



Ministério da
Ciência e Tecnologia



INPE-15217-TDI/1311

**ANÁLISE INTEGRADA DE DADOS AMBIENTAIS
UTILIZANDO TÉCNICAS DE CLASSIFICAÇÃO E
AGRUPAMENTO DE MICROARRANJOS DE DNA**

Heloisa Musetti Ruivo

Dissertação de Mestrado do Curso de Pós-Graduação em Computação Aplicada,
orientada pelo Dr. Fernando Manuel Ramos, aprovada em 17 de dezembro de 2007

Registro do documento original:

<<http://urlib.net/sid.inpe.br/mtc-m17@80/2007/12.14.12.09>>

INPE
São José dos Campos
2008

PUBLICADO POR:

Instituto Nacional de Pesquisas Espaciais - INPE

Gabinete do Diretor (GB)

Serviço de Informação e Documentação (SID)

Caixa Postal 515 - CEP 12.245-970

São José dos Campos - SP - Brasil

Tel.:(012) 3945-6911/6923

Fax: (012) 3945-6919

E-mail: pubtc@sid.inpe.br

CONSELHO DE EDITORAÇÃO:

Presidente:

Dr. Gerald Jean Francis Banon - Coordenação Observação da Terra (OBT)

Membros:

Dr^a Maria do Carmo de Andrade Nono - Conselho de Pós-Graduação

Dr. Haroldo Fraga de Campos Velho - Centro de Tecnologias Especiais (CTE)

Dr^a Inez Staciarini Batista - Coordenação Ciências Espaciais e Atmosféricas (CEA)

Marciana Leite Ribeiro - Serviço de Informação e Documentação (SID)

Dr. Ralf Gielow - Centro de Previsão de Tempo e Estudos Climáticos (CPT)

Dr. Wilson Yamaguti - Coordenação Engenharia e Tecnologia Espacial (ETE)

BIBLIOTECA DIGITAL:

Dr. Gerald Jean Francis Banon - Coordenação de Observação da Terra (OBT)

Marciana Leite Ribeiro - Serviço de Informação e Documentação (SID)

Jefferson Andrade Ancelmo - Serviço de Informação e Documentação (SID)

Simone A. Del-Ducca Barbedo - Serviço de Informação e Documentação (SID)

REVISÃO E NORMALIZAÇÃO DOCUMENTÁRIA:

Marciana Leite Ribeiro - Serviço de Informação e Documentação (SID)

Marilúcia Santos Melo Cid - Serviço de Informação e Documentação (SID)

Yolanda Ribeiro da Silva Souza - Serviço de Informação e Documentação (SID)

EDITORAÇÃO ELETRÔNICA:

Viveca Sant´Ana Lemos - Serviço de Informação e Documentação (SID)



Ministério da
Ciência e Tecnologia



INPE-15217-TDI/1311

**ANÁLISE INTEGRADA DE DADOS AMBIENTAIS
UTILIZANDO TÉCNICAS DE CLASSIFICAÇÃO E
AGRUPAMENTO DE MICROARRANJOS DE DNA**

Heloisa Musetti Ruivo

Dissertação de Mestrado do Curso de Pós-Graduação em Computação Aplicada,
orientada pelo Dr. Fernando Manuel Ramos, aprovada em 17 de dezembro de 2007

Registro do documento original:

<<http://urlib.net/sid.inpe.br/mtc-m17@80/2007/12.14.12.09>>

INPE
São José dos Campos
2008

Dados Internacionais de Catalogação na Publicação (CIP)

Ruivo, Heloisa Musetti.

R858an Análise integrada de dados ambientais utilizando técnicas de classificação e agrupamento de microarranjos de DNA / Heloisa Musetti Ruivo. – São José dos Campos : INPE, 2008.

98 p. ; (INPE-15217-TDI/1311)

Dissertação (Mestrado em Computação Aplicada) – Instituto Nacional de Pesquisas Espaciais, São José dos Campos, 2007.

Orientador : Dr. Fernando Manuel Ramos.

1. Agrupamento. 2. Classificação. 3. Biologia molecular. 4. Limnologia. 5. Climatologia. I. Título.

CDU 504:519.687

Copyright © 2008 do MCT/INPE. Nenhuma parte desta publicação pode ser reproduzida, armazenada em um sistema de recuperação, ou transmitida sob qualquer forma ou por qualquer meio, eletrônico, mecânico, fotográfico, reprográfico, de microfilmagem ou outros, sem a permissão escrita do INPE, com exceção de qualquer material fornecido especificamente com o propósito de ser entrado e executado num sistema computacional, para o uso exclusivo do leitor da obra.

Copyright © 2008 by MCT/INPE. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, microfilming, or otherwise, without written permission from INPE, with the exception of any material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use of the reader of the work.

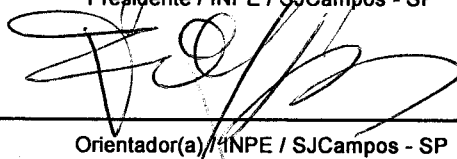
Aprovado (a) pela Banca Examinadora
em cumprimento ao requisito exigido para
obtenção do Título de Mestre em
Computação Aplicada

Dr. José Demisio Simões da Silva



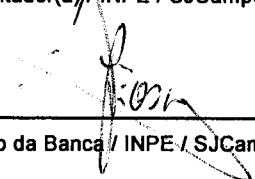
Presidente / INPE / SJCampos - SP

Dr. Fernando Manuel Ramos



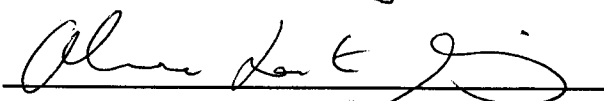
Orientador(a) / INPE / SJCampos - SP

Dr. Reinaldo Roberto Rosa



Membro da Banca / INPE / SJCampos - SP

Dr. Alexandre Souto Martinez



Convidado(a) / USP / Ribeirão P. / Ribeirão Preto - SP

Aluno (a): Heloisa Musetti Ruivo

São José dos Campos, 17 de Dezembro de 2007

“A ciência não é uma ilusão, mas seria uma ilusão acreditar que poderemos encontrar noutra lugar o que ela não nos pode dar”.

SIGMUND FREUD

AGRADECIMENTOS

Gostaria de agradecer primeiramente ao meu orientador, o Prof. Dr. Fernando Manuel Ramos, pela confiança depositada em mim para desenvolver este trabalho, pela sua dedicação e disponibilidade em sua função de orientador, e pelo seu estímulo em desenvolver um trabalho multidisciplinar. Sempre cativante, ele sabe transmitir sua energia positiva e seu conhecimento aos alunos. Tudo isso torna seu papel decisivo na elaboração desta dissertação.

Agradeço também aos colaboradores científicos: Dr. Gilvan Sampaio de Oliveira do CP-TEC/INPE sempre solícito, no fornecimento de dados climatológicos; Dr. Ivan Berguier Tavares de Lima da Empresa Brasileira de Pesquisa Agropecuária (Embrapa) - Pantanal, e Dr. Donato Seiji Abe do Instituto Internacional de Ecologia (IIE) em suas cooperações na área limnológica; e ao Dr. Eduardo Reis do Instituto de Química da USP em sua contribuição na área biológica.

À cooperação de muitos alunos da pós-graduação do INPE, dentre eles a Laurita que foi meu help biológico, sempre com muita boa vontade. Em especial à Aline Soterroni que passou seu conhecimento adquirido no mestrado sabendo compartilhar humildemente sua sabedoria.

Ao meu marido Reinaldo, pela compreensão de minha ausência em alguns momentos como também pela colaboração com o material necessário. Às minhas filhas Carla e Julia que me apoiaram e respeitaram esta minha decisão.

E finalmente a CAPES pelo apoio financeiro concedido a esta pesquisa.

RESUMO

O crescente “dilúvio de dados” na área de ciências ambientais gera um gargalo na sua análise e interpretação. Esta tendência requer cada vez mais o emprego de técnicas estatísticas e computacionais avançadas de extração do conhecimento. Na biologia molecular experimental, por exemplo, os microarranjos de DNA são, hoje em dia, uma das tecnologias chave em estudos genômicos e geram gigabytes de dados de expressão gênica. Por este motivo a bioinformática é uma das áreas pioneiras no tratamento de vastos volumes de informação. Esta dissertação tem por objetivo mostrar que é possível transpor técnicas computacionais que são utilizadas atualmente na bioinformática, para a área ambiental. As aplicações realizadas investigaram, inicialmente, quais foram os fatores climáticos associados à grande seca de 2005 na Amazônia. Como outra aplicação, procurou-se identificar quais foram as variáveis físico-químicas que controlam a emissão de gases de efeito estufa em reservatórios de hidrelétricas. Em ambas as aplicações, grandes volumes de dados originários de diferentes fontes foram organizados como se fossem experimentos de microarranjos. Os resultados obtidos comprovam que métodos de análise da bioinformática podem ser extremamente úteis na área ambiental.

INTEGRATED DATA ENVIRONMENTAL ANALYSIS USING CLASSIFICATION TECHNOLOGY AND CLUSTERING OF DNA MICROARRAY

ABSTRACT

The growing flood of data in the environmental sciences generates a bottleneck in this information extraction as well as in its analysis and interpretation. This tendency requests the employment of computational techniques and advanced statistics analysis increasingly. In the experimental molecular biology, for instance, DNA microarrays, nowadays, one of the key technologies in gene expression studies, and they generate gigabytes of gene expression data making it one of the pioneering areas in the treatment of this vast information. The objective of this dissertation is to show the possibility of transpose computational techniques currently used in the bioinformatic, into the environmental area. Initially the accomplished applications investigated, which were the climatic component for the great drought of Amazonia in 2005. As another application, we have been identified the physiochemical variables that control the emission of greenhouse effect in hydroelectric dams. In both applications, great volumes of original data of different sources were organized as microarray experiments. The results shows that methods of analysis of the bioinformatic can be extremely useful in the environmental area.

SUMÁRIO

Pág.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

1	INTRODUÇÃO	23
2	TÉCNICAS COMPUTACIONAIS DA BIOINFORMÁTICA NA ANÁLISE DE DADOS	25
2.1	Biologia Molecular	25
2.1.1	Conceitos Fundamentais	25
2.1.2	A Tecnologia de Microarranjos	27
2.2	Técnicas de Agrupamento e Classificação	29
2.3	Pacote BRB-ArrayTools	33
2.4	Validação do BRB-ArrayTools em biologia molecular	39
3	APLICAÇÃO EM CLIMATOLOGIA	43
3.1	Dados Analisados	44
3.1.1	Dados em Grade	44
3.1.2	Séries Temporais	46
3.2	Resultados	50
3.2.1	Casos Estudados	50
3.2.2	Análise dos Resultados	56
4	APLICAÇÃO EM LIMNOLOGIA	61
4.1	Projeto Carbono Furnas	62
4.2	Banco de Dados Analisado	64
4.3	Resultados	66
5	CONCLUSÃO	77
	REFERÊNCIAS BIBLIOGRÁFICAS	79
A	- Relação de parâmetros analisados no Projeto Climatológico	83

B - Relação de parâmetros analisados no Projeto Carbono Furnas . . .	91
--	----

LISTA DE FIGURAS

	<u>Pág.</u>
2.1 Estrutura física da célula	26
2.2 Molécula de DNA produzida pelas substâncias químicas chamadas nucleotídeos que ocorrem em pares: adenina (A) com timina (T), e guanina (G) com citosina (C).	26
2.3 Origem das proteínas: a informação é copiada, base a base, para uma fita de DNA dentro da fita do RNA mensageiro (mRNA), e transferida para fora do núcleo (dentro do citoplasma) às organelas chamadas ribossomos. Aqui o mRNA direciona a montagem do aminoácido que dá origem à proteína.	27
2.4 Construção de Microarray.	29
2.5 Diagrama bi-dimensional (dendograma).	32
2.6 Tabela representativa do Preditor Multivariável - Para cada array excluído da amostra, é apresentado o número de genes que apresentam significância em relação ao restante e se o preditor o classificou corretamente.	40
2.7 Clusterização dos 52 genes (colunas) em relação ao alto e baixo GS (linhas). Genes em verde sobrepõem-se aos encontrados na análise publicada em Reis et al. (2004).	40
2.8 Representação dos pacientes em escala multidimensional.	41
2.9 Matriz de expressão dos 56 genes (colunas) em relação ao alto e baixo GS (linhas). Genes rotulados com círculos verde sobrepõem-se aos encontrados na Figura 2.6.	41
3.1 Seca do Amazonas - 2005.	43
3.2 Região analisada: 140W à 0W, 40N à 40S.	45
3.3 Localização das três regiões analisadas	46
3.4 Mapa do Pacífico Sul, evidenciando Darwin na Austrália e Tahiti, uma das ilhas do Pacífico.	47
3.5 Série temporal SOI.	47
3.6 <i>North Atlantic Oscillation</i>	48
3.7 PDO - Temperatura da superfície do mar na época de inverno.	49
3.8 Agrupamento do índice integrado.	51
3.9 Localização geográfica dos parâmetros mais relevantes na análise do índice integrado.	52
3.10 Agrupamento do índice de vazão do Rio Amazonas em Óbidos.	52
3.11 Localização geográfica dos parâmetros mais relevantes na análise do Rio Amazonas em Óbidos.	53

3.12	Agrupamento do índice de vazão do Rio Amazonas em Óbidos utilizando-se 3 classes: $[-1; -0.1]$, $[-0.1; 0.2]$ e $[0.2; 1]$.	54
3.13	Localização geográfica dos parâmetros mais relevantes na análise do Rio Amazonas em Óbidos (3 classes).	54
3.14	Agrupamento do índice de vazão do Rio Amazonas em Óbidos utilizando-se 3 classes: $[-1; \textit{mediana} - 0.03]$, $[\textit{mediana} - 0.031; \textit{mediana} + 0.03]$ e $[\textit{mediana} + 0.03; 1]$.	55
3.15	Localização geográfica dos parâmetros mais relevantes na análise do Rio Amazonas em Óbidos (3 classes - caso2).	56
3.16	Série de vazão do Rio Amazonas em Óbidos nos períodos compreendidos entre Jan/2000 e Dez/2006.	57
3.17	Agrupamento do índice de vazão do Rio Amazonas em Óbidos utilizando-se 3 classes - períodos de Jun à Nov em 2000 e 2005.	58
3.18	Localização geográfica dos parâmetros mais relevantes encontrados na Figura .3.17	58
3.19	Histograma das parâmetros mais relevantes encontrados nas abordagens estudadas.	59
3.20	Localização geográfica dos parâmetros mais relevantes encontrados nas abordagens estudadas.	59
4.1	O Efeito Estufa	62
4.2	Vista esquemática dos processos lentos e rápidos do ciclo de carbono. Aqui é mostrado como ocorre a velocidade de trocas de carbono entre reservatórios, afetando todo o ciclo (INPE et al., 2006).	63
4.3	Represas pertencentes ao Projeto Furnas analisadas (INPE et al., 2006).	65
4.4	Agrupamento - Fluxo CH_4 (bolha), interface água-atmosfera. Análise com 12 campanhas.	67
4.5	Histograma dos parâmetros relevantes - Fluxo CH_4 (bolha), interface água-atmosfera.	67
4.6	Agrupamento dos parâmetros em comum nas duas etapas - Fluxo CH_4 (bolha), interface água-atmosfera.	68
4.7	Agrupamento - Fluxo CO_2 (bolha), interface água-atmosfera. Análise com 12 campanhas.	69
4.8	Histograma dos parâmetros relevantes - Fluxo CO_2 (bolha), interface água-atmosfera.	69
4.9	Agrupamento dos parâmetros em comum nas duas etapas - Fluxo CO_2 (bolha), interface água-atmosfera.	70
4.10	Gráfico comparativo das medidas normalizadas de CO_2 e CH_4 (bolha), interface água-atmosfera.	70

4.11 Gráfico em escala multidimensional - Fluxo CO_2 e CH_4 (bolha), interface água-atmosfera.	71
4.12 Agrupamento - Fluxo CH_4 interface sedimento-água. Análise com 12 campanhas.	72
4.13 Histograma dos parâmetros relevantes - Fluxo CH_4 , interface sedimento-água.	72
4.14 Agrupamento dos parâmetros em comum nas duas etapas - Fluxo CH_4 , interface sedimento-água.	72
4.15 Gráfico em escala multidimensional - Fluxo CH_4 , interface sedimento-água. . .	73
4.16 Agrupamento - Fluxo CO_2 , interface sedimento-água. Análise feita com as 12 campanhas.	74
4.17 Histograma dos parâmetros relevantes - Fluxo CO_2 , interface sedimento-água.	74
4.18 Agrupamento dos parâmetros em comum nas duas etapas - Fluxo CO_2 , interface sedimento-água.	75
4.19 Imagem da Represa Manso.	75

LISTA DE TABELAS

	<u>Pág.</u>
4.1 Campanhas por reservatório	64

LISTA DE ABREVIATURAS E SIGLAS

MA	–	Microarranjo, Microarray
DNA	–	Ácido desoxirribonucléico
mRNA	–	Ribonucléico mensageiro
SOM	–	Self-organized map
GS	–	Grau de Gleason
PC	–	Componente principal
ENSO	–	El Niño Southern Oscillation
SOI	–	Southern Oscillation Index
MEI	–	Multivariate ENSO Index
NAO	–	North Atlantic Oscillation
PDO	–	Pacific Decadal Oscillation
SST	–	Temperatura da Superfície do Mar
SLP	–	Pressão ao Nível do Mar
GEE	–	Gases do Efeito Estufa
CH_4	–	Metano
CO_2	–	Gás Carbônico
N_2O	–	Óxido Nitroso
N_2	–	Nitrogênio
O_2	–	Oxigênio
COT	–	Carbono Orgânico Total
COD	–	Carbono Orgânico Dissolvido
DIC	–	Carbono Inorgânico Dissolvido
DOC	–	Carbono Orgânico Dissolvido
POC	–	Carbono Orgânico Particulado (reduzido a partículas)
NT	–	Nitrogênio Total
NOT	–	Nitrogênio Orgânico Total
PT	–	Fósforo Total
PH	–	Potencial hidrogeniônico
FITO	–	Fitoplâncton

1 INTRODUÇÃO

Uma das conseqüências da crescente preocupação mundial com o meio ambiente é o aumento acentuado do volume de dados disponíveis para a comunidade científica. Alguns pesquisadores já se referem a este fenômeno recente como “dilúvio de dados” (HEY; TREFETHEN, 2003). Trata-se naturalmente de um fenômeno positivo, mas que tem gerado uma crescente demanda por ferramentas computacionais para análise e extração de conhecimento de bancos de dados cada vez maiores e mais complexos. Por exemplo, na Europa, os satélites da Agência Espacial Européia (ESA) geram atualmente por volta de 100 Gigabytes por dia de dados (HEY; TREFETHEN, 2003). Apenas no *European Centre for Medium Range Weather Forecasting* (ECMWF), no Reino Unido, aproximadamente 4 milhões de campos meteorológicos são adicionados diariamente ao seu banco de dados. Este padrão é similar nos EUA, onde estima-se que em 2007 serão produzidos 15 Petabytes de dados.

Uma das áreas pioneiras no tratamento de grandes volumes de dados é a Bioinformática. O genoma humano, por exemplo, contém por volta de 3.2 Gigabases ou, aproximadamente, um Gigabyte de informação. Se considerarmos as 100 mil proteínas e os 32 milhões de aminoácidos expressos pelo genoma humano, esta quantidade de informação cresce para 200 Gigabytes. E se incluirmos as medidas de difração de raio X da estrutura destas macromoléculas, o volume de dados expande-se e atinge rapidamente a casa dos Petabytes (HEY; TREFETHEN, 2003). Por este motivo, desde os seus primórdios a Bioinformática desenvolveu técnicas computacionais eficazes para a análise de grandes volumes de dados (AMARATUNGA; CABRERA, 2004). Na biologia molecular experimental, por exemplo, os Microarranjos (ou Microarrays - MA) de DNA são hoje em dia uma das tecnologias chave em estudos genômicos. Os MA permitem um monitoramento dos níveis de expressão de milhares de genes simultaneamente. Através da análise destes dados é possível agrupar os genes com base nas semelhanças existentes entre os seus perfis de expressão nas diversas condições analisadas. Esta abordagem tem sido utilizada por exemplo, para diagnosticar e classificar tumores em subgrupos relevantes, assim como para identificar marcadores moleculares que permitam prever a evolução clínica ou a resposta à quimioterapia em novos pacientes diagnosticados com câncer (AMARATUNGA; CABRERA, 2004).

Este trabalho insere-se no contexto acima delineado, e tem por objetivo demonstrar que é possível transpor técnicas computacionais que são utilizadas atualmente na bioinformática, em particular na análise de experimentos de MA, para a área ambiental. Para isso o pacote BRB-ArrayTools (SIMON; LAM, 2006) foi adaptado e aplicado ao estudo de dois problemas ambientalmente relevantes em climatologia e em limnologia. O pacote BRB-ArrayTools é uma ferramenta computacional para análise de dados de MA de pacientes

com câncer. Trata-se de um software livre desenvolvido pelo *Biometric National Cancer Institute*, que permite processar, agrupar, classificar e visualizar dados de expressão gênica de vários experimentos simultaneamente.

Na primeira aplicação realizada neste trabalho foram investigados que os fatores meteorológicos são responsáveis pela grande seca de 2005 na Amazônia. Na segunda, busca-se identificar quais são as variáveis físico-químicas que controlam a emissão de gases de efeito estufa em reservatórios de hidroelétricas. Em ambas aplicações, grandes volumes de dados, originários de diferentes fontes, foram organizados como se fossem experimentos de MA e, em seguida, analisados sem a imposição de restrições a priori. Apesar do foco aqui ser a transposição da ferramenta computacional para a área ambiental, os resultados das aplicações realizadas são extremamente relevantes e merecem ser destacados como contribuições desta dissertação.

Esta dissertação organiza-se da seguinte maneira. O Capítulo 2 revisa alguns conceitos fundamentais em Bioinformática e Biologia Molecular, úteis ao entendimento dos demais capítulos, além de descrever o pacote computacional BRB-ArrayTools. Os Capítulos 3 e 4 apresentam os resultados das aplicações em Climatologia e Limnologia. As conclusões desta dissertação e sugestões de trabalhos futuros encontram-se no Capítulo 5.

2 TÉCNICAS COMPUTACIONAIS DA BIOINFORMÁTICA NA ANÁLISE DE DADOS

Neste capítulo serão analisadas as técnicas computacionais utilizadas na análise de dados da biologia molecular, que serão posteriormente empregadas na área ambiental.

2.1 Biologia Molecular

A Biologia Computacional diz respeito à utilização de técnicas e ferramentas de computação para a resolução de problemas da Biologia. Nesse contexto, a computação pode ser aplicada na resolução de problemas como comparação de seqüências (de DNA, RNA e proteínas), montagem de fragmentos de DNA, reconhecimento de genes, identificação e análise da expressão de genes, e determinação da estrutura de proteínas (BALDI; BRUNAK, 2001). Recentemente, Simpson (1999) considerou que enquanto a determinação da estrutura do DNA foi o coroamento da pesquisa biológica em meados do século XX, a decifração de seu conteúdo informacional é a grande aventura da entrada do novo milênio.

A biologia molecular retrata o estudo das células e moléculas. Nosso organismo é composto por diversas células que são definidas como a unidade fundamental dos seres vivos, ou a menor unidade capaz de manifestar as propriedades de um ser vivo. As células são capazes de sintetizar seus componentes, de crescer e de se multiplicar. A seguir serão explicados alguns conceitos biológicos necessários para entender as técnicas de análise de dados de MA.

2.1.1 Conceitos Fundamentais

De acordo com a organização estrutural, as células são divididas em duas classes: procarióticos (seres unicelulares, por exemplo, bactérias) e eucarióticos (seres vivos complexos, como os humanos). Elas estão envolvidas numa membrana - chamada membrana celular - que contém uma substância rica em água, chamada citoplasma. O citoplasma compreende todo o volume da célula, com exceção do núcleo (DOMANY, 2003).

O núcleo é o cérebro das células eucarióticas e controla todas as suas atividades, representando assim o centro de coordenação celular. Todas as células contêm a informação hereditária (genética) codificada em moléculas de ácido desoxirribonucléico (DNA) como pode-se observar na Figura 2.1 (DOMANY, 2003).

As moléculas de DNA são compostas por dois filamentos complementares enrolados entre si na forma de uma dupla hélice (Figura 2.2). Cada filamento consiste de uma seqüência linear de bases nitrogenadas conectada a outro filamento por pontes de hidrogênio.

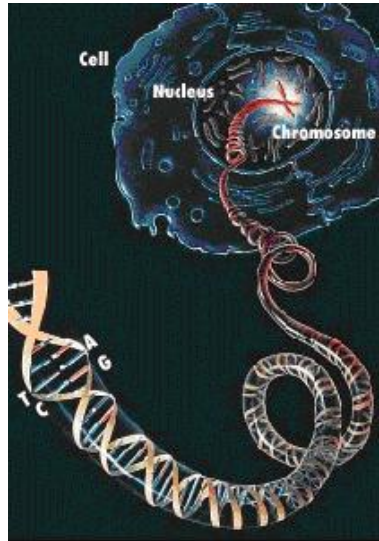


Figura 2.1 - Estrutura física da célula

Fonte: www.medonline.com.br/med-ed/med6/gene.gif

Existem quatro tipos de bases nitrogenadas: citosina, guanina, adenina e timina.

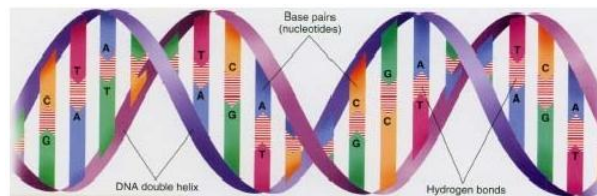


Figura 2.2 - Molécula de DNA produzida pelas substâncias químicas chamadas nucleotídeos que ocorrem em pares: adenina (A) com timina (T), e guanina (G) com citosina (C).

As informações genéticas são codificadas na sequência linear em que a base dos dois filamentos é ordenada dentro da molécula de DNA. Os segmentos codificados do DNA são chamados *genes*. Os genes especificam a estrutura das proteínas que são macromoléculas responsáveis por realizar todo o trabalho importante num organismo (DOMANY, 2003). A propriedade mais importante dos genes está no fato de que eles codificam proteínas, componentes essenciais de todo ser vivo. Os genes dão origem às proteínas em dois passos (Figura ??). Primeiro, o DNA é transcrito ao ribonucleico mensageiro (mRNA) em um processo chamado transcrição e depois é transformado em proteínas, em um processo chamado tradução. A transcrição de DNA é a reprodução de uma fita de DNA em uma sequência de RNA complementar. Numa célula característica, cerca de 10.000 a 20.000 genes são expressos simultaneamente. O nível de expressão do gene é um número que

mede a quantidade de mRNA associado a um gene particular (HAUTANIEMI, 2003). Está relacionado com a quantidade de proteína que o gene produz. Os principais objetivos da análise de expressão gênica são revelar padrões presentes na expressão dos genes de diferentes tecidos (conjuntos de dados compostos por centenas de milhares de medidas, com padrões de similaridade e dissimilaridade) e apresentar resultados da análise em uma forma amigável e de fácil compreensão (CARVALHO, 2003).

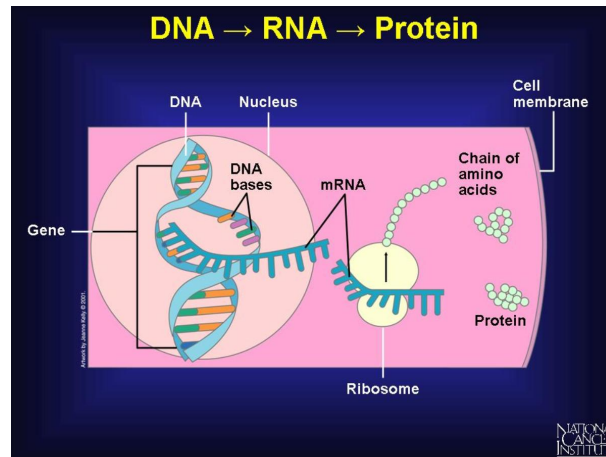


Figura 2.3 - Origem das proteínas: a informação é copiada, base a base, para uma fita de DNA dentro da fita do RNA mensageiro (mRNA), e transferida para fora do núcleo (dentro do citoplasma) às organelas chamadas ribossomos. Aqui o mRNA direciona a montagem do aminoácido que dá origem à proteína.

O Genoma é o conjunto de genes de uma espécie. É a coleção de todas as formulações químicas das proteínas que um organismo precisa e produz. O genoma humano possui entre 30.000 e 40.000 genes.

2.1.2 A Tecnologia de Microarranjos

Um dos maiores avanços no estudo do genoma é a utilização de MA de DNA para medir o nível de expressão de milhares de genes simultaneamente (KRUTOVSKII; NEALE, 2001). A idéia fundamental é comparar níveis de expressão do gene entre duas amostras de tecidos, uma normal e outra com tumor (HAUTANIEMI, 2003). A tecnologia de MA é um processo baseado em hibridização que possibilita observar a concentração de mRNA de uma amostra de células analisando a luminosidade de sinais fluorescentes. Hibridização é o processo bioquímico onde duas fitas de ácido nucléico com seqüências complementares se combinam (DANTAS, 2004). Uma lâmina de MA tem, em cada posição (spot), pedaços de cDNA (DNA complementar) de um gene que se quer estudar. Uma única lâmina consiste de um conjunto ordenado de dezenas a centenas de milhares de grupos distintos de milhões

de moléculas únicas de DNA cujas sequências são conhecidas; cada grupo é fixado a uma superfície rígida, normalmente de vidro, numa posição previamente definida.

A metodologia de preparação consiste em:

- cultivam-se células em 2 soluções distintas:
 - uma correspondendo à situação tida como padrão - controle,
 - outra correspondendo à situação a estudar - condição ou teste.
- os ambientes de desenvolvimento das culturas apenas deverão distinguir-se nos aspectos a serem estudados;
- extrai-se o mRNA das 2 culturas;
- marca-se o mRNA de uma das culturas com corante fluorescente vermelho e a outra com corante fluorescente verde;
- mistura-se o mRNA das 2 culturas;
- hibridiza-se o microarray de DNA com a mistura do mRNA das 2 culturas.

Como resultado tem-se que em cada ponto do microarray de DNA encontram-se moléculas de DNA que hibridizam apenas com um dos mRNA's recolhidos, da cultura de controle e/ou da cultura condição. A intensidade e cor final da hibridização de cada moléculas depende do nível de concentração do mRNA recolhido de cada uma das culturas. Esta intensidade pode ser verificada por análise da radiação fluorescente verde e vermelho que se encontra associado a cada uma das culturas (DANTAS, 2004).

A Figura 2.4 ilustra a hibridização de um MA com duas amostras de mRNA, cada uma marcada com um corante fluorescente que emite luz em comprimentos de onda diferentes; em geral coloração verde e coloração vermelha. A partir das regras de pareamento de bases de Watson-Crick, o mRNA marcado (em solução) hibridiza com o cDNA correspondente depositado no MA. Neste processo de hibridização, ocorre um pareamento das moléculas complementares, a partir do qual em cada um dos "spots" da lâmina, que referencia um certo gene, tem-se as proporções de mRNA nas duas amostras testadas. Dessa forma a intensidade de fluorescência em cada spot "aceso" está relacionada à abundância do

respectivo mRNA na solução. Ou seja, os spots da lâmina que contêm genes mais expressos na amostra marcada com o corante cy3 devem aparecer na imagem como círculos verdes intensos; caso contenham genes mais expressos na amostra com cy5, aparecerão como círculos vermelhos; se a expressão for a mesma, devem aparecer amarelos. Em seguida, a lâmina é digitalizada (KRUTOVSKII; NEALE, 2001).

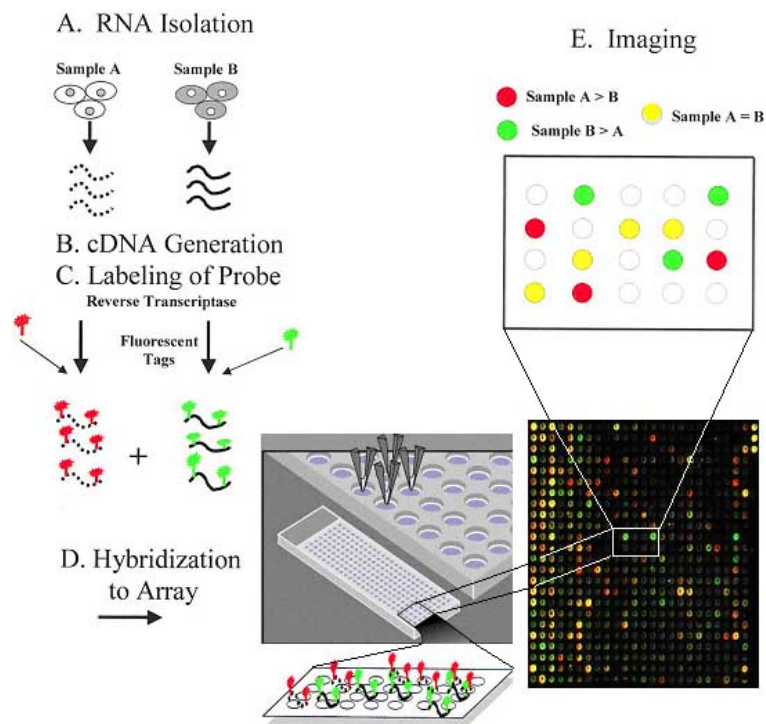


Figura 2.4 - Construção de Microarray.

Expressão gênica é o processo que envolve a conversão da informação contida nos genes em proteína. Sua análise fornece informações importantes sobre as funções de uma célula (SOUTO et al., 2004). Experimentos de microarranjos de DNA estão sendo usados para melhorar a classificação de diagnóstico de doenças, em especial o câncer, seu tratamento e desenvolvimento de novas terapias.

2.2 Técnicas de Agrupamento e Classificação

A grande quantidade de dados armazenada, cria a necessidade de se ter técnicas que permitam a sua automatização e análise de forma inteligente. As técnicas que buscam transformar os dados armazenados em conhecimento desempenham tarefas de classificação e agrupamento dos dados, ou ainda, de descoberta de regras de associação entre eles.

Em um contexto mais geral, as técnicas de agrupamento e classificação objetivam realizar

uma separação ótima entre objetos de uma coleção, permitindo a descoberta de novos padrões, previamente desconhecidos. O resultado da segmentação, independentemente da ferramenta utilizada, pode ser interpretado eficientemente por um especialista na área de origem dos dados sob análise (ANDRADE, 2004). A facilidade de visualização, resultante do agrupamento, favorece a análise.

Dentre os métodos capazes de fazer a classificação, podem-se citar as populares árvores de decisão, as máquinas de suporte de vetores (Support Vector Machines, SVM), os métodos estatísticos, as redes neurais, os algoritmos genéticos e as meta-heurísticas de uma forma geral; estas técnicas vêm sendo amplamente exploradas na literatura (STEINER et al., 2006).

Estudos de técnicas de classificação têm conduzido a modelos matemáticos abstratos, que fornecem a base teórica para o modelo classificador. O problema de classificação é basicamente o de particionar o espaço de característica em regiões. Mas nem sempre isto é possível, e conseqüentemente o problema de classificação passa a ser um problema de decisão estatística. Duda e Hart (1973) estabelecem algumas técnicas de classificação à teoria estatística, dentre elas: problemas de classificação em termos de teoria de decisão; classificação parametrizada e o estudo de caminhos do uso de amostras para determinar o classificador diretamente; técnicas de treinamento não supervisionado e classificação.

Métodos de agrupamento e classificação têm sido usados numa grande variedade de disciplinas científicas e de engenharia que incluem: reconhecimento de padrões, teoria do conhecimento, astrofísica, imagens médicas e processamento de dados, como análise de dados de satélites (BLATT et al., 1997). O objetivo é particionar os dados de acordo com suas características naturais presentes. As técnicas de agrupamento podem ser divididas em duas classes: supervisionada e não supervisionada. No agrupamento supervisionado, vetores são classificados em relação a um vetor de referência conhecido. No agrupamento não supervisionado, não existe vetor de referência a ser relacionado.

Usualmente, as técnicas de agrupamento são aplicadas sobre grandes conjuntos de dados de modo não supervisionado com objetivo de identificar padrões que permitam extrair algum conhecimento. Os grupos são criados de maneira que padrões em um mesmo agrupamento sejam mais similares entre si do que com padrões de um outro agrupamento. Os padrões são determinados de forma a obter-se homogeneidade dentro dos grupos e heterogeneidade entre eles. Após os dados serem agrupados, o usuário precisa visualizar e identificar os agrupamentos (DOMANY, 2003).

Na área médica, técnicas de agrupamento e classificação são aplicadas em dados de MA com o objetivo de reduzir a dimensão do dado e fornecer uma visualização de um experi-

mento complexo. Através da análise destes dados é possível agrupar os genes com base nas semelhanças existentes entre os seus perfis de expressão nas diversas condições analisadas. Esta abordagem tem sido utilizada para diagnosticar e classificar tumores em subgrupos relevantes, assim como para identificar marcadores moleculares que permitam prever a evolução clínica ou a resposta à quimioterapia em novos pacientes diagnosticados com câncer (HAUTANIEMI, 2003).

Além da área médica, outra área de aplicação das técnicas de agrupamento e classificação é a de análise de imagens de satélite. Na área agrícola, por exemplo, os grupos criados pelo agrupamento de uma imagem de satélite de uma plantação, permitem elucidar a distribuição de diferentes atividades agrícolas e medir suas produtividades. A contribuição em botânica está em se extrair amostras de vegetação devidamente agrupadas que facilitam a descrição da ecologia de comunidades de plantas nativas, e indicam áreas para desenvolvimento agrícola ou conservação (ANDRADE, 2004).

Segundo Hanai et al. (2006), existem vários métodos de agrupamento e classificação não supervisionados para análise de dados de MA: agrupamento hierárquico, k -médias (LIKAS et al., 2003), mapas auto-organizáveis, dentre outros. Estes têm sido usados para caracterizar os padrões de expressão. No passo inicial do método de agrupamento hierárquico, cada gene ou amostra constitui um grupo (*cluster*). Os próximos passos consistem em agrupar os pares mais próximos de acordo com sua similaridade. O agrupamento é feito até que todos os genes ou amostras fiquem no mesmo grupo. O mais importante é descobrir quais grupos serão unidos em cada nível hierárquico. Esta escolha é feita baseada na distância entre os grupos. Usualmente, as distâncias métricas utilizadas são a distância Euclidiana e o coeficiente de correlação. É necessário também escolher o método de concatenação (*linkage*), sendo os mais comuns: o *single linkage* (menor distância), o *complete linkage* (maior distância) e *average linkage* (distância média). A saída do agrupamento hierárquico é um diagrama bi-dimensional conhecido como dendograma (AMARATUNGA; CABRERA, 2004; SIMON et al., 2003) conforme ilustrado na Figura 2.5.

Outro importante objetivo no estudo de MA de DNA é a identificação de genes que são diferentemente expressos entre classes pré-definidas. Esta identificação com funções desconhecidas pode levar a um melhor entendimento das funções destes genes. Métodos de comparação de classes são supervisionados pois utilizam a informação de que amostra pertence a qual classe. A teoria estatística utilizada permite estimar a probabilidade de se ver esta diferença tão grande quanto observada. O método mais comumente usado é o t-estatístico (AMARATUNGA; CABRERA, 2004) que mede a razão da variação de expressão do gene entre o entre-classes e o interior-classes. O t-estatístico é então convertido para probabilidade, conhecido como p-value, que representa a probabilidade de se observar em

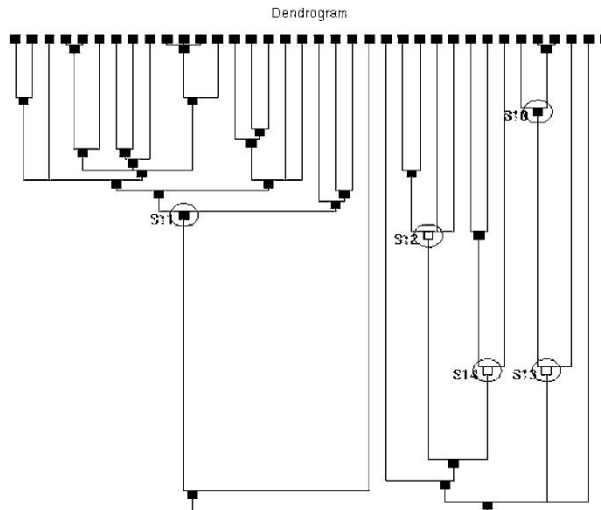


Figura 2.5 - Diagrama bi-dimensional (dendograma).

hipótese nula, um t-estatístico tão grande quanto observado no dado real (AMARATUNGA; CABRERA, 2004).

Em Eisen et al. (1998), utiliza-se uma técnica de agrupamento hierárquico do tipo *average-linkage* na análise de níveis de expressão de genes coletados de amostras de soro com evolução temporal (0 à 24hs). Neste método, as relações entre os genes são representadas por uma árvore onde o comprimento dos galhos reflete o grau de similaridade entre os objetos. Na seqüência, esta árvore é usada para deduzir a história da evolução das seqüências comparadas.

Uma outra técnica usual de agrupamento e classificação é o mapa auto-organizável (*self-organized map* - SOM) ou mapa de Kohonen, que é um algoritmo de agrupamento em redes neurais (AMARATUNGA; CABRERA, 2004). Este permite uma melhor visualização e identificação de agrupamentos similares como também da correlação entre as amostras. O SOM transforma dados de alta dimensão em imagens de uma ou duas dimensões, onde o agrupamento pode ser identificado. É um algoritmo versátil e a visualização dos resultados pode ser feita de diferentes maneiras.

Embora vários métodos de agrupamento organizem tabelas (de medidas de expressão de genes) de forma útil, o resultado desta grande massa de dados fica difícil de ser assimilada. É usual combinar métodos de agrupamento com técnicas de representação gráfica do dado primitivo, apresentando cada ponto com uma cor que, quantitativamente e qualitativamente, reflete as observações do experimento original. O produto final é a representação gráfica do dado complexo (de expressão do gene) através de uma ordenação estatística, permitindo à especialistas (como biólogos) uma assimilação e exploração do dado de uma

maneira natural intuitiva (EISEN et al., 1998).

Em Khan et al. (2001) é desenvolvido um método de classificação de câncer para definir categorias de diagnósticos baseado no perfil de expressão dos genes através de MA, utilizando rede neural artificial. Neste trabalho, também são ranqueado os genes que contribuem para esta classificação, e esses genes definem o menor conjunto que classifica corretamente as amostra dentro de categorias de diagnóstico. O algoritmo inicia fazendo uma filtragem dos dados e em seguida, reduz a dimensionalidade por Análise de Componente Principal (PCA). Com esta camada de entrada, a rede é treinada tendo como camada de saída as categorias de câncer. Os resultados foram satisfatórios e fornecem uma técnica alternativa para detecção de assinaturas de expressão de genes.

A análise não supervisionada de agrupamento hierárquico também foi utilizada em Markretsov et al. (2004) para ordenar dados de MA em pacientes com câncer de mama. Esta análise organiza os casos de acordo com a similaridade ou dissimilaridade dos perfis de expressão dos genes, deixando os similares mais próximos. Utilizou algoritmos de *average linkage* e *complete linkage*. O resultado do agrupamento foi apresentado utilizando-se o programa *Cluster* e a saída pode ser visualizada através da interface gráfica *TreeView*, que exibe graficamente as conexões em duas dimensões. *Cluster* e *TreeView* são programas livres que podem ser obtidos no site <http://rana.lbl.gov/EisenSoftware.htm>.

Em resumo, o SOM é muito eficiente tanto no agrupamento quanto na descoberta de correlação entre amostras. Agrupamento hierárquico é útil para obter uma visão completa do dado obtido (pelo MA). Os resultados do agrupamento devem ser interpretados cuidadosamente na tentativa de identificar um caminho regulatório, uma vez que algoritmos de agrupamento não agrupam genes que são causalmente conectados, e sim com níveis de expressão correlatos (HAUTANIEMI, 2003).

2.3 Pacote BRB-ArrayTools

Nesta dissertação será estudado em detalhes o pacote **BRB-ArrayTools** versão 3.4.0, desenvolvido pelo *Biometric Research Branch of the Division of Cancer Treatment and Diagnosis of the National Cancer Institute*, sob a direção do Dr. Richard Simon. Trata-se de um software livre, voltado para análise de dados de MA de DNA, e está disponível no site <http://linus.nci.nih.gov/brb/download.html>, onde diversas informações como documentação completa e publicações de artigos que incluem resultados feitos pelo software são fornecidas. É compatível com versões do Windows 98/2000/NT/XP ou superiores e projetado para ser usado como um *add-in* do Excell 2000 ou superior.

BRB-ArrayTools contém utilitários para processar dados de expressão em vários expe-

rimentos, visualizá-los, agrupá-los, classificá-los, dentre outras funções. O software foi desenvolvido por estatísticos experientes em análise de dados de MA, mas possui uma interface gráfica que facilita a utilização por biólogos. Mais detalhes podem ser encontrados em [Simon e Lam \(2006\)](#).

- Visão Geral dos Comandos

Será apresentado a seguir uma breve descrição dos comandos do programa **BRB-ArrayTools** ([SIMON; LAM, 2006](#)) que foram utilizados neste trabalho. Após a importação dos dados em formato texto ou Excel, o programa faz uma ligação com bancos de dados genômicos para que as anotações dos genes apareçam nos relatórios gerados pelos comandos requeridos. Estes dados podem ser selecionados na opção de filtragem, que fornece também opções para normalização.

Na importação, o arquivo de entrada possui em cada coluna os dados de amostras onde cada linha representa o nível de expressão gênica. Para a execução dos comandos é preciso inicialmente que as amostras estejam classificadas, e para isso é necessário que haja um arquivo com as classes pré-estabelecidas em cada amostra. O programa conduz as comparações e análise de predição das classes de acordo com o que foi informado neste arquivo.

Antes de se comparar os valores de expressão dos genes entre as amostras, é necessário que se faça uma normalização dos dados, pois existe um provável desequilíbrio de intensidade entre as amostras de RNA. O objetivo da normalização é o de ajustar o valor de expressão do gene em todas as amostras tal que os genes que não são diferentemente expressos tenham valores similares nos arrays.

Com o projeto pronto, os comandos de análise utilizados neste trabalho encontram-se resumidos abaixo:

a) *Clustering*

Cria um *cluster dendrogram* (agrupamento hierárquico) e um gráfico da imagem colorida para os genes selecionados (ou todos). Pode-se também fazer um agrupamento por amostras. A análise de agrupamento pode ser baseada em todos os genes ou apenas em um sub-grupo específico pré determinado pelo conhecimento das classes. Faz-se também uma interface com o software *Cluster 3.0* e *TreeView*, produzido pelo grupo *Stanford*.

b) *Multidimensional Scaling of Samples*

Produz uma visualização tri-dimensional das amostras. Cada amostra é representada por um ponto e a distância entre os pontos é proporcional à

desigualdade dos perfis de expressão representados por estes pontos. Esta distância pode ser calculada utilizando as distâncias métricas Euclidiana ou correlação. O BRB-ArrayTools utiliza os três primeiros componentes principais como eixos da figura gerada. A componente principal é uma combinação linear ortogonal dos genes. O primeiro componente é a combinação linear dos genes com maior variância sobre as amostras de todas as outras combinações lineares. O segundo componente principal é a combinação linear dos genes que é perpendicular ao primeiro com maior variância. De maneira análoga, calcula-se o terceiro.

c) *Class Comparison*

Tem o objetivo de determinar se o perfil de expressão do gene difere entre amostras selecionadas de classes pré definidas e identifica qual gene é diferentemente expresso entre as classes. No estudo do câncer, as classes frequentemente representam distintas categorias de tumores tanto em relação ao estágio do tumor, quanto em relação à presença de mutação genética, ou resposta à terapias. Uma característica de Class Comparison é que as classes são pré definidas independente do perfil de expressão.

d) *Class Prediction*

Class Prediction é similar ao Class Comparison exceto que a ênfase está em se desenvolver um modelo estatístico que pode prever a que classe uma nova amostra pertence, baseada em seu perfil de expressão gênica. Class Prediction é importante em problemas médicos de classificação de diagnóstico, predição de prognóstico e seleção de tratamento (SIMON et al., 2003).

Algumas funções matemáticas são também utilizadas pelo programa na análise estatística dos dados, são elas:

a) *Leave One Out Cross Validation*

Validação cruzada do tipo *Leave one out* é aplicada devido à ausência de um banco de dados para legitimá-los. São incluídos no preditor os genes diferentemente expressos dentro da classe, com nível de significância menor que o limiar.

b) *Nearest Neighbour*

Vizinho mais próximo baseia-se na determinação de qual perfil de expressão na amostra é mais similar ao perfil de expressão apontado como preditor. Utiliza distância Euclidiana como métrica. O cálculo de K-vizinhos mais próximos é similar. Por exemplo, para $k = 3$, o perfil de expressão teste

é comparado a todas as outras amostras e os três perfis mais similares ao teste são detectados. Em seguida, a classe predita é aquela que aparece mais vezes.

c) *Compound Covariate Predictor*

O preditor composto é um método utilizado para predição de classes usando dados de microarray aplicados em genes diferentemente expressos entre duas classes conhecidas. O valor para cada amostra é dado por:

$$c_j = \sum_{i=1}^G t_i x_{ij}, \quad (2.1)$$

onde:

t_i = t-estatístico para os dois grupos em relação ao gene i ,

x_{ij} = medida da amostra j no gene i ,

G = conjunto de genes selecionados,

j = amostra.

Após o cálculo de cada c_j , obtém-se:

$$C_t = \frac{\bar{c}^{(1)} + \bar{c}^{(2)}}{2}, \quad (2.2)$$

onde:

$\bar{c}^{(1)}$ é a média dos valores de c para classe 1,

$\bar{c}^{(2)}$ é a média dos valores de c para classe 2.

Uma nova espécie é predita Classe 1 se o valor de c for próximo de $\bar{c}^{(1)}$, e predita Classe 2 se o valor de c for próximo de $\bar{c}^{(2)}$.

d) *Distância métrica: Correlation*

A correlação centrada entre dois experimentos é definida como:

$$\frac{\sum_{i=1}^N (X_i - X_{avg})(Y_i - Y_{avg})}{\sqrt{\sum_{i=1}^N (X_i - X_{avg})^2 \sum_{i=1}^N (Y_i - Y_{avg})^2}} \quad (2.3)$$

onde o índice i da somatória varia de 1 ao número de genes nos dois experimentos, e X_{avg} e Y_{avg} são as médias dos genes nos experimentos X e Y, respectivamente.

e) *Distância métrica: Euclidiana*

A distância Euclidiana entre dois experimentos é dada por:

$$d = \sqrt{\sum_{i=1}^N (X_i - Y_i)^2}, \quad (2.4)$$

onde o índice i da somatória varia de 1 ao número de genes nos dois experimentos X e Y.

f) *t-teste*

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_p^2 \left(\frac{1}{J_1} + \frac{1}{J_2} \right)}}$$

Onde:

$$s_p^2 = \frac{(J_1 - 1)s_1^2 + (J_2 - 1)s_2^2}{J_1 + J_2 - 2}$$

e

$$s_i^2 = \frac{1}{J_i - 1} \sum_{j=1}^{J_i} (x_{ij} - \bar{x}_i)^2$$

para $i=1,2$

Sendo:

\bar{x}_1 = média dos genes classe 1

\bar{x}_2 = média dos genes classe 2

J_1 = quantidade de amostras classe 1

J_2 = quantidade de amostras classe 2

g) *t-estatístico*

$$t = \frac{[J_1(\bar{x}_1 - \bar{x})^2 + J_2(\bar{x}_2 - \bar{x})^2 + \dots + J_I(\bar{x}_I - \bar{x})^2] / (I - 1)}{s_p^2}$$

Onde:

$$s_p^2 = \frac{1}{J_1 + J_2 + \dots + J_I - I} \sum_{i=1}^I \sum_{j=1}^{J_i} (x_{ij} - \bar{x}_i)^2$$

e

$$\bar{x} = \frac{1}{J_1 + J_2 + \dots + J_I} \sum_{i=1}^I \sum_{j=1}^{J_i} (x_{ij})$$

h) *Nível descritivo*

O p-valor é conhecido na estatística como nível descritivo e está associado ao que se chama de testes de hipóteses. O papel fundamental da hipótese na pesquisa científica é sugerir explicações para os fatos. Uma vez formuladas as hipóteses, estas devem ser comprovadas ou não através do estudo com a ajuda de testes estatísticos. Num teste estatístico são formuladas duas hipóteses chamadas hipótese nula (H_0) e hipótese alternativa (H_1). Hipótese nula é aquela que é colocada à prova, enquanto que hipótese alternativa

é aquela que será considerada como aceitável, caso a hipótese nula seja rejeitada.

Todo teste de hipótese possui erros associados a ele. Um dos mais importantes é chamado “erro do tipo I” que corresponde à rejeição da hipótese nula quando esta for verdadeira. A probabilidade do erro do tipo I chama-se nível de significância e é expressa através da letra grega α . Os níveis de significância usualmente adotados são 5%, 1% e 0,1%. Formalmente, o nível descritivo (p) é definido como o “menor nível de significância α que pode ser assumido para se rejeitar (H_0)”, porém esta interpretação não é simples até mesmo para os estatísticos. Considerando, de maneira muito generalizada, que os pesquisadores ao rejeitarem a hipótese nula costumam dizer que existe “significância estatística” ou que o resultado é “estatisticamente significativo”, poderíamos definir o nível descritivo (p) como a “probabilidade mínima de erro ao concluir que existe significância estatística”.

É importante ressaltar que o nível de significância α é um valor arbitrado previamente pelo pesquisador, enquanto que o nível descritivo (p) é calculado de acordo com os dados obtidos. Fixado α e calculado o “p”, a pergunta é: “será que posso dizer com segurança que o resultado é estatisticamente significativo?”. Para responder à esta questão é necessário avaliar se a probabilidade de erro é “aceitável” ou não, isto é, se o “valor do p” é pequeno o suficiente para concluir que existe “significância estatística” dentro de uma margem de erro tolerável. Mas saber “o que é pequeno ou grande” depende do nível de significância adotado, portanto a decisão do pesquisador sempre estará baseada na comparação entre os dois valores. Se o valor do p for menor que o nível de significância α deve-se concluir que o resultado é significativo, pois o erro está dentro do limite fixado. Por outro lado, se o valor de p for superior à α significa que o menor erro que podemos estar cometendo ainda é maior do que o erro máximo permitido, o que nos levaria a concluir que o resultado é não significativo, pois o risco de uma conclusão errada seria acima do que se deseja assumir (PAES, 1998).

i) p-valor

As amostras J_1 e J_2 são randomicamente permutadas e calcula-se t novamente após cada permutação, denotado por t^* :

$$p - value = \frac{1 + \#\text{permutacoes randomicas onde } |t^*| \geq |t|}{1 + \#\text{permutacoes randomicas}}.$$

2.4 Validação do BRB-ArrayTools em biologia molecular

Antes de realizar a aplicação na área ambiental, o software BRB-ArrayTools foi validado em uma aplicação em biologia molécula. Os dados foram fornecidos pelo Departamento de Bioquímica do Instituto de Química da USP, que utilizou outra ferramenta computacional para analisá-los (REIS et al., 2004).

Nesta aplicação foram analisados dados de expressão gênica de 27 tumores de próstata agrupados em um MA contendo cerca de 4.000 *spots*. Em Reis et al. (2004) foram identificados 56 genes capazes de separar grupos de amostras de tumores de próstata em função de suas características histopatológicas, isto é, o grau de Gleason - GS, uma métrica clínica para o prognóstico de evolução do câncer de próstata. Tumores com GS baixo ($GS \leq 6$) são geralmente menos agressivos, enquanto tumores com GS alto ($GS \geq 9$) são mais agressivos (REIS et al., 2004).

A Figura 2.6 apresenta a tabela com os resultados obtidos nesta dissertação a partir da aplicação de dois métodos distintos de classificação (*compound covariate predictor* e *nearest neighbor predictor* (SIMON et al., 2003; SIMON; LAM, 2006)) para a identificação de assinaturas de expressão gênica capazes de discriminar amostras de tumores com baixo GS (5 e 6) de amostras com alto GS (9 e 10). Foram utilizados dados de expressão de 3576 genes medidos em 11 amostras de tumor de próstata. A significância estatística dos preditores foi estimada a partir de uma validação do tipo *leave-one-out cross-validation*. Aplicando-se um nível de significância de 0.0005 foram identificados 52 genes correlacionados com a distinção GS baixo *vs* GS alto. O nível de significância foi adotado após vários testes, onde o resultado mais satisfatório foi obtido com o valor de 0.0005.

A análise de agrupamento foi feita utilizando-se os resultados da predição de classes da Tabela 2.6. O agrupamento utilizou os métodos *centered correlation* e *average linkage* (SIMON; LAM, 2006; SIMON et al., 2003). Os resultados estão apresentados na Figura 2.7.

A Figura 2.8 apresenta o resultado da análise em escala multidimensional (*multidimensional scaling*) com correlação centrada (SIMON; LAM, 2006; SIMON et al., 2003), onde cada ponto representa um paciente e cada cor uma classe. Nesta análise foram considerados os 52 genes mais significativos e todos os 27 pacientes. Observa-se uma nítida separação entre os pacientes com Gleason alto (GS9 e GS10) e baixo (GS5 e GS6), o que corrobora os resultados anteriores indicando a existência de padrões distintos de expressão gênica no conjunto de dados.

Este resultado é semelhante ao obtido em Reis et al. (2004). Neste trabalho utilizou-se uma

	Array id	Class label	Mean Number of genes in classifier	Compound Covariate Predictor Correct?	1-Nearest Neighbor	3-Nearest Neighbors Correct?
1	88sG5	G5G6	42	YES	YES	YES
2	97sG6	G5G6	42	YES	YES	YES
3	43sG6	G5G6	41	YES	YES	YES
4	52sG6	G5G6	33	YES	YES	YES
5	110sG6	G5G6	26	YES	YES	YES
6	114sG6	G5G6	29	YES	YES	YES
7	4sG9	G9G10	35	YES	YES	YES
8	565sG10	G9G10	45	NO	NO	NO
9	525sG10	G9G10	33	YES	YES	YES
10	549sG10	G9G10	23	YES	YES	YES
11	9sG10	G9G10	39	YES	YES	YES
Mean percent of correct classification:				91	91	91

Figura 2.6 - Tabela representativa do Preditor Multivariável - Para cada array excluído da amostra, é apresentado o número de genes que apresentam significância em relação ao restante e se o preditor o classificou corretamente.

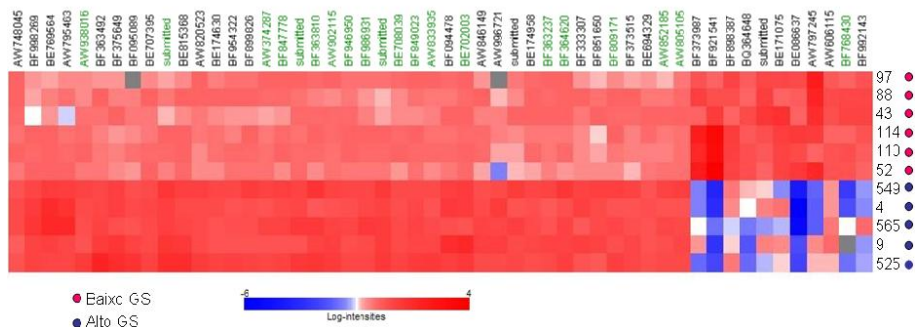


Figura 2.7 - Clusterização dos 52 genes (colunas) em relação ao alto e baixo GS (linhas). Genes em verde sobrepõem-se aos encontrados na análise publicada em Reis et al. (2004).

classificação supervisionada com correlação de Pearson, seguida de *bootstrap resampling* com 10.000 permutações para identificar genes significativamente alterados entre amostras com alto e baixo GS. Esta abordagem possibilitou a identificação de 56 genes com significância relevante ($p \leq 0.001$) (Figura 2.9), posteriormente agrupados hierarquicamente utilizando distância Euclidiana. Dos 56 genes preditores identificados na análise original, 20 deles coincidem com os obtidos na predição de classes obtidas pelo BRB-ArrayTools. Esta sobreposição sugere que estes 20 genes constituam um subconjunto robusto de genes classificadores no câncer de próstata para ser caracterizado em mais detalhe. Os resultados obtidos com o pacote BRB-ArrayTools, não obstante seguem uma outra abordagem, produzindo resultados semelhantes, como é possível perceber comparando-se as Figuras 2.7 e

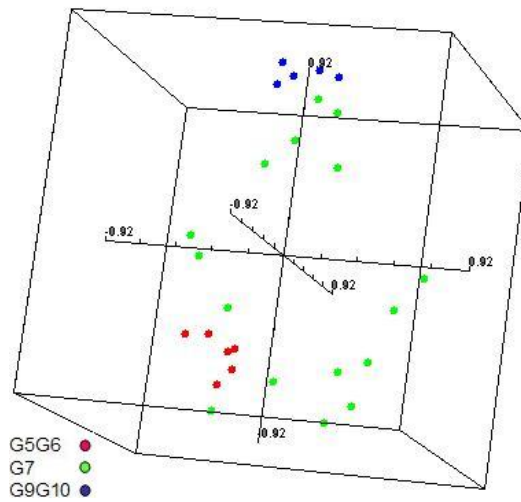


Figura 2.8 - Representação dos pacientes em escala multidimensional.

2.9. Estes resultados foram considerados satisfatórios pelo grupo de pesquisa da IQ/USP que forneceu os dados.

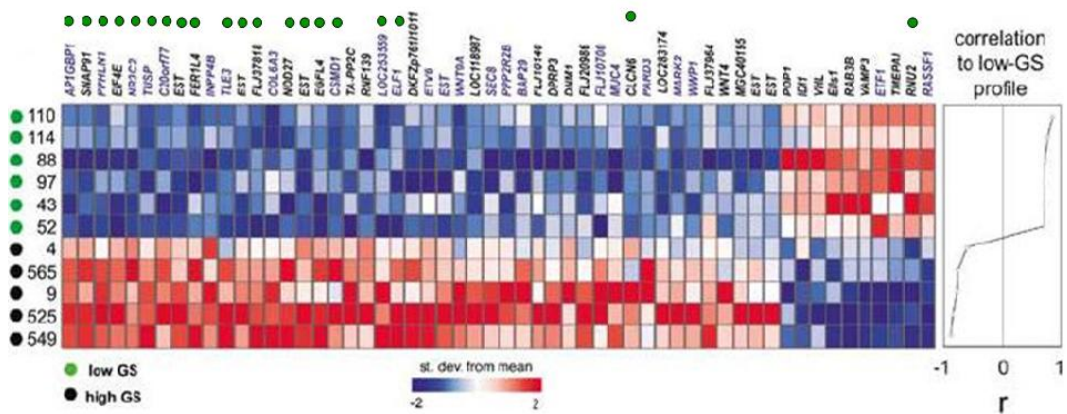


Figura 2.9 - Matriz de expressão dos 56 genes (colunas) em relação ao alto e baixo GS (linhas). Genes rotulados com círculos verde sobrepõem-se aos encontrados na Figura 2.6.

Fonte: (REIS et al., 2004)

3 APLICAÇÃO EM CLIMATOLOGIA

Além de demonstrar a possibilidade de transpor o BRB-ArrayTools para a área ambiental, o objetivo maior deste capítulo é investigar quais foram as variáveis climáticas que causaram a grande seca na Amazônia em 2005 (Figura 3.1). Segundo Marengo et al. (2008), a região da Amazônia passou pelo mais intenso episódio de aridez dos últimos 100 anos. Esta seca afetou severamente a população residente no principal canal do Rio Amazonas como também dos afluentes oeste e sudoeste do Rio Solimões, e do Rio Madeira. A navegação por estes rios foi suspensa pois o volume de água caiu a níveis historicamente baixos. A seca deixou inúmeras pessoas sem alimentação devido à falta de transporte, afetou a agricultura e a geração de hidroeletricidade, prejudicando diretamente ou indiretamente a população que mora em boa parte da bacia Amazônica.



Figura 3.1 - Seca do Amazonas - 2005.

Fonte: www.bacaninha.com.br/pg.php?id=152a=1

A região Amazônica possui precipitação média anual de aproximadamente 2.200 mm por ano, embora haja regiões (na fronteira entre Brasil, Colômbia e Venezuela, e próxima a Foz do Rio Amazonas) em que o total anual pode ultrapassar 3.500 mm por ano. O setor sul, que abrange a região afetada pela seca, tem período de chuvas compreendido entre novembro e março, sendo que o período de seca ocorre entre os meses de maio e setembro. Os meses de abril e outubro são meses de transição entre um regime e outro.

Ao se analisar os dados de precipitação no setor sul da Amazônia, verifica-se que durante a estação chuvosa de 2005, que na realidade estendeu-se de dezembro de 2004 a março de 2005, as chuvas apresentaram valores até 350 mm menores que a média histórica. Isto contribuiu para que os níveis dos rios desta região estivessem com valores bem abaixo da

média no final da estação chuvosa de verão e no início do período de estiagem, que ocorre de maio a setembro. Além disso em 2005, observou-se uma precipitação média menor que a usual durante todos os meses deste ano.

Um dos possíveis fatores responsáveis por esta seca intensa estaria relacionado à temperatura da superfície do mar no Atlântico Tropical Norte, que esteve acima da média nos 12 meses anteriores ao episódio da seca. Assim, o movimento ascendente do ar que normalmente se forma sobre o Atlântico Tropical Norte, ficou mais intenso em 2005, o que fez com que os movimentos descendentes correspondentes sobre o sudoeste da Amazônia fossem mais intensos do que a média, dificultando a formação de nuvens e, portanto, de chuva na região. Adicionalmente, a seca agravou-se devido ao anticiclone do Atlântico Sul que se tornou mais intenso, estendendo-se até o continente e gerando uma região de estabilidade atmosférica que não favoreceu a formação de chuva no sul da Amazônia (MARENGO et al., 2008).

3.1 Dados Analisados

Para se analisar a seca ocorrida em 2005 na Amazônia, foram utilizados dados climatológicos provenientes de diferentes fontes. Todos os dados são mensais e cobrem o período de janeiro de 2000 a dezembro de 2006 (84 meses). Dentro do espírito da analogia com a análise de MA, cada mês de dados representa um “paciente”, e uma coluna na base de dados. Já cada grandeza climatológica (temperatura, velocidade do vento, vazão de rio, etc.), corresponde a um “gene”, e uma linha na base de dados. Estendendo a analogia, pode-se imaginar a seca de 2005 como uma “doença”, e os fatores climáticos causadores do fenômeno, os genes reguladores ainda desconhecidos. Todos os dados utilizados foram tabulados em termos de anomalias (isto é, o valor corrente menos a média do mês para os 7 anos considerados), normalizadas para variarem no intervalo $[-1, 1]$. Este procedimento garante que todos os dados terão, em princípio, o mesmo peso na análise. Descreve-se em detalhe a seguir cada conjunto de dados utilizado.

3.1.1 Dados em Grade

Foram utilizados dois conjuntos de dados em grade fornecidos pelo CPTEC/INPE. O primeiro refere-se a dados mensais de precipitação interpoladas para o Brasil, com resolução de 0.25 graus de latitude e longitude. São 176 pontos em x , a partir de $76^{\circ}W$, e 160 pontos em y , a partir de $34^{\circ}S$. Dentro deste conjunto, foram utilizados dados de um quadrilátero com coordenadas $15^{\circ}S$ a $5^{\circ}S$ e $75^{\circ}W$ a $50^{\circ}W$, que cobre a região afetada pela seca de 2005. Deste subconjunto de 40×60 pontos, foram extraídos valores mensais de precipitação média regional.

O segundo conjunto são dados globais de reanálise (isto é, assimilados e integrados), tomados ao nível de superfície, com resolução espacial de $2.5^\circ \times 2.5^\circ$, com exceção da temperatura da superfície do mar, cuja resolução é de $1^\circ \times 1^\circ$. O método de reanálise foi criado por um complexo sistema de programas, bibliotecas, documentos e banco de dados envolvendo alguns passos que incluem tradução, reformatação, controle de qualidade, análise, previsão, pós-processamento, e arquivamento (KANAMITSU et al., 2002). Da grade global selecionou-se uma subregião com coordenadas 140W a 0W, e 40N a 40S, conforme ilustrado na Figura 3.2. Dentro desta subregião, foram calculados valores médios mensais de cada grandeza, em quadriláteros de 20° de longitude por 20° de latitude. Os dados deste segundo conjunto compreendem os seguintes parâmetros climáticos:



Figura 3.2 - Região analisada: 140W à 0W, 40N à 40S.

- Geopotencial - O geopotencial em algum ponto na atmosfera é definido como o trabalho que deve ser feito contra o campo gravitacional da Terra para elevar uma massa de um quilograma do nível do mar até o ponto considerado (HOLTON, 2004). O geopotencial ao nível do mar é zero;
- Movimento vertical - é a componente vertical da velocidade do vento (CP-TEC/INPE, 2006);
- Pressão atmosférica ao nível médio do mar;
- Radiação de onda longa emergente;
- Temperatura da superfície do mar;
- Temperatura do ar;
- Componente zonal do vento - componente ao redor de círculos latitudinais do vento (WALLACE; HOBBS, 2006);

- Componente meridional do vento - fatia norte-sul através da atmosfera (WALLACE; HOBBS, 2006);
- Umidade relativa.

A mineração dos dados brutos fornecidos pelo CPTEC foi feita com software livre Grads - *Grid Analysis and Display System* (<http://www.iges.org/grads/>), ferramenta interativa utilizada para manipulação e visualização de dados científicos da Terra.

3.1.2 Séries Temporais

a) *Séries Hidrológicas*

Neste trabalho foram utilizadas séries de vazão dos rios Madeira e Amazonas, medidas nos municípios de Humaitá, Manicoré e Óbidos (ver Fig. 3.3), provenientes da rede hidrometeorológica da Agência Nacional de Águas (ANA), operada pelo Serviço Geológico do Brasil.



Figura 3.3 - Localização das três regiões analisadas

Fonte: cprm.gov.br/rehi/amazonialegal/Bacias_de_controle.pdf

b) *Southern Oscillation Index (SOI)*

A série temporal do SOI foi obtida no site <http://www.cru.uea.ac.uk/cru/data/soi.htm>. O SOI é a diferença na pressão atmosférica medida entre as regiões orientais (Tahiti) e ocidentais (Darwin, Austrália) do Oceano Pacífico (Figura 3.4). Quando a pressão é elevada em Darwin, ela é baixa no Tahiti e vice-versa. O El Niño e o seu evento simétrico a La Niña representam as fases extremas da Oscilação do Pacífico Sul. Durante episódios de El Niño, o SOI assume um valor absoluto elevado mas negativo devido à pressão inferior à média no Tahiti e superior à média em Darwin (Figura 3.5). Durante episódios de La Niña, o SOI assume um valor positivo elevado devido aos valores da pressão do ar acima da média no Tahiti e abaixo da média em Darwin. Episódios de El Niño ocorrem a cada 2 a 7 anos, aproximadamente. (SCHLANGER, 2006).

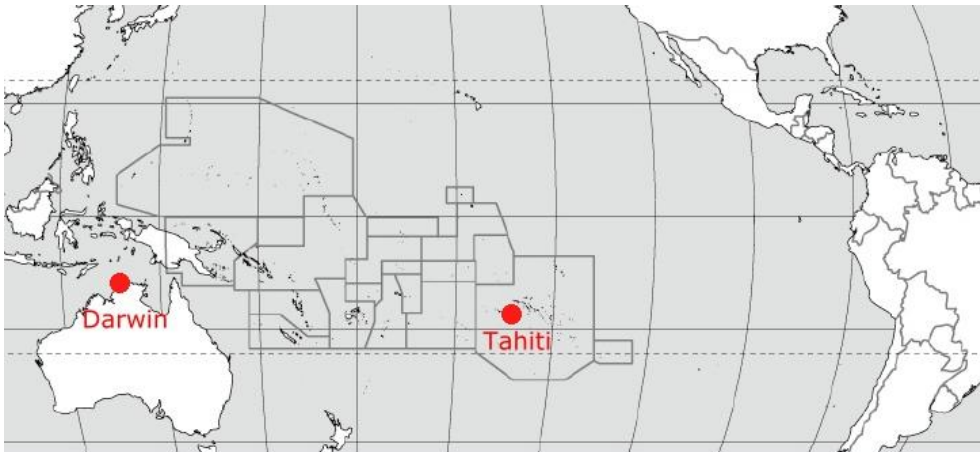
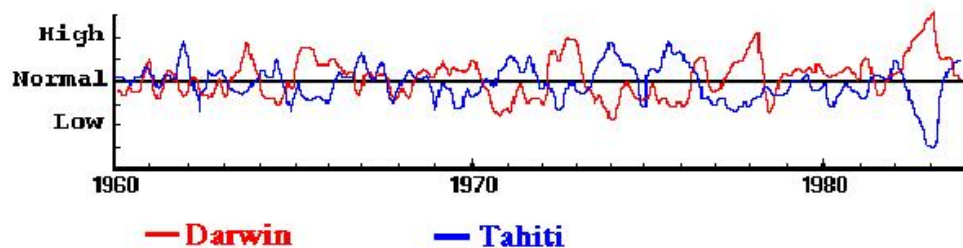


Figura 3.4 - Mapa do Pacífico Sul, evidenciando Darwin na Austrália e Tahiti, uma das ilhas do Pacífico.

Fonte: (SCHLANGER, 2006)



Pressão ao nível do mar 1960-1984
fonte: U. S. Army Topographic Engineering Center

Figura 3.5 - Série temporal SOI.

c) *North Atlantic Oscillation (NAO)*

A série temporal do NAO foi obtida no site <http://www.ldeo.columbia.edu/NAO/main.html>. Este índice controla a variabilidade climática de uma região do Atlântico Norte que estende-se de uma parte central da América do Norte até a Europa e grande quantidade do Norte da Ásia conforme ilustrado na Figura 3.6. O índice é baseado na diferença da pressão de superfície entre altas subtropicais (arquipélago de Azores localizado no Oceano Atlântico) e baixas subpolares (Iceland, ilha localizada no Norte do Oceano Atlântico). A NAO está associada a fenômenos de seca e inundação no Hemisfério Norte, mas seu impacto na Amazônia ainda é pouco estudado.

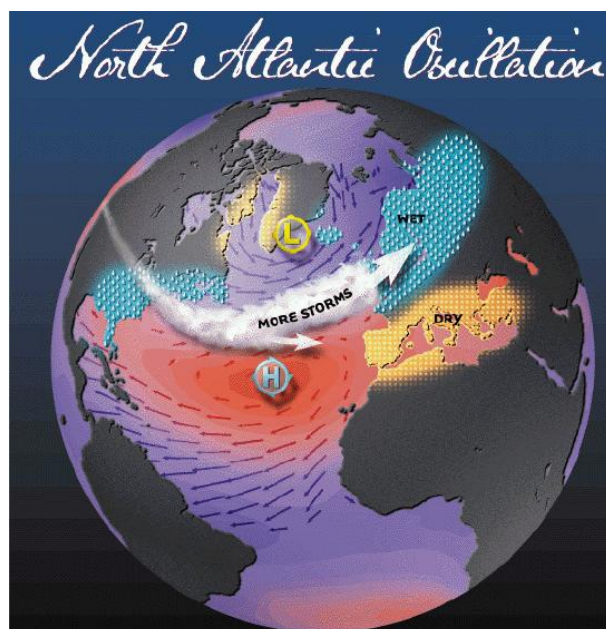


Figura 3.6 - *North Atlantic Oscillation*

Fonte: (OSCILLATION, 2005)

d) *Pacific Decadal Oscillation (PDO)*

A série temporal do PDO obtida no site <http://jisao.washington.edu/pdo/>, é definida como padrão de vida longa do El Niño na variação climática do Pacífico. Seu índice é definido como o componente principal dominante da variabilidade mensal de temperatura da superfície do mar do Pacífico Norte. Segundo Mantua (1999), fases extremas de PDO têm sido classificadas como quentes ou frias, de acordo com as anomalias de temperaturas oceânicas no Oceano Pacífico nordeste e tropical. A Figura 3.7 ilustra as fases do PDO dos padrões de anomalia durante

o frio e o calor na região do Pacífico. As cores indicam a temperatura da superfície do mar. Quando a temperatura da superfície do mar é irregularmente fria no interior do Pacífico Norte e quente em direção à costa do Pacífico (warm phase), e quando a pressão ao nível do mar (SLP) é abaixo da média sobre o Pacífico Norte, os respectivos índices tem valores positivos. Já quando as anomalias climáticas comportam-se de forma inversa (cool phase), os índices tem valores negativos.

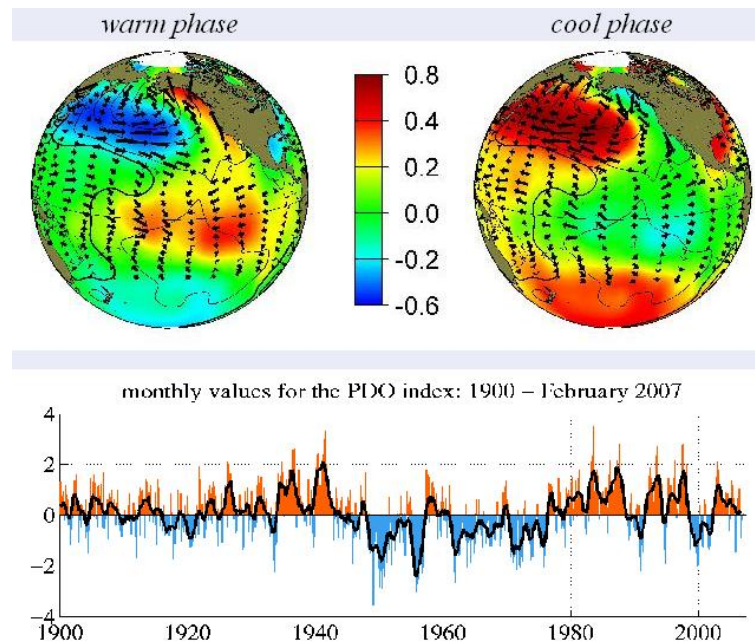


Figura 3.7 - PDO - Temperatura da superfície do mar na época de inverno.

Fonte: (OSCILLATION, 2000)

e) *Sea Surface Temperature (SST)*

Os índices SST correspondem à temperatura média da superfície do mar medidas nas seguintes regiões: Atlântico Norte (5-20N, 60-30W), Atlântico Sul (0-20S, 30W-10E) e Faixa Tropical Global (10S-10N). Estes dados foram obtidos no site <http://www.cpc.ncep.noaa.gov/data/indices/>.

Um total de 269 parâmetros foram armazenados em uma planilha EXCEL, software de base do BRB-ArrayTools. A lista com todas as variáveis analisadas está descrita no Apêndice A.

3.2 Resultados

Após completada a planilha com os dados normalizados, a importação dos dados climatológicos no programa BRB-ArrayTools foi feita de maneira semelhante aos dados gênicos, diferindo apenas na conexão com o *Gene Bank*, que é opcional, logo não foi executada. Como o objetivo desta aplicação é analisar o período de seca (mais precisamente, de vazões decrescentes), foram considerados apenas os meses de julho de 2000 a novembro de 2006. Como dito anteriormente, a planilha utilizada nas análises foi gerada de maneira que cada linha represente uma variável climática (gene), e cada coluna uma média mensal (paciente).

A extração do conhecimento do banco de dados é realizada através de “projetos”. Um projeto necessita a definição das “classes” que nortearão as operações de classificação e agrupamento. Um exemplo típico é a divisão em duas classes: acima ou abaixo da mediana da propriedade. Cada projeto busca responder a uma pergunta padronizada do tipo “Quais são as variáveis climáticas, dentre as consideradas no banco de dados, responsáveis pela propriedade X (por exemplo, vazão média mensal do rio Amazonas em Óbidos) pertencer a uma determinada classe?”. Naturalmente, perguntas diferentes levam a resultados diferentes. Nesta aplicação, foram considerados diferentes propriedades X: i) a média aritmética das vazões do rio Madeira em Humaitá, Manicoré e do Amazonas em Óbidos (doravante chamada de “índice integrado”); e ii) a vazão do rio Amazonas em Óbidos (que por estar a jusante, contém “informação” de vários afluentes inclusive o Madeira). Os principais resultados obtidos estão apresentados a seguir.

3.2.1 Casos Estudados

a) Análise do Índice Integrado (2 classes)

Como dito acima, este parâmetro compreende a média aritmética das vazões dos Rios Madeira e Amazonas próximos a Humaitá, Manicoré e Óbidos. Após a manipulação dos dados, foram consideradas duas classes, acima ou abaixo da mediana. Após vários testes, o resultado mais significativo foi adquirido utilizando-se um p-valor de 0.02 (isto é, uma probabilidade de 2% de ocorrerem falsos positivos) na opção *Class Comparison* do pacote BRB. Ao atribuir-se um p-valor maior, obtém-se muitos parâmetros na classificação e a chance de ocorrerem falsos positivos aumenta muito. Já para p-valor menor, apesar de diminuir a probabilidade de falsos positivos, aparecem poucas variáveis climatológicas na análise. Por isso é necessário fazer vários testes até que possa se obter um p-valor satisfatório. Os resultados obtidos incluem as 13 variáveis climatológicas (além do próprio índice integrado), selecionadas pela ferramenta computacional como sendo capazes de

explicar a seca de 2005. O resultado da operação de agrupamento pode ser visualizado na Figura 3.8, onde a ordem dos parâmetros segue o agrupamento obtido pela ferramenta computacional. Dentre as variáveis selecionadas, além de parâmetros como a temperatura da superfície do mar numa determinada coordenada, estão naturalmente incluídas as séries de vazão em Óbidos e em Humaitá (mas não em Manicoré). A cor de cada quadrado na matriz de agrupamento indica se uma determinada variável climática está abaixo (tons de azul) ou acima (tons de vermelho) de sua respectiva mediana. A cor da legenda das colunas indica se a vazão integrada naquele mês pertence à classe 1 (azul, abaixo da mediana) ou à classe 2 (vermelho, acima da mediana). A Figura 3.9 mostra disposição geográfica das variáveis relevantes para explicar a seca. Nesta figura, a cor das setas indica novamente a posição do valor da variável em relação a sua própria mediana, tendo por base o mês de setembro de 2005, o auge da seca.

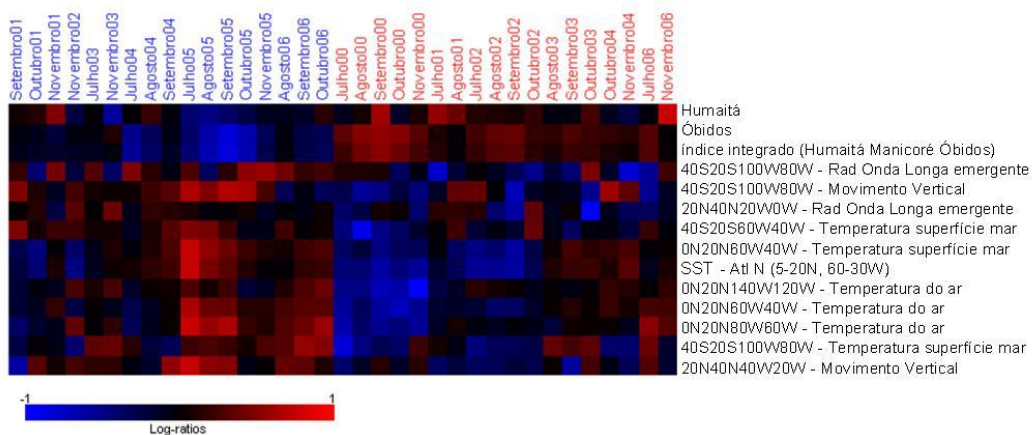


Figura 3.8 - Agrupamento do índice integrado.

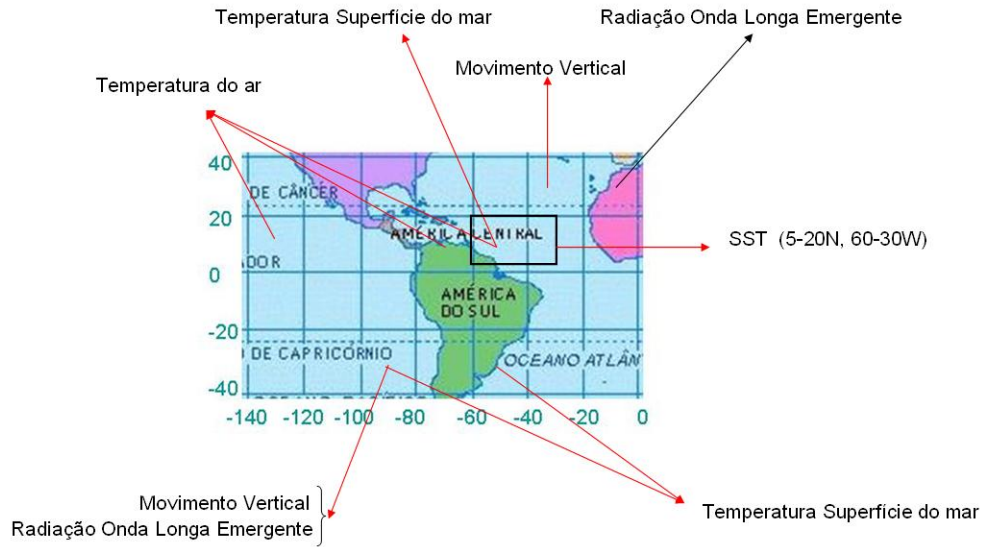


Figura 3.9 - Localização geográfica dos parâmetros mais relevantes na análise do índice integrado.

b) Análise da Vazão em Óbidos (2 classes)

Por motivos análogos ao caso anterior, utilizou-se um p-valor de 0.02. Os resultados obtidos estão apresentados nas Figuras 3.10 e 3.11. O processo de análise encontrou 13 variáveis relevantes, muitas delas idênticas às encontradas no projeto anterior, como seria de se esperar. Desta vez, nem o índice integrado nem as vazões em Humaitá e Manicoré foram selecionados pelo algoritmo, para este p-valor. No entanto, observa-se que como no caso anterior, as variáveis selecionadas permitem diferenciar bem as épocas secas das chuvosas.

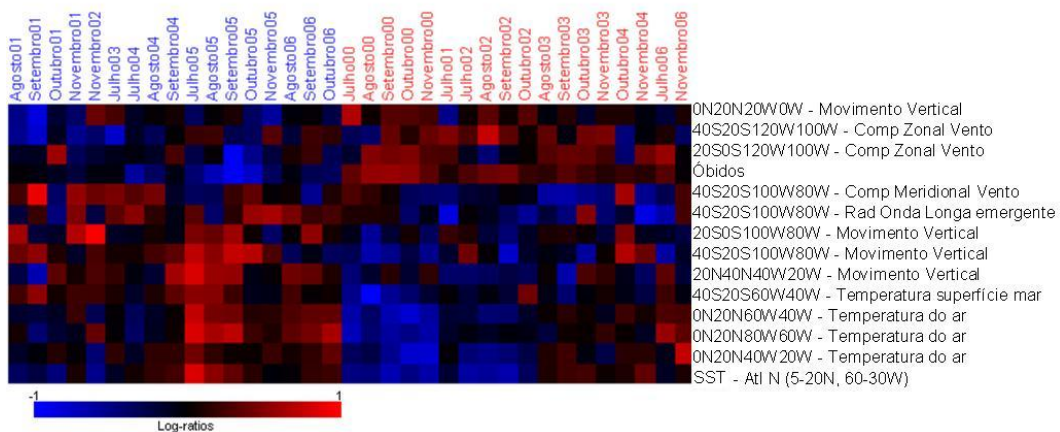


Figura 3.10 - Agrupamento do índice de vazão do Rio Amazonas em Óbidos.

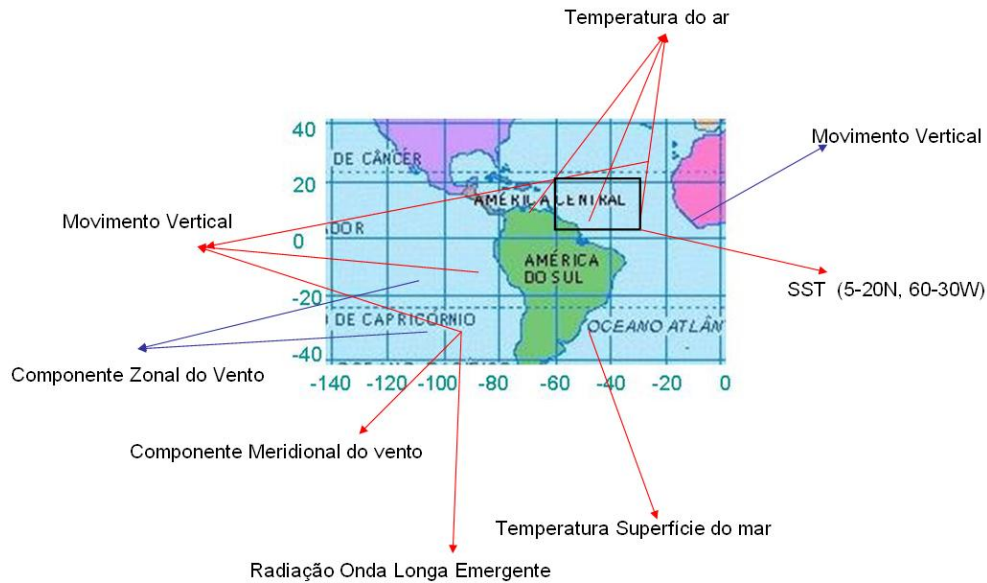


Figura 3.11 - Localização geográfica dos parâmetros mais relevantes na análise do Rio Amazonas em Óbidos.

c) Análise da Vazão em Óbidos (3 classes, caso 1)

Neste projeto foram consideradas três classes, definidas da seguinte maneira: classe 1 (tons de azul) = valores em $[-1; -0.1]$, classe 2 (preto) = valores em $[-0.1; 0.2]$, e classe 3 (tons de vermelho) = valores em $[0.2; 1]$. A análise foi feita utilizando-se p-valor de 0.02. Os resultados estão apresentados nas Figuras 3.12 e 3.13. Desta vez foram selecionadas 21 variáveis mas a qualidade do agrupamento é apenas razoável, apesar de permitir ainda a diferenciação entre época seca e chuvosa. Convém ressaltar que pela primeira vez o algoritmo selecionou o nível médio de precipitação na região afetada pela seca como uma das variáveis explicativas do fenômeno. Estes resultados sugerem que a discriminação em 3 classes requer muito mais informação que os casos anteriores. Outra conclusão é que nem sempre variáveis aparentemente relacionadas com um fenômeno ou processo (por exemplo, a precipitação média na Amazônia) permitem uma boa classificação e agrupamento da informação contida no banco de dados. Outros fatores como o p-valor, a definição das classes e, provavelmente, o nível de ruído dos dados são também determinantes nesta escolha.

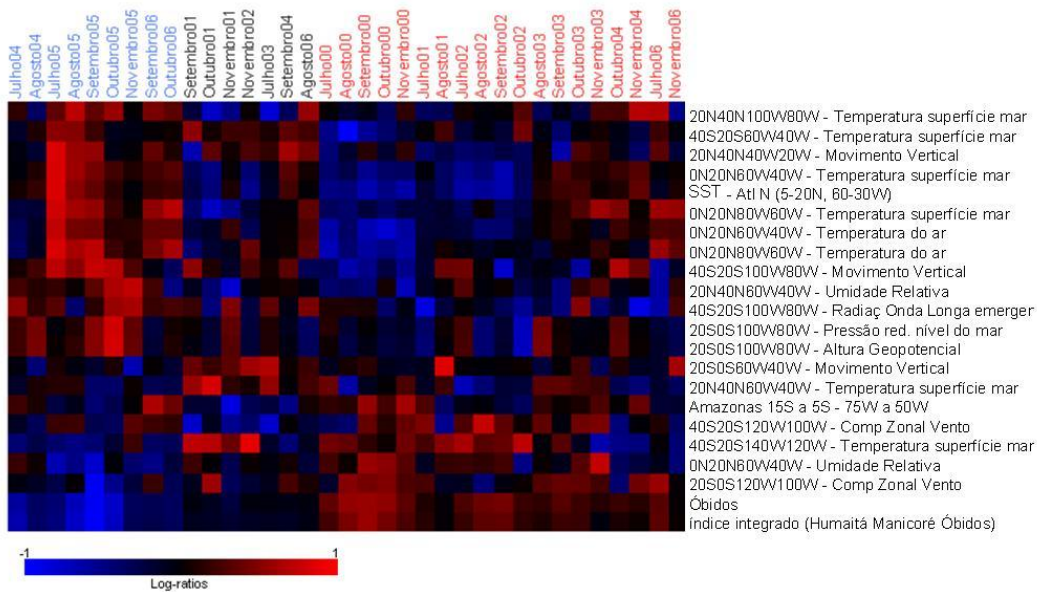


Figura 3.12 - Agrupamento do índice de vazão do Rio Amazonas em Óbidos utilizando-se 3 classes: [-1; -0.1], [-0.1; 0.2] e [0.2; 1].

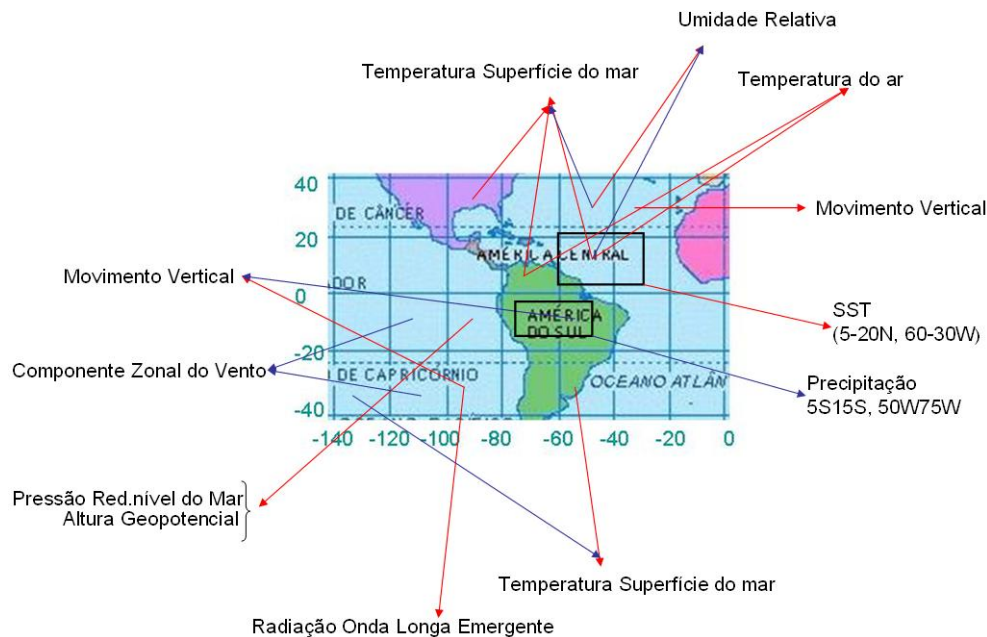


Figura 3.13 - Localização geográfica dos parâmetros mais relevantes na análise do Rio Amazonas em Óbidos (3 classes).

d) Análise de Vazão em Óbidos (3 classes, caso 2)

Aqui, também, foram consideradas três classes mas definidas de maneira diferente: classe 1 (tons de azul) = valores entre -1 e (mediana-0.03), classe 2

(preto)= valores entre (mediana-0.03) e (mediana+0.03), e classe 3 (tons de vermelho)= valores entre (mediana+0.03) e 1. A análise foi feita utilizando-se um p-valor de 0.01, metade do valor utilizado anteriormente. Os resultados estão apresentados nas Figuras 3.14 e 3.15. Neste caso observa-se o agrupamento, baseado em 24 parâmetros, tornou-se mais nítido que o anterior, apresentando uma forte diferenciação entre os meses de seca e chuva. A conclusão imediata é que definição das classes é uma etapa crítica na execução das tarefas de classificação e agrupamento, dependendo do conhecimento a priori da natureza do problema analisado e de uma certa dose de bom senso. Note-se também que mesmo com a utilização de um p-valor mais restrito, o número de variáveis selecionadas aumentou em relação ao caso anterior.

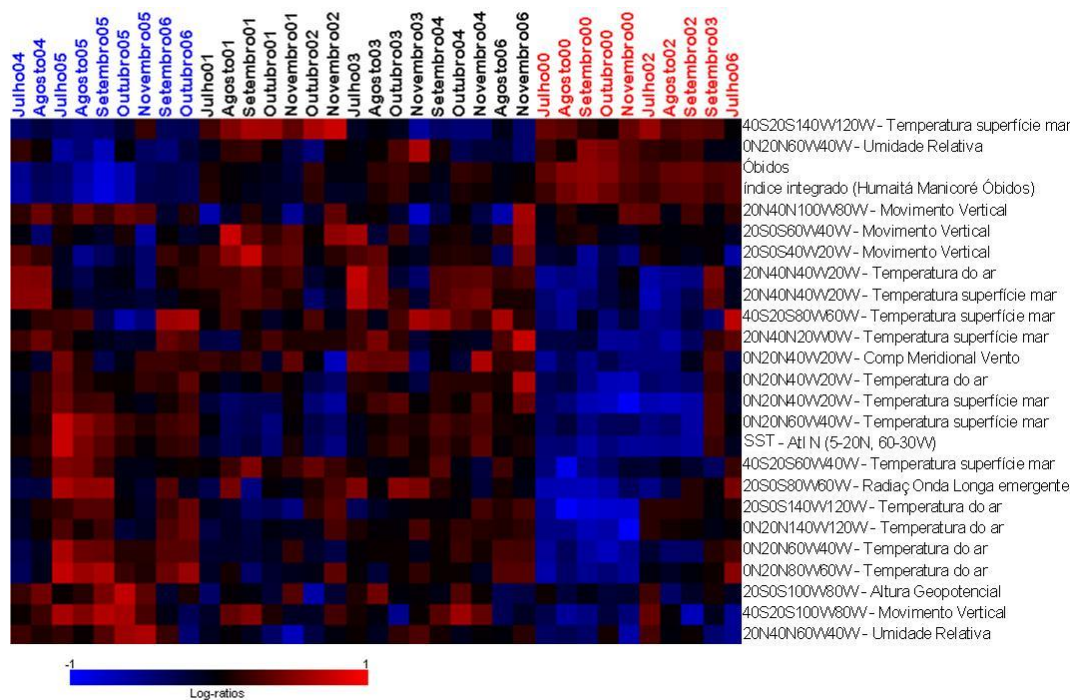


Figura 3.14 - Agrupamento do índice de vazão do Rio Amazonas em Óbidos utilizando-se 3 classes: $[-1; mediana - 0.03]$, $[mediana - 0.031; mediana + 0.03]$ e $[mediana + 0.03; 1]$.

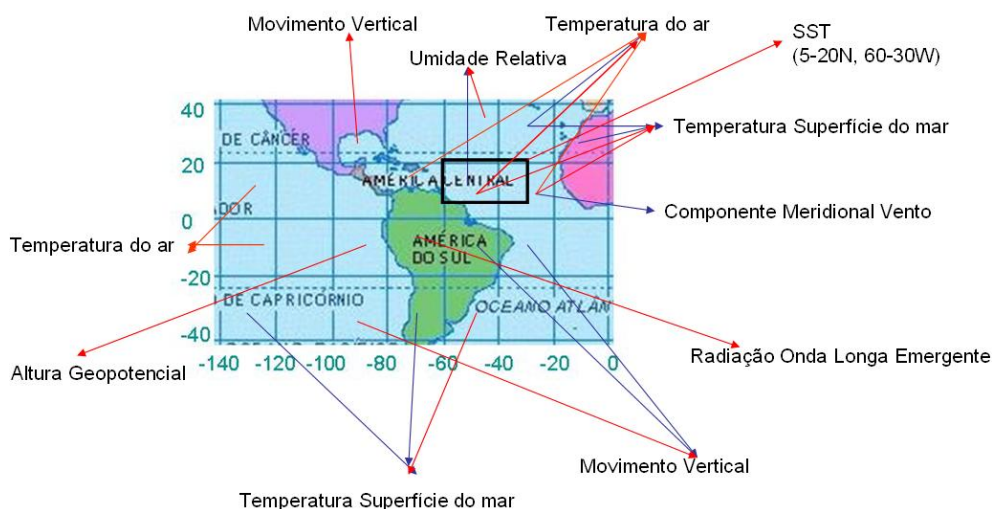


Figura 3.15 - Localização geográfica dos parâmetros mais relevantes na análise do Rio Amazonas em Óbidos (3 classes - caso2).

e) Análise comparativa Ano Seco x Ano Chuvoso

Nesta análise foram selecionadas da Figuras 3.14 as colunas referentes aos meses de julho a novembro, dos anos 2000 (estiagem moderada) e 2005 (estiagem extrema). Estes dois períodos estão destacados na Figura 3.16, que apresenta a série de vazão mensal do rio Amazonas em Óbidos, no período de janeiro de 2000 à dezembro de 2006. O objetivo aqui é demonstrar que o resultado anterior permite diferenciar claramente situações extremas de chuva e seca na região analisada. Os resultados estão apresentados nas Figuras 3.17 e 3.18.

3.2.2 Análise dos Resultados

Percebe-se nas análises anteriores que algumas variáveis, como a temperatura da superfície do mar na região do Atlântico Norte Tropical, aparecem de maneira repetida nos resultados. Para quantificar esta constatação, construiu-se um histograma que mostra o número de vezes que uma variável climática foi selecionada pelo algoritmo, de um máximo possível de 4. Os resultados estão apresentados nas Figuras 3.19 e 3.20.

Nossos resultados mostram que o índice SST no Atlântico Norte compreendido entre as coordenadas 5-20N, 60-30W aparece em todos os casos analisados. Outros parâmetros como a temperatura da superfície do mar na costa sul brasileira parecem ter um papel relevante e merecem ser analisadas em detalhe pelos especialistas. Por outro lado, o índice SOI, relacionado ao fenômeno El Niño/La Niña não foi selecionado uma única vez.

Outra constatação, já antecipada, é que a escolha das classes é um item fundamental

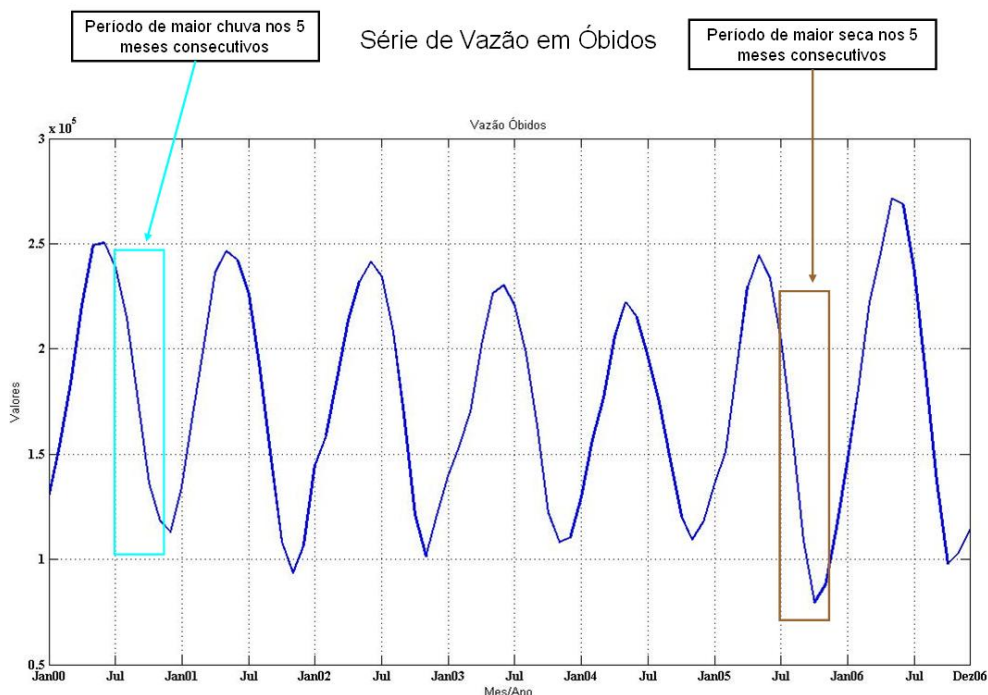


Figura 3.16 - Série de vazão do Rio Amazonas em Óbidos nos períodos compreendidos entre Jan/2000 e Dez/2006.

na análise. No caso presente, os resultados mais significativos foram obtidos na análise de vazão em Óbidos (3 classes, caso 2), e permitiram distinguir claramente dois anos extremos em termos de precipitação (veja Figura 3.15). Desta análise, parâmetros como a temperatura da superfície do mar na região do Atlântico Norte entre as coordenadas 20W e 60W e a temperatura do ar na mesma região aparecem como variáveis chave para explicar a seca de 2005. Neste período, observa-se que a baixa umidade relativa na região do Atlântico Norte, com valores próximos do mínimo, e o fraco movimento vertical sobre a região do Amazonas são também relevantes.

Para validar os resultados aqui apresentados é preciso recorrer a literatura especializada. [Marengo et al. \(2008\)](#) atribuem o aumento da SST no Atlântico Norte tropical como o principal responsável pela seca de 2005, na ausência do fenômeno El Niño. Os autores apontam, também, o fraco movimento vertical sobre a região do Amazonas como um dos possíveis causadores da seca. Estes resultados corroboram as nossas análises, sobretudo no que se refere ao papel fundamental da SST na seca de 2005. Outra análise, publicada em [Trenberth e Shea \(\)](#), também destaca o papel da SST na região do Atlântico Norte (10°N - 20°N) na seca da Amazônia. Estes dois trabalhos corroboram alguns dos resultados obtidos neste capítulo e reforçam a impressão de que a transposição para a Climatologia de técnicas de análise de grandes quantidades de dados biológicos, como a ferramenta

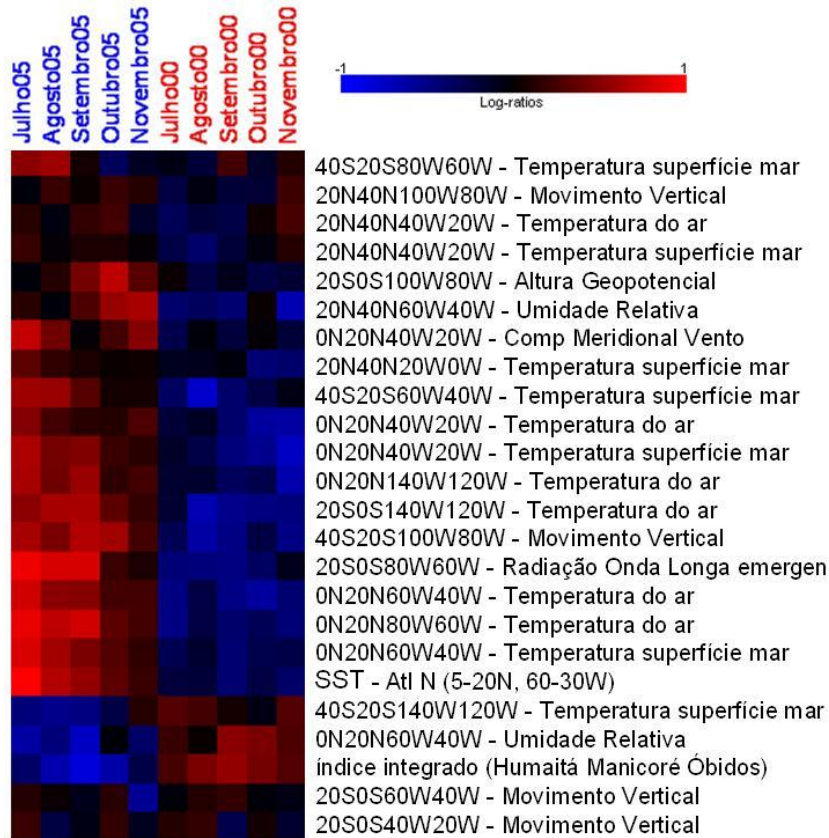


Figura 3.17 - Agrupamento do índice de vazão do Rio Amazonas em Óbidos utilizando-se 3 classes - períodos de Jun à Nov em 2000 e 2005.

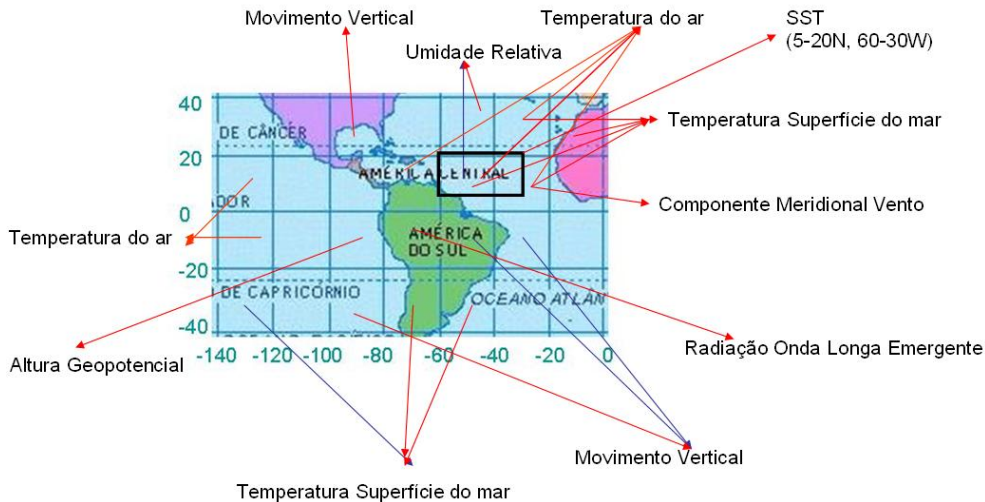


Figura 3.18 - Localização geográfica dos parâmetros mais relevantes encontrados na Figura .3.17

BRB-ArrayTools, é plenamente viável. Cabe ressaltar que outros resultados obtidos neste trabalho, como variáveis climatológicas não citadas nos trabalhos de [Marengo et al. \(2008\)](#), [Trenberth e Shea \(\)](#), devem também ser analisados com mais detalhes por profissionais

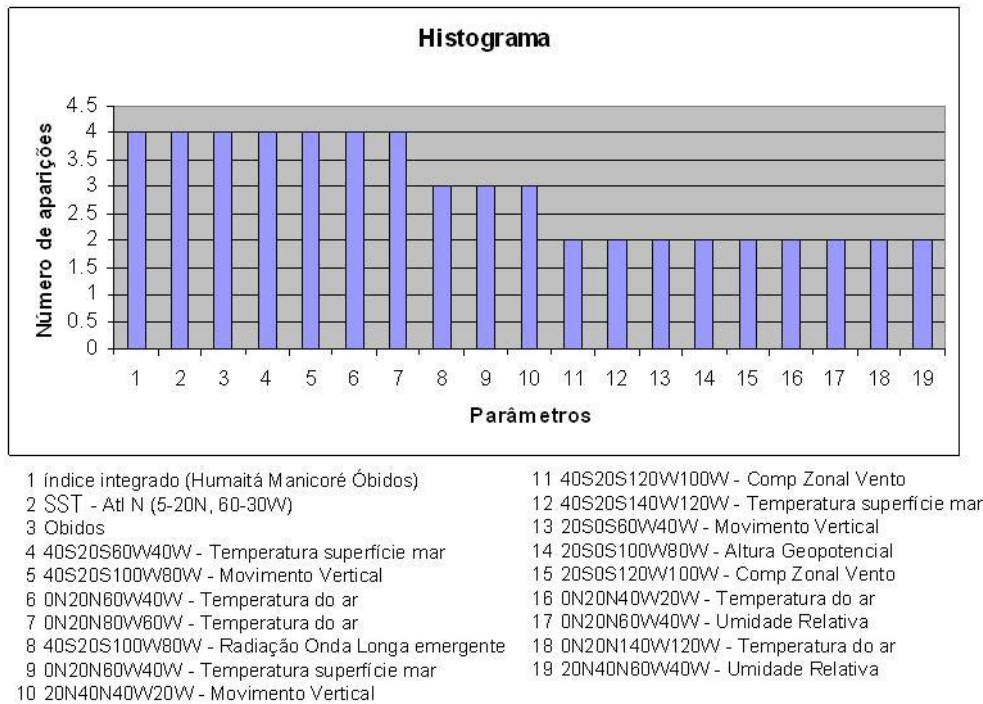


Figura 3.19 - Histograma das parâmetros mais relevantes encontrados nas abordagens estudadas.

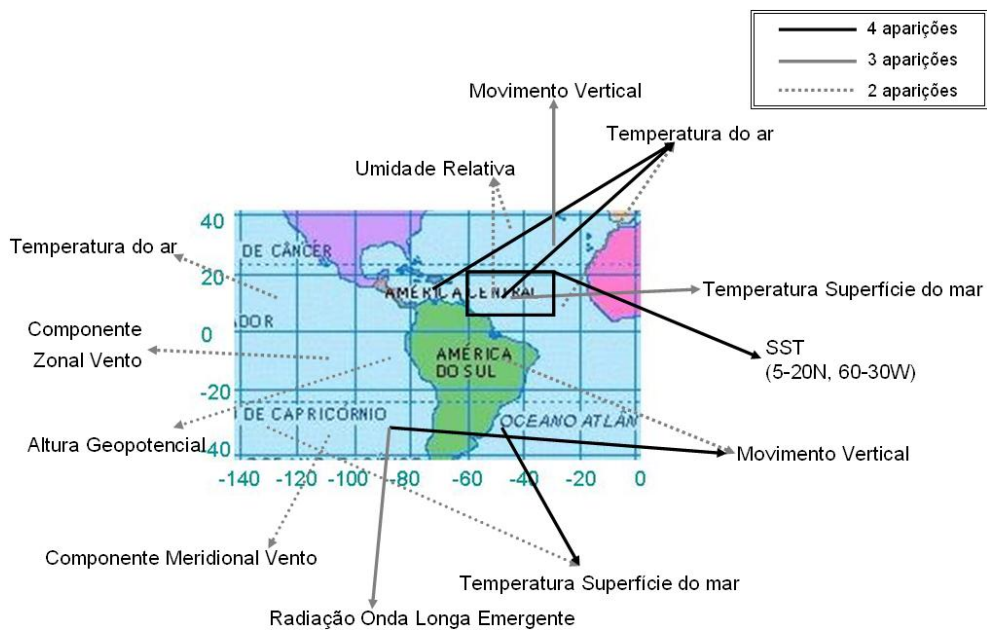


Figura 3.20 - Localização geográfica dos parâmetros mais relevantes encontrados nas abordagens estudadas.

da área.

4 APLICAÇÃO EM LIMNOLOGIA

A Limnologia é o estudo das reações funcionais e produtividade das comunidades bióticas de lagos, rios, reservatórios e região costeira em relação aos parâmetros físicos, químicos e bióticos ambientais. Os estudos dos ecossistemas aquáticos remontam a Grécia Antiga, sendo inicialmente listagens de organismos. Apenas no final do século XIX passaram a ser sistematicamente estudados com um estruturado ferramental teórico e metodológico.

A limnologia apresenta um ilimitado campo de atuação na pesquisa básica (estrutura e função dos ecossistemas aquáticos) e aplicada (controle da qualidade e quantidade da água, usos múltiplos de lagos e reservatórios, etc). Também tem um importante papel no monitoramento e recuperação dos corpos de água. Na atualidade, uma das atuações mais significativas do limnólogo diz respeito ao controle da eutrofização (processo decorrente do excesso de nutrientes básicos adicionados ao corpo de água). Hoje pode ser considerada uma das mais importantes áreas da pesquisa em ecologia no Brasil (ESTEVES, 1988).

Como última aplicação desta dissertação, analisou-se o banco de dados do Projeto Balanço de Carbono Furnas. Além de demonstrar a viabilidade do emprego do pacote BRB-ArrayTools em um problema de limnologia, o objetivo científico desta aplicação foi identificar quais são os fatores relevantes que controlam a emissão de gases de efeito estufa (GEE) em reservatórios.

As mudanças climáticas transformaram-se em um dos temas de maior relevância mundial. Pesquisas recentes confirmam que o aquecimento global nos últimos 50 anos é consequência do aumento das concentrações de GEE (Figura 4.1), originado principalmente da queima de combustíveis fósseis. Como resultado, é prevista a ocorrência de eventos climáticos extremos e são esperados impactos na circulação e no volume (elevação do nível) dos oceanos, nos regimes pluviométricos, na agricultura e na estrutura e produtividade dos ecossistemas, com perda de biodiversidade e alteração nos ciclos do carbono e nutrientes (INPE et al., 2006).

Grandes represas apresentam um papel central no desenvolvimento da civilização. Em várias nações, eles são responsáveis pela maior fonte de energia. A partir da última década, a comunidade científica tem questionado se os reservatórios destinados à geração hidrelétrica contribuem ou não para o aumento do efeito estufa. Pesquisas recentes sobre a produção e emissão de GEE em reservatórios indicam que estes sistemas, sob certas condições, podem apresentar emissões consideráveis, particularmente de metano, gás carbônico e óxido nitroso (LIMA et al., 2008; INPE et al., 2006). Os lagos das usinas recebem e produzem dois tipos de carbono: o inorgânico e o orgânico. O primeiro encontrado em maior quantidade,

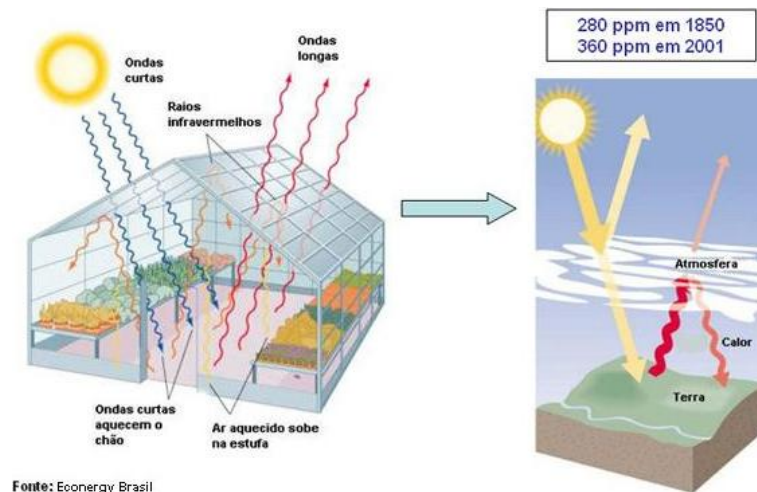


Figura 4.1 - O Efeito Estufa

tem origem principalmente nas trocas gasosas entre a água e a atmosfera. O carbono orgânico tem diversas origens: além de sua formação dentro do próprio corpo d'água por meio da fotossíntese e da cadeia alimentar, o mesmo pode ser canalizado pela bacia de drenagem, através da vegetação morta e da adubação de plantações, levado pela água das chuvas para os rios que deságuam nos reservatórios. Outras entradas de carbono orgânico acontecem pelo despejo de esgotos domésticos e industriais e, também, pela vegetação que foi submersa com o enchimento dos lagos (BAMBACE et al., 2007).

Muitos trabalhos vem sendo desenvolvidos no contexto de análise de emissão de metano em represas hidrelétricas. Em Lima (2005), por exemplo, o autor investigou a emissão de metano em forma de bolhas nas represas brasileiras de Tucuruí e Samuel, e verificou que a profundidade da represa influencia na oxidação do CH_4 . Em Ramos et al. (2006) foram investigadas as características estatísticas de fluxos de ebulição de metano em outros reservatórios brasileiros de Manso e Corumbá.

4.1 Projeto Carbono Furnas

O Projeto Carbono Furnas tem por objetivo determinar as emissões de GEE dos reservatórios de Furnas Centrais Elétricas S.A. Envolve o Departamento de Meio Ambiente (DMA.T) de Furnas Centrais Elétricas S.A.; o Instituto Alberto Luiz Coimbra de Pós-Graduação e Pesquisas de Engenharia - COPPE/UFRJ, o Instituto Nacional de Pesquisas Espaciais - INPE, a Universidade Federal de Juiz de Fora - UFJF e o Instituto Internacional de Ecologia - IIE. O projeto é composto por quatro sub-projetos desenvolvidos em paralelo:

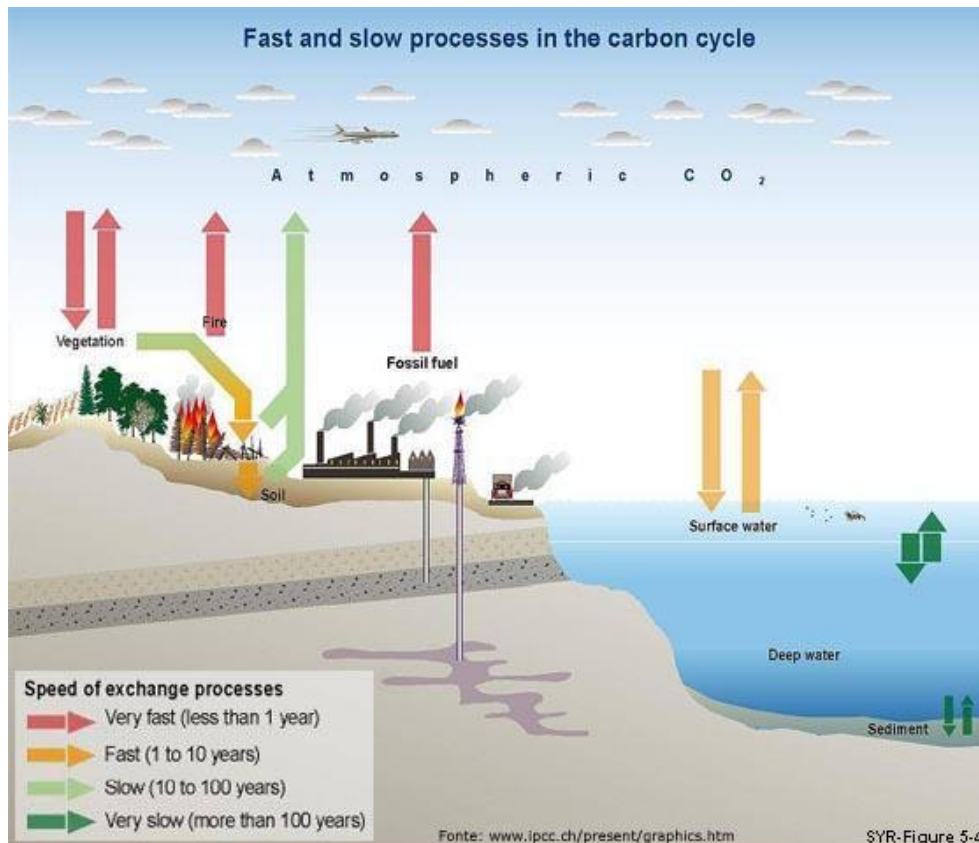


Figura 4.2 - Vista esquemática dos processos lentos e rápidos do ciclo de carbono. Aqui é mostrado como ocorre a velocidade de trocas de carbono entre reservatórios, afetando todo o ciclo (INPE et al., 2006).

a) Aquisição de dados micrometeorológicos e limnológicos em tempo real

A coleta destes dados é feita pelo Sistema Integrado de Monitoração Ambiental (SIMA). Trata-se de uma instrumentação montada sobre uma bóia, desenhado para a coleta de dados e a monitoração em tempo real de sistemas hidrológicos. Esse sistema foi desenvolvido a partir de uma parceria entre a Universidade do Vale do Paraíba (UNIVAP) e o INPE e as medições são feitas pelo INPE.

b) Estimativa de Fluxos de CO_2 , CH_4 e N_2O na interface água-atmosfera e coluna d'água

É um programa de coletas de amostras feito pela UFRJ e o INPE de gás emitido na interface água-atmosfera, tanto sob a forma de bolhas como por difusão.

c) Ciclo de Carbono na coluna d'água

Nos ambientes aquáticos, a maior parte do carbono está presente nas formas inorgânica, e orgânica dissolvidas e a avaliação desses dois processos biológicos, associada a parâmetros físicos e químicos, é fundamental para a compreensão e construção do modelo do ciclo do carbono e suas implicações nas emissões observadas. Estas medições são feitas pela UFJF.

- d) Estimativa de Fluxos de CO_2 , CH_4 e nitrogênio (N_2) na interface água-sedimento

Uma grande parte dos GEE é originada da decomposição da matéria orgânica presente nos sedimentos anóxicos (sem oxigênio), os quais constituem-se em componente fundamental nas transformações de carbono e nitrogênio nos ambientes aquáticos. A coleta destes sedimentos é feita pelo IIE.

4.2 Banco de Dados Analisado

A cada ano são realizadas campanhas de medida pelas equipes do projeto Furnas em dois reservatórios diferentes, durante as estações seca e chuvosa. Nas análises aqui realizadas foram consideradas as campanhas indicadas na Tabela 4.1, nos reservatórios de Serra da Mesa (GO), Manso (MT), Corumbá (GO) e Itumbiara (GO/MG). A localização geográfica destes reservatórios, todos localizados no bioma do Cerrado, está indicada no Figura 4.3. Estas campanhas foram escolhidas por serem as mais completas já realizadas.

Tabela 4.1 - Campanhas por reservatório

Represas	Período		
Serra de Mesa	Novembro/2003	Março/2004	Julho/2004
Manso	Novembro/2003	Março/2004	Julho/2004
Corumbá	Novembro/2004	Março/2005	Julho/2005
Itumbiara	Novembro/2004	Março/2005	Julho/2005

Os dados foram retirados do banco de dados do projeto, disponível mediante senha no site <http://www.dpi.inpe.br/sima/>. No total foram considerados 166 variáveis ambientais (ver Apêndice B), medidas, em princípio, durante as 12 campanhas experimentais. Novamente, dentro do espírito da analogia com a análise de MA, cada campanha (por exemplo, Manso, julho/2004) representa um “paciente”, e uma coluna na base de dados. Já cada propriedade ambiental (pH, fluxo de CO_2 , etc.), corresponde a um “gene”, e uma linha na base de dados. Estendendo a analogia, pode-se imaginar a emissão de GEE como uma “doença”, e os fatores ambientais causadores do fenômeno, os genes reguladores ainda desconhecidos.



Figura 4.3 - Represas pertencentes ao Projeto Furnas analisadas (INPE et al., 2006).

Dependendo da resolução espacial e/ou temporal utilizada, valores médios são computadas de modo que cada variável ambiental esteja associada a apenas um valor numérico por campanha. Os valores faltantes, são substituídos pelas medianas calculadas sobre todas as campanhas realizadas. Em seguida os dados foram normalizados para variarem no intervalo $[-1, 1]$.

Como na aplicação anterior, a extração do conhecimento do banco de dados é realizada através de “projetos”, que necessitam a definição das “classes” que nortearão as operações de classificação e agrupamento. As classes foram pré-determinadas adotando-se o critério da mediana. Em outras palavras, para analisar “Fluxo CH_4 (bolha), interface água-atmosfera”, por exemplo, foi calculada a mediana desta variável em todas as campanhas. A campanha (coluna na tabela) que tiver o valor desta variável menor que a mediana, é classificada como classe 1, caso contrário, como classe 2. Os projetos considerados respondem a uma pergunta do tipo: “Quais são as variáveis ambientais, dentre as consideradas no banco de dados, responsáveis pela emissão do gás X pertencer a uma determinada classe?”. Especificamente foram analisados os fluxos difusivos e ebulitivos de CH_4 e CO_2 nas interfaces água-atmosfera e sedimento-água. Ou seja, dentre todos os parâmetros contidos no projeto, pretende-se identificar aqueles que influenciam estas emissões.

Neste trabalho foi adotada uma abordagem constituída de três etapas complementares para cada projeto. Primeiramente foi analisada uma planilha contendo todas as 12 campanhas. Como segunda etapa, para verificar a robustez dos resultados da primeira fase, eliminou-se sucessivamente uma campanha (coluna do banco de dados) e repetiu-se 12 vezes as operações de classificação. Os resultados finais foram condensados em um histograma. Por fim, realiza-se a operação de agrupamento com 12 campanhas, mas considerando apenas as variáveis ambientais que mais apareceram no histograma (>5 , por exemplo) e que já haviam sido selecionadas na primeira análise. O p-valor utilizado em todas as análises foi 0.05. Nesta análise também utilizou-se o fato de ao atribuir-se um p-valor maior, obtém-se muitos parâmetros na classificação e a chance de ocorrerem falsos positivos aumenta muito. Já para p-valor menor, apesar de diminuir a probabilidade de falsos positivos, aparecem poucas variáveis climatológicas na análise. Por isso é necessário fazer vários testes até que possa se obter um p-valor satisfatório.

4.3 Resultados

Os resultados dos quatro projetos considerados estão apresentados a seguir:

a) *Fluxo CH_4 (bolha), interface água-atmosfera*

Na análise de *Fluxo CH_4 (bolha), interface água-atmosfera*, a primeira etapa do projeto, contendo as 12 campanhas, gerou o agrupamento ilustrado na Figura 4.4. Na análise do mesmo parâmetro na segunda etapa do projeto, foram encontrados 30 parâmetros relevantes onde foram selecionados os 17 que apareciam em no mínimo 6 vezes do total de 12 repetições (Figura 4.5). A Figura 4.6 mostra a operação de agrupamento final onde observa-se um nítido agrupamento das campanhas que estão acima da mediana (roxo), e das que estão abaixo da mediana (rosa). Em especial, destaca-se o estoque de CO_2 na interface sedimento-água, que aparece em todas as análises realizadas. Este resultado ilustra a estreita correlação entre os processos de geração de metano e dióxido de carbono na coluna d'água dos reservatórios.

Observa-se, também, que nas duas primeiras campanhas de Corumbá e nas três de Itumbiara, que estão abaixo da mediana em termos de fluxo ebulitivo de CH_4 na interface água-atmosfera, a concentração de fósforo total na água é baixa. Segundo Esteves (1988) o fósforo é importante devido à sua participação no processo do metabolismo dos seres vivos, sendo assim um elemento indispensável ao crescimento das algas. Como conseqüência, observa-se um baixo valor na densidade total do fitoplâncton (biótico superfície), ou seja baixo índice de vegetais presentes na superfície.

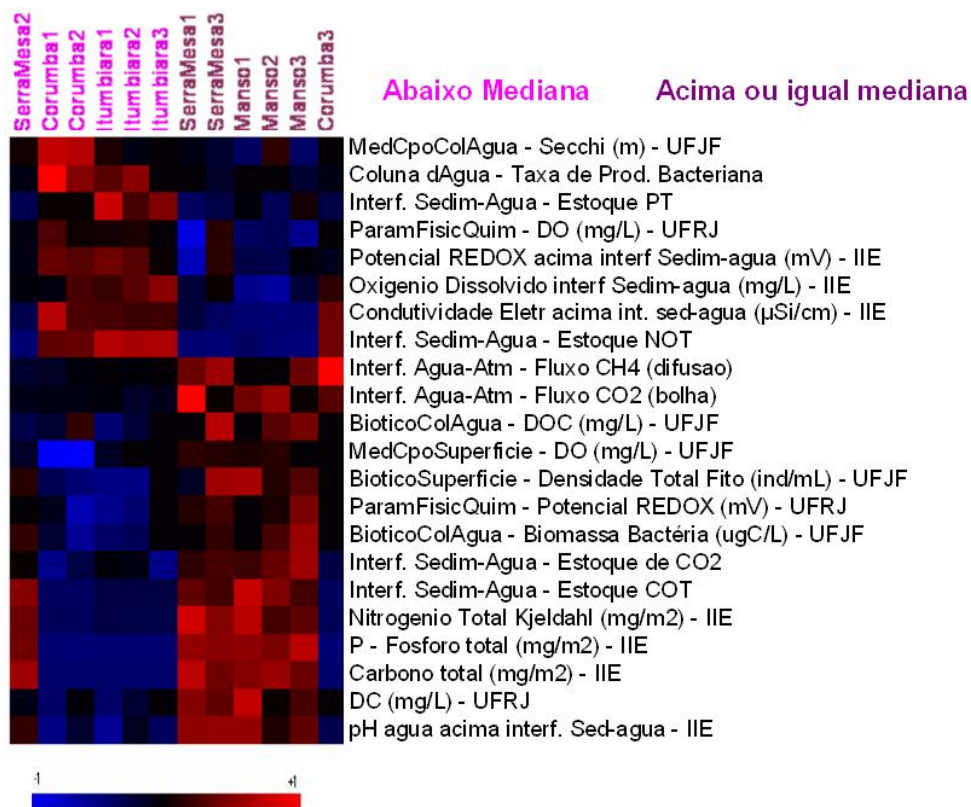
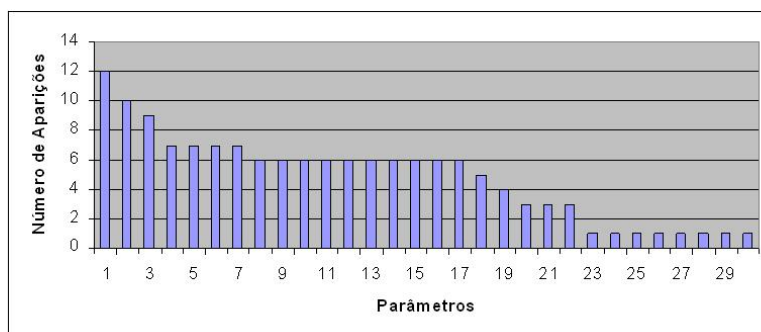


Figura 4.4 - Agrupamento - Fluxo CH_4 (bolha), interface água-atmosfera. Análise com 12 campanhas.



- | | |
|---|---|
| 1 Interf. Sedim-Agua - Estoque de CO2 | 16 Coluna d'Agua - Taxa de Prod. Bacteriana |
| 2 Interf. Agua-Atm - Fluxo CO2 (bolha) | 17 Interf. Sedim-Agua - Estoque NOT |
| 3 Interf. Agua-Atm - Fluxo CH4 (difusao) | 18 BioticoColAgua - DOC (mg/L) - UFJF |
| 4 BioticoColAgua - Densidade Total Zoo (ind/L) - UFJF | 19 Condutividade Eletr acima int. sed-agua - IIE |
| 5 Oxigenio Dissolvido acima interf Sedim-agua - IIE | 20 MedCpoColAgua - Turbidez (NTU) - UFJF |
| 6 P - Fosforo total (mg/m2) - IIE | 21 MedCpoColAgua - Secchi (m) - UFJF |
| 7 Nitrogenio Total Kjeldahl (mg/m2) - IIE | 22 Interf. Sedim-Agua - Estoque PT |
| 8 ParamFisicQuim - DO (mg/L) - UFRJ | 23 MedCpoColAgua - DO (mg/L) - UFJF |
| 9 ParamFisicQuim - Potencial REDOX (mV) - UFRJ | 24 BioticoSuperficie - Biomassa Carbono Total Fito - UFJF |
| 10 DC (mg/L) - UFRJ | 25 Sulfato (mg-S/m2) - IIE |
| 11 MedCpoSuperficie - DO (mg/L) - UFJF | 26 Acetato (mg/m2) - IIE |
| 12 BioticoSuperficie - Densidade Total Fito (ind/mL) - UFJF | 27 Carbono total (mg/m2) - IIE |
| 13 BioticoColAgua - Biomassa Bactéria (ugC/L) - UFJF | 28 Materia Organica Sedimento (mg/m2) - IIE |
| 14 pH agua acima interf. Sed-agua - IIE | 29 Coluna d'Agua - Estoque COD |
| 15 Potncial RODOX acima interf Sedim-agua - IIE | 30 Interf. Sedim-Agua - Estoque COT |

Figura 4.5 - Histograma dos parâmetros relevantes - Fluxo CH_4 (bolha), interface água-atmosfera.

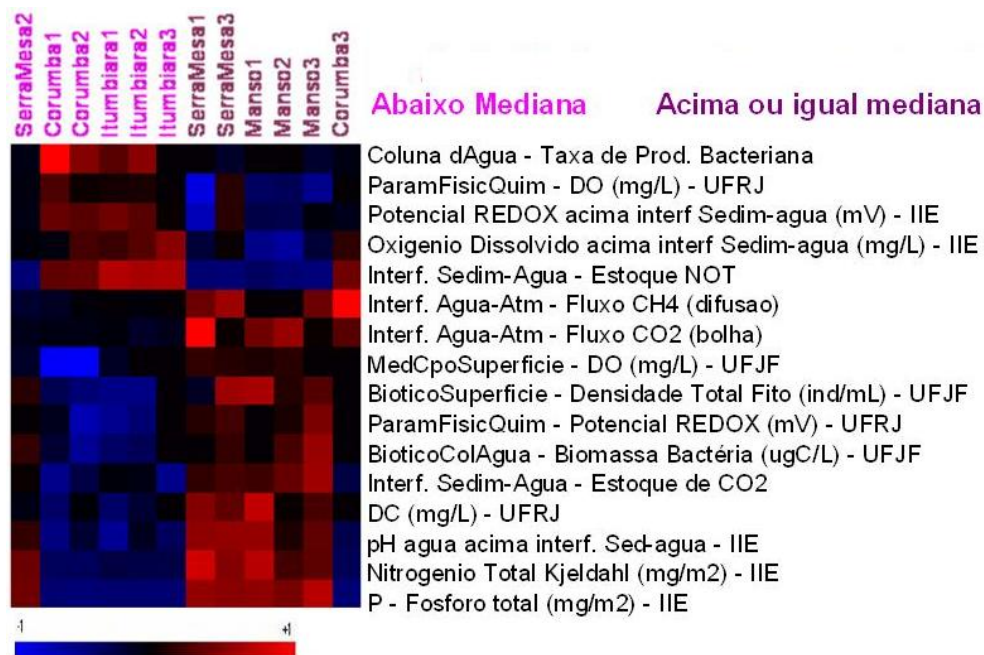


Figura 4.6 - Agrupamento dos parâmetros em comum nas duas etapas - Fluxo CH_4 (bolha), interface água-atmosfera.

Analogamente, nas três campanhas de Manso e na terceira de Serra de Mesa (fluxo de CH_4 bolha abaixo da mediana), observam-se altos valores de fósforo e altos valores de densidade total fitoplâncton (FITO). Ressalta-se que o padrão de distribuição de oxigênio em ecossistemas aquáticos é, normalmente, inverso ao do metano. Este fato pode ser observado na linha referente à variável “oxigênio dissolvido (DO)” que está anticorrelacionada com o fluxo de metano.

b) *Fluxo CO_2 (bolha), interface água-atmosfera*

Na análise de *Fluxo CO_2 (bolha), interface água-atmosfera*, o projeto contendo as 12 campanhas gerou o agrupamento ilustrado na Figura 4.7. Na análise do mesmo parâmetro numa outra abordagem, os 11 projetos geraram 30 parâmetros relevantes onde foram selecionados os 17 que apareciam em no mínimo 6 projetos (Figura 4.8). A Figura 4.9 mostra o agrupamento dos parâmetros mais relevantes em comum entre as duas abordagens estudadas. Estes resultados são idênticos aos obtidos no projeto anterior, com o fluxo de metano. A razão para isto está na forte correlação positiva entre os dois fluxos, ilustrada na Figura 4.10.

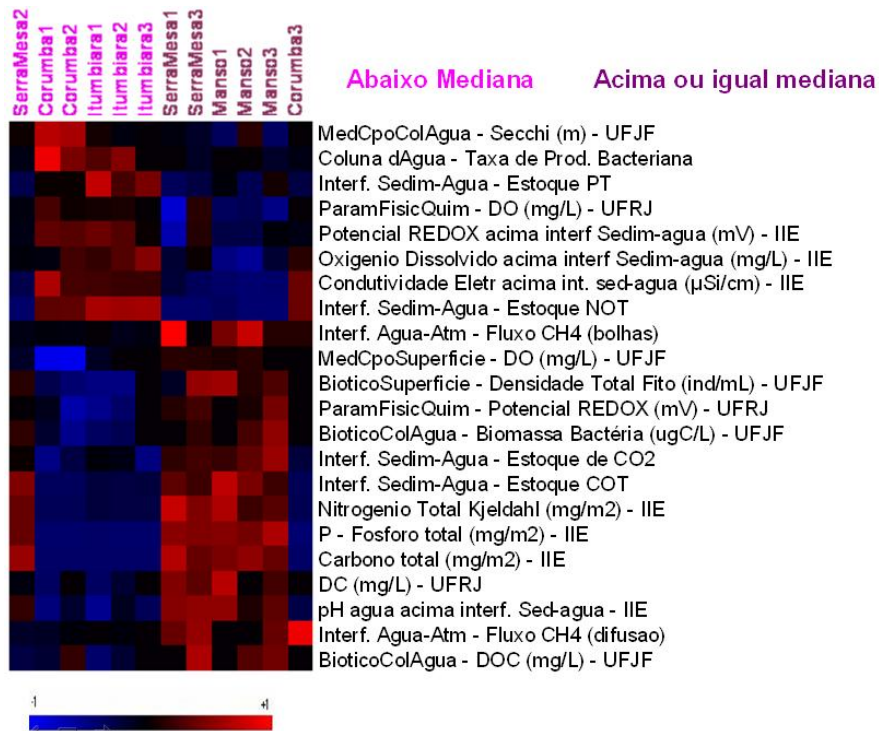
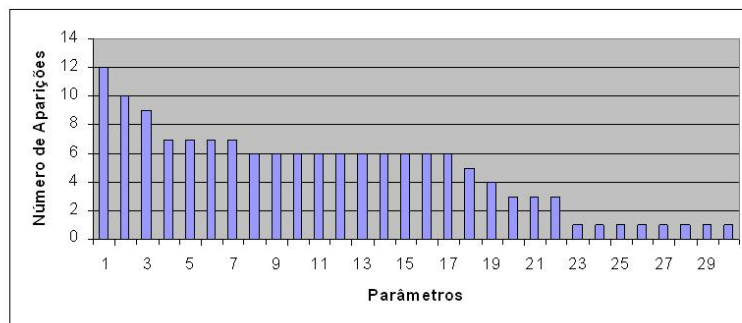


Figura 4.7 - Agrupamento - Fluxo CO_2 (bolha), interface água-atmosfera. Análise com 12 campanhas.



- | | |
|---|---|
| 1 Interf. Sedim-Agua - Estoque de CO2 | 16 Coluna dAgua - Taxa de Prod. Bacteriana |
| 2 Interf. Agua-Atm - Fluxo CH4 (difusao) | 17 Interf. Sedim-Agua - Estoque NOT |
| 3 Interf. Agua-Atm - Fluxo CH4 (bolhas) | 18 BioticoColAgua - DOC (mg/L) - UFJF |
| 4 BioticoColAgua - Densidade Total Zoo (ind/L) - UFJF | 19 Condutividade Eletr acima int. sed-agua - IIE |
| 5 Oxigenio Dissolvido acima interf Sedim-agua - IIE | 20 MedCpoColAgua - Turbidez (NTU) - UFJF |
| 6 P - Fosforo total (mg/m2) - IIE | 21 MedCpoColAgua - Secchi (m) - UFJF |
| 7 Nitrogenio Total Kjeldahl (mg/m2) - IIE | 22 Interf. Sedim-Agua - Estoque PT |
| 8 ParamFisicQuim - DO (mg/L) - UFRJ | 23 MedCpoColAgua - DO (mg/L) - UFJF |
| 9 ParamFisicQuim - Potencial REDOX (mV) - UFRJ | 24 BioticoSuperficie - Biomassa Carbono Total Fito - UFJF |
| 10 DC (mg/L) - UFRJ | 25 Sulfato (mg-S/m2) - IIE |
| 11 MedCpoSuperficie - DO (mg/L) - UFJF | 26 Acetato (mg/m2) - IIE |
| 12 BioticoSuperficie - Densidade Total Fito (ind/mL) - UFJF | 27 Carbono total (mg/m2) - IIE |
| 13 BioticoColAgua - Biomassa Bactéria (ugC/L) - UFJF | 28 Materia Organica Sedimento (mg/m2) - IIE |
| 14 pH agua acima interf. Sed-agua - IIE | 29 Coluna dAgua - Estoque COD |
| 15 Potencial REDOX acima interf Sedim-agua - IIE | 30 Interf. Sedim-Agua - Estoque COT |

Figura 4.8 - Histograma dos parâmetros relevantes - Fluxo CO_2 (bolha), interface água-atmosfera.

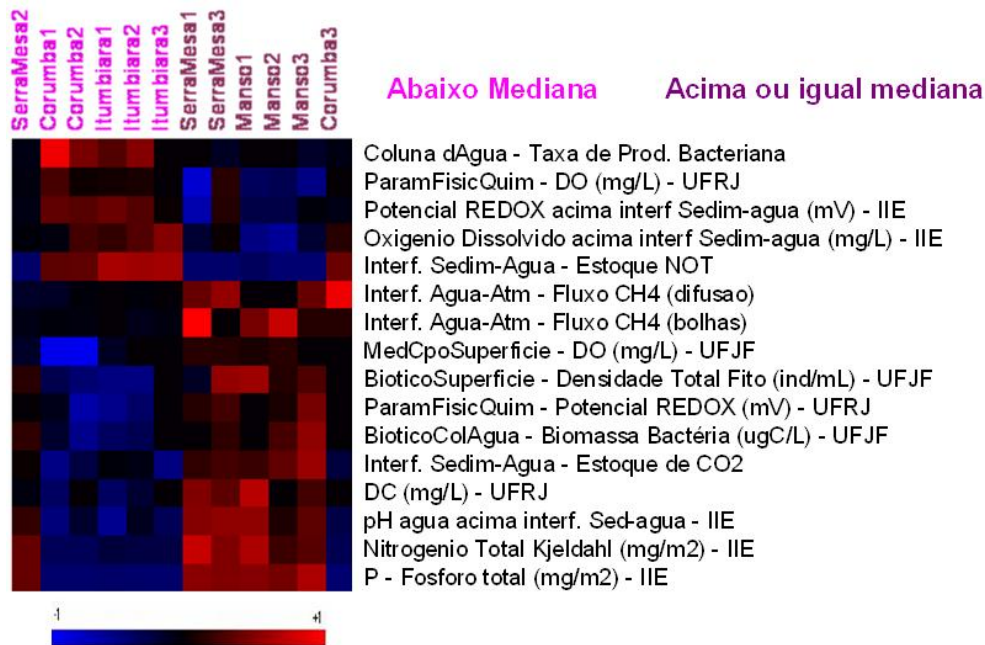
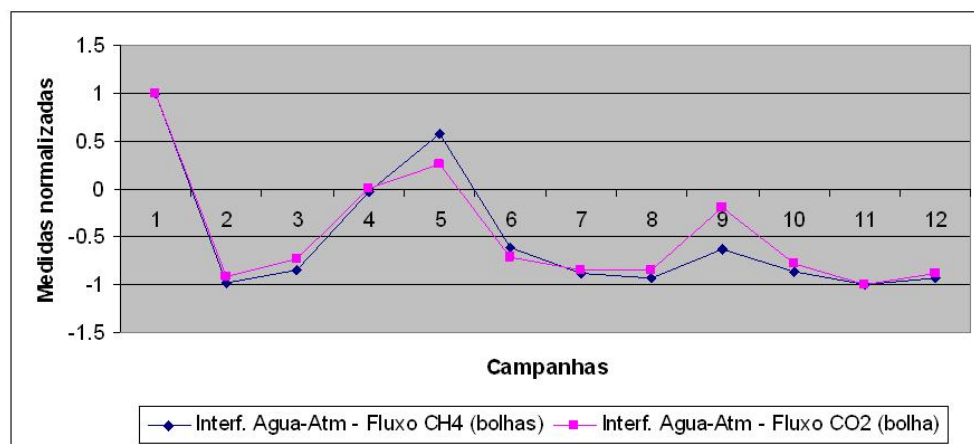


Figura 4.9 - Agrupamento dos parâmetros em comum nas duas etapas - Fluxo CO_2 (bolha), interface água-atmosfera.



Campanhas	
1 SerraMesa1	7 Corumba1
2 SerraMesa2	8 Corumba2
3 SerraMesa3	9 Corumba3
4 Manso1	10 Itumbiara1
5 Manso2	11 Itumbiara2
6 Manso3	12 Itumbiara3

Figura 4.10 - Gráfico comparativo das medidas normalizadas de CO_2 e CH_4 (bolha), interface água-atmosfera.

Visualizando-se o agrupamento em 3D (via a função Multidimensional Scaling of Samples), observa-se uma nítida separação das campanhas que estão acima

da mediana das que estão abaixo (Figura 4.11), o que corrobora os resultados obtidos nas três etapas de análise.

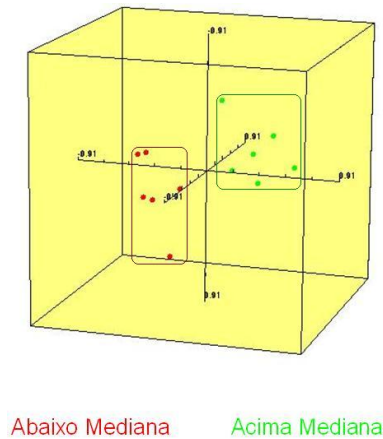


Figura 4.11 - Gráfico em escala multidimensional - Fluxo CO_2 e CH_4 (bolha), interface água-atmosfera.

c) Fluxo CH_4 , interface sedimento-água

Na análise de Fluxo CH_4 , interface sedimento-água, a primeira etapa do projeto, contendo as 12 campanhas, gerou o agrupamento ilustrado na Figura 4.12. Na análise do mesmo parâmetro na segunda etapa do projeto, foram encontrados 11 parâmetros relevantes onde foram selecionados os 3 que apareciam em no mínimo 6 projetos (Figura 4.13). A Figura 4.14 mostra a operação de agrupamento final onde observa-se um nítido agrupamento das campanhas que estão acima da mediana (roxo), e das que estão abaixo da mediana (rosa). O número baixo de parâmetros selecionados (em comparação com os dois projetos anteriores) explica-se pelo fato do sedimento ser pouco ou nada susceptível às variáveis ambientais que caracterizam a interface água-atmosfera (o inverso, como visto, não é verdadeiro). Em outras palavras, as relações de causalidade vão no sentido sedimento \rightarrow água \rightarrow ar. Ainda assim é possível agrupar as campanhas como demonstra a análise em escala 3D (Figura 4.15). No entanto, apenas três campanhas são comuns às obtidas na análise de fluxo de metano na interface água-atmosfera.



Figura 4.12 - Agrupamento - Fluxo CH_4 interface sedimento-água. Análise com 12 campanhas.

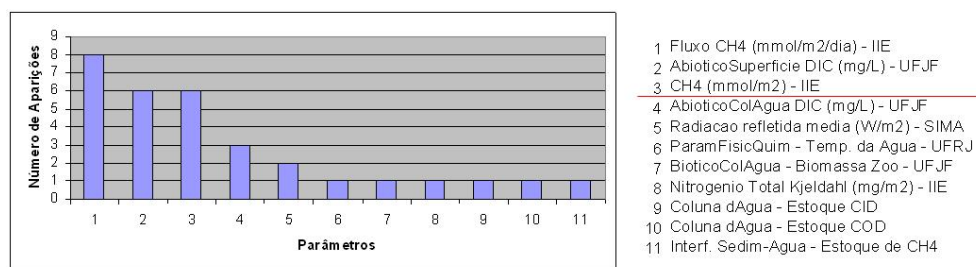


Figura 4.13 - Histograma dos parâmetros relevantes - Fluxo CH_4 , interface sedimento-água.



Figura 4.14 - Agrupamento dos parâmetros em comum nas duas etapas - Fluxo CH_4 , interface sedimento-água.

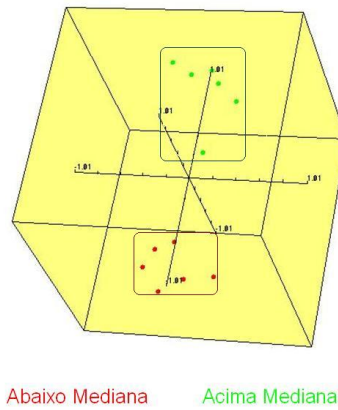


Figura 4.15 - Gráfico em escala multidimensional - Fluxo CH_4 , interface sedimento-água.

d) Fluxo CO_2 , interface sedimento-água

Nesta última análise, de *Fluxo CO_2 , interface sedimento-água*, o projeto contendo as 12 campanhas gerou o agrupamento ilustrado na Figura 4.16. Analisando-se consecutivamente 11 campanhas, foram selecionados 26 parâmetros e escolhidos 10 variáveis relevantes, conforme Figura 4.17. A Figura 4.18 mostra o agrupamento resultante. Excepcionalmente neste caso, o programa selecionou um número muito pequeno de parâmetros, mas que ainda assim permitem classificar os reservatórios. Neste caso, quatro campanhas são comuns às obtidas na análise de fluxo de dióxido de carbono na interface água-atmosfera.

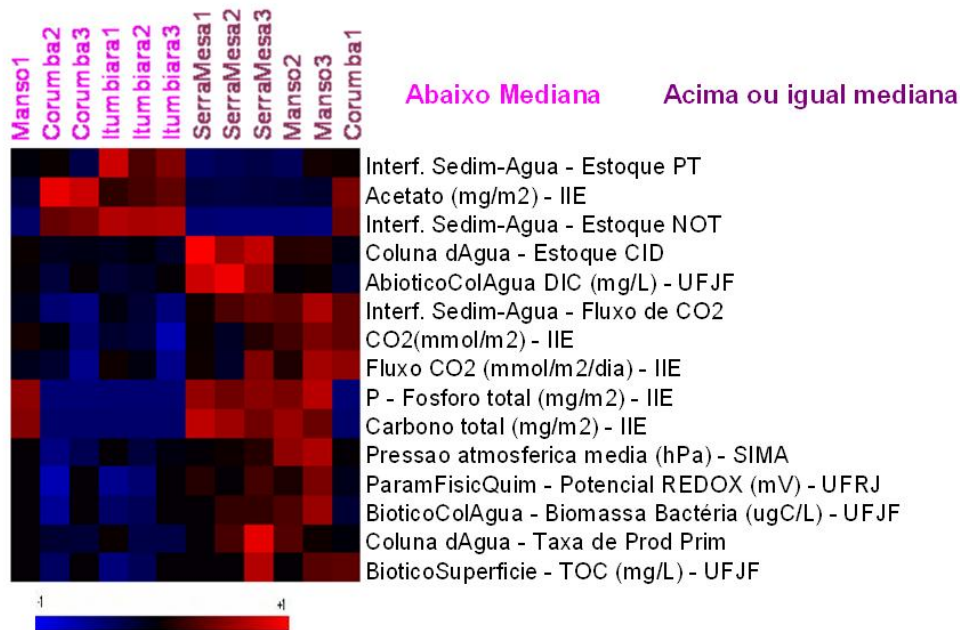
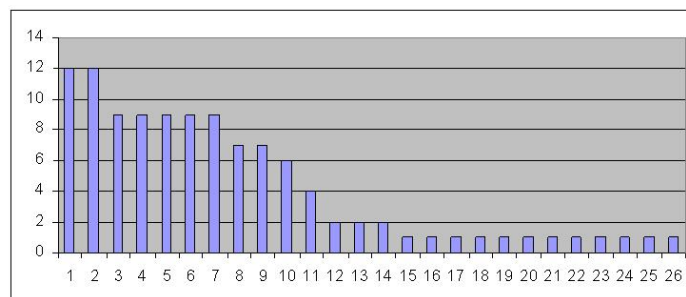


Figura 4.16 - Agrupamento - Fluxo CO_2 , interface sedimento-água. Análise feita com as 12 campanhas.



- | | |
|---|--|
| 1 Interf. Sedim-Agua - Estoque PT | 13 Temp. da água 20m media (C) - SIMA |
| 2 Interf. Sedim-Agua - Estoque NOT | 14 Interf. Sedim-Agua - Fluxo de N20 |
| 3 CamaraSolo - CO2 (mg/m2/dia) - UFRJ | 15 Conc. de NH4+ media (mg/l) - SIMA |
| 4 MedCpoSuperficie - Cond. (uS/cm) - UFJF | 16 Temp. da sonda media (C) - SIMA |
| 5 MedCpoColAgua - Cond. (uS/cm) - UFJF | 17 Vel. meridional da corrente media (cm/s) - SIMA |
| 6 BioticoSuperficie - Clorofila A (ugC/L) - UFJF | 18 CamaraSolo - CH4 (mg/m2/dia) - UFRJ |
| 7 BioticoColAgua - Clorofila A (ugC/L) - UFJF | 19 MedCpoSuperficie - Material em suspensao - UFJF |
| 8 AbioticoSuperficie PT (mM) - UFJF | 20 Fluxo Carbono - Carbono OrgAnico Excretado - UFJF |
| 9 Oxigenio Dissolvido acima interf Sedim-agua - IIE | 21 BioticoSuperficie - TOC (mg/L) - UFJF |
| 10 AbioticoColAgua PT (mM) - UFJF | 22 AbioticoSuperficie NT (mM) - UFJF |
| 11 CO2(mmol/m2) - IIE | 23 AbioticoColAgua NT (mM) - UFJF |
| 12 Conc. de NO3- media (mg/l) - SIMA | 24 pH agua acima interf. Sed-agua - IIE |
| | 25 Interf. Agua-Atm - Fluxo N2O (bolha) |
| | 26 Coluna d'Agua - Taxa de Respiracao Bact |

Figura 4.17 - Histograma dos parâmetros relevantes - Fluxo CO_2 , interface sedimento-água.

De uma maneira geral observando-se os resultados apresentados, conclui-se que o BRB-ArrayTools apresentou um desempenho satisfatório, digno de ser analisado mais detalhadamente por limnólogos. Convém lembrar que existe muito ruído e lacunas nos dados,



Figura 4.18 - Agrupamento dos parâmetros em comum nas duas etapas - Fluxo CO_2 , interface sedimento-água.

além de uma forte variabilidade interna (removida pelo processo de filtragem pela média), devido a diferenças de local e horário da medição.

De certa forma a importância da abordagem proposta nesta dissertação é ressaltada pelo fato de que, até hoje, os diferentes grupos de pesquisa do projeto apenas analisavam seus próprios dados. Este trabalho reúne todos os dados e faz uma análise global, que até então não havia sido feita. Para um trabalho futuro, pretende-se considerar o local e hora da medida na análise. Isto porque as represas são muito grandes, com formas irregulares (ver Figura 4.19), e as medidas analisadas foram colhidas nos mais diversos pontos.

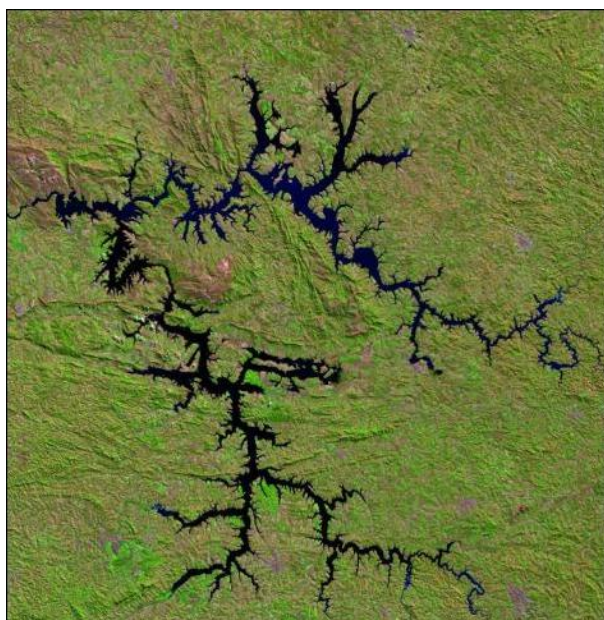


Figura 4.19 - Imagem da Represa Manso.

5 CONCLUSÃO

Uma das conseqüências da crescente preocupação mundial com o meio ambiente é o aumento acentuado do volume de dados disponíveis para a comunidade científica. Esta dissertação teve por objetivo principal demonstrar a viabilidade do uso de técnicas computacionais, que são utilizadas atualmente na análise de experimentos de microarranjos de DNA, na área ambiental. Para este fim, o pacote BRB-ArrayTools (SIMON; LAM, 2006) foi adaptado, validado e aplicado ao estudo de dois problemas ambientalmente relevantes em climatologia e em limnologia.

Na primeira aplicação, foram investigados os fatores climáticos responsáveis pela grande seca de 2005 na Amazônia. Os resultados obtidos indicam que a temperatura da superfície do mar na região do Atlântico Tropical Norte tem um impacto direto no regime caudal do rio Amazonas e alguns de seus afluentes. Outros parâmetros como a temperatura da superfície do mar na costa sul brasileira parecem ter um papel relevante e merecem ser analisadas em detalhe pelos especialistas. Por outro lado, o fenômeno El Niño/La Niña não foi selecionado uma única vez nas análises realizadas. Estes resultados, de um modo geral, são corroborados pela literatura especializada. Ressalte-se, no entanto, que pela primeira vez um volume de dados tão grande é utilizado para analisar o fenômeno da seca na Amazônia.

Na segunda aplicação, analisou-se o banco de dados do Projeto Balanço de Carbono Furnas com o objetivo de identificar os fatores ambientais relevantes que controlam a emissão de gases de efeito estufa (GEE) em reservatórios. Os resultados obtidos permitiram a classificação e o agrupamento das campanhas sem grandes dificuldades, conforme observado nas figuras geradas, apesar das lacunas e do ruído contido nos dados, e ilustram a estreita correlação entre os processos de geração de metano e dióxido de carbono na coluna d'água dos reservatórios.

Num sentido mais amplo, os resultados desta dissertação corroboram a conjectura que está na origem deste trabalho, a saber, de que métodos da bioinformática para o tratamento de grandes volumes dados podem ser extremamente úteis na área ambiental. Como sugestão de trabalhos futuros, propõe-se a completa transposição e adaptação do pacote BRB-ArrayTools para a área ambiental, com a modificação do jargão típico da biologia molecular e a introdução de novas funcionalidades porventura necessárias para as suas novas aplicações. Mais especificamente, pretende-se, na área climatológica, fazer um estudo sistemático de todas as bacias hidrográficas brasileiras, identificando quais são as variáveis ambientais mais relevantes para sua hidrologia. No caso da aplicação limnológica, almeja-se completar o banco de dados com as outras campanhas do Projeto Furnas,

como Marimbondo, Estreito, Porto Colômbia, Mascarenhas de Moraes e Funil.

REFERÊNCIAS BIBLIOGRÁFICAS

- AMARATUNGA, D.; CABRERA, J. **Exploration and analysis of DNA microarray and protein array data**. New Jersey: Wiley Interscience, 2004. 246 p. [23](#), [31](#), [32](#)
- ANDRADE, L. P. **Procedimento interativo de agrupamento de dados**: Dissertação de (mestrado em engenharia). Rio de Janeiro, Brasil: Universidade Federal do Rio de Janeiro, 2004. 201 p. [30](#), [31](#)
- BALDI, P.; BRUNAK, S. **Bioinformatics: The machine learning approach**. New York: Cambridge University Press, 2001. Segunda Edição. [25](#)
- BAMBACE, L. A. W.; RAMOS, F. M.; LIMA, I. B. T.; ROSA, R. R. Mitigation and recovery of methane emission from tropical hydroelectric dams. **Energy**, v. 32, p. 1038–1046, 2007. [62](#)
- BLATT, M.; WISEMAN, S.; DOMANY, E. Data clustering using a model granular magnet. **Neural Computation**, v. 9, p. 1805–1842, 1997. [30](#)
- CARVALHO, A. C. P. L. F. Computação bioinspirada. Revista Eletrônica de Ciências, n. 22, 2003. Disponível em: <http://cdcc.sc.usp.br/ciencia/artigos/art_22/computacaobioinspirada.html>. Acesso em: 23 jan. 2007. [27](#)
- CPTEC/INPE. **Glossário**. 2006. Disponível em: <<http://www7.cptec.inpe.br/glossario/>>. Acesso em: 02 fev. 2009. [45](#)
- DANTAS, D. O. **Uma técnica automática baseada em morfologia matemática para medida de sinal em imagens de cDNA**: Dissertação de (mestrado em ciência da computação). São Paulo: Universidade de São Paulo, 2004. [27](#), [28](#)
- DOMANY, E. Cluster analysis of gene expression data. **Journal of Statistical Physics**, v. 110, p. 1117–1139, 2003. [25](#), [26](#), [30](#)
- DUDA, R. O.; HART, P. E. **Pattern classification and scene analysis**. California: Wiley-interscience, 1973. 482 p. [30](#)
- EISEN, M. B.; SPELLMAN, P. T.; BROWN, P. O.; BOTSTEIN, D. Cluster analysis and display of genome-wide expression patterns. **PNAS**, v. 95, p. 14863–14868, 1998. [32](#), [33](#)

ESTEVEES, F. A. **Fundamentos de limnologia**. Rio de Janeiro - Brasil: Editora Interciência LTDA/FINEP, 1988. 575 p. Primeira Edição. 61, 66

HANAI, T.; HAMADA, H.; OKAMOTO, M. Application of bioinformatics for dna microarray data to bioscience, bioengineering and medical fields. **Journal of Bioscience and Bioengineering**, v. 101, 2006. 31

HAUTANIEMI, S. **Studies of microarray Data analysis with applications for human cancers**. Tampere, Finland: Tampere University of Technology, 2003. 87 p. 27, 31, 33

HEY, T.; TREFETHEN, A. The data deluge: An e-science perspective. **Grid Computing - Making the Global Infrastructure a Reality**, Wiley and Sons, p. 809–824, 2003. 23

HOLTON, J. R. **An Introduction to Dynamic Meteorology**. Seattle, WA: Academic Press, 2004. 535 p. 45

INPE; UFJF; COPPETEC; IEE. **O balanço de carbono nos reservatórios de FURNAS centrais elétricas S. A.** 2006. Disponível em: <<http://www.dsr.inpe.br/projetofurnas>>. Acesso em: 15 abr. 2007. 16, 61, 63, 65

KANAMITSU, M.; EBISUZAKI, W.; WOOLLEN, J.; YANG, S.; HNILO, J.; FIORINO, M.; POTTER, G. Ncep-doe amip-ii reanalysis (r-2). **American Meteorological Society**, v. 83, p. 1631–1643, 2002. 45

KHAN, J.; WET, J. S.; RINGNÉR, M.; SAAL, L. H.; LADANYI, M.; WESTTERMANN, F.; BERTHOLD, F.; SCHWAB, M.; ANTONESCU, C. R.; PETERSON, C.; MELTZER, P. S. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. **Nature Medicine**, v. 7, p. 673–679, 2001. 33

KRUTOVSKII, K. V.; NEALE, D. B. **Forest genomics for conserving adaptive genetic diversity**. Rome: Food and Agriculture Organization of the United Nations - FAO, 2001. Disponível em: <<ftp://ftp.fao.org/docrep/fao/004/x6884e/x6884e00.pdf>>. 27, 29

LIKAS, A.; VLASSIS, N.; VERBEEK, J. J. The global k-means clustering algorithm. **The journal of the pattern recognition society**, v. 36, p. 451–461, 2003. 31

LIMA, I. B. T. Biogeochemical distinction of methane releases from two amazon hydroreservoirs. **Chemosphere**, v. 59, p. 1697–1702, Dezembro 2005. 62

LIMA, I. B. T.; RAMOS, F. M.; BAMBACE, L. A. W.; ROSA, R. R. Methane emissions from large dams as renewable energy resources: A developing nation perspective. **Springer Science**, v. 13, p. 193–206, 2008. [61](#)

MAKRETSOV, N. A.; HUNTSMAN, D. G.; NIELSEN, T. O.; YORIDA, E.; PEACOCK, M.; CHEANG, M. C. U.; DUNN, S. E.; HAYES, M.; RIJN, M. van de; BAJDIK, C.; GILKS, C. B. Hierarchical clustering analysis of tissue microarray immunostaining data identifies prognostically significant groups of breast carcinoma. **Clinical Cancer Research**, v. 10, p. 6143–6151, 2004. [33](#)

MANTUA, N. J. The pacific decadal oscillation and climate forecasting for north america. **Climate Risk Solutions**, v. 1, p. 10–13, 1999. [48](#)

MARENGO, J. A.; NOBRE, C. A.; TOMASELLA, J.; OYAMA, M. D.; OLIVEIRA, G. S.; OLIVEIRA, R.; CAMARGO, H.; ALVES, L. M.; BROWN, I. F. The drought of amazonia in 2005. **Journal of Climate**, v. 21, n. 3, 2008. [43](#), [44](#), [57](#), [58](#)

OSCILLATION, N. N. A. **Climate analysis group**. NOAA/Lamont-Doherty Earth Observatory NAO Pamphlet, 2005. Disponível em: <http://www.ldeo.columbia.edu/NAO/main.html>. Acesso em: 15 ago. 2007. [48](#)

OSCILLATION, P. T. P. D. **The Pacific decadal oscillation**. 2000. Disponível em: <http://jisao.washington.edu/pdo/>. Acesso em: 15 ago. 2007. [49](#)

PAES, A. T. Itens essenciais em bioestatística. *Arq. Bras. Cardiol.*, v. 71, p. 575–580, 1998. [38](#)

RAMOS, F. M.; LIMA, I. B. T.; ROSA, R. R.; MAZZI, E. A.; CARVALHO, J. C.; RASERA, M. F. F. L.; OMETTO, J. P. H. B.; ASSIREU, A. T.; STECH, J. Extreme event dynamics in methane ebullition fluxes from tropical reservoirs. **Geophysical Research Letters**, v. 33, p. L21404–94, 2006. [62](#)

REIS, E. M.; NAKAYA, H. I.; LOURO, R.; CANAVEZ, F. C.; FLATSCHART, A. V. F.; ALMEIDA, G. T.; EGIDIO, C. M.; PAQUOLA, A. C.; MACHADO, A. A.; FESTA, F.; YAMAMOTO, D.; ALVARENGA, R.; SILVA, C. C. da; BRITO, G. C.; SIMON, S. D.; MOREIRA-FILHO, C. A.; LEITE, K. R.; CAMARA-LOPES, L. H.; CAMPOS, F. S.; GIMBA, E.; VIGNAL, G. M.; EL-DORRY, H.; SOGAYAR, M. C.; BARCINSKI, M. A.; SILVA, A. M. da; VERJOVSKI-ALMEIDA, S. Antisense intronic non-coding rna levels correlate to the degree of tumor differentiation in prostate cancer. **Oncogene**, v. 23, p. 6684–6692, 2004. [15](#), [39](#), [40](#), [41](#)

SCHLANGER, V. **Condições Metereológicas**. Hungarian Meteorological Service: ESPERE - Environmental Science Published for Everybody Round the Earth, 2006. Disponível em: <http://www.atmosphere.mpg.de/enid/2__Principais_sistemas_de_vento/-_El_Ni_o___SOI_4z1.html>. Acesso em: 19 jul. 2007. 47

SIMON, R.; LAM, A. P. **BRB-ArrayTools - version 3.4 - User's manual**. National Cancer Institute, 2006. 108 p. Disponível em: <<http://linus.nci.nih.gov/~brb/download.html>>. Acesso em: 10 jan. 2002. 23, 34, 39, 77

SIMON, R. M.; KORN, E. L.; MCSHANE, L. M.; RADMACHER, M. D.; WRIGHT, G. W.; ZHAO, Y. **Statistics for biology and health**. New York: Springer-Verlag, 2003. 199 p. 31, 35, 39

SIMPSON, A. J. G. The human genome project and its implication for human health. In: **Anais do I congresso brasileiro de biossegurança**. Rio de Janeiro: [s.n.], 1999. 25

SOUTO, M. C. P.; LORENA, A. C.; DELBEM, A. C. B.; CARVALHO, A. C. P. L. F. **Técnicas de aprendizado de máquina para problemas de biologia molecular**. Universidade de São Paulo - São Carlos: Editora SBC, 2004. 103–152 p. III Jornada de Mini-Curso de Inteligência Artificial - Livro Texto, capítulo Técnicas de Aprendizado de Máquina para Problemas de Biologia Molecular. 29

STEINER, M. T. A.; SOMA, N. Y.; SHIMIZU, T.; NIEVOLA, J. C.; NETO, P. J. S. Study of a medical problem using kdd, with emphasis on exploratory data analysis. **Gest. Prod.**, v. 13, p. 325–337, 2006. 30

TRENBERTH, K. E.; SHEA, D. J. **Atlantic hurricane and natural variability in 2005**. Geophysical Research Letters. Disponível em: <<http://www.cgd.ucar.edu/cas/trenberth.pdf/TrenberthSheaHurricanes2006GRL026894.pdf>>. Acesso em: 22 set. 2007. 57, 58

WALLACE, J. M.; HOBBS, P. V. **Atmospheric science: an introductory survey**. Massachusetts: Academic Press, 2006. 483 p. 45, 46

A - Relação de parâmetros analisados no Projeto Climatológico

Amazonas 15°S a 5°S - 75°W a 50°W

20N40N140W120W - Comp Meridional Vento

20N40N140W120W - Umidade Relativa

20N40N140W120W - Comp Zonal Vento

20N40N140W120W - Temperatura do ar

20N40N140W120W - Temperatura superfície mar

20N40N140W120W - Radiação Onda Longa emergente

20N40N140W120W - Pressão red. nível do mar

20N40N140W120W - Movimento Vertical

20N40N140W120W - Altura Geopotencial

20N40N120W100W - Comp Meridional Vento

20N40N120W100W - Umidade Relativa

20N40N120W100W - Comp Zonal Vento

20N40N120W100W - Temperatura do ar

20N40N120W100W - Temperatura superfície mar

20N40N120W100W - Radiação Onda Longa emergente

20N40N120W100W - Pressão red. nível do mar

20N40N120W100W - Movimento Vertical

20N40N120W100W - Altura Geopotencial

20N40N100W80W - Comp Meridional Vento

20N40N100W80W - Umidade Relativa

20N40N100W80W - Comp Zonal Vento

20N40N100W80W - Temperatura do ar

20N40N100W80W - Temperatura superfície mar

20N40N100W80W - Radiação Onda Longa emergente

20N40N100W80W - Pressão red. nível do mar

20N40N100W80W - Movimento Vertical

20N40N100W80W - Altura Geopotencial

20N40N80W60W - Comp Meridional Vento

20N40N80W60W - Umidade Relativa

20N40N80W60W - Comp Zonal Vento

20N40N80W60W - Temperatura do ar

20N40N80W60W - Temperatura superfície mar

20N40N80W60W - Radiação Onda Longa emergente

20N40N80W60W - Pressão red. nível do mar

20N40N80W60W - Movimento Vertical

20N40N80W60W - Altura Geopotencial
20N40N60W40W - Comp Meridional Vento
20N40N60W40W - Umidade Relativa
20N40N60W40W - Comp Zonal Vento
20N40N60W40W - Temperatura do ar
20N40N60W40W - Temperatura superfície mar
20N40N60W40W - Radiação Onda Longa emergente
20N40N60W40W - Pressão red. nível do mar
20N40N60W40W - Movimento Vertical
20N40N60W40W - Altura Geopotencial
20N40N40W20W - Comp Meridional Vento
20N40N40W20W - Umidade Relativa
20N40N40W20W - Comp Zonal Vento
20N40N40W20W - Temperatura do ar
20N40N40W20W - Temperatura superfície mar
20N40N40W20W - Radiação Onda Longa emergente
20N40N40W20W - Pressão red. nível do mar
20N40N40W20W - Movimento Vertical
20N40N40W20W - Altura Geopotencial
20N40N20W0W - Comp Meridional Vento
20N40N20W0W - Umidade Relativa
20N40N20W0W - Comp Zonal Vento
20N40N20W0W - Temperatura do ar
20N40N20W0W - Temperatura superfície mar
20N40N20W0W - Radiação Onda Longa emergente
20N40N20W0W - Pressão red. nível do mar
20N40N20W0W - Movimento Vertical
20N40N20W0W - Altura Geopotencial
0N20N140W120W - Comp Meridional Vento
0N20N140W120W - Umidade Relativa
0N20N140W120W - Comp Zonal Vento
0N20N140W120W - Temperatura do ar
0N20N140W120W - Temperatura superfície mar
0N20N140W120W - Radiação Onda Longa emergente
0N20N140W120W - Pressão red. nível do mar
0N20N140W120W - Movimento Vertical
0N20N140W120W - Altura Geopotencial
0N20N120W100W - Comp Meridional Vento

0N20N120W100W - Umidade Relativa
0N20N120W100W - Comp Zonal Vento
0N20N120W100W - Temperatura do ar
0N20N120W100W - Temperatura superfície mar
0N20N120W100W - Radiação Onda Longa emergente
0N20N120W100W - Pressão red. nível do mar
0N20N120W100W - Movimento Vertical
0N20N120W100W - Altura Geopotencial
0N20N100W80W - Comp Meridional Vento
0N20N100W80W - Umidade Relativa
0N20N100W80W - Comp Zonal Vento
0N20N100W80W - Temperatura do ar
0N20N100W80W - Temperatura superfície mar
0N20N100W80W - Radiação Onda Longa emergente
0N20N100W80W - Pressão red. nível do mar
0N20N100W80W - Movimento Vertical
0N20N100W80W - Altura Geopotencial
0N20N80W60W - Comp Meridional Vento
0N20N80W60W - Umidade Relativa
0N20N80W60W - Comp Zonal Vento
0N20N80W60W - Temperatura do ar
0N20N80W60W - Temperatura superfície mar
0N20N80W60W - Radiação Onda Longa emergente
0N20N80W60W - Pressão red. nível do mar
0N20N80W60W - Movimento Vertical
0N20N80W60W - Altura Geopotencial
0N20N60W40W - Comp Meridional Vento
0N20N60W40W - Umidade Relativa
0N20N60W40W - Comp Zonal Vento
0N20N60W40W - Temperatura do ar
0N20N60W40W - Temperatura superfície mar
0N20N60W40W - Radiação Onda Longa emergente
0N20N60W40W - Pressão red. nível do mar
0N20N60W40W - Movimento Vertical
0N20N60W40W - Altura Geopotencial
0N20N40W20W - Comp Meridional Vento
0N20N40W20W - Umidade Relativa
0N20N40W20W - Comp Zonal Vento

0N20N40W20W - Temperatura do ar
0N20N40W20W - Temperatura superfície mar
0N20N40W20W - Radiação Onda Longa emergente
0N20N40W20W - Pressão red. nível do mar
0N20N40W20W - Movimento Vertical
0N20N40W20W - Altura Geopotencial
0N20N20W0W - Comp Meridional Vento
0N20N20W0W - Umidade Relativa
0N20N20W0W - Comp Zonal Vento
0N20N20W0W - Temperatura do ar
0N20N20W0W - Temperatura superfície mar
0N20N20W0W - Radiação Onda Longa emergente
0N20N20W0W - Pressão red. nível do mar
0N20N20W0W - Movimento Vertical
0N20N20W0W - Altura Geopotencial
20S0S140W120W - Comp Meridional Vento
20S0S140W120W - Umidade Relativa
20S0S140W120W - Comp Zonal Vento
20S0S140W120W - Temperatura do ar
20S0S140W120W - Temperatura superfície mar
20S0S140W120W - Radiação Onda Longa emergente
20S0S140W120W - Pressão red. nível do mar
20S0S140W120W - Movimento Vertical
20S0S140W120W - Altura Geopotencial
20S0S120W100W - Comp Meridional Vento
20S0S120W100W - Umidade Relativa
20S0S120W100W - Comp Zonal Vento
20S0S120W100W - Temperatura do ar
20S0S120W100W - Temperatura superfície mar
20S0S120W100W - Radiação Onda Longa emergente
20S0S120W100W - Pressão red. nível do mar
20S0S120W100W - Movimento Vertical
20S0S120W100W - Altura Geopotencial
20S0S100W80W - Comp Meridional Vento
20S0S100W80W - Umidade Relativa
20S0S100W80W - Comp Zonal Vento
20S0S100W80W - Temperatura do ar
20S0S100W80W - Temperatura superfície mar

20S0S100W80W - Radiação Onda Longa emergente
20S0S100W80W - Pressão red. nível do mar
20S0S100W80W - Movimento Vertical
20S0S100W80W - Altura Geopotencial
20S0S80W60W - Comp Meridional Vento
20S0S80W60W - Umidade Relativa
20S0S80W60W - Comp Zonal Vento
20S0S80W60W - Temperatura do ar
20S0S80W60W - Temperatura superfície mar
20S0S80W60W - Radiação Onda Longa emergente
20S0S80W60W - Pressão
20S0S80W60W - Movimento Vertical
20S0S80W60W - Altura Geopotencial
20S0S60W40W - Comp Meridional Vento
20S0S60W40W - Umidade Relativa
20S0S60W40W - Comp Zonal Vento
20S0S60W40W - Temperatura do ar
20S0S60W40W - Temperatura superfície mar
20S0S60W40W - Radiação Onda Longa emergente
20S0S60W40W - Pressão red. nível do mar
20S0S60W40W - Movimento Vertical
20S0S60W40W - Altura Geopotencial
20S0S40W20W - Comp Meridional Vento
20S0S40W20W - Umidade Relativa
20S0S40W20W - Comp Zonal Vento
20S0S40W20W - Temperatura do ar
20S0S40W20W - Temperatura superfície mar
20S0S40W20W - Radiação Onda Longa emergente
20S0S40W20W - Pressão red. nível do mar
20S0S40W20W - Movimento Vertical
20S0S40W20W - Altura Geopotencial
20S0S20W0W - Comp Meridional Vento
20S0S20W0W - Umidade Relativa
20S0S20W0W - Comp Zonal Vento
20S0S20W0W - Temperatura do ar
20S0S20W0W - Temperatura superfície mar
20S0S20W0W - Radiação Onda Longa emergente
20S0S20W0W - Pressão red. nível do mar

20S0S20W0W - Movimento Vertical
20S0S20W0W - Altura Geopotencial
40S20S140W120W - Comp Meridional Vento
40S20S140W120W - Umidade Relativa
40S20S140W120W - Comp Zonal Vento
40S20S140W120W - Temperatura do ar
40S20S140W120W - Temperatura superfície mar
40S20S140W120W - Radiação Onda Longa emergente
40S20S140W120W - Pressão red. nível do mar
40S20S140W120W - Movimento Vertical
40S20S140W120W - Altura Geopotencial
40S20S120W100W - Comp Meridional Vento
40S20S120W100W - Umidade Relativa
40S20S120W100W - Comp Zonal Vento
40S20S120W100W - Temperatura do ar
40S20S120W100W - Temperatura superfície mar
40S20S120W100W - Radiação Onda Longa emergente
40S20S120W100W - Pressão red. nível do mar
40S20S120W100W - Movimento Vertical
40S20S120W100W - Altura Geopotencial
40S20S100W80W - Comp Meridional Vento
40S20S100W80W - Umidade Relativa
40S20S100W80W - Comp Zonal Vento
40S20S100W80W - Temperatura do ar
40S20S100W80W - Temperatura superfície mar
40S20S100W80W - Radiação Onda Longa emergente
40S20S100W80W - Pressão red. nível do mar
40S20S100W80W - Movimento Vertical
40S20S100W80W - Altura Geopotencial
40S20S80W60W - Comp Meridional Vento
40S20S80W60W - Umidade Relativa
40S20S80W60W - Comp Zonal Vento
40S20S80W60W - Temperatura do ar
40S20S80W60W - Temperatura superfície mar
40S20S80W60W - Radiação Onda Longa emergente
40S20S80W60W - Pressão red. nível do mar
40S20S80W60W - Movimento Vertical
40S20S80W60W - Altura Geopotencial

40S20S60W40W - Comp Meridional Vento
40S20S60W40W - Umidade Relativa
40S20S60W40W - Comp Zonal Vento
40S20S60W40W - Temperatura do ar
40S20S60W40W - Temperatura superfície mar
40S20S60W40W - Radiação Onda Longa emergente
40S20S60W40W - Pressão red. nível do mar
40S20S60W40W - Movimento Vertical
40S20S60W40W - Altura Geopotencial
40S20S40W20W - Comp Meridional Vento
40S20S40W20W - Umidade Relativa
40S20S40W20W - Comp Zonal Vento
40S20S40W20W - Temperatura do ar
40S20S40W20W - Temperatura superfície mar
40S20S40W20W - Radiação Onda Longa emergente
40S20S40W20W - Pressão red. nível do mar
40S20S40W20W - Movimento Vertical
40S20S40W20W - Altura Geopotencial
40S20S20W0W - Comp Meridional Vento
40S20S20W0W - Umidade Relativa
40S20S20W0W - Comp Zonal Vento
40S20S20W0W - Temperatura do ar
40S20S20W0W - Temperatura superfície mar
40S20S20W0W - Radiação Onda Longa emergente
40S20S20W0W - Pressão red. nível do mar
40S20S20W0W - Movimento Vertical
40S20S20W0W - Altura Geopotencial

Humaitá

Manicoré

Óbidos

SOI

Wind - 200 millibar Zonal Winds Equator (165°W-110°W)

PDO

NAO-update

sst - Atl N (5-20N, 60-30W)

sst - Atl S (0-20S, 30W-10E)

sst - Global Tropics (10S-10N, 0-360)

slp 0N20N120W100W (-) 20S0S120W100W

sst 0N20N120W100W (-) 20S0S120W100W
slp 0N20N100W80W (-) 20S0S100W80W
sst 0N20N100W80W (-) 20S0S100W80W
12S3N77O65O - Precipitação
índice integrado (Humaitá Manicoré Óbidos)

Os parâmetros estão explicados no capítulo 3, exceto as diferenças entre slp (pressão reduzida ao nível do mar) e sst. Estes representam a diferença entre os índices: Pressão reduzida ao nível do mar, e Temperatura da superfície mar nas coordenadas indicadas. O último parâmetro, índice integrado (Humaitá Manicoré Óbidos), representa uma média entre as vazões dos rios próximos aos municípios citados.

B - Relação de parâmetros analisados no Projeto Carbono Furnas

- Universidade Federal de Juiz de Fora - **UFJF**

Medidas são feitas na coluna d'água.

AbioticoColAgua DIC (mg/L) - quantidade de Carbono Inorgânico Dissolvido em forma de estrutura não viva medido na coluna d'água.

AbioticoColAgua NT (mM) - quantidade de Nitrogênio Total em forma de estrutura não viva medido na coluna d'água.

AbioticoColAgua PT (mM) - quantidade de Fósforo Total em forma de estrutura não viva medido na coluna d'água.

AbioticoSuperficie DIC (mg/L) - quantidade de Carbono Inorgânico Dissolvido em forma de estrutura não viva medido na superfície.

AbioticoSuperficie NT (mM) - quantidade de Nitrogênio Total em forma de estrutura não viva medido na superfície.

AbioticoSuperficie PT (mM) - quantidade de Fósforo Total em forma de estrutura não viva medido na superfície.

BioticoColAgua - DOC (mg/L) - quantidade de Carbono Orgânico Dissolvido em forma de estrutura viva medido na coluna d'água.

BioticoColAgua - TOC (mg/L) - quantidade de Carbono Orgânico Total em forma de estrutura viva medido na coluna d'água.

BioticoColAgua POC (mg/L) - quantidade de Carbono Orgânico Particulado em forma de estrutura viva medido na coluna d'água.

BioticoColAgua - Densidade Bactéria (10^6 cels/mL) - quantidade de Bactéria na coluna d'água em forma viva.

BioticoColAgua - Biomassa Bactéria (ugC/L)

BioticoColAgua - Clorofila A (ugC/L) - quantidade de microorganismos vegetais na água.

BioticoColAgua - Biomassa Carbono Total Fito (ugC/L) - quantidade de Carbono Total (orgânico e inorgânico) presente nos organismos fitoplancteanos.

BioticoColAgua - Densidade Total Fito (ind/mL) - quantidade de fitoplâncton na coluna d'água.

BioticoColAgua - Biomassa Zoo (ugC/L)

BioticoColAgua - Densidade Total Zoo (ind/L)

BioticoSuperficie - DOC (mg/L)

BioticoSuperficie - TOC (mg/L)

BioticoSuperficie POC (mg/L)

BioticoSuperficie - Densidade Bacteria (10^6 cels/mL)

BioticoSuperficie - Biomassa Bactéria (ugC/L)

BioticoSuperficie - Clorofila A (ugC/L)

BioticoSuperficie - Biomassa Carbono Total Fito (ugC/L)

BioticoSuperficie - Densidade Total Fito (ind/mL)

BioticoSuperficie - Biomassa Zoo (ugC/L)

BioticoSuperficie - Densidade Total Zoo (ind/L)

Fluxo Carbono - Producao Fitoplanctonica (mgC/m2/d) - o quanto de carbono o fitoplancton tirou do meio e transformou em biomassa.

Fluxo Carbono - Carbono Orgânico Excretado (mgC/m2/d) - quantidade de Carbono Organico excretado.

Fluxo Carbono - Respiração Fito (mgC/m2/d)

Fluxo Carbono - Produção Bacteriana (mgC/m2/d) - o quanto de carbono a população bacteriana tirou do meio e transformou em biomassa.

Fluxo Carbono - RespiraCAo Bacteriana (mgC/m2/d) - o quanto é emitido de Carbono para o meio (feita a noite).

Fluxo Carbono - Taxa de sedimentação (g/m2/d)

MedCpoColAgua - Temp. da Agua - temperatura da água na coluna água medida em °C.

MedCpoColAgua - DO (mg/L) - quantidade de Oxigênio Dissolvido na coluna água.

MedCpoColAgua - Secchi (m) - mede transparência da água. Trata-se de uma medida de turbidez mais clássica.

MedCpoColAgua - Ph - quantidade de PH na coluna água.

MedCpoColAgua - Turbidez (NTU) - o quanto a água está clara ou escura. A turbidez mostra o quanto a luz pode penetrar.

MedCpoColAgua - Cond. (uS/cm) - Condutividade. Mede o quanto de sais está dissolvido na água.

MedCpoColAgua - Material em suspensao (mg/L) - mede o quanto de matéria, por exemplo poeira, folhas,... , está suspensa na água.

MedCpoColAgua - Intensidade Luminosa (uM/cm2/s)

MedCpoSuperficie - Temp. da Agua

MedCpoSuperficie - DO (mg/L)

MedCpoSuperficie - Secchi (m)

MedCpoSuperficie - Ph - quantidade de PH na coluna água.

MedCpoSuperficie - Turbidez (NTU) - o quanto a água está clara ou escura.

MedCpoSuperficie - Cond. (uS/cm)

MedCpoSuperficie - Material em suspensao (mg/L)

- Instituto Internacional de Ecologia - **IIE**

Medidas são feitas no sedimento.

Horiba-Sedimento - pH

Horiba-Sedimento - Cond. (uS/cm)

Horiba-Sedimento - DO (mg/L)

Horiba-Sedimento - Temp. da água (C)

Horiba-Sedimento - TDS (g/L)

Horiba-Sedimento - Potencial REDOX (mV)

Horiba-Sedimento - Turbidez (NTU)

Conc. Gas Agua - CH₄ (mM)

Conc. Gas Agua - CO₂ (mM)

VarFisicQuimAgua -F- = quantidade de flúor.

VarFisicQuimAgua Cl- = quantidade de Cloro.

VarFisicQuimAgua N-NO₂- = quantidade de Nitrogênio em forma de Nitrato.

VarFisicQuimAgua Br = quantidade de Bromo.

VarFisicQuimAgua N-NO₃- = quantidade de Nitrogênio na forma de Nitrito.

VarFisicQuimAgua P-PO₄³⁻ = quantidade de fósforo na forma de perclorato.

VarFisicQuimAgua SSO₄²⁻ -

VarFisicQuimAgua S-SO₄²⁻ =

VarFisicQuimAgua Na = quantidade de Sódio.

VarFisicQuimAgua N-NH₄ = quantidade de nitrogênio na forma de sulfato de amônia.

VarFisicQuimAgua K = quantidade de potássio.

VarFisicQuimAgua Mg = quantidade de Magnésio.

VarFisicQuimAgua Ca = quantidade de Cálcio.

VarFisicQuimAgua Clorofila = quantidade de Clorofila.

VarFisicQuimAgua Feofitina =

VarFisicQuimAgua Secchi =

CH₄ (mmol/m²)

CO₂(mmol/m²)

Fluxo CH₄ (mmol/m²/dia)

Fluxo CO₂ (mmol/m²/dia)

Materia Organica Sedimento (mg/m²)

Nitrogenio Total Kjeldahl (mg/m²) - O Nitrogênio Kjeldahl é a soma dos nitrogênios orgânico e amoniacal. Ambas as formas estão presentes em detritos de nitrogênio orgânico oriundos de atividades biológicas naturais. O nitrogênio

Kjeldahl total pode contribuir para a completa abundância de nutrientes na água e sua eutrofização. Os nitrogênios amoniacal e orgânico são importantes para avaliar o nitrogênio disponível para as atividades biológicas. A concentração de Nitrogênio Kjeldahl Total em rios que não são influenciados pelo excesso de insumos orgânicos variam de 1 a 0,5 mg/L.

P - Fosforo total (mg/m²)

Carbono total (mg/m²)

Acetato (mg/m²)

Nitrito ($\mu\text{g} - \text{N}/\text{m}^2$)

Nitrato ($\mu\text{g} - \text{N}/\text{m}^2$)

Sulfato (mg-S/m²)

Amônio (mg-S/m²)

Secchi (m)

Temp agua acima interf Sedim-agua

Potencial REDOX da agua acima interf Sedim-agua (mV)

Oxigenio Dissolvido na agua acima interf Sedim-agua (mg/L)

Condutividade Eletr agua acima int. sed-agua ($\mu\text{Si}/\text{cm}$)

pH agua acima interf. Sed-agua

- Instituto Alberto Luiz Coimbra de Pós-Graduação e Pesquisa de Engenharia - **COPPE/UFRJ**

Medidas são feitas na superfície.

Bolhas - CO₂ (mg/m²/dia)

Bolhas - O₂ (mg/m²/dia)

Bolhas - N₂ (mg/m²/dia)

Bolhas - CH₄ (mg/m²/dia)

Bolhas - N₂O (mg/m²/dia)

CamaraSolo - CH₄ (mg/m²/dia)

CamaraSolo - CO₂ (mg/m²/dia)

CamaraSolo - N₂O (mg/m²/dia)

DC (mg/L) - quantidade de Carbono Dissolvido.

DOC (mg/L) - quantidade de Carbono Orgânico Dissolvido.

POC (mg/L) - quantidade de Carbono Orgânico Particulado (reduzido a partículas).

Difusao - (CH₄) (mg/m²/dia) - mede o fluxo de (CH₄) por processo difusivo.

Difusao - (CO₂) (mg/m²/dia) - mede o fluxo de (CO₂) por processo difusivo.

Difusao - (N_2O) (mg/m²/dia) - mede o fluxo de (N_2O) por processo difusivo.
 Difusao - Temp. do ar (C)
 Difusao - Temp. da agua (C)
 Difusao - pH
 Difusao - Vel. do vento (m/s)
 GasesemBolhas - (CO_2) (ml) - quantidade de Gás Carbônico aparece na forma de bolhas no transporte.
 GasesemBolhas - (O_2) (ml) - quantidade de oxigênio aparece na forma de bolhas no transporte.
 GasesemBolhas - (N_2) (ml) - quantidade de Nitrogênio aparece na forma de bolhas no transporte.
 GasesemBolhas - (CH_4) (ml) - quantidade de Metano aparece na forma de bolhas no transporte.
 ParamFisicQuim - Temp. do ar ($^{\circ}C$)
 ParamFisicQuim - Temp. da Agua (C)
 ParamFisicQuim - pH
 ParamFisicQuim - Potencial REDOX (mV) Potencial de redução de oxigênio.
 ParamFisicQuim - DO (mg/L) - quantidade de Oxigênio Dissolvido.

- Sistema Integrado de Monitoração Ambiental **SIMA**

Temp. da agua 2m média = temperatura média da água numa profundidade de 2m.
 Temp. da agua 5m média = temperatura média da água numa profundidade de 5m.
 Temp. da agua 20m média = temperatura média da água numa profundidade de 20m.
 Temp. da agua 40m média = temperatura média da água numa profundidade de 40m.
 Umidade relativa do ar media (%)
 Pressao atmosferica media (hPa)
 Radiacao incidente media (W/m²)
 Radiacao refletida media (W/m²)
 Vel. meridional da corrente media (cm/s)
 Vel. zonal da corrente media (cm/s)
 Temp. da sonda media (C)
 Condutividade media (mS/cm)
 Porcentagem de DO media (%)

Conc. de DO media (mg/l)
pH media
Conc. de NH₄⁺ media (mg/l)
Conc. de NO₃⁻ media (mg/l)
Turbidez media (NTU)
Clorofila media (ug/l)
Bateria da sonda media (V)

Para um melhor entendimento, seguem algumas explicações fornecidas por membros do projeto Furnas:

Abiótico = lugar SEM seres vivos.

Biótico = lugar COM seres vivos.

Biomassa = matéria orgânica viva existente em determinado espaço.

Disco de Secchi = dispositivo circular para medir visualmente a transparência da água.

pH = potencial hidrogeniônico. O valor do pH é um número aproximado entre 0 e 14 que indica se uma solução é ácida ($pH < 7$), neutra ($pH = 7$), ou básica/alcalina ($pH > 7$).

Fitoplâncton (FITO) = comunidade de plantas marinhas microscópicas, em sua maioria fotossintetizante, que vive em suspensão na coluna d'água. Absorvem o dióxido de carbono atmosférico com a fotossíntese. Apesar de individualmente microscópico, a clorofila do fitoplâncton tingem coletivamente as águas dos reservatórios.

Potencial REDOX = Potencial de redução de oxigênio.

Além das medidas acima foi utilizada uma tabela integrada que contém medidas das diversas entidades, obtida no site <http://www.dpi.inpe.br/sima/>.