sid.inpe.br/mtc-m19/2013/04.16.19.46-TDI

FERRAMENTA DE AUTOMAÇÃO PARA DESCOBERTA DE CONHECIMENTO EM BANCO DE DADOS NOAA

Wanderson Gomes de Almeida

Dissertação de Mestrado em Computação Aplicada, orientada pelos Drs. Fernando Manuel Ramos, e Walter Abrahão dos Santos, aprovada em 07 de maio de 2013.

URL do documento original:
<http://urlib.net/8JMKD3MGP7W/3DTLCCH>

INPE
São José dos Campos
2013

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) 3208-6923/6921
Fax: (012) 3208-6919
E-mail: pubtc@sid.inpe.br

CONSELHO DE EDITORAÇÃO E PRESERVAÇÃO DA PRODUÇÃO INTELECTUAL DO INPE (RE/DIR-204):
Presidente:
Marciana Leite Ribeiro - Serviço de Informação e Documentação (SID)
Membros:
Dr. Antonio Fernando Bertachini de Almeida Prado - Coordenação Engenharia e Tecnologia Espacial (ETE)
Drª Inez Staciarini Batista - Coordenação Ciências Espaciais e Atmosféricas (CEA)
Dr. Gerald Jean Francis Banon - Coordenação Observação da Terra (OBT)
Dr. Germano de Souza Kienbaum - Centro de Tecnologias Especiais (CTE)
Dr. Manoel Alonso Gan - Centro de Previsão de Tempo e Estudos Climáticos (CPT)
Drª Maria do Carmo de Andrade Nono - Conselho de Pós-Graduação
Dr. Plínio Carlos Alvalá - Centro de Ciência do Sistema Terrestre (CST)
BIBLIOTECA DIGITAL:
Dr. Gerald Jean Francis Banon - Coordenação de Observação da Terra (OBT)
REVISÃO E NORMALIZAÇÃO DOCUMENTÁRIA:
Marciana Leite Ribeiro - Serviço de Informação e Documentação (SID)
Yolanda Ribeiro da Silva Souza - Serviço de Informação e Documentação (SID)
EDITORAÇÃO ELETRÔNICA:
Maria Tereza Smith de Brito - Serviço de Informação e Documentação (SID)
Luciana Manacero - Serviço de Informação e Documentação (SID)

sid.inpe.br/mtc-m19/2013/04.16.19.46-TDI

FERRAMENTA DE AUTOMAÇÃO PARA DESCOBERTA DE CONHECIMENTO EM BANCO DE DADOS NOAA

Wanderson Gomes de Almeida

Dissertação de Mestrado em Computação Aplicada, orientada pelos Drs. Fernando Manuel Ramos, e Walter Abrahão dos Santos, aprovada em 07 de maio de 2013.

URL do documento original:
<http://urlib.net/8JMKD3MGP7W/3DTLCCH>

INPE
São José dos Campos
2013

Dados Internacionais de Catalogação na Publicação (CIP)

Almeida, Wanderson Gomes de.
Al64f Ferramenta de automação para descoberta de conhecimento em banco de dados NOAA / Wanderson Gomes de Almeida. – São José dos Campos : INPE, 2013.
xxii + 132 p. ; (sid.inpe.br/mtc-m19/2013/04.16.19.46-TDI)

Dissertação (Mestrado em Computação Aplicada) – Instituto Nacional de Pesquisas Espaciais, São José dos Campos, 2013.
Orientadores : Drs. Fernando Manuel Ramos, e Walter Abrahão dos Santos.

1. Descoberta de Conhecimento em Banco de Dados (KDD). 2. Engenharia de Software. 3. Mineração de Dados. 4. Sistemas de Informação Geográfica (GIS). I.Título.

CDU 004.4

Esta obra foi licenciada sob uma Licença Creative Commons Atribuição-NãoComercial 3.0 Não Adaptada.

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported License.

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

Computação Aplicada

Dr. Eduardo Martins Guerra

Presidente / INPE / São José dos Campos - SP

Dr. Fernando Manuel Ramos

Orientador(a) / INPE / SJCampos - SP

Dr. Walter Abrahão dos Santos

Orientador(a) / INPE / São José dos Campos - SP

Dr. Haroldo Fraga de Campos Velho

Membro da Banca / INPE / São José dos Campos - SP

Dr. Luiz Alberto Vieira Dias

Convidado(a) / ITA / São José dos Campos - SP

Este trabalho foi aprovado por:

( ) maioria simples

(X) unanimidade

Aluno (a): ***Wanderson Gomes de Almeida***

“O Senhor é o meu pastor, nada me faltará. Ensina-me a fazer a vossa vontade, pois sois o meu Deus. Que vosso Espírito de bondade me conduza pelo caminho reto”.

Salmos 22,1; 142,10

AGRADECIMENTOS

Gostaria de agradecer primeiramente a Deus, pela vida, por estar sempre em nossos caminhos, nos iluminando e nos guiando às escolhas certas. Os sinceros agradecimentos às instituições e pessoas, as quais tornaram possível o desenvolvimento deste presente trabalho.

A meus pais e irmãos pelo apoio, amor e carinho. Pela minha esposa que tanto me incentivou e apoiou em minhas decisões. Aos colegas e amigos pelo estímulo, amizade, carinho, críticas e sugestões durante o curso.

Ao Prof. Dr. Fernando Manuel Ramos e ao Dr. Walter Abrahão dos Santos, pela confiança depositada em mim, pela dedicação e disponibilidade na função de orientadores, e pelo estímulo em desenvolver este trabalho de mestrado.

Agradeço também a Dr. Heloisa Musetti Ruivo pelo incentivo e colaboração no processo de desenvolvimento deste trabalho. Aos demais professores, alunos, pesquisadores e técnicos do INPE pelo ensino, receptividade, acolhida, apoio técnico e administrativo.

Em especial ao eterno Dr. José Demisio Simões da Silva pelos seus ensinamentos e orientação no curso de pós-graduação em Computação Aplicada e pelo seu grande exemplo de amizade e dedicação profissional.

Finalmente, às instituições CNPq e CAPES pelo apoio financeiro concedido a este trabalho de pesquisa e desenvolvimento de software.

RESUMO

O estudo de eventos climáticos é importante na tomada de decisão em políticas públicas, devido ao seu impacto social e econômico. No entanto, esse estudo lida com uma massa considerável de dados. O desenvolvimento de ferramentas eficazes para apoiar este processo é altamente recomendado, pois aumentam a produtividade e confiabilidade, especialmente se eles fornecem análise e visualização de fenômenos. Este trabalho apresenta uma ferramenta que aborda as fases de Descoberta de Conhecimento em Bases de Dados, do inglês *Knowledge Discovery in Databases* (KDD) de séries espaço-temporal, que são: pré-processamento, mineração de dados e pós-processamento. A ferramenta foi desenvolvida em Java empregando soluções de código majoritariamente abertos e *scripts* em MATLAB. Um estudo de caso é apresentado por meio de um conjunto de dados da grande seca de 2005 na Amazônia do Centro Nacional para Pesquisa Atmosférica da Administração Nacional Oceânica e Atmosférica (NOAA). A ferramenta concebida automatiza o processo KDD que era antes realizado manualmente, portanto, contribuindo para uma maior produtividade na análise e visualização dos dados.

AN AUTOMATION TOOL FOR KNOWLEDGE DISCOVERY IN NOAA DATABASES

ABSTRACT

The study of weather events is important in decision making in public policies due to its both social and economic impacts. Nevertheless, such study deals with a considerable data mass. The development of effective tools for supporting this process is highly recommended as they increase productivity and reliability, especially if they provide phenomena analysis and visualization. This work presents a tool addressing the phases of Knowledge Discovery in Databases (KDD) for time-spatial series, which are: pre-processing, data mining and post-processing. The tool has been developed in Java employing majorly open source solutions and MATLAB. A case study is presented using a dataset of the 2005 Amazon Great Drought from the National Center for Atmospheric Research of the National Oceanic and Atmospheric Administration (NOAA). The conceived tool automates the KDD process which was before manually performed therefore contributing to higher productivity on data analysis and visualization.

LISTA DE FIGURAS

Pág.

Figura 2.1 - KDD e suas três etapas essenciais..... 8
Figura 2.2 – Estrutura do Arquivo de Dados NetCDF. 22
Figura 3.1 – Construção de Microarranjos. 27
Figura 3.2 – Exemplo de Calculo de p-valor..... 31
Figura 3.3 – Métricas de distância entre os grupos..... 36
Figura 3.4 – Agrupamento hierárquico e representação por cores. 38
Figura 3.5 – Agrupamento usando BRB-ArrayTools na área Ambiental. 40
Figura 4.1 - Diagrama de Caso de Uso do ERB-ArrayTools. 42
Figura 4.2 - Esquema de integração dos subsistemas..... 43
Figura 4.3 - Diagrama de Classe da tela Principal do ERB-ArrayTools. 44
Figura 4.4 - Diagrama de Classe em alto nível dos subsistemas integrados. .. 47
Figura 4.5 - Diagrama de Pacotes do ERB-ArrayTools..... 48
Figura 4.6 - Diagrama de Classe do subsistema ProduceNetCDF. 55
Figura 4.7 - Exemplo pictórico de consultas geométricas usando PostGIS. 57
Figura 4.8 - Tabelas espaço-temporal criadas pelo ChronosGIS..... 58
Figura 4.9 - Diagrama de Classe do subsistema ChronosGIS..... 61
Figura 4.10 - Diagrama de Pacotes desenvolvidos em Matlab. 62
Figura 4.11 - Diagrama de Pacotes dos gráficos estatísticos..... 65
Figura 4.12 - Diagrama de Classe do Subsistema StatisticalAnalysis. 67
Figura 4.13 - Diagrama de Classe dos subsistemas ToolsUI e IDV..... 68
Figura 4.14 – Arquitetura do Projeto do ERB-ArrayTools..... 70
Figura 4.15 – Script build.xml para geração do ERB-ArrayTool.jar 72
Figura 4.16 – Script NSIS para criar programa de instalação. 73
Figura 4.17 – Script runERB.bat para execução do sistema..... 74
Figura 5.1 – Programa de instalação do ERB-ArrayTools..... 75
Figura 5.2 – Estrutura do ERB-ArrayTools instalado no Windows 7. 76
Figura 5.3 – Instalação dos Scripts na pasta statisticalAnalysis..... 77
Figura 5.4 – Interface gráfica do ERB-ArrayTools mapeada para o KDD. 78
Figura 5.5 – Principais botões e parâmetros do ERB-ArrayTools. 79
Figura 5.6 – Tela de conexão com o BDG. 80
Figura 5.7 – Tela obtida após abrir o arquivo NetCDF. 80
Figura 5.8 – Fluxo de telas ao realizar a carga no BDG..... 82
Figura 5.9 – Subjanela com resultados da SQL usada. 83
Figura 5.10 – Dados Pré-processados para Mineração. 84
Figura 5.11 – Descritor de Dados para Mineração..... 84
Figura 5.12 – Telas do Matlab com os scripts e as variáveis definidas..... 85
Figura 5.13 – Subjanela com resultados obtidos..... 86
Figura 5.14 – Arquivos de texto e NetCDF's obtidos pela DM ao salvar..... 87
Figura 5.15 – Diferentes tipos de visualização..... 88
Figura 5.16 – Índice Integrado agrupado pelo Matlab..... 88
Figura 5.17 – Variável air visualizada no GridView do subsistema ToolUI..... 89

Figura 5.18 – Variável air visualizada no ImageView do subsistema ToolUI. .. 90
Figura 5.19 – Variável air visualizada em 3D do subsistema IDV. 90
Figura 6.1 – Coordenadas de 0W à 140W e 40N à 40S. 93
Figura 6.2 – Vento meridional visualizado pelo GridView e ImageView. 94
Figura 6.3 – Vento meridional visualizado em 3D pelo IDV. 94
Figura 6.4 – Temperatura do ar na superfície pelo GridView e ImageView. 95
Figura 6.5 – Temperatura do ar na superfície em 3D pelo IDV. 95
Figura 6.6 – Temperatura da superfície do mar pelo GridView e ImageView. . 95
Figura 6.7 – Temperatura na superfície do mar em 3D pelo IDV. 96
Figura 6.8 – Localização das três regiões analisadas. 96
Figura 6.9 – Agrupamento no Índice Integrado pelo Matlab. 98
Figura 6.10 – Agrupamento em Óbidos pelo Matlab. 98
Figura 6.11 – P-valor de vwnd no Índice Integrado pelo GridView e IDV. 99
Figura 6.12 – P-valor de vwnd em Óbidos pelo GridView e IDV. 99
Figura 6.13 – P-valor de air no Índice Integrado pelo GridView e IDV. 99
Figura 6.14 – P-valor de air em Óbidos pelo GridView e IDV. 100
Figura 6.15 – P-valor de sst no Índice Integrado pelo GridView e IDV. 100
Figura 6.16 – P-valor de sst em Óbidos pelo GridView e IDV. 100
Figura A.1 – Gráfico quântico normal de t. 115
Figura A.2 – Histograma dos resultados do método t-test. 116
Figura A.3 – Gráfico do p-valor no Índice Integrado utilizando o FDR. 117
Figura A.4 – Gráfico do p-valor em Óbitos utilizando o FDR. 118
Figura A.5 – Gráfico de teste de significância no Índice Integrado. 119
Figura A.6 – Gráfico de teste de significância em Óbidos. 120
Figura A.7 – Visualização da dispersão dos dados mais significativos. 121
Figura A.8 – Agrupamento no Índice Integrado. 122
Figura A.9 – Agrupamento em Óbidos. 123
Figura A.10 – Dados da variável ambiental vwnd em janeiro de 2000. 124
Figura A.11 – P-valor de vwnd no Índice Integrado e em Óbitos. 125
Figura A.12 – Dados de vwnd e seu p-valor sobrepostos. 126
Figura A.13 – Dados da variável ambiental air em janeiro de 2000. 127
Figura A.14 – P-valor de air no Índice Integrado e em Óbitos. 128
Figura A.15 – Dados de air e seu p-valor sobrepostos. 129
Figura A.16 – Dados da variável ambiental sst em janeiro de 2000. 130
Figura A.17 – P-valor de sst no Índice Integrado e em Óbitos. 131
Figura A.18 – Dados de sst e seu p-valor sobrepostos. 132

LISTA DE TABELAS

Pág.

Tabela 4.1 - Ajuste da Matriz de Dados. 50
Tabela 4.2 - Granularidade Espacial Aplicada na Matriz de Dados. 51
Tabela 4.3 - Matriz de Dados Brutos. 52
Tabela 4.4 - Matriz de Médias Aritméticas. 52
Tabela 4.5 - Matriz de Dados com Anomalia. 52
Tabela 4.6 - Matriz de Dados com Anomalia e Normalizados. 52
Tabela 4.7 - Descritor de um subconjunto de dados com tamanho par. 54
Tabela 4.8 - Descritor de um subconjunto de dados com tamanho impar. 54
Tabela 4.9 - Estatísticas espaço-temporal obtida pelo ChronosGIS. 60

LISTA DE SIGLAS E ABREVIATURAS

AEM	Análises de Estatísticas Matriciais
APIs	Application Programming Interfaces
BC	Biologia Computacional
BD	Base de Dados
BDG	Bases de Dados geográficos
CPTEC	Centro de Previsão de Tempo e Clima
DM	Mineração de Dados
DMG	Mineração de Dados Geográficos
FDR	*False Discovery Rate*
IDEs	*Integrated Development Environment*
IDV	*Integrated Data Viewer*
INPE	Instituto Nacional de Pesquisas Espaciais
INTC-MC	Instituto Nacional de Ciência e Tecnologia para Mudanças Climáticas
KDD	Descoberta de Conhecimento em Banco de Dados
KDDG	*Knowledge Discovery in Databases Geographical*
MA	Microarranjos
NetCDF	*Network Common Data Form*
NOAA	Administração Nacional do Oceano e da Atmosfera
SGBD	Sistemas de Gerenciamento de Banco de Dados
SID	Serviço de Informação e Documentação
SIG	Sistemas de Informação Geográfica
SPG	Serviço de Pós-Graduação
SQL	*Structure Query Language*
TDI	Teses e Dissertações Internas
UML	*Unified Modeling Languague*

LISTA DE SÍMBOLOS

air	Temperatura do ar na superfície
hgtoct	Altura Geopotencial
olr	Radiação da Onda Longa Emergente
rhum	Umidade Relativa
slp	Pressão ao Nível do Mar
sst	Temperatura da superfície do Mar
uwnd	Vento Zonal
vstm	Movimento Vertical
vwnd	Vento Meridional
F	Método estatístico F-teste
J_1	Quantidade de amostras da classe 1
J_2	Quantidade de amostras da classe 2
N	Número de variáveis
$\mathrm{p\text{-}valor}$	Probabilidade de encontrar em hipótese nula um t-estatístico tão grande quanto o encontrado no dado real
t	Método estatístico t-teste
$\mathrm{X, Y}$	Experimentos
$\overline{x_1}$	Média das amostras da classe 1
$\overline{x_2}$	Média das amostras da classe 2
$\mathrm{X_{avg}}$	Média das variáveis no experimento X
$\mathrm{x_i}$	Variáveis do experimento X
$\mathrm{Y_{avg}}$	Média das variáveis no experimento Y
$\mathrm{y_i}$	Variáveis do experimento Y

SUMÁRIO

Pág.

1 INTRODUÇÃO 1
1.1. Contextualização 1
1.2. Motivação 3
1.3. Objetivos 4
1.4. Abordagem da Solução 5
1.5. Organização da Dissertação 6
2 DESCOBERTA DE CONHECIMENTO EM BASE DE DADOS 7
2.1. Etapas do Processo de Descoberta de Conhecimento 7
2.2. Tipos de Conhecimento 12
2.3. Formas de Aquisição do Conhecimento 13
2.4. Tipos de Aprendizado em Base de Dados 14
2.5. Descoberta de Conhecimento em Banco de Dados Geográficos 16
2.6. Relevância dos Dados Geográficos 19
2.7. Formato de Dados NetCDF 20
3 TÉCNICAS DE DM UTILIZADAS NA BIOLOGIA COMPUTACIONAL 25
3.1. Tecnologia de Microarranjos de DNA 25
3.2. Técnica de Classificação 28
3.3. Técnica de Agrupamento 32
3.3.1. Método de Particionamento 33
3.3.2. Método Hierárquico 35
3.4. BRB-ArrayTools 39
4 ARQUITETURA DO SISTEMA PARA KDD GEOGRÁFICO 41
4.1. Levantamento do Fluxo de Trabalho do KDD 41
4.2. Modelagem do Sistema ERB-ArrayTools 43
4.2.1. O Subsistema ProduceNetCDF 49
4.2.2. O Subsistema ChronosGIS 56
4.2.3. O Subsistema StatisticalAnalysis 62
4.2.4. Os Subsistemas ToolsUI e IDV 68
4.3. Estrutura de Arquivos do Projeto 69
4.4. Implantação do ERB-ArrayTools 73

5 TUTORIAL DO SISTEMA ERB-ARRAYTOOLS 75
5.1. Instalação do ERB-ArrayTools 75
5.2. Funcionalidades do ERB-ArrayTools..... 78
5.2.1. Pré-processamento 80
5.2.2. Mineração de Dados 85
5.2.3. Pós-processamento 87
6 ESTUDO DE CASO 91
6.1. Objetivos, Hipóteses e Planejamento do Estudo de Caso 91
6.2. Delimitação do Conjunto de Dados para Estudo de Caso..... 92
6.3. Visualização do Conjunto de Dados de Entrada 94
6.4. Resultados Sumários do Processo de KDD 96
6.5. Avaliação do Estudo de Caso utilizando o ERB-ArrayTools..... 101
7 CONSIDERAÇÕES FINAIS 103
7.1. Trabalhos Relacionados..... 103
7.2. Conclusões..... 105
7.3. Trabalhos Futuros 106
7.4. Sumário das principais contribuições 107
REFERÊNCIAS BIBLIOGRÁFICAS 109
APÊNDICE A 115

1 INTRODUÇÃO

Devido à crescente preocupação mundial com as questões ambientais, as temperaturas médias globais do ar e dos oceanos estão sendo constantemente monitoradas. Com este monitoramento, é possível identificar evidências do aumento da temperatura do sistema climático global resultando em um volume de dados cada vez maior.

Nas subseções seguintes, será apresentado: (1) a contextualização da abordagem utilizada; (2) a motivação deste trabalho ao utilizar o processo de Descoberta de Conhecimento em Banco de Dados Geográficos; (3) os principais objetivos alcançados; (4) a solução proposta para a realização do processo de KDD e o estudo de caso aplicado; e (5) finalmente a organização esta dissertação.

1.1. Contextualização

A previsão climática é realizada por diversas instituições como o Centro de Previsão de Tempo e Clima (CPTEC) do INPE no Brasil. Normalmente as instituições empregam modelos numéricos para descrever o comportamento das condições físico-químicas da atmosfera, que são executados em grandes sistemas computacionais gerando os possíveis estados futuros da atmosfera. Estas atividades apoiam estudos de impactos e vulnerabilidades do clima e permitem antecipar cenários de extremos climáticos do estado atmosférico.

A Administração Nacional do Oceano e da Atmosfera (NOAA) é uma agência de grande importância para as áreas da meteorologia e do clima. A NOAA foi criada especificamente para o estudo e observação da atmosfera e dos recursos naturais disponibilizando dados e informações para previsões meteorológicas, avisos de tempestades severas e monitoramento do clima via web (NOAA, 2011). Os dados científicos disponibilizados são armazenados em bancos de dados, cada vez maiores, de centros de pesquisa e serviços operacionais em formatos específicos, como NetCDF, HDF, HDFEOS entre

outros. Estes formatos não são exclusivos para meteorologia, sendo utilizados em diversas áreas científicas.

Com o aquecimento do planeta, fenômenos climáticos e meteorológicos extremos como as secas, inundações, tempestades severas, ventanias e incêndios florestais tornam-se cada vez mais frequentes (IPCC, 2007). Estas mudanças climáticas envolvem um dinamismo mais complexo do que a simples elevação da média térmica, por exemplo. Por este motivo, tipicamente estes fenômenos possuem um comportamento de reação em cadeia que deve ser avaliado em profundidade através da interpretação dos dados quantitativos e qualitativos com impacto em perdas e prejuízos sócio-ambiental. Esta interpretação requer técnicas eficientes que buscam transformar o “dilúvio de dados” armazenados de diversas séries históricas em conhecimento para posterior análise.

Para a transformação destes dados em conhecimento, foram desenvolvidos os sistemas geo-espaciais chamados Sistemas de Informação Geográfica (SIG). Estes sistemas são capazes de armazenar, manipular e analisar dados geográficos, ou seja, objetos e fenômenos do mundo real em que a localização geográfica é uma característica inerente e indispensável para tratá-los (CÂMARA, 1996).

Um ferramental que facilite estudos de mudanças climáticas é essencial pela diversidade de operações e formatos de dados envolvidos que usualmente necessitam de uma fase de pré-processamento geralmente tediosa. O desenvolvimento de softwares neste contexto deve gerar resultados que satisfaçam as necessidades de usuários deste domínio.

Este trabalho apresenta uma ferramenta de automação para o processo de Descoberta de Conhecimento em Banco de Dados, do inglês Knowledge Discovery in Databases (KDD) (BOTIA et al., 2002). Para o processo de Mineração de Dados, do inglês *Data Mining* (DM), o subsistema chamado *StatisticalAnalysis* conecta ao Matlab através da biblioteca MatlabControl

(2012) para execução de scripts que contém funções responsáveis pela aplicação do método estatístico t-teste e geração das visualizações dos resultados obtidos.

1.2. Motivação

O processo de KDD permite a descoberta da informação útil e implícita em grandes bases de dados. Este processo é dividido em várias etapas que inclui a gestão de algoritmos de mineração de dados utilizados para extrair e interpretar padrões dos dados.

As ferramentas de KDD utilizam diversos algoritmos para identificar relacionamentos e padrões que estão implícitos nos dados. Estes representam o conhecimento acerca da Base de Dados (BD) em análise e das entidades nela contidas. Uma das etapas que necessita da participação do usuário é a decisão sobre se os padrões encontrados refletem ou não o conhecimento útil (BOGORNY, 2003).

A Mineração de Dados é uma das etapas deste processo de KDD. Existem várias técnicas que são aplicadas nesta etapa, dentre elas podemos citar o agrupamento, a classificação e as regras de associação. A identificação de agrupamentos tem sido a motivação de muitas pesquisas na área.

Até o momento, os grandes processos na área de KDD restringem-se quase que exclusivamente à busca pelo conhecimento em dados armazenados em base de dados relacionais (KOPERSKI et al., 1997). Entretanto, em várias bases de dados organizacionais existe uma dimensão espacial cuja semântica não é interpretada utilizando os algoritmos tradicionais de mineração de dados.

A descoberta de conhecimento em Bases de Dados Geográficos (BDG) está ligada à extração de características e padrões espaciais interessantes, à identificação de relacionamentos entre dados espaciais e não-espaciais, às

restrições entre objetos geográficos e outras características que não estão explicitamente armazenadas nesses bancos de dados (BOGORNY, 2003).

As características espaciais dos dados em análise envolvem operações mais complexas e demoradas, mas que podem ser de grande importância para o processo de KDD. Esta análise requer a utilização de técnicas específicas, que permitam a inclusão da semântica espacial.

Os algoritmos de mineração de dados baseiam-se nas técnicas existentes, mas são um pouco diferentes dos tradicionais, sendo capazes de incluir a semântica espacial no processo de KDD, ou, na integração dos SIG com ferramentas de KDD permitindo a manipulação dos dados espaciais e não-espaciais (KOPERSKI et al., 1996). A principal diferença entre os KDD's convencionais e os KDD's geográficos está nos relacionamentos espaciais existentes entre as entidades do mundo real (NEVES et al., 2001).

1.3. Objetivos

Este trabalho objetiva contribuir para a automação do processo de descoberta de conhecimento resultando na diminuição do tempo entre a entrada dos dados, pré-processamento, mineração e pós-processamento se comparado a este mesmo processo manualmente realizado como em Ruivo (2008).

Para isso, foi desenvolvida uma ferramenta *stand-alone* que lida com cenários de “dilúvio de dados”. Com isso, foi totalmente eliminado a ferramenta *Excel* e seu *plug-in* BRB-ArrayTools do processo KDD, os quais foram muito utilizados por Ruivo (2008) e Ruivo (2013).

A ferramenta desenvolvida é composta por diferentes subsistemas responsáveis pela execução das fases do processo de KDD. O subsistema responsável pela DM realiza a conexão com Matlab para execução de *scripts* desenvolvidos a partir da evolução do código legado de um projeto

descontinuado chamado *Array Statistical Analysis System* (ASAS) (CARVALHO et al., 2011).

Como *benchmark*s serão utilizados, os dados da grande seca de 2005 na Amazônia. A estrutura da ferramenta foi instalada no sistema operacional Windows7 por meio de um programa de instalação criado a partir de um *script* de linguagem *Nullsoft Install System Scriptable* (NSIS, 2013) utilizando a ferramenta HM NIS Edit (RODRIGUEZ, 2013).

1.4. Abordagem da Solução

Neste trabalho, foi desenvolvido um aplicativo em Java, chamado ERB-ArrayTools, como protótipo capaz de realizar as fases de pré-processamento, mineração e pós-processamento do processo de KDD para os estudos da meteorologia e do clima.

Para isso, foi desenvolvido um subsistema, aqui denominado *ProduceNetCDF*, responsável pela automatização da fase de pré-processamento de dados georreferenciados obtidos da NOAA no formato NetCDF, do inglês *Network Common Data Form*. Este subsistema também permite o empacotamento dos resultados obtidos pelo sistema ERB-ArrayTools neste mesmo formato.

Geralmente, as atividades desta fase consomem tempo e demandas tediosas no processamento e manuseio manual dos dados de séries temporais geo-referenciados. Por isso, foi desenvolvido um subsistema chamado *ChronosGIS* (ALMEIDA et al., 2011) responsável pela carga dos dados pré-processados no Banco de Dados Geográfico (DBG).

Para a mineração e pós-processamento foi desenvolvido um subsistema em java chamado *StatisticalAnalysis* para executar códigos Matlab a fim de obter os resultados estatísticos tanto em forma de tabela quanto em forma gráfica. Para a visualização destes resultados também foram adaptados e integrados

as interfaces das ferramentas ToolsUI e o *Integrated Data Viewer* (IDV) visando obter dois novos subsistemas para visualização de arquivos NetCDF's.

Como estudo de caso, o subsistema *StatisticalAnalysis* realizou a mineração sobre o conjunto de dados globais de reanálise referente ao período da seca ocorrida em 2005 na Amazônia, o qual foi fornecido pelo CPTEC/INPE e utilizado no processo de análise em Ruivo (2008). Para guiar o processo de mineração de dados, foram utilizadas as séries temporais referentes a vazão do rio Madeira em Humaitá, Manicoré e do rio Amazonas em Óbidos.

1.5. Organização da Dissertação

Esta dissertação está organizada da seguinte forma: o Capítulo 2 apresenta o processo de KDD e a sua aplicação em Banco de Dados Ambientais Geo-referênciados; o Capítulo 3 apresenta algumas técnicas de Mineração de Dados utilizadas na Biologia Computacional; o Capítulo 4 trata da arquitetura do sistema para o KDD Geográfico; o Capítulo 5 apresenta um tutorial do sistema ERB-ArrayTools; o Capítulo 6 apresenta um estudo de caso relatando o resultado encontrado; e o Capítulo 7 apresenta os trabalhos relacionados, as conclusões, os trabalhos futuros e, em seguida, o sumário das principais contribuições finalizando esta dissertação.

2 DESCOBERTA DE CONHECIMENTO EM BASE DE DADOS

A necessidade de recolher e armazenar grandes quantidades de dados, de diversos tipos, formatos e origem, resultou na incapacidade humana de analisar, sintetizar e extrair o conhecimento a partir desses dados. Surgiram então, ferramentas de banco de dados utilizadas para armazenamento e utilização de destes dados, mas a compreensão e análise dos mesmos requer a utilização de ferramentas apropriadas que automatizem o processo de análise dos dados e descoberta de conhecimento (FAYYAD et al., 1996).

A Descoberta de Conhecimento em Banco de Dados, do inglês *Knowledge Discovery in Data Bases* (KDD) é o processo total de descobrir informação implícita e útil em grandes bases de dados, enfatizando a aplicação de alto-nível de técnicas para mineração de dados (BOTÍA et al., 2002). Os fundamentos de diversas áreas são princípios associados ao KDD, das quais pode ser citada a inteligência artificial, a aprendizagem automática, o reconhecimento de padrões e a estatística. O objetivo da integração das teorias, métodos e algoritmos destas diferentes áreas é a extração do conhecimento a partir de grandes bases de dados.

Para a busca de padrões nos dados, são utilizados algoritmos denominados algoritmos de Mineração de Dados, do inglês *Data Mining* (DM). Dentre as várias etapas do processo KDD as que mais se destacam são a gestão dos algoritmos de DM e a interpretação dos padrões encontrados pelos mesmos. A interpretação dos padrões pelo especialista humano ou sistema especialista com conjunto de regras pode dar suporte à tomada de decisão. A seguir são apresentadas as principais etapas do processo KDD.

2.1. Etapas do Processo de Descoberta de Conhecimento

Conforme mencionado anteriormente, o principal objetivo do processo de KDD é extração de regras e informações implícitas de grandes bases de dados (HAN et al., 2001). Para isso, este processo é composto por três fases,

mostrado na Figura 2.1: (1) Pré-Processamento que contempla a limpeza, integração, a seleção e a transformação; (2) Mineração de Dados, que contempla os algoritmos de mineração e avaliação dos dados para o descobrimento de padrões; e (3) Pós-Processamento que contempla os padrões, a visualização e interpretação do conhecimento descoberto.

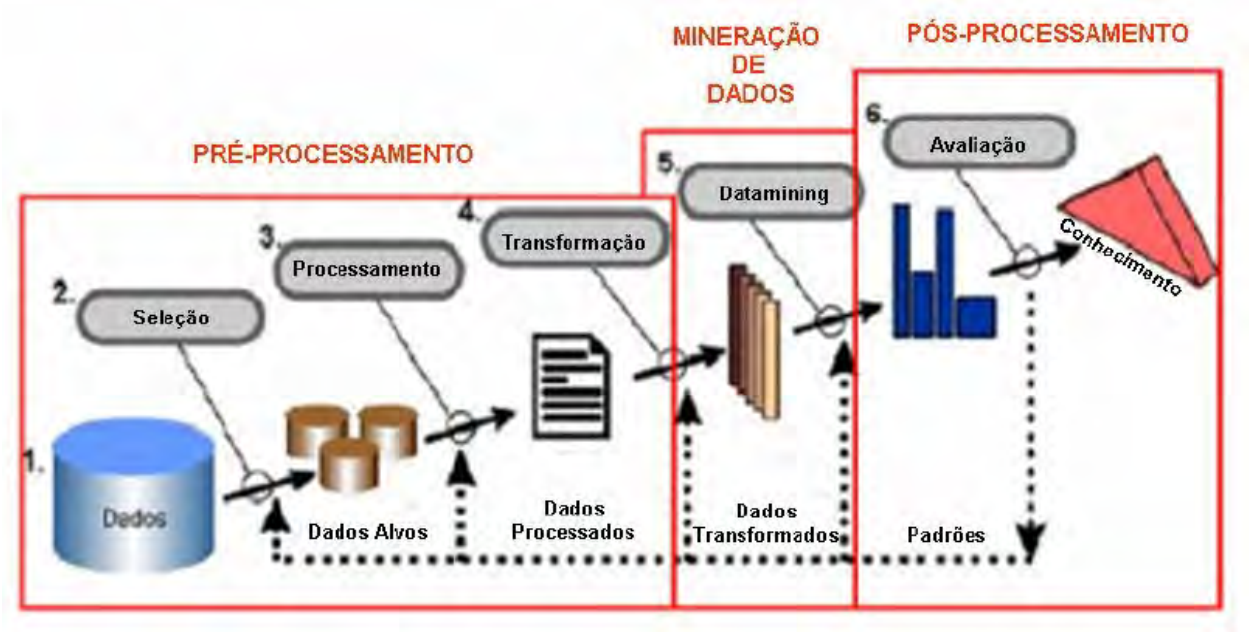


Figura 2.1 - KDD e suas três etapas essenciais
Fonte: Alves (2005)

Para a execução do processo KDD, é necessária a intervenção do engenheiro do conhecimento para a execução dos procedimentos como a seleção e a limpeza dos dados até que sejam obtidos dados expressivos, além de envolver numerosos passos com várias decisões a serem tomadas pelo especialista e pelo engenheiro (ALVES, 2005). Cada uma destas etapas deve ser executada corretamente, pois elas são fundamentais para que sejam alcançados os objetivos estabelecidos.

a) Definição e entendimento do Problema

O processo de KDD inicia a partir da definição e do entendimento do problema que se deseja resolver fazendo uma análise das atividades, a fim de atingir os

objetivos propostos. O problema deve ser definido pelo engenheiro do conhecimento em parceria com o especialista da área para chegar aos resultados positivos e úteis.

b) Pré-processamento

Na fase de pré-processamento são utilizadas as funções responsáveis pela captação, organização e tratamento dos dados. O objetivo desta etapa é a disponibilização de uma base de dados integra, consolidada e coerente para os algoritmos da fase seguinte, a mineração de dados. Caso seja necessário, poderão ser removidos os ruídos, coletadas as informações para a modelagem, definidas as estratégias para o manuseio e o tratamento dos campos que não influenciam na solução das perguntas que se deseja responder.

Nesta fase, também é definido a estratégia para resolver o problema da ausência ou a não disponibilidade de dados. O procedimento a ser realizado pode variar de uma simples migração que inclui as conversões de tipos e formatos de dados dependendo da complexidade, das fontes e do repositório destino dos dados.

Após a extração dos repositórios, os dados em geral não estão prontos para a mineração. Os dados precisam ser formatados durante e depois do processo de extração. Segundo Bogorny (2003), o pré-processamento é uma das fases mais demoradas do processo de KDD podendo consumir cerca de 80% dos esforços necessários para concluir todo o processo. As tarefas mais realizadas são: (1) determinação dos objetivos da descoberta definindo o problema claramente; (2) limpeza dos dados eliminando os ruídos e a inconsistência; (3) integração dos dados podendo combiná-los a partir de múltiplas fontes; (4) seleção de dados gerando uma amostra do banco de dados a partir dos dados relevantes identificados e agrupados para mineração de dados; e (5) transformação de dados realizando a conversão para um formato interpretável pelas ferramentas de mineração de dados.

c) Mineração de Dados

A DM é a principal etapa do processo de KDD, sendo responsável pela busca efetiva por conhecimentos úteis. Nesta etapa, aplicam-se técnicas e algoritmos em busca do conhecimento sobre os dados do problema. Dentre as principais técnicas existentes citam-se: Redes Neurais, Algoritmos Genéticos, Modelos Estatísticos e probabilísticos (GOLDSCHMIDT, 2005).

A escolha da técnica de DM depende do tipo de tarefa a ser realizada no KDD, como por exemplo, a Descoberta de Associação, a Classificação, a Regressão, a Clusterização, a Sumarização, a Detecção de Desvios, a Descoberta de Sequências e a Previsão de Séries Temporais.

O objetivo da técnica de DM é a realização da análise dos dados para encontrar padrões e regularidades no conjunto de dados. A ideia é encontrar "ouro" em lugares inesperados, ou seja, o software de DM busca extrair padrões que ainda não foram encontrados ou que são tão óbvios que ninguém os notou antes (NORVIG, 1995).

A DM tem a finalidade de ajustar modelos ou determinar padrões a partir dos dados observados, além de selecionar, explorar e modelar grandes conjuntos de dados para encontrar padrões de comportamento (TZAFESTAS, 1989). Com o ajuste dos padrões, a DM encontra o conhecimento inferido, tornando-se uma poderosa ferramenta de auxílio para tomada de decisão.

O processo de análise inicia com um conjunto de dados menores como amostra usando uma metodologia para criar uma representação da estrutura dos dados durante o tempo em que o conhecimento é adquirido. Adquirindo o conhecimento inicial, o processo é estendido a um conjunto de dados maiores trabalhando na suposição de que o conjunto de dados maior tem uma estrutura semelhante aos dados da amostra. A inspiração do processo está na analogia de uma operação de mineração, onde grandes quantidades de materiais de baixo valor são peneirados para achar algo de real valor (REZEDE, 2003).

Para verificar se os resultados obtidos correspondem aos dados originais, deve ser realizado um processo de interpretação que pode ser realizada por árvores indutivas e modelos de regras (IAN; WITTEN, 2000). Para o agrupamento dos resultados pode ser utilizado gráficos e/ou tabelas. A avaliação é realizada através da validação estatística e de teste significativo junto ao especialista para avaliar a relevância dos resultados (KAMBER, 2001). Na maioria dos casos, a validação é feita através da observação e da experimentação.

O classificador é responsável por fornecer uma determinada decisão para um dado problema. Normalmente, diferentes classificadores têm precisões diferentes para o mesmo conjunto de dados. Tradicionalmente o melhor classificador para uma aplicação é aquele que fornece a maior precisão ou menor erro para uma aplicação. Porém, esta escolha pode não ser adequada, pois informações valiosas encontradas por outros classificadores podem estar sendo desprezadas quando se escolhe apenas um único classificador (SINOARA et al. 2002). Para a geração de classificadores são utilizados algoritmos de paradigma de indução de regras como, por exemplo, as Árvores de Decisão (AD), os conjuntos de regras de produção e as listas de decisão.

d) Pós-processamento

Finalmente, a etapa de pós-processamento é responsável pelo tratamento do conhecimento obtido na DM visando a utilidade do conhecimento descoberto (FAYYAD et al., 1996). As principais funções desta fase são a elaboração, organização e a simplificação do conhecimento obtido, a construção de gráficos, diagramas, relatórios demonstrativos e entre outros cruciais para uma possível tarefa de tomada de decisão.

Nesta etapa, os resultados obtidos pela fase de DM são analisados e verificados se há necessidade ou não de retornar a alguma fase anterior. Após a identificação e transformação dos padrões em conhecimento, estes serão utilizados para compreender os fenômenos observados e para apoiar as decisões humanas (ALVES, 2005).

O pós-processamento é realizado através das etapas de avaliação e apresentação dos padrões. Estas etapas são responsáveis pela identificação e análise de padrões interessantes que representam conhecimento extraído, bem como, a forma que eles serão apresentados ao usuário.

Na etapa da apresentação do conhecimento, são utilizas técnicas de visualização e representação do conhecimento para mostrar ao analista o conhecimento minerado de uma forma clara. Os conhecimentos encontrados pelo DM são consolidados em forma de relatórios demonstrativos com documentação e explicação das informações relevantes de cada etapa do processo de KDD (FREITAS, 2002).

A grande importância da visualização dos dados está em várias situações em que pode ser necessária a interação entre o ser humano e o processo KDD. Um exemplo é a análise prévia dos dados que irão fazer ou não parte do processo, onde serão feitas consultas usando ferramentas de análise ou mesmo visualização dos dados através de gráficos, ícones e figuras (CARVALHO, 1999).

A ferramenta, que será apresentada nos capítulos seguintes, foi desenvolvida utilizando subsistemas como biblioteca de funções para permitir maior interatividade do operador e do especialista com a execução iterativa do processo de KDD.

2.2. Tipos de Conhecimento

Após a execução do processo KDD, diversos tipos de conhecimento podem ser obtidos. Segundo Addrians (1997), o conhecimento encontrado pode ser classificado como:

- Conhecimento superficial: informação facilmente recuperada por uma ferramenta de consulta;
- Conhecimento multidimensional: informação que permite a análise dos

dados por uma ferramenta de processamento analítico podendo, na maioria dos casos, extrair o conhecimento rapidamente por ferramentas de consulta como, por exemplo, SQL (*Structure Query Language*).

- Conhecimento oculto: informação que pode ser encontrada por algoritmos de reconhecimento de padrões ou aprendizado de máquina. Também poderiam ser utilizada linguagem de consultas SQL para a extração deste tipo de conhecimento, a desvantagem está no consumo excessivo do tempo de processamento;
- Conhecimento profundo: informação armazenada na base de dados que só pode ser localizada se alguém informar onde ela está contida. A diferença entre o conhecimento oculto e conhecimento profundo é que o oculto poderá ser encontrado por ferramentas de KDD, já o conhecimento profundo somente poderá ser encontrado por meio de indicação de pistas que indiquem ou apontem o conhecimento implícito.

2.3. Formas de Aquisição do Conhecimento

Os algoritmos de extração de padrões formam o núcleo de um sistema KDD. Algumas formas de extração de padrões da base de dados são (FELDENS, 1997):

- Correlação: influência de alguma forma de um elemento da base de dados sobre outro elemento. O valor dessa influência pode ser previamente especificado;
- Dependência: dependência entre dois elementos. O comportamento de um elemento pode ser afetado pelas características de outro elemento. Por exemplo, para que haja um parto normal ou cesariano a pessoa deve ser do sexo feminino.
- Detecção de sequência: dependência em relação ao tempo. A

dependência acontece quando um procedimento precede o outro, ou quando ele somente pode ser repetido depois de um intervalo mínimo de tempo.

- Classificação: descrição dos conceitos e uso de algoritmo para detecção dos padrões. O algoritmo identifica a classe que o elemento pertence, já a forma de descrição de conceitos não depende do uso de algoritmos;
- Regressão linear: Mapeamento de um item de dado para uma variável de precisão com valor real pelo algoritmo. Os dados podem ser discriminados pela combinação dos atributos de entrada;
- Detecção de desvio: Procura de elementos ou ocorrências que estão fora do conjunto de dependências, sequências ou descrições de conceitos pelo algoritmo, ou seja, procura os elementos que estão fora do Padrão. As anomalias são detectadas na base de dados tornando evidentes os problemas de qualidade e fraudes.

2.4. Tipos de Aprendizado em Base de Dados

A informação a ser localizada pode não estar contida explicitamente na base de dados, podendo ser inferida. As duas principais técnicas de inferência de dados são a dedução e a indução (BOGORNY, 2003).

Na técnica de dedução, a informação é extraída da base de dados a partir da utilização de operadores dos próprios Sistemas de Gerenciamento de Banco de Dados (SGBD) resultando em descrições corretas em relação ao mundo descrito na base de dados.

Na técnica de indução a busca de padrões ou regularidades é realizada percorrendo a base de dados a fim de encontrar combinações de valores para certos atributos que compartilham características comuns. Cada regularidade

encontrada forma uma regra que prevê o valor de um atributo a partir dos valores de outros atributos.

O aprendizado indutivo trata da criação de um modelo em que os objetos e eventos similares são agrupados em classes (BOGORNY, 2003). A classe, ou rótulo do exemplo, é um atributo especial utilizado para descrever o conceito que se deseja induzir. O comportamento dos elementos de cada uma destas classes é caracterizado por um conjunto de regras criadas para cada uma delas. De acordo com Avila (1998) e Silva (2008), as técnicas de aprendizado indutivo são:

- Aprendizado supervisionado ou aprendizado de exemplos. Nesta técnica, as classes e exemplos de cada classe do sistema são fornecidos para encontrar a descrição, que se refere às propriedades comuns nos exemplos, de cada classe. Este tipo de aprendizado pode ser comparado a um professor que orienta seus alunos. Este "professor" é o conhecimento prévio dos conceitos (hipóteses) ou classes (rótulos) que estão sendo descritas pelo conjunto de exemplos de treinamento guiando o processo de aprendizado. Este conceito induzido é visto como um classificador.
- Aprendizado semi-supervisionado. Nesta técnica, uma parte dos dados usados no treinamento é classificada, já a outra parte é composta por dados não-rotulados. Este paradigma é útil em casos onde determinadas amostras do conjunto de treinamento não disponibiliza informação suficiente para a indução de uma regra-geral. Por isso, é utilizado o conjunto de teste como fonte extra de informação para solucionar o problema.
- Aprendizado não-supervisionado ou aprendizado por observação. Nesta técnica, o sistema encontra a classe dos objetos a partir das propriedades em comum entre eles. Neste tipo de aprendizado, o indutor analisa as amostras fornecidas e determina se alguns deles

podem ou não serem agrupados para a formação de grupos, do inglês, *clusters.* Após a criação dos agrupamentos, normalmente, é realizado uma análise para definir o que cada agrupamento representa no contexto do problema abordado.

Dentre os tipos de sistemas de aprendizado existentes, pode ser citado o aprendizado de máquina e o aprendizado em mineração de dados. O sistema de aprendizado de máquina utiliza informações de uma amostra de dados de treinamento cuidadosamente selecionados (AVILA, 1998). A formação desta amostra depende do tipo de técnica de aprendizado a ser utilizada. Para a técnica de aprendizado supervisionado, o sistema busca as descrições das classes definidas pelo usuário. Para a técnica de aprendizado não-supervisionado, o sistema gera um conjunto de novas classes com suas descrições.

O sistema de aprendizado em mineração de dados utiliza informações de dados em uma base inteira e não em uma amostra (HALMENSCHLAGER, 2000), o que resulta num processo mais complexo e demorado.

Apesar das semelhanças, a principal diferença entre estes dois tipos de aprendizado é que uma base de dados é construída de acordo com a necessidade da aplicação e não com as necessidades da mineração de dados (BOGORNY, 2003). O motivo está na ausência das propriedades ou atributos que simplificariam a tarefa de aprendizado em mineração de dados. Já no aprendizado de máquina isso não acontece, pois todas as informações ou atributos necessários são adicionados ao conjunto de treinamento.

2.5. Descoberta de Conhecimento em Banco de Dados Geográficos

O processo KDD também pode ser aplicado à exploração de dados georreferenciados, ou seja, dados que são associados às referências a objetos geográficos, localizações ou partes de uma divisão territorial (BOGORNY, 2003). Este processo trata da extração de padrões espaciais e características

interessantes. Exemplo de características extraídas são os relacionamentos espaciais e relacionamentos existentes entre dados espaciais e dados descritivos, a construção de uma base de conhecimento espacial e a descoberta de conhecimento não explicitamente armazenado na base de dados (KOPERSKI et al., 1997).

A principal diferença entre a análise de dados espaciais e não espaciais está no fato das entidades geográficas envolvidas poderem ser afetadas por características de entidades vizinhas. Os fatores que contribuem para que haja a influência mútua entre duas entidades são a topologia, a distância e a direção existente entre elas.

Os dados armazenados em bases de dados convencionais estão relacionados, porém são independentes. Já os dados armazenados em BDG são interdependentes por estarem geograficamente relacionados uns com aos outros. Um exemplo é a cidade de São José dos Campos que está dentro do estado de São Paulo, o qual também está dentro do Brasil.

De acordo com Koperski, Han e Adhikari (1997), a interdependência dos dados pode prejudicar o processo KDD, pois os algoritmos de DM consideram os dados de forma independente. Para solucionar este problema, os pesquisadores da área de SIG e KDD estenderam as técnicas tradicionais de DM para suportar dados espaciais.

Com a integração de dados espaciais e dados não-espaciais, a semântica associada à localização dos objetos do mundo real e a análise dessas localizações fazem com que a utilidade do conhecimento obtido no processo KDD seja largamente melhorada (SANTOS, 2001).

A Descoberta de Conhecimento em Bases de Dados Geográficas, do inglês, *Knowledge Discovery in Databases Geographical* (KDDG) é o processo KDD tradicional adaptado para a extração de padrões ou regularidades espaciais nos dados, relacionamentos existentes entre dados espaciais e dados não-

espaciais, ou outras características implícitas em BDG (BOGORNY, 2003). O papel fundamental desempenhado por este processo está na percepção das características não-espaciais associadas aos dados espaciais.

O processo KDDG pode ser aplicado em diversas áreas como: nas áreas de sensoriamento remoto, em bases de dados de imagens e em outra áreas onde são utilizados dados espaciais. As tarefas comumente incluídas ao processo KDDG são (BOGORNY, 2003):

- Descrição da distribuição espacial dos dados e sua caracterização espacial – Esta caracterização consiste na descrição das propriedades espaciais e não-espaciais comuns aos objetos em análise. Além de considerar as propriedades dos objetos alvo de estudo, também são consideradas as propriedades dos objetos vizinhos. Para se determinar o conjunto de registros (atributo e valor) e o conjunto de objetos, é verificado se a frequência relativa de incidência nesse conjunto e nos seus vizinhos é diferente da frequência relativa verificada nos demais registros do BD;
- Verificação das características não-espaciais em regiões geográficas através da análise espacial discriminante – Esta análise discriminante permite realçar padrões espaciais de dados não-espaciais através da comparação da variação dos atributos não-espaciais em diferentes regiões geográficas. Um exemplo da aplicação de uma regra discriminante é a comparação do número de mortes causadas por neoplasias em diversas regiões geográficas;
- Estabelecimento de relações entre dados espaciais e não-espaciais usando associação espacial – Esta associação espacial permite identificar tanto a relação existente entre um conjunto de objetos espaciais e não-espaciais quanto entre dois conjuntos de objetos espaciais definindo a associação ou implicação que existe entre os mesmos. Para aplicar uma regra de associação espacial, pelo menos

um predicado espacial deve ser integrado. Este predicado pode estar associado a relações do tipo direção, distância ou topologia.

2.6. Relevância dos Dados Geográficos

Segundo Faria (1998) o termo dado espacial denota qualquer tipo de dado que descreve fenômenos aos quais esteja associada alguma dimensão espacial. Os dados geográficos são compostos principalmente por atributo, localização e tempo. O componente atributo descreve as propriedades temáticas de uma entidade geográfica, como por exemplo, o nome, já o componente localização informa a localização espacial do fenômeno associado às propriedades geométricas e topológicas. Finalmente, o componente tempo descreve os períodos em que os valores daqueles dados geográficos são válidos.

O mundo real é frequentemente modelado segundo a visão do modelo de campos e modelo de objetos (FRANK; GOODCHILD, 1990) (GOODCHILD, 1992) (COUCLELIS, 1992). O modelo de campos representa o mundo como uma superfície contínua, onde os fenômenos geográficos observados variam conforme as diferentes distribuições.

Um campo representa uma função matemática cujo domínio é uma abstração da região geográfica e o contradomínio é o conjunto de valores que o campo pode tomar. A formalização de um campo dá-se pela seguinte função:

$$\phi: p \rightarrow v, \text{ onde} \tag{2.1}$$
$$p \in \{\textit{pontos que formam uma região } R\} \text{ e}$$
$$v \in \{\textit{valores que podem ser associados } R\}$$

Caso se deseje tratar o aspecto temporal do campo, basta considerar como domínio da função o conjunto de pares *(p,t)*. A função seria:

$$\phi: (p,t) \rightarrow v, \text{ onde} \tag{2.2}$$
$$t = \text{valor de tempo}$$

Um exemplo é o campo definido para representar a variação de temperatura de uma região. Este campo será modelado como uma função cujo domínio é uma abstração da região e o contradomínio é um conjunto de valores ou intervalos de valores da temperatura. A descrição da variação do fenômeno geográfico é enfatizada por esta abordagem sem se preocupar com a identificação das entidades independentes (CÂMARA et al., 1996).

Já o modelo de objetos representa o mundo como uma superfície composta por objetos identificáveis, com geometria e características próprias. Estes objetos não representam necessariamente qualquer fenômeno geográfico específico e podem inclusive representar a mesma localização geográfica.

De acordo com Câmara et al. (1996), os campos dos modelos são representados no formato matricial e os objetos geográficos no formato vetorial. O formato matricial é composto por uma matriz de células de tamanhos regulares, onde a cada célula é associado um conjunto de valores representando as características geográficas da região correspondente. O formato vetorial descreve os dados espaciais com uma combinação de formas geométricas como pontos, linhas e polígonos.

2.7. Formato de Dados NetCDF

Diversas organizações de pesquisas ambientais disponibilizam em *sites* dados geográficos no formato de séries espaço-temporais, dentre elas a Administração Nacional do Oceano e da Atmosfera (NOAA), a qual fornece dados empacotados no formato de NetCDF. Como todas as informações fornecidas pela NOAA são referenciadas pela latitude e longitude, a geoinformática desempenha um papel fundamental, uma vez que possibilita o processamento espacial para a solução proposta.

O formato de dados NetCDF é organizado em um conjunto de interfaces com funções de acesso a dados armazenados na forma de matrizes, denominadas estruturas multidimensionais para representação de dados científicos.Um

arquivo NetCDF contendo dados que definem informações para uma aplicação particular é denominado “conjunto de dados”, do inglês *dataset*, e apresenta as seguintes características (UNIDATA, 2012):

- Autodescrição: inclui informações sobre os dados ou metadados que ele contém.
- Portabilidade: independente da arquitetura e da forma de armazenar números inteiros, caracteres e números de ponto flutuante.
- Escalabilidade: um pequeno subconjunto de um grande conjunto de dados pode ser acessado de forma eficiente.
- Agregação: permite anexar os dados a estruturas de outros arquivos NetCDF sem redefinir o arquivo original.
- Concorrência: vários processos de leitura podem ter acesso simultâneo ao mesmo arquivo.
- Compatibilidade com versões anteriores: o acesso a todas as versões anteriores dos dados NetCDF será provido pelas versões atuais e futuras do software.

Para um melhor armazenamento e agrupamento dos dados, o arquivo NetCDF pode ser organizado em: dimensões, variáveis e atributos. Um nome e um número de identificação pode ser atribuído aos arquivos para tornar possível a identificação das relações entre dados e atribuir significado aos campos de dados existentes no *dataset* (UNIDATA, 2012).

As dimensões do NetCDF correspondem aos tamanhos dos vetores e matrizes dos dados armazenados e podem ser de dois tipos: estáticas ou dinâmicas. Dimensões estáticas são os dados nos quais não haverá mais anexação, ou seja, o volume de dados não vai crescer ao longo do tempo, como por exemplo, os vetores de nível, latitude e longitude. No caso de variáveis dinâmicas, o volume de dados pode crescer ao longo do tempo de maneira indiscriminada, como por exemplo, o vetor de tempo.

Um arquivo NetCDF é composto por uma determinada variável georreferenciada. Esta variável é a estrutura, em forma de matriz, responsável pelo armazenamento dos dados geográficos. A dimensão desta matriz está relacionada com o tamanho de determinados vetores, como por exemplo os vetores de nível, latitude, longitude e tempo. Os atributos não interferem de maneira alguma nos dados e apenas agregam informações aos dados (metadados). A Figura 2.2 apresenta a estrutura de um típico arquivo de dados NetCDF.

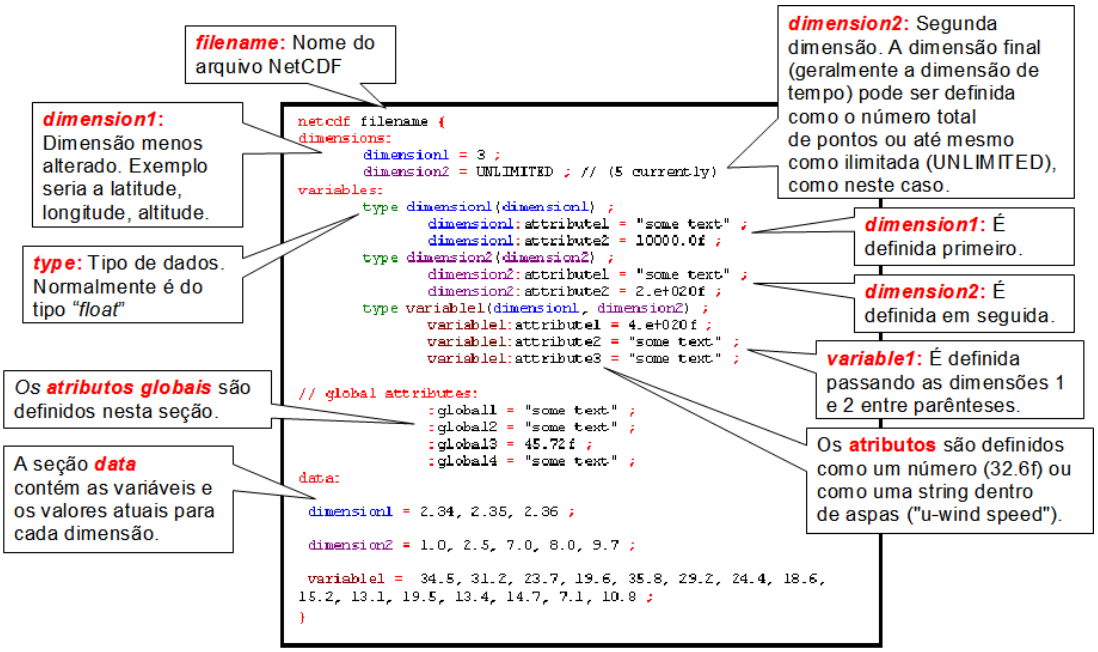


Figura 2.2 – Estrutura do Arquivo de Dados NetCDF.

Fonte: BADC (2012)

Um exemplo de variável ambiental de um arquivo NetCDF é a Temperatura do Ar (**`Air`**) na superfície da Terra com as medidas referentes ao período de janeiro de 1948 à dezembro de 2010 com uma resolução espacial de 2.5 graus em ambas as dimensões de latitude e longitude com uma resolução temporal de 1 mês. As datas são codificadas com o número de horas (ou dias) a partir de uma data no formato YYYY-MM-DD hh:mm:s.ds em diante. Apesar de seguir o mesmo formato, os arquivos NetCDF's disponibilizados não seguem o mesmo padrão, surgindo a necessidade de conversão do formato para leitura e

compreensão humana e a padronização dos dados georreferenciados para a mesma escala espaço temporal.

Os arquivos NetCDF's estão disponíveis em ESRL (2011). Os nomes destes arquivos geralmente são compostos de abreviaturas de variáveis, período de amostragem e estatística. Os períodos de amostragens incluem “dia” e “mês”. As estatísticas incluem abreviação do significado, como “ltm” que significa média a longo prazo, do inglês *Long Term Mean* e "inter.std" que significa desvio padrão interanual, do inglês *Interannual Standard Deviation*. Assim, o nome de um arquivo NetCDF pode ter a seguinte estrutura:

`(variável).(período de amostragem).(estatísticas).nc`

A ideia de produzir uma consistente reanálise dos dados atmosféricos surgiu a partir das análises operacionais que são afetadas por mudanças nos modelos, técnicas de análise, assimilação e uso de observações. Por isso, uma biblioteca foi desenvolvida por um grupo de pesquisadores do *Unidata Program Center*, em Boulder, Colorado para manipulação de arquivos no formato NetCDF (ESRL, 2011).

Esta biblioteca, seu executável e seu respectivo código fonte estão disponibilizados, no site da Unidata (2012), em diversas linguagens como C, C++, Fortran 77, Fortran 90, Java entre outras. Este trabalho utiliza esta biblioteca na implementação de subsistemas responsáveis pelo pré-processamento e pós-processamento do processo KDD.

Embora o formato de arquivo NetCDF apresente muitas vantagens como a portabilidade, escalabilidade e processos concorrentes, não é tão adequado para processamento, manipulação e consulta de dados espaciais. Uma alternativa para a solução desse tipo de problema é a exportação dos dados contidos no *dataset* do NetCDF para um banco de dados geográficos. Este processo de exportação, também foi aplicado neste trabalho como um subsistema para aumentar a produtividade do processo de KDD.

3 TÉCNICAS DE DM UTILIZADAS NA BIOLOGIA COMPUTACIONAL

A Biologia Computacional (BC) trata, dentre outras abordagens, do uso de técnicas e ferramentas computacionais para solucionar problemas da Biologia como: a comparação de sequências 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).

As principais técnicas de DM encontradas na literatura são: generalização, associação espacial, aproximação e agregação, classificação e clusterização (BOGORNY, 2003). Para aplicação destas técnicas, surgiram diversos algoritmos de DM. A maior parte destes algoritmos implementam as técnicas de classificação e agrupamento.

A principal abordagem deste trabalho é a aplicação das técnicas de classificação e agrupamento no processo de DM com inspiração na temática da BC. As próximas seções descrevem a tecnologia de microarranjos de DNA e algumas das técnicas de DM utilizadas na BC.

3.1. Tecnologia de Microarranjos de DNA

A aplicação da BC torna possíveis estudos genômicos envolvendo análise de grandes conjuntos de dados de informações derivadas de diversos experimentos biológicos. Neste sentido, surge a técnica de Microarranjos (MA) de DNA para o monitoramento dos níveis de expressão de milhares de genes simultaneamente sob uma condição particular (RUIVO, 2013).

O principal objetivo é comparar os níveis de expressão do gene entre duas amostras de tecidos como, por exemplo, uma normal e outra com tumor (HAUTANIEMI, 2003). Esta tecnologia origina a partir de um experimento que consisti de uma lâmina de vidro sobre a qual são fixadas as moléculas de DNA de forma ordenada em determinados locais, chamados “*spots*”.

Em outras palavras, define-se que MA é um conjunto de materiais genéticos fixadas em diversas posições (*spots*) de uma lâmina de vidro. Desta forma, o conjunto MA pode conter milhares de *spots* e cada um deles pode conter milhões de cópias de moléculas de material genético que correspondem a um determinado gene.

Em síntese, 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 (RUIVO, 2013). A hibridização é definida como um processo bioquímico, no qual duas fitas de ácidos nucleicos se combinam a partir de suas sequências complementares.

O processo de hibridização consiste em marcar o mRNA em solução hibridizada com o seu cDNA correspondente sendo depositado no MA a partir das regras de pareamento de bases de *Watson-Crick*. Com isso, um pareamento das moléculas complementares ocorre em cada um dos *spots* da lâmina que referencia um determinado gene. Estes *spots* contem as proporções de mRNA nas duas amostras testadas. O resultado é a intensidade de fluorecência em cada *spot* “aceso” que está relacionada à abundância do respectivo mRNA na solução (KRUTOVSKII; NEALE, 2001).

A Figura 3.1 ilustra a hibridização de um MA com duas amostras de mRNA, cada amostra está marcada com um corante fluorescente que emite luz em comprimentos de onda diferentes, geralmente as cores são verde e vermelha.

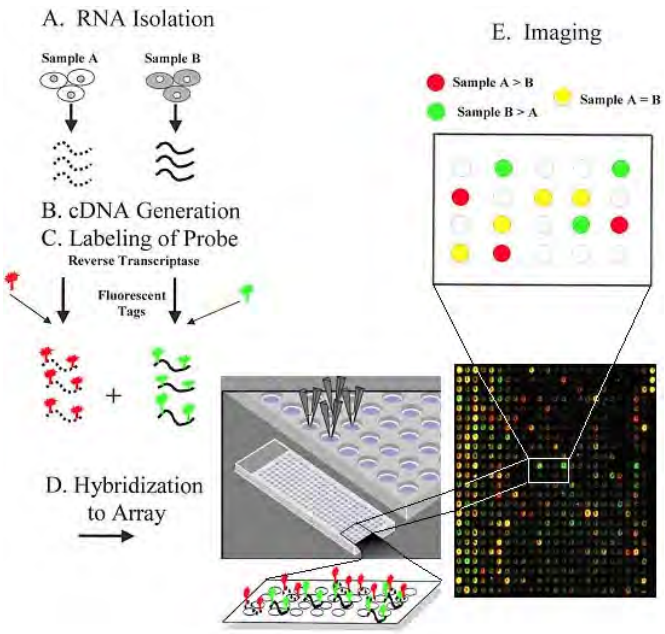


Figura 3.1 – Construção de Microarranjos.

Fonte: Ruivo (2013)

De acordo com Ruivo (2008), 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, é realizada a digitalização da lâmina.

A expressão gênica é o processo responsável pela conversão da informação localizada nos genes em proteína. A análise destes genes possibilita obter informações importantes sobre as funções de uma determinada célula. Experimentos de MA são utilizados para melhorar a classificação do diagnóstico de doenças, como do câncer, seu tratamento, e desenvolvimento de novas terapias (SOUTO et al., 2004).

Dentre as diversas maneiras de medir um determinado MA, a aplicação mais conhecida é a análise da expressão de um conjunto de genes de uma célula numa condição particular em comparação com o mesmo conjunto de genes de uma célula de referência dita normal.

A técnica popularmente utilizada é o de comparação de classes, do inglês *class comparison* (SIMON; LAM, 2006). Esta técnica consiste em identificar os genes que são diversamente expressos em determinadas classes. Um exemplo é o estudo genômico com MA's relacionado com a comparação e a classificação de níveis de expressão genética de amostra de tecidos normais e afetados por uma dada patologia como, por exemplo, um tumor.

A técnica de agrupamento, do inglês *clustering* (SIMON; LAM, 2006), também é muito utilizada em análise de conjuntos de dados de grande porte. Esta técnica pode ser aplicada em genes ou amostras com padrões similares de expressão, medidos segundo alguma métrica de correlação ou semelhança.

3.2. Técnica de Classificação

A técnica de classificação consiste em associar um dado, armazenado no BD explorado, como parte de uma determinada classe dentre vária pré-definidas ou a serem descobertas. O modelo de classificação definido pelo algoritmo de DM permite determinar em qual classe cada elemento do BD se enquadra (BOGORNY, 2003). Para cada classe está associada uma descrição correspondente. Esta descrição é o padrão único de valores dos atributos previsores.

A classificação tem como objetivo encontrar regras que dividem o conjunto de dados em grupos. Nesta técnica, um dado é caracterizado pelo comportamento de um grupo onde está inserido. Se tratando de BDG, os dados podem ser classificados não somente pelas características espaciais, mas também pelos atributos descritivos ou funções espaciais. A função do classificador não é apenas classificar, mas também explicitar o conhecimento extraído da BD de forma compreensível (BREIMAN et al., 1984).

Na comparação de classes entre grupos de amostras, é verificado se há diferenças estatísticas significativas entre atributos. O objetivo de grande importância no estudo de MA na biologia molécula é a identificação de genes

que são diferentemente expressos entre classes pré-definidas. Esta identificação com características desconhecidas pode levar a um melhor entendimento das características destes genes.

Os métodos de comparação de classes são supervisionados, pois usam a informação de que amostra faz parte de qual classe. O método estatístico mais utilizado é o t-teste (AMARATUNGA; CABRERA, 2004). Este método tem como objetivo medir diferenças entre duas classes na variabilidade de expressão do gene em unidades de variância. Para medir mais de duas classes é utilizado o F-teste. O método estatístico t-teste é calculado por:

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{S_p^2 \left(\frac{1}{J_1} + \frac{1}{J_2} \right)}}, \qquad (3.1)$$

onde

$$S_p^2 = \frac{(J_1 - 1)\, S_1^2 + (J_2 - 1)\, S_2^2}{J_1 + J_2 - 2}$$

e

$$S_i^2 = \frac{1}{J_i - 1} \sum_{j=1}^{J_i} (x_{ij} - \overline{x_i})^2$$

para

sendo:

— média das amostras da classe 1;

— média das amostras da classe 2;

quantidade de amostras da classe 1;

J_2 = quantidade de amostras da classe 2.

O método F-teste é calculado por:

$$F = \frac{[J_1(\overline{x_1} - \overline{x})^2 + J_2(\overline{x_2} - \overline{x})^2 + \ldots + J_I(\overline{x_I} - \overline{x})^2] / (I - 1)}{S_p^2}, \quad (3.2)$$

onde

$$S_p^2 = \frac{1}{J_1 + J_2 + \ldots + (J_I - I)} \sum_{i=1}^{I} \sum_{j=1}^{J_i} (x_{ij} - \overline{x_i})^2$$

e

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

O próximo passo é a obtenção do p-valor a partir da conversão do t-teste para probabilidade (RICE, 2006). Em Amaratunga e Cabrera (2004) define-se que o p-valor é a probabilidade de encontrar em hipótese nula um t-estatístico tão grande quanto o encontrado no dado real. Esta probabilidade está associada aos testes de hipóteses sendo conhecida na estatística como nível descritivo (MATTHEWS; FAREWELL, 1988).

O objetivo da hipótese na pesquisa científica é sugerir explicações para os fatos encontrados na interpretação dos dados em analise. Ao obter as hipóteses, estas devem ser analisadas para comprovar ou não através do estudo com testes estatísticos (RUIVO, 2013). Para cada teste, são criadas duas hipóteses chamadas de hipótese nula (H0) e hipótese alternativa (H1). A Hipótese nula é usada para comprovar o estudo. Já a hipótese alternativa é usada como aceitável, caso a hipótese nula seja rejeitada por não comprovar este estudo.

Os t-testes são conhecidos como “testes paramétricos”, pois tratam de uma distribuição normal que consiste no calculo da média e do desvio padrão de seus parâmetros (RUIVO, 2013). Uma aproximação de distribuição normal pode não obter bons resultados quando vários dados são muito diferentes da normalidade ou quando se busca valores muito pequenos do p-valor.

Uma solução é utilizar o método da permutação para estimar os p-valores sem uma distribuição subjacente. Este método consiste em (AMARATUNGA; CABRERA, 2004):

- Passo 1: Obtêm-se duas classes de amostras, classe 1 e classe 2:
  - ➔ J1 elementos da classe 1;
  - ➔ J2 elementos da classe 2.
- Passo 2: Calcula-se o t-teste usando a eq. 3.1.
- Passo 3: Obtém-se o t-teste.
- Passo 4: Permutam-se aleatoriamente os elementos entre classes 1 e 2:
  - ➔ Elementos da classe 1 são aleatoriamente rotulados como classe 2;
  - ➔ Elementos da classe 2 são aleatoriamente rotulados como classe 1.
- Passo 5: Calcula-se novamente o t-teste com os rótulos temporários obtidos.
- Passo 6: Obtém-se o t*.

A Figura 3.2 é um exemplo da representação do funcionamento deste algoritmo para a obtenção do p-valor em um banco de dados usando 126 permutações aleatórias.

	Classe 1					Classe 2				t-teste
Dado original	-0.18	-0.10	-0.13	0.30	-0.14	0.15	0.84	0.66	-0.52	t=3.64
	:					:				
Dado permutado 1:	-0.18	-0.10	-0.13	0.30	-0.14	0.15	0.84	0.66	-0.52	t*=3.64
Dado permutado 2:	-0.18	-0.10	-0.13	0.30	0.15	-0.14	0.84	0.66	-0.52	t*=2.15
Dado permutado 3:	-0.18	-0.10	-0.13	0.15	0.84	0.30	-0.14	0.66	-0.52	t*=0.83
Dado permutado 4:	-0.18	-0.10	-0.13	-0.14	0.15	0.30	0.84	0.66	-0.52	t*=5.48
	:					:				
	:					:				
Dado permutado 124:	0.30	-0.14	0.84	0.66	-0.52	-0.18	-0.10	-0.13	0.15	t*=2.48
Dado permutado 125:	0.30	0.15	0.84	0.66	-0.52	-0.18	-0.10	-0.13	-0.14	t*=-4.49
Dado permutado 126:	-0.14	0.15	0.84	0.66	-0.52	-0.18	-0.10	-0.13	0.30	t*=2.48

Figura 3.2 – Exemplo de Calculo de p-valor.

Fonte: Amaratunga e Cabrera (2004).

Observa-se que a cada passo foi calculado um novo t* e que foram obtidos três valores onde |t*| ≥ |t|, resultando em um p-valor de 0,0238 para o conjunto de dados (gene) em questão.

$$\text{p-valor} = \frac{3}{126} = 0.0238$$

Ao utilizar teste de hipótese, surge erro associado, chamado “erro do tipo I”. Este tipo de erro corresponde à rejeição da hipótese nula mesmo sendo verdadeira. A probabilidade de ocorre este erro é chamada de nível de significância expressa pela letra grega α. Os níveis de significância normalmente aplicados são 5%, 1% e 0,1% (AMARATUNGA; CABRERA, 2004).

A interpretação formalmente aceita, porém não muito simples para os estatísticos, é que o nível descritivo (p) representa o “menor nível de significância α que pode ser assumido para se rejeitar a hipótese nula (H0)” (RUIVO, 2013). Considerando que os pesquisadores rejeitem a hipótese nula, ao concluir que existe “significância estatística” ou que o resultado obtido é “estatisticamente significante”, o nível descritivo (p) pode ser definido, de forma generalizada, como a “probabilidade mínima de erro ao concluir que existe significância estatística”.

3.3. Técnica de Agrupamento

A técnica de agrupamento ou clusterização é utilizada para identificar grupos de dados semelhantes. O agrupamento destes dados pode ser baseado em funções de distância e ser específicas para diferentes contextos da aplicação. Um exemplo é o agrupamento de casas de uma área, definido pela sua categoria, área construída, e localização geográfica.

Tanto o agrupamento quanto a classificação, têm como objetivo realizar uma separação entre os dados de um BD de forma que permita a descoberta de novos padrões antes desconhecidos. Independente da ferramenta usada, o

resultado obtido pode ser interpretado com maior eficiência pelo especialista da área de origem dos dados (RUIVO, 2013). A aplicação do agrupamento facilita a visualização dos padrões obtidos favorecendo a análise.

A principal tarefa da clusterização é agrupar um conjunto de dados em subconjuntos, de acordo com os critérios apropriados (NEVES et al., 2001). Desta forma, os subconjuntos são compostos de elementos que têm um alto grau de semelhança ou similaridade. Já quaisquer outros elementos pertencentes a grupos distintos tenham pouca semelhança entre si.

Essa técnica tem sido muito utilizada em análise exploratória de dados espaciais e em procedimentos de regionalização devido sua habilidade de identificar estruturas diretamente dos dados sem que haja um conhecimento prévio dos mesmos (BOGORNY, 2003).

Segundo Neves, Freitas e Camara (2001), os critérios mais utilizados por esta técnica e pelo algoritmo *k-means*, tratado a seguir, são a homogeneidade e a separação. Na primeira, os dados de um mesmo grupo devem ser os mais similares possíveis. Na segunda, os dados de diferentes grupos devem ser os mais distintos possíveis.

A qualidade dos grupos, do inglês *clusters*, depende das definições estabelecidas pelo usuário como a escolha dos atributos, medidas de dissimilaridade, critérios de agrupamento, escolha do algoritmo e definição da quantidade de grupos, entre outros. A dissimilaridade, em geral, é usada para avaliar o grau de semelhança entre dois dados durante o processo de agrupamento. Na maioria das vezes, essa medida representa a distância entre dois objetos.

3.3.1. Método de Particionamento

Os métodos de particionamento têm como objetivo encontrar a melhor partição do n dados em k *clusters*. Na maioria das vezes, os k grupos encontrados por

estes métodos são de melhor qualidade em comparação com os produzidos pelos métodos hierárquicos. Outra vantagem é o seu maior desempenho, o que tem ocasionado cada vez mais investigações e usos de algoritmos de particionamento (NG; HAN, 1994).

Um dos algoritmos de agrupamento mais utilizado e bastante difundido é o *k-means*. Este algoritmo tem como base um ponto central representado pela média dos atributos dos dados. A área de grande aplicação é em Sensoriamento Remoto, onde tem a finalidade de executar procedimento de classificação não supervisionada de imagens de satélite (BOGORNY, 2003).

Para a configuração do *k-means*, é exigido a definição prévia do número de *clusters* e do posicionamento do centro de cada *cluster* k no espaço de atributos. O centro do *cluster* é denominado Centróide, o qual representa o ponto médio mais central do *cluster*. Esse algoritmo é sensível ao ruído, porém em termos de desempenho é relativamente eficiente para grandes BD. Os principais passos para execução do algoritmo *k-means* são (BOGORNY, 2003):

- Passo 1: seleção de n dados para serem centros inicias dos k *clusters*.
- Passo 2: cada dado é associado a um *cluster*, para o qual a dissimilaridade entre o dado e o centro do *cluster* é menor que as demais.
- Passo 3: os centros dos *clusters* são recalculados, redefinindo cada um, em função dos atributos de todos os dados pertencentes ao *cluster*.
- Passo 4: retorna ao passo 2 até que os centros dos *clusters* se estabilizem.

A cada interação, os centros dos *clusters* são reavaliados no passo 3 devido o agrupamento dos dados em função do centro do *cluster* mais próximo provocando o deslocamento dos centros médios no espaço de busca. A execução é interrompida quando não existe mais deslocamento das médias ou existe uma insignificante realocação de dados entre os *clusters*.

O objetivo do algoritmo *k-means* é minimizar a distância dos elementos para obter um conjunto de forma iterativa (RUIVO, 2013). Um exemplo de *k-means* em modelos matemáticos é a aplicação em Grafos representando relações entre objetos. Segundo Linden (2009) um grafo é um conjunto representado

por G = (V, E), onde V é um conjunto finito de pontos, geralmente chamados de nós ou vértices, e E é uma relação entre vértices, ou seja, um conjunto de pares em VxV.

3.3.2. Método Hierárquico

O método hierárquico consiste em decompor o BD na forma de árvore dividindo-a recursivamente em grupos de dados menores. As duas formas de decomposição são os divisivos e os aglomerativos (NEVES et al., 2001). O primeiro realizado no sentido *top-down* e o segundo no sentido *botton-up* (HAN et al., 2001).

No sentido *top-down*, a decomposição inicia a partir de todos os dados no mesmo *cluster* que vai sendo dividido sucessivamente até que cada cluster obtido contenha apenas um único dado. No sentido inverso, o *botton-up*, a aglomeração inicia a partir de vários *clusters,* dos quais cada *cluster* é um dado, que vai sendo fundido os dois *clusters* mais próximos (similares), a cada passo do procedimento, até obter somente um grande *cluster* contendo todos os dados.

O processo *botton-up* é chamado hierárquico, pois admiti obter vários níveis de agrupamento. Os passos para execução do algoritmo hierárquico aglomerativo são (BOGORNY, 2003):

- Passo 1: iniciar com n *clusters*, cada um contendo um dado.
- Passo 2: calcular a dissimilaridade entre os dados.
- Passo 3: procurar o par de *clusters* com menor dissimilaridade.
- Passo 4: recalcular a dissimilaridade do *cluster* fundido com os demais *clusters*.
- Passo 5: repetir os passo 3 e 4, n-1 vezes.

Para o cálculo da dissimilaridade entre *clusters*, é utilizado um ponto central que representa o *cluster* definido pelos valores médios dos atributos dos dados membros de cada grupo. A dissimilaridade entre dois *clusters* é igual a menor dissimilaridade existente entre dois dados quaisquer, considerando que estes dados pertencem aos *clusters* envolvidos (BOGORNY, 2003). A desvantagem

na aplicação deste método é a produção de *clusters* de forma análoga criando o efeito de encadeamento.

Como mencionado anteriormente, a técnica de agrupamento aplicado neste trabalho permite escolher a métrica de distância entre os grupos, a qual envolve os agrupamentos (RUIVO, 2013):

- Ligação Média (*Average Linkage*) – defini a distância entre dois *clusters* como a média das distâncias entre todos os pares de elementos, um do primeiro *cluster* e um do segundo.
- Ligação Completa (*Complete Linkage*) – defini a distância entre dois *clusters* como a distância máxima entre um elemento do primeiro conjunto e um elemento no segundo.
- Ligação Unica (*Single Linkage*) – defini a distância entre os dois *clusters* como a distância mínima entre um elemento do primeiro conjunto e um elemento no segundo.

A Figura 3.3 é uma representação desta métrica definindo a distância entre perfis de expressão de duas amostras.

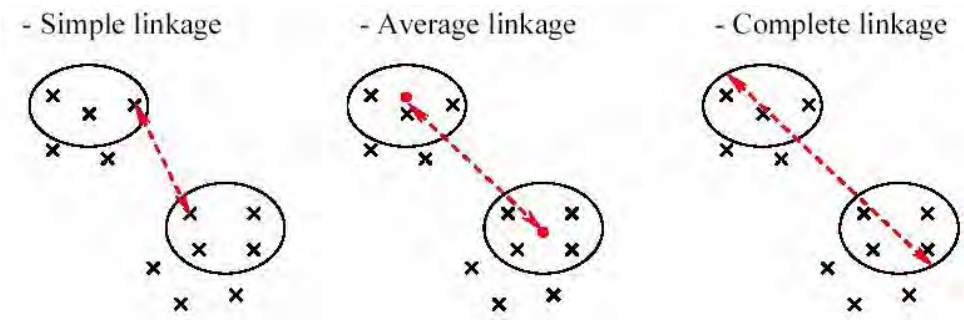


Figura 3.3 – Métricas de distância entre os grupos.
Fonte: Ruivo (2013)

Segundo Ruivo (2013), o agrupamento hierárquico produz uma fusão sequencial aninhada das variáveis em estudo tendo como base algumas métricas de correlação ou semelhança. Dentre as métricas existentes, as mais conhecidas são a correlação de *Pearson* e a distância Euclidiana.

Para a representação da fusão aninhada, é utilizado um “dendrograma”. No nível mais baixo deste dendrograma, cada variável faz parte de um *cluster* individual. Para a construção do dendrograma, é utilizado o algoritmo hierárquico aglomerativo com a decomposição do BD no sentido *botton-up* como apresentado anteriormente. No nível mais alto, existe um conjunto contendo todas as variáveis.

A fórmula matemática para a correlação de *Pearson* entre dois experimentos X e Y é dada por (SIMON; LAM, 2006):

$$\frac{\sum_{i=1}^{N}(x_i - X_{avg})(y_i - Y_{avg})}{\left[\left(\sum_{i=1}^{N}(x_i - X_{avg})^2\right)\left(\sum_{i=1}^{N}(y_i - Y_{avg})^2\right)\right]^{1/2}}, \tag{3.3}$$

onde N é o número de variáveis, X_{avg} é a média das variáveis no experimento X, e Y_{avg} é a média das variáveis no experimento Y. A correlação de *Pearson* é uma medida se similaridade muito utilizada entre dois vetores. Para a distância Métrica, basta calcular um menos a correlação.

A distância Euclidiana entre dois vetores é dada por (SIMON; LAM, 2006):

$$\sqrt{\sum_{i=1}^{N}(x_i - y_i)^2}, \tag{3.4}$$

onde x_i são variáveis do experimento X e y_i são variáveis do experimento Y.

Diversas aplicações biológicas empregam com sucesso os métodos hierárquicos, porém não existe uma revisão dos *clusters* durante a execução do procedimento (BOGORNY, 2003). Em outras palavras, no método hierárquico aglomerativo a fusão de dois dados em um mesmo *cluster,* não permite a separação destes dados novamente, resultando na permanência no mesmo *cluster* até o final do procedimento. Seguindo a mesma analogia, no

hierárquico divisivo a separação de dois dados não permite o agrupamento no mesmo *cluster* novamente.

Podem-se associar métodos de agrupamento com técnicas de representação gráfica, a fim de obter uma boa assimilação da grande massa de dados em forma de tabelas organizadas por estes métodos. A representação gráfica resultante apresenta cada ponto com uma cor que reflete as observações do experimento original de forma quantitativa e qualitativa. Esta representação com o dado complexo é obtido através de uma ordenação estatística que permite aos especialistas uma assimilação e exploração do dado de uma forma natural e intuitiva (EISEN et al., 1998).

Como resultado do agrupamento é obtido uma imagem colorida em forma de matriz. O traçado das cores varia de vermelho a verde representando a variação da maior para a menor expressão de cada variável em cada amostra. Nas linhas horizontais, estão representadas as variáveis e nas colunas estão representadas as amostras obtidas. Os tons de vermelho representam o grau de aumento da amplitude e os tons de verde representam o grau de diminuição de amplitude (SIMON; LAM, 2006). A Figura 3.4 é exemplo deste típico clustergrama aplicado em BC.

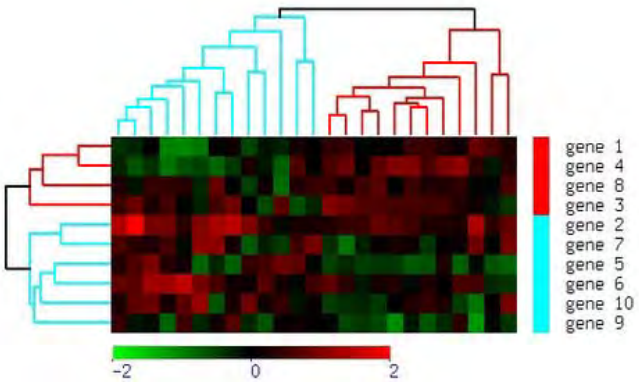


Figura 3.4 – Agrupamento hierárquico e representação por cores.
Fonte: Simon e Lam (2006)

3.4. BRB-ArrayTools

Uma das ferramentas de classificação comumente utilizadas na área da BC é o BRB-ArrayTools da Linha de Pesquisa Biométrica, do inglês *Biometric Research Branch*, da Divisão do Tratamento e Diagnóstico de Câncer do Instituto Nacional do Câncer sobre a direção do Dr. Richard Simon disponível em Simon (2013).

Esta ferramenta é um software livre desenvolvida por estatísticos experientes em análise de dados de MA de DNA para processar, visualizar, agrupar, classificar, dentre outras funções realizada sobre os dados de expressão em vários experimentos. A sua implementação foi desenvolvida no sistema de estatística R, em programas em C e Fortran e em aplicações java. Foi utilizado o programa Visual Basic para integrar os componentes e ocultar a complexidade dos métodos de análise e visualização.

Esta ferramenta funciona como um *add-in* para a sua integração com o programa Excel do sistema operacional Windows da Microssoft. Os dados a serem analisados são inseridos em planilhas no Excel, onde são descritos os valores da expressão e fornece os fenótipos que foram especificados pelo usuário para as amostras da matriz.

Assim como o BRB-Arraytools, neste trabalho foi desenvolvido um subsistema que emprega o princípio de Microarranjos de DNA para a mineração de dados georreferênciados. De forma análoga ao princípio do MA, as variáveis com suas coordenadas são os genes e os pacientes são os meses no seu respectivo ano ao aplicar o processo de classificação implementado na DM.

Mais detalhes sobre o BRB-ArrayTools e sua aplicação de forma adaptada para área ambiental podem ser encontrados em Ruivo (2008) e em Ruivo (2013). Um exemplo de agrupamento utilizando a ferramenta BRB-ArrayTools adaptado para a área ambiental inicialmente proposto por Ruivo (2008) é apresentado na Figura 3.5.

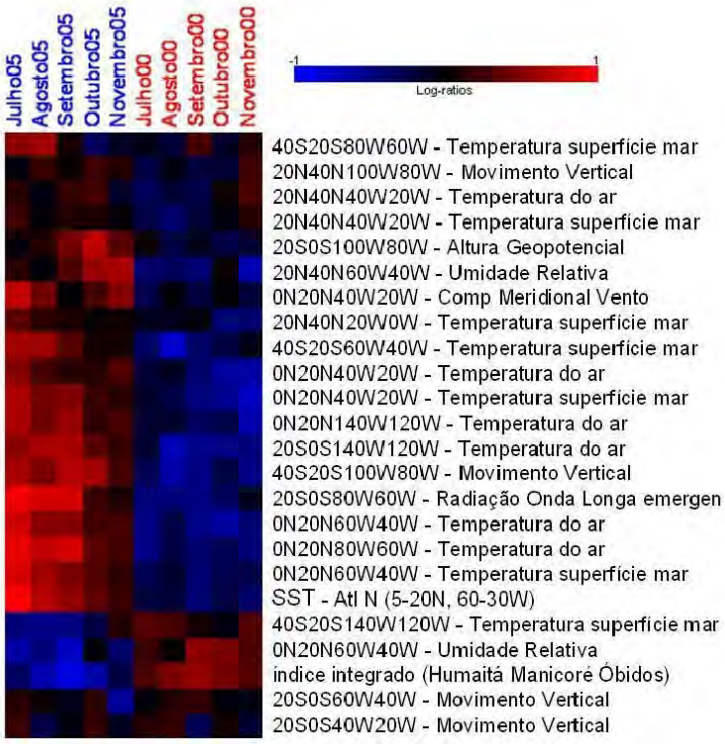


Figura 3.5 – Agrupamento usando BRB-ArrayTools na área Ambiental.

Fonte: Ruivo (2008)

4 ARQUITETURA DO SISTEMA PARA KDD GEOGRÁFICO

O grande aumento do volume de dados ambientais gerados por satélites de monitoração e redes de sensores dificulta a análise e a interpretação destes dados pela comunidade científica. Com isso, surge a necessidade do desenvolvimento de ferramentas capazes de auxiliá-los no processamento e eventual tomada de decisões para a solução de problemas ambientais.

Este capítulo descreve a concepção de um sistema, denominada ERB-ArrayTools, para automação do processo de KDD na área ambiental. Foi escolhido este nome por tratar de uma ferramenta que aplica o método estatístico t-teste para classificação e agrupamento de conjunto de dados geográficos na fase de DM, como apresentado no capítulo 3.

Os diagramas UML (2012) apresentados neste capítulo foram criadas utilizando a ferramenta Umbrello (2012) na versão 2.0, já o diagrama das tabelas do BDG, criadas pelo subsistema *ChronosGIS*, foi obtido pela ferramenta SGBD PostgreSQL.

4.1. Levantamento do Fluxo de Trabalho do KDD

O sistema foi desenvolvido mediante entrevistas com Ruivo (2008) no intuito de se identificar gargalos no processo manual de KDD e automatizar a análise de dados ambientais abordado em seu trabalho. O processo que antes despendia dias de forma manual, com o sistema obtido pode ser executado em apenas alguns minutos para a mesma massa de dados.

Os requisitos para o desenvolvimento deste sistema foram derivados do fluxo de trabalho do especialista neste domínio. O diagrama de Caso de Uso da Figura 4.1 apresenta os requisitos do sistema e enfoca dois atores: (1) O usuário que solicita os requisitos levantados; e (2) O especialista que analisa os requisitos obtidos por meio das visualizações a serem geradas pelo sistema.

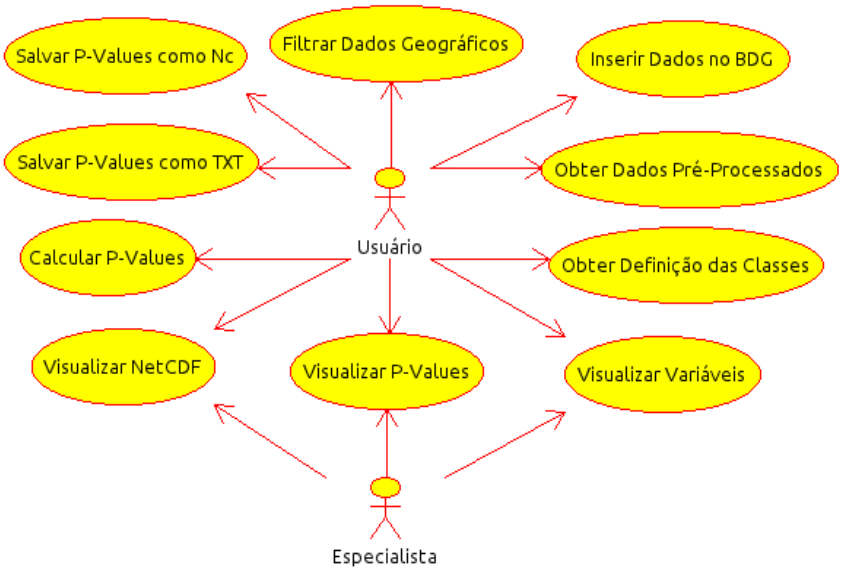


Figura 4.1 - Diagrama de Caso de Uso do ERB-ArrayTools.

Para a implementação dos requisitos solicitados, optou-se pela utilização de subsistemas como biblioteca de funções visando obter maior agilidade e flexibilidade no processo de desenvolvimento do sistema proposto. Para isso, foi necessário investigar diferentes ferramentas e bibliotecas desenvolvidas em Java na busca de funcionalidades que melhor adéquam aos requisitos.

Para a integração destas funcionalidades ao ERB-ArrayTools, foi preciso adaptar as interfaces das ferramentas encontradas como subsistemas. Além das funcionalidades obtidas, outras também tiveram que ser desenvolvidas para atender o levantamento de requisitos. As funcionalidades destes subsistemas são:

- ProduceNetCDF – Lê e produz arquivos NetCDF's; normaliza e cria anomalias sobre os dados; e define o descritor de dados com as classes a partir de uma série temporal na fase de pré-processamento.
- ChronosGIS – Cria as tabelas, caso não exista, e armazenar os dados pré-processados no BDG.
- StatisticalAnalysis – Realiza a permutação do método estatístico t-teste na fase de DM e gera as visualizações dos resultados estatísticos na fase de pós-processamento utilizando o Matlab.

- ToolsUI e IntegratedDataViewer – Visualizam na fase de pós-processamento os arquivos NetCDF gerados com o conjunto de dados pré-processados e os resultados estatísticos obtidos.

A Figura 4.2 apresenta em alto nível um esquema de integração destes quatro grandes subsistemas.

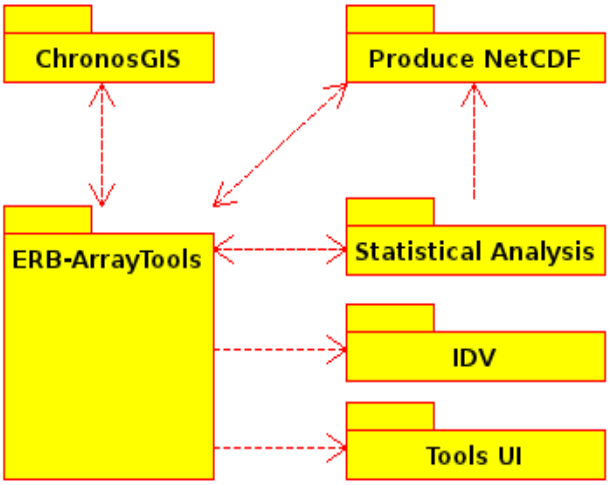


Figura 4.2 - Esquema de integração dos subsistemas.

4.2. Modelagem do Sistema *ERB-ArrayTools*

O subsistema agregador ERB-ArrayTools com a interface principal para a interação com o usuário e os demais subsistemas foram desenvolvidos em Java utilizando o ambiente de desenvolvimento NetBeans na versão 7.2 (NETBEANS, 2013).

Nas classes, cujo nome inicia com a sigla ERB, contém a implementação das telas do subsistema principal para a interação com o usuário. As demais classes estão relacionadas com as funcionalidades do sistema. A Figura 4.3 apresenta o diagrama de classe deste subsistema com algumas classes diretamente relacionados com a classe principal *ERBMainView*.

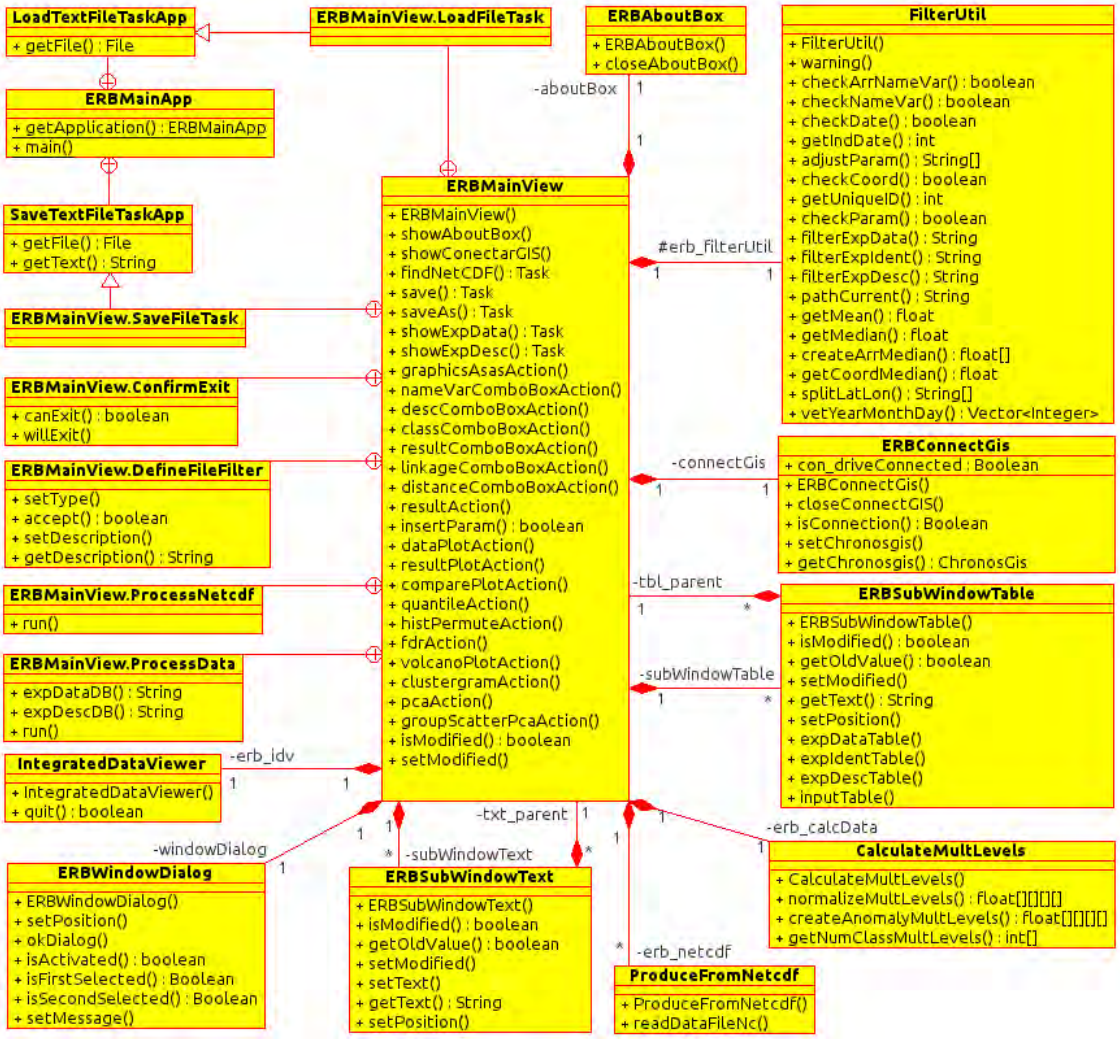


Figura 4.3 - Diagrama de Classe da tela Principal do ERB-ArrayTools.

A implementação de todo o sistema seguiu a estrutura padrão de desenvolvimento do NetBeans. As principais classes que compõem o subsistema principal são: ERBMainApp, ERBMainView, ERBConnectGIS, ERBSubWindowTable, ERBSubWindowText, ERBWindowDialog, ERBAboutBox e FilterUtil.

A classe ERBMainApp contém o método principal "*main*" para execução do sistema, a classe LoadTextFileTaskApp para abrir um arquivo e a classe SaveTextFileTaskApp para salvar um arquivo.

A classe ERBMainView contém diversas funções utilizadas para a integração dos demais subsistemas e interação com o usuário. Além destas funções também é composta por diferentes classes, as quais são:

- LoadTextFileTask – Herda da classe LoadTextFileTaskApp e permite abrir arquivo NetCDF ou arquivo texto contendo um conjunto de dados ou descritor de dados. Caso haja falha ao abrir um arquivo, uma mensagem de erro é retornada;
- SaveTextFileTask – Herda da classe SaveTextFileTaskApp e permite salvar em formato de texto diretamente ou através do subsistema StatisticalAnalysis para salvar os resultados estatísticos em formato NetCDF e em formato texto pelo Matlab. Caso haja falha ao salvar um arquivo, uma mensagem de erro é retornada;
- ProcessNetCDF – Permite realizar o processo de carga dos dados pré-processado no BDG utilizando o subsistema ChronosGIS. A barra de processos é acionada possibilitando o usuário acompanhar o processo de carga dos dados. Nesta classe são invocados os métodos do ChronosGIS responsáveis pela criação de tabelas e inserção ou atualização dos dados no BDG.
- ProcessData – Permite obter o conjunto de dados do BDG utilizando o subsistema ChronosGIS e gera o descritor de dados utilizando a classe CalculateMultLevels do subsistema ProduceNetCDF. Assim como a classe ProcessNetCDF, a barra de processos também é acionada possibilitando o usuário acompanhar o processo de obtenção do conjunto de dados ou de geração do descritor de dados.
- DefineFileFilter – filtra os arquivos a serem localizados em uma pasta pelos métodos que permitem abrir arquivo e salvar arquivo. O tipo “.txt” permite localizar apenas arquivos do tipo texto, já o tipo “.nc” permite localizar apenas arquivos do tipo NetCDF.
- ConfirmExit – verifica se há alguma modificação e confirma com o

usuário se deseja salvar antes de fechar o sistema ERB-ArrayTools.

Na classe ERBConnectGIS está implementada uma janela de diálogo que utiliza o subsistema ChronosGIS para a conexão do sistema com o SGBD PostgreSQL. Nesta janela deve ser informado a URL, o nome do usuário e a senha para a conexão com o BDG.

Na classe ERBSubWindowTable está implementada uma subjanela contendo uma tabela para visualização do conjunto de dados ou do descritor de dados.

Na classe ERBSubWindowText está implementada uma subjanela contendo uma área de texto para a visualização da estrutura do arquivo NetCDF aberto ou dos resultados estatísticos básicos obtidos por consulta SQL no BDG pelo subsistema ChronosGIS.

Na classe ERBWindowDialog está implementada uma janela de diálogo com dois “RadioButton” e um botão OK com a finalidade de interagir com o usuário para a escolha entre duas opções. Esta janela de opções é utilizada para perguntar ao usuário se deseja criar anomalia ou apenas normalizar o conjunto de dados e se deseja obter o conjunto de dados do BDG ou do arquivo texto.

Na classe ERBAboutBox está implementada uma janela de diálogo que contém informações sobre o sistema ERB-ArrayTools como o número da versão, o nome do desenvolvedor e a *Homepage*.

Na classe FilterUtil estão implementados os métodos de filtragem espaço-temporal do conjunto de dados, além dos diversos outros métodos responsáveis pela manipulação da descrição espaço-temporal deste conjunto. Os métodos desta classe são utilizados pelos subsistemas ERB-ArrayTools, ChronosGIS, ProduceNetCDF e StatisticalAnalysis.

As demais classes utilizadas pela classe principal ERBMainView, como IntegratedDataView, ProduceFormNetCDF e CalculateMultLevels, serão

apresentadas nas próximas subseções. A Figura 4.4 apresenta as classes dos subsistemas integrados à classe *ERBMainView* do sistema ERB-ArrayTools.

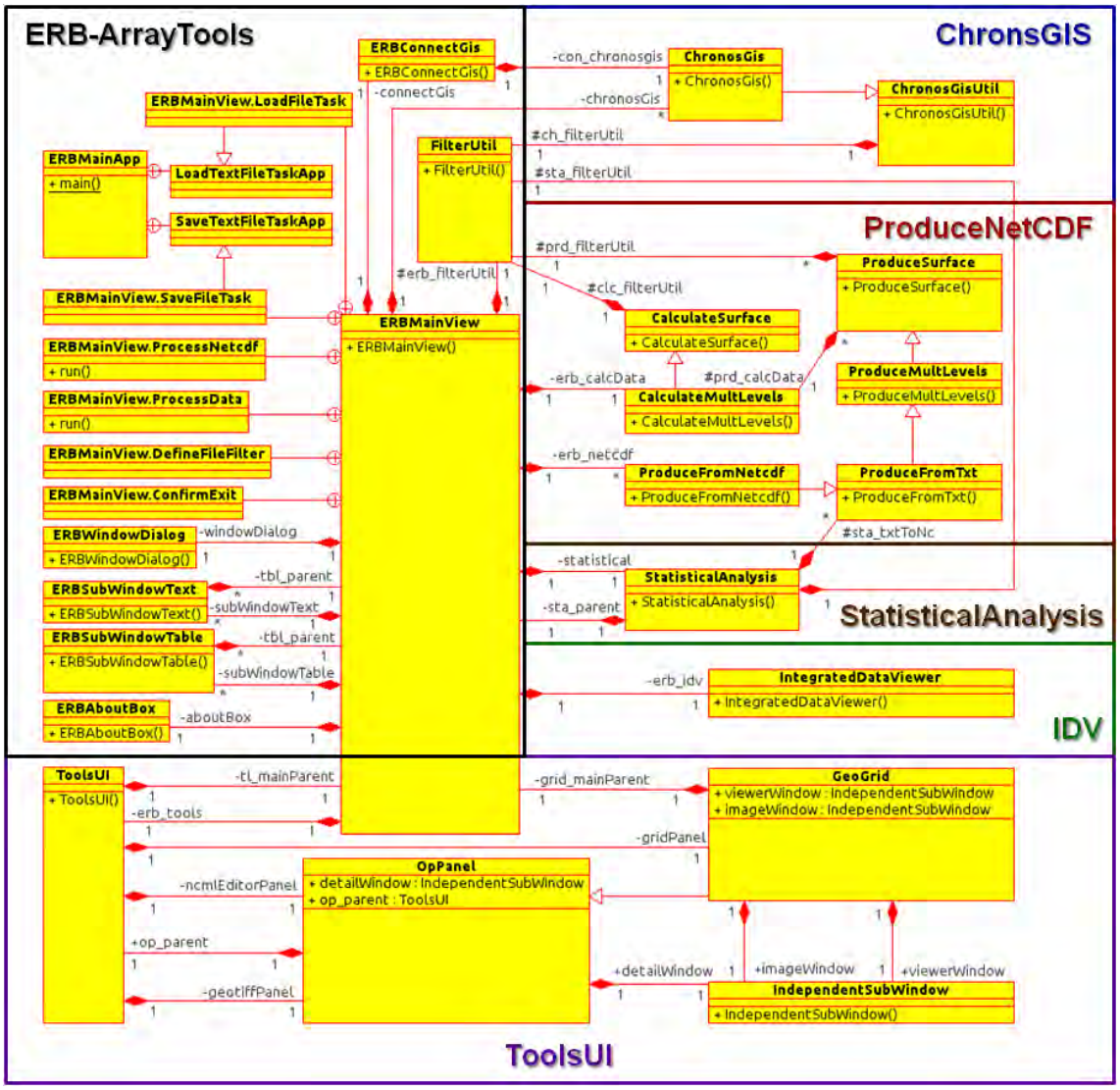


Figura 4.4 - Diagrama de Classe em alto nível dos subsistemas integrados.

Os subsistemas de visualização de arquivos NetCDF's e os *scripts* de código Matlab para visualização de gráficos estatísticos, têm suas próprias telas de forma independente, as quais facilitam a integração destas funcionalidades ao sistema principal. Nas subseções seguintes será apresentado o funcionamento de cada um dos subsistemas do ERB-ArrayTools.

A Figura 4.5 apresenta o diagrama de pacotes mostrando a cooperação destes subsistemas com o ERB-ArrayTools.

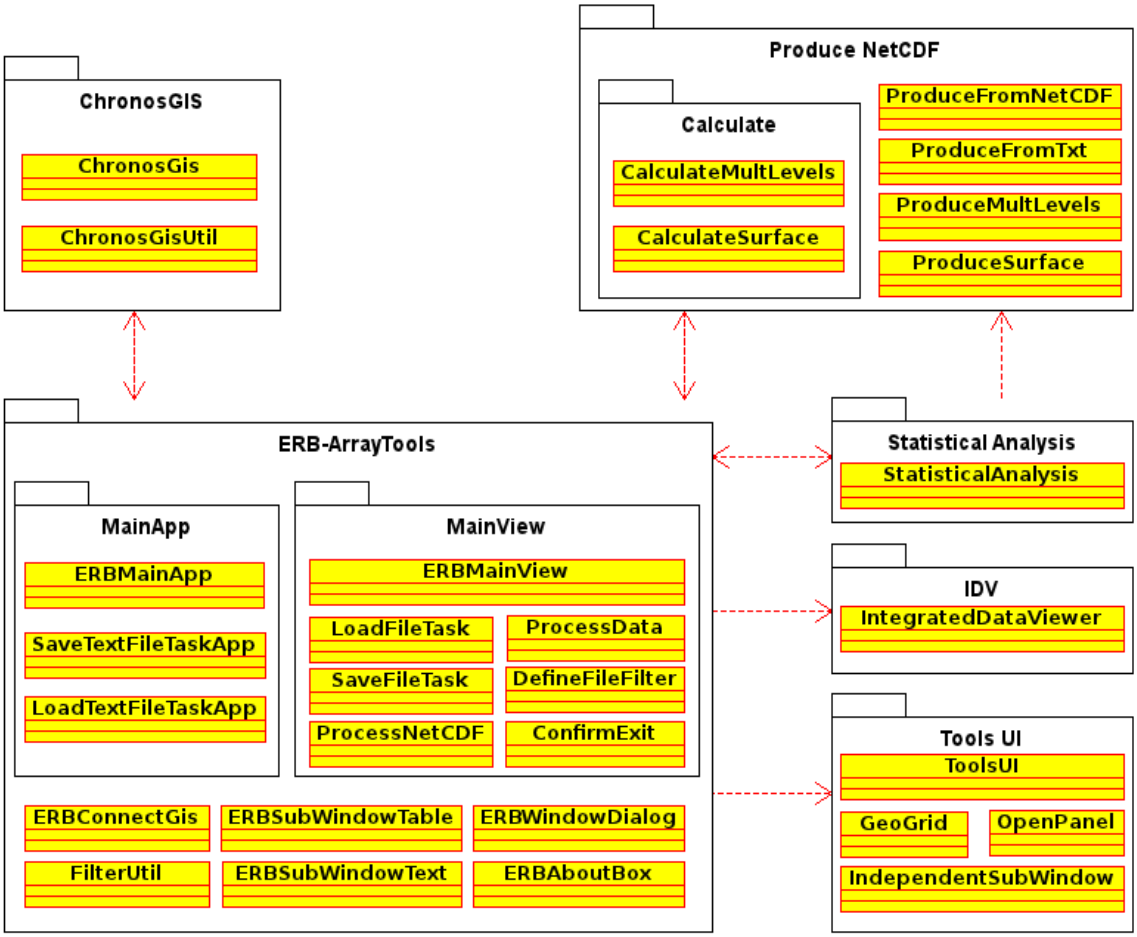


Figura 4.5 - Diagrama de Pacotes do ERB-ArrayTools.

De forma sucinta, esta cooperação consiste em:

- O subsistema ChronosGIS permite a conexão com o SGBD, criação das tabelas, caso não exista, e a inserção dos dados pré-processados no BDG;
- O subsistema ProduceNetCDF lê arquivos NetCDF's e utiliza o pacote Calculate para normalizar e criar anomalias sobre o conjunto de dados obtido. Em seguida, este conjunto é carregado no BDG pelo ERB-ArrayTools utilizando o subsistema ChronosGIS;

- O ERB-ArrayTools utiliza o subsistema ProduceNetCDF para definir o descritor de dados com as classes a partir de uma série temporal do BDG ou do arquivo texto para a realização do processo de DM pelo subsistema StatisticalAnalysis;
- O subsistema StatisticalAnalysis é invocado pelo ERB-ArrayTools para realizar a permutação do método estatístico t-teste na fase de DM, visualizar os resultados estatísticos obtidos, salvar em arquivo texto a tabela de resultados pelo Matlab e produzir novos arquivos NetCDF's contendo estes resultados e a variável ambiental pré-processada através da funcionalidade ProduceFromTxt do subsistema ProduceNetCDF;
- Os subsistemas ToolsUI e IDV são invocados pelo ERB-ArrayTools para a visualização dos novos arquivos NetCDF's sobre o mapa. O subsistema ToolsUI é utilizado para visualização em forma de grade. O subsistema IDV disponibiliza diversos tipos de visualização, porém neste trabalho está focado no uso deste subsistema para visualização de arquivos NetCDF's em 3D.

4.2.1. O Subsistema *ProduceNetCDF*

Como mencionado anteriormente, o subsistema *ProduceNetCDF* foi desenvolvido utilizando a biblioteca NetCDF-Java para a realização do pré-processamento de um conjunto de dados geográficos. Para isso, o sistema permite realizar a filtragem dos dados brutos de um arquivo NetCDF conforme os seguintes parâmetros: (1) o nome da variável juntamente com seu nível, caso seja de multinível; (2) as datas inicial e final de sua série temporal; (3) o passo para a definição de sua granularidade espacial; e (4) as coordenadas (latitude e longitude) inicial e final.

Nesta fase de pré-processamento, surge a necessidade de padronizar as dimensões de diferentes variáveis ambientais dos *dataset* obtidos do NetCDF

original. Ao carregar os dados, o sistema verifica como o tempo e as coordenadas estão definidos no arquivo NetCDF. O padrão adotado para o tempo é em número de horas. Caso não esteja, o tempo é ajustado para o seu armazenamento no NetCDF pré-processado. Para o armazenamento no BDG, este tempo é convertido para dia, mês e ano. Porém, este sistema foi implementado para o carregamento de dados mensais. Com isso, a informação do tempo referente ao dia sempre é fixado no dia primeiro.

O padrão adotado para todas as coordenadas geográficas é a variação de 2.5 em 2.5 graus. Caso não esteja neste padrão, as coordenadas são ajustadas juntamente com o dimensionamento de sua matriz de dados aplicando a média aritmética, como exemplificado na Tabela 4.1.

Tabela 4.1 - Ajuste da Matriz de Dados.

Dados de Janeiro de 1999				
Vetor de Passos	Coordenada Original	Dados de SST	Coordada Desejada	Dados Ajustados
2	0.5W 41.5N	-0,10173	0.0W 40.0N	-0,01313
	1.5W 40.5N	0,07547		
3	2.5W 39.5N	0,00356	2.5W 37.5N	-0,00988
	3.5W 38.5N	-0,02616		
	4.5W 37.5N	-0,00704		
2	5.5W 36.5N	-0,19539	5.0W 35.0N	-0,10851
	6.5W 35.5N	-0,02162		
3	7.5W 34.5N	-0,04771	7.5W 32.5N	0,04555
	8.5W 33.5N	0,24828		
	9.5W 32.5N	-0,06393		
2	10.5W 31.5N	-0,18094	10.0W 30.0N	-0,33977
	11.5W 30.5N	-0,49859		
3	12.5W 29.5N	-0,40182	12.5W 27.5N	-0,22988
	13.5W 28.5N	-0,19257		
	14.5W 27.5N	-0,09523		

O mesmo método aplicado para o ajuste das coordenadas e da dimensão da matriz de dados, também é aplicado para definir a granularidade espacial. Porém, o passo é fixo para todas as coordenadas da matriz de dados. Por exemplo, se o passo, definido pelo usuário, for 8 a granularidade será definida obtendo o ponto médio de cada 8 coordenadas.

De forma prática, se as 8 primeiras coordenadas são de 0.0W à 20.0W e de 40.0N à 22.5N o ponto médio será de 7.5W e 32.5N, e assim sucessivamente. Caso seja um número ímpar como o passo 9, o ponto médio será a coordenada central de 10.0W e 30.0N, e assim sucessivamente. Da mesma forma, a cada passo é calculada a média aritmética dos dados das coordenadas correspondentes para cada mês, como representado na Tabela 4.2. Se o passo for 1, a granularidade não será aplicada mantendo as coordenas e a matriz de dados como o original.

Tabela 4.2 - Granularidade Espacial Aplicada na Matriz de Dados.

Dados de Janeiro de 1999				
Passo	Coordenada Original	Dados de SLP	Coordada Desejada	Dados Ajustados
8	0.0W 40.0N	0,18562	7.5W 32.5N	0,08227
	2.5W 37.5N	-0,08378		
	5.0W 35.0N	-0,13729		
	7.5W 32.5N	0,22947		
	10.0W 30.0N	0,34532		
	12.5W 27.5N	0,14738		
	15.0W 25.0N	-0,10056		
	17.5W 22.5N	0,07198		
8	20.0W 20.0N	-0,00808	27.5W 12.5N	-0,02814
	22.5W 17.5N	-0,02088		
	25.0W 15.0N	-0,03731		
	27.5W 12.5N	-0,04451		
	30.0W 10.0N	-0,08106		
	32.5W 7.5N	0,02583		
	35.0W 5.0N	0,00135		
	37.5W 2.5N	-0,06044		
8	40.0W 0.0N	-0,07894	47.5W 7.5S	0,00328
	42.5W 2.5S	-0,06086		
	45.0W 5.0S	0,06000		
	47.5W 7.5S	0,05258		
	50.0W 10.0S	0,12133		
	52.5W 12.5S	0,04031		
	55.0W 15.0S	-0,01110		
	57.5W 17.5S	-0,09710		

Por meio de uma janela de diálogo, o sistema ERB-ArrayTools pergunta ao usuário se deseja apenas normalizar ou criar anomalia sobre estes dados. Em seguida um novo arquivo NetCDF pré-processado é gerado e carregado na BD conforme a filtragem.

A anomalia consiste em subtrair o dado bruto de um mês do determinado ano pela média do mês correspondente de todos os anos. Em seguida, esta matriz é normalizada. A normalização é feita a partir do maior valor absoluto encontrado na matriz de dados mantendo o sinal negativo ou positivo do respectivo valor a ser normalizado. Por exemplo, se os números são -5.0, 10.5, -30.0, 9.5 o maior número em módulo, que dividirá os demais, será o 30.0 tendo como resultado -0.16, 0.35, -1, 0.32 e assim por diante. As Tabelas 4.3 à 4.6 mostram um exemplo de como é calculado esta anomalia com a normalização em seguida.

Tabela 4.3 - Matriz de Dados Brutos.

Dados Brutos							
Coordenadas	Unique ID	Janeiro99	Fevereiro99	Janeiro00	Fevereiro00	Janeiro01	Fevereiro01
sst_7.5W32.5N	59	21,02	19,59	-49,08	-4,11	-10,48	28,79
sst_27.5W32.5N	60	-34,93	16,83	-1,53	-35,67	-18,13	46,03
sst_47.5W32.5N	61	-16,32	-4,43	-4,72	-4,93	11,28	33,77

Tabela 4.4 - Matriz de Médias Aritméticas.

Médias			
Coordenadas	Unique ID	Janeiro	Fevereiro
sst_7.5W32.5N	59	-12,85	14,76
sst_27.5W32.5N	60	-18,20	9,07
sst_47.5W32.5N	61	-3,25	8,13

Tabela 4.5 - Matriz de Dados com Anomalia.

Dados com Anomalia							
Coordenadas	Unique ID	Janeiro99	Fevereiro99	Janeiro00	Fevereiro00	Janeiro01	Fevereiro01
sst_7.5W32.5N	59	33,87	4,83	-36,23	-18,87	2,37	14,03
sst_27.5W32.5N	60	-16,73	7,77	16,67	-44,73	0,07	36,97
sst_47.5W32.5N	61	-13,07	-12,57	-1,47	-13,07	14,53	25,63
Maior valor absoluto:				**44,73**			

Tabela 4.6 - Matriz de Dados com Anomalia e Normalizados.

Dados Normalizados							
Coordenadas	Unique ID	Janeiro99	Fevereiro99	Janeiro00	Fevereiro00	Janeiro01	Fevereiro01
sst_7.5W32.5N	59	0,76	0,11	-0,81	-0,42	0,05	0,31
sst_27.5W32.5N	60	-0,37	0,17	0,37	-1,00	0,00	0,83
sst_47.5W32.5N	61	-0,29	-0,28	-0,03	-0,29	0,32	0,57

A matriz de dados pré-processados, denominado Conjunto de Dados de Expressão, pode ser obtida como uma tabela no formato apresentado da tabela anterior por meio de consulta SQL. Isto é feito de acordo com a filtragem espaço-temporal utilizando o subsistema ChronosGIS abordado no tópico seguinte. Esta solução possibilita salvar esta tabela em arquivo texto. Adicionalmente, o ERB-ArrayTools permite ao usuário abrir um arquivo texto com uma matriz de dados neste formato.

Na coluna das coordenadas, consta o nome da variável ambiental a ser analisada, o nível de altitude em Hecto Pascal, caso seja uma variável de multinível, e as coordenadas de longitude e latitude. Como exemplo, tem-se a seguinte forma vwnd_850hPa_7.5W32.5N para variáveis de multiníveis ou slp_7.5W32.5N para variáveis de superfície.

Esta informação do conjunto de dados é utilizada tanto na atualização da tela principal quanto na geração do arquivo de dados NetCDF com os resultados dos p-valores para posterior visualização pelos subsistemas ToosUI e IDV em forma de grade sobre o mapa da região correspondente.

Outro parâmetro muito importante para a fase seguinte de DM é o chamado “Descritor de Dados”. Para a geração do descritor o usuário deve escolher uma variável ambiental em uma determinada coordenada a fim de guiar o processo de mineração. Este descritor pode tanto ser obtido via um arquivo texto quanto gerado pelo ERB-ArrayTools a partir do BDG ou do arquivo texto com a matriz de dados lidos. Este descritor gerado também pode ser salvo para aplicação em outros conjuntos de dados.

A geração do descritor consiste em: (1) a partir do subconjunto de dados escolhido, denominado *Pacient Array*, o ordenar este subconjunto; (2) obter sua mediana; e (3) em seguida definir suas classes. Se o tamanho deste subconjunto de dados for par, a mediana será a média aritmética dos dois dados centrais do subconjunto ordenado. Se for impar, a mediana será o dado central.

Caso o dado pré-processado tiver valor menor do que a mediana, irá pertencer a classe 1. Caso ele tiver valor maior do que a mediana, irá pertencer a classe 2. As Tabelas 4.7 e 4.8 representam a geração do Descritor de Dados.

Tabela 4.7 - Descritor de um subconjunto de dados com tamanho par.

Dados Pré-processados

Time/Coord	prec_70.0W7.5S
Janeiro00	0,77
Fevereiro00	-0,29
Marco00	0,37
Abril00	0,01
Janeiro01	-0,81
Fevereiro01	0,64
Marco01	-0,51
Abril01	-0,61
Janeiro02	0,66
Fevereiro02	-0,10
Marco02	-0,73
Abril02	0,46

Dados Ordenados

Time/Coord	prec_70.0W7.5S	Classes
Janeiro01	-0,81	1
Marco02	-0,73	1
Fevereiro00	-0,29	1
Fevereiro02	0,10	1
Marco00	0,37	1
Abril02	0,46	1
Marco01	0,51	2
Abril01	0,61	2
Fevereiro01	0,64	2
Janeiro02	0,66	2
Janeiro00	0,77	2
Abril00	0,82	2
Mediana:	**0,48**	

Descritor de Dados

Pacient Array	prec_70.0W7.5S
Janeiro00	2
Fevereiro00	1
Marco00	1
Abril00	2
Janeiro01	1
Fevereiro01	2
Marco01	2
Abril01	2
Janeiro02	2
Fevereiro02	1
Marco02	1
Abril02	1

Tabela 4.8 - Descritor de um subconjunto de dados com tamanho impar.

Dados Pré-processados

Time/Coord	prec_70.0W7.5S
Janeiro00	0,77
Fevereiro00	-0,29
Marco00	0,37
Janeiro01	-0,81
Fevereiro01	0,64
Marco01	-0,51
Janeiro02	0,66
Fevereiro02	-0,10
Marco02	-0,73

Dados Ordenados

Time/Coord	prec_70.0W7.5S	Classes
Janeiro01	-0,81	1
Marco02	-0,73	1
Fevereiro02	0,10	1
Fevereiro00	0,29	1
Marco00	0,37	2
Marco01	0,51	2
Fevereiro01	0,64	2
Janeiro02	0,66	2
Janeiro00	0,77	2
Mediana:	**0,37**	

Descritor de Dados

Pacient Array	prec_70.0W7.5S
Janeiro00	2
Fevereiro00	1
Marco00	2
Janeiro01	1
Fevereiro01	2
Marco01	2
Janeiro02	2
Fevereiro02	1
Marco02	1

A Figura 4.6 apresenta o diagrama de classe do subsistema *ProduceNetCDF*.

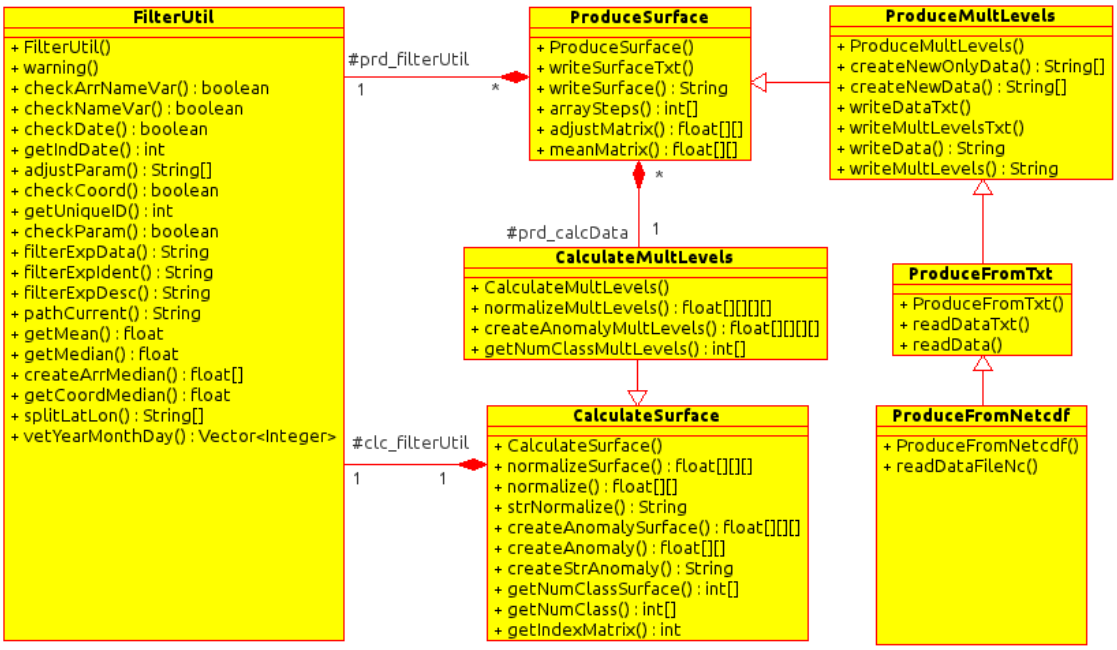


Figura 4.6 - Diagrama de Classe do subsistema ProduceNetCDF.

Na classe ProduceFromNetCdf estão implementados os métodos responsáveis pela leitura de toda estrutura de um arquivo NetCDF tanto de superfície quanto de multinível. Esta classe possibilita a produção de arquivos NetCDF's ao herdar os métodos da classe ProduceFromTxt.

Na classe ProduceFromTxt estão implementados os métodos responsáveis pela leitura de uma tabela espaço-temporal com dados ambientais tanto de uma variável *String* quanto de arquivo texto. Esta classe possibilita a produção de arquivos NetCDF's tanto de superfície quanto de multinível ao herdar os métodos da classe ProduceMultLevels.

Na classe ProduceMultLevels estão implementados os métodos responsáveis pela produção de arquivos NetCDF's de multinível, além dos métodos já herdados da classe ProduceSurface.

Na classe ProduceSurface estão implementados os métodos responsáveis pela produção de arquivos NetCDF's de superfície. Dentre os diversos métodos

implementados, pode ser citado: os métodos para o ajuste e aplicação da granularidade espacial; os métodos para aplicação da anomalia e normalização utilizando as classes do pacote *Calculate* deste subsistema; e os métodos responsáveis pela inclusão dos resultados estatísticos obtidos pelo subsistema StatisticalAnalysis.

O pacote *Calulate* é composto por duas classes CalculateMultLevels e CalculateSurface. Na classe CalculateMultLevels estão implementados diversos métodos responsáveis pela normalização e criação de anomalias no conjunto de dados de multinível. Nesta classe também incluem os métodos herdados da classe CalculateSurface.

Na classe CalculateSurface estão implementados diversos métodos responsáveis pela normalização e criação de anomalias no conjunto de dados de superfície. Nesta classe também estão implementados métodos responsáveis pela definição de classes a partir dos dados de uma série temporal para a execução do cálculo estatístico pelo subsistema *StatisticalAnalysis*. A classe FilterUtil do subsistema ERB-ArrayTools foi utilizado para a manipulação e ajuste da descrição espaço-temporal do conjunto de dados.

4.2.2. O Subsistema *ChronosGIS*

O subsistema *ChronosGIS* é uma adaptação do software de mesmo nome desenvolvido por Almeida et al. (2011). Este subsistema foi aprimorado e teve sua interface adaptada para integrar ao ERB-ArrayTools. O principal objetivo deste subsistema é a realização da carga da série de dados espaço-temporal de um arquivo NetCDF oriundo da NOAA (2012) para um banco de dados PostGIS facilitando o acesso e manipulação destes dados.

Este subsistema permite: (1) extrair do arquivo NetCDF o conjunto de dados da variável ambiental e sua informação espaço-temporal; (2) armazená-los no

banco de dados PostGIS; e (3) obter estatísticas básicas das variáveis ambientais armazenadas através de consulta SQL.

O PostGIS é um *plug-in* de código aberto robusto desenvolvido para ser uma extensão espacial do SGBD PostgreSQL destinada a trabalhar com dados geo-referenciados e multidimensionais (ALMEIDA et al., 2011). Este *plug-in* visa promover o desenvolvimento de padrões que facilitem a interoperabilidade entre Sistemas de Informações Geográficas. A Figura 4.7 apresenta alguns exemplos de consultas geométricas que envolvem o mapeamento e a seleção de áreas alcançadas por sentenças SQL no PostGIS.

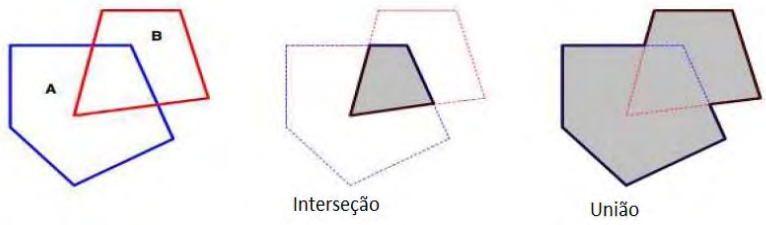


Figura 4.7 - Exemplo pictórico de consultas geométricas usando PostGIS.

Os *datasets* utilizados têm as seguintes informações:

- Para os *datasets* de superfície são consideradas três variáveis que definem a dimensão espaço-temporal. Estas variáveis são a *lat* (latitude), *lon* (longitude) e *time* (tempo). Para os *datasets* de multinível, além destas variáveis, também está incluído a variável referente ao nível de nome *level*. O *time* tem a dimensão ilimitada e as demais dimensões têm tamanho fixo.
- A variável seguinte é do tipo *float* e contêm os dados geográficos. Seu nome representa uma variável ambiental e é estruturada conforme sua característica, como por exemplo, a variável de nome *air* que contém dados de Temperatura do Ar.
- Além do nome, cada uma das variáveis possui um tipo associado: *float* para *lat*, *lon*, *level* e *air* e *double* para *time*. Elas também possuem

uma forma dada pelas dimensões nomeadas entre parênteses ao lado de cada nome de variável.

Uma vez que os *datasets* podem ser tridimensionais para dados de superfície ou quatro-dimensionais para dados de multinível, foi decidido um modelo para os dados utilizando um SGBD, o qual permite trabalhar com dados espaciais. Neste caso, as coordenadas *lat* e *lon* podem ser consideradas como um tipo *point* (ponto) no BDG, o que permite armazenar a dimensão do tempo e o valor da variável ambiental para cada ponto lido.

Porém, o *ChronosGIS* foi originalmente desenvolvido para armazenar dados de superfície no BD. Por isso, a solução mais prática adotada ao carregar o *dataset* de multinível, foi considerar que o valor da variável referente ao nível seria parte do nome da variável ambiental ao adicionar o campo referente a esta variável no BD. Esta abordagem permite diversas consultas SQL sobre os dados e torna possível sua aplicação no processo de KDD. A Figura 4.8 mostra as Tabelas criadas no banco de dados PostGIS utilizando dados espaço-temporal.

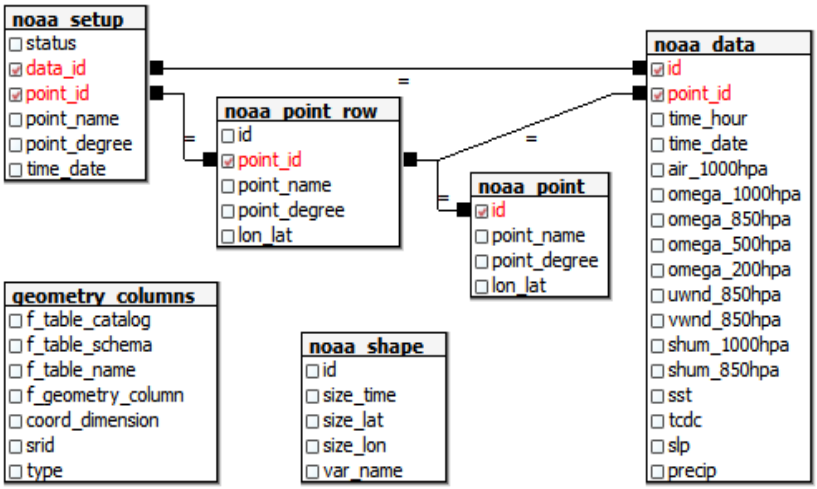


Figura 4.8 - Tabelas espaço-temporal criadas pelo ChronosGIS.

A tabela *geometry_columns* é uma tabela padrão do BDG criada para listar a informação de determinadas tabelas como: o catálogo (*f_table_catalog*), o

esquema das tabelas (*f_table_schema*) (*public*), o nome das tabelas (*f_table_name*) (*noaa_point* e *noaa_point_row*), o nome da coluna que irá conter os pontos geométricos referentes às coordenadas (*f_geometry_column*) (*lon_lat*), a dimensão das coordenadas (*coord_dimension*) (2), o *srid* (-1) e o tipo das tabelas (*type*) (*POINT*). Além desta tabela, há outras definidas como padrão pelo PostGIS para a manipulação de banco de dados geográficos.

Na tabela *noaa_shape* é inserido o identificador (*id*), o nome da variável ambiental (*var_name*) e as dimensões do conjunto de dados do arquivo NetCDF a ser carregado, como o tamanho do tempo (*size_time*), latitude (*size_lat*) e longitude (*size_lon*).

Na tabela *noaa_point* é inserido o identificador do ponto geométrico referente à coordenada (*id*), o nome do ponto (*point_name*) ("P91.23"), a coordenada geográfica em graus deste ponto (*pont_degree*) ("7.5W32.5N") e o ponto geométrico (*lon_lat*).

A tabela *noaa_point_row* é utilizada para auxiliar na verificação das coordenadas a serem inseridas na tabela *noaa_point* evitando a inserção ou atualização das mesmas coordenadas já inseridas anteriormente. Nesta tabela é inserido o identificador do ponto geométrico (*id*), a chave estrangeira para a tabela *noaa_point* (*point_id*), o nome do ponto (*point_name*), a coordenada (*point_degree*) e o ponto geométrico (*lon_lat*) da ultima coordenada inserida na tabela *noaa_point*. Com isso, apenas novas coordenadas serão inseridas na tabela *noaa_point* diminuindo o tempo de processamento a cada carregamento dos dados geográficos.

Na tabela noaa_data é inserido o identificador de cada dado ambiental (*id*), a chave estrangeira para a tabela *noaa_point* (*point_id*), o tempo em horas corridas (*time_hour*), a data (*time_date*) e em seguida as colunas referentes às variáveis ambientais de multinível e de superfície (*air_1000hpa*, *sst*, *slp*,...).

A tabela noaa_setup é utilizada para auxiliar na verificação do conjunto de

dados a ser inserido na tabela *noaa_data* evitando a inserção ou atualização do mesmo conjunto de dados de uma determinada variável ambiental já inserido anteriormente. Nesta tabela é inserido: o identificador do estado do processo de carga na tabela *noaa_data* (*status*); as chaves estrangeiras para as tabelas *noaa_data* (*data_id*) e *noaa_point* (*point_id*); o nome do ponto (*point_name*); a coordenada (*point_dregree*); e a data (*time_date*).

O identificador da coluna *status* desta tabela é definido como 1 para o primeiro dado carregado e como 2 para o ultimo dado obtido do processo interrompido pelo usuário ou como 3 para o ultimo dado obtido do processo concluído com sucesso. Com isso, apenas novos conjuntos de dados serão inseridos na tabela *noaa_data* diminuindo o tempo de processamento.

Além da função de armazenar dados espaço-temporais no PostGIS, o ChronosGIS também permite efetuar estatísticas básicas por meio de sentenças SQL's utilizando a localização geográfica e o tempo inicial e final destes dados. Anteriormente, esta funcionalidade não era possível com os dados geográficos empacotados no formato NetCDF. A Tabela 4.9 mostra um exemplo de dados estatísticos de temperatura do ar com o nível de 1000hPa obtidos pela consulta SQL abaixo.

```
SELECT AVG(air_1000hPa) temperatura_media, MIN(air_1000hPa) temperatura_min,
      MAX(air_1000hPa) temperatura_max, STDDEV(air_1000hPa) desvio_padrao,
      VARIANCE(air_1000hPa) variancia
FROM   noaa_point, noaa_data
WHERE  (noaa_point.id = noaa_data.point_id)
      AND point_degree BETWEEN '27.5W12.5N' AND '67.5W27.5S'
      AND time_date BETWEEN '2000/01/01' AND '2006/12/01'
```

Tabela 4.9 - Estatísticas espaço-temporal obtida pelo ChronosGIS.

	Média	**Mínima**	**Máxima**	**Desvio Padrão**	**Variância**
Temperaturas	0.0201	-1	0.9345	0.1544	0.0238

A Figura 4.9 apresenta o diagrama de classe do subsistema ChronosGIS.

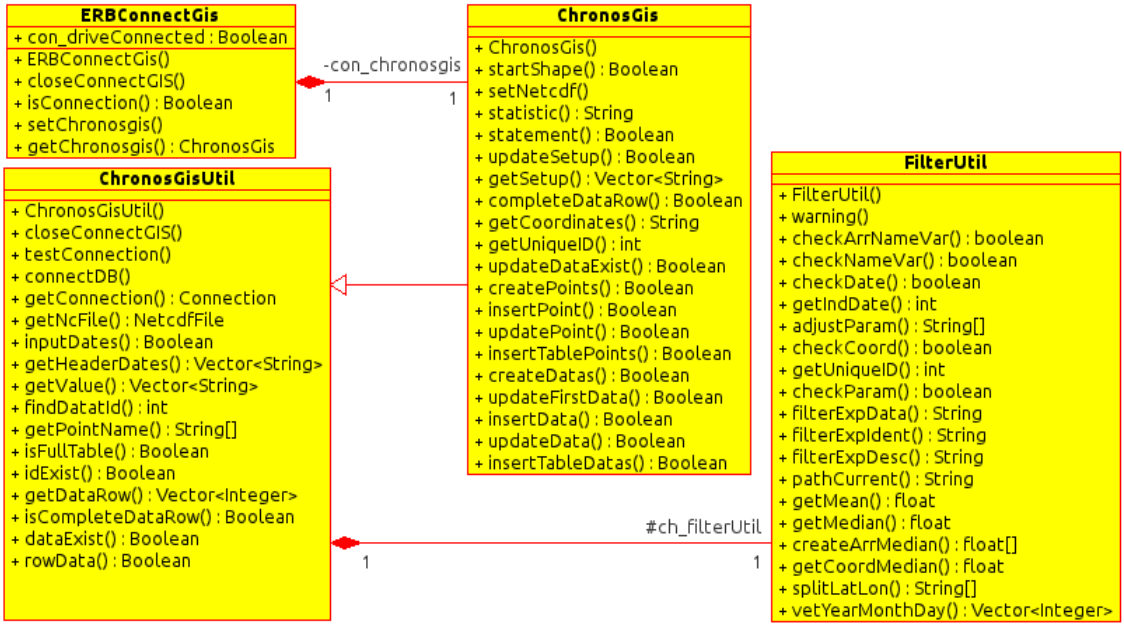


Figura 4.9 - Diagrama de Classe do subsistema ChronosGIS.

Na classe ChronosGis estão implementados os métodos responsáveis pela carga dos dados ambientais de um arquivo NetCDF em um BDG. Esta classe permite: abrir um arquivo NetCDF; obter o conjunto de dados e sua informação espaço-temporal; criar as tabelas de coordenadas geográficas e de dados ambientais em uma série temporal; preencher ou atualizar as tabelas criadas; e obter estatísticas básicas por meio de consulta SQL. Outros métodos essenciais para o funcionamento do subsistema ChronosGIS são herdados da classe ChronosGisUtil.

Na classe ChronosGisUtil estão implementados: o método de conexão com o SGBD; o método de teste de conexão; métodos de retorno e consultas em tabelas do BDG; e uma janela de interação com o usuário para a visualização do processo de importação de arquivos NetCDF no BDG.

Como dito anteriormente, a classe ERBConnetGIS foi utilizado para conexão com o SBDG PostgreSQL e a classe FilterUtil foi utilizado para a manipulação das informações espaço-temporal do conjunto de dados geográficos.

4.2.3. O Subsistema *StatisticalAnalysis*

O subsistema *StatisticalAnalysis* foi desenvolvido em Java para integrar o ERB-ArrayTools ao Matlab através de métodos que utilizam a biblioteca MatlabControl (2012) na versão 4.0 para acesso às funções dos *scripts* de código Matlab. Estes *scripts* em foram desenvolvidos a partir da evolução do código legado de um projeto descontinuado chamado *Array Statistical Analysis System* (ASAS) (CARVALHO et al., 2011). O diagrama de pacotes da Figura 4.10 representa a integração dos *scripts* para a obtenção dos p-valores.

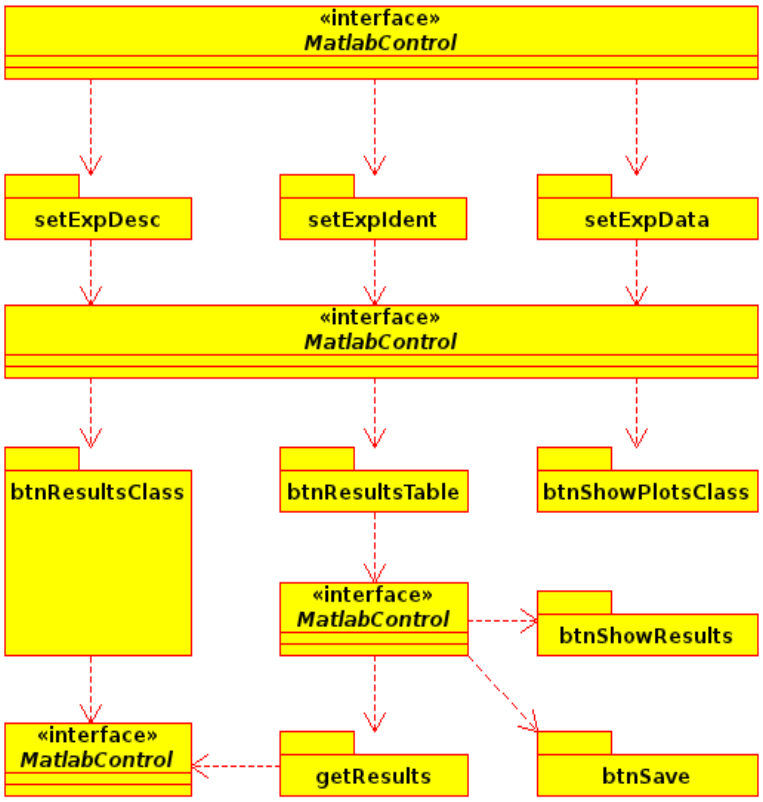


Figura 4.10 - Diagrama de Pacotes desenvolvidos em Matlab.

As funcionalidades da biblioteca MatlabControl permite o *StatisticalAnalysis* agrupar as funções dos *scripts* para a realização do processo de DM, como proposto inicialmente por Ruivo (2008), sobre uma tabela de dados no formato apresentado na subseção 4.2.1.

Assim como o BRB-ArrayTools, o ASAS também foi desenvolvido para processar, visualizar, agrupar, classificar, dentre outras funções realizada sobre os dados de MA de DNA. Para isso, esta ferramenta utiliza a biblioteca de funções do Matlab para aplicar o método estatístico t-teste sobre dados matriciais, como apresentado no capítulo 3, e gerar visualizações dos resultados estatísticos obtidos.

O *script* `setExpIdent` recebe a *string* da primeira coluna da tabela de dados e define uma variável do tipo vetor para receber as descrições espaciais no formato: nome da variável ambiental, nível, caso seja de multinível, e coordenada geográfica, como por exemplo “vwnd_850hPa_7.5W32.5N” ou “slp_7.5W32.5N”. Para isso, são removidos os caracteres tabulação (‘\t’) e salto de linha (‘\n’) da *string* recebida.

O *script* `setExpData` recebe *strings* com os dados e o cabeçalho da tabela de dados e define variáveis do tipo matriz para receber os dados ambientais e do tipo vetor para receber as descrições da série temporal no formato de mês e ano, como por exemplo “Janeiro99”. Assim como o *script* `setExpIdent`, são removido os caracteres de tabulação e salto de linha.

O *script* `setExpDesc` recebe *strings* com a primeira coluna, o cabeçalho e as classes da tabela de descritor de dados e define variáveis do tipo matriz para receber as classes e do tipo vetor para receber as descrições espaciais e as descrições da série temporal. Neste *script* também são removido os caracteres de tabulação e salto de linha.

O *script* `btnResultsTable` realiza o calculo estatístico através das funções `mattest` do script `btnShowHistPermute` e `mafdr` do script `btnShowFDR` e retorna as matrizes de resultados obtidos de todos os tipos de agrupamentos do descritor de dados e um vetor com os nomes das variáveis estatísticas. Cada um destes nomes corresponde a uma determinada coluna da matriz com seus respectivos resultados.

Neste caso, existe apenas um tipo de agrupamento A versos B para obtenção dos resultados, pois o critério usado na criação do descritor foi a definição de dois tipos de classes de uma série temporal, classe A de valor 1 para acima da mediana e classe B de valor 2 para abaixo da mediana.

O *script* `btnResultsClass` utiliza o *script* `btnResultsTable` para calcular e retornar os resultados estatísticos e seus respectivos nomes obtidos de um tipo de agrupamento escolhido pelo usuário. Os resultados obtidos por estes *scripts* são visualizados em forma de tabela pelo subsistema *StatisticalAnalysis*.

Da mesma forma que na tabela de dados, na primeira coluna da tabela de resultados terá a descrição espacial e no cabeçalho terá o nome das variáveis estatísticas obtidas. No capítulo a seguir, será apresentado um tutorial sobre o funcionamento do sistema ERB-ArrayTools, o qual mostra um exemplo para obtenção da tabela de resultados neste formato.

O *script* `getResults` retorna um vetor de *strings* contendo as tabelas de resultados estatísticos. Como dito anteriormente, neste caso existe apenas um tipo de agrupamento, então o vetor retornado terá apenas uma posição com uma tabela de resultados. Este *script* é utilizado no método do subsistema *StatisticalAnalysis* responsável pela geração de arquivos NetCDF's com os resultados estatísticos.

O *script* `btnSave` é utilizado para salvar os resultados estatísticos em forma de tabela em arquivo texto.

O *script* `btnShowResults` é utilizado para visualizar o conjunto de dados e os resultados estatísticos de uma variável ambiental utilizando gráficos de contorno e em 3D.

O *script* `btnShowPlotsClass` utiliza o mesmo processo aplicado no *script* `btnResultsClass` para diversas formas de visualizações dos resultados estatísticos. No Apêndice A deste trabalho apresenta as visualizações deste

script e do *script* `btnShowResults`, o qual mostra em gráficos o conjunto de dados e os p-valores das variáveis ambientais vwnd, air e sst.

No diagrama de pacotes da Figura 4.11, está representado o *script* `btnShowPlotsClass` realizando a chamada de diversos outros scripts para a visualização dos p-valores e dos demais resultados estatísticos obtidos.

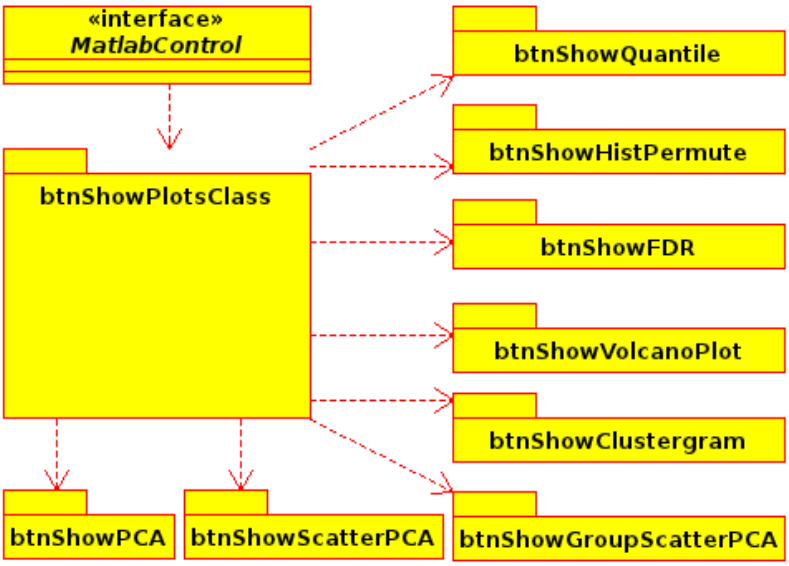


Figura 4.11 - Diagrama de Pacotes dos gráficos estatísticos.

Os *scripts* a seguir utilizam as funções da biblioteca padrão do Matlab voltadas para a aplicação na Biologia Computacional:

- `btnShowQuantile` – gera o gráfico quantil normal do t-teste. Este *script* utiliza a função `mattest` que realiza o método estatístico t-teste de duas amostras para avaliar a expressão diferencial de genes de duas condições experimentais ou fenótipos.
- `btnShowHistPermute` – gera os gráficos de histograma dos t-testes e dos p-valores utilizando a função `mattest`.
- `btnShowFDR` – gera o gráfico da estimativa da taxa de falsa descoberta. Este *script* utiliza a função `mafdr` que realiza a estimava

da taxa de falsa descoberta, do inglês *False Discovery Rate* (FDR) de genes diferencialmente expressos a partir de duas condições experimentais ou fenótipos.

- `btnShowVolcanoPlot` – gera o gráfico de teste de significância dos dados mais significativos. Este *script* utiliza a função `mavolcano` obtida da adaptação da função `mavolcanoplot`. A função `mavolcano` permite o retorno dos resultados estatísticos com a opção de mostrar ou não a janela de visualização do gráfico *Volcano Plot*. Desta forma, estes resultados poderão ser utilizados por outros *scripts* para visualização individualmente. A função `mavolcanoplot` e seu resultante `mavolcano` geram um gráfico de dispersão dos dados de MA traçando a taxa de significância versus a taxa de expressão gênica.
- `btnShowClustergram` – gera o dendrograma com o agrupamento hierárquico das variáveis mais significativas. Este *script* utiliza a função `clustergram` que calcula o agrupamento hierárquico, exibi o dendrograma e o mapa de calor, e retorna o objeto clustergrama que contém os dados da análise do agrupamento obtido.
- `btnShowPCA` – utiliza a função `mapcaplot` que executa a ferramenta *Principal Component Visualization* para a visualização da dispersão dos dados mais significativos. Esta função gera o gráfico de dispersão para análise do subsistema principal sobre os dados do microarranjo.
- btnShowScatterPCA – gera o gráfico de dispersão dos dados geográficos. Este *script* utiliza as funções `princomp` para a análise do subsistema principal sobre os dados e `scatter` para a geração do gráfico de dispersão.
- btnShowGroupScatterPCA - gera o gráfico de dispersão dos dados geográficos com grupos coloridos. Este *script* também utiliza a função `princomp`, além das funções `clusterdata` para a construção de

grupos aglomerados a partir dos dados e `gscatter` para a geração do gráfico de dispersão por grupo.

Como os scripts btnShowScatterPCA e btnShowGroupScatterPCA têm funcionalidades semelhantes, neste trabalho foi integrado ao ERB-ArrayTools apenas o script btnShowGroupScatterPCA para a visualização do gráfico de dispersão dos dados agrupados com colores diferentes. A Figura 4.12 apresenta o diagrama de classe do subsistema *StatisticalAnalysis*.

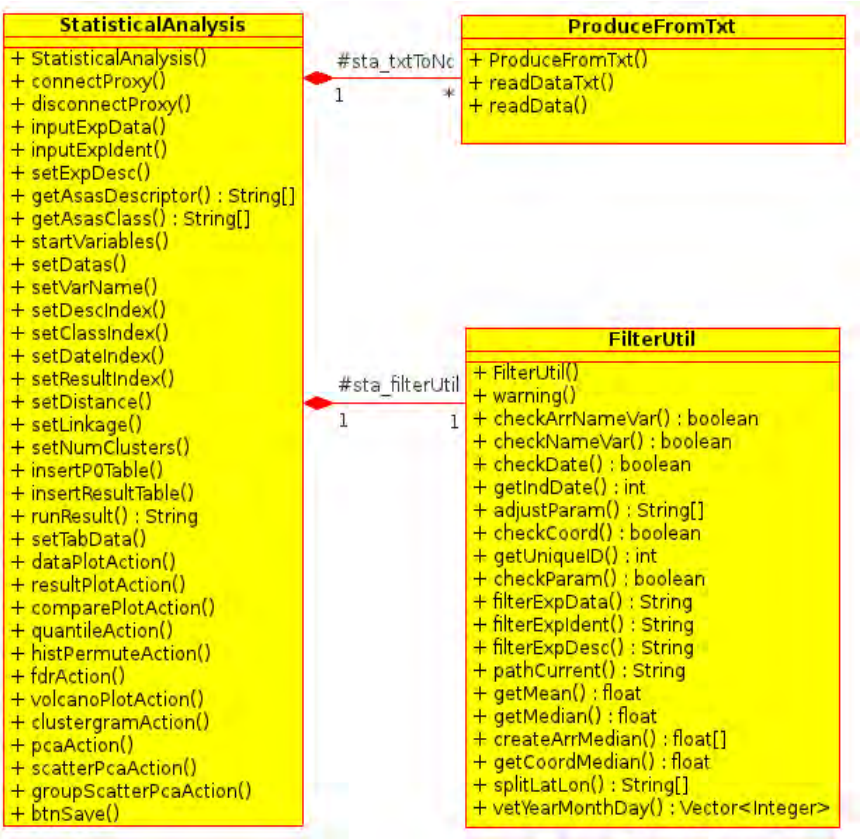


Figura 4.12 - Diagrama de Classe do Subsistema StatisticalAnalysis.

Na classe StatisticalAnalysis estão implementados diversos métodos para: conexão e desconexão com o Matlab, definição de variáveis essenciais para o funcionamento dos *scripts*; execução das funções dos *scripts* para a realização do método estatístico t-teste; obtenção e visualização dos resultados estatísticos em forma gráfica pelo Matlab e em uma tabela da subjanela

implementada nesta classe; criação de arquivo texto com os resultados obtidos em forma de tabela; e exportação do conjunto de dados pré-processado das variáveis ambientais com seu respectivo resultado estatístico para arquivos NetCDF's através da classe ProduceFromTxt do subsistema ProduceNetCDF.

Assim como o subsistema ChronosGIS, este subsistema também utiliza a classe FilterUtil para a manipulação das informações espaço-temporal do conjunto de dados geográficos para a realização da mineração de dados.

4.2.4. Os Subsistemas ToolsUI e IDV

Os subsistemas ToolsUI e *Integrated Data Viewer*, ou simplesmente IDV, são interfaces desenvolvidas para a integração do ERB-ArrayTools às ferramentas responsáveis pela visualização de arquivos NetCDF's em forma de grade sobre um mapa. O diagrama de classes destes subsistemas está representado na Figura 4.13.

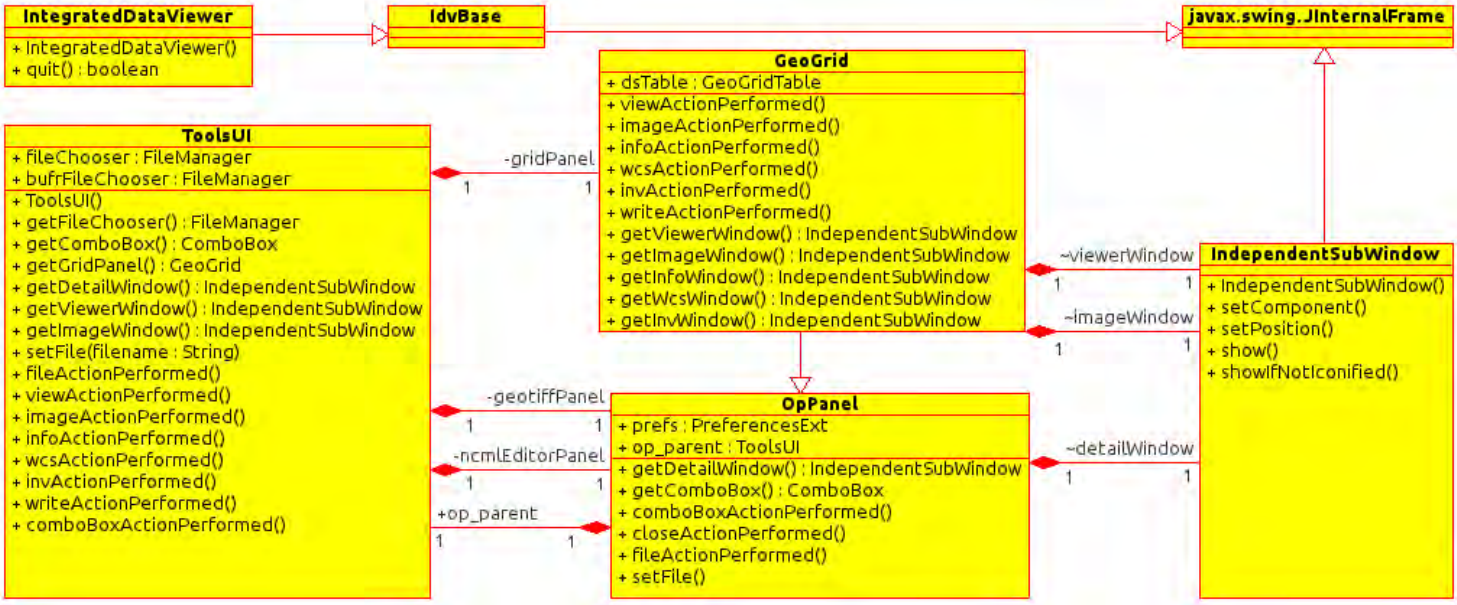


Figura 4.13 - Diagrama de Classe dos subsistemas ToolsUI e IDV.

O subsistema ToolsUI integra as funcionalidades GridView e ImageView da ferramenta de mesmo nome desenvolvida pela Unidata (NETCDF-JAVA,

2012). A classe OpPanel contém métodos responsáveis por localizar e abrir um arquivo NetCDF. A classe GeoGrid herda os métodos desta classe e constrói os objetos responsáveis pelas funcionalidades *Grid View* para a visualização dos dados em cores e *Image View* para visualização em escala de cinza.

A classe IndependentSubWindow herda as funcionalidades da classe JInternalFrame e é utilizada na construção destes objetos para sua disponibilização como subjanela. A classe ToolsUI contém o métodos responsáveis pela integração destas subjanelas no sistema principal ERB-ArrayTools.

A ferramenta IDV (2012), também desenvolvida pela Unidata (2012), foi adaptada e integrada como um subsistema para disponibilizar diversos tipos de visualização, dentre elas a visualização dos arquivos NetCDF's em 3D sobre um mapa. A principal ideia foi utilizar todas as telas e funcionalidades desta ferramenta como se fosse parte do ERB-ArrayTools.

O problema de utilizá-lo como uma das telas, foi que ao encerrar o IDV, automaticamente encerrava-se todo o sistema. Por isso, a solução encontrada foi alterar a classe *IntegratedDataViewer,* na qual estão localizados os principais métodos responsáveis pelo seu funcionamento. Basicamente, a sua classe pai *IdvBase* foi alterada para herdar da classe *JinternalFrame.* Desta forma, foi possível alterar na classe *IntegratedDataViewer* a função *System.exit(false)* do método *quit* para *setVisible(false)*.

4.3. Estrutura de Arquivos do Projeto

Para o funcionamento das visualizações mais arrojadas utilizando o subsistema IDV, foi necessária a instalação do *Java Development Kit* versão 6 e da API Java 3D disponibilizados pela Oracle nos sites (JAVAJDK, 2012) e (JAVA3D, 2012). Na Figura 4.14 está a arquitetura do projeto utilizado para o desenvolvimento do sistema ERB-ArrayTools.

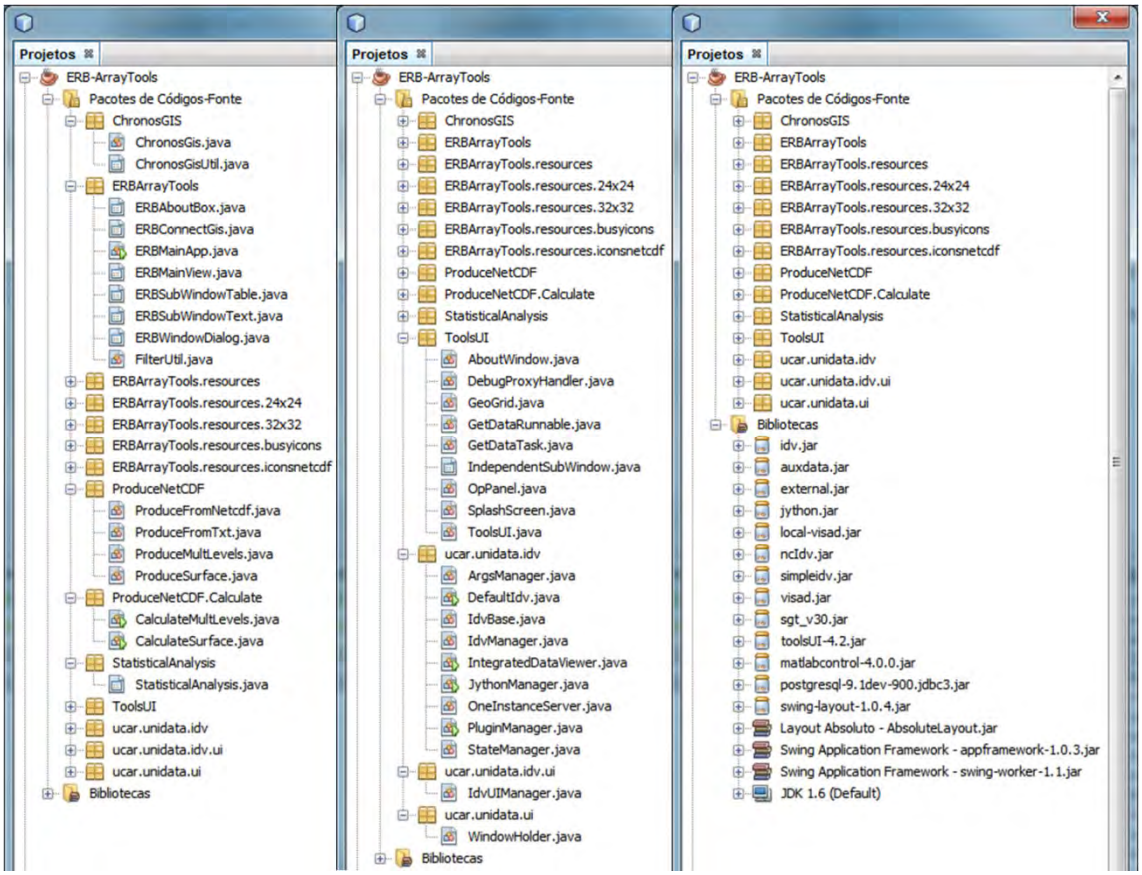


Figura 4.14 – Arquitetura do Projeto do ERB-ArrayTools.

A pasta **`ERBArrayTools`** contém as classes do subsistema principal para integração com os demais subsistema e interação com o usuário.

A pasta **`ProduceNetCDF`** contém as classes responsáveis pela leitura e pré-processamento de arquivos NetCDF's. A subpasta **`Calculate`** contém as classes responsáveis pela criação de anomalias e normalização dos dados ambientais e definição do descritor de dados de uma série temporal.

A pasta **`ChronosGIS`** contém as classes que permitem ao usuário realizar a conexão, carga e obtenção de dados pré-processados do banco de dados geográfico.

A pasta **`StatisticaAnalysis`** contém a classe responsável pela integração do ERB-ArrayTools com o Matlab para a aplicação da mineração de dados utilizando a permutação do métodos estatístico t-teste.

A pasta **`toolsUI`** contém as classes responsáveis pela integração das funcionalidades *GridView* e *ImageView* para a visualização dos novos arquivos NetCDF's com p-valores obtidos pela mineração de dados.

A pasta **`ucar.unidata.idv`**, contém as classes responsáveis pela integração da ferramenta *Integrated Data Viewer* juntamente com suas principais dependências. As demais dependências dos subsistemas estão nas respectivas bibliotecas adicionadas ao projeto.

Para o funcionamento do subsistema IDV, foi necessário alterar a biblioteca idv.jar de forma que o sistema acesse a classe alterada e não o arquivo .class da biblioteca. Para isso, foi removido, por exemplo, o *IntegratedDataViewer.class* da biblioteca para sua substituição pela nova classe modificada.

Como padrão a ferramenta de desenvolvimento NetBeans gera o pacote de execução .jar com apenas os arquivos .class do projeto do ERB-ArrayTools estando separado das demais bibliotecas. Para solucionar este problema, foi alterado o arquivo build.xml do projeto responsável pela criação do arquivo executável ERB-ArryTools.jar. Com isso, todas as dependências foram incluídas neste executável em Java corrigindo o erro que obtia ao executar o ERB-ArrayTools fora do ambiente de desenvolvimento NetBeans. A Figura 4.15 mostra o código XML do arquivo build.xml para a solução deste problema.

```xml
<?xml version="1.0" encoding="UTF-8"?>
<project name="ERB-ArrayTools" default="default" basedir=".">
       <description>Builds, tests, and runs the project
        ERB-ArrayTools.</description>
       <import file="nbproject/build-impl.xml"/>
       <!-- Essa é a parte que cria apenas um JAR com todas as bibliotecas -->
       <target name="-post-jar">
       <!-- Foi atribuído ao parâmetro value a String "ERB-ArrayTools"
            para criar o JAR com o nome de 'ERB-ArrayTools.jar' -->
              <property name="store.jar.name" value="ERB-ArrayTools"/>
              <property name="store.dir" value="store"/>
              <property name="store.jar"
               value="${store.dir}/${store.jar.name}.jar"/>
              <echo message="Packaging ${store.jar.name} into a single JAR
               at ${store.jar}"/>
              <delete dir="${store.dir}"/>
              <mkdir dir="${store.dir}"/>
              <jar destfile="${store.dir}/temp_final.jar"
               filesetmanifest="skip">
                     <zipgroupfileset dir="dist" includes="*.jar"/>
                     <zipgroupfileset dir="dist/lib" includes="*.jar"/>
                     <manifest>
                            <attribute name="Main-Class" value="${main.class}"/>
                     </manifest>
              </jar>
              <zip destfile="${store.jar}">
                     <zipfileset src="${store.dir}/temp_final.jar"
                      excludes="META-INF/*.SF, META-INF/*.DSA, META-
                      INF/*.RSA"/>
              </zip>
              <delete file="${store.dir}/temp_final.jar"/>
       </target>
</project>
```

Figura 4.15 – Script build.xml para geração do ERB-ArrayTool.jar

Diversas outras APIs de arquivos NetCDF desenvolvidos em Java também podem ser adaptadas como biblioteca de funções. Exemplos de APIs disponibilizadas pela EPIC (2012) são:

- DapperM – Uma interface de Matlab para Dapper - Serviço de dados de um OPeNDAP *in-situ*;
- ncBrowse – Navegador e visualizador de NetCDF e OpeNDAP;
- MultiView – Visualizador de meta arquivos Plot Plus;
- Java OceanAtlas (JOA) – Aplicativo para visualização e manipulação de dados de perfis oceanográficos;
- NdEdit – Ferramenta interativa para selecionar graficamente e sub-configurar grandes coleções de dados *in-situ*.

4.4. Implantação do ERB-ArrayTools

Para a instalação do ERB-ArrayTools, foi criado um programa de instalação utilizando uma ferramenta chamada HM NIS Edit na versão 2.0.3 disponível em Rodriguez (2013). Esta ferramenta é um software livre muito utilizado para edição de *scripts* em *Nullsoft Install System Scriptable* (NSIS, 2013). Um tutorial sobre a criação de *scripts* NSIS pode ser encontrado em Mosavi (2013). A Figura 4.16 mostra um exemplo criado a partir do *script* NSIS original para a geração do instalador do ERB-ArrayTools para Windows7.

```
Name "${PRODUCT_NAME} ${PRODUCT_VERSION}"
OutFile "InstallERB.exe"
InstallDir "$PROGRAMFILES\ERB-ArrayTools"
InstallDirRegKey HKLM "${PRODUCT_DIR_REGKEY}" ""
ShowInstDetails show
ShowUnInstDetails show

Section "SeçãoPrincipal" SEC01
       SetOutPath "$PROGRAMFILES\ERB-ArrayTools"
       SetOverwrite try
       File "ERB-ArrayTools\runERB.bat"
       File "ERB-ArrayTools\runERB_terra_32x29.ico"
       ; Shortcuts
       CreateDirectory "$SMPROGRAMS\ERB-ArrayTools"
       CreateShortCut "$SMPROGRAMS\ERB-ArrayTools\ERB-ArrayTools.lnk"
       "$PROGRAMFILES\ERB-ArrayTools\runERB.bat" ""
       "$PROGRAMFILES\ERB-ArrayTools\runERB_terra_32x29.ico"
       "0" SW_SHOWNORMAL
       CreateShortCut "$DESKTOP\ERB-ArrayTools.lnk"
       "$PROGRAMFILES\ERB-ArrayTools\runERB.bat" ""
       "$PROGRAMFILES\ERB-ArrayTools\runERB_terra_32x29.ico"
       "0" SW_SHOWNORMAL
       CreateShortCut "$STARTMENU\ERB-ArrayTools.lnk"
       "$PROGRAMFILES\ERB-ArrayTools\runERB.bat" ""
       "$PROGRAMFILES\ERB-ArrayTools\runERB_terra_32x29.ico"
       "0" SW_SHOWNORMAL
SectionEnd

Section -AdditionalIcons
       SetOutPath $INSTDIR
       WriteIniStr "$INSTDIR\${PRODUCT_NAME}.url" "InternetShortcut" "URL"
       "${PRODUCT_WEB_SITE}"
       CreateShortCut "$SMPROGRAMS\ERB-ArrayTools\Website.lnk"
       "$INSTDIR\${PRODUCT_NAME}.url" ""
       "$PROGRAMFILES\ERB-ArrayTools\runERB_32x28.ico"
       "0" SW_SHOWNORMAL
       CreateShortCut "$SMPROGRAMS\ERB-ArrayTools\Uninstall.lnk"
       "$INSTDIR\uninst.exe"
SectionEnd
```

Figura 4.16 – Script NSIS para criar programa de instalação.

Com a integração da ferramenta IDV ao sistema ERB-ArrayTools, foi necessário a criação de um *script* chamado runERB.bat com os mesmos parâmetros de execução utilizados e mesma estrutura de instalação criada para esta ferramenta. A Figura 4.17 apresenta o script utilizado para execução do sistema com algumas especificações na forma de comentários.

```
@echo off
REM ##############################################################################
REM Script: runERB
REM
REM Purpose: script to launch the IDV
REM
REM Syntax: runERB <idv options>
REM
REM Notes: In past versions of the IDV, users had to change this script to
REM manipulate memory settings. The IDV now configures the appropriate memory.
REM Users can also change the memory via the Preferences menu. In exceptional
REM situations where the IDV may not start due to memory issues, it may be
REM necessary to bootstrap the memory size. In this case, please uncomment the
REM idv_memory section below and subsequently choose memory via the Preferences
REM menu. Be sure to comment it out that after setting the memory via the
REM Preferences if you want the preference to take effect.
REM ##############################################################################
Setlocal
FOR /F "tokens=*" %%i IN ('jre\bin\java -cp ERB-ArrayTools.jar
ucar.unidata.idv.IdvCommandLinePrefs %* 2^>NUL') DO SET %%i
REM IF %idv_memory%.==. (
REM echo IDV failed to start. Please contact support-idv@unidata.ucar.edu
REM GOTO end)
REM Stripping quotes
set idv_memory=%idv_memory:"=%
REM See important note about this above. To bootstrap the IDV memory, uncomment
REM the line below and set to a value in megabytes.
REM set idv_memory=512
REM To avoid IDV OutOfMemory problems, it may be necessary to increase the
REM MaxPermSize in the Java
REM Virtual Machine. The default MaxPermSize is 64m. To increase it to 128m
REM please comment out next
REM line and uncomment the line following that.
@echo on
jre\bin\java -Xmx%idv_memory%m -Didv.enableStereo=false -jar ERB-ArrayTools.jar %*
REM jre\bin\java -Xmx%idv_memory%m -XX:MaxPermSize=128m -Didv.enableStereo=false -
jar
ERB-ArrayTools.jar %*
@echo off
REM Use the line below instead if you want to use the D3D version of Java 3D
REM jre\bin\java -Xmx%idv_memory%m -Dj3d.rend=d3d -jar ERB-ArrayTools.jar %*
Endlocal
:end
```

Figura 4.17 – Script runERB.bat para execução do sistema.

5 TUTORIAL DO SISTEMA ERB-ARRAYTOOLS

O resultado do desenvolvimento da arquitetura apresentada no capítulo 4 foi o sistema ERB-ArrayTools, o qual é capaz de aplicar o método estatístico t-teste sobre os dados espaço-temporais de forma análoga à técnica de análise de Microarranjos de DNA na Biologia Computacional. Neste capítulo, é apresentado um tutorial do sistema ERB-ArrayTools demonstrando suas principais funcionalidades para a interação com o usuário.

5.1. Instalação do ERB-ArrayTools

Nesta seção, será abordado o processo de instalação do sistema mostrando suas principais modalidades de interação com o usuário. Antes da instalação do ERB-ArrayTools, o usuário deve instalar em seu sistema operacional Windows o programa SGBD PostgreSQL juntamente com seu *plug-in* PostGIS para armazenamento de dados geográficos no BDG e o programa Matlab de igual ou superior a versão 7.9 (R2009b) para a execução das funções estatísticas escritas em código Matlab. A Figura 5.1 apresenta o programa de instalação do ERB-ArrayTools.

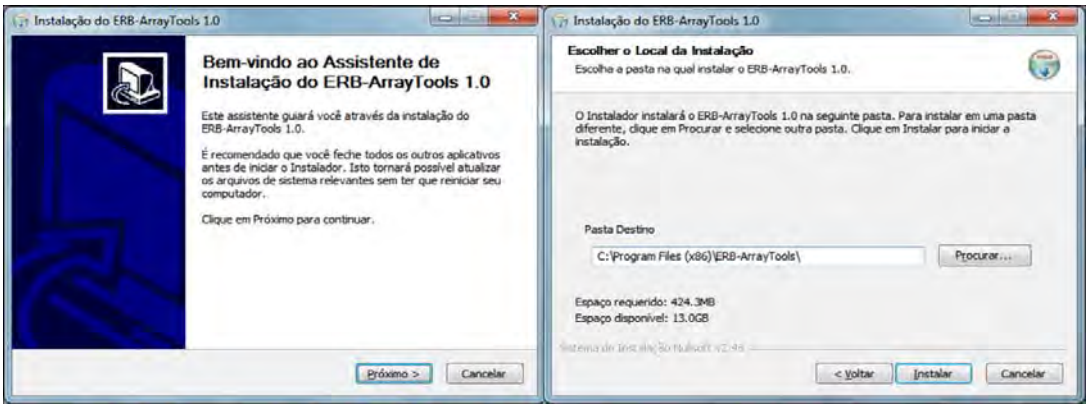


Figura 5.1 – Programa de instalação do ERB-ArrayTools.

Ao executar o instalador, basta seguir os passos indicados para que a estrutura do ERB-ArrayTools seja instalada na pasta Arquivo de Programa do Windows. A estrutura adotada para sua instalação é composta basicamente por:

- Arquivos de dependências do subsistema IDV na pasta **`.install4j`**;
- Java-jre versão 6 com a API Java 3D na pasta **`jre`** para execução de forma independente com recursos gráficos em 3D;
- Exemplos na pasta **`ArquivosERB`**;
- Licença das bibliotecas do sistema na pasta **`licenses`**;
- *Scripts* na pasta **`statisticalAnalysis`** contendo funções responsáveis pela obtenção dos resultados e das visualizações gráficas pelo Matlab;
- **`Link`** para o *site* a ser criado para o sistema;
- Programa principal **`ERB-ArrayTools.jar`**;
- *Script* de execução **`runERB.bat`**;
- Programa de desinstalação do sistema **`uninst.exe`**.

A Figura 5.2 apresenta esta estrutura instalada na pasta **`ERB-ArrayTools`** criada dentro da pasta **`Arquivos de Programas`** do Windows7 e a Figura 5.3 mostra os *scripts* do Matlab na pasta **`statisticalAnalysis`**.

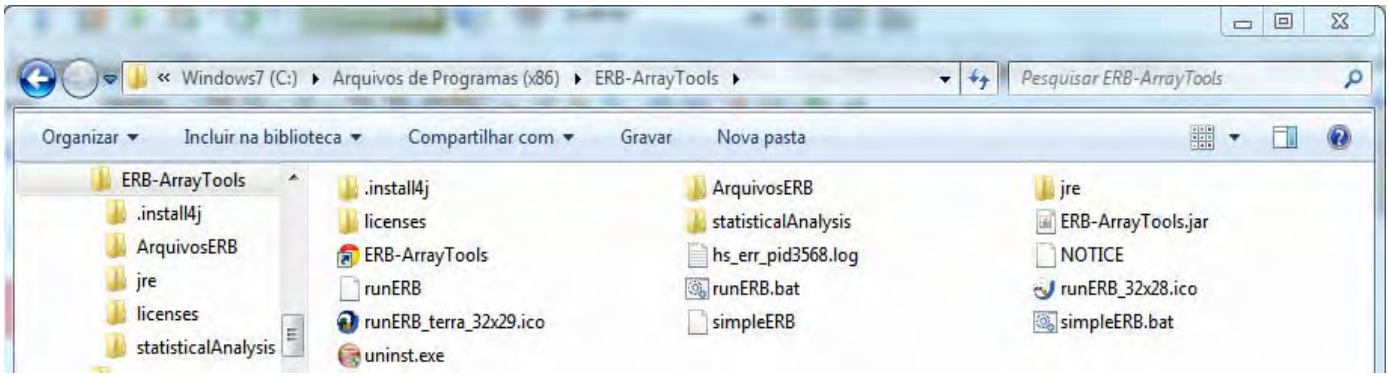


Figura 5.2 – Estrutura do ERB-ArrayTools instalado no Windows 7.

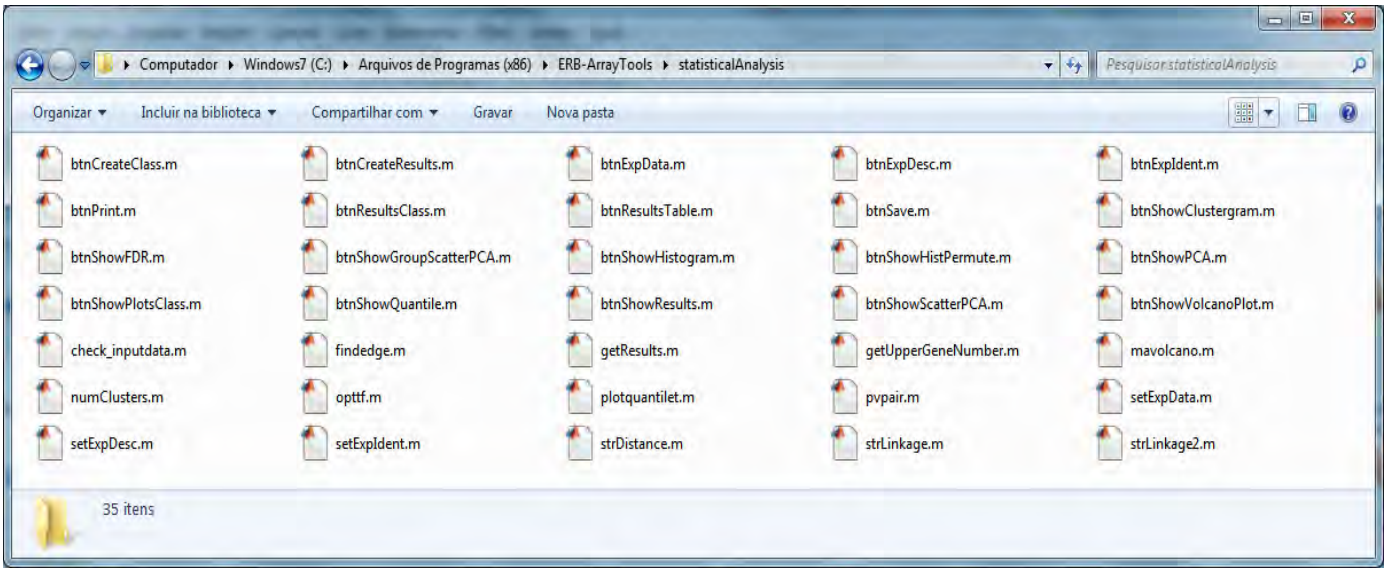


Figura 5.3 – Instalação dos *Scripts* na pasta statisticalAnalysis.

Assim como outros instaladores, o instalador do ERB-ArrayTools extrai todos os arquivos e subpastas desta estrutura na pasta **`Arquivos de Programas`** e cria no menu iniciar os atalhos para o *script* **`runERB.bat`**, para o link de acesso ao site do INPE e para o desinstalador **`uninst.exe`**. Além da criação dos atalhos no menu iniciar, também cria um atalho para o **`runERB.bat`** na área de trabalho facilitando ainda mais o acesso ao ERB-ArrayTools pelo usuário.

No termino da instalação, o usuário tem a opção de iniciar em seguida o programa ERB-ArrayTools. Automaticamente, o Matlab é executado inicializando as variáveis: **`expIdent`** que irá conter a identificação dos dados geográficos, **`expData`** que irá conter a matriz destes dados e **`expDesc`** que irá conter o descritor de dados. Estas variáveis seguem o formato apresentado na subseção 4.2.1 do capítulo anterior. O sistema separa a tabela com o conjunto de dados em duas variáveis, a primeira coluna das coordenadas é atribuída a variável **`expIdent`** e as demais colunas é atribuída a variável **`expData`**.

5.2. Funcionalidades do ERB-ArrayTools

O ERB-ArrayTools é um sistema agregador que interage com o usuário e integra os demais subsistemas. Nesta integração diversas formas de visualização, foram disponibilizadas, dentre elas, uma utiliza sub-janelas para melhor usabilidade. Na Figura 5.4, são apresentadas as funcionalidades da interface gráfica com o usuário (GUI) do ERB-ArrayTools mapeada no processo de KDD.

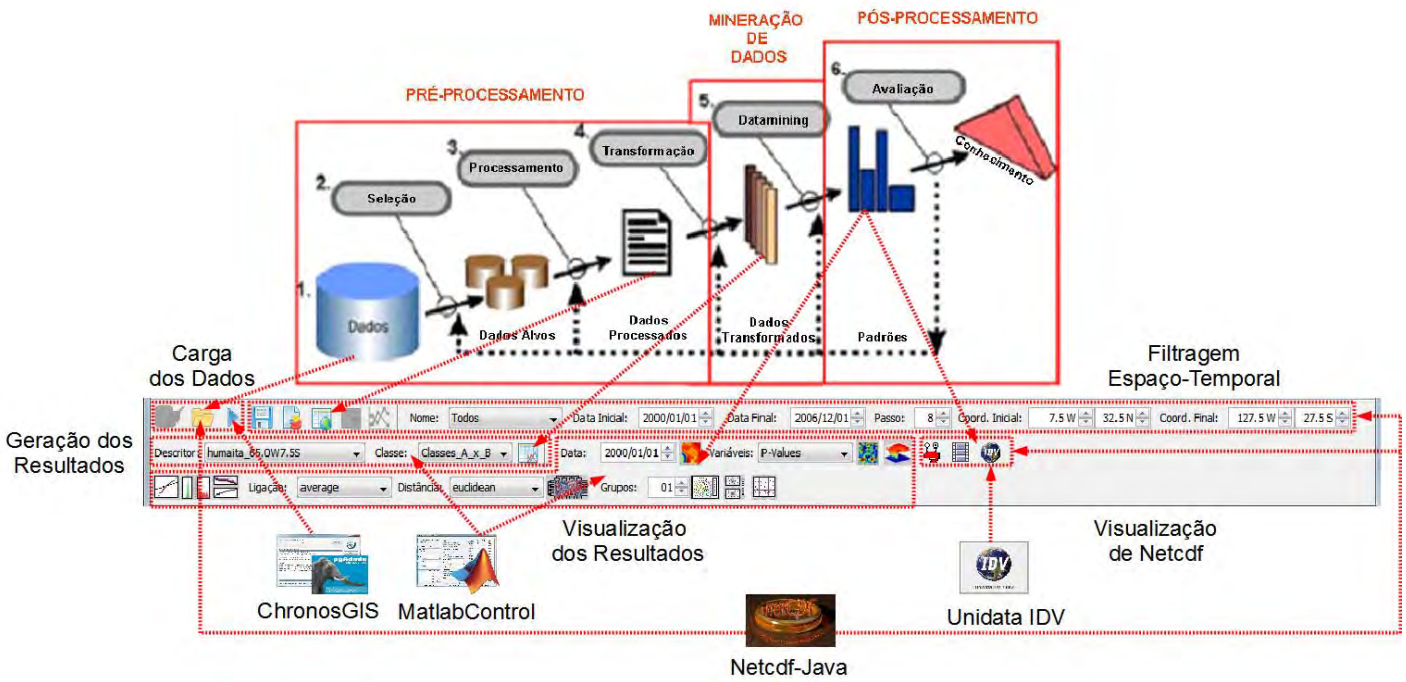


Figura 5.4 – Interface gráfica do ERB-ArrayTools mapeada para o KDD.

Basicamente a sequencia de atividades do ERB-ArrayTools se aproxima do fluxo de trabalho de um especialista do domínio do processo de KDDG, o qual segue as seguintes etapas:

- Filtrar os dados ambientais conforme sua dimensão espaço-temporal;
- Realizar o pré-processamento dos dados brutos de arquivos NetCDF (.nc) ou de uma tabela em arquivo texto (.txt) aplicando a normalização ou anomalia;

- Inserir os dados pré-processados obtidos a partir de arquivos NetCDF no banco de dados geográfico. Estes dados deverão compor, por meio de consulta SQL, o conjunto de dados a ser enviados para a próxima fase de DM.
- Obter de arquivo texto o Descritor de Dados com as classes a serem aplicadas no método de classificação da fase de DM ou defini-lo a partir de série temporal do conjunto de dados pré-processados.
- Calcular os p-valores usando o método t-teste na fase de DM.
- Salvar os p-valores obtidos em arquivo texto;
- Exportar os dados pré-processados juntamente com os p-valores obtidos no formato NetCDF;
- Visualizar o conjunto de dados das variáveis e os seus p-valores obtidos, tanto em forma de tabela quanto em forma de grade;
- Visualizar os arquivos NetCDF's obtidos com as variáveis ambientais e seus respectivos p-valores sobre um mapa com a localização correspondente.

Estas atividades são apoiadas pelos principais botões do ERB-ArrayTools apresentados na Figura 5.5.

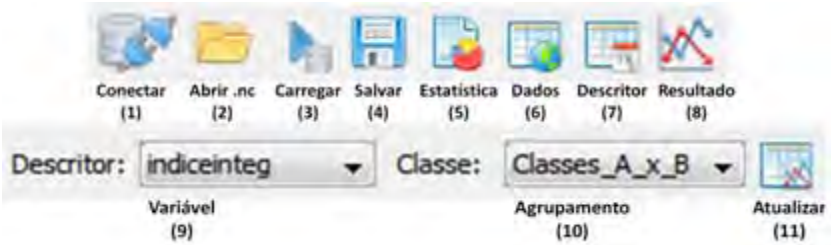


Figura 5.5 – Principais botões e parâmetros do ERB-ArrayTools.

A seguir será apresentado o fluxo de execução do sistema abordando as três fases tradicionais do processo KDD: pré-processamento, mineração de dados e pós-processamento.

5.2.1. Pré-processamento

Após a instalação e execução do sistema, o usuário deve pressionar o botão **Conectar** para abrir a tela de conexão com o BDG no SGBD PostgreSQL, como mostrada na Figura 5.6, informando a URL, o nome do usuário e senha. Ao fazer conexão com o banco de dados, as variáveis existentes no BDG são carregadas na tela principal.

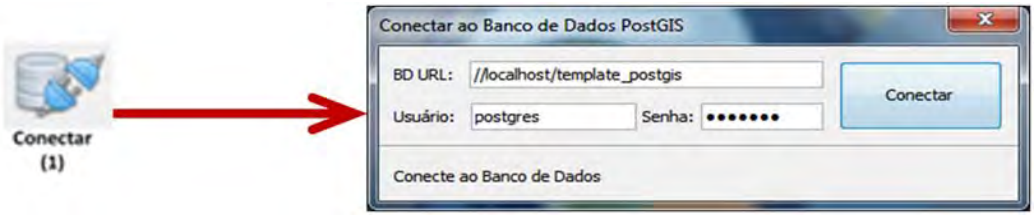


Figura 5.6 – Tela de conexão com o BDG.

Para inserir dados geográficos, o usuário deve pressionar o botão **AbrirNetCDF** e escolher os arquivos NetCDF's a serem carregados no BDG. Em seguida uma subjanela é aberta mostrando a estrutura deste arquivo como demonstrado na Figura 5.7.

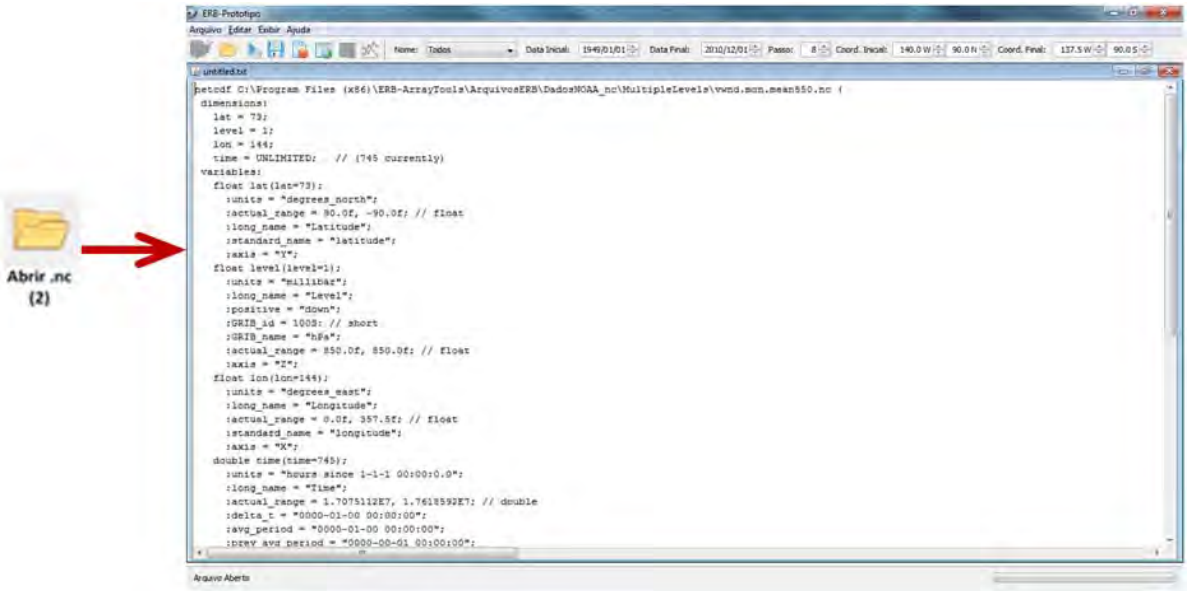


Figura 5.7 – Tela obtida após abrir o arquivo NetCDF.

Para este trabalho, foram utilizados arquivos NetCDF obtidos do site da NOAA, o qual permite ao usuário escolher, configurar e baixar estes arquivos na dimensão desejada. Desta forma, o ERB-ArrayTools aceita a inserção de conjunto de dados ambientais tanto de superfície quanto de multinível. Entretanto, permite a abertura e inserção de arquivos NetCDF's configurados com apenas um nível de cada vez.

O próximo passo é informar as datas inicial e final, a granularidade espacial desejada e as coordenadas inicial e final para a realização do processo de carga dos dados geográficos no BDG.

Ao pressionar o botão **Carregar**, abre-se uma janela para que seja informada a localização de onde será salvo o arquivo NetCDF pré-processado. A janela de opções permite ao usuário escolher entre criar anomalia ou apenas normalizar os dados brutos. Em seguida, uma tela de diálogo informa ao usuário que o arquivo pré-processado foi gerado com sucesso e o sistema inicia o processo de carga dos dados.

Este processo consiste em criar tabelas no BDG, caso não existam, e inserir ou atualizar os dados pré-processados neste banco de dados. O usuário pode acompanhar a qualquer momento este processo pela barra de *status*. No termino da carga dos dados, uma tela de diálogo informa que o processo foi concluído e a tela principal é atualizada com a nova variável ambiental inserida no BDG. Na Figura 5.8 está representado o fluxo de execução deste processo de carga no BDG.

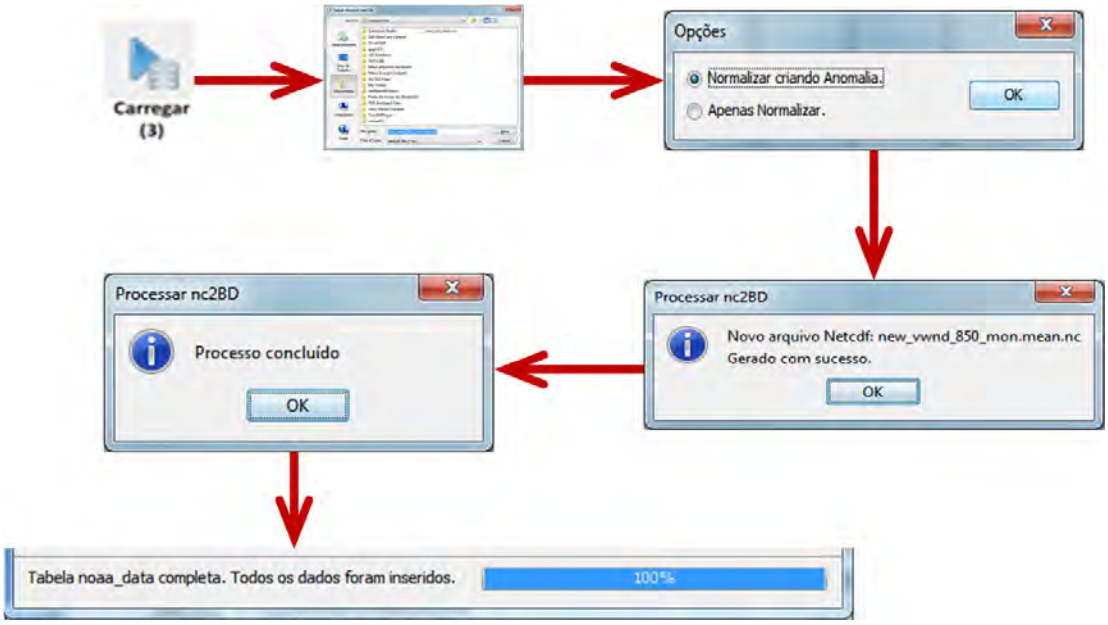


Figura 5.8 – Fluxo de telas ao realizar a carga no BDG.

Caso o usuário volte a realizar a carga dos dados da mesma variável com anomalia, o programa verifica na tabela do banco de dados o campo desta variável a fim de informar ao usuário que estes dados pré-processados já existem na BDG evitando, assim, a carga do mesmo conjunto de dados novamente. Da mesma forma, esta verificação é realizada para o caso da carga do mesmo conjunto de dados apenas normalizado.

Ao pressionar no botão **Estatística**, o sistema efetua cálculos estatísticos utilizando sentenças SQL sobre dados ambientais com eventual filtragem espaço-temporal realizada pelo usuário. O resultado da SQL é impresso em forma de texto em uma subjanela como demonstrado na Figura 5.9. Este resultado pode ser eventualmente salvo em arquivo texto pelo botão **Salvar**. Isto permite ao especialista de domínio maior flexibilidade e visibilidade estatística sobre a massa de dados.

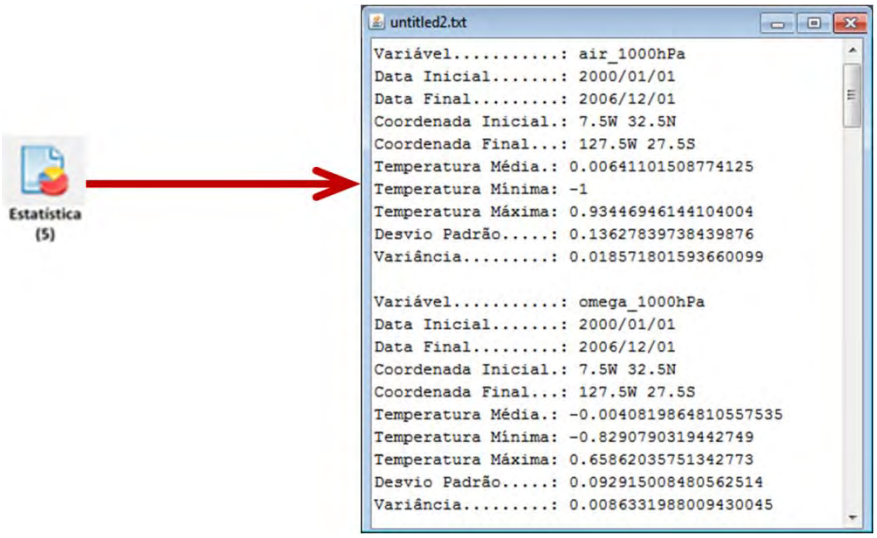
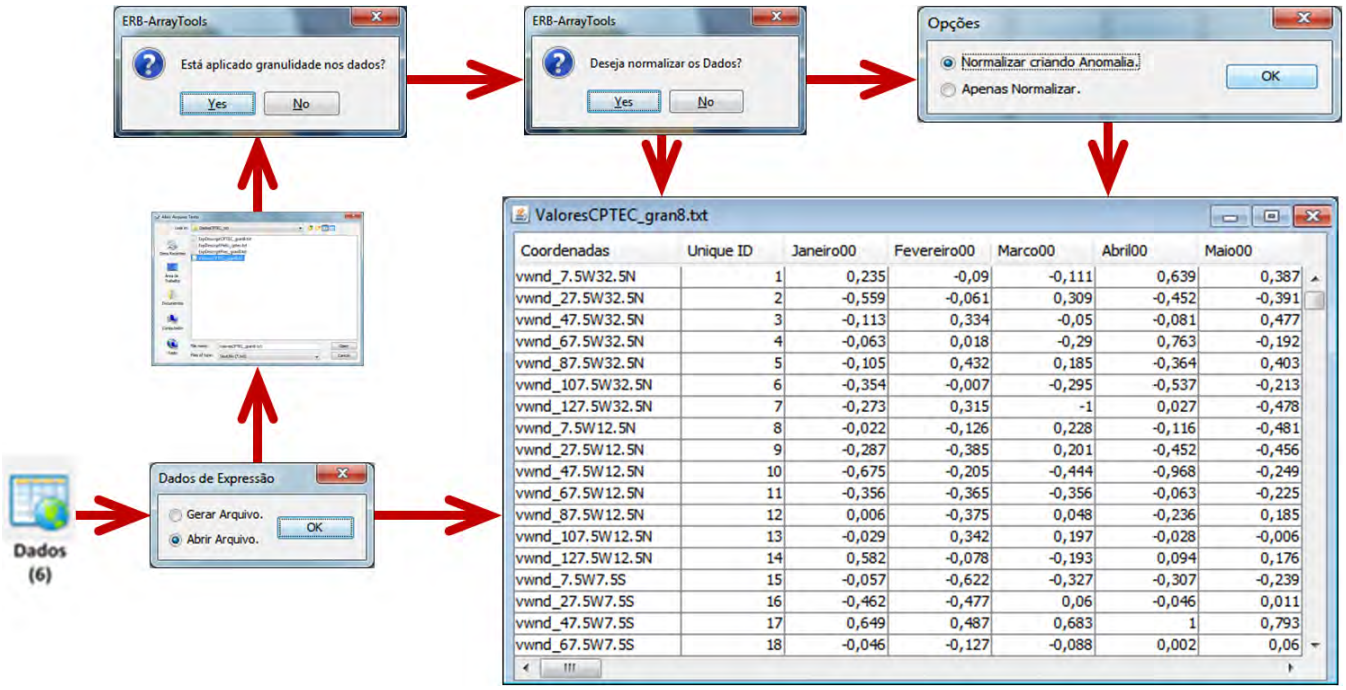


Figura 5.9 – Subjanela com resultados da SQL usada.

Ainda nesta etapa de pré-processamento, o usuário escolhe se deseja obter o conjunto de dados pré-processados do BDG ou de um arquivo texto ao pressionar no botão **Dados**. Casos escolha abrir um arquivo texto externo, o usuário será questionado se deseja aplicar a granularidade e, em seguida, se deseja normalizar os dados.

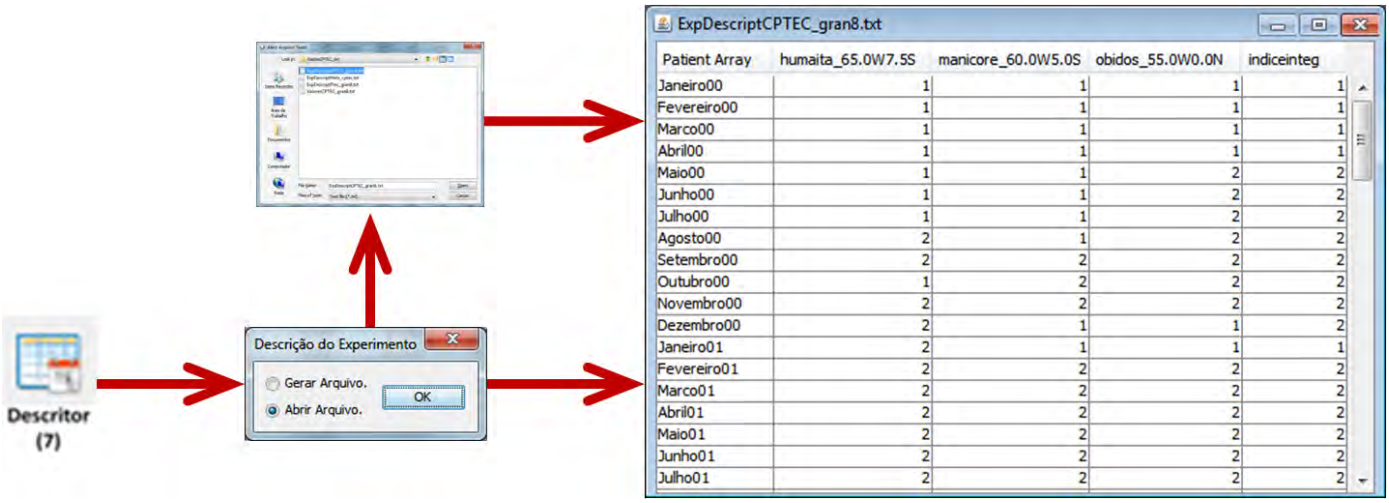
Em seguida, os dados são mostrados na tela em forma de uma tabela espaço-temporal onde no cabeçalho das colunas estão os meses com seus respectivos anos, na primeira coluna está a descrição das variáveis ambientais deste conjunto e na segunda coluna está o id único para cada linha da matriz de dados. Esta tabela de dados poderá ser salva pelo usuário em arquivo texto neste mesmo formato. A Figura 5.10 representa o fluxo para obtenção deste conjunto de dados.



Coordenadas	Unique ID	Janeiro00	Fevereiro00	Marco00	Abril00	Maio00
vwnd_7.5W32.5N	1	0,235	-0,09	-0,111	0,639	0,387
vwnd_27.5W32.5N	2	-0,559	-0,061	0,309	-0,452	-0,391
vwnd_47.5W32.5N	3	-0,113	0,334	-0,05	-0,081	0,477
vwnd_67.5W32.5N	4	-0,063	0,018	-0,29	0,763	-0,192
vwnd_87.5W32.5N	5	-0,105	0,432	0,185	-0,364	0,403
vwnd_107.5W32.5N	6	-0,354	-0,007	-0,295	-0,537	-0,213
vwnd_127.5W32.5N	7	-0,273	0,315	-1	0,027	-0,478
vwnd_7.5W12.5N	8	-0,022	-0,126	0,228	-0,116	-0,481
vwnd_27.5W12.5N	9	-0,287	-0,385	0,201	-0,452	-0,456
vwnd_47.5W12.5N	10	-0,675	-0,205	-0,444	-0,968	-0,249
vwnd_67.5W12.5N	11	-0,356	-0,365	-0,356	-0,063	-0,225
vwnd_87.5W12.5N	12	0,006	-0,375	0,048	-0,236	0,185
vwnd_107.5W12.5N	13	-0,029	0,342	0,197	-0,028	-0,006
vwnd_127.5W12.5N	14	0,582	-0,078	-0,193	0,094	0,176
vwnd_7.5W7.5S	15	-0,057	-0,622	-0,327	-0,307	-0,239
vwnd_27.5W7.5S	16	-0,462	-0,477	0,06	-0,046	0,011
vwnd_47.5W7.5S	17	0,649	0,487	0,683	1	0,793
vwnd_67.5W7.5S	18	-0,046	-0,127	-0,088	0,002	0,06

Figura 5.10 – Dados Pré-processados para Mineração.

Ao pressionar no botão **Descritor**, o usuário poderá abrir o descritor de dados a partir de um arquivo texto ou criá-lo a partir do conjunto de dados do BDG ou do arquivo. Com isso, este descritor é visualizado como demonstrado no fluxo da Figura 5.11. Ao obter o descritor, também poderá salvá-lo em arquivo.



Patient Array	humaita_65.0W7.5S	manicore_60.0W5.0S	obidos_55.0W0.0N	indiceinteg
Janeiro00	1	1	1	1
Fevereiro00	1	1	1	1
Marco00	1	1	1	1
Abril00	1	1	1	1
Maio00	1	1	2	2
Junho00	1	1	2	2
Julho00	1	1	2	2
Agosto00	2	1	2	2
Setembro00	2	2	2	2
Outubro00	1	2	2	2
Novembro00	2	2	2	2
Dezembro00	2	1	1	2
Janeiro01	2	1	1	1
Fevereiro01	2	2	2	2
Marco01	2	2	2	2
Abril01	2	2	2	2
Maio01	2	2	2	2
Junho01	2	2	2	2
Julho01	2	2	2	2

Figura 5.11 – Descritor de Dados para Mineração.

5.2.2. Mineração de Dados

Ao pressionar no botão **Resultados**, surgem as barras de ferramentas de visualização e atualização dos resultados e o Matlab é acionado para realizar a Mineração de Dados utilizando o conjunto de dados pré-processados e o descritor de dados definidos na fase anterior. A Figura 5.12 mostra duas telas do Matlab: a primeira contém os códigos com as funções responsáveis por aplicar a permutação do método estatístico t-teste e a segunda contém as variáveis utilizadas por estas funções.

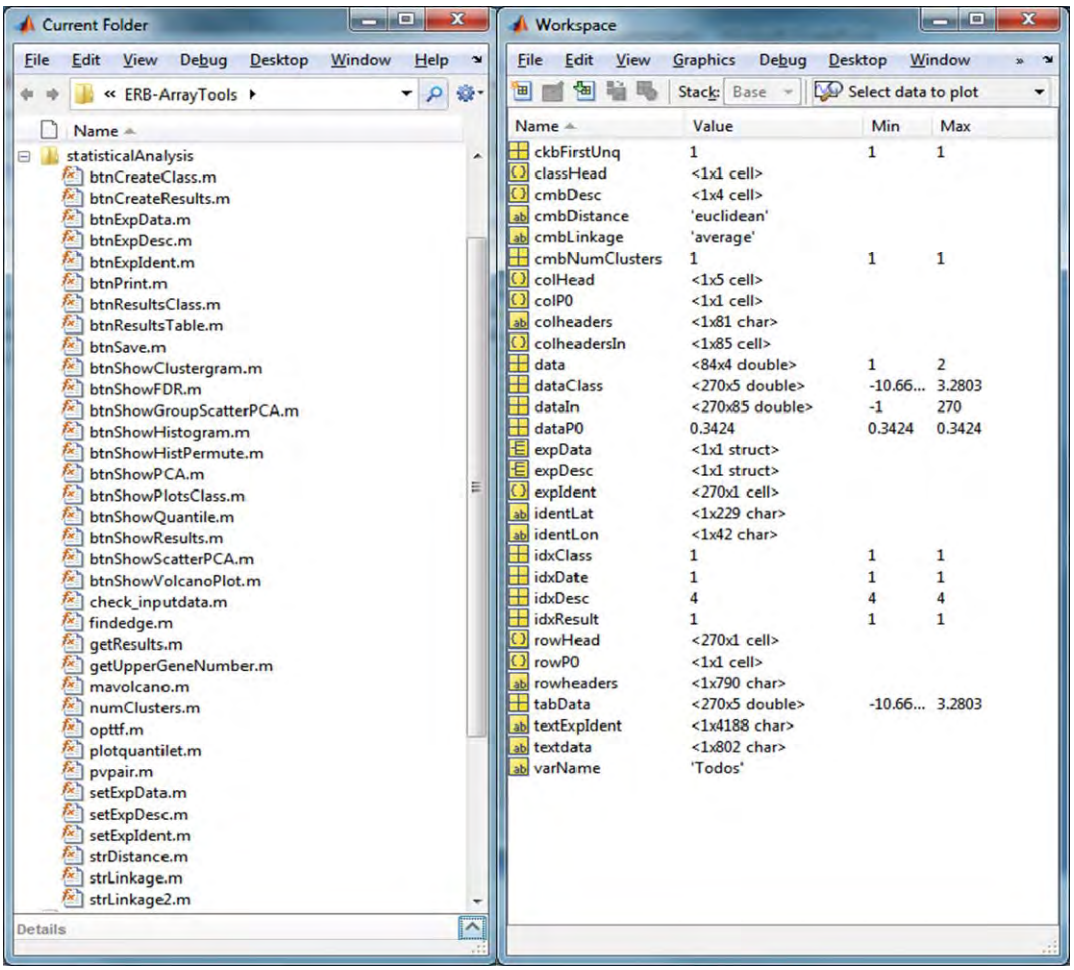
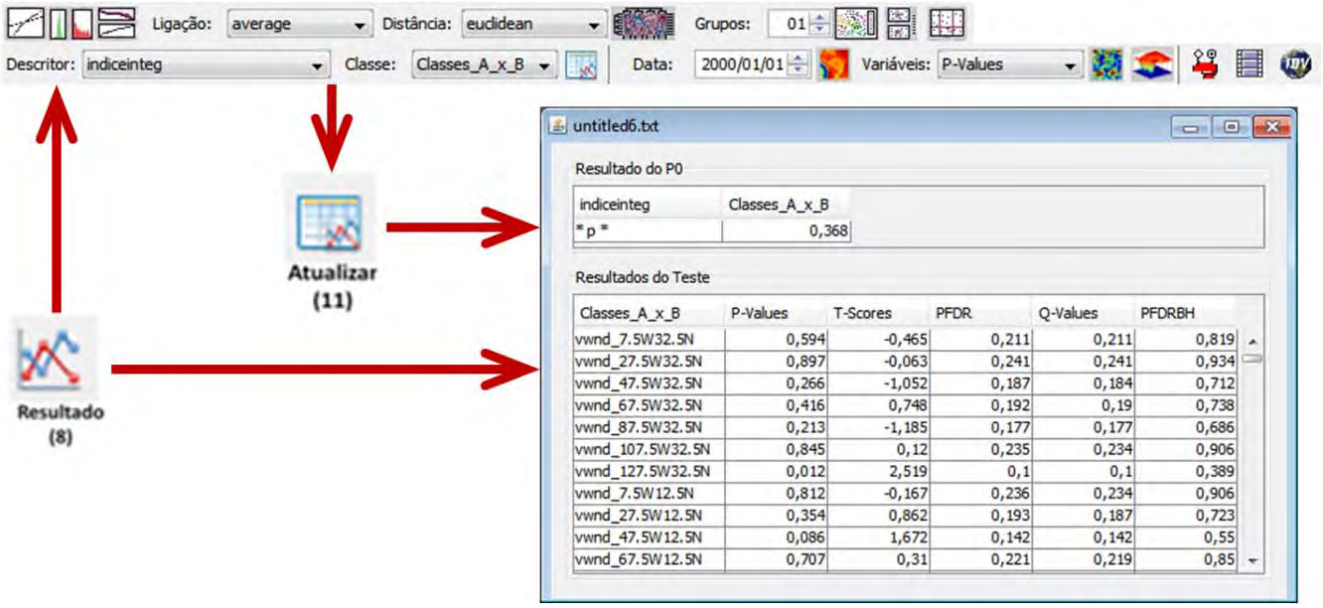


Figura 5.12 – Telas do Matlab com os *scripts* e as variáveis definidas.

Na barra de atualização, há o botão **Atualizar** com funcionalidade semelhante ao botão **Resultados**, o qual é utilizado para recalcular os resultados obtidos ou calcular novos resultados caso se escolha outra **Variável** do descritor e outro **Agrupamento**. Feita a mineração, os resultados estatísticos são obtidos e visualizados numa subjanela como mostrado na Figura 5.13.



Classes_A_x_B	P-Values	T-Scores	PFDR	Q-Values	PFDRBH
vwnd_7.5W32.5N	0,594	-0,465	0,211	0,211	0,819
vwnd_27.5W32.5N	0,897	-0,063	0,241	0,241	0,934
vwnd_47.5W32.5N	0,266	-1,052	0,187	0,184	0,712
vwnd_67.5W32.5N	0,416	0,748	0,192	0,19	0,738
vwnd_87.5W32.5N	0,213	-1,185	0,177	0,177	0,686
vwnd_107.5W32.5N	0,845	0,12	0,235	0,234	0,906
vwnd_127.5W32.5N	0,012	2,519	0,1	0,1	0,389
vwnd_7.5W12.5N	0,812	-0,167	0,236	0,234	0,906
vwnd_27.5W12.5N	0,354	0,862	0,193	0,187	0,723
vwnd_47.5W12.5N	0,086	1,672	0,142	0,142	0,55
vwnd_67.5W12.5N	0,707	0,31	0,221	0,219	0,85

Figura 5.13 – Subjanela com resultados obtidos.

No parâmetro **Agrupamento**, mostrado na Figura 6.5, define-se o critério de agrupamento de classes. Como por exemplo, no parâmetro classes_A_x_B temos duas classes, A e B, cujos descritores estão respectivamente, abaixo e a cima da mediana. Neste caso em particular, foram definidas apenas duas classes com o surgimento de um único parâmetro de agrupamento de classes A versos B para a geração de apenas uma tabela de p-valores.

Caso haja mais de duas classes no descritor ao se aplicar algum outro critério, poderá ser criado mais de um parâmetro de agrupamento de classes para a obtenção de mais de uma tabela de p-valores. Por exemplo, o surgimento de três tabelas de p-valores diferentes ao utilizar os parâmetros de agrupamentos classes_A_x_B, classes_A_x_C, classes_B_x_C.

5.2.3. Pós-processamento

Ao pressionar o botão **Salvar**, o usuário define o nome da tabela de resultados estatísticos a ser salvo em arquivo texto pelo Matlab. Em seguida o ERB-ArrayTools exporta o conjunto de dados pré-processado de cada variável ambiental com seus respectivos resultados estatísticos para um arquivo NetCDF. O nome deste arquivo será composto pelos nomes: (1) da tabela de resultados definido pelo usuário, (2) da variável, (3) do Descritor de Dados usado e (3) do tipo de agrupamento utilizado no processo de mineração. Um exemplo é o nome resultados_vwnd_indiceinteg_a_x_b.nc. As telas da Figura 5.14 informam que os arquivos obtidos foram salvos com sucesso.

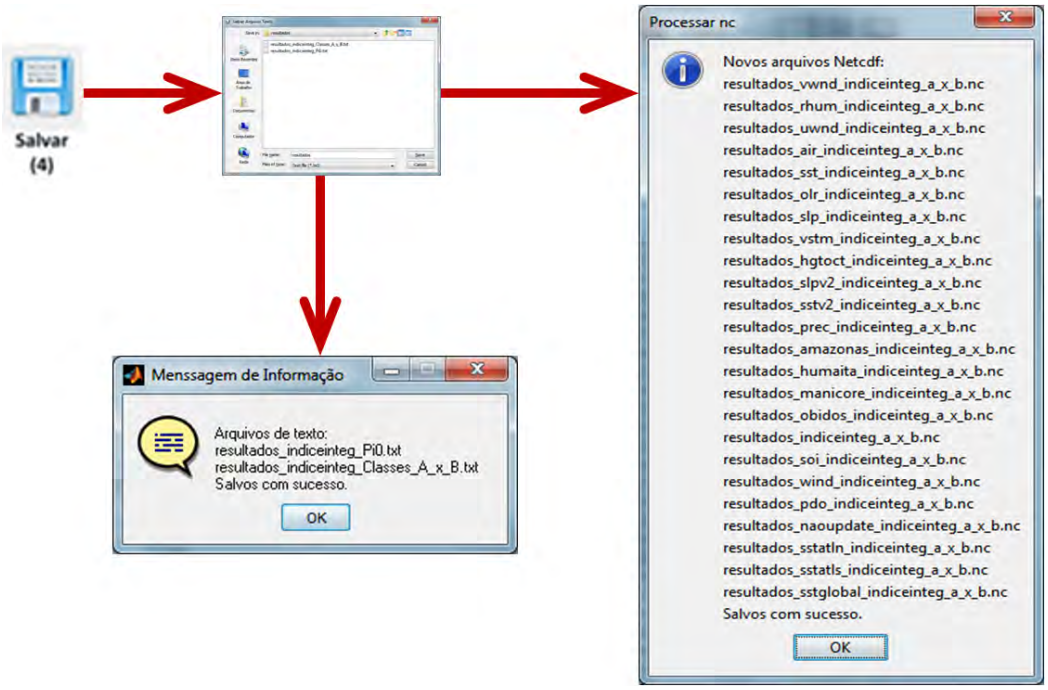


Figura 5.14 – Arquivos de texto e NetCDF's obtidos pela DM ao salvar.

Neste trabalho foi enfatizado na fase pós-processamento o uso das funcionalidades: (1) *Clustergram* do Matlab; (2) GridView e ImageView do subsistema ToolsUI; e a (3) visualização em 3D do subsistema IDV. As demais funcionalidades, mostradas na Figura 5.15, serão apresentadas no apêndice A.

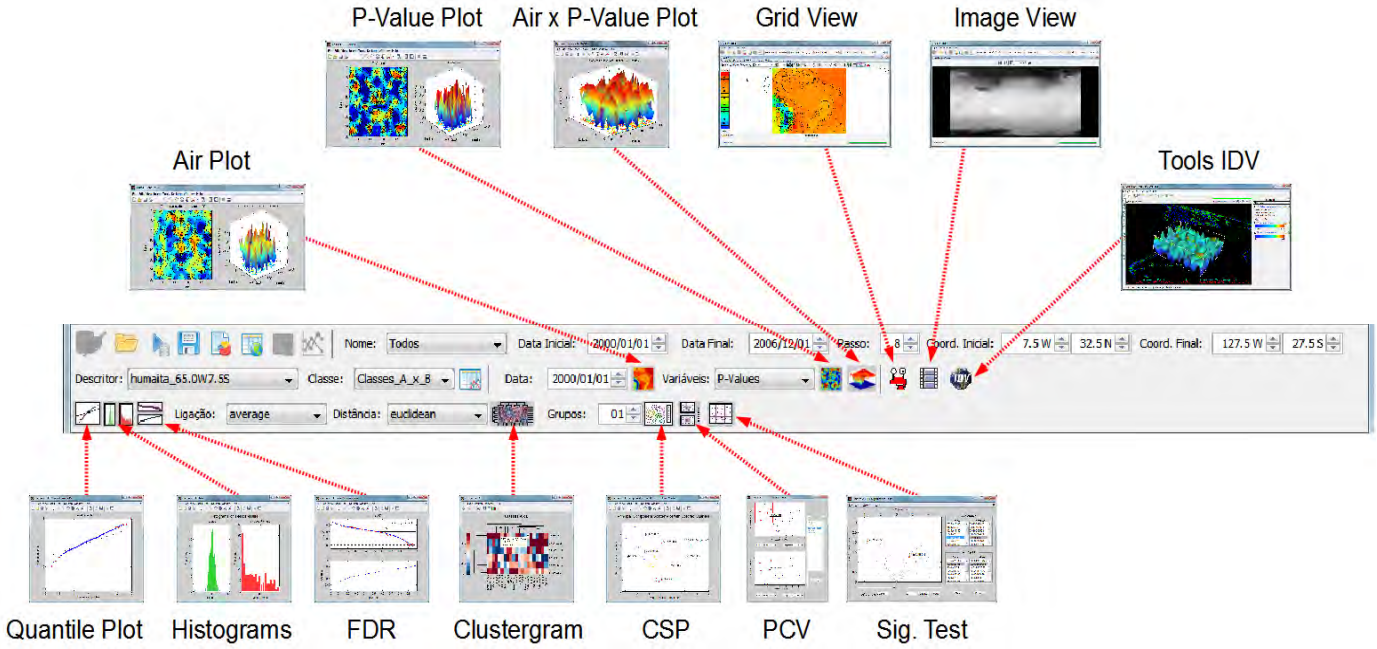


Figura 5.15 – Diferentes tipos de visualização.

Para o agrupamento utilizando o *Clustergram* é necessário definir os parâmetros de ligação e de distância a ser aplicado. Na janela *Clustergram*, mostrado na Figura 5.16, o usuário tem acesso a diversos recursos, entre eles habilitar a legenda e visualizar uma parte da matriz de dados.

Figura 5.16 – Índice Integrado agrupado pelo Matlab.

Com a utilização da funcionalidade GridView, o usuário tem a possibilidade de visualizar um arquivo NetCDF em forma de grade sobre um mapa. Nesta subjanela, mostrada na Figura 5.17, o usuário pode escolher se deseja visualizar o conjunto de dados pré-processado da variável ambiental ou algum resultado estatístico, como por exemplo, os p-valores calculados a partir desta variável. Além disso, esta funcionalidade permite visualizar a variável ambiental ao longo de sua série-temporal e em um determinado nível.

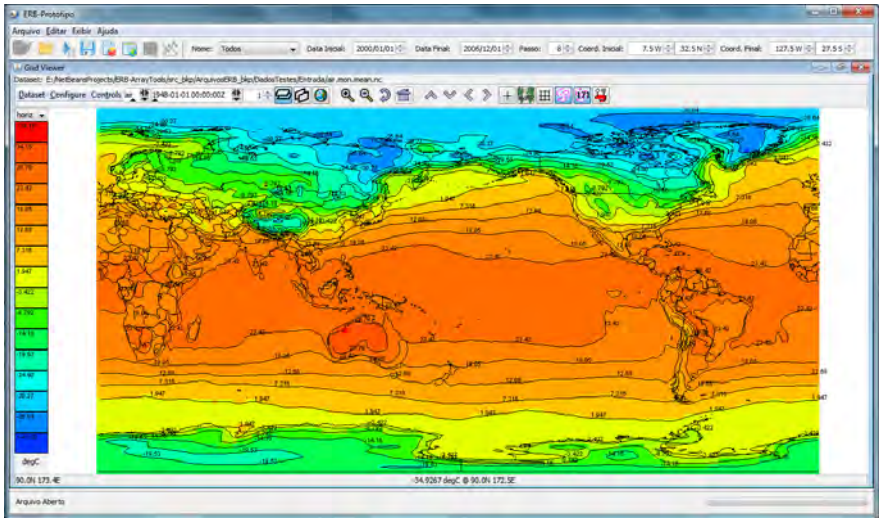


Figura 5.17 – Variável air visualizada no GridView do subsistema ToolUI.

A funcionalidade ImageView, mostrada na Figura 5.18, permite o usuário visualizar apenas a variável ambiental de um arquivo NetCDF em escala de cinza ao logo de sua série temporal. Por isso, esta funcionalidade não permite a visualização das novas variáveis estatísticas dos arquivos NetCDF's produzidos pelo ERB-ArrayTools.

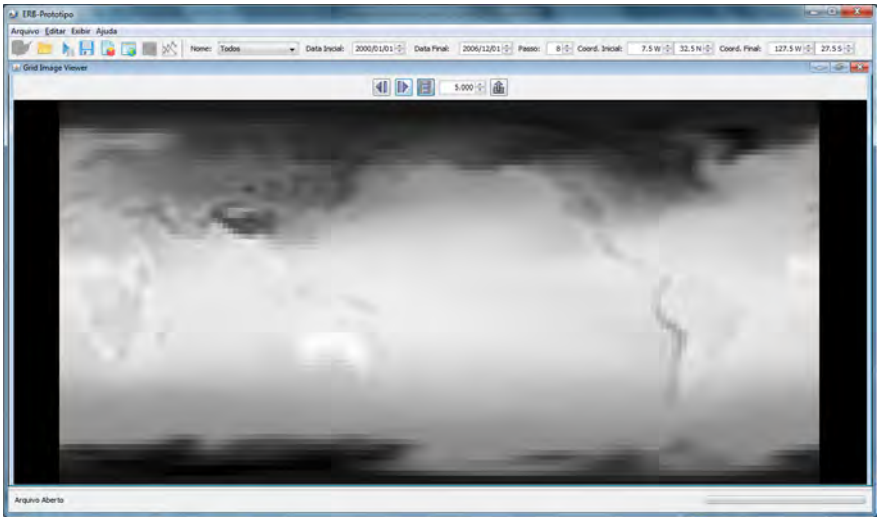


Figura 5.18 – Variável air visualizada no ImageView do subsistema ToolUI.

Ao pressionar o botão IDV, o usuário invoca a ferramenta *Integrated Data Viewer*, mostrada na Figura 5.19, a qual contém, dentre diversas funcionalidades, o recurso de visualização em três dimensões. Pelo menu *File*, o usuário abre o arquivo NetCDF para o armazenamento que permitirá a visualização dos dados geográficos sobre mapa. Pelo menu *Data*, o usuário escolhe as variáveis listando-as ao lado do mapa. Utilizando esta lista, há possibilidade da visualização sobreposta de duas ou mais variáveis e em três dimensões ao longo de uma série-temporal.

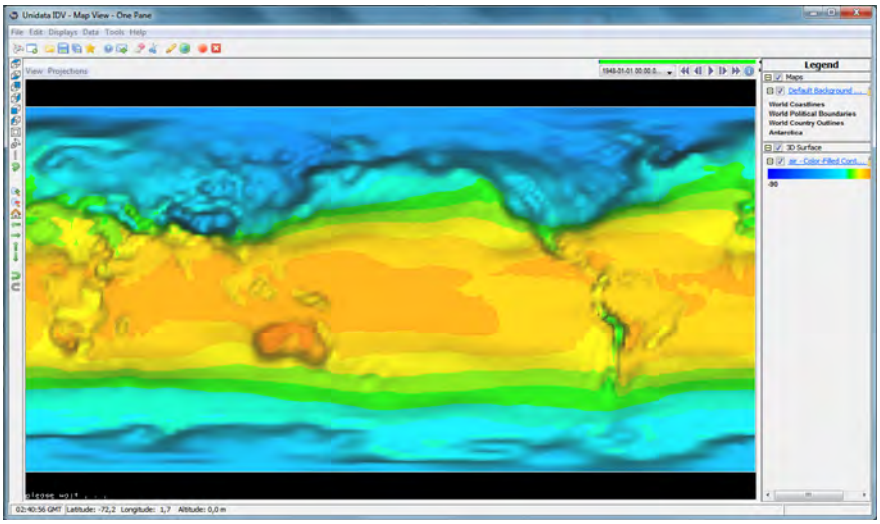


Figura 5.19 – Variável air visualizada em 3D do subsistema IDV.

6 ESTUDO DE CASO

Neste capítulo, é apresentado um estudo de caso que não enfatiza a precisão da análise de eventos ambientais extremos, mas sim o funcionamento de um ferramental para um especialista deste domínio.

6.1. Objetivos, Hipóteses e Planejamento do Estudo de Caso

O objetivo deste estudo de caso é fundamentalmente apresentar as visualizações do conjunto de dados de entrada e dos p-valores obtidos avaliando o desempenho do sistema ERB-ArrayTools em termos do tempo gasto para a realização do processo de KDD de forma automatizada comparado ao processo manual realizado por Ruivo (2008).

Este processo manual consistia em realizar: (1) o pré-processamento detalhado na subseção 4.2.1 do capítulo 4 para um grande conjunto de dados brutos utilizando o programa *Excel* do sistema operacional *Windows*; (2) a mineração de dados utilizando o *add-in* BRB-ArrayTools adicionado ao *Excel*; e (3) o pós-processamento para a visualização dos p-valores utilizando as funcionalidades do BRB-ArrayTools, a ferramenta GRADS (2013) e entre outras. Além disso, também havia a necessidade de alternância de sistema operacional entre *Windows* e *Linux* para a utilização de diferentes ferramentas para uma boa análise destes resultados.

O sistema ERB-ArrayTools foi concebido inicialmente sobre a hipótese de que métodos individualmente empregados por um operador do processo de KDD podem ser formalizados numa sequência de atividades e encapsuladas no código de uma ferramenta. Com isso, aumentaria eficiência do processo de descoberta de conhecimento resultando na diminuição do tempo entre a entrada dos dados, pré-processamento, mineração e pós-processamento se comparado a este processo de forma manual.

Este estudo de caso foi planejado da seguinte forma:

- Utilizar um conjunto de dados de entrada referente ao período da seca ocorrida em 2005 na Amazônia, sendo apresentado na seção 6.2;
- Pré-processar este conjunto de dados e visualizar as variáveis ambientais Vento Meridional, Temperatura do Ar na Superfície e Temperatura da Superfície do Mar utilizando as funcionalidades GridView, ImageView e IDV, sendo apresentadas na seção 6.3;
- Minerar este conjunto pré-processado e realizar o pós-processamento para a visualização dos p-valores destas variáveis utilizando as funcionalidades *Clustergram*, GridView e IDV, sendo apresentados na seção 6.4;
- Apresentar, finalmente, os resultados obtidos a um especialista no domínio da aplicação e comparar qualitativamente o tempo obtido da execução automatizada do processo de KDD pelo sistema ERB-ArrayTools em relação ao processo manual realizado em Ruivo (2008), mais detalhes na seção 6.5.

6.2. Delimitação do Conjunto de Dados para Estudo de Caso

Como mencionado anteriormente, o conjunto de dados utilizado na execução deste sistema contém dados globais de reanálise fornecidos pelo CPTEC/INPE em forma de tabela em um arquivo texto e utilizados no processo de análise em Ruivo (2008). As variáveis ambientais referentes a este conjunto são tomadas em nível de superfície com resolução espacial de 2.5º x 2.5º, somente a variável de temperatura da superfície do mar contém resolução de 1º x 1º.

Os dados são todos mensais e compreendem o período de janeiro de 2000 à dezembro de 2006 sendo um total de 84 meses. Esta matriz de dados compreende uma sub-região com coordenadas 140W a 0W, e 40N a 40S. Assim como em Ruivo (2008), foram calculados valores médios mensais de cada grandeza dentro da sub-região da Figura 6.1 em quadriláteros de 20º de longitude por 20º de latitude.

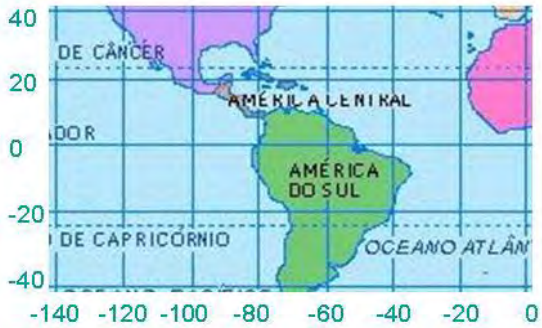


Figura 6.1 – Coordenadas de 0W à 140W e 40N à 40S.
Fonte: Ruivo (2008)

Para os nomes das variáveis ambientais que compõem este conjunto foram utilizadas as definições obtidas da NOAA, as quais são listadas a seguir:

- Altura Geopotencial (**`hgtoct`**) em algum ponto na atmosfera;
- Movimento Vertical (**`vstm`**) que define o componente vertical do vento (ômega) (CPTEC/INPE, 2006);
- Vento Zonal (**`uwnd`**) que define o componente ao redor dos círculos latitudinais do vento (WALLACE; HOBBS, 2006);
- Pressão Atmosférica ao Nível do Mar (**`slp`**);
- Radiação da Onda Longa Emergente (**`olr`**);
- Umidade Relativa (**`rhum`**);
- Vento Meridional (**`vwnd`**) na fatia norte-sul através da atmosfera (WALLACE; HOBBS, 2006);
- Temperatura do Ar na Superfície (**`air`**);
- Temperatura da Superfície do Mar (**`sst`**).

A variável ambiental **`hgtoct`** define o trabalho que deve ser feito contra o campo gravitacional da Terra para elevar uma massa de um quilograma do nível do mar até um determinado ponto (HOLTON, 2004). A altura geopotencial no nível do mar é zero.

6.3. Visualização do Conjunto de Dados de Entrada

Devido à restrição de espaço e escopo, serão apresentados neste trabalho, apenas as variáveis **`vwnd`**, **`air`**, **`sst`** e os seus respectivos p-valores. A Figura 6.2 mostra a variável ambiental **`vwnd`** utilizando os recursos de visualização em grade do subsistema ToolsUI à esquerda (a) pelo GridView e à direita (b) pelo ImageView; a Figura 6.3 mostra esta variável utilizando o recurso de visualização em 3D pelo subsistema IDV; analogamente, as Figuras 6.4 e 6.5 mostram a variável **`air`**; e as Figuras 6.6 e 6.7 mostram a variável **`sst`**.

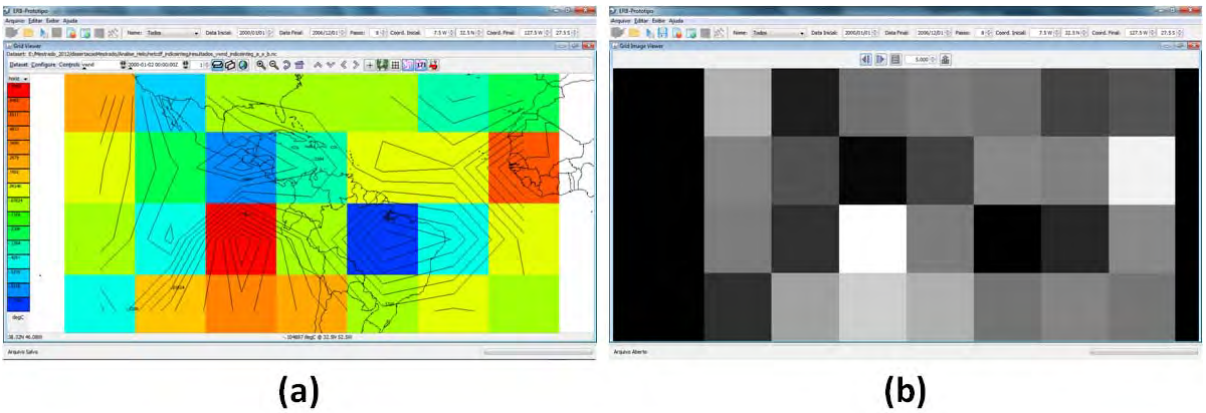


Figura 6.2 – Vento meridional visualizado pelo GridView e ImageView.

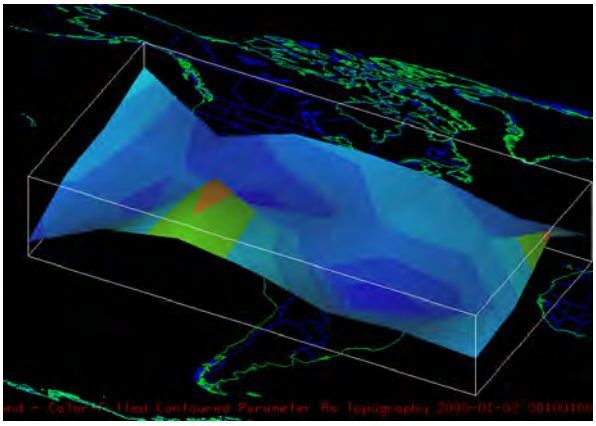


Figura 6.3 – Vento meridional visualizado em 3D pelo IDV.

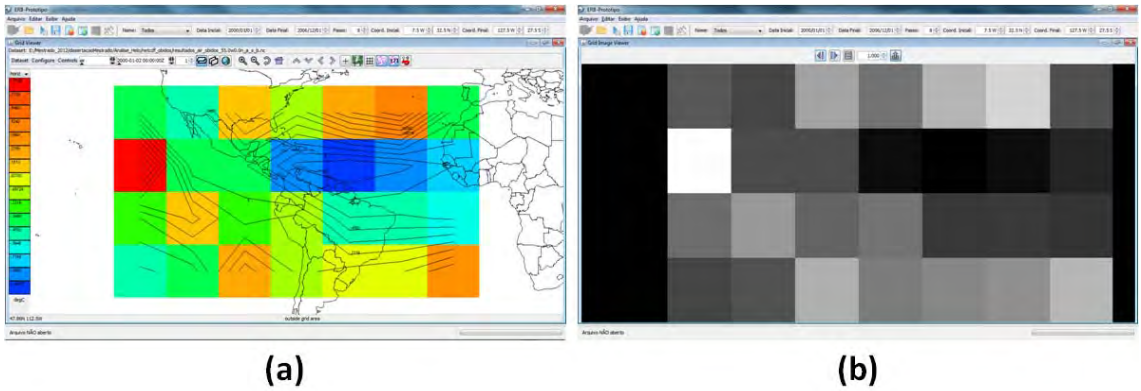


Figura 6.4 – Temperatura do ar na superfície pelo GridView e ImageView.

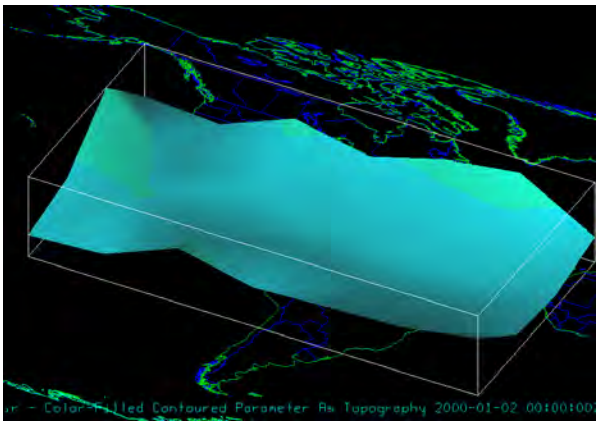


Figura 6.5 – Temperatura do ar na superfície em 3D pelo IDV.

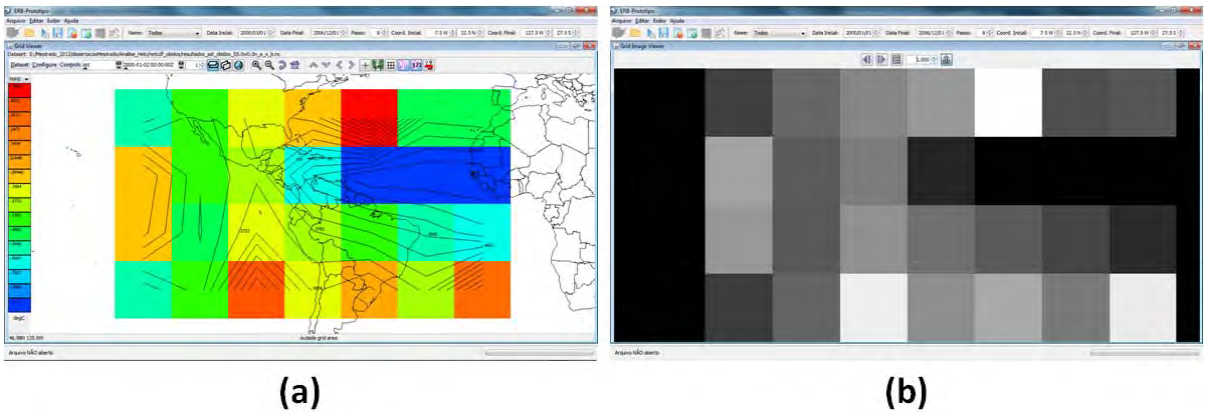


Figura 6.6 – Temperatura da superfície do mar pelo GridView e ImageView.

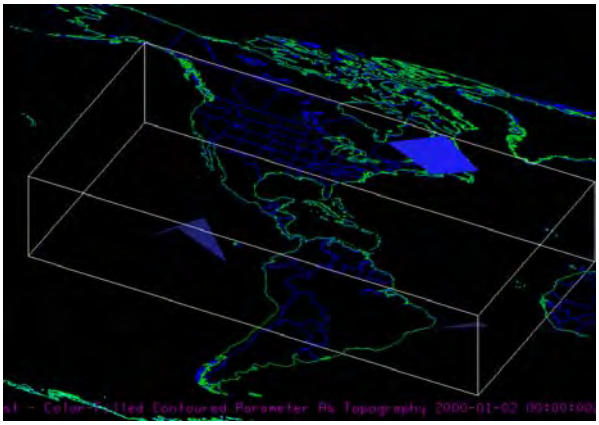


Figura 6.7 – Temperatura na superfície do mar em 3D pelo IDV.

6.4. Resultados Sumários do Processo de KDD

Aplicando a analogia de análise com Microarranjos, o Conjunto de Dados de Expressão foi organizado de forma que cada linha representa uma variável ambiental como se fosse um gene e cada coluna uma média mensal como se fosse um paciente. Em seguida, o Descritor de Dados com as “classes” foi definido a partir das séries temporais da vazão dos rios Madeira e Amazonas provenientes da rede hidrometeorológica da Agência Nacional de Águas (ANA), operada pelo Serviço Geológico do Brasil. As medidas destas séries são referentes aos municípios de Humaitá, Manicoré e Óbidos, como mostrados na Figura 6.8.

Figura 6.8 – Localização das três regiões analisadas.

Fonte: CPRM (2013) apud Ruivo (2008)

Estas séries são usadas para guiar o processo de classificação e agrupamento na DM a fim de extrair o conhecimento do conjunto de dados. Neste estudo de caso, foram utilizadas as propriedades levantadas por Ruivo (2008), a saber:

- A série temporal da vazão do índice integrado, a qual é definida pela média aritmética das séries das vazões do rio Madeira em Humaitá, Manicoré e o rio Amazonas em Óbidos;
- A série temporal da vazão do rio Amazonas em Óbidos, a qual contém a "informação" de vários afluentes inclusive o Madeira.

Assim como a NOAA definiu nomes para as variáveis, também optou-se por atribuir nomes destas e de outras séries temporais para a inclusão no conjunto de dados e geração dos NetCDF's com os dados pré-processados e seus respectivos p-valores. Para as sub-regiões abordadas foram mantidos seus nomes em letra minúscula e sem acento, ou seja, **`humaita`**, **`manicore`** e o **`obidos`**. Para o Índice Integrado foi definido **`indiceinteg`**.

Utilizando o subsistema *ProduceNetCDF*, foram obtidas duas classes após a manipulação dos dados desta série temporal sendo que uma está acima e outra abaixo da mediana. Definidas as classes, foi obtido o resultado do processo de agrupamento utilizando o Matlab através do subsistema *StatisticalAnalysis*.

Na matriz de agrupamento obtido pela funcionalidade *Clustergram* do Matlab, mostrada nas Figuras 6.9 e 6.10, a cor de cada quadrado indica que uma determinada variável ambiental está abaixo da mediana quando está com tons de azul ou acima da mediana quando está com tons de vermelho.

Da mesma forma que as cores, o número em um determinado quadrado indica o quanto a respectiva variável está acima ou abaixo da mediana de acordo com a legenda ao lado. Na Figura 6.9 está o agrupamento com as variáveis mais relevantes obtidos para a região do Índice integrado e na Figura 6.10 está o agrupamento para a região de Óbidos.

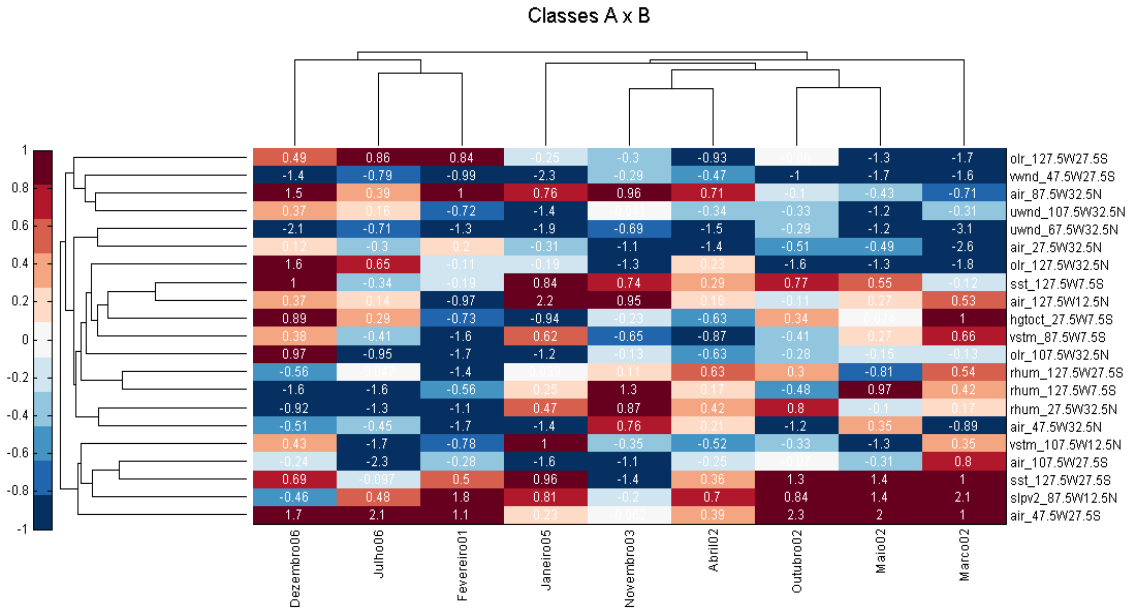


Figura 6.9 – Agrupamento no Índice Integrado pelo Matlab.

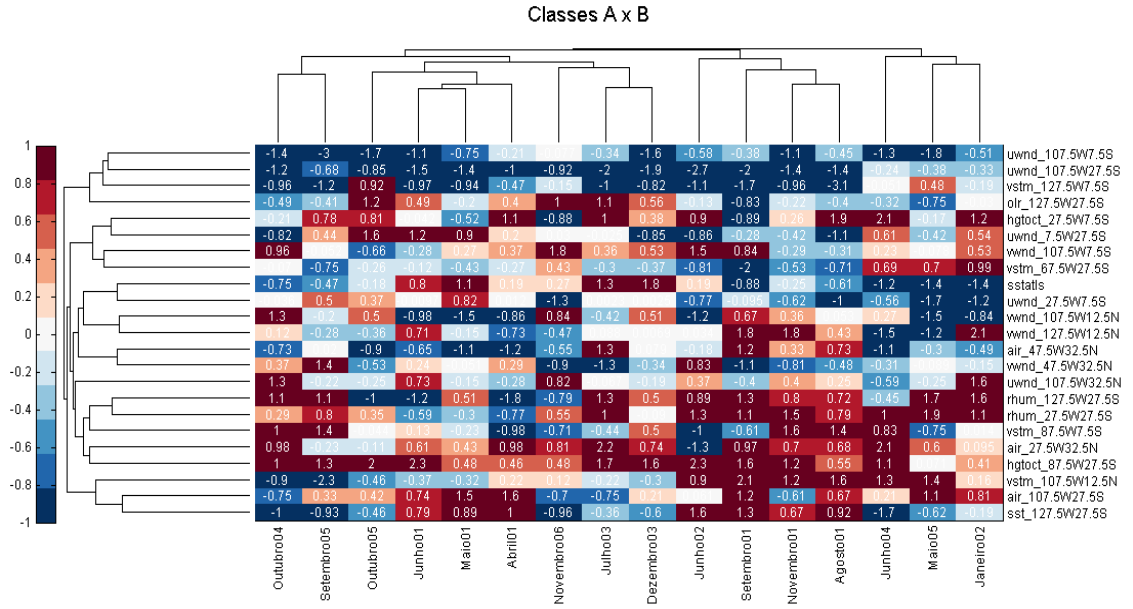
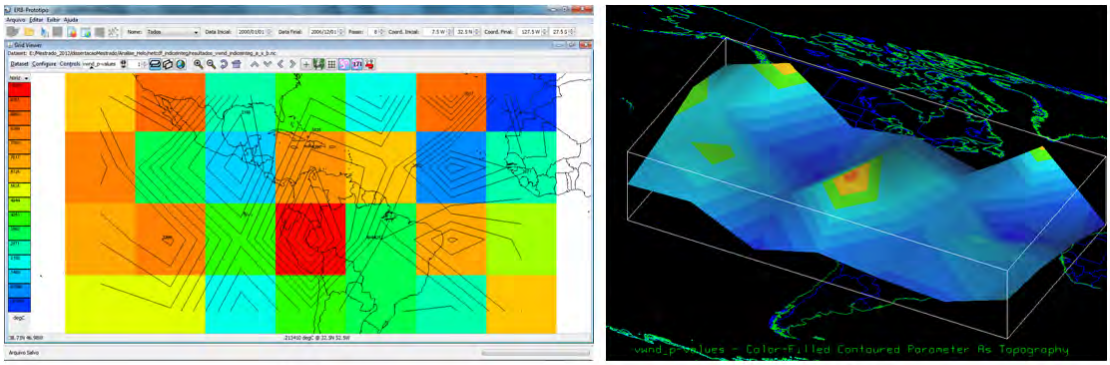


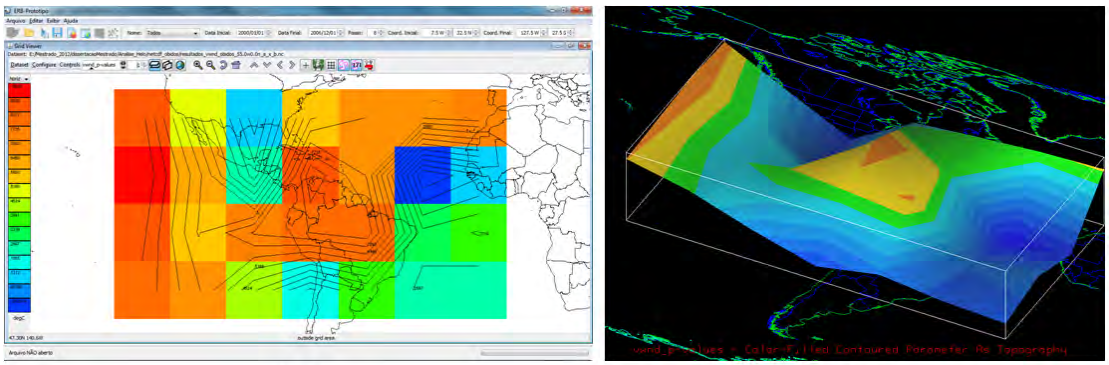
Figura 6.10 – Agrupamento em Óbidos pelo Matlab.

A Figura 6.11 apresenta os p-valores da variável **vwmd** no Índice Integrado em grade à esquerda (a) pelo GridView e em 3D à direita (b) pelo IDV. Analogamente, a Figura 6.12 apresenta os p-valores desta variável em Óbidos.



(a) (b)

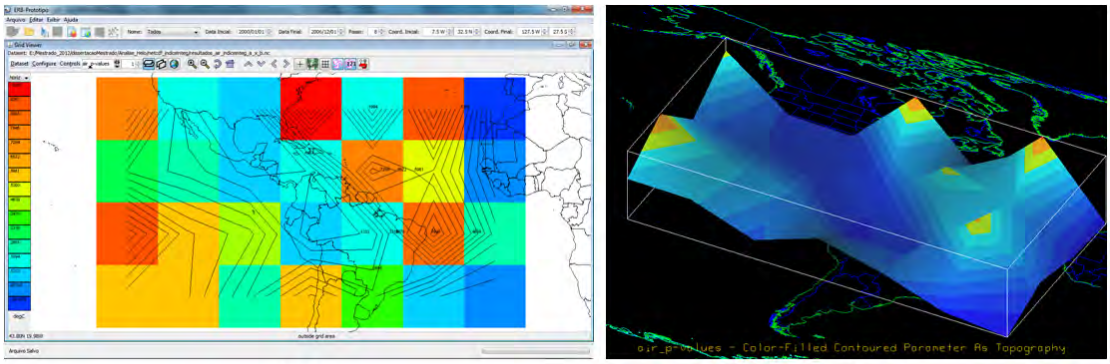
Figura 6.11 – P-valor de vwnd no Índice Integrado pelo GridView e IDV.



(a) (b)

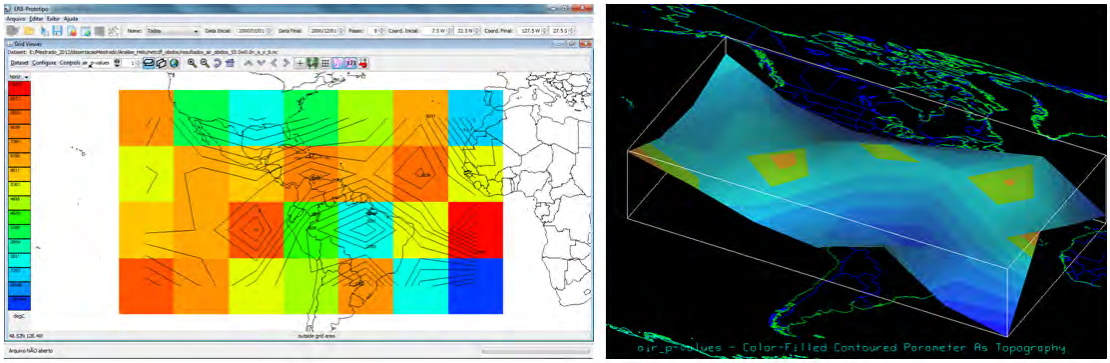
Figura 6.12 – P-valor de vwnd em Óbidos pelo GridView e IDV.

A Figura 6.13 apresenta os p-valores da variável **`air`** no Índice Integrado em grade à esquerda (a) pelo GridView e em 3D à direita (b) pelo IDV. Analogamente, a Figura 6.14 apresenta os p-valores desta variável em Óbidos.



(a) (b)

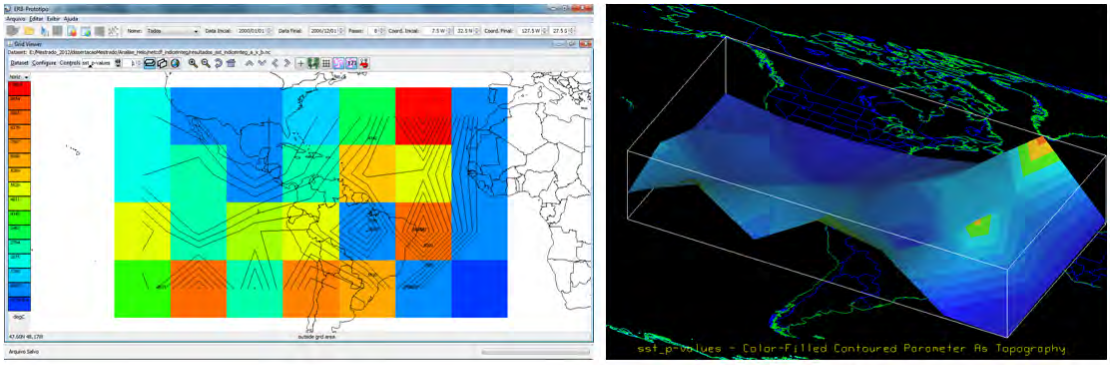
Figura 6.13 – P-valor de air no Índice Integrado pelo GridView e IDV.



(a) (b)

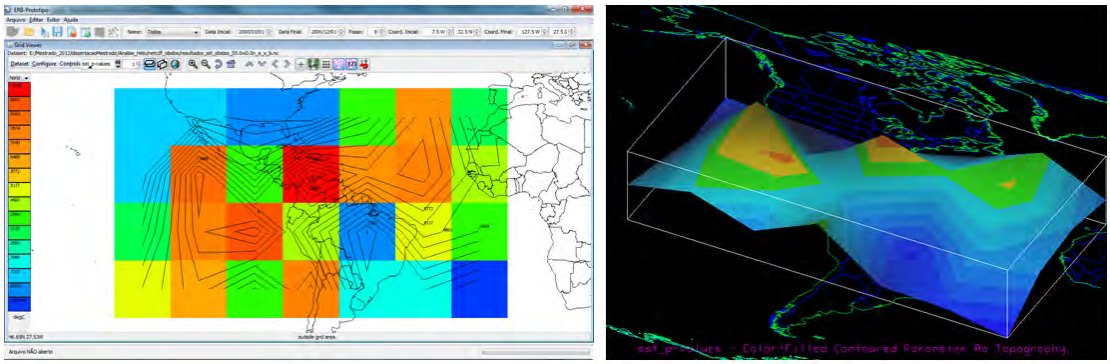
Figura 6.14 – P-valor de air em Óbidos pelo GridView e IDV.

A Figura 6.15 apresenta os p-valores da variável **sst** no Índice Integrado em grade à esquerda (a) pelo GridView e em 3D à direita (b) pelo IDV. Analogamente, a Figura 6.16 apresenta os p-valores desta variável em Óbidos.



(a) (b)

Figura 6.15 – P-valor de sst no Índice Integrado pelo GridView e IDV.



(a) (b)

Figura 6.16 – P-valor de sst em Óbidos pelo GridView e IDV.

6.5. Avaliação do Estudo de Caso utilizando o ERB-ArrayTools

O conjunto de dados ambientais utilizando neste trabalho foi o mesmo conjunto utilizado em Ruivo (2008) para a grande seca de 2005 na Amazônia. A utilização deste conjunto é importante para a obtenção de uma métrica qualitativa temporal da execução automatizada do processo de KDD comparado ao processo manualmente realizado por Ruivo (2008).

Mediante consulta ao especialista de domínio na área, foi obtida uma boa avaliação das diferentes visualizações das variáveis ambientais e dos resultados estatísticos obtidos pelo processo de KDD. Além disto, a avaliação dos resultados feita pelo especialista foi satisfatória, o que torna possível a avaliação de fenômenos.

Finalmente, a comparação qualitativa do tempo do processo de KDD automatizado pelo sistema ERB-ArrayTools em relação ao processo manual realizado em Ruivo (2008), mostra que os resultados foram obtidos em um espaço de tempo da ordem de minutos enquanto que em Ruivo (2008) estes foram da ordem de dias. Cabe ressaltar que, a comparação realizada em ambos os casos, considera apenas o tempo de manipulação dos dados no processo de KDD não englobando o tempo de obtenção de dados da NOAA.

7 CONSIDERAÇÕES FINAIS

Este capítulo apresenta os trabalhos relacionados, as conclusões, trabalhos futuros e a lista das principais contribuições desse trabalho de mestrado.

7.1. Trabalhos Relacionados

A viabilidade da aplicação de um pacote de funções da biologia computacional na área ambiental é destacada no trabalho proposto por Ruivo (2008). 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 fisico-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 (MA). Os resultados obtidos comprovam que métodos de análise da bioinformática podem ser úteis na área ambiental.

Em Ruivo (2013) são utilizados metodologias de mineração de dados para analisar fenômenos meteorológicos extremos. O objetivo é identificar os fatores climatológicos relevantes que influenciaram tais eventos extremos. Duas metodologias são avaliadas: classificação estatística e árvores de decisão. O método estatístico está implementado na ferramenta computacional BRB-ArrayTools, software desenvolvido na área de bioinformática, sendo adaptado e aplicado na área ambiental. Algoritmos de árvores de decisão, estão implementados no software WEKA (2013), cujo objetivo é utilizar os atributos mais relevantes gerados pela análise estatística e gerar um classificador de árvore de decisão.

Neste trabalho, foi desenvolvida a ferramenta ERB-ArrayTools capaz de automar todas as fases do processo de KDD, as quais eram anteriormente realizadas de forma manual por Ruivo (2008) e Ruivo (2013). Esta ferramenta

permite processar, visualizar, agrupar, classificar, dentre outras funções realizada sobre o conjunto de dados de arquivos NetCDF's.

No trabalho desenvolvido por Papa (2009) foram estudas as técnicas supervisionadas, as quais se caracterizam pelo total conhecimento dos rótulos das amostras da base de dados. Além disso, foi proposto um novo método para classificação supervisionada de padrões baseada em Floresta de Caminhos Ótimos, do inglês *Optimum-Path Forest* (OPF), a qual modela o problema de reconhecimento de padrões como sendo um grafo, onde os nós são as amostras e os arcos definidos por uma relação de adjacência.

Já a ferramenta ERB-ArrayTools utiliza a técnica supervisionada que aplica o método de comparação de classes. Um descritor de dados com as classes é definido a partir de uma série temporal para guiar o processo de classificação e agrupamento dos dados geográficos na fase de mineração de dados.

O autor Bogorny (2003) descreve algumas das principais técnicas de DM e dentro dessas técnicas, investigou alguns algoritmos que empregam dados espaciais para a realização do processo de Descoberta de Conhecimento em Bases de Dados Geográficos. Buscou também, fazer um levantamento das ferramentas que implementam essas técnicas em dados espaciais.

No presente trabalho, o processo de descoberta de conhecimento foi aplicado sobre o conjunto de dados espaço-temporal referente à grande seca de 2005 na Amazônia de forma análoga à técnica de análise de MA . Para isso, o ERB-ArrayTools aplica a mineração de dados utilizando a permutação do método estatístico t-teste pelo Matlab para obtenção dos p-valores, como apresentado no capítulo 3.

7.2. Conclusões

Este trabalho lidou diretamente com o problema de automação de processos na Descoberta de Conhecimento em Bancos de Dados (KDD) ambientais cujo volume cresce a um ritmo acelerado devido à relevância que questões ambientais afetam a sociedade atual e aos avanços nas tecnologias de monitoramento ambiental.

O sistema ERB-ArrayTools foi desenvolvido especificamente para processar, visualizar, agrupar e classificar o conjunto de dados espaço-temporal tanto de arquivos NetCDF's quanto de arquivos texto para a área de Climatologia (Ruivo, 2008). Este sistema poderá ser utilizado em qualquer outro estudo de caso que considere conjuntos que estejam organizados de acordo com as dimensões nível, latitude, longitude e tempo como apresentado na subseção 4.2.1 do capítulo 4.

Para que este sistema aceite, por exemplo, o conjunto de dados da área de Limnologia (Ruivo, 2008) que utiliza outro critério de organização, será necessária a adaptação dos subsistemas existentes e criação novos subsistemas que atendam os novos requisitos especificados pelo especialista neste domínio.

A adoção de subsistemas neste trabalho foi providencial na montagem da ferramenta ERB-ArrayTools disponibilizada. Esta última integrou diversos recursos isolados e proveu suporte às fases do processo de KDD. Isto envolveu uma extensa investigação e experimentação de novas ferramentas para sua integração como biblioteca de funções, adaptação de existentes e mesmo escrita de novos subsistemas respeitando todos os detalhes de interface entre os mesmos e o agregador.

No ERB-ArrayTools, foram integrados cinco subsistemas: (1) ProduceNetCDF, para o pré-processamento; (2) ChronosGIS, para a carga dos dados pré-processados; (3) StatisticaAnalysis, para a mineração dos dados e pós-processamento pelo Matlab; (4) ToolsUI, para o pós-processamento com visualização em grade; e (5) IDV, para o pós-processamento com visualização em 3D.

Neste trabalho, primou-se pela área de visualização de dados científicos como coadjuvante ao processo de KDD, pois com uma grande massa de dados, é muito fácil perder o senso comum dos mesmos. Por isso, a visualização adicionou uma dimensão extra ao processo de KDD, a fim de auxiliar a interpretação dos dados extraídos e permitir a experimentação de cenários acelerando a tarefa de análise de mudanças climáticas por exemplo.

Finalmente, como não foi escopo deste o trabalho a metodologia da análise estatística ambiental *per se* buscou-se o objetivo da provisão de um ambiente automatizado, amigável, interativo e de grande usabilidade para especialistas de análise do domínio de eventos ambientais. Testes realizados com um especialista da área comprovaram que este objetivo foi alcançado, pois aumentou a velocidade e a flexibilidade na montagem de hipótese de cenários de teste e no fluxo de trabalho que antes era manual, sujeito a erros e tedioso.

7.3. Trabalhos Futuros

Há diversas possibilidades de melhorias e tratamento de exceções que poderão ser implementados em trabalhos futuros. Neste sistema também podem ser agregadas diversas outras funcionalidades, como:

- Melhorias na interface principal de acordo com as necessidades do usuário;
- Maior exploração do subsistema ChronosGIS para armazenamento e acesso à base de dados;
- Desenvolver novas funções em Octave, Silab ou em Java para substituir as funções estatísticas prontas da biblioteca do Matlab tornando o ERB-ArrayTools mais independente em sua fase de DM;
- Realização de cálculos estatísticos com ou sem a utilização de scripts como é feito no GRADS (2013);
- Criação de um subsistema para a busca de dados direto da NOAA;

- Adaptar o recurso ImageView do subsistema ToolsUI para o seu funcionamento como o GridView, a fim de permitir a visualização em escala de cinza tanto da variável ambiental como das novas variáveis estatísticas dos arquivos NetCDF's obtidos pelo ERB-ArrayTools.
- Adição de novas formas de visualização e melhorar as já existentes;
- Empregar GOOGLEMAPS (2013) para gráficos sobrepostos;
- Tornar a ferramenta ERB-ArrayTools multiplataforma;
- Aplicar o Desenvolvimento Baseado em Componentes de forma que cada subsistema seja um componente independente com sua própria interface para comunicação com o sistema agregador.

7.4. Sumário das principais contribuições

A lista a seguir apresenta no julgar do autor as principais contribuições do desenvolvimento desse trabalho de mestrado:

- Automação do processo KDD contribuindo para a diminuição do tempo de manipulação dos dados de dias para minutos;
- A ferramenta ERB-ArrayTools contribuiu para a automação do processo de KDD resultando na transformação do conhecimento tácito em conhecimento explicito, pois os métodos que antes eram individualmente empregados por um operador do processo, agora são formalizados numa sequência de atividades e encapsulados no código desta ferramenta.
- Aumento da produtividade e flexibilidade da configuração do KDD mediante desenvolvimento de uma interface de trabalho mais amigável e com maior usabilidade por especialistas de domínio;
- Emprego de subsistemas como biblioteca de funções, a qual acelerou o tempo de desenvolvimento e aumentou a sua confiabilidade sistêmica;

- Utilização de um Sistema Gerenciador de Banco de Dados Geográfico para apoiar as atividades da fase de pré-processamento do KDD;
- Utilização da interface Java-Matlab aumentando o repertorio de emprego desta ferramenta;
- Criação de uma nova função, chamada mavolcano.m baseada na função mavolcanoplot.m da biblioteca padrão do Matlab. Esta nova função permite maior flexibilidade e interatividade entre cálculos estatísticos e suas eventuais visualizações gráficas;
- Adaptação das ferramentas ToolsUI e IDV como subsistemas objetivando melhor visualização de arquivos NetCDF's em forma de grade sobre mapas. Esta solução mostrou-se melhor do que a alternativa da ferramenta GRADS (2013);
- Modificação da herança da classe IntegratedDataViewer desacoplando o fechamento do processo de visualização do IDV com o restante do ERB-ArrayTools;
- Empacotamento de dependências para o correto funcionamento do subsistema IDV;
- Criação de um instalador para todo o aplicativo ERB-ArrayTools executável em várias versões do Windows;
- Artigos técnicos nos anais dos eventos do Workshop de Computação Aplicada (WORCAP) do INPE nos anos de 2011 e 2012.

REFERÊNCIAS BIBLIOGRÁFICAS

ALMEIDA, W. G. D.; BORGES, M. E. G.; SANTOS, P. L. B. D.; ANOCHI, J. A.; SANTOS, W. A. D. Automação do pré-processamento de séries temporais de dados da NOAA para mineração de dados. In: WORKSHOP DOS CURSOS DE COMPUTAÇÃO APLICADA DO INPE, 11. (WORCAP), 2011, São José dos Campos. **Anais...** São José dos Campos: INPE, 2011.

ALVES, F. C. **Aplicação de técnicas de mineração de dados a uma base de um sistema gerenciador de informações para UTI.** 112p. Dissertação (Pós-Graduação em Tecnologia em Saúde) – Pontifícia Universidade Católica do Paraná, Curitiba, 2005.

AMARATUNGA, D.; CABRERA, J. **Exploration and analysis of DNA microarray and protein array data.** New Jersey: Wiley Interscience, 2004. 246 p.

AVILA, B. C. Data mining. In: ESCOLA DE INFORMÁTICA DA SBC-REGIONAL SUL, 6., 1998, Blumenau. **Anais...** Curitiba: Champagnat, 1998. p.87-106

BADC. **The CF metadata convention.** British Atmospheric Data Centre. Disponível em <http://badc.nerc.ac.uk/help/formats/netcdf/index_cf.html>. Acesso em: 01 nov. 2012.

BALDI, P.; BRUNAK, S. **Bioinformatics:** the machine learning approach. New York: Cambridge University Press, 2001. Segunda Edição.

BOGORNY, V. **Algoritmos e ferramentas de descoberta de conhecimento em bancos de dados geográficos.** Porto Alegre: PPGC da UFRGS, 2003.

BOTÍA, J. A. G. M. VELASCO, J. R. SKARMETA, A. F. **A Generic dataminig system. Basic desing and implementation guidelines.** Departamento de Informática, Inteligencia Artificial y Electrónica, Universidad de Murcia, 2002.

BREIMAN, L.; FRIEDMAN, J. H.; OLSHEN, R. A.; STOLE, C. J. **Classification and regression trees.** Monterey, CA.: Wadsworth and Brooks, 1984. 358 p.

CÂMARA, G. CASANOVA, M. A. HEMERLY, A. S. MAGALHÃES, G. C. MEDEIROS, C. M. B. **Anatomia de sistemas de informação geográfica.** Campinas, Instituto de Computação, UNICAMP, 1996.

CARVALHO, D.R., **Data mining através de indução de regras e algoritmos genéticos.** Dissertação (Mestrado em Informática Aplicada) – Pontifícia Universidade Católica do Paraná. Curitiba - PR, 1999.

CARVALHO, O. L. JÚNIOR, C. R. RUIVO, H. M. **Array Statistical Analysis System (A.S.A.S).** Código-fonte legado de projeto conjunto da Universidade Federal de Itajubá (UNIFEI) e Instituto Nacional de Pesquisas Espaciais (INPE), 2011.

COLLAB. **Ocean share collaborative tool for integrated browse of data from multiple archives.** Disponível em: <http://www.epic.noaa.gov/collab/>. Acessado em 30 de agosto de 2012.

COUCLELIS, H. People manipulate objects (but cultivate fields): beyond the raster-vector debate in GIS. **In Proc International Conference on GIS** - From Space to Territory: Theories and Methods of Spatial Reasoning, pages 65-77. Springer Verlag Lecture Notes in Computer Science 639, 1992.

Companhia de Pesquisa de Recursos Minerais (CPRM). **Acompanhamento por Bacia Hidrográfica.** Serviço Geológico do Brasil. Disponível em: < http://www.cprm.gov.br/rehi/amazonialegal/Bacias_de_controle.pdf >. Acesso em: 02 fev. 2013.

INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS. Centro de Previsão de Tempo e Estudos Climáticos (INPE/CPTEC). **Glossário.** 2006. Disponível em: <http://www7.cptec.inpe.br/glossario/>. Acesso em: 02 fev. 2013.

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.

EPIC. **Pacific marine environmental laboratory.** Disponível em: <http://www.epic.noaa.gov/epic/software/ep_java.htm> Acesso em: 30 ago. 2012.

ESRL. **Earth system research laboratory.** Physical Sciences Division. Disponível em: <http://www.esrl.noaa.gov>. Acesso em: 15 jul. 2011.

FARIA, G. **Um banco de dados espaço-temporal para desenvolvimento de aplicações em sistemas de informação geográfica.** X Escola de Computação, Instituto de Computação, UNICAMP, 1996.

FAYYAD, U.M. ET AL. **Advances in knowledge discovery and data mining.** California: AAAI Press, 1996.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. **From data mining to knowledge discovery:** an overview. In: Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P; Uthhurusany, R., Eds., Advances in knowledge discovery. Cambridge: MIT Press, 1996, p. 1-36.

FELDENS, M. A. **Engenharia da descoberta de conhecimento em bases de dados:** estudo e aplicação na área da saúde. Dissertação. Porto Alegre: PPGC da UFRGS, 1997.

FRANK, A. GOODCHILD, M. **Two perspectives on geographical data modelling**. National Center for Geographic Information and Analysis, 1990. p. 90-11. Technical Report.

FREITAS, A. A. **Data mining and knowledge discovery with evolutionary algorithms.** Springer, 2002. Natural Computing Series.

GOLDSCHMIDT, R.; PASSOS, E. **Data mining:** um guia prático. 1. Ed. Campus, 2005.

GOODCHILD, M. ET AL. Integrating GIS and spatial data analysis: problems and possibilities. **International Journal of Geographical Information Systems**, v. 6, n.5, p. 407-424, 1992.

GOOGLEMAPS. **Static maps API V2 developer guide.** Disponível em: <https://developers.google.com/maps/documentation/staticmaps/index>. Acesso em: 5 maio 2013.

GRADS. **Grid Analysis and Display System (GrADS).** Disponível em: <http://www.iges.org/grads>. Acesso em: 27 fev. 2013.

HALMENSCHLAGER, C. **Utilização de agentes na descoberta de conhecimento.** Porto Alegre: PPGC da UFRGS, 2000. (TI-955).

HAN, J.; KAMBER, M.; TUNG, A. K. H. **Spatial clustering methods in data mining:** a survey. school of computing science, Simon Fraser University, Burnaby, BC Canada, v5a156, 2001.

HAUTANIEMI, S. **Studies of microarray data analysis with applications for human cancers**. Tampere, Finland: Tampere University of Technology, 2003.

HOLTON, J. R. **An introduction to dynamic meteorology.** Seattle, WA: Academic Press, 2004. 535 p.

IAN, H. WITTEN, E.F. **Data mining -** pratical machine learning tools and techniques with Java implementations. Morgan Kaufmann Publishers ed. 2000.

IDV. **Integrated data viewer.** Disponível em: <http://www.unidata.ucar.edu/software/idv/#home >. Acesso em: 30 ago 2012.

IPCC. **Cambio climático 2007:** informe de síntesis. Grupo Intergubernamental de Expertos sobre el Cambio Climático [Equipo de redacción principal: Pachauri, R.K. y Reisinger, A. (directores de la publicación)], Ginebra, Suiza, 2007.

JAVAJDK. **Java SE 6 update 35.** Disponível em: <http://www.oracle.com/technetwork/java/javase/downloads/index.html>. Acesso em 30 ago. 2012.

JAVA3D. **Java 3D 1.1.3 API install notes.** Disponível em: <http://www.oracle.com/technetwork/java/javase/java3d-install-139068.html>. Acesso em: 30 ago. 2012.

KAMBER, J.H.M. **Data mining -** concepts and techniques**.** Morgan Kaufmann Publishers, 2001.

KOPERSKI, K.; HAN, J. Discovery of spatial association rules in geographic information databases**.** In: INT. SYMP. ON LARGE SPATIAL DATABASES (SSD'95), 4., 1995, Portland, ME. **Proceedings...** Portland, 1995. p.47-66.

KOPERSKI, K; ADHIKARY, J.; HAN, J. Spatial data mining: progress and challenges. In: SIGMOD WORKSHOP ON RESEARCH ISSUES ON DATA MINING AND KNOWLEDGE DISCOVERY, 1996, Montreal. **Proceedings…** Montreal, 1996

KOPERSKI, K.; HAN, J.; ADHIKARI, J. Mining knowledge in geographical data. **Communications of the ACM**, v. 26, 1997.

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>. Acesso em: 22 jul. 2006.

LINDEN, R. Técnicas de agrupamento**. Revista de Sistemas de Informação da FSMA**, 2009. Disponível em: <http://www.fsma.edu.br/si/edicao4/FSMA_SI_2009_2_Tutorial.pdf>. Acesso em: 27 set. 2012.

MATLABCONTROL. **A Java API to interact with MATLAB.** Disponível em: <http://code.google.com/p/matlabcontrol/> Acesso em: 30 ago. 2012.

MATTHEWS, D.E. FAREWELL, V.T. **Using and understanding medical statistics.** New York: Karger, 1988. p. 17-19.

MOSAVI, H. **Creating an installer.** code project – for those who code. Disponível em: <http://www.codeproject.com/Articles/24187/Creating-an-Installer>. Acesso em: 02 abr. 2013.

NETBEANS. **NetBeans IDE the smarter and faster way to code.** Disponível em: <https://netbeans.org/> Acesso em: 06 mar. 2013.

NEVES, M. C.; FREITAS, C. C.; CAMARA, G. **Mineração de dados em grandes bancos de dados geográficos.** INPE, 2001. Relatório Técnico.

NETCDF-JAVA. **NetCDF-Java library.** Unidata Program Center. UCAR Community Programs. University Corporation for Atmospheric Research – UCAR. Disponível em: <http://www.unidata.ucar.edu/software/netcdf-java/>. Acesso em: 30 ago. 2012.

NG, R. T.; HAN, J. Efficient and effecctive clustering methods for spatial data mining. In: INTERNATIONAL CONFERENCE ON VERY LARGE DATA BASE, 20., 1994, Santiago, 1994.

NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION (NOAA). **About NOAA.** United States Department of Commerce. Disponível em: <http://www.noaa.gov/about-noaa.html >. Acesso em: 30 ago. 2012.

NORVIG, S. J. R. A. P. **Artificial intelligence -** a modern approach. A Simom & Shuster Company, 1995.

Nullsoft Scriptable Install System (NSIS). **Main Page.** Disponível em: <http://nsis.sourceforge.net/Main_Page>. Acesso em: 02 abr. 2013.

PAPA, J. P. **Classificação supervisionada de padrões utilizando floresta de caminhos ótimos.** 2008. Tese (Doutorado em Ciência da Computação) – Instituto de Computação da Universidade Estadual de Campinas, Campinas, UNICAMP, 2009.

RICE, J. A. **Mathematical statistics and data analysis.** 3. ed: Duxbury Advanced, 2006.

RODRIGUEZ, H. M. **HM NIS EDIT:** a free NSIS Editor/IDE. Disponível em: <http://hmne.sourceforge.net/index.php>. Acesso em: 02 abr. 2013.

RUIVO, H. M. **Análise integrada de dados ambientais utilizando técnicas de classificação e agrupamento de micro arranjos de DNA.** 2008. 98 p. (INPE-15217-TDI/1311). Dissertação (Mestrado em Computação Aplicada) -

Instituto Nacional de Pesquisas Espaciais, São José dos Campos, 2007. Disponível em: <http://urlib.net/sid.inpe.br/mtc-m17@80/2007/12.14.12.09>. Acesso em: 01 jul. 2013.

RUIVO, H. M. **Metodologias de mineração de dados em análise climática.** 2013. 121 p. (sid.inpe.br/mtc-m19/2013/02.14.13.23-TDI). Tese (Doutorado em Computação Aplicada) - Instituto Nacional de Pesquisas Espaciais (INPE), São José dos Campos, 2013. Disponível em: <http://urlib.net/8JMKD3MGP7W/3DHME8E>. Acesso em: 01 jul. 2013.

SANTOS, M. **Padrão:** um sistema de descoberta de conhecimento em bases de dados georreferenciadas. Tese de doutorado. Universidade do Minho, 2001.

SILVA, M. M. **Uma abordagem evolucionária para o aprendizado semi-supervisionado em máquinas de vetores de suporte.** Dissertação (Mestrado em Engenharia Elétrica) – Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal de Minas Gerais, 2008.

SIMON, R. **BRB-ArrayTools developed by:** Richard Simon & BRB-ArrayTools Development Team. Biometric Research Branch (BRB). Division of Cancer Treatment and Diagnosis. Institute Cancer National. Disponível em: <http://linus.nci.nih.gov/BRB-ArrayTools.html > Acesso em: 02 fev. 2013.

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: 27 fev. 2011.

SINOARA, R. A. J.B.P. REZENDE, S. O. **Combinação de classificadores no processo data mining.** São Paulo: Instituto de Ciências Matemáticas e de Computação - USP, 2002.

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, 2004. Disponível em: <http://www.dimap.ufrn.br/~marcilio/ENIA2003/jaia2003-14-08.pdf>.

TZAFESTAS, S.G., **Knowledge-based system diagnosis, supervision and control.** New York: Plenum Press, 1989.

UMBRELLO. **Umbrello UML modeller.** Disponível em: <http://uml.sourceforge.net/>. Acesso em: 27 maio 2012.

THE UNITED MODELING LANGUAGE – UML**. Object management group**. Disponível em: <http://www.uml.org/>. Acesso em: 27 de maio de 2012.

UNIDATA. **Unidata overview.** Disponível em: <http://www.unidata.ucar.edu>. Acesso em: 30 ago. 2012.

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

WEKA. **WEKA The university of waikato.** Disponível em: http://www.cs.waikato.ac.nz/ml/weka. Acesso em: 01 maio 2013.

APÊNDICE A

Os gráficos das Figuras deste apêndice foram obtidas pelo subsistema *StatisticalAnalysis* utilizando a biblioteca *MatlabControl*. O usuário poderá visualizar os gráficos dos resultados estatísticos de uma determinada variável ou de todas as variáveis ambientais. As Figuras A.1, A.2, A.3 e A.4 são gráficos obtidos de todas as variáveis do conjunto de dados.

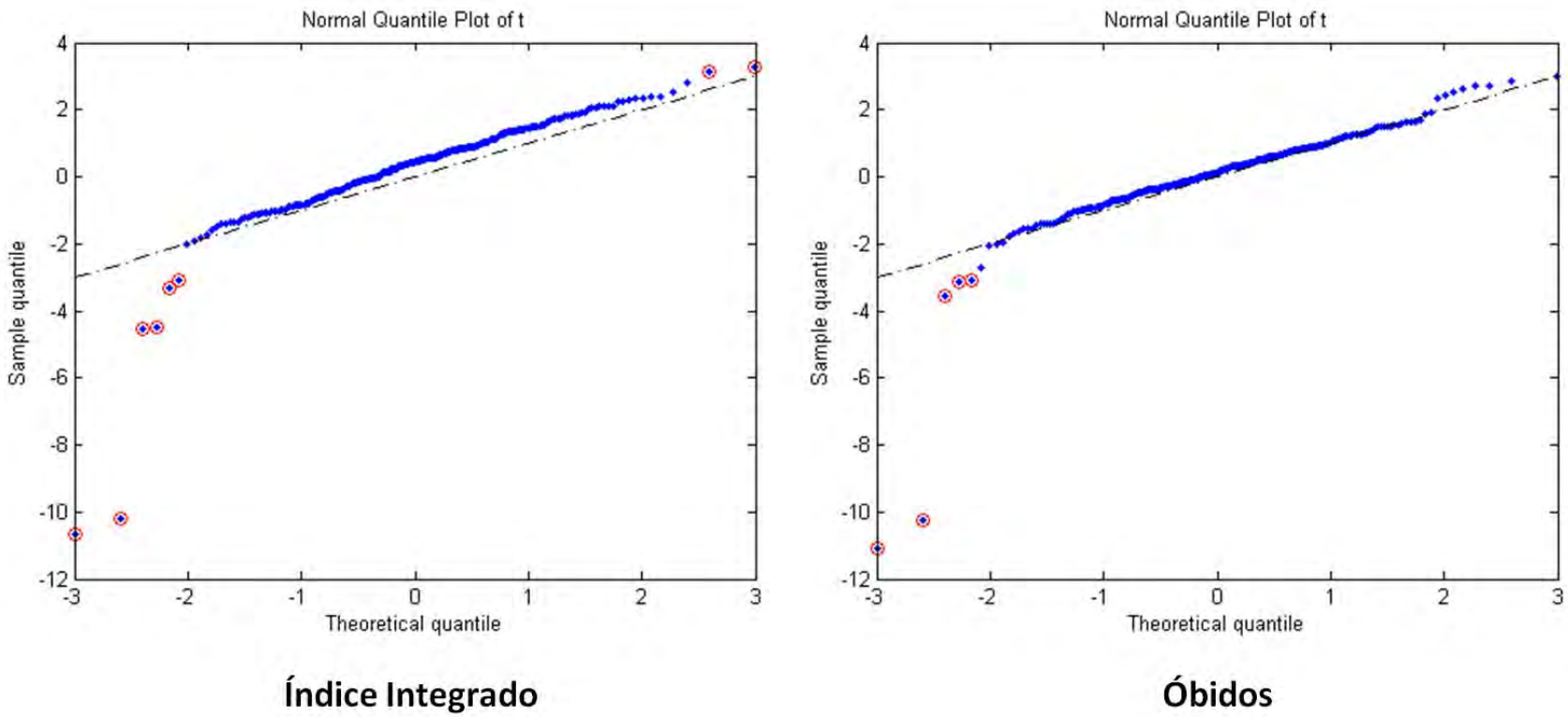


Figura A.1 – Gráfico quântico normal de t.

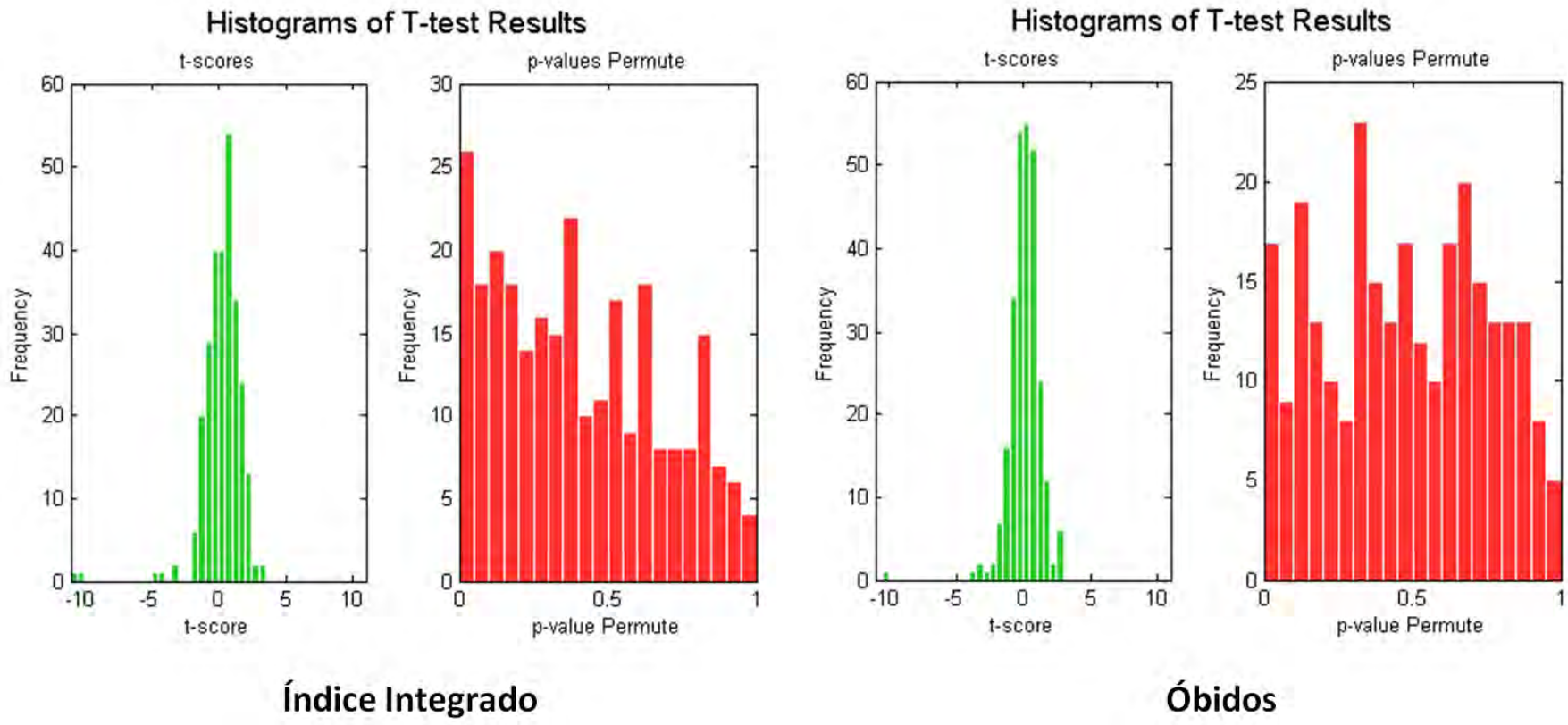
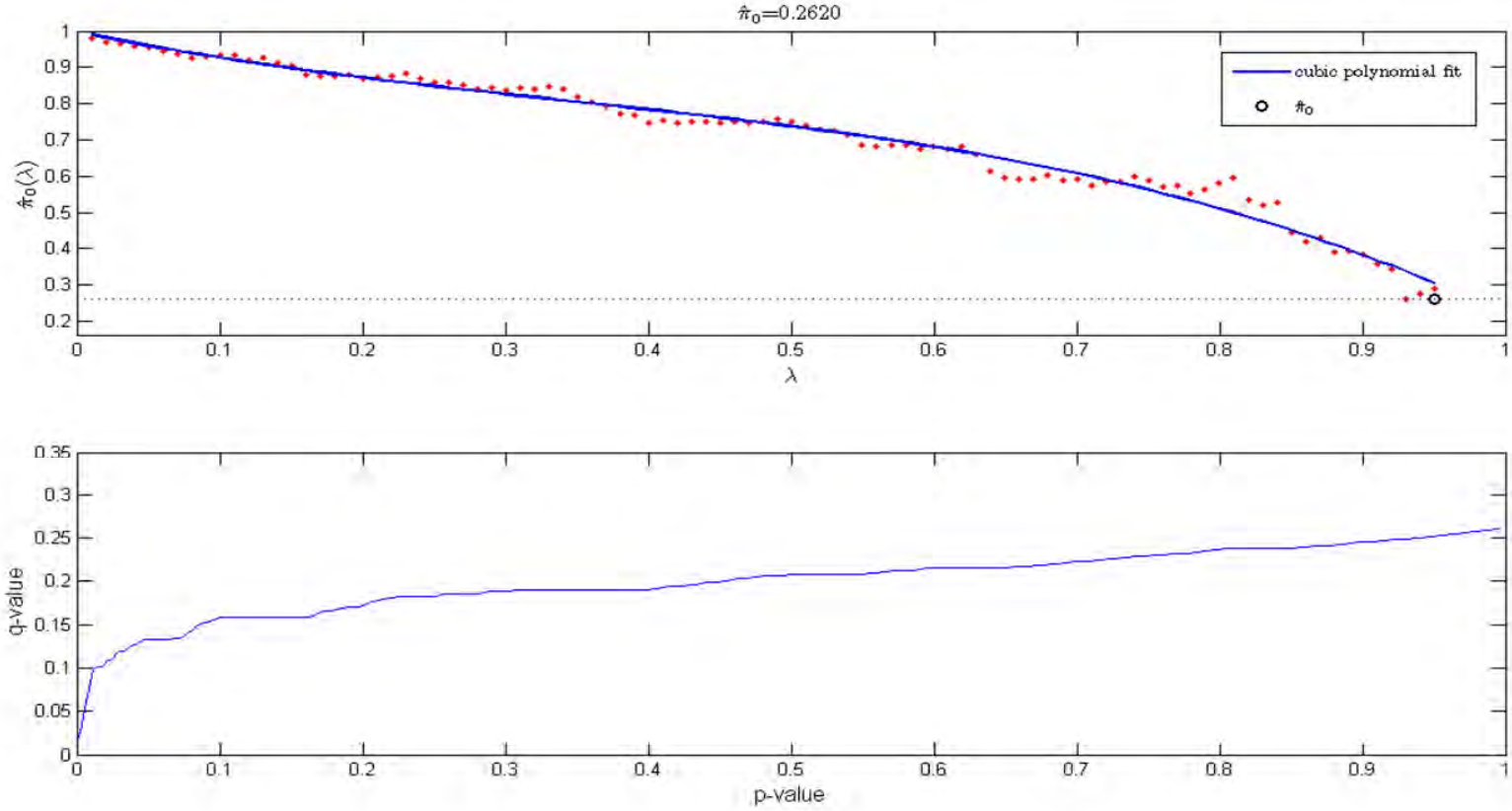
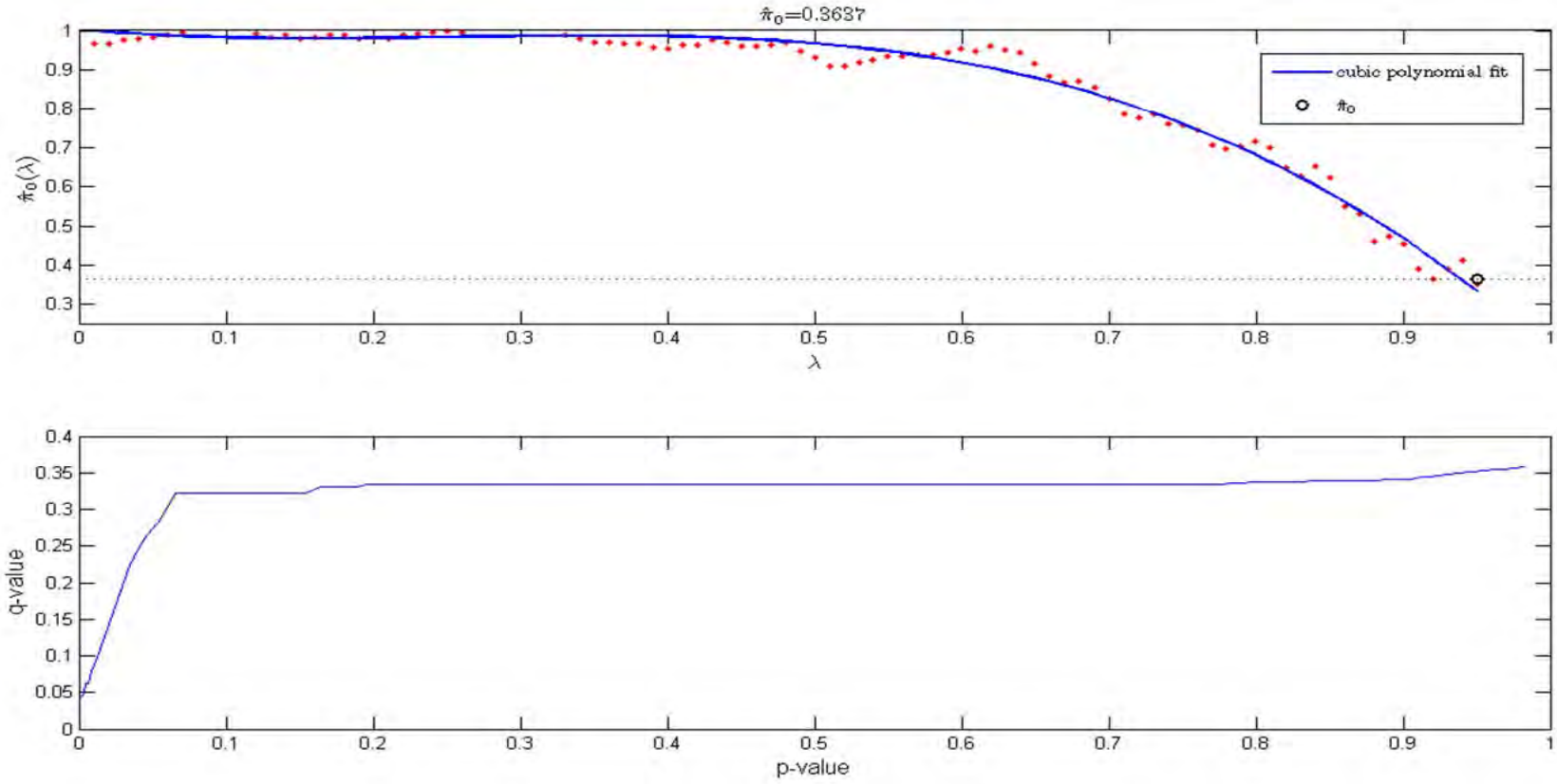


Figura A.2 – Histograma dos resultados do método t-test.



Índice Integrado

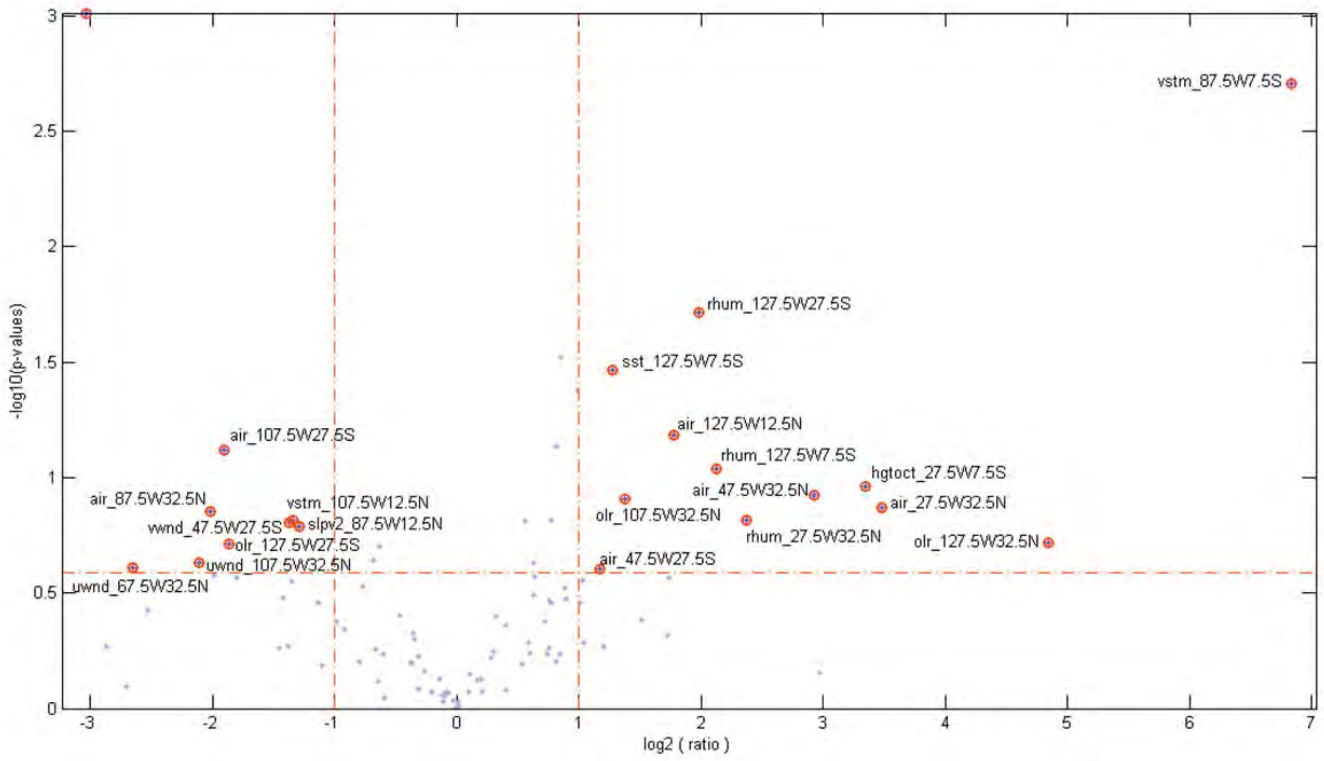
Figura A.3 – Gráfico do p-valor no Índice Integrado utilizando o FDR.



Óbidos

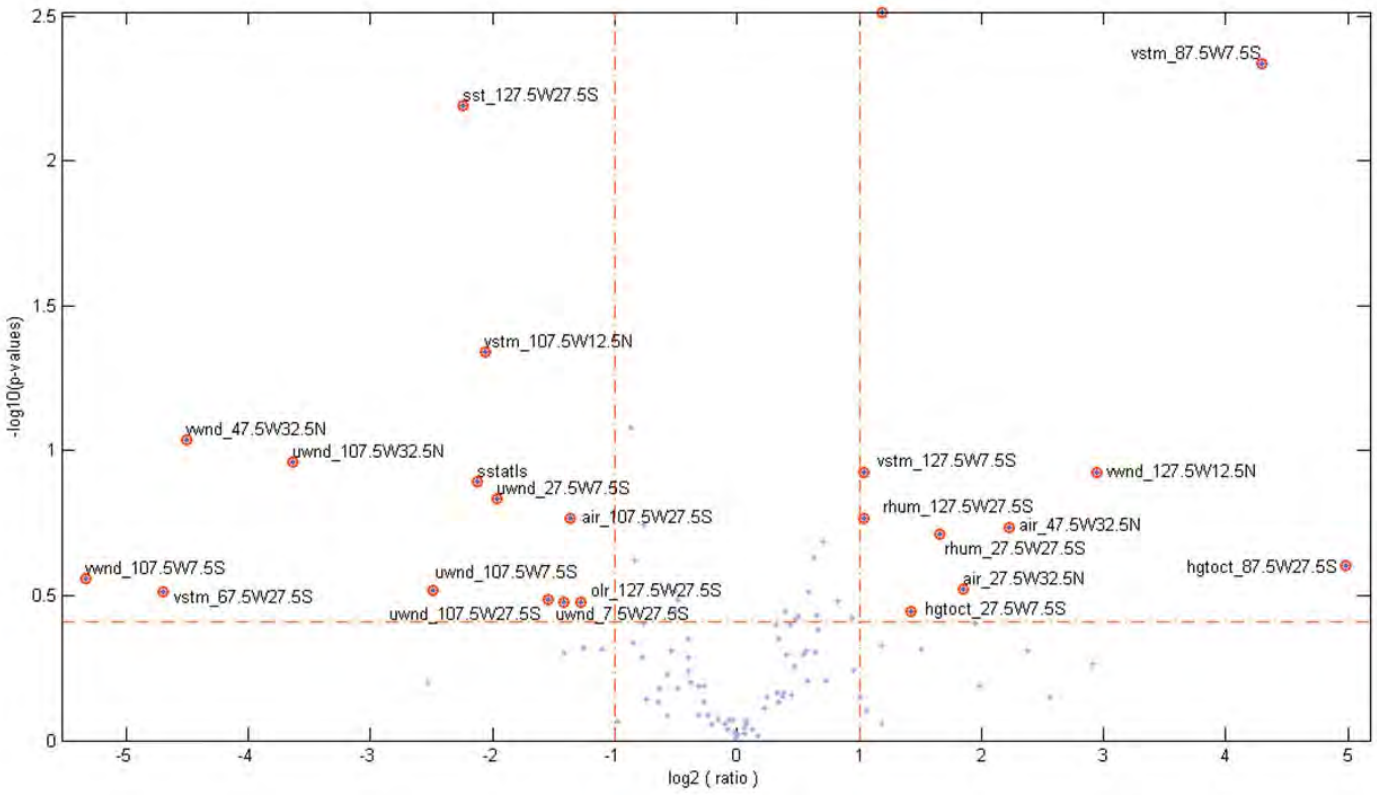
Figura A.4 – Gráfico do p-valor em Óbitos utilizando o FDR.

As Figuras A.5 e A.6 mostram a funcionalidade *Significance Test* usada para visualizar os dados mais significativos.



Índice Integrado

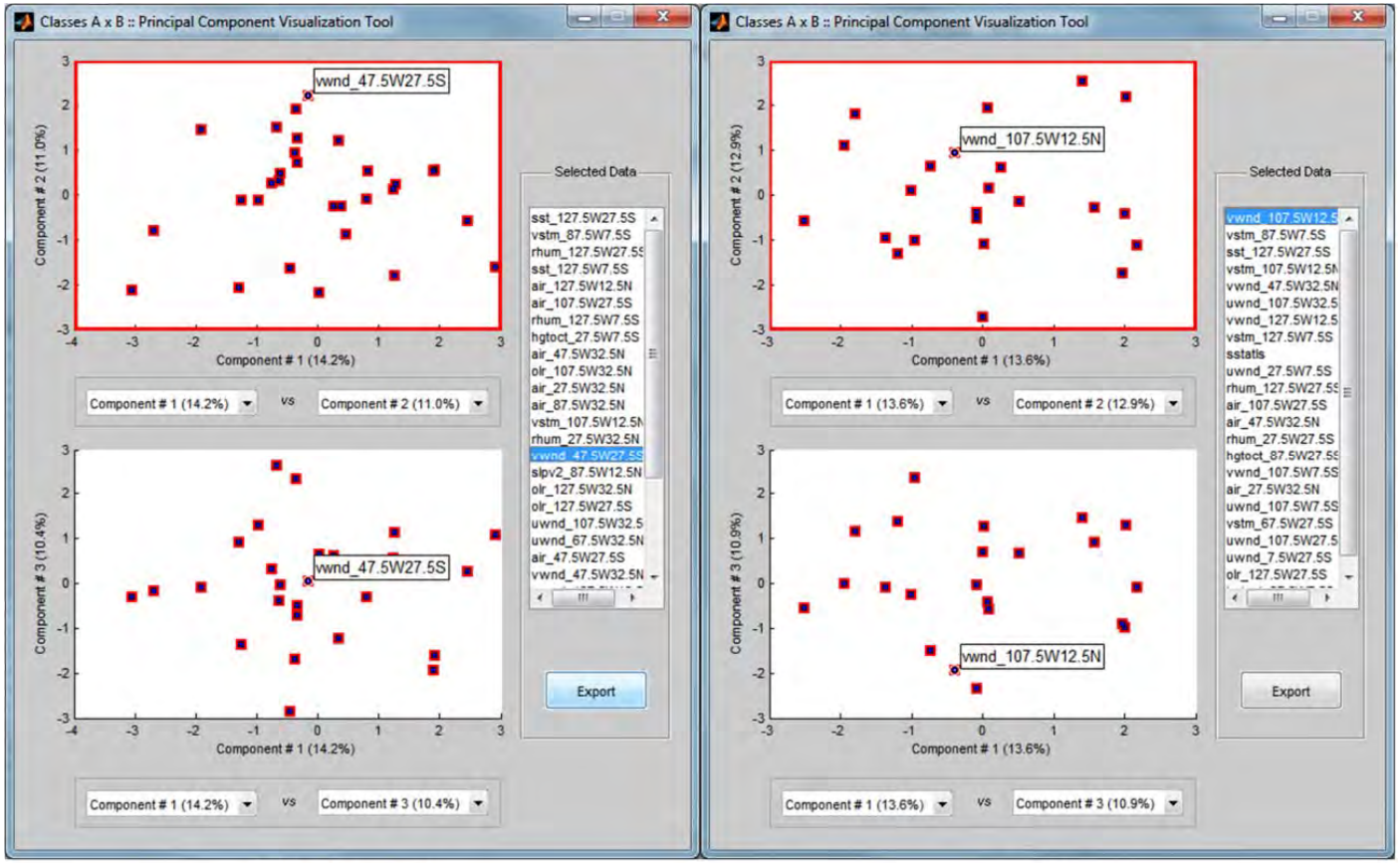
Figura A.5 – Gráfico de teste de significância no Índice Integrado.



Óbidos

Figura A.6 – Gráfico de teste de significância em Óbidos.

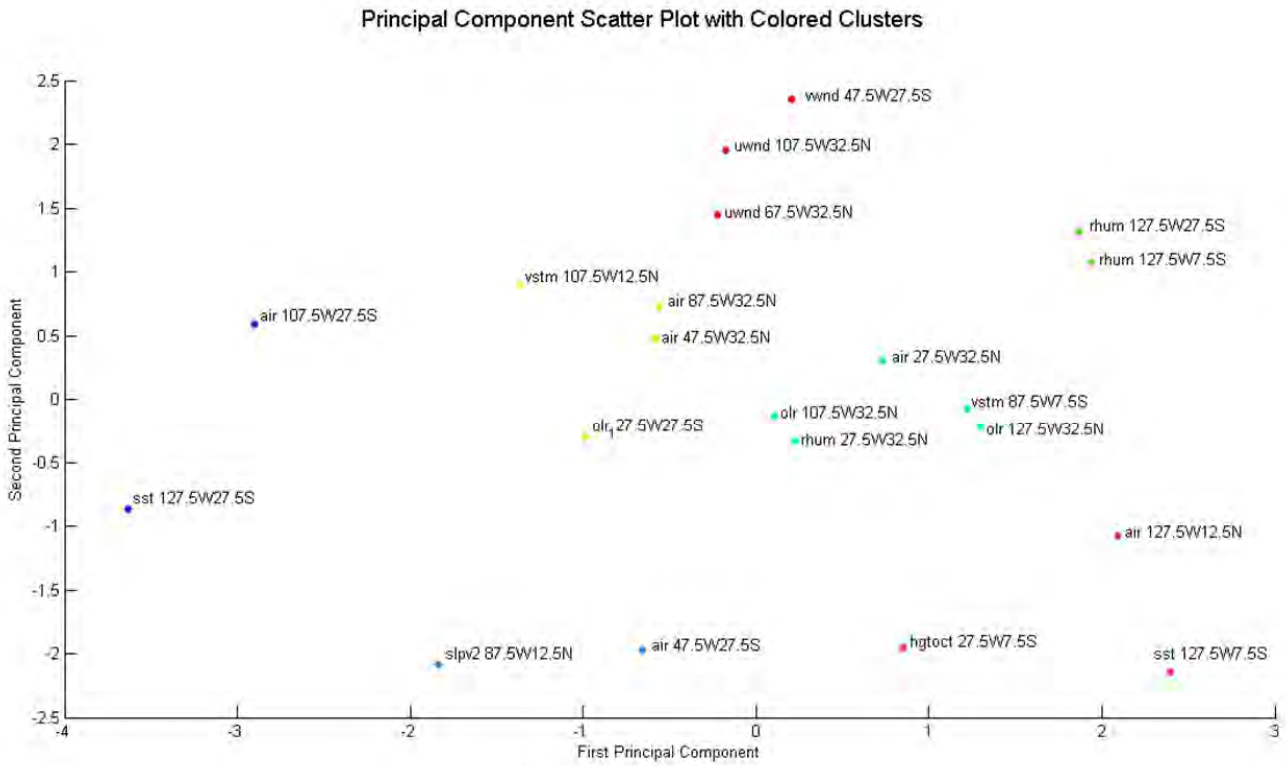
A ferramenta *Principal Component Visualization Tools* do Matlab é utilizada para visualizar a dispersão dos dados mais significativos, os quais são identificados pela sua descrição, como mostrado na Figura A.7.



Índice Integrado **Óbidos**

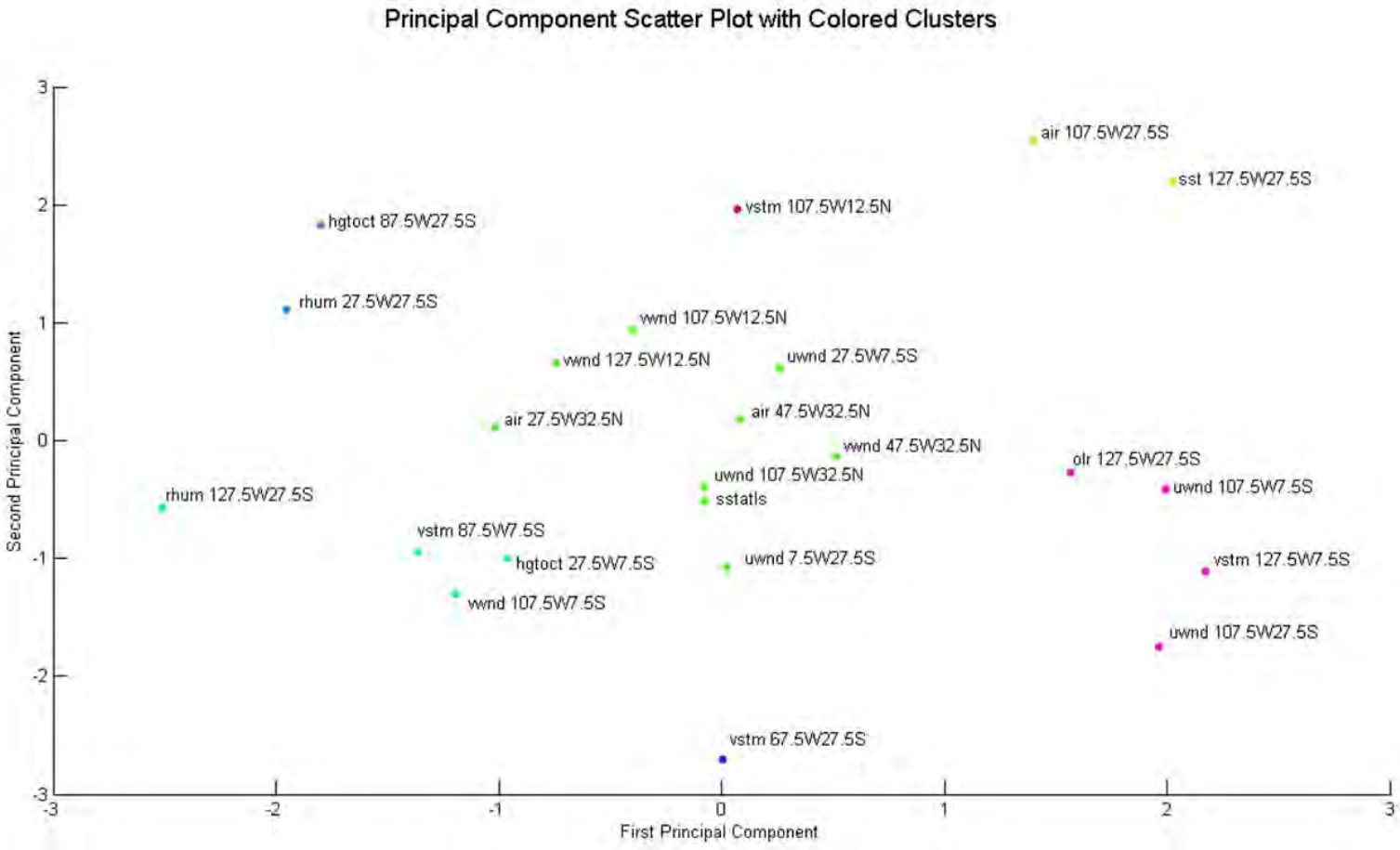
Figura A.7 – Visualização da dispersão dos dados mais significativos.

Para o agrupamento pela funcionalidade *Principal Component Scatter Plot*, o usuário pode definir o número de grupos a ser visualizado de forma colorida, como mostrado nas Figuras A.8 e A.9. Nesta funcionalidade, é possível a definição e organização da descrição de cada ponto que representa o conjunto de dados mais significativos de uma determinada variável ambiental.



Índice Integrado

Figura A.8 – Agrupamento no Índice Integrado.



Óbidos

Figura A.9 – Agrupamento em Óbidos.

Utilizando o Matlab, também é possível visualizar uma variável e seu respectivo p-valor em três dimensões e sobre postos, como mostrado nas Figuras a seguir.

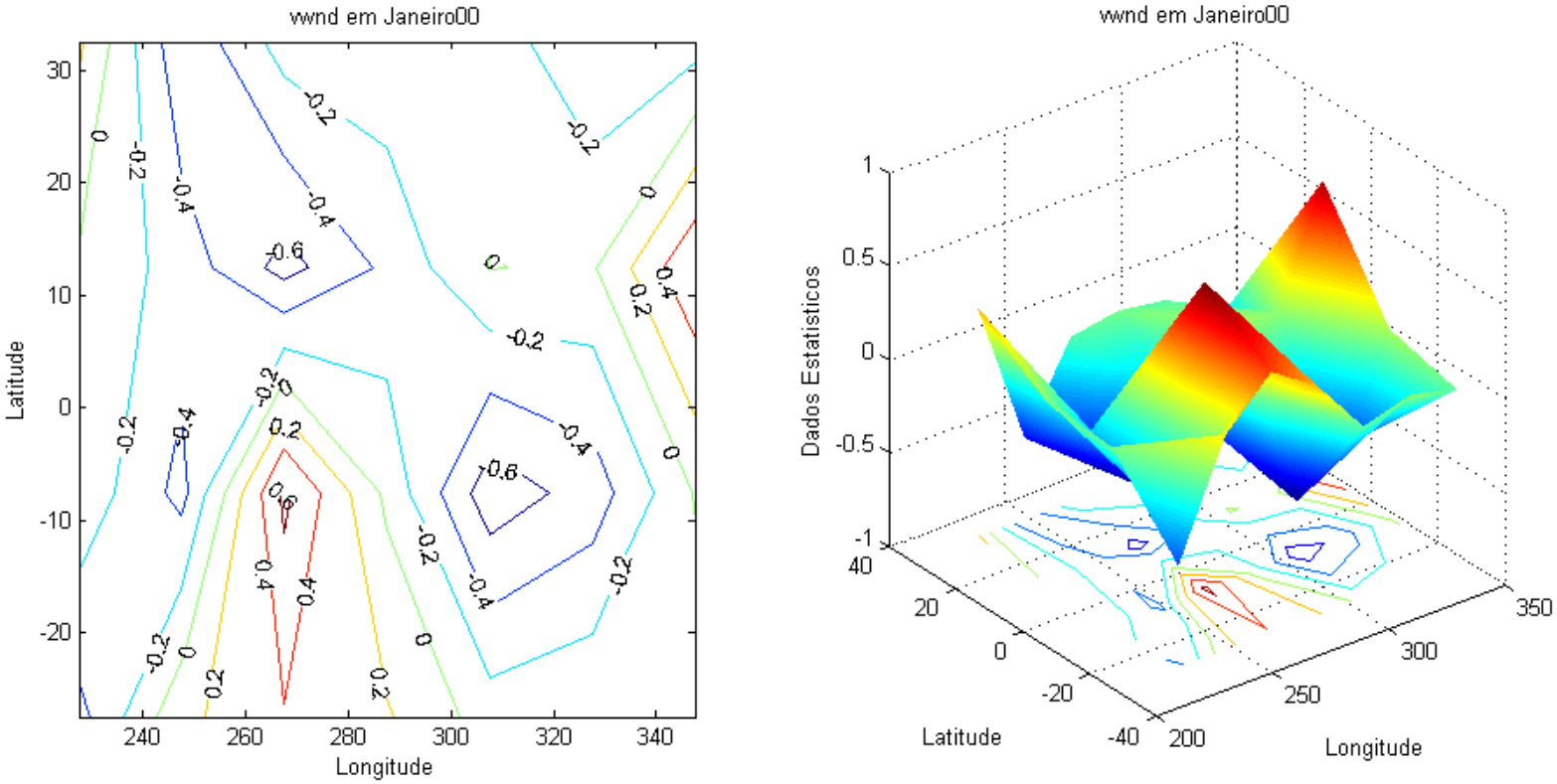


Figura A.10 – Dados da variável ambiental vwnd em janeiro de 2000.

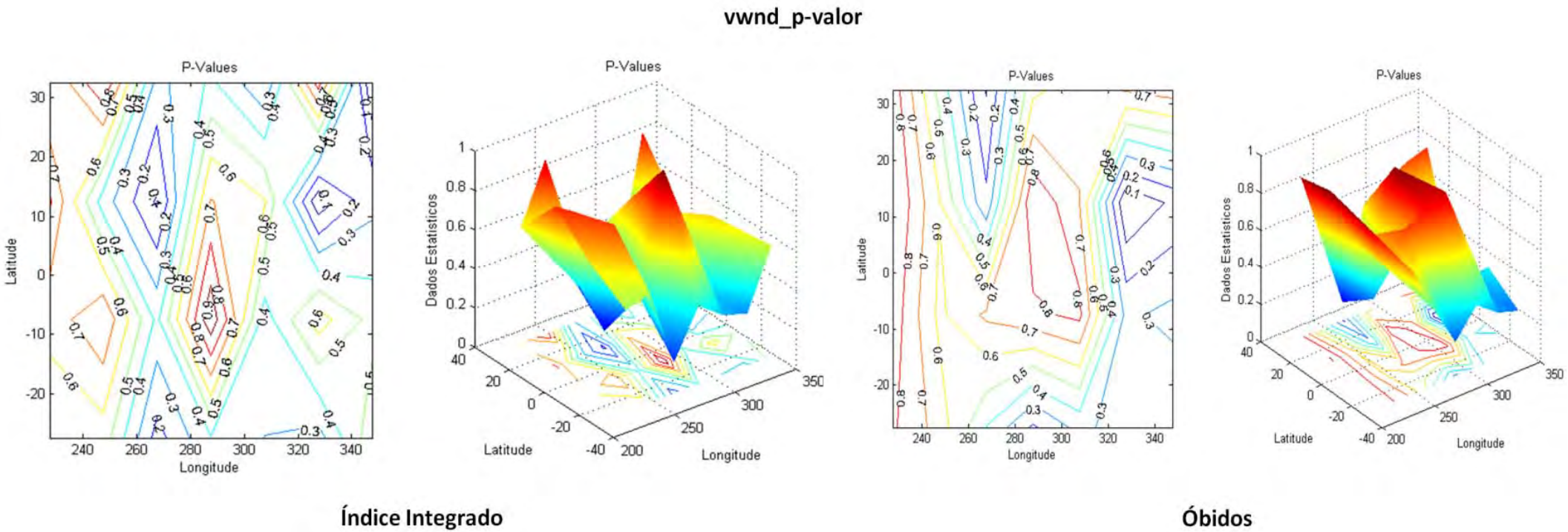


Figura A.11 – P-valor de vwnd no Índice Integrado e em Óbitos.

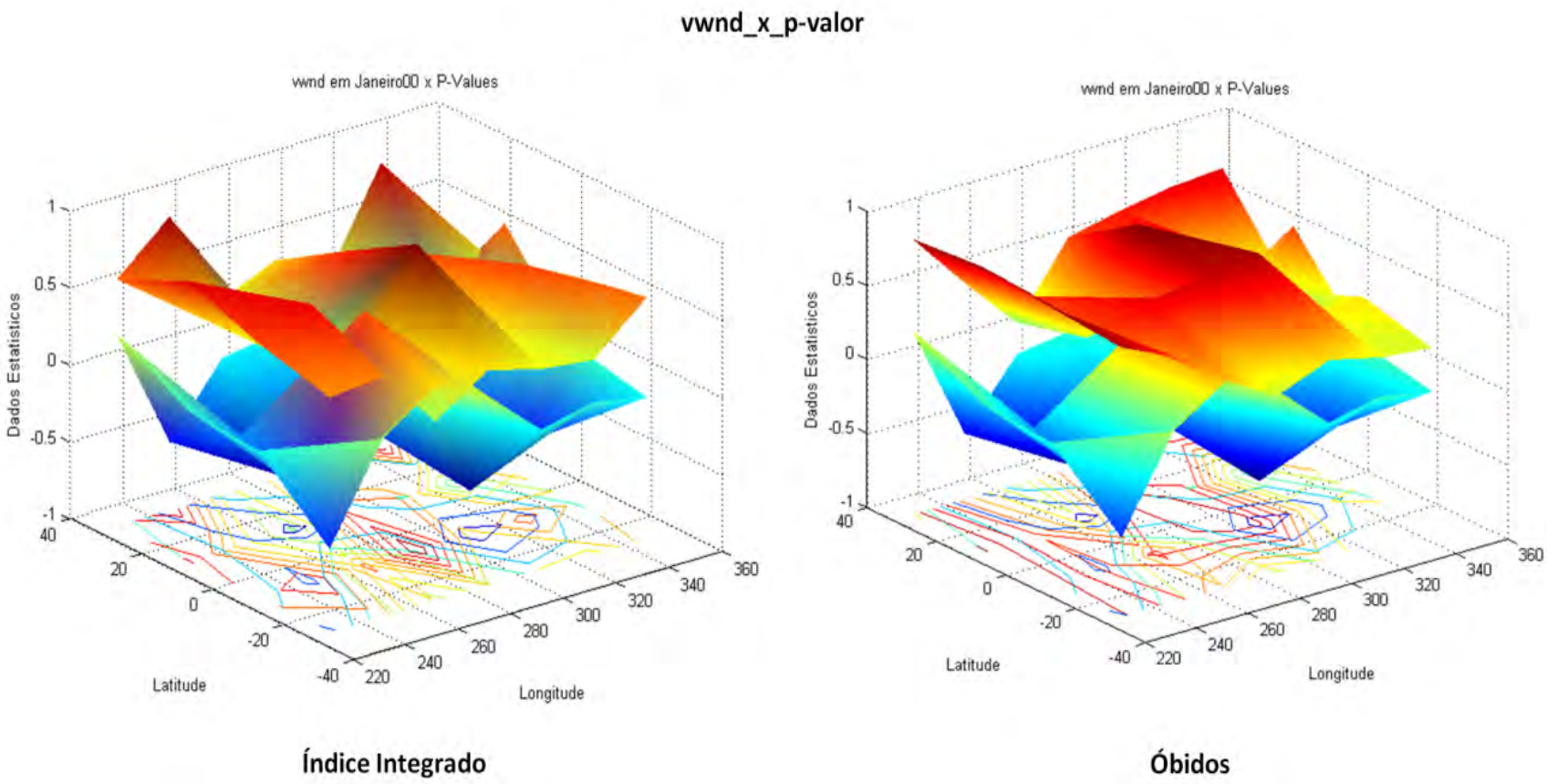


Figura A.12 – Dados de vwnd e seu p-valor sobrepostos.

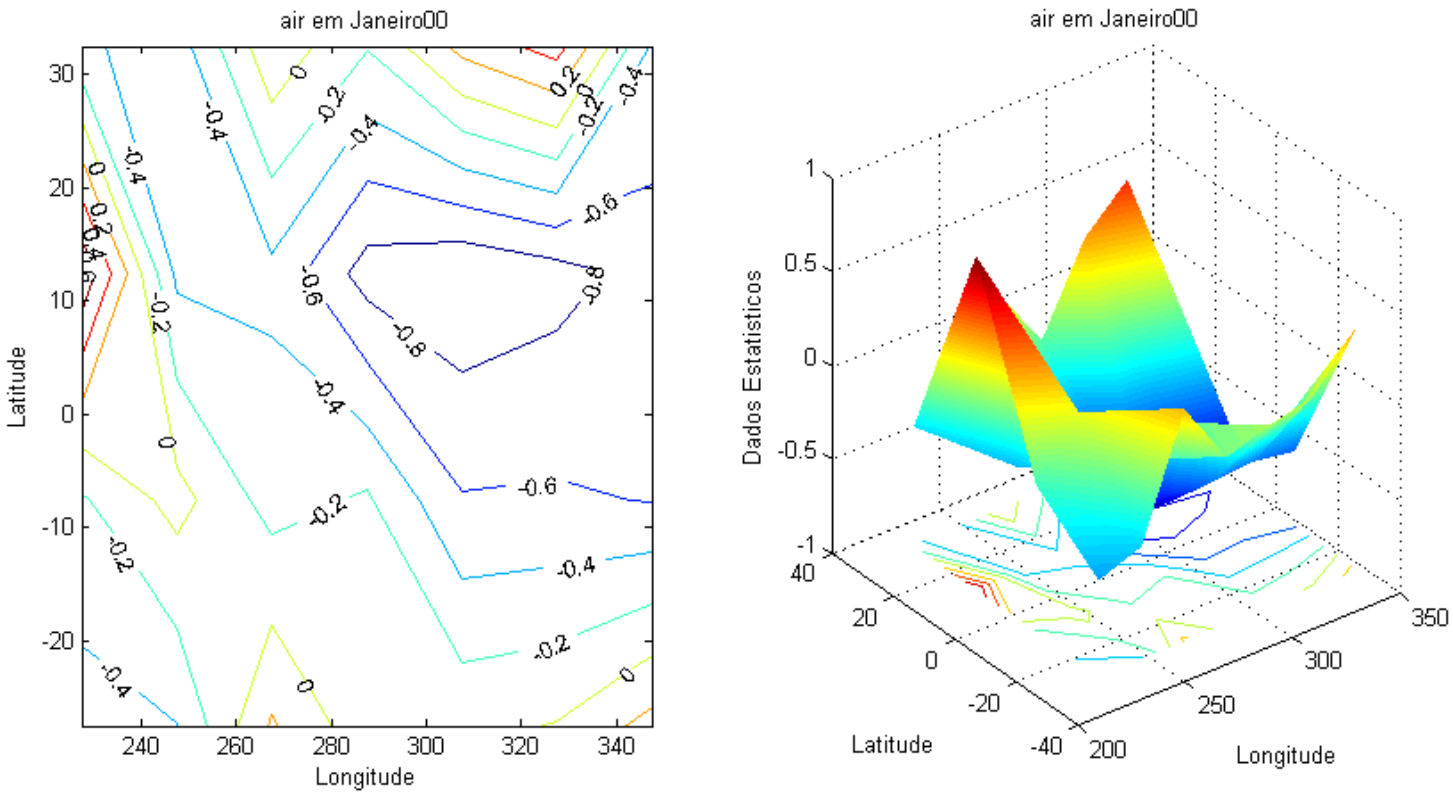


Figura A.13 – Dados da variável ambiental air em janeiro de 2000.

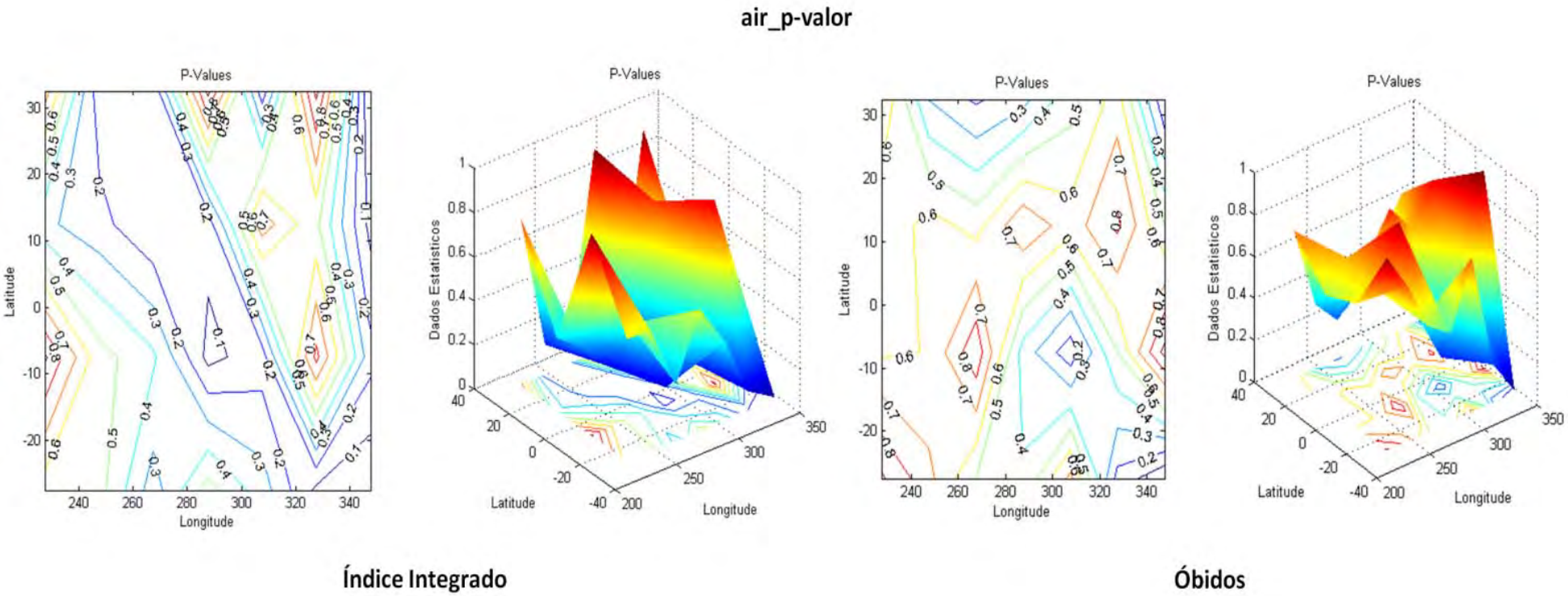
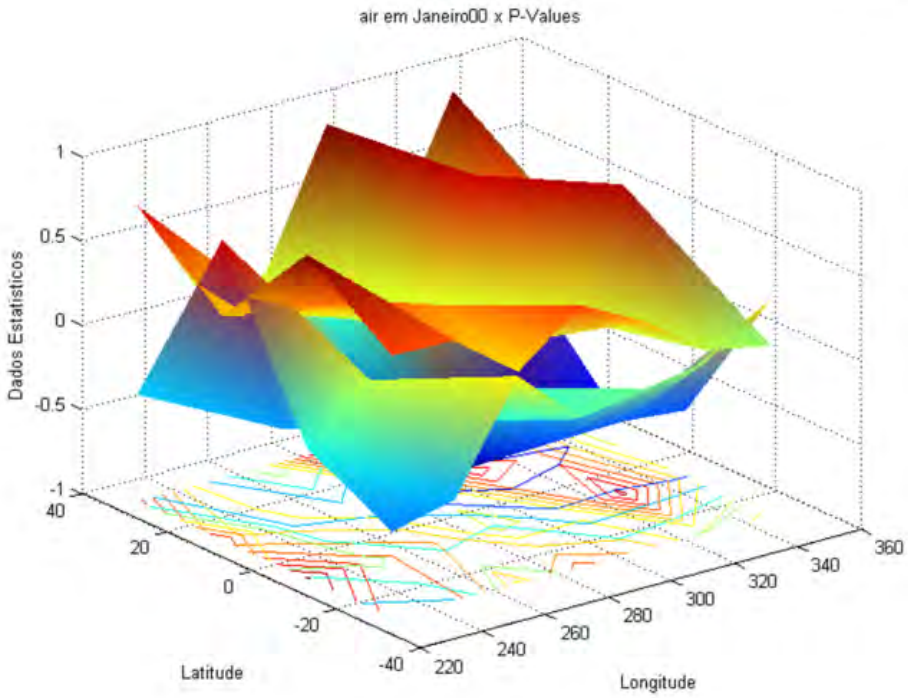
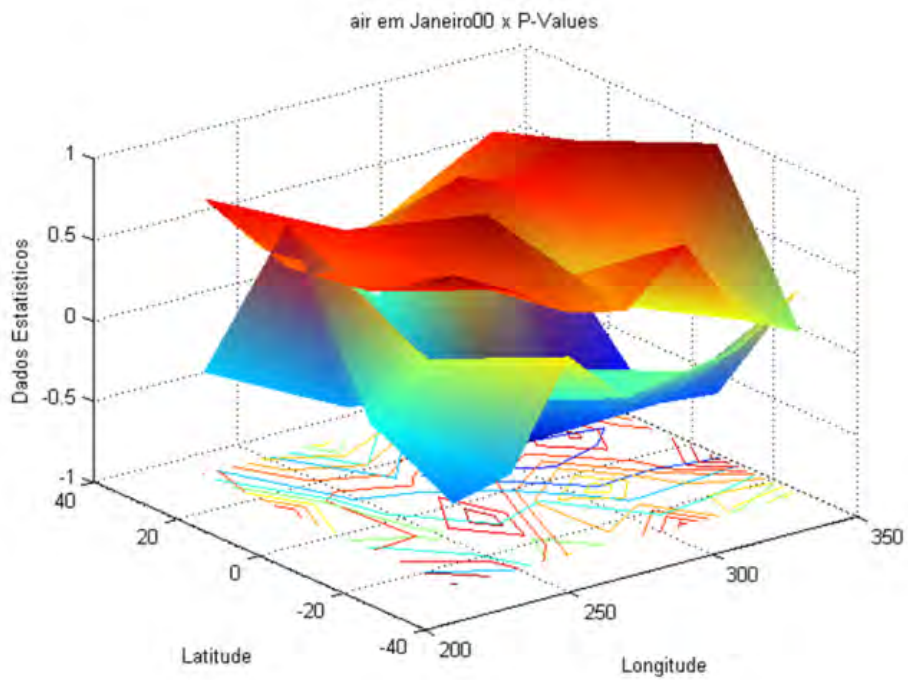


Figura A.14 – P-valor de air no Índice Integrado e em Óbitos.

air_x_p-valor



Índice Integrado



Óbidos

Figura A.15 – Dados de air e seu p-valor sobrepostos.

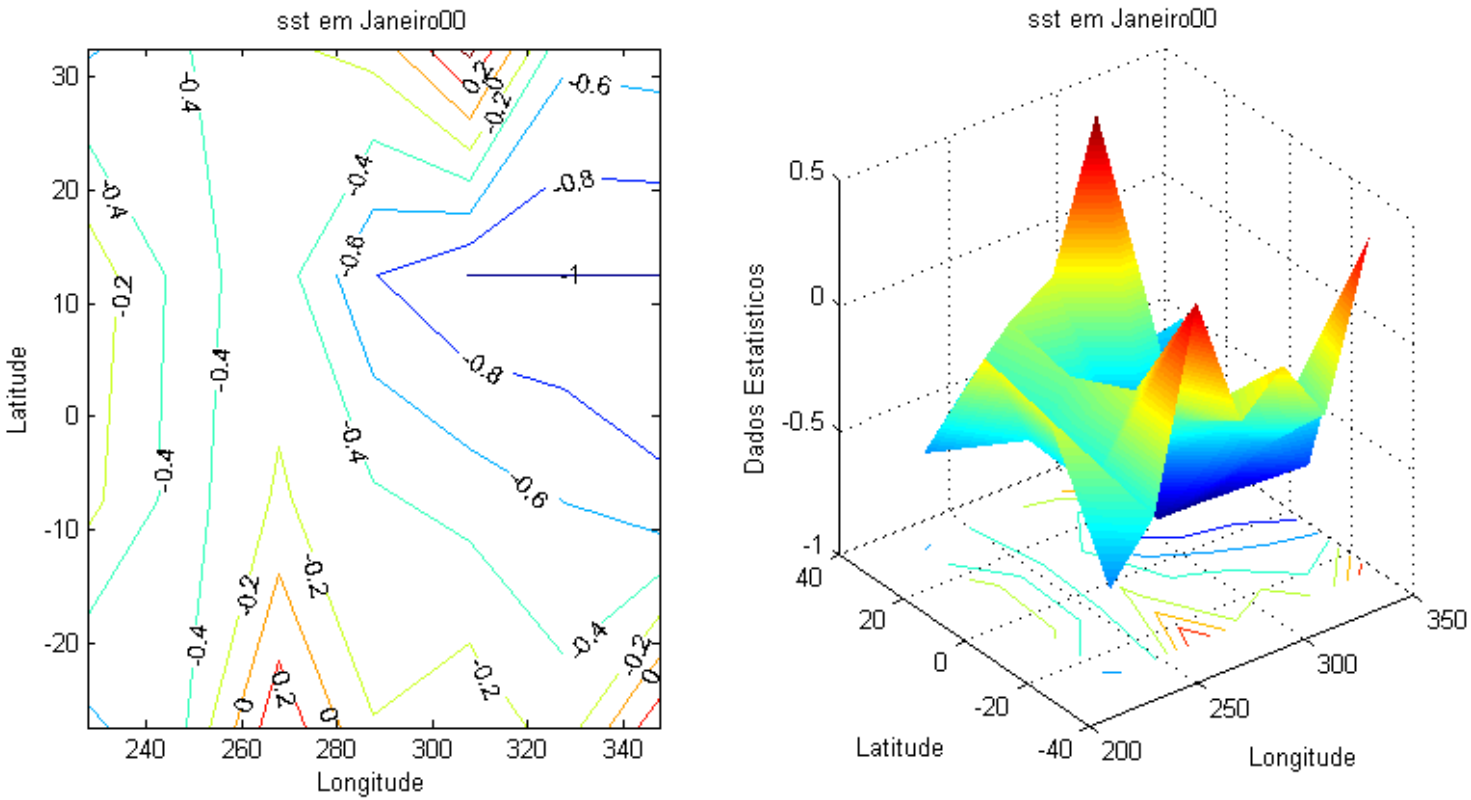


Figura A.16 – Dados da variável ambiental sst em janeiro de 2000.

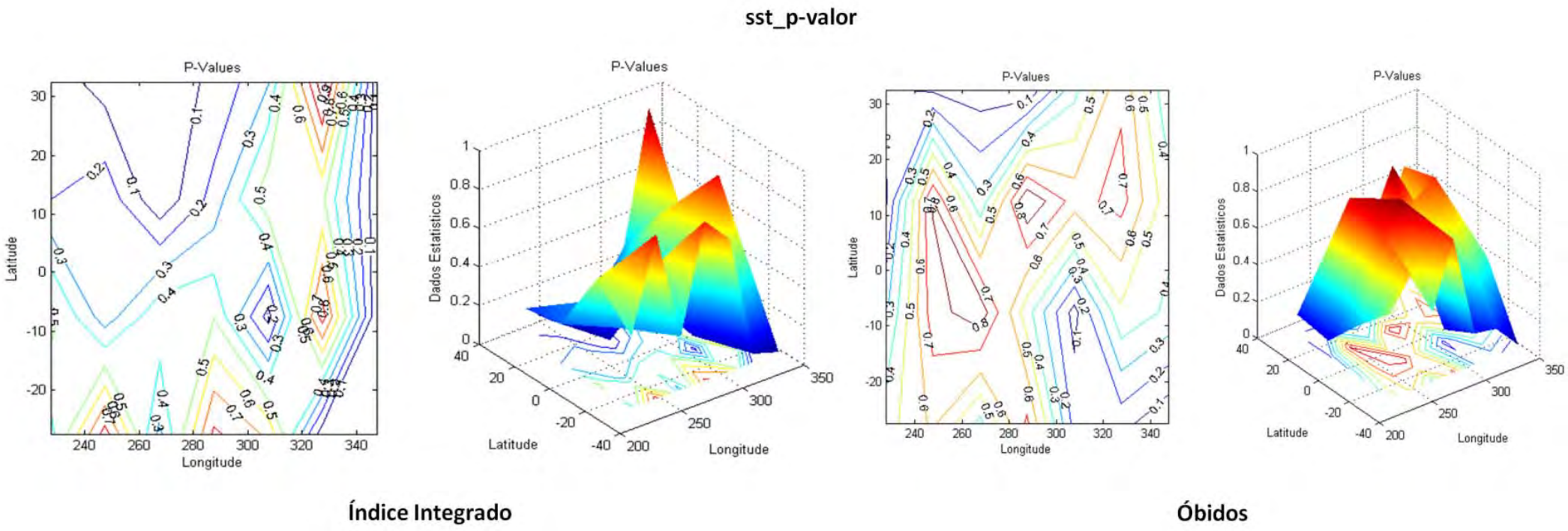


Figura A.17 – P-valor de sst no Índice Integrado e em Óbitos.

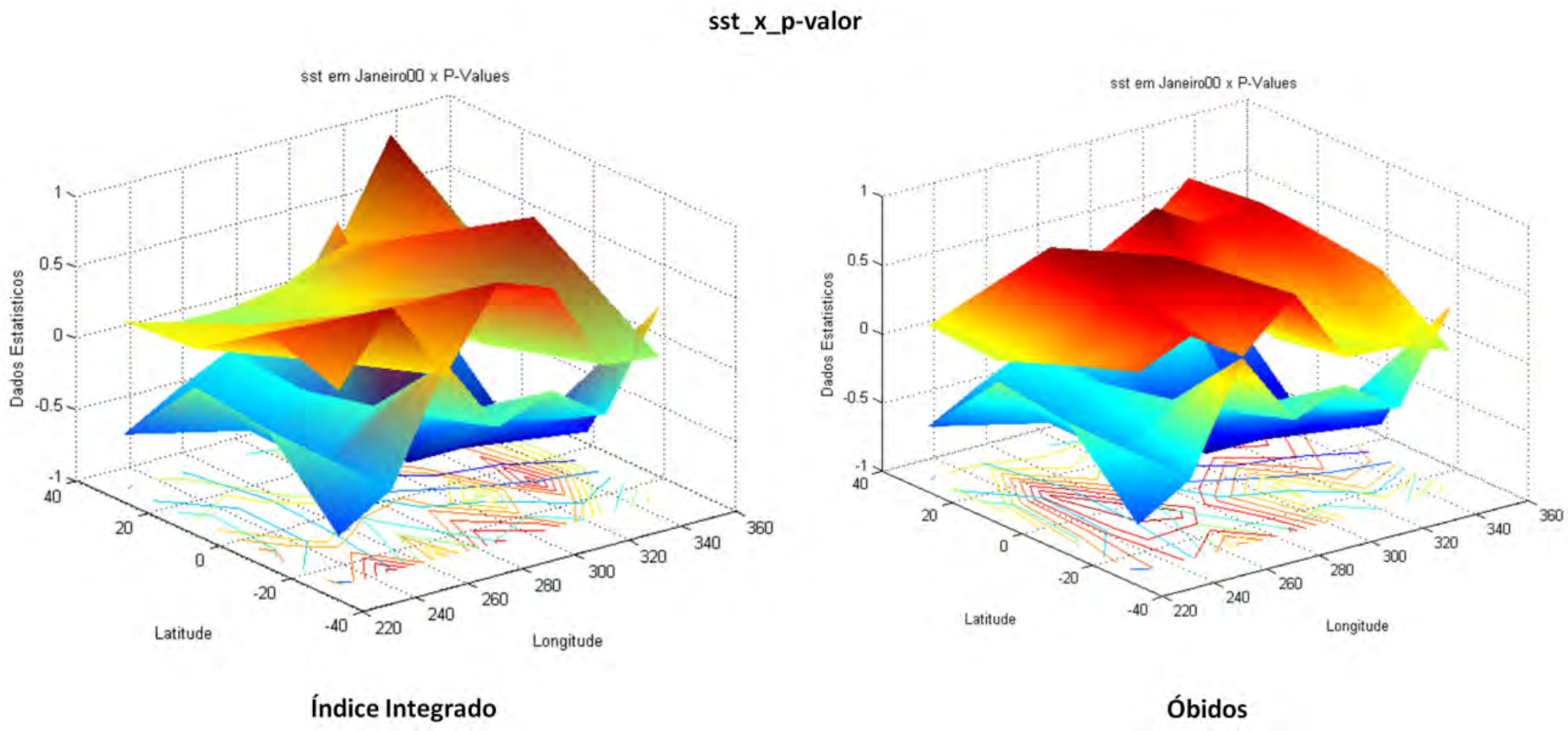


Figura A.18 – Dados de sst e seu p-valor sobrepostos.