MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E INOVAÇÃO
INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS

sid.inpe.br/mtc-m21b/2014/08.25.14.21-TDI

UM AMBIENTE VIRTUAL COLABORATIVO DE COMPUTAÇÃO CIENTÍFICA PARA ANÁLISE AVANÇADA DE SÉRIES TEMPORAIS

Murilo da Silva Dantas

Tese de Doutorado do Curso de Pós-Graduação em Computação Aplicada, orientada pelos Drs. Reinaldo Roberto Rosa e Nilson Sant'Anna, aprovada em 29 de agosto de 2014.

URL do documento original:
<http://urlib.net/8JMKD3MGP5W34M/3GT9JQB>

INPE
São José dos Campos
2014

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

MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E INOVAÇÃO
INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS

sid.inpe.br/mtc-m21b/2014/08.25.14.21-TDI

UM AMBIENTE VIRTUAL COLABORATIVO DE COMPUTAÇÃO CIENTÍFICA PARA ANÁLISE AVANÇADA DE SÉRIES TEMPORAIS

Murilo da Silva Dantas

Tese de Doutorado do Curso de Pós-Graduação em Computação Aplicada, orientada pelos Drs. Reinaldo Roberto Rosa e Nilson Sant'Anna, aprovada em 29 de agosto de 2014.

URL do documento original:
<http://urlib.net/8JMKD3MGP5W34M/3GT9JQB>

INPE
São José dos Campos
2014

Dados Internacionais de Catalogação na Publicação (CIP)

Dantas, Murilo da Silva.

D235a Um ambiente virtual colaborativo de computação científica para análise avançada de séries temporais / Murilo da Silva Dantas. – São José dos Campos : INPE, 2014.

xxvi + 175 p. ; (sid.inpe.br/mtc-m21b/2014/08.25.14.21-TDI)

Tese (Doutorado em Computação Aplicada) – Instituto Nacional de Pesquisas Espaciais, São José dos Campos, 2014.

Orientadores : Drs. Reinaldo Roberto Rosa e Nilson Sant'Anna.

1. Séries temporais. 2. DFA. 3. Sistemas colaborativos. 4. Computação em nuvem. 5. Ambientes virtuais. I.Título.

CDU 004.8

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

Computação Aplicada

Dr. Fernando Manuel Ramos

Presidente / INPE / SJCampos - SP

Dr. Reinaldo Roberto Rosa

Orientador(a) / INPE / SJCampos - SP

Dr. Nilson Sant'Anna

Orientador(a) / INPE / SJCampos - SP

Dra. Karine Reis Ferreira

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

Dr. José Carlos Becceneri

Membro da Banca / INPE / SJCampos - SP

Dr. Maurício José Alves Bolzam

Convidado(a) / UFG / Goiás - GO

Dra. Paula Rodrigues Teixeira Coelho

Convidado(a) / Univ. Cruzeiro do Su / São Paulo - SP

Este trabalho foi aprovado por:

( ) maioria simples

(X) unanimidade

Aluno (a): ***Murilo da Silva Dantas***

São José dos Campos, 29 de Agosto de 2014

"O conhecimento é como um jardim: se não for cultivado, não pode ser colhido."

Provérbio africano

“O conselho da sabedoria é: procure obter sabedoria; use tudo o que você possui para adquirir entendimento.”

Provérbio judeu

Dedico este trabalho a Deus e à Kéola, minha esposa tão amada!

AGRADECIMENTOS

Tantos anos dedicados aos estudos são frutos de muitas pessoas especiais que passaram ou ainda estão em minha vida.

A primeira delas é o meu Deus. A Ele que resolveu em Seu infinito amor relaciona-se comigo, ser meu amigo e Pai por toda a eternidade. Sem Ele, certamente eu jamais poderia ser quem eu sou, nem fazer o que preciso fazer. A Ele toda a minha honra prestada!

Em especial, também, agradeço a minha esposa, Kéola. Ela é sem dúvidas a mulher mais fantástica que conheço. Obrigado pelo amor e carinho de todos os dias. Sem você, jamais estaria onde estou. Meus dias são perfeitos, pois você está neles!

Minha família é meu alicerce! Agradeço também a meus pais, Lourival e Marli; a meus irmãos, Arthur e Diêgo; as minhas cunhadas, Sarah e Katiane; a meus sogros, Cleones e Lucinélia; e a todos os Duarte da Silva e os Dantas das Virgens, dos quais me orgulho fazer parte!

Reinaldo Rosa é uma daquelas pessoas que levamos para o resto de nossas vidas! Obrigado por me acolher aqui quando cheguei, pela amizade ao longo desses anos e pelo privilégio que tenho de ser orientado por um dos grandes pesquisadores deste país!

Quando assisti a uma palestra de Nilson Sant’Anna em seu retorno do pós-doc, pensei: “este cara é o cara!”. Não pensei duas vezes em chamá-lo para a orientação. Muito obrigado por seu esmero e talento, além do privilégio de me aceitar em seu time de orientandos.

Sou muito grato aos meus orientadores, pois eles me inspiram!

Agradeço também às seguintes pessoas:

Ao Moacyr Cereja, por sua capacidade em programação e envolvimento neste trabalho. Obrigado por tudo, amigo!

Aos professores da CAP e aos que direta ou indiretamente forjaram a minha vida acadêmica. Muito obrigado por todo esclarecimento!

Aos colegas do INPE, principalmente àqueles que se tornaram amigos, como Rudinei, Laurita, Thalita, Sóstenes, Juliana e Flávia.

A minha linda igreja, a Primeira Igreja Batista em São José dos Campos-SP, minha segunda casa, em especial, aos meus pastores e aos meus amigos do coração: Carlito, Leila, Fabiano, Viviam, Fábio, Natália, Yan, Fábio Anderson e Andrei.

À FATEC-SJC pela oportunidade de ser um de seus professores, pelo incentivo financeiro e pelas pessoas que a tornam um lugar tão favorável ao desenvolvimento de novas tecnologias. Lá, tenho tecido minha missão na educação de uma nação.

À Banca examinadora pelas sugestões de melhorias para este trabalho.

RESUMO

Esta tese aborda resultados da pesquisa de doutorado na área de sistemas colaborativos com ênfase na análise avançada de séries temporais. Apresenta um estudo e uma pesquisa inédita sobre tendências e demandas para sistemas disponibilizados na Internet com ferramentas matemáticas e estatísticas validadas e acopladas num único ambiente computacional, disponível via web usando conceitos atuais de sistemas colaborativos para pesquisa científica, como Computação em Nuvem e Arquitetura Orientada a Serviço. No contexto do INPE, uma pesquisa como essa, através de um ambiente computacional próprio, sugere um novo paradigma na análise de dados em ciências espaciais e ambientais. A inserção de dados em ambientes virtuais colaborativos, dependendo da aplicação, é capaz de proporcionar serviços de monitoramento e previsão de forma mais direta e compacta utilizando tecnologias de dispositivos móveis, disponíveis em larga escala e a baixo custo. Para isso, a pesquisa foi desenvolvida em duas frentes: a primeira está relacionada à pesquisa de ferramentas de análise de dados que, segundo critérios definidos na tese, serão aplicadas por desenvolvedores de ambientes colaborativos para análise avançada de séries temporais; já a segunda, está relacionada à pesquisa de metodologias e propriedades de ambientes virtuais que possam incorporar a análise identificada na primeira parte. Portanto, esta tese apresenta uma revisão bibliográfica inédita acerca de ambientes desta natureza, destacando também as principais tecnologias envolvidas num ambiente virtual. Além disso, como o contexto é o de análise de séries temporais, é apresentado como estudo de caso as propriedades de técnicas convencionais na classificação de séries curtas em comparação com a Análise de Flutuação Destendenciada (DFA). Esta técnica foi escolhida como um exemplo de ferramenta de análise avançada de dados medidos diretamente no domínio do tempo. A DFA cumpre dois critérios para uma análise avançada de séries temporais em ambientes virtuais: ser robusta também para séries temporais curtas e apresentar como resultado um único parâmetro. Com isso, a pesquisa produz como principal resultado um protótipo para um novo ambiente virtual colaborativo dedicado à análise avançada de séries temporais.

A SCIENTIFIC COMPUTING COLLABORATIVE VIRTUAL ENVIRONMENT FOR ADVANCED ANALYSIS OF TIME SERIES

ABSTRACT

This thesis presents the results of doctoral research in the area of collaborative systems' area with emphasis on advanced time series analysis. Presents a study and an unpublished research on trends and demands for systems available on the Internet with mathematical and statistical validated tools and coupled into a single computing environment, available via the Web using current concepts of collaborative systems for scientific research, such as Cloud Computing and Service-Oriented Architecture. In the context of INPE, a search like this, through their own computing environment, suggests a new paradigm in data analysis in space and environmental sciences. Entering data in collaborative virtual environments, depending on the application, is able to provide monitoring and forecast of more direct and compact form using mobile technologies available on a large scale and at low cost. Furthermore, this research was conducted on two fronts: the first is related to the research data analysis which, according to criteria defined in theory, be applied by developers of collaborative environments for advanced time series analysis tools; the second is related to research methods and properties of virtual environments that can incorporate the analysis identified in the first part. Therefore, this thesis presents a new literature review concerning the environments of this nature, also highlighting key technologies involved in a virtual environment. Moreover, as the context of this work is the analysis of time series it is presented, as a case study, the properties of conventional techniques in the classification of short series compared with Detrended Fluctuation Analysis (DFA). This technique chosen as an example of a tool for data analysis directly in the time domain. The DFA meets two criteria for an advanced time series analysis in virtual environments: being also robust for short time series and present the result as a single parameter. Finally, the research produces as main result a prototype for a new virtual collaborative environment dedicated to the advanced time series analysis.

LISTA DE FIGURAS

Pág.

LISTA DE TABELAS

Pág.

LISTA DE SIGLAS E ABREVIATURAS

API	Do Inglês, *Application Programming Interface*
AR	Modelos Autorregressivos
ARMA	Modelos Mistos
BDA	Do inglês, *Brazilian Decimetric Array*
BPEL	Do inglês, *Business Process Execution Language*
CSEO	Do inglês, *Computational Science and Engineering On-line*
DCA	Do inglês, *Data Centre Alliance*
DFA	Do inglês, *Detrended Fluctuation Analysis*
TC	Do inglês, *Technology Centre*
DFA	Do inglês, *Detrended Fluctuation Analisys*
EADA	Equipe de Análise de Dados e Algoritmos
EC2	Do inglês, *Elastic Computing Cloud*
EES	Equipe de Engenharia de Software
EGR	Equipe de Gestão de Recursos
EMBRACE	Estudo e Monitoramento Brasileiro do Clima Espacial
ENPC	Do francês, *École Nationale des Ponts et Chaussées*
ERAD	Equipe de Redes e Alto Desempenho
EURO-VO	Do inglês, *European Virtual Observatory*
FC	Do inglês, *Facility Centre*
fMRI	Do inglês, *Functional Magnetic Resonance Imaging*
GEODISE	Do inglês, *Grid Enabled Optimisation and Design Search for Engineering*
GPA	Do inglês, *Gradient Pattern Analysis*
GSA	Do inglês, *Gradient Spectral Analysis*
HTML	Do inglês, *HyperText Markup Language*
HTTP	Do inglês, *HyperText Transfer Protocol*
HTTPS	Do inglês, *HyperText Transfer Protocol Secure*
IaaS	Do inglês, *Infraestructure as a Service*
IDE	Do inglês, *Integrated Development Environment*

IMAGINE	Do inglês, *Integrated Multidisciplinary Autonomous Global Innovation Networking Environment*
INPE	Instituto Nacional de Pesquisas Espaciais
INRIA	Do francês, *Institut National de Recherche en Informatique et en Automatique*
IP	Do inglês, *Internet Protocol*
IVOA	Do inglês, *International Virtual Observatory Alliance*
JVM	Do inglês, *Java Virtual Machine*
LAC	Laboratório Associado de Computação e Matemática Aplicada
LAF	Laboratório de Agricultura e Floresta em Sensoriamento Remoto
MA	Modelos de Médias Móveis
MAS	Do inglês, *Multifractal Analysis Spectra*
NSF	Do inglês, *National Science Foundation*
NVO	Do inglês, *National Virtual Observatory*
OASIS	Do inglês, *Organization for the Advancement for Structured Information Standards*
PaaS	Do inglês, *Platform as a Service*
PDF	Do inglês, *Portable Document Format*
PSD	Do inglês, *Power Spectrum Density*
QoS	Do inglês, *Quality of Service*
RMS	Do inglês, *Root Mean Square*
SaaS	Do inglês, *Software as a Service*
SBSE	Do inglês, *Search-based Software Engineering*
SMS	Do inglês, *Short Message Service*
SOA	Do inglês, *Service Oriented Architecture*
SOAP	Do inglês, *Simple Object Access Protocol*
ST	Série Temporal
STC	Série Temporal Curta
TCP	Do inglês, *Transmission Control Protocol*
TI	Tecnologia da Informação
UDDI	Do inglês, *Universal Description, Discovery and Integration*

VLADA	Do inglês, *Virtual Laboratory for Advanced Data Analysis*
VLK	Do inglês, *Virtual Library Knowledge*
VO	Do inglês, *Virtual Observatory*
W3C	Do inglês, *World Wide Web Consortium*
WSDL	Do inglês, *Web Service Description Language*
XML	Do inglês, *eXtended Markup Language*

SUMÁRIO

1 INTRODUÇÃO

Um dos principais aspectos socioeconômicos do mundo moderno é o desenvolvimento da ciência. A ciência, como sabemos, também é parte inerente da cultura ocidental e o seu desenvolvimento está intimamente ligado à inovação tecnológica, em especial àquela relacionada a novos instrumentos. Assim, há uma relação estreita entre ciência e tecnologia, pois os instrumentos tecnológicos permitem que a ciência avance e a ciência possibilita o desenvolvimento de novas tecnologias (KUHN, 2007; STOKES, 2005).

Várias áreas da ciência, principalmente a Ciência da Computação e suas aplicações, têm passado atualmente por uma intensa transformação (LUZ, 2012). Essa nova forma de fazer ciência está intimamente ligada à exploração de dados, seja na sua coleta, na sua transmissão ou na sua análise e divulgação. Ela é descrita como o "quarto paradigma"[1] da ciência segundo James Nicholas Gray[2] citado em HEY et al., (2009). Esse paradigma é definido como *e-Science* e faz uso intenso de tecnologias computacionais.

Além da tecnologia, a ciência avança por meio da colaboração entre os pesquisadores, seja através de relacionamento informal ou através de projetos e acordos formais. As relações sociais entre os cientistas determinam as ligações entre as ideias dos mesmos (ZIMAN, 1979; BERNAL, 1939). A produção científica está ligada à maneira como os cientistas se comportam, se organizam e se relacionam.

Projetar um ambiente colaborativo envolve muitas ações e, eventualmente,

[1] Os paradigmas científicos na tentativa de descrever os fenômenos naturais são apresentados no livro homônimo de HEY et al., (2009). São eles: (1) a Ciência Empírica, cultivada há milhares de anos, que tem como fundamentação a experiência e a observação; (2) a Ciência Teórica, cultivada há algumas centenas de anos, que usa modelos e generalizações; (3) a Ciência Computacional, cultivada há algumas décadas, que, devido ao crescente poder computacional, permite a simulação de fenômenos complexos; e (4) a Ciência da Exploração dos Dados (*e-Science* ou *electronic science*), que é o atual paradigma e busca unificar a experimentação, a teoria e a simulação através de sistemas computacionais.

[2] Gray foi um cientista da computação americano ganhador do Prêmio Turing de 1998. Ele foi um pesquisador especialista em bancos de dados, processamento de transações e desenvolvimento de sistemas.

uma pesquisa alinhada com a necessidade. O'BRIEN (2000) faz algumas recomendações para o sucesso de um sistema colaborativo: (i) indicação de especialistas para utilização do sistema; (ii) possibilidade de mensurar o desempenho individual da equipe através do sistema; (iii) incentivo do uso do sistema para as tarefas para os quais foi projetado em detrimento de ferramentas alternativas; (iv) mapeamento e análise os dados fornecidos; e (v) estabelecimento de regras de conduta para o uso do ambiente.

No contexto desta pesquisa, considera-se uma ferramenta de análise um componente de software, um programa ou rotina que processe uma massa de dados de modo a obter um resultado esperado em algum modelo científico apropriado.

Em geral, pesquisas que envolvem grandes quantidades de dados, seja na sua geração, transmissão ou manipulação, precisam de uma colaboração mais intensa entre os pesquisadores, pois têm uma demanda maior por especialistas em áreas diferentes. Atualmente, o conceito de *Big Data*[3] (WHITE, 2009 e MIKE, 2011) em Tecnologia da Informação consiste em conjuntos de dados que crescem muito rapidamente e exigem ferramentas avançadas para captura, armazenamento, busca, compartilhamento, visualização e, sobretudo, análise.

Essa demanda de bases de dados crescentes é alimentada pela necessidade na análise de dados em identificar tendências em negócios, prevenir doenças, monitoramento ambiental ou combater crimes (CUKIER, 2010). Cientistas de diferentes áreas do conhecimento – como: Meteorologia, Genômica, Sistemas Complexos, Biologia, Astronomia, Clima Espacial, Sistemas Terrestres e Mudanças Globais do Clima – deparam-se com este volume crescente de dados.

Uma característica atual de bases de dados com rápido crescimento é a dificuldade em trabalhar com elas usando pacotes de visualização ou análise

[3] A expressão pode ser traduzida como "grandes volumes de dados". Optou-se em utilizar a expressão inglesa por ser a mais difundida.

usando micro computador de mesa. Ao invés disso, a manipulação desses dados requer software massivamente paralelo executado em dezenas ou até milhares de servidores e/ou compartilhado através de uma arquitetura baseada em serviço (JACOBS, 2009).

O INPE é uma das instituições envolvidas intensamente com o problema da geração de grandes volumes de dados – coletados por diferentes tipos de sensores, nas áreas de ciências, tecnologias e aplicações espaciais e ambientais – gerando uma grande demanda por novos métodos, modelos e algoritmos, baseados em técnicas numéricas e estatísticas. Estas atividades demandam, em sua maioria, recursos computacionais para tratamento, processamento, armazenamento e extração da informação, de maneira automatizada, viabilizando o uso dessas informações para responder às necessidades e os desafios específicos das áreas de aplicações.

Grande parte dos dados científicos é coletada na forma de série temporal definida como “qualquer conjunto de observações ordenadas no tempo” (MORETTIN e TOLOI, 2006). No contexto desta tese, a série temporal será denominada de ST e uma série temporal curta, a partir deste ponto será denominada de STC, que é um conjunto ordenado insuficiente de medidas discretas para uma caracterização fenomenológica baseada em técnicas estatísticas (DANTAS, 2008).

Existem muitas técnicas e métodos para a análise de ST’s. Esta tese aborda técnicas cujo desenvolvimento ou aplicação estão fundamentados em conceitos da teoria de sistemas complexos, enfatizando o caráter avançado deste tipo de análise. Tais técnicas analisam, em geral, flutuações relacionando os processos de fenômenos específicos do sistema estudado, objetivando entender os regimes dinâmicos ligados a processos periódicos, intermitentes, caóticos e turbulentos que envolvam transições de fase (KANTZ e SCHREIBER, 1997). Há vários algoritmos que permitem a análise de não linearidades e escalas, como o cálculo das dimensões generalizadas

(dimensão de correlação e dimensão Hausdorff), juntamente com seus expoentes dinâmicos (Lyapunov, Hurst, etc.) (FREITAS, 2012). Em geral, todos os pacotes de análise de dados suportam uma ampla variedade de métodos tradicionais. Entretanto, são limitados para lidar com séries temporais provenientes de processos não lineares e complexos utilizando ferramentas avançadas, como as listadas na Tabela 1.1. Essa limitação se traduz como um campo de pesquisa de um ambiente colaborativo para esse tipo de análise.

Tabela 1.1 – Uma lista de referência de técnicas relativamente recentes que têm sido utilizadas, especialmente em ciências físicas, para análise avançada de séries temporais ao longo das últimas décadas.

Técnica	**Sigla**	**Principal resultado**	**Referências**
Análise de Flutuação Destendenciada	DFA[4]	Expoente de Escala	VERONESE et al., 2011; BARONI et al., 2010; HU et al., 2001; e PENG et al., 1994.
Gráfico de Recorrência	RP[5]	Padrão da Dinâmica	MARWAN et al., 2008 e ECKMANN et al., 1987.
Análise Espectral Multifractal	MSA[6]	Espectro de Singularidade	BAI e ZHU, 2010 e BULDYREV et al., 1995.
Análise de Padrões Gradientes	GPA[7]	Coeficiente de Assimetria Gradiente	ROSA et al., 2008; ASSIREU et al., 2002; RAMOS et al. 2000; ROSA et al. 1999.
Análise Espectral de Padrões-Gradiente	GSA[8]	Espectro de Assimetria Gradiente	DANTAS, 2008

Ao propor um modelo para tal ambiente, esta tese foi desenvolvida considerando as seguintes motivações:

1. *A quantidade crescente de dados disponibilizados para análise*. O reflexo disso é a capacidade tecnológica mundial *per capita* para armazenar informações que praticamente dobrou a cada 40 meses

[4] Do inglês, *Detrended Fluctuation Analysis*.
[5] Do inglês, *Recurrence Plot.*
[6] Do inglês, *Multifractal Spectral Analysis*.
[7] Do inglês, *Gradient Pattern Analysis.*
[8] Do inglês, *Gradient Spectral Analysis.*

desde a década de 1980 (SEGARAN; HAMMERBACHER, 2009; HELLERSTEIN, 2008);

2. *A demanda crescente de colaboração científica em projetos de pesquisas que envolvam grandes quantidades de dados*. Atualmente há muitos consórcios multidisciplinares em torno da exploração de dados. Casos como os observatórios virtuais, os ambientes de *e-Science* e de simulação em larga escala, como os apresentados no próximo capítulo;

3. *A modernização das redes de dados (celulares e banda larga)*. Tais serviços estão numa evolução cada vez mais rápida aumentando suas possibilidades de uso pelos cientistas;

4. *A popularização da computação móvel.* O impacto da miniaturização e a revolução da digitalização nas telecomunicações e na eletrônica têm possibilitado uma difusão em larga escala dos dispositivos móveis, como *notebooks*, *netbooks*, celulares, *smartphones*, *tablets*, etc. Há uma previsão de que em 2020 o mundo tenha cerca de 55 bilhões de dispositivos desta natureza (SIQUEIRA, 2011). Esse crescimento prevê que a computação móvel se consolidará como a tecnologia que mais se expandiu e alcançou tantos seres humanos em tão pouco tempo, permitindo o acesso remoto a bases de dados e a ambientes virtuais de análise desses dados disponíveis na nuvem;

5. *O conceito de recursos mínimos*. Em geral, ambientes para análise de dados exigem do usuário equipamentos de grande desempenho e programas licenciados proprietários instalados nestes equipamentos. É uma vantagem importante o cientista não precisar destes equipamentos e programas para efetuar uma análise de dados, mas somente os dados (que podem inclusive ser inseridos no ambiente virtual previamente) e o acesso à Internet através de um dispositivo móvel;

6. *As possibilidades de desenvolvimento com Arquitetura Orientada a Serviços (SOA).* Esta arquitetura permite a criação de componentes de software que sejam reutilizáveis e extensíveis com alta capacidade de integração. Nela as interfaces são bem definidas através dos componentes disponibilizados como serviços. Isso permite o acoplamento de novas funcionalidades mais rapidamente do que em aplicações monolíticas;

7. *A demanda de ambientes virtuais robustos que permitam a análise avançada de ST's através de rotinas automatizadas.* De modo geral, há um uso crescente dos computadores no processamento, visualização e análise de dados envolvendo novos métodos com aprimoramento de arquiteturas computacionais, banco de dados, algoritmos e técnicas matemáticas para análise. Isso tem proporcionado novos conhecimentos científicos, melhor caracterização do padrão de variabilidade temporal em fenômenos dinâmicos e automatização de sistemas de controle e de alerta com monitoramento em tempo real de vários processos observados na natureza. O avanço tecnológico dos sistemas computacionais nas últimas décadas possibilitou o aprimoramento de técnicas e metodologias de análise de ST's. Portanto, a automatização do uso de técnicas de análise de ST's em ambientes virtuais têm se tornado um tópico recente de pesquisa em Computação Aplicada. Muitos desses ambientes podem ser complexos o suficiente para explorarem também os benefícios da Computação Paralela, da Computação Distribuída e da Computação em Nuvem.

A partir das motivações acima, foi identificado um conjunto de características desejáveis, listado na Tabela 1.2, para ambientes virtuais colaborativos robustos para a análise avançada de séries temporais. Além disso, foi realizado um levantamento sobre ambientes virtuais colaborativos disponibilizados na *Web* para computação científica. A Tabela 1.3 apresenta os ambientes eleitos e descritos no Capítulo 2.

Tabela 1.2 – Lista de características desejáveis em um ambiente virtual colaborativo robusto para a análise avançada de séries temporais

Identificação	Descrição
1	Interface: simples e amigável.
2	Multiusuário: uso simultâneo por vários usuários.
3	Customização: personalização de acordo com o perfil do usuário, oferecendo uma interface adequada para cada tipo.
4	Ferramentas: disponibilização de técnicas de análise básica e avançada de ST's e STC's automatizadas.
5	Escalabilidade: permissão para a inclusão de novas técnicas e métodos de acordo com um processo pré-estabelecido e amplo.
6	Controle: permissão por parte do usuário para um controle total de experimentos com inclusão de novos dados inclusive.
7	Usabilidade: possibilidade de conhecimento especialista baseado em regras para a definição de ajustes no ambiente de acordo com as preferências do usuário.
8	Padronização: diminuir a necessidade de suporte técnico e a curva de aprendizado.
9	Integração: uso de padrões abertos através da Internet para suportar a integração e interoperabilidade com outros pacotes de softwares, locais ou remotos, permitindo o uso de aplicações remotas.
10	Duração: controle adequado em experimentos com tarefas de longa duração.
11	Escalonamento: escolha de tarefas de acordo com a disponibilidade de recursos.
12	Validação: fluxo adequado de validação de novas ferramentas por usuários especialistas do ambiente.
13	Baixo custo: acessível ao usuário.
14	Dados: capacidade de análise de grandes quantidades de dados a partir de múltiplos acessos simultâneos.
15	Distribuição: execução de módulos em diferentes locais no planeta.
16	Mobilidade: acesso a partir de diferentes dispositivos móveis.
17	Cliente leve: acesso remoto à interface. Possibilidade de usar o sistema sem programas pré-instalados.
18	SOA: disponibilização dos componentes como serviços.
19	*Web*: acesso ao sistema através de interface *web*.
20	Multilinguagem: possibilidade de integração do código com diferentes linguagens computacionais.
21	Documentação: possibilidade de manter documentação científica relacionada.

Tabela 1.3 – Sistemas eleitos nas buscas. Os itens em cinza são para plataforma PC e os demais para plataforma totalmente *Web*.

Identificação	Sistema
A	Matlab
B	Octave
C	Scilab
D	R-Project
E	NVO
F	EURO-VO
G	IVOA
H	Chimera
I	GEODISE
J	GridPP
K	GridChem
M	LAF

Tabela 1.4 – Presença das características desejáveis (linhas) nos ambientes avaliados (colunas): os itens em cinza são para plataforma PC e os demais para plataforma totalmente *Web*.

	A	B	C	D	E	F	G	H	I	J	K	L
1	X	X	X	X					X		X	X
2	X	X	X	X	X	X	X		X	X	X	X
3											X	
4	X	X	X	X								X
5	X	X	X	X	X	X	X		X	X	X	
6	X	X	X	X				X	X	X		X
7	X		X	X	X	X	X					
8	X	X	X	X	X	X	X	X	X	X	X	X
9	X	X	X	X					X	X		
10										X	X	
11							X			X	X	
12												
13		X	X	X	X	X	X	X	X	X	X	X
14	X	X	X	X	X	X	X		X	X		X
15	X	X	X	X	X	X	X		X	X	X	
16	X	X			X	X	X					X
17	X	X			X	X						X
18	X	X	X	X	X	X	X		X			
19	X	X	X	X	X	X	X					X
20	X	X	X	X			X					
21	X	X		X	X	X	X			X	X	X

Muitos sistemas foram encontrados nas buscas efetuadas. Para determinar quais seriam testados, como ponto de partida para o novo ambiente proposto, foram usados os seguintes critérios: estarem na lista dos três primeiros encontrados em cada busca e terem uma interação ativa com o usuário, através de simulações e uso de ferramentas disponibilizadas que permitissem o desenvolvimento de pesquisa científica colaborativa. A Tabela 1.4 apresenta uma relação entre as características desejáveis para um ambiente virtual colaborativo de computação científica para análise avançada de séries temporais da Tabela 1.2 e os sistemas escolhidos da Tabela 1.3.

É possível notar através da Tabela 1.4 que nenhum dos sistemas possui todas as características mínimas consideradas, provavelmente devido à amplitude requerida. Tais ambientes não atendem às necessidades listadas na Tabela 1.2 e detalhadas no Capítulo 4, reforçando a tese de que, para um sistema dessa natureza, é apropriado o uso de componentes especialistas distribuídos e construídos sobre uma plataforma aberta e em software livre. Assim, o ambiente proposto apresenta-se como uma contribuição científica inédita para a área de Computação Aplicada.

Baseado nas considerações acima, esta tese tem por objetivo principal apresentar como pesquisa em nível de doutorado uma infraestrutura de software colaborativo orientado a serviços distribuídos para a análise avançada tanto de séries temporais longas, quanto de séries temporais curtas, envolvendo diferentes linguagens e plataformas computacionais. Para isso, a pesquisa especificou: (i) os critérios que permitiram avaliar a adequabilidade de ferramentas de análise avançada de ST’s em ambientes virtuais; (ii) uma avaliação dos modelos de desenvolvimento e de metodologias da Engenharia de Software que direcionaram a proposta do protótipo, fornecendo uma arquitetura de referência para ambientes colaborativos dessa natureza; (iii) uma estrutura para conexão de serviços de análise de ST’s; e (iv) um portal para acesso a serviços de análise de ST’s.

Este objetivo foi atingido a partir das seguintes atividades:

- Estudo dos ambientes disponíveis na *web* de computação científica;
- Avaliação das tecnologias disponíveis para construção de ambientes dessa natureza;
- Estudo comparativo entre técnicas de análise de séries temporais para uso em ST's longas e curtas;
- Pesquisa acerca dos critérios para a escolha de técnicas avançadas de análise de ST's em ambientes virtuais;
- Pesquisa acerca das propriedades de Engenharia de Software que serviram como diretrizes para o desenvolvimento de ambientes virtuais para análise avançada de séries temporais; e
- Desenvolvimento e teste de um protótipo.

Um ambiente virtual colaborativo de computação científica para análise avançada de séries temporais pode atender demandas, tanto no escopo de pesquisa ambiental e espacial do INPE, quanto no de uma rede de parceiros, com intercâmbio de técnicas, algoritmos e bases de conhecimento formando uma nuvem de serviços. Tal ambiente pode desenvolver soluções inteligentes e eficientes que permitam a redução, o gerenciamento, a integração e a disseminação de informações, muitas vezes heterogêneas, provenientes de diferentes fontes. Além disso, facilitaria a extração e a interpretação das informações úteis para as áreas de ciências, tecnologias e aplicações espaciais e ambientais a partir de séries brutas de dados.

O restante desta tese está organizado como segue: o Capítulo 2 apresenta de forma mais detalhada o conceito de colaboração científica e suas motivações, além dos tipos de ferramentas utilizadas de maneira geral na colaboração científica. Discute os principais conceitos e tecnologias atuais passíveis de utilização num ambiente virtual colaborativo como o proposto e também expõe uma revisão bibliográfica acerca do escopo da tese. O capítulo finaliza apresentando alguns sistemas virtuais colaborativos para a computação

científica encontrados na Internet; o Capítulo 3 contém uma discussão conceitual sobre as técnicas de análise de séries temporais, enfatizando a DFA utilizada neste trabalho; o Capítulo 4 discute os requisitos para o ambiente proposto; no Capítulo 5 é apresentada a arquitetura do protótipo proposto e suas principais funcionalidades; o Capítulo 6 apresenta o protótipo em seu estado funcional atual; no Capítulo 7 são apresentados testes efetuados em dados ambientais usando o protótipo; e o último capítulo destaca as considerações finais e perspectivas futuras para o trabalho proposto. O Apêndice A contém um exemplo de casos de uso; o Apêndice B contém um exemplo de script de banco de dados do protótipo e, finalmente, o Anexo A contém material das publicações e da participação em eventos relacionados ao desenvolvimento desta tese.

2 COLABORAÇÃO CIENTÍFICA E AMBIENTES VIRTUAIS

Neste capítulo é apresentado o conceito da colaboração científica. Em seguida, são destacadas as motivações e os tipos de ferramentas para este tipo de colaboração. São também apresentados conceitos e tecnologias para ambientes virtuais colaborativos e, na sequência, é exposta uma revisão bibliográfica sobre ambientes virtuais colaborativos. Finalmente, são apresentados alguns ambientes virtuais colaborativos para computação científica disponíveis na Internet envolvendo diferentes áreas do conhecimento, a partir de um levantamento sobre tais ambientes, boa parte, disponibilizados na *web*.

2.1. Definição de Colaboração

A palavra “colaboração” vem do latim “*collaborare*” e está associada como “ajuda, cooperação, participação, coadjuvação”. O conceito de colaboração científica é amplo e não há um consenso na comunidade acadêmica. A mesma não só se refere à realização de tarefas, mas também ao compartilhamento do significado dessas tarefas de pesquisa entre os pesquisadores (SONNENWALD, 2008). A colaboração nesse sentido envolve revisão de trabalhos, normas de instituições envolvidas, políticas nacionais ou internacionais de pesquisa científica, paradigmas científicos, dentre outros aspectos.

A colaboração ocorre entre cientistas de um departamento ou programa de pós-graduação (SILVA et al., 2006; MAIA e CAREGNATO, 2008) ou entre diferentes departamentos de uma mesma universidade (WANG et al., 2005). Também é realizada entre grupos de pesquisa e comunidades de uma área do conhecimento (HOU, KRETSCHMER e LIU, 2008); entre diferentes instituições e pode envolver setores como, governo, universidade e iniciativa privada (LETA, GLÄNZEL e THIJS, 2006) em diferentes regiões geográficas e países (WANG et al., 2005; ZHANG e GUO, 1997).

Os primeiros estudos sobre colaboração científica são do final da década de 1950 (SONNENWALD, 2008). Desde então diversos pesquisadores têm estudado esse tema envolvendo setores da sociedade, instituições de pesquisa, diferentes áreas do conhecimento ou países. Atualmente, a colaboração científica é um importante tema de pesquisa acadêmica. Desde o ano 2000 há inclusive uma rede internacional de pesquisa, denominada COLLNET[9], sobre como a colaboração se dá em diferentes áreas do conhecimento.

Entretanto, o conceito de colaboração vem se expandindo com o tempo, principalmente com o apoio de sistemas computacionais. Há algumas expressões usadas para este conceito, como *groupware* e *Computer Supported Cooperative Work* (CSCW) desde o final da década de 1970 (JOHNSON-LENZ e JOHNSON-LENZ, 1998). Tais expressões são usadas para especificar tanto sistemas que apoiem a colaboração, quanto seus efeitos (PIMENTEL e FUKS, 2011). É a partir dessas expressões que surgiu no Brasil a expressão "sistemas colaborativos", que reflete a ideia de um "sistema baseado em computador para dar suporte a grupos de pessoas engajadas numa tarefa comum e que provê uma interface para um ambiente compartilhado" (ELLIS et al, 1991).

2.2. Motivações para a Colaboração Científica

É possível mencionar alguns motivos pelos quais se dá a colaboração científica: a interdisciplinaridade da Ciência; o alto custo de equipamentos e laboratórios possivelmente compartilháveis; os fatores sociais de vínculos profissionais; o desenvolvimento das redes de computadores e da Internet; a ampliação de eventos científicos; o aumento da produtividade; a racionalização do uso da mão-de-obra científica e do tempo dispensado à pesquisa; a obtenção e/ou ampliação de financiamentos, recursos, equipamentos especiais e materiais; o desejo de aumentar a própria experiência através da experiência

[9] Website: www.collnet.de.

de outros cientistas; a escrita colaborativa e o desenvolvimento de softwares que facilitam o trabalho em equipe (VANZ, 2009; BEAVER e ROSEN, 1978).

A colaboração científica pode estar relacionada também com a coautoria. PRICE (1976) apresenta um estudo baseado nos dados do *Chemical Abstracts*, de 1910 a 1960 e mostra que o número de artigos em coautoria passou de menos de 20%, em 1910, para mais de 60%, em 1960. Trabalhos mais recentes ratificam essa tendência: BALANCIERI et al. (2005), CRONIN (2005), KRETSCHMER (2004), SILVA (2002), dentre outros. Porém, a coautoria é apenas um aspecto da colaboração científica. Nem sempre os autores são os responsáveis pelo trabalho de pesquisa. Também é comum na Ciência os coautores honorários (KATZ e MARTIN, 1997). Além disso, nem sempre colaboradores publicam juntos (BORDONS e GÓMEZ, 2000).

2.3. Tipos de Ferramentas para a Colaboração Científica

A comunicação entre pesquisadores e a interação entre experimentos também são determinantes na colaboração. Atualmente, a Internet é o meio mais favorável à comunicação e ao desenvolvimento de pesquisas baseadas em técnicas de análise compartilhadas. Motivado pelo desenvolvimento das redes de computadores, a Internet facilita a interação e a colaboração entre cientistas.

De maneira geral, a Internet oferece muitas ferramentas que estimulam a colaboração científica, principalmente através de seu paradigma denominado *Web* 2.0 ou *Web* Social. A *Web* 2.0 estimula a colaboração, a disseminação da comunicação e a interação muitos-para-muitos, representando um novo padrão de interação (MOURA, 2009; PRIMO, 2008; ANTOUN, 2008; CAVALCANTI e NEPOMUCENO, 2007; CASTELLS, 2003). Nela há uma participação mais ativa do usuário “na criação, seleção e troca de informações, podendo não só acessar, mas também alterar o conteúdo a qualquer momento” (AMBINDER e MARCONDES, 2011). Veja a diferença em termos de interação entre Web 1.0 e Web 2.0 na Figura 2.1. Ambinder e Marcondes (2011) destacam ainda

algumas ferramentas da *Web* 2.0 com potencial para colaboração científica: *Blogs*, *Wikis*, Sites de Redes Sociais, *Folksonomias* ou "nuvens de tags", Compartilhamento de Vídeos, Compartilhamento de Apresentações ou *Slides* e Serviços de *Microblogs*.

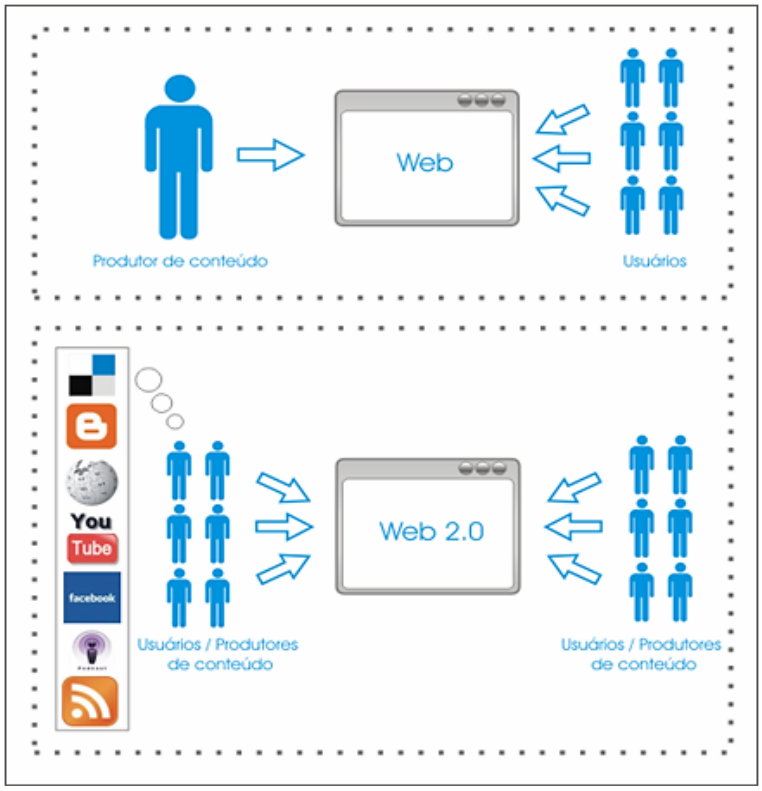


Figura 2.1 – Diferença de interação com o ambiente entre Web 1.0 e Web 2.0

Fonte: adaptada de COZIC (2007).

As ferramentas supracitadas podem ser utilizadas de maneira independente ou agrupadas, mas contribuem para a colaboração científica apenas de maneira mais geral e em termos de comunicação entre os parceiros. Para uma colaboração mais específica envolvendo compartilhamento de dados e técnicas de análise, existem sistemas sofisticados utilizados por comunidades científicas. Esses sistemas têm se beneficiado dos avanços tecnológicos na área digital e de sensores, que têm permitido a construção de meios de aquisição de informação cada vez mais precisos gerando grandes volumes de dados. Tais dados precisam ser armazenados, manipulados e tratados, para

serem processados e permitir a extração de informação para geração de conhecimento através da análise de dados.

2.4. Conceitos e Tecnologias para Ambientes Virtuais Colaborativos

Atualmente os computadores usam a rede como um requisito fundamental para o funcionamento de diversas aplicações. As máquinas estão cada vez menores e, com o avanço tecnológico, com aplicações mais distribuídas e interdependentes. Um ambiente virtual colaborativo precisa ser uma aplicação distribuída. Para que isso seja possível na análise avançada de séries temporais, é preciso o entendimento de alguns conceitos e tecnologias relacionadas, conforme a seguir.

2.4.1. Engenharia de Software

A Engenharia de Software permite o tratamento adequado a um dos desafios mais críticos em sistemas distribuídos que é a sua heterogeneidade. Tais ambientes devem ser aplicações capazes de lidar com a diversidade crescente de hardware e software disponibilizados. Isso é possível através de técnicas para construção de softwares confiáveis e flexíveis que lidam com sistemas legados antigos. Um exemplo relacionado ao ambiente aqui proposto é o acoplamento de aplicações legadas já disponibilizadas pelos usuários em outras máquinas conectadas à rede.

A Engenharia de Software também provê mecanismos mais assertivos para garantir transparência ao usuário. À medida que as pessoas ficam mais dependentes dos programas computacionais nas diversas áreas da experiência humana, mais se exige transparência naquilo que a aplicação se propõe a executar, não só no nível de programação, mas também no seu descritivo e

isso exige padronização e sistemática. LEITE (2011) faz uma discussão mais específica usando a ferramenta 5W2H[10].

Extrair do usuário as reais necessidades de requisitos para o software de forma mais precisa, diminui o impacto de ações corretivas durante o processo de desenvolvimento. A Engenharia de Software prevê ações como essa, assim como uma atenção especial por parte do desenvolvedor no quesito de usabilidade: o usuário está cada vez mais interessado em um menor tempo de resposta do software, uma maior integração com a rede e com Internet, além de interfaces mais intuitivas.

Há também uma crescente preocupação sobre a maneira como as aplicações voltadas para a Internet são desenvolvidas e sobre sua qualidade em longo prazo. A chamada *Web Engineering*[11] é uma disciplina oriunda da Engenharia de Software que defende uma abordagem processual e sistemática para o desenvolvimento, a implantação e a manutenção de alta qualidade de aplicações para a *web* através de princípios científicos, de gestão e de engenharia. Para uma visão mais aprofundada sobre os princípios e funções da Engenharia *Web*, bem como uma avaliação das semelhanças e diferenças entre o desenvolvimento de sistemas tradicionais e de sistemas baseados na *web*, veja MURUGESAN et al. (2001).

2.4.2. Computação em Nuvem

A Computação em Nuvem permite um intercâmbio de dados e aplicações entre os usuários e a infraestrutura computacional disponibilizada de modo que o

[10] 5W2H trata-se de uma ferramenta de gestão que funciona como um *checklist* de atividades que precisam ser desenvolvidas com o máximo de clareza possível por parte dos colaboradores de uma empresa. É assim chamada por causa das 7 perguntas efetuadas em inglês acerca daquilo que está em análise (What: o que será feito (etapas); Why: por que será feito (justificativa); Where: onde será feito (local); When: quando será feito (tempo); Who: por quem será feito (responsabilidade); How: como será feito (método) e How much: quanto custará fazer (custo)). Essa ferramenta pode ser utilizada no contexto do desenvolvimento de software.

[11] A expressão pode ser traduzida como “Engenharia *Web*”. Optou-se em utilizar a expressão inglesa por ser a mais difundida.

usuário possa acessá-los de maneira simples, rápida e de qualquer local. Um ambiente virtual para computação científica pode usufruir bastante deste modelo com cooperação intensa entre pesquisadores parceiros. Seu principal objetivo é proporcionar serviços computacionais ou de Tecnologia da Informação (TI) de acordo com o uso e prover serviços tanto a usuários finais interessados em usar aplicações disponibilizadas na nuvem, quanto a instituições públicas e privadas interessadas em disponibilizar tais aplicações ou terceirizar sua área de TI.

É possível descrever a Computação em Nuvem explorando alguns aspectos fundamentais como seus modelos de serviço ou suas abordagens de implantação (MELL e GRANCE, 2009). Em termos de modelos de serviço, os aspectos são: (i) Infraestrutura como Serviço ou IaaS[12], que fornece recursos computacionais como rede, processamento, armazenamento, etc., para que o usuário possa implantar e executar vários softwares; (ii) Plataforma como Serviço ou PaaS[13], que utiliza a infraestrutura de nuvem para criar e implantar novas aplicações, usando linguagens computacionais e ferramentas suportadas pelo provedor, que é a organização que fornece o serviço; e (iii) Software como Serviço ou SaaS[14], que provê aplicações à nuvem para serem usadas sob demanda, onde o usuário não controla a infraestrutura. Veja a Figura 2.2 para um esquema que relaciona esses modelos com os diferentes tipos de usuário.

Já em termos de abordagens de implantação, a nuvem pode ser: (i) Pública, acessível a todos os usuários; (ii) Privada, exclusiva de uma instituição; (iii) Comunitária, compartilhada entre instituições de acordo com interesses comuns; e (iv) Híbrida, qualquer tipo de combinação entre as categorias anteriores (MARCON JR et al., 2010).

Várias tarefas computacionais têm migrado dos computadores de mesa e dos servidores para a nuvem computacional (ERICKSON et al., 2009). Ambientes

[12] Do inglês, *Infraestructure as a Service*.
[13] Do inglês, *Platform as a Service*.
[14] Do inglês, *Software as a Service*.

de Computação em Nuvem têm se tornado mais comuns e afetado diretamente os desenvolvedores de software e de hardware, os usuários e a área de TI das organizações (HAYES, 2008).

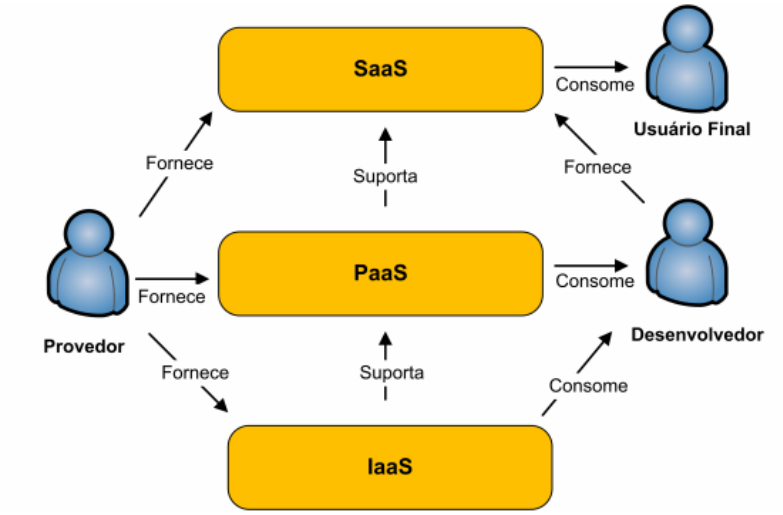


Figura 2.2 – Modelos de Computação em Nuvem

Fonte: SOUSA (2009).

Existem inúmeras vantagens ao migrar aplicações tradicionais para um modelo baseado em nuvem através de uma arquitetura baseada em serviços: economia em servidores, armazenamento, rede, licenças de software, energia, resfriamento e bens materiais; redução de trabalho na administração de sistemas; redução do tempo de configuração; diminuição de equipes de trabalho; desenvolvimento de aplicações com ciclo de vida mais curto e consequente redução do tempo de disponibilização de novos produtos e serviços no mercado; maior confiabilidade com custos menores; redução de custos com manutenção e com atualizações de hardware e infraestrutura (ZHANG et al., 2010).

Um ambiente colaborativo robusto e interinstitucional precisa oferecer acesso a dispositivos móveis, escalabilidade e disponibilidade. Os modelos de Computação em Nuvem oferecem vantagens para isso, permitindo que recursos adicionais sejam demandados mesmo com um aumento inesperado

do número de usuários. O uso desses modelos é percebido na ciência inclusive para computação de alto desempenho (OGRIZOVIC et al., 2010). Alguns pesquisadores têm avaliado o desempenho em clusters virtualizados.

2.4.3. Tipos de Arquiteturas

Uma arquitetura computacional define como os componentes de um sistema são combinados. A escolha da arquitetura para o sistema colaborativo precisa atender às necessidades desse ambiente. A seguir são apresentadas alguns tipos de arquiteturas.

Cliente-Servidor

Este é o principal padrão de interação entre aplicações cooperativas. É a arquitetura mais frequentemente encontrada em ambientes em rede. A interação cliente-servidor “forma a base da comunicação de rede e serve como alicerce para os serviços de aplicação” (COMER, 2006). Nesse contexto, um cliente é uma aplicação que envia para outra aplicação solicitação de dados ou computação e aguarda a resposta para continuar a execução demandada pelo usuário; já um servidor é a aplicação que fornece esses serviços, após receber a requisição e elaborar uma resposta.

Software em Camadas

A característica mais marcante dessa arquitetura é a hierarquia entre as camadas. Nesse esquema, a camada superior usa os serviços da camada inferior que responde às requisições da camada de cima. Num ambiente com muitas camadas, há o ocultamento das camadas internas diminuindo o acoplamento e aumentando a reusabilidade. Quanto maior o número de camadas, menor o desempenho devido ao *overhead* e ao processamento entre as camadas. No contexto de rede é combinado com a arquitetura cliente-servidor.

Software em Objetos Distribuídos

Ambientes com esta arquitetura permitem operações com objetos remotos. O objeto é uma abstração ou entidade que possui um conjunto de dados e métodos para manipulação desses dados. Tais métodos permitem a leitura ou a modificação do estado (conjunto de dados) do objeto. Com isso, tais métodos formam uma interface que permite que as especificações de acesso às operações do objeto sejam públicas ocultando a implementação de tais operações e das informações. Assim, se houver alterações internas, o objeto requisitante sequer é afetado se a interface não for modificada.

Arquitetura Orientada a Serviço

A Arquitetura Orientada a Serviço (SOA[15]) suporta baixo acoplamento, o que permite flexibilidade e interoperabilidade. Com os serviços em rede, pode integrar aplicações independentemente da tecnologia. É um conjunto de componentes que podem ser invocados e ter descrições de interfaces publicadas e descobertas.

O conceito de serviço é o de funções de negócios implementadas em software e acessíveis através das suas interfaces (PAPAZOGLOU, 2003). A função da interface é fornecer o mecanismo pelo qual os serviços se comunicam com outros serviços, e apresentar o conjunto de operações disponíveis para invocação dos clientes do serviço.

Além disso, essa arquitetura oferece dois princípios fundamentais para ambientes de grande porte: a modularidade (subdivisão de tarefas) e o encapsulamento (abstração). Dada a necessidade de integração entre diferentes aplicações remotas, um ambiente colaborativo distribuído com essa arquitetura pode ter os seguintes benefícios: componentes simplificados, facilidade de montar solução como uma nuvem de serviços, facilidade de integração de aplicações heterogêneas, flexibilidade e agilidade nas mudanças e proteção dos componentes devido a alterações na tecnologia.

[15] Do inglês, *Service-Oriented Architecture*.

2.4.4. *Web Services*

Os *Web Services*[16] são definidos pelo *World Wide Web Consortium* (W3C) como um software projetado para suportar a interação entre componentes de software em máquinas remotas. É uma das abordagens para uma Arquitetura Orientada a Serviço que mais tem sido aceita pelo mercado (WANG et al., 2005).

Em relação às abordagens tradicionais como as tecnologias de objetos distribuídos, o *Web Service* fornece mais baixo acoplamento, provendo um ambiente mais dinâmico e adaptável a mudanças. Além disso, através dos *Web Services*, os problemas relacionados a ambientes distribuídos têm sido resolvidos com o desenvolvimento de padrões e tecnologias abertas através de parcerias na indústria ou de consórcios como o W3C e a *Organization for the Advancement for Structured Information Standards* (OASIS).

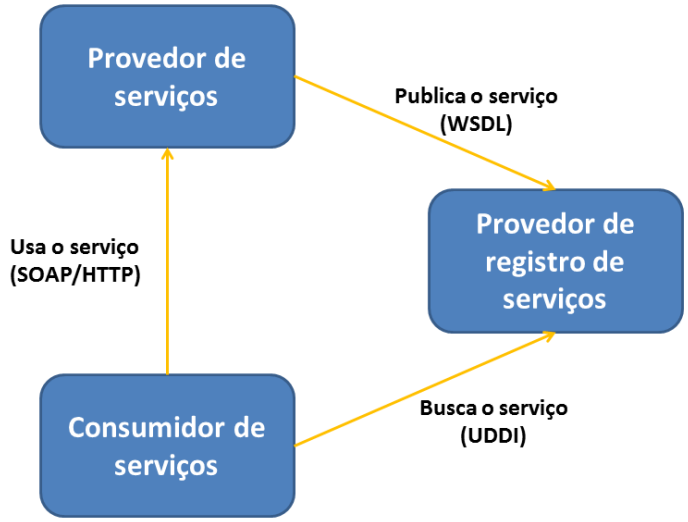


Figura 2.3 – Arquitetura dos *Web Services*

A interoperabilidade entre os *Web Services* é fornecida através de um conjunto de protocolos: (i) o *Simple Object Access Protocol* (SOAP), que fornece um mecanismo simples para troca de informações entre os serviços em formato de

[16] A expressão pode ser traduzida como “Serviços *Web*”. Optou-se em utilizar a expressão inglesa por ser a mais difundida.

um documento XML. Com as mensagens SOAP os serviços podem se comunicar usando diversos padrões de troca de mensagens, o que permite satisfazer a grande variedade de aplicações distribuídas; (ii) o *Web Service Description Language* (WSDL), que é um formato XML para descrever os serviços disponíveis, o formato de mensagens e o protocolo de rede usado para comunicação com o serviço. Esse protocolo permite então a publicação do serviço num provedor de registro de serviço para que possa ser descoberto; e o (iii) *Universal Description, Discovery and Integration* (UDDI), que é uma especificação para registro de *Web Services*. Essa especificação permite a consulta e atualização de repositórios de dados sobre *Web Services*, através de metadados e protocolos específicos.

Os protocolos supracitados permitem a comunicação entre os *Web Services* independente de sua localização e especificam uma descrição dos serviços e da infraestrutura onde o serviço está implantado. Para o transporte do *Web Service* pode ser usado o protocolo *Hypertext Transfer Protocol* (HTTP) ou *Hypertext Transfer Protocol Secure* (HTTPS), ou ainda protocolos proprietários. Na Figura 2.3 é possível visualizar um esquema da arquitetura dos *Web Services* relacionando esses protocolos.

A prática da análise avançada de ST's fomenta o desenvolvimento de muitas técnicas que são ou serão objetos de pesquisa. Como consequência, esses novos métodos terão de ser incorporados e disponibilizados rapidamente após testes e liberação para uso. Essa integração rápida só é econômica e tecnicamente possível através do uso de técnicas e conceitos computacionais modernos e de sistemas bem projetados através da Engenharia de Software e orientados por uma arquitetura que forneça extensibilidade, reusabilidade e portabilidade, como a SOA e os *Web Services*.

2.5. Revisão Bibliográfica

Utilizar ambientes virtuais para a colaboração científica atraiu um grande interesse na comunidade de ciência da computação, em meados da década de

1990. Alguns trabalhos acadêmicos discutiram o potencial de um *framework* para simulações e computação científica baseado na *Web* como: Kuljis e Paul (2001), Page et al. (2000), Bruzzone et al. (1999) e Fishwick et al. (1998). Porém, o seu desenvolvimento foi bastante limitado, devido à falta de aplicações reais, de acordo com Kuljus e Paul (2001).

A partir da década de 2000, o interesse no desenvolvimento de ambientes baseados na *web* para o desenvolvimento de pesquisa envolvendo computação científica foi revitalizado. Isso é evidenciado na literatura especialmente com as recentes aplicações de Computação em Grade, Virtualização, Serviços *Web* e Computação em Nuvem.

Afsarmanesh et al. (2001) aborda a criação desses ambientes, denominando-os de “laboratórios virtuais científicos”, como fundamentais para o desenvolvimento da ciência na atualidade, pois os cientistas precisam lidar com grandes quantidades de dados, com suporte para a colaboração distribuída e com questões de desempenho. Os autores consideram que a definição de laboratório virtual deve ser a de um ambiente generalizado, aberto e multidisciplinar favorecendo a colaboração e não limitado à solução de apenas um problema específico. Assim, discutiu os requisitos e uma arquitetura de referência para tais sistemas.

A necessidade da criação de sistemas baseados na *web* que virtualizem ações conjuntas na pesquisa científica compartilhando instrumentos científicos, dados experimentais, simulações numéricas e ferramentas de análise de dados, também está preconizado em Allen et al. (2002). Neste trabalho os autores previram a necessidade de portais *web* construídos a partir de serviços em aplicações distribuídas, sugerindo cenários em Astrofísica envolvendo caros instrumentos de pesquisa e grandes quantidades de dados necessárias para as simulações. Alguns exemplos de portais com diversos recursos são abordados no trabalho.

Foster et al. (2002) apresentam um sistema de dados virtuais chamado *Chimera* que combina um conjunto de dados gerados através de simulações, com um interpretador virtual de dados que traduz as solicitações do usuário através da definição dos dados no sistema e de consultas em bancos de dados. Os autores destacam a importância dos dados e sistemas virtuais, que podem aumentar significativamente a usabilidade de sistemas de gerenciamento de dados científicos, automatizando a pesquisa e favorecendo a colaboração a partir de dados pré-armazenados.

Hey e Trefethen (2003) apresentam o termo *e-Science* como o conjunto de serviços computacionais e de dados disponibilizados como infraestrutura para a pesquisa científica. Apresentam vários exemplos de demandas científicas da Bioinformática e da Física para sistemas computacionais acessíveis na *web* com alto poder de processamento e que permitam a colaboração e o compartilhamento de dados, juntamente com disponibilização de ferramentas de análise. Os mesmos autores apresentam uma descrição dos requisitos para uma infraestrutura mais rápida e melhor para diferentes necessidades de pesquisa em computação científica em Hey e Trefethen (2005), chamada de *Cyberinfraestrutura* ou *e-Infraestrutura*, que compõe um conjunto de programas computacionais e serviços. Esta infraestrutura permite a condução de pesquisas de maneira segura em ambientes controlados compartilhando recursos de forma distribuída. O trabalho conclui que há um vasto espaço para o desenvolvimento de aplicações virtuais para a computação científica, como os Observatórios e Laboratórios Virtuais e seus *e-experiments*.

Já em Roure e Hendler (2004), o termo *e-Science* foi tratado como uma junção de dois conceitos apontados como fundamentais para o futuro da Internet: a *Web* Semântica e as Grades Computacionais. São discutidas as peculiaridades de cada um dos conceitos, com o foco de uni-los de maneira complementar numa nova ideia, chamada “Grade Semântica”. Os autores destacam o programa britânico de *e-Science* que, desde 2001, tem o foco em desenvolver infraestrutura computacional e colaboração global em áreas-chave da ciência.

Engelen (2003) discute a importância da Computação em Grade para a solução colaborativa de problemas de grande magnitude em termos de dados e processamento. Segundo o autor, a relação entre Grade e *Web Services*[17] é recente e tem criado as chamadas Aplicações em Grade Orientadas a Serviço. O artigo estuda aspectos de usabilidade, interoperabilidade e desempenho dos *Web Services* em Grade para a computação científica.

O trabalho de Zhuge (2004) aborda a influência crescente da Internet na pesquisa científica. Muitos cientistas têm encontrado meios de compartilhar sua pesquisa e seus dados ou de trabalhar de forma colaborativa através da rede. Entretanto, apesar da quantidade de informação disponível atualmente na *web*, por ela não refletir semântica compreensível em nível de máquina, a *web* tem dificuldade em suportar serviços inteligentes. Para isso, o autor defende o desenvolvimento de "Grades de Conhecimento", que sintetiza conhecimento a partir de dados na rede através de métodos de mineração e permite que motores de busca façam referências e tirem conclusões a partir desses dados. É apresentada uma aplicação com essas características desenvolvida na Academia de Ciências da China, o IMAGINE (*Integrated Multidisciplinary Autonomous Global Innovation Networking Environment*).

Truong (2004) apresenta o desenvolvimento de um ambiente de simulação integrado escalável baseado na *web* chamado *Computational Science and Engineering On-line* (CSEO). Este ambiente permite o desenvolvimento e compartilhamento de pesquisas usando ferramentas computacionais atuais e consultando bases privadas ou públicas de dados. Neste artigo o CSEO fornece um ambiente integrado para modelagem multi-escala de sistemas complexos reativos. Um exemplo particular demonstra como os resultados obtidos em simulações de Química Quântica são usados para calcular propriedades termodinâmicas e cinéticas de uma reação química, que subsequentemente são utilizados na simulação de um reator de combustão. Já

[17] A expressão pode ser traduzida como "serviços *web*". Optou-se em utilizar a expressão inglesa por ser a mais difundida.

em Truong et al. (2006), o CSEO é apresentado de forma mais completa e atualizada, onde os autores enfatizam a inovação no fato de o software permitir o desenvolvimento de pesquisa individual ou colaborativa no mesmo ambiente.

Bernholdt et al. (2006) propõem uma arquitetura baseada em componentes para computação científica de alto desempenho. Baseado em componentes, é uma abordagem natural da Engenharia de Software para ambientes modernos de computação científica. A habilidade para a fácil reutilização e junção de componentes em várias aplicações traz benefícios significativos em termos de produtividade na criação de software de simulação. Os autores discutem ainda outras características importantes para um ambiente de computação científica, como: a interoperabilidade e disponibilização de ambientes de programação em linguagens computacionais diferentes, como C, C++, Fortran, Java, Matlab, Octave, Perl, Python, Tcl, etc; mínimo *overhead* entre os componentes; componentes paralelizados para incremento de desempenho; componentes distribuídos, abrangendo os conceitos de *Web Services*; e Computação em Grade.

Em Shneiderman (2007), o autor discute as diferenças entre a Ciência *Web* e a Ciência da Computação tradicional, como é conhecida atualmente. Para ele, o futuro da pesquisa científica está atrelado ao desenvolvimento da Ciência *Web*. O trabalho aborda a definição da nova disciplina denominada “ciência dos sistemas de informação descentralizados”, que inclui tecnologias emergentes, como *Web* Semântica, ontologias, *Web Services*, dentre outras.

Keahey (2007) et al. descrevem um estudo de caso no uso de máquinas virtuais para implantar ambientes customizados em sistemas remotos. Usam um ambiente de programação em C++ e Fortran preparado para desenvolver pesquisa científica em Física Nuclear. O maior ganho constatado na pesquisa foi o rápido crescimento do experimento com a escalabilidade oferecida. O autor discute a tímida adoção de soluções similares pela comunidade científica devido à carência de experiência com a tecnologia.

Keahey et al. (2008) apresentam um projeto para aplicações científicas colaborativas na *Web* chamado "*Science Clouds*", cujos principais objetivos são: facilitar o desenvolvimento de projetos científicos e educacionais em ambiente de Computação em Nuvem usando o *Elastic Computing Cloud* (EC2) da empresa Amazon; e compreender melhor o potencial e os desafios que a Computação em Nuvem representa para essas comunidades, além do que pode ser feito para superá-los.

Hoffa et al. (2008) exploram o conceito de Computação em Nuvem aplicado na pesquisa envolvendo computação científica, em detrimento da Computação em Grade: no artigo, a definição de Computação em Nuvem engloba Computação em Grade com Virtualização. Os autores comparam seus experimentos em quatro cenários distintos: (i) uma máquina local; (ii) múltiplas máquinas virtuais, cada uma como um computador independente; (iii) um *cluster* local e (iv) um *cluster* virtual. Os autores concluem que os ambientes virtuais oferecem maior escalabilidade e que soluções baseadas na Computação em Nuvem oferecem redução no *overhead* de escalonamento entre processos.

Para o desenvolvimento de ambientes virtuais de pesquisa, os *Web Services* são fundamentais. Paolini e Bhattacharjee (2008) enfatizam essa importância quando discutem o uso de tais componentes de software para prover infraestrutura de computação de dados termoquímicos de substâncias, além de permitir a integração de ferramentas entre diferentes fabricantes.

Hazelhurst (2008) estuda o uso de clusters virtuais como uma ferramenta para a computação científica colaborativa. O autor classifica o uso de sistemas que exijam alto desempenho em sistemas: (i) de grande porte, com alto custo e risco de se tornar ocioso; (ii) de pequeno porte, que não atendem às demandas reais de computação; e (iii) de compartilhamento de *clusters*, que esbarra na disponibilidade e no acesso ao equipamento. Assim, um modelo alternativo é avaliado usando o EC2 da Amazon, que permite a escalabilidade no uso dos recursos e na diminuição significativa do custo da pesquisa científica

envolvendo sistemas de alto desempenho. Reher et al. (2010) também destacam essa possibilidade de *clusters* virtuais, mas questionam a baixa performance oferecida devido à virtualização, à dependência da rede e às distâncias. No artigo é apresentada uma análise das capacidades computacionais e de rede do EC2 para experimentos científicos que calculam propriedades eletrônicas e óticas de sistemas complexos e que exigem computação de alto desempenho. Em termos de capacidade de processamento, os *clusters* virtuais oferecem resultados compatíveis com as necessidades gerais dos usuários. Porém, a rede que interliga as máquinas virtuais tem latência e características semelhantes a uma rede *gigabit*[18], o que é muito aquém dos grandes *clusters* de pesquisa acadêmica, segundo o próprio autor.

Em 2009, Vecchiola et al. também defendem que ambientes colaborativos de pesquisa em nuvem ofereçam maior Qualidade de Serviço (QoS[19]) em relação às soluções tradicionais como *clusters* e supercomputadores, que são mais difíceis de configurar, manter e operar e são mais caros. Para os autores, a Computação em Nuvem é um novo modelo para desenvolver pesquisa que necessite do compartilhamento dinâmico de: infraestrutura e recursos caros, grandes massas de dados e aplicações. Neste trabalho é apresentada uma solução corporativa de Computação em Nuvem, o Aneka, usada para a classificação de dados de expressão gênica e para determinar quais partes do cérebro reagem em resposta a um estímulo através da Ressonância Magnética Funcional (fMRI). Foram usadas as infraestruturas da Amazon como estudo de caso.

Finalmente, SRIRAMA et al. (2010) apresentam um projeto denominado *SciCloud* que estuda a criação de nuvens privadas em universidades para o uso em pesquisas científicas colaborativas que exijam computação de alto desempenho. O projeto usa o Eucalyptus como software de nuvem e

[18] Rede cuja taxa de transmissão seja de 1 *gigabits* por segundo.
[19] Do inglês, *Quality of Service*.

desenvolveu várias imagens otimizadas para aplicações em Bioinformática, Computação Distribuída e *Web Services* Móveis.

2.6. Ambientes para Plataforma *Desktop*

Conforme o levantamento mencionado no Capítulo 1 sobre ambientes virtuais colaborativos para a análise de dados disponibilizados na *Web*, são apresentados a seguir os sistemas eleitos.

2.6.1. Matlab

O Matlab[20] é um sistema que permite efetuar cálculo científico através de uma linguagem computacional de alto nível e de um ambiente de programação de fácil utilização. É um ambiente em constante evolução que permite uma maior produtividade em termos de desenvolvimento de programas, comparado com C/C++ ou Java.

O termo “Matlab” origina da conjugação dos termos *MATrix* e *LABoratory*. Isso se deve ao fato de a linguagem ser capaz de manipular matrizes sem a necessidade de prévio dimensionamento. Além disso, o ambiente integra análise numérica, processamento de ST’s e visualização através de gráficos 2-D ou 3-D. Tanto o problema, quanto a solução são expressos em linguagem matemática, ao contrário das linguagens mais tradicionais.

2.6.2. GNU Octave

O GNU Octave[21] é um sistema que permite a computação numérica oferecendo suporte à solução de problemas científicos. É uma linguagem interpretada de alto nível com muitos recursos para a solução de problemas lineares, não lineares, cálculo de integrais, polinômios, funções ordinárias e integração numérica de equações diferenciais. O ambiente também oferece recursos gráficos para a visualização e possui comandos muito similares com o Matlab, o que permite a portabilidade entre seus códigos.

[20] Website: www.mathworks.com.
[21] Website: www.octave.org.

Criado em 1988, o Octave começou como uma ferramenta especializada para auxiliar na solução de problemas envolvendo um Projeto de Reatores Químicos, na Universidade de Winsconsin-Madison e na Universidade do Texas, oferecendo mais flexibilidade que as linguagens tradicionais como Fortran e C bastante usadas em engenharia. Sua primeira versão final foi lançada em 1994 e atualmente tem sido usada em aplicações de pesquisa científica e aplicações comerciais.

2.6.3. Scilab

Assim como o Octave, o Scilab[22] é distribuído gratuitamente e é muito semelhante ao Matlab. É um ambiente e uma linguagem de alto nível com códigos interpretados para solução de problemas científicos através de computação numérica.

Foi desenvolvido em 1990 por pesquisadores do INRIA (*Institut National de Recherche en Informatique et en Automatique*) e do ENPC (*École Nationale des Ponts et Chaussées*). Atualmente é mantido pelo consórcio *Scilab Enterprises* e é bastante utilizado em aplicações comerciais e educacionais no mundo.

2.6.4. R

O R[23] também é um ambiente de programação integrado gratuito para cálculos científicos. Tem muitas funções de estatística e a linguagem foi criada por Ross Ihaka e Robert Gentleman na Universidade de Auckland, Nova Zelândia, a partir da linguagem S da Bell Laboratories. A linguagem é muito popular para a análise de dados e análise estatística, sendo considerada de alto nível como as linguagens supracitadas.

[22] Website: www.scilab.org.
[23] Website: www.r-project.org.

O ambiente tem versões para várias plataformas e a linguagem é extensível a partir do desenvolvimento de novas funções, sendo muitas delas desenvolvidas na própria linguagem. Além disso, a linguagem permite a interação com códigos escritos em outra linguagem (C, C++ e Fortran) ligados em tempo de execução.

Existem muitas interfaces gráficas para o R, como o RStudio, o JGR e o SciViews-R e há extensões para o IDE Eclipse. Além disso, códigos em C ou em Java podem manipular objetos R e os dados podem ser visualizados através de gráficos estáticos nativamente e de gráficos interativos através de pacotes adicionais.

2.7. Ambientes para Plataforma *Web*

2.7.1. NVO

O *National Virtual Observatory*[24] (NVO) é um projeto norte-americano de Observatório Virtual (VO[25]) que disponibiliza um ambiente com dados astronômicos de telescópios em terra e no espaço com enorme quantidade de dados disponíveis.

O ambiente permite a análise cruzada de vários tipos de observações do mesmo objeto. Possui muitas ferramentas para interação com o usuário especializado, uma vasta documentação com sistemas de ajuda e um livro (GRAHAM et al., 2008) que permitem um melhor entendimento do uso das ferramentas disponibilizadas, da estrutura e das tecnologias computacionais sobre as quais o sistema está operando.

[24] Website: www.us-vo.org.
[25] Do inglês, *Virtual Observatory*.

2.7.2. EURO-VO

O *European Virtual Observatory*[26] (EURO-VO) trata-se de um consórcio europeu de instituições que disponibilizam dados astronômicos, programas de análise e armazenamento das informações. Possui tutoriais, artigos relacionados com os temas e muitos programas e bibliotecas computacionais.

O EURO-VO está organizado em três grandes áreas: o *Facility Centre* (FC), que é uma estrutura de suporte central para facilitar a ampla aceitação de ferramentas para VO pela comunidade e fornecer uma interface com os usuários; o *Data Centre Alliance* (DCA), que provê um fórum para a aceitação de padrões do VO, o compartilhamento de melhores práticas para provedores de dados, a consolidação dos requisitos operacionais para as ferramentas e a identificação das exigências científicas de programas de interesse estratégico nacional que requerem tecnologias e serviços de VO; e o *Technology Centre* (TC), que consiste nos projetos de desenvolvimento e pesquisa tecnológica coordenados e conduzidos de forma distribuída através das organizações membros.

2.7.3. IVOA

O principal objetivo do *International Virtual Observatory Alliance*[27] (IVOA) é fornecer estrutura administrativa, softwares e ferramentas para exploração de dados de diversos parceiros de forma cooperativa e interoperável. Seus grupos de trabalho desenvolvem padrões de dados e programas computacionais para uso e manipulação dos dados.

2.7.4. Chimera

O Chimera (FOSTER et al., 2002) é um sistema virtual que combina um conjunto de dados gerados através de simulações, com um interpretador virtual para traduzir as solicitações do usuário através da definição dos dados no

[26] Website: www.euro-vo.org.
[27] Website: www.ivoa.net.

sistema e de consultas em bancos de dados. Ele está acoplado a uma Grade de Dados para executar de acordo com a demanda de computação a partir de consultas nos bancos de dados. O sistema tem tido sucesso em simulações voltadas à colisão de partículas na Física e busca de dados panorâmicos do céu para os aglomerados de galáxias (FOSTER et al., 2002). Na Figura 2.4 é apresentado um esquema de sua arquitetura.

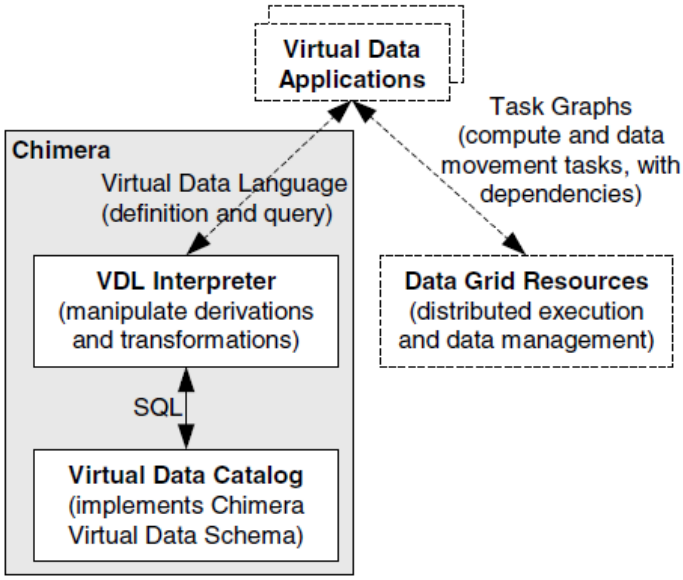


Figura 2.4 – Esquema da arquitetura do Chimera

Fonte: Foster et al. (2002).

2.7.5. GEODISE

O GEODISE[28] é uma colaboração entre as Universidades de Southampton, Oxford e Manchester, juntamente com as empresas BAe Systems, Rolls-Royce e Fluent. O projeto visa proporcionar acesso a ferramentas modernas de otimização e de busca, acesso a recursos de computação e de dados distribuídos e a um repositório de conhecimento inteligente. Seu website possui informações sobre a tecnologia empregada, área para *downloads* e acesso ao ambiente.

[28] Website: www.geodise.org.

2.7.6. GRIDPP

O GridPP[29] é uma colaboração de físicos de partículas e cientistas da computação do Reino Unido e do CERN para experimentos de computação científica, mais especificamente, física de partículas, numa rede de computação distribuída em todo o Reino Unido. O GridPP cria e gerencia uma infraestrutura computacional para os físicos de partículas do Reino Unido, usando software de código aberto, aplicações e *middleware*. No momento, o projeto envolve 17 instituições do Reino Unido e também contribui para muitos projetos relacionados.

2.7.7. GridChem

O GridChem[30] é um ambiente de *e-Science* que visa atender a uma necessidade crescente de recursos computacionais da comunidade de química computacional. O sistema visa oferecer uma interface fácil de usar e também aproveitar os sistemas de infraestrutura desenvolvidos pela *National Science Foundation* (NSF) dos Estados Unidos e outras agências para oferecer um ambiente virtual para a computação científica com ênfase na pesquisa de estruturas moleculares das substâncias.

2.7.8. LAF

O Laboratório de Agricultura e Floresta em Sensoriamento Remoto[31] (LAF) é um ambiente virtual que fornece ferramentas para a visualização de ST's derivadas de imagens de sensoriamento remoto e ferramentas para análise e filtragem dessas séries usando Transformada *Wavelet*[32] Discreta (FREITAS et al., 2011). O laboratório é integrado com o Google Maps[33] de modo que uma

[29] Website: www.gridpp.ac.uk.
[30] Website: www.gridchem.org.
[31] Website: www.dsr.inpe.br/laf/series/en/index.html.
[32] Uma vez que há controvérsias para o termo em português (ondeletas x ondaletas), convencionou-se utilizar o termo em inglês, *wavelet*.
[33] Website: maps.google.com.

mudança no uso do solo pode ser rapidamente detectada. O foco de estudo é a cobertura na América do Sul.

Tais ambientes apresentam possibilidades de colaboração científica em diferentes áreas do conhecimento. Isso se dá ao permitir acesso e manipulação de grandes bases de dados, e ao possuir também material que permita o aprendizado dos conceitos de modo geral. Diante destes ambientes descritos, apenas o último, o LAF, se propõe à investigação de ST’s considerando a aplicação de técnicas mais avançadas, recentes e robustas. Mesmo assim, este ambiente não atende todas as demandas apresentadas no capítulo introdutório.

3 TÉCNICAS PARA A ANÁLISE DE SÉRIES TEMPORAIS

Essa pesquisa focou sua abordagem na disponibilização de técnicas para a análise de ST's. Tal análise tem sido importante para diversas comunidades científicas na ampla área da ciência de dados. Algumas comunidades interessadas em análise avançada de ST's como as de Física Estatística e as de Dinâmica Não Linear, possuem várias aplicações em ciências espaciais e ambientais. Com isso, o objetivo deste capítulo é apresentar inicialmente a importância de técnicas avançadas para a análise de séries temporais, abordando os principais fatores determinantes na escolha das técnicas: o problema do tamanho e a necessidade de tratamento diferenciado na análise de séries com padrões de variabilidade não lineares e/ou não estacionários. Em seguida, são apresentadas duas técnicas convencionais para a análise e a técnica escolhida como canônica que serviu para validar o protótipo.

3.1. Fatores Determinantes na Escolha das Técnicas

3.1.1. Tamanho da Série

Dados reais podem ser coletados e organizados no formato de uma ST e, apesar dos avanços em sensores, sistemas de coleta, transmissão e armazenamento de dados, existem problemas intrínsecos ao tipo de observação ou experimento que acabam limitando seu tamanho. No caso em que dispomos de ST reais, ou seja, obtidas através de medições em campo, é razoavelmente comum nos depararmos com sinais que apresentem descontinuidades significativas (ver Figura 3.1) causadas principalmente por falhas diversas nos equipamentos de coleta de dados. Essas falhas resultam em perda de informação sobre o padrão de variabilidade ao longo de um período de interesse, gerando um conjunto de amostras curtas que, na maioria das vezes, são descartadas da análise devido ao restrito número de pontos, comprometendo sua interpretação estatística.

Na Figura 3.1 são apresentados segmentos de uma série de medidas diárias da temperatura da água de um reservatório de Furnas (INPE, 2008). Note que, devido a problemas instrumentais ou de transmissão de dados, o conjunto de medidas destacado em vermelho possui apenas 75 medidas de um conjunto total contendo 384. Mesmo numa resolução horária teríamos 1800 medidas na região de destaque de um total de 9216.

Conforme definição de série temporal curta no Capítulo 1, em geral, qualquer técnica que dependa de um número N mínimo de medidas sequenciais para que o resultado da análise possa ser considerado robusto requer um número significativo de pontos (SCHREIBER, 1998 e OSBORNE e PROVENZALE, 1989). Assim, o tamanho da ST em termos de pontos de medida é crucial para que técnicas estatísticas possam ser aplicadas sem comprometer a correta interpretação dos momentos estatísticos usuais (média, variância, assimetria e curtose). Mesmo as técnicas mais sofisticadas como algoritmos para cálculo de leis de escalas e leis de potência, em geral baseadas na Transformada de

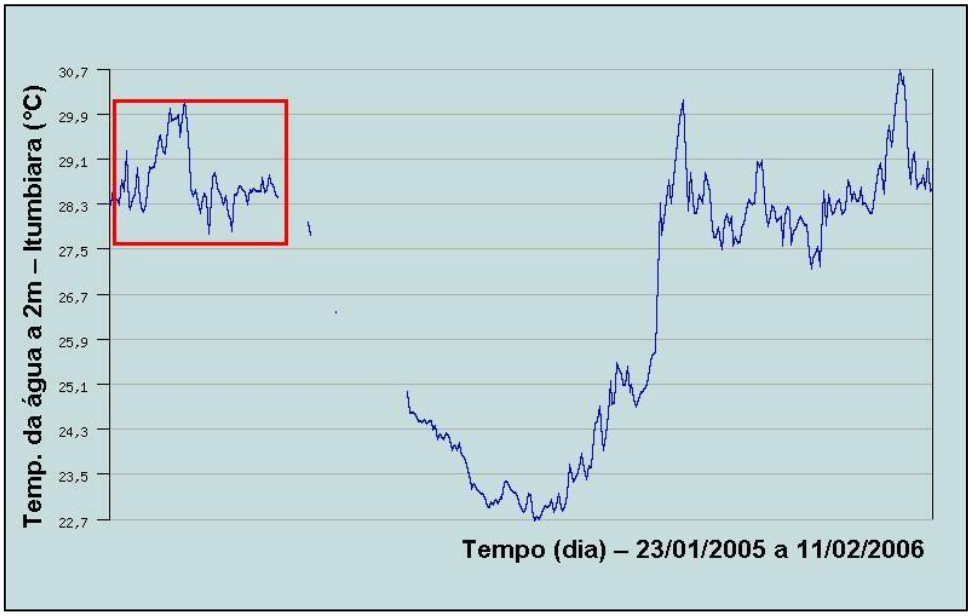


Figura 3.1 – Exemplo de série temporal com descontinuidades

Fonte: Adaptado de INPE (2008).

Fourier do sinal, ficam seriamente comprometidas quando a ST é curta. Alguns estudos sobre caracterização de processos dinâmicos a partir de ST, em geral, não mencionam essa limitação, pois os autores assumem que as ST's medidas ou simuladas são amostras com validade estatística. Nesses casos, STC's são completamente descartadas das análises mais sofisticadas onde se busca interpretar um padrão de variabilidade não linear a partir de possíveis mecanismos físicos subjacentes.

Normalmente, STC's não servem para a aplicação de diversas metodologias que buscam, por exemplo, caracterizar intermitência e turbulência (FRISH, 1995), fenômenos de auto-organização fora do equilíbrio (BAK *et al*., 1988), difusões anômalas (SWINNEY e TSALLIS, 2004), caoticidade (PEITGEN *et al*., 1992), regimes reativo-difusivos (CROSS e HOHENBERG, 1993), entre outros.

3.1.2. Não Linearidade e Não Estacionariedade

De um modo bastante geral, podemos dizer que uma ST é não linear quando responde de maneira diferente a estímulos grandes ou pequenos. É o paradigma da não linearidade: pequenas causas levam a grandes efeitos.

A princípio, um comportamento não linear pode ser considerado como algum fator de ruído externo adicionado ao sistema. Porém, de acordo com a Teoria do Caos, adições aleatórias não são as únicas fontes de irregularidade no sistema. Sistemas caóticos não lineares podem produzir dados muito irregulares com equações puramente determinísticas de movimento em um sistema autônomo, ou seja, sem dependência de adições de ruído no tempo (KANTZ e SCHREIBER, 2004).

Segundo Morettin (1999), os principais métodos de análise de ST estão baseados geralmente nos conceitos de estacionariedade da ST e linearidade do processo dinâmico. Em geral, as metodologias baseadas em modelos autorregressivos (AR), de médias móveis (MA) e mistos (ARMA) são

classificadas como “lineares”, pois são consideradas apropriadas apenas para processos lineares que geram ST aproximadamente estacionárias.

Uma ST é dita estacionária “quando ela se desenvolve no tempo aleatoriamente ao redor de uma média constante, refletindo alguma forma de flutuação estável” (MORETTIN e TOLOI, 2006). A forma mais fraca, mas não evidente de estacionariedade exige que todos os parâmetros que são relevantes para a dinâmica de um sistema têm que ser fixos e constantes durante o período de medição (KANTZ e SCHREIBER, 2004). Para a maioria das ferramentas matemáticas e computacionais de análise, as ST’s devem refletir este comportamento, já que estes procedimentos supõem que elas sejam estacionárias.

Sinais não estacionários são muito comuns particularmente em fenômenos naturais. A Figura 3.2 ilustra a diferença entre os padrões de variabilidade de ST’s. Os itens (a) e (b) correspondem à mesma série estacionária, porém com visualização da média e da faixa de variabilidade dinâmica em janelas de 256 e 128 pontos, respectivamente. De maneira equivalente, os itens (c) e (d) correspondem à ST não estacionária, com visualização também em 256 e 128 pontos, respectivamente.

É usual, na análise de ST’s não estacionárias por métodos convencionais, transformar as mesmas em ST’s estacionárias, tomando as diferenças sucessivas da série original até que a faixa dinâmica da flutuação torne-se estável em torno de uma média.

A primeira diferença de $A(t)$ é definida por

$$\Delta A(t) = A(t) - A(t-1), \tag{3.1}$$

e a segunda diferença é

$$\Delta^2 A(t) = A(t) - 2.A(t-1) + A(t-2)\,, \tag{3.2}$$

onde $A(t)$ é a ST de amplitudes.

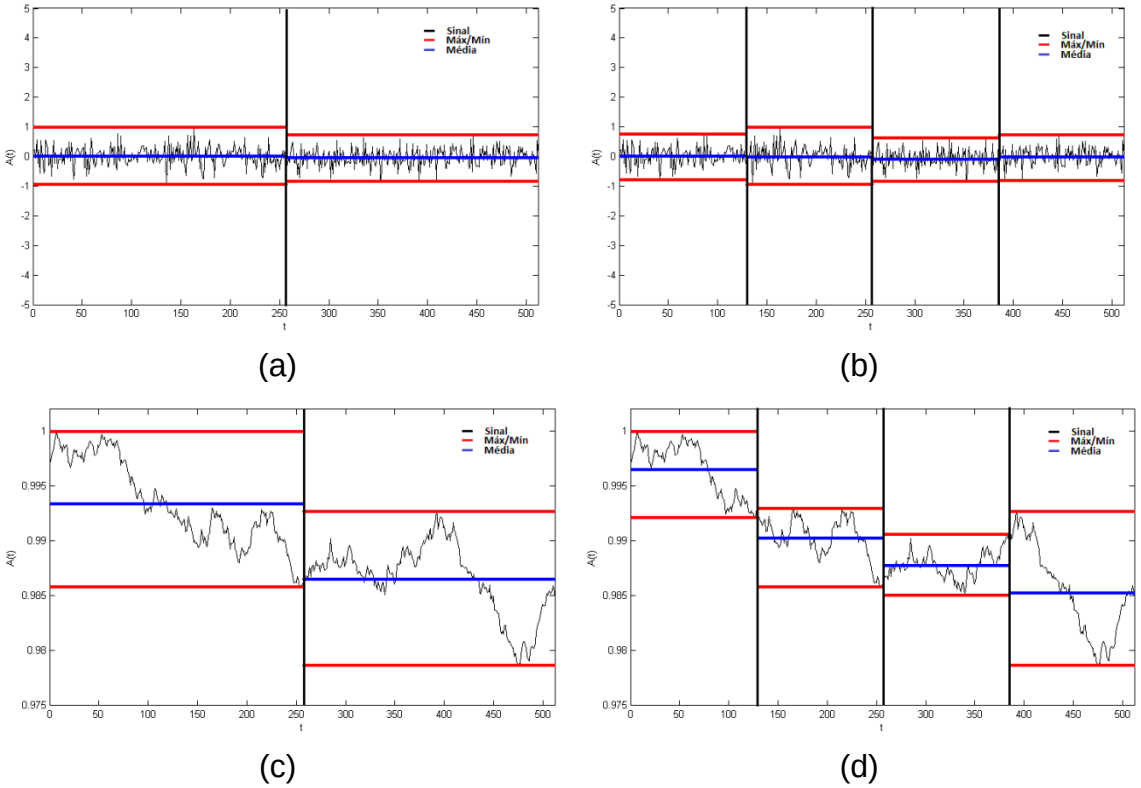


Figura 3.2 – Exemplo de série: (a) e (b) estacionária; e (c) e (d) não estacionária

Apesar da comodidade da transformação acima ou de outras transformações similares, NELSON (1976) concluiu que transformações neste sentido não melhoram a qualidade da análise, tanto para previsões de valores futuros, como para caracterizações sobre o processo físico subjacente. Ao invés disso, a transformação de estacionariedade de uma série introduz um erro nas previsões decorrente da tal transformação, desqualificando o método. Granger e Newbold (1976), estudando este problema, concluíram que “previsões em dados transformados estão contaminadas e deveriam, portanto, ser ajustadas, o que não é feito por grande parte dos programas computacionais”. No contexto da aplicação de modelos autorregressivos, Plosser (1979) conclui que

“parece ser preferível fazer a previsão usando diretamente o modelo sazonal ao invés de ajustar a série e depois utilizar um modelo não sazonal de análise”.

3.2. Métodos Convencionais de Análise

3.2.1. Momentos Estatísticos

Para qualquer ST X e qualquer inteiro positivo k, a esperança $E(X^k)$ é denominada k-ésimo momento de X, ou momento de ordem k. Suponha que X seja uma ST com $E(X) = \mu$. Para qualquer inteiro positivo k, a esperança $E[(X-\mu)^k]$ é denominada k-ésimo momento central de X ou k-ésimo momento em torno da média. Os momentos de uma ST podem ser gerados a partir de uma função geradora de momentos $M_x(t)$ disposta na Equação 3.3:

$$M_x(t) = E[e^{tx}], \tag{3.3}$$

considerando X uma ST e t números reais da função.

A partir da função acima é possível gerar todos os momentos. Seja $M_x^k(t)$ a k-ésima derivada de $M_x(t)$. Então:

$$\begin{aligned} M_x^1(0) &= \left[\frac{d}{dt}E(e^{tx})\right]_{t=0} \\ &= E\left[\left(\frac{d}{dt}e^{tx}\right)_{t=0}\right] \\ &= E[(Xe^{tx})_{t=0}] \\ &= E(X). \end{aligned} \tag{3.4}$$

Analogamente, temos:

$$
\begin{aligned}
M_x^k(0) &= \left[\frac{d^k}{dt^k} E(e^{tx}) \right]_{t=0} \\
&= E\left[\left(\frac{d^k}{dt^k} e^{tx} \right)_{t=0} \right] \\
&= E[(X^k e^{tx})_{t=0}] \\
&= E(X^k).
\end{aligned}
\tag{3.5}
$$

Portanto, $M_x^1(0) = E(X), M_x^2(0) = E(X^2), M_x^3(0) = E(X^3)$ e assim por diante.

Estes momentos são nomeados pela ordem, podendo ser de 1ª, 2ª, 3ª, 4ª, etc., sendo que os de ordem superior à 4ª ordem são pouco usados. O momento de 1ª ordem mede o valor médio dos dados e é definido, assim como os demais, para uma variável discreta, como:

$$E\{X\} = \frac{1}{N} \sum_{n=1}^{N} x_n \ . \tag{3.6}$$

O momento de 2ª ordem mede a dispersão dos dados em torno do 1º momento e é definido como:

$$E\{(X - E\{X\})^2\} = \frac{1}{N} \sum_{n=1}^{N} (x_n - E\{X\})^2 \ . \tag{3.7}$$

O momento de 3ª ordem mede a simetria dos dados em torno do 1º momento e é definido como:

$$E\{(X - E\{X\})^3\} = \frac{1}{N} \sum_{n=1}^{N} (x_n - E\{X\})^3 \ . \tag{3.8}$$

Já o momento de 4ª ordem mede a acuidade ou a concentração dos dados em torno do 1º momento e é definido como:

$$E\{(X - E\{X\})^4\} = \frac{1}{N}\sum_{n=1}^{N}(x_n - E\{X\})^4 \ . \qquad (3.9)$$

3.2.2. Densidade Espectral de Potência

A Densidade Espectral de Potências (PSD[34]) de uma ST pode ser considerada a metodologia convencional mais utilizada para classificar um processo estocástico de acordo com a lei de potência obtida a partir do seu espectro. A PSD $P(\omega)$ é definida como o quadrado do módulo de $A(\omega)$ (série no domínio da frequência) e indica a influência da frequência ω na ST. Ela é obtida a partir da Transformada de *Fourier* (TF) da função de autocorrelação, que descreve como um sinal está evoluindo no tempo e informa o quanto o valor de um ponto da ST é capaz de influenciar seus vizinhos. Supondo-se uma ST X_t discreta com média μ, sua autocorrelação $R(k)$ é definida como:

$$R(k) = \frac{E[(X_t - \mu)(X_{t+k} - \mu)]}{\sigma^2}, \qquad (3.10)$$

onde k é o deslocamento no tempo e σ^2 a variância da ST. O valor da autocorrelação varia entre 1 (correlação perfeita) e -1 (anticorrelação perfeita) e pode ser 0 quando não há correlação.

De acordo com a Teoria da Análise Harmônica, podemos considerar genericamente que uma ST qualquer é um somatório de sinais periódicos com diferentes escalas. Fazer a análise espectral desta ST é verificar a influência de cada escala presente na mesma em relação às demais. Se a ST é periódica, o seu espectro pode ser representado como uma combinação linear (denominada como Série de Fourier) de oscilações cujas frequências são múltiplos inteiros da frequência básica, ω. Quando a ST é não periódica, o espectro de frequências varia continuamente e, portanto, para representar a ST

[34] Do inglês, *Power Spectral Density*.

em termos dessas frequências, usa-se a Transformada de Fourier (PAPOULIS, 1962).

Para efeitos de definição, a Transformada de Fourier na forma contínua pode ser escrita como:

$$A(\omega) = \int_{-\infty}^{\infty} e^{-i\omega t} A(t) dt \,, \tag{3.11}$$

onde $A(t)$ é a ST na forma:

$$A(n) = A(t_n), t_n = n\Delta t \,, \tag{3.12}$$

sendo que as medidas são realizadas em intervalos de tempo regulares, Δt.

No caso discreto, a chamada Transformada Discreta de Fourier de uma ST pode ser definida pela série abaixo:

$$A_k = \frac{1}{\sqrt{N}} \sum x(n) e^{i\frac{2\pi nk}{N}}, k = 1, 2, ..., N \,, \tag{3.13}$$

onde $x(n)$ é a série discreta a ser transformada, N é o período da série e $e^{i\frac{2\pi nk}{N}}$ é o kernel da transformada.

A partir da Equação 3.11, a PSD é definida como:

$$P(\omega) = \mid A(\omega) \mid^2 . \tag{3.14}$$

Já para o caso discreto, a PSD é definida por

$$P(\omega) = \mid A_k \mid^2 , \tag{3.15}$$

sendo que a ST da Equação 3.12 está definida no domínio do tempo e a série da Equação 3.13, no domínio das frequências. Do ponto de vista

computacional, o cálculo desta transformada pode ser obtido com maior eficiência para sinais longos, usando o algoritmo da Transformada Rápida de Fourier (FFT) (COOLEY e TUKEY, 1965).

A PSD em função de uma determinada frequência está relacionada com esta frequência através de uma lei de potência:

$$P(\omega_k) \sim \omega_k^{-\beta}. \tag{3.16}$$

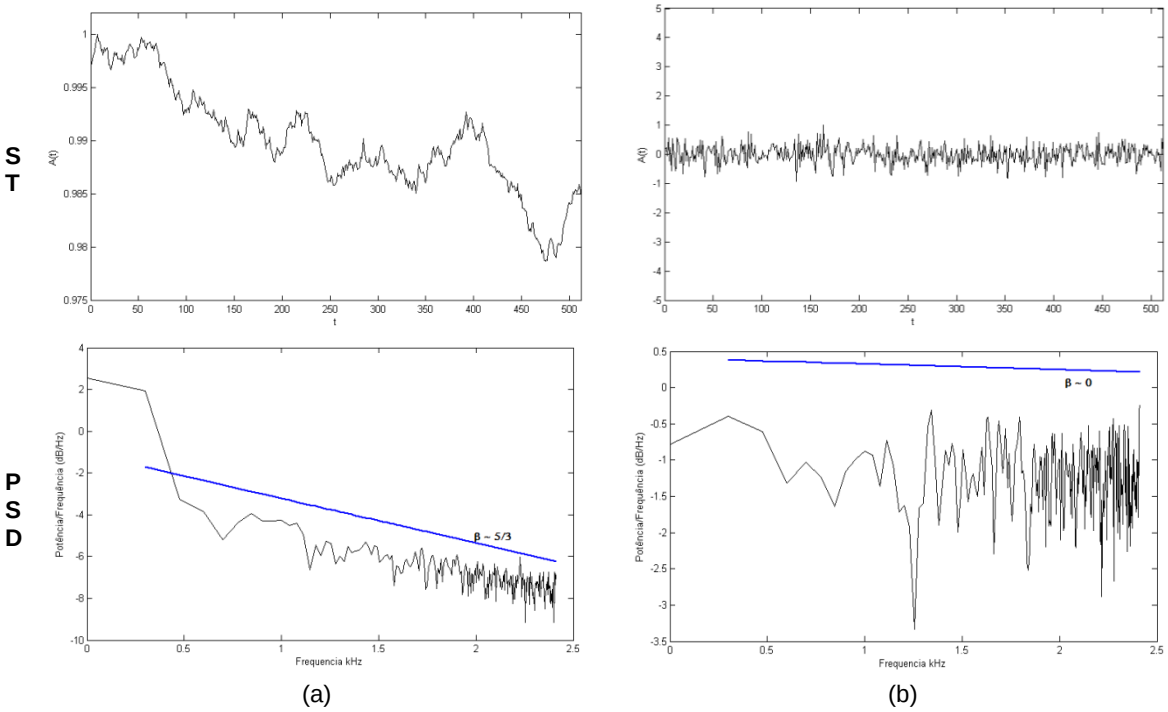


Figura 3.3 – Obtenção do Espectro de Potências: (a) ST estocástica com $\beta \sim \frac{5}{3}$ e (b) ST pseudoaleatória $\beta \sim 0$

Através de diversas medidas experimentais, tem sido constatado que a potência β está diretamente ligada às leis universais de escala, indicando que cada processo possui sua lei de potência ou "classe de universalidade"

definida[35] (FRISH, 1995; SALMON, 1982; GARRET e MUNK, 1979; ORSZAG, 1977; LESLIE, 1973 e PANCHEV, 1971).

Assim, cada valor de β determina um tipo de ruído randômico e cada tipo de ruído randômico corresponde a um tipo de processo estocástico. Ou seja, diferentes séries que possuem β aproximados dentro de uma margem de erro, indicam pertencer ao mesmo processo estocástico, fazendo de β um classificador natural de padrões de variabilidade temporal de ST estocásticas. A Figura 3.3 mostra a obtenção da PSD de ST's estudadas nesta tese: (a) estocástica com $\beta \sim \frac{5}{3}$ e (b) pseudoaleatória com $\beta \sim 0$.

3.3. Análise de Flutuação Destendenciada

A Análise de Flutuação Destendenciada (DFA[36]) é apresentada como uma metodologia alternativa para análise de padrões de variabilidade tanto de ST, quanto de STC introduzindo um quantificador robusto para classificação desses padrões, com diferentes graus de autocorrelação, autoafinidade, intermitência e persistência.

3.3.1. Algoritmo da DFA

O objetivo da DFA é determinar a autoafinidade de um sinal, através da computação do chamado *expoente de escala*. Este expoente foi introduzido por PENG *et al.* (1994) e pode ser utilizado para a caracterização de padrões temporais oriundos de um processo estocástico com efeito de memória ou correlação de longo alcance. A DFA tem sido utilizada em análise de ST não estacionárias a partir de dados biológicos, ambientais e de física espacial.

[35] Por exemplo: a classe de universalidade para a turbulência plenamente desenvolvida é caracterizada pela lei $\omega^{-5/3}$ em um espectro de energia contra o número de onda (FRISH, 1995).
[36] Do inglês, *Detrended Fluctuations Analysis*.

Para a obtenção do expoente de escala através da relação entre a flutuação média do sinal e o tamanho da janela do sinal a ela correspondente, a DFA é composta de algumas operações computacionais sobre a ST:

1. A primeira operação consiste na remoção do valor médio do sinal, subtraindo-o da ST. Em seguida, é feita uma integração do sinal. Essas duas operações estão representadas pela Equação 3.17.

$$C(k) = \sum_{i=1}^{k} (A(i) - < A >), \; k = 1, 2, 3, ..., \; N \tag{3.17}$$

   onde $A(i)$ corresponde às amplitudes do sinal e $< A >$ é seu valor médio.

   Quando a diferença entre sinal e a média é calculada, obtêm-se valores relativos à média, ou seja, maiores ou menores do que ela. Assim, a integração do sinal sem o seu valor médio resulta na obtenção de um sinal de tendências.

2. Na próxima operação, o sinal de N pontos é dividido em janelas iguais não sobrepostas de tamanho n. Pode acontecer de N não ser múltiplo inteiro de n. Então, o sinal será janelado novamente a partir de N, obtendo assim, $2N_n$ subintervalos.

3. Na janela é calculada uma reta de regressão, por meio do método dos mínimos quadrados, a qual representa a tendência do sinal naquele intervalo de tempo:

$$p_j^m(k) = b_{j_0} + b_{j_1} k + ... + b_{j_{m-1}} k^{m-1} + b_{j_m} k^m, \; m = 1, 2, ... \tag{3.18}$$

   onde m é interpretado como a *ordem de destendenciamento*, denotado por DFA^m, b é o subintervalo e k o número de pontos.

4. O próximo passo é retirar a tendência calculando a série de desvio acumulado em cada subintervalo.

$$C_j(k) = C(k) - p_j^m(k),\ m = 1, 2, \ldots \tag{3.19}$$

e então a variância dos subintervalos $2N_n$ é calculada:

$$F^2(j,n) = < C_j^2(i) > = \frac{1}{n}\sum_{i=1}^{n}[C((j-1)n+i) - p_j^m(i)]^2 \text{ , para } j = 1, 2, \ldots, N_n \text{, e} \tag{3.20}$$

$$F^2(j,n) = < C_j^2(i) > = \frac{1}{n}\sum_{i=1}^{n}[C(N-(j-N_n)n+i) - p_j^m(i)]^2 \text{ , para } j = N_n+1, N_n+2, \ldots, 2N_n \tag{3.21}$$

5. A próxima operação fornece a função de flutuação da DFA. É calculada a média das variâncias e sua raiz quadrada e, por isso, é também chamada de RMS (do inglês, *Root Mean Square*, ou Raiz da Média Quadrática). A flutuação é dada pela equação:

$$F(n) = \sqrt{\frac{1}{2N_n}\sum_{j=1}^{2N_n} F^2(j,n)}\ , \tag{3.22}$$

6. Essa flutuação é calculada novamente recursivamente para diferentes tamanhos de janela, para que possa ser verificada a sua característica fractal[37], e para que uma relação entre a amplitude das flutuações destendenciadas e o tamanho das janelas possa ser obtida.

Em geral, na presença de flutuações na forma de lei de potência $F(n) = Kn^{\alpha}$, o valor de $F(n)$ aumenta linearmente com o aumento do tamanho da janela. Assim, usando a regressão linear por meio do método dos mínimos quadrados

[37] A característica marcante nos fractais é a propriedade de autossimilaridade, que permite que suas formas geométricas sejam subdivididas de tal forma que os segmentos resultantes dessa divisão se assemelhem a cópias reduzidas da forma original.

de $F(n)$, $\log F(n) = \log K + \alpha \log n$. No gráfico log-log, obtém-se a inclinação α que é o expoente de escala.

3.3.2. O que a DFA mede

Quando a DFA foi desenvolvida por Peng e seus colaboradores (PENG et al., 1994), eles estavam interessados em distinguir entre as complexas flutuações intrínsecas ao sistema nervoso no comando das ações vitais do corpo humano, daquelas advindas do meio e que também exercem influência sobre a frequência cardíaca.

A pesquisa mostrou uma característica de fractais nas flutuações do sinal que são intrínsecas ao sistema, pois ocorrem ao longo de todo o sinal, independentemente da escala observada. Já as flutuações extrínsecas, ou do meio, caracterizou-se por apresentar efeitos locais e de curto prazo, implicando no surgimento de diferentes tendências ao longo do sinal. Através da DFA, essas tendências são subtraídas da ST para que seja realizada a sua análise.

Algumas doenças apresentam mudanças na característica da correlação, quando comparadas à invariância à escala observada nos sinais de coração de indivíduos normais. A DFA revelou-se, portanto, uma análise importante na caracterização de eventos locais que não apresentam correlação de longo prazo num sinal.

Assim, o objetivo principal da DFA é quantificar correlações de longo alcance num sinal, assim como o coeficiente de Hurst, porém de maneira mais generalizada, através do expoente de escala.

O expoente de escala pode ser calculado para sinais com estatísticas ou dinâmicas subjacentes não lineares ou variantes com o tempo. O valor de α aproxima-se de 0,5 para grandes valores de janela, quando o sinal apresenta apenas correlações de curto-prazo e esse valor de α é o esperado para um

sinal completamente descorrelacionado, ou seja, com características de ruído branco.

Se o valor de α estiver entre 0,5 e 1, o sinal apresenta correlação de lei de potência de longo alcance, o que significa que um grande intervalo (ou evento) tende a ser seguido por outro grande intervalo e não por um intervalo pequeno; essa correlação é também chamada de *correlação persistente*. Já quando α varia entre 0 e 0,5, observa-se que os intervalos tendem a se alternar entre grandes e pequenos, o que também caracteriza uma correlação de lei de potência, porém diferente da anterior, também chamada de *correlação antipersistente* ou de *anticorrelação*. Os casos em que $\alpha = 1$ são denominados de ruído rosa, ou ruído $\frac{1}{f}$ caracterizado por apresentar uma relação de amplitude que é o inverso da frequência em sua PSD. Por isso denota uma relação de lei de potência no sinal. Já as ST's para as quais são obtidos valores de $\alpha \geq 1$, apresentam características de correlação, porém uma correlação não exponencial. ST's com valor de $\alpha = \frac{3}{2}$ indicam que o sinal comporta-se como ruído Browniano. Além disso, o expoente α serve como representante da suavidade da ST: para valores maiores de α, mais suave é o sinal.

3.3.3. DFA em Ambientes Virtuais

De acordo com o que foi exposto na subseção anterior, a DFA congrega algumas características muito significativas para sua escolha como ferramenta canônica para testes num ambiente virtual: (i) é uma generalização de uma ferramenta muito utilizada em análise de ST's, o coeficiente de Hurst, sendo que a principal diferença entre eles é que o expoente de escala obtido pela DFA realiza também a análise de sinais não estacionários e não lineares; (ii) conforme está demonstrado no Capítulo 7, apresenta maior estabilidade na análise de STC's, caracterizando uma correlação com precisão maior do que outras ferramentas convencionais em tais séries; (iii) possui unicidade no

parâmetro de saída, o que permite classificar a série; (iv) possui um algoritmo de fácil compreensão, implementação e adaptação para diferentes linguagens computacionais, o que facilita a certificação da ferramenta; (v) pode ser utilizado de forma nativa ao ambiente, ou executado como um serviço num ambiente colaborativo; e (vi) é uma ferramenta muito utilizada. Só o artigo do Peng possui mais de 1500 citações de acordo com o Google Acadêmico[38] e a DFA possui ao menos 4800 trabalhos relacionados no mesmo site usando a expressão inglesa. Portanto, a DFA é uma ferramenta de análise avançada de ST's estratégica para o uso em um ambiente como o proposto nesta tese.

[38] Website: scholar.google.com.br.

4 REQUISITOS PARA O AMBIENTE

Sistemas colaborativos estão cada vez mais presentes no mundo e são caracterizados fortemente pela heterogeneidade. Tais sistemas exigem aplicações capazes de lidar com a diversidade crescente de hardware e software disponibilizados. Além disso, um ambiente colaborativo requer equilíbrio com métodos de produção mais eficientes para atender à crescente demanda por software ou serviços e à necessidade de redução do tempo de operacionalização dos mesmos no sistema. Este capítulo apresenta as demandas para um ambiente virtual colaborativo de computação científica para análise avançada de séries temporais. Em seguida, relaciona os tipos de usuários envolvidos com suas devidas funções no sistema.

4.1. Demandas do Ambiente

Para a especificação de um ambiente virtual colaborativo de computação científica para análise avançada de séries temporais é preciso considerar algumas características desejáveis conforme citado no Capítulo 1. Tais características podem ser agrupadas em demandas de: infraestrutura, metodologia e linguagens, colaboração e análise de dados.

4.1.1. Demandas de Infraestrutura

As demandas de infraestrutura referem-se aos requisitos computacionais e de conectividade do ambiente. Tais aspectos servem de sustentação para a aplicação como um todo permitindo seu funcionamento colaborativo.

Em termos computacionais, o ambiente deve estar estruturado a partir de um *hardware* que permita abrigar o sistema gerenciador de banco de dados e atender suas consultas aos dados com eficiência. Além disso, o conjunto de *hardware* deve permitir a execução apropriada de um servidor de aplicações sobre o qual o ambiente executará suas regras de negócio. O equipamento

também deve suportar *softwares* responsáveis pela publicação de páginas *web* que servirão de interface para o ambiente.

Em relação à segurança, o ambiente deve fornecer controle de acesso às suas diferentes áreas, garantindo a privacidade de cada usuário. Para isso, cada usuário deve ser responsabilizado pelo uso de sua senha de acesso ao ambiente e a aplicação deve reconhecer a distinção entre os usuários e seus perfis de acesso.

Todos os experimentos efetuados com as séries temporais devem ser armazenados para possível acompanhamento posterior. Assim, o ambiente não só deve suportar a inclusão de novas ferramentas de análise, juntamente com toda a sua documentação, mas também armazenar o histórico das análises efetuadas pelo usuário, inclusive mantendo as séries carregadas. Isso se reflete no requisito de controle dos dados e dos experimentos por parte do usuário.

Outro requisito para o ambiente objeto de estudo desta tese é a disponibilização de seus componentes como serviços acessíveis através de uma Arquitetura Orientada a Serviço. Por meio desta arquitetura torna-se possível disponibilizar acesso às ferramentas de análise, através de um protocolo definido, além de permitir uma evolução de plataforma aberta para o ambiente.

Finalmente, o ambiente deve ser multiplataforma operando possivelmente em diferentes tipos de hardware através de Sistemas Operacionais baseados no padrão UNIX (Linux, Solaris, Mac OS X, etc.) ou no padrão Microsoft (Windows XP, 7, 8, etc.). Deve ser compatível também com protocolos de rede bem difundidos no mercado como a suíte TCP/IP para que esteja disponibilizado na *web*, acessível a partir de diferentes dispositivos, sejam computadores pessoais, *notebooks*, *tablets*, *smartphones*, etc., atendendo também ao requisito da mobilidade.

4.1.2. Demandas de Metodologia e de Linguagens

As demandas de metodologia e de linguagens configuram os requisitos técnicos relacionados com o desenvolvimento e acoplamento de novas ferramentas de análise de dados ao sistema. Tais requisitos referem-se aos métodos e às técnicas ligados a este desenvolvimento.

Uma característica requerida em termos de metodologia é o uso de técnicas adequadas de engenharia de software em sua elaboração. O uso de tais técnicas diminui a incidência de erros e de reformulações não previstas. Além disso, a adoção de procedimentos que garantam a qualidade do ambiente é fundamental.

Como o ambiente pressupõe uma integração de ferramentas de análise de dados oriundas de diferentes fontes colaborativas, cada uma destas ferramentas pode ser desenvolvida usando uma metodologia peculiar ao pesquisador programador da técnica. A disposição da ferramenta no ambiente dependerá do processo de validação descrito na próxima subseção.

Entretanto, é requerido também que o ambiente forneça mecanismos de padronização. O processo de validação de novas ferramentas deve prever uma análise criteriosa do processo e das técnicas de desenvolvimento da nova ferramenta a ser inserida.

Finalmente, é preciso considerar o fator multilinguagem no ambiente proposto através desta pesquisa. Esta demanda está relacionada com a escalabilidade citada acima e com a distribuição abordada a seguir. É muito importante que um ambiente colaborativo desta natureza não limite sua expansão ao uso restrito de uma determinada linguagem, pois diferentes pesquisadores de técnicas de análise de dados possuem suas preferências neste aspecto, além de formação computacional variada.

4.1.3. Demandas de Colaboração

Em termos de colaboração, o ambiente proposto não só permite a interação entre grupos formais de pesquisa. O ambiente é também caracterizado como colaborativo devido à disponibilização das ferramentas de análise para os diversos usuários. Todos devem ter acesso à ferramenta, ao seu código fonte e à literatura científica que a fundamenta.

Assim, para que o ambiente seja colaborativo é necessário que seja multiusuário. Ou seja, o ambiente deve permitir, controlar e coordenar a ação simultânea de vários usuários.

É requerida também a integração do ambiente com outros pacotes de software de análise de dados, como Matlab, Octave, programas em C, C++, etc. Ou seja, o ambiente deve usar padrões abertos através da Internet para operar em cooperação com outros sistemas, sejam locais ou remotos. Essa característica permite o ambiente ter várias ferramentas remotas acopladas sendo utilizadas de maneira transparente pelo usuário a partir de diferentes localidades, caracterizando o requisito da distribuição.

Um requisito igualmente importante é o do baixo custo financeiro. Tal requisito pode ser atendido com o uso de padrões abertos de comunicação para o desenvolvimento do ambiente, evitando a necessidade de licenças de software. A adoção de tais padrões atende também o requisito da escalabilidade, permitindo a inclusão de novas técnicas e métodos pelos usuários do ambiente.

Para atender as demandas de colaboração é preciso também considerar o requisito do processo de validação das técnicas de análise na inserção de novas ferramentas. Esse requisito é fundamental para o ambiente ser escalável e distribuído, além de permitir a execução de ferramentas de análise em várias linguagens computacionais mantendo sua coesão e sua coerência. Sem a

devida validação computacional e científica, o ambiente pode abrigar ferramentas de análise sem o devido rigor necessário.

A documentação relacionada às ferramentas desenvolvidas e acopladas ao ambiente é muito importante para a colaboração. As publicações científicas, as séries temporais de exemplo, gráficos relacionados às análises e aos dados, algoritmos, documentação técnicas, etc. são exemplos de documentos que devem estar disponibilizados desde o processo de validação da ferramenta no ambiente.

4.1.4. Demandas de Análise de Dados

As demandas de análise de dados caracterizam os requisitos que são identificados como atividades de maior interação do usuário com o ambiente. Eles estão relacionados com o acesso às informações das análises, com a inclusão de novas ferramentas de análise e sua documentação, com a execução das ferramentas, além da disponibilização dos dados e resultados.

Para isso, um dos requisitos a serem atendidos pelo ambiente é o de uma interface simples, amigável e intuitiva. Através desta interface, o usuário poderá interagir com o ambiente de acordo com o seu perfil. O perfil do usuário definirá suas permissões de acesso ao ambiente, determinando a customização do ambiente para o usuário, que deve funcionar como um portal *web* permitindo acesso às ferramentas, às informações relacionadas, aos dados e às informações do ambiente, como sua evolução, eventos relacionados, disponibilização de novos recursos, etc.

O ambiente deverá permitir a definição de ajustes no ambiente de acordo com as preferências do usuário. É preciso que estas preferências alimentem um subsistema de conhecimento especialista baseado em regras, atendendo ao requisito da usabilidade do ambiente.

Um aspecto fundamental para o ambiente em questão é a disponibilização de acesso remoto através de uma interface universal a diferentes tipos de dispositivos. Este requisito é denominado de cliente leve, pois permite o uso do ambiente sem programas pré-instalados, modificações no dispositivo ou robustez do hardware local. Tal requisito é fundamental para a popularização e ampla adoção do ambiente.

Além disso, alguns experimentos podem ter tarefas que exijam mais tempo de execução. Assim, o ambiente precisa atender ao requisito de duração controlando adequadamente as tarefas relacionadas a tais experimentos. Seu sistema operacional e servidor de aplicações precisam lidar com esta necessidade, além do escalonamento de tarefas de acordo com a disponibilidade de recursos computacionais.

Para executar análises de séries temporais, o ambiente precisa ser capaz de receber através de um processo de *upload*, armazenar e avaliar grandes quantidades de dados a partir de múltiplos acessos simultâneos. Este requisito está relacionado ao requisito das ferramentas disponibilizadas. O ambiente deve oferecer acesso a técnicas de análise básica e avançada de séries temporais e de séries temporais curtas de maneira automatizada, bem como permitir a reutilização de suas partes. Finalmente, precisa também permitir que o usuário tenha um controle total dos experimentos.

4.2. Tipos de Usuários Envolvidos

Para que o ambiente virtual proposto seja funcional, é necessária a definição de perfis de usuários conforme suas diversas atribuições. A Figura 4.1 apresenta o diagrama de contexto do ambiente virtual ilustrando a interação entre o mesmo, seus tipos de usuários e o ambiente externo.

O ambiente virtual deve ser composto por diversos *sites* remotos. Estas instalações em instituições parceiras, quando interligadas, servirão como uma nuvem privada para fornecer alta disponibilidade do sistema. Em cada *site*

deve ser disponibilizada a infraestrutura mínima para que o ambiente possa operar de maneira funcional. O ambiente de maneira geral precisa ter acesso à Internet para que a intercomunicação entre os *sites* seja transparente aos usuários.

O *site* inicial é o INPE. Com a evolução do projeto há a previsão da adesão de outras instituições como parceiras através de acordos firmados de cooperação. Essas instituições poderão estar unidas através de uma rede de parceiro caracterizando um possível consórcio internacional de cooperação.

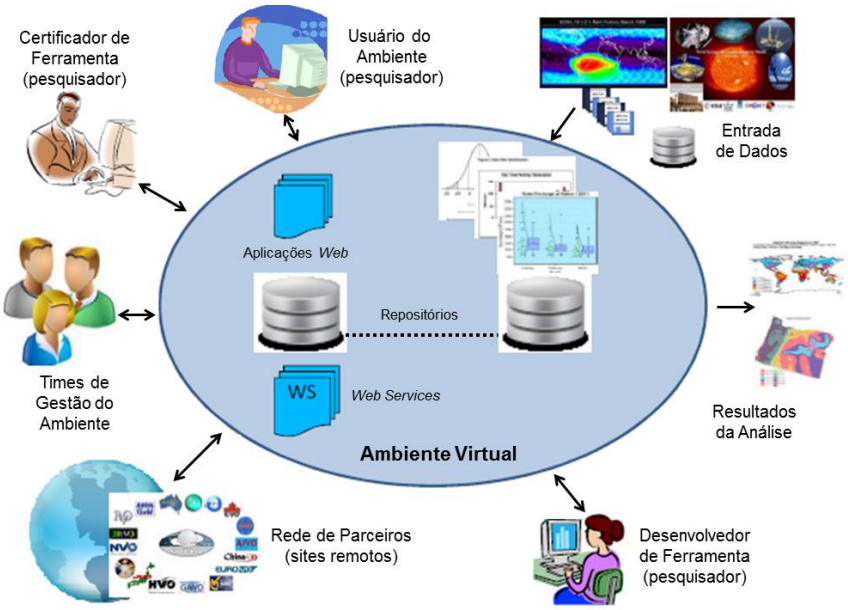


Figura 4.1 – Diagrama de contexto do ambiente virtual

Em sua estrutura interna o ambiente está preparado para armazenar os dados inseridos pelos usuários para a análise. Ele também possui estrutura para salvaguardar os dados das análises e informações dos usuários. Estes repositórios estão interligados com *web services* e aplicações *web* que formalizam a interface com os usuários.

Além da estrutura do ambiente, a ser apresentada com maiores detalhes nos próximos capítulos, e da sua relação com os dados a serem analisados, é possível apresentar os tipos de usuários relacionados.

4.2.1. Usuário do Ambiente

Este tipo de usuário é o pesquisador de alguma área científica, interessado em analisar suas séries temporais através de ferramentas avançadas disponibilizadas no ambiente. Este é o usuário final do ambiente. Ele não é necessariamente especialista em modelos matemáticos ou no desenvolvimento de algoritmo científico, mas conhece a fonte de seus dados e usa os mecanismos de análise do ambiente para processá-los.

O perfil dos usuários do ambiente deve permitir que sejam montados experimentos através das ferramentas de análise. Para que o pesquisador execute sua análise, é necessário que ele seja capaz de carregar as STs no ambiente e escolher as ferramentas de análise. O ambiente deve retornar para o usuário o resultado da execução, com um resumo dos tempos de execução e com a possibilidade de exportação do resultado para tipos comuns de arquivos, como arquivos em texto, PDF ou XML. O contexto da execução deve ficar armazenado como um experimento executado. O usuário pode executar um experimento num momento bem posterior à montagem. Ou seja, ao criar o experimento, o usuário não deve ser obrigado a executá-lo imediatamente. Em cada fase, o experimento deve ser categorizado com um estado como, iniciado, executando ou finalizado. Além disso, o usuário deve ter acesso a todas as informações das ferramentas disponibilizadas no ambiente. Assim, ele terá acesso a toda documentação científica que fundamenta a ferramenta cadastrada no ambiente.

4.2.2. Desenvolvedor de Ferramenta

Este tipo de usuário é um especialista em modelos matemáticos ligados à sua área de atuação. Sua ação principal é no desenvolvimento de algoritmos e na

inserção destes algoritmos no ambiente para que sejam disponibilizados para uso. O desenvolvimento dessas ferramentas pode ser numa linguagem diferente da utilizada no ambiente, desde que a plataforma hospedeira da ferramenta permita a sua execução remota através de um serviço disponibilizado pelo ambiente.

O desenvolvimento das ferramentas de análise é feito fora do ambiente virtual. O perfil do desenvolvedor de ferramenta deve permitir ao usuário com essa característica que ele cadastre e disponibilize ferramentas de análise de dados no ambiente. Para que isso aconteça, o desenvolvedor precisa inserir informações mínimas sobre a ferramenta, como artigos científicos, *websites*, figuras, código fonte do algoritmo, etc. A partir desse cadastro, o usuário deve ser capaz também de cadastrar a ferramenta com mais informações e acoplá-la ao ambiente. Através deste cadastro o usuário deve ser capaz de configurar a localização da ferramenta na rede, indicando o IP do computador hospedeiro da ferramenta, com sua porta de comunicação. Além disso, o usuário deve especificar o tipo de arquivo que contém os dados a serem analisados e o tipo de saída em relação ao processamento da análise. Toda essa ação deve ser passível de edição para correção de erros de configuração ou mudança de alguma informação.

4.2.3. Certificador de Ferramenta

Para que a ferramenta esteja disponível no sistema, ela precisa passar por um certificador de ferramenta. Este usuário é um pesquisador na área científica relacionada e ativo no ambiente, capacitado e indicado para analisar a funcionalidade da ferramenta. É preciso verificar se a produção acadêmica que fundamenta o algoritmo é coerente cientificamente e se o algoritmo faz o que informa que faz. Essa análise faz parte do processo de validação da ferramenta candidata a ser inserida no ambiente. Esse processo pode variar no quesito tempo de análise e deve envolver mais de um pesquisador especialista na área em questão.

O perfil do certificador de ferramenta deve contemplar para o usuário certificador a possibilidade de visualizar as ferramentas recém-cadastradas no sistema e analisá-las. Sendo assim, o usuário certificador deve ter acesso às informações cadastradas acerca da ferramenta em análise e deve ser capaz de executar experimentos para testar as funcionalidades da ferramenta de análise. Através de seu perfil, portanto, este usuário deve alterar a condição da ferramenta como certificada, liberando-a para o uso no sistema. Até essa liberação, a ferramenta não deve estar disponível aos demais usuários. Para que um usuário certificador inicie seu trabalho, o ambiente deve escolher um dos usuários cadastrados de acordo com a especialidade e das informações contidas no cadastro da ferramenta. O certificador deve ser avisado através de um e-mail previamente cadastrado. Ele também deve ter acesso ao desenvolvedor da ferramenta para dialogarem sobre a ferramenta e este processo de validação.

4.2.4. Equipes de Gestão

Estes usuários estão mais ligados à estrutura do ambiente. A primeira é a Equipe de Gestão de Recursos (EGR). Essa equipe é a que deve gerenciar o ambiente em si, cadastrando usuários atribuindo suas funções, monitorando as ações e as necessidades do ambiente visando sua disponibilidade e crescimento. Ela é a responsável para configurar o ambiente e prepará-lo para a utilização. Além disso, é esta equipe que gerará relatórios periódicos do ambiente em termos de uso, manutenção, configurações, solução de problemas, prazos, certificação de ferramentas e modificações no ambiente.

A segunda é a Equipe de Redes e Alto Desempenho (ERAD) que analisa o uso do equipamento e da rede de dados. Esta equipe é responsável por desenvolver ações que permitam que o ambiente esteja sempre disponível, estruturando protocolos e verificando também a segurança de acesso ao ambiente. Esta equipe também é responsável por relatar eventos relacionados ao desempenho e à comunicação que possibilitem o acesso ao ambiente.

A terceira equipe de gestão é a Equipe de Engenharia de Software (EES). Esta equipe é a responsável por remodelar novos recursos disponíveis para o ambiente. Esta equipe deve contar com analistas e programadores experientes para o crescimento do ambiente e para corrigir seus eventuais defeitos de codificação. Sua interação com o ambiente pode estar relacionada à definição de requisitos e testes de novos recursos a serem disponibilizados.

Finalmente, a quarta equipe é a Equipe de Análise de Dados e Algoritmos (EADA). Esta equipe é responsável por dar suporte ao desenvolvimento de ferramentas de análise para o sistema e ao processo de certificação e validação de ferramentas. Os usuários Certificadores de Ferramenta farão parte deste grupo. Este grupo deve compor um fórum público, transparente e colaborativo, análogo às equipes trabalho de código aberto de projetos de desenvolvimento de software.

Cada tipo de usuário representa um perfil específico de acesso, conforme a utilização possível do ambiente pelo usuário. No caso dos usuários ligados às equipes de gestão, o perfil é o administrativo. Neste perfil todas as funcionalidades e informações do ambiente estão disponibilizadas. É através deste perfil que podem ser feitas as configurações do ambiente, a criação, edição ou exclusão dos demais tipos de usuários e o acompanhamento de suas ações. Este usuário deve poder executar tudo o que estiver disponível no ambiente e toda modificação gerada em termos de disponibilização de novos recursos e nova configuração deve estar documentada e disponível no próprio ambiente virtual, mantendo todo o seu histórico.

5 ESTRUTURA DO AMBIENTE

O ambiente especificado e proposto nesta tese considera as tendências na construção de *frameworks* científicos, conforme abordado no Capítulo 2. Este ambiente também atende aos requisitos apresentados no Capítulo 4, promove uma junção de tecnologias de maneira inédita para promover uma nova abordagem para colaboração científica no âmbito da análise de dados e considera a colaboração com processos mediados e bem definidos integrados pela rede e pela Internet.

A colaboração científica é evidenciada a partir dos conceitos explanados no Capítulo 2. Ela refere-se a dois aspectos: o primeiro, no sentido da disponibilização das ferramentas de análise. Cada tipo ou grupo de usuário interage e colabora para a disponibilização de uma nova técnica de análise no ambiente. Já o segundo aspecto refere-se ao uso da ferramenta em si. Com o algoritmo disponibilizado, diferentes grupos de pesquisa podem interagir a partir do ambiente colaborando para a evolução da técnica de análise, além de desenvolverem ciência conjuntamente em suas áreas de pesquisa.

A estrutura do ambiente deve prever também processos bem definidos. O mais importante deles é o da validação de ferramentas, que exemplifica a capacidade do ambiente de participar ativamente de um processo utilizado. Ou seja, há elementos pertencentes ao ambiente que executam ações específicas para controlar o fluxo de uma atividade.

Além disso, todos os componentes relacionados devem ser conhecidos pelo ambiente: os processos, as ferramentas disponibilizadas, a documentação do ambiente, os documentos científicos relacionados às ferramentas de análise, o histórico da evolução do ambiente e os usuários com suas respectivas ações na interação com o ambiente.

Sendo assim, essas demandas podem ser agrupadas em três camadas subjacentes que suportem tais necessidades delineadas na Figura 5.1: a

Camada de Interface com o Usuário, responsável pela interação do usuário com o ambiente; a Camada de Serviços, responsável pela prestação de serviços e a Camada de Armazenamento de Dados, responsável pela integração do ambiente através dos dados do ambiente e dos usuários. Para viabilizar esta estrutura no ambiente virtual, há uma junção entre os produtos de software instalados que dão suporte à estrutura e as aplicações desenvolvidas necessárias à arquitetura apresentada.

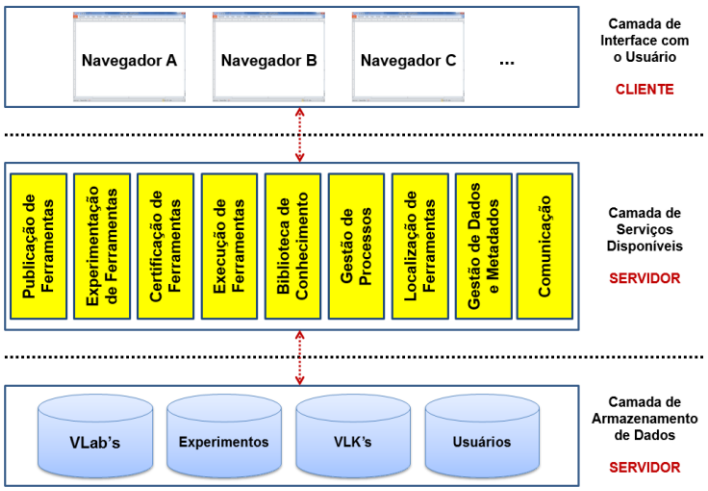


Figura 5.1 – Arquitetura do Ambiente

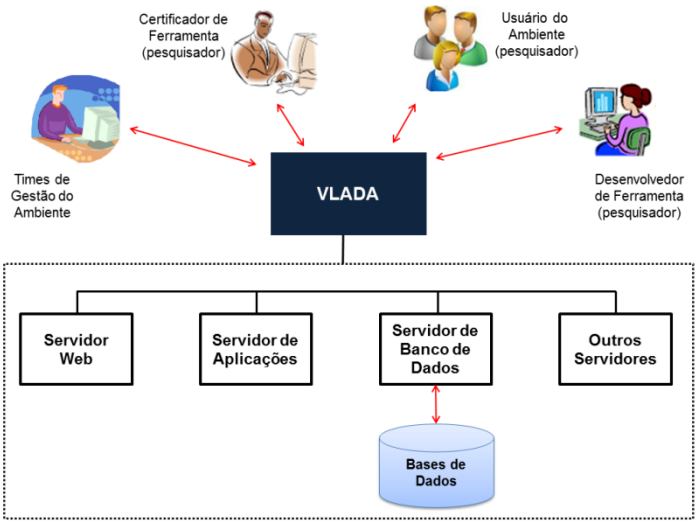


Figura 5.2 – Interação entre o VLADA e outros servidores no ambiente

Este conjunto de aplicações desenvolvidas para o ambiente de análise colaborativa de dados é denominado de VLADA como uma sigla para *Virtual Laboratory for Advanced Data Analysis*. Foi determinado o uso da expressão em inglês para facilitar a internacionalização do projeto. Além dos serviços oferecidos, o VLADA interage com outros sistemas como, por exemplo, o Servidor de Aplicação e o Sistema de Gerenciamento de Banco de Dados, conforme mostrado na Figura 5.2.

5.1. Camada de Interface com o Usuário

Os serviços do ambiente precisam ser acessíveis a partir de um portal *web*, conforme sua especificação de requisitos no Capítulo 4. Esse acesso não deve exigir a instalação de interfaces ou sistemas adicionais no dispositivo cliente, através dos quais o usuário interagiria com o ambiente. Além disso, a interface precisa ser intuitiva, de uso comum, simples e executada em modo gráfico com ambiente de janelas e amplamente difundida em diferentes plataformas. Para atender a essas exigências o navegador *web* foi escolhido para oferecer interface ao usuário, permitindo que ele interaja com os serviços disponíveis na camada inferior.

Os navegadores permitem a interação do usuário com arquivos da Internet. Eles compõem o aplicativo cliente para uma aplicação *web* interagindo com aplicativos servidores, também denominados servidores *web*, através principalmente do protocolo HTTP. Eles têm a capacidade de interpretar e executar diferentes tipos de arquivos de maneira nativa, ou através de *plug-ins*, além de interagir com outros protocolos de rede, inclusive oferecendo criptografia. Uma lista com os navegadores mais famosos historicamente está destacada na Tabela 5.1.

O ambiente proposto é voltado para a *web*. Trata-se de um *website* com ligações para os serviços, expondo-os em uma área comum facilitando o acesso aos recursos pelo usuário final. A Internet está relacionada com mudanças de paradigmas na computação científica e na análise de dados.

Vários aspectos das tecnologias relacionadas com a Internet são fundamentais para o ambiente virtual proposto.

Tabela 5.1 – Navegadores mais conhecidos no mundo

Navegador	Criador	Ano	Plataforma
Chrome	Google	2008	Múltiplas
Firefox	Mozilla	2004	Múltiplas
Safari	Apple	2003	Apple
I. Explorer	Microsoft	1995	Windows
Netscape	Netscape	1994	Múltiplas
Opera	Telenor	1994	Múltiplas
Mosaic	Marc e Eric	1993	Unix
WorldWideWeb	Tim Berners-Lee	1990	NeXTSTEP

O primeiro deles é o acesso a computadores remotos através da rede e de seus protocolos. O uso do navegador, conforme explicitado acima, também promove a facilidade de uso do ambiente, já que muitas funcionalidades são amplamente difundidas no acesso cotidiano à Internet. O terceiro aspecto é a modelagem do ambiente no formato de um portal *web*, o que aumenta a sua usabilidade. Atualmente, o uso de um *website* está no cotidiano da maioria das pessoas interessadas num ambiente como o proposto. Já o quarto aspecto é a facilidade para a portabilidade. Através dos inúmeros protocolos disponíveis, é possível projetar um ambiente a partir de plataformas heterogêneas.

5.2. Camada de Serviços Disponíveis

Como visualizado na Figura 5.1 e descrito no Capítulo 4, o ambiente proposto precisa oferecer um conjunto de serviços. Esses serviços devem permitir ao usuário atingir seu objetivo de fazer a análise de dados de maneira colaborativa com outros cientistas, usando diferentes ferramentas de análise a partir das informações disponibilizadas nas bases de dados. A seguir são apresentados os serviços disponibilizados no ambiente.

5.2.1. Publicação de Ferramentas

Este serviço permite a disponibilização das ferramentas no ambiente pelos pesquisadores desenvolvedores. Ele é responsável por manter as informações das ferramentas publicadas no ambiente, armazenando as informações dos laboratórios virtuais e ferramentas relacionadas. Através de uma API denominada de “ToolExecutor”, o ambiente permite a publicação de ferramentas desenvolvidas em qualquer linguagem, desde que o ambiente nativo da linguagem ofereça uma interface de linha de comandos para a sua execução. Além disso, o computador hospedeiro deve permitir esse tipo de acesso.

As ferramentas publicadas no ambiente podem estar disponibilizadas localmente nos servidores da aplicação ou remotamente em outros equipamentos. O acionamento da ferramenta se dá através do seu endereço informado no cadastro, contendo número IP e número da porta do serviço, e de um serviço de execução disponível no seu hospedeiro.

5.2.2. Experimentação de Ferramentas

O serviço de experimentação de ferramentas possibilita a análise prévia do funcionamento da ferramenta candidata ao acoplamento no ambiente. É através deste serviço que o processo de validação poderá ser executado. Ele armazena os dados de experimentos das ferramentas de modo a possibilitar a sua certificação. Também fornece a aplicação para publicar as ferramentas de análise, definindo os formatos aceitos e acompanhamento no processo de validação, possibilitando sua certificação e liberação para uso no ambiente. O pesquisador responsável pela certificação da ferramenta pode executá-la também acompanhando a documentação científica cadastrada. Ao avaliar o algoritmo, o certificador pode analisar o funcionamento da ferramenta, indicar ajustes ou solicitar mais informações até que a ferramenta esteja apta à ampla divulgação.

Para isso, o serviço de experimentação de ferramentas permite a criação de experimentos de análise de dados usando a metáfora de laboratórios de ciências. Cada laboratório agrupa um tipo de ferramentas de análise. O serviço de experimentação deve permitir, portanto, a criação, visualização, edição e exclusão de experimentos; a entrada dos dados a serem analisados; e a execução do experimento que está intrinsecamente ligada à execução de uma ferramenta de análise de dados.

5.2.3. Certificação de Ferramentas

Uma proposta mais inovadora do ambiente é a possibilidade da disponibilização das ferramentas de análise de séries temporais oriundas de diferentes grupos de pesquisa, incrementando a colaboração. Entretanto, é preciso considerar que é importante o estabelecimento de um padrão de confiabilidade da ferramenta com métricas objetivas, como por exemplo, difusão da técnica, corretude algorítmica e objetivo claro da solução.

Assim, para que o ambiente forneça mecanismos de análise de dados com alto rigor científico, é preciso certificar a credibilidade da ferramenta candidata. O serviço de certificação de ferramentas impõe um fluxo de trabalho definido para essa tarefa, conforme passos a seguir:

1º passo: o usuário desenvolvedor de ferramenta recebe acesso ao ambiente;

2º passo: o desenvolvedor cadastra a ferramenta no ambiente, que não a disponibiliza para todos os usuários, mas apenas para usuários certificadores de ferramenta.

3º passo: o grupo de usuários certificadores relacionado com aquele tipo de ferramenta recebe a notificação por e-mail de que a ferramenta foi disponibilizada. As ferramentas devem estar categorizadas por áreas denominadas como laboratórios virtuais e cada usuário Certificador de Ferramenta precisa estar relacionado a um dos laboratórios.

4º passo: a Equipe de Análise de Dados e Algoritmos deve designar até três pesquisadores para o processo de certificação para interagir com a ferramenta criando experimentos, analisando sua documentação, sugerindo melhorias ou correções até a finalização do processo que é marcada com a liberação da ferramenta para uso no ambiente.

5º passo: com a ferramenta certificada é efetivada sua publicação no ambiente para o uso geral pelos demais tipos de usuários.

5.2.4. Execução de Ferramentas

O serviço de execução de ferramentas permite que as ferramentas disponibilizadas no ambiente sejam acionadas pelos usuários. Um *thread* deve ser criado para atender cada solicitação de usuário. Essa execução pode ser local ou remota através de protocolos como o IP, TCP, SOA e HTTP. Este serviço deve empacotar o comando acionador da ferramenta juntamente com seus parâmetros e os enviar para o *site* hospedeiro da ferramenta. Neste hospedeiro é preciso ter um serviço de recepção instalado apto a receber a requisição.

A solicitação deve então ser desempacotada e encaminhada para o serviço de terminal da linguagem na qual a ferramenta foi desenvolvida. Este serviço deve interpretar a solicitação e encaminhar a execução da ferramenta conforme os parâmetros enviados. Após a execução da ferramenta, o resultado da análise deve ser encaminhado em formato de texto para o ambiente que retornará para o usuário interessado. Tal resultado poderá ser exportado em formatos diversos como XML, XLS ou PDF, formatos aceitos pela comunidade científica.

5.2.5. Biblioteca de Conhecimento

Este serviço consiste de uma base de dados com informação acerca das ferramentas contidas no ambiente, bem como recursos de busca para a melhor localização deste conhecimento. Tais recursos devem estar relacionados às

ferramentas, metodologias e outros assuntos disponíveis no ambiente. Essa base de dados poderá ser consultada a qualquer tempo pelos usuários do ambiente através de um mecanismo de buscas com possibilidades de filtragem.

Através deste serviço diferentes tipos de documentos podem estar disponibilizados no ambiente, como artigos científicos, códigos-fontes, séries temporais de exemplo, tutoriais, resultados esperados em formato de texto ou de imagens, figuras, gráficos de exemplo, vídeos, áudios, etc. Cada documento deve estar associado com uma ferramenta disponibilizada no ambiente para que os usuários, desde os certificadores de ferramenta aos usuários finais, consigam compreender o uso da ferramenta e ajudar a avançar cientificamente a técnica de análise referida.

5.2.6. Gestão de Processos

O ambiente virtual precisa apoiar processos relacionados com a análise de séries temporais. O serviço de gestão de processos deve ser capaz de configurar, instanciar, acompanhar e finalizar processos, dispondo dos meios pelos quais os mesmos são tratados pelo ambiente. Este serviço deve garantir que um processo siga seu fluxo preestabelecido dentro do ambiente, através de ações executadas para avisar e relatar os usuários interessados no referido processo sobre seu andamento e eventuais ocorrências.

5.2.7. Localização de Ferramentas

Este serviço permite através de *threads* atuantes nas bases de dados a busca pelas ferramentas disponíveis por meio de palavras chave e categorias. Esta busca deve estar associada inicialmente com o laboratório virtual relacionado em que a ferramenta está registrada, mas isso não é restritivo. Assim como na busca de documentação, deve ser possível buscar as ferramentas conforme filtros estabelecidos pelo usuário.

5.2.8. Gestão de Dados e Metadados

O serviço de Gestão de Dados e Metadados é o responsável por gerenciar os usuários e seus perfis, as configurações do ambiente e os relatórios de uso. Ele mantém o repositório de usuários com seus dados pessoais como nome, sobrenome, e-mail, etc. e efetua o controle de acesso que consulta o perfil e autoriza o acesso a funcionalidades do ambiente. Além disso, este serviço deve permitir o gerenciamento da configuração do ambiente liberando sua edição e também o gerenciamento de seus relatórios de uso.

Através deste serviço, o ambiente efetua a salvaguarda, a atualização e a recuperação dos dados relacionados aos experimentos, às ferramentas, aos usuários e à documentação relativa às ferramentas. Este serviço deve interagir com os sistemas gerenciadores de banco de dados que estão na Camada de Armazenamento de Dados.

5.2.9. Comunicação

O ambiente deve prover um serviço de comunicação entre os usuários conforme a consecução dos processos. Este serviço deve contemplar a troca de mensagens entre os usuários, considerando um remetente e um destinatário. Além disso, cada mensagem deve conter o assunto a ser tratado e o seu conteúdo.

O serviço de comunicação deve permitir a troca de mensagens entre os usuários através do envio de e-mail por meio de serviço de SMTP. Deve permitir também a configuração deste serviço para o envio de mensagens internamente. Tais mensagens podem ser previamente automatizadas de acordo com os eventos ou elaboradas pelos usuários através de uma interface adequada de criação e edição de mensagens.

5.3. Camada de Armazenamento de Dados

Esta camada do ambiente comporta os componentes relacionados com o armazenamento e recuperação de informações. As informações devem estar categorizadas em quatro grandes grupos de dados: os laboratórios virtuais com as respectivas ferramentas disponibilizadas conforme um agrupamento definido, também denominados como *Virtual Laboratory*, ou VLab[39]; os experimentos criados pelos usuários com seus respectivos conjuntos de dados a serem analisados pelas ferramentas; as bibliotecas virtuais categorizando os tipos de informações relacionadas às ferramentas, também denominados como *Virtual Library Knowledge* ou VLK[40], e os dados de usuário, como seu perfil, suas ações, suas comunicações, etc.

Os componentes desta camada podem estar em plataformas heterogêneas, centralizadas ou distribuídas, integrando e gerenciando as informações do ambiente. O uso e o acesso a esses dados devem ser transparentes para os usuários independentemente da infraestrutura adotada. Além disso, é preciso considerar o acesso concorrente e simultâneo aos dados pelos diversos usuários interessados.

Esta camada deve contemplar tráfego simultâneo de grande volume de dados ocasionado pelo uso crescente do ambiente. Também deve ser considerada a eficiência de armazenamento e recuperação dos diversos tipos de informações mencionados acima.

Toda informação do ambiente precisa ser tratada com segurança. Esta camada deve oferecer mecanismos de controle que permitam a restrição de acesso ao ambiente conforme os perfis dos usuários. Este acesso ao ambiente pode ser local, ou seja, a partir da rede local do *site* onde o ambiente estará disponível, ou remoto com acesso através da Internet.

[39] O Capítulo 6 apresenta um protótipo designando os VLabs.
[40] O Capítulo 6 apresenta um protótipo designando os VLKs.

Finalmente, esta camada é responsável por garantir alta disponibilidade do ambiente, conforme a disponibilização dos dados. O fator da disponibilidade também está relacionado com a recuperação consistente dos dados após uma falha lógica ou física.

6 UM PROTÓTIPO DO AMBIENTE

Este capítulo apresenta um protótipo do ambiente proposto nesta tese. O desenvolvimento e a montagem deste protótipo precisam de estrutura de rede, computadores servidores e computadores apropriados e configurados para o desenvolvimento do ambiente. O protótipo apresentado a seguir está instalado no Laboratório de Computação e Matemática Aplicada (LAC) no INPE. Este capítulo descreve inicialmente a infraestrutura de hardware usada para o protótipo. Em seguida, é destacada a infraestrutura de software que dá suporte ao ambiente. Finalmente, são abordadas também as principais atividades relacionadas com a construção do protótipo, que é apresentado na sequência em suas principais interfaces.

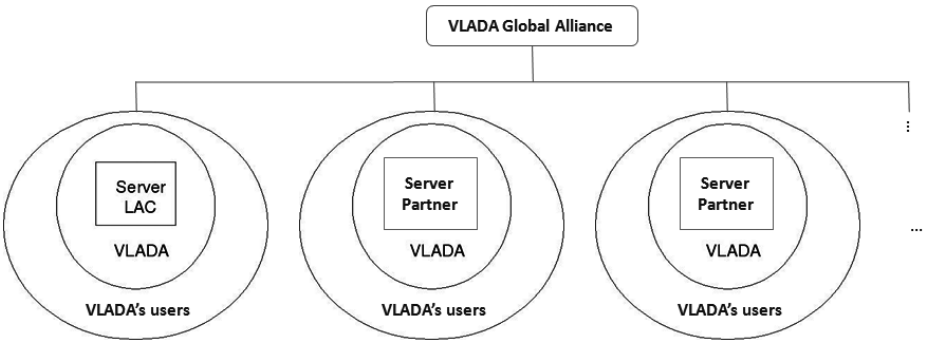


Figura 6.1 – Perspectiva de uma Aliança Global para o VLADA

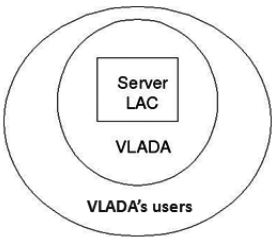


Figura 6.2 – Protótipo VLADA no INPE

A Figura 6.1 ilustra o plano de o VLADA ser concebido num contexto de comunidade colaborativa para ser um ambiente virtual aberto. Ele pode agregar novos parceiros para ampliar a disponibilidade das ferramentas com hardware

e software adicionais a fim de formar uma rede global de ambientes virtuais para a análise de dados, aumentando a quantidade e a variedade de ferramentas de análise em grade. O hardware e o software mínimos para um protótipo local expansível foram disponibilizados pelo LAC-INPE, conforme esquema da Figura 6.2, acessíveis apenas à comunidade do INPE inicialmente.

O portal *web* do VLADA é denominado de VLADA *Web*. Ele irá fornecer uma interface fácil de usar para acessar o ambiente e seus recursos. Nesta fase de prototipação as técnicas disponibilizadas foram a DFA e a média aritmética. O VLADA *Web* conduz o usuário do INPE para a aplicação da DFA numa ST. De maneira resumida, o usuário deve efetuar *login* no ambiente que permite acesso às ferramentas disponibilizadas, juntamente com sua documentação. O usuário poderá então fazer o *upload* da sua ST e efetuar a sua análise no contexto de um experimento. A Figura 6.3 ilustra essa operação.

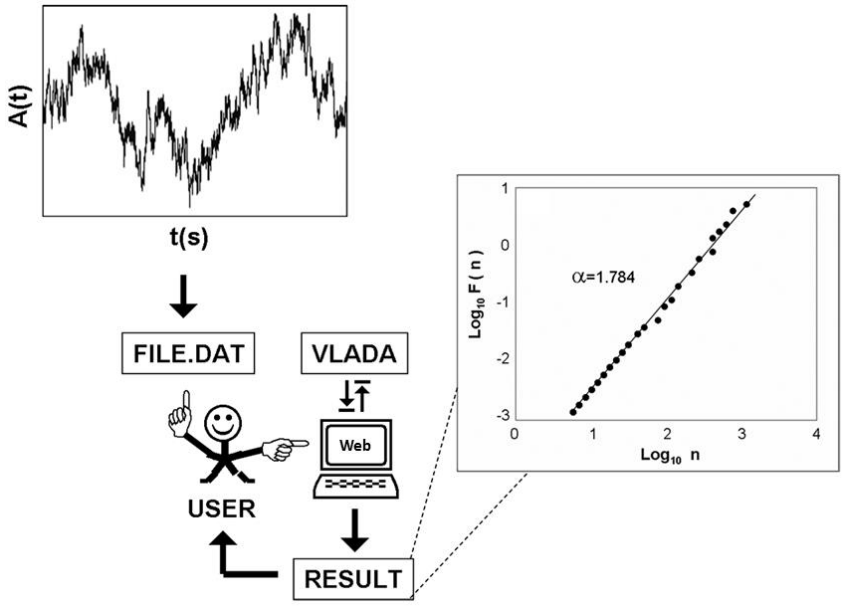


Figura 6.3 – Ilustração da operação do ambiente pelo usuário

6.1. Infraestrutura de Hardware

A infraestrutura de hardware contempla uma rede de dados composta pelos *switches* de alta densidade que proveem a estrutura do LAC/INPE. Esta estrutura está conectada à rede corporativa do INPE monitorada de maneira automatizada. Ela também obedece aos padrões estabelecidos de cabeamento estruturado para categoria 5e.

O *rack* principal contém os *switches* do *backbone* de 1 Gbps (um gigabit por segundo) e é a partir dele que os servidores se conectam à rede e à Internet. A Figura 6.4 apresenta um diagrama dos equipamentos mínimos para a operação em larga escala representados pelos elementos coloridos. Além disso, o diagrama possui elementos tracejados sugerindo sua evolução.

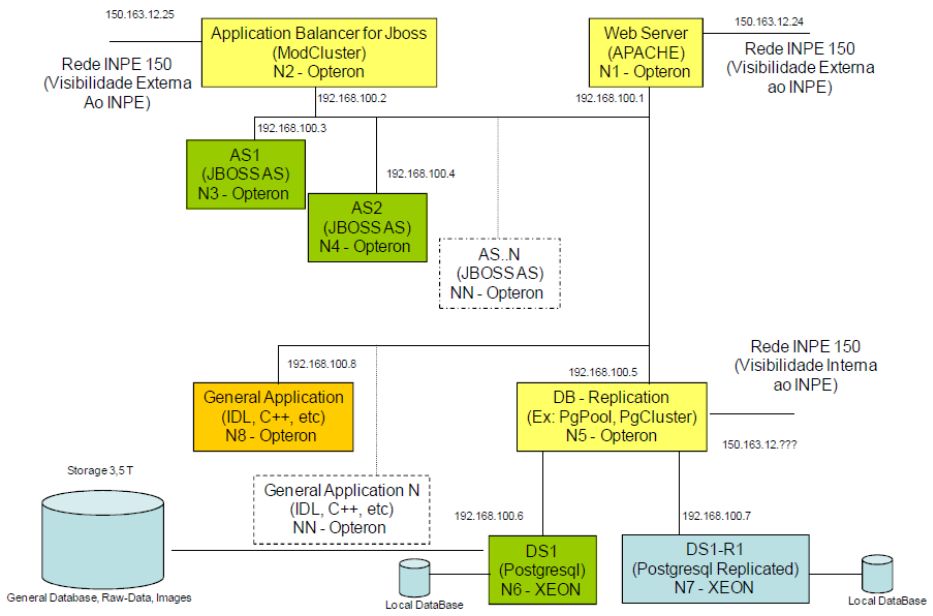


Figura 6.4 – Diagrama dos equipamentos mínimos para a operação

Conforme pode ser observado também na Figura 6.4, o ambiente pressupõe a estrutura para servidores de dados, de aplicação e *web*, sendo que os dois primeiros tipos de servidores preveem balanceamento de carga entre equipamentos. Isso pode garantir disponibilidade e desempenho satisfatório.

Atualmente, estes equipamentos estão disponibilizados no INPE através de um cluster de três servidores HP 2U com a seguinte configuração: 2 processadores *quadcore*, 32 GB de RAM e 4 discos rígidos de 1 TB cada (ver Figura 6.5). Além disso, a infraestrutura é composta por quatro sistemas de armazenamento, ou *storages*, com capacidade para 3,6 TB cada (ver Figura 6.6) conectados aos servidores através de uma *Storage Area Network* (SAN).

Figura 6.5 – *Cluster* do LAC/INPE

Figura 6.6 – *Storage* do LAC/INPE

6.2. Infraestrutura de Software

6.2.1. Ferramentas para Estrutura do Ambiente

Seguindo a proposta de estabelecer um ambiente virtual livre, todos os softwares adotados para este ambiente são livres e de código aberto. Para o protótipo em questão, o servidor de aplicações e *web* escolhido foi o JBoss[41] 6.0. Este servidor é uma aplicação livre de código fonte aberto mantida atualmente pela RedHat desde 2006. Ele é escrito em Java e tem sua estrutura baseada no J2EE (*Java Platform Enterprise Edition*), fornecendo um ambiente completo para que o VLADA possa usufruir de serviços já disponibilizados, como acesso a banco de dados, mecanismos de autenticação, segurança e *clustering* dos servidores de aplicação.

Já o sistema de gerenciamento de banco de dados escolhido foi o PostgreSQL[42] 8.4 para gerenciar e armazenar os dados do ambiente. Ele é um servidor de banco de dados objeto-relacional de código aberto que funciona em diversos sistemas operacionais. Mesmo sendo um sistema livre, este servidor é reconhecido como um forte concorrente de mercado para produtos comerciais como MS SQL Server e Oracle Database. Com amplo desenvolvimento, ele suporta completamente chaves estrangeiras, *joins*, *views*, *triggers*, procedimentos armazenados (em vários idiomas) e SQL. Suporta aplicações em diversas interfaces de programação, como: C, C++, Java, .Net, Perl, Python, Ruby, Tcl, ODBC, etc. Sua estrutura suporta alta escalabilidade para grande número de usuários e grandes quantidades de dados.

Para o protótipo também foram instalados serviços para suporte de informações sobre o sistema computacional e *backup* dos dados. Tais serviços operam em plataforma *web*. Para isso o servidor *web* escolhido foi o Apache. Este software é mantido pela Fundação Apache Software, uma entidade sem fins lucrativos que dá suporte legal, organizacional e financeiro a mais de 140 projetos de software livre (APACHE, 2013). Com vários programadores

[41] Website: www.jboss.org.
[42] Website: www.postgresql.org.

envolvidos em seu desenvolvimento colaborativo, o servidor *web* Apache é um produto robusto utilizado livremente por diversos projetos em vários níveis de atuação de mercado. O início de seu desenvolvimento refere-se a meados de 1994 no *National Center for Supercomputing Applications* (NCSA) por Rob McCool (APACHE, 2013). Foi lançado como Apache 1.0 em 1995 com muitos recursos na forma de módulos padrões e em 1996 tornou-se o servidor *web* mais popular na Internet, de acordo com a empresa Netcraft[43], empresa que efetua pesquisas periódicas sobre a Internet, seus produtos e serviços. Atualmente, o Apache está em sua versão 2.4.

Além destas aplicações servidoras, o protótipo faz uso de Padrões de Projeto com a utilização do *framework* Hibernate para persistência dos dados, permitindo o mapeamento dos atributos entre a base tradicional de dados relacionais e o modelo objeto da aplicação; do *framework* Spring para a camada de negócio e, para a camada de aplicação, do *framework* Struts ou Java Server Faces e de Ajax.

6.2.2. Ferramentas para o Desenvolvimento

Foram usadas algumas ferramentas específicas no desenvolvimento do protótipo. Para a modelagem Orientada a Objetos, foi usado o programa MagicDraw[44]. Esta é uma aplicação reconhecida por sua capacidade de modelar processos de negócios e arquitetura de um sistema, proporcionando engenharia de código e do banco de dados. Ela suporta UML 2, o padrão XMI para armazenamento de dados e diversas linguagens de programação. Sua escolha foi baseada na capacidade de integração com diversas ferramentas, como ambientes de desenvolvimento, requisitos, testes, banco de dados e outros.

Juntamente com as demais aplicações, foi usado o Enterprise Architect, que é uma ferramenta gráfica multiusuária projetada para a utilização na construção de ambientes de grande porte. Ele utiliza diversos padrões abertos, como UML

[43] Website: www.netcraft.com.
[44] Website: www.nomagic.com.

e SysML, dentre outros. Esta aplicação é reconhecida pelo seu desempenho em manipular grandes modelos de maneira eficiente, além de permitir o compartilhamento de projetos possibilitando o desenvolvimento colaborativo. Oferece muitas possibilidades, como a simulação do ambiente em desenvolvimento; o acompanhamento de todas as fases na construção de um sistema, dos requisitos à implantação; a simplificação da gestão do processo de desenvolvimento; a geração de documentação e de código fonte, inclusive através de engenharia reversa; a depuração, a compilação e a visualização da aplicação; dentre outras funcionalidades.

Para a gestão do protótipo foi usado o e-WebProject (SANT'ANNA et al., 2002 e SANT'ANNA, 2000), que é um ambiente de engenharia de software centrado em processos. Seu principal objetivo é fornecer um conjunto de ferramentas integradas para o trabalho cooperativo, suporte à execução de processos de engenharia de software, gestão do conhecimento organizacional e apoio à infraestrutura organizacional. Diversos trabalhos (CEREJA JR. et al, 2003; GENVIGIR, 2004; ABDALA, 2004 e LAHOZ, 2004) deram continuidade ao trabalho desenvolvido por SANT'ANNA (2000) tendo como base esse ambiente, agregando módulos e serviços ampliando seu alcance.

Finalmente, para o desenvolvimento também foi utilizado o IDE (do inglês, *Integrated Development Environment*) NetBeans[45]. Ele oferece editores, analisadores de código e conversores que permitem o desenvolvimento integrado e produtivo de aplicações em diversas linguagens computacionais, como Java, C/C++, XML e HTML, PHP, Groovy, Javadoc, JavaScript e JSP (NETBEANS, 2013). É um ambiente muito conhecido e utilizado pela comunidade de desenvolvimento de aplicações de grande porte.

6.3. Estrutura do VLADA

Todos os serviços do VLADA estão interligados a partir de seus componentes. Eles refletem a arquitetura proposta no Capítulo 5 e estão interconectados

[45] Website: netbeans.org.

através de protocolos bem definidos. A Figura 6.7 apresenta um diagrama de implantação representando uma perspectiva física do VLADA com tais componentes.

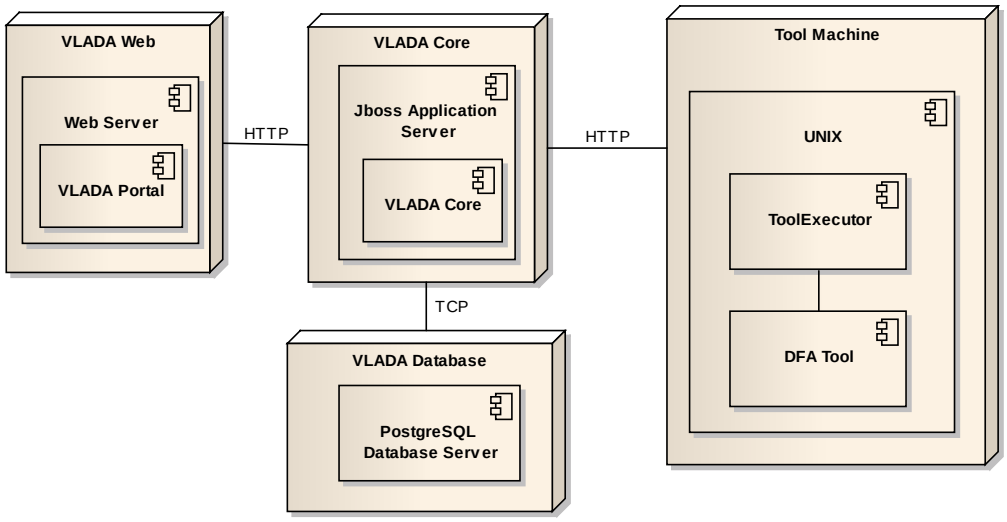


Figura 6.7 – Perspectiva física do VLADA

O componente que mais interage com o usuário é o VLADA *Web*. Ele contém a aplicação cliente do ambiente que permite publicação de ferramentas, experimentos, revisão e localização de documentos. Esta aplicação cliente é o VLADA Portal, uma aplicação *Web* que propicia as ações supracitadas para o usuário.

As regras de negócio do ambiente estão no nó denominado VLADA Core. É nele que estão os serviços expostos pelo VLADA, como o publishTool, para publicar ferramentas; o createExperiment, para criar experimentos; o executeExperiment, para executar experimentos, entre outros. Há ainda o VLADA Database que contém o Sistema Gerenciador de Banco de Dados gerenciando e armazenando os dados do ambiente; o Tool Machine, que abriga alguma instância de ferramenta como a DFA, a GSA ou a Média Aritmética e o Tool Executor, que expõe acesso a alguma Ferramenta na *Web*. Neste trabalho, é evidenciada a DFA Tool, uma instância da ferramenta DFA. Assim, as ferramentas de análise podem ser acessadas através da interface, conforme sequência ilustrada na Figura 6.8. As rotinas da interface estão

dispostas em JSF e o VLADA Core foi desenvolvido em J2EE. Já as ferramentas de análise de dados, usualmente escritos em C, C++, Java, Python ou em linguagens de alto nível, devem ser executadas pelo VLADA Tool Executor, desenvolvido em Java.

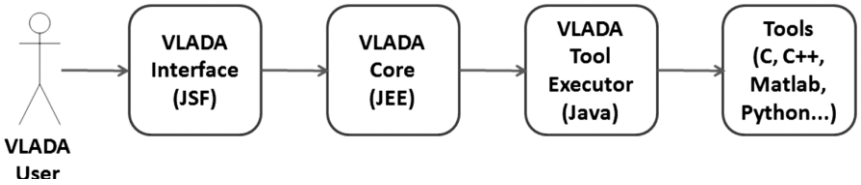


Figura 6.8 – Sequência de acesso aos artefatos de softwares

É possível também esquematizar o ambiente a partir de um ponto de vista lógico. A Figura 6.9 apresenta um diagrama de componentes representando essa perspectiva do VLADA.

O VLADA precisa ter uma estrutura final que permita a organização dos procedimentos de análise de dados. Apesar do foco desta pesquisa ser voltada para o estudo de séries temporais, o protótipo foi projetado para disponibilizar as ferramentas de análise em seis grandes grupos denominados Laboratórios Virtuais, ou no inglês, conforme adotado no protótipo, *Virtual Laboratories* (VLabs). São eles:

- VLab1: ferramentas de visualização e análise padrão de dados;
- VLab2: ferramentas avançadas de estatística;
- VLab3: ferramentas avançadas para análise de séries temporais;
- VLab4: ferramentas avançadas de análise espaço-temporal;
- VLab5: ferramentas avançadas para sistemas de dados multivariados;
- VLab6: técnicas avançadas de mineração de dados.

O VLab1 corresponde à visualização e à análise padrão de dados. Alguns exemplos de análise padrão podem ser citados, como: o cálculo de momentos estatísticos, os histogramas, as funções de autocorrelação e outros encontrados em muitos livros didáticos de estatística básica. Assim, como

mostrado na Figura 6.10, o VLab1 deverá englobar a maior quantidade de técnicas do ambiente.

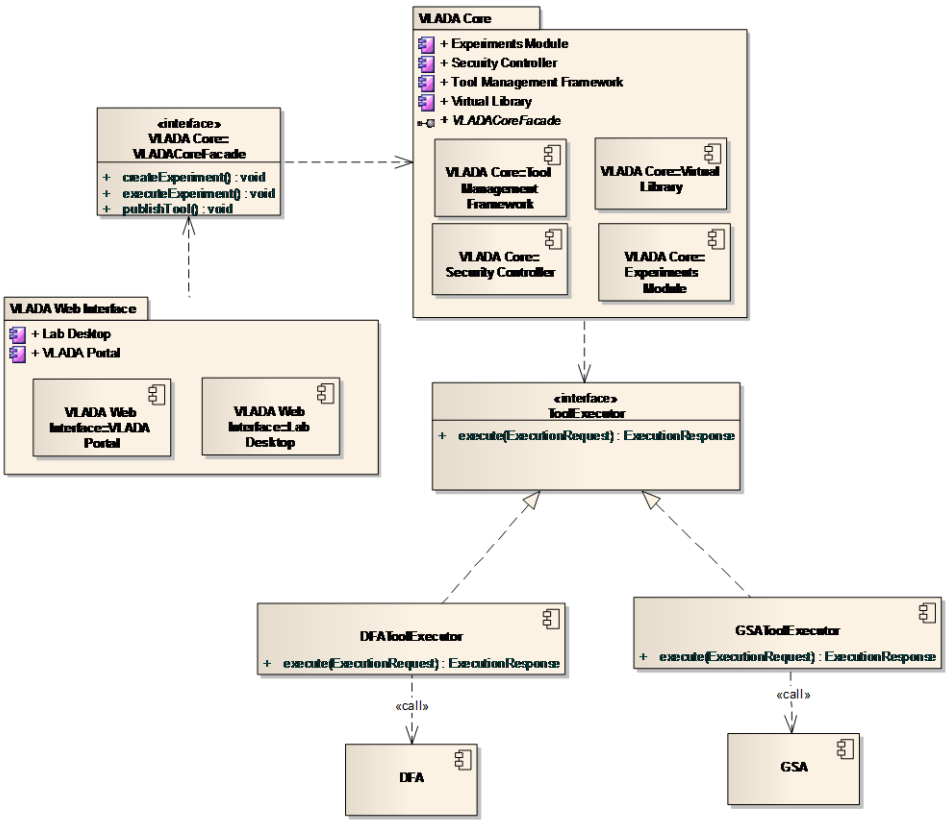


Figura 6.9 – Perspectiva lógica do VLADA

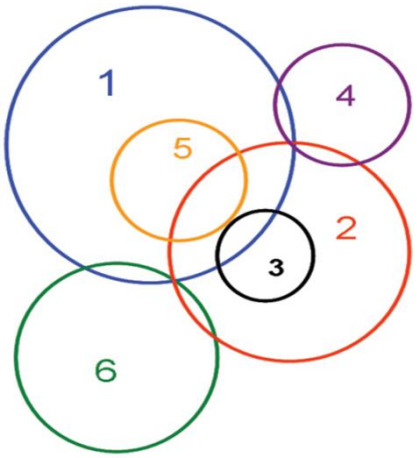


Figura 6.10 – Domínios do VLADA e relações de complementaridade

Apesar de a proposta ser de um ambiente para técnicas avançadas de análise, é suposto que todas as técnicas do VLab1 sejam conhecidas e de muita utilização na comunidade acadêmica. Sendo assim, devido a fins de completude do ambiente, ferramentas designadas para o VLab1 deverão ser aceitas, considerando o VLADA como um projeto aberto colaborativo de longo prazo.

Os VLabs foram assim idealizados a partir das possíveis grandes áreas de análise avançada de dados, com exceção do VLab1. Na estratégia de desenvolvimento do protótipo, foi escolhido adotar o VLab3 como o primeiro VLab com uma ferramenta disponibilizada no ambiente, sendo que foi escolhida a ferramenta do DFA devido ao foco de pesquisa na análise de STs e às motivações explicitadas nos capítulos 1 e 3.

Cada ferramenta disponibilizada deve conter todas as informações que possam ser necessárias para orientar a sua certificação e seu uso na análise virtual, além de tornar tão fácil quanto possível o uso do ambiente por usuários não tão familiarizados com ferramentas analíticas. Desta forma, os usuários serão capazes de acessar de forma independente um repositório de conhecimento composto de três Bibliotecas Virtuais de Conhecimento (VLK do inglês, *Virtual Library Knowledge*):

- VLK1: repositório de texto básico, referências científicas e *websites*;
- VLK2: repositório de pacotes de software e de códigos fontes;
- VLK3: repositório de dados exemplos e materiais gráficos ilustrativos.

6.3.1. Tarefas Realizadas

Para o desenvolvimento do protótipo, muitas tarefas foram executadas. Essas tarefas foram agrupadas em seis fases. A seguir são listadas as tarefas, conforme as fases de desenvolvimento.

Fase 1: Gerenciamento de Dados e de Metadados

- Revisar, analisar e especificar requisitos de controle de acesso;

- Elaborar o modelo do banco de dados do controle de acesso;
- Projetar o banco de dados do controle de acesso;
- Desenvolver o banco de dados de controle de acesso;
- Testar o banco de controle de acesso;
- Criar a análise da aplicação de autenticação e autorização baseada em perfis;
- Projetar a aplicação de autenticação e autorização baseada em perfis;
- Desenvolver a aplicação de autenticação e autorização baseada em perfis;
- Testar a aplicação de autenticação e autorização baseada em perfis;
- Criar a análise da aplicação de gerenciamento da configuração do ambiente;
- Projetar a aplicação de gerenciamento da configuração do ambiente;
- Desenvolver a aplicação de gerenciamento da configuração do ambiente;
- Testar a aplicação de gerenciamento da configuração do ambiente;
- Criar a análise da aplicação de controle de sessões e relatórios do ambiente;
- Projetar a aplicação de controle de sessões e relatórios do ambiente;
- Desenvolver a aplicação de controle de sessões e relatórios do ambiente;
- Testar a aplicação de controle de sessões e relatórios do ambiente;
- Efetuar testes de integração;
- Elaborar a documentação;
- Instalar o serviço.

Fase 2: Publicação e Experimentação de Ferramentas

- Revisar, analisar e especificar requisitos da publicação de ferramentas e da criação de experimentos;
- Elaborar o modelo do banco de dados do repositório de laboratórios virtuais;

- Projetar o banco de dados do repositório de laboratórios virtuais;
- Desenvolver o banco de dados do repositório de laboratórios virtuais;
- Testar o repositório de laboratórios virtuais;
- Elaborar o modelo do banco de dados do repositório de experimentos;
- Projetar o banco de dados do repositório de experimentos;
- Desenvolver o banco de dados do repositório de experimentos;
- Testar o banco de dados do repositório de experimentos;
- Criar a análise da aplicação para publicação de ferramentas;
- Projetar a aplicação para publicação de ferramentas;
- Desenvolver a aplicação para publicação de ferramentas;
- Testar a aplicação para publicação de ferramentas;
- Criar a análise da aplicação para validação e certificação de ferramentas;
- Projetar a aplicação para validação e certificação de ferramentas;
- Desenvolver a aplicação para validação e certificação de ferramentas;
- Testar a aplicação para validação e certificação de ferramentas;
- Efetuar testes de integração no processo de certificação e validação;
- Elaborar a documentação no processo de certificação e validação;
- Instalar o serviço.

Fase 3: Localização de Ferramentas

- Revisar, analisar e especificar requisitos para a localização de ferramentas;
- Criar a análise da aplicação de busca de ferramentas por palavra chave e categorias;
- Projetar a aplicação de busca de ferramentas por palavra chave e categorias;
- Desenvolver a aplicação de busca de ferramentas por palavra chave e categorias;

- Testar a aplicação de busca de ferramentas por palavra chave e categorias;
- Efetuar testes de integração;
- Elaborar a documentação;
- Instalar o serviço.

Fase 4: Biblioteca Virtual de Conhecimento

- Revisar, analisar e especificar requisitos para a biblioteca de recursos;
- Elaborar o modelo do banco de dados do repositório de recursos;
- Projetar o banco de dados do repositório de recursos;
- Desenvolver o banco de dados de repositório de recursos;
- Testar o banco do repositório de recursos;
- Criar a análise da aplicação de busca por palavra-chave;
- Projetar a aplicação de busca por palavra-chave;
- Desenvolver a aplicação para a ferramenta de busca por palavra-chave;
- Testar a aplicação de busca por palavra-chave;
- Efetuar testes de integração;
- Elaborar a documentação;
- Instalar o serviço.

Fase 5: Execução e Certificação de Ferramentas, Gestão de Processos e Comunicação

- Revisar, analisar e especificar requisitos para a execução e certificação de ferramentas, gestão de processos e comunicação;
- Criar a análise da aplicação de carga de dados, acionamento de ferramentas, gestão de processos e comunicação;
- Projetar a aplicação de carga de dados, acionamento de ferramentas, gestão de processos e comunicação;
- Desenvolver a aplicação de carga de dados, acionamento de ferramentas, gestão de processos e comunicação;

- Testar a aplicação de carga de dados, acionamento de ferramentas, gestão de processos e comunicação;
- Criar a análise da aplicação de exportação dos resultados para os formatos pdf, csv, xls e xml;
- Projetar a aplicação de exportação dos resultados para os formatos pdf, csv, xls e xml;
- Desenvolver a aplicação de exportação dos resultados para os formatos pdf, csv, xls e xml;
- Testar classes da aplicação de exportação dos resultados para os formatos pdf, csv, xls e xml.
- Efetuar testes de integração;
- Elaborar a documentação;
- Instalar os serviços.

Fase 6: Portal VLADA

- Revisar, analisar e especificar requisitos do portal;
- Projetar o portal;
- Desenvolver o portal;
- Testar o portal;
- Efetuar testes de integração;
- Elaborar a documentação;
- Instalar o serviço.

6.3.2. Artefatos de Software Produzidos

A partir das tarefas mencionadas acima, foram produzidos alguns artefatos de software, conforme destaque a seguir nas respectivas fases.

Fase 1: Gerenciamento de Dados e de Metadados

- Repositório de usuários;
- Aplicação de autenticação e autorização baseada em perfis;

- Aplicação de gerenciamento da configuração e relatórios de uso do ambiente.

Fase 2: Publicação e Experimentação de Ferramentas

- Repositório de laboratórios virtuais;
- Repositório de experimentos;
- Aplicação para publicação de ferramentas;
- Aplicação para validação e certificação de ferramentas.

Fase 3: Localização de Ferramentas

- Aplicação de busca de ferramentas por palavra chave e categorias.

Fase 4: Biblioteca Virtual de Conhecimento

- Repositório de recursos;
- Aplicação de busca por palavra-chave;
- Aplicação de publicação de imagens, documentos e textos que documentam as ferramentas de análise de dados.

Fase 5: Execução e Certificação de Ferramentas, Gestão de Processos e Comunicação

- Aplicação de carga de dados e acionamento de ferramentas;
- Aplicação de certificação de ferramentas;
- Aplicação de gestão de processos;
- Aplicação de comunicação;
- Aplicação de exportação dos resultados para os formatos pdf, csv, xls e xml.

Fase 6: Portal VLADA

- Interface Web com links e recursos de integração.

6.4. Protótipo em Execução

6.4.1. Interfaces de Acesso

A Figura 6.11 apresenta a página inicial do *website* do ambiente protótipo. Este *website* está hospedado no endereço http://lacteste.sir.inpe.br/vlada/index.php. Ele oferece um *menu* de opções, conforme destaque em vermelho na Figura 6.11. Entre as opções estão uma breve explicação do protótipo, dos laboratórios virtuais, das bibliotecas de conhecimento, dos parceiros do projeto, dos grupos de trabalho, de como o usuário pode obter ajuda e tirar dúvidas, além de opção de contato e de acesso ao ambiente de análise.

Figura 6.11 – Portal protótipo do ambiente VLADA

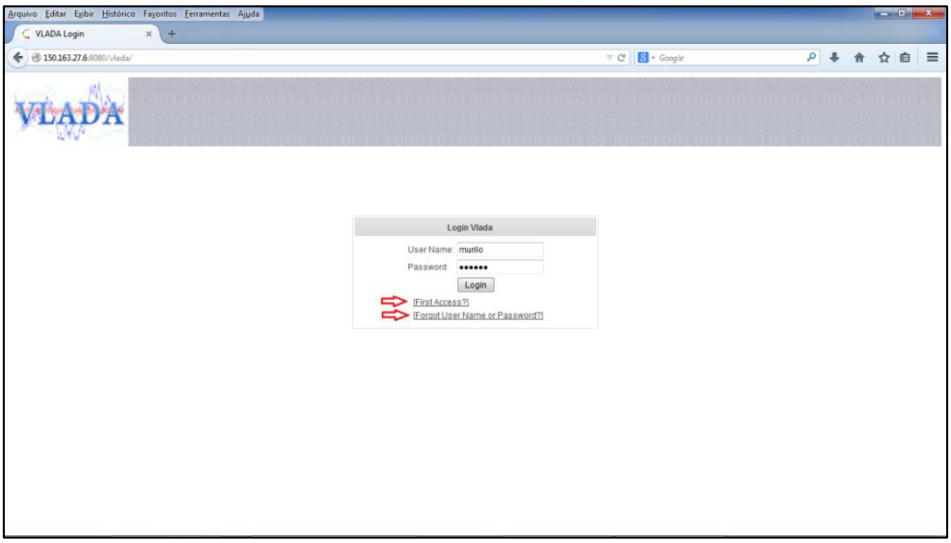


Figura 6.12 – Tela de login

Ao clicar em “Login” no Portal é possível acessar o protótipo do ambiente virtual. Nesta tela é possível efetuar três ações: fazer o login, cadastrar-se no caso de um primeiro acesso e recuperar a senha. Na Figura 6.12 é possível visualizar essas três possibilidades.

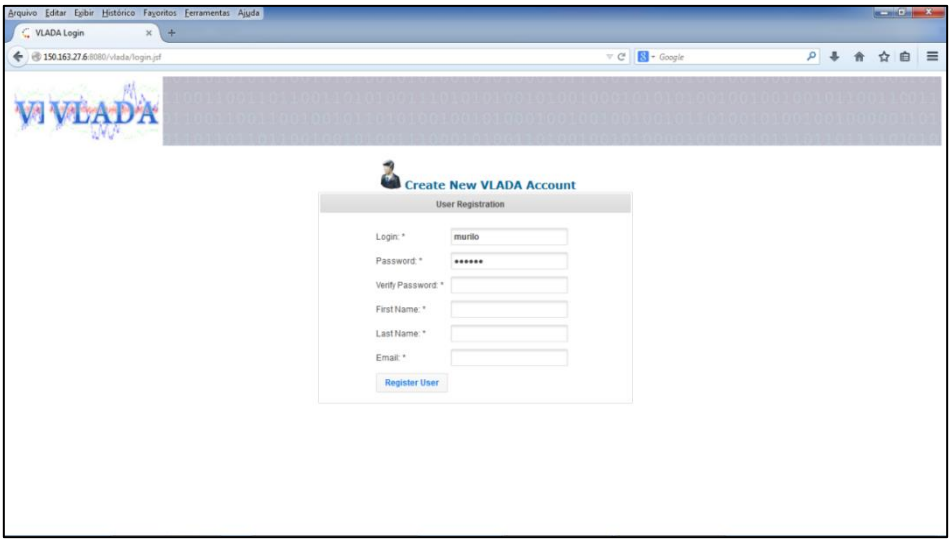


Figura 6.13 – Tela de cadastro para novos usuários

No caso de um primeiro acesso, é possível criar um novo usuário do VLADA, que por padrão terá o perfil de apenas pesquisador para o uso das ferramentas. A Figura 6.13 apresenta o formulário de cadastro ao clicar na opção “First Access” da tela anterior. Já para recuperar a senha basta clicar em “Forgot User Name or Password”. A Figura 6.14 ilustra a tela de como recuperar esses dados. O ambiente confere se o e-mail existe e envia os dados de acesso para o usuário. Isso implica que o e-mail deve ser único no ambiente.

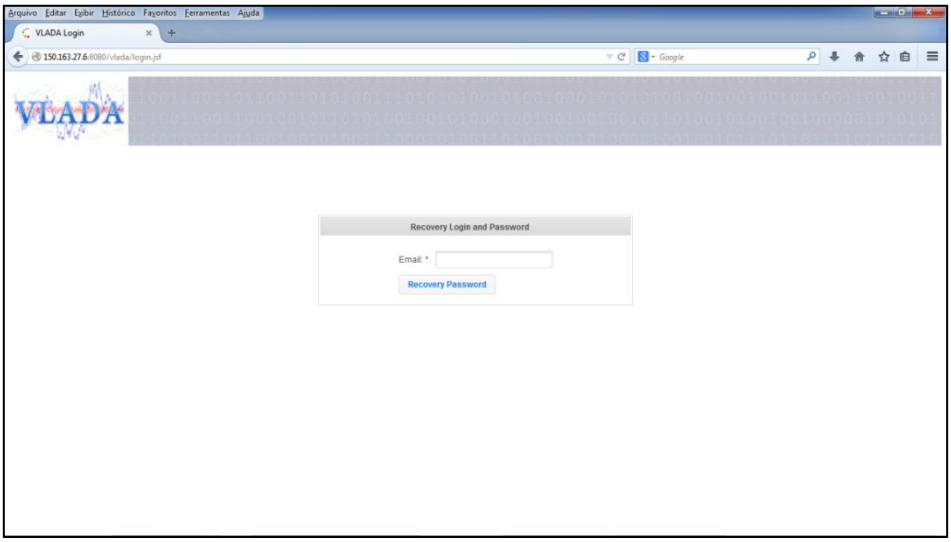


Figura 6.14 – Tela de recuperação de dados de acesso

6.4.2. Usuário do Ambiente

Uma vez acessando o ambiente, o usuário estará usando-o conforme seu perfil de acesso. O usuário pesquisador verá os laboratórios organizados e disponíveis. Para acessar as respectivas ferramentas, basta clicar sobre o bloco correspondente ao laboratório conforme a Figura 6.15.

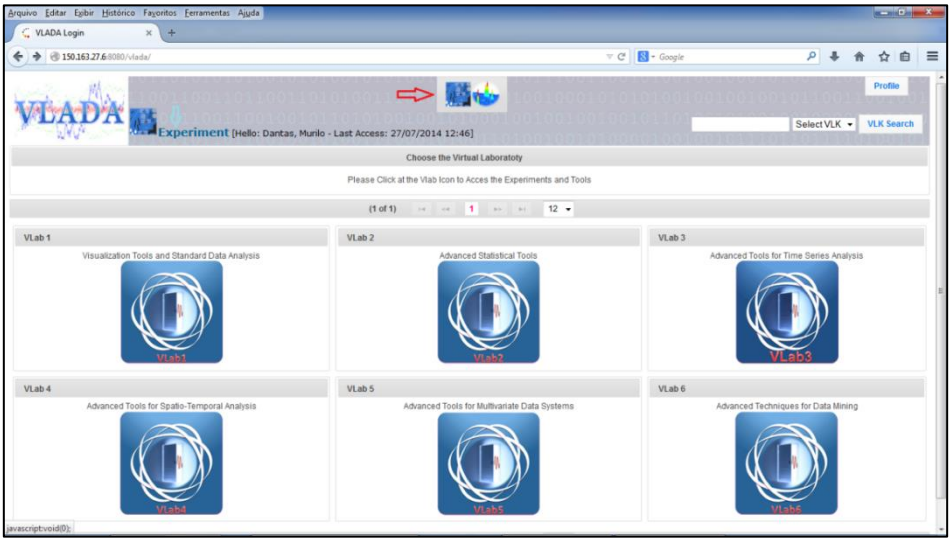


Figura 6.15 – Primeira visualização do usuário pesquisador

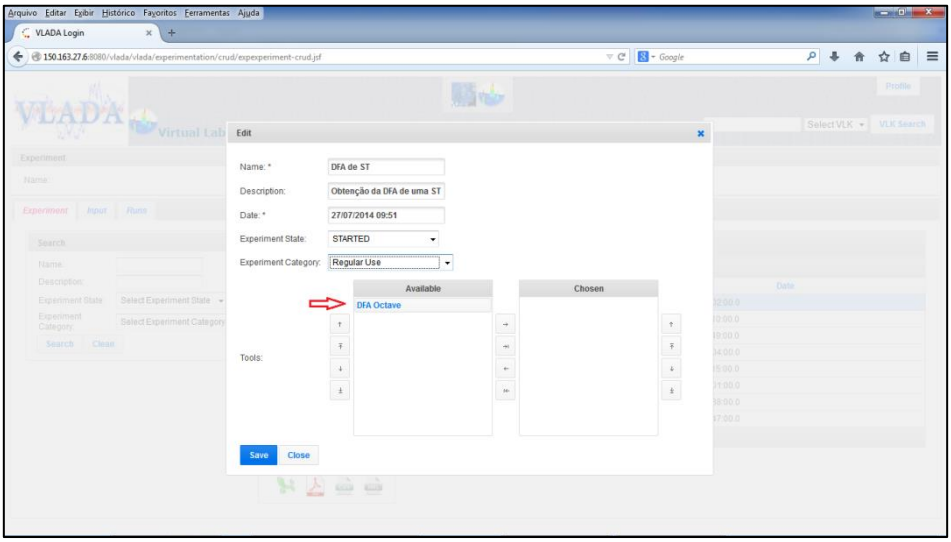


Figura 6.16 – Montagem de um experimento

Na parte superior central na Figura 6.16 é possível o usuário escolher se deseja fazer um experimento através do ícone em destaque, ou se deseja escolher outro VLab. Ao selecionar um VLab, é possível montar um experimento. Há vários campos a serem preenchidos em destaque para a ferramenta disponível para o experimento (algoritmo da DFA desenvolvida em Octave). Veja Figura 6.17.

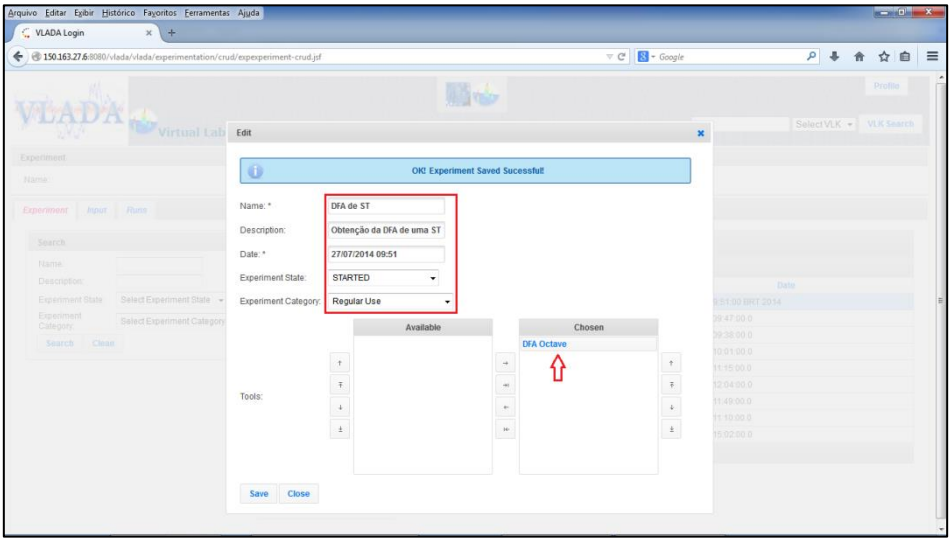


Figura 6.17 – Mensagem de sucesso na montagem de experimento

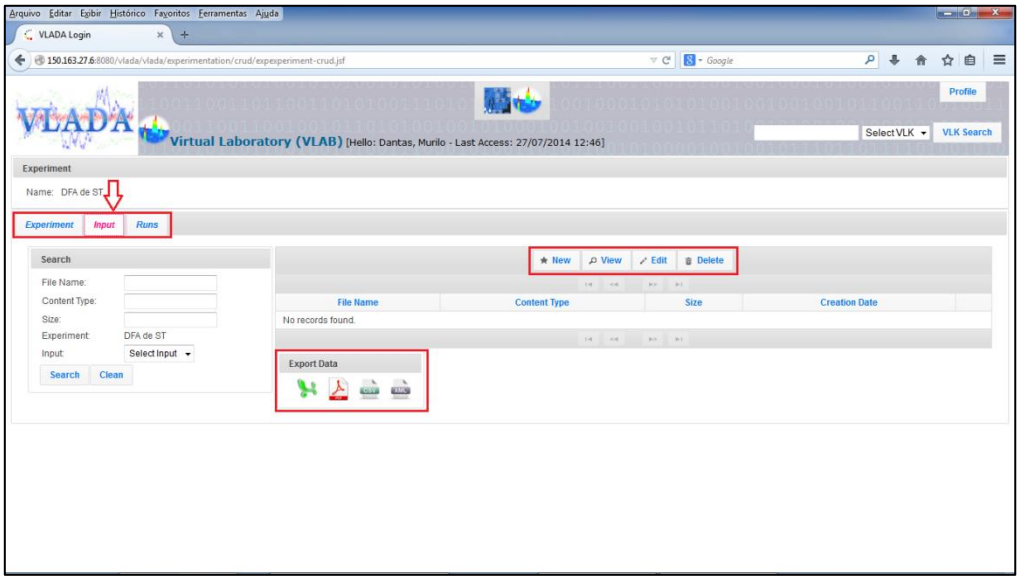


Figura 6.18 – Inserção de dados no experimento para a análise

A Figura 6.17 destaca o sucesso da operação de montagem de experimento que terá um nome, uma descrição, uma data registrada de montagem, um estado e uma categoria selecionada.

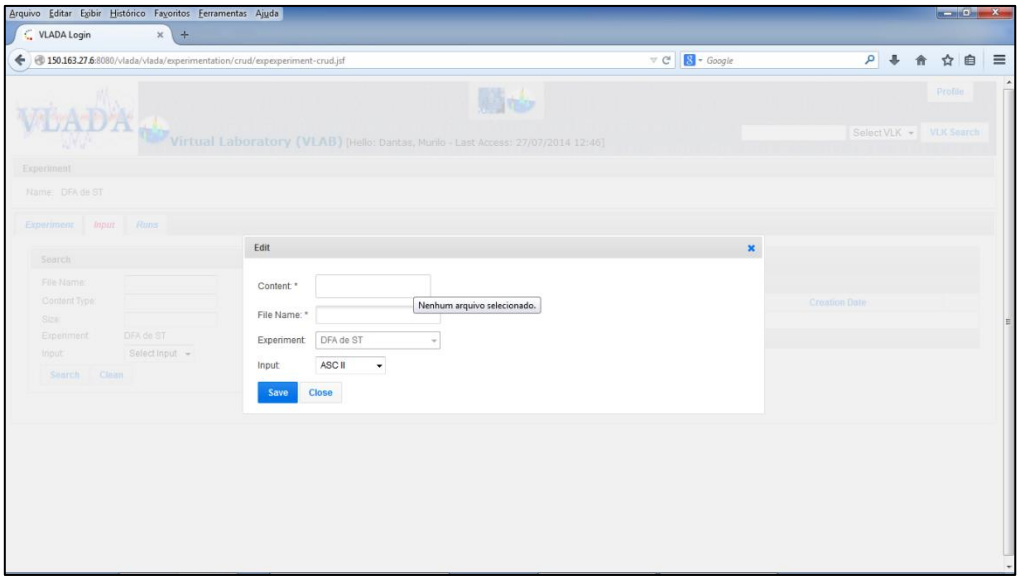


Figura 6.19 – Carregando o arquivo com a ST para a análise

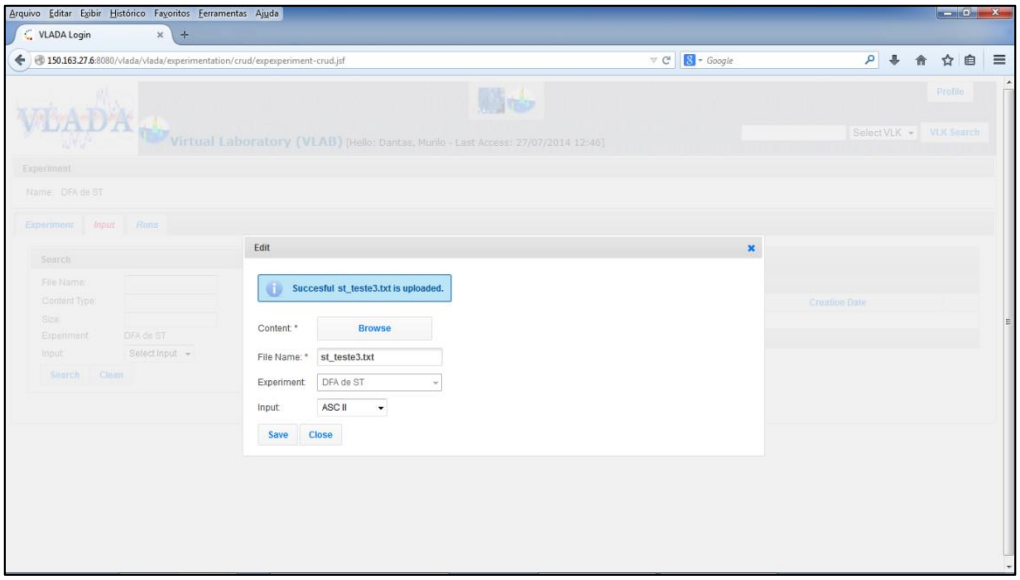


Figura 6.20 – Arquivo contendo ST carregado

Após a criação do experimento, é preciso inserir os dados a serem analisados. A sequência para a análise é, portanto, criar o experimento, inserir os dados para a análise e executar o experimento (Figura 6.18). Na tela de entrada de dados, é possível inserir, visualizar, editar ou apagar um arquivo.

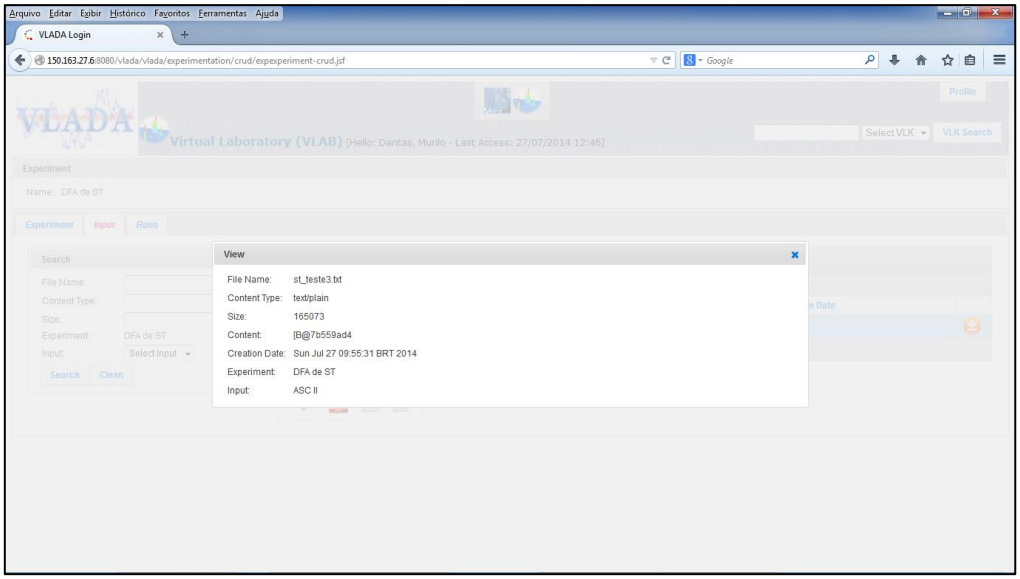


Figura 6.21 – Visualização das informações do arquivo da ST

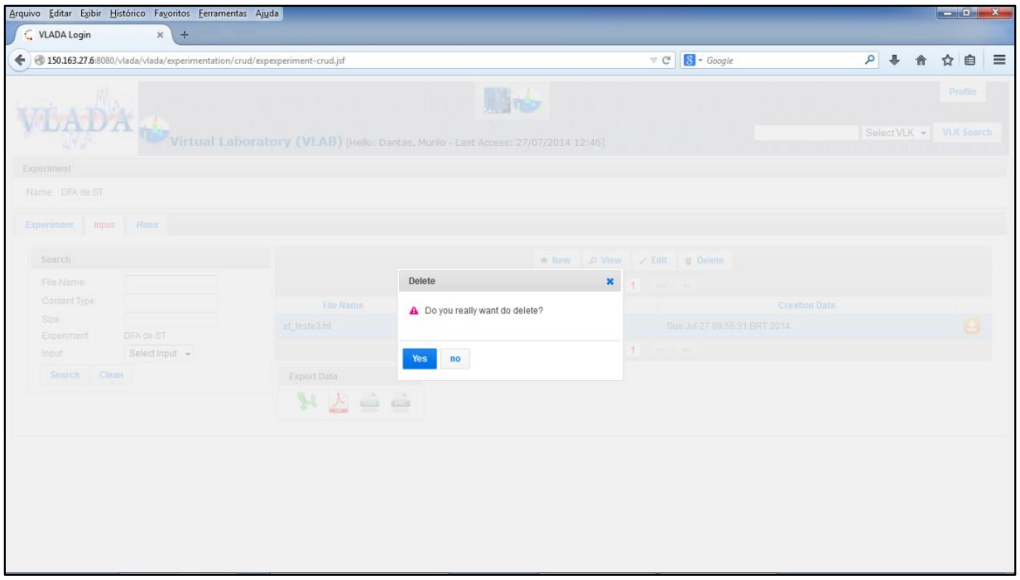


Figura 6.22 – Interação para deleção de arquivo de ST

Em todas as telas está disponível a exportação dos dados nos seguintes formatos: XLS, PDF, CSV e XML. Veja Figura 6.13 para essas manipulações. As Figuras 6.19 e 6.20 mostram o detalhe de inserção do arquivo e as Figuras 6.21 e 6.22 mostram a possibilidade de visualização e de exclusão, respectivamente.

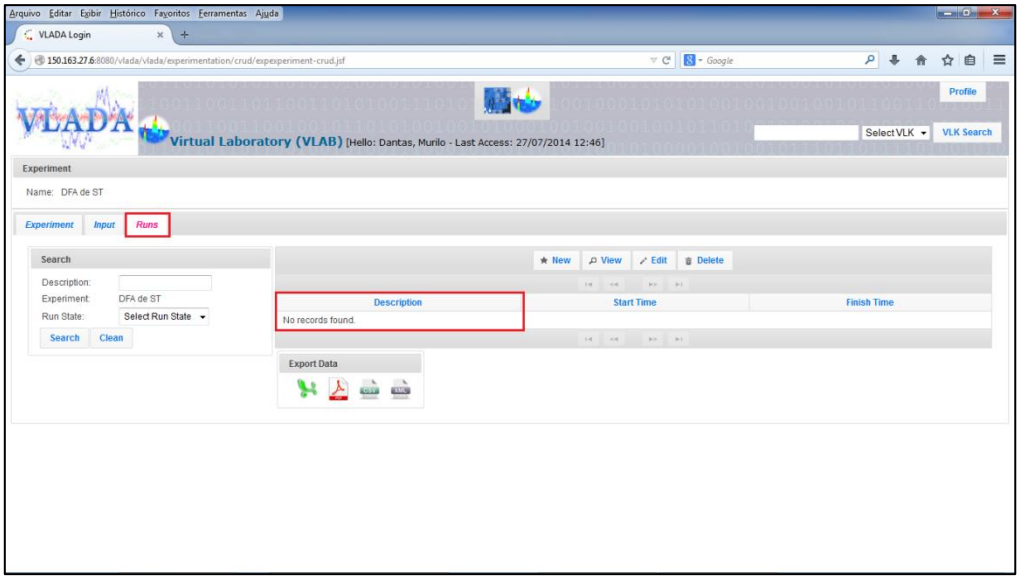


Figura 6.23 – Interface para a criação de execuções do experimento

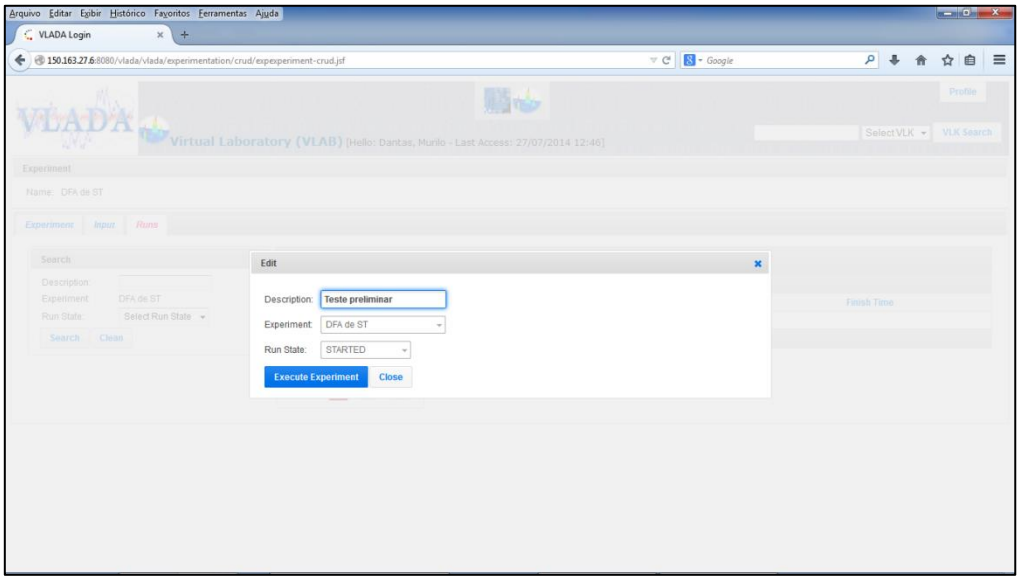


Figura 6.24 – Criação de uma nova execução de um experimento

Finalmente, é possível executar o experimento. Para isso é preciso criar uma instância do mesmo para possíveis análises posteriores. A Figura 6.23 mostra essa interface que é muito similar à de entrada de dados. A Figura 6.24 ilustra a criação de uma nova execução e a Figura 6.25 mostra a execução finalizada com sucesso. É possível visualizar o resultado conforme a Figura 6.26.

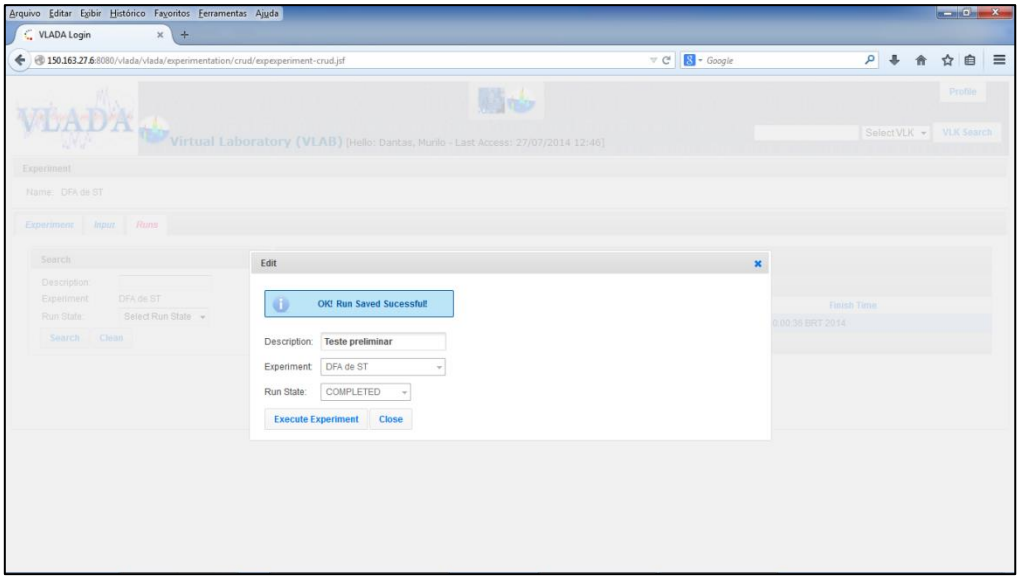


Figura 6.25 – Execução criada com sucesso

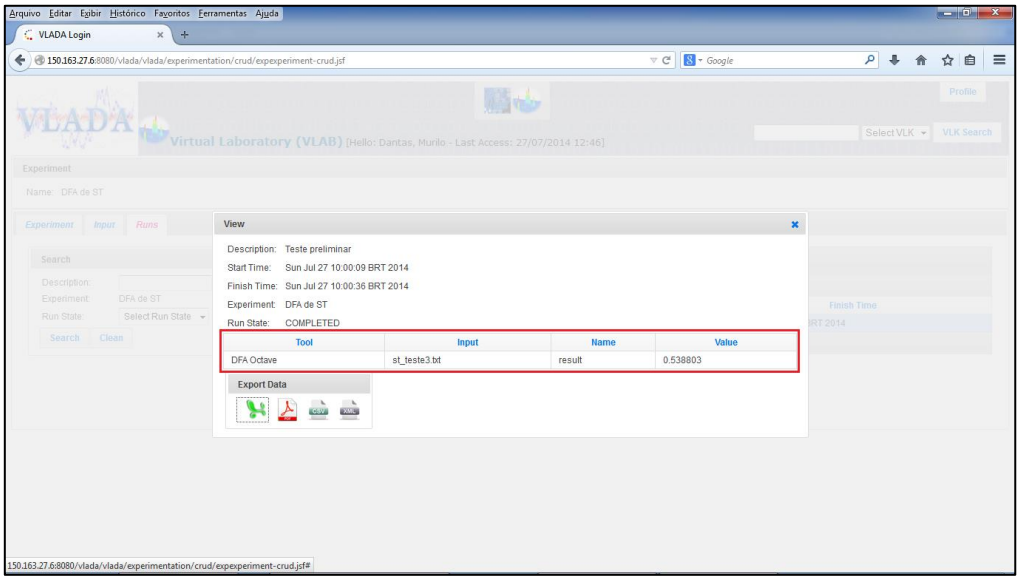


Figura 6.26 – Visualização do resultado da execução do experimento

Conforme mencionado acima, é possível exportar os dados em alguns formatos conhecidos. A Figura 6.27 mostra um exemplo de exportação e a Figura 6.28, mostra o arquivo exportado aberto.

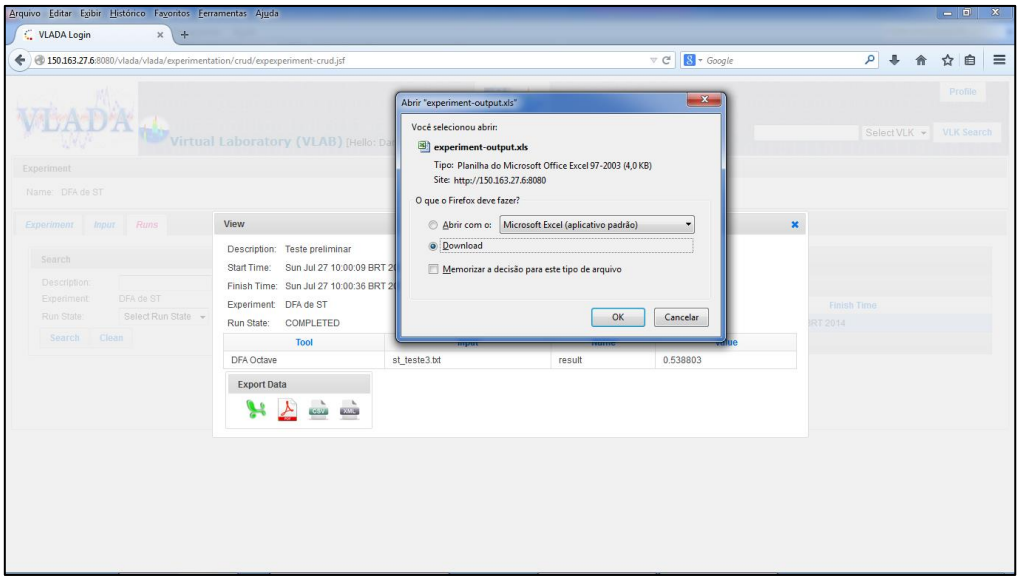


Figura 6.27 – Exemplo de exportação para o formato xls (MS Excel)

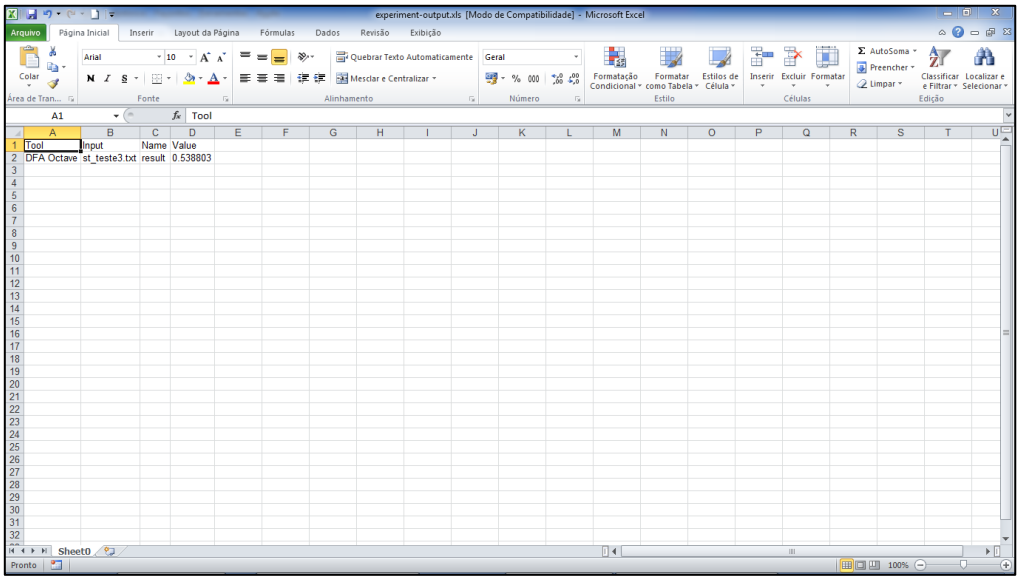


Figura 6.28 – Arquivo exportado para MS Excel

É possível também editar um experimento e incluir mais ferramentas, por exemplo. As Figuras 6.29 a 6.32 mostram que esta ação é possível, colocando em destaque algumas funcionalidades da interface.

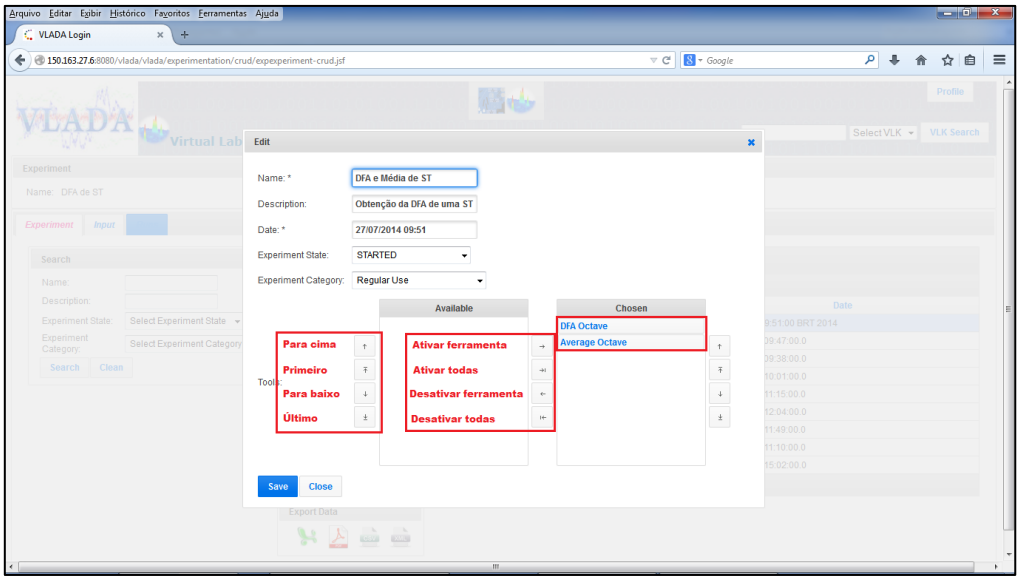


Figura 6.29 – Editando experimento inserindo outra ferramenta

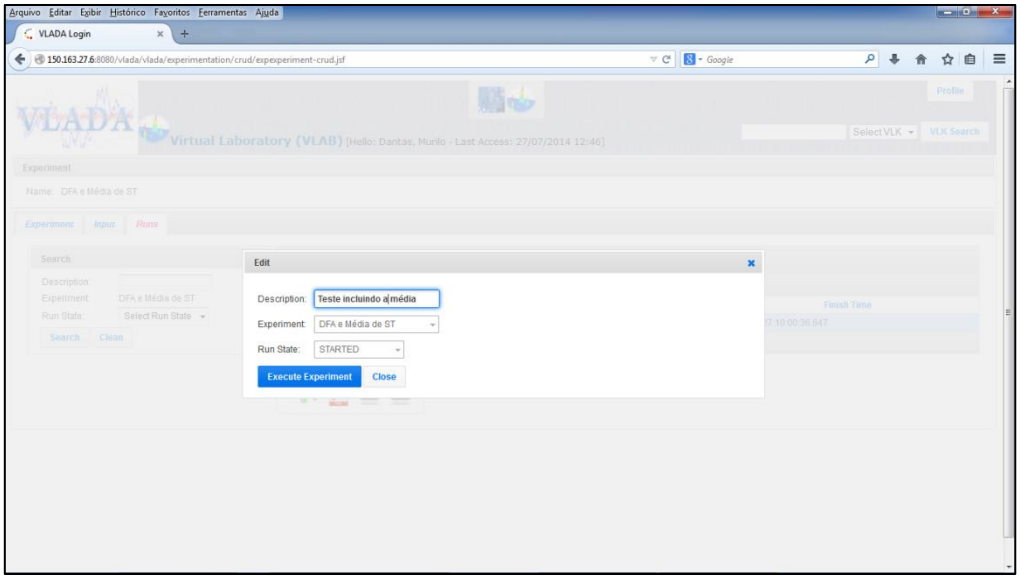


Figura 6.30 – Detalhe da edição do experimento

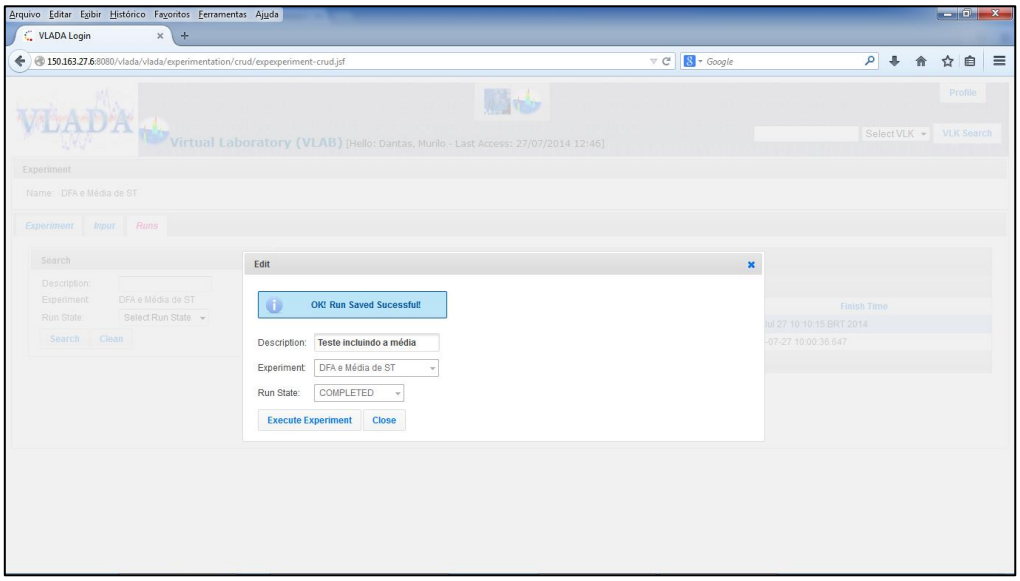


Figura 6.31 – Mensagem de edição finalizada

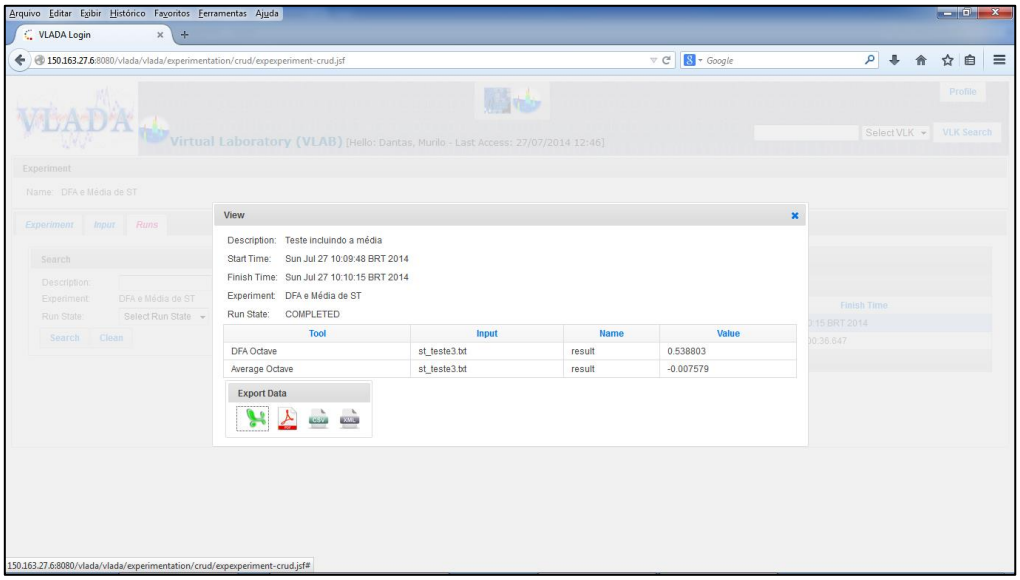


Figura 6.32 – Resultado do experimento editado

6.4.3. Administrador

Conforme exposto no Capítulo 4, o ambiente deve prover diferentes perfis de acesso aos usuários. A Figura 6.33 apresenta esses perfis em destaque.

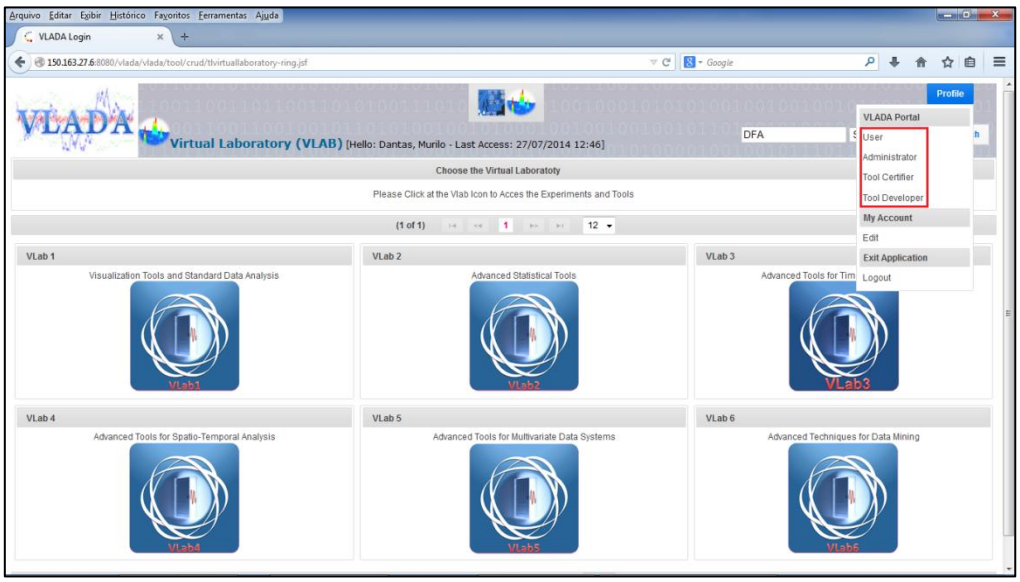


Figura 6.33 – Perfis para ambiente VLADA

Há muitas funções que são inerentes à configuração por parte do administrador (ver Tabela 6.1). A Figura 6.34 mostra a função de categoria de ferramentas para classificação posterior das ferramentas, mas também destaca todas as funções do administrador do ambiente.

Tabela 6.1 – Funções que os administradores do ambiente podem executar

Ícone	Função	Opções atuais
	Gestão de usuário	Formulário com dados e perfil.
	Categoria de ferramenta	Pacote comercial; Pacote livre ou Desenvolvimento em casa.
	Tipo de ferramenta	Estatística regular ou Heurística.
	Estado de ferramenta	Submetido; Certificado ou Rejeitado.
	Tipo de arquivo de entrada	Texto ou Imagem.
	Tipo de arquivo de entrada	Texto; Gráfico ou Imagem.
	Família de ferramentas	DFA ou Momentos estatísticos.
	Categoria de experimentos	Uso regular ou Certificação regular.
	Estado do experimento	Iniciado; Em execução ou Finalizado.
	Estado de execução	Iniciado; Em execução ou Finalizado.
	Estado de certificação	Submetido; Em avaliação; Certificado ou Rejeitado.
	Subtipo de entrada	ASCII; JPEG ou PNG.
	Subtipo de saída	TXT; JPEG ou PNG.
	Natureza da publicação	GNU
	Tipo de publicação	Artigo; Anais; Revista; Código fonte; Tese; Periódico científico ou Livro.
	Categoria de publicação	Análise de Séries Temporais.
	Língua	Inglês; Português; Francês ou Alemão.
	País	Brasil; Inglaterra ou Estados Unidos.
	Licença	Sem licença ou GNU.
	Tipo de divulgação	Meio magnético; Impresso ou Website.
	Laboratório Virtual	Ferramentas de Visualização e Análise de Dados Padrão; Ferramentas de Estatística Avançada; Ferramentas Avançadas para Análise de Séries Temporais; Ferramentas Avançadas para Análise Espaço-Temporal; Ferramentas Avançadas para Sistemas com Dados Multivariados ou Técnicas Avançadas para Mineração de Dados.
	Biblioteca Virtual de Conhecimento	Técnicas de Análises de Dados; Pacotes e Códigos de Dados e Repositório ou Exemplos de Dados.

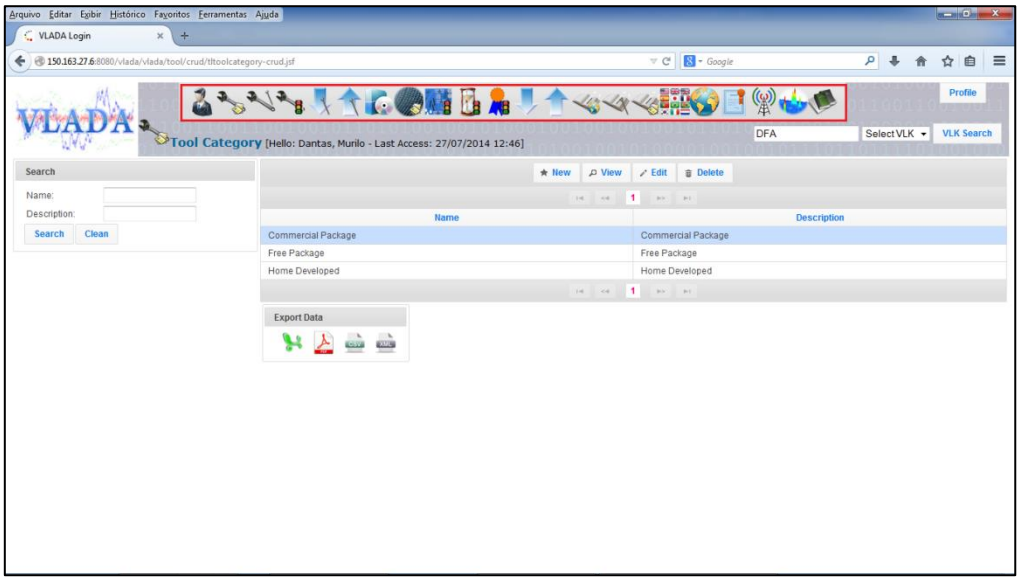


Figura 6.34 – Categoria de ferramentas. Funções do administrador

6.4.4. Desenvolvedor de Ferramenta

O usuário desenvolvedor de ferramentas computacionais deve produzir a sua aplicação fora do ambiente. Toda a documentação científica relacionada deve ser disponibilizada no ambiente através da interface fornecida em seu perfil. A Figura 6.35 mostra o cadastro de publicações.

Conforme visualizado na Figura 6.35, é possível cadastrar vários tipos de publicações relacionadas à ferramenta submetida. A Figura 6.36 mostra a possibilidade de acoplar imagens a cada ferramenta disponibilizada. Já a Figura 6.37 demonstra a possibilidade de disponibilizar documentos relacionados.

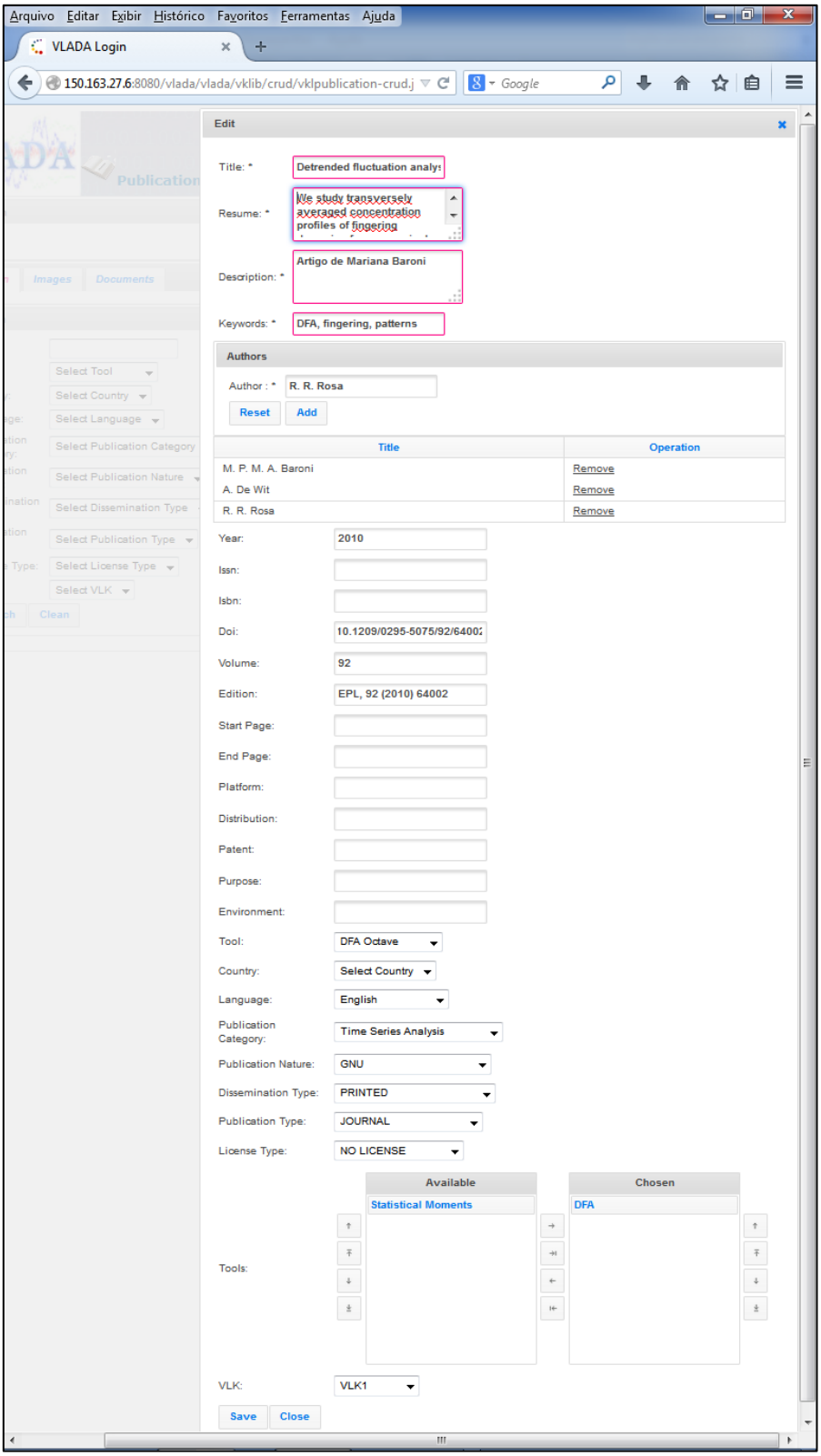


Figura 6.35 – Cadastro de publicação para a ferramenta

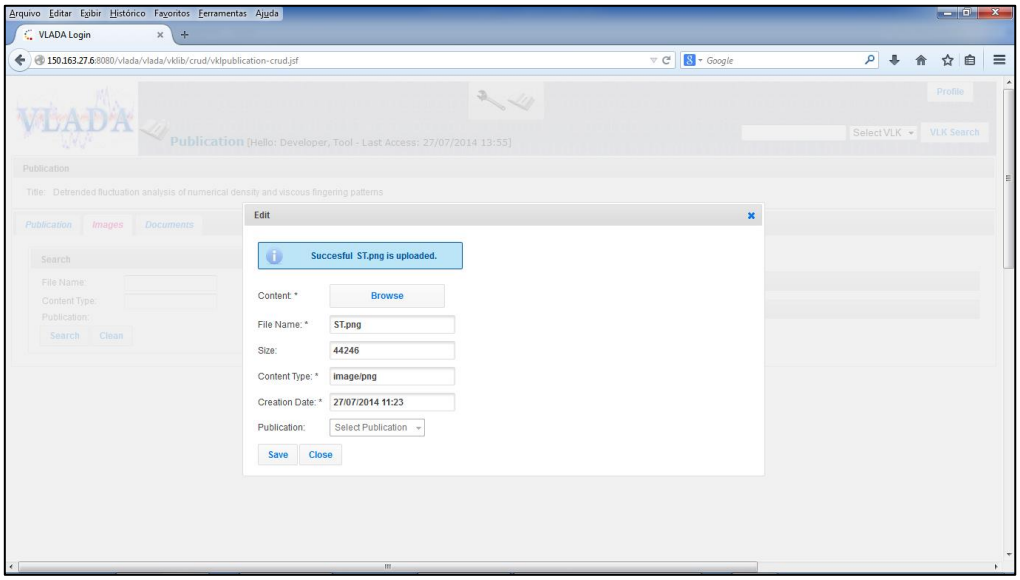


Figura 6.36 – Cadastro de imagens relacionadas à ferramenta

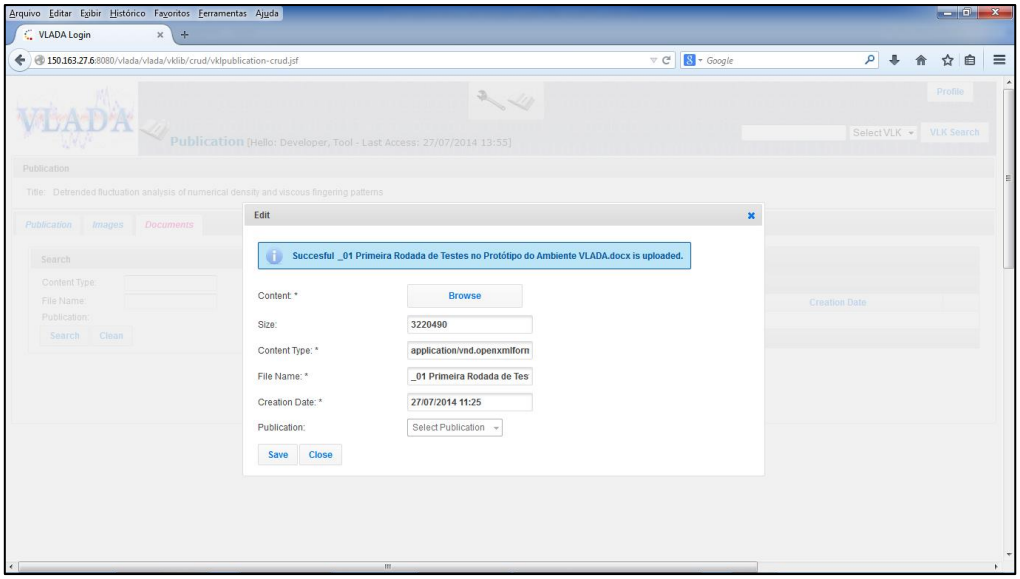


Figura 6.37 – Cadastro de documentos relacionados à ferramenta

A Figura 6.38 demonstra o cadastro de ferramentas que prevê a indicação do serviço da aplicação disponibilizada através de endereço e porta da aplicação.

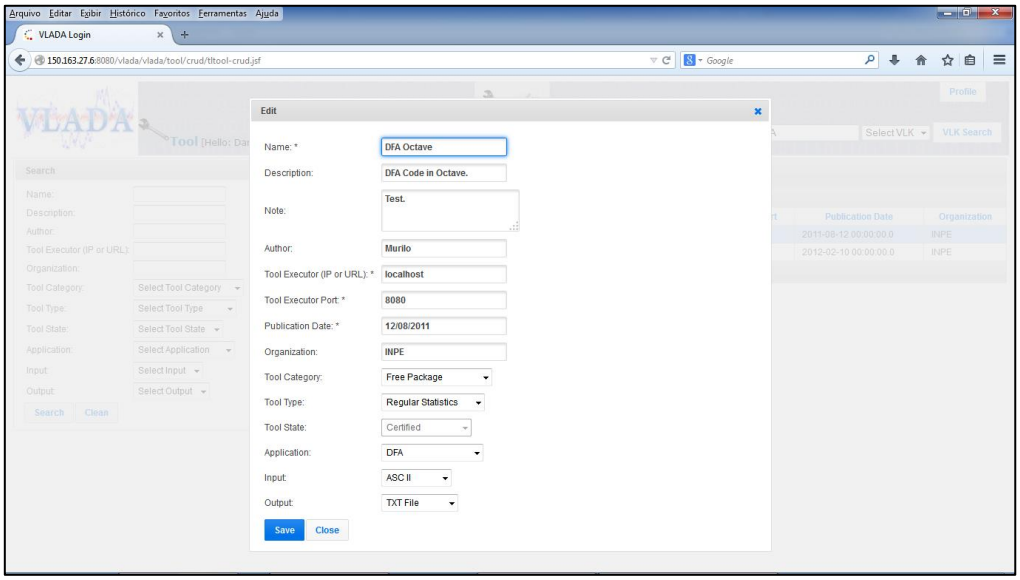


Figura 6.38 – Cadastro de ferramenta

6.4.5. Certificador de Ferramenta

Conforme a especificação no Capítulo 4, quando cadastrada no ambiente, a ferramenta computacional de análise precisa passar por um processo de certificação que a validará científica e algoritmicamente através da documentação cadastrada. Os estados possíveis de uma ferramenta genérica de análise, desde o seu carregamento inicial para a certificação até a publicação que permite o uso, são mostrados como um diagrama de estado na Figura 6.39. A disponibilização da documentação juntamente com a ferramenta é obrigatória e esta análise é efetuada pelo usuário certificador de ferramentas. Este usuário é um pesquisador especialista da área científica da ferramenta capaz de julgar com seus pares sua relevância e exatidão algorítmica.

A ferramenta cadastrada ficará disponível para um ou mais certificadores de ferramentas designados pela Equipe de Análise de Dados e Algoritmos. Nesta interface, o usuário certificador poderá executar experimentos de testes na ferramenta até finalizar o seu parecer (veja Figuras 6.40 e 6.41). Durante todo o processo, a ferramenta não poderá ser utilizada pelos demais usuários do ambiente. O usuário desenvolvedor interagirá com o certificador dirimindo

eventuais dúvidas e efetuando correções solicitadas. Esta interação ocorrerá por mensagens de e-mail.

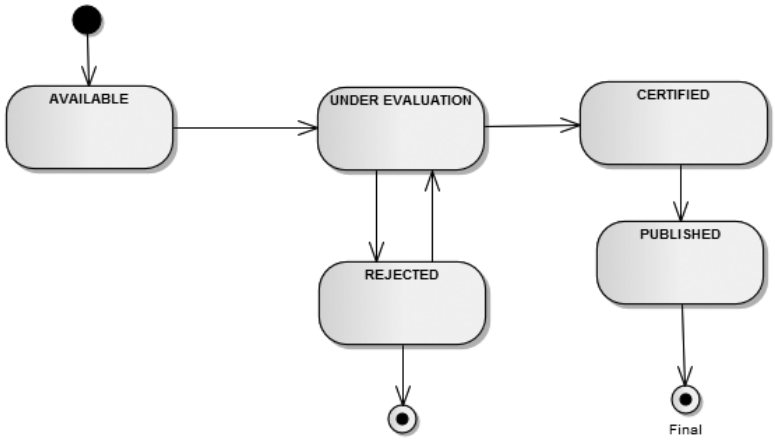


Figura 6.39 – Diagrama de estado para disponibilização de ferramentas

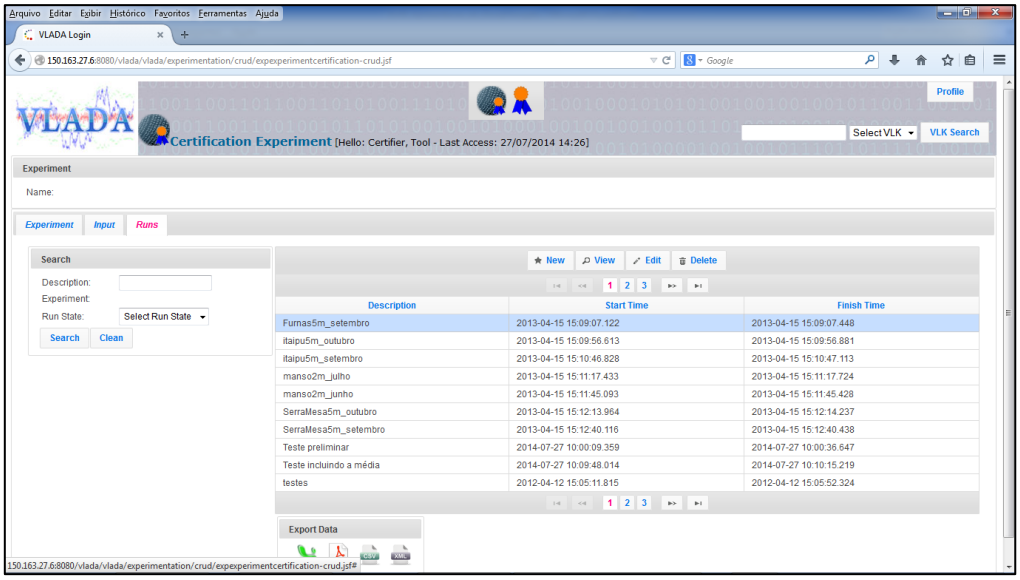


Figura 6.40 – Acesso aos experimentos no processo de certificação

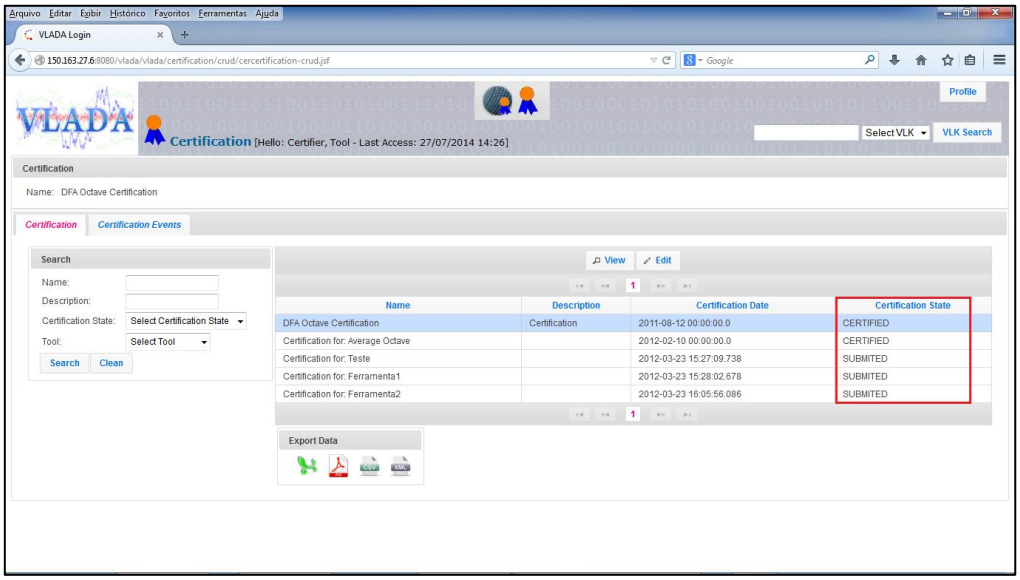


Figura 6.41 – Interface expressando o estado de certificação de uma ferramenta

6.4.6. Acesso via Dispositivos Móveis

O protótipo VLADA também cumpriu o requisito da mobilidade ao fornecer uma interface usual acessível através de dispositivos móveis (Figuras 6.42 a 6.50). Essa capacidade libera o pesquisador de usar ferramentas a partir de computadores de mesa com alto poder de processamento, além da necessidade de licenças de alto valor econômico para a análise de dados.

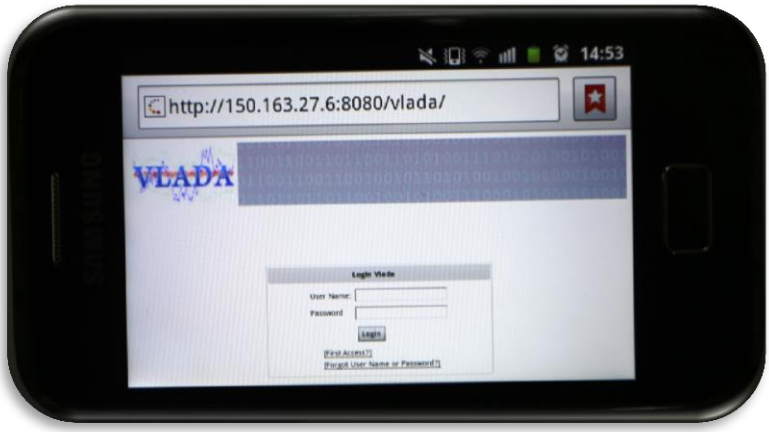


Figura 6.42 – Acesso ao protótipo através de um dispositivo móvel

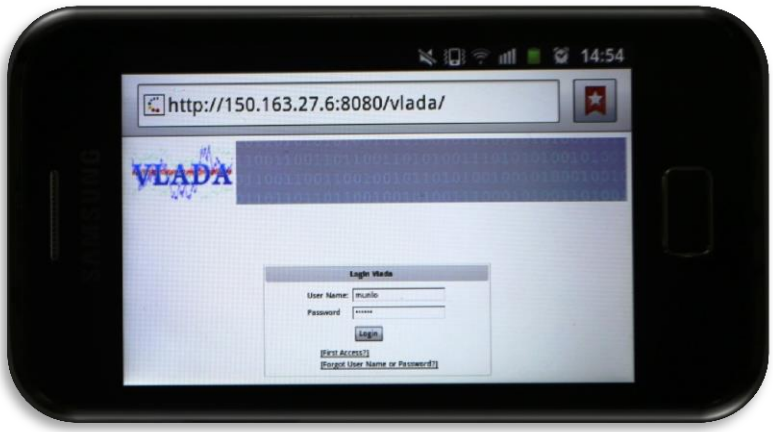


Figura 6.43 – Uso de login e senha no acesso ao VLADA

Figura 6.44 – Disponibilização do ambiente após o login

Figura 6.45 – Criação de um experimento através do celular

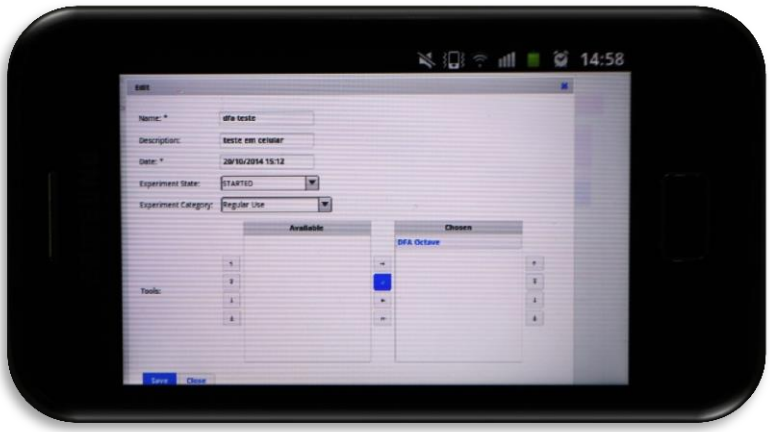


Figura 6.46 – Especificação do experimento associando com uma ferramenta

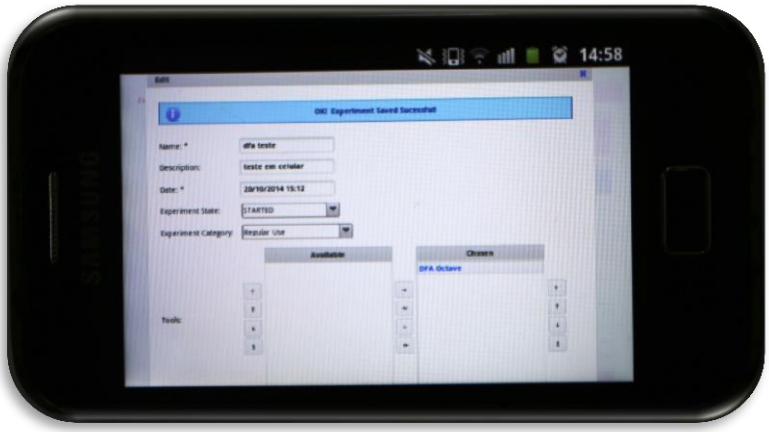


Figura 6.47 – Finalização da configuração do experimento no celular

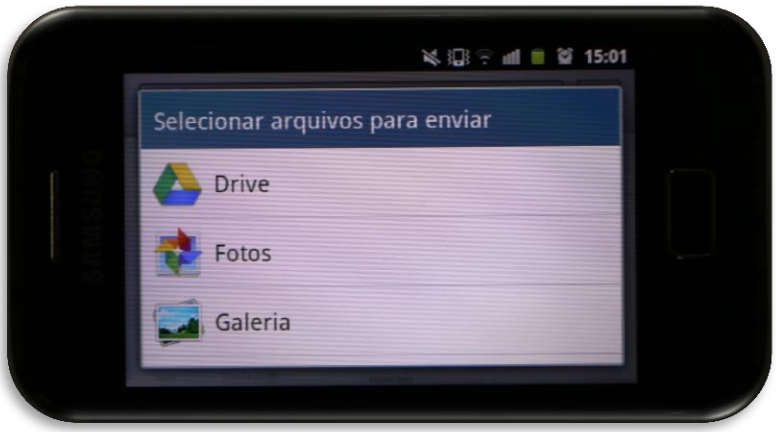


Figura 6.48 – Escolha do arquivo contendo a ST a ser analisada

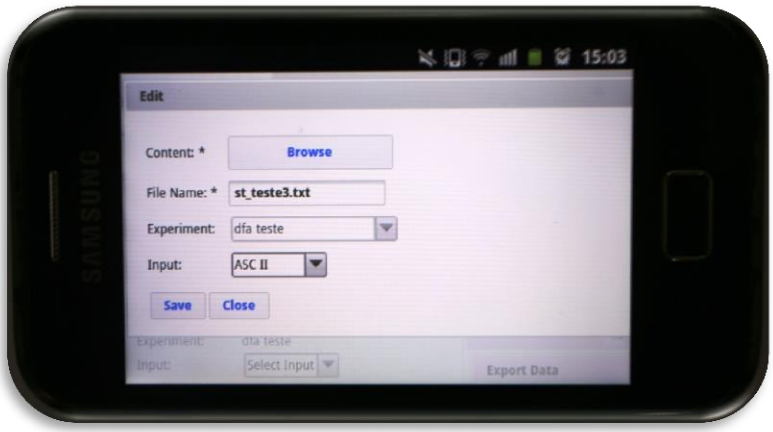


Figura 6.49 – Upload da ST a ser analisada pelo VLADA via celular

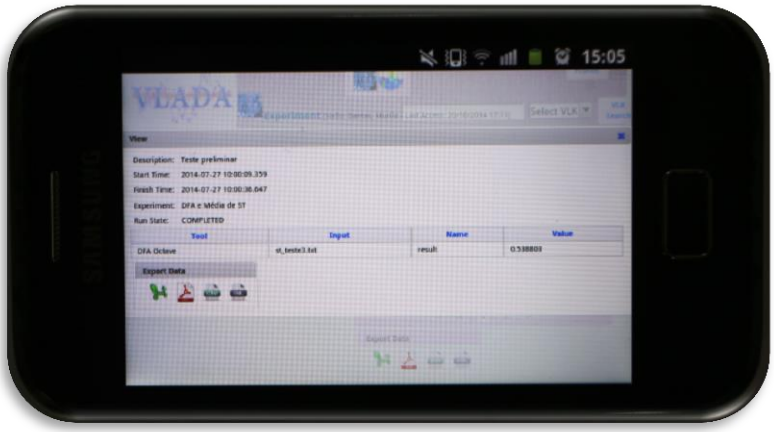


Figura 6.50 – Apresentação do resultado da análise no experimento

Através desta interface, o pesquisador pode inserir sua série temporal a ser analisada a partir de qualquer lugar que permita a conexão do dispositivo móvel e portátil à rede. Isso reitera o paradigma de análise de dados usando ferramentas matemáticas, estatísticas e computacionais avançadas disponibilizadas na nuvem. O usuário poderá então analisar rapidamente seus dados e gerar resultados no ambiente virtual.

6.4.7. Atendimento às características desejáveis

No capítulo introdutório foi abordada uma série de características desejáveis em um ambiente virtual colaborativo para análise avançada de séries

temporais. Tais características estão listadas na Tabela 1.2. A Tabela 6.2 a seguir mostra o que já é oferecido no protótipo VLADA.

Tabela 6.2 – Características desejáveis em um ambiente virtual colaborativo robusto para a análise avançada de séries temporais contempladas no VLADA

Identificação	Característica	VLADA
1	Interface	X
2	Multiusuário	X
3	Customização	X
4	Ferramentas	X
5	Escalabilidade	X
6	Controle	X
7	Usabilidade	X
8	Padronização	X
9	Integração	X
10	Duração	-
11	Escalonamento	Em parte
12	Validação	Em parte
13	Baixo custo	X
14	Dados	X
15	Distribuição	-
16	Mobilidade	X
17	Cliente leve	X
18	SOA	X
19	*Web*	*X*
20	Multilinguagem	X
21	Documentação	X

Praticamente todas as características estão presentes no VLADA. Entretanto, quatro delas não estão plenamente desenvolvidas. A característica da Duração (item 10) e a do Escalonamento (item 11) exigem o uso de ferramentas de orquestração com coordenação centralizada usando linguagens e sistemas específicos para isso. Estas características certamente serão contempladas numa nova fase de desenvolvimento do protótipo.

A característica da Validação (item 12) foi de certo modo prejudicada devido às políticas de segurança do INPE que impediram o uso pleno da funcionalidade de enviar e-mail. Todo o processo de validação está sustentado na comunicação entre o usuário Desenvolvedor e os Certificadores de sua ferramenta. Com o protótipo se tornando operacional internamente no INPE,

mais pessoas estarão utilizando e testando o ambiente. Certamente essa característica será contemplada numa nova fase do desenvolvimento do protótipo.

Finalmente, a característica da Distribuição (item 15) também possui aspectos dependentes do conceito de orquestração a ser incorporado ao ambiente, além do uso e divulgação do próprio ambiente como uma ferramenta para a pesquisa computacional científica colaborativa.

7 ANÁLISES USANDO O PROTÓTIPO

Durante o desenvolvimento desta tese foi possível executar testes relacionados ao tema. Portanto, este capítulo complementa o capítulo anterior na exposição dos resultados desta tese, apresentando algumas análises executadas já no protótipo do ambiente. Sendo assim, na primeira parte do capítulo é apresentada uma análise de séries temporais geradas artificialmente por computador na tentativa de classificação conforme sua dinâmica de flutuação. Já na segunda parte, são apresentados alguns estudos de caso na análise de séries temporais.

7.1. Análise de STC's Artificiais

7.1.1. Geração das Séries Temporais

Conforme discussão apresentada no Capítulo 3, as técnicas de análise tidas como avançadas neste trabalho são aquelas que consideram a não linearidade e a não estacionariedade da série temporal, além de se mostrar estável na análise de séries curtas. Tais técnicas são muito importantes para a análise de ST’s, pois boa parte dos dados reais coletados está no formato de séries temporais curtas, com poucos dados.

É possível comparar a estabilidade de diferentes técnicas variando o tamanho da série temporal e avaliando a variabilidade de uma de suas características marcantes. Para que isso fosse executado de maneira controlada, foram geradas séries artificiais a partir de técnicas recentemente discutidas na literatura científica. Apesar de essas ST’s terem sido geradas artificialmente, a inspiração para a discussão dos problemas vem de dados reais, que são aqueles coletados na natureza, sendo os de maior interesse para a análise avançada no contexto do INPE.

Neste trabalho foi comparada a estabilidade da lei de potência obtida pela PSD (técnica convencional) e a do expoente de escala obtido pela DFA (técnica

avançada). Antes de apresentar o resultado das análises é preciso demonstrar o processo de geração das séries artificiais.

Processos estocásticos estão associados ao comportamento de variáveis aleatórias cuja dinâmica é regida por uma grande quantidade de graus de liberdade e, portanto, são chamados de processos de alta dimensão.

O espectro de um processo estocástico pode ser representado por uma série discreta de frequências e a sua função aleatória correspondente pode ser escrita como uma série discreta para tempos $t_i = i\Delta t$ com $i = 1, 2, \dots, M$, permitindo que a ST possa ser escrita como uma simples superposição de oscilações harmônicas, dada por (OSBORNE e PROVENZALE, 1989):

$$A(t_i) = \sum_{k=1}^{M/2} [P(\omega_k)\Delta\omega]^{1/2} \cos(\omega_k t_i + \emptyset_k), i = 1, \dots, M, \quad (7.1)$$

onde $\omega_k = k\Delta\omega$, com $k = 1, \dots, M/2$, $\Delta\omega = 2\pi/Mt$, $\emptyset_k$ sendo as fases escolhidas aleatoriamente e M sendo o número de pontos da ST.

Uma ST estocástica pode ser obtida a partir da Equação 7.1, utilizando o algoritmo de MALAMUD e TURCOTE (1999) com um valor específico de β. No caso de $\beta = \frac{5}{3}$, tem-se uma lei de potência típica de padrões estocásticos de alta dimensão. Na Equação 7.1, $P(\omega) = 1/\omega^{\beta}$.

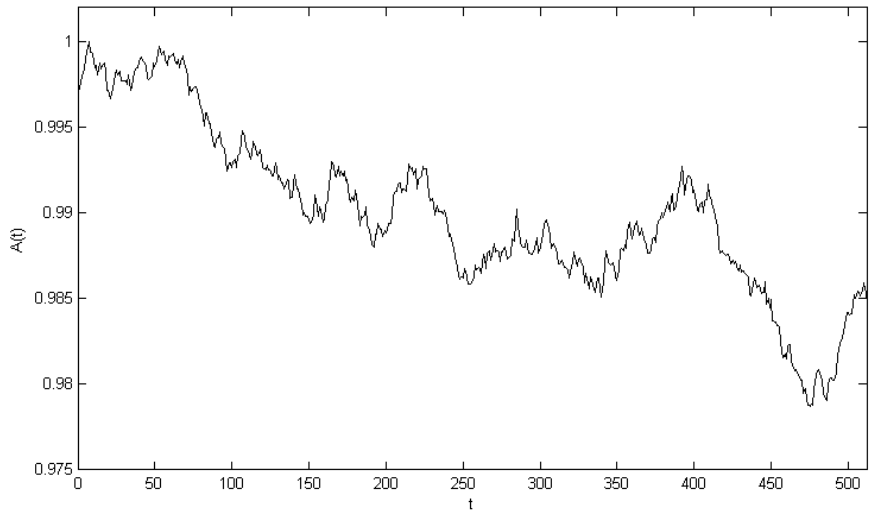


Figura 7.1 – Exemplo de STC estocástica com 512 pontos: $\boldsymbol{P(\omega) \sim \omega^{-5/3}}$

O conceito de lei de potência, associado a um espectro de energias, permite caracterizar um padrão de variabilidade complexa, isto é, cujas densidades espectrais são proporcionais a $\omega^{-\beta}$, com β assumindo diferentes valores a partir do tipo de processo considerado (por exemplo, turbulência), num intervalo de algumas ordens de grandeza da frequência. Sinais gerados por processos estocásticos do tipo $\omega^{-\beta}$ são encontrados na física, meteorologia, biologia, engenharia, economia, etc. Na Figura 7.1 é possível visualizar uma ST deste tipo com 512 pontos e $\beta = \frac{5}{3}$. Esta série representa um exemplo de série não estacionária, conforme definição no Capítulo 3.

De forma geral, o comportamento destes sinais está associado com correlações que decaem lentamente, e o interesse por estes processos tem sido recentemente renovado pelo desenvolvimento das fenomenologias caóticas e estocásticas, nas quais um comportamento aparentemente estocástico pode resultar de um mecanismo determinístico caótico com dependência de longo alcance. De outra perspectiva, a lei de potência dos espectros indica que um sinal existe em todas as escalas e, portanto, não tem uma escala característica, resultando num aspecto de auto similaridade no contexto estocástico (ABRY et al., 1995).

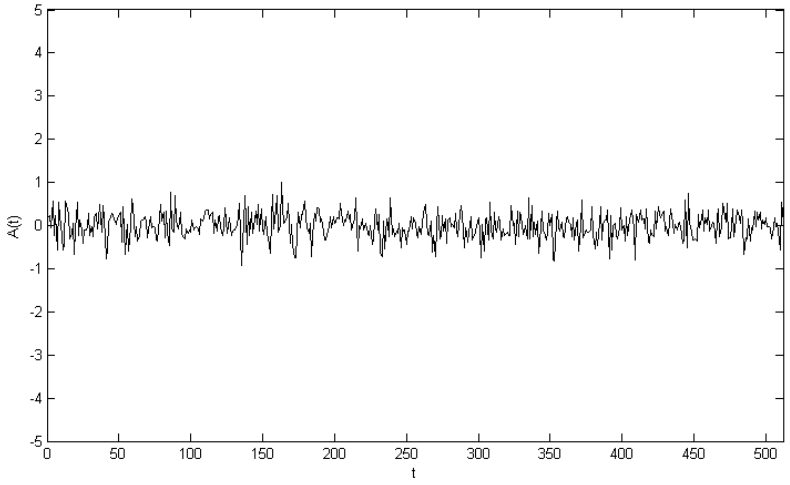


Figura 7.2 – Exemplo de STC pseudoaleatória

Um processo físico comum que pode ser simulado a partir da Equação 7.1 é o movimento fracionário Browniano (fBm). O fBm foi introduzido por MANDELBROT e VAN NESS (1968) como um meio de representar processos estocásticos não estacionários que exibem dependência de longo alcance e/ou que têm propriedade de auto similaridade. Outra série utilizada neste trabalho é uma ST estacionária caracterizada como um ruído pseudoaleatório. Essa ST pode ser visualizada na Figura 7.2.

Através das ST's apresentadas nas Figuras 7.1 e 7.2 podemos exemplificar medidas que caracterizam tais séries. Na Tabela 7.1 abaixo, temos os valores para os momentos estatísticos dessas séries.

Tabela 7.1 – Momentos estatísticos das ST's estocástica e pseudoaleatória

Momento	**ST estocástica ($\boldsymbol{\beta = \frac{5}{3}}$)**	**ST pseudoaleatória**
Média	0,990000	-0,005900
Variância	0,000024	0,094300
Assimetria	0,098800	-0,116600
Curtose	2,700100	3,106400

7.1.2. Estabilidade das Técnicas

Bruce Malamud e Donald Turcotte, realizando estudos de persistência em ST (MALAMUD e TURCOTTE, 1999), concluíram que para sinais não estacionários, a robustez das medidas estatísticas do sinal começa a ser comprometida com a diminuição do número de pontos da amostra.

Para efeitos de demonstração desta perda de robustez da análise estatística e consequente ineficiência dos métodos convencionais de análise, a PSD foi usada para classificar STC. Para isso, foram geradas STs estocásticas com 2^{17} pontos (para visualização do padrão, ver Figura 7.3), com $\beta = \frac{5}{3}$, também através do algoritmo de Bruce Malamud e Donald Turcotte (MALAMUD e TURCOTTE, 1999), a partir da Equação 7.1. Calculado o valor de β, a série foi dividida ao meio. Em seguida, o valor de β foi calculado para a primeira metade

da sequência anterior. Essas operações foram realizadas iterativamente até a ST possuir 128 pontos. O valor de β foi calculado para um conjunto de mil ST do mesmo tipo de sinal ($\beta = \frac{5}{3}$), a partir do qual foi calculado o valor médio de β e seu desvio padrão, conforme Tabela 7.2.

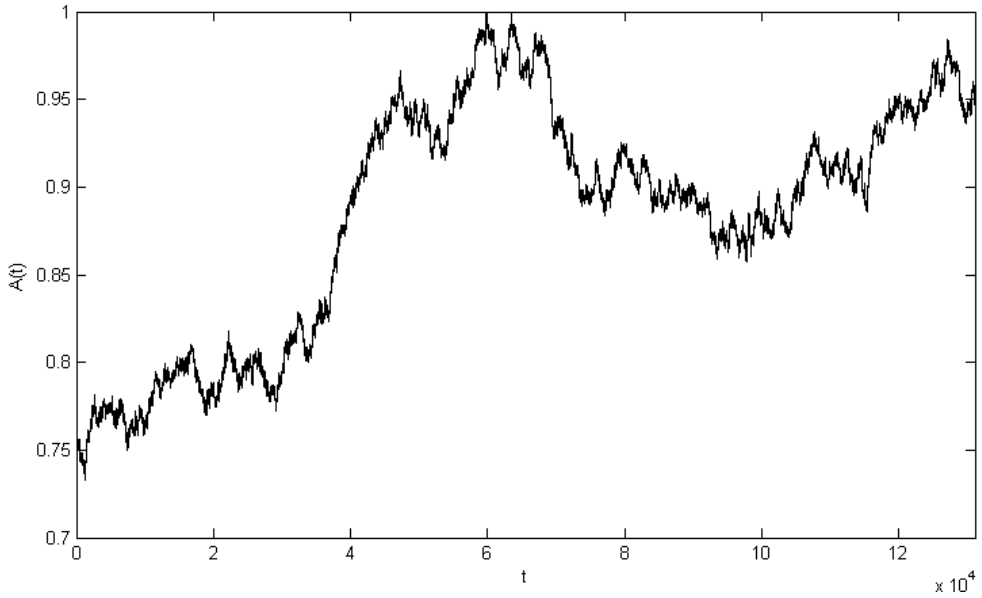


Figura 7.3 – ST estocástica com 2^{17} pontos e $\boldsymbol{\beta = \frac{5}{3}}$

Tabela 7.2 – Valores de $\boldsymbol{\beta}$ relacionados com o tamanho da série

Tamanho da série	$< \boldsymbol{\beta} >$	**Desvio padrão**
131072	1,68	0,007
65536	1,68	0,014
32768	1,68	0,028
16384	1,68	0,061
8192	1,76	0,109
4096	1,79	0,230
2048	1,75	0,451
1024	1,86	0,904
512	1,81	1,812
256	2,86	2,213
128	2,36	3,291

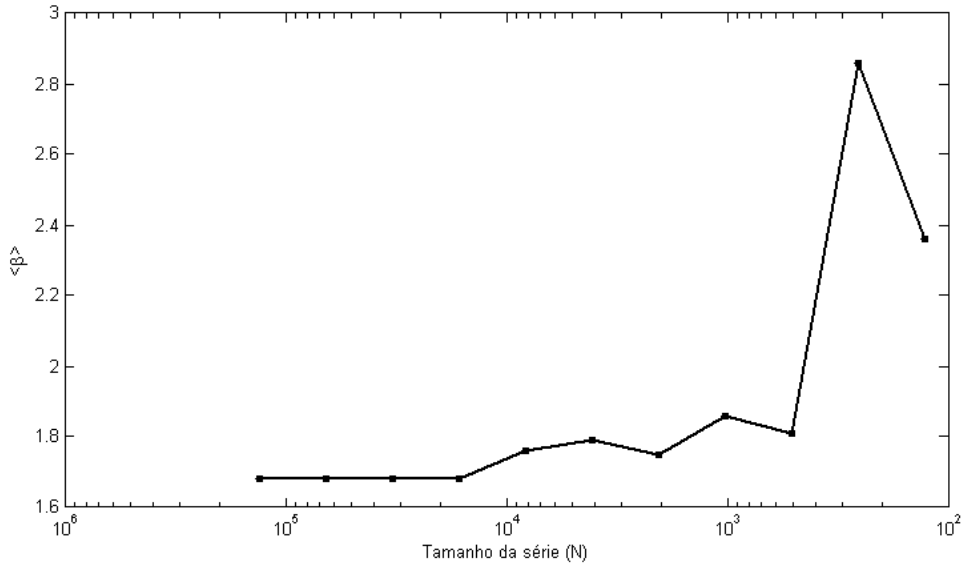


Figura 7.4 – Relação entre o tamanho da série e o valor de $< \boldsymbol{\beta} >$

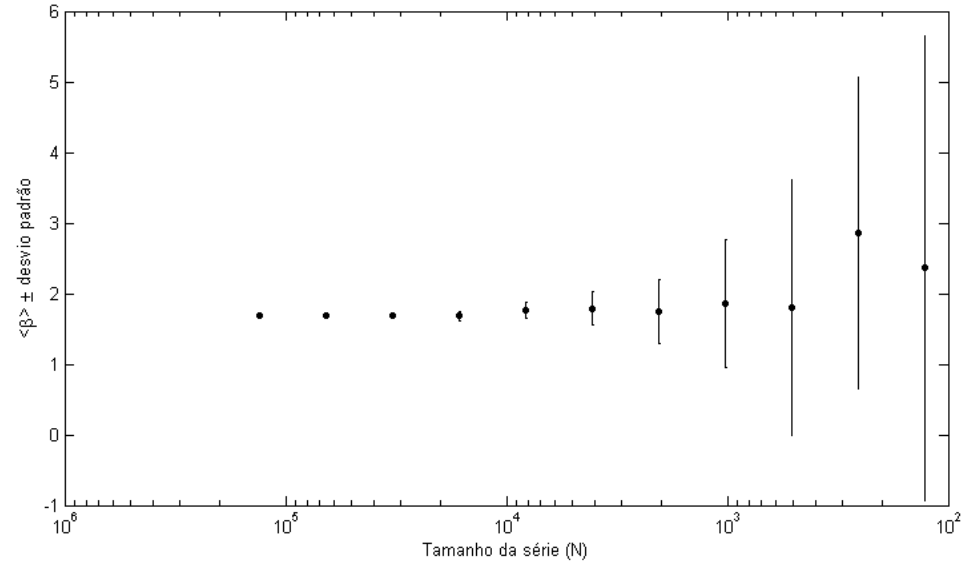


Figura 7.5 – Erros no cálculo de $\boldsymbol{\beta}$. Para ST menores, $\boldsymbol{\beta}$ com maior variação

Para a obtenção de $< \beta >$ foi considerado o uso de mil ST's, pois de acordo com DANTAS (2011), é uma quantidade razoável conforme critério de convergência abordado neste trabalho para padrões estocásticos.

Como pode ser visto na Tabela 7.2 e nas Figuras 7.4 e 7.5, a confiabilidade do valor de β está ligada diretamente com o tamanho do sinal devido à dependência estrita da média da autocorrelação. Como a média de uma ST diverge quando seu tamanho é da ordem de até 10^4, consequentemente sua autocorrelação e sua PSD vão variar excessivamente.

O mesmo teste foi realizado com a DFA para a demonstração de sua estabilidade em classificar STC's. Portanto, foram usadas as mesmas STs estocásticas com 2^{17} pontos, com $\beta = \frac{5}{3}$. Em seguida, o valor de α foi calculado iterativamente para a primeira metade da série anterior. O valor de α foi calculado para um conjunto de mil ST do mesmo tipo de ST ($\beta = \frac{5}{3}$), a partir do qual foi obtido o valor médio de α e seu desvio padrão, conforme Tabela 7.3.

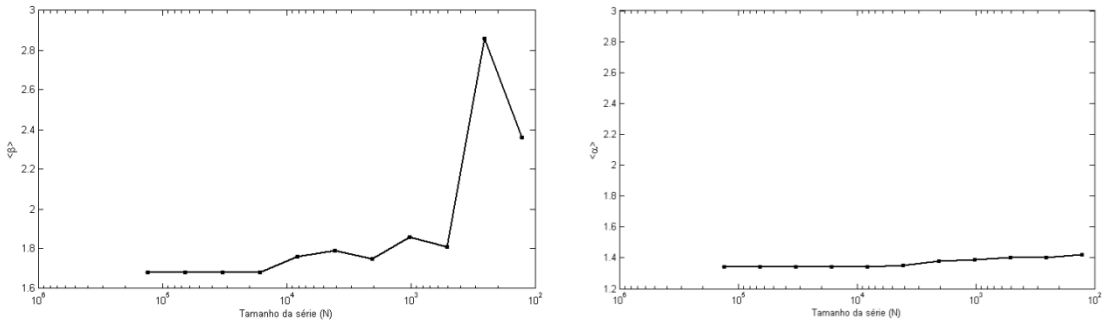
De acordo com o teorema de Wiener-Khinchin (KAY e MARPLE, 1981), é possível mostrar que os dois expoentes β (da PSD) e α (da DFA) estão relacionados por $\beta = 2\alpha - 1$. Assim, para o movimento Browniano fracionário, temos $1 \leq \beta \leq 3$, e, consequentemente, $1 \leq \alpha \leq 2$. Logo, para $\beta = \frac{5}{3}$, temos $\alpha = \frac{4}{3}$.

Tabela 7.3 – Valores de $\boldsymbol{\alpha}$ relacionados com o tamanho da série

Tamanho da série	$< \boldsymbol{\alpha} >$	**Desvio padrão**
131072	1,34	0,006
65536	1,34	0,008
32768	1,34	0,010
16384	1,34	0,011
8192	1,34	0,013
4096	1,35	0,021
2048	1,38	0,025
1024	1,39	0,032
512	1,40	0,040
256	1,40	0,046
128	1,42	0,057

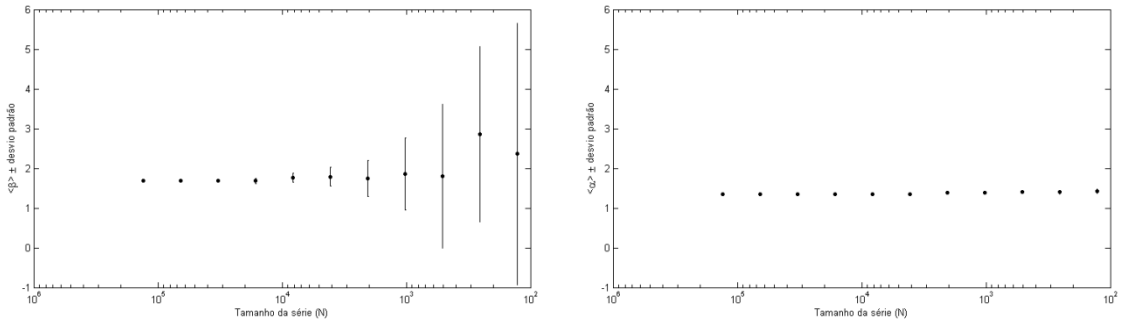
Como pode ser visto na Tabela 7.3 e nas Figuras 7.6 e 7.7, a medida de α é robusta em relação à variação do tamanho do sinal. Nestas figuras são

comparados os resultados de β e α e seus respectivos desvio-padrões para as ST's computadas.



(a) (b)

Figura 7.6 – Variação com o tamanho da ST de: (a) $< \boldsymbol{\beta} >$ e (b) $< \boldsymbol{\alpha} >$



(a) (b)

Figura 7.7 – Erros no cálculo de: (a) $\boldsymbol{\beta}$ e (b) $\boldsymbol{\alpha}$

Ao compararmos esses resultados preliminares, é possível distinguir facilmente a diferença de robustez entre as ferramentas na análise de STC's. Essa robustez da DFA deve-se ao fato de a ferramenta não ter uma dependência estrita da média que possui uma variação crescente à medida que tomamos menos pontos de uma ST, acentuando a partir de $N \sim 10^4$ (VERONESE et al, 2011).

Na análise com PSD, entre as séries de 2^{17} e as de 2^{7} pontos tem-se uma variação de 40,5% entre os respectivos valores de β e uma variação de 46914,3% entre a variância das maiores séries e a das menores. Em contrapartida, na DFA a variação do expoente α entre as STs de 2^{17} e as de 2^{7} pontos não passou de 6,0% e a variação entre a variância das maiores séries e a das menores não passou de 850,0%. Ao relacionarmos essas métricas podemos afirmar que a PSD é 6,7 vezes menos estável em relação à variação do índice e 55,2 vezes menos robusta em relação à diferença de variância com a diminuição da ST. Isso demonstra a estabilidade da DFA em análise de STC's na precisão da detecção da correlação e justifica a escolha da DFA como técnica canônica para testes em um ambiente virtual.

7.2. Análise de Séries Ambientais

7.2.1. Fonte dos Dados

Além dos testes em séries artificiais foram realizados testes com séries ambientais. Os dados utilizados foram fornecidos por Furnas SA (FURNAS, 2013), instituição parceira do INPE em diversas pesquisas. Neste projeto, os dados são obtidos em formato de séries temporais através de vários instrumentos que compõem o Sistema Integrado de Monitoramento Ambiental (SIMA), que é um conjunto de hardware e software feito para a coleta e monitoramento de dados hidrológicos (STEVENSON et al. 1993), desenvolvido em uma parceria entre o INPE e a Universidade do Vale do Paraíba (UNIVAP) (ver Figura 7.8). Informações adicionais sobre o SIMA estão disponíveis em FURNAS (2013) e Stech et al (2006).

Várias informações podem ser medidas pelo SIMA. Assim como a direção do fluxo da água, outras medidas mensuráveis são: concentração de clorofila (mg/l), o pH, a turbidez (NTU), a concentração de O_2 dissolvido (mg/l), a condutividade elétrica (mS/cm), a concentração de nitrato (mg/l), a concentração de amoníaco (mg/l), a temperatura da água (°C), pressão

atmosférica (hPa), umidade relativa (%), temperatura do ar (°C), direção (oNV) e velocidade do vento

(m/s) e a radiação solar (W/m^2).

Figura 7.8 – SIMA instalado no reservatório de Serra da Mesa

7.2.2. Análise de Dados Ambientais I

Com o intuito de testar o VLADA, foi aplicado um estudo, com o objetivo de comparar a complexidade de sistemas limnológicos, caracterizando o padrão de variabilidade de suas variáveis ambientais medidas pelo SIMA. Neste estudo, os dados usados foram obtidos do Lago Curuaí e dos reservatórios Serra da Mesa e Tucuruí. Com base na motivação analítica do projeto, ou seja, séries temporais curtas, quatro variáveis foram selecionadas: média de pressão atmosférica (hPa), temperatura média do ar (°C), umidade relativa média do ar (%) e velocidade do vento (m/s). As séries selecionadas correspondem a variáveis medidas contiguamente no tempo.

Os dados coletados a partir do Lago Curuaí e dos reservatórios de Serra da Mesa e Tucuruí foram extraídos de um banco de dados on-line fornecidos pelo Projeto Balanço do Carbono de Furnas Centrais Elétricas (INPE, 2013) e

correspondem a medidas com resolução diária. Em nossa análise foram utilizados dois períodos comuns sem falhas nos dados para as variáveis escolhidas abrangendo um total de 589 medidas em cada período. A primeira abrange o período de 27/04/2005 a 12/06/2006 e o segundo período de 07/12/2006 a 17/07/2008. As Figuras 7.9 a 7.12 mostram os dados normalizados do segundo período escolhido.

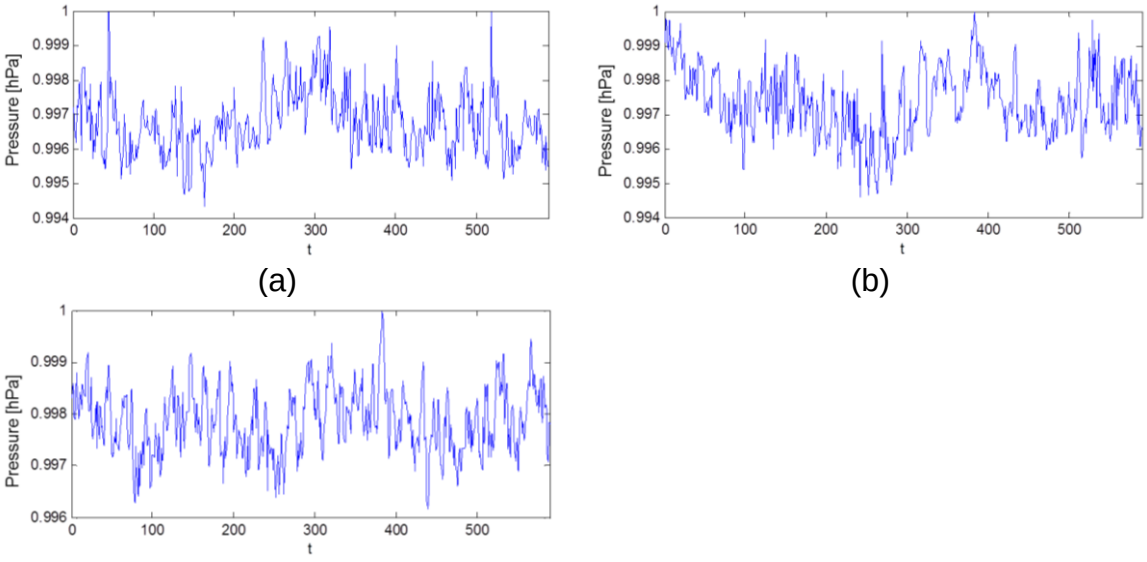


Figura 7.9 – Pressão Atmosférica média normalizada (hPa): (a) Lago Curuaí, (b) Serra da Mesa e (c) Tucuruí

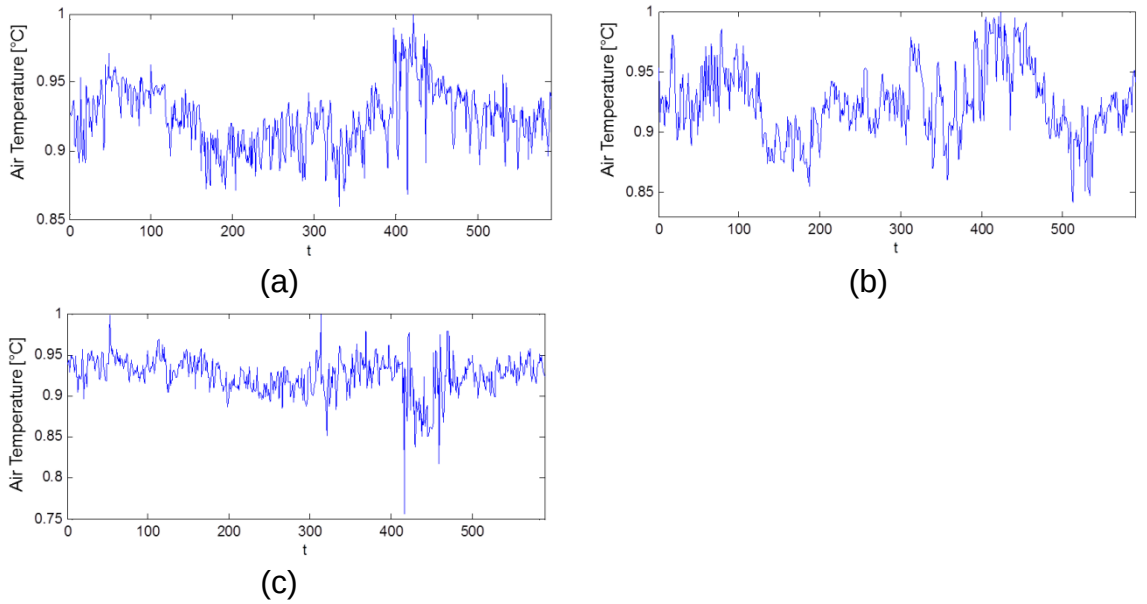


Figura 7.10 – Temperatura do ar média normalizada (°C): (a) Lago Curuaí, (b) Serra da Mesa e (c) Tucuruí

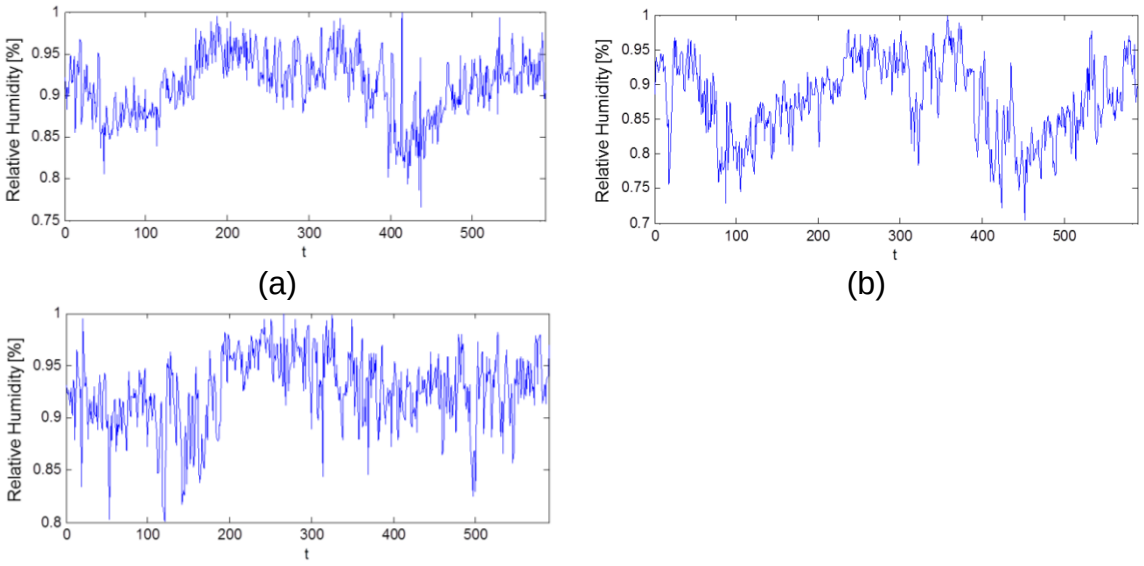


Figura 7.11 – Humidade relativa média normalizada (%): (a) Lago Curuaí, (b) Serra da Mesa e (c) Tucuruí

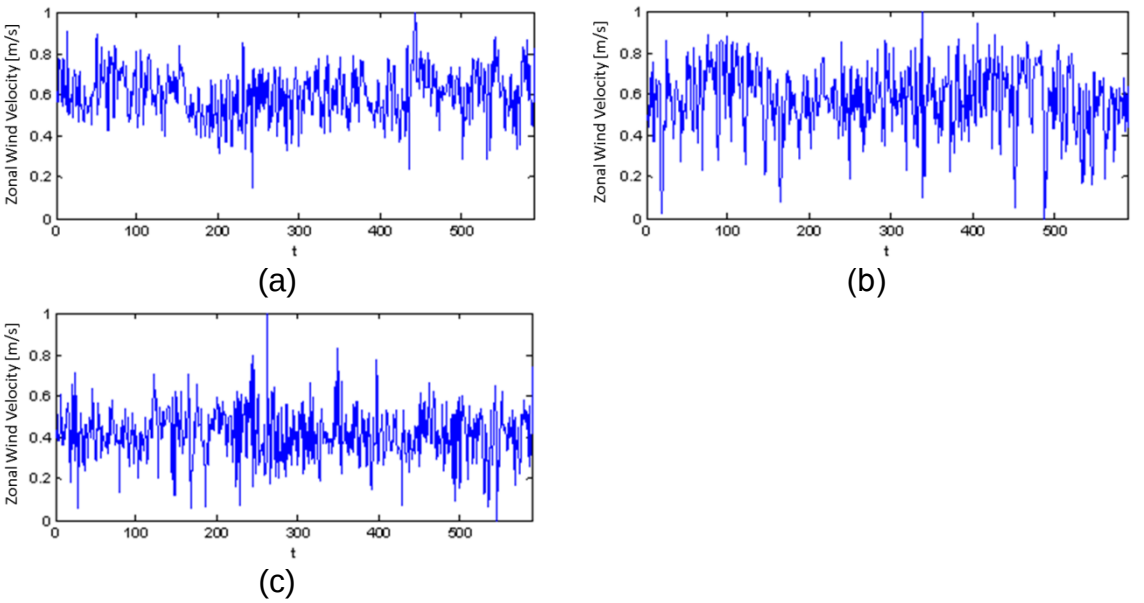


Figura 7.12 – Velocidade do vento média normalizada (m/s): (a) Lago Curuaí, (b) Serra da Mesa e (c) Tucuruí

Como pode ser observado na Tabela 7.4, foram obtidas duas medidas para cada variável nos três sistemas escolhidos devido aos dois períodos de dados em comum.

Tabela 7.4 – Medidas da DFA das variáveis escolhidas em cada sistema

Variável	Curuaí	Serra da Mesa	Tucuruí
Pressão atmosférica 1	1,0825	1,0067	0,7767
Pressão atmosférica 2	1,0579	1,1688	0,9816
Temperatura do ar 1	1,3649	1,1382	1,3649
Temperatura do ar 2	1,1923	0,9885	0,9812
Humidade relativa 1	1,3430	1,3437	1,2205
Humidade relativa 2	1,1886	1,1737	1,2822
Velocidade do vento 1	0,5119	0,5610	0,8473
Velocidade do vento 2	0,8227	0,7005	0,6016

A partir destes resultados é possível identificar algumas semelhanças entre reservatórios. No primeiro período escolhido, identificado pelo número 1, temos uma maior proximidade entre Curuaí e Serra da Mesa em relação à pressão atmosférica, umidade relativa do ar e velocidade do vento. Durante este período, a temperatura do ar de Curuaí é mais próxima em Tucuruí. No segundo período, há uma maior variabilidade dos valores do expoente de escala. Há apenas uma aproximação do índice de temperatura do ar entre a Serra da Mesa e Tucuruí.

7.2.3. Análise de Dados Ambientais II

Este segundo estudo envolvendo séries temporais ambientais e o VLADA envolveu a análise comparativa entre temperatura da água a cinco metros de profundidade para quatro reservatórios com propriedades limnológicas e meteorológicas diferentes. A DFA foi aplicada sobre todas as séries temporais disponibilizadas no banco de dados. Elas foram selecionadas segundo um critério de qualidade mínima, isto é, sem falhas. Tais séries foram inseridas como arquivos de entrada na interface do VLADA.

Para este teste foram selecionadas cerca de 40 séries temporais de cada sistema contendo cerca de 600 pontos cada, representando médias diárias,

totalizando 2 anos de dados. Na Figura 7.13 são apresentados exemplos de quatro séries temporais normalizadas dessa variável para os quatro sistemas analisados.

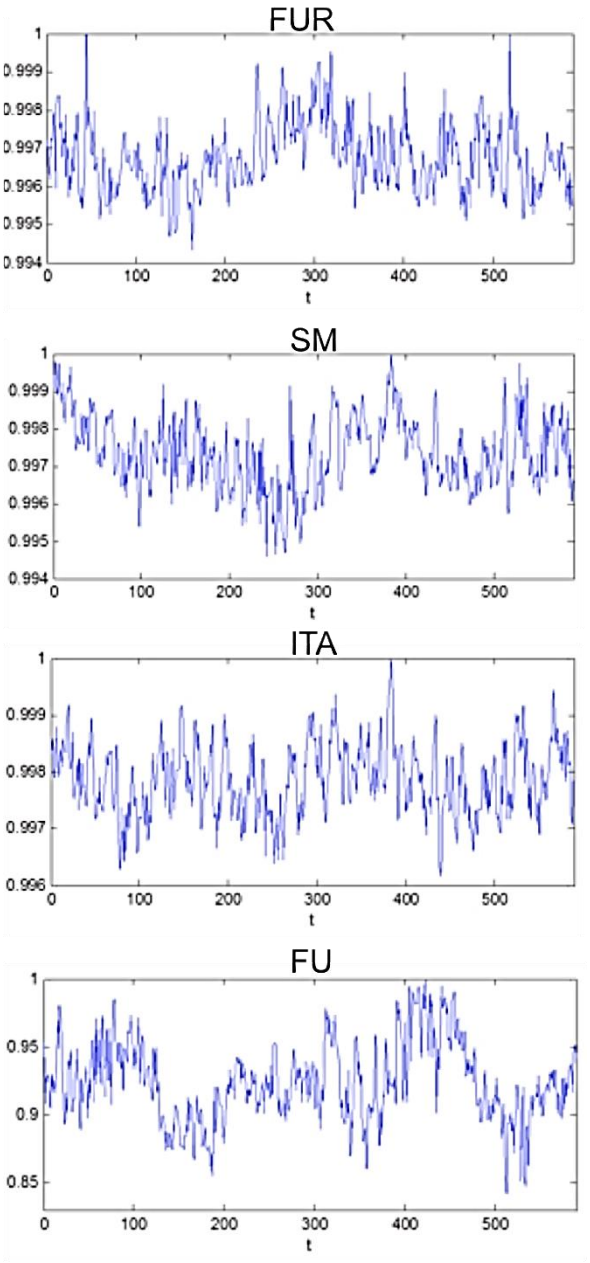


Figura 7.13 – Exemplos de ST's normalizadas da temperatura da água: Furnas (FUR), Serra da Mesa (SM), Itaipu (ITA) e Funil (FU)

A Tabela 7.5 apresenta as medidas da DFA para a temperatura da água na profundidade de cinco metros para os quatro sistemas limnológicos supracitados. Os valores discriminados foram obtidos a partir da média de 40 séries temporais de cada reservatório.

Tabela 7.5 – Medidas da DFA média da temperatura da água a 5 metros

Reservatório	$< \alpha >$
Furnas	1,3024
Serra da Mesa	0,6109
Itaipu	0,3142
Funil	0,0298

Como se trata de reservatórios que podem apresentar características atmosféricas e limnológicas ligeiramente diferentes, o resultado indica a possibilidade de o expoente de escala da DFA poder ser usado como um caracterizador de classificação das propriedades de cada reservatório. Cabe agora investigar outras profundidades, outras variáveis e outros reservatórios para resultados mais contundentes.

Os resultados obtidos através do ambiente em nuvem VLADA foram os mesmos que aqueles obtidos utilizando computação local. A análise em ambiente virtual não impôs erros e não alterou a complexidade computacional da operação. O tempo de execução da análise sobre mais de 200 séries, repetidas várias vezes, variou apenas cerca de alguns segundos (entre 2 e 3).

7.2.4. Análise de Dados Espaciais e Nova Modelagem

Diferentes sensores associados à ionosfera, magnetosfera, sol e à própria superfície terrestre geram uma grande quantidade de dados heterogêneos que são usados para prover informações sobre o estado do clima espacial para a sociedade. As redes de sensores localizadas na superfície da Terra, em geral, estão em locais geográficos distintos implicando na necessidade de transmissão para um centro de dados, de onde podem ser utilizados pelos especialistas de clima espacial.

Por outro lado, diferentes centros regionais também detêm infraestrutura de hardware e software que pode ser integrada a outros sistemas, formando uma rede para a aplicação. Essa integração, no entanto, precisa ser realizada de forma que os dados disponibilizados nesses centros sejam compartilhados de forma transparente, com alta disponibilidade e robustez, garantindo-se a integridade e segurança da informação. Além disso, os centros podem compartilhar serviços para o objetivo final da aplicação. O Programa de Clima Espacial do INPE, do qual o LAC participa, possui características intrínsecas de um sistema distribuído.

Alguns algoritmos foram usados dentro do conceito de ambiente virtual com tais características para analisar 25 arquivos provenientes do projeto Estudo e Monitoramento Brasileiro do Clima Espacial (EMBRACE). Tais dados são oriundos de séries temporais da atividade solar observada a partir do instrumento CALISTO instalado junto ao radiotelescópio *Brazilian Decimetric Array* (BDA), em Cachoeira Paulista-SP.

No intuito de emular um ambiente com 10 frames virtuais de CPU’s distribuídas em nuvem, foram realizados experimentos preliminares usando uma extensão do Matlab, o Simulink. Através dessas aplicações foram obtidos os tempos de resposta para diversos tamanhos de arquivos analisados por três diferentes técnicas. Todas as aplicações resultaram em tempos compatíveis com um ambiente virtual de nuvem como a proposta do VLADA. O tempo mínimo de resposta, considerando a montagem de bancos de dados em tempo real para monitoramento de processos não lineares investigados foi de no máximo 0,25 segundos.

Na Tabela 7.6 são mostrados os tempos médios de resposta para três algoritmos desenvolvidos em Matlab e acopláveis ao VLADA: DFA, GEV e GSA, sendo os dois últimos descritos logo a seguir. A configuração do Simulink Cloud Computing distribui de forma paralela uma entrada sequencial de séries temporais recebidas em tempo real. O tempo na tabela é o tempo médio para processamento de uma série temporal.

Tabela 7.6 – Tempo médio de resposta da operação em ambiente simulado Simulink com frame de 10 CPU's virtuais. Teste em ST's de diferentes tamanhos

Tamanho N pontos (bytes)	**Tempo DFA (seg.)**	**Tempo GEV (seg.)**	**Tempo GSA (seg.)**
500 (100 KB)	0,05	0,07	0,09
1000 (200 KB)	0,07	0,10	0,12
5000 (1000 KB)	0,12	0,15	0,18
10000 (2000 KB)	0,16	0,18	0,21

A DFA foi amplamente apresentada no Capítulo 3. Já o modelo estatístico *Generalized Exteme Value* (GEV) permite encontrar a melhor função densidade de probabilidades para os histogramas das amplitudes medidas nas diversas variáveis relacionadas. O GEV é descrito por três parâmetros de ajuste estatístico que permite a modelagem de flutuações não gaussianas, típicas de eventos extremos, onde a média não é o valor com maior energia de acumulação da série. Dados do fluxo de metano, coletados anteriormente nos reservatórios de Serra da Mesa e Manso, foram bem descritos pelo modelo GEV (Ramos et al, 2006).

O terceiro algoritmo é o da Análise de Espectros-Gradientes (GSA) (do inglês, *Gradient Spectral Analysis*), que também foi testado. Quando se analisa séries temporais curtas, a modelagem estatística torna-se degradada. Dessa forma, é necessário aplicar técnicas que sejam robustas, já que a medida de autocorrelação do sinal é prejudicada pelo baixo número de medidas. A GSA congrega a Análise de Padrões-Gradiente (GPA) com a análise multirresolução por *wavelets*, possuindo assim o potencial para buscar informações no domínio das frequências (escalas) quando a informação no domínio do tempo é insatisfatória. A técnica decompõe as assimetrias da série no domínio das frequências, gerando um espectro que apresenta comportamentos típicos para diferentes processos estocásticos (DANTAS, 2008).

Esses resultados se mostraram promissores e permitiram a modelagem de uma possível atualização da interface do protótipo, conforme pode ser observado na Figura 7.14.

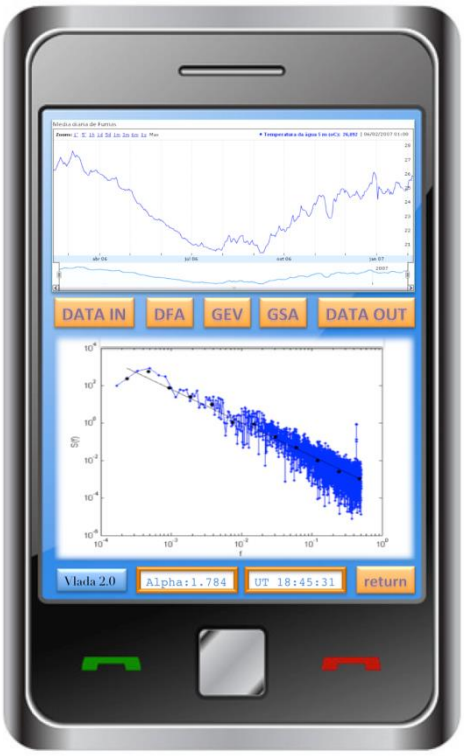


Figura 7.14 – Modelagem de atualização da interface do protótipo VLADA

O comando “DATA IN” seleciona a série temporal a ser visualizada e analisada, previamente arquivada na nuvem. Com os comandos “DFA”, “GEV”e “GSA” é possível selecionar o tipo de análise desejada a ser realizada através do ambiente sobre a série de entrada visualizada na parte superior do dispositivo móvel. Ao clicar num dos comandos de análise e posteriormente o “DATA OUT”, o sistema retorna o gráfico de saída, o valor do expoente de escala (α) e o tempo universal no qual foi realizada a análise. Para selecionar outro dado é necessário antes pressionar o comando “RETURN”.

8 CONSIDERAÇÕES FINAIS

Este capítulo apresenta reflexões finais a respeito do trabalho desenvolvido. São delineadas as conclusões sobre o trabalho fundamentadas nos resultados e nos estudos de caso apresentados. São destacadas as contribuições desta tese e os possíveis trabalhos de pesquisa que podem ser fomentados a partir deste estudo.

8.1. Conclusões do Trabalho

Uma das primeiras tarefas de pesquisa executadas no contexto deste trabalho foi uma pesquisa exaustiva sobre ambientes virtuais para a análise avançada de dados, em especial, de séries temporais, considerando séries temporais curtas. Esta pesquisa demonstrou vários aspectos importantes. O primeiro deles é o de não haver um ambiente livre que permita pesquisa em caráter colaborativo. O segundo aspecto reside no desenvolvimento proprietário de alguns ambientes, o que limita o conceito de comunidade colaborativa de desenvolvimento. Este caráter proprietário é certamente um dos principais fatores de insucesso de muitas propostas. Finalmente, o terceiro aspecto aponta para a demanda crescente de colaboração científica em projetos que envolvam grande quantidade de dados.

Há muitas motivações para o investimento científico num ambiente como o proposto por esta tese: a disseminação do uso das redes de dados possibilitando usos diversificados; a computação móvel acessível através de dispositivos miniaturizados e cada vez mais baratos; a possibilidade de usar o ambiente sem a necessidade de operar equipamentos de alto desempenho e sem a necessidade de aquisição de novos programas; o amadurecimento das tecnologias para ambientes baseados em serviço e nuvem e a necessidade de análise de séries temporais com ferramentas populares e, sobretudo, avançadas.

Conforme os aspectos e as motivações supracitados e a necessidade da análise avançada de dados cada vez mais automatizada e colaborativa, foi

realizado um estudo sobre a alternativa que possa se mostrar mais viável em termos de baixo custo, capacidade evolutiva, interoperabilidade e uso de padrões abertos e de software livre. Deste estudo concluiu-se que a estrutura mais adequada para um ambiente desta natureza é a combinação da arquitetura cliente-servidor em camadas com a orientação a serviço, com o uso de componentes especialistas distribuídos e construídos sobre uma plataforma aberta.

De acordo com os estudos realizados sobre ambientes virtuais para a análise avançada de séries temporais, foi definida então uma infraestrutura de software colaborativo orientado a serviços distribuídos para a análise avançada tanto de séries temporais longas, quanto de séries temporais curtas, envolvendo diferentes linguagens e plataformas computacionais. Sendo assim, a pesquisa especificou: os critérios que permitiram avaliar a adequabilidade de ferramentas de análise avançada de ST's em ambientes virtuais; uma proposta de protótipo, fornecendo uma arquitetura de referência para ambientes colaborativos dessa natureza; uma estrutura para conexão de serviços de análise de ST's; e um portal para acesso a serviços de análise de ST's.

A arquitetura em camadas e orientada a serviços se mostrou promissora para o crescimento e desenvolvimento do ambiente, uma vez que fornece regras de negócio consistentes e uma estrutura bem definida e padronizada, facilitando o projeto e o desenvolvimento.

A estrutura para conexão de serviços caracterizou-se como fundamental para a integração e a disponibilização de ferramentas locais e remotas para o usuário. É a partir desse conjunto de serviços disponibilizados em colaboração que diferentes técnicas de análise de dados podem ser desenvolvidas e disponibilizadas como ferramentas de pesquisa científica.

O protótipo, apesar de ainda ser uma estrutura simplificada de software, tem atendido a demanda dos testes efetuados. Eles atendem a necessidade de interfaces padronizadas e centralizadas que facilitem a análise avançada de séries temporais. Também fornecem personalização, autenticação

centralizada, integra diferentes conteúdos numa página *web* e são baseados em padrões abertos. Eles atendem a maioria das necessidades identificadas neste trabalho.

Em resumo, a infraestrutura de software apresentada neste trabalho mostrou-se adequada para a análise avançada de séries temporais num contexto de computação em nuvem de baixo custo e com alta capacidade de evolução colaborativa. Este ambiente tem potencial para alterar o perfil atualmente preponderante proprietário para a adoção de ambientes de plataforma aberta comunitária e baseada em software livre.

8.2. Principais Contribuições da Tese

Esta tese demonstrou como construir um ambiente colaborativo para análise avançada de séries temporais, seguindo o conceito de interoperabilidade, de baixo custo e com potencial de evolução usando uma infraestrutura baseada em plataforma aberta e software livre. Isso foi possível com a construção de um modelo de negócio baseado em serviços distribuídos disponibilizados numa estrutura de serviços e acessados por um portal *web*. Essa arquitetura distribuída de serviços permitirá a evolução política e administrativa do ambiente para um consórcio de ambientes virtuais em grade de alto desempenho para a análise colaborativa e avançada de dados.

Fez parte do desenvolvimento desta pesquisa a construção do protótipo para a validação das ideias. É importante ressaltar as principais dificuldades e o que foi realizado em termos de desenvolvimento do protótipo:

- Projeto e desenvolvimento do ambiente nas seis fases explicitadas no Capítulo 6;
- Implantação nos servidores do LAC;
- Disponibilização da documentação no ambiente e-WebProject do LAC e arquivamento físico (impresso e em CD) dos documentos das etapas na secretaria do LAC;

- Cinco rodadas de testes com 165 solicitações de correções ao todo, sendo que a maioria foi efetivada e algumas foram identificadas como melhorias posteriores;
- Sete implantações de modificações.

Portanto, o processo de desenvolvimento do protótipo envolveu um sem-número de reuniões. Além disso, é importante ressaltar a não disponibilidade imediata para a compra dos servidores e do *rack* por parte do INPE; os processos rígidos de segurança impostos pelo Serviço Corporativo de Tecnologia da Informação do INPE e o compromisso de manter a documentação do projeto atualizada, padronizada e coerente. Estes aspectos foram considerados no cômputo do tempo de pesquisa e de desenvolvimento do protótipo.

No aspecto da análise de dados, o ambiente inova por definir uma solução para as regras de negócio baseada em *Web Services*, artefatos independentes de hardware e software que podem ser disponibilizados e acoplados via Internet. Inova também na composição de serviços abertos e distribuídos de análise de dados, criando uma nova geração de ambientes de análises, antes monolíticos e proprietários.

Em termos de arquitetura, o ambiente inova com sua infraestrutura cliente-servidor em camadas, combinada com a orientação a serviços, definindo um padrão de referência para ambientes desta natureza. Esta arquitetura facilita a evolução e manutenção do ambiente.

O processo de colaboração em ambientes virtuais também é uma inovação para a área de análise de dados. A proposta do ambiente é inovadora quando permite a integração de novas técnicas pela comunidade do ambiente, mas não de maneira indiscriminada. Há um processo definido para a validação das ferramentas antes da disponibilização das mesmas para os usuários. Esse processo potencializará escrita de código como maior correção, além da discussão científica que permeará o procedimento de publicação da ferramenta.

O ambiente também inova com o conceito de análise de dados nas nuvens ao oferecer uma interface preparada para acesso através de dispositivos móveis sem a necessidade de instalação de programas extras ou de hardware de alto desempenho, com licenças proprietárias de uso. Finalmente, o acoplamento remoto de ferramentas através do Executor de Ferramentas também inova, pois descentraliza o processamento da análise, permitindo um crescimento do ambiente com menor velocidade de *upgrade* de hardware central.

8.3. Trabalhos Futuros

Em termos de pesquisas futuras, elas são consideradas para a consolidação do ambiente e sua evolução em médio e em longo prazo. São elas:

- O desenvolvimento de normas internacionais e melhores práticas para o VLADA, com o intuito de permitir a criação de um comitê internacional para a gestão do ambiente[46]. Essa regulamentação vai facilitar o intercâmbio internacional de recursos VLADA e permitir a padronização dos instrumentos de análise e do formato dos dados;
- A discussão sobre o cadastramento e a execução de ferramentas que estejam instaladas na estrutura do VLADA. É preciso definir esse processo de maneira similar ao processo de certificação e implantação de ferramentas remotas, com a liberação de acesso via protocolo de rede (FTP, por exemplo) para o carregamento do software no sistema. Além disso, é preciso definir os critérios para garantir a segurança do sistema;
- A estruturação da alta disponibilidade permitindo a replicação dos serviços do VLADA, facilitando a internacionalização do ambiente de modo a permitir serviços e usuários globais;
- A aplicação do conceito de orquestração de *Web Services* através de um mecanismo de coordenação central, usando uma linguagem como a BPEL (*Business Process Execution Language*), por exemplo;

[46] Em vários eventos e reuniões científicos ao longo do desenvolvimento desta tese foi observado que vários grupos de pesquisa têm mantido o interesse na evolução do ambiente.

- A consideração do acúmulo de um número elevado de solicitações que gera um enfileiramento das solicitações dos serviços. A revisão da arquitetura deverá contemplar um Sistema de Agendamento, Enfileiramento e Monitoramento da Execução das ferramentas;
- O aprimoramento do sistema de busca para procura dentro de publicação de textos, imagens e documentos para a melhor disseminação dos conhecimentos contidos no VLADA;
- A consideração do conceito de atributos de qualidade de processo que permitam a escolha automática de ferramentas em tempo de execução baseada em fatores de qualidade, como o tempo de execução de tarefa, por exemplo. Essa pesquisa pode ser essencial para a continuidade do VLADA, pois com o seu crescimento e disseminação será possível ter no ambiente, muitas instâncias duplicadas ou replicadas da mesma ferramenta. A definição da ferramenta a ser utilizada pelo usuário dependerá dessa capacidade de escolha.

REFERÊNCIAS BIBLIOGRÁFICAS

ABDALA, M. A. D. **Uma abordagem para a gerência das modificações e configuração em um ambiente para o desenvolvimento e gestão de projetos de software**, (INPE-14099-TDI/1078). Dissertação de Mestrado em Computação Aplicada - Instituto Nacional de Pesquisas Espaciais, São José dos Campos, 2004. Disponível em: http://mtc-m17.sid.inpe.br/col/sid.inpe.br/jeferson/2004/10.01.16.56/doc/publicacao.pdf. Acessado em: maio de 2012.

ABRY, P.; GONÇALVÈS, P.; FLANDRIN, P. Wavelets, spectrum analysis and 1/f processes. In: ANTONIADS; OPPENHEIM (Editors). **Lectures notes in statistics 103:** Wavelet and Statistics. New York: Springer-Verlag, 1995, p. 15-29.

AFSARMANESH, H.; KALETAS, E.C.; BENABDELKADER, A.; GARITA, C.; HERTZBERGER, L.O. A reference architecture for scientific virtual laboratories. **Elsevier**, v. 17, n. 8, p. 999–1008, 2001.

ALLEN, G.; SEIDEL, E.; SHALF, J. Scientific computing on the Grid. **Scientific Computing**, 2002. Available at: http://www-vis.lbl.gov/Publications/2002/LBNL-51039-Byte2002.pdf. Access in: 25/08/14

AMBINDER, D. M.; MARCONDES, C. H. As potencialidades da web semântica e web 2.0 para a ciência da informação e os novos formatos de publicações eletrônicas para a pesquisa acadêmico-científica. **Edicic**, 2011. v. 1, n. 4, 2011.

ANTOUN, H. **Web 2.0:** participação e vigilância na era da comunicação distribuída. Rio de Janeiro: Mauad, 2008.

APACHE. **Apache HTTP server project**. Disponível em: https://httpd.apache.org/. Acessado em: julho de 2013.

BALANCIERI, R.; BOVO, A.B.; KERN, V.M.; PACHECO, R.C.S.; BARCIA, R.M. A análise de redes de colaboração científica sob as novas tecnologias de informação e comunicação: um estudo na Plataforma Lattes. **Ciência da Informação**, Brasília, v.34, n.1, 2005.

BARONI, M.P.M.A.; DE WIT, A.; ROSA, R.R. Detrended fluctuation analysis of numerical density and viscous fingering patterns. **EPL 92**, 64002, doi: 10.1209/0295-5075/92/64002, 2010.

BEAVER, D. B.; ROSEN, R. Studies in scientific collaboration: part I – the professional origins of scientific co-authorship. **Scientometrics**, Amsterdam, v.1, n. 1, p. 65-84, 1978.

BERNAL, J. D. **The social function of science**. London: G. Routledge, 1939.

BERNHOLDT, D.E.; ALLAN, B.A.; ARMSTRONG, R. et al. A component architecture for high-performance scientific computing. **International Journal of High Performance Computing Applications**, v. 20, n. 2, p.163-202, 2006.

BORDONS, M.; GÓMEZ, I. Collaboration networked in science. In: CRONIN, ; ATKINS. (Eds.). **The web of knowledge:** a festschrift in honor of Eugene Garfield. New Jersey: ASIS, 2000. p. 197-214.

BRUZZONE, A. G.; UHRMACHER, A.; PAGE, E. H. The society for computer simulation international. In: INTERNATIONAL CONFERENCE ON WEB-BASED MODELING AND SIMULATION, 1999, San Francisco. **Proceedings…** San Francisco: The Society for Computer Simulation, 1999.

CASTELLS, M. **A galáxia da *Internet***: reflexões sobre a *internet*, os negócios e a sociedade. Rio de Janeiro: J. Zahar, 2003.

CAVALCANTI, M.; NEPOMUCENO, C. **O conhecimento em rede**. Rio de Janeiro: Elsevier, 2007.

CEREJA JR., M. G.; SANT'ANNA, N.; BORREGO FILHO, L.F.; GENVIGIR, E.C.; LUQUE, L.; TAVARES, R.P.; CASILLO, B.H. **(**PDE) Process Definition Environment: uma ferramenta para apoio à definição de processos de software no ambiente e-WebProject. In: CONGRESSO INTERNACIONAL DE TECNOLOGIA DE SOFTWARE, 14., 2003, Curitiba. **Anais...** Curitiba : CITS Centro Internacional de Tecnologia de Software, 2003.

COZIC, F. **Web x Web 2.0**. Paris, 2007. Disponível em: http://blog.aysoon.com/le-Web20-illustre-en-une-seule-image. Acessado em: maio de 2012.

COMER, D. **Interligação de redes com TCP/IP**. Rio de Janeiro: Elsevier, 2006. V. 1.

CRONIN, B. **The hand of science**: academic writing and rewards. Oxford: Scarecrow Press, 2005.

CUKIER, K. Data, data everywhere. **The Economist.** Disponível em: http://www.economist.com/specialreports/displaystory.cfm?story_id=15557443. Acessado em: 18 de Setembro de 2011.

DANTAS, M.S.; ROSA, R.R.; SANT'ANNA, N.; CEREJA JR, M.G.; VERONESE, T.B.; BIANCHI, S.; ROSA, J.C.; ALEXIEV, K.M.; SILVA, J.D.S. The VLADA white paper: building an active Virtual Lab for Advanced Data

Analysis. **Journal of Computational Interdisciplinary Sciences, v.** 2, n.1, p. 47-56, 2011.

DANTAS, M.S. **Análise espectral de padrões-gradiente de séries temporais curtas**. (INPE-15676-TDI/1450). Dissertação (Mestrado em Computação Aplicada) – Instituto Nacional de Pesquisas Espaciais, São José dos Campos, 2008. Disponível em: http://URLIB.NET/SID.INPE.BR/MTC-M18@80/2009/02.05.10.55. Acessado em: 16 de setembro de 2011.

ELLIS, C. A.; GIBBS, S. J.; REIN, G. Groupware: some issues and experiences. **Communications of the ACM**, v. 34, n.1, p. 39-58, 1991.

ENGELEN, R.A.V. Pushing the SOAP envelope with web services for scientific computing. In: INTERNATIONAL CONFERENCE ON WEB SERVICES (ICWS), 2003, Erfurt, Germany. **Proceedings…** Erfurt: IEEE Computer Society, 2003.

ERICKSON, J. S.; SPENCER, S.; RHODES, M.; BANKS, D.; RUTHERFORD, J.; SIMPSON, E.; BELROSE, G. Content-centered collaboration spaces in the cloud. **IEEE Internet Computing**, v. 13, n. 5, p. 34–42, 2009.

FISHWICK, P.A.; HILL, D.R.C.; SMITH, R. In: INTERNATIONAL CONFERENCE ON WEB-BASED MODELING AND SIMULATION, 1998, San Diego. **Proceedings…** San Diego: The Society for Computer Simulation International, 1998.

FOSTER, I.; VOCKLER, J.; WILDE, M.; ZHAO, Y. Chimera: avirtual data system for representing, querying, and automating data derivation. In: INTERNATIONAL CONFERENCE ON SCIENTIFIC AND STATISTICAL DATABASE MANAGEMENT, 14, 2002. **Proceedings...** [S.l.]: IEEE Computer Society, 2002.

FREITAS, R.M. de. **Laboratório virtual para visualização e caracterização do uso e cobertura da terra utilizando imagens de sensoriamento remoto**. (INPE- 02.24.17.32-TDI). Tese (Doutorado em Computação Aplicada) – Instituto Nacional de Pesquisas Espaciais, São José dos Campos, 2012. Disponível em: http://urlib.net/8JMKD3MGP7W/3BDRG5P. Acessado em: fevereiro de 2013.

FREITAS, R.M.; ARAI, E.; ADAMI, M.; FERREIRA, A.S. Virtual laboratory of remote sensing time series: visualization of MODIS EVI2 data set over South America. **Journal of Computational Interdisciplinary Sciences**, v. 2, n. 1, p. 57-64, 2011.

FURNAS CENTRAIS ELÉTRICAS S/A (FURNAS). **Banco de dados**. Disponível em: http://www.dpi.inpe.br/sima. Acessado em: novembro de 2013.

GENVIGIR, E. C. **Um modelo de processo da engenharia de requisitos aplicáveis a ambientes de software centrado em processos**. (INPE-14630-TDI/1201). Dissertação (Mestrado em Computação Aplicada) - Instituto Nacional de Pesquisas Espaciais, São José dos Campos, 2004. Disponível em: http://mtc-m18.sid.inpe.br/sid.inpe.br/jeferson/2004/07.05.16.17?mirror=sid.inpe.br/mtc-m18@80/2008/03.17.15.17.24&metadatarepository=sid.inpe.br/jeferson/2004/07.05.16.17.40. Acessado em: maio de 2012.

GRAHAM, M.J., FITZPATRICK, M.J., MCGLYNN, T.A. **The national virtual observatory**: tools and techniques for astronomical research. Astronomical Society of the Pacific, 2007. v. 382. ISBN: 978-1-58381-327-0.

HAZELHURST, S. Scientific computing using virtual high-performance computing: a case study using the Amazon Elastic Computing Cloud. In: ANNUAL CONFERENCE OF THE SOUTH AFRICAN INSTITUTE OF COMPUTER SCIENTISTS AND INFORMATION TECHNOLOGISTS ON IT RESEARCH IN DEVELOPING COUNTRIES (SAICSIT), 2008, Wilderness, South Africa. **Proceedings…** Wilderness: ACM, 2008.

HAYES, B. Cloud computing. **Communications of the ACM**, v.51, n.7, p. 9–11, 2008.

HELLERSTEIN, J. **Parallel programming in the age of big data**. Disponível em: http://gigaom.com/2008/11/09/mapreduce-leads-the-way-for-parallel-programming/. Acessado em: 18 de Setembro de 2011.

HEY, T.; TREFETHEN, A. Cyberinfrastructure for e-Science. **Science,** v. 308, n. 5723, p. 817-821, 2005. DOI: 10.1126/science.1110410.

HEY, T.; TREFETHEN, A. e-Science and its implications. **Philos Trans A Math Phys Eng Sci**. v.361,p. 1809-25. 2003.

HOFFA, C.; MEHTA, G.; FREEMAN, T. On the Use of Cloud Computing for Scientific Workflows. In: FOURTH IEEE INTERNATIONAL CONFERENCE ON e-SCIENCE, In: IEEE FOURTH INTERNATIONAL CONFERENCE ON eSCIENCE (eScience'08), 4., 2008, Indianapolis, Indiana. **Proceedings…** Indianápolis: IEEE, 2008.

HOU, H.; KRETSCHMER, H.; LIU, Z. The structure of scientific collaboration networks in Scientometrics. **Scientometrics**, Amsterdam, v. 75, n. 2, p. 189-202, 2008.

HU K.; IVANOV P.C.; CHEN Z.; CARPENA P.; STANLEY H.G. Effect of trends on detrended fluctuation analysis. **Phys. Rev. E**, v. 64, 011114, 2001.

INTERNATIONAL BUSINESS MACHINES (IBM). **BigData.** Disponível em: http://www-01.ibm.com/software/data/bigdata/. Acesso em: 16 de Setembro de 2011.

INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS (INPE). **Projeto balanço do carbono de furnas centrais elétricas.** Período: 23/01/2005 a 11/02/2006. Disponível em: http://www.dpi.inpe.br/sima/bancos. Acesso em: maio de 2013.

JACOBS, A. **The pathologies of big data**. Disponível em: http://queue.acm.org/detail.cfm?id=1563874. Acessado em: 17 de Setembro de 2011.

JOHNSON-LENZ, P.; JOHNSON-LENZ, T. Groupware: coining and defining it. **ACM SIGGROUP Bulletin**, v. 19, n. 3, p. 58, 1998.

KATZ, J. S.; MARTIN, B. R. What is research collaboration? **Research Policy**, Amsterdam, n. 26, p. 1-18, 1997.

KAY, S.M.; MARPLE,S.L. Spectrum analysis – a modern perspective. **Proceedings of the IEEE**, v. 69, n. 11, 1981.

KEAHEY, K.; FIGUEIREDO, R.; FORTES, J.; FREEMAN, T.; TSUGAWA, M. Science Clouds: Early Experiences in Cloud Computing for Scientific Applications. **Cloud computing and its applications**, v. 2008, p. 825-830, 2008.

KEAHEY, K.; FREEMAN, T.; LAURET, J.; OLSON, D. Virtual workspaces for scientific applications. **Journal of Physics: Conference Series**, v. 78, 2007.

KRETSCHMER, H. Author productivity and geodesic distance in bibliographic co-authorship networks, and visibility on the Web. **Scientometrics**, Amsterdam, v.60, n.3, 2004.

KUHN, T.S. **A estrutura das revoluções científicas**. São Paulo: Perspectiva, 2007.

KULJIS, J.; PAUL, R.J. An appraisal of web-based simulation: whither we wander. **Simul. Practice Theory**, v.9, n.1, p. 37-54, 2001.

LAHOZ, C. H. N. **Uma abordagem para a gerência da qualidade em um ambiente de engenharia de software centrado em processo**. (INPE-11550-TDI/958). Dissertação (Mestrado em Computação Aplicada) - Instituto Nacional de Pesquisas Espaciais, São José dos Campos, 2004. Disponível em: http://mtc-m16.sid.inpe.br/col/sid.inpe.br/jeferson/2004/05.06.14.08/doc/publicacao.pdf. Acessado em: maio de 2012.

LEITE, J.C.S.P. **Transparência:** desafios para a engenharia de software. Disponível em: http://jcspl.worldpress.com/. Acessado em: 16 de setembro de 2011.

LETA, J.; GLÄNZEL, W.; THIJS, B. Science in Brazil. Part 2: sectoral and institutional research profiles. **Scientometrics**, Amsterdam, v. 67, n. 1, p. 87-105, 2006.

LUZ, E.F.P. **Meta-heurísticas paralelas na solução de problemas inversos**. (INPE- 02.22.17.13-TDI). Tese (Doutorado em Computação Aplicada) – Instituto Nacional de Pesquisas Espaciais, São José dos Campos, 2012. Disponível em: http://urlib.net/8JMKD3MGP7W/3BDGLGH. Acessado em: agosto de 2012.

MAIA, M.F.S.; CAREGNATO, S.E. Co-autoria como indicador de redes de colaboração científica. **Perspectivas em Ciência da Informação**, Belo Horizonte, v. 13, n. 2, p. 18-31, maio/ago, 2008.

MALAMUD, B. D.; ~~e~~ TURCOTE, D. L. Self-affine time series: measures of weak and strong persistence. **Journal of Statistical Planing and Inference**, v. 80, p. 173-196, 1999.

MANDELBROT, B.B.; VAN NESS, J.W. Fractional brownian motions, fractional noises and applications. **SIAM Review**, v.10, n.4, 1968, p.422-437.

MARCON JR, A.; LAUREANO, M.; SANTIN, A.; MAZIERO, C. Aspectos de segurança e privacidade em ambientes de computação em nuvem. In: SIMPÓSIO BRASILEIRO EM SEGURANÇA DA INFORMAÇÃO E DE SISTEMAS COMPUTACIONAIS, 10, 2010, **Minicurso...** Disponível em: http://professor.ufabc.edu.br/~joao.kleinschmidt/aulas/seg2011/nuvem.pdf. Acesso em: 15 de junho de 2012.

MELL, P.; GRANCE, T. **The NIST definition of cloud computing.** Gaithersburg, MD,: National Institute of Standards and Technology, 2009.

MIKE2.0. **Big data definition**. Disponível em: http://mike2.openmethodology.org/wiki/Big_Data_Definition. Acessado em: 15 de setembro de 2011.

MORETTIN, P. A., TOLOI, C. M. C. **Análise de séries temporais**. 2. ed. São Paulo: Edgard Blucher, 2006.

MOURA, M.A. Informação e conhecimento em redes virtuais de cooperação científica: necessidades, ferramentas e usos. **DataGramaZero:** revista de ciência da informação, Rio de Janeiro, v. 10, n. 2, abr. 2009. Disponível em: http://www.dgz.org.br/abr09/Art_02.htm. Acesso em: 10 de julho de 2012.

MURUGESAN, S.; DESHPANDE, Y.; HANSEN, S.; GINIGE, A. Web Engineering: a New Discipline for Development of Web-Based Systems. **Springer Berlin**, v. 2016, p. 3-13, 2001.

O'BRIEN, W. J. Implementation issues in project websites: a practitioner's viewpoint. **Journal of Management in Engineering**, ASCE. v. 16, n. 3, p. 34-39, mai 2000.

OGRIZOVIC, D.; SVILICIC, B.; TIJAN, E. Open source science clouds. In: INTERNATIONAL CONVENTION ON INFORMATION AND COMMUNICATION TECHNOLOGY, ELECTRONICS AND MICROELECTRONICS (MIPRO 2010), 33, 2010. Opatija, Croatia. **Proceedings...** Opatija, IEEE, 2010.

OSBORNE, A. R.; PROVENZALE, A. Finite correlation dimension for stochastic systems with power-law spectra. **Physica D: Nonlinear Phenomena**, v.35, n. 3, p. 357-381, 1989.

PAGE, E.H.; BUSS, A.; FISHWICK, P.A.; HEALY, K.J.; NANCE, R. E.; PAUL, R. J. Web-based simulation: revolution or evolution? **ACM Trans. Model. Comput. Simul**. v. 10, n. 1, p.3-17, 2000.

PAOLINI, C.P., BHATTACHARJEE, S. A Web service infrastructure for thermochemical data. **J. Chem. Inf. Model**, v. 48, n. 7, p.1511-23, 2008. doi: 10.1021/ci700457p.

PAPAZOGLOU, M.P. Service-Oriented Computing: Concepts, Characteristics and Directions. In: WEB INFORMATION SYSTEMS ENGINEERING WORKSHOPS (WISE' 03), 4, 2003, Rome. **Proceedings…**Rome: IEEE, 2003. p. 3-12.

PENG, C.K.; BULDYREV, S.V.; HAVLIN, S.; SIMONS, M.; STANLEY, H.E.; GOLDBERGER, A.L. Mosaic organization of DNA nucleotides. **Physical Review E,** v. 49, p. 1685-1689, 1994.

PENG, C.K.; SHLOMO, H.; STANLEY, H.E.; GOLDBERGER, A.L. Quantification of scaling exponents and crossover phenomena in nonstationary heartbeat time series. **AIP Chaos**, v. 5, p. 82-87, 1995.

PIMENTEL, M.; FUKS, H. (orgs.). **Sistemas colaborativos**. Rio de Janeiro: Elsevier, 2011.

PRICE, D.J.S. **O desenvolvimento da ciência:** análise histórica, filosófica, sociológica e econômica. Rio de Janeiro: Livros Técnicos e Científicos, 1976.

PRIMO, A. **Interação mediada por computador.** Porto Alegre: Sulina, 2008.

RAMOS et al., Extreme event dynamics in methane ebullition fluxes from tropical reservoirs, **Geophysical Research Letters**, v. 33, L21404, 2006. Doi:10.1029/2006GL027943

REHR, J.J.; VILA, F.D.; GARDNER, J.P.; SVEC, L.; PRANGE, M. Scientific Computing in the Cloud. **Computing in Science and Engineering,** v. 12, n. 3, p. 34-43, 2010. doi:10.1109/MCSE.2010.70.

ROURE, D.; HENDLER J.A. E-Science: the grid and the semantic web. **IEEE Computer Society**, p. 65-71, 2004.

SANT'ANNA N.; CEREJA Jr, M. G.; BORREGO FILHO, L. F.; LUQUE L.; CASILLO, B. H. e-WebProject - Um ambiente integrado para o apoio ao desenvolvimento e gestão de projetos de software. In: CONGRESSO INTERNACIONAL DE TECNOLOGIA DE SOFTWARE: QUALIDADE E PRODUTIVIDADE NO GERENCIAMENTO DE PROJETOS, 13, 2002, 19-21 junho, Curitiba, Paraná, Brasil. **Anais...** 2002. p. 163-174.

SANT'ANNA, N. **Um ambiente integrado para o apoio ao desenvolvimento e gestão de projetos de software para sistemas de controle de satélites**. (INPE-8306-TDI/765). Tese (Doutorado em Computação Aplicada) – Instituto Nacional de Pesquisas Espaciais, São José dos Campos, 2000. Disponível em: http://mtc-m05.sid.inpe.br/rep/dpi.inpe.br/lise/2002/03.28.20.03?mirror=sid.inpe.br/banon/2001/04.03.15.36.19&metadatarepository=dpi.inpe.br/lise/2002/03.28.20.03.51. Acessado em: maio de 2010.

SEGARAN, T.; HAMMERBACHER, J. **Beautiful Data.** 1. ed. Sebastopol: O'Reilly Media, 2009.

SHNEIDERMAN, B. Web Science: a provocative invitation to computer science. **Communications of the ACM**, v. 50, n.6, p. 25-27, 2007.

SILVA, A.B.O. et al. Estudo da rede de coautoria e da interdisciplinaridade na produção científica com base nos métodos de análise de redes sociais: avaliação do caso do Programa de pós-graduação em Ciência da Informação - PPGCI/UFMG. **Encontros Bibli**, Florianópolis, v. especial, p.179-194, 2006.

SILVA E.L. Rede científica e a construção do conhecimento. **Informação e sociedade: estudos**, João Pessoa, v.12, n.1, 2002.

SIQUEIRA, E. Mundo terá 55 bilhões de dispositivos móveis em 2020. **Estadão**. Disponível em: http://blogs.estadao.com.br/ethevaldo-siqueira/2011/02/21/55-bilhoes-de-dispositivos-moveis/. Acessado em: 15 de setembro de 2011.

SOMMERVILLE, I. **Engenharia de software**. 8. ed. São Paulo: Pearson Education, 2007.

SONNENWALD, D. H. Scientific collaboration. **Annual Review of Information Science and Technology**, New York, v. 42, n. 1, p. 643-681, 2008.

SOUSA, F. R. C.; MOREIRA, L. O.; e MACHADO, J. C. Computação em nuvem: conceitos, tecnologias, aplicações e desafios. In: MOURA, R. S.; SOUZA, F. V.; OLIVEIRA, A. C. (Orgs.). **Escola regional de computação** (Ceará, Maranhão e Piauí), ERCEMAPI 2009, 1. Ed.Piauí: EDUFPI, 2009.

SRIRAMA, S.; BATRASHEV, O.; VAINIKKO, E. SciCloud: scientific computing on the cloud. In: IEEE/ACM INTERNATIONAL CONFERENCE ON CLUSTER, CLOUD AND GRID COMPUTING, 10, 2010, Melbourne, Victoria, Australia **Proceedings…** Melbourne: IEEE Computer Society, 2010.

STECH, J.L.; LIMA, I.B.T.; NOVO, E.M.L.M.; SILVA, C.M.; ASSIREU, A.T.; LORENZZETTI, J.A.; CARVALHO, J.C.; BARBOSA, C.C.; ROSA, R.R. telemetric monitoring system for meteorological and limnological data acquisition in aquatic environments. **Verh Internat, Verein Limnol.**, v.29, 2006.

STEVENSON, M.R.; LORENZZETTI, J.A.; STECH, J.L.; ARLINO, P.R.A. SIMA: an integrated environmental monitoring system. IN: SIMPÓSIO BRASILEIRO DE SENSORIAMENTO REMOTO, 7, 1993, Curitiba. **Anais...** São José dos Campos: INPE, 1993. v. 4, p. 300-310. Printed, On-line. ISBN 978-85-17-00055-3. (INPE-7922-PRE/3758). Disponível em: <http://urlib.net/sid.inpe.br/iris@1912/2005/07.20.01.08.41>. Acesso em: 27 ago. 2014.

STOKES, D. **O quadrante de Pasteur**: a ciência básica e a inovação tecnológica. Campinas: Ed. da UNICAMP, 2005.

TRUONG, T.N.; NAYAK, M.; HUYNH, H.H. et al. Computational Science and Engineering Online (CSE-Online): a cyber-infrastructure for scientific computing. **J. Chem. Inf. Model.**, v. 46, n.3, p. 971-84, 2006.

TRUONG, T.N. An integrated web-based grid-computing environment for research and education in computational science and engineering. In: ANNUAL SIMULATION SYMPOSIUM (ANSS'04), 37, 2004, Arlington, Virginia. **Proceedings...** IEEE Computer Society, 2004.

VANZ, S.A.S. **As redes de colaboração científica no Brasil:** 2004-2006. Porto Alegre, 2009. 204 f. Tese (Doutorado em Comunicação e Informação) – Faculdade de Biblioteconomia e Comunicação, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2009.

VECCHIOLA, C.; PANDEY, S.; BUYYA R. High-performance cloud computing: a view of scientific applications. In: INTERNATIONAL SYMPOSIUM ON PERVASIVE SYSTEMS, ALGORITHMS, AND NETWORKS (ISPAN '09), 10, 2009. **Proceedings…** Washington, DC: IEEE Computer Society, 2009.

VERONESE T.B.; ROSA R.R.; BOLZAN M.J.A.; FERNANDESC F.C.R.; SAWANT H.S.; KARLICKY, M. Fluctuation analysis of solar radio bursts associated with geoeffective X-class ares. **Journal of Atmospheric and Solar-Terrestrial Physics**, doi:10.1016/j.jastp.2010.09.030, 2011.

WANG, Y. et al. Scientific collaboration in China as reflected in coauthorship. **Scientometrics**, Amsterdam, v. 62, n. 2, p. 183-198, 2005.

WEI, Z., PIERRE, G., CHI, C.-H. Scalable transactions for web applications in the cloud. **Euro-Par**, p. 442–453, 2009.

WHITE, T. **Hadoop:** the definitive guide**.** 1. ed. O'Reilly Media. 2009.

YALAMANCHILLI, N.; COHEN, W. Communication performance of java-based parallel virtual machines. **Concurrency: Practice and Experience**, John Wiley & Sons, Ltda., United States of America. v. 10, n. 11-13, p. 1189–1196, Sept-Nov.1998.

ZHANG, Q.; CHENG, L.; BOUTABA, R. Cloud computing: state-of-the-art and research challenges. **Journal of Internet Services and Applications**, p. 7-18, 2010.

ZHANG, H.; GUO, H. Scientific research collaboration in China. **Scientometrics**, Amsterdam, v. 38, n. 2, p. 309-319, 1997.

ZHUGE, H. China's e-Science knowledge grid environment. **IEEE Intelligent Systems**, v.19, n. 1, p. 13-17, 2004.

ZIMAN, J. M. **Conhecimento público.** Belo Horizonte: Itatiaia. São Paulo: Ed. da USP, 1979.

APÊNDICE A – EXEMPLO: CASOS DE USO

Este apêndice apresenta os casos de uso do Controle de Acesso.

A.1 Atores

Tool Publisher: desenvolvedor ou autor que publica uma ferramenta para realização de experimentos no Laboratório Virtual;

Researcher: pesquisador que realiza experimentos, isto é, carrega os dados, executa ferramentas e analisa resultados;

User: usuário que acessa o sistema com as permissões concedidas ao seu perfil.

A.2 Diagrama de Contexto

A Figura A.1 apresenta o Diagrama de Contexto para o Controle de Acesso.

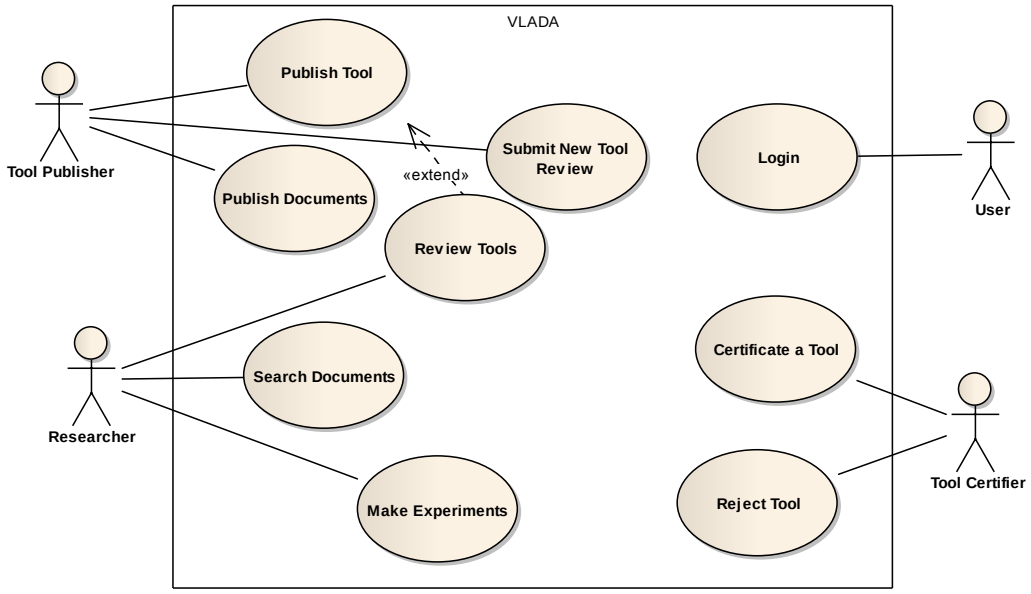


Figura A.1 – Diagrama de Contexto do Controle de Acesso

A.3 Funcionalidades

Publicar Ferramentas: o desenvolvedor publica a ferramenta na WEB e informa os dados de autoria da ferramenta e acesso no laboratório.

Fazer Experimentos: o pesquisador carrega série de dados, informa nome do experimento, faz diversas execuções com as ferramentas disponíveis e analisa os dados.

Login: autenticação e autorização de funcionalidades de acordo com o perfil de usuários.

Gestão de Usuários: funcionalidades para gerenciamento de usuários e perfis.

A.4 Descrição dos Casos de Uso

Descrição Caso de Uso VLADA-UC-01 – Deploy Tool

Identificador do Caso de Uso	VLADA-UC-01
Nome do Caso de Uso	Publicar Ferramentas
Atores	Tool Publisher
Prioridade	[ ] baixa [ ] média [X] alta
Pré Condição	Não se aplica
Requisitos Especiais	Não se aplica
Primeiro Cenário – Publicar Ferramentas	
Fluxo de Eventos:	
Fluxo Principal 1. Este caso de uso inicia-se quando o desenvolvedor acessa o *menu* publicar ferramentas do portal VLADA. 2. O sistema apresenta um formulário com dados de autoria como, Nome da Ferramenta, Instituição, URL para acesso entre outros. 3. O usuário seleciona a opção publicar. 4. O sistema apresenta a mensagem de sucesso e disponibiliza a ferramenta para experimentação.	
Fluxos Alternativos No passo 3, se não forem informados campos obrigatórios o sistema deve informar. No passo 3, deve ser feita uma verificação de disponibilidade da ferramenta.	
Pós Condição: O sistema cria uma ferramenta NÃO certificada com situação.	

Descrição Caso de Uso VLADA-UC-02 – Make Experiments

Identificador do Caso de Uso	VLADA-UC-02
Nome do Caso de Uso	Realizar Experimentos
Atores	Researcher
Prioridade	[ ] baixa [ ] média [X] alta
Pré Condição	Deve haver pelo menos 1 ferramenta disponível para Experimentação.
Requisitos Especiais	Os experimentos devem ser filtrados por usuário.
Primeiro Cenário – Criar Experimentos	
Fluxo de Eventos:	
Fluxo Principal 1. Este caso de uso inicia-se quando o pesquisador seleciona um experimento da lista de experimentos do usuário. 2. O sistema apresenta um formulário com dados de autoria como, Nome do Experimento, Carregar Arquivo, Formato do Arquivo e Descrição. 3. O Usuário pressiona o botão criar. 4. O Sistema apresenta a lista uma lista atualizada dos experimentos.	
Fluxos Alternativos No passo 3, se não forem informados os campos obrigatórios, o sistema deve informar.	
Pós Condição: Não se aplica.	
Segundo Cenário – Executar Experimentos	
Fluxo de Eventos:	
Fluxo Principal 1. Este caso de uso inicia-se quando o pesquisador seleciona um experimento da lista de experimentos. 2. O sistema apresenta uma tela com os dados do experimento e a lista de execuções com um formulário para selecionar uma ferramenta e o botão para executar o experimento. 3. O usuário pressiona o botão executar. 4. O sistema apresenta a lista uma lista atualizada dos resultados de experimentos.	
Fluxos Alternativos Não se aplica.	
Pós Condição: Não se aplica.	

Descrição Caso de Uso VLADA-UC-01 – Login

Identificador do Caso de Uso	VLADA-UC-01
Nome do Caso de Uso	Logan
Atores	User
Prioridade	[ ] baixa [ ] média [X] alta
Pré Condição	Não se aplica
Requisitos Especiais	Não se aplica
Primeiro Cenário – Publicar Ferramentas	
Fluxo de Eventos:	
Fluxo Principal 1. Este caso de uso inicia-se quando o desenvolvedor acessa a página inicial do aplicativo. 2. O sistema apresenta um formulário com usuário e senha. 3. O usuário pressiona *login*. 4. O sistema apresenta uma tela inicial com a funcionalidade padrão do perfil do usuário.	
Fluxos Alternativos No passo 3, deve ser encaminhada uma notificação de falha no caso de insucesso no processo de autenticação.	
Pós Condição: O sistema cria uma ferramenta NÃO certificada com situação.	

Descrição Caso de Uso VLADA-UC-02 – User Management

Identificador do Caso de Uso	VLADA-UC-02
Nome do Caso de Uso	User Management
Atores	User
Prioridade	[ ] baixa [ ] média [X] alta
Pré Condição	O caso de Uso VLADA-UC-01 deve ser executado com usuário de perfil Administrador.
Requisitos Especiais	Não se Aplica.
Primeiro Cenário – Consultar Usuários	
Fluxo de Eventos:	
Fluxo Principal 1. Este caso de uso inicia-se quando o pesquisador informa uma palavra chave no formulário de busca e pressiona “Buscar”. 2. O sistema apresenta uma lista com dados de usuários e opções de edição, exclusão, criação e visualização.	
Fluxos Alternativos Não se aplica.	
Pós Condição: Não se aplica	

Segundo Cenário – Editar/Criar
Fluxo de Eventos:
Fluxo Principal 1. Este caso de uso inicia-se quando o pesquisador seleciona “Editar um usuário” ou pressiona o botão “Novo”. 2. O sistema apresenta uma tela com os dados do usuário. 3. O usuário informa os dados e pressiona o botão “Salvar”. 4. O sistema apresenta uma lista atualizada dos usuários.
Fluxos Alternativos Não se aplica.
Pós Condição: Não se aplica
Segundo Cenário – Apagar
Fluxo de Eventos:
Fluxo Principal 1. Este caso de uso inicia-se quando o pesquisador seleciona “Apagar um usuário” da lista de usuários. 2. O sistema apresenta um formulário de confirmação da operação. 3. O usuário pressiona “Ok”. 4. O sistema apresenta a lista uma lista atualizada dos usuários.
Fluxos Alternativos Não se aplica.
Pós Condição: Não se aplica.

APÊNDICE B – EXEMPLO: SCRIPT DE BANCO DE DADOS

Este apêndice apresenta o script para a criação do Banco de Dados para o Controle de Acesso, escrito em SQL para PostgreSQl versão 7 ou 8.

```
CREATE TABLE CDA_APPLICATION (
        CAP_ID numeric(10) NOT NULL,
        CAP_NAME varchar(150) NOT NULL,
        CAP_DESCRIPTION varchar(200),
        CAP_KEY varchar(150),
        CAP_ICON varchar(300),
        CAP_SHORT_ICON varchar(200)
);

CREATE TABLE CDA_PERMISSION (
        CDA_CDR_ID numeric(10) NOT NULL,
        CDA_CUC_ID numeric(10) NOT NULL,
        CDE_DEFAULT_USE_CASE numeric(1) DEFAULT 0 NOT NULL
);

CREATE TABLE CDA_PROFILE (
        CDA_CDU_ID BIGINT NOT NULL,
        CDA_CDR_ID numeric(10) NOT NULL,
        CDP_DEFAULT_ROLE numeric(1) DEFAULT 0 NOT NULL
);

CREATE TABLE CDA_ROLE (
        CDR_ID numeric(10) NOT NULL,
        CDR_NAME varchar(200) NOT NULL,
        CDR_DESCRIPTION varchar(200),
        CDR_NOTE varchar(200),
        CDA_CAP_ID numeric(10) NOT NULL
);

CREATE TABLE CDA_USE_CASE (
        CUC_ID numeric(10) NOT NULL,
        CUC_ACTION varchar(200) NOT NULL,
        CUC_NAME varchar(100) NOT NULL,
        CUC_DESCRIPTION varchar(200),
        CUC_ACTIVE_ICON varchar(200) NOT NULL,
        CUC_INACTIVE_ICON varchar(200) NOT NULL,
        CUC_MESSAGE_KEY varchar(200) NOT NULL,
        CDA_CAP_ID numeric(10) NOT NULL
);

CREATE TABLE CDA_USER_ACCESS (
        CUA_LOGIN_TIME timestamp NOT NULL,
        CUA_LOGOUT_TIME timestamp,
        CUA_SESSION_ID varchar(150) NOT NULL,
        CUA_IP varchar(30) NOT NULL,
        CDA_CDU_ID bigint NOT NULL
);

ALTER TABLE CDA_APPLICATION ADD CONSTRAINT PK_CDA_APPLICATION
```

```
PRIMARY KEY (CAP_ID);
ALTER TABLE CDA_PERMISSION ADD CONSTRAINT PK_CDA_PERMISSION
PRIMARY KEY (CDA_CDR_ID, CDA_CUC_ID);
ALTER TABLE CDA_PROFILE ADD CONSTRAINT PK_CDA_PROFILE
PRIMARY KEY (CDA_CDU_ID, CDA_CDR_ID);
ALTER TABLE CDA_ROLE ADD CONSTRAINT PK_CDA_ROLE
PRIMARY KEY (CDR_ID);
ALTER TABLE CDA_USE_CASE ADD CONSTRAINT PK_CDA_USE_CASE
PRIMARY KEY (CUC_ID);
ALTER TABLE CDA_APPLICATION
ADD CONSTRAINT UQ_CAP_NAME UNIQUE (CAP_NAME);
ALTER TABLE CDA_ROLEADD CONSTRAINT UQ_CDR_NAME UNIQUE (CDR_NAME);
ALTER TABLE CDA_USE_CASE
ADD CONSTRAINT UQ_CUC_ACTION UNIQUE (CUC_ACTION);
ALTER TABLE CDA_USER
ADD CONSTRAINT UQ_CDU_NAME UNIQUE (CDU_NAME);
CREATE INDEX CDA_CUA_LOGIN_TIME_INDEX
ON CDA_USER_ACCESS (CUA_LOGIN_TIME);
CREATE INDEX CDA_CUA_LOGOUT_TIME_INDEX
ON CDA_USER_ACCESS (CUA_LOGOUT_TIME);
ALTER TABLE CDA_PERMISSION ADD CONSTRAINT FK_CDA_PERMISSION_CDA_ROLE
FOREIGN KEY (CDA_CDR_ID) REFERENCES CDA_ROLE (CDR_ID);
ALTER TABLE CDA_PERMISSION ADD CONSTRAINT
FK_CDA_PERMISSION_CDA_USE_CASE FOREIGN KEY (CDA_CUC_ID) REFERENCES
CDA_USE_CASE (CUC_ID);
ALTER TABLE CDA_PROFILE ADD CONSTRAINT FK_CDA_PROFILE_CDA_ROLE
FOREIGN KEY (CDA_CDR_ID) REFERENCES CDA_ROLE (CDR_ID;
ALTER TABLE CDA_PROFILE ADD CONSTRAINT FK_CDA_PROFILE_CDA_USER
FOREIGN KEY (CDA_CDU_ID) REFERENCES CDA_USER (CDU_ID);
ALTER TABLE CDA_ROLE ADD CONSTRAINT FK_CDA_ROLE_CDA_APPLICATION
FOREIGN KEY (CDA_CAP_ID) REFERENCES CDA_APPLICATION (CAP_ID);
ALTER TABLE CDA_USE_CASE ADD CONSTRAINT
FK_CDA_USE_CASE_CDA_APPLICATION FOREIGN KEY (CDA_CAP_ID) REFERENCES
CDA_APPLICATION (CAP_ID);
ALTER TABLE CDA_USER_ACCESS ADD CONSTRAINT
FK_CDA_USER_ACCESS_CDA_USER FOREIGN KEY (CDA_CDU_ID) REFERENCES
CDA_USER (CDU_ID);
```

ANEXO A – PUBLICAÇÕES E PARTICIPAÇÃO EM EVENTOS RELACIONADOS AO TRABALHO DE TESE

Nas próximas páginas estão inclusas as publicações e as participações em eventos, frutos dessa pesquisa.

Journal of Computational Interdisciplinary Sciences (2011) 2(1): 47-56

Printed version ISSN 1983-8409 / Online version ISSN 2177-8833
http://epacis.org

The VLADA white paper: building an active Virtual Lab for Advanced Data Analysis

Murilo da S. Dantas[1,2], Reinaldo R. Rosa[1], Nilson Sant'Anna[1], Moacyr G. Cereja Jr[3], Thalita B. Veronese[1], Silvia Bianchi[1], Julia C. Rosa[1], Kiril M. Alexiev[4] and José D.S. da Silva[1]

Manuscript received on November 11, 2010 / accepted on March 15, 2011

ABSTRACT

This technical white paper describes the design and initial implementation of a virtual environment for straightforward and robust data analysis intended for students and researchers acting in science and technology. The Virtual Laboratory for Advanced Data Analysis (VLADA) aims to fill a growing demand for scientific mathematical and statistical tools validated and coupled with appropriate high performance computing infrastructure into a single computing environment available on the Web using advanced parallel processing and object-oriented programming. This work proposes to provide: (i) a detailed study on the feasibility of building a such virtual environment with large international access, and (ii) a description of a preliminary single prototype including a standard method for advanced time series analysis. The main steps taken to develop such a laboratory, including preliminary software engineering implementation, are shown in this paper.

Keywords: computational data analysis, virtual systems, advanced parallel processing, object-oriented programming, software engineering.

1 INTRODUCTION

In recent decades, due to technological advancement in sensor data acquisition, the observation and collection of scientific data became almost automatic procedures in many applications. Similarly, the development of advanced mathematical tools in computer enabled the automation of data analysis methods. Hence, the advanced mathematical analysis of massive data bases is rapidly becoming a key component of scientific research and related technological applications. More specifically, there is an increasing use of computers in data processing, visualization and analysis involving new methods to improve data mining in order to provide new scientific knowledge, variability pattern characterization, system identification, control and warning from real-time monitoring. Therefore, there are many computational packages available for scientific data analysis to assist scientists in this effort. Most of them are free software packages (GNU-style) with appropriate syntax that can be downloaded and installed on personal computers as stand-alone (offline, non-Internet) programs only.

Correspondence to: Murilo da S. Dantas – E-mail: murilo.dantas@lac.inpe.br
[1] Lab for Computing and Applied Mathematics (LAC), National Institute for Space Research (INPE), S.J. dos Campos, SP, Brazil.
[2] FATEC, S.J. dos Campos, SP, Brazil.
[3] SESIS Sistemas de Engenharia de Software Ltda, S.J. dos Campos, SP, Brazil.
[4] Institute of Information and Communication Technologies (IICT), Bulgarian Academy of Sciences, Sofia, Bulgaria.

The usual personal computers normally have limited computing resources by using only one or by combining two multicore processors and graphics card with graphics accelerator. Thus, the development of online avaliable electronic packages and corresponding cyber-infrastructure where researchers can perform advanced data analysis are recently required. Note that, for analysing data in a such virtual environment, there is no need to install any scientific package and the task can be performed using a low-cost basic plataform as standard personal computers. In fact, online virtual environments where people can work and interact in a somewhat realistic manner have been gaining great interest and potential as sites for virtual universities, observatories and labs [1]. On this basis, VLADA has been designed as the first virtual data analysis computing environment for the World Wide Web scientific community. The pilot prototype of VLADA includes time series analysis. In the wide range of data science, time series analysis has been important in a number of different scientific communities, the most important of which are statistical physics and nonlinear dynamics with several applications in environmental, space and material sciences, genomics and econometrics.

An easy-to-use virtual advanced time series analysis package should contain a specific selection of common statistical, plotting and modelling functions as, for example, Detrended Fluctuation Analysis (DFA) [2, 3]. It is usually applied for robust characterization of long-range correlation in relatively short nonlinear time series (e.g., [4, 5]). In general, all data analysis packages support a wide variety of traditional methods, but are limited when handling nonlinear time series using advanced tools. Some examples of such mathematical tools are given in the Table 1. Considering that DFA is the most commonly used technique and its implementation is relatively easy, the feature chosen to be implemented in the prototype for VLADA is the computation of the scaling exponent for characterization of long-range correlations in nonlinear time series. In particular, applications of the DFA-technique have therefore gained increasing popularity. Thus, DFA has been selected as the canonical advanced data analysis technique in the prototype of VLADA (see Appendix).

The rest of the paper is organized as follows: Section 2 gives a technical overview of the system. The implemented software engineering resources are described in Section 3. Finally, in Section 4 we discuss the expected short-term results and outline some concluding remarks and challenges that may motivate the further steps in this project.

2 DESCRIPTION OF THE VIRTUAL LABORATORY

2.1 A technical overview

The VLADA working group (VWG) has been organized into the following four initial teams: Management Resources Team (MRT), High Performance Networks Team (HPNT), Software Engineering Team (SET) and Data Analysis Algorithm Team (DAAT). In this logical collection of basic work tasks, it is worthly to mention that the main goal of the DAAT is to deploy, certify and develop, more efficiently, useful data analysis algorithms in a transparent and public collaborative forum (analogous to the teams working on open-source software development projects). It should be emphasized that the success of the project critically depends on the VWG Long-Term Strategic Plan which should provide a virtual laboratory that must have a final structure that is as close as possible to an actual laboratory for data analysis procedures. Ongoing improvements, both in software and hardware, and in quality and speed of internet connections, will enable us to enhance the virtual laboratory prototype and complement it with new elements which are defined as software modules, called Virtual Labs (VLab). A possible structure of VLADA, containing six virtual labs, is the following:

- VLab1: Visualization Tools and Standard Data Analysis;
- VLab2: Advanced Statistical Tools;
- VLab3: Advanced Tools for Time Series Analysis;
- VLab4: Advanced Tools for Image and Spatio-temporal Analysis;
- VLab5: Advanced Tools for Multivariate Data Systems;
- VLab6: Advanced Data Mining Techniques.

VLab1 is divided into visualization and standard data analysis as, for instance, the calculation of statistical moments, histograms and autocorrelation functions. Once the data is activated in VLADA specifying the required analysis, the chosen tasks from a large list of traditional data analysis routines (found in many text books of classical statistics) are generated automatically by VLab1. Hence, as shown in Figure 1, VLab1 is expected to enclose the major amount of standard techniques. Moreover, all techniques in this module are supposed to be well known in the academic community. Consequently, the VLab1 implementation procedure although simple, is quantitatively more complex and, from the point of view of a mature user, its merit will be hard to be assessed from a virtual environment. However, due to academic and completeness purposes of this project, we understand that VLab1 should be part of the system in a long term open project as VLADA is.

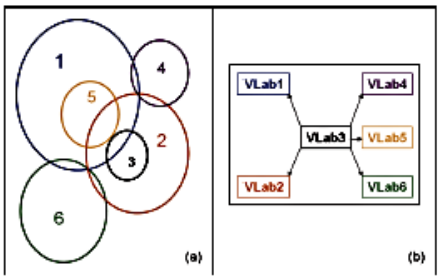


Figure 1 – (a) The VLab domains and (b) the strategy of implementation after VLab 3.

Table 1 – A reference list of relatively recent techniques that have been used, specially in physical sciences, for advanced time series analysis over the last twenty years.

Technique	Main Output	Year [Main Ref.]
Detrended Fluctuation Analysis (DFA)	Scaling Exponent	1992 [2-9]
Multifractal Spectral Analysis (MSA)	Singularity Spectra	1992 [10-12]
Gradient Spectral Analysis (GSA)	Gradient Asymmetry Spectra	1999 [13-15]
Trajectory Parallel Method (TPM)	Trajectory Curve	1999 [16,17]

VLab2 may be a complementary open module to pursue as much as possible the most important advanced statistical tools, previously validated by the DAAT. A schematic overview of all VLabs as interconnected and complementary software modules is given in Figure 1, where VLab3 has been selected to be the initial core of the VLab system. After development of the prototype module (VLab3), the complementary modules (1, 2, 4, 5 and 6) will be simultaneously developed, so that the entire systems will be expanded further. This is illustrated in Figure 1(b). Besides the DFA, other some typical advanced techniques which would be compatible with VLab3 are listed in Table 1.

In addition, each VLab must contain the all necessary information to be carried out the virtual analysis and it must be as easy as possible to help those users who may not be familiar with the available analytical tools. In this way, the users will be able to access independently a Virtual Knowledge Repository (VKR) composed of three virtual knowledge libraries (VKLib):

- VKLib1: Techniques for Data Analysis: Basic Text, References and Links;
- VKLib2: Packages and Codes;
- VKLib3: Samples and Data Repository.

These multi-module elements might be used to enable data analysers (scientists and students) to perform virtual data analysis as a straightforward procedure. Therefore, the architecture implementation group should explore the variety of possible network topologies and discuss methods of collecting information (data and analytical requirements) from users such as scheduled and on-demand harvesting, taking into account management overhead, adaptability and timeliness. It should also be determined how to combine searching and harvesting services as part of a comprehensive solution. Some multi-tier network topologies are available performing combination of repositories, registries, and access points. A multiple repository with single registry, corresponding appropriate topology and archive perspective should be the first to be addressed. The mathematical tools for data analysis will define a kernel accessed through an interface, as shown in Figure 2.

VLADA has been designed to work as a dedicated system that allows the registration of users through a robust interface. The registered users perform a Basic Logical Procedure (BLP) for Data Analysis Services (DAS) based on Laboratory Modules (VLab) and Knowledge Library Modules (VKLib), as shown in Figure 3. In this figure, an important part of the process is the step 6. In this last procedure the users could send a standard statement to VLADA, reporting the quality of the obtained results.

As it can be seen in Figure 4, VLADA has been designed to be an active collaborative open project. It may require additio-

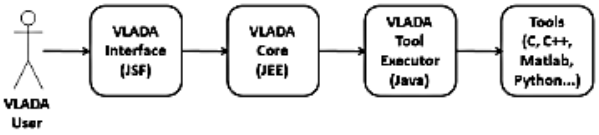


Figure 2 – The routines at the core of VLADA are being developed in JEE, while the interface will be developed in JSF. The data analysis algorithms, usually written in $C/C++$ or in higher level scientific computing languages, will be executed from the VLADA Tool Executer developed in JAVA.

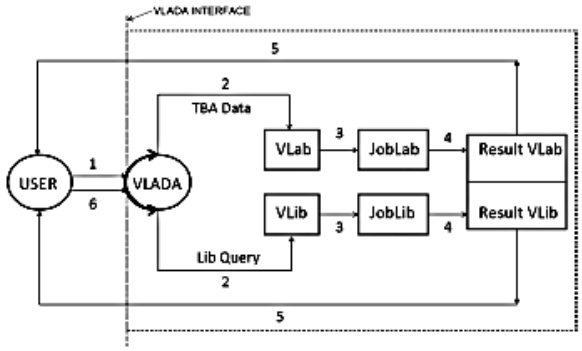


Figure 3 – A registered user access VLADA for VLab and/or VKLib services. VLADA returns to the user the result after Job module response. Two possible inputs from users are "to be analyzed data" (TBA Data) and "Library Information Requirement" (LI Query). The six basic logical steps are identified. The user is an active component only for steps 1, 2 and 6. In the step 6 the user returns to VLADA the service evaluation.

nal partners, for client advances, in the allocation of hardware and software development in order to form a global network of virtual labs, increasing the quantity and variety of analysis tools and computational infrastructure. Nevertheless, as shown in Figure 5, a preliminary virtual server environment for routine integrity should be designed to optimize server utilization in real time allowing management of physical and virtual resources. The minimal software and hardware requirements for an expandable local prototype based on LAC-INPE network and LAC-INPE community has been designed based on DFA technique and INPE's users only.

2.2 The prototype from the user's interface perspective

The VLADA prototype website (VPW) will provide an easy to use interface to access the VLab3 and its respective VKLib prototype resources. Those resources will be restricted to the VLab3 where the only technique avaliable will be DFA. The VPW will lead the INPE's user through step-by-step instructions for doing common tasks with the application of DFA. To getting started with VLADA the first step is to get the username and password using the register link. The registered user will access the Command Line Interface (CLI), which provides immediate access to the VLab (DFA) and VKLib (on DFA). The web page for each service introduces its capabilities and provides online documentation of the task.

The default task using VLab is to post your own time series, in ASCII-like format, to be analyzed by means of the DFA technique. During this procedure the user will get a WebServiceId which should be used in order to get the result of his analysis in the ResultVLab. For users without expertise on DFA, the files from VKLib are distributed as a GZip-compressed *tar* file containing source code and complete documentation on DFA. An outline of a virtual analysis, via VLab3 based on DFA, using the VPW may be seen in Figure 6.

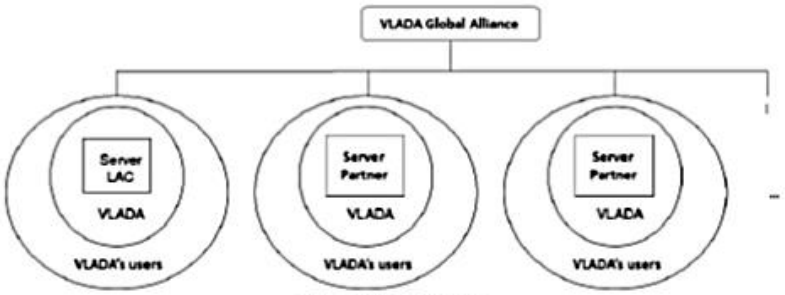


Figure 4 – VLADA Global Alliance.

Figure 5 – VLADA prototype based on LAC-INPE.

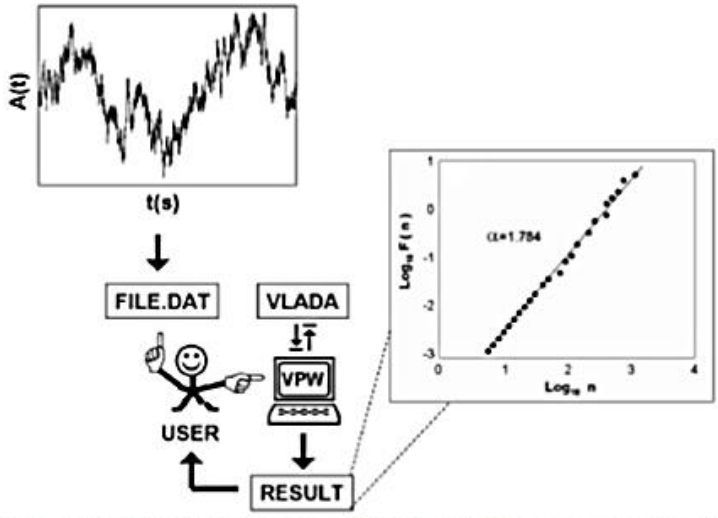


Figure 6 – A virtual Detrended Fluctuation Analysis using the VLADA prototype web site. In this example, the input is a generic time series: Amplitude × time (in seconds) written as a *.dat* format. The output is the DFA slope with the respective scaling exponent used to characterize the persistence level of the fluctuations observed in a nonlinear time series (see Appendix).

The Lab for Computing and Applied Mathematics (LAC) at INPE has been providing an appropriate initial environment to deploy the VLADA prototype. This environment is composed by 4 storages HP with 3.6 TB of storage space and 3 servers HP 2U with the following configuration: 2 processors quad-core, 32 GB RAM and 4 TB hard drives. To allow multiple tests for the VLADA prototype performance the initial software engineering infrastructure has been developed by the SE team and will be described in the next section.

3 SOFTWARE ENGINEERING RESOURCES

Selected process techniques have been used by SET to improve the quality of the VLADA prototype development effort. The documented collection of policies, processes and procedures developed/chosen by SET follows the rigid protocols for software development methodology (SDM) and system development life cycle (SDLC). In this project, we are defining some issues regarding the state of art in Software Engineering. These elements are related to VLADA's functional and non-functional requirements, architecture, open source components, etc. Some examples of these elements are:

- VLADA environment should have a very sophisticated "Access Control System", in such a way that users can have profiles, roles and permissions. This component will allow users to effectively and efficiently fulfill their jobs (high level of Usability) while accomplishes security requirements.
- We should provide a very robust architecture to keep VLADA working reliably as long as possible. In order to materialize this issue, we can use, for example, "Design Patterns", dividing the system architecture in "Layers" and using Software Engineering best practices.
- VLADA's architecture will need some components like a data base management system and an application server. For these components, we are considering PostgreSQL and JBOSS, both open source components.

The initial VLADA's software project planning provides a set of diagrams to depict software structures graphically. Figure 7 shows the main components of VLADA (high level software modules, interfaces and kernels), the *VLADA-ComponentDiagram*, having the access flow from tool executor for two possible advanced analytical tools listed in Table 1. The distribution of software modules in the hardware infra-structure is shown in Figure 8, the so-called *VLADA-DeploymentDiagram*. The possible states of a generic analytical tool, from its initial deposition to its certification, is shown as the *VLADA-ToolStateChartDiagram* (Fig. 9). The proposed functionalities defining the VLADA system are organized in the *VLADA-UseCaseDiagram* shown in Figure 10. These diagrams are shown as products that have been implemented in the present phase of this project.

The SET diagrams provide a single framework for organizing, relating, and viewing several diverse aspects of the project. It has been successfully used providing: a framework for project planning, identification of intermediate and final deliverables, a systematic method for deriving a work breakdown structure and a framework for tracking progress in terms of completed/not completed status of all activities. They are also supporting the tracing of requirements through all stages of VLADA software development.

4 EXPECTED RESULTS AND CONCLUDING REMARKS

The initial VLADA's team aspires to cooperate with others in the development of international standards for VLADA. We intend to extend the interactions of INPE, FATEC-SJC and IICT-BAC with partners already interested in this project, as groups from University of Western Ontario (Canada), Royal Institute of Technology (Sweden), Russian Academy of Science (Russia), Massachusetts Institute of Technology (USA), University of Louisiana (USA), and Universite de Caen (France). It will facilitate the international exchange of VLADA resources and allow for the standardization of the analytical tools and data format. It is also intended that, at some point in next year, the VLADA prototype will be accessible through the Internet.

For this open on-line implementation, we must evaluate computer languages, hardware, technical and algorithmic complexity and interface with users and, thus, propose scenarios for providing the most appropriate virtual environment. Formally, the next steps to be taken are the following:

- To establish criteria for the development of large systems in parallel to provide a cloud of tools for advanced data analysis;
- To make a detailed survey of technologies for software and hardware to be used in the project with growing perspectives.
- To implement analysis tools according to the technologies of software and hardware tested.

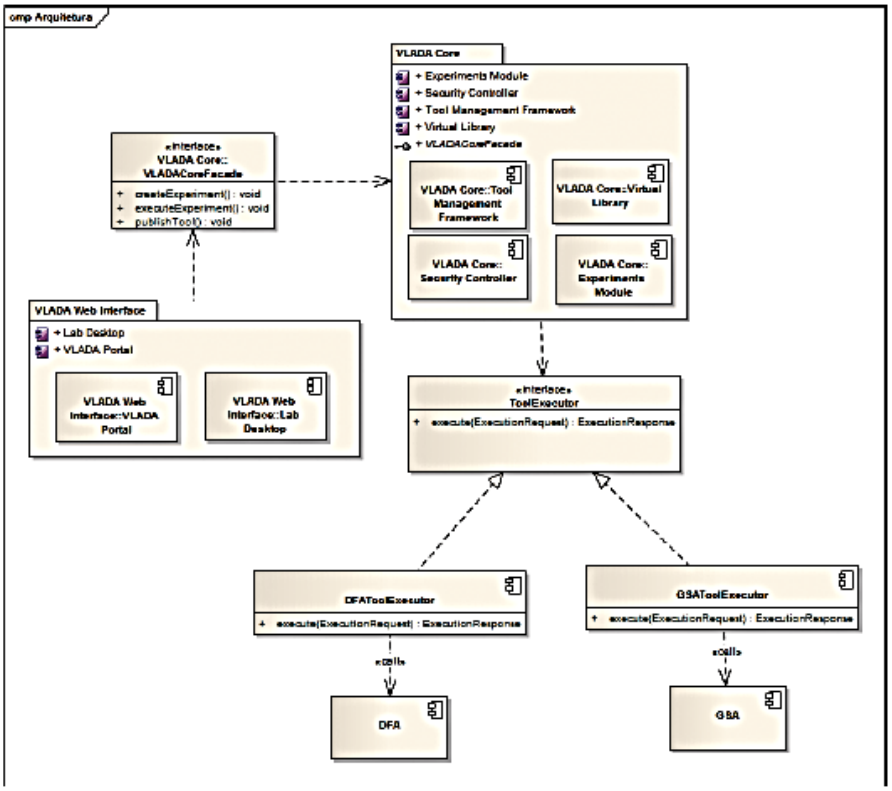


Figure 7 – VLADA ComponentDiagram.

- To develop the prototype of VLADA based on the available hardware described in Section 2.

It is expected that the results of this project will be of direct interest and use to the scientific community, especially to researchers who want to perform straightforward time series analysis having both the minimal hardware and software requirements at the moment when the analysis should be performed. Also, it will be important for academic purposes, including non-specialist users, but have no affinity with mathematical or computational technology for the development of data analysis tools. That is, VLADA attempts to offer a contribution to data analysis, developing new methodologies and providing a distributed virtual environment for the application of advanced tools not easily found in conventional data analysis packages. As a result of this effort, we will provide a new and complementary infrastructure to allow multiple users to search, discover, view, and share technical and instructional content on advanced data analysis.

ACKNOWLEDGMENTS

This work has been partially supported by LAC-INPE. The authors are very grateful to M.D.Todorov, E.F.P. da Luz, R.R. de Carvalho, H.F. Campos Velho and H.V. Capelato for fruitful technical discussions on VLADA, and to N.L. Vijaykumar and an anonymous reviewer, for helpful suggestions and comments in improving this manuscript.

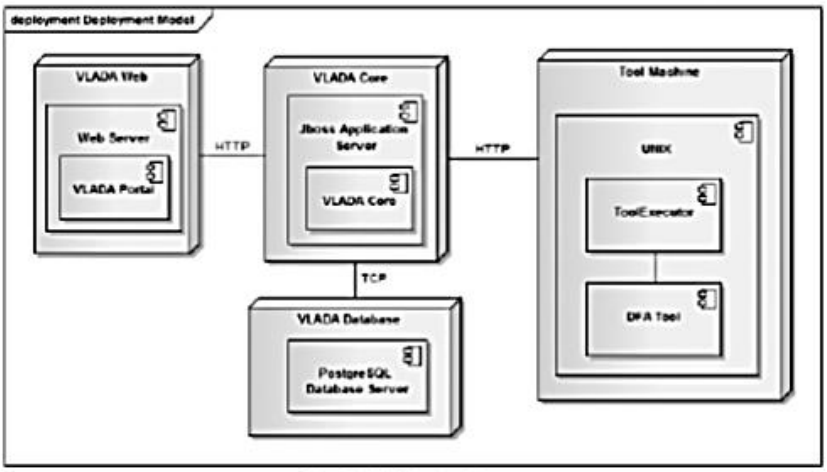


Figure 8 – VLADA DeploymentDiagram.

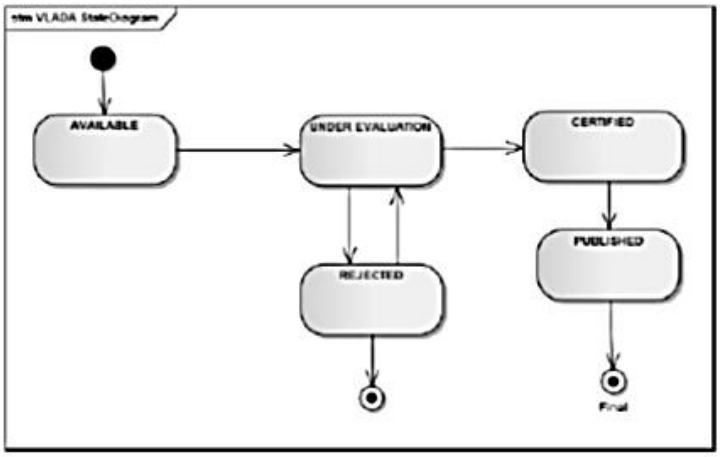


Figure 9 – VLADA ToolStateChartDiagram.

APPENDIX – Detrended Fluctuation Analysis

Detrended Fluctuation Analysis (DFA) measures the so-called *scaling exponents* from non-stationary time series. It is useful for characterizing variability patterns that appear to be due to long-range temporal correlations. The DFA technique has been used, in the last twenty years, to compute the values of the scaling exponents in several applications from physiological data to signals in physics and finance (e.g., [6, 7, 4, 8]).

The standard DFA algorithm ([9]), is composed of four main computational operations starting here on a discrete series of amplitudes $\{A_i\}$:

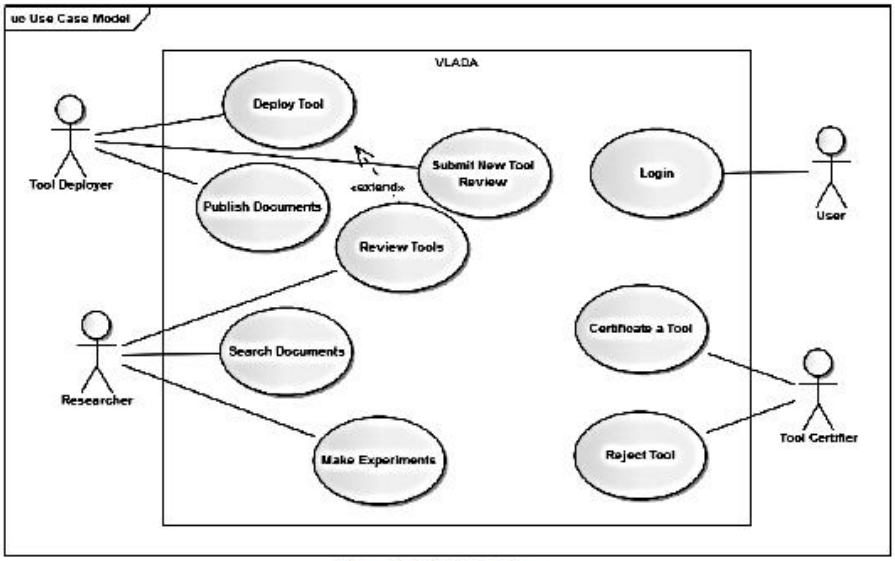


Figure 10 – VLADA UseCaseDiagram.

1. *Discrete Integration and Windowing*: Compute the cumulative representation of $\{A_i\}$ as

$$y(k) = \sum_{i=1}^{k} (A_i - \langle A \rangle),$$

with $k = 1, 2, \ldots, N$, where

$$\langle A \rangle = \sum_{i=1}^{N} A$$

is the average of $\{A_i\}$. Using an arbitrary local window of length n, divide $y(k)$ into non-overlapping $N_n = int(N/n)$ sub-interval y_j $(j = 1, 2, \ldots, N_n)$. Note that each sub-interval y_j has length n and N may not be the integer multiple of n. Then, the series $y(k)$ is divided once more from the opposite side to make sure all points are addressed, performing at the end of this operation $2N_n$ sub-intervals on each profile.

2. *Fitting and Variance*: In each sub-interval, calculate the least-square fits $p_j^m(k)$ where m is interpreted as the *order of the detrended trend*, and compute the cumulative deviation series in every sub-interval, where the trend has been subtracted: $y_j(k) = y(k) - p_j^m(k)$. Then, calculate the variance of the $2N_n$ sub-intervals for $j = 1, 2, \ldots, N_n$ and $j = N_n + 1, N_n + 2, \ldots, 2N_n$.

3. *Fluctuation*: Calculate the average of all the variances and the square root. Then get the fluctuation function $F(n)$:

$$F(n) = \left[\frac{1}{2N_n} \sum_{j=1}^{2N_n} F^2(j, n) \right]^{1/2}. \quad (1)$$

4. *Scaling Exponent*: Perform again, recursively, computation from windowing to calculation of corresponding $F(n)$ with different $n([N/4] > n = 2m + 2)$ box lengths. In the presence of power law: $F(n) = Kn^{\alpha}$, $F(n)$ increases linearly with increasing n. Then, get the slope α using the linear least-square regression on the double log plot $\log F(n) = \log K + \alpha \log n$ (see Fig. 6). The scaling exponent $\alpha = 0.5$ characterizes that the fluctuations are uncorrelated. When $\alpha > 0.5$ the auto-correlation is persistent. Values of $\alpha < 0.5$ corresponds to long-term anticorrelations, meaning that large values are most likely to be followed by small

values and vice versa characterizing anti-persistence. Higher values of alpha characterizes stronger auto-correlations in the signal ($\alpha > 1$ indicates a non-stationary local average of the data).

REFERENCES

[1] BAINBRIDGE WS. 2007. The Scientific Research Potential of Virtual Worlds. Science 317(5837): 472–476. DOI: 10.1126/science.1146930.

[2] PENG CK, BULDYREV S, GOLDBERGER A, HAVLIN S, SCIORTINO F, SIMONS M & STANLEY HE. 1992. Long-range correlations in nucleotide sequences. Nature, 356: 168.

[3] HU K, IVANOV PC, CHEN Z, CARPENA P & STANLEY HG. 2001. Effect of trends on detrended fluctuation analysis. Phys. Rev. E, 64: 011114.

[4] BARONI MPMA, DE WIT A & ROSA RR. 2010. Detrended fluctuation analysis of numerical density and viscous fingering patterns. EPL 92, 64002. DOI: 10.1209/0295-5075/92/64002.

[5] VERONESE TB, ROSA RR, BOLZAN MJA, FERNANDESC FCR, SAWANT HS & KARLICKY M. 2011. Fluctuation analysis of solar radio bursts associated with geoeffective X-class flares. Journal of Atmospheric and Solar-Terrestrial Physics, in press. doi:10.1016/j.jastp.2010.09.030.

[6] BUNDE A et al. Phys. Rev. E, 85: 3736.

[7] BULDYREV SV, GOLDBERGER AL, HAVLIN S, MANTEGNA RN, MATSA CK & PENG C-K et al. 1995. Phys. Rev. E, 51: 5084.

[8] BAI MY & ZHU HB. 2010. Physica A, 389: 1883.

[9] PENG C-K, BULDYREV SV, HAVLIN S, SIMONS M, STANLEY HE & GOLDBERGER AL. 1994. Phys. Rev. E, 49: 1685.

[10] MUZY JF, BACRY E & ARNEODO A. 1991. Wavelets and Multifractal formalism for singular signals: Application to turbulence data. Phys. Rev. Lett., 67(25): 3515–3518.

[11] MALLAT SG & HWANG WL. 1992. Singularity Detection and Processing with Wavelets. IEEE Trans. on Information Theory, 38: 617–643.

[12] BOLZAN MJA, ROSA RR & SAHAI Y. 2009. Multifractal analysis of low-latitude geomagnetic fluctuations. Annales Geophysicae, 27: 569–576.

[13] ROSA RR, SHARMA AS & VALDIVIA JA. 1999. Characterization of asymmetric fragmentation patterns in spatially extended systems. Int. J. Mod. Phys. C, 10: 147–163.

[14] ROSA RR, KARLICKY M, VERONESE TB, VIJAYKUMAR NL, SAWANT HS, BORGAZZI AI, DANTAS MS, BARBOSA EBM, SYCH RA & MENDES O. 2008. Gradient pattern analysis of short solar radio bursts. Adv. Space Res., 42(5): 844–851.

[15] DANTAS MS. 2010. Análise Espectral de Padrões-Gradiente de Séries Temporais Curtas, (INPE-15676-TDI/1450). Dissertation (MSc in Applied Computing) – National Institute for Space Research, São José dos Campos, 2009. Available at: <http://URLIB.NET/SID.INPE.BR/MTC-M18@80/2009/02.05.10.55>. Acessed: 20 MAR. 2010.

[16] FUJIMOTO Y & IOKIBE T. 1999. Measurement of determinism in time series by chaotic approach and its applications. Int. J. of Advanced Computational Intelligence, 3(1): 50–55.

[17] FUJIMOTO Y & IOKIBE T. 2000. Evaluation of deterministic property of time series by the method of surrogate data and the trajectory parallel measure method. IEICE Trans. Fundamentals, vol. E83-A, No 2: 343–349.

Biogeosciences (B)

B32A
A3
Wednesday
1030

Linking Hydrology and Nutrient Cycling in Large Wetland Ecosystems I

Presiding: L O Sternberg, Biology, University of Miami, Coral Gables; V Engel, Everglades National Park, Homestead

B32A-01

TWO ANALYSIS OF HYDRO-ECOLOGICAL SEASONAL VARIATION IN THE PANTANAL WETLAND, BRAZIL

*Calheiros, D F (debora@cpap.embrapa.br), Embrapa Pantanal, Corumbá, Brazil
Oliveira, M D (mmarci@cpap.embrapa.br), Embrapa Pantanal, Corumbá, Brazil
Dantas, M (murilodantas06@gmail.com), INPE, São José dos Campos, Brazil
Rosa, R R (reinaldo.rosa@pq.cnpq.br), INPE, São José dos Campos, Brazil
Futter, M N (martyn.futter@vatten.slu.se), Swedish University of Agricultural Sciences, Uppsala, Sweden

The Pantanal wetland is one of the largest wetlands in the world (ca. 140.000 km2), considered as a National Heritage by the Brazilian Federal Constitution and Humanity Heritage and Biosphere Reserve by the United Nations. For understanding the hydrology, biogeochemistry, and ecology of the extensive floodplains of tropical regions, such as Pantanal Wetland, information is necessary regarding the spatial and temporal variability of inundation patterns, and also essential to managing this complex ecosystem. Most of the time series of water quality data collected from this natural system result in partial data set, thus compromising the performance of usual statistical analysis. The main goal of this research is to apply two different methods: 1- a new computational method for short time series analysis, showing non-linear behavior in the time, amplitude and frequency domains to understand the hydro-ecological functioning of this river-floodplain system. The Gradient Spectral Analysis (GSA), combines two mathematical techniques, the so-called Gradient Pattern Analysis (GPA) and the Wavelet Multiresolution Analysis (WMA). The GSA classifies different non-linear regimes taking into account short time series samples generated from dynamic processes previously associated with chaotic and stochastic models, and classifies the pattern variability from high to low scaling and asymmetry fluctuation. We classified short times series (240 events) related to long term hydro-ecological research of the Paraguay River, the main river of the Pantanal floodplain. The data set is based on hydrological monitoring at Ladário Station, carried out daily by the Brazilian Navy since 1900, and water quality monitoring of 5 main limnological parameters, carried out monthly during 20 years (1989-2009). The preliminary results showed that the dynamics of Dissolved Oxygen and Electrical Conductivity were correlated directly with the River Level, the dynamics of Dissolved Carbonic Gas and Total Phosphorus were very close, but Total Nitrogen presented a larger distance, showing high asymmetry; 2- an analysis correlating the seasonal variation of discharges and yearly discharge variation and nutrients concentration of Paraguay River. The data base set is also based on river level recording at Ladário Station (LD), and its level vs. discharge equation, as well as on a other upper river station (São Francisco - SF) with discharge values measured "in situ" with the variation of 6 main nutrient parameters (NTotal, NO3, NO2, NH4, PTotal, and PO4) sampled monthly during 1989-2009. The preliminary results showed that the dynamics of nutrients as significantly correlated with the discharge dynamics, but in a better way in the SF station. Thus, both methods of hydro-ecological understanding of nutrients dynamics can be correlated for further purpose on climate change forecasting and correlation with primary and secondary aquatic production. Grants: LTER/CNPq Institutions: Embrapa Pantanal, INPE, Swedish University of Agricultural Sciences

B32A-02

The effects of hydro-biogeochemical processes on groundwater chemistry including nutrient concentrations in an oligotrophic wetland: Examples from the Florida Everglades

*Price, R M (pricer@fiu.edu), Earth and Environment, Florida International University, Miami, FL, USA
Sullivan, P L (sullivap@fiu.edu), Earth and Environment, Florida International University, Miami, FL, USA
Zapata, X (xzapa001@fiu.edu), Earth and Environment, Florida International University, Miami, FL, USA

Concentrations of nutrients and other chemical constituents are often elevated in groundwater as compared to surface water due to a variety of hydro-biogeochemical processes including but not limited to water-rock interactions, organic matter remineralization, evapotranspiration processes, and seawater intrusion. The interaction of the high nutrient groundwater with either plant roots or the surface water can lead to biogeochemical hotspots in an otherwise oligotrophic ecosystem, which in turn can lead to landscape patterning. The objective of this presentation is to describe the hydro-biogeochemical processes that lead to higher nutrient and chemical concentrations in groundwater in two regions of the Florida Everglades. The first region is located in the freshwater ridge-slough-tree island continuum of the Everglades, while the other is located along the coastal mangrove ecotone. In each region, groundwater and surface water levels were monitored along with concentrations of ions, nutrients, and the stable isotopes of oxygen and hydrogen. Within the tree islands, the process of evapotranspiration results in concentrating ions within the groundwater. Though groundwater nutrient concentrations beneath tree islands are also elevated, the mechanism driving the enrichment is likely due to organic remineralization within the soil zone. Along the coastal mangrove ecotone, seawater intrudes into the underlying aquifer causing salinization of the groundwater. Water-rock interactions such as calcium carbonate dissolution and ion exchange reactions release phosphorus to the groundwater. The phosphorus is then available to the mangroves roots that penetrate into the groundwater or to the overlying surface water through groundwater discharge.

www.fiu.edu/~pricer

B32A-03

2009 LTER All Scientists Meeting
"Integrating Science and Society in a World of Constant Change, Sept. 14-16th 2009, Estes Park CO"

Gradient Spectral Analysis for Short Time Series of Hydro-Ecological Seasonal Variation in the Pantanal Wetland, Brazil

Poster Disciplines/Format: Altered moisture regimes Ecological Modeling ILTER Site Description

Poster Number: 28
Presenter/Primary Author: Débora Calheiros
Co-Authors: Oliveira, M.D.
Dantas, M.
Rosa, R.R.

The Pantanal wetland is one of the largest wetlands in the world (ca. 140.000 km2). Most of the time series collected from this natural system result in partial data set, specially water quality, thus compromising the performance of usual statistical analysis. The main goal of this research is to apply a new computational methodology for short time series analysis, showing non-linear behaviour in the time, amplitude and frequency domains to understand the hydro-ecological functioning of this river-floodplain system. The applied methodology, called Gradient Spectral Analysis (GSA), combines two mathematical techniques, the so-called Gradient Pattern Analysis (GPA) and the Wavelet Multiresolution Analysis (WMA). The GSA classifies different non-linear regimes taking into account short time series samples generated from dynamical processes previously associated with chaotic and stochastic models. We classified short times series (240 events) related to long-term hydro-ecological research of the Paraguay River, the main river of the Pantanal floodplain. From preliminary results, we correlate the seasonal variation of the river level records with the variation of 25 parameters of water quality, sampled monthly during 20 years (1989-2009). There is a good indication that this analysis will be robust enough for the prediction of behaviour variables and further application on climate change forecasting and correlation with primary and secondary aquatic production.

Related Materials and Graphics:
Paraguay River basin.doc

P1.jpg

Session: Hydrological understanding and modelling of flood pulse dynamics (2)
Presentation: Oral

Gradient Spectral Analysis of Hydro-Ecological Seasonal Variation in the Pantanal Wetland, Brazil

Calheiros D.F. [1], Oliveira M.D.[1], Dantas M.[2], Rosa R.R. [2]

[1] Embrapa Pantanal
[2] INPE

The Pantanal wetland is one of the largest wetlands in the world (ca. 140.000 km2), considered as a National Heritage by the Brazilian Federal Constitution and Humanity Heritage and Biosphere Reserve by the United Nations. For understanding the hydrology, biogeochemistry, and ecology of the extensive floodplains of tropical regions, such as Pantanal Wetland, information is necessary regarding the spatial and temporal variability of inundation patterns, and also essential to managing this complex ecosystem. Most of the time series of water quality data collected from this natural system result in partial data set, thus compromising the performance of usual statistical analysis. The main goal of this research is to apply a new computational methodology for short time series analysis, showing non-linear behavior in the time, amplitude and frequency domains to understand the hydro-ecological functioning of this river-floodplain system. The applied methodology, called Gradient Spectral Analysis (GSA), combines two mathematical techniques, the so-called Gradient Pattern Analysis (GPA) and the Wavelet Multiresolution Analysis (WMA). The GSA classifies different non-linear regimes taking into account short time series samples generated from dynamic processes previously associated with chaotic and stochastic models, and classifies the pattern variability from high to low scaling and asymmetry fluctuation. We classified short times series (240 events) related to long term hydro-ecological research of the Paraguay River, the main river of the Pantanal floodplain. The data set is based on hydrological monitoring at LadÆrio Station, carried out daily by the Brazilian Navy since 1900, and the monitoring of 30 parameters of water quality, sampled monthly during 20 years (1989-2009). Initially we correlated the seasonal variation of the river level records with the variation of 5 main limnological parameters. The preliminary results showed that the dynamics of Dissolved Oxygen and Electrical Conductivity were correlated directly with the River Level, the dynamics of Dissolved Carbonic Gas and Total Phosphorus were very close, but Total Nitrogen presented a larger distance, showing high asymmetry. Based on these first round results, there is a good indication that this analysis will be robust enough for the prediction of behavior variables for further application on climate change forecasting and correlation with primary and secondary aquatic production. Grant: LTER/CNPq

Correspondence to: Dr DØbora Calheiros debora@cpap.embrapa.br