aquático. Geralmente o oxigênio dissolvido se reduz ou desaparece, quando a água recebe grandes quantidades de substâncias orgânicas biodegradáveis encontradas, por exemplo, no esgoto doméstico, em certos resíduos industriais, no vinhoto, e outros (CARMOUZE, 1994).

### **2.19.1.4 Salinidade**

No caso das baía e águas costeiras, pode-se dizer que esse parâmetro auxilia no entendimento do grau de renovação das águas. Nos estudos ambientais, a salinidade pode representar o equilíbrio do ecossistema, uma vez que os organismos precisam assimilar a sua variação natural (MAYR *et al.*, 1989).

### **2.19.1.5 Nitrogênio Total**

As águas naturais, em geral, contêm nitratos em solução e, além disso, principalmente tratando-se de águas que recebem esgotos, podem conter quantidades variáveis de compostos mais complexos, ou menos oxidados, tais como: compostos orgânicos quaternários, amônia e nitritos. Em geral, a presença destes denuncia a existência de poluição recente, uma vez que essas substâncias são oxidadas rapidamente na água, graças principalmente à presença de bactérias nitrificantes. Por essa razão,

constituem um importante índice da presença de despejos orgânicos recentes (CARMOUZE, 1994).

Os compostos de nitrogênio, um dos elementos mais importantes no metabolismo de ecossistemas aquáticos, possuem comportamento químico complexo, em virtude dos vários estágios que o nitrogênio pode assumir e dos impactos que a mudança do seu estado de oxidação podem trazer sobre os organismos vivos.

#### **2.19.1.6 Fósforo Total**

Representa a soma de todas as formas de fósforo disponíveis em um corpo d'água, e assim como os parâmetros anteriores, pode auxiliar no nível e comprometimento trófico de uma região. Os compostos de fósforo são um dos mais importantes fatores limitantes à vida dos organismos aquáticos e a sua economia, em uma massa d'água, é de importância fundamental no controle ecológico das algas. Despejos orgânicos, especialmente esgotos domésticos, bem como alguns tipos de despejos industriais, podem enriquecer as águas com esse elemento.

#### **2.19.1.7 pH**

Parâmetro “símbolo” representando o potencial hidrogeniônico. Essa grandeza indica a acidez, neutralidade ou alcalinidade de uma solução líquida, estando relacionada com as concentrações de ácidos fracos como a amônia, e com potencial de avaliação na qualidade da água (BARROS, 2001).

A condutividade elétrica é a capacidade que a água possui de conduzir corrente elétrica. Este parâmetro está relacionado com a presença de íons dissolvidos na água, que são partículas carregadas eletricamente. Quanto maior for a quantidade de íons dissolvidos, maior será a condutividade elétrica da água. O parâmetro condutividade elétrica não determina, especificamente, quais os íons que estão presentes em determinada amostra de água, mas pode contribuir para possíveis reconhecimentos de impactos ambientais que ocorram na bacia de drenagem ocasionados por lançamentos de resíduos industriais, mineração, esgotos, etc. (CARMOUZE, 1994).

### **2.19.1.8 Temperatura da Água**

Nos ecossistemas aquáticos continentais, a quase totalidade da propagação do calor ocorre por transporte de massa d'água, sendo a eficiência deste em função da ausência ou presença de camadas de diferentes densidades. As diferenças de temperatura geram camadas d'água com diferentes densidades, que em si já formam uma barreira física, impedindo que se misturem, e se a energia do vento não for suficiente para misturá-las, o calor não se distribui uniformemente, criando a condição de estabilidade térmica. Quando ocorre este fenômeno, o ecossistema aquático está estratificado termicamente. Os estratos formados freqüentemente estão diferenciados física, química e biologicamente (CARMOUZE, 1994).

A importância desse parâmetro está associada ao comportamento da variável em águas estuarinas, sendo influenciado pela temperatura das águas fluviais, escoamento de água salgada oceânica, condições climatológicas e profundidade do estuário (CUNHA, 1982).

### **2.19.1.9 Profundidade da Estação**

Esse parâmetro está diretamente ligado à circulação da baía. Em teoria, as áreas mais profundas apresentam maior circulação de água, aumentando a possibilidade de renovação da água.

## **2.19.2 Parâmetros Microbiológicos**

### **2.19.2.1 Clorofila-a**

É a concentração de pigmentos fotossintetizantes, extensivamente utilizada para estimar a biomassa de fitoplâncton, podendo ser utilizada como indicador do nível de eutrofização do corpo d'água por permitir a avaliação da produtividade primária do fitoplâncton (MAYR *et al.*, 1989).

### **2.19.2.2 Abundância Bacteriana**

As bactérias são a base de muitas cadeias alimentares existentes no mar. Se a taxa de introdução de matéria orgânica no mar exceder a taxa de ação bacteriana, que depende da temperatura ambiente, da disponibilidade do oxigênio, das correntes marinhas e de outros fatores, haverá um acúmulo de matéria orgânica, que favorecerá, a princípio, as plantas, resultando no aumento da abundância de fito e zooplânctons, além de beneficiar inúmeras outras cadeias alimentares. Mas, em contrapartida, se o acúmulo de nutrientes for excessivo, ocorrerá a eutrofização. No mar, a eutrofização está associada ao desenvolvimento da maré vermelha, ou seja, um rápido aumento da população de fitoplânctons que faz com que o mar perca sua coloração original, ficando vermelho e, às vezes, amarelo ou marrom (NASSER, 2001).

# CAPÍTULO 3

## ESTUDO DE CASO

O estudo de caso proposto utilizou dados adquiridos no projeto de pesquisa denominado “Avaliação Ambiental da Baía de Guanabara”, desenvolvido pela PETROBRAS S.A., com a coordenação realizada pelo CENPES – Centro de Pesquisas e Desenvolvimento Leopoldo Américo Miguez de Mello, mais especificamente na gerência AMA – Avaliação e Monitoramento Ambiental.

Com uma proposta integradora, o projeto atua de forma multidisciplinar (Tabela 2 ), envolvendo diversas universidades e instituições de pesquisa, focados em avaliar o ecossistema estuarino da Baía de Guanabara, área com a presença de inúmeras instalações da PETROBRAS, no Rio de Janeiro. Durante o período de julho de 2005 a junho de 2007 foram realizadas coletas de dados ambientais de diversos compartimentos da baía, como água, sedimento e biota, com o objetivo de gerar o conhecimento necessário para caracterizar as variações sazonais e, por conseguinte, a dinâmica natural deste ecossistema.

Apesar de todo o cenário de variáveis disponível no projeto, este estudo aborda somente os dados do compartimento água, avaliando os resultados da hidroquímica, considerando os parâmetros físico-químicos e biológicos da água.

### 3.1 Resumo Metodológico

O estudo em questão envolveu uma seqüência de procedimento metodológicos ate a geração dos “novos conhecimentos”, o esquema apresentado na Figura 41 retrata todas as fases envolvidas no estudo de caso, desde a geração das amostras em campo, processamento das alíquotas em laboratório e modelagem/carga do banco de dados. Seguindo pela fase de mineração dos dados, que teve inicio a partir de uma análise exploratória dos dados na busca de padrões espaço-temporal considerando ambiente uni e multivariado. Com o entendimento necessário do comportamento das variáveis e suas inter-relações buscou-se através de técnicas de classificação supervisionada (árvore de

decisão) avaliar a possibilidade de criação de um modelo de regras de classificação utilizando um zoneamento de qualidade de água definido por Mayr *et al*, 1989 como conhecimento *a priori*. Com o interesse em avaliar uma possibilidade de agrupamento espacial dos dados, realizou-se também uma análise de classificação não supervisionada (sobreposição ponderada e *fuzzy c-medias*), considerando a ausência de uma bibliografia que apresente valores de referencias para as regras classes de qualidade da água para a área de estudo, optou-se em definir os limiares das classes de qualidade por métodos matemáticos respeitando uma ordenação ambiental.

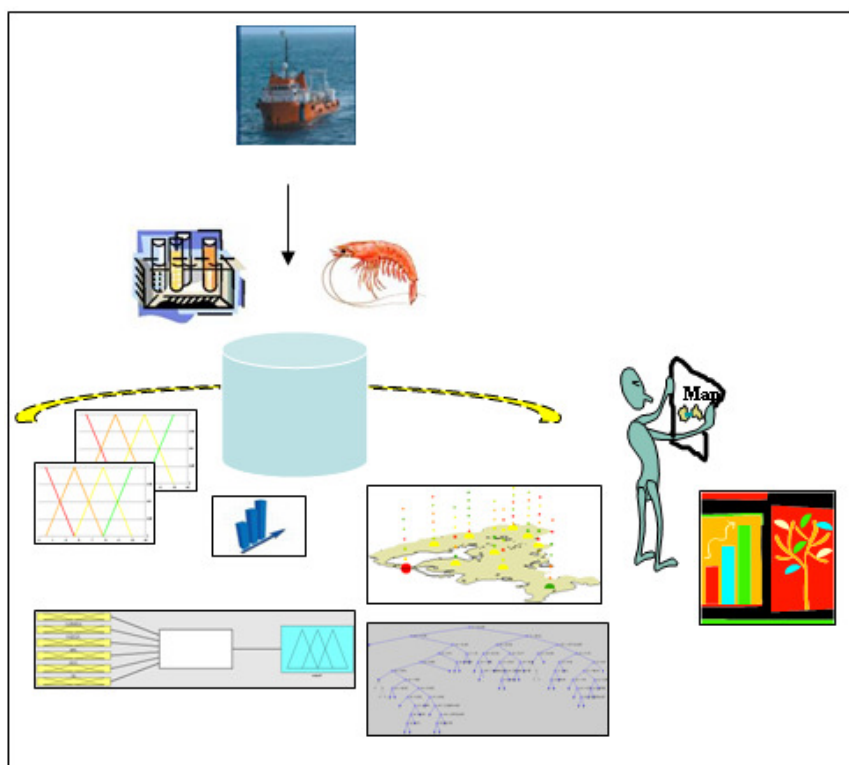


Figura 12: Esquema dos processos envolvidos até a geração do conhecimento.

### 3.2 Área de Estudo

Situada na região sudeste do Brasil, a Baía de Guanabara está localizada entre paralelos 22° 24' e 22° 57' de latitude sul e entre os meridianos 42° 33' e 43° 19' de longitude leste (Figura 13). Está inserida no contexto da região metropolitana do Rio de Janeiro, sendo considerado um ecossistema de extrema importância para o desenvolvimento da região e seu entorno. A presença de planícies, colinas, maciços costeiros, praias, costão rochoso, ilhas e escarpas da Serra do Mar caracterizam a

geografia física da região. Faz parte de uma bacia hidrográfica que ocupa uma área total de 4.600 Km<sup>2</sup> (MAYR, 1989), com seu espelho d'água (400 Km<sup>2</sup>), e estende-se por 28 Km no sentido norte-sul e até 27 Km de largura (leste-oeste). Possui um canal central com mais de 20 Km de comprimento por 4 Km de largura (boca), com sua profundidade variando entre 15 a 58 metros, enquanto o restante da baía apresenta uma profundidade média de 5,7 metros (RIBEIRO, 1996).

Tabela 2: Apresenta as áreas de atuação do projeto de avaliação ambiental da Baía de Guanabara.

<b>Compartimento Água da baía:</b>
1 Hidroquímica
3 Ictioplacton
4 Costões
5 Praia
<b>Compartimento Sedimento da baía:</b>
6 Geologia
7 HC/coprostanol
8 Metais/AVS
9 Ecotoxicologia
<b>Compartimento Biota da baía:</b>
10 Endofauna
11 Arrasto Ictiofauna
12 Arrasto Epifauna
13 Foraminíferos / Diatomáceas
<b>Compartimento Manguezal</b>
14 Físico-química do sedimento / topografia
15 HC
16 Metais/AVS
17 Vegetação
18 Histopatológica
19 Endofauna
20 Caranguejo
21 Aves
22 Replântio

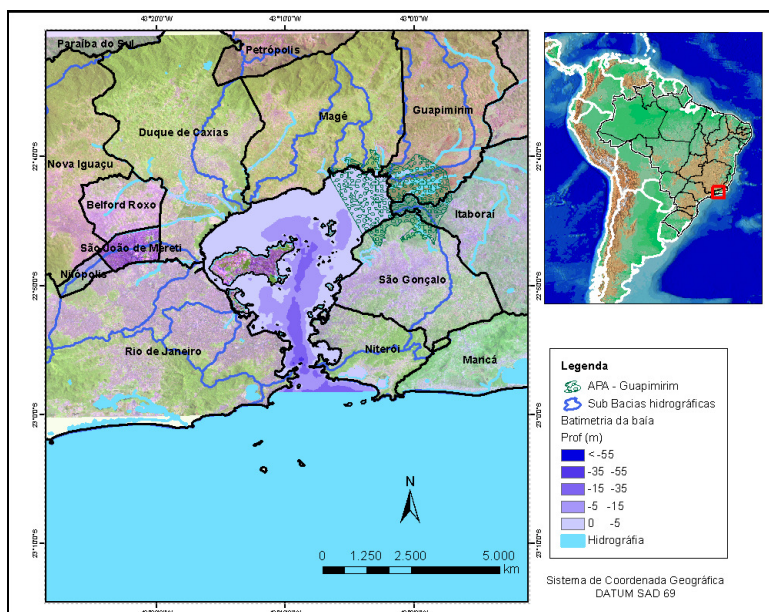


Figura 13: Apresentação geográfica da área de estudo.

### 3.3 Coleta dos Dados

Os dados analisados foram gerados durante o período de dois anos de coleta (2005-2007) ininterruptos, com uma periodicidade quinzenal, totalizando 48 expedições. A malha amostral foi composta por 10 estações distribuídas sobre o espelho d'água da baía, e o desenho amostral foi planejado buscando a melhor representação espacial dentro das possibilidades/orientações do projeto (Figura 14).

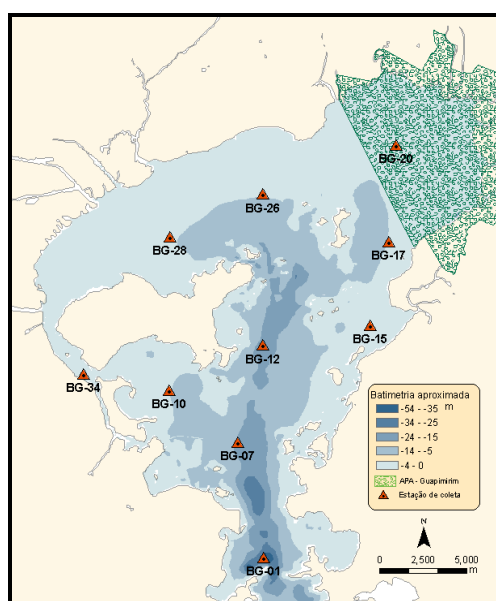


Figura 14: Malha amostral.

Para todas as campanhas de coleta de água (expedições), a equipe responsável pela execução da tarefa adotou um padrão para as práticas de coleta, conservação, transporte e análise das amostras, garantindo uma qualidade e rastreabilidade dos dados.

As alíquotas das amostras para as análises foram retiradas da superfície (0,5 m de profundidade) e fundo (1,0 m acima do fundo de acordo com a profundidade local). Todo o procedimento de coleta e geração dos parâmetros físico-químico avaliados foram de responsabilidade da Universidade Federal do Rio de Janeiro, Equipe do Instituto de Biologia - Departamento de Biologia Marinha, sob coordenação do Dr. Rodolfo Paranhos.

Quanto a coleta de água foi realizada em condições de maré onde a profundidade da estação não era superior a 2 m de lamina da água, nesse estudo adotou-se os mesmos valores (repetidos) para os parâmetros de fundo e superfície, as estações BG-15, BG-17 e BG-20 em alguns campanhas apresentaram essa condição particular, a Tabela 3 apresenta os valores médio da profundidade em metros para cada estação de coleta ao longo das 48 campanhas.

Tabela 3: Relação das estações com a profundidade média

Estação de coleta	Profundidade média da estação (m)
BG-01	21.6
BG-07	20.6
BG-10	5.0
BG-12	25.8
BG-15	2.2
BG-17	3.4
BG-20	2.9
BG-26	4.0
BG-28	4.0
BG-34	7.1

### 3.4 Banco de Dados

As características integradoras e multidisciplinares do projeto de Caracterização Ambiental da Baía de Guanabara, incentivou a PETROBRAS S.A. a promover uma

estrutura para que um dos produtos esperados do projeto fosse um banco de dados georreferenciado.

O desenvolvimento da modelagem do banco de dados espacial (geodatabase) dimensionou um conjunto de atributos (metadados) que permite maior rastreabilidade de uso, armazenando detalhes como:

- informações de campo, caderneta de campo – coordenadas geográficas, dia e hora da coleta, condição climática, equipe, condição de maré, entre outras;
- informações associadas ao processo de triagem, laboratório e resultado – métodos e processos de análise, resultados analíticos, dados brutos;
- informações secundárias (dados derivados) produzidos a partir de interpretação especialista – índices, médias, padrões, variações espaço-temporais, entre outros.

O banco de dados do projeto tem seu conceito fundamentado em “*Spatial Data Warehouse- SDW*” que é definido por Bohorquez (2000) como “uma grande base de dados espacial com procedimentos, desde a população dos dados, a transformação e padronização, a fixação do valor temporal, assim como toda a infra-estrutura para a consulta, dentro de uma perspectiva holística orientada à área de negócios (...) para servir de base para a tomada de decisão”.

Em termos genéricos, pode-se entender o SDW como um grande banco de dados espacial corporativo, com o objetivo de ser base de consultas da organização, e não o repositório para aplicações transacionais do dia-a-dia da empresa (BOHORQUEZ, 2000).

Por se tratar de uma grande estrutura para armazenamento de dados disponíveis para consulta de múltiplos usuários, é de conveniência que o SDW tenha sua estrutura de metadados habilitada, vinculando ao dado um conjunto de outras informações que não são adequadas para o modelo tabular.

Para o banco em questão, o padrão de metadados adotado foi o FGDC - Federal Geographic Data Committee, adaptado pela ESRI. Esse padrão atende de forma satisfatória o uso, preservando informações relevantes para o compartilhamento do dado.

A complexidade de informações disponível no modelo de metadados adotado (FGDC) condicionou o uso de um padrão mínimo de preenchimento, servindo como critério mínimo para a carga do dado no banco. Para cada nível de informação do banco, o metadados apresenta os seguintes itens:

- Palavras chaves;
- Abstract;
- Propósito;
- Informações Suplementares;
- Descrição;

Para cada campo da tabela uma descrição:

- Alias (nome do campo apresentado para o usuário no momento da consulta);
- Definição;

A disponibilidade de consulta dos metadados no padrão FGDC (ESRI) apresenta-se dividida em três categorias:

- *Description* (Descrição) – armazenando informações descritivas sobre o layer (Figura 14);
- *Spatial* (Espacial) – contendo informações sobre as características espaciais do dado (Figura 15);
- *Attributes* (Atributos) – descreve informações detalhas sobre cada campo da tabela (Figura 16).

The image shows a web-based metadata viewer interface. At the top, there are three tabs: 'Description', 'Spatial', and 'Attributes'. The 'Description' tab is selected and highlighted in blue. Below the tabs, the content is organized into several sections:

- Keywords:** Theme: Baía de Guanabara, Qualidade de água, nutrientes, clorofila, eutrofização
- Description:**
  - Abstract:** Foram realizadas 48 campanhas de coleta de amostras de água na baía de Guanabara entre julho de 2005 e junho de 2007, e os resultados obtidos permitem avaliar o grau de eutrofização deste importante ecossistema.
  - Purpose:** A partir de coletas frequentes (quinzenais) em 10 locais da baía de Guanabara pretende-se avaliar as condições de qualidade das águas deste importante ecossistema aquático do litoral do Estado do Rio de Janeiro.
  - Supplementary Information:** Prof. Rodolfo Paranhos, Laboratório de Hidrobiologia, Instituto de Biologia da UFRJ.
- Status of the data:** (with a horizontal line below it)
- Time period for which the data is relevant:** (with a horizontal line below it)
- Publication Information:** (with a horizontal line below it)
- Data storage and access information:**
- Details about this document:**

Figura 15: Aba de exibição das descrições dos metadados.

Description	Spatial	Attributes
<p><b>Horizontal coordinate system</b>            Projected coordinate system name: SAD_1969_UTM_Zone_23S            Geographic coordinate system name: GCS_South_American_1969  <a href="#">Details</a></p>		
<p><b>Altitude System Definition</b>            Resolution: 0.000003            Encoding Method: Explicit elevation coordinate included with horizontal coordinates</p>		
<p><b>Bounding coordinates</b></p> <p><b>Horizontal</b></p> <p><b>In decimal degrees</b>            West: -43.290620            East: -42.948725            North: -22.655040            South: -22.942264</p> <p><b>In projected or local coordinates</b>            Left: 675635.949970            Right: 710345.529956            Top: 7493209.909994            Bottom: 7461839.179948</p>		
<p><b>Lineage</b></p> <p><b>FGDC lineage</b>  <a href="#">Process step 1</a>  <a href="#">Process step 2</a>  <a href="#">Process step 3</a></p>		
<p><b>Spatial data description</b></p> <p><b>Vector data information</b></p> <p><b>ESRI description</b>  <b>HIDROBIOLOGIA_RESULTADOS</b>            ESRI feature type: Simple            Geometry type: Point            Topology: FALSE            Feature count: 1147            Spatial Index: TRUE            Linear referencing: FALSE</p>		

Figura 16: Aba de exibição dos metadados para as informações referentes ao sistema de coordenada do layer.

Description	Spatial	Attributes
<p><b>Details for Qualidade de água da Baía de Guanabara</b>            Type of object: Feature Class            Number of records: 1147  <a href="#">Description</a>            Estudo da qualidade da água na Baía de Guanabara</p>		
<p><b>Attributes</b></p> <p><b>OBJECTID</b></p> <p><b>Shape</b></p> <p><b>Estacao</b></p> <p><b>Campanha</b></p> <p><b>nivel</b>            Alias: nivel            Data type: String            Width: 50            Precision: 0            Scale: 0            Definition:            Podendo ser superfície ou fundo</p> <p><b>DataColeta</b></p> <p><b>HoraColeta</b></p> <p><b>HoraColeta_fim</b></p> <p><b>Coletor</b></p> <p><b>Prof_Estacao</b>            Alias: Profundidade local (m)            Data type: Integer            Width: 4            Precision: 0            Scale: 0            Definition:            Profundidade do local de coleta</p> <p><b>Prof_Secchi</b></p> <p><b>Prof_Coleta</b></p> <p><b>Temperatura_Ambiente</b></p> <p><b>CondicaoClimatica</b></p>		

Figura 17: A aba do metadados associada aos atributos da tabela exibe toda a estrutura do campo, assim como seu alias e sua definição.

Seguindo o processo baseado no conceito KDD, após a geração dos dados prosseguiu-se na seleção dos dados a serem investigados.

### 3.5 Análise Exploratória dos Dados

A análise exploratória dos dados teve início com uma breve visualização do conteúdo a ser avaliado no banco e o entendimento do seu propósito, bem como a importância de seus atributos no contexto da análise. Nesta fase de consulta as informações disponíveis no metadados do banco espacial auxiliaram o entendimento das variáveis a serem exploradas, a seguir a relação dos atributos do banco assim como os itens disponíveis no metadados, para cada campo da tabela de dados:

- 1) Nome do campo no modelo de dados: MPS  
*Alias:* Material Particulado em Suspensão (mg L-1)  
*Data type:* Double; Width: 8; Precision: 0; Scale: 0  
*Definition:* Material Particulado em Suspensão (mg L-1)
  
- 2) Nome do campo no modelo de dados: AMON\_AG  
*Alias:* Nitrogênio Amoniacal ( $\mu\text{M}$  - N-NH<sub>3</sub>/NH<sub>4</sub><sup>+</sup>)  
*Data type:* Double; Width: 8; Precision: 0; Scale: 0  
*Definition:* Nitrogênio Amoniacal ( $\mu\text{M}$  - N-NH<sub>3</sub>/NH<sub>4</sub><sup>+</sup>)
  
- 3) Nome do campo no modelo de dados: OD\_AG  
*Alias:* Oxigênio dissolvido (mL.L-1)  
*Data type:* Double; Width: 8; Precision: 0; Scale: 0  
*Definition:* Oxigênio dissolvido (mL.L-1)
  
- 4) Nome do campo no modelo de dados: AB\_BAC  
*Alias:* Abundância Bacteriana (cel.mL-1)  
*Data type:* Double; Width: 8; Precision: 0; Scale: 0

*Definition:* Abundância Bacteriana (cel.mL-1)

5) Nome do campo no modelo de dados: CLOR\_A

*Alias:* Clorofila a ( $\mu\text{g.L-1}$ )

*Data type:* Double; Width: 8; Precision: 0; Scale: 0

*Definition:* Clorofila a ( $\mu\text{g.L-1}$ )

6) Nome do campo no modelo de dados: SAL\_AG

*Alias:* Salinidade (S)

*Data type:* Double; Width: 8; Precision: 0; Scale: 0

*Definition:* Salinidade (S)

7) Nome do campo no modelo de dados: NIT\_TOT\_AG:

*Alias:* Nitrogênio Total ( $\mu\text{M N}$ )

*Data type:* Double; Width: 8; Precision: 0; Scale: 0

*Definition:* Nitrogênio Total ( $\mu\text{M N}$ )

8) Nome do campo no modelo de dados: FOS\_TOT\_AG:

*Alias:* Fósforo Total ( $\mu\text{M P}$ )

*Data type:* Double; Width: 8; Precision: 0; Scale: 0

*Definition:* Fósforo Total ( $\mu\text{M P}$ )

9) Nome do campo no modelo de dados: PH

*Alias:* pH

*Data type:* Double; Width: 8; Precision: 0; Scale: 0

*Definition:* potencial hidrogeniônico

10) Nome do campo no modelo de dados: TEMP\_AG

*Alias:* Temp. da água ( $^{\circ}\text{C}$ )

*Data type:* Double; Width: 8; Precision: 0; Scale: 0

*Definition:* Temperatura do ar ( $^{\circ}\text{C}$ )

11) Nome do campo no modelo de dados: PROF\_SEC

*Alias:* Profundidade máxima de visualização do disco de Secchi (m)

*Data type:* Double; Width: 8; Precision: 0; Scale:

*Definition:* Medida obtida pela profundidade máxima de visualização do disco de Secchi (m) inferindo um valor a transparência da água.

12) Nome do campo no modelo de dados: PROF\_COL

*Alias:* Profundidade (m)

*Data type:* Double; Width: 8; Precision: 0; Scale:

*Definition:* Profundidade da amostragem (m)

13) Nome do campo no modelo de dados: PROF\_EST

*Alias:* Profundidade local (m) da estação

*Data type:* Double; Width: 8; Precision: 0; Scale:

*Definition:* Profundidade do local de coleta

14) Nome do campo no modelo de dados: CLASS\_MAYR

*Alias:* Qualidade da água (Mayr 1989, et al)

*Data type:* Integer; Width: 4; Precision: 0; Scale: 0

*Definition:* Valores associados às áreas de classificação da qualidade de água proposta por Mayr et al, 1989.

15) Nome do campo no modelo de dados: DIST\_C\_M

*Alias:* Distância do limite da classe (m)

*Data type:* Integer; Width: 4; Precision: 0; Scale: 0

*Definition:* Distância calculada em metros (m) entre o ponto amostral (estação) e o limite mais próximo da borda de classificação da qualidade d'água proposta por Mayr et al, 1989.

16) Nome do campo no modelo de dados: Coord\_X

*Alias:* UTM (X)

*Data type:* Double; Width: 8; Precision: 0; Scale: 0

*Definition:* Sistema de coordenada UTM (X) - DATUM SAD69

17) Nome do campo no modelo de dados: Coord\_Y

*Alias:* UTM (Y)

*Data type:* Double; Width: 8; Precision: 0; Scale: 0

*Definition:* Sistema de coordenada UTM (Y) - DATUM SAD69

18) Nome do campo no modelo de dados: CAMP\_COL

*Alias:* Campanha de coleta de dados

*Data type:* String; Width: 35

*Definition:* Nomenclatura da serie de coletas de campo (Campanha de coleta de amostras)

19) Nome do campo no modelo de dados: DATA\_COL

*Alias:* Data da Campanha

*Data type:* date

*Definition:* Data da coleta das amostras de água

18) Nome do campo no modelo de dados: EST\_COL

*Alias:* Estação de coleta

*Data type:* String; Width: 35

*Definition:* Ponto de coleta de amostras (Estação de coleta de água)

Com o conhecimento das variáveis disponíveis no banco, selecionou as variáveis com potencial exploração ambiental para a geração de uma estatística básica. Nesse momento o fator temporal foi desconsiderado. Essa abordagem univariada estabeleceu entendimento geral para a análise apresentando os valores máximos, mínimos e o desvio padrão de cada atributo da tabela, e esses resultados são apresentados na tabela 3.

Tabela 4: Valores da estatística básica das variáveis.

Parâmetro	Análise Descritiva dos dados					
	Valid N	Mean	Median	Minimum	Maximum	Std.Dev.
PROF_EST	960	10	5	1.0	31	9
PROF_SEC	960	1	1	0.0	8	1
PROF_COL	960	5	1	1.0	30	7
TEMP_AG	944	24	24	16.0	32	2
SALIN_AG	959	30	31	11.5	37	4
PH	960	8	8	7.2	9	0
OD_AG	958	4	4	0.0	13	2
FOS_TOT_AG	960	5	4	0.5	38	5
AMON_AG	956	30	12	0.0	473	54
NIT_TOT_AG	945	98	61	4.7	1124	117
MPS	940	39	31	1.7	341	31
CLOR_A	954	38	20	0.2	535	54
AB_BAC	960	14884895	12365194	120000.0	96800000	12189998

Com um total de 960 amostras (10 estações X 48 tempos X dois níveis de coleta – superfície e fundo), observou-se que a base de dados apresentava algumas variáveis com um número menor de medições. As variáveis TEMP\_AG, SALIN\_AG, OD\_AG, AMON\_AG, NIT\_TOT\_AG, MPS, CLOR\_A apresentam esse problema. Segundo informação disponível nos metadados, essas ausências correspondem a problemas associadas ao processo analítico.

Um número baixo de medições existentes no banco pode dificultar o entendimento da variação espaço-temporal da base de dados, assim como a definição de qualquer modelo matemático. Por isto, uma imputação de dados coerentes foi utilizada para resolver esses “vazios” do banco sem comprometer o resultado da análise.

Das inúmeras técnicas disponíveis no mercado para a imputação de dados perdidos em base de dados, optou-se por utilizar os valores médio, também conhecidos como métodos de imputação única. Esta prática é bastante usada pela sua facilidade de implementação. Entretanto, existem desvantagens na sua utilização, como a subestimação da variabilidade e a impossibilidade da utilização de outras variáveis do próprio conjunto de dados para melhorar o processo de imputação.

Assim, a primeira inferência aplicada ao banco foi a imputação do valor correspondente a média, que foi calculada a partir dos valores da campanha anterior e posterior. Essa estratégia foi utilizada na intenção de manter o padrão da série temporal. A Tabela 4 apresenta uma nova estatística básica com os valores imputados.

Tabela 5: Apresenta a estatística básica já considerando a substituição

dos valores perdidos por valores médios.

Parâmetro	Análise Descritiva dos dados					
	Valid N	Mean	Median	Minimum	Maximum	Std.Dev.
PROF_EST	960	10	5	1.0	31	9
PROF_SEC	960	1	1	0.2	8	1
PROF_COL	960	5	1	1.0	30	7
TEM_AG	960	24	24	16.0	32	2
SAL_AG	960	30	31	11.5	37	4
PH	960	8	8	7.2	9	0
OD_AG	960	4	4	0.0	13	2
FOS_TOT	960	5	4	0.5	38	5
AMON_AG	960	31	12	0.0	473	56
NIT_TOT_AG	960	98	62	4.7	1124	117
MPS	960	39	31	1.7	341	30
CLOR_A	960	38	20	0.2	535	54
AB_BAC	960	14884895	12365194	120000.0	96800000	12189998

Comparando o resultado obtido com o anterior (Tabela 3) não observa-se alteração no padrão de comportamento dos dados para os limites apresentados.

A avaliação da distribuição dos valores de cada variável físico-química a partir do gráfico de histograma (ANEXO I), assim como a curva normal, demonstrou bom comportamento para os dados. A curva normal apresentou seu pico quase sempre coincidente com o intervalo de valores com maior frequência (forma de sino).

A partir desse ponto, a abordagem investigativa passou a considerar a questão do tempo e espaço presente na base de dados.

Na busca para identificar semelhança no comportamento dos dados, o ANEXO II, apresenta os gráficos com a distribuição dos valores das variáveis considerando cada estação de coleta e sua variabilidade ao longo dos 48 campanhas amostrais.

Seguindo a investigação por padrões, a seguir avaliou-se a distribuição temporal de cada variável ao longo dos dois anos de trabalho (campanhas de coleta - C01 até C48) considerando separadamente cada nível de coleta, superfície (0,5 meio metro de profundidade) e fundo (1 metro acima do leito da baía). Com a interpretação dos gráficos do ANEXO III (gráficos *Box plot* com a média de cada parâmetro por nível de coleta) nota-se pequena influencia temporal no comportamento das variáveis. Uma leve alternância dos valores sugere o efeito das épocas secas e chuvosas, concordando com o conhecimento consolidado pelos trabalhos de Mayr et al. (1989) e Villac (1990).

Assim, para os meses com maior nível de precipitação a influência da pluviosidade é marcante, principalmente na salinidade, com ocorrências das maiores concentrações durante o inverno, que representa o período de seca. Com base nos estudos citados e nos resultados obtidos, foi criado um campo na tabela de atributo com registro da época do ano baseado na variação temporal dos períodos de seca e o de chuva. A Tabela 5 associa o agrupamento definido pelo período úmido/chuvoso (verão) e seco (inverno) a seu período de coleta. Para a época úmida, obteve-se 28 tempos, sendo que seis (6) ocorreram em 2005, quatorze (14) em 2006 e oito (8) em 2007; já o período de seca totaliza 20 campanhas, sendo que seis (6) em 2005, dez (10) em 2006 e quatro (4) em 2007, lembrando que as coletas foram realizadas quinzenalmente.

Os gráficos de *box plot* (ANEXO IV) com a distribuição dos valores médios de cada variável físico-química, considerando separadamente as amostras obtidos no fundo e na superfície, a variação dos valores conforme o período de seca e chuva. Para a maioria das variáveis, os menores valores foram para as coletas de fundo e o maiores para a superfície, com exceção da salinidade. Considerando os períodos de seca e chuva, nota-se também um comportamento levemente distinto com valores maiores na época de chuva e menores para o período de seca, esse comportamento só não foi observado nas variáveis PH e da CLOR\_A.

Tabela 6: Relação campanha, época do ano, mês e ano da coleta.

Campanha	Época	Mês da Coleta	Ano da Coleta
C01	SECO	JUL	2005
C02	SECO	JUL	2005
C03	SECO	AGO	2005
C04	SECO	AGO	2005
C05	SECO	SET	2005
C06	SECO	SET	2005
C07	CHUVOSO	OUT	2005
C08	CHUVOSO	OUT	2005
C09	CHUVOSO	NOV	2005

C10	CHUVOSO	NOV	2005
C11	CHUVOSO	DEZ	2005
C12	CHUVOSO	DEZ	2005
C13	CHUVOSO	JAN	2006
C14	CHUVOSO	JAN	2006
C15	CHUVOSO	FEV	2006
C16	CHUVOSO	FEV	2006
C17	CHUVOSO	MAR	2006
C18	CHUVOSO	MAR	2006
C19	CHUVOSO	ABR	2006
C20	CHUVOSO	ABR	2006
C21	SECO	MAI	2006
C22	SECO	MAI	2006
C23	SECO	JUN	2006
C24	SECO	JUN	2006
C25	SECO	JUL	2006
C26	SECO	JUL	2006
C27	SECO	AGO	2006
C28	SECO	AGO	2006
C29	SECO	SET	2006
C30	SECO	SET	2006
C31	CHUVOSO	OUT	2006
C32	CHUVOSO	OUT	2006
C33	CHUVOSO	NOV	2006
C34	CHUVOSO	NOV	2006
C35	CHUVOSO	DEZ	2006
C36	CHUVOSO	DEZ	2006
C37	CHUVOSO	JAN	2007
C38	CHUVOSO	JAN	2007
C39	CHUVOSO	FEV	2007
C40	CHUVOSO	FEV	2007
C41	CHUVOSO	MAR	2007
C42	CHUVOSO	MAR	2007
C43	CHUVOSO	ABR	2007
C44	CHUVOSO	ABR	2007
C45	SECO	MAI	2007
C46	SECO	MAI	2007
C47	SECO	JUN	2007
C48	SECO	JUN	2007

A base de dados avaliada tem sua estrutura definida em quatro dimensões X, Y, Z e T, sendo X e Y as coordenadas geográficas, Z a profundidade de coleta, e T o tempo de coleta. A primeira investigação exploratória sugeriu/permitiu a definição de quatro (4) cenários, considerando as situações envolvendo a profundidade de coleta (Z) e o tempo de coleta (T) :

- época chuvosa – fundo;
- época chuvosa – superfície;
- época seca – fundo ;

- época seca – superfície.

O ANEXO V apresenta uma série de gráficos de histograma com a distribuição da frequência dos dados para cada variável físico-química e sua porcentagem de ocorrência. Os dados estão organizados de forma a permitir a comparação entre época do ano e nível de coleta (profundidade), reforçando o agrupamento de dados estabelecido nas análises do ANEXO IV.

Com duas dimensões (nível de coleta: superfície e fundo e época do ano: chuvosa e seca) já exploradas de forma univariada, a abordagem é direcionada para o entendimento da distribuição e variação espacial dos dados. Nesse contexto, foram criados mapas temáticos (Figuras 18 à 26) com valores médios para cada variável explorada, orientado o entendimento espacial e suas tendências.

Nesta abordagem, a temperatura média da água (TEMP\_AG) observada na Figura 18, sugere que as amostras coletadas na superfície, no período de chuva são maiores que as do período de seca para todas as estações amostradas. No fundo o padrão foi o mesmo, com exceção das estações BG01, BG-07, BG-12, situadas no canal central, mais profundo, onde as médias do período de seca foram maiores.

O mapa de salinidade (SAL\_AG) apresenta maior valor de média para o período de seca, tanto na coleta de superfície, quanto na coleta de fundo (Figura 19).

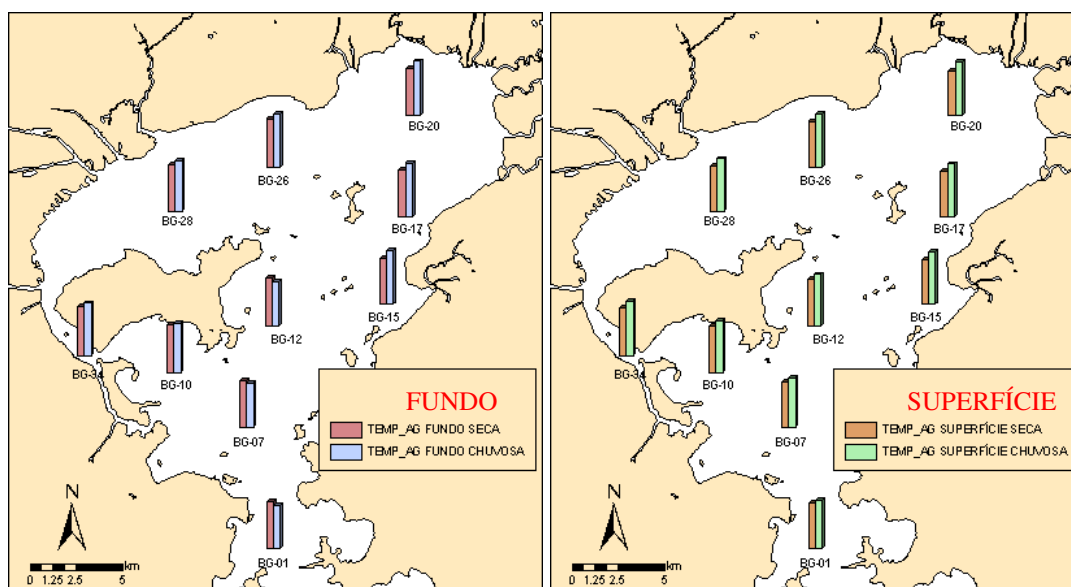


Figura 18: Mapa temático qualitativo com valores médios para TEMP\_AG.

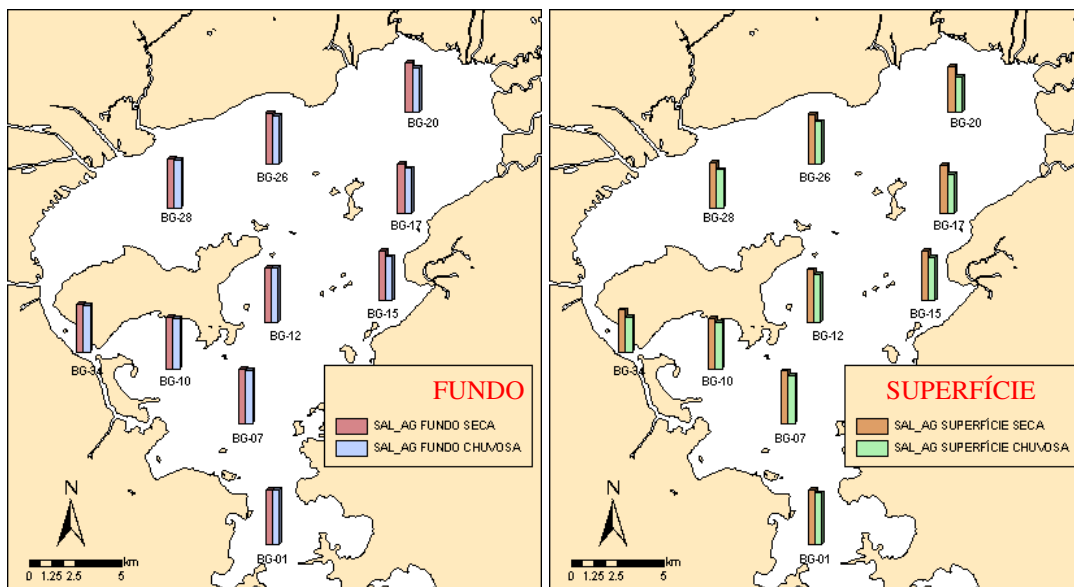


Figura 19: Mapa temático qualitativo com valores médios para SAL\_AG.

A Figura 20 apresenta o mapa das estações com os valores de oxigênio da água. Nas coletas de fundo, o OD\_AG manteve-se maior em todo o período de seca. Para as médias da superfície não foi observado padrão espacial; os valores alternavam ora para maiores valores no período de seca, ora para maiores valores no chuvoso.

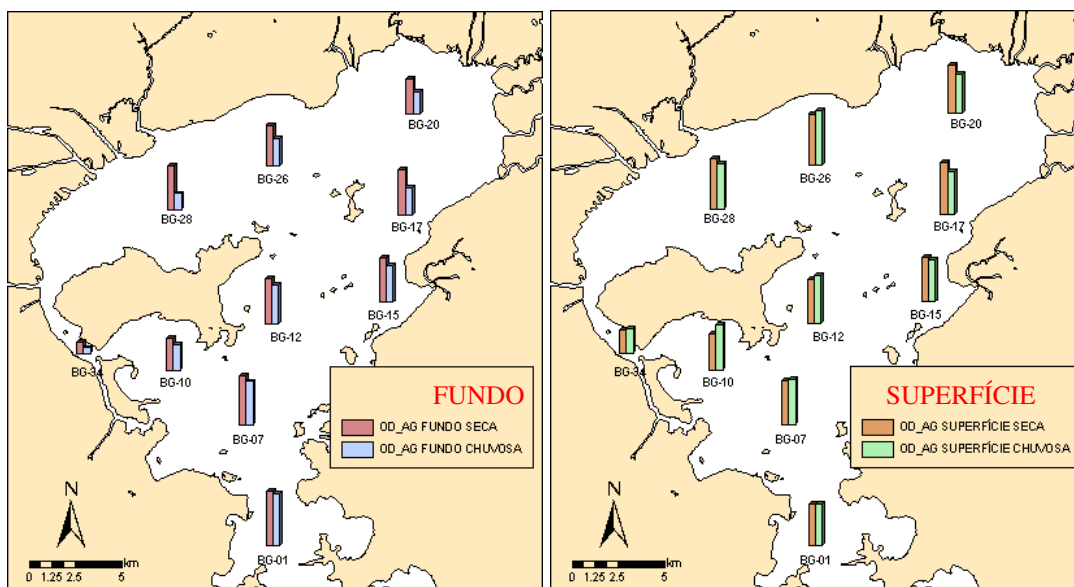


Figura 20: Mapa temático qualitativo com valores médios para OD\_AG.

O mapa da Figura 21, com os dados médio de Fósforo Total (FOS\_TOT\_AG), demonstram que para todo o período de coleta as maiores médias foram encontradas nas

bordas da baía (águas mais rasas de menor circulação/renovação). A estação BG-34 concentrou as maiores médias para essa variável.

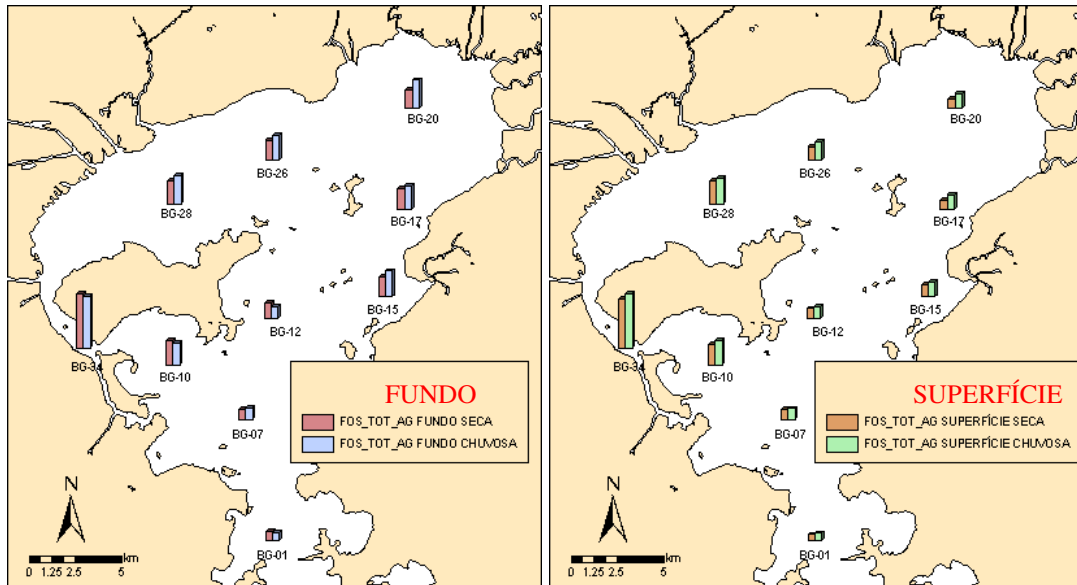


Figura 21: Mapa temático qualitativo com valores médios para FOS\_TOT\_AG.

A Figura 22, contendo a representação da Amônia (AMON\_AG), e a Figura 23, com o Nitrogênio Total (NIT\_TOT), apresentam as maiores médias associadas ao período de seca em todos os níveis de coleta (superfície e fundo). Uma observação espacial sugere ainda que as maiores médias em ambos os períodos e níveis de coleta estão na região oeste da baía, nas estações BG-34, BG-28 e BG-10.

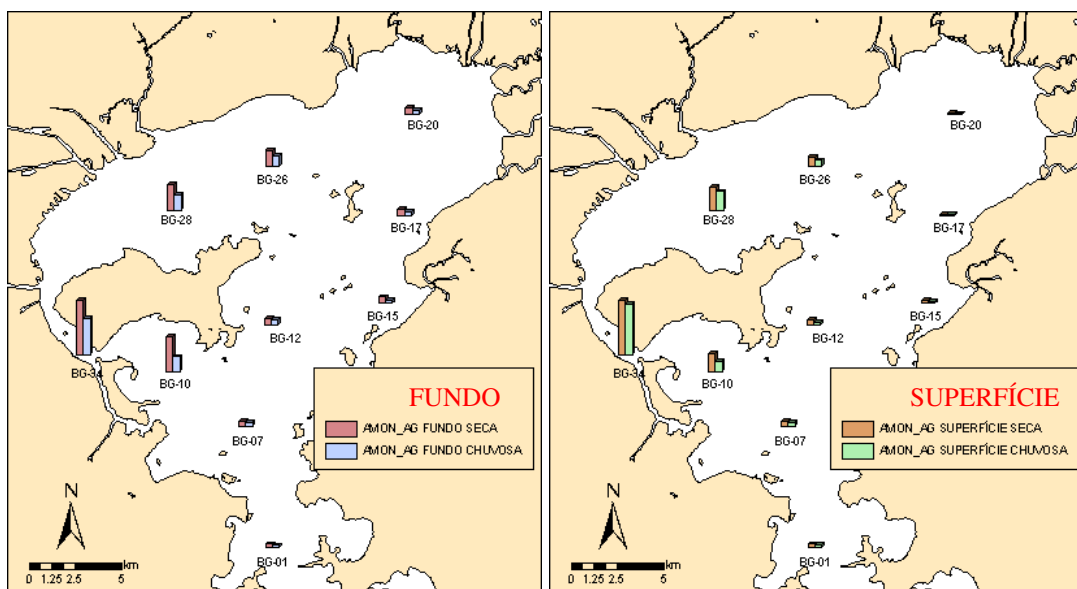


Figura 22: Mapa temático qualitativo com valores médios para AMON\_AG.

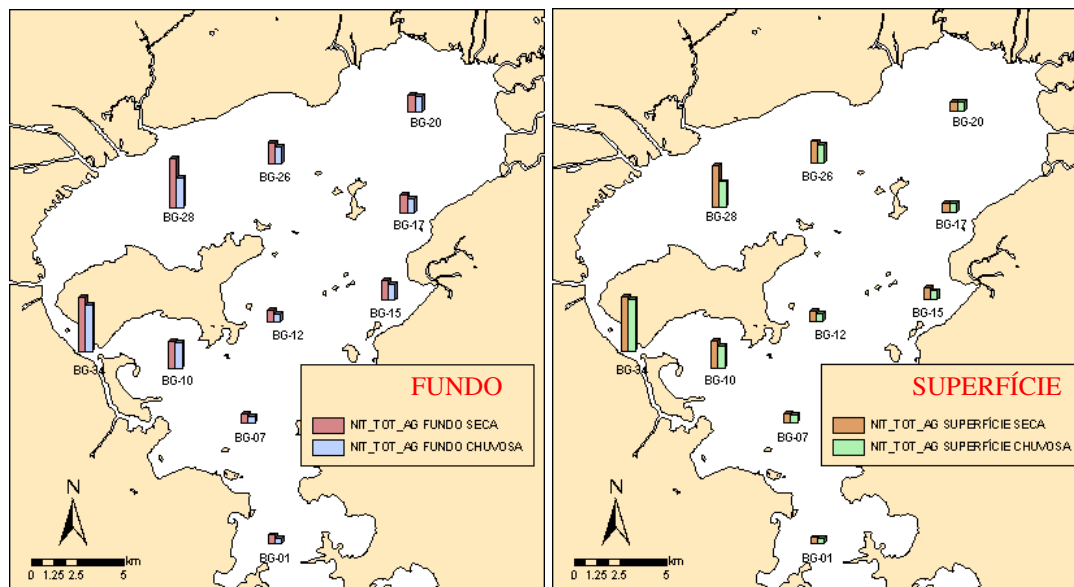


Figura 23: Mapa temático qualitativo com valores médios para NIT\_TOT\_AG.

O mapa de Material Particulado em Suspensão (MPS), exibido na Figura 24, apresentou visualmente desempenho similar entre as coletas realizadas no fundo e na superfície, sendo que os maiores valores são no período de seca. Espacialmente não foi observado padrão que mereça atenção.

O mapa da Figura 25, com os valores médios observados de Clorofila a (CLOR\_A), demonstra que os menores valores são encontrados no canal central, estações BG-01, BG-07 e BG-12, em ambos os níveis de coleta. Quanto à época do ano, pode-se dizer que o período de chuva teve maiores médias que a época de seca, contribuindo com o entendimento obtido pelos ANEXOS IV e V.

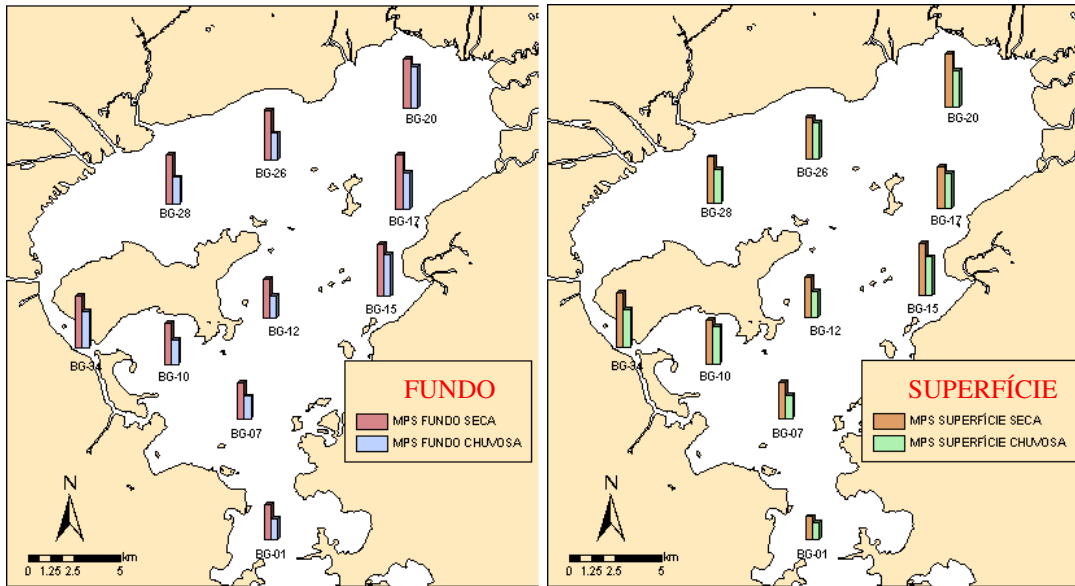


Figura 24: Mapa temático qualitativo com valores médios para MPS.

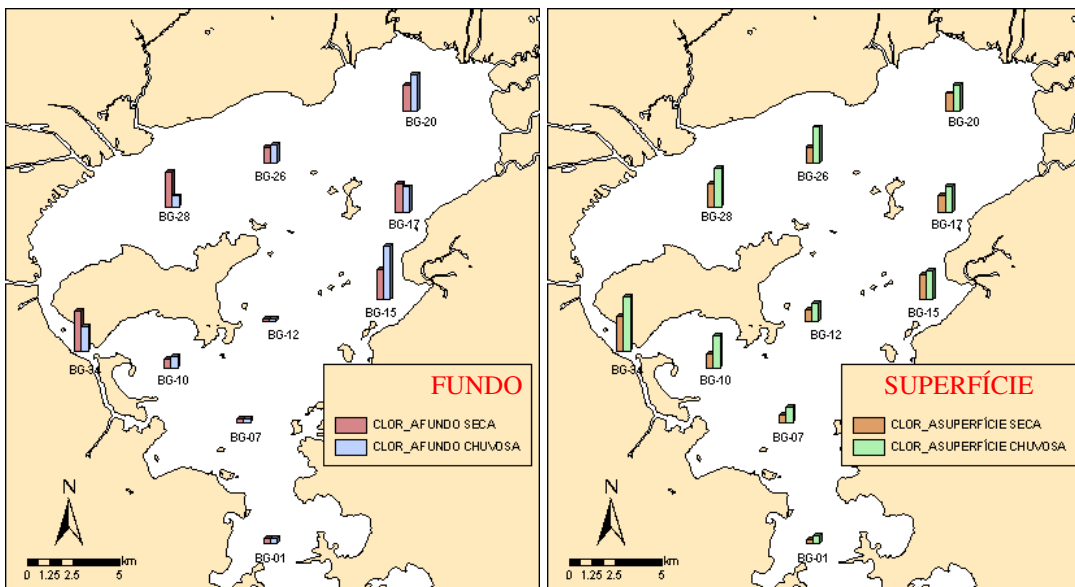


Figura 25: Mapa temático qualitativo com valores médios para CLOR\_A.

Os dados médios apresentados para a Abundancia Bacteriana (AB\_BAC) na Figura 26 mostram uma tendência onde os valores da época de chuva foram maiores que os valores da época de seca em ambos os níveis de coleta, com uma leve queda nas médias para a região do canal central.

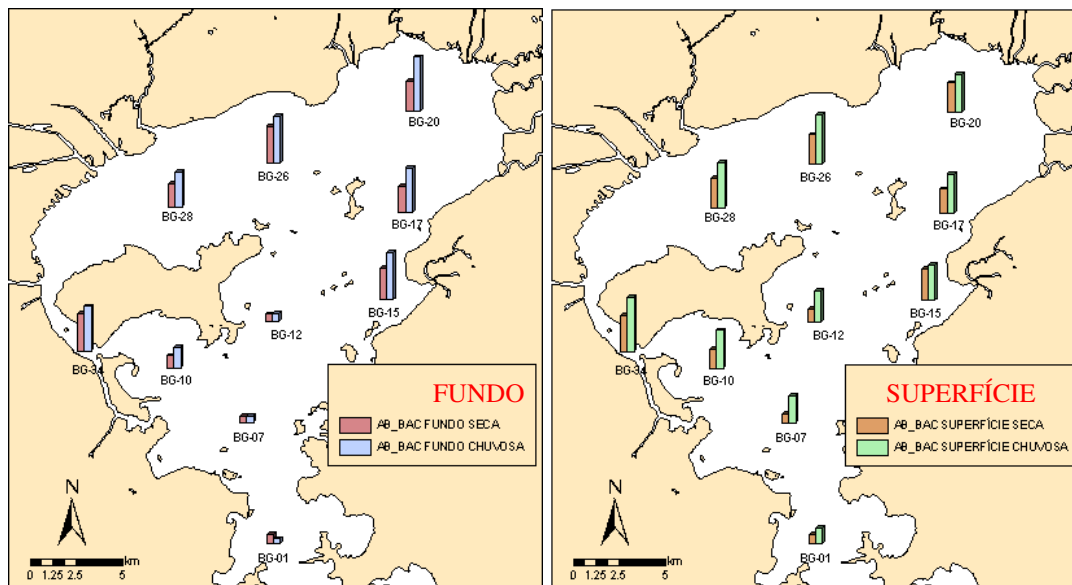


Figura 26: Mapa temático qualitativo com valores médios para AB\_BAC.

Todas essas observações contextualizam o cenário da área de estudo que tem como característica uma complexa geometria, com diversas ilhas em seu interior, aportes fluviais e de água marinha. Os resultados preliminares obtidos até o momento apresentam variações dos valores para as épocas do ano (temporal – seca e chuvosa) para o nível da estação (profundidade da coleta – fundo e superfície) assim como para o posicionamento das estações onde para a maioria dos parâmetros o comportamento das estações BG-01, BG-07 e BG-12 eram similares assim como para as estações BG-15, BG-17 e BG-20 porém as demais estações apresentam uma maior variação de comportamento quando comparada as demais estações.

A partir do entendimento obtido pela análise univariada (espaço-temporal), inicia-se a abordagem multivariada, ou seja, a avaliação da relação de dependência que pode existir entre as variáveis no espaço e no tempo.

### 3.6 Matriz de Correlação

A análise de correlação pode ser considerada como uma continuação da análise exploratória dos dados, tratando os dados de forma bivariada. Avalia-se a hipótese de dependência linear entre cada par de variável e suas interações básicas.

O produto da análise é obtido pelo coeficiente de correlação  $\rho$  de Spearman que indica o grau de relação entre as variáveis. Esse valor é obtido dividindo a covariância de duas variáveis pelo produto de seus desvios padrão. Quanto maior o valor absoluto da correlação calculada para duas variáveis, mais correlacionadas estas são, ou seja, a informação que elas trazem é muito parecida (redundância).

Podem existir dois tipos de correlação: positiva e negativa. O efeito é que, quando os valores de uma das variáveis aumentam, os da outra diminuem e vice e versa. Quando o valor da correlação é próximo ou igual a zero, as variáveis são pouco ou não correlacionadas, respectivamente.

Desta forma, a análise de correlação fornece informações significativas para as análises subseqüentes, sejam elas componentes principais, classificações, clusterizações, dentre outras.

A Tabela 6 apresenta a matriz de correlação criada sem levar em conta as dimensões mapeadas anteriormente (espaço e tempo), a qual demonstra um baixo potencial de correlação entre as variáveis. As correlações com valores mais próximos a 1 podem ser consideradas significativas.

O baixo potencial medido pelo coeficiente de correlação entre as variáveis para toda a base motivou a investigação da matrizes de correlação para cada um dos quatro (4) cenários do estudo identificado preliminarmente.

A matriz de correlação apresentada na Tabela 7 avalia o potencial de coeficiente de correlação entre as variáveis exploradas considerando somente os dados coletados na superfície no período da seca, onde se observa alguma correlação entre suas variáveis. FOS\_TOT e CLOR\_A foram os parâmetros de menor correlação entre si. Um total de 200 medições correspondem aos dados de coleta de superfície, no período de seca.

Com 280 medições, a matriz de correlação gerada para os dados das coletas de superfície para a época chuvosa (Tabela 8) mostra que as variáveis são menos correlacionadas quando comparadas com a matriz gerada para o período de seca.

As Tabelas 9 e 10 apresentam a matriz de correlação para os dados de coleta de fundo. O desempenho obtido pelo coeficiente de correlação entre as variáveis é similar aos dados de superfície, com um baixo potencial de correlação entre os dados.

Tabela 7: Matriz de correlação considerando todas as variáveis da base de dados. Correlação significativa para  $p < ,05000$ .

Variáveis	Matriz de correlação: BASE TOTAL N=960													
	PROF_EST	PROF_COL	TEM_AG	SAL_AG	PH	OD_AG	FOS_TOT	AMON_AG	NIT_TOT_AG	MPS	CLOR_A	AB_BAC	X	Y
<b>PROF_EST</b>	1	0.5957	-0.3185	0.4324	0.0504	0.1285	-0.2921	-0.164	-0.255	-0.2096	-0.2563	-0.4113	-0.1564	-0.6669
<b>PROF_COL</b>	0.5957	1	-0.4119	0.4649	-0.0661	-0.0857	-0.3136	-0.1646	-0.2566	-0.2123	-0.3123	-0.4333	-0.0931	-0.4023
<b>TEM_AG</b>	-0.3185	-0.4119	1	-0.5317	0.1163	0.0721	0.3502	0.1021	0.1547	0.0336	0.3613	0.5359	0	0.2659
<b>SAL_AG</b>	0.4324	0.4649	-0.5317	1	0.0094	-0.1228	-0.5062	-0.2969	-0.3268	-0.1696	-0.4564	-0.5295	-0.0476	-0.3962
<b>PH</b>	0.0504	-0.0661	0.1163	0.0094	1	0.5407	-0.2379	-0.3565	-0.1191	0.0966	0.2047	-0.0266	0.2963	0.0304
<b>OD_AG</b>	0.1285	-0.0857	0.0721	-0.1228	0.5407	1	-0.093	-0.2041	-0.0289	0.1699	0.3051	0.0599	0.2001	-0.0328
<b>FOS_TOT</b>	-0.2921	-0.3136	0.3502	-0.5062	-0.2379	-0.093	1	0.7383	0.7191	0.286	0.5763	0.4718	-0.4654	0.0773
<b>AMON_AG</b>	-0.164	-0.1646	0.1021	-0.2969	-0.3565	-0.2041	0.7383	1	0.6919	0.1288	0.2584	0.2287	-0.5487	-0.0393
<b>NIT_TOT_AG</b>	-0.255	-0.2566	0.1547	-0.3268	-0.1191	-0.0289	0.7191	0.6919	1	0.1264	0.398	0.2839	-0.4686	0.0719
<b>MPS</b>	-0.2096	-0.2123	0.0336	-0.1696	0.0966	0.1699	0.286	0.1288	0.1264	1	0.3852	0.138	0.0398	0.1729
<b>CLOR_A</b>	-0.2563	-0.3123	0.3613	-0.4564	0.2047	0.3051	0.5763	0.2584	0.398	0.3852	1	0.4436	-0.0777	0.1641
<b>AB_BAC</b>	-0.4113	-0.4333	0.5359	-0.5295	-0.0266	0.0599	0.4718	0.2287	0.2839	0.138	0.4436	1	0.001	0.3536
<b>X</b>	-0.1564	-0.0931	0	-0.0476	0.2963	0.2001	-0.4654	-0.5487	-0.4686	0.0398	-0.0777	0.001	1	0.3823
<b>Y</b>	-0.6669	-0.4023	0.2659	-0.3962	0.0304	-0.0328	0.0773	-0.0393	0.0719	0.1729	0.1641	0.3536	0.3823	1

Tabela 8: Apresenta a matriz de correlação considerando somente os dados de coleta na superfície, no período de seca. Correlação significativa para  $p < ,05000$ .

Variáveis	Matriz de correlação: SUPERFÍCIE SECA N=200												
	PROF_EST	TEM_AG	SAL_AG	PH	OD_AG	FOS_TOT	AMON_AG	NIT_TOT_AG	MPS	CLOR_A	AB_BAC	X	Y
<b>PROF_EST</b>	1	-0.0194	0.4736	0.0418	-0.0577	-0.2848	-0.1822	-0.2553	-0.2029	-0.3193	-0.5082	-0.1388	-0.6587
<b>TEM_AG</b>	-0.0194	1	-0.0162	0.2184	0.1505	0.2121	0.0948	0.0834	0.0023	0.1677	0.1254	-0.2766	-0.05
<b>SAL_AG</b>	0.4736	-0.0162	1	0.3204	0.0455	-0.5774	-0.5148	-0.4064	-0.3139	-0.3983	-0.4503	0.1902	-0.4308
<b>PH</b>	0.0418	0.2184	0.3204	1	0.634	-0.3569	-0.4697	-0.1683	0.1195	0.1605	-0.1716	0.3676	0.075
<b>OD_AG</b>	-0.0577	0.1505	0.0455	0.634	1	-0.2608	-0.4194	-0.0961	0.0552	0.2673	-0.0225	0.2837	0.2134
<b>FOS_TOT</b>	-0.2848	0.2121	-0.5774	-0.3569	-0.2608	1	0.8364	0.6659	0.2139	0.4234	0.3553	-0.6622	-0.008
<b>AMON_AG</b>	-0.1822	0.0948	-0.5148	-0.4697	-0.4194	0.8364	1	0.666	0.0663	0.2117	0.1685	-0.6749	-0.0706
<b>NIT_TOT_AG</b>	-0.2553	0.0834	-0.4064	-0.1683	-0.0961	0.6659	0.666	1	0.019	0.2325	0.1333	-0.5389	0.0668
<b>MPS</b>	-0.2029	0.0023	-0.3139	0.1195	0.0552	0.2139	0.0663	0.019	1	0.4431	0.0681	-0.0014	0.1931
<b>CLOR_A</b>	-0.3193	0.1677	-0.3983	0.1605	0.2673	0.4234	0.2117	0.2325	0.4431	1	0.3653	-0.1327	0.1901
<b>AB_BAC</b>	-0.5082	0.1254	-0.4503	-0.1716	-0.0225	0.3553	0.1685	0.1333	0.0681	0.3653	1	-0.0319	0.3888
<b>X</b>	-0.1388	-0.2766	0.1902	0.3676	0.2837	-0.6622	-0.6749	-0.5389	-0.0014	-0.1327	-0.0319	1	0.3823
<b>Y</b>	-0.6587	-0.05	-0.4308	0.075	0.2134	-0.008	-0.0706	0.0668	0.1931	0.1901	0.3888	0.3823	1

Tabela 9: Apresenta a matriz de correlação considerando somente os dados de coleta na superfície, no período de chuva. Correlação significativa para  $p < ,05000$ .

Variáveis	Matriz Correlação SUPERFÍCIE CHUVOSA												
	N=280												
	PROF_EST	TEM_AG	SAL_AG	PH	OD_AG	FOS_TOT	AMON_AG	NIT_TOT_AG	MPS	CLOR_A	AB_BAC	X	Y
PROF_EST	1	-0.3444	0.4559	0.0585	0.0457	-0.3037	-0.156	-0.2718	-0.3066	-0.2882	-0.3253	-0.1669	-0.6715
TEM_AG	-0.3444	1	-0.354	0.126	0.2213	0.3085	0.1659	0.2567	0.2229	0.3192	0.4561	-0.0386	0.3223
SAL_AG	0.4559	-0.354	1	0.1382	-0.0212	-0.3811	-0.3094	-0.3523	-0.1057	-0.257	-0.3172	-0.0201	-0.479
PH	0.0585	0.126	0.1382	1	0.4863	-0.3095	-0.4026	-0.1711	0.0634	0.162	-0.1387	0.1748	0.0163
OD_AG	0.0457	0.2213	-0.0212	0.4863	1	-0.2175	-0.2784	-0.1983	0.249	0.1732	-0.015	0.1156	0.0704
FOS_TOT	-0.3037	0.3085	-0.3811	-0.3095	-0.2175	1	0.7776	0.8402	0.3315	0.5829	0.4603	-0.5839	0.0571
AMON_AG	-0.156	0.1659	-0.3094	-0.4026	-0.2784	0.7776	1	0.7996	0.069	0.3006	0.3362	-0.6205	-0.0611
NIT_TOT_AG	-0.2718	0.2567	-0.3523	-0.1711	-0.1983	0.8402	0.7996	1	0.1756	0.5465	0.4088	-0.6107	0.0433
MPS	-0.3066	0.2229	-0.1057	0.0634	0.249	0.3315	0.069	0.1756	1	0.509	0.2708	0.004	0.2369
CLOR_A	-0.2882	0.3192	-0.257	0.162	0.1732	0.5829	0.3006	0.5465	0.509	1	0.3605	-0.2189	0.1873
AB_BAC	-0.3253	0.4561	-0.3172	-0.1387	-0.015	0.4603	0.3362	0.4088	0.2708	0.3605	1	-0.169	0.3021
X	-0.1669	-0.0386	-0.0201	0.1748	0.1156	-0.5839	-0.6205	-0.6107	0.004	-0.2189	-0.169	1	0.3823
Y	-0.6715	0.3223	-0.479	0.0163	0.0704	0.0571	-0.0611	0.0433	0.2369	0.1873	0.3021	0.3823	1

Tabela 10: Apresenta a matriz de correlação considerando somente os dados de coleta de fundo para o período de seca. Correlação significativa para  $p < ,05000$ .

Variáveis	Matriz de correlação: FUNDO SECO N=200												
	PROF_EST	TEM_AG	SAL_AG	PH	OD_AG	FOS_TOT	AMON_AG	NIT_TOT_AG	MPS	CLOR_A	AB_BAC	X	Y
<b>PROF_EST</b>	1	0.007	0.6508	0.1567	0.2253	-0.3213	-0.2028	-0.3276	-0.1437	-0.3309	-0.4102	-0.1453	-0.6616
<b>TEM_AG</b>	0.007	1	-0.0357	-0.0309	-0.1491	0.1684	0.0464	0.0046	-0.1133	0.0295	0.0663	-0.1823	0.0051
<b>SAL_AG</b>	0.6508	-0.0357	1	0.256	0.267	-0.441	-0.2563	-0.238	-0.232	-0.3534	-0.4365	-0.0387	-0.5237
<b>PH</b>	0.1567	-0.0309	0.256	1	0.5532	-0.5187	-0.3574	-0.2564	-0.0503	-0.0093	-0.123	0.3905	-0.0369
<b>OD_AG</b>	0.2253	-0.1491	0.267	0.5532	1	-0.4898	-0.3259	-0.2682	-0.1399	-0.032	-0.0863	0.3152	-0.1129
<b>FOS_TOT</b>	-0.3213	0.1684	-0.441	-0.5187	-0.4898	1	0.4612	0.5474	0.3229	0.2272	0.3087	-0.4476	0.1016
<b>AMON_AG</b>	-0.2028	0.0464	-0.2563	-0.3574	-0.3259	0.4612	1	0.4413	0.1086	0.0479	0.1356	-0.4636	-0.004
<b>NIT_TOT_AG</b>	-0.3276	0.0046	-0.238	-0.2564	-0.2682	0.5474	0.4413	1	-0.0331	0.1026	0.1408	-0.4363	0.1458
<b>MPS</b>	-0.1437	-0.1133	-0.232	-0.0503	-0.1399	0.3229	0.1086	-0.0331	1	0.2279	-0.0327	0.0352	0.1189
<b>CLOR_A</b>	-0.3309	0.0295	-0.3534	-0.0093	-0.032	0.2272	0.0479	0.1026	0.2279	1	0.177	-0.0136	0.2239
<b>AB_BAC</b>	-0.4102	0.0663	-0.4365	-0.123	-0.0863	0.3087	0.1356	0.1408	-0.0327	0.177	1	0.045	0.3333
<b>X</b>	-0.1453	-0.1823	-0.0387	0.3905	0.3152	-0.4476	-0.4636	-0.4363	0.0352	-0.0136	0.045	1	0.3823
<b>Y</b>	-0.6616	0.0051	-0.5237	-0.0369	-0.1129	0.1016	-0.004	0.1458	0.1189	0.2239	0.3333	0.3823	1

Tabela 11: Apresenta a matriz de correlação considerando somente os dados de coleta de fundo para o período de chuva. Correlação significativa para  $p < ,05000$ .

Variáveis	Matriz de correlação: FUNDO CHUVOSO N=280												
	PROF_EST	TEM_AG	SAL_AG	PH	OD_AG	FOS_TOT	AMON_AG	NIT_TOT_AG	MPS	CLOR_A	AB_BAC	X	Y
<b>PROF_EST</b>	1	-0.6817	0.6813	-0.0325	0.4496	-0.4979	-0.2124	-0.3144	-0.2997	-0.4098	-0.5981	-0.1666	-0.6721
<b>TEM_AG</b>	-0.6817	1	-0.6022	0.0047	-0.3405	0.4699	0.2064	0.2334	0.2271	0.3391	0.6162	0.1707	0.5626
<b>SAL_AG</b>	0.6813	-0.6022	1	-0.2306	0.1587	-0.4637	-0.1398	-0.2009	-0.386	-0.492	-0.5696	-0.2968	-0.562
<b>PH</b>	-0.0325	0.0047	-0.2306	1	0.4322	-0.2079	-0.3548	-0.1478	0.149	0.2313	0.0691	0.3674	0.064
<b>OD_AG</b>	0.4496	-0.3405	0.1587	0.4322	1	-0.5159	-0.4422	-0.3222	0.007	0.0762	-0.2164	0.2828	-0.3949
<b>FOS_TOT</b>	-0.4979	0.4699	-0.4637	-0.2079	-0.5159	1	0.7079	0.5808	0.3273	0.4318	0.4797	-0.2849	0.3033
<b>AMON_AG</b>	-0.2124	0.2064	-0.1398	-0.3548	-0.4422	0.7079	1	0.6117	0.114	0.0274	0.1525	-0.5832	0.0157
<b>NIT_TOT_AG</b>	-0.3144	0.2334	-0.2009	-0.1478	-0.3222	0.5808	0.6117	1	0.0712	0.1469	0.2997	-0.385	0.1158
<b>MPS</b>	-0.2997	0.2271	-0.386	0.149	0.007	0.3273	0.114	0.0712	1	0.5249	0.3467	0.1888	0.2279
<b>CLOR_A</b>	-0.4098	0.3391	-0.492	0.2313	0.0762	0.4318	0.0274	0.1469	0.5249	1	0.4948	0.2881	0.268
<b>AB_BAC</b>	-0.5981	0.6162	-0.5696	0.0691	-0.2164	0.4797	0.1525	0.2997	0.3467	0.4948	1	0.1953	0.5298
<b>X</b>	-0.1666	0.1707	-0.2968	0.3674	0.2828	-0.2849	-0.5832	-0.385	0.1888	0.2881	0.1953	1	0.3823
<b>Y</b>	-0.6721	0.5626	-0.562	0.064	-0.3949	0.3033	0.0157	0.1158	0.2279	0.268	0.5298	0.3823	1

Uma forma de interpretar os resultados obtidos pela correlação de Pearson é a relação apresentada na Tabela 11. Comparando os valores das matrizes obtidas com a proposta da tabela, concluímos que em sua maioria, a correlação entre as variáveis é relativamente fraca para todos os cenários avaliados.

Tabela 12: Interpretação para as faixas de valores obtidas em  $p$  pela matriz de correlação.

<i>Valor de p (+ ou -)</i>	<i>Interpretação</i>
<i>0,00 a 0,19</i>	correlação bem fraca
<i>0,20 a 0,39</i>	correlação fraca
<i>0,40 a 0,69</i>	correlação moderada
<i>0,70 a 0,89</i>	correlação forte
<i>0,90 a 1,00</i>	correlação muito forte

### 3.7 Análise dos Componente Principais – ACP

Análise de Componentes Principais (ACP) é uma técnica que consiste em tentar reduzir o numero de variáveis, sendo que essas novas variáveis devem conter o máximo possível de informação das variáveis originais, sendo bastante utilizada na tentativa de reduzir a dimensão dos dados.

A análise dos componentes retorna um pequeno número de combinações lineares (componentes principais) do conjunto de variáveis. Esses componentes devem manter o máximo possível da informação contida nas variáveis originais. Os componentes são extraídos na ordem do mais explicativo para o menos explicativo. Teoricamente o número de componentes é sempre igual ao número de variáveis, entretanto, alguns poucos componentes são responsáveis por grande parte da explicação total do problema, podendo ser utilizados na análise em substituição às variáveis originais.

O Gráfico 1 apresenta os autovalores calculados, no eixo Y da esquerda estão às variâncias e no eixo Y da direita estão os valores acumulativos. A linha acima dos blocos representa a soma proporcional dos valores acumulativos, o eixo X o autovalor.

Observa-se que a primeira variável, do novo sistema de coordenadas, representa 30 % da variância total dos dados.

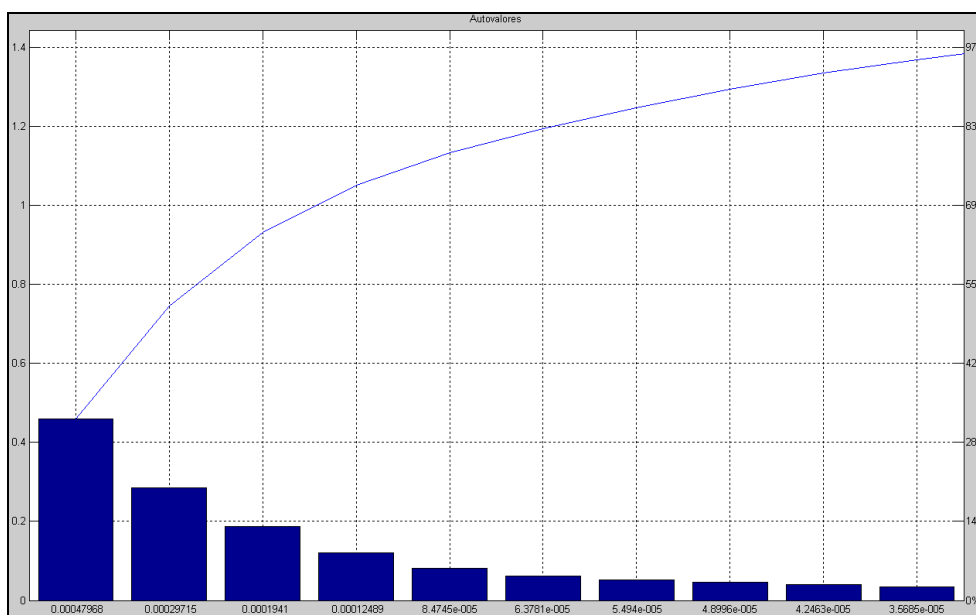


Gráfico 1: Gráfico de Pareto, que tem por finalidade obter melhor visualização quando se necessita priorizar diversos itens, descrevendo a contribuição de cada fator de correlação entre as variáveis.

Avaliando os 10 primeiros componentes, chega-se a 96% da variância total dos dados, uma redução pouco significativa comparando com a dimensão original dos dados.

O entendimento da composição estrutural dos componentes principais a partir do gráfico de Pareto (Gráfico 1) para autovalores calculados, identifica uma baixa possibilidade de simplificação no dimensionamento das variáveis. A baixa inter-relação entre todas as variáveis, simultaneamente, é confirmada no Gráfico 2, onde os valores em sua maioria apresentam-se alternados e distantes da meta ( $\pm 1$ ), considerando todos os atributos.

A Tabela 12 apresenta a avaliação das componentes principais, com os autovalores e a contribuição das diferentes variáveis para o cálculo. O eixo vertical lista as variáveis e o eixo horizontal indica as componentes principais.

A primeira componente “Fator 1” é composta de associações positivas e negativas entre as variáveis, sendo que o OD\_AG é a variável que apresenta menor contribuição positiva (+). As demais variáveis não são esclarecedoras, apresentando uma baixa

correlação positiva para as variáveis PROF\_EST, SAL\_AG, PH, OD e X e negativa para: TEMP\_AG, FOS\_TOT, AMON\_AG, NIT\_TOT\_AG, MPS, CLOR\_A, AB\_BAC e Y.

O entendimento do resultando obtido pela análise dos componentes principais não sugestia eficiência em trabalhar com a nova matriz de dados gerada pelos autovalores. Assim o uso das variáveis originais foi mantido nos processos de análises seqüentes.

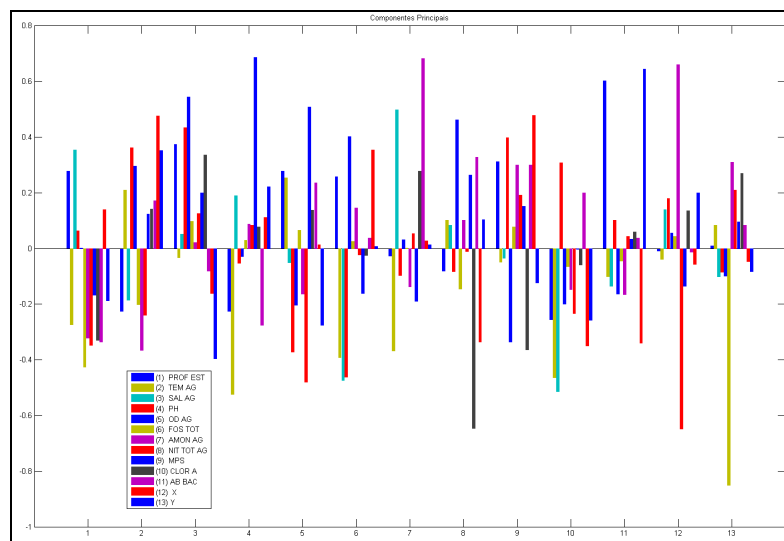


Gráfico 2: Correlação entre as variáveis. Observa-se a grande alternância nas barras sugerindo baixa correlação entre as variáveis estudadas.

Tabela 13: Apresenta os valores para as componentes principais.

Variável	FATOR 1	FATOR 2	FATOR 3	FATOR 4	FATOR 5	FATOR 6	FATOR 7	FATOR 8	FATOR 9	FATOR 10	FATOR 11	FATOR 12	FATOR 13
(1) PROF_EST	0.2781	-0.2263	0.3747	-0.227	0.2783	0.2591	-0.0244	-0.0781	0.312	-0.2575	0.6037	-0.0067	0.0102
(2) TEMP_AG	-0.2745	0.2115	-0.0307	-0.5257	0.2541	-0.3932	-0.3684	0.1022	-0.0462	-0.4654	-0.0995	-0.0372	0.0854
(3) SAL_AG	0.356	-0.1871	0.0533	0.1905	-0.0494	-0.4745	0.4995	0.0859	-0.0323	-0.5141	-0.1371	0.1406	-0.1037
(4) PH	0.0655	0.3621	0.434	-0.0515	-0.3732	-0.4628	-0.0955	-0.0801	0.3982	0.3099	0.103	0.1806	-0.0822
(5) OD_AG	0.0024	0.2962	0.5451	-0.027	-0.2048	0.4021	0.0325	0.4639	-0.3377	-0.2015	-0.1645	0.0574	-0.0977
(6) FOS_TOT	-0.4268	-0.2027	0.0986	0.0309	0.0661	0.0273	0.0033	-0.1469	0.0789	-0.0623	-0.0426	0.0456	-0.8517
(7) AMON_AG	-0.3227	-0.3664	0.0221	0.0882	-0.1641	0.1472	-0.1386	0.104	0.3017	-0.1489	-0.1663	0.6612	0.3108
(8) NIT_TOT_AG	-0.3489	-0.24	0.1275	0.0853	-0.4813	-0.0218	0.0547	-0.0082	0.1929	-0.2357	0.045	-0.6499	0.2114
(9) MPS	-0.1686	0.1242	0.2019	0.6862	0.5091	-0.1633	-0.1906	0.2641	0.1524	-0.002	0.036	-0.1368	0.0964
(10) CLOR_A	-0.3319	0.142	0.3361	0.079	0.1392	-0.0221	0.2797	-0.6474	-0.3657	-0.0577	0.0614	0.1364	0.2717
(11) AB_BAC	-0.338	0.1731	-0.0787	-0.277	0.2365	0.0398	0.6827	0.3298	0.3016	0.2019	0.0381	-0.0109	0.0851
(12) X	0.1417	0.477	-0.1638	0.1137	0.015	0.3542	0.0295	-0.3365	0.4783	-0.3514	-0.3401	-0.0543	-0.0446
(13) Y	-0.1883	0.3524	-0.3963	0.2227	-0.2768	0.0094	0.0157	0.106	-0.1256	-0.2588	0.6445	0.2019	-0.0809

### 3.8 Classificação Supervisionada

A etapa de classificação supervisionada dos dados fez uso de um conhecimento *a priori* para estabelecer as classes de qualidade de água para cada estação de coleta. Em 1989, Mayr *et al.*, realizou uma avaliação propondo uma divisão da baía em cinco áreas, adotando critérios de avaliação para as condições químicas da água e grau de poluição associados a respostas biológicas.

A proposta de divisão de Mayr *et al.* (1989) é apresentada na Figura 27 e aqui descrita por Villac (1990) com a seguinte proposta de divisão:

Área 1: Abrange a região do canal central.

Área 2: Compreende principalmente as enseadas de Botafogo e Jurujuba, as quais são as mais estudadas, destacando-se a região da Urca.

Área 3: Abrange a região leste da Ilha do Governador e o litoral leste da baía, na altura de São Gonçalo foi menos estudada. Estas duas regiões encontram-se lado a lado do canal central e provavelmente sofrem influência indireta das águas oceânicas que penetram na baía através deste canal.

Área 4: Abrange a região do fundo da baía ao norte e leste de Paquetá e Ilha do Governador, é a menos estudada. Apesar de existirem costões rochosos nesta área e a presença marcante de manguezais, a área 4 (juntamente com a Ilha de Paquetá - área 1 e a área 3), está sujeita ao impacto direto de possíveis vazamentos dos oleodutos presentes na baía.

Área 5: Abrange a região oeste da Baía de Guanabara, foi mais bem estudada que a área 4 e é considerada a mais degradada em termos de poluição orgânica e industrial.

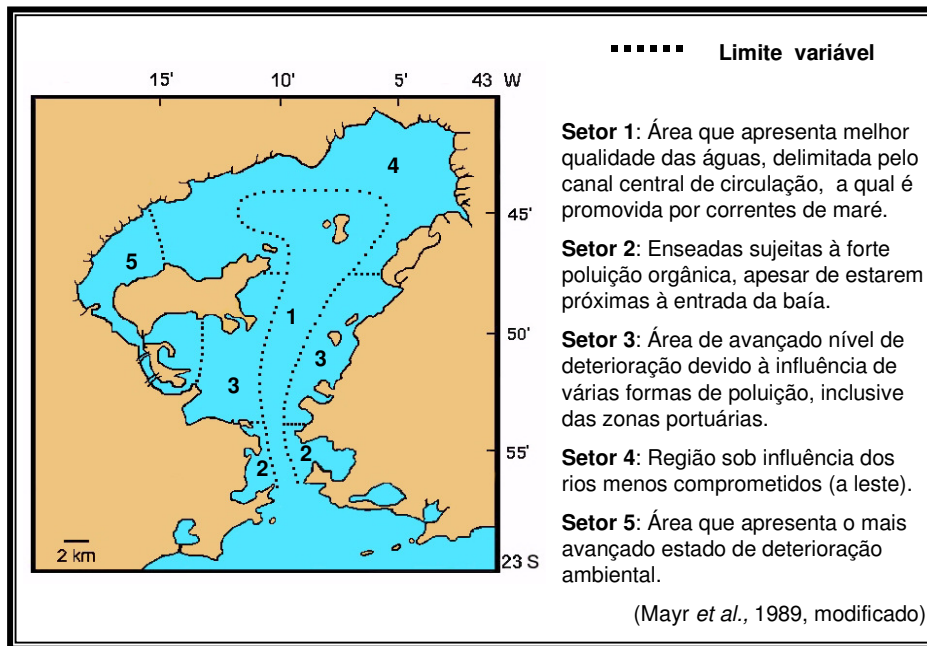


Figura 27: Representação das cinco áreas de qualidade ambiental proposta por Mayer *et al.* (1989) e adotada no presente estudo como conhecimento *a priori*.

Para o desenvolvimento da análise, a informação de classificação da qualidade da água foi incorporada na base de dados do presente trabalho, após realizado a digitalização e o georeferenciamento do mapa de Mayer *et al.* (1989). Com isso, uma análise espacial (*spatial join*) foi estabelecida entre o mapa criado e as estações de coleta (pontos). O produto desta análise de sobreposição é o acréscimo da informação de classes definida por Mayer *et al.* (1989) na base de dados. A Figura 28 apresenta o posicionamento destas estações sob as classes e a distância calculada da estação até a “borda” da classe mais próxima.

A Figura 29 representa de forma temática as estações de coleta do projeto e a sua associação nas classes definidas por Mayer *et al.*, (1989). Nota-se que a classe 2 não tem associação com malha amostral do projeto. Pela falta de representação espacial da malha do projeto para a área 2, esta classe foi desconsiderada como conhecimento *a priori*, não sendo considerada para o teste de classificação. Como parte da contextualização do cenário de estudo, as faixas batimétricas são também apresentadas pelo mapa da Figura 29.

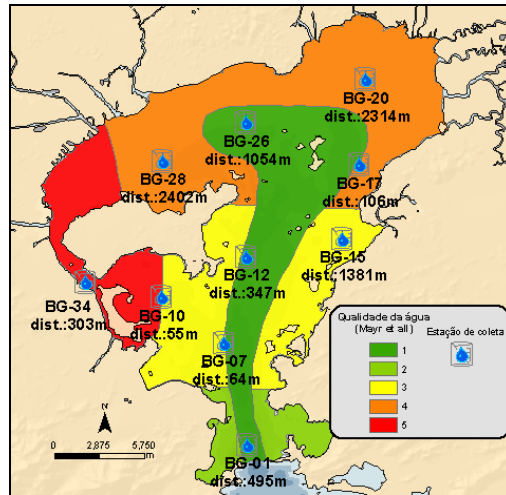


Figura 28: Posicionamento das estações de coleta em relação ao mapa de classes de qualidade da água (Mayr *et al.*, 1989), e a distância de cada estação a borda mais próxima da classe do mapa.

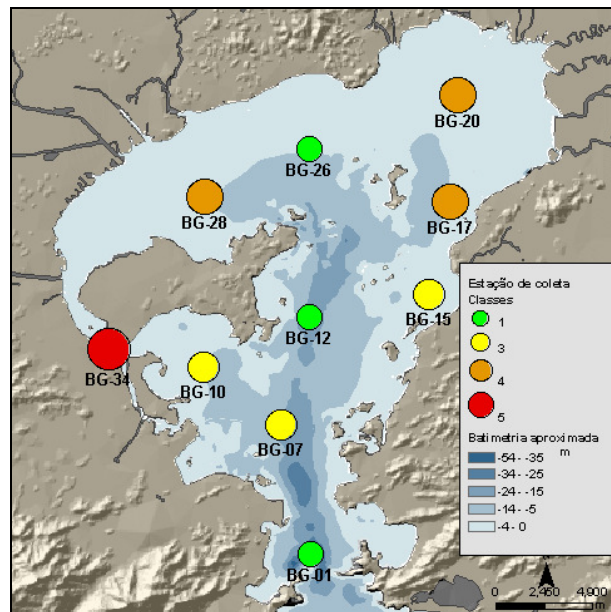


Figura 29: Mapa apresentando o resultado da análise espacial com a distribuição das classes (conhecimento *a priori*) para as estações.

O alinhamento estabelecido entre a metodologia utilizada por Mayr *et al.*, (1989) e a desenvolvida neste trabalho só tem potencial de comparação para os dados de coleta de superfície.

A Tabela 13 apresenta a quantidade de medições por classe estabelecida. A Classe 5 é representada somente pela estação BG-34, tendo apenas 48 medições, uma

medida a cada unidade de tempo (campanha de coleta); as demais classes são representadas por 3 estações cada, totalizando 144 medidas em cada classe.

Antes da avaliação do algoritmo de classificação realizou-se uma nova abordagem univariada descrevendo as faixas dos valores para cada classe estabelecida por Mayr *et al.*, 1989, a partir de gráficos de histograma com *plot* da curva normal. O eixo Y do lado direito apresenta a porcentagem de ocorrência do valor, o eixo Y do lado esquerdo representa a quantidade de ocorrência e o eixo X descreve os valores das variáveis associadas à classe correspondente.

A partir da análise visual do Gráfico 3, onde a temperatura da água (TEMP\_AG) é representada, observa-se um pequeno acréscimo no valor da temperatura no sentido da Classe 1 para a Classe 5, porém com pouca diferença nos valores das maiores frequências, entre as classes.

Seguindo o padrão de observação analítica, no Gráfico 4 para os dados de OD\_AG observa-se que a Classe 5 tem sua maior frequência para os menores valores, e as outras classes não apresentaram variações de comportamento significativas.

Tabela 14: Apresenta a sumarização considerando o número de estações e a quantidade de medições para cada classe.

Classes	Quantidade de Estações	Total de medições	Estações associadas
C= 1	3	144	BG-01, BG-12, BG-26
C= 3	3	144	BG-07, BG-10, BG-15
C= 4	3	144	BG-17, BG-20, BG-28
C= 5	1	48	BG-34

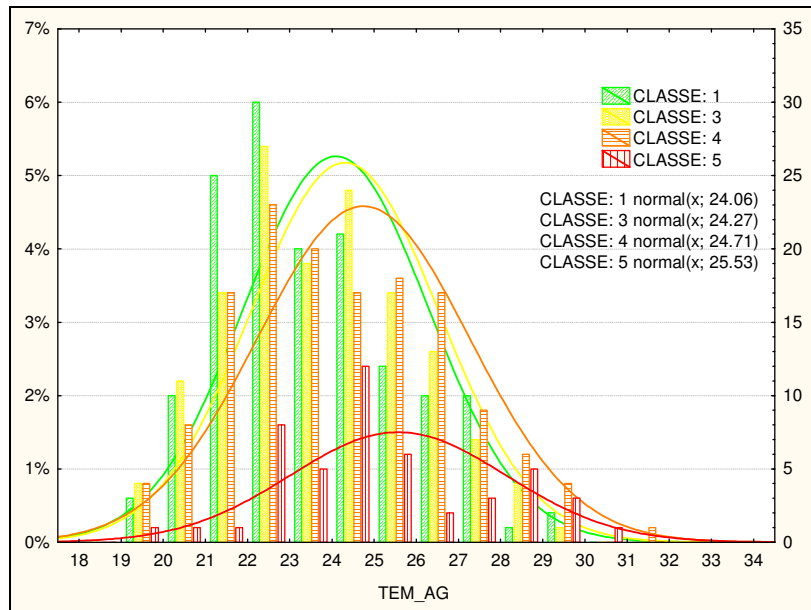


Gráfico 3: Descreve as informações coletadas para parâmetro de temperatura da água na superfície agrupado pelas classes estabelecidas pelo projeto referência.

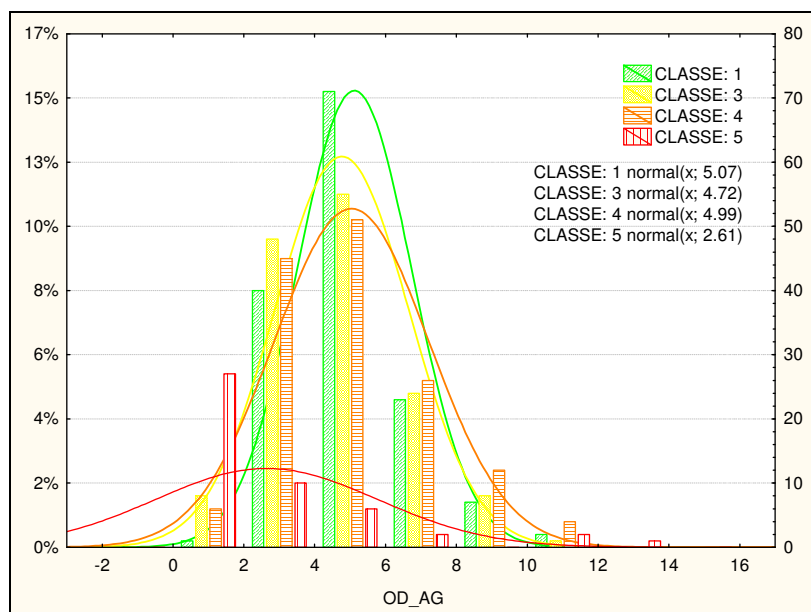


Gráfico 4: Histograma gerado a partir dos dados de OD\_AG.

O histograma da variável SAL\_AG, apresentado no Gráfico 5, mostra que as classes 4 e 5 representam as maiores freqüência nos menores valores da distribuição, e as classes 3 e 1 apresentam padrão similar de comportamento, avaliando a distribuição de sua freqüência junto com a curva da normal.

O gráfico associado aos dados de fósforo total (Gráfico 6) apresenta a classe 5 com um comportamento bastante distinto das outras classes, com faixas de valores e maiores freqüências associadas aos maiores valores. As classes 3 e 4 apresentam uma similaridade de comportamento associados a faixa de seus valores mais freqüentes. Já a classe 1 apresenta sua maior freqüência de valor concentrados sob a mesma faixa de valor.

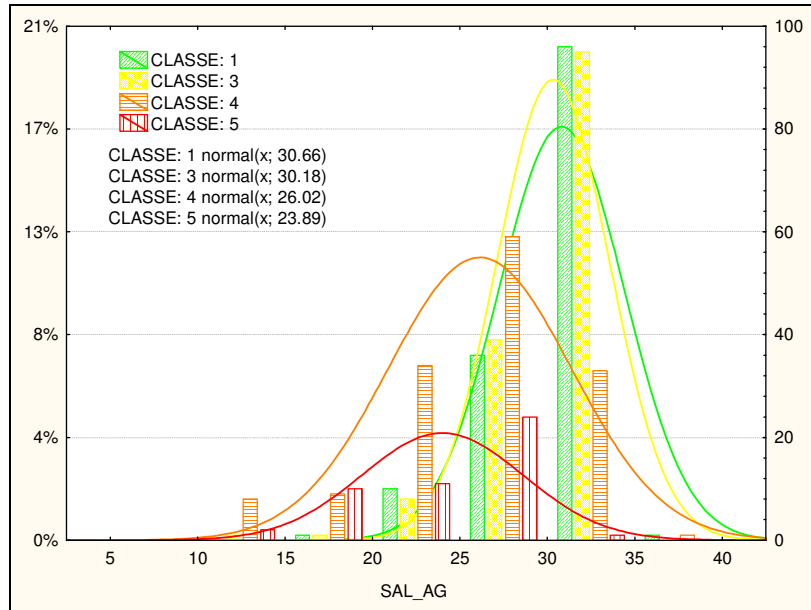


Gráfico 5: Histograma gerado a partir dos dados de SAL\_AG para as coletas de superfície.

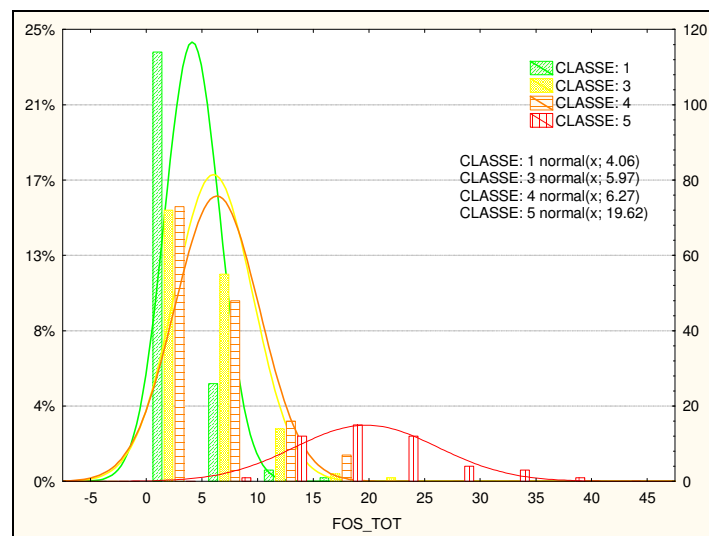


Gráfico 6: Histograma apresentado os dados de FOS\_TOT.

Na avaliação do gráfico de histograma de pH (Gráfico 7) nota-se que as classes 1, 3 e 4 apresentam sua distribuição de valores semelhantes para as faixas de frequência de ocorrência. A classe 5 apresenta suas maiores frequências nas faixas de menores valores de ocorrências.

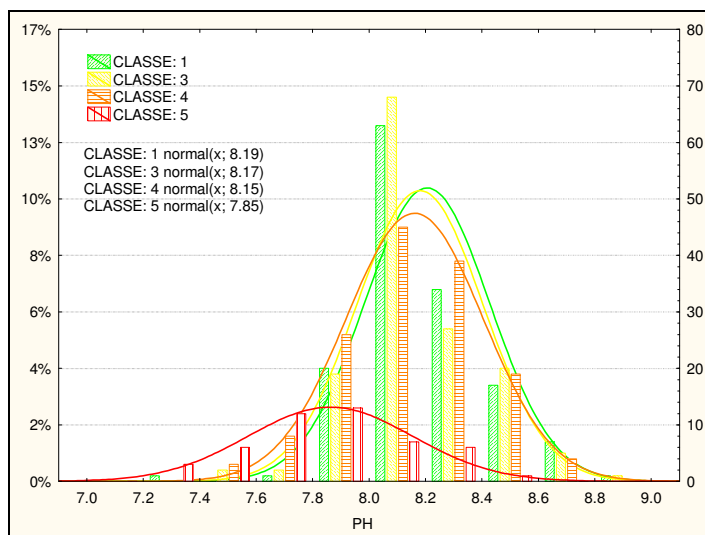


Gráfico 7: Gráfico de histograma gerado para os dados do parâmetro pH agrupados pelas classes de água.

O histograma de AMON\_AG (Gráfico 8) mostra uma alta frequência de ocorrência para os baixos valores nas Classes 1, 3 e 4. A Classe 5 tem sua maior frequência para os maiores valores.

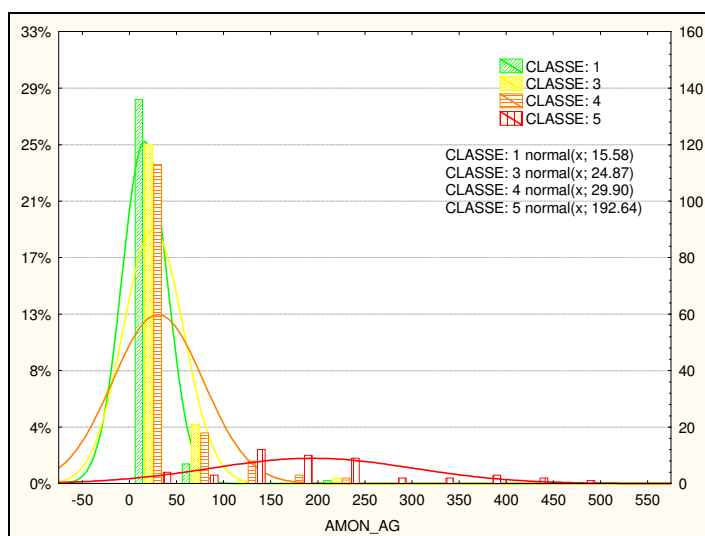


Gráfico 8: Histograma da frequência de ocorrência do parâmetro de AMON\_AG.

O nitrogênio total da água (Gráfico 9) segue um comportamento similar a outros parâmetros já avaliados, onde a classe 5 tem um comportamento distinto das outras, com a frequência média de valores ora maior, ora menor, que todas as outras classes.

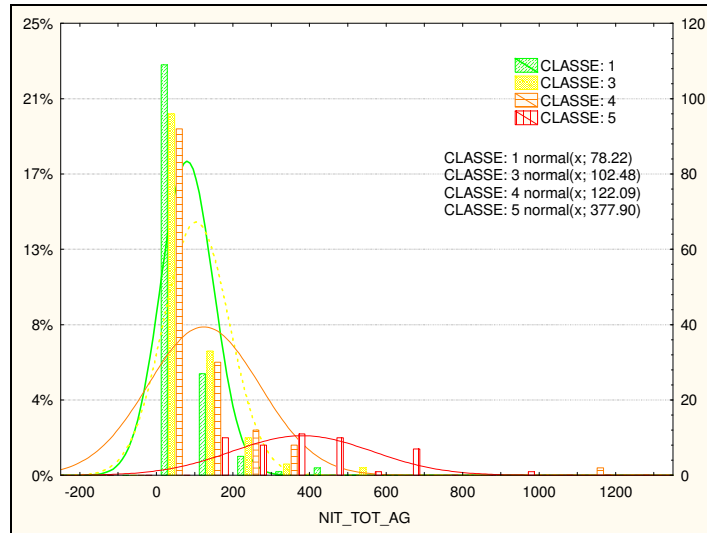


Gráfico 9: Gráfico da distribuição dos valores para NIT\_TOT\_AG por frequência de ocorrência.

O Gráfico 10, com a avaliação da variável MPS, mostra uma semelhança no padrão de distribuição dos valores para todas as classes.

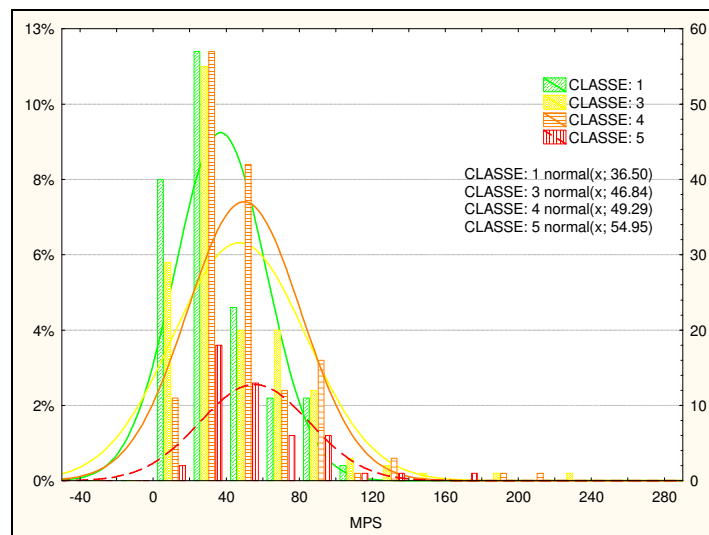


Gráfico 10: Histograma de frequência dos valores considerados para o parâmetro MPS.

O Gráfico 11 do histograma da clorofila a apresenta as maiores frequências nos menores valores nas classe 1, 2 e 3, e a classe 5 está distribuída por toda a faixa de valor.

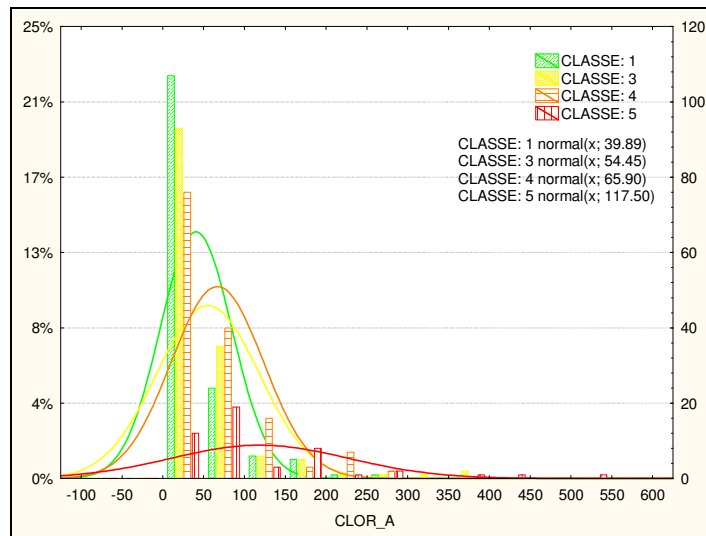


Gráfico 11: Valores do parâmetro de CLOR\_A apresentados por gráfico de histograma gerado para as coletas de superfície.

O histograma apresentado no Gráfico 12 avalia as classes para o parâmetro de abundância bacteriana. A classe 1 destaca-se por apresentar suas maiores freqüências nos menores valores, similar a alguns parâmetros onde as classes seqüentes ganham amplitude gradativamente.

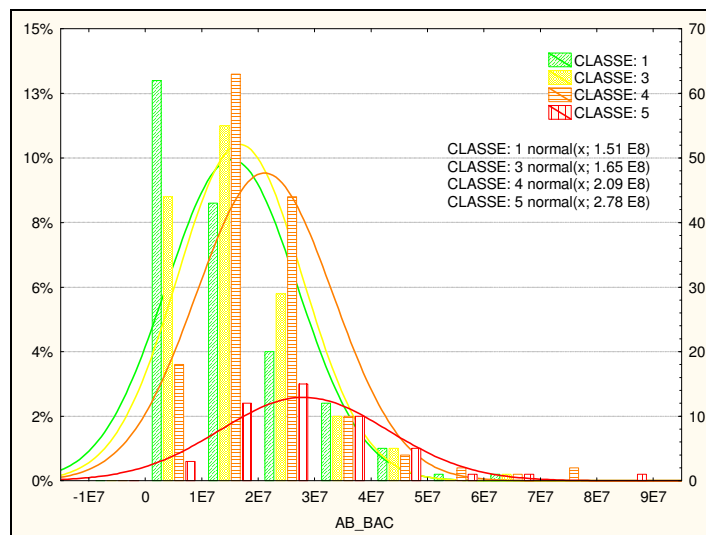


Gráfico 12: Histograma gerado a partir dos dados de AB\_BAC para as coletas de superfície.

A avaliação dos gráficos demonstra que existe uma sobreposição dos valores entre as faixas de cada classe. A Classe 5 representada somente pela BG-34 é que

apresentou maior diferença de distribuição e frequência de valores, comparando com as demais classes. Os valores da Classe 1 representada pelas as estações do canal central BG-01, BG-07 e BG-12 normalmente tinham a sua distribuição e frequência com menor variação, mantendo esse padrão para quase todas as variáveis. As demais estações que representam as Classe 3 e 4 foram as mais sobreposta em termos da distribuição e frequência de valores, essa abordagem não permitiu o apontamento de tendência no comportamento espacial dessas estações.

Após o entendimento dos Gráficos de 3 a 12, iniciou-se o teste de potencial de geração de uma modelo numérico de classificação que resolva em classes as propriedades multivariada das estações de coleta, a classificação supervisionada propriamente dita, foi implementada a partir de *scripts* do software MATLAB.

Cenário do modelo de classificação supervisionada:

- Método algoritmo de Árvore de Decisão
- Variáveis consideradas: 11 (Tabela 14)
- Número de classes: 4 (segundo a sobreposição das estações de coleta no mapa de Mayr *et al.*, 1989.)
- Número total de registros 480 (somente dados de superfície): 48 campanhas
- Número de registros para treinamento: 336 (seleção aleatória)
- Número de registros para teste: 144
- associação classe Mayr e classificador (Tabela 15)

Tabela 15: Variáveis e sua codificação no ambiente da classificação supervisionada.

TEM_AG	SAL_AG	PH	OD_AG	FOS_TOT	AMON_AG	NIT_TOT_AG	MPS	CLOR_A	AB_BAC	CLASS_MARY
X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	Conhecimento <i>a priori</i>

Tabela 16: Relação entre as classes do conhecimento a priori e adotadas pelo classificador (associação estabelecida por análise espacial).

Classe Mayr <i>et al</i>	Classificador
Classe 1	Classe 1
Classe 2	
Classe 3	Classe 2
Classe 4	Classe 3
Classe 5	Classe 4

O Gráfico 13 apresenta a distribuição dos dados para todas as classes no conjunto de treinamento do algoritmo de classificação. Observa-se que para o conjunto de dados definido como conjunto de treinamento, os valores avaliados como Classe 4 são os mais “separáveis” enquanto as demais classes apresentam certa confusão em sua separação, esse resultado reforça as interpretações anteriores de diferença da BG-34. Já o gráfico 14 representa o resultado do modelo de classificação desenvolvido, aplicado sobre o conjunto de dados de teste.

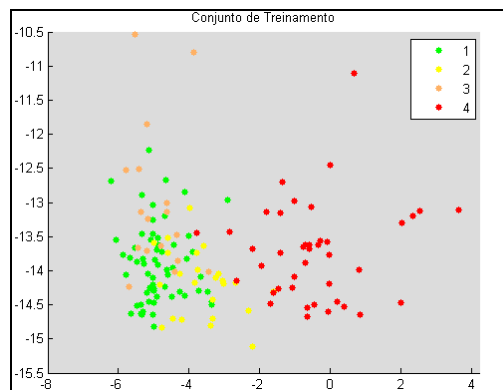


Gráfico 13: Distribuição das classes para a seleção de dados utilizada para treinar o modelo do classificador.

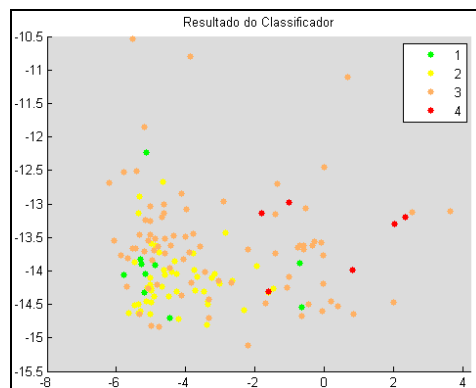


Gráfico 14: Resultado obtido a partir do teste do algoritmo de classificação no conjunto de dados selecionados.

O valor obtido pelo índice global da curva ROC (Gráfico 15) avaliando o desempenho da área (AUC), que é de 0.572, é considerado baixo, sendo que 1 é o valor

máximo da área e quanto maior a área, melhor o resultado do modelo classificador avaliado.

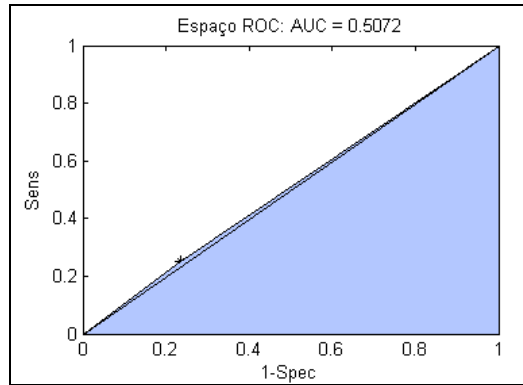


Gráfico 15: Apresenta a área da curva ROC – Critério de validação do modelo.

O cálculo de eficiência (AUC) para cada classe (Gráfico 16) demonstra que as classes 1, 3 e 4 têm melhor solução com 0.53 de área (mais resolvidas), sendo que a classe 2 apresentou menor área (menor solução).

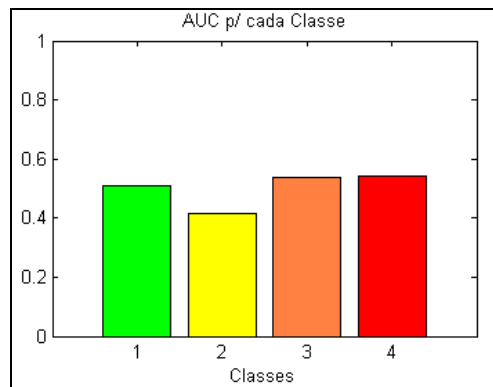


Gráfico 16: AUC calculada para cada classe.

A matriz de confusão (Tabela 16) apresenta de forma quantitativa os números de registros classificados para cada classe a partir do algoritmo treinado. Um total de 144 registros foram utilizados para testar o modelo classificador: para a Classe 1, o algoritmo classificou corretamente apenas 7 registros, os restante dos registros da Classe 1 foram classificados como Classe 2 e 3. O modelo definiu regras de classificação pouco precisa para a Classe 2, com confusão para a Classe 3; a Classe 3

apresentou melhor desempenho no classificador, com acerto de 14 dos 17 registros, e a Classe 4 apresenta confusão na classificação para a Classe 3.

Tabela 17: Matriz de confusão criada a partir do resultado do classificador.

	Classe 1 - predita	Classe 2 - predita	Classe 3 - predita	Classe 4 - predita
Classe 1 (real)	7	29	25	0
Classe 2 (real)	1	10	12	1
Classe 3 (real)	0	3	14	0
Classe 4 (real)	2	5	30	5

A integração do SIG com o modelo do classificador viabilizou uma possibilidade de análise espacial do desempenho do algoritmo de classificação supervisionada. O resultado da matriz de confiabilidade proporcional de cada classe foi associado a base de dados e apresentado aqui como mapa (Figura 30) e corresponde à confiança no resultado da classificação para cada classe.

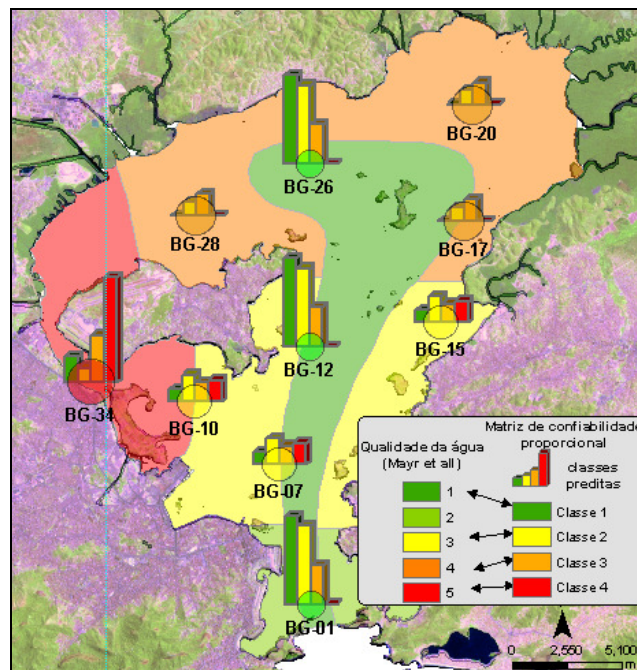


Figura 30: Representação das estações de coleta com os valores obtidos na matriz de confiança de classificação; para comparação visual estão sobrepostos as áreas definidas por Mayr *et al.*, 1989.

A Figura 31 corresponde ao modelo numérico criado pelo algoritmo de árvore de decisão, criado para esse problema de classificação supervisionada.

O modelo do algoritmo da árvore de decisão é composto de ramos e nós (internos e terminais), cada nó interno indica o teste em um atributo que se divide em ramos representando um resultado do teste para a condição do modelo, e os nós terminais (folhas) representam classes definidas pelo modelo

A Figura 31 apresenta a estrutura da árvore (modelo) criado para esse problema de classificação, em destaque e como exemplo o ponto branco assinalado é folha (nó terminal) originalmente classe 1. Os quadros dispostos junto ao gráfico auxiliam o entendimento do modelo no nó:

- Quadro 1: folha da árvore com sua classe predita (nó numero 45);
- Quadro 2: probabilidade de confusão entre as classes no nó;
- Quadro 3: Definição da regra de classificação para o nó;
- Quadro 4: Total de pontos testados para esse nó e sua classificação original.

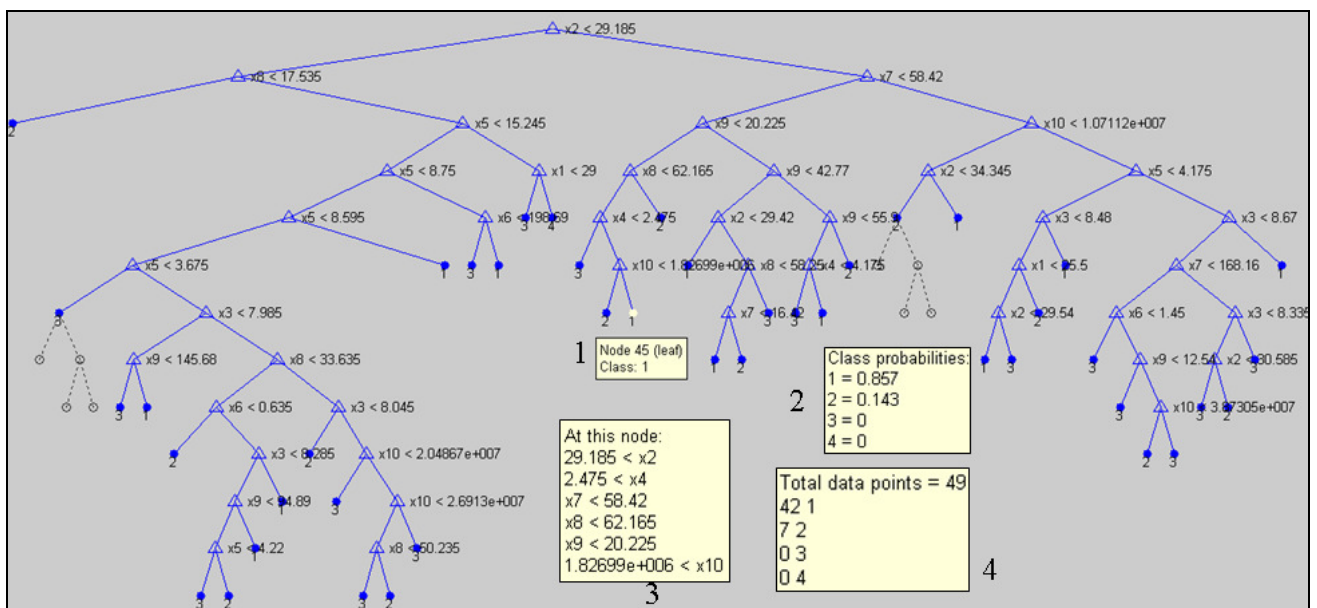


Figura 31: Modelo de árvore de decisão criado pelo classificador.

Para este exemplo, a regra do nó 45 criada pelo modelo (quadro 3 da Figura 31) é desenvolvido a partir de condições seqüentes para as variáveis, então se o valor de X2 (SAL\_AG) for menor que 29,185 e X4 (OD\_AG) for menor que 2,475 e X7 (NIT\_TOT\_AG) for menor que 58,43 e X8 (MPS) for menor que 62,165 e X9

(CLOR\_A) for menor que 20,225 e X10 (AB\_BAC) for maior que 1,82699+006, então sua estação é classe 1.

A árvore de decisão é construída recursivamente, de cima para baixo, localizando o atributo de maior importância, ou seja, aquele que apresenta significância maior para a classificação das amostras disponíveis. Este atributo apresenta o maior ganho para o modelo.

A árvore de decisão permite um último recurso em seu modelo de regras, a poda. A tentativa de reduzir (podar) os ramos da estrutura da árvore busca reduzir o número de regras (nós terminais) sem perder a eficiência do classificador. A Figura 31 apresenta os ramos da árvore representados por linhas pontilhadas, é o primeiro nível de poda do modelo que pode corresponder a uma otimização das regras do modelo de classificação.

Antes da conclusão do baixo desempenho dos dados para problemas de classificação supervisionada, foram testados outros algoritmos de classificação como: redes neurais (RBF - *Radial Basis Function* ou função de base radial e MLP - *Perceptron Multi-Camadas* ou preceptor de múltiplas camadas), máquina de vetor de suporte (MSV - *Support Vector Machines*) e classificação bayesiana, porém todos apresentaram baixa eficiência na classificação.

O uso do método de árvore de decisão como algoritmo classificador teve inspiração na didática das regras definidas pelo modelo do classificador e sua facilidade de ser implantado.

O erro global do modelo com base no percentual de acertos foi apenas de 38%, considerado muito baixo, desmotivando a investigação por algoritmos de classificação supervisionada e definição de um modelo de classificação para a baía.

Mesmo com essa condição limitante do modelo, de forma preliminar observamos que a estação BG-34 (Classe 4) foi prioritariamente classificada pelo modelo como classe 4 porém ao longo do tempo (48 campanhas) essa estação apresentou sua classificação associada a outras classes. A Classe 2 representada pelas estações BG-07, BG-10 e BG-15 foi a classe mais confusa para o modelo gerando associação a todas as demais classes ao longo do tempo. A Classe 1 das estações BG-01, BG-07 e BG-26 foi classificada prioritariamente como Classe 1, porém o modelo apresentou certa confusão com as Classes 2 e 3. As estações BG-17, BG-20 e BG-28 (Classe 3) tiveram sua classificação na maior parte do tempo associada a classe correta porém em alguns momentos o modelo definiu essas estações como Classe 2.

### 3.9 Classificação Não-Supervisionada

O modelo de classificação não supervisionada busca extrair as combinações (clusters) possíveis/disponível no banco de dados a partir das características e interações entre as variáveis de entrada.

Os testes realizados com os modelos de classificação não supervisionada, apresentados a seguir só consideram os dados coletados na superfície.

Nesta linha, a primeira consideração na abordagem foi identificar a possibilidade de quantos grupos (*clusters*) existe na base de dados. Essa prática pode ser desenvolvida utilizando de métricas de desempenho. A Tabela 17 demonstra os valores obtidos por algumas métricas utilizadas para teste, considerando a condição máxima de 5 possibilidades de *clusters*.

O resultado obtido pelos índices são poucos esclarecedores na sugestão de agrupamentos para a base de dados, já que as métricas de maximização (Critério Fisher, Índice PBM, Calinski-Harabasz) indicam diferentes possibilidades. O Critério Fisher tem melhor ajuste para 4 classes, PBM para 2 classes e Calinski-Harabasz para 3 classes, o índice de minimização Xie-Bie sugere 2 classes.

Para o cenário teste de avaliação do classificador não-supervisionado estão sendo considerados somente os dados de coleta de superfície, inferindo-se a existência de 4 classes (*clusters*).

Tabela 18: Apresenta os valores das métricas de validação utilizada, em verde os índices que destacam a eficiência do agrupamento.

	<b>2 cluster</b>	<b>3 cluster</b>	<b>4 cluster</b>	<b>5 cluster</b>
<b>Critério Fisher</b>	0.5800	0.5500	0.5900	0.2300
<b>Índice PBM</b>	0.5400	0.3891	0.2478	0.3126
<b>Índice Xie-Beni</b>	0.4717	0.9527	1.3004	1.1922
<b>Calinski-Harabasz</b>	114.0124	161.3332	125.7884	104.5297

No contexto de avaliação de qualidade de água, a melhor opção seria fazer uso do critério de avaliação legal da área de estudo, no caso do Brasil, a resolução CONAMA 357/2005 “Dispõe sobre a classificação dos corpos de água e diretrizes ambientais para o seu enquadramento,...”. O fator “restritivo” de qualidade da lei considera um valor de tolerância máximo para estabelecer o limite das classes, sendo somente duas: própria ou imprópria (regra CRISP).

O fato da legislação brasileira avaliar a qualidade de corpos hídricos a partir de valores de tolerância máximos, ou seja, duas classes, sugere o uso de um classificador CRISP, onde a resposta do modelo seria: “é poluído” ou “não é poluído” (0,1). A existência no mundo de agências ambientais, tais como USEPA (Americana), EEA (Europa), JEMAI (Japão), que adotam critério de faixa de valores como tolerância para a avaliação ambiental que motivou o teste a partir de um critério graduado de importância, estabelecendo faixas de valores de importância para cada classe de cada parâmetro, definindo assim um índice sintético de qualidade.

Em outras palavras, essa parte do trabalho, pode ser entendida como sendo uma avaliação de agrupamentos (*clusters*) das estações pelas semelhanças nos critérios (parâmetros/regras ambientais) definidos.

Ainda, a legislação está fundamentada em poluentes orgânicos persistentes (POPs), e os dados disponível tem foco mais ecológico. Essa particularidade, associada ao número baixo de variáveis da base, compatível com as solicitadas na lei, direcionamos o trabalho para uma abordagem comparativa com mapa de produzido por Mayr *et al*, 1989, ainda podendo sugerir uma nova condição de agrupamento das estações da baía.

A primeira investigação na classificação não-supervisionada foi desenvolvida a partir de técnicas de geoprocessamento denominada sobreposição ponderada (ESRI 2000) e aplicada ao banco de dados espacial.

Essa técnica consiste em classificar cada variável segundo um critério de importância e sobrepor a importância das classes através de operações matemáticas e espaciais. O produto é um mapa índice “sintético” de qualidade que considera o critério das classes de entrada do modelo.

A abordagem proposta testou as faixas dos valores das classes segundo dois critérios matemáticos:

- definição dos valores das classes por intervalos iguais;
- definição dos valores das classes por método de quebras naturais;

Sendo que o critério de quebras naturais apresentou melhor desempenho para essa proposta de estudo.

Cenário da classificação não-supervisionada:

- Variáveis consideradas: 10 (Tabela 16)
- número de classes: 4
- número total de registros 480 (somente dados de superfície): 48 campanhas

Na análise de sobreposição ponderada é necessário definir um fator de importância para cada variável de entrada do modelo, que pode ser igual ou não para todas as variáveis, dependendo do estudo de caso.

$$\text{var1} * 0.1 + \text{var2} * 0.1 + \text{var2} * 0.1 + \text{var3} * 0.1 + \text{var4} * 0.1 + \text{var5} * 0.1 + \dots + \text{var10} * 0.1 = \text{índice}$$

Fórmula 6: Estrutura da fórmula de cálculo da sobreposição ponderada para 10 variáveis.

A Figura 32 apresenta o mapa resultante do modelo classificador ponderado para 4 classes, tendo os valores das classes definidas pelo método matemático de intervalos iguais respeitando uma ordenação de importância ambiental para as regras dos parâmetros para cada classes. Ainda no mapa à classificação utilizada como conhecimento *a priori* (Mayr *et al.*, 1989). As barras representam de forma qualitativa a soma da frequência das classificações consideradas para cada estação ao longo das 48 campanhas

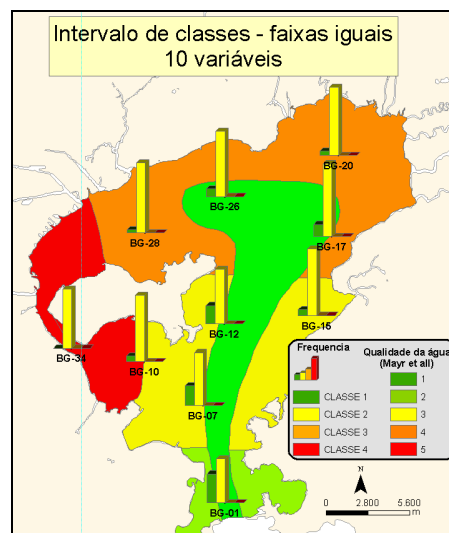


Figura 32: Modelo de classificação ponderado com definição das classes por intervalos iguais; as barras representam a soma das ocorrências para cada estação.

Outra abordagem testada para o modelo de sobreposição ponderado foi utilizando outras regras de fatiamento para as regras das classes de cada variável. Considerando 4 classes e as suas regras de fatiamentos definidas pelo método de quebras naturais a Figura 33 traz o resultado desse modelo.

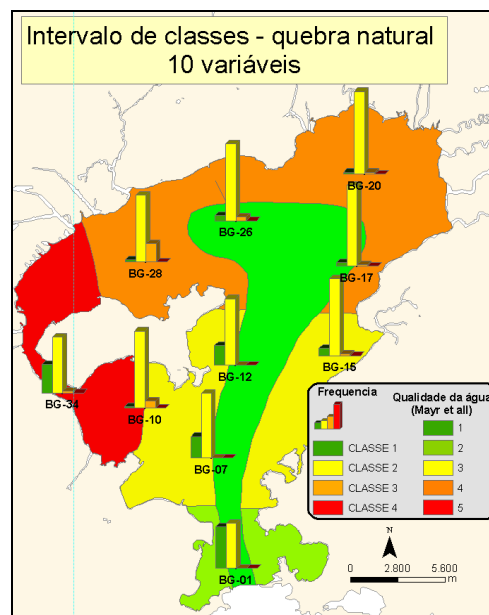


Figura 33: Modelo de classificação ponderado com definição das classes por quebras naturais.

Considerando os mapas da Figura 32 e 33, houve uma generalização do resultado do classificador ponderado com maior frequência na Classe 2 em quase todo o tempo e do espaço para ambas as regras definidas. Essa condição pode ter sido criada pela condição das regras definidas para cada classe.

Esse resultado motivou a busca por um modelo que pudesse trazer uma divisão menos homogênea para as estações da baía, considerando as mesmas faixas dos valores (regras) das classes, estabelecidas no estudo de caso.

Nessa linha optou-se em realizar uma nova abordagem considerando um número menor de variáveis, o critério adotado para a escolha dessas variáveis foram os seis (6) primeiros parâmetros definidos como “prioritário” pelo modelo de classificação

supervisionada, a árvore de decisão, sendo eles na ordem (SAL\_AG, MPS, NIT\_TOT\_AG, FOS\_TOT, CLOR\_A, AB\_BAC ).

A Figura 34 apresenta o resultado da segunda abordagem do classificador ponderado considerando apenas 6 variáveis com o uso da técnica de intervalos iguais para fatiar as classes. O fator de peso do modelo foi de 0.166 para cada variável.

O mapa produto da classificação de sobreposição ponderada (não-supervisionada) considerando 6 variáveis e técnica de fatiamento das classes por quebras naturais (Figura 35), mostra uma melhor associação do classificador (classes das estações) com as áreas das classes estabelecida por Mary em 1989, onde o canal central é marcado pelas estações de melhor qualidade BG-01, BG-07e BG-12, a BG-34 com o pior cenário de classificação. As estações BG-15, BG-17 e BG-20 com uma semelhança em suas classificações e as demais estações com um padrão ainda não muito bem resolvido.

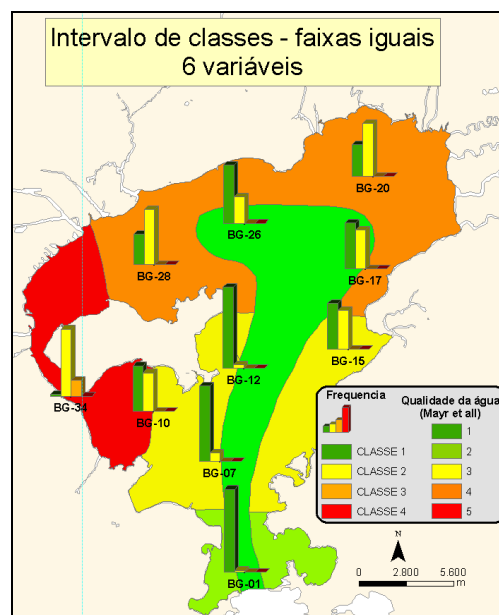


Figura 34: Modelo de classificação ponderado com definição das classes por intervalos iguais, considerando somente seis variáveis.

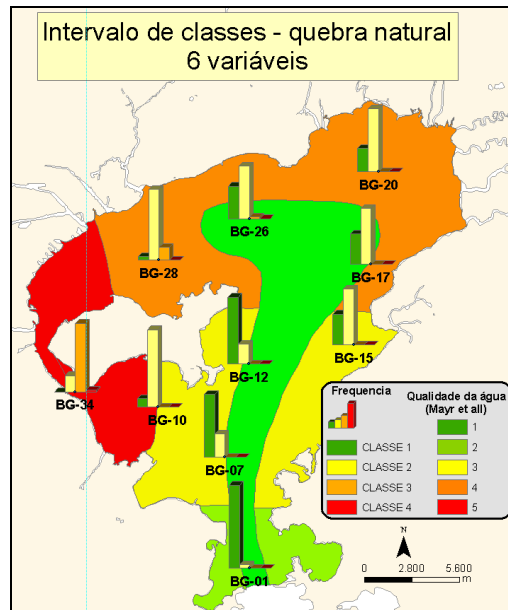


Figura 35: Modelo de classificação ponderado com definição das classes por quebras naturais, considerando somente seis variáveis.

A Tabela 18 apresenta o valor para cada faixa de classe de cada parâmetro, considerando o método de partição do dado por quebras naturais, o que obteve melhor relação com a proposta de zoneamento conhecida. A técnica de quebras naturais é conhecida por estabelecer fronteira de agrupamento através da localização de depressões sobre um histograma, a qual tende a produzir classes contendo um grande número de valores similares.

Tabela 19: As faixas de valores estabelecidas pelo método de quebras naturais, consideradas como base das regras do modelo classificador não-supervisionado; em negrito as seis variáveis prioritárias.

Parâmetro	Classe 1	Classe 2	Classe 3	Classe 4
TEM_AG	20 - 23	23 - 25	25 - 27	27 - 32
SAL_AG	<b>30.59 - 37.03</b>	<b>26.17 - 30.59</b>	<b>19.83 - 26.17</b>	<b>11.51 - 19.83</b>
PH	7.24 - 7.79	7.79 - 8.09	8.09 - 8.35	8.35 - 8.84
OD_AG	7.44 - 12.96	4.91 - 7.44	2.72 - 4.91	0.01 - 2.72
FOS_TOT	<b>0.73 - 5.54</b>	<b>5.54 - 10.95</b>	<b>10.95 - 18.81</b>	<b>18.81 - 38.33</b>
AMON_AG	0.02 - 44.92	44.92 - 131.93	131.93 - 264.33	264.33 - 472.62
NIT_TOT_AG	<b>5.18 - 122.72</b>	<b>122.72 - 284.21</b>	<b>284.21 - 557.59</b>	<b>557.59 - 1123.93</b>
MPS	<b>1.90 - 32.47</b>	<b>32.47 - 64.00</b>	<b>64.01 - 123.50</b>	<b>123.51 - 239.39</b>
CLOR_A	<b>0.40 - 42.50</b>	<b>42.51 - 109.73</b>	<b>109.73 - 234.19</b>	<b>234.19 - 535.14</b>
AB_BAC	<b>902071.07 - 13989330.17</b>	<b>13989330.18 - 27157652.47</b>	<b>27157652.48 - 48245275.59</b>	<b>48245275.60 - 81660899.65</b>

A ordem de valores das Classes foi determinada pela fragilidade ambiental do parâmetro para o ecossistema, onde a Classe 1 sugere melhor qualidade e Classe 4, pior qualidade ambiental.

Estabelecido o cenário das classes e suas condições, evoluímos o modelo de classificação não supervisionada para uma abordagem *fuzzy*. Esse método foi desenvolvido em duas fases:

- definição de regras de classificação pelo algoritmo *fuzzy* com a função de aproximação *c-medias* para cada classe de variável (cálculo do valor de pertinência para cada classe de cada variável – geração do índice de qualidade).
- Operação entre os conjuntos *fuzzy* criados para “fusão” entre as variáveis – índice sintético, permitindo que o resultado seja apresentado de forma lingüística pela suas classes de qualidade.

O limiar utilizado para particionar cada classe *fuzzy* (Tabela 19) foi calculado com base no valor médio estipulado de cada faixa para cada parâmetro estabelecido pelo método de quebras naturais (Tabela 18).

Tabela 20: Apresenta as classes utilizadas para definir as regras de partição fuzzy, em negrito as variáveis consideradas pelo classificador fuzzy.

Parâmetro	Classe 1	Classe 2	Classe 3	Classe 4
TEM_AG	21.5	24	26	29.5
<b>SAL_AG</b>	<b>33.81</b>	<b>28.38</b>	<b>23</b>	<b>15.67</b>
PH	7.52	7.94	8.22	8.6
OD_AG	10.2	6.18	3.82	1.37
<b>FOS_TOT</b>	<b>3.14</b>	<b>8.25</b>	<b>14.88</b>	<b>28.57</b>
AMON_AG	22.47	88.43	198.13	368.48
<b>NIT_TOT_AG</b>	<b>63.95</b>	<b>203.47</b>	<b>420.9</b>	<b>840.76</b>
<b>MPS</b>	<b>17.19</b>	<b>48.24</b>	<b>93.76</b>	<b>181.45</b>
<b>CLOR_A</b>	<b>21.45</b>	<b>76.12</b>	<b>171.96</b>	<b>384.67</b>
<b>AB_BAC</b>	<b>7445700.6</b>	<b>20573491</b>	<b>37701464</b>	<b>64953088</b>

Com base nos valores apresentados na Tabela 19, segue a apresentação dos gráficos com as regras de partição fuzzy para cada uma das seis (6) variáveis definidas como prioritárias pelo modelo de árvore de decisão e adotadas como melhor solução nesse

estudo de caso, onde o eixo Y representa o valor de pertinência fuzzy e o eixo X o valor da variável.

As regras fuzzy, com função triangular, foram aplicadas na base de dados a partir de rotinas/scripts elaborados no software Python versão 2.4.

Os Gráficos 17 a 22 apresentam as regras dos limiares *fuzzy* para cada um dos seis (6) parâmetros utilizados pelo modelo de classificação não supervisionado.

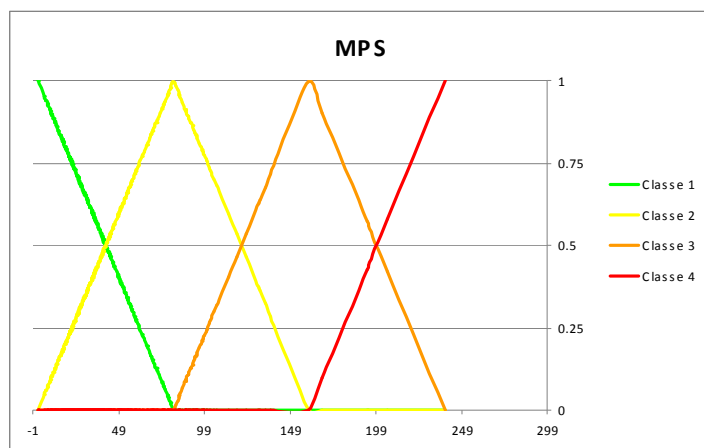


Gráfico 17: Partição *fuzzy* referente às classes de importância ambiental para a variável de MPS – material particulado em suspensão.

Importante lembrar que para o parâmetro de salinidade da água, apresentado no Gráfico 18, tem para a classe 4 os menores valores de salinidade, associa a água salgada que entra na baía a uma melhor qualidade ambiental.

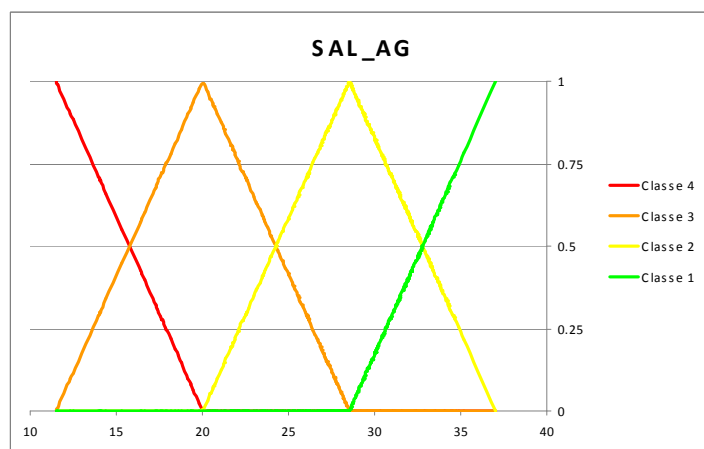


Gráfico 18: Partição *fuzzy* para a variável de salinidade da água.

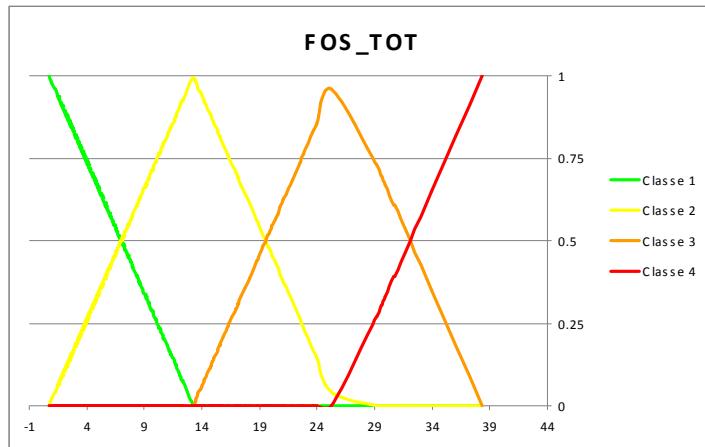


Gráfico 19: Classe de partição *fuzzy* utilizada no modelo de classificação não-supervisionado para a variável de fósforo total.



Gráfico 20: Partições *fuzzy* para a variável de clorofila-a para as coletas de água superficial.

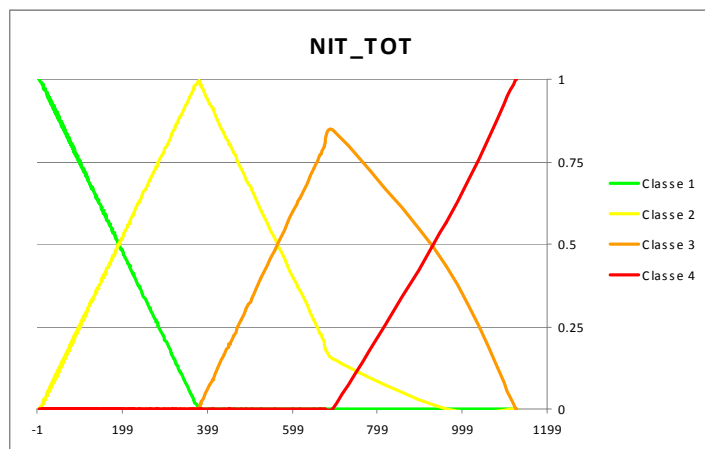


Gráfico 21: Partição *fuzzy* para a variável de nitrogênio total.

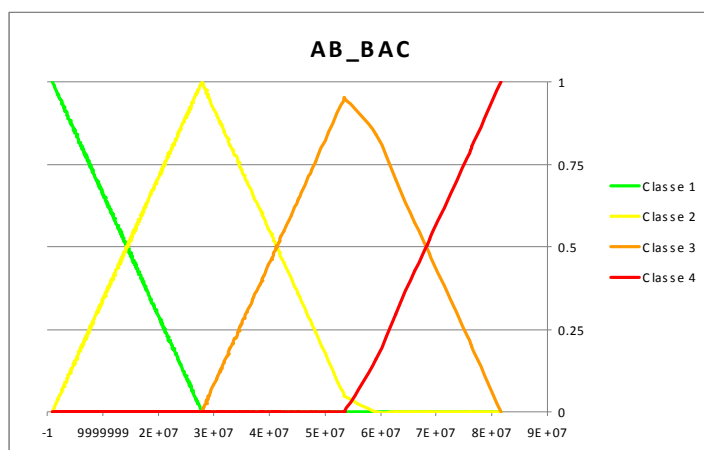


Gráfico 22: Pertinência associadas as classes de importância ambiental, partição *fuzzy* para a variável de abundancia bacteriana.

Com a definição das regras fuzzy para cada variável iniciou-se a segunda fase do processo fuzzy, corresponde a inferência das pertinências fuzzy, obtendo um produto que sintetiza todas as informações de entrada. O processo de defuzzificação foi semelhante ao adotado no modelo de sobreposição ponderada: onde a soma das pertinências das classes foi multiplicada pelo fator de peso de importância da variável, neste caso sendo igual para todas as variáveis.

A implantação do modelo de regras *fuzzy* na base do SIG viabilizou que cada registro do banco tenha atributos com a informação de sua pertinência para cada classe de cada variável. A Figura 36 apresenta um visão do banco de dados com as informações calculadas pelo modelo de classificação *fuzzy*.

No exemplo da Figura 36, os campos da tabela “FOS\_TOT\_CLASSE\_1, 2, 3 e 4”, respectivamente, correspondem ao valor de pertinência calculado pelo modelo fuzzy considerando as regras de cada classe para o parâmetro de fósforo total. O modelo se repete para cada variável estudada. Os campos “C1\_FUZZY, C2, C3 e C4”, respectivamente, são os valores de pertinência *fuzzy* associados ao produto da integração entre as seis (6) variáveis (definidas como prioritárias pelo estudo de caso) avaliadas.

Estacao	FOS TOT CLASSE 1	FOS TOT CLASSE 2	FOS TOT CLASSE 3	FOS TOT CLASSE 4	ANO	Campanha	DataColeta	FOS TOT	C1 FUZZY	C2 FUZZY	C3 FUZZY	C4 FUZZY
BG-26	0.591489	0.408511	0	0	2005	C01	14/07/2005	5.85	0.261247	0.510558	0.185532	0.042653
BG-10	0.337756	0.662244	0	0	2005	C01	14/07/2005	9.03	0.309253	0.457543	0.153029	0.080176
BG-26	0.717553	0.282447	0	0	2005	C01	14/07/2005	4.27	0.445629	0.489371	0.065	0
BG-20	0.841223	0.158777	0	0	2005	C01	14/07/2005	2.72	0.435845	0.381493	0.12484	0.057821
BG-15	0.676064	0.323936	0	0	2005	C01	14/07/2005	4.79	0.44789	0.353854	0.181654	0.016602
BG-07	0.840426	0.159574	0	0	2005	C01	14/07/2005	2.73	0.507396	0.318089	0.174515	0
BG-34	0	0.14734	0.85266	0	2005	C01	14/07/2005	23.95	0.133835	0.441015	0.41533	0.009819
BG-12	0.784574	0.215426	0	0	2005	C01	14/07/2005	3.43	0.540654	0.332604	0.126742	0
BG-01	0.808298	0.111702	0	0	2005	C01	14/07/2005	2.13	0.57186	0.23904	0.180606	0.008494
BG-17	0.848202	0.150798	0	0	2005	C01	14/07/2005	2.82	0.530598	0.35409	0.109071	0.00625
BG-12	0.850798	0.149202	0	0	2005	C02	28/07/2005	2.6	0.625342	0.267783	0.092821	0.014054
BG-01	0.938362	0.060638	0	0	2005	C02	28/07/2005	1.49	0.67084	0.196035	0.125557	0.007588
BG-17	0.873936	0.126064	0	0	2005	C02	28/07/2005	2.31	0.533965	0.302502	0.183533	0
BG-26	0.655319	0.344681	0	0	2005	C02	28/07/2005	5.05	0.431453	0.384989	0.145874	0.037683
BG-07	0.893085	0.106915	0	0	2005	C02	28/07/2005	2.07	0.639687	0.260313	0.088494	0.011506
BG-10	0.224468	0.775532	0	0	2005	C02	28/07/2005	10.45	0.522479	0.377521	0.038919	0.061081

Figura 36: Visualização de parte da tabela do banco obtida pela análise.

Diferentemente do modelo de classificação de sobreposição ponderada, onde a classe é apresentada de forma direta (classe 1 ou classe 2 ou classe 3 ou classe 4), o modelo *fuzzy* apresenta um grau de pertinência para cada classe do resultado do modelo.

A particularidade do resultado do modelo fuzzy em apresentar sua resposta em valores de pertinência por classe, ou seja, um problema de 4 classes são 4 respostas, gera uma dificuldade na representação da resposta em mapa temático, superada com a criatividade de apresentar o produto em forma de gráfico de barras por classe. Cada barra do gráfico associado a uma cor que representa uma classe, a altura da barra representa o grau de pertinência da classe para a estação.

O padrão temporal identificado anteriormente pela análise univariada (época: seca e chuvosa) entrou como condição de teste para o modelo classificador não-supervisionado. O ANEXO VI contém um mapa para cada campanha de coleta e apresenta o resultado comparativo entre a classificação *fuzzy* e a sobreposição ponderada utilizando o fatiamento das classes pelo método de quebra natural e considerando somente os parâmetros selecionados como prioritários e os dados de superfície.

A Figura 37 compara a classificação proposta por Mayr com o produto integrado de todas as campanhas (48 tempos de coleta) do classificador não supervisionada *fuzzy*. O gráfico de barras representa a soma das pertinências de cada classe para cada estação, e o círculo representa a classe (Mayr *et al.*, 1989) associada ao ponto de coleta.

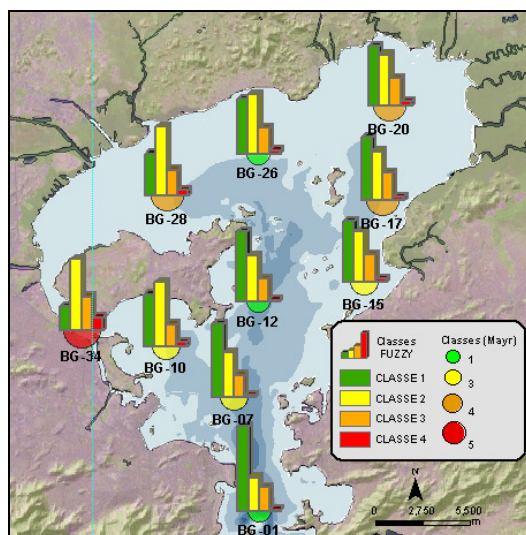


Figura 37: Representação da variabilidade espaço-temporal do classificador fuzzy.

Para melhor compreensão do comportamento do resultado de cada modelo, considera-se o momento de tempo da Figura 38 (campanha 05) e a estação BG-34. A Tabela 20 representa os valores das 6 variáveis consideradas no ponto amostral BG-34 durante a Campanha 05, as cores representam a classe correspondente a faixa de classificação de cada parâmetro, como apresentado na Tabela 20 . Assim, o modelo de sobreposição ponderada classificou a estação BG-34 como Classe 3. O modelo *fuzzy* classificou a mesma estação como tendo maior possibilidade de pertinência na Classe 2 seguido da Classe 3 e Classe 1 e 4, respectivamente. Os valores de pertinência de cada classe é apresentado na Tabela 21. As regras das classes *fuzzy* com seus limiares esta disponível nos Gráficos 17 a 22.

Tabela 21: Valor dos parâmetros da estação BG-34, campanha 05 (superfície).

VARIÁVEL	VALOR DA VARIÁVEL
SAL_AG	25,98
CLOR	175,35
FOS_TOT	21,08
AB_BAC	29078467,24
NIT_TOT_AG	283,35
MPS	73,33

Tabela 22: Resultado do valor de pertinência de cada classe do modelo fuzzy.

Classe fuzzy	Pertinência
C1	0.061462
C2	0.773628
C3	0.16091
C4	0.004

Considerando o cenário dos modelos apresentados observou que os cada técnica investigada apresentam particularidades a sua representação e na forma sintética de apresentar e entender seus indicadores ( índice de qualidade).

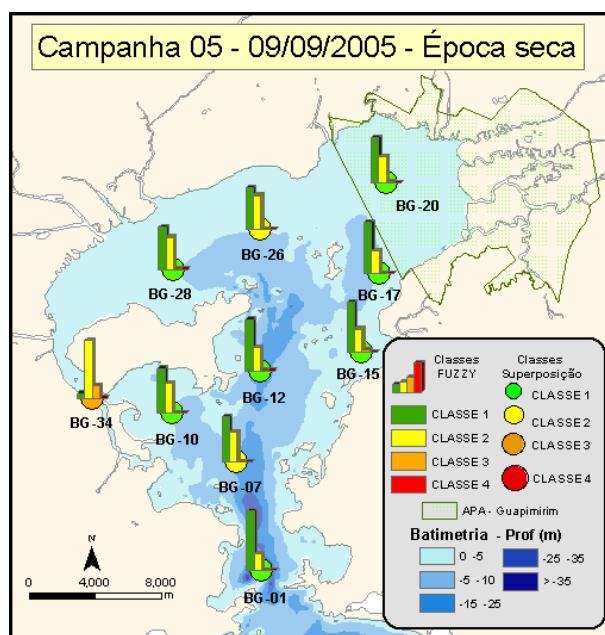


Figura 38: Mapa com a representação dos resultados do modelo fuzzy e sobreposição ponderada em um mesmo tempo (campanha 05). As barras gráficas do mapa do modelo *fuzzy* estão representando o valor de pertinência da estação para cada classe e os pontos representam a classificação do modelo de superposição; as cores são correspondentes as classes sinalizando o comprometimento ambiental.

Optou-se por não associar técnicas de geoestatística ao processo pelo motivo da malha de coleta de dados não representar de forma regular todo o espelho d'água da baía, por existirem barreiras geográficas (ilhas) e pela complexa da dinâmica da circulação desse sistema estuarino tropical, como descrito por alguns autores.

### 3.10 Modelos dos Classificadores (não-supervisionada)

Uma representação esquemática dos modelos de classificação utilizados no exemplo da figura 38 é detalhado pelas Figuras 39 e 40. A Figura 39 apresenta o esquema do modelo de classificação de sobreposição ou superposição ponderado onde cada estação tem o valor entrada da variável para a classe determinada e multiplicada por com um fator de importância, somado aos demais parâmetros, gerando assim uma saída única.

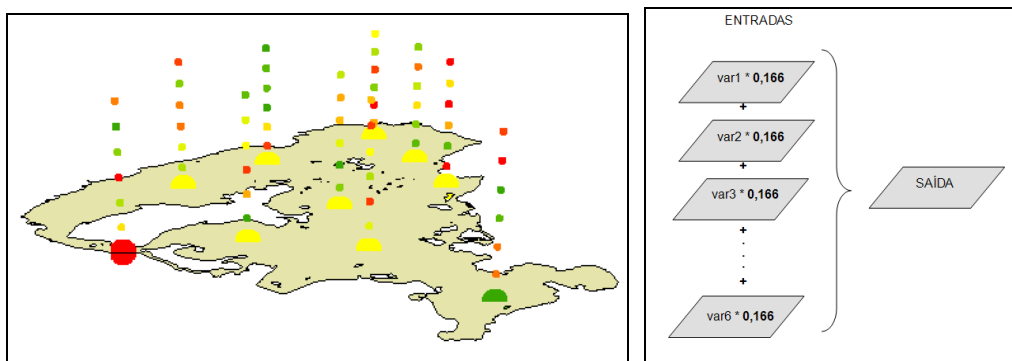


Figura 39: Representação esquemática do modelo de classificação ponderado.

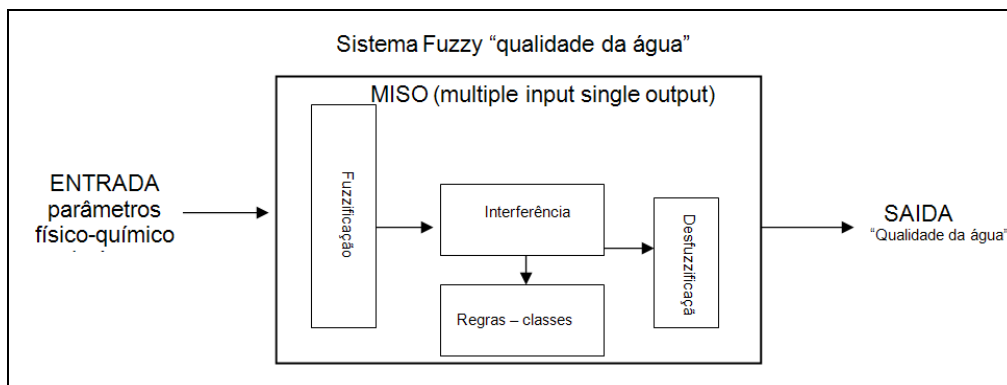


Figura 40: Representação esquemática do modelo de classificação fuzzy.

O modelo *fuzzy* utilizado tem sua representação esquemática na Figura 40, onde consideramos um modelo de múltiplas variáveis de entradas e uma única saída, apresentando ainda as fases do processo como: a fuzzificação das classes (generalização), a interferência, a aplicação das regras de multivariadas (conversão) e a desfuzzificação dos resultados em classes (especificação).

# CAPÍTULO 4

## DISCUSSÃO

A Tabela 22 apresenta uma forma consolidação para entendimento dos resultados dos classificadores não supervisionados avaliados, os métodos testados, assim como a consideração sobre o número de parâmetros testados para cada estação. O valor da tabela representa a quantidade de vezes que a estação de coleta foi classificada para aquele período (época e ano: seca ou chuvosa). O campo da tabela “QN 6” corresponde ao método de particionamento das classes por quebras naturais considerando os seis parâmetros prioritários utilizando a técnica de classificação sobreposição ponderada. O campo “QN 10” representa o número de classificação, considerando todos os 10 parâmetros investigados de cada estação para o método fatiamento das classes de quebra natural. Os campos iniciados com “II” correspondem à técnica de sobreposição ponderada com definição das classes com a técnica estatística de intervalos iguais, sendo que o número que segue ao método corresponde a quantidade de variável considerada pelo modelo classificador. As informações apresentadas nas colunas “FUZZY” correspondem a maior classe de pertinência do modelo, e seguindo as mesmas orientações quanto ao número de parâmetros envolvidos na avaliação do modelo. Quando a célula da tabela não apresentar valor significa que o modelo não classificou a estação no tempo e classe determinada.

Tabela 23: Consolidação espaço-temporal da avaliação dos classificadores utilizados; as cores da tabela fazem relação com a classe que ocorreu com maior frequência considerando todos os classificadores não supervisionados avaliados.

ESTAÇÃO	EPOCA	ANO	CLASSE	QN 6	QN 10	II 10	II 6	FUZZY 10	FUZZY 6
BG-01	CHUVOSA	2005	1	6	4	2	6	6	6
BG-01	CHUVOSA	2005	2		2	4			
BG-01	CHUVOSA	2006	1	14	7	4	14	14	14
BG-01	CHUVOSA	2006	2		7	10			
BG-01	CHUVOSA	2007	1	7	1	1	7	8	8
BG-01	CHUVOSA	2007	2	1	7	7	1		
BG-01	SECA	2005	1	5	3	4	6	6	6
BG-01	SECA	2005	2	1	3	2			
BG-01	SECA	2006	1	10	8	6	10	10	10
BG-01	SECA	2006	2		2	4			
BG-01	SECA	2007	1	4		2	4	4	4
BG-01	SECA	2007	2		4	2			

BG-07	CHUVOSA	2005	1	6	1	3	6	6	6
BG-07	CHUVOSA	2005	2		5	3			
BG-07	CHUVOSA	2006	1	9	3	2	13	12	14
BG-07	CHUVOSA	2006	2	5	11	12	1	2	
BG-07	CHUVOSA	2007	1	5			7	7	8
BG-07	CHUVOSA	2007	2	3	8	8	1	1	
BG-07	SECA	2005	1	3	2	3	5	6	6
BG-07	SECA	2005	2	3	4	3	1		
BG-07	SECA	2006	1	10	6	5	10	10	10
BG-07	SECA	2006	2		4	5			
BG-07	SECA	2007	1	2			2	3	3
BG-07	SECA	2007	2	2	4	4	2	1	1
BG-10	CHUVOSA	2005	1	1			5	3	4
BG-10	CHUVOSA	2005	2	5	6	6	1	3	2
BG-10	CHUVOSA	2006	1			1	6	4	6
BG-10	CHUVOSA	2006	2	14	12	13	8	10	8
BG-10	CHUVOSA	2006	3		2				
BG-10	CHUVOSA	2007	1	1			5	1	3
BG-10	CHUVOSA	2007	2	7	6	8	3	7	5
BG-10	CHUVOSA	2007	3		2				
BG-10	SECA	2005	1	1	1	1	1	2	1
BG-10	SECA	2005	2	5	5	5	5	4	5
BG-10	SECA	2006	1	2		2	8	6	5
BG-10	SECA	2006	2	8	10	8	2	4	5
BG-10	SECA	2007	1				1	1	
BG-10	SECA	2007	2	4	4	4	3	3	4
BG-12	CHUVOSA	2005	1	4		2	6	5	6
BG-12	CHUVOSA	2005	2	2	6	4		1	
BG-12	CHUVOSA	2006	1	12	1	2	13	13	13
BG-12	CHUVOSA	2006	2	2	13	12	1	1	1
BG-12	CHUVOSA	2007	1	7			8	8	8
BG-12	CHUVOSA	2007	2	1	8	8			
BG-12	SECA	2005	1	3	2	2	5	6	6
BG-12	SECA	2005	2	3	4	4	1		
BG-12	SECA	2006	1	10	8	6	10	9	10
BG-12	SECA	2006	2		2	4		1	
BG-12	SECA	2007	1	1			4	4	4
BG-12	SECA	2007	2	3	4	4			
BG-15	CHUVOSA	2005	1	2			5	5	6
BG-15	CHUVOSA	2005	2	4	6	6	1	1	
BG-15	CHUVOSA	2006	1	5			7	8	14
BG-15	CHUVOSA	2006	2	9	14	14	7	6	
BG-15	CHUVOSA	2007	1	2			1	5	7
BG-15	CHUVOSA	2007	2	6	7	8	7	2	1
BG-15	CHUVOSA	2007	3		1			1	
BG-15	SECA	2005	1	2	1	1	2	6	6
BG-15	SECA	2005	2	4	5	5	4		
BG-15	SECA	2006	1	5	3	3	8	6	7
BG-15	SECA	2006	2	5	7	7	2	4	3
BG-15	SECA	2007	1	1			3	2	3
BG-15	SECA	2007	2	3	4	4	1	2	1
BG-17	CHUVOSA	2005	1	1			2	5	6

BG-17	CHUVOSA	2005	2	5	6	6	4	1	
BG-17	CHUVOSA	2006	1	3			7	10	13
BG-17	CHUVOSA	2006	2	11	14	14	7	4	1
BG-17	CHUVOSA	2007	1				2	7	7
BG-17	CHUVOSA	2007	2	8	7	8	6	1	1
BG-17	CHUVOSA	2007	3		1				
BG-17	SECA	2005	1	3	1	2	3	6	6
BG-17	SECA	2005	2	3	5	4	3		
BG-17	SECA	2006	1	8		5	9	10	8
BG-17	SECA	2006	2	2	10	5	1		2
BG-17	SECA	2007	1	2	1	1	3	4	4
BG-17	SECA	2007	2	2	3	3	1		
BG-20	CHUVOSA	2005	1	2			1	3	6
BG-20	CHUVOSA	2005	2	4	6	6	5	3	
BG-20	CHUVOSA	2006	1	2			4	11	13
BG-20	CHUVOSA	2006	2	12	13	14	10	2	1
BG-20	CHUVOSA	2006	3		1			1	
BG-20	CHUVOSA	2007	1				2	6	8
BG-20	CHUVOSA	2007	2	8	8	8	6	2	
BG-20	SECA	2005	1	1			2	6	6
BG-20	SECA	2005	2	5	6	6	4		
BG-20	SECA	2006	1	8	1	3	9	8	8
BG-20	SECA	2006	2	2	9	7	1	2	2
BG-20	SECA	2007	1					2	3
BG-20	SECA	2007	2	4	4	4	4	2	1
BG-26	CHUVOSA	2005	1	3			3	4	5
BG-26	CHUVOSA	2005	2	3	6	6	3	2	1
BG-26	CHUVOSA	2006	1	3			9	1	7
BG-26	CHUVOSA	2006	2	11	14	14	5	13	7
BG-26	CHUVOSA	2007	1	1			6	4	5
BG-26	CHUVOSA	2007	2	7	6	8	2	4	3
BG-26	CHUVOSA	2007	3		2				
BG-26	SECA	2005	1	3	1	1	3	3	3
BG-26	SECA	2005	2	3	5	5	3	3	3
BG-26	SECA	2006	1	8	2	4	9	9	9
BG-26	SECA	2006	2	2	8	6	1	1	1
BG-26	SECA	2007	1				3		1
BG-26	SECA	2007	2	3	4	4	1	4	3
BG-26	SECA	2007	3	1					
BG-28	CHUVOSA	2005	1				2	2	3
BG-28	CHUVOSA	2005	2	6	6	6	4	4	3
BG-28	CHUVOSA	2006	1				6	2	4
BG-28	CHUVOSA	2006	2	13	11	14	8	12	10
BG-28	CHUVOSA	2006	3	1	3				
BG-28	CHUVOSA	2007	1				1		
BG-28	CHUVOSA	2007	2	5	2	8	7	8	8
BG-28	CHUVOSA	2007	3	3	6				
BG-28	SECA	2005	1	1	1	1	2	1	2
BG-28	SECA	2005	2	4	5	5	4	5	4
BG-28	SECA	2005	3	1					
BG-28	SECA	2006	1	1		1	5	1	
BG-28	SECA	2006	2	8	10	9	5	9	10

BG-28	SECA	2006	3	1					
BG-28	SECA	2007	1				1		
BG-28	SECA	2007	2	3	3	4	3	4	4
BG-28	SECA	2007	3	1	1				
BG-34	CHUVOSA	2005	1				1		
BG-34	CHUVOSA	2005	2	2	4	5	5	5	6
BG-34	CHUVOSA	2005	3	4	2	1		1	
BG-34	CHUVOSA	2006	2	3	3	11	12	11	12
BG-34	CHUVOSA	2006	3	10	10	3	2	2	1
BG-34	CHUVOSA	2006	4	1	1			1	1
BG-34	CHUVOSA	2007	2			3	4	4	3
BG-34	CHUVOSA	2007	3	8	8	5	4	4	4
BG-34	CHUVOSA	2007	4						1
BG-34	SECA	2005	2	2	2	6	6	6	5
BG-34	SECA	2005	3	4	4				1
BG-34	SECA	2006	1					1	
BG-34	SECA	2006	2	1	5	10	8	9	8
BG-34	SECA	2006	3	9	5		2		2
BG-34	SECA	2007	2	1	2	4	3	4	4
BG-34	SECA	2007	3	3	2		1		

As considerações e entendimento da Tabela 22 consolidada os resultados dos classificadores não supervisionados apresentam o seguinte padrão: a BG-01, BG-07 prioritariamente foram classificadas como classe 1, porém a classe 2 esteve presente em menor número para todos os anos e épocas. A BG-10 foi classificada na maioria do tempo como classe 2, e a classe 1 esteve presente, em menor número, durante o período chuvoso nos anos de 2006 e 2007; o classificador ponderado com método de fatiamento de quebra natural “QN 10”, considerando 10 variáveis, classificou a estação como classe 3 em dois momentos. A estação BG-12 teve sua classificação determinada como classe 1 por todos os classificadores ao longo do tempo, porém alguns momentos intercalados para a classe 2. A BG-15, BG-17 e BG-20 tiveram sua maioria das avaliações como classe 2 com momentos de classe 1, e uma medição de classe 3 no período chuvoso, sendo que para as duas primeiras estações esta última medida ocorreu em 2007, para a estação BG-20, em 2006. A BG-26 foi prioritariamente classificada como classe 2 com ocorrência de classe 1 para todos os períodos, com dois momentos de classe 3, ambos em 2007, um em cada época do ano (seca/chuvosa). A BG-28 teve sua classificação alternando entre as classe 1, classe 2 (maior número) e classe 3, para todas as épocas e anos, com exceção do período chuvoso de 2005 que apresentou somente as classes 1 e 2. Finalizando, a BG-34 teve sua classificação prioritária distribuída entre as classes 2 e 3, sendo que esta última em maior número; em dois

momentos em 2005 na época chuvosa e em 2006 na época seca a estação foi classificada como classe 1 pelos métodos “II 6” e “FUZZY 10”, respectivamente; em dois tempos foi classificada como classe 4, um no período chuvoso de 2006 e outro para a mesma época em 2007.

A Figura 41 apresenta uma consolidação espacial da variabilidade de classificação não supervisionada (sobreposição ponderada) de cada estação ao longo do tempo, considerando somente os dados de superfície. As barras representam a presença da classe na estação, sendo que a barra maior corresponde a classe de predominância; essa representação considera uma média dos resultados entre todos os classificadores de sobreposição ponderada.

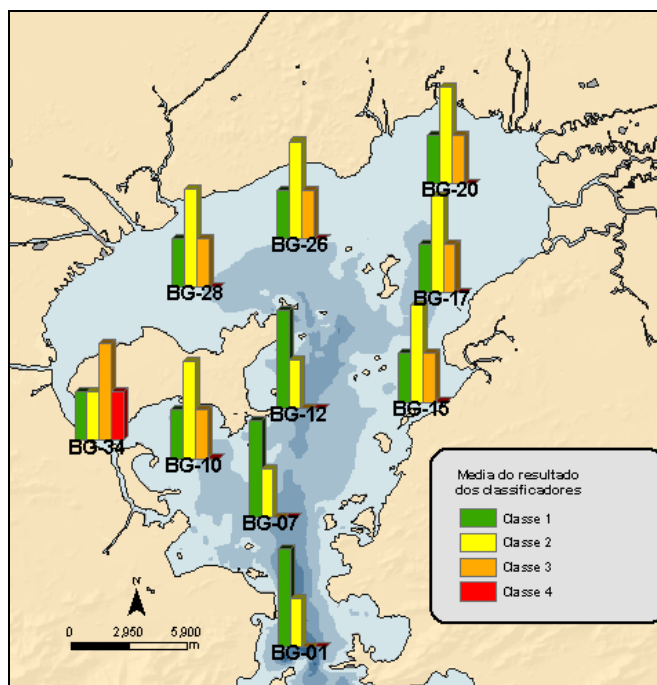


Figura 41: Representação da variabilidade de classificação não supervisionada das estações para o método de sobreposição ponderada.

Uma interpretação espacialmente consolidada dos dados classificados de forma não supervisionada pelo algoritmo *fuzzy* não destacam diferenças significativas no índice do classificador para épocas do ano (chuva e seca) conforme apresentado na Figura 42, porem observa-se uma maior presença da Classe 4 nas estações BG-34, BG-28, BG-20, BG-17, BG-15 e BG-10 para a época chuvosa.

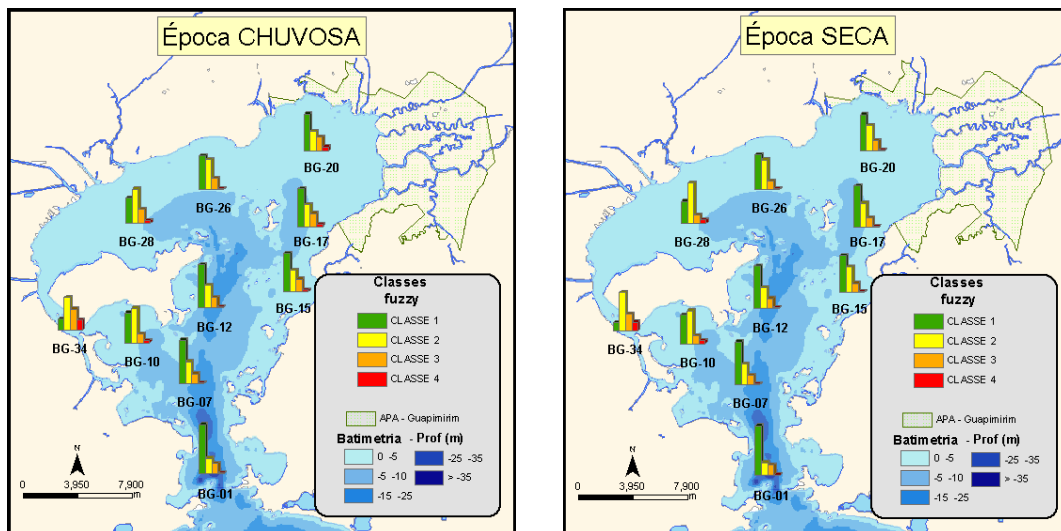


Figura 42: Consolidação do resultado do classificador não supervisionado fuzzy ( dados somente de superfície considerando as 6 parâmetros definidos como prioritários e para as coletas de superfície.

Se considerarmos o resultado dos modelos de classificação não supervisionado avaliados, segundo as regras estabelecidas para cada faixa de valores associadas a classe de “qualidade ambiental” e levando em conta a malha amostral disponível, podemos observar uma tendência no comportamento das estações onde:

- BG-01, BG-07, BG-12 estão associadas a estações de melhor qualidade;
- BG-15, BG-17, BG-20 estão associadas a uma qualidade intermediária;
- BG-10, BG-28, BG-34 estão associadas a uma pior qualidade;
- BG-26 apresenta característica instável entre as classes.

De forma preliminar a Figura 43 consolida uma ilustração da interpretação de zoneamento para o resultado dos classificadores não supervisionado. As estações BG-01, BG-07 e BG-12 foram quase sempre classificadas como Classe 1 (de melhor qualidade) as estações BG-15, BG-17 e BG-20 foram classificadas em sua maioria como Classe 2 ou 3 (qualidade intermediária) . As estações BG-34, BG-28 e BG-10 como sendo as de piores condições para o classificador prioritariamente associadas as Classe 3 e 4 a BG-34 ainda teve pior desempenho dessas três (3) estações. A Estação BG-26 foi a que apresentou maior variabilidade de classificação, com associações a todas as classes ao longo de tempo porem com menor frequência nas Classe 1 e 4.

A Figura 43 deve ser analisada com critério por diversos fatores: as estações apresentaram oscilação nas classificações durante o tempo; poucos pontos distribuídos na baía; critérios das classes de qualidade estabelecidos por modelo matemático; estamos tratando do compartimento água (móvel/dinâmico) que sofre a influência de diversos fatores externos que contribuem para a sua variabilidade (maré, correntes, chuvas, atividade antropica entre outras) não consideradas nessa avaliação.

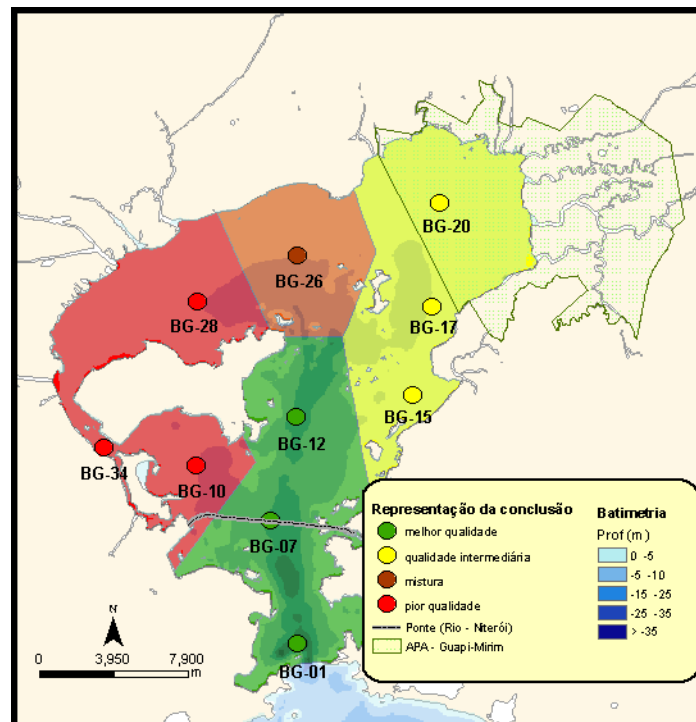


Figura 43: Zoneamento preliminar das estações, sugerido a partir da interpretação dos classificadores não supervisionados.

O modelo de dados espacial adotado mostrou-se bastante flexível, o compatível com a inserção dos resultados do modelo na base de dados, dando maior rastreabilidade ao processo de análise.

## CAPÍTULO 5

### CONCLUSÃO

O modelo de dados implementado mostrou-se eficiente para o propósito de uso do projeto, o modelo de dados espacial foi elaborado em geodatabase, formato proprietário da família dos softwares ESRI. Como ferramental de SIG, o ArcGIS 9.3 atendeu à necessidade de manipulação do banco, análises espaciais e geração de mapas. O banco de dados comercial proporcionou em um único ambiente tecnológico toda estrutura necessária para o uso SIG, sem necessidade de novos desenvolvimentos.

Vários foram os benefícios do projeto em adotar o uso de modelo de dados espacial desde seu início, dentre os quais pode-se destacar a facilidade para a consulta, acesso aos metadados (nativo do modelo de dados), comparação de resultados, integração de dados espacialmente, apresentação dos resultados das análises, integração com outros aplicativos, entre outros.

A etapa de exploração univariada da base de dados foi fundamental, fornecendo um entendimento do complexo cenário da análise, onde cada variável foi detalhada em todas as dimensões presentes nos dados (X, Y, Z e T), sendo X e Y a posição no espaço, Z o nível de coleta da estação, e T o tempo da coleta ou campanha. Nesta abordagem os dados apresentaram comportamento temporal definido pela época do ano seca e chuvosa esse padrão de variação se manteve para as coletas de fundo e superfície.

O classificador supervisionado avaliado – árvore de decisão – apresentou desempenho abaixo do satisfatório, o que pode ser explicado pelo pequeno número de registros disponíveis para treinar o algoritmo de classificação e a diferença metodológicas envolvida na informação estabelecida como conhecimento *a priori*.

As métricas de desempenho atestam o baixo rendimento do classificador supervisionado, situação que não motivou a implementação do modelo de regras determinada pelo algoritmo de classificação utilizado, árvore de decisão.

Na classificação não supervisionada, a comparação entre as técnicas utilizadas para estabelecer corte para faixa de valor para cada classe, mostrou melhor eficiência no método de quebras naturais, se comparado o resultado do classificador com a proposta

feita por Mayr *et al.*, 1989 sendo também essa a condição que suportou a decisão pelo uso de quatro classes.

O desenvolvimento do modelo classificação ponderada ou superposição mostrou-se eficiente para uma abordagem de resposta direta. As classes foram definidas com alguma eficiência ao longo do tempo quando comparado com a proposta de Mayr *et al.*, 1989. O modelo de classificação foi implementado na base de dados com simplicidade.

Considerando a possibilidade de entendimento e uso dos resultados dos classificadores não supervisionados, pode-se dizer que o classificador ponderado apresenta uma resposta direta com relação ao classificador *fuzzy*, porém o classificador *fuzzy* traz mais informações sobre o conjunto de variáveis de entrada definindo a classe de maior pertinência e suas outras possibilidades.

A possibilidade de apresentar o resultado do classificador não supervisionado a partir do valor de pertinência para cada classe permite entender a complexidade do cenário e a fragilidade da classificação em modelos ambientais. Ao invés de apontar uma única classe como resposta, o modelo o *fuzzy* determina em sua resposta a classe de maior pertinência apresentando também as pertinências para outras classes, como exemplo conjunto de entrada pode ser classe boa pela maior pertinência mas pode ter “características” de uma classe de pior qualidade com pertinências menores, com sugestão de alerta ou de degradação.

Avaliando temporalmente cada campanha com o critério de classificação ponderada não foi encontrado nenhuma campanha (tempo) onde as estações foram representadas por todas as classes de qualidade, normalmente a resposta apresentava predominância de duas classes para esse método de classificação.

O SIG, como ferramenta de apoio aos processos ambientais, foi fundamental para suportar a tarefa de gestão/organização da informação, permitindo o rápido acesso aos dados e facilitando os processos associados às análises, como o uso em séries temporais de dados, em análises multivariadas na exploração e/ou integração das informações; assim recursos computacionais e técnicas avançadas de mineração de dados são grandes aliados nessa evolução tecnológica.

## CAPITULO 6

### RECOMENDAÇÕES

Recomendação para futuros desenvolvimentos estão relacionados a definir faixas de valores das classes de qualidade da água a partir de critérios ambientais mais fundamentados para a área de estudo, o *background* natural da baía e sua variabilidade natural. Avaliar a compatibilidade entre os parâmetros utilizados, métodos de coleta para a integração de informações geradas por outros projetos aumentando assim a serie temporal e malha amostral da área de estudo. Buscar a integração com outras variáveis que alteram de forma significativa a condição da baía como: circulação/correntes, mares, precipitação entre outras variáveis biológicas.

## REFERÊNCIAS BIBLIOGRÁFICAS

ANDREOZZI, V. “Modelos lineares generalizados”. CEUL, Fev/2008.

ANDREW MACDONALD. *Building a geodatabase*. Ed. Redlands, version 9.3, ESRI Press, California, 2001.

APHA, AWWA, WEF. *Standard methods for examination of water and wastewater*. 19<sup>th</sup> ed., APHA/AWWA/WEF, Washington, DC, 1995.

BACKER, E. *Computer Assisted reasoning in cluster analysis*. Prentice Hall, New York, 1995.

BAPTISTELLA, B. *O uso de redes neurais e regressão linear múltipla na engenharia de avaliações: determinação dos valores venais de imóveis urbanos*. Dissertação de Mestrado, UFPR, Curitiba, PR, Brasil, 2005.

BARCZAK, C.L.; VERDINELLI, M.A.; VERDINELLI, M.E.P.; FIGUEIREDO, L.F.G. A lógica difusa e métodos de análise em sistemas de informação geográfica. In: *Anais do III Congresso Brasileiro de Cadastro Técnico Multifinalitário*, Florianópolis, SC, 1998.

BARILIS, E.; PSAILA, G. “Designing templates for mining association rules”, *Journal of Intelligent Information Systems*, 9: 7-32, 1997.

BARROS,R.S. *Estimação de parâmetros físico-químicos da água com suporte do Sensoriamento Remoto – Estudo de caso; Baía de Guanabara, Rio de Janeiro*. Dissertação de Mestrado, UFRJ, Rio de Janeiro, RJ, Brasil, 2002.

BENITEZ, R. M.; “Modelos de previsão para séries temporais univariadas”, *Revista Alcance*, Itajaí, v. 8, n. 7, p. 55-71, 2001.

BEZDEK, J.C., EHRLICH, R.; FULL, W. “FCM: The fuzzy c-means clustering algorithm”, *Computer & Geosciences*, v. 10, n. 2-3, pp. 191-203, 1984.

BRANCO, S. M. (1986). *Hidrobiologia aplicada à engenharia sanitária*, São Paulo, 3 ed., CETESB/ASCETESB

CÂMARA, G., MEDEIROS, F.S. *Geoprocessamento em projetos ambientais*. 1ª ed, São José dos Campos: INPE, 190 p, 1998.

CÂMARA, G.; DAVIS, C.; MONTEIRO, A.M. *Introdução à ciência da geoinformação*, São José dos Campos: INPE. Disponível em <<http://www.dpi.inpe.br/gilberto/livro/introd.htm>>. Acesso em: ago. 2004.

CAMARGO, E.; MONTEIRO, A.M.; FELGUEIRAS, C.; DRUCK, S. “Integração da geoestatística e sistemas de informação geográfica: uma necessidade”. In: *Congresso e feira para usuários de Geoprocessamento da América Latina*, GISBRASIL, Salvador, 1999.

CAMARGO, E. “Apostila do curso: Geoestatística e Aplicações em Geoprocessamento”, *XIII Simpósio Brasileiro de Sensoriamento Remoto*, Florianópolis, SC, 2007.

CARMOUZE, J. P. (1994). *O Metabolismo dos ecossistemas aquáticos: fundamentos teóricos, métodos de estudo e análises químicas*. São Paulo - Editora Edgard Blücher – FAPESP. 253p.

CAZZELA, S. “Árvores e tabelas de decisão”. Notas de aula da Universidade do Vale do Rio dos Sinos – UNISINOS, disciplina: Sistemas de Apoio a Decisão, 2007.

CUNHA, E. M. S. “Caracterização e planejamento ambiental do estuário Potengi”, *Coleção Textos Acadêmicos*, n. 285, 211 pp, UFRN, Natal, 1982.

DRUMMOND, I.N. *Implementação do método de classificação contínua fuzzy k-médias no ambiente TerraLib*. Monografia do curso de Introdução ao Geoprocessamento, INPE, São José dos Campos, 2003.

ESRI. *Understanding GIS - The Arc/Info Method*. Rev 6, ed. Redlands, California: ESRI Press, 1992.

FEEMA. *Qualidade de Água da Baía de Guanabara 1990/1997*. FEEMA, PDBG – Programas Ambientais Complementares, 1998.

FEEMA. *Projeto de recuperação gradual do ecossistema da Baía de Guanabara*. V. I, FEEMA, Rio de Janeiro, 1990.

FIGUEIRA, E. *Bases cartográficas para GIS*. ExpoGeoBrasil 1999, Curso C4, Ed. Espaço Geo, Curitiba, Paraná, 33pp., 1999.

FIGUEIREDO, L.F.G.; VERDINELLI, M.E.P.; VERDINELLI, M.A.; BARCZAK, C. “Cadastro técnico ambiental, sistemas de informação geográfica e lógica fuzzy: ferramentas conjugadas para a gestão ambiental”. In: *Anais do III Congresso Brasileiro de Cadastro Técnico Multifinalitário*, 1998.

GARRIDO, A.M.F.; COSTA, H.R.; BARRETO, M.K. *Avaliação biológica dos pontos críticos da Baía de Guanabara*. Fundação Estadual de Engenharia do Meio Ambiente, pp. 337-359, 1978.

GIMENES, E. *Data Mining – Data Warehouse. A importância da mineração de dados em tomadas de decisão*. Monografia, Faculdade de Tecnologia de Taquaritinga, Taquaritinga, São Paulo, Brasil, 2000.

GONZALEZ, G.R.A.; EVSUKOFF, A.G.; SANTOS, R.C.P.; SOBRAL, A.P.B.; SILVA, J.A. “Fuzzy geo-processing for characterization of social groups: an application to a Brazilian mid-size city”. In: *Seventh International Conference on Data Mining*, 2006.

HOSMER, D.G., LEMESHOW, S. *Applied Logistic Regression*, New York: Wiley-Interscience, 131 p., 1998.

IBGE. Noções básicas de Cartografia. Disponível em: [http://www.ibge.gov.br/home/geociencias/cartografia/manual\\_nocoos/indice.htm](http://www.ibge.gov.br/home/geociencias/cartografia/manual_nocoos/indice.htm). Acesso em: fev / 2004.

ISAAKS, E.H.; SRIVASTAVA, M. *An introduction to applied geostatistics*. New York, Oxford University Press, 560 p., 1989.

JAIN, A.J. “Statistical pattern recognition: a review”. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 22, n. 1 (Jan), 2000.

KLIR, J.G; YUAN, B. *Fuzzy sets and fuzzy logic: theory and applications*. Prentice-Hall PTR, 1st ed, 592 p., 1995.

MACHADO, L.E. e LADEIRA, M. “Um Estudo de Limpeza em Base de Dados Desbalanceada com Sobreposição de Classes”, In: *Anais do VI Encontro Nacional de Inteligência Artificial*, IME, Rio de Janeiro, 2007.

MAIDMENT, D.R. “GIS and Hydrologic Modelling”, *GIS in Hydraulics, Hydrology and Water Resources*, v. 1, pp. 59-102, 1992.

MARTINS, D.M.S. “Métodos de Agrupamento de Dados”. Monografia, COPPE/UFRJ, Rio de Janeiro, RJ, Brasil, 2005.

MAYR, L.M., TENENBAUM, D.R., VILLAC, M.C., PARANHOS, R., NOGUEIRA, C.R., BONECKER, S.L.C., BONECKER, A.C.T. *Hydrobiological characterization of Guanabara Bay*. In: MAGOON, O., NEVES, C., (eds.), *Coastlines of Brazil*. New York, American Society of Civil Engineers, 1989.

MAYR, L.M. *Avaliação ambiental da Baía de Guanabara com o suporte do geoprocessamento*. Tese Doutorado, Instituto de Geociências, UFRJ, Rio de Janeiro, RJ, Brasil, 1998.

MILLIGAN, G., COOPER, M. "An examination of procedures for determining the number of clusters in a data set". *Pshychometrika*, v. 50, n. 2, pp. 159–179, 1985.

MOITA NETO, J.M. Estatística multivariada: uma visão didática-metodológica. 2004, disponível em: <[http://criticanarede.com/cien\\_estatistica.html](http://criticanarede.com/cien_estatistica.html)>. Acesso em: jan/2009.

MONARD, M. C., BARANAUSKAS, J. A. *Indução de regras e árvores de decisão*. In: Rezende, S. O. (org), *Sistemas inteligentes: fundamentos e aplicações*, capítulo 5, ed. Manole, Rio de janeiro, 2003.

NASSER, V.L. Estudo da qualidade de água na Baía de Guanabara utilizando técnicas de sensoriamento remoto e análise geoestatística. Tese de Mestrado, COPPE/UFRJ, Rio de Janeiro, RJ, Brasil, 2001.

NAVEGA, S. "Princípios Essenciais do Data Mining". In: Anais do Infoimagem, Cenadem, 2002. Disponível em: <[www.intelliwise.com/snavega](http://www.intelliwise.com/snavega)>, acesso em: set 2004.

NÉIA, M. *Cartografia Digital*, Fator – Sagres Editora, Guarulhos, São Paulo, 56 p., 1997.

PEREIRA, A.V.G.; TAVARES, G.C.O.; MARTINS, J.M.P.N.; COELHO, M.P.S. *Metadados*. Disponível em: <<http://www.isa.utl.pt/dm/sig/sig20002001/TemaMetadados/trabalho.htm>>. Acesso em: fev / 2009.

PRADO, H.A. , ENGEL, P.M., FILHO, H.C. "Rough clustering: an alternative to find meaningful clusters by using the reducts from a dataset". In: *Rough Sets and Current Trends in Computing: Third International Conference, RSCTC 2002*, v. 2475 of LNCS, pp. 234 – 238, Berlin, 2002.

PRADO, P. I., LEWINSOHN, T.M., CARMO, R.L., HOGAN, D.J. "Ordenação Multivariada na Ecologia e seu Uso em Ciências Ambientais", *Ambiente e Sociedade*, v. 10, pp. 69-83, Campinas, SP, 2002.

PAKHIRA, M. K., BANDYOPADHYAY, S.; MAULIK, U. “Validity index for crisp and fuzzy clusters”. *Pattern Recognition*, v. 37 (3), pp. 487–501, 2004.

RODRIGUES, M. “Geoprocessamento.” In: *Anais do 5º Encontro Nacional de Engenheiros Cartógrafos*, v. 1, pp. 144-160, Presidente Prudente, 1988.

ROVERE, E.E.L. “Opções tecnológicas e de gestão ambiental para a despoluição da Baía de Guanabara”. Projeto Ensino e Pesquisa, Processo: 62.0009/99-3, PADCTIII/CIAMB – 4ª Rodada, 2005.

SANDRI, S.; CORREA, C. “Lógica Nebulosa”. *V Escola de Redes Neurais*, Conselho Nacional de Redes Neurais, pp. c073-c090, ITA, São José dos Campos, SP, 1999.

SANTOS, R.C.P. *Avaliação de Métodos Baseados em Sistemas FUZZY para Mineração de Dados Georeferenciados*. Dissertação de Mestrado, COPPE/UFRJ, Rio de Janeiro, Brasil, 2006.

SILVA, F.C. *Análise ROC*. 2006. Disponível em: [http://www.dpi.inpe.br/~felipe/works/inpe/spr/roc\\_analyzes.pdf](http://www.dpi.inpe.br/~felipe/works/inpe/spr/roc_analyzes.pdf), acesso em: nov/2007.

SILVA, M.P.S. *Mineração de padrões de mudança em imagens de sensoriamento remoto*. Tese de Doutorado, INPE, São José dos Campos, SP, Brasil, 2006.

SOARES, Z.O.; FRANCA, L.B.P.; UTCHITEL, S. *Fitoplâncton e fatores abióticos na Baía de Guanabara, Rio de Janeiro, subsídios para o controle da poluição*. Cadernos FEEMA, série Congressos, Rio de Janeiro, 09/81, 30 p, 1981.

SOUZA, F. J. *Modelos Neuro-Fuzzy Hierárquicos*. Tese de Doutorado, PUC-Rio, Rio de Janeiro, RJ, Brasil, 1999.

SOUZA, M. *Data Mining iMaster*, 2003. Disponível em [www.imaster.com.br](http://www.imaster.com.br). Acesso em: jul / 2007.

ORMSBY, T.; NAPOLEON, E.; BURKE, R.; GROESSL, C.; FEASTER, L. *Getting to Know ArcGIS Desktop*, ed. Redlands, California: ESRI Press, 2001.

VALENTIN, J.L.; TENENBAUM, D.R.; BONECKER, A.C.T.; BONECKER, S.L.C; NOGUEIRA, C.R.; PARANHOS, R.; VILLAC, M.C. “Caractéristiques hydrobiologiques de la Baie de Guanabara (Rio de Janeiro, Brésil)”. *J. Res. Océanographique*, 24 (1): 33-41, 1999.

VASCONCELOS, B.S. *Mineração de regras de classificação com sistemas de banco de dados objeto-relacional estudo de caso: regras de classificação de litofácies de poços de petróleo*. Dissertação de mestrado, UFCG, Campina Grande, Paraíba, 2002.

WELLING, M. “Fisher linear discriminant analysis”. Department of Computer Science University of Toronto, Canadá, 2005, disponível em: <[http://www.ics.uci.edu/~welling/classnotes/papers\\_class/Fisher-LDA.pdf](http://www.ics.uci.edu/~welling/classnotes/papers_class/Fisher-LDA.pdf)>. Acesso em: jun/2007.

ZADEH, L. A. Fuzzy sets. *Information and Control*, 8 (3), pp. 338-353, 1969.

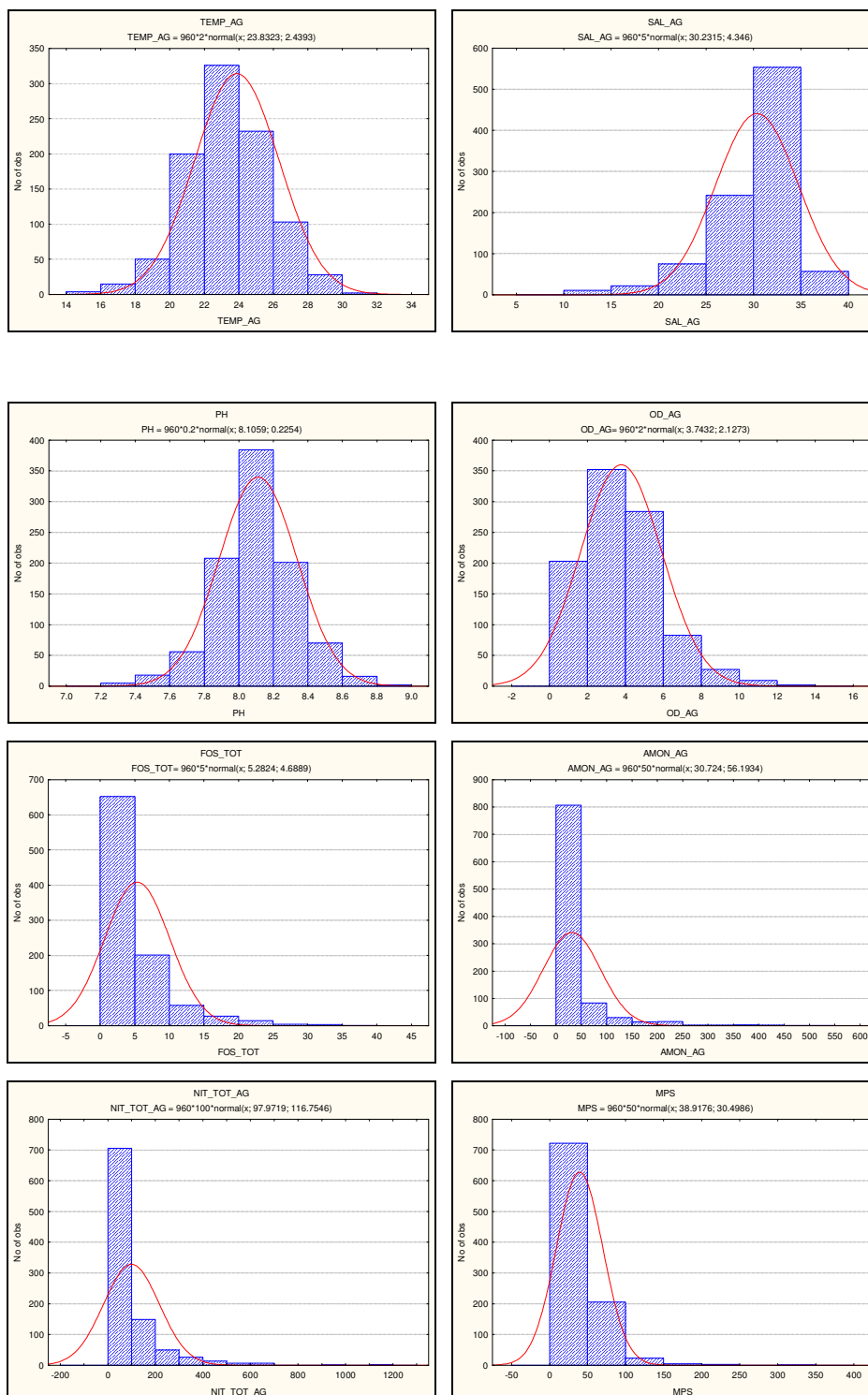
ZANUSSO, M.B. *Data mining*. Disponível em: <[http://www.dct.ufms.br/~mzanusso/Data\\_Mining.htm](http://www.dct.ufms.br/~mzanusso/Data_Mining.htm)>. Acesso em fev/2009.

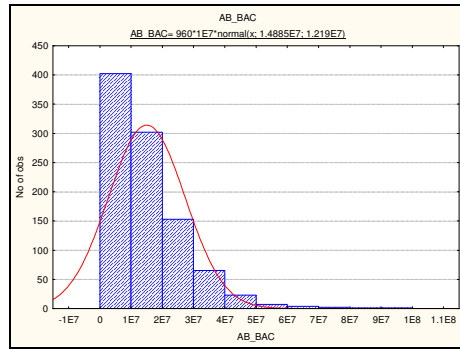
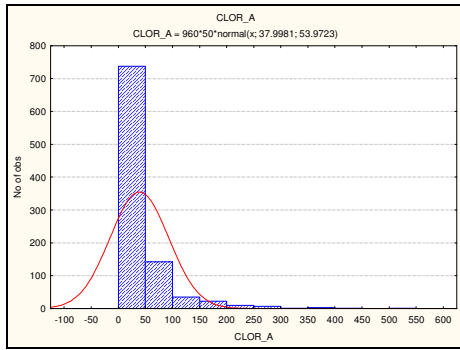
ZEILER, M. *Modeling our world: the ESRI guide to geodatabase desing*, p. 46, 1999.

ZHANG, G.P. “Neural networks for classification: a survey”. *IEEE Transactions on Systems, Man, And Cybernetics: Reviews*, v. 30, n. 4, pp. 451-462, 2000.

# ANEXO I

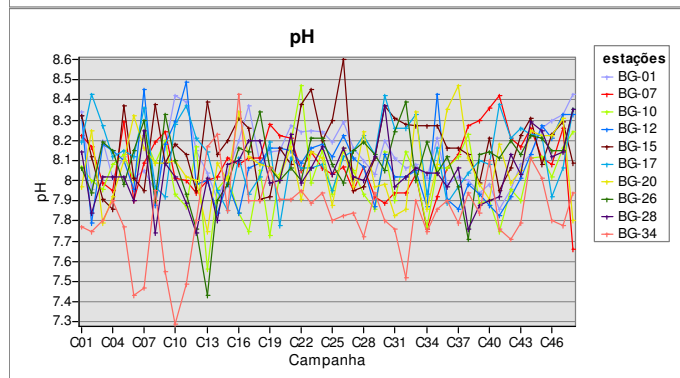
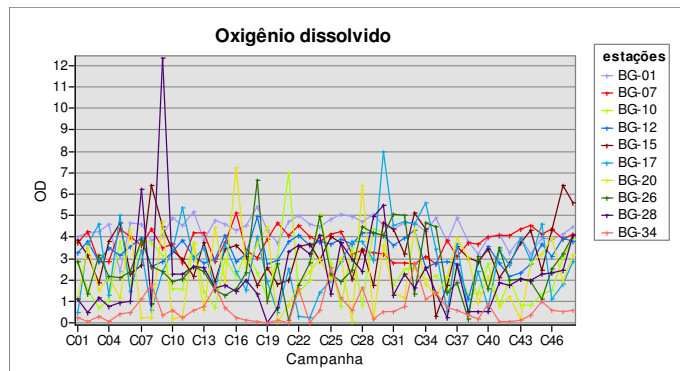
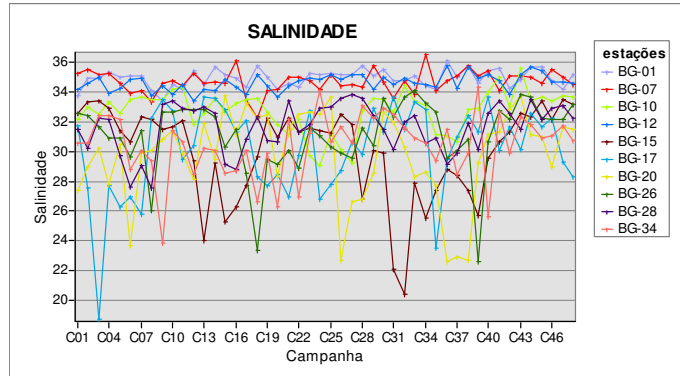
Gráfico de histograma e curva normal de cada variável físico-química avaliada no estudo de caso.

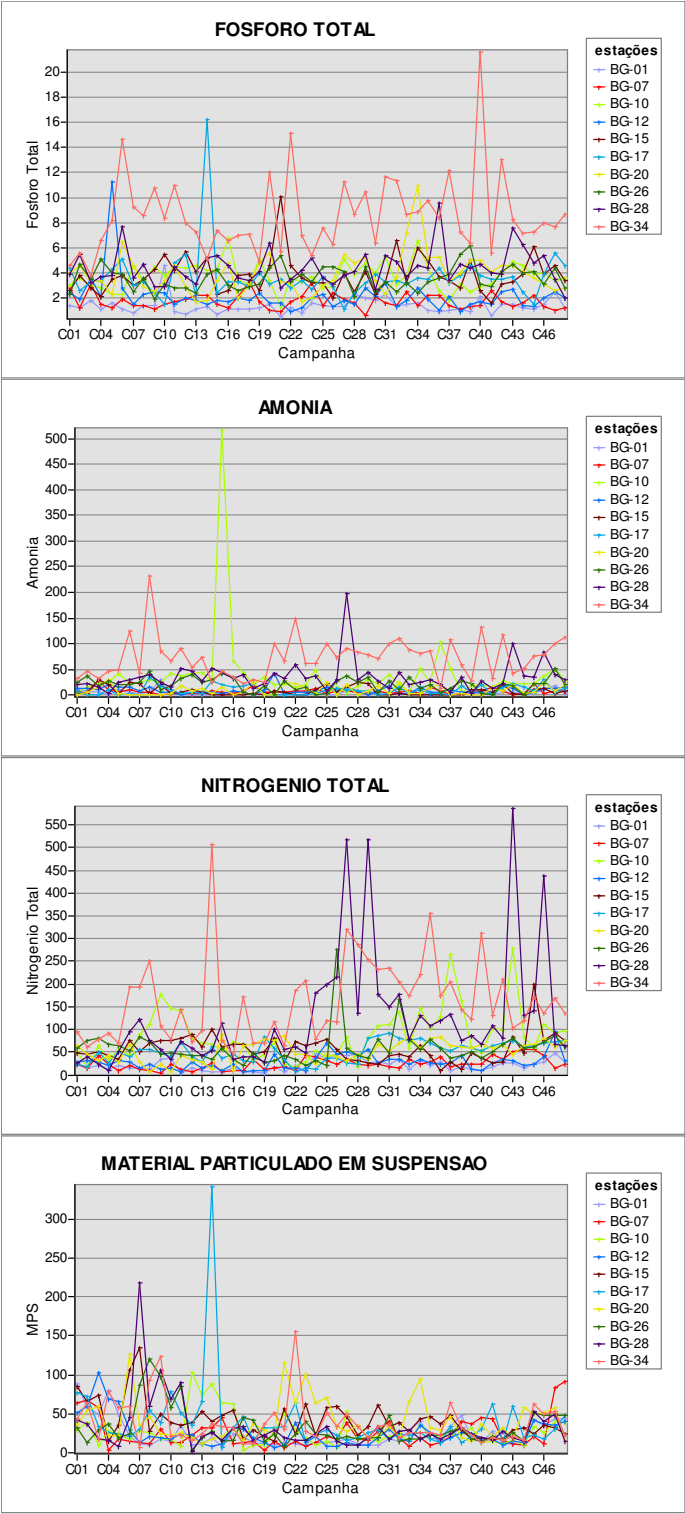


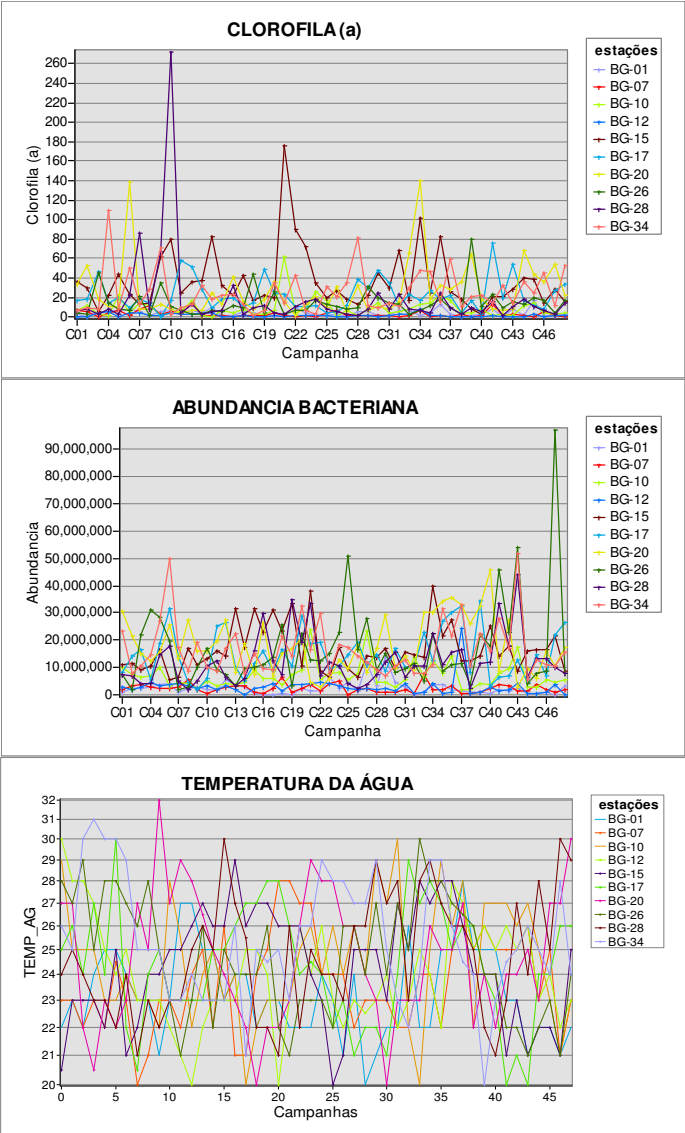


## ANEXO II

Gráficos com o comportamento das variáveis selecionadas por estação ao longo das 48 campanhas. O eixo Y do gráfico representa o valor da variável e o eixo X, as campanhas; as cores representam cada estação de coleta detalhada pela legenda do gráfico.

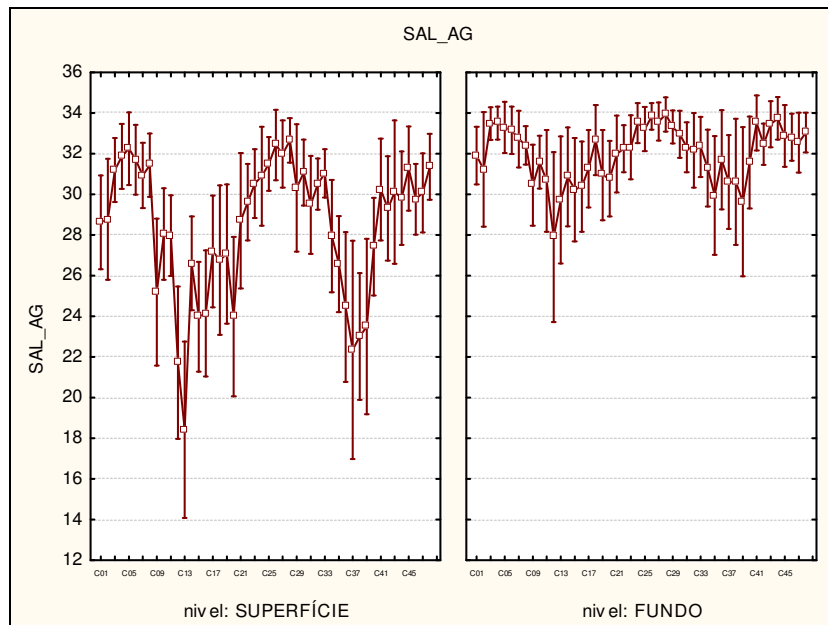
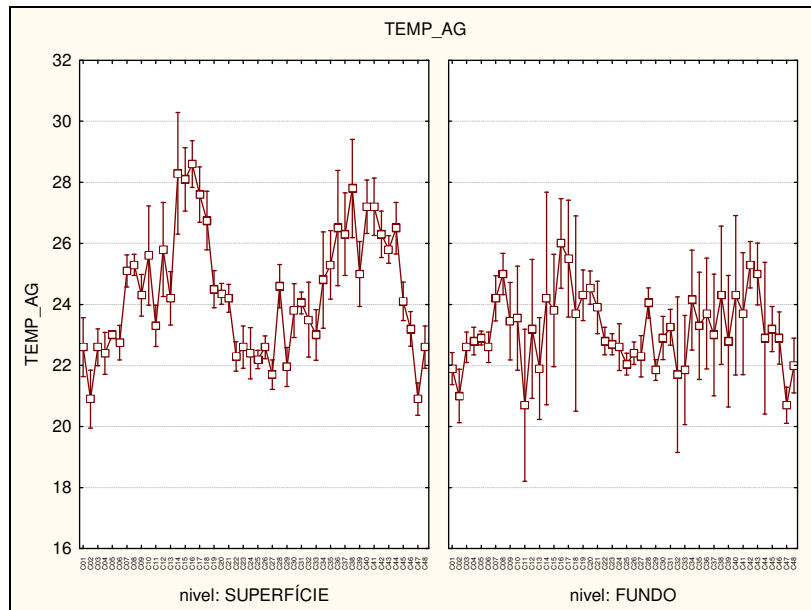


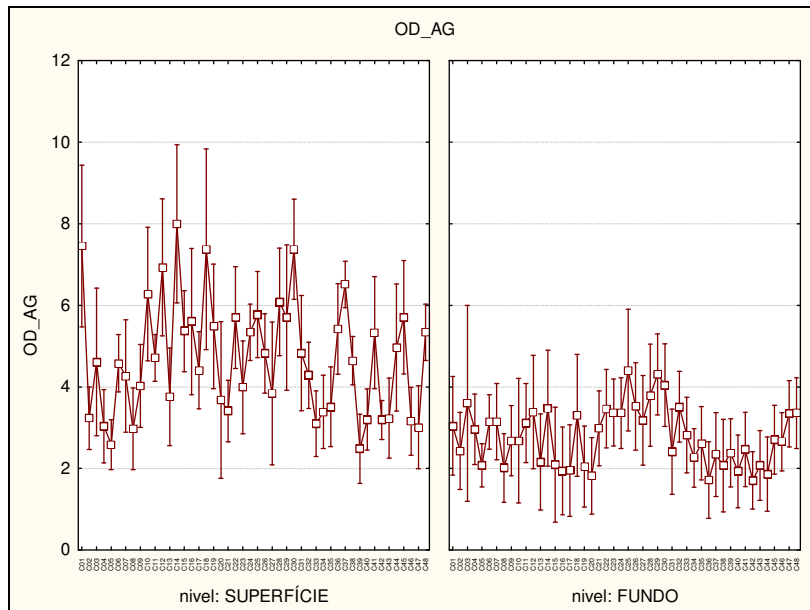
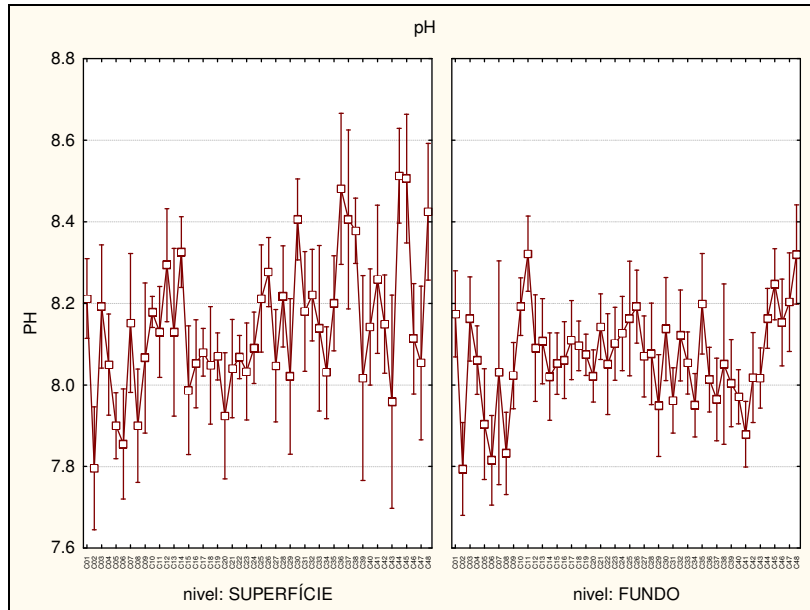


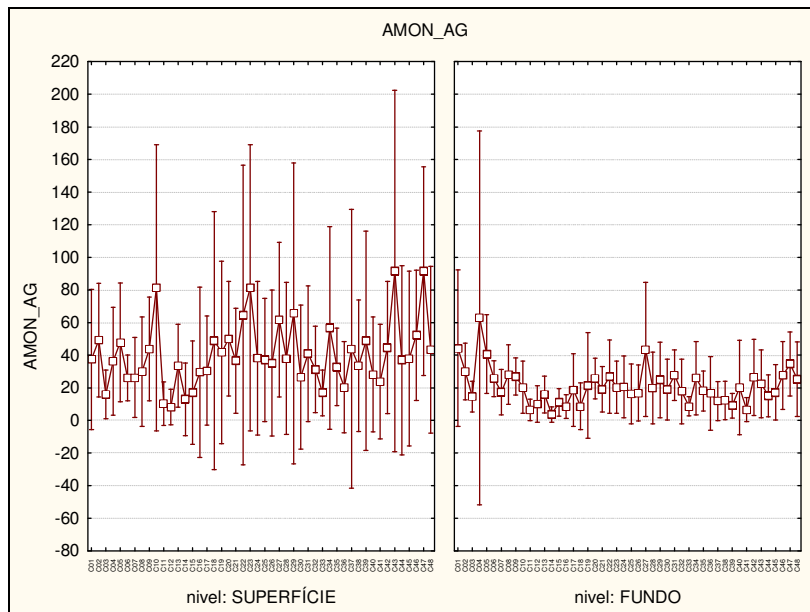
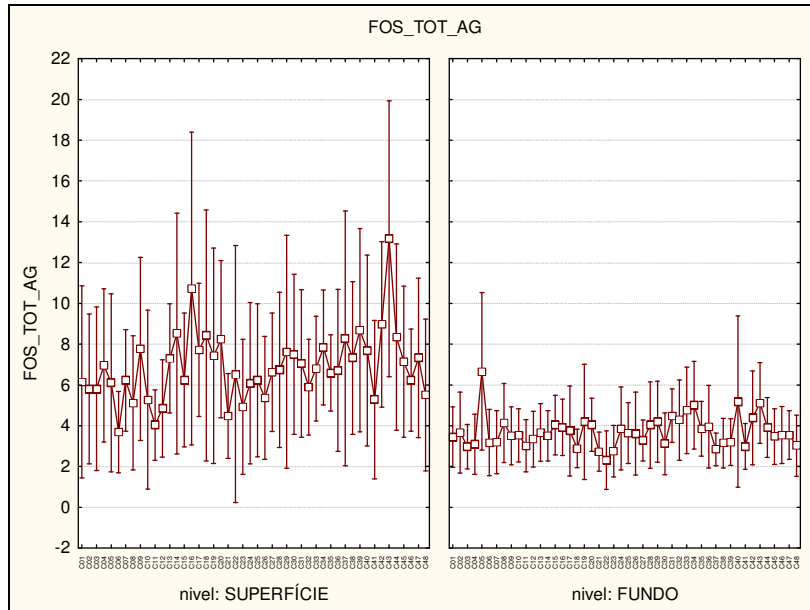


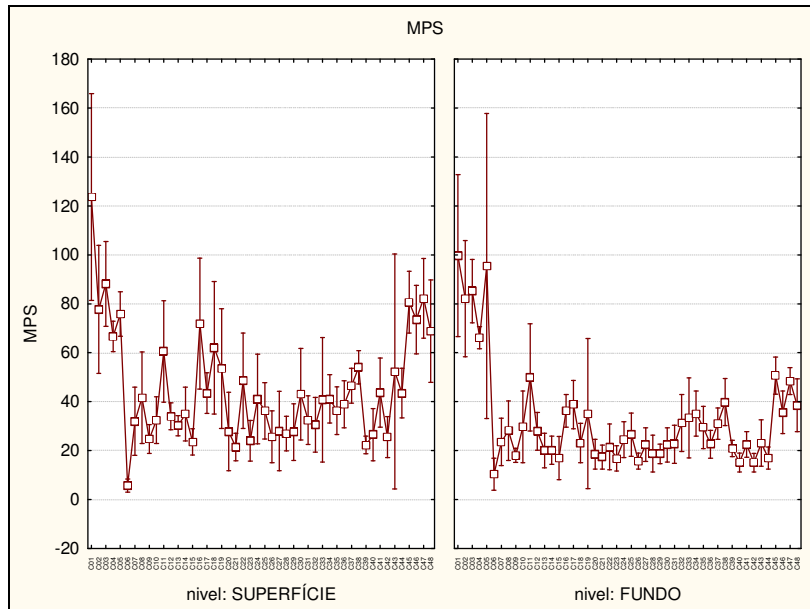
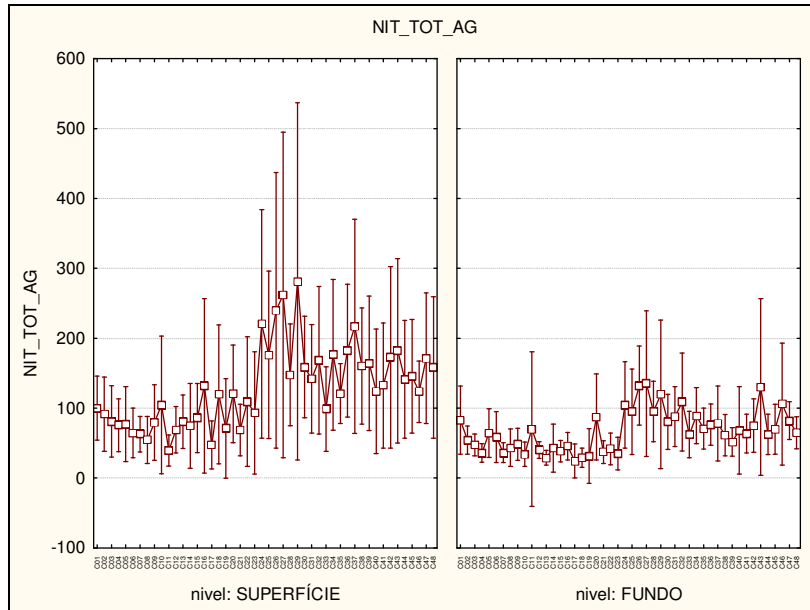
### ANEXO III

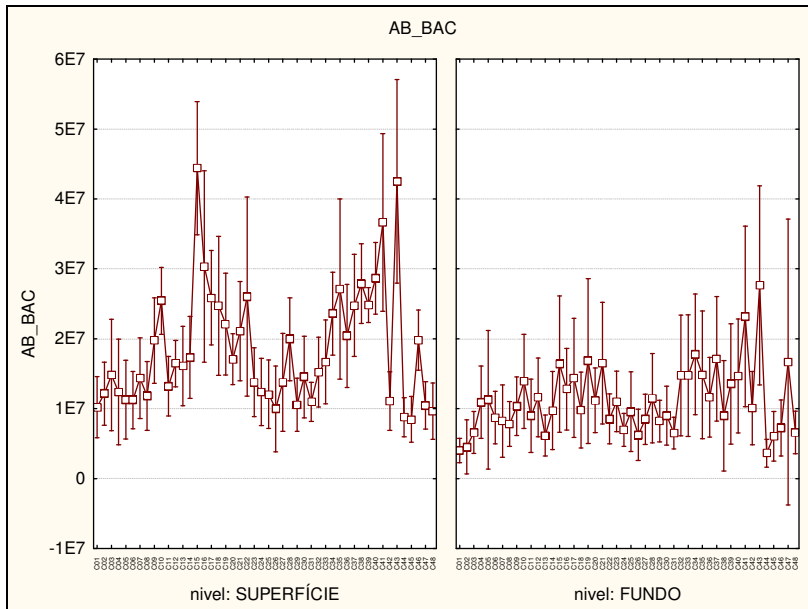
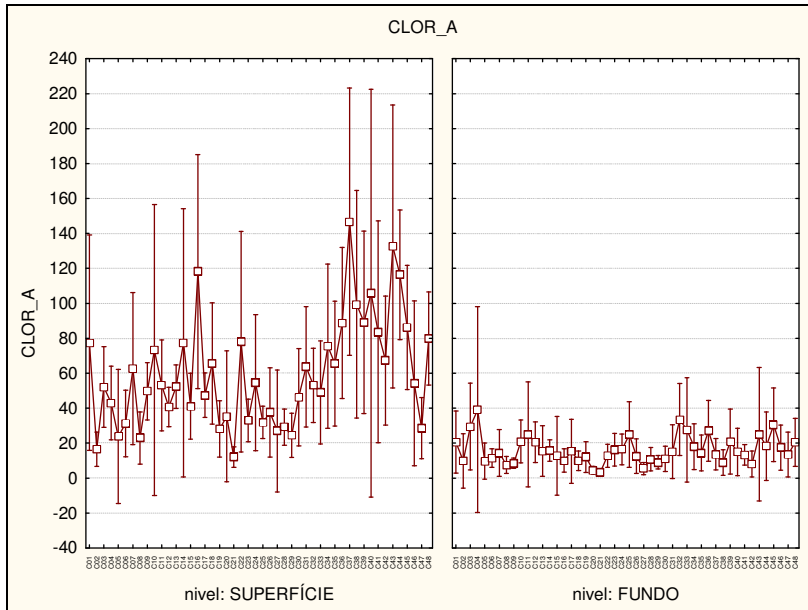
Gráficos da distribuição temporal das variáveis nas 48 campanhas (C01 até C048). O eixo x representa as campanhas e o eixo Y, os valores da variável; a simbologia  $\square$  representa o valor médio para campanha e  $\pm$  representa o intervalo de confiança para os valores considerados.





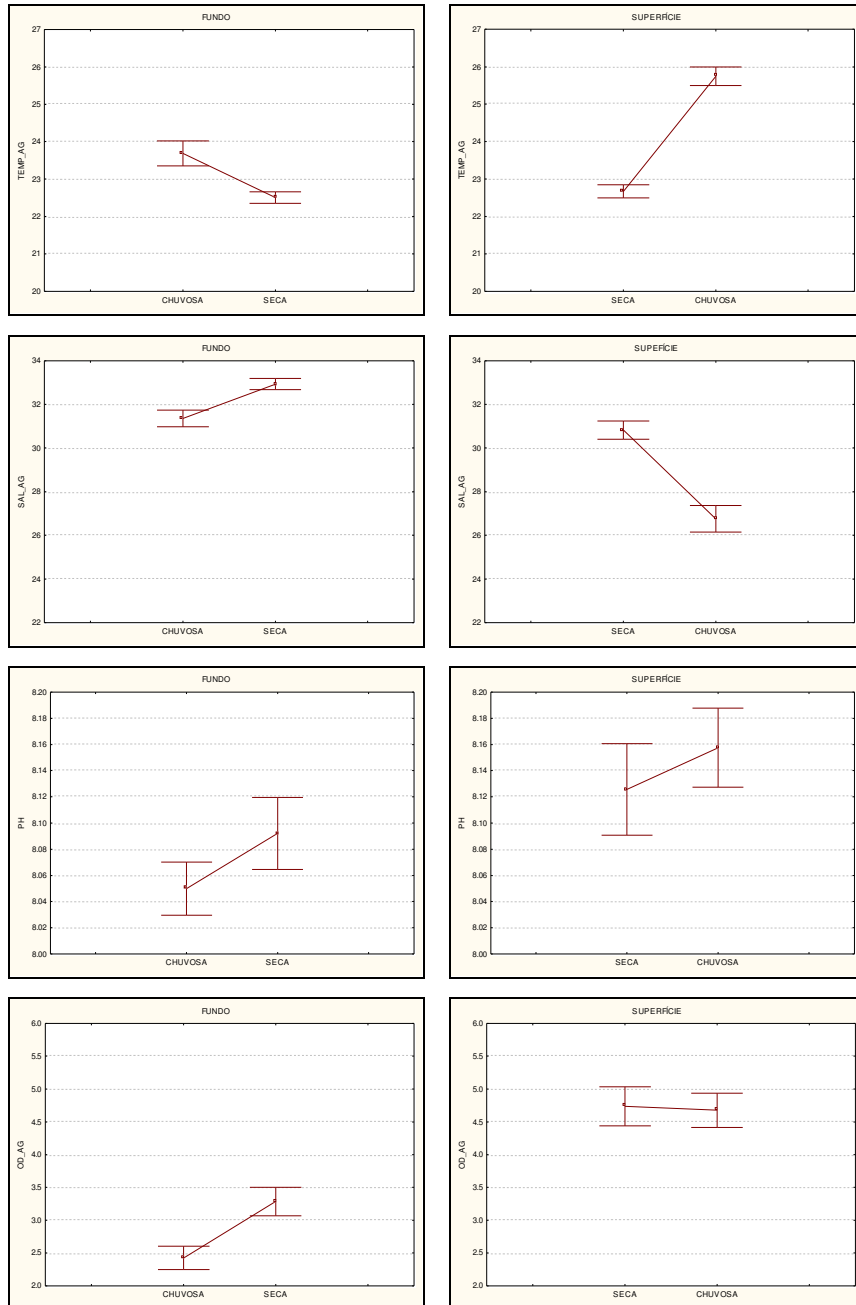


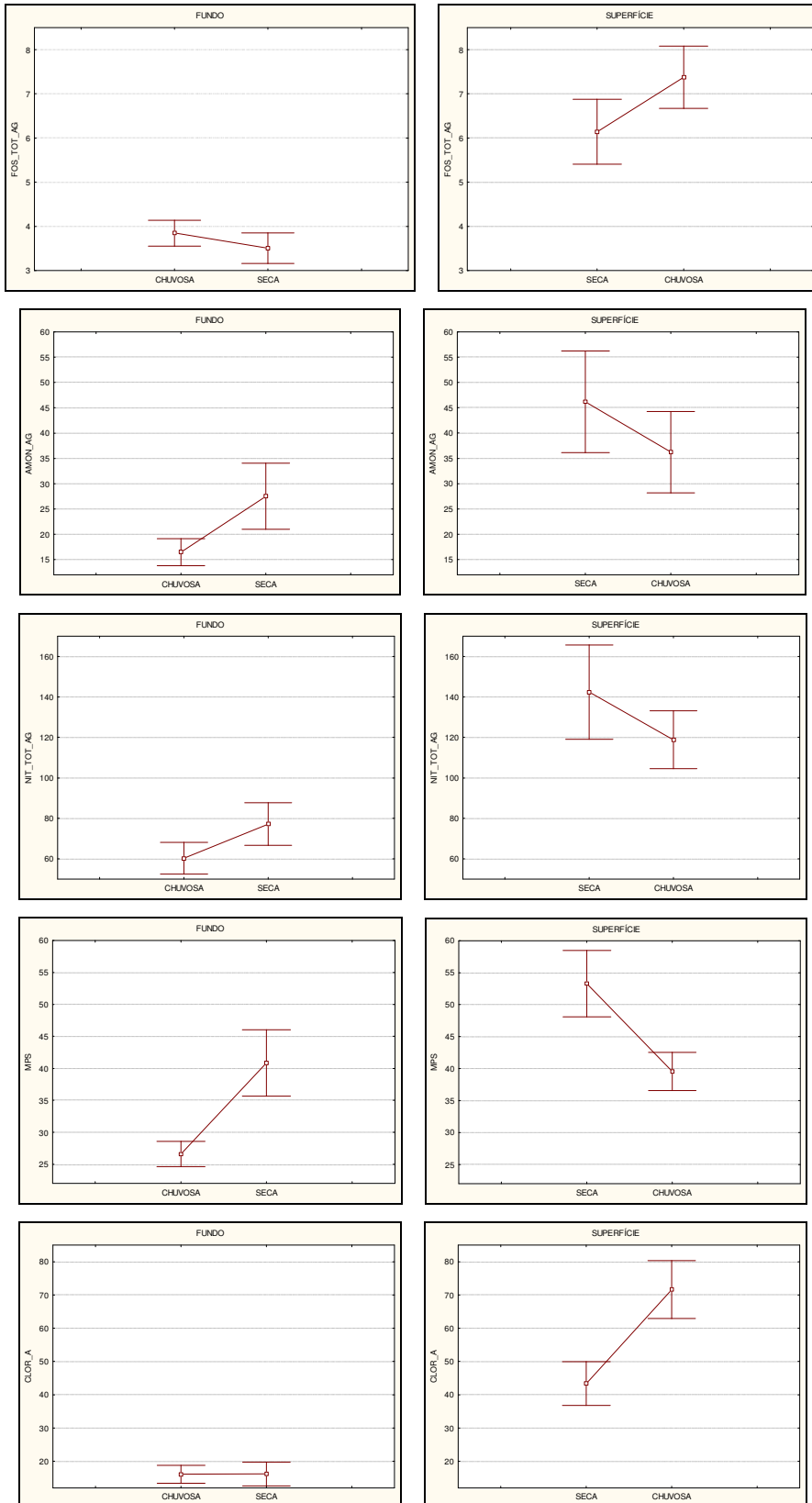


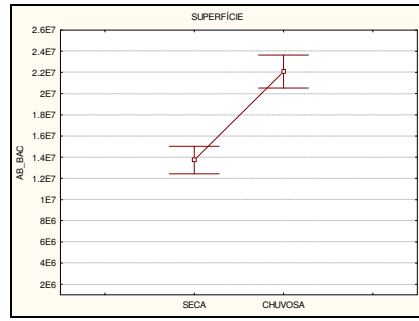
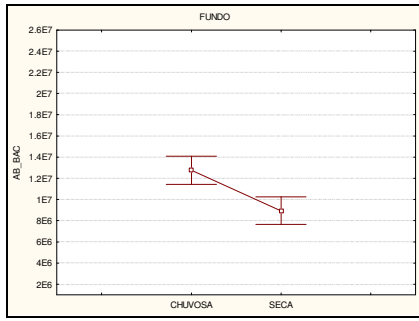


## ANEXO IV

Representação da variação média dos valores para cada variável hidrobiológica. Considerando o período (chuvoso e seca) e o nível da coleta (superfície e fundo). O eixo “Y” do gráfico corresponde ao valor da variável e o eixo “X” agrupa os dados em época chuvosa e seca.

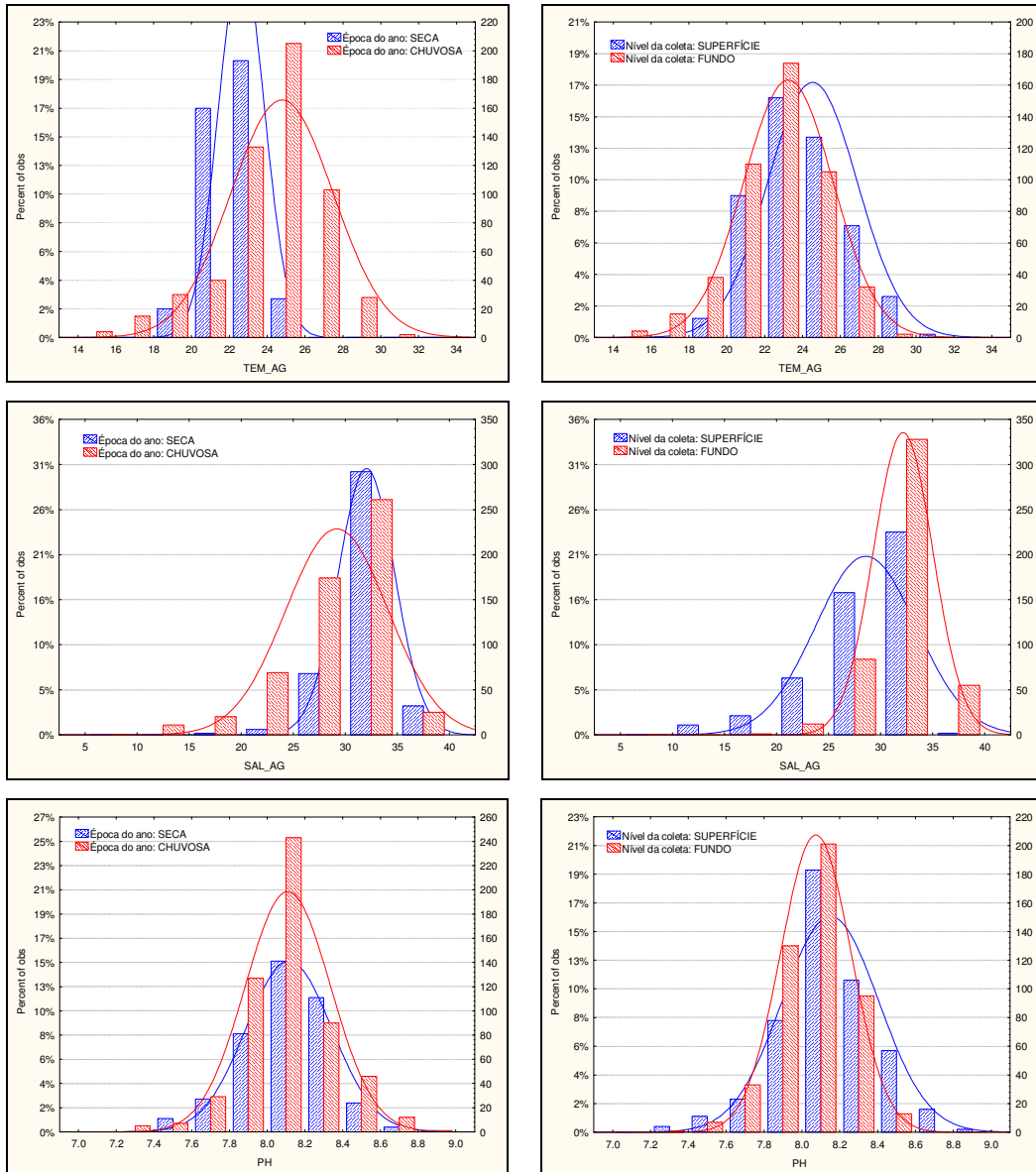


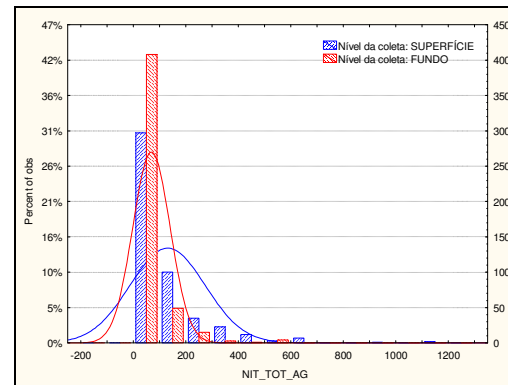
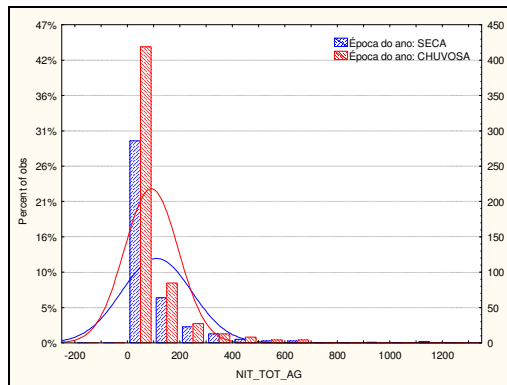
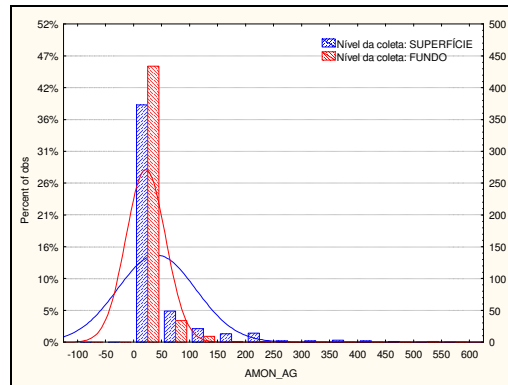
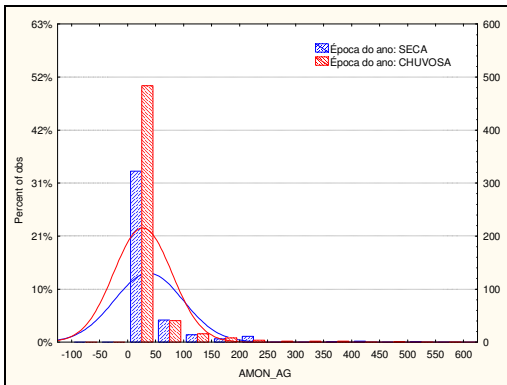
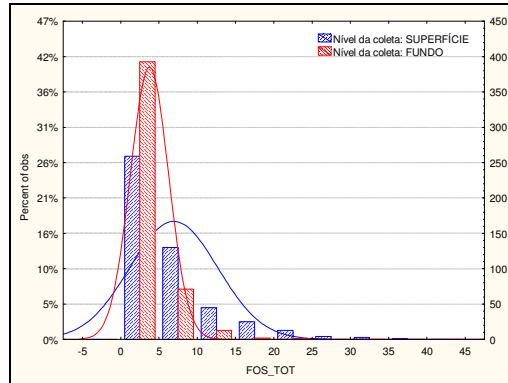
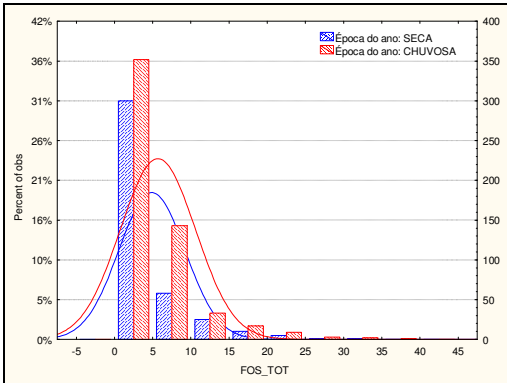
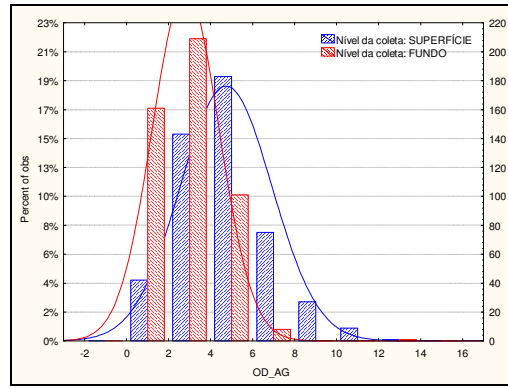
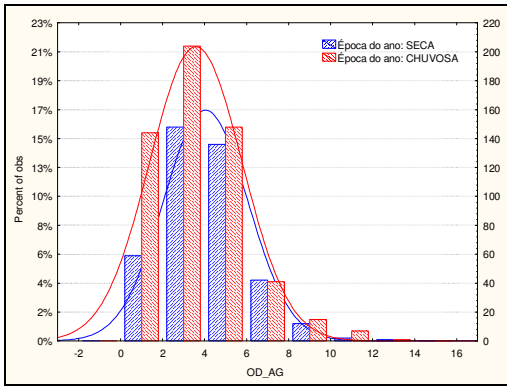


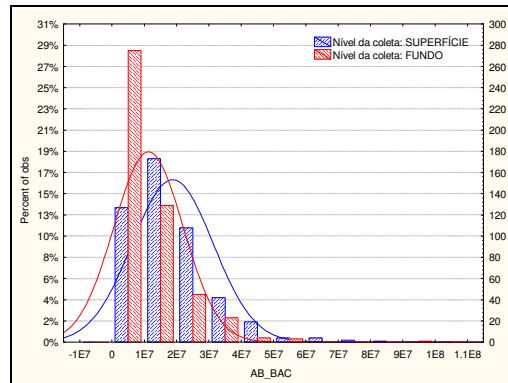
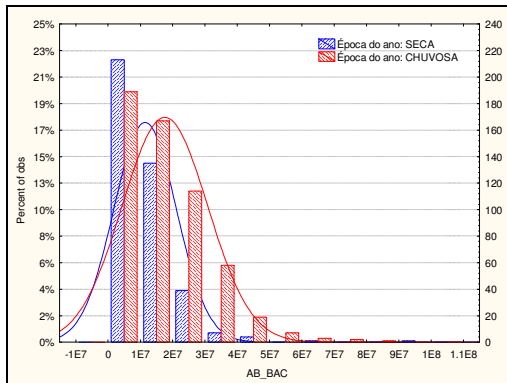
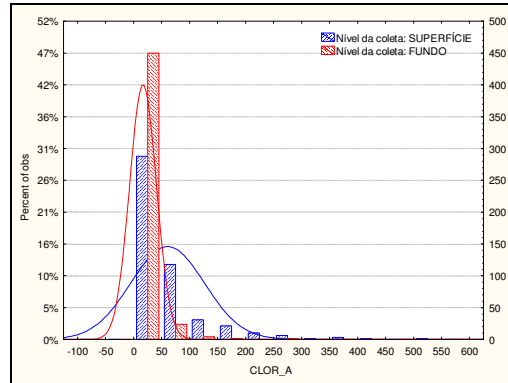
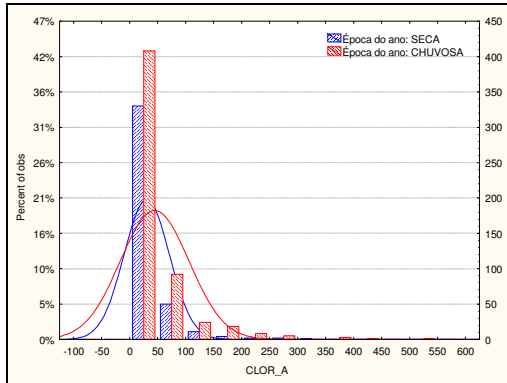
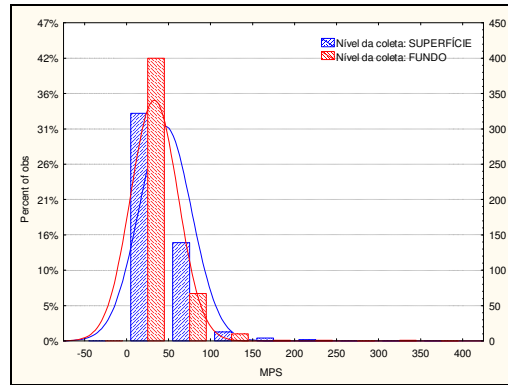
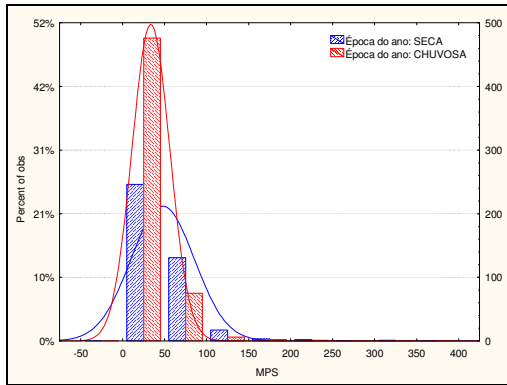


## ANEXO V

Gráficos com histogramas da distribuição dos valores para cada variável segundo o nível de coleta e a época do ano e sua curva normal. O eixo Y do lado direito apresenta a porcentagem de ocorrência, o eixo Y do lado esquerdo o número total de ocorrência e o eixo X corresponde ao valor do parâmetro.







## ANEXO VI

Apresenta os mapas comparando os resultados do classificador não-supervisionado *fuzzy* ao resultado do modelo de sobreposição ponderada, considerando a particularidade das regras das classes (Tabelas 18 – sobreposição ponderada e 19 - *fuzzy*) definidas para cada método de classificação não supervisionada, avaliando somente os dados coletados na superfície.

