sid.inpe.br/mtc-m19/2013/02.14.13.23-TDI

METODOLOGIAS DE MINERAÇÃO DE DADOS EM ANÁLISE CLIMÁTICA

Heloisa Musetti Ruivo

Tese de Doutorado do Curso de Pós-Graduação em Computação Aplicada, orientada pelos Drs. Fernando Manuel Ramos, Haroldo Fraga de Campos Velho, e Gilvan Sampaio aprovada em 19 de fevereiro de 2013.

URL do documento original:
<http://urlib.net/8JMKD3MGP7W/3DHME8E>

INPE
São José dos Campos
2013

PUBLICADO POR:

Instituto Nacional de Pesquisas Espaciais - INPE
Gabinete do Diretor (GB)
Serviço de Informação e Documentação (SID)
Caixa Postal 515 - CEP 12.245-970
São José dos Campos - SP - Brasil
Tel.:(012) 3208-6923/6921
Fax: (012) 3208-6919
E-mail: pubtc@sid.inpe.br

sid.inpe.br/mtc-m19/2013/02.14.13.23-TDI

METODOLOGIAS DE MINERAÇÃO DE DADOS EM ANÁLISE CLIMÁTICA

Heloisa Musetti Ruivo

Tese de Doutorado do Curso de Pós-Graduação em Computação Aplicada, orientada pelos Drs. Fernando Manuel Ramos, Haroldo Fraga de Campos Velho, e Gilvan Sampaio aprovada em 19 de fevereiro de 2013.

URL do documento original:
<http://urlib.net/8JMKD3MGP7W/3DHME8E>

INPE
São José dos Campos
2013

Dados Internacionais de Catalogação na Publicação (CIP)

Ruivo, Heloisa Musetti.

R859m Metodologias de mineração de dados em análise climática / Heloisa Musetti Ruivo. – São José dos Campos : INPE, 2013.

xx + 101 p. ; (sid.inpe.br/mtc-m19/2013/02.14.13.23-TDI)

Tese (Doutorado em Computação Aplicada) – Instituto Nacional de Pesquisas Espaciais, São José dos Campos, 2013.

Orientadores : Drs. Fernando Manuel Ramos, Haroldo Fraga de Campos Velho, e Gilvan Sampaio.

1. mineração de dados. 2. análise estatística. 3. extração de conhecimento. 4. árvore de decisão. 5. seca na Amazônia. 6. precipitação extrema. 7. mudanças climáticas. I.Título.

CDU 519.72

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

Computação Aplicada

Dr. Nandamudi Lankalapalli Vijaykumar

Presidente / INPE / SJCampos - SP

Dr. Fernando Manuel Ramos

Orientador(a) / INPE / SJCampos - SP

Dr. Haroldo Fraga de Campos Velho

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

Dr. Gilvan Sampaio de Oliveira

Orientador(a) / CCST/INPE / Cachoeira Paulista - SP

Dr. Stephan Stephany

Membro da Banca / INPE / SJCampos - SP

Dr. João Carlos Carvalho

Convidado(a) / ANA / Brasília - DF

Dr. Tércio Ambrizzi

Convidado(a) / USP / São Paulo - SP

Este trabalho foi aprovado por:

( ) maioria simples

(X) unanimidade

Aluno (a): *Heloisa Musetti Ruivo*

São José dos Campos, 19 de Fevereiro de 2013

"A coisa mais bela que podemos sentir é o mistério. É a fonte da verdadeira arte, ciência e religião.".

Einstein

AGRADECIMENTOS

Gostaria de agradecer primeiramente aos meus orientadores, Prof. Dr. Fernando Manuel Ramos e Prof. Haroldo Fraga de Campos velho, pela confiança depositada em mim para desenvolver este trabalho. Admiro a disponibilidade e dedicação prestada a mim durante todos estes anos de pesquisa.

Agradeço também ao meu co-orientador Dr. Gilvan Sampaio de Oliveira do CCST/INPE com sua contribuição, análise e comentários da tese. Sua participação foi muito importante e abriu um grande leque para trabalhos futuros.

À cooperação de muitos alunos da pós-graduação do INPE e do CPTEC que com muita humidade compartilharam seus conhecimentos. Uma colega que merece destaque é Christiane F. Lemos Lima, do Instituto Federal do Maranhão, que me ajudou muito em uma etapa de minha pesquisa. Agradeço também ao LAC e ao INPE pelos recursos disponibilizados

Especialmente ao apoio, à confiança e à força de minha mãe, pessoa que admiro muito.

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

RESUMO

A ciência tem disponível atualmente uma grande quantidade de dados de diversas origens, particularmente em meteorologia e ciência do clima, que nos permitem desenvolver estratégias eficazes de descoberta de conhecimento. Na área ambiental, eventos extremos meteorológicos acarretam um impacto na vida humana e na economia. Este trabalho utiliza metodologias de mineração de dados para analisar fenômenos meteorológicos extremos. O objetivo é identificar os fatores climatológicos relevantes que influenciaram tais eventos. Duas metodologias serão avaliadas: classificação estatística; e árvores de decisão. O método estatístico está implementado na ferramenta computacional BRB-ArrayTools, software desenvolvido na área de bioinformática, aqui adaptado e aplicada à área ambiental. Algoritmos de árvores de decisão, estão implementados no software WEKA, cujo objetivo é utilizar os atributos mais relevantes gerados pela análise estatística e gerar um classificador de árvore de decisão. Estas técnicas foram testadas para análise das grandes secas do Amazonas ocorridas em 2005 e 2010, e na precipitação extrema ocorrida em Santa Catarina em 2008, onde se pode apontar alguns parâmetros climatológicos responsáveis pelos eventos. A literatura indica que mudanças climáticas (decorrente do aquecimento global antropogênica, vulcanismo, alteração do eixo de rotação do planeta, alteração da atividade solar e/ou outras causas geofísicas ou geológicas) alteram a intensidade e frequência dos eventos extremos. O objetivo aqui é analisar os eventos extremos citados e contribuir com um conjunto de ferramentas de avaliação com o Instituto Nacional de Ciência e Tecnologia para Mudanças Climáticas (INTC-MC).

DATA MINING METHODOLOGIES IN CLIMATE ANALYSIS

ABSTRACT

Science has currently available a large amount of data from various sources, particularly in meteorology and climate science, allowing us to develop effective strategies for knowledge discovery. In the environmental area, extreme weather events imply an impact on human life and economy. This thesis uses data mining methodologies to analyze extreme meteorological phenomena. The goal is to identify the relevant climatological factors which influenced such events. Two methods will be evaluated: statistical classification and decision trees. The statistical method is implemented in the computional tool called BRB-ArrayTools, developed by statisticians experienced in the biology, here adapted and applied to environmental problems. Decision tree algorithms (implemented in WEKA), whose purpose is to use the most relevant attributes generated by the statistical analysis and generate a decision tree classifier. These techniques were tested to analyze the great droughts in Amazon occurred in 2005 and 2010, and in an extreme precipitation event occurred in Santa Catarina in 2008, which may point out some climatological parameters responsible for these events. The literature indicates that climate change (due to anthropogenic global warming, volcanism, changes in the planet rotation axis, changes in solar activity and/or other geophysical or geological causes) alter the intensity and frequency of extreme events. The goal here is to analyze the extreme events mentioned and contribute with a set of assessment tools with National Institute of Science and Technology for Climate Change (INTC-MC).

LISTA DE FIGURAS

Pág.

LISTA DE TABELAS

Pág.

SUMÁRIO

Pág.

1 INTRODUÇÃO

Consolida-se cada vez mais o consenso na comunidade científica de que a civilização humana tornou-se uma força geofísica. A geração de energia da atual sociedade industrial é baseada majoritariamente em carbono (carvão e petróleo). É energia para as indústrias, residências e transporte (marítimo, automotivo e aéreo). Um dos subprodutos desta base tecnológica para geração de energia é a liberação de enormes quantidades de gases na atmosfera. Alguns destes gases, como o dióxido de carbono (CO_2) e o metano (CH_4) tem a propriedade de absorver energia radiante na faixa de comprimento de onda que é emitida pela superfície do planeta. O aumento da concentração destes gases na atmosfera, faz aumentar a quantidade de energia retida no planeta. Isto é comumente conhecido como *efeito estufa*, que seria a principal causa do aumento da temperatura média do planeta Terra.

Há evidências experimentais do aquecimento do sistema climático global, através do monitoramento das temperaturas médias globais do ar e dos oceanos. Observações indiretas do aquecimento global são o derretimento das geleiras e a elevação do nível global médio dos oceanos. Tais evidências são obtidas por meio de observações em todos os continentes e da maior parte dos oceanos, que mostram que muitos sistemas naturais estão sendo afetados pelas mudanças climáticas regionais, principalmente pelos aumentos de temperatura (IPCC, 2007).

Questiona-se sobre o aumento da temperatura pelos gases de efeito estufa relacionados a componente antropogênica. O atual estágio da ciência do clima indica a importância desta componente no aquecimento. Sofisticados modelos computacionais, que simulam o clima, indicam que um fator relevante para o aumento médio da temperatura do planeta está ligado ao crescimento da concentração destes gases de origem antropogênica.

Em um planeta mais aquecido, os fenômenos climáticos e meteorológicos extremos como secas, inundações, tempestades severas, ventanias e incêndios florestais se tornam mais frequentes. Segundo IPCC (2007) secas mais intensas e mais longas foram observadas sobre áreas mais amplas desde 1970, especialmente nos trópicos e subtrópicos. Neste período também se observou um aumento da atividade intensa dos ciclones tropicais no Atlântico Norte correlacionado com os aumentos das temperaturas da superfície do mar nos trópicos.

Podem-se identificar os desafios científicos tomando-se o século XX como referência. Antes do século XX, o desafio consistia em se descobrir as leis fundamentais da

natureza: determinar a natureza da dinâmica dos corpos e dos fluidos, como se dá o transporte da energia térmica e sua relação com o trabalho mecânico, as leis do eletromagnetismo e da ótica, a marcha evolucionária dos seres vivos e a forma de manipular quantidades infinitas na matemática. O desafio do século XX foi gerar soluções para as equações da física-matemática que modelam os fenômenos naturais. Uma metodologia desenvolvida no século XX foi baseada em métodos numéricos para os computadores digitais, gerando soluções aproximadas dos modelos matemáticos. Aerodinâmica computacional, projeto de moléculas pela química computacional e a moderna previsão numérica do tempo são somente tres exemplos do enorme impacto que o avanço científico da simulação computacional vem causando.

O avanço tecnológico de sensores, dados de satélites, simulação computacional provê um vertiginoso crescimento da quantidade e da qualidade dos dados. Muitos cientistas apontam que a análise de dados emerge como um dos principais desafios científicos do século XXI. O problema tem chamado tanto a atenção que muitos opinam que estamos no limiar da criação de uma "ciência dos dados". É claro que a ciência da computação está no centro desta nova questão. Sendo assim, surgiu a necessidade do desenvolvimento de novas técnicas e ferramentas que pudessem transformar, de maneira inteligente e automática, os dados processados em informações úteis e em conhecimento. As técnicas estatísticas e técnicas de redução de dimensão formam um contingente de ferramentas para análise de dados. Com o desenvolvimento da computação, outros métodos estão também disponíveis, como os métodos da inteligência artificial. De maneira mais geral, mineração de dados é umas das formas de agrupar este conjunto de técnicas para extrair informações ou conhecimento a partir dos dados. Análise de dados climatológicos, experimentais ou simulados, é hoje um capítulo importante do desafio da ciência dos dados no século XXI. Aliando este progresso aos problemas ambientais ocorridos nas últimas décadas, este trabalho fará uso desta quantidade de dados disponível e de ferramentas computacionais, com o objetivo de apontar variáveis climatológicas responsáveis por eventos climáticos extremos.

1.1 Objetivo e Contribuições

A mineração de dados surgiu da necessidade do emprego de técnicas e ferramentas que permitissem extrair informações de maneira automática de dados disponíveis. Há grande desenvolvimento desta tecnologia, especialmente em aplicações em bancos de dados reais.

Existem vários métodos de mineração de dados, que objetivam encontrar padrões,

como regras de associação ou classificação, para detectar relacionamentos sistemáticos entre variáveis, detectando assim novos subconjuntos de dados. Um dos métodos mais recentes de classificação desenvolvidos pela comunidade de computação é a construção de árvores de decisão, que são intuitivas e de fácil interpretação.

Análise de dados sempre foi um tema de interesse. Uma das áreas em que análise de dados se tornou fundamental desde seus primórdios foi a astronomia. Mais recentemente, outra área que tem necessidade de tratamento de grandes volumes de dados é a bioinformática. Em genética, com a criação de métodos experimentais sofisticados para descobrir o nível de expressão dos genes, biólogos e cientistas computacionais logo perceberam a necessidade de uma abordagem holística para analisar conjuntos com grande quantidade de dados. A evolução da tecnologia permitiu o monitoramento da expressão de milhares de genes simultaneamente, onde o foco se deslocou do estudo de experimentos simples, para a análise de todo o conjunto de observações.

Na biologia molecular experimental, por exemplo, os microarranjos (MA) de DNA são hoje em dia uma das tecnologias chave em estudos genômicos. Os MA permitem um monitoramento do nível de expressão de milhares de genes simultaneamente. Existem várias técnicas computacionais de classificação, agrupamento, predição de dados, dentre outras, utilizadas em bioinformática que permitem reduzir o banco de dados identificando os genes mais significantes.

Atualmente, a análise de dados ambientais depende fortemente do conhecimento *a priori* do meteorologista ou técnico que realiza o estudo. Este, com base na sua experiência e a partir de um vasto conjunto de opções, determina quais serão as variáveis ambientais analisadas em busca de eventuais correlações que expliquem os fenômenos ambientais e, se possível, aumentem sua previsibilidade. Devido ao tamanho e a complexidade dos bancos de dados ambientais atuais, a necessidade do uso de abordagens mais gerais e menos subjetivas na extração do conhecimento torna-se cada vez mais premente.

O objetivo desta tese é demostrar que técnicas que já são de uso corrente em biologia molecular na extração de conhecimento de grandes conjuntos de dados genômicos podem ser adaptadas para a análise de bases dados ambientais. Para alcançar este objetivo, foi estruturado um banco de dados ambientais com milhares de variáveis de reanálise e diferentes índices meteorológicos, e implementadas diferentes técnicas de classificação, agrupamento e visualização, algumas nunca utilizadas no contexto ambiental antes. Logo, os resultados apresentados são provenientes de dados reais, não simulados. Além disto a técnica de classificação foi utilizada como redutor de

dados para aplicação em árvore de decisão. Estas são as contribuições originais desta tese.

1.2 Aplicações

A mudança climática envolve um dinamismo mais complexo do que a simples elevação da média térmica. A comunidade científica acredita, por exemplo, que a frequência e a severidade dos eventos extremos crescerão devido ao aquecimento global. Segundo Barbieri et al. (2009) os desastres naturais que ocorrem no Brasil são, na sua maioria, de origem atmosférica, sendo a precipitação o elemento atmosférico que mais contribui para a ocorrência de desastres no Sul do Brasil. Chuvas intensas acarretam inundações que ainda podem ser agravadas por outros fenômenos como ventos fortes, deslizmentos de terra, granizo, entre outros. Por outro lado, os baixos índices de precipitação são responsáveis pelas secas que também ocasionam elevados prejuízos para diversos segmentos da economia, afetando diretamente a sociedade.

A nível sazonal, uma estação chuvosa fraca ou extremamente forte, acompanhada por relativas altas temperaturas devido a anomalias de temperatura da superfície do mar no Oceano Pacífico tropical (El Niño, La Niña), ou no Atlântico tropical ou subtropical, podem ter fortes impactos sobre a população e em setores econômicos dependendo da disponibilidade de água. Os efeitos da seca podem, por exemplo, prejudicar as atividades agrícolas e a geração hidrelétrica nas áreas afetadas. Com a redução persistente da precipitação nessas áreas, os lagos secam, as vazões dos rios diminuem e o abastecimento de água potável é reduzido, dificultando as opções de conservação e esgotando as reservas de água potável (MARENGO, 2009a).

As perdas em vidas humanas e os prejuízos materiais associados colocaram no topo da agenda do país a busca de meios de identificar as causas de certos eventos extremos e de avaliar até que ponto eles podem ser previstos e mitigados. O INPE sedia o Instituto Nacional de Ciência e Tecnologia para Mudanças Climáticas (INCT-MC), que é uma das maiores redes de pesquisa em mudanças climáticas do país. Um dos objetivos do INCT-MC é detectar e atribuir as causas das transformações ambientais que ocorrem no Brasil e na América do Sul. Esta tese se insere neste contexto particular e propõe-se a analisar dois eventos extremos recentes de grande impacto na opinião pública brasileira: as grandes secas de 2005 e 2010 na Amazônia e a ocorrência de precipitação extrema em Santa Catarina ocorrida em 2008.

1.2.1 Secas na Amazônia

A floresta Amazônica tem importante papel no clima do planeta, em particular devido ao grande estoque de carbono, tanto na biomassa como no solo, bem como ao transporte de energia e umidade devido à intensa convecção registrada na região. As variações climáticas na região Amazônica podem estar relacionadas às mudanças climáticas globais decorrentes de causas naturais como a variações da intensidade solar, da inclinação do eixo de rotação da Terra ou da excentricidade da órbita terrestre, a atividades vulcânicas, a variações da composição química da atmosfera, entre outras (NOBRE et al., 2007). Além das causas naturais, fatores antropogênicos como desflorestamento e queimadas, provocam mudanças climáticas nesta região.

O aquecimento global representa um risco para o ciclo hidrológico na Amazônia, uma vez que o aumento de temperatura provocará uma maior evaporação e maior transpiração das plantas, o que levará a uma aceleração do ciclo hidrológico (CASE, 2006). Se, além disso, a precipitação diminuir durante a estação seca, o impacto das mudanças climáticas no regime hidrológico na Amazônia será agravado (NIJSSEN et al., 2001).

O efeito do desmatamento e das mudanças climáticas afeta o ciclo hidrológico em todas as escalas de tempo. Em escalas de anos a décadas, teleconexões nos padrões de circulação global atmosférica, ocasionadas pela interação oceano-atmosfera, afetam a hidrologia de algumas regiões, especialmente nos trópicos (NIJSSEN et al., 2001). Outro fator importante é o fogo. A floresta densa Amazônica era praticamente impenetrável ao fogo, mas devido à combinação da fragmentação florestal, desmatamentos e aquecimento devido ao aquecimento global, aliada a prática agrícola predominante que utiliza fogo intensamente, esse quadro está rapidamente mudando, e a frequência de incêndios florestais vem crescendo a cada ano. A assombrosa velocidade com que tais alterações estão ocorrendo, em comparação àquelas dos processos naturais em ecossistemas, introduz séria ameaça à mega-diversidade de espécies da flora e da fauna dos ecossistemas, em especial da Amazônia, com o provável resultado de sensível empobrecimento biológico (NOBRE; ASSAD, 2005).

Em 2005 e 2010 a Amazônia sofreu duas das mais intensas secas já registradas na região, com graves consequências socioambientais. Ao se analisar, por exemplo, os dados de precipitação no setor sul da Amazônia, verifica-se que durante a estação chuvosa de 2005 (que na realidade estendeu-se de dezembro de 2004 a março de 2005), as chuvas apresentaram valores até 350 mm menores que a média histórica. Isto contribuiu para que os níveis dos rios desta região estivessem com valores bem

abaixo da média no final da estação chuvosa de verão e no início do período de estiagem, que ocorre de maio a setembro. Além disso, em 2005, observou-se uma precipitação média menor que a usual durante todos os meses deste ano. Um dos possíveis fatores responsáveis por esta seca intensa estaria relacionado à temperatura da superfície do mar no Atlântico Tropical Norte, que esteve acima da média nos 12 meses anteriores ao episódio da seca (MARENGO et al., 2008a).

Apenas cinco anos após o evento de 2005, uma nova intensa seca atingiu a Amazônia. A seca de 2010 afetou uma grande área que compreendia o noroeste, centro e sudoeste da Amazônia, incluindo partes da Colômbia, do Peru e do norte da Bolívia. Poucas nuvens e menos chuvas também se traduziram em temperaturas mais altas e baixas históricas no nível do principal afluente, o Rio Negro (MARENGO et al., 2011).

Lewis et al. (2011) comparam as secas de 2005 e 2010 usando uma medida de intensidade de seca que se correlaciona com a mortalidade de árvores. Em 2010 o aumento de mortalidade das árvores alcançou 3.2 milhões km^2 comparados com os 2.5 milhões km^2 medidos em 2005. Com isto, eles afirmam, é possível que secas repetidas possam ter importante impacto em escala de décadas no ciclo global de carbono.

1.2.2 Precipitação extrema em Santa Catarina

O Estado de Santa Catarina (SC), localizada no sul do Brasil é propenso à ocorrência de vários fenômenos climáticos severos, tais como ventos fortes, chuvas de granizo, enchentes, inundações e até mesmo tornados. Um dos fenômenos mais frequentes que causam desastres naturais em Santa Catarina são as inundações, com mais de mil ocorrências em um período de 21 anos. Neste cenário, o povo de Santa Catarina está familiarizado com inundações, principalmente nas áreas atingidas com frequência, havendo em alguns municípios sistemas de aviso e outras medidas preventivas, a fim de minimizar o impacto causado por esse tipo de evento (MARCELINO et al., 2005).

Chuva intensa é, obvimente, o fator causador destes desastres na região. As águas das chuvas podem sofrer evaporação imediata, infiltração e escoamento superficial. Devido ao relevo acidentado da Serra do Mar, a tendência maior é de escoamento imediato das águas para riachos e rios, o que causa as inundações. A localização geográfica de Santa Catarina e seu próprio desenho da costa, vizinha ao oceano, já o coloca em condições de receber grande quantidade de umidade através da brisa marítima (DIAS, 2009).

Em novembro de 2008 na região do baixo vale do rio Itajaí houve um desses even-

tos extremos. Este evento foi consequência de uma elevada quantidade de chuva localizada no tempo e espaço, que ocasionou enchentes e diversos deslizamentos nas encostas. O período de maior precipitação concentrou-se entre 20 e 24 de novembro de 2008. Segundo Dias (2009), não há registro de um novembro tão chuvoso nas regiões da Grande Florianópolis, Vale do Itajaí e Litoral Norte como observado em 2008, quando diversos recordes históricos foram quebrados. Em Blumenau e Joinville, os totais do mês ficaram em torno de 1000 mm (equivalente a 1.000 litros/m^2), para uma média climatológica mensal de aproximadamente 150 mm. Este evento afetou 1.5 milhões de pessoas e resultou em 120 vitimas e 69 mil pessoas desabrigadas. Os deslizamentos e as inundações causadas pelas tempestades bloquearam quase todas as estradas da região e interromperam a distribuição de água e eletricidade em milhares de casas (MARENGO, 2009b).

Ainda segundo Dias (2009), tais chuvas foram causadas pelo estabelecimento de um bloqueio atmosférico no oceano Atlântico, acompanhado por um vórtice ciclônico em altitude (entre 4000 m e 5000 m), localizado entre o leste de Santa Catarina e o leste do Paraná, que favoreceu a ascensão do ar úmido ao longo da Serra do Mar.

1.3 Organização da Tese

Esta tese organiza-se em seis capítulos. Após esta introdução, o capítulo 2 descreve os métodos de análise de dados utilizados nesta tese. O capítulo 3 faz uma descrição dos dados e a estratégia da análise. Os capítulos 4 e 5 apresentam os resultados das aplicações na Seca da Amazônia e na precipitação extrema em Santa Catarina respectivamente, com os comentários. Finalmente, o capítulo 6 apresenta a conclusão de tese.

2 METODOLOGIAS DE MINERAÇÃO DE DADOS

Técnicas avançadas de mineração de dados são ainda pouco utilizadas em estudos meteorológicos e ambientais. Poucos trabalhos na literatura reportam o uso de técnicas avançadas de extração de conhecimento para explorar de modo sistemático e abrangente os grandes bancos de dados ambientais existentes atualmente. Graças aos avanços da bioinformática nas últimas décadas, o panorama na área biológica é mais promissor.

Os sistemas biológicos são paradigmas da complexidade e do comportamento não-linear, fato que muitas vezes faz a análise de dados biológicos uma tarefa delicada. Felizmente, bioestatísticos têm à sua disposição uma grande gama de métodos e ferramentas para lidar com um vasto conjunto de problemas, que vão de falsos positivos a "outliers". Este capítulo aborda as metodologias empregadas nesta tese, algumas correntemente utilizadas na área de biologia molecular. Em particular, são descritos o embasamento conceitual de mineração de dados, as metodologias de classificação, agrupamento e árvore de decisão, e os conceitos computacionais e estatísticos utilizados nas tecnologias descritas.

A grande quantidade de dados gerados, coletados ou armazenados, obtidos por operações diárias ou pesquisas científicas, requer um processo automatizado para descobrir padrões, exceções, tendências ou correlações entre eles. A necessidade cresceu a tal ponto que criou-se uma nova disciplina: *mineração de dados*. Muitos acreditam que ainda teremos a necessidade de um desenvolvimento ainda maior e que a mineração de dados será somente um dos capítulos de uma área mais ampla: *ciência dos dados* (ou *data science*, como se registra na literatura internacional).

Mineração de dados é uma fase na descoberta de conhecimento em bancos de dados (Knowledge Discovery in Databases - KDD) que procura por uma série de padrões escondidos nos dados, frequentemente envolvendo uma aplicação iterativa e repetitiva de métodos de mineração de dados particulares. O objetivo de todo o processo de KDD é tornar os padrões compreensíveis às pessoas, visando facilitar uma melhor interpretação dos dados existentes (FAYYAD et al., 1996). O conceito de mineração de dados está se tornando cada vez mais popular como uma ferramenta de descoberta de informações e de estruturas de conhecimento.

Figura 2.1 - Visão geral das etapas que compõem o processo de KDD.

A descoberta de conhecimento em bancos de dados se dá em 6 etapas (Figura 2.1):

1. Definição de metas;
2. Seleção ou segmentação dos dados apropriados para a análise de acordo com algum critério;
3. Pré-processamento: eliminação de ruídos e erros, verificação da falta de dados; convenções para construção de uma base de dados consistente;
4. Transformação: redução dos dados;
5. Mineração de Dados;
6. Interpretação/Avaliação dos dados.

Existem várias definições para esta tecnologia, mas segundo Chen et al. (1996), mineração de dados é um extenso campo de pesquisa, que associa técnicas e conceitos de diversas áreas como sistemas de banco de dados, sistemas baseados em conhecimento, inteligência artificial, aprendizado de máquina, aquisição de conhecimento, estatística, banco de dados espaciais e visualização de dados. As várias tarefas desenvolvidas em mineração de dados têm como objetivo primário a predição e/ou a descrição. A predição usa atributos para predizer os valores futuros de uma ou mais variáveis (atributos) de interesse. A descrição contempla o que foi descoberto nos dados sob o ponto de vista da interpretação humana.

O desenvolvimento de tecnologias de banco de dados e descoberta de conhecimento, abriu grandes avenidas na pesquisa de ciência da informação. Em Yuan (1998) os autores examinam o estado da atual tecnologia de mineração de dados em descoberta

do conhecimento (DM/KD), identificando necessidades especiais para DM e KD geoespaciais, e discutem desafios de pesquisa e potencial impacto em GIScience (Geographic Information Science). A primeira característica relevante em DM/KD é que os repositórios de dados geoespaciais tendem a ser muito volumosos. O volume de dados foi o primeiro fator na transição de algumas agências federais via mecanismos físicos (Compact Disk) ou eletrônicos. A revolução do dado geográfico digital está criando novos tipos de formatos de dados além dos tradicionais "vetores" e "rasters" (mapa de bits). Outra característica do dado geoespacial está relacionada à fase cíclica de coleta de dados, como por exemplo, o padrão de dispersão de epidemia na saúde, ou de resíduo tóxico. Outra característica está na origem do dado em si, que muitas vezes é acessada por ponteiros. Principalmente dados que circulam pela internet que às vezes são dificilmente localizados, resgatados e analisados. Segundo os autores, isto se deve ao fato de que a internet carece de um catálogo abrangente ou índices. Sumarizando, o desenvolvimento de ferramentas de DM/KD pode ser sustentado por uma sólida base geográfica. A emergência de infraestrutura de dados geoespaciais globais e nacionais tem sido feita neste sentido.

Exemplos de técnicas de mineração de dados podem ser encontrados em Sattari et al. (2012) que afirmam que o futuro próximo, não será significantemente diferente do passado recente. Sendo assim, as regras de decisão extraídas de dados anteriores, permitirão uma predição precisa do que possa acontecer no futuro. Em seu trabalho, são usados dados mensais de precipitação, vento, umidade e temperatura, bem como valores SPI (Standardized Precipitation Index), medidos em uma estação meteorológica situada em uma região do Alaska, com o objetivo de determinar a possibilidade de períodos de seca e chuva com a ajuda de técnicas de árvore de decisão. Os dados foram obtidos de 18 estações meteorológicas na província do Ankara. Estas estações contêm valores mensais medidos entre 1926 e 2006. Os valores de SPI são classificatórios, definindo 8 categorias de severidade da seca (Tabela 2.1). SPI foi obtido da subtração da quantidade de chuva na estação i no período de tempo X_i pela quantidade média de chuva $\overline{X_i}$, dividido pelo desvio padrão (σ), conforme fórmula abaixo:

$$SPI = \frac{X_i - \overline{X_i}}{\sigma}$$

O algoritmo de árvore de decisão utilizado foi o C5.0, uma extensão do sistema conhecido ID3 (QUINLAN, 1986). Os resultados obtidos mostraram que a província de Ankara tem um clima normal ou quase normal árido, e que a quantidade de precipitação em todos os meses pode ser considerada fator determinante para tal

Tabela 2.1 - Indice SPI de categoria de severidade da seca e label classificatório.

valores SPI	Categoria seca	label
≥ 2	Umidade extrema	EW
1.50 1.99	Umidade severa	SW
1.00 1.49	Umidade moderada	MW
0.99 0	Ligeriamente úmido	MiW
$0 \ -0.99$	Ligeiramente seco	MiD
$-1.00 \ -1.49$	Seca moderada	MD
$-1.50 \ -1.99$	Seca severa	SD
≤ 2	Seca extrema	ED

aridez. O estudo examinou todos os meses do ano, além de uma avaliação anual dos 81 anos de dados. Esta avaliação apresentou o atributo precipitação como o critério mais relevante na árvore de decisão, ou seja, as regras se basearam apenas na quantidade de precipitação da região. Os outros atributos (vento, umidade e temperatura) não apareceram nas árvores geradas.

Outro exemplo de aplicação de tecnologia de mineração de dados foi publicado por Hui et al. (2004) que utilizou árvores de decisão em dados de sensoriamento remoto para extrair *wetland* (ecossistemas naturais que ficam parcial ou totalmente inundados durante o ano) de imagens de satélites, (Landsat 5/Thematic Mapper (TM)) em uma área de da planície de Yinchuan (China). Os autores utilizaram esta técnica devido ao aumento do uso desta tecnologia nos últimos anos, para classificar dados de sensoriamento remoto. Para isto, a área estudada foi classificada em três tipos: *river wetland* (rios), *lake wetland* (lagos), e *paddy fields wetland* (plantações de arroz). Os atributos utilizados na análise foram: a diferença normalizada do indice de vegetação (NDVI), forma, tamanho, tonalidade, sombra, textura, e a transformação Cap Tasseled (TC). Antes da aplicação em árvores de decisão, foi feita uma classificação supervisionada utilizando a máxima verossimilhança e os resultados foram comparados com os obtidos pela árvore de decisão. A árvore gerada está ilustrada na Figura 2.2. Os resultados desta análise indicaram que a árvore de decisão tem um excelente desempenho em comparação com o classificador de máxima verossimilhança.

O trabalho de Gagne et al. (2009) utiliza árvores de decisão para classificar tipos de tempestades baseado em dados de refletividade em radar meteorológico. Os dados usados originam-se de simulações e observações de radares. São 250 simulações geradas em ambiente propício a tempestade. As observações de radares foram feitas a 4 km de altura do solo. Os tipos de atributos testados foram: morfológicos,

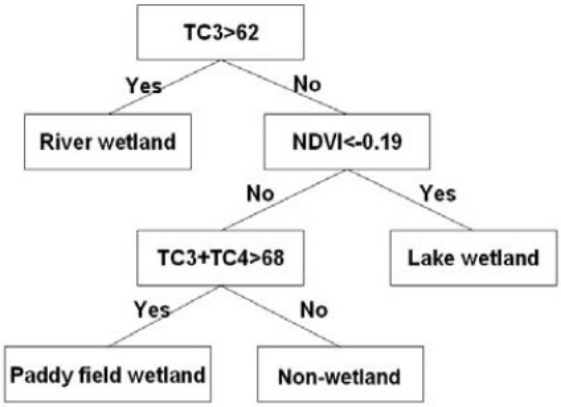


Figura 2.2 - Árvore de decisão gerada para o ano de 1999. TC3 representa tonalidade, TC4 representa sombra.
Fonte: Hui et al. (2004)

refletivos e ventos. As árvores de decisão foram geradas baseadas na permutação dos atributos (morfológicos e refletivos, morfológicos e vento, refletivos e vento). A árvore obtida pelo algoritmo J4.8 (WEKA) apresenta o atributo "area" como o de maior ganho de informação, indicando que o tamanho geral da tempestade é o fator mais importante na separação de tipos de tempestades. O experimento mostrou que árvores de decisão são métodos viáveis para determinar automaticamente tipos de tempestades, comparados a outros algoritmos padrões de aprendizado de máquinas. As árvores são também capazes de predizer o tipo de tempestade em um conjunto de teste independente com pequena perda de desempenho.

Em Ruivo (2008), foram utilizados metodologias de classificação e agrupamento na aplicação de bancos de dados ambientais para investigar quais foram os fatores climáticos associados à grande seca de 2005 na Amazônia, e identificar quais foram as variáveis fisico-químicas que controlam a emissão de gases de efeito estufa em reservatórios de hidrelétricas. Em ambas as aplicações, grandes volumes de dados originários de diferentes fontes foram organizados como se fossem experimentos de microarranjos. As metodolgias empregadas foram as mesmas utilizadas nesta tese e serão descritas a seguir.

2.1 Classificação

Classificação consiste em associar um dado como pertencente a uma determinada classe dentre várias previamente definidas ou a serem descobertas. Cada classe corresponde a um padrão único de valores dos atributos previsores. Esse padrão único

pode ser considerado como a descrição da classe. O modelo de classificação construído é utilizado para predizer classes de novos casos que serão incluídos em um banco de dados.

Um classificador extraído de um conjunto de dados objetiva predizer um valor, e entender a relação existente entre os atributos previsores e a classe. Para que se cumpra esta relação é exigido do classificador que ele não apenas classifique, mas também explicite o conhecimento extraído da base de dados de forma compreensível (BREIMAN et al., 1984).

Na comparação de classes entre grupos de amostras checa-se se existem diferenças estatísticas significativas entre atributos. Na biologia molecular, por exemplo, um importante objetivo no estudo de MA de DNA é a identificação de genes que são diferentemente expressos entre classes pré-definidas. Esta identificação com funções desconhecidas pode levar a um melhor entendimento das funções destes genes.

Métodos de comparação de classes são supervisionados, pois utilizam a informação de que amostra pertence a qual classe. O modelo estatístico utilizado permite estimar a probabilidade de se ver esta diferença tão grande quanto observada. O método mais comumente usado é o t-teste (AMARATUNGA; CABRERA, 2004) que mede diferenças entre duas classes na variabilidade de expressão do gene em unidades de variância. Para mais classes, utiliza-se o F-teste. Grandes valores absolutos da estatística t ou F sugerem que as diferenças observadas entre as classes não são devidas ao acaso e que a hipótese nula pode ser rejeitada. Matematicamente, a estatística t é calculada por:

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{s_p^2 \left(\frac{1}{J_1} + \frac{1}{J_2}\right)}} \tag{2.1}$$

onde
$s_p^2 = \frac{(J_1-1)s_1^2+(J_2-1)s_2^2}{J_1+J_2-2}$
e
$s_i^2 = \frac{1}{J_i-1}\sum_{j=1}^{J_i}(x_{ij} - \overline{x_i})^2$
para i=1,2
sendo
$\overline{x_1}$= média das amostras classe 1
$\overline{x_2}$= média das amostras classe 2
J_1= quantidade de amostras classe 1

J_2= quantidade de amostras classe 2.

Para o F-teste tem-se:

$$F = \frac{[J_1(\overline{x_1} - \overline{x})^2 + J_2(\overline{x_2} - \overline{x})^2 + \ldots + J_I(\overline{x_I} - \overline{x})^2]/(I-1)}{s_p^2} \tag{2.2}$$

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

O resultado do t-teste é normalmente convertido para probabilidade, conhecido como p-valor, que representa a probabilidade de se observar em hipótese nula, um t-estatístico tão grande quanto observado no dado real (AMARATUNGA; CABRERA, 2004; SIMON et al., 2003).

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

Apesar da sua popularidade, o uso do t-teste em comparação de classe pode ser problemático. Por exemplo, as estimativas de variância podem ser distorcidas por genes que têm uma variabilidade muito baixa. Estes genes estão associados a um valor alto de t-estatística e serão erroneamente selecionados como diferencialmente expressos. Outra desvantagem vem de sua aplicação em amostras muito pequenas ou não normalmente distribuídas. Isso levou ao desenvolvimento de muitas alternativas inovadoras (JEANMOUGIN et al., 2010).

Uma distribuição normal pode ser completamente caracterizada pela média e desvio padrão de seus parâmetros, com isto, t-testes são exemplos do que é conhecido como "testes paramétricos". Uma aproximação de distribuição normal pode não ser muito boa quando a variabilidade dos dados difere muito da normalidade ou quando se

está interessado em valores muito pequenos do p-valor. Uma abordagem alternativa para se estimar p-valores sem assumir uma distribuição subjacente é o método da permutação. Neste método, as amostras são permutadas de classes aleatoriamente e o t-teste é calculado em cada permutação. Supondo duas classes de amostras, classe 1 e classe 2 aonde tem-se J_1 elementos da classe 1 e J_2 elementos da classe 2. O resultado do t-teste usando o método da permutação é obtido calculando-se primeiramente o t-teste pela eq. 2.1. Em seguida permutam-se aleatoriamente os elementos entre as classes 1 e 2, onde elementos de classe 1 são aleatoriamente rotulados como classe 2, e elementos de classe 2 são aleatoriamente rotulados como classe 1. Usando estes rótulos temporários, o t-teste é calculado novamente e designado como t^*.

A figura 2.3 ilustra um exemplo de cálculo do p-valor em um banco de dados dividido em duas classes com 126 permutações aleatórias. A cada passo, foi calculado o novo t^* e observa-se que foram obtidos 3 valores onde $|t^*| \geq |t|$. Assim, obteve-se um p-valor de 0.0238 para o registro (gene) em questão.

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

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

Figura 2.3 - Exemplo de cálculo de p-valor.
Fonte: Amaratunga e Cabrera (2004)

Todo teste de hipótese possui erros associados a ele. Um dos mais importantes é o chamado "erro do tipo I" que corresponde à rejeição da hipótese nula quando esta for verdadeira. A probabilidade do erro do tipo I chama-se nível de significância e é expressa através da letra grega α. Os níveis de significância usualmente adotados são 5%, 1% e $0,1\%$.

Formalmente, o nível descritivo (p) é definido como o "menor nível de significância α que pode ser assumido para se rejeitar a hipótese nula (H_0)", porém esta interpreta-

ção não é simples até mesmo para os estatísticos. Considerando, de maneira muito generalizada, que os pesquisadores ao rejeitarem a hipótese nula costumam dizer que existe "significância estatística" ou que o resultado é "estatisticamente significante", poderíamos definir o nível descritivo (p) como a "probabilidade mínima de erro ao concluir que existe significância estatística".

2.2 Agrupamento

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

Agrupamento hierárquico produz uma fusão sequencial aninhada das variáveis em estudo, baseado em algumas métricas de correlação ou semelhança, como a correlação de Pearson ou a distância Euclidiana. A fusão aninhada é representada por um "dendrograma". No nível mais baixo do dendrograma, cada variável é um membro de um *cluster* (grupo) individual. No primeiro passo, as variáveis mais semelhantes são fundidas em um *cluster*. Em seguida, as próximas duas variáveis mais similares são unidas em outro *cluster*. Em cada passo, os dois grupos mais semelhantes (incluindo conjuntos unitários) são unidos para formar um grande grupo. No nível mais alto do dendrograma, existe um conjunto contendo todas as variáveis.

A correlação de Pearson entre dois experimentos X é Y é definida por:

$$\frac{\sum_{i=1}^{N}(x_i - X_{avg})(y_i - Y_{avg})}{[(\sum_{i=1}^{N}(x_i - X_{avg})^2)(\sum_{i=1}^{N}(y_i - Y_{avg})^2)]^{1/2}} \tag{2.3}$$

onde N é o número de variáveis, X_{avg} é a média das variáveis no experimento X, e Y_{avg} é a média das variáveis no experimento Y. A correlação de Pearson é uma medida de similaridade comumente usada entre dois vetores, portanto, um menos a correlação é usada como distância métrica.

A distância Euclidiana entre dois vetores é dada por:

$$\sqrt{\sum_{i=1}^{N}(x_i - y_i)^2}, \tag{2.4}$$

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

Além destas métricas, podemos citar ainda a métrica de Minkowski, que tem a distância Euclidiana como um caso particular. Outras versões populares da métrica de Minkowski são as distâncias de Manhattan e de Chebyshev. Outra métrica de distância cujo cálculo é extremamente simples é a distância Camberra que calcula a soma das diferenças fracionárias entre as coordenadas de pares de objetos (LINDEN, 2009).

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

O $K - means$ é também uma heurística de agrupamento não hierárquico que busca minimizar a distância dos elementos a um conjunto de forma iterativa. Podemos citar ainda os Grafos que são modelos matemáticos que representam relações entre objetos. Um grafo é um conjunto dado por G=(V, E), onde V é um conjunto finito de pontos, normalmente denominados de nós ou vértices e E é uma relação entre vértices, ou seja, um conjunto de pares em VxV (LINDEN, 2009).

Neste trabalho, a técnica de agrupamento foi usada para estudos feitos em uma das análises. Para este cálculo, é possível também fazer uma escolha da métrica de distância entre os grupos: *average linkage*, *complete linkage* ou *single linkage* (Figura 2.4). Esta métrica define a distância entre os perfis de expressão de duas amostras. Durante o agrupamento hierárquico, no entanto, o algoritmo deve calcular a distância entre os agrupamentos formados no passo anterior. Com agrupamento de *average linkage*, a distância entre os dois grupos é tomada como a média das distâncias entre todos os pares de elementos, um do primeiro grupo e um do segundo. Com agrupamento de *complete linkage*, a distância entre os dois grupos é tomada como a distância máxima entre um elemento do primeiro conjunto e um elemento no segundo. Com agrupamento de *single linkage*, a distância entre os dois grupos

é tomado como a distância mínima entre um elemento do primeiro conjunto e um elemento no segundo.

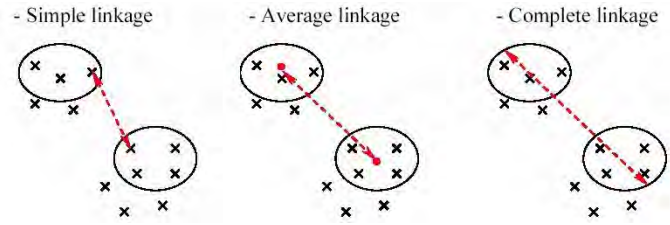


Figura 2.4 - Ilustração de métricas de distância entre os grupos.

Embora vários métodos de agrupamento organizem tabelas de forma útil, o resultado desta grande massa de dados fica difícil de ser assimilada. É usual combinar métodos de agrupamento com técnicas de representação gráfica do dado primitivo, apresentando cada ponto com uma cor que, quantitativamente e qualitativamente, reflete as observações do experimento original. O produto final é a representação gráfica do dado complexo através de uma ordenação estatística, permitindo à especialistas uma assimilação e exploração do dado de uma maneira natural intuitiva (EISEN et al., 1998).

O resultado do agrupamento produz uma imagem colorida em forma de matriz. O traçado das cores varia de vermelho a verde, variando da maior a menor expressão de cada variável em cada amostra, onde as linhas horizontais representam as variáveis e as colunas representam as amostras. Tons de vermelho são usados para representar graus de amplitude crescente (Figura 2.5). Da mesma forma, os tons de verde são usados para representar graus de diminuição da amplitude (SIMON; LAM, 2006).

2.3 Árvores de decisão

Métodos de árvore de decisão representam um tipo de algoritmo de aprendizado de máquina que utiliza uma abordagem dividir-para-conquistar para classificar casos usando uma representação baseada em árvores. Esta filosofia baseia-se na sucessiva divisão do problema em vários subproblemas de menores dimensões, até que uma solução para cada um dos problemas mais simples possa ser encontrada.

Uma árvore de decisão é um modelo representado graficamente por nós e ramos (Figura 2.6), parecidos com uma árvore, mas invertida. O nó raiz é o primeiro nó da árvore e fica no topo da estrutura. Cada nó contém um teste sobre um ou mais

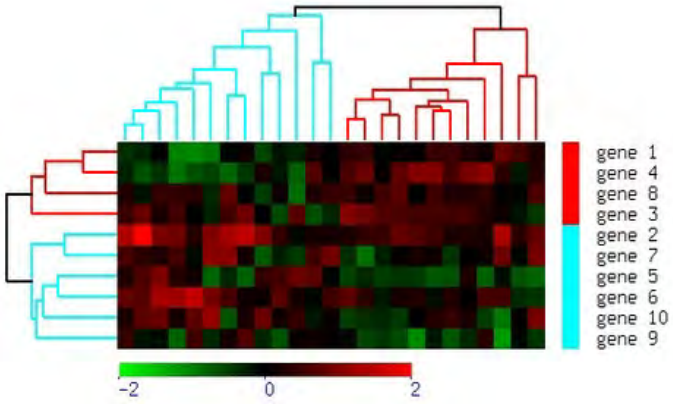


Figura 2.5 - Ilustração de agrupamento hierárquico e representação por cores.

atributos (parâmetros) e os resultados deste teste formam os ramos das árvores (WITTEN; FRANK, 2005). Cada nó folha, nas extremidades da árvore, representa um valor de predição para o atributo meta (MEIRA, 2008).

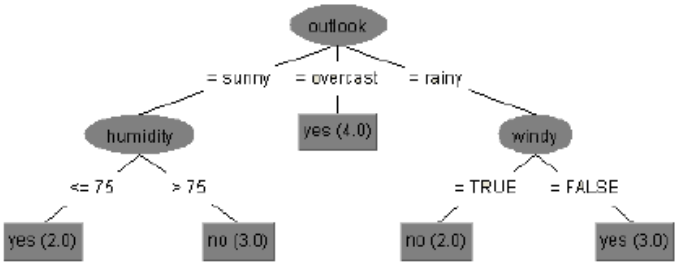


Figura 2.6 - Exemplo de árvore de decisão simples baseada no problema modelo weather, gerada com a ferramenta Weka.
Fonte: Quinlan (1993)

O conhecimento representado em árvores de decisão pode ser extraído na forma de regras de classificação "se-então". Uma regra é criada para cada caminho entre a raiz e um nó folha. Os testes de valor de atributo ao longo do caminho formam uma conjunção no antecedente da regra, e o nó transforma-se no consequente da regra. Essas regras podem ser mais fáceis de compreender, especialmente se a árvore de decisão for muito grande (HAN; KAMNER, 2001).

Fundamentados neste princípio, o algoritmo subdivide recursivamente o conjunto de treinamento em vários subconjuntos até que cada um deles contemple apenas uma classe, ou até que um critério de parada seja alcançado. Ao decidir qual atributo

será utilizado em cada subdivisão, um teste estatístico é adotado como critério de quebra.

Depois de construída, a árvore pode ser usada para classificar exemplos cuja classe é desconhecida. Para classificar um exemplo, testam-se os valores de seus atributos segundo a árvore de decisão. Um caminho é traçado a partir do nó raiz, descendo pelos ramos de acordo com os resultados dos testes, até chegar a um nó folha, que representa a classe de predição (HAN; KAMNER, 2001).

O critério para escolha do atributo que divide o conjunto de exemplos em cada repetição é um dos aspectos principais do processo do método. Entre os critérios mais conhecidos e usados tem-se o ganho de informação e a razão de ganho, definidos com base na teoria da informação (QUINLAN, 1993). O ganho de informação é uma medida usada para selecionar o atributo de teste em cada nó de decisão de uma árvore. O atributo com maior ganho de informação é escolhido como atributo de teste de cada nó, em cada iteração do processo. Este atributo minimiza a informação necessária para classificar os exemplos das partições resultantes da divisão. Tal abordagem ligada à teoria da informação minimiza o número de testes esperados para classificar um exemplo e garante que uma árvore simples seja encontrada (HAN; KAMNER, 2001).

2.3.1 Algoritmos de geração de árvores

Existem várias implementações utilizando algoritmos de indução (construção) baseados em árvores de decisão conhecidos na literatura. O algoritmo ID3, desenvolvido por Quinlan (QUINLAN, 1993; QUINLAN, 1986) é um dos mais populares. A construção da árvore é realizada de cima para baixo ("top-down"), com o objetivo de escolher o melhor atributo para um nó da árvore. É um processo recursivo que se inicia na raiz, e após ter escolhido um atributo para um nó, aplica o mesmo algoritmo aos descendentes desse nó, até que certos critérios de parada sejam verificados.

Um dos problemas existentes na indução de árvores de decisão é o chamado problema de *Overfitting*, isto é sobre-ajustamento da árvore aos dados de treinamento, obtendo um desempenho quase perfeito nesses, mas um desempenho pobre nos novos dados. Isto ocorre devido a presença de ruído nos dados, ou à existência de um conjunto de atributos inadequado ou insuficiente.

O algoritmo C4.5 (QUINLAN, 1993) é um método melhorado do ID3 que combate o problema do overfitting, utilizando uma estratégia de "poda da árvore". Existem

duas estratégias de combate ao problema do sobre-ajustamento, que pressupõe que a árvore tem uma complexidade inadequada e irreal. O algoritmo J4.8 é uma implementação do C4.5, que constrói um modelo de árvore de decisão baseado num conjunto de dados de treinamento, e usa esse modelo para classificar outras instâncias num conjunto de testes. Este é utilizado quando se quer conhecer qual classe de valores o algoritmo de aprendizagem prediz para cada registro. O tempo computacional depende da complexidade da árvore gerada (QUINLAN, 1993).

O J4.8 utiliza um conjunto de treinamento T para a construção da árvore. O conjunto T é composto por uma coleção de casos de teste cujas classes são bem conhecidas. Cada caso representa um objeto O_i que é definido por um conjunto de atributos A_{ij}. A estrutura da árvore é construída nó por nó. Primeiro escolhe-se um determinado atributo para ser testado no nó. O teste executado no nó X irá gerar as saídas $U_1; U_2; \ldots; U_p$, com um ramo representando cada saída. Por sua vez, essas saídas particionam T em p subconjuntos $T_1, T_2, \ldots, T_p$, onde cada subconjunto contém todos os casos que tem saída U_i para o atributo testado. Esse procedimento é aplicado recursivamente para cada saída U_i, ou seja, um novo nó é colocado na extremidade de cada ramo gerando novas saídas que dividirão cada subconjunto T_i em novos subconjuntos. O processo recursivo só termina quando todos os subconjuntos gerados são compostos apenas por casos de uma mesma classe.

Para a escolha dos atributos a serem testados, o J4.8 utiliza uma grandeza chamada "taxa de ganho" para selecionar o atributo que tenha o maior poder de discriminação entre as classes para cada nó. A taxa de ganho mede a quantidade de informação gerada pelo teste de um atributo específico que seja relevante para classificação de um objeto. Assim o algoritmo seleciona os atributos que irão gerar uma árvore simples e eficiente.

Para entender melhor o conceito de informação, considere-se o seguinte exemplo: o resultado da corrida de dois cavalos "iguais" é menos incerto do que o resultado de uma corrida com 8 cavalos "iguais". Antes da corrida, temos a incerteza em relação ao resultado, depois de sabermos o resultado temos a informação. Então, antes de um acontecimento ocorrer, temos a Incerteza, e depois de tomarmos conhecimento da ocorrência, o que recebemos chama-se Informação. A entropia pode ser entendida como uma medida matemática da informação necessária para descrever uma variável aleatória, ou como uma medida de incerteza acerca desta (COUTINHO, 2004).

O ganho de informação mede a redução da entropia (SHANNON, 1948) causada pela partição dos exemplos de acordo com os valores do atributo, ou seja, representa a

diferença entre a quantidade de informação necessária para uma predição correta e as correspondentes quantidades acumuladas dos segmentos resultantes após a introdução de um novo teste para o valor de determinado atributo (BERTHOLD; HANS, 1999).

A construção de uma árvore de decisão é guiada pelo objetivo de diminuir a entropia. A introdução da entropia no processo de construção de árvores de decisão visa a criação de árvores menores e mais eficazes na classificação. A forma de obtenção da entropia de Shannon é explicada a seguir (SHANNON, 1948).

Seja C um conjunto de objetos que pertencem a duas clases: P e N. Seja p o número total de objetos pertencentes a classe P, e n o número total de objetos pertencentes a classe N. Para cada nó da árvore serão determinadas as probabilidades de um exemplo pertencente à classe P como p/(p+n), e à classe N como n/(p+n).

Assim, a entropia é definida pela quantidade de informação necessária para decidir se um exemplo pertence a P ou a N, segundo a expressão:

$$\text{entropia}(p,n) = -\frac{p}{p+n}\log_2\left(\frac{p}{p+n}\right) - \frac{n}{p+n}\log_2\left(\frac{n}{p+n}\right) \text{ para } p,n \neq 0 \quad (2.5)$$

$$\text{entropia}(p,n) = 0 \quad \text{caso contrário.} \quad (2.6)$$

A entropia dos segmentos descendentes de um nó pai da árvore é acumulada de acordo com o número de exemplos cobertos pela ramificação. Para avaliar o quanto de desordem é reduzido através de um novo teste, basta calcular a entropia em cada novo segmento (nó filho) criado por cada ramo, onde cada ramo é associado com um valor do atributo sendo testado.

Se o atributo A com um domínio $\{v_1, \ldots, v_N\}$ é usado como raiz da árvore de decisão, a árvore terá então N partições de T, $\{T_1, \ldots, T_N\}$, onde T_i conterá aqueles exemplos em T que possuam o valor v_i de A. Dado que T_i contém p_i exemplos de P e n_i exemplos de N, a expectativa de informação requerida para a subárvore T_i é dada pela $entropia(p_i, n_i)$. A medida de ganho de informação, *ganho(A)*, obtida pela partição associada com o atributo em A, é dada pela expressão:

$$\text{ganho}(A) = \text{entropia}(p, n) - \text{entropia\_ponderada}(A) \tag{2.7}$$

onde a entropia_ponderada(A) é dada por:

$$\text{entropia\_ponderada}(A) = \sum_{i=1}^{N} \frac{p_i + n_i}{p + n} \text{entropia}(p_i, n_i) \tag{2.8}$$

sendo N é o número de partições criadas pelo teste.

O algoritmo examina todos os atributos candidatos e escolhe o atributo A que maximiza o ganho de informação para rotular o nó atual da árvore, e repete o processo de forma recursiva para dar continuação à construção da árvore de decisão nos subconjuntos residuais $T_1, \ldots, T_N$.

Para o cálculo do limiar em atributos discretos, o algoritmo utiliza uma divisão binária. Suponha que o artributo A seja utilizado como um nó teste. Os valores de A são colocados em ordem crescente $\{a_1, a_2, ..., a_m\}$. O conjunto é dividido em duas partes. São elas: A_1, cujos elementos possuem valores $\{a_1, a_2, ..., a_i\}$ e A_2, com valores $\{a_{i+1}, a_{i+2}, ..., a_m\}$ para o atributo A. Assim temos no máximo $m-1$ possíveis divisões. Considerando cada a_i $i \in \{1, 2, ...m-1\}$ como um limiar, é calculado o ganho para a respectiva divisão de acordo com as fórmulas 2.5 e 2.7, sempre para o atributo A. Uma vez determinado que o maior ganho foi obtido para o limiar candidato a_j $j \in \{1, 2, ...m-1\}$, muitos algoritmos escolhem o ponto médio entre a_j e a_{j+1} como limiar, ou seja:

$$\text{limiar} = \frac{a_j + a_{j+1}}{2}, \qquad j \in \{1, 2, ...m-1\} \tag{2.9}$$

O algoritmo J4.8 no entanto escolhe o maior valor de A em todo o conjunto de treinamento que não exceda o intervalo médio abaixo:

$$\text{limiar} = \max \left\{ j | j <= \frac{a_j + a_{j+1}}{2} \right\} \tag{2.10}$$

Isto garante que qualquer valor de limiar usado na árvore de decisão esteja dentro do banco de dados (WITTEN; FRANK, 2005; BERZAL et al., 2004).

Ainda dentro da teoria da informação, foram formuladas propostas de generalização

da entropia, chegando a reduzirem-se à mesma expressão em alguns casos particulares, tais como a formulação de Rényi (RÉNYI, 1961) e Tsallis (TSALLIS, 1988). Em 1961, Alfréd Rényi evidenciou, através de um conjunto de postulados, que outras quantidades poderiam ajustar-se igualmente ou talvez melhor a uma medida de informação, surgindo a idéia de entropia generalizada. Rényi propôs uma entropia parametrizada ($\alpha > 0$ e $\alpha \neq 0$) que contém a entropia de Shannon como caso limite, chamada entropia de Rényi (LIMA et al., 2010).

$$I_\alpha = \frac{1}{1-\alpha} \log(\sum_{i=1}^{n} p_i^\alpha) \tag{2.11}$$

Esta tem propriedades similares a entropia de Shannon, como aditividade, e ter seu máximo em $ln(n)$ para $p_i = 1/n$, mas contém o parâmetro adicional α que a torna mais sensível na forma da probabilidade de distribuição.

Alguns anos depois, Constantino Tsallis (TSALLIS, 1988) definiu uma forma generalizada de entropia, tendo como caso particular a entropia de Boltzmann-Gibbs-Shannon.

$$S_\alpha = \frac{1}{1-\alpha} \log(1 - \sum_{i=1}^{n} p_i^\alpha) \tag{2.12}$$

Em Lima et al. (2010) é feito um estudo comparativo do uso das entropias de Shannon, Rényi e Tsallis, com o objetivo de encontrar alternativas mais eficientes aplicadas a um Sistema Tolerante a Intrusões. Árvores de Decisão foram utilizadas em modelos de classificação de problemas relacionados à detecção de intrusos em redes de computadores, apresentando bons resultados. Os resultados experimentais mostraram que as entropias de Tsallis e Rényi aplicadas a este problema, constroem árvores mais compactas e eficientes, se comparadas com a entropia de Shannon.

2.4 Softwares empregados

2.4.1 Tecnologia de Microarranjos

Estudos genômicos normalmente envolvem a análise de grandes conjuntos de dados de informações derivadas de vários experimentos biológicos. Graças a avanços como a tecnologia de microarranjos (MA) de DNA, tornou-se possível o monitoramento dos níveis de expressão de milhares de genes simultaneamente sob uma condição

particular.

Um MA é tipicamente uma lâmina de vidro sobre a qual as moléculas de DNA são fixadas de uma forma ordenada em locais específicos, chamados de "spots". Um MA pode conter milhares de spots e cada spot pode conter alguns milhões de cópias de moléculas de material genético que correspondem a um dado gene. Os spots são impressos na lâmina de vidro por um robô ou sintetizados localmente pelo processo de fotolitografia.

Simplificadamente, a tecnologia de MA é um processo baseado em hibridização que possibilita observar a concentração de mRNA de uma amostra de células analisando a luminosidade de sinais fluorescentes. Hibridização é o processo bioquímico onde duas fitas de ácido nucléico com sequências complementares se combinam (DANTAS, 2004).

A Figura 2.7 ilustra a hibridização de um MA com duas amostras de mRNA, cada uma marcada com um corante fluorescente que emite luz em comprimentos de onda diferentes; em geral coloração verde e coloração vermelha. A partir das regras de pareamento de bases de Watson-Crick, o mRNA marcado (em solução) hibridiza com o cDNA correspondente depositado no MA. Neste processo de hibridização, ocorre um pareamento das moléculas complementares, a partir do qual em cada um dos spots da lâmina, que referencia um certo gene, tem-se as proporções de mRNA nas duas amostras testadas. Dessa forma a intensidade de fluorescência em cada spot "aceso" está relacionada à abundância do respectivo mRNA na solução (KRUTOVSKII; NEALE, 2001).

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

MAs podem ser utilizados para medir a expressão do gene de várias maneiras, mas uma das aplicações mais conhecidas é comparar a expressão de um conjunto de genes a partir de uma célula mantida numa condição particular para o mesmo conjunto de genes a partir de uma célula de referência dita normal.

Um método popular de análise de dados de expressão gênica é a identificação de genes que são diversamente expressos dentre classes definidas a priori (em inglês,

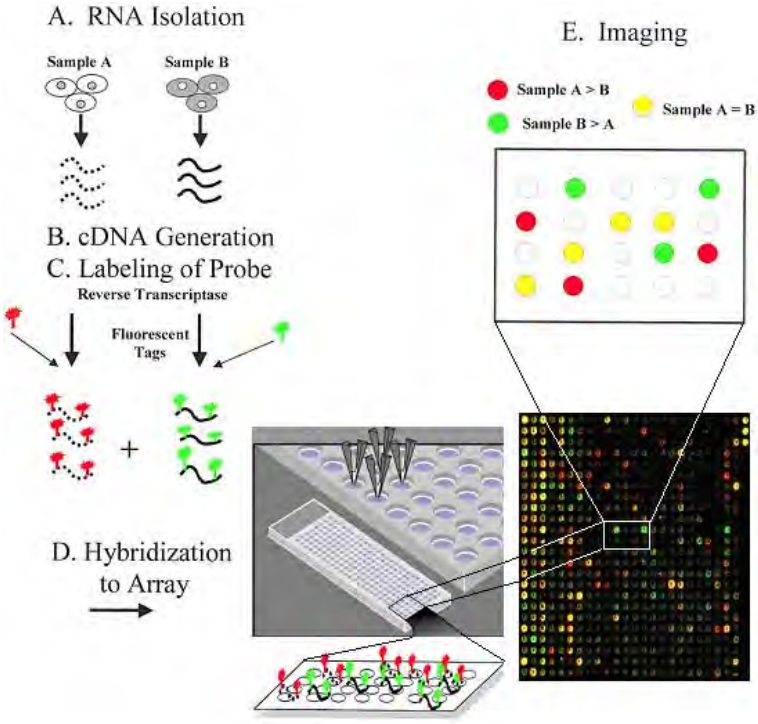


Figura 2.7 - Construção de Microarray.
Fonte: Krutovskii e Neale (2001)

"class comparison"). Assim, por exemplo, em um típico estudo genômico com MAs, níveis de expressão gênica de amostras de tecidos normais e afetados por uma dada patologia (um tumor, por exemplo) são comparados e classificados.

Outra abordagem também muito utilizada, sobretudo na análise de conjuntos de dados de grande porte, envolve o agrupamento ("clustering") de genes ou amostras com padrões similares de expressão, medidos segundo alguma métrica de correlação ou semelhança. Estas duas famílias de técnicas serão detalhadas a seguir.

2.4.2 BRB-ArrayTools

As ferramentas de classificação e agrupamento de dados utilizada neste trabalho provêm do pacote **BRB-ArrayTools** versão 3.7.0, desenvolvido por *Biometric Research Branch of the Division of Cancer Treatment and Diagnosis of the National Cancer Institute.*

Trata-se de um software livre, voltado para análise de dados de MA de DNA, e está disponível no site *http://linus.nci.nih.gov/ brb/download.html*, onde diversas informações como documentação completa e publicações de artigos que incluem re-

sultados feitos pelo software são fornecidas. É compatível com versões do Windows $98/2000/NT/XP$ ou superiores e projetado para ser usado como um *add-in* do Excelm; 2000 ou superior.

BRB-ArrayTools contém utilitários para processar dados de expressão em vários experimentos, visualizá-los, agrupá-los, classificá-los, dentre outras funções. O software foi desenvolvido por estatísticos experientes em análise de dados de MA, mas possui uma interface gráfica que facilita a utilização por biólogos. Mais detalhes podem ser encontrados em Simon e Lam (2006).

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

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

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

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

a) *Class Comparison*

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

genética, ou resposta à terapias. Uma característica de Class Comparison é que as classes são pré definidas independente do perfil de expressão.

b) *Clustering*

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

2.4.3 WEKA

A ferramenta utilizada neste trabalho para geração de árvores de decisão é o algoritmo J4.8 disponível no pacote WEKA, desenvolvido na Universidade de Waikato na Nova Zelândia (WITTEN; FRANK, 2000). O pacote WEKA é formado por um conjunto de implementações de diversos algoritmos de aprendizagem. Trata-se de um sistema que possui licença pública GNU, é fácil de instalar e sua implementação é feita em JAVA tornando-o facilmente portável. O WEKA possui várias funcionalidades, incluindo classificação, agrupamento e busca por regras de associações aplicadas ao conjunto de dados.

Para utilização do algoritmo J4.8 é necessário ajustar alguns parâmetros para melhora dos resultados. São eles:

- U : usa a árvore sem poda,
- -C : escolhe o fator de confiança inicial para a poda (Default: 0.25),
- -M : escolhe o número mínimo de instâncias por folha. (Default: 2) ,
- -R : usa a poda com redução de erro,
- -N : escolhe o número de partições para a poda com redução de erro. Uma partição é utilizada como conjunto de poda (Default: 3),
- -B : usa árvore binária,
- -S : não utiliza criação de subárvores,
- -L : não apaga a árvore depois de construída.

O suporte (-M) corresponde ao número mínimo de elementos permitidos em um determinado nó, seja ele um nó-intermediário ou terminal. Caso um determinado nó possua um número inferior de elementos que o suporte, esse nó sofre poda e é eliminado. Nessa implementação o suporte é implementado como número absoluto de registros em um determinado nó.

A confiança (-C) corresponde a assertividade da árvore quando usada em sua totalidade, ou seja, avalia a precisão das regras geradas. Uma regra é uma expressão na forma X => Y, onde X é chamado de antecedente e Y denominado consequente da regra. Tanto X como Y podem ser formados por conjuntos de dados. O fator de confiança é calculado pela razão X/Y, onde X é o número de registro que satisfaz o antecedente e o consequente da regra (número de casos que a regra cobre) e Y é o número total de registros que satisfazem o antecedente da regra(total de casos aplicáveis).

Na tela de saída do resultado é apresentada a matriz de confusão (*Confusion Matrix*) que indica que instâncias foram classificadas de forma correta e incorreta por classe. Os resultados são apresentados sob a forma de tabela de duas entradas: uma das entradas é constituída pelas classes desejadas, a outra pelas classes previstas pelo modelo. As células são preenchidas com o número de instâncias que correspondem ao cruzamento das entradas. Se houve 100% de classificação correta podemos esperar uma matriz de confusão onde todo elemento fora das diagonais é igual a zero.

Ainda na tela de saída são apresentados os seguintes índices:

- Kappa Statistic;
  O Kappa Statistic (K)é um indice que compara o valor encontrado nas observações com aquele que se pode esperar do acaso. É o valor calculado dos resultados encontrados na observações e relatado como um decimal (0 a 1). Quanto menor o valor de kappa menor a confiança da observação, o valor 1 implica a correlação perfeita, difícil de ser encontrada. O índice é calculado pela fórmula abaixo, onde P(A) é a probabilidade de se classificar as observações de modo correto, e P(E) é a probailidade hipotética de se ocorrer a classificação.
  $$K = \frac{P(A) - P(E)}{1 - P(E)} \tag{2.13}$$

- Root Mean-Squared Error;
  Este valor é calculado pela média da raiz quadrada da diferença entre o

valor predito e o valor correto. O índice é calculado pela fórmula abaixo, onde p_i é o valor predito, e a_i é o valor correto.

$$\sqrt{\frac{(p_1 - a_1)^2 + \ldots + (p_n - a_n)^2}{n}} \tag{2.14}$$

- Mean Absolute Error;
  É a média da diferença entre os valores atuais e os preditos em todos os casos. O índice é calculado pela fórmula abaixo, onde p_i é o valor predito, e a_i é o valor correto.

$$\frac{|p_1 - a_1| + \ldots + |p_n - a_n|}{n} \tag{2.15}$$

- Root Relative Squared Error;
  Reduz o quadrado do erro relativo na mesma dimensão da quantidade sendo predita incluindo raiz quadrada. O índice é calculado pela fórmula abaixo, onde p_i é o valor predito, e a_i é o valor correto.

$$\sqrt{\frac{(p_1 - a_1)^2 + \ldots + (p_n - a_n)^2}{(a_1 - \overline{a})^2 + \ldots + (a_n - \overline{a})^2}} \tag{2.16}$$

- Relative Absolute Error;
  É o erro total absoluto. O índice é calculado pela fórmula abaixo, onde p_i é o valor predito, e a_i é o valor correto.

$$\frac{|p_1 - a_1| + \ldots + |p_n - a_n|}{|a_1 - \overline{a}| + \ldots + |a_n - \overline{a}|} \tag{2.17}$$

- True Positives (TP);
  São os valores classificados verdadeiramente positivos.

- False Positives (FP);
  São os dados classificados erroneamente como positivos pelo classificador.

- Precisão (precision);
  É o valor da predição positiva, ou seja, número de casos positivos por total de casos cobertos.

- Recall (Cobertura);
  É o valor da cobertura de casos muito influenciada pela sensitividade e pouco pela especificidade. É calculada por número de casos cobertos pelo número total de casos aplicáveis.

- F-measure;
  É calculada por 2TP/2TP+FP+FN, onde TP (true positives) são os verdadeiros positivos, FP (False positives) são os falsos positivos, FN (false negative) são os falsos negativos ou (2*recall*precision/recall+precision).

- ROC Curve (Curva ROC);
  A curva ROC plota os números de casos positivos no eixo vertical(x), e os expressa como uma porcentagem do total do número de casos positivos, contra o número de casos negativos incluídos no exemplo, expressos como uma porcentagem do total de casos negativos no eixo horizontal.

3 DESCRIÇÃO DOS DADOS E ESTRATÉGIA DE ANÁLISE

Neste capítulo, foram aplicadas as metodologias de mineração de dados descritas no capítulo anterior com a finalidade de identificar as variáveis que influenciaram dois tipos de eventos climáticos extremos ocorridos no Brasil na última década. Os eventos analisados são os episódios de seca na Amazônia ocorridos nos anos de 2005 e 2010, e o episódio de grande volume de precipitação com ocorrência de enchentes em Santa Catarina no ano de 2008 (MARENGO et al., 2008a; TOMASELLA et al., 2011; COELHO et al., 2012). Os bancos de dados com as variáveis climatológicas analizadas estão descritos na seção a seguir. Trabalhos utilizando as metodologias empregadas nesta tese foram apresentados em congressos. Em Ruivo et al. (2012b) foram apresentados os resultados na análise das secas ocorridas na Amazonia em 2005 e 2010, e em Ruivo et al. (2012a), os resultados da análise do evento de precipitação extrema ocorrido em Santa Catarina em 2008.

Resumidamente, nestas aplicações foram calculados os p-valores, pelo método de permutação (SIMON et al., 2003), de todas as variáveis ambientais por meio de comparação de classes, definidas de acordo com o problema analisado. Estes campos de probabilidade são apresentados em seu contexto geográfico, na forma de mapas de p-valores para cada conjunto de variáveis climatológicas semelhantes (temperatura da superfície do mar, por exemplo). As variáveis que apresentaram menor p-valor em cada evento analisado, foram utilizadas como entrada de um algoritmo gerador de árvores de decisão, obtendo assim um classificador de eventos extremos. Este classificador em forma de árvore de decisão tem o objetivo de determinar qual a variável com maior ganho de informação, ou seja, que variável meteorológica tem maior relevância na análise de eventos extremos em estudo.

Além dos pacotes BRB-Array Tools e WEKA, existem outras ferramentas disponíveis capazes de realizar esta análise. Dentre elas podemos citar as Funções Ortogonais Empíricas (Empirical Orthogonal Functions - EOF) introduzidas por Lorenz e Project (1956) em estudos meteorológicos, com a finalidade de encontrar uma maneira eficaz de extrair uma representação simplificada ou compacta de um conjunto de dados. Outra ferramenta utilizada no trabalho de Anochi (2010) usa a teoria dos conjuntos aproximativos para o estudo de padrões climáticos sazonais.

3.1 Descrição dos dados

Para a análise dos fenômenos climáticos mencionados nesse trabalho, foram utilizadas séries temporais extraídas da base de dados do NCEP/NCAR (National Cen-

ters for Environmental Prediction / The National Center for Atmospheric Research) (KALNAY et al., 1996). Trata-se de um conjunto de dados globais de reanálise (isto é, assimilados e integrados) em grade, com resolução espacial de $2,5°$ x $2,5°$, com exceção da temperatura da superfície do mar, cuja resolução é de $2°$ x $2°$. Os dados para a análise das secas na Amazônia são mensais e cobrem o período de janeiro de 1999 a dezembro de 2010 (144 meses). Já para a enchente em SC trabalhou-se com pêntadas (médias de 5 dias) e o mesmo período da seca.

O processo de reanálise utiliza um complexo sistema de programas, bibliotecas, documentos e bancos de dados, e envolve etapas como tradução, reformatação, controle de qualidade, análise, predição, pós-processamento, e arquivamento. Bancos de dados de reanálise são criados por assimilação de observações climáticas usando o mesmo modelo climático durante o período completo de reanálise, para reduzir os efeitos de alteração dos modelos na estatística do clima. As observações são provenientes de diferentes origens, como navios, satélites, aviões, estações em terra e radares, dentre outras (KALNAY et al., 1996).

A Tabela 3.1 ilustra as variáveis climatológicas em grade utilizadas. Na primeira coluna aparece o nome das variáveis, na segunda coluna a unidade, na terceira coluna o nível (superfície ou acima da superfície) das variáveis utilizadas na análise da seca na Seca na Amazônia, e na quarta coluna o nível das variáveis utilizadas na análise do evento de precipitação extrema em Santa Catarina.

Tabela 3.1 - Descrição das variáveis usadas nas análises

Variáveis	**Unidade**	**seca Amazonas**	**prec. extrema SC**
Temp. superfície do mar	^0C	superfície	superfície
Pressão nível do mar	Pa	1000hPa	1000hPa
Temp. do ar na superfície	^0C	superfície	superfície
Umidade específica	g/kg	850, 1000hPa	850, 1000hPa
Omega	Pa/s	100, 200, 300, 400, 500, 600, 850, 1000 hPa	100, 200, 300, 400, 500, 600, 700, 850, 1000 hPa
Altura Geopotencial	m	1000hPa	1000hPa
Vento Zonal	m/s	200, 500, 850hPa	200, 500, 850, 1000hPa
Vento Meridional	m/s	200, 500, 850hPa	200, 500, 850, 1000hPa
Cobertura de nuvens	%	superfície	superfície
Fluxo de calor sensível	W/m^2	superfície	não utilizado
Fluxo de calor latente	W/m^2	superfície	não utilizado

Alem dos dados em grade, outras séries temporais foram utilizadas, como El Niño (ENOS) por exemplo (LABORATORY, 2007). A descrição de todas as variáveis encontram-se no Apêndice.

Para visualização dos dados em grade, utilizou-se o pacote GrADS (Grid Analysis and Display System). Este fornece um ambiente integrado para acesso, manipulação e exibição de dados. Para visualização dos dados em ponto de grade utilizam-se dois arquivos: um arquivo descritor (ctl) que descreve as características e informações dos arquivos de dados, tais como, as coordenadas geográficas (latitude e longitude), tempo, nível, número e nome das variáveis, e um arquivo de dados no formato binário (DOTY, 2009).

3.2 Procedimento geral de análise

A seguir, uma breve descrição dos procedimentos executados com os dados desde sua importação do site *http://www.esrl.noaa.gov/psd/data/gridded/data.ncep.reanalysis.html* até a aplicação das ferramentas de mineração de dados.

a) Primeiramente foi definido o período e a área de análise, e em seguida utilizou-se um script em GrADS para transformar os dados adquiridos em arquivo binário.

b) Em seguida foi feito um programa em Matlab para transformar todas as séries temporais geradas pelo GrADS em uma única planilha de dados Excel. Nesta planilha as séries originais são transformadas em séries de anomalias, calculadas com referência às médias mensais (secas na Amazônia) ou pêntadas (precipitação extrema em SC) do período analisado (1999-2010).

c) Posteriormente, foram definidas as séries classificatórias que foram distintas para cada aplicação (explicadas detalhadamente nas aplicações).

d) Com os dados tabulados no formato amigável, é possível montar o projeto no BRB-ArrayTools. A classificação (explicado no item seguinte) foi feita usando-se a opção Class Comparisson do BRB-Array Tools e setando-se o limiar máximo ($\alpha = 1$). Sendo assim, foi emitido um relatório com todas as variáveis e respectivos p-valores.

   Em estudos de microarranjos, os resultados são apresentados em uma lista com genes que são mais semelhantes entre si em termos de nível de expressão, acompanhados por seus respectivos p-valores. Genes com menores

p-valores são mais significantes. Aliando-se a isto e com o objetivo de colocar esta análise em um contexto geográfico e de fácil interpretação, os resultados são apresentados em termos de campos de p-valores.

e) Para imprimir as imagens foram gerados arquivos separados para cada variável que contém as coordenadas e os respectivos p-valores. Em seguida, foram feitos programas em fortran, para converter estes arquivos texto em arquivos binários, para finalmente executar um script na ferramenta GrADS, que lê os arquivos binários e gera as imagens dos p-valores produzidas.

f) Após a análise dos resultados da classificação, com as variáveis mais significativas, foi feito um agrupamento para falcilitar a interpretação do evento extremo na primeira aplicação.

g) Finalmente, para as duas aplicações, foi feito um estudo das variáveis mais relevantes na análise a fim de se gerar o classificador de árvores de decisão. Para isso, o cálculo dos p-valores foi utilizado como redutor de dados. O objetivo das árvores de decisão é apontar a variável de maior impacto nas análises, e os limiares adotados pelas árvores. É importante saber também se a árvore "acerta" a classificação durante o teste nos períodos de eventos extremos. O parâmetro classificatório foi o mesmo utilizado no programa BRB-ArrayTools (mais detalhes, nos próximos capítulos).

Uma etapa importante no projeto é definir qual variável será utilizada para a classificação. Cada projeto busca responder a uma pergunta padronizada do tipo "Quais são as variáveis climatológicas, dentre as consideradas no banco de dados, responsáveis pela propriedade X pertencer a uma determinada classe?". Naturalmente, perguntas diferentes levam a resultados diferentes. Por se tratar de eventos extremos, procura-se avaliar a propriedade que está mais diretamente ligado ao evento. Como estamos estudando seca e enchente, optou-se por usar a precipitação acumulada média em uma determinada área como classificadora.

Com o objetivo de aplicar a comparação de classes da bioiformática, procedemos com a seguinte analogia: variáveis climatológicas substituem genes, e séries temporais substituem sequências de experimentos ou amostras. Finalmente, doenças ou outro processo bioquímico, correspondem ao fenômeno climatológico. Mais especificamente, investigamos as causas (*genes*) dos eventos em questão (seca AM e precipitação extrema em SC) utilizando séries temporais de variáveis climatológicas.

Faz-se necessário definir como será feita a classificação. O mais comum tem sido pela mediana, ou seja: calcula-se a mediana da série classificatória, compara-se cada valor mensal desta série com a mediana - se menor que a mediana, classe C1, se maior, classe C2. Estas classificações podem ser alteradas a qualquer momento pelo usuário. Neste trabalho, foram feitas classificações diferentes para cada análise.

Os algoritmos classificadores de árvores de decisão objetivam criar regras. Neste trabalho, os algoritmos foram executados utilizando-se o software *WEKA* que tem disponível vários algoritmos de árvore de decisão. O algoritmo empregado foi o J4.8.

No conjunto de dados, as instâncias representam as médias dos períodos (meses no caso da Amazônia, e pêntadas no caso de Santa Catarina), e os atributos, as variáveis climatológicas. A classificação baseou-se na mesma série temporal adotada no BRB ArrayTools, diferindo apenas nas classes, que variaram para cada estudo, bem como os arquivos de treinamento e teste.

4 SECA NA AMAZÔNIA

A região de análise selecionada compreende uma sub-região com coordenadas 140°W a 0°W, e 40°N a 40°S, conforme ilustrado na Figura 4.1. Foram utilizadas 44.268 variáveis ambientais.

Figura 4.1 - Região analisada: 140°W a 0°W, e 40°N a 40°S.

4.1 Classificação e Agrupamento

A classificação foi baseada no intervalo entre o menor e o maior valor da anomalia da precipitação acumulada numa região da Amazônia em uma área delimitada pela região de coordenadas 4°S e 8°S, 68°W e 72°W (região delimitada pelo quadradinho vermelho na Figura 4.1). Esta região localizada no oeste da Amazônia foi fortemente afetada pelas secas de 2005 e 2010 (LEWIS et al., 2011). A figura 4.2 ilustra a série temporal em anomalia da precipitação acumulada nesta região. Observa-se que em 2005 e 2010 a precipitação esteve abaixo da média. A série de precipitação foi obtida do site da NASA extraído de http://disc2.nascom.nasa.gov/Giovanni/tovas/rain.3B43_-V6_anom.shtml. São dados mensais de precipitação acumulada com resolução de $0,25°$ x $0,25°$.

Para um estudo mais específico de seca, foram definidas três classes: seca, neutra e chuvosa. Elas foram definidas de tal forma a se capturar os principais episódios de seca ocorridos durante o período em estudo. Com este objetivo dividiu-se o intervalo entre o maior e o menor valor de anomalia da precipitação em três partes, atribuindo à seca e à chuva 37% dos extremos da série para cada, e o restante 26% para classe neutra, ou seja:
classe seca = $36,6\%$ do intervalo entre a menor e maior anomalia = 67 meses,

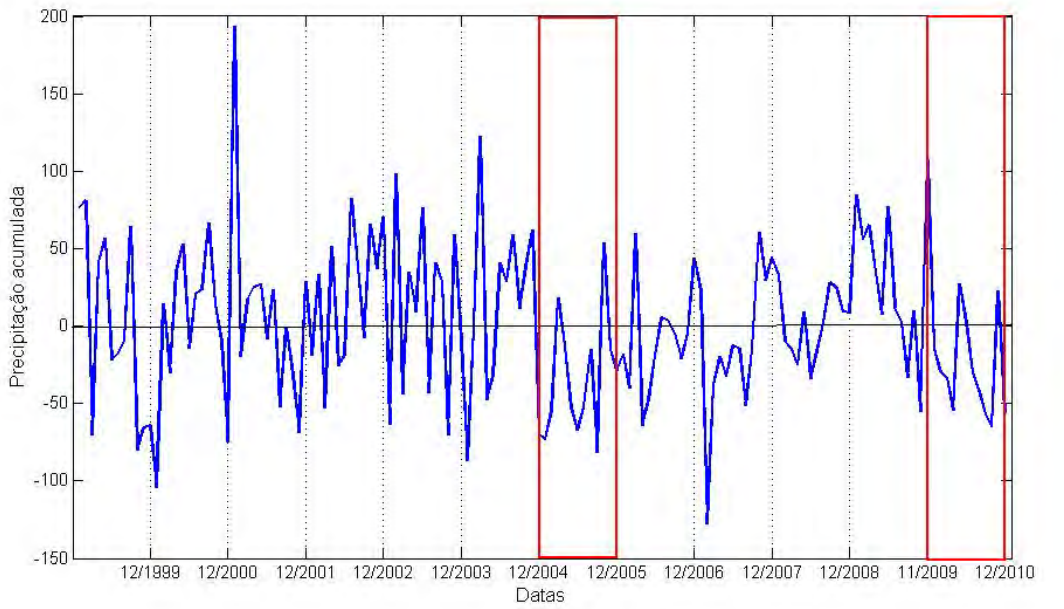


Figura 4.2 - Precipitação acumulada em anomalia na região em estudo.

classe neutra $= 26,8\%$ do intervalo entre a menor e maior anomalia $= 67$ meses, classe chuvosa $= 36,6\%$ do intervalo entre a menor e maior anomalia $= 10$ meses.

Dentro deste contexto, a comparação de classes objetiva determinar quais dentre as 44.268 variáveis se comportam diferentemente entre as diferentes classes: seca, neutra e chuvosa. Os resultados apresentados das Figuras 4.3 a 4.8 correspondem a comparação entre seca x neutra e são apresentados em campos de p-valores. Resultados adicionais encontram-se no Anexo A (Figuras .1 a .8).

Os p-valores em cada ponto de grade podem ser interpretados como a probabilidade que a correlação observada em cada variável da grade e a precipitação acumulada na área delimitada pelo quadrado vermelho é produto do acaso. Claramente, padrões coerentes de baixos p-valores (áreas escuras) chamam a atenção. Observa-se que o déficit de chuva é consistente com o aumento generalizado da temperatura da superfície do mar no Atlântico tropical Norte que se estende até o oeste da África (Figura 4.3). Isto pode ser observado pela evolução temporal de 1999 a 2010 da anomalia da temperatura da superfície do mar em um ponto de grade com baixo p-valor (estrela vermelha na Figura 4.3). Este condição de anomalia quente é acompanhada pelo aumento dos níveis de temperatura do ar e umidade específica sobre o Atlântico tropical Norte, observado nas Figuras 4.4 e 4.5. Estes resultados confirmam estudos anteriores de Marengo et al. (2011), Tomasella et al. (2011), Lewis et al. (2011), Yoon

e Zeng (2010), Marengo et al. (2008b), Cox et al. (2008), Marengo et al. (2008a), Zeng et al. (2008), Malhi et al. (2008) que apontam o aquecimento da temperatura do Atlântico Norte como o principal fator responsável pelas secas de 2005 e 2010. Durante estes eventos extremos, as anomalias da temperatura da superfície do mar alcançaram 1°C e $0,9$°C respectivamente, dois recordes desde 1902 (MARENGO et al., 2011).

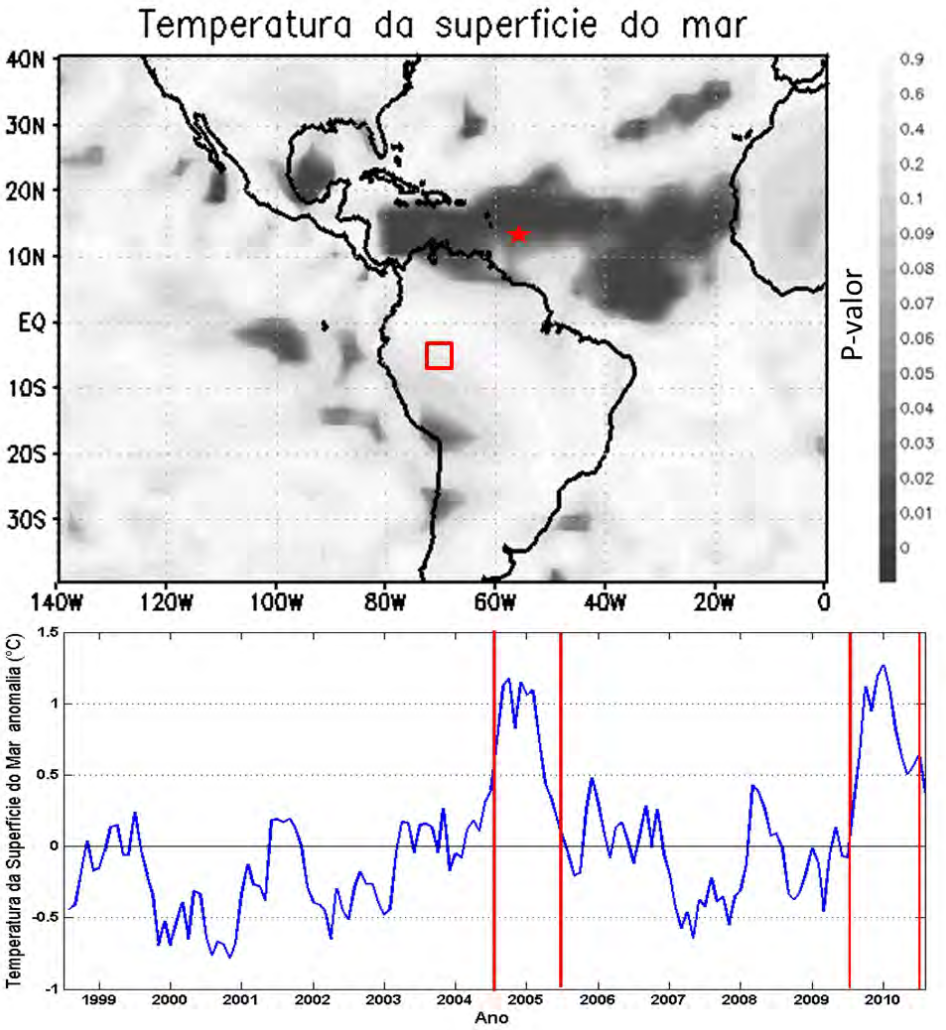


Figura 4.3 - Campos de p-valor para a temperatura da superfície do mar em anomalia. Abaixo: evolução temporal da temperatura da superfície do mar em anomalia a $12,5$°N$-55,5$°W, ponto de grade localizado em região de baixo p-valor (estrela vermelha), de 1999 a 2010.

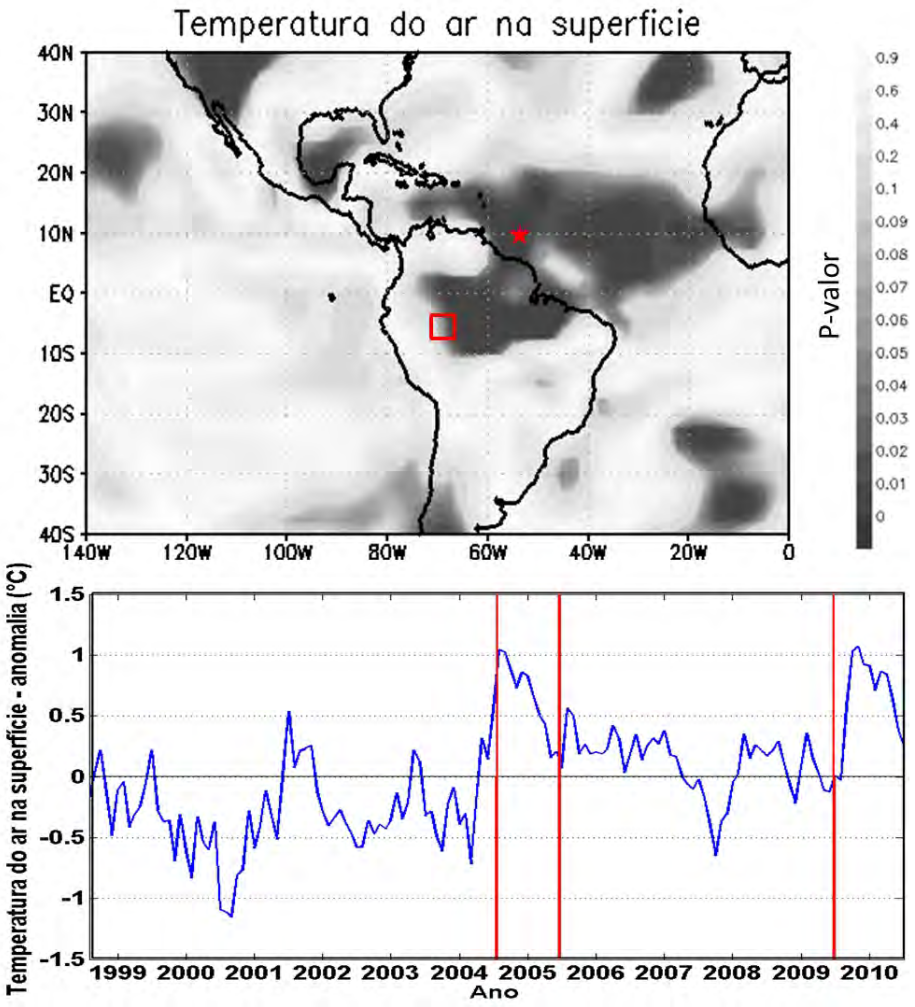


Figura 4.4 - Campos de p-valor para a temperatura do ar na superfície em anomalia. Abaixo: evolução temporal da temperatura do ar na superfície em anomalia a 10°N−55°W, ponto de grade localizado em região de baixo p-valor (estrela vermelha), de 1999 a 2010.

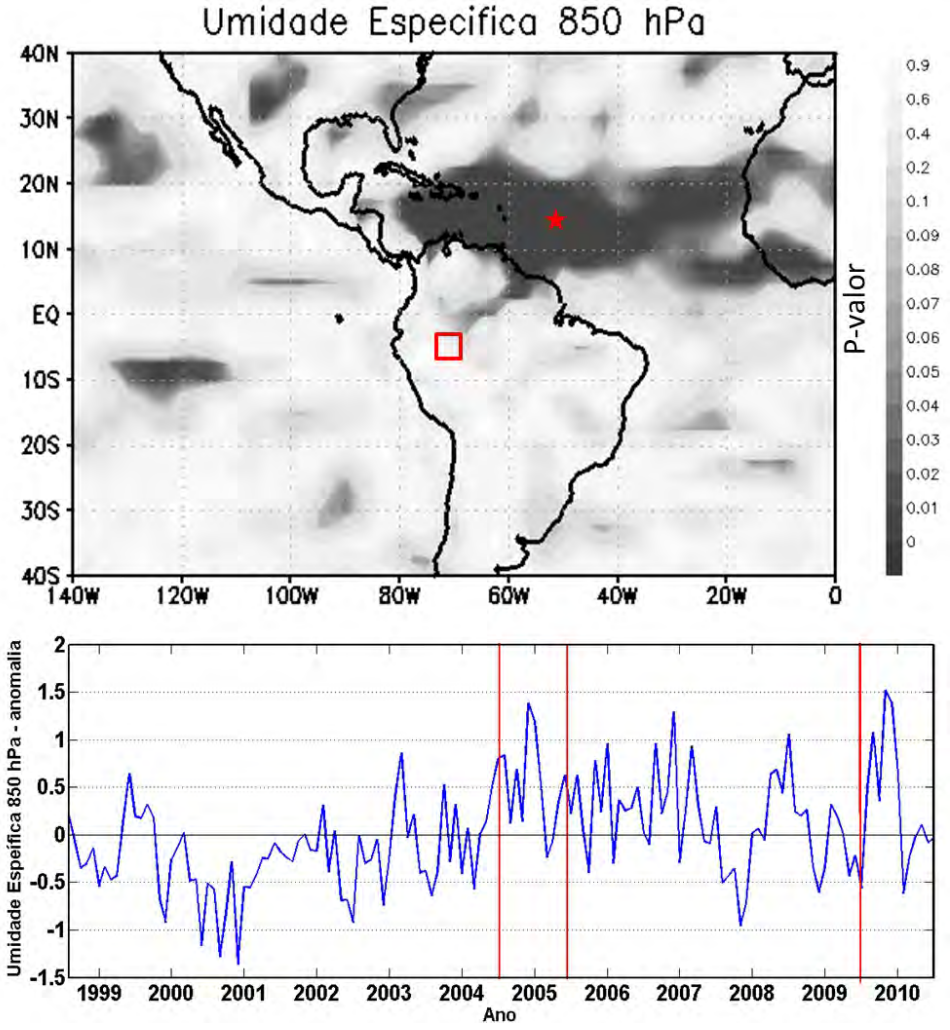


Figura 4.5 - Campos de p-valor para umidade específica em a nomalia a 850 hPa. Abaixo: evolução da umidade específica em anomalia a 15°N−50°W, ponto de grade localizado em região de baixo p-valor (estrela vermelha), de 1999 a 2010.

As figuras 4.6 e 4.7 apresentam campos de p-valores para vento zonal e meridional a 850 hPa (média da anomalia de Agosto e Setembro de 2010 sobrepostas), e anomalia da pressão ao nível do mar. Na parte inferior da Figura 4.7 tem-se a diferença de pressão entre os dois pontos de grade com baixos p-valores, um no oceano e outro no continente. Observa-se que esta diferença de pressão é negativa em 2005 e 2010 resultando em uma diminuição do fluxo de umidade do oceano para o continente, logo um déficit de precipitação. Estes resultados são consistentes com um padrão de circulação caracterizado pelo transporte de ventos alíseos mais fraco que o normal para o sul da Amazônia, como sugerido por alguns autores (MARENGO et al., 2008b; MARENGO et al., 2011; ZENG et al., 2008; COELHO et al., 2012). Observa-se que a diferença de pressão entre os pontos de grade 15°N-50°W e 5°S-60°W permanecem

negativas durante a maior parte de 2005 e 2010, resultando em redução de fluxo de umidade do Atlântico tropical Norte e menos precipitação sobre a Amazônia.

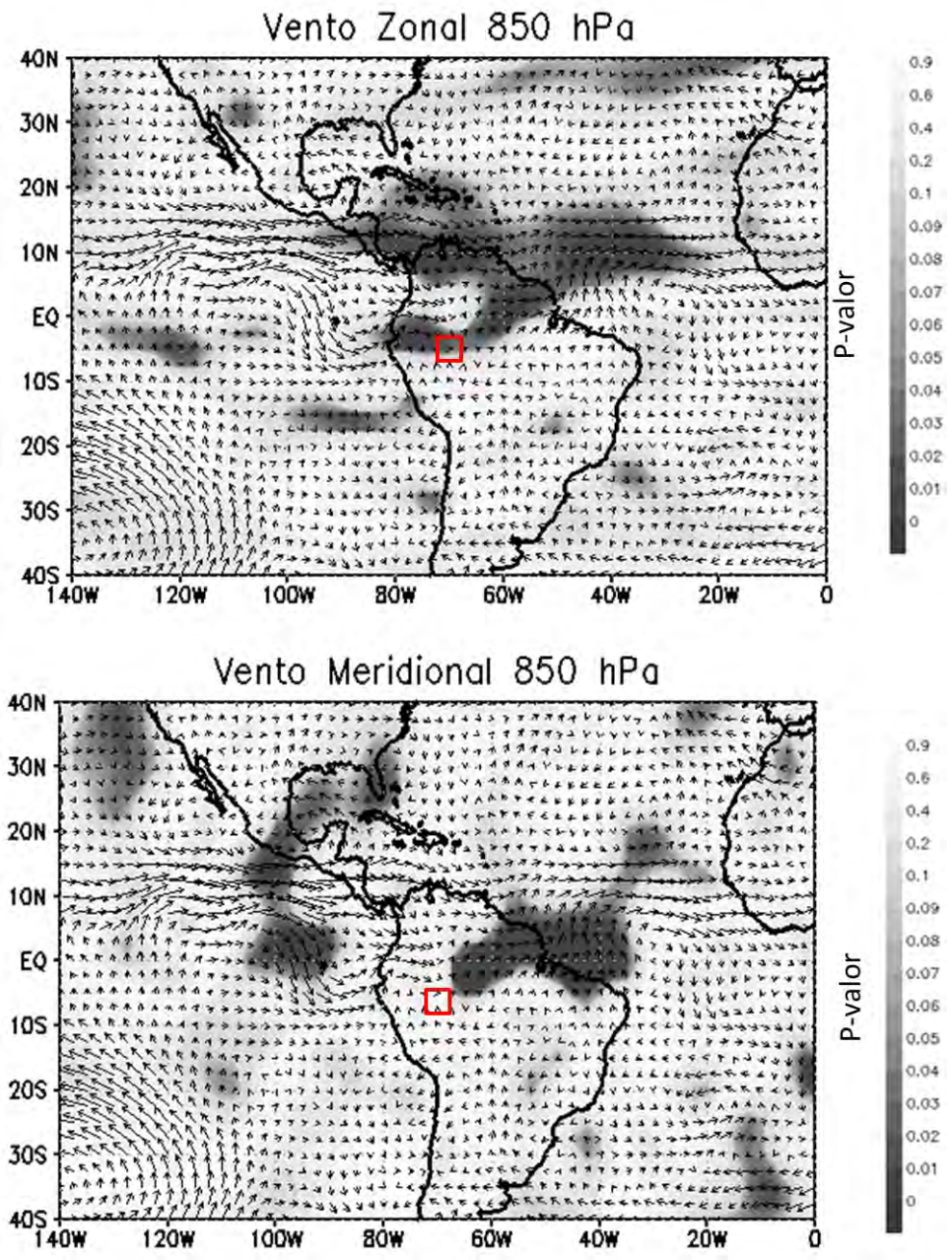


Figura 4.6 - Campos de p-valor para vento zonal (acima) e meridional (abaixo) em anomalia a 850 hPa; média das anomalias dos ventos entre Agosto e Setembro de 2010 sobrepostas.

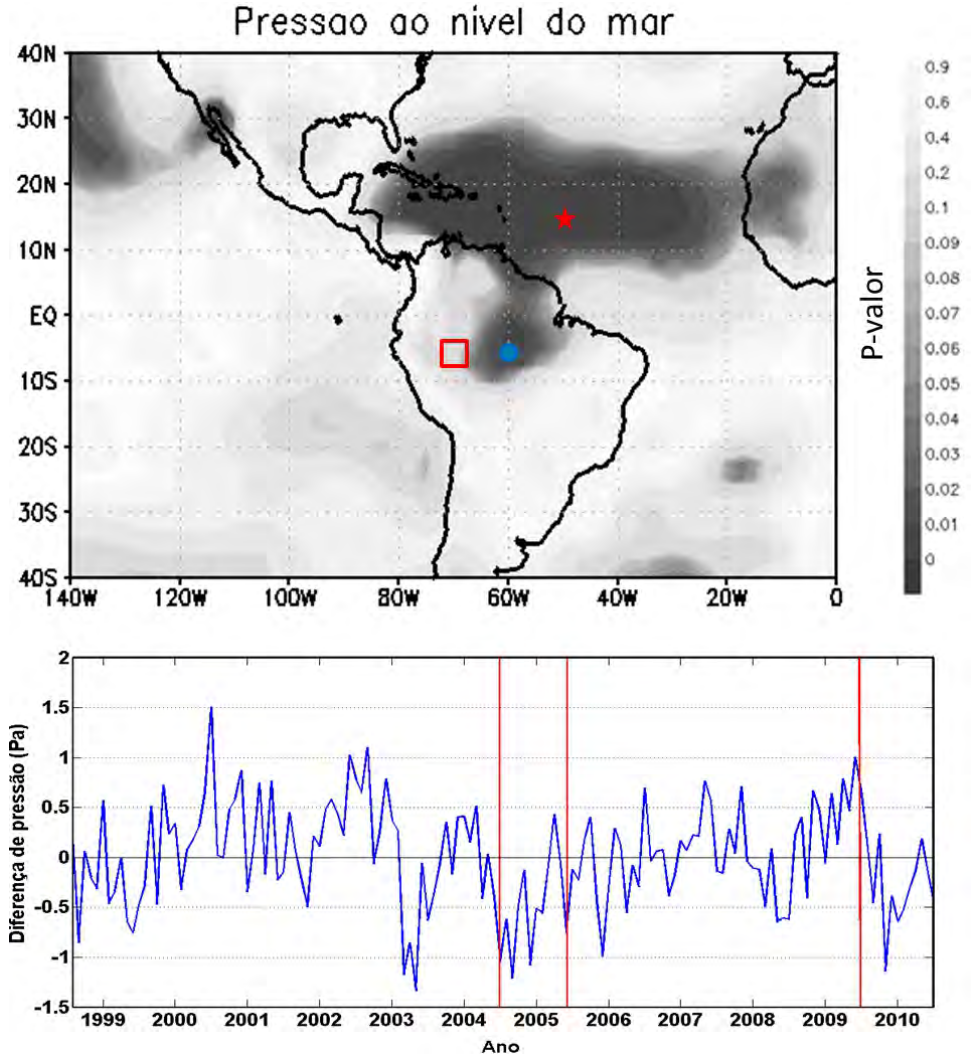


Figura 4.7 - Campos de p-valor para pressão ao nível do mar em anomalia. Abaixo: diferença de pressão entre os dois pontos de grade com baixos p-valores 15°N−50°W (estrela vermelha) e 5°S−60°W (círculo azul), from 1999 to 2010.

A chuva sobre a Amazônia é fortemente impactada por estruturas de larga-escala, como a circulação Leste-Oeste de Walker e o mecanismo de circulação do meridiano local, conhecido como célula de Hadley do Atlântico (COOK et al., 2010; YOON; ZENG, 2010). A influência do Atlântico sobre e Amazônia é modulada pelas variações sazonais e interanuais na robustez e na posição da zona de convergência intertropical (ITCZ) que é influenciada por mudanças na temperatura da superfície do mar no Atlântico Tropical. Conjecturou-se por Cook et al. (2010), Yoon e Zeng (2010), Tomasella et al. (2011), Marengo et al. (2011), Marengo et al. (2008b), Zeng et al. (2008) que os recentes episódios de seca intensa na região estão ligadas ao deslocamento para noroeste da ITCZ. Em 2010, por exemplo, a ITCZ foi deslocada de sua posição climática por aproximadamente cinco graus para o norte (MARENGO et

al., 2011). Este cenário pode ser observado na Figura 4.8 que mostra os campos de p-valores para omega a 500 hPa, junto com campos de anomalia em duas sub-áreas da região, para Agosto-Setembro 2009 e 2010. Estes dois anos são usados aqui como paradigma de anos com precipitação acumuladas acima e abaixo da média climática respectivamente. Anomalias negativas indicam movimento ascendente de vento.

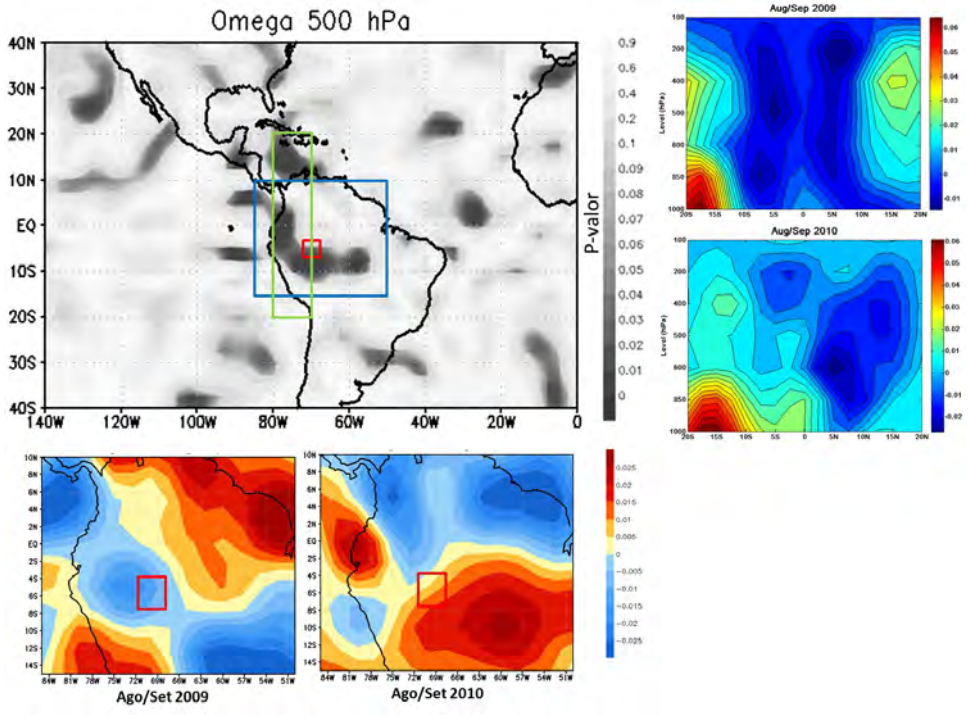


Figura 4.8 - Campos de p-valor para omega em anomalia a 500 hPa. Abaixo: média da anomalia em Agosto/Setembro sobre o quadrado azul, para 2009 e 2010. À direita, uma secção transversal das anomalias de omega para as longitudes em verde, de 1000 a 100 hPa para 2009 e 2010.

Não foram observados variações da temperatura da superfície do mar no Pacífico durante esta temporada. Em 2010 a seca sobre a floresta Amazônica iniciou no princípio do verão austral durante o El Niño e foi subsequentemente intensificada com o aquecimento do Atlântico tropical Norte (MARENGO et al., 2011).

Fortes eventos de El Niño geralmente causam déficit sobre a Amazônia central e nordeste (TOMASELLA et al., 2011; COOK et al., 2010). Estudos anteriores (COX et al., 2008; TOMASELLA et al., 2011; MARENGO et al., 2008b; MARENGO et al., 2008a) mostram que não existe correlação entre as secas de 2005 e 2010 que atingiram principalmente o sudoeste da Amazônia, e o fenômeno El Niño. Em concordância com estes autores, este trabalho não detectou a influência de eventos tipo El Niño

(veja Tabela 4.1). No entanto, a pequena área de baixos p-valores claramente visível na figura 4.3, a oeste das Ilhas Galápagos, destaca a necessidade de uma análise mais aprofundada (por exemplo, com as ferramentas de correlação defasada; ver (YOON; ZENG, 2010)) para melhor entender o complexo mecanismo de conexão da precipitação na Amazônia com a TSM no Atlântico Tropical Norte e as atividades de ENOS no Pacífico.

Tabela 4.1 - Relação dos indices extras analisados com os respectivos p-valores

Série temporal	**p-valor**
Equatorial SOI	0,5189
Equatorial Eastern Pacific SLP	0,7216
Stand Tahiti - Stand Darwin SLP	0,9511
SST Niño	0,2808
SST North Atlantic	0,0070
SST South Atlantic	0,9696
North Atlantic Oscillation - NAO	0,7220
Pacific Decadal Oscillation - PDO	0,1234

Os resultados acima podem ser sumarizados por meio do agrupamento hierárquico de um subconjunto de parâmetros significativos, selecionados dentre os resultados obtidos pela classificação. Este subconjunto inclui as 50 variáveis com menor p-valor e está ilustrado na Figura 4.9. A Figura 4.10 ilustra o agrupamento das 4 variáveis com menor p-valor nas figuras 4.3 a 4.8, totalizando 28 variáveis. Para uma melhor interpretação, as colunas foram rearranjadas em ordem cronológica. As cores representam as intensidades da anomalia normalizada.

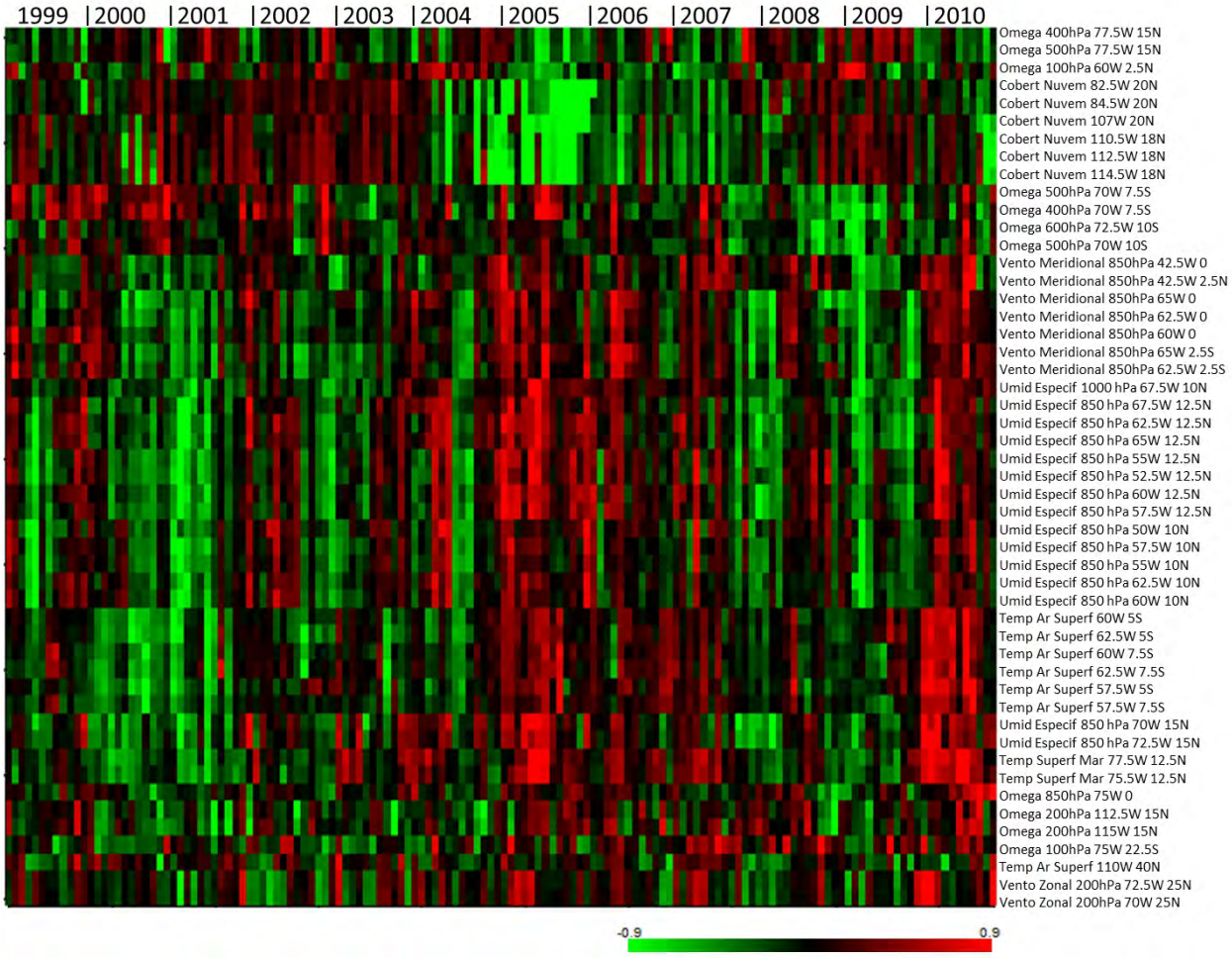


Figura 4.9 - Agrupamento hierarchico das cinquenta variáveis com menor p-valor para os 12 anos em estudo (de 1999 a 2010).

Os resultados do agrupamento para o período de 1999 a 2010 estão ilustrados na Figura 4.10, e o agrupamento de 2009 e 2010 estão ilustrados na Figura 4.11. As Figuras 4.10 e 4.11 apresentam no topo respectivamente a anomalia da precipitação acumulada, e a vazão mensal do Rio Madeira. Observa-se que diferenças na precipitação acompanham muito bem o comportamento dos níveis de expressão das variáveis climatológicas. A predominância de cores verde e vermelha alternadas dependem se o ano era "seco" ou "chuvoso". Observa-se que em Fevereiro de 2010, seis meses antes de os níveis dos rios no sudoeste da Amazônia atingirem seus mínimos, variáveis como temperatura da superfície do mar nas coordenadas $75,5°$W-$12,5°$N, vento zonal a 850 hPa em $75°$W-$7,5°$N, e omega a 500 hPA em $77,5°$W-$15°$N (em negrito na Figura 4.11) mudam abruptamente como se o sistema climático alternasse entre dois estados diferentes. Padrão similar se observa no final de 2004.

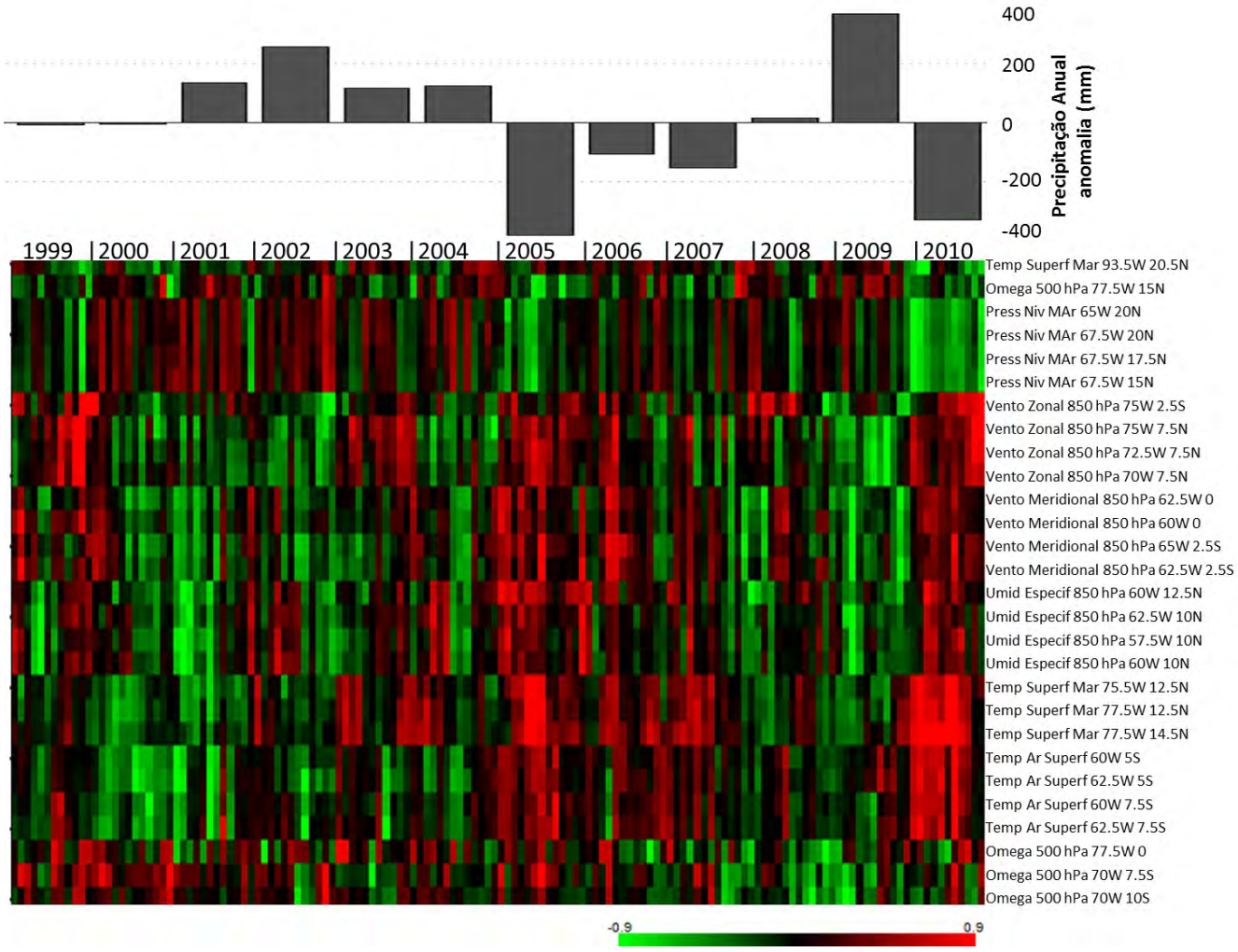


Figura 4.10 - Agrupamento hierarchico das quatro variáveis com menor p-valor em cada um dos sete campos de p-valor apresentados nas figuras anteriores, para o período de 1999-2010. As colunas foram rearranjadas e inseridas em ordem cronológica. As cores representam a anomalia normalizada da intensidade.

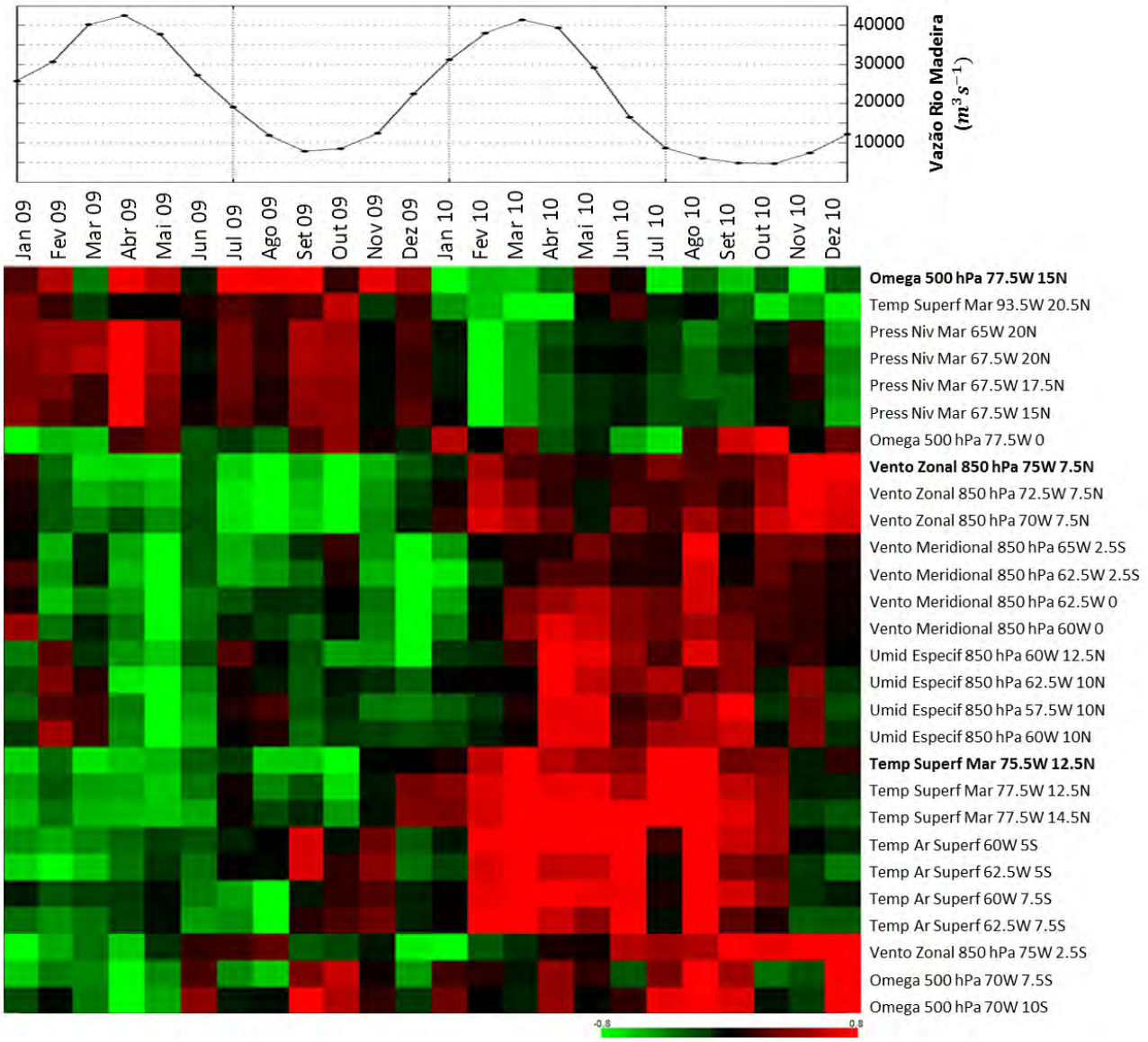


Figura 4.11 - Agrupamento hierarchico das quatro variáveis com menor p-valor em cada um dos sete campos de p-valor apresentados nas figuras anteriores, para o período de 2009 e 2010. As colunas foram rearranjadas e inseridas em ordem cronológica. As cores representam a anomalia normalizada da intensidade.

Neste sentido, é notável como um pequeno subconjunto de variáveis climatológicas pode ser usado para reduzir efetivamente a complexidade do conjunto de dados original. Esta complexidade pode ser ainda mais reduzida através da representação gráfica em baixa dimensão ilustrada na Figura 4.12, que usa apenas as variáveis em negrito da Figura 4.11. Observa-se que os anos chuvosos e secos de 2002/2009 e 2005/2010, respectivamente, estão claramente segregados em diferentes regiões do gráfico.

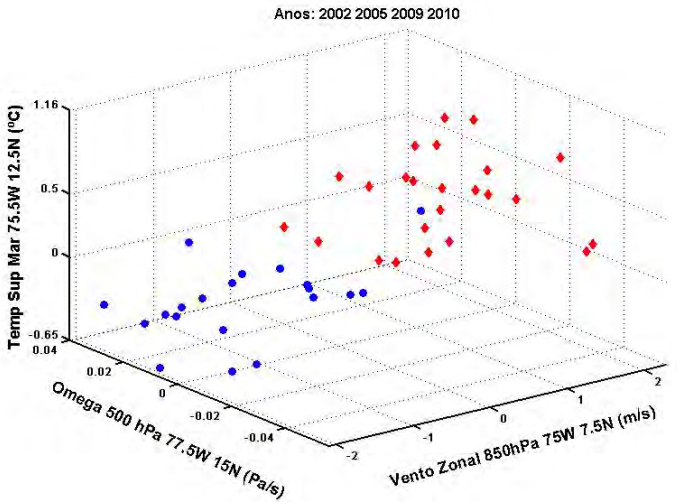


Figura 4.12 - Gráfico do agrupmento tri-dimensional da tempertura da superfície do mar a 75, 5°W-12, 5°N versus vento zonal a 850 hPa e 75°W-7, 5°N versus omega a 500 hPa e 77, 5°W-15°N, para 2002/2009 (anos "secos") e 2005/2010 (anos "umidos").

Coelho et al. (2012) demonstraram que seria possível prever, com um mês de antecedência, um estação seca mais intensa que o usual na Amazônia. Os resultados apresentados aqui, apesar de limitados a 12 anos e a dois eventos extremos neste intervalo, sugerem que esta previsão é viável com vários meses de antecedência; veja Figura 4.11. Em 2005 e 2010, o Rio Amazonas obteve seu menor nível por décadas, com vários afluentes completamente secos e mais de 20 municípios declararem estado de emergência. Seria muito importante se estes períodos anômalos de seca pudessem ser previstos com mais antecedência permitindo que governo e comunidades pudessem tomar ações apropriadas de mitigação.

4.2 Árvores de decisão

Foram feitos testes para verificar quais os arquivos ideais (treinamento e teste) a serem executados pelas árvores de decisão através do algoritmo J4.8. Para isso, foram usadas as variáveis com os menores p-valores resultantes da análise do BRB-ArrayTools. Como a primeira grande seca ocorreu em 2005, o conjunto de treinamento compreende as séries temporais de 1999 a 2004, e para os testes as séries temporais de 2005 a 2010. A classificação foi feita analisando a mesma série de precipitação da aplicação com o BB-ArrayTools que foi dividida em 2 classes de acordo com a mediana: pouco (valores abaixo da mediana) e muito (valores acima da mediana), representando respectivamente pouca chuva e muita chuva. Para o treinamento do algoritmo J4.8, foram utilizados os parâmetros default C=0,25 (fa-

tor de confiança) e M=2 (número de instâncias por folha). De todos os testes feitos, apresenta-se abaixo 5 casos ilustrativos:

- Caso1
  Utilizadas as variáveis com p-valores menores que 0,001, ou seja as variáveis com uma probabilidade menor de 0,1% de serem falso-positivas, totalizando 182 variáveis;

- Caso2
  Utilizadas as variáveis com p-valores menores que 0,005, ou seja as variáveis com uma probabilidade menor de 0,5% de serem falso-positivas, totalizando 825 variáveis;

- Caso3
  Utilizadas as variáveis com p-valores menores que 0,0005, ou seja as variáveis com uma probabilidade menor de 0,05% de serem falso-positivas, totalizando 104 variáveis;

- Caso4
  De todas as variáveis climatológicas analisadas, não em coordenadas (por exemplo omega 850 hPa, Temperatura do ar, Vento Zonal), foram selecionadas no máximo 10 em coordenadas daquelas com menor p-valor e que tivessem a condição do p-valor $<$ 0,001, totalizando 120 variáveis;

- Caso5
  As 28 variáveis representadas na Figura 4.10.

A Tabela 4.2 mostra a porcentagem de acertos dos casos de testes citados acima considerando-se dois conjuntos de testes: geral - utiliza todo o arquivo de teste (2005 a 2010), e secaAM - utiliza apenas os meses de Julho a Dezembro de 2005 e 2010 nos testes.

Com os resultados apresentados na Tabela 4.2, observa-se que nos casos 1 e 4 existe uma porcentagem de acerto geral próxima de 70%, e de 83% quando se analisam os períodos de maior seca. Mais precisamente, a árvore "erra" em apenas 2 meses durante os segundos semestres de 2005 e 2010. No caso 1 foram necessários 62 variáveis a mais do que o caso 4 para se identificar a seca. Isto significa que existe informação redundante das variáveis. Logo, não é necessário um grande número de variáveis para se criar uma árvore de decisão. O caso 3 é muito restrito e com isso,

Tabela 4.2 - Porcentagem de acerto nos testes executados

Casos	**conjuntos teste**	**%acertos**
1	geral	69,44%
	secaAM	83,33%
2	geral	62,5%
	secaAM	50%
3	geral	70,83%
	secaAM	75%
4	geral	69,44%
	secaAM	83,33%
5	geral	66,66%
	secaAM	83,33%

pode-se observar que alguns eventos não foram identificados. O caso 2, que contém o maior número de variávei dentre todos analisados, foi o que obteve o menor número de acertos no período de seca, com isto conclui-se que o banco de dados em questão inclui muitos eventos com informação não relevante para a identificação da seca profunda. O caso 5 também obteve uma boa porcentagem de acertos no período da seca, mas como foi utilizado um banco de dados arbitrário, não pode ser generalizado na análise.

De todos os testes executados, o caso 4 foi apontado como o melhor classificador, e com a vantagem que a árvore gerada é a mesma do caso 1, ilustrada na Figura 4.13. Observa-se que a variável com maior ganho de informação apontada pelo classificador em questão foi Vento Zonal a 200 hPa, até então não citada na literatura. Por outro lado não existe na literatura um classificador tipo árvore de decisão que utilize um grande número de variáveis climatológicas, todas distintas.

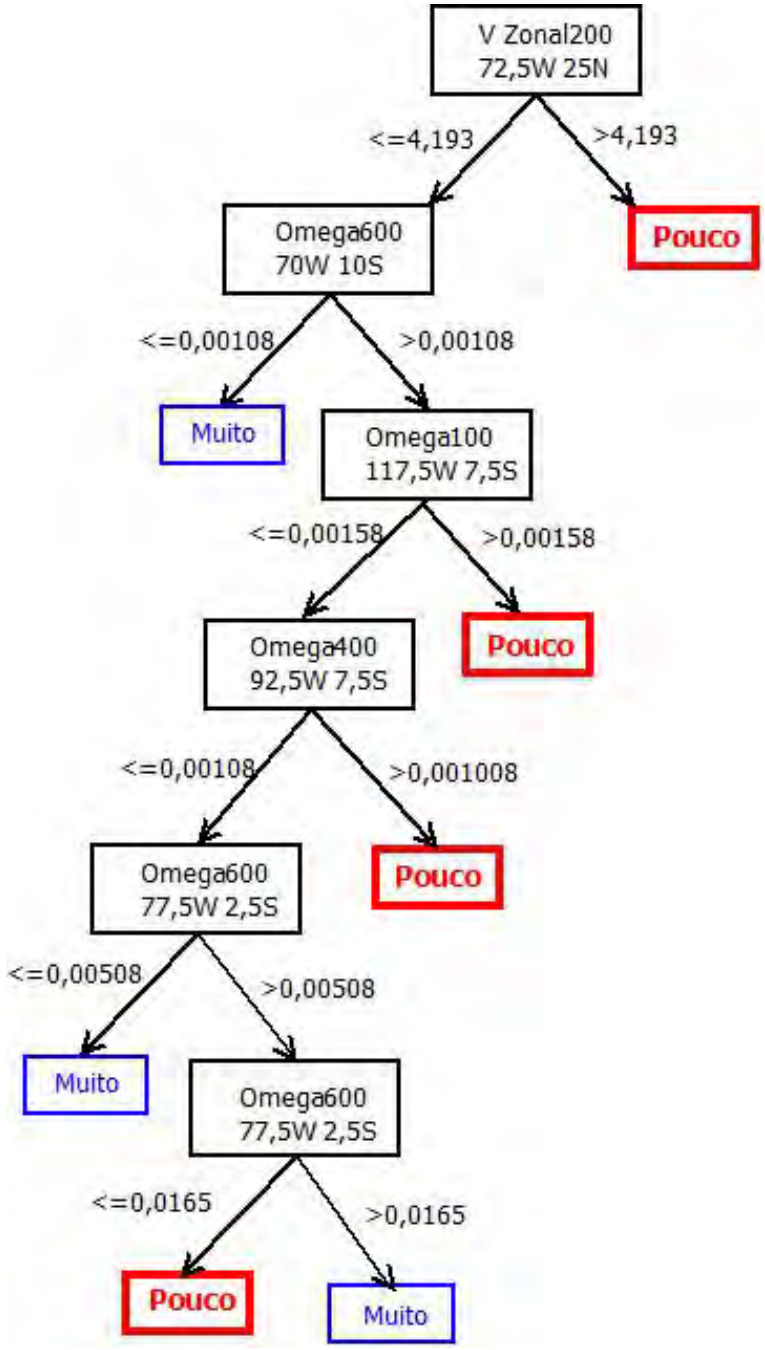


Figura 4.13 - Árvore de decisão gerada com entropia de Shannon para o caso 4. Arquivo de treinamento compreende os anos de 1999 a 2004, e arquivo de teste, os anos de 2005 a 2010.

A Tabela 4.3 detalha os acertos e erros nos períodos de Julho a Dezembro de 2005 e 2010 para o Caso 4, ou seja, na primeira coluna aparecem os meses de maior seca. Na segunda coluna, é apontanda a saída desejada de acordo com a classificação em relação a mediana (acima da mediana, chove muito, abaixo chove pouco). Na terceira coluna é apresentada a saída gerada pela árvore de decisão. Observa-se que a árvore

acerta o período de seca de acordo com as variáveis treinadas.

Tabela 4.3 - Análise das saídas nos períodos de seca

Mês Ano	**saida desejada**	**saída Shannon**
Julho/2005	Pouco	`Pouco`
Agosto/2005	Pouco	`Pouco`
Setembro/2005	Pouco	`Pouco`
Outubro/2005	Muito	`Muito`
Novembro/2005	Pouco	Muito
Dezembro/2005	Pouco	`Pouco`
Julho/2010	Pouco	`Pouco`
Agosto/2010	Pouco	`Pouco`
Setembro/2010	Pouco	Muito
Outubro/2010	Pouco	`Pouco`
Novembro/2010	Muito	`Muito`
Dezembro/2010	Pouco	`Pouco`

5 PRECIPITAÇÃO EXTREMA EM SANTA CATARINA

A região de análise selecionada compreende uma subregião com coordenadas $30°W$ a $60°W$, e $20°S$ a $50°S$, conforme ilustrado na Figura 5.1. Foram utilizadas 4.036 variáveis ambientais. Diferente dos episódios de seca, precipitações intensas como as de SC desenvolvem-se em períodos muito curtos, dificultando as ações de prevenção e mitigação. Por este motivo, optou-se por trabalhar com pêntadas, ou seja, foram calculadas as médias de cinco dias dentro do período de 1999 a 2010. Sendo assim, temos 73 valores anuais para cada variável no período. Em seguida foram calculadas as anomalias dentro do período, ou seja, calculou-se a média para cada pêntada no ano (73 médias), e em seguida subtraiu-se o valor da pêntada da média.

Figura 5.1 - Região analisada em destaque.

5.1 Classificação

A classificação foi baseada na precipitação medida em uma região fortemente afetada pela enchente em SC (círculo vermelho na Figura 5.1). Os dados de precipitação são provenientes das estações pluviométricas administradas pela Agência Nacional de Águas (ANA), obtidos diretamente do endereço *http://ana.gov.br/portalsnirh/*. O site fornece dados de estações espalhadas pelo Brasil. Dentro das opções de estação, foram selecionadas 5 estações pertencentes a região afetada pela enchente, como também aquelas que contém todos os dados diários dentro do período em análise. Em seguida, foi calculada a média da precipiação diária das cinco estações. Abaixo, estão listadas as estações escolhidas e suas respectivas coordenadas geográficas.

- Latitude: -26,4642 Longitude: -49,0867 - Municipio - Jaraguá do Sul SC
- Latitude: -27,0403 Longitude: -49,3814 - Municipio - Apaiuna SC

- Latitude: -26,4239 Longitude: -49,2925 - Municipio - Corupá SC
- Latitude: -26,0356 Longitude: -48,85 - Municipio - Garuva SC
- Latitude: -26,7408 Longitude: -49,2706 - Municipio -Rio dos Cedros SC

Para obter a série temporal da precipitação, foi calculada a média de cada pêntada nas 5 estações acima, e a anomalia dentro do período, gerando assim apenas uma série temporal ilustrada na Figura 5.2. O trecho de série delimitada entre as barras vermelhas representa a época de maior chuva. A classificação foi baseada no intervalo entre o menor e o maior valor da série em anomalia gerada. Para uma análise mais específica, dividiu-se o intervalo de anomalia em 3 sub-intervalos: maior anomalia e 8 ([43,5, 8])- representando chuva abundante, 8 e 0 ([8, 0])- representando chuva moderada, 0 e menor anomalia ([0, -11,2])- representando chuva fraca. Esta divisão objetivou identificar períodos mais relevantes de chuva.

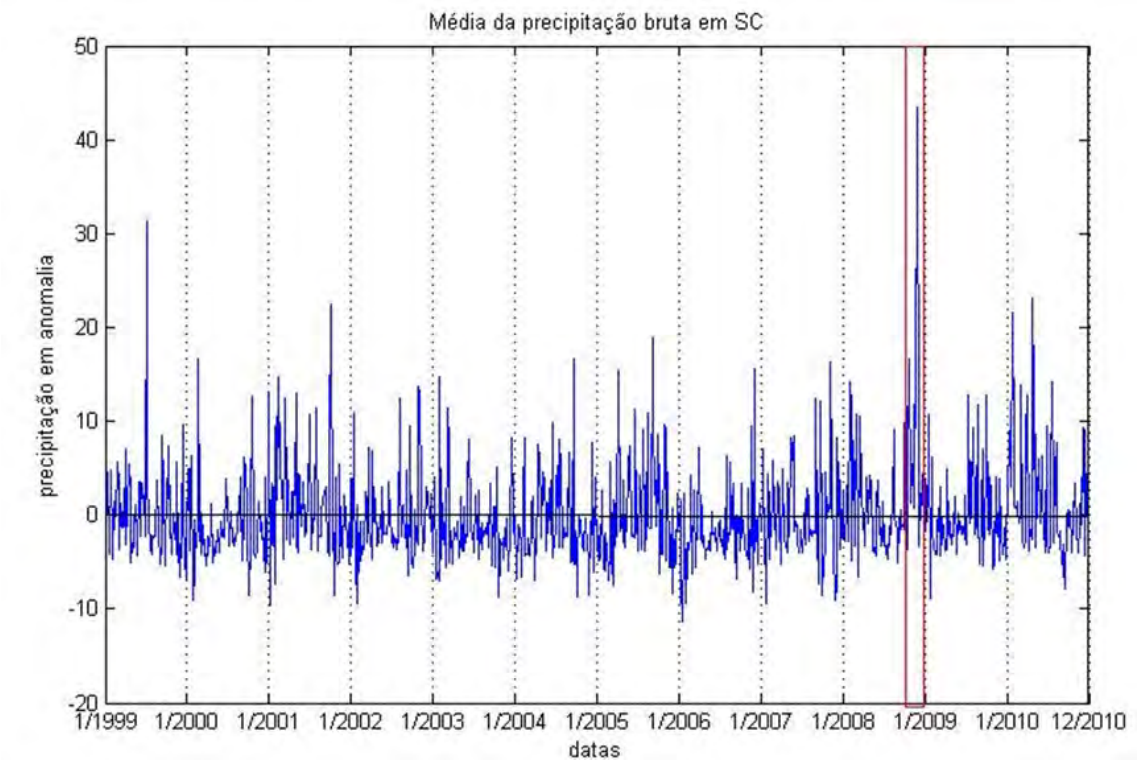
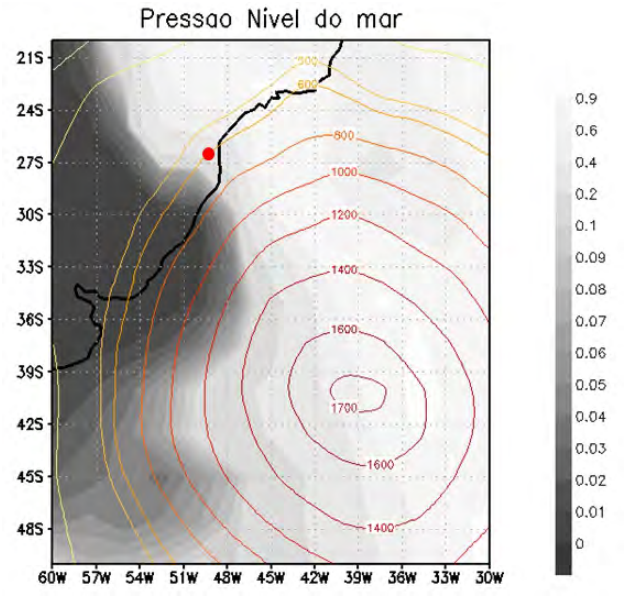


Figura 5.2 - Precipitação acumulada em anomalia na região em estudo.

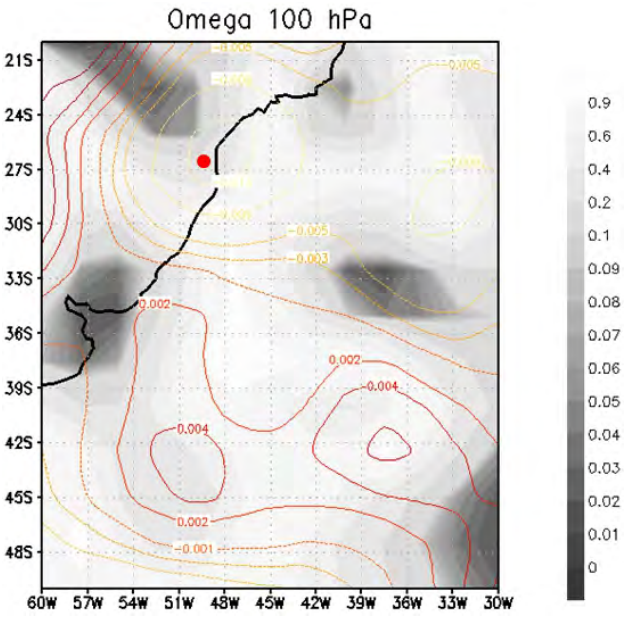
Dentro deste novo contexto, a comparação de classes objetiva determinar quais dentre as 4.036 variáveis se comportam diferentemente entre as diferentes classes: chuva abundante, moderada e fraca. Os resultados apresentados nas Figuras 5.3 a 5.10 correspondem a comparação entre chuva abundante x chuva moderada, considerando sempre valores positivos de anomalia de chuva. As isolinhas coloridas representam o dado em anomalia da pêntada com maior índice de chuva que foi de 22 a 26 novembro 2008 (intervalo da série entre barras vermelhas na Figura 5.2). Resultados adicionais encontram-se no Anexo B (Figuras .1).

Ressalta-se que nos resultados as regiões com tonalidades mais escuras estão associadas aos menores p-valores e, portanto, às variáveis mais significativas. Nestes resultados (Figuras 5.3 a 5.7), chamam a atenção a densa área escura de omega que se estende do Oceano Atlântico até o litoral de SC. Observa-se pelas isolinhas que no período da enchente, omega está negativo no continente (movimento vertical ascendente) e positivo no oceano (movimento vertical descendente). Como é sabido, movimento ascendente favorece a precipitação. As áreas escuras resultantes da análise de omega se repetem em todos os níveis de pressão. Os resultados acima indicam que este processo de movimento descendente sobre o oceano acoplado ao

movimento ascendente sobre o continente está diretamente relaciondo ao alto nível de precipitação no Vale do Itajaí durante o episódio de enchente aqui estudado.

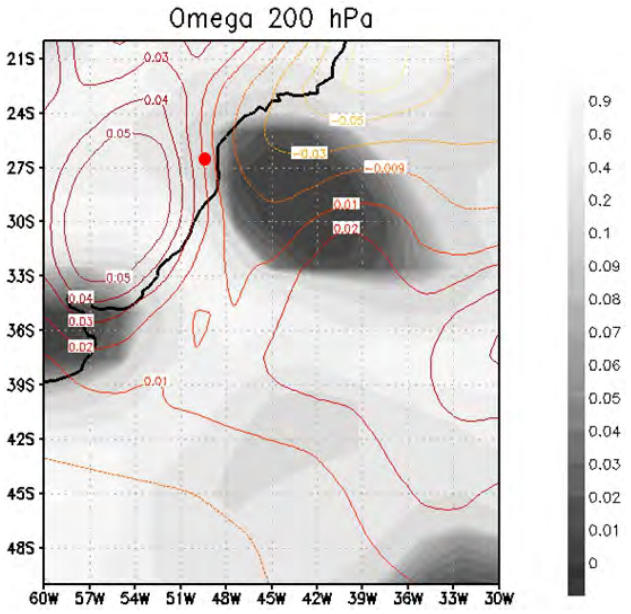


(a)

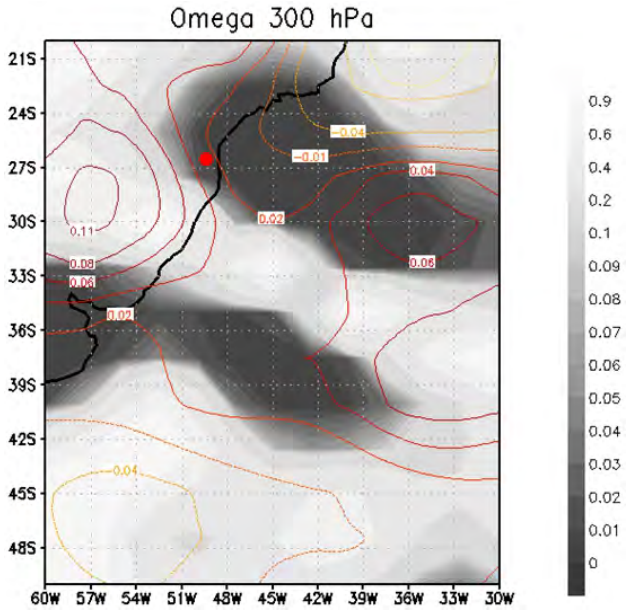


(b)

Figura 5.3 - Representação em p-valores da influência das variáveis climatológicas no evento de precipitação extrema em Santa Catarina - 2008

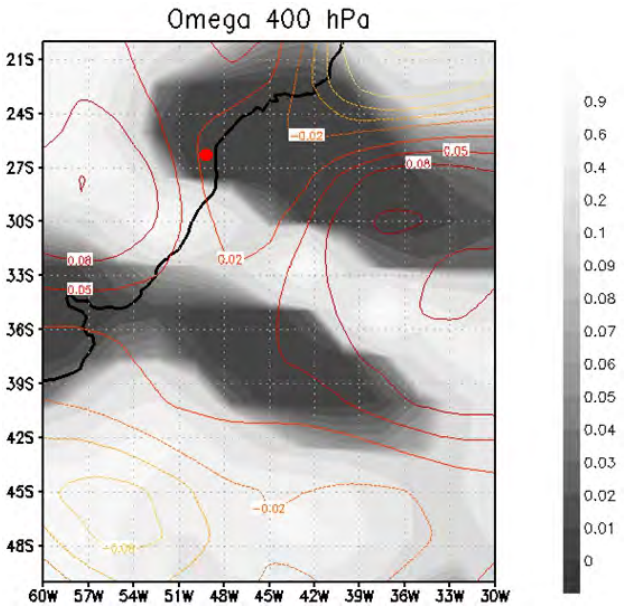


(a)

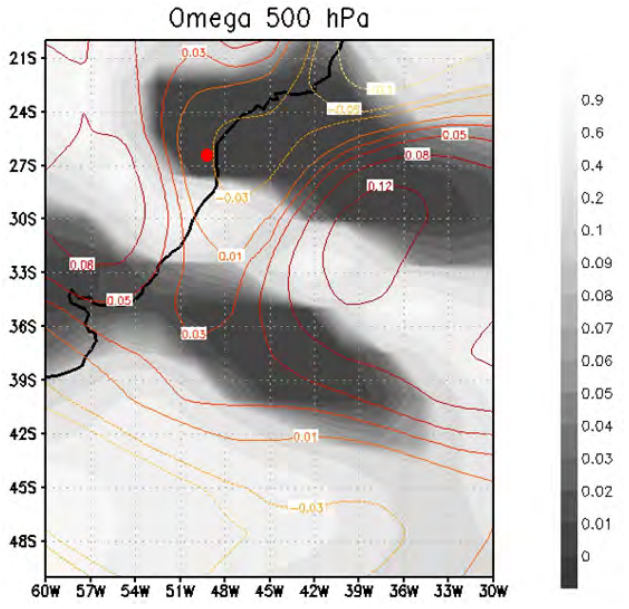


(b)

Figura 5.4 - Representação em p-valores da influência das variáveis climatológicas no evento de precipitação extrema em Santa Catarina - 2008

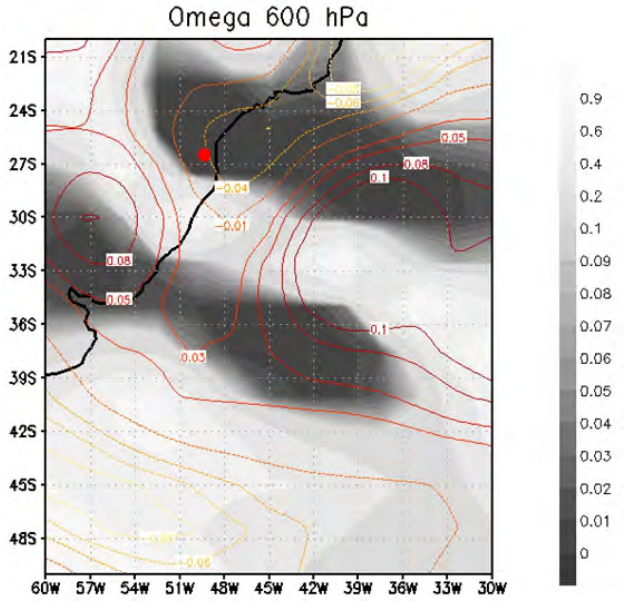


(a)

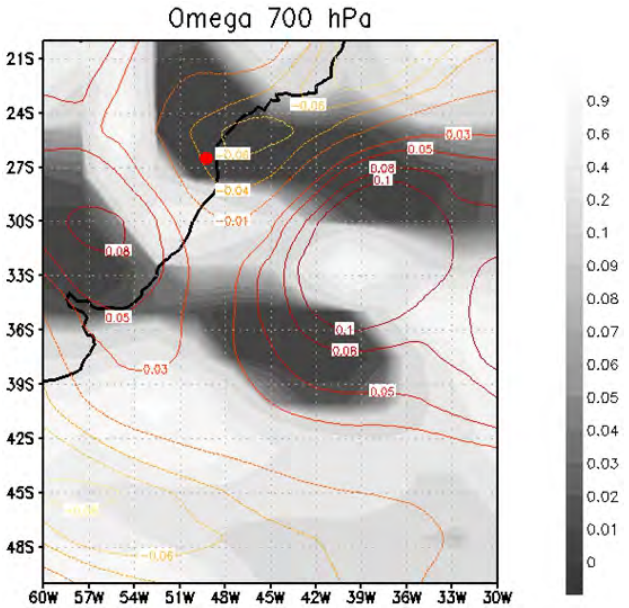


(b)

Figura 5.5 - Representação em p-valores da influência das variáveis climatológicas no evento de precipitação extrema em Santa Catarina - 2008

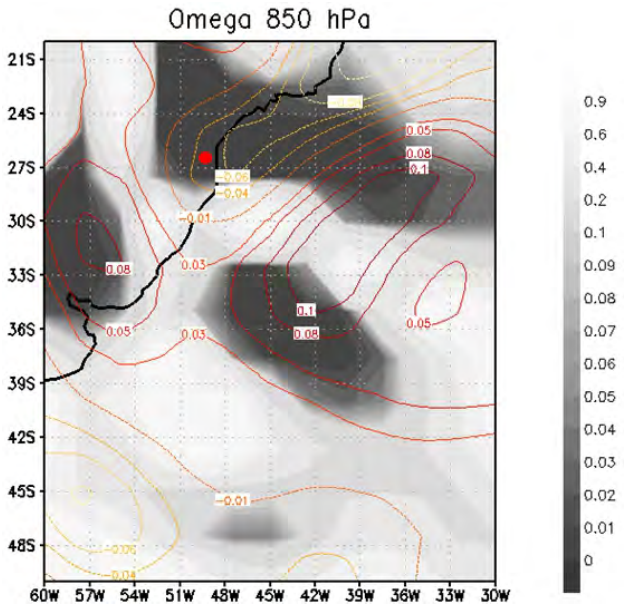


(a)

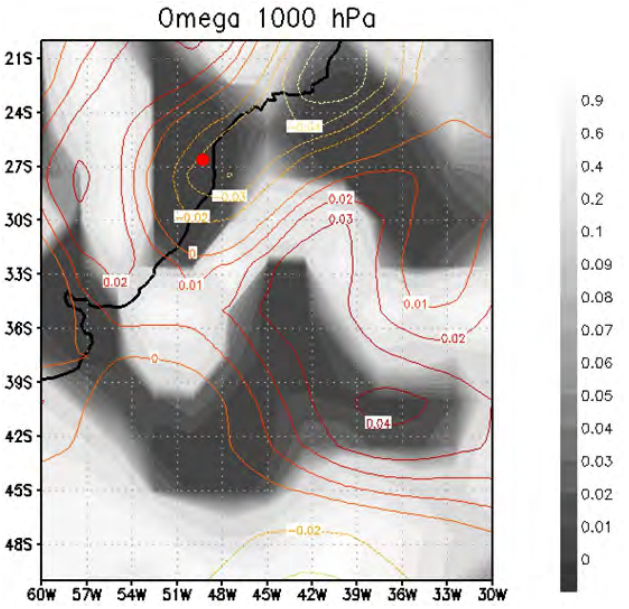


(b)

Figura 5.6 - Representação em p-valores da influência das variáveis climatológicas no evento de precipitação extrema em Santa Catarina - 2008



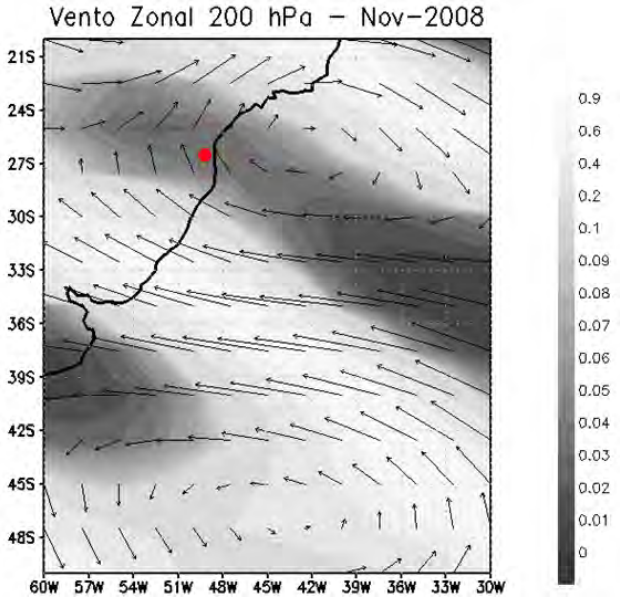
(a)



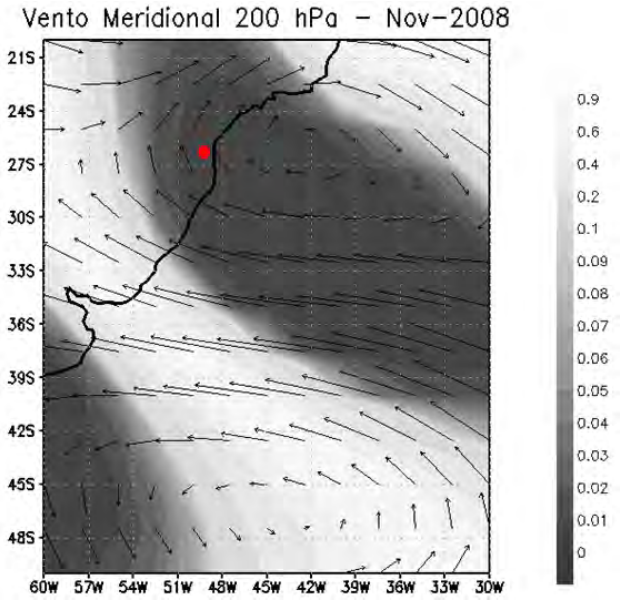
(b)

Figura 5.7 - Representação em p-valores da influência das variáveis climatológicas no evento de precipitação extrema em Santa Catarina - 2008

Esta análise é corroborada pelos resultados das Figuras 5.8, 5.9, 5.10 e 5.11. Nestas figuras as setas representam a resultante do vento zonal e meridional medido também de 22 a 26 novembro 2008. Segundo Dias (2009), Severo et al. (2008) a localização de um anticiclone de bloqueio no oceano Atlântico em altitude entre 4000 m e 5000 m (com ventos que giram no sentido anti-horário no Hemisfério Sul, como esquematizado na Figura 5.10) determinou a ocorrência de ventos de leste sobre boa parte da costa da Região Sul. Esses ventos, devido à orientação Norte-Sul da costa, incidiram mais diretamente sobre o litoral de SC, transportando, portanto, a umidade típica do oceano para o continente. As características topográficas da região atuaram como um mecanismo adicional que aumentou as quantidades de chuva durante o levantamento do ar que ascendia pelas escarpas dos morros e serras causando o esfriamento e a condensação do ar. Como consequência disso, chuvas de fraca ou moderada intensidade atingiram continuamente a região litorânea de SC. Severo et al. (2008) concluiram que as chuvas de novembro estiveram associadas a um sistema atmosférico relativamente raso, com uma extensão vertical de alguns quilômetros e com o mecanismo transportador da umidade localizado rente à superfície e sobre o Oceano Atlântico, como pode ser observado na Figura 5.12.

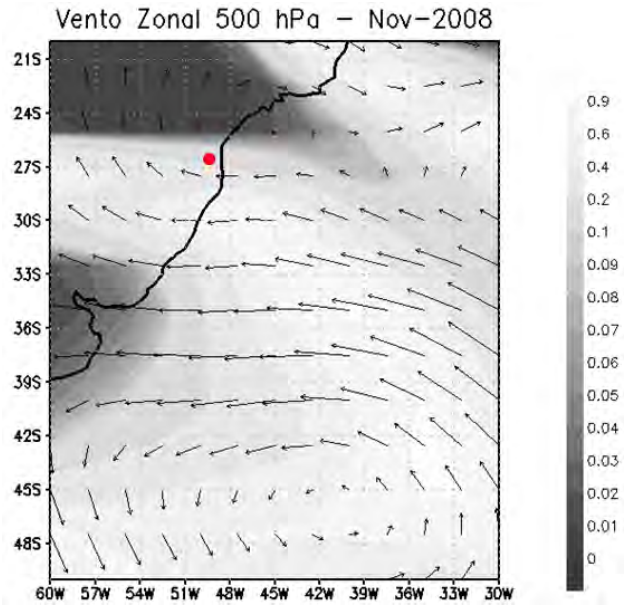


(a)

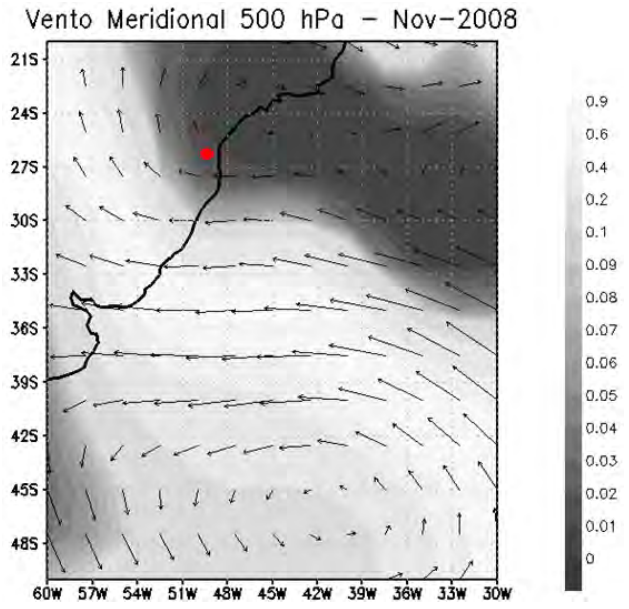


(b)

Figura 5.8 - Representação em p-valores da influência das variáveis climatológicas no evento de precipitação extrema em Santa Catarina - 2008; média das anomalias dos ventos na pêntada de 22 a 26 de novembro sobrepostas.

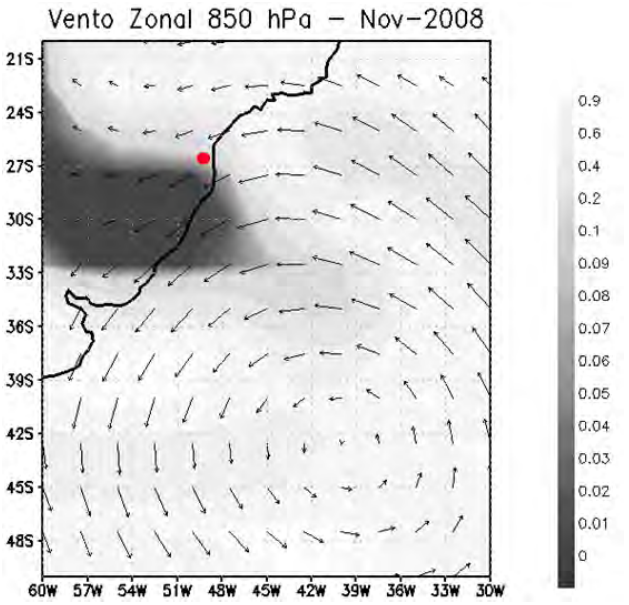


(a)

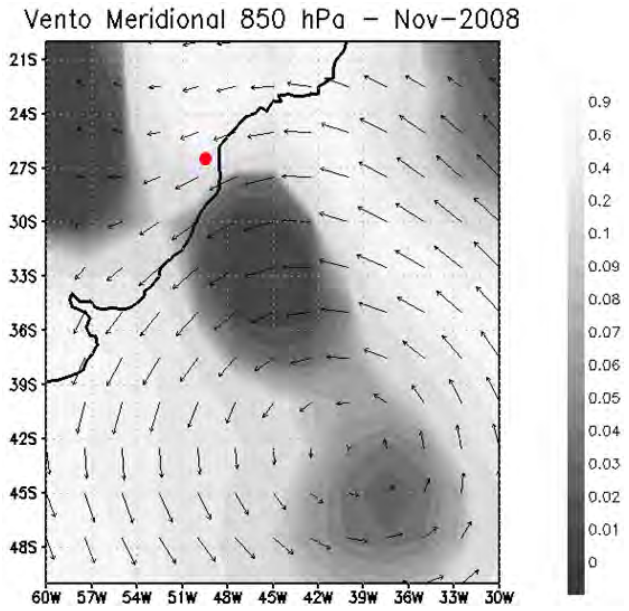


(b)

Figura 5.9 - Representação em p-valores da influência das variáveis climatológicas no evento de precipitação extrema em Santa Catarina - 2008; média das anomalias dos ventos na pêntada de 22 a 26 de novembro sobrepostas.

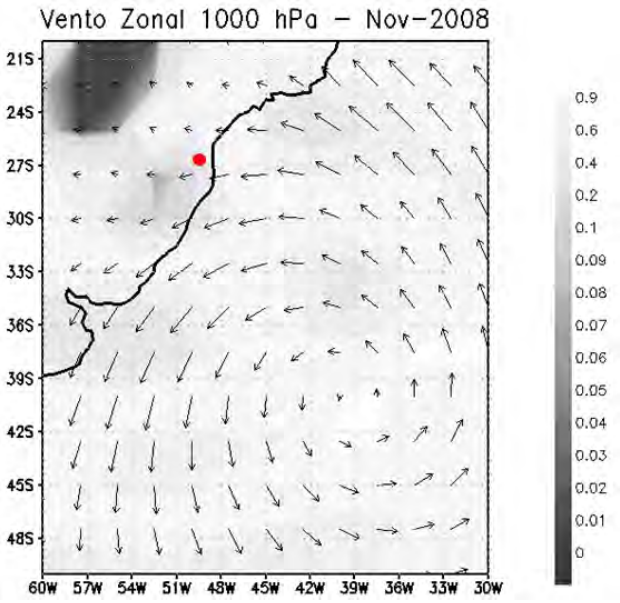


(a)

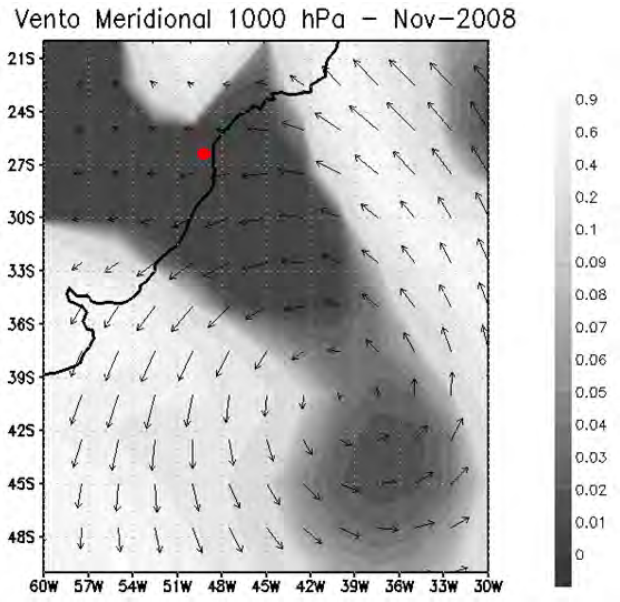


(b)

Figura 5.10 - Representação em p-valores da influência das variáveis climatológicas no evento de precipitação extrema em Santa Catarina - 2008; média das anomalias dos ventos na pêntada de 22 a 26 de novembro sobrepostas.

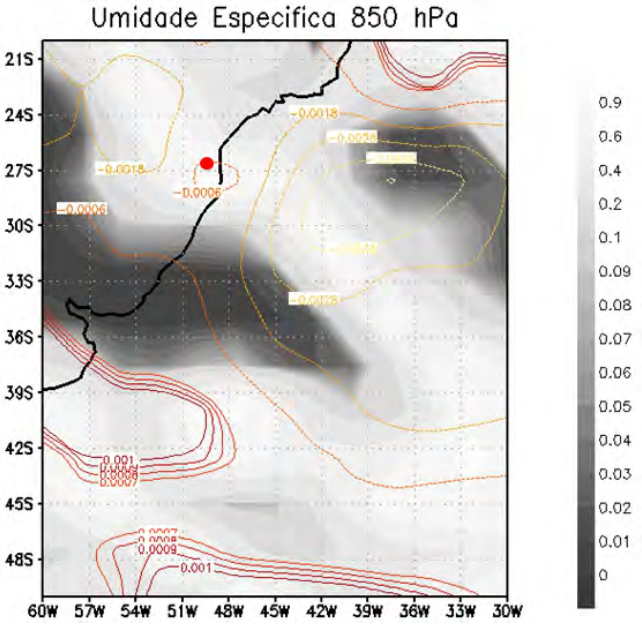


(a)

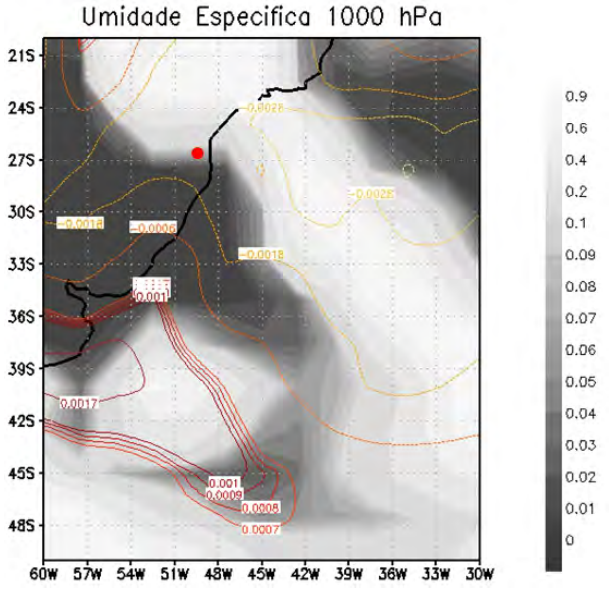


(b)

Figura 5.11 - Representação em p-valores da influência das variáveis climatológicas no evento de precipitação extrema em Santa Catarina - 2008; média das anomalias dos ventos na pêntada de 22 a 26 de novembro sobrepostas.



(a)



(b)

Figura 5.12 - Representação em p-valores da influência das variáveis climatológicas na cheia de precipitação extrema em Santa Catarina - 2008; média das anomalias dos ventos na pêntada de 22 a 26 de novembro sobrepostas.

5.2 Árvores de decisão

Para determinar os atributos a serem utilizados no classificador em árvores de decisão, novamente foram usadas as variáveis mais significativas decorrentes da análise em p-valores, ou seja, variáveis com menor p-valor. Foram feitos diversas análise

variando os conjuntos de treinamento e teste. Novamente, a classificação foi baseada na série temporal de precipitação (Figura 5.2) aonde se observa que em meados de 1999 houve uma cheia significativa. Para determinar os arquivos de treinamento e teste, foi excluído o ano de 1999 do conjunto de treinamento em duas análises, e em uma terceira o mesmo foi inserido para ver como se comporta o classificador quando existe um episódio de enchente mais fraco. Desta forma dividimos as análises em 3 partes:

- Análise 1:
  Conjunto de treinamento = 1999 a 2006 (2/3 do banco de dados); conjunto de teste = 2007 a 2010 (1/3 do banco de dados). Os resultados estão na Tabela 5.1;

- Análise 2:
  Conjunto de treinamento = 2000 a 2007; conjunto de teste = 1999, 2008 a 2010. Os resultados estão na Tabela 5.2;

- Análise 3:
  Conjunto de treinamento = 2000 a 2006; conjunto de teste = 1999, 2007 a 2010. Os resultados estão na Tabela 5.3.

Em cada análise, foram testados 4 tipos de classificação observando-se a série de precipitação. Os casos testados foram:

- class=mediana:
  Classificação de acordo com a mediana. Se valor da precipitação em anomalia na pêntada for menor que a mediana, assume-se que chove pouco, senão, chove muito;

- class=4:
  Classificação de acordo com o valor 4 da série. Se valor da precipitação em anomalia na pêntada for menor que 4, assume-se que chove pouco, senão, chove muito;

- class=8:
  Classificação de acordo com o valor 8 da série. Se valor da precipitação em anomalia na pêntada for menor que 8, assume-se que chove pouco, senão,

chove muito;

- class=12:
  Classificação de acordo com o valor 12 da série. Se valor da precipitação em anomalia na pêntada for menor que 12, assume-se que chove pouco, senão, chove muito;

Para o treinamento do algoritmo J4.8, foram utilizados os parâmetros default C=0,25 (fator de confiança) e M=8 (número de instâncias por folha). Nas Tabelas 5.1 a 5.3 a primeira coluna identifica as variáveis testadas que compõem os seguintes casos:

- Caso1
  Utilizadas as 50 variáveis com menores p-valores;
- Caso2
  Utilizadas as variáveis com p-valores menores que 0,001, ou seja as variáveis com uma probabilidade menor de 0,1% de serem falso-positivas, totalizando 179 variáveis;
- Caso 3
  De todas as variáveis climatológicas analisadas, não em coordenadas (por exemplo omega 850 hPa, Temperatura do ar, Vento Zonal), foram selecionadas as 10 com menor p-valor, no máximo 10 em coordenadas, e aquelas com p-valor $<$ 0,001 totalizando 94 variáveis;
- Caso 4
  De todas as variáveis climatológicas analisadas, não em coordenadas (por exemplo omega 850 hPa, Temperatura do ar, Vento Zonal), foram selecionadas as 10 com menor p-valor, no máximo 5 em coordenadas, totalizando 50 variáveis.

As Tabelas 5.1 a 5.3 apresentam a porcentagem de acertos dos casos de testes citados acima considerando-se tres conjuntos de testes: “Geral” - utiliza todo o arquivo de teste (primeira linha de cada caso), “Chuva forte” - analisa os eventos (pêntadas) que tiveram um valor de precipitação em anomalia maior que 20 em cada caso (6 pêntadas), e “Chuva extrema” - analisa os eventos que tiveram um valor de precipitação em anomalia maior que 30 (2 pêntadas), em cada caso.

Tabela 5.1 - Análise 1: treinamento = 1999 a 2006, teste = 2007 a 2010. Porcentagem de acerto nos testes executados.

Casos	**Testes**	**class=mediana**	**class = 4**	**class = 8**	**class = 12**
1	Geral	64,75%	78,77%	87,33%	94,52%
	Chuva forte	83,33%	33,33%	33,33%	0,00%
	Chuva exrema	50,00%	50,00%	50,00%	0,00%
2	Geral	60,07%	72,95%	88,36%	94,52%
	Chuva forte	100%	83,33%	50,00%	0,00%
	Chuva exrema	100%	100%	50,00%	0,00%
3	Geral	64,38%	75,00%	88,70%	94,52%
	Chuva forte	83,33%	50,00%	33,33%	0,00%
	Chuva exrema	100%	50,00%	50,00%	0,00%
4	Geral	65,41%	78,77%	88,70%	94,52%
	Chuva forte	100,00%	50,00%	33,33%	0,00%
	Chuva exrema	100,00%	50,00%	50,00%	0,00%

Tabela 5.2 - Análise 2: treinamento = 2000 a 2007, teste = 1999, 2008 a 2010. Porcentagem de acerto nos testes executados.

Casos	**Testes**	**class=mediana**	**class = 4**	**class = 8**	**class = 12**
1	Geral	63,36%	70,40%	89,38%	95,20%
	Chuva forte	83,33%	33,33%	33,33%	0,00%
	Chuva exrema	100,00%	50,00%	50,00%	0,00%
2	Geral	60,62%	73,97%	88,70%	94,52%
	Chuva forte	66,66%	33,33%	33,33%	0,00%
	Chuva exrema	50,00%	0,00%	0,00%	0,00%
3	Geral	63,36%	74,66%	87,33%	93,83%
	Chuva forte	83,33%	33,33%	66,66%	0,00%
	Chuva exrema	50,00%	0,00%	50,00%	0,00%
4	Geral	64,04%	81,51%	89,38%	93,49%
	Chuva forte	83,33%	16,66%	33,33%	0,00%
	Chuva exrema	100%	0,00%	50,00%	0,00%

De uma maneira geral, pode-se concluir que o melhor classificador para determinar o evento extremo de enchente encontra-se em class=mediana, pois quanto mais se restringe a classificação, ou seja, quanto mais restritos os episódios de chuva, poucos eventos são analisados, e com isto o classificador fica incapacitado de "acertar". Com poucos eventos de chuva, não existe uma frequencia dentro de uma das classes logo, o limiar não pode ser facilmente determinado.

Tabela 5.3 - Análise 3: treinamento = 2000 a 2006, teste = 1999, 2007 a 2010. Porcentagem de acerto nos testes executados.

Casos	**Testes**	**class=mediana**	**class = 4**	**class = 8**	**class = 12**
1	Geral	64,93%	78,63%	89,59%	95,07%
	Chuva forte	100,00%	33,33%	16,66%	33,33%
	Chuva exrema	100,00%	50,00%	0,00%	0,00%
2	Geral	62,74%	76,99%	89,59%	93,42%
	Chuva forte	50,00%	50,00%	16,66%	0,00%
	Chuva exrema	50,00%	50,00%	0,00%	0,00%
3	Geral	58,36%	78,36%	89,59%	95,34%
	Chuva forte	50,00%	33,33%	16,66%	0,00%
	Chuva exrema	50,00%	50,00%	0,00%	0,00%
4	Geral	64,93%	79,18%	89,59%	95,07%
	Chuva forte	100,00%	33,33%	16,66%	33,33%
	Chuva exrema	100,00%	50,00%	0,00%	0,00%

Observa-se que a Análise 1 (Tabela 5.1) tem uma porcentagem de acertos maior do que as outras análises pois contém o ano de 1999 no arquivo de treinamento, ou seja, com uma enchente em menor escala no conjunto de treinamento, o classificador tem uma maior porcentagem de acerto nos eventos extremos.

Os resultados apresentados na Tabela 5.3 são os mais satisfatórios pois compreendem uma porcentagem de 100% de acerto nos episódios de chuva forte quando se utiliza um conjunto pequeno de atributos (casos 1 e 4). Isto significa que pode-se definir um classificador de chuva extrema com apenas 50 variáveis climatológicas. Embora a Tabela 5.2 também tenha apontado quase 100% de acertos nos Casos 1 e 4, observa-se que o ano de 2007 consta no arquivo de treinamento, o que não ocorre na Análise 2. Se observarmos a série temporal da precipitação, podemos concluir que o ano de 2007 foi muito mais chuvoso do que o de 2006 na área analisada, e por este fato o classificador identificou melhor as chuvas extremas.

A figura 5.13 ilustra a árvore de decisão gerada para a Análise 3, casos 1 e 4. Observa-se que a variável com maior ganho de informação apontada pelo classifivador foi Omega 500 hPa com coordenadas longitudinais de 50 Oeste, e latitudinais de 25 Sul. Estas coordenadas estão praticamente na região afetada pela enchente. Se observarmos nas outras folhas da árvore, as variáveis mais relevantes continuam sendo, em sua maioria, Omega em outras altitudes e todas próximas à região de estudo. Embora se saiba que a variável omega, que é responsável pelo tranporte de

vento vertical, tem fundamental importancia nos eventos de chuva, e observando os resultados das imagens geradas pelos p-valores, obtem-se a amplitude da influencia desta variável em um período de chuva.

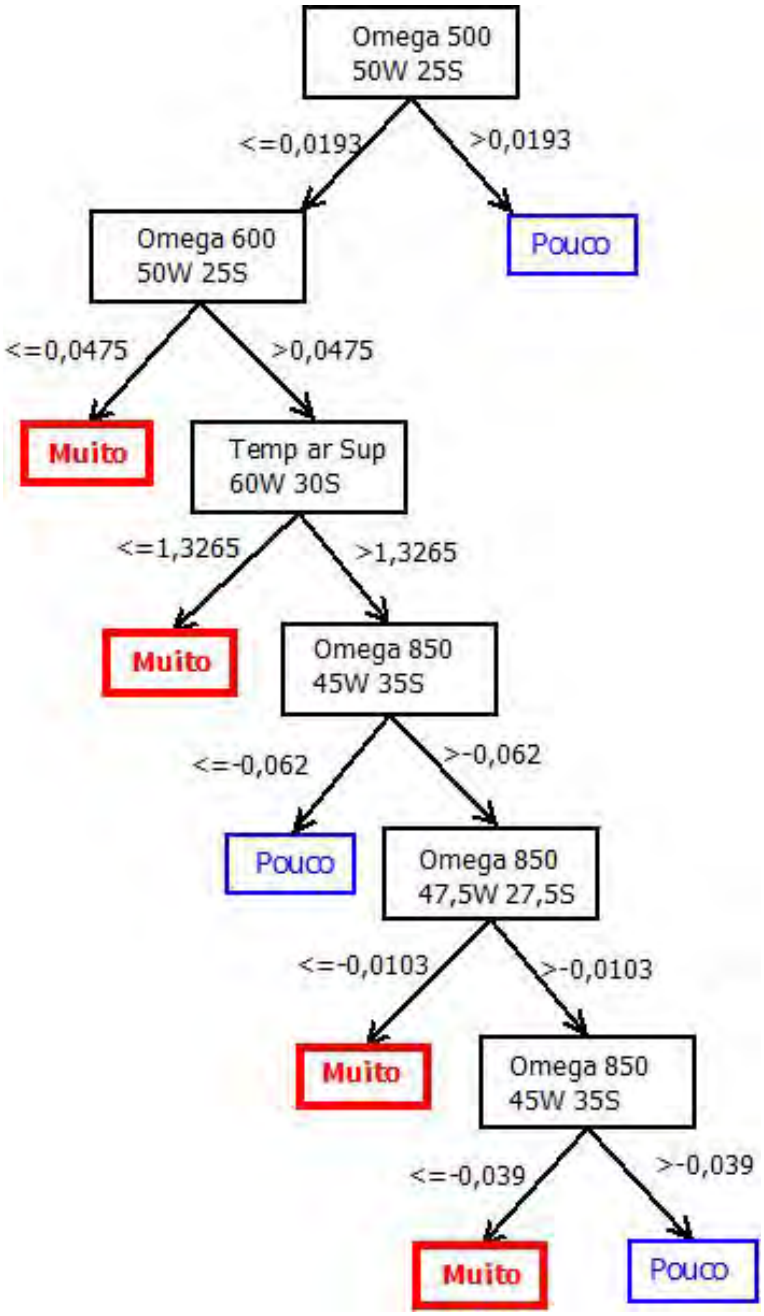


Figura 5.13 - Árvore de decisão gerada com entropia de Shannon para os casos 1 e 4. Arquivo de treinamento compreende os anos de 2000 a 2006, e arquivo de teste, os anos de 1999, 2007 a 2010, classificação pela mediana da precipitação.

6 CONCLUSÃO

O volume de dados gerados em quase todas as áreas da ciência e das engenharias, particularmente em meteorologia e ciência do clima, está aumentando cada vez mais. Um dos desafios gerados por este "dilúvio de dados" é o desenvolvimento de estratégias eficazes de descoberta de conhecimento.

O clima é um sistema não linear com dependências espaciais e temporais complexas. De modo similar, os sistemas biológicos são paradigmas de complexidade fato este que, frequentemente, tornam a análise de dados biológicos um desafio. Felizmente, a bioestatística dispõe de métodos estatísticos e ferramentas computacionais robustas e eficazes para análise de dados. Neste trabalho, mostrou-se que técnicas estatísticas utilizadas rotineiramente na análise de dados genômicos podem ser empregadas com sucesso em análise climática. Deste modo, foram investigados dois eventos climáticos extremos ocorridos no Brasil na última década – as secas de 2005 e 2010 na Amazônia e a precipitação extrema de 2008 em SC – com o objetivo de identificar as suas possíveis causas e a viabilidade de identificação de fenômenos similares.

Por um lado, os resultados obtidos na primeira aplicação indicaram que a temperatura da superfície do mar na região do Atlântico Tropical Norte teve um impacto direto no regime de chuva na região afetada. Por outro lado, o fenômeno El Niño/La Niña não foi selecionado uma única vez nas análises realizadas. Estes resultados, de um modo geral, são corroborados pela literatura especializada. Ressalte-se, no entanto, que pela primeira vez um volume de dados tão grande foi utilizado para analisar o fenômeno da seca na Amazônia. A árvore de decisão gerada pela metodologia empregada identificou os períodos de estiagem que ocorreram nos dois semestres de seca extrema.

A aplicação das técnicas em precipitação extrema apontou a magnitude de algumas variáveis climatológicas, dentro da área analisada, que podem estar correlacionadas com o aumento de precipitação durante o período analisado. Esta precipitação extrema foi muito menos analisada na literatura do que as secas intensas de 2005 e 2010 na região da Amazônia. Os mapas de p-valor identificaram a relevância das variáveis *omega* e das componentes do *vento zonal e meridional* em todos os níveis durante a análise. A árvore de decisão identificou corretamente os períodos de acentuada precipitação nas pêntadas onde foram identificadas chuvas abundantes.

Em um sentido mais amplo, os resultados desta tese mostram que a técnica estatística de p-valor é útil tanto na biologia molecular quanto na área ambiental.

Entretanto os mapas de p-valor tornam mais fáceis a identificação da dependência espacial da variável meteorológica com o evento climatológico associado. Além disto, as árvores de decisão geradas pela metodologia empregada podem ser uma ferramenta poderosa para detectar prováveis eventos extremos. Tais abordagens não foram encontradas até então na literatura para este tipo de análise de eventos extremos climáticos.

A principal mensagem transmitida é que devido a complexidade dos bancos de dados climáticos disponíveis atualmente, métodos de mineração de dados e técnicas de visualização tornam-se críticas na habilidade de se descobrir padrões climáticos recorrentes e persistentes, e conexões escondidas entre regiões geográficas distantes. Naturalmente, não se advoga aqui ingenuamente o uso irrestrito e sem a supervisão de especialistas de ferramentas computacionais para estabelecer relações casuais entre variáveis ambientais. Em vez disso, argumenta-se que uma abordagem holística, com o emprego de técnicas avançadas de mineração de dados, permite aos pesquisadores explorarem melhor os grandes bancos de dados climáticos disponíveis atualmente.

Como sugestões de trabalhos futuros, propõe-se a completa transposição e adaptação do pacote BRB-ArrayTools para a área ambiental com a modificação do jargão típico da biologia molecular e a introdução de novas funcionalidades porventura necessárias para as suas novas aplicações. Além disso, pretende-se automatizar todo o processo de descoberta de conhecimento em banco de dados tornando este trabalho mais operacional. Todos os passos executados neste trabalho foram realizados de acordo com as etapas da descoberta de conhecimento. Cada etapa foi minuciosamente executada e a "receita" foi devidamente registrada com o objetivo de gerar um aplicativo operacional para análise de outros eventos extremos. Este aplicativo disponibilizará opções mais abrangentes, a saber: área geográfica de análise, período, tipos de variáveis, bem como dar uma opção maior na escolha da variável classificatória que rege a análise. Existe atualmente um grande número de variáveis climatológicas disponíveis e que não foram utilizadas aqui, mas que devem ser testadas em sua forma mais ampla, independente do resultado. A não relevância de algumas variáveis é também importante na análise de eventos.

Outra sugestão de trabalho futuro na área climatológica seria fazer um estudo sistemático de todas as bacias hidrográficas brasileiras, identificando quais são as variáveis ambientais mais relevantes para sua hidrologia. Ainda na área de mineração de dados, outra sugestão é a avaliação de outras metodologias de redução de dados

para comparar com a tecnologia estatística de *p-valor* abordada neste trabalho. O objetivo é comparar diferentes metodologias para a análise de eventos meteorológicos extremos. Por exemplo, uma das tecnologias de redução de dados a ser analisada é a *"Teoria dos Conjuntos Aproximativos"*, anteriormente utilizada por Anochi (2010) com o objetivo de selecionar variáveis meteorológicas para os modelos de redes neurais artificiais de previsão climática sazonal de precipitação. Pretende-se também investigar o uso de lógica fuzzy, mais precisamente, utilizar um algoritmo fuzzy de árvore de decisão baseado no C4.5, chamado "fuzzyDT". Este algoritmo foi desenvolvido por Cintra et al. (2012) devido ao fato de que árvores de decisão fuzzy combinam poderosos modelos de árvores de decisão com a interpretação e habilidade do processamento de incertezas e imprecisão de sistemas fuzzy.

REFERÊNCIAS BIBLIOGRÁFICAS

AMARATUNGA, D.; CABRERA, J. **Exploration and analysis of DNA microarray and protein array data**. New Jersey: Wiley Interscience, 2004. 246 p. 14, 15, 16, 18

ANOCHI, J. **Modelos baseados em redes neurais para o estudo de padrões climáticos sazonais a partir de dados tratados com a teoria dos conjuntos aproximativos.** 187 p. Dissertação (Mestrado) — Instituto Nacional de Pesquisas Espaciais, São José dos Campos, Brasil, 2010. 33, 79

BARBIERI, D. W.; FERREIRA, C. C.; SAITO, S. M.; SAUSEN, T. M.; HANSEN, M. A. F. Relação entre os desastres naturais e as anomalias de precipitação para a região sul do brasil. In: EPIPHANIO, J. C. N.; GALVÃO, L. S. (Ed.). **Anais...** São José dos Campos: Instituto Nacional de Pesquisas Espaciais (INPE), 2009. p. 3527–3534. ISBN 978-85-17-00044-7. 4

BERTHOLD, M. R.; HANS, D. J. **Intelligent data analysis**. New York: Springer-Verlag, 1999. 530 p. 23

BERZAL, F.; CUBETO, J. C.; MARíN, N.; SáNCHEZ, D. Building multi-way decision trees with numerical attributes. **Information Sciences,**, v. 165, p. 73–90, 2004. ISSN 0020-0255. 24

BREIMAN, L.; FRIEDMAN, J. H.; OLSHEN, R. A.; STOLE, C. J. **Classification and regression trees**. Monterey, Ca.: Wadsworth and Brooks, 1984. 358 p. 14

CASE, M. **Climate Change Impacts in the Amazon: Review of scientific literature**. WWF Climate Change Programme, 2006. 13 p. Disponível em: <http://assets.panda.org/downloads/amazon_cc_impacts_lit_review_final_2.pdf>. 5

CHEN, M. S.; HAN, J.; YU, P. S. Data mining: An overview from a database perspective. **IEEE Trans. on Knowl. and Data Eng.**, IEEE Educational Activities Department, Piscataway, NJ, USA, v. 8, n. 6, p. 866–883, 1996. ISSN 1041-4347. 10

CINTRA, M. E.; MONARD, M. C.; CAMARGO, H. A. Fuzzydt - a fuzzy decision tree algotithm based on c4.5. In: II CBSF - SEGUNDO CONGRESSO BRASILEIRO DE SISTEMAS FUZZY. **Anais...** Rio Grande do Norte - Natal, 2012. p. 199–211. 79

COELHO, C. A. S.; CAVALCANTI, I. A. F.; COSTA, S. M. S.; FREITAS, S. R.; ITO, E. R.; LUZ, G.; SANTOS, A. F.; NOBRE, C. A.; MARENGO, J. A.; PEZZA, A. B. Climate diagnostics of three major drought events in the amazon and illustrations of their seasonal precipitation predictions. **Meteorol. Appl.**, v. 19, p. 237–255, 2012. 33, 43, 51

COOK, B.; ZENG, N.; YOON, J.-H. Climatic and ecological future of the amazon: likelihood and causes of change, earth syst. **Dynam. Discuss.**, v. 1, p. 63–101, 2010. 45, 46

COUTINHO, D. P. **Teoria da Informação**: conceito de entropia e sua aplicação. Instituto Superior de Engenharia de Lisboa, 2004. Disponível em: <http://www.deetc.isel.ipl.pt/analisedesinai/ccd/docs/CCD0304FolhasApoio_parte1.pdf>. 22

COX, P. M.; HARRIS, P. P.; HUNTINGFORD, C.; BETTS, R. A. e. a. Increasing risk of amazonian drought due to decreasing aerosol pollution. **Nature**, v. 453, p. 212—215, 2008. 40, 41, 46

CPTEC/INPE. **Glossário**. 2006. Disponível em: <http://www7.cptec.inpe.br/glossario/>. Acesso em: 02 fev. 2009. 89

DANTAS, O. D. **Uma técnica automática baseada em morfologia matemática para medida de sinal em imagens de cDNA**. Dissertação de (Mestrado em Ciência da Computação) — Universidade de São Paulo, São Paulo, 2004. 26

DIAS, M. A. F. S. **As chuvas de novembro de 2008 em Santa Catarina: um estudo de caso visando à melhoria do monitoramento e da previsão de eventos extremos**. São José dos Campos: INPE, 2009. 67 p. Disponível em: <http://urlib.net/8JMKD3MGP7W/36C4PG2>. Acesso em: 15 mar. 2010. 6, 7, 65

DOTY, B. **Grid Analysis and Display System (GrADS) - Center for Ocean - Land-Atmosphere Studies (COLA)**. 2009. Disponível em: <http://grads.iges.org/grads/head.html>. Acesso em: 10 mar. 2012. 35

EISEN, M. B.; SPELLMAN, P. T.; BROWN, P. O.; BOTSTEIN, D. Cluster analysis and display of genome-wide expression patterns. **PNAS**, v. 95, p. 14863–14868, 1998. 19

FAYYAD, U.; PIATESKY-SHAPIRO, G.; SMYTH, P.; UTHURUSAMY, R. **Advances in knowledge discovery and data mining**. California: The MIT Press, 1996. 560 p. 9

GAGNE, D. I.; MCGOVERN, A.; BROTZGE, J. Classification of convective areas using decision trees. **J. Atmos. Oceanoc Technol.**, v. 26, p. 1341–1353, 2009. 12

HAN, J.; KAMNER, M. **Data mining: concepts and techniques**. San Francisco: Morgan Kaufmann Publishers, 2001. 550 p. 20, 21

HUI, Y.; RONGQUN, Z.; XIANWEN, L. Extracting wetland using decision tree classification. In: 8TH WSEAS INTERNATIONAL CONFERENCE ON APPLIED COMPUTER AND APPLIED COMPUTATIONAL SCIENCE. **Proceedings...** 2004. p. 240–245. ISBN 978-960-474-075-8. Disponível em: <http://www.wseas.us/e-library/conferences/2009/hangzhou/ACACOS/ACACOS39.pdf>. 12, 13

IPCC. **Cambio climático 2007: Informe de síntesis**. Ginebra, Suiza: Grupo Intergubernamental de Expertos sobre el Cambio Climático [Equipo de redacción principal: Pachauri, R.K. y Reisinger, A. (directores de la publicación)], 2007. 104 p. 1

JEANMOUGIN, M.; REYNIES, A. de; MARISA, L.; PACCARD, C.; NUEL, G.; GUEDJ, M. Should we abandon the t-test in the analysis of gene expression microarray data: a comparison of variance modeling strategies. **PLoS ONE**, v. 5, n. 9, p. 6684–6692, 2010. 15

KALNAY, E.; KANAMITSU, M.; KISTLER, R.; COLLINS, W.; DEAVEN, D. e. a. The ncep/ncar 40-year reanalysis project. **Bull. Amer. Meteor. Soc.**, v. 77, p. 437–470, 1996. 34

KRUTOVSKII, K. V.; NEALE, D. B. **Forest genomics for conserving adaptive genetic diversity**. Rome: Food and Agriculture Organization of the United Nations - FAO, 2001. Disponível em: <ftp://ftp.fao.org/docrep/fao/004/x6884e/x6884e00.pdf>. Acesso em: 22 jul. 2006. 26, 27

LABORATORY, N. E. S. R. **Multivariate ENSO index (MEI)**. U.S. Department of Commerce: National Oceanic and Atmospheric Administration, 2007. Disponível em: <http://www.cdc.noaa.gov/people/klaus.wolter/MEI/>. Acesso em: 19 jul. 2010. 35

LEWIS, S. L.; BRANDO, P. M.; PHILLIPS, O. L.; HEIJDEN, G. M. F. V. D.; D., N. The 2010 amazon drought. **Science**, v. 331, p. 554, 2011. 6, 39, 40, 41

LIMA, C.; ASSIS, F.; SOUZA, C. Decision tree based on shannon, rényi and tsallis entropies for intrusion tolerant systems. In: INTERNET MONITORING AND PROTECTION (ICIMP), 2010 FIFTH INTERNATIONAL CONFERENCE ON. **Proceeding...** 2010. p. 117 –122. Disponível em: <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5476870>. 25

LINDEN, R. **Técnicas de Agrupamento**. Revista de Sistemas de Informação da FSMA, 2009. Disponível em: <http://www.fsma.edu.br/si/edicao4/FSMA_SI_2009_2_Tutorial.pdf>. Acesso em: 27 set. 2012. 18

LORENZ, E.; PROJECT, M. I. of T. S. F. **Empirical orthogonal functions and statistical weather prediction**. Massachusetts Institute of Technology, Dept. of Meteorology, 1956. (Scientific report). Disponível em: <http://books.google.com.br/books?id=q7sJAQAAIAAJ>. 33

MALHI, Y.; ROBERTS, J. T.; BETTS, R. A.; KILLEEN, T. J. e. a. Climate change, deforestation, and the fate of the amazon. **Science**, v. 319, p. 169–172, 2008. 40, 41

MANTUA, N. J.; HARE, S. R. The pacific decadal oscillation. **Journal of Oceanography**, v. 58, p. 35–44, 2002. 91

MARCELINO, I. P. V. O.; NASCIMENTO, E. L.; FERREIRA, N. J. **Tornadoes in santa catarina state (southern brazil): event documentation, meteorological analysis and vulnerability assessment**. 2005. Disponível em: <http://urlib.net/sid.inpe.br/iris@1912/2006/01.13.11.42>. Acesso em: 19 jul. 2012. 6

MARENGO, J. A. Impactos de extremos relacionados com o tempo e o clima - impactos sociais e econômicos. In: RAIGOZA, D. (Ed.). **Anais...** Cachoeira Paulista, 2009. v. 8. 4

_____. Impactos de extremos relacionados com o tempo e o clima - impactos sosiais e econômicos. In: . [S.l.]: Mudanças Climaticas - INPE, 2009. v. 8. 7

MARENGO, J. A.; NOBRE, C. A.; SHOU, S. C.; TOMASELLA, J. e. a. Riscos das mudanÇas climÁticas no brasil. In: . [S.l.]: Projeto colaborativo realizado pelo Centro de Ciência do Sistema Terrestre (CCST) do Instituto Nacional de Pesquisas

Espaciais (INPE) do Brasil e o Met Office Hadley Centre (MOHC) do Reino Unido, 2011. 6

MARENGO, J. A.; NOBRE, C. A.; TOMASELLA, J.; OYAMA, M. D.; OLIVEIRA, G. S.; OLIVEIRA, R.; CAMARGO, H.; ALVES, L. M.; BROWN, I. F. The drought of amazonia in 2005. **Journal of Climate**, v. 21, n. 3, 2008. 6, 33, 40, 41, 46

MARENGO, J. A.; NOBRE, C. A.; TOMASELLA, J.; CARDOSO, M. F.; OYAMA, M. C. Hydro-climatic and ecological behaviour of the drought of amazonia in 2005. **Phil. Trans. R. Soc. B.**, v. 363, p. 1773–1778, 2008. 40, 41, 43, 45, 46

MARENGO, J. A.; TOMASELLA, J.; ALVES, L. M.; SOARES, W. R.; RODRIGUEZ, D. A. The drought of 2010 in the context of historical droughts in the amazon region. **Geophys Res Let**, v. 38, 2011. 40, 41, 43, 45, 46

MEIRA, C. A. A. **Processo de descoberta de conhecimento em baes de dados para a análise e o alerta de doenças de culturas agrícolas e suas aplicações na ferrugem do cafeeiro**. Tese (Doutorado em Engenharia Agrícola) — Universidade Estadual de Campinas, Campinas, 2008. Acesso em: 2009. 20

NIJSSEN, B.; O'DONNELL, G.; HAMLET, A.; LETTENMAIER, D. Hydrologic sensitivity of global rivers to climate change. **Climate Change**, v. 50, p. 143–175, 2001. 5

NOBRE, C. A.; ASSAD, E. Mudança ambiental no brasil. em terra na estufa. **edição especial Scientific American Brasil**, n. 12, p. 70–75, 2005. 5

NOBRE, C. A.; SAMPAIO, G.; SALAZAR, L. Mudanças climáticas e amazônia. **Ciência e Cultura**, v. 59, n. 3, 2007. 5

QUINLAN, J. R. Induction of decision trees. **Machine Learning**, n. 1, p. 81–106, 1986. 11, 21

_____. **C4.5: programs for machine learning.** San Francisco: Morgan Kaufmann Publishers, 1993. 20, 21, 22

RéNYI, A. On mesures of entropy and information. In: 4TH BERKELEY SYMPOSIUM ON MATHEMATICAL STATISTICS AND PROBABILITY. **Proceding...** [S.l.], 1961. v. 1, p. 547–561. 25

RUIVO, H. M. **Análise integrada de dados ambientais utilizando técnicas de classificação e agrupamento de microarranjos de DNA**. Dissertação de (Mestrado em Ciência da Computação) — Instituto Nacional de Pesquisas Espaciais, São Paulo, 2008. 13

RUIVO, H. M.; RAMOS, F. M.; VELHO, H. F. C. Data mining to identify extreme meteorological events in brazil. In: **XII International Conference on Integral Methods in Science and Engineering (IMSE)**. [S.l.: s.n.], 2012. 33

RUIVO, H. M.; VELHO, H. F. C.; SAMPAIO, G.; RAMOS, F. M. Analysis of the 2005 and 2010 amazon droughts using a genome-wide knowledge extraction approach. In: **2nd Conference on Computational Interdisciplinary Science (CCIS)**. [S.l.: s.n.], 2012. 33

SATTARI, M. T.; ANLI, A. S.; APAYDIN, H.; KODAL, S. Decision trees to determine the possible drought periods in ankara. **Atmosféra**, v. 25, p. 65–83, 2012. 11

SEVERO, D. L.; CORDERO, A.; M., T.; SILVA, H. **Análise hidrometeorológica da catástrofe no Vale do Itajaí em 2008**. 2008. Disponível em: <http://www.cbmet2010.com/anais/artigos/409_76562.pdf>. Acesso em: 10 out. 2012. 65

SHANNON, C. E. A mathematical theory of communication. **Reprinter with corrections from The Bell System Technical Journal**, v. 27, p. 379–423, 1948. 22, 23

SIMON, R.; LAM, A. P. **BRB-ArrayTools - version 3.4 - User's manual**. National Cancer Institute: [s.n.], 2006. 108 p. Disponível em: <http://linus.nci.nih.gov/~brb/download.html>. Acesso em: 10 jan. 2002. 19, 28

SIMON, R. M.; KORN, E. L.; MCSHANE, L. M.; RADMACHER, M. D.; WRIGHT, G. W.; ZHAO, Y. **Statistics for biology and health**. New York: Springer-Verlag, 2003. 199 p. 15, 33

SOUTO, M. C. P.; LORENA, A. C.; DELBEM, A. C. B.; CARVALHO, A. C. P. L. F. **Técnicas de aprendizado de máquina para problemas de biologia molecular**. Universidade de São Paulo - São Carlos: [s.n.], 2004. Disponível em: <http://www.dimap.ufrn.br/~marcilio/ENIA2003/jaia2003-14-08.pdf>. 26

TOMASELLA, J.; BORMA, L. S.; MARENGO, J. A.; RODRIGUEZ, D. A. e. a. The droughts of 1996-1997 and 2004-2005 in amazon: hydrological response in the river main-stem. **Hydrological Processes**, v. 25, p. 1228–1242, 2011. 33, 40, 41, 45, 46

TSALLIS, C. Possible generalization of boltzmann-gibbs statistica. In: . [S.l.]: Journal of Statistical Physics, 1988. v. 52, p. 479–487. 25

WALLACE, J. M.; HOBBS, P. V. **Atmospheric science: an introdictory survey**. Massachusetts: Academic Press, 2006. 483 p. 89, 90

WITTEN, I. H.; FRANK, E. **Data mining: practical machine learning tools and techniques.** San Francisco: Morgan Kaufmann Publishers, 2005. 525 p. 20, 24

WITTEN, I. H.; FRANK, E. S. **Data mining: practical machine learning tools and techniques with java implementation.** California: Morgan Kaufmann Publishers, 2000. 29

YOON, J. H.; ZENG, N. An atlantic influence on amazon rainfall. **Clim. Dyn.**, v. 34, p. 249–264, 2010. 40, 41, 45, 47

YUAN, B. B. M. G. H. M. M. Geospatial data mining and knowledge discovery. **University Consortium for Geographic Information Science (UCGIS) White Paper on Emergent Research Theme**, v. 71, p. 575–580, 1998. 10

ZENG, N.; YOON, J.; MARENGO, J. A.; SUBRAMANIAM, A. e. a. Causes and impacts of the 2005 amazon drought. **Environ Res Lett**, v. 3, 2008. 40, 41, 43, 45

APÊNDICE

A seguir, uma descrição de todas as variáveis utilizadas obtidas em: *National Weather Service Glossary* (http://w1.weather.gov/glossary/), *Centro de Previsão de Tempo e Estudos Climáticos - Glossário* (http://www.cptec.inpe.br/glossario.shtml) e *Instituto Nacional de Meteorologia* (INMET) (http://www.inmet.gov.br/).

Variáveis em grade utilizadas nas duas análises:

- Temperatura da superfície do mar (TSM)
  Temperatura média da água próxima a superfície.

- Pressão ao nível do mar
  Pressão da atmosfera ao nível do mar em um dado local. Pressão é a força por unidade de área, exercida pelo peso da atmosfera sobre um ponto localizado na superfície da Terra ou acima da mesma.

- Temperatura do ar na superfície
  Temperatura da mistura de gases compreendida próximo à superfície da terra .

- Umidade Específica
  Em um sistema de ar úmido, representa a relação entre a massa de vapor de água e a massa total do sistema.

- Omega
  Termo usado para descrever a velocidade de transporte vertical do vento. Valores positivos de omega representam movimento vertical descendente e valores negativos representam movimento vertical ascendente.

- Altura Geopotencial
  Nível de pressão de um ponto a uma altura acima do nível do mar. O geopotencial em algum ponto na atmosfera é definido como o trabalho que deve ser feito contra o campo gravitacional da Terra para elevar uma massa de um quilograma do nível do mar até o ponto considerado (CPTEC/INPE, 2006).

- Vento zonal
  A componente zonal do vento é o movimento leste-oeste através da atmosfera (WALLACE; HOBBS, 2006).

- Vento meridional
  A componente meridional do vento é o movimento norte-sul através da atmosfera (WALLACE; HOBBS, 2006).

- Cobertura de nuvens
  Conjunto de partículas sólidas de água ou de gelo visíveis na atmosfera, acima da superfície da terra.

- Fluxo de calor sensível
  Fluxo de calor, da superfície da terra até a atmosfera, que não está associado com as fases de trocas de água. É a componente de balanço de energia da superfície.

- Fluxo de calor latente
  Fluxo associado com a evaporação ou condensação de vapor de água na superfície. É a componente de balanço de energia da superfície.

- Precipitação
  Processo em que vapor de água na atmosfera condensa para formar gotículas de água que caem para a Terra como chuva, granizo, neve, etc.

Séries temporais utilizadas na análise da Seca da AM:

- Indice de oscilação Sul (SOI)
  Indica o desenvolvimento e a intensidade dos fenômenoa El Niño ou La Niña no Oceano Pacífico. O SOI é calculado usando as diferenças de pressão entre Tahiti (Polinesia Francesa) e Darwin (Austrália), extraído de *http://www.cpc.ncep.noaa.gov/data/indices/*.

- Temperatura da superfície do mar - Niño
  Média da temperatura no oceano na região delimitada: 5^0Na5^0S, e 150^0Wa90^0W, extraído de *http://www.cpc.ncep.noaa.gov/data/indices/*.

- Temperatura da superfície do mar - Atlântico Norte
  Média da temperatura no oceano na região delimitada: 5^0Na20^0N, e 30^0Wa60^0W, extraído de *http://www.cpc.ncep.noaa.gov/data/indices/*.

- Temperatura da superfície do mar - Atlântico Sul
  Média da temperatura no oceano na região delimitada: 0^0a20^0S, e 30^0Wa10^0E, extraído de *http://www.cpc.ncep.noaa.gov/data/indices/*.

- Oscilação do Atlantico Norte (NAO)
  Diferença da pressão normalizada entre duas estações do Atlântico Norte: uma de pressão baixa localizada próxima de Islândia, e outra de pressão alta sobre Açores - cadeia de ilha no leste do Oceano Atlântico (http://www.ldeo.columbia.edu/NAO).

- Oscilação de décadas do Pacífico (PDO)
  Componente principal da variabilidade da temperatura mensal da superfície do mar no Pacífico Norte (MANTUA; HARE, 2002).

Séries temporais utilizadas na análise da precipitação em SC obtidas atraves das séries de reanálise da NOAA:

- Gibraltar-Islândia
  Diferença da pressão ao nível do mar entre Gibraltar e Islândia, semelhante ao índice NAO;

- Nino1 (90W-80W 0-10S)
  Média da temperatura da superfície do mar na região delimitada pelas coordenadas 90W-80W e 0-10S;

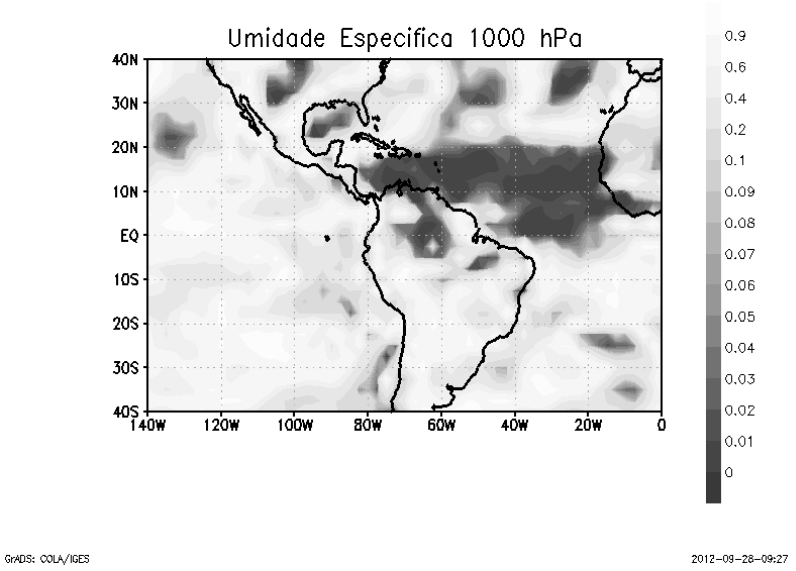
- Nino2 (150W-90W 5N-5S)
  Média da temperatura da superfície do mar na região delimitada pelas coordenadas 150W-90W e 5N-5S;

- Nino3 (170W-120W 5N-5S)
  Média da temperatura da superfície do mar na região delimitada pelas coordenadas 170W-120W e 5N-5S;

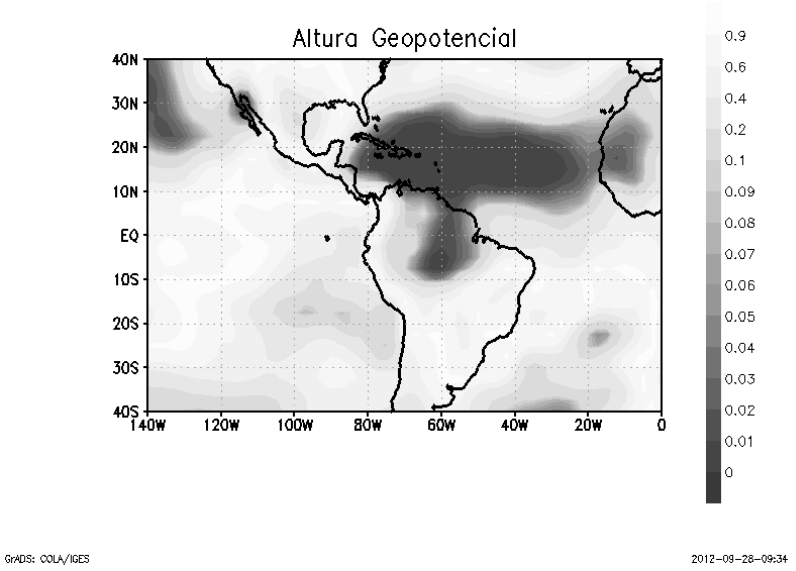
- Nino4 (160W-150W 5N-5S)
  Média da temperatura da superfície do mar na região delimitada pelas coordenadas 160W-150W e 5N-5S;

ANEXO A - RESULTADOS ADICIONAIS - SECA AMAZONAS

Os resultados apresentados das Figuras .1 a .8 referem-se a análise da seca na Amazônia. Correspondem a comparação entre seca x neutra e são apresentados em campos de p-valores.

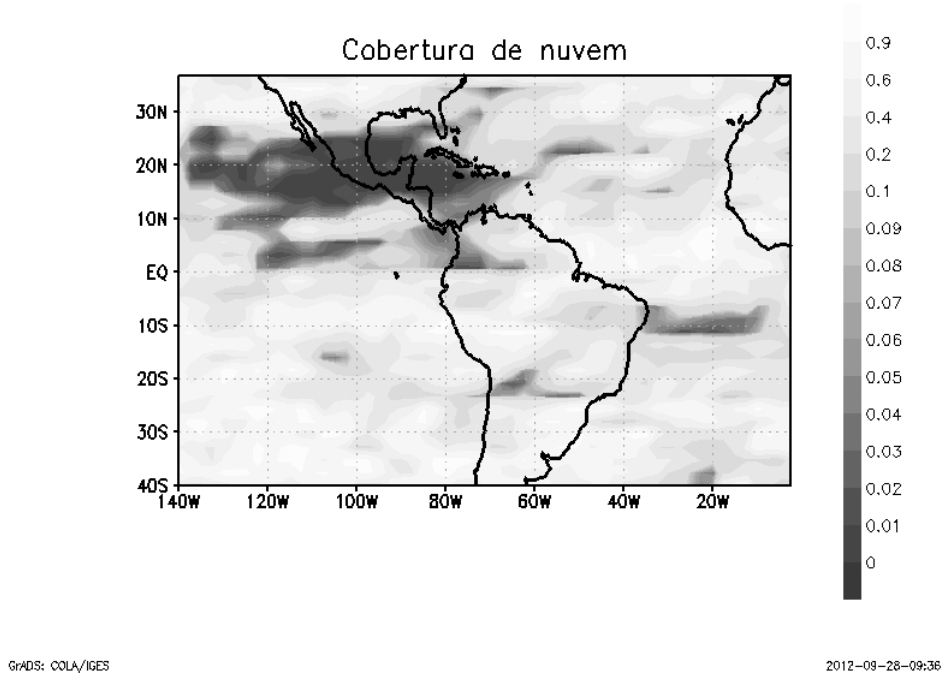


(a)

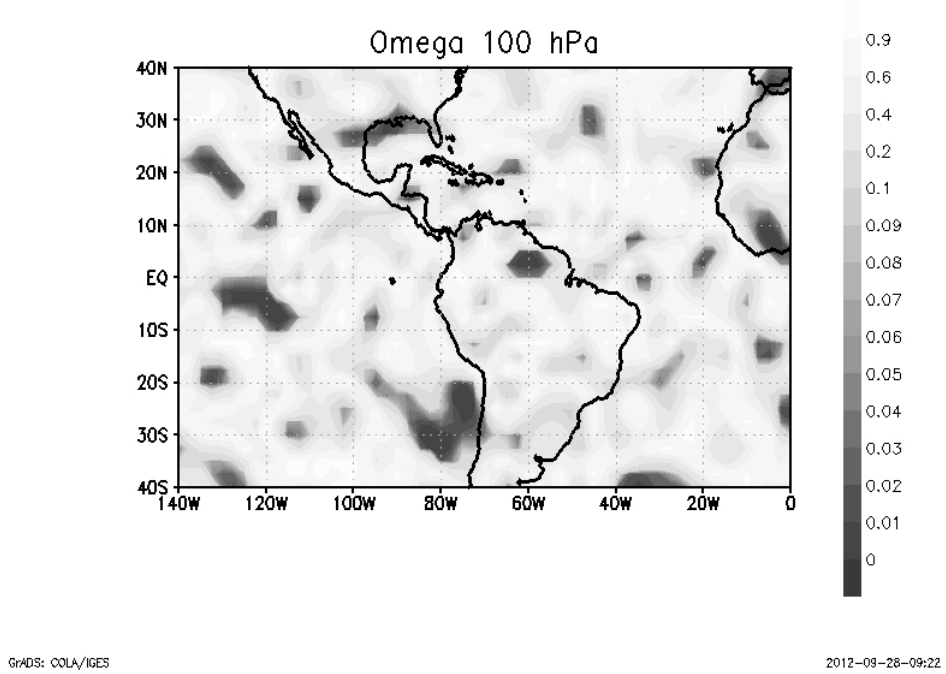


(b)

Figura .1 - Representação em p-valores da influência das variáveis climatológicas na seca da Amazonia - sec XXI.

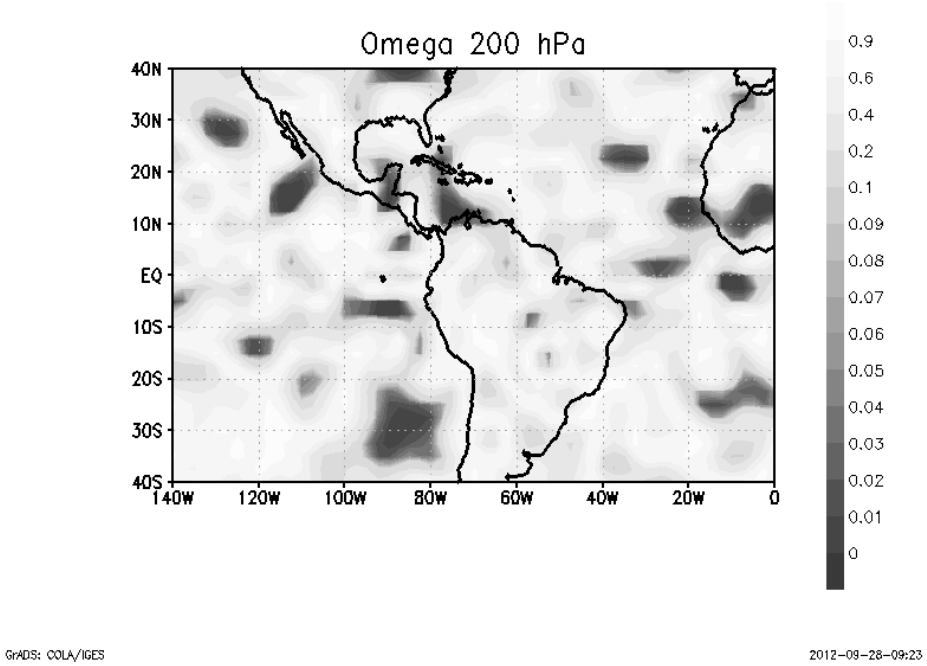


(a)

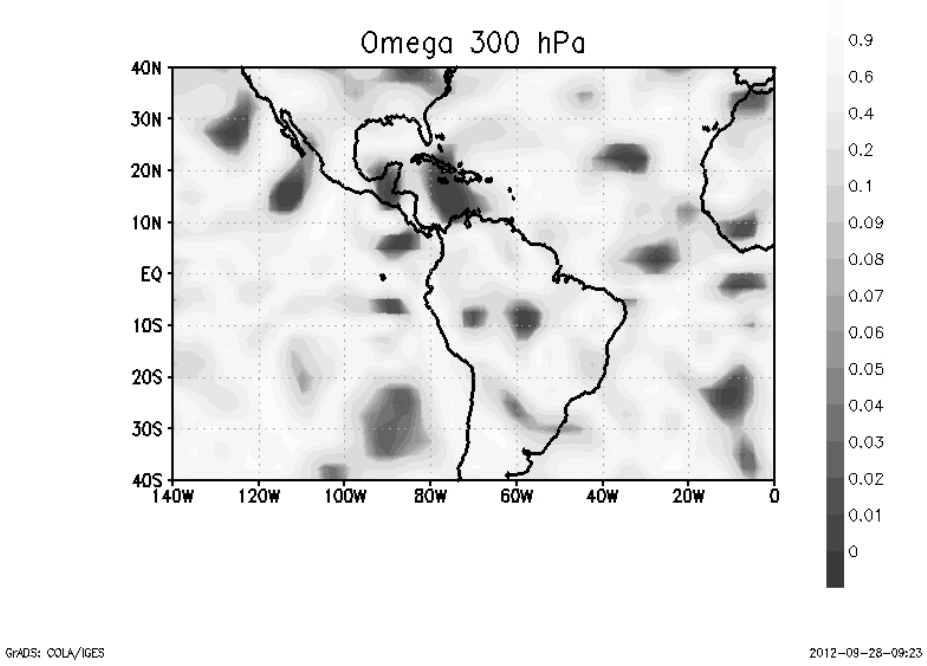


(b)

Figura .2 - Representação em p-valores da influência das variáveis climatológicas na seca da Amazonia - sec XXI.

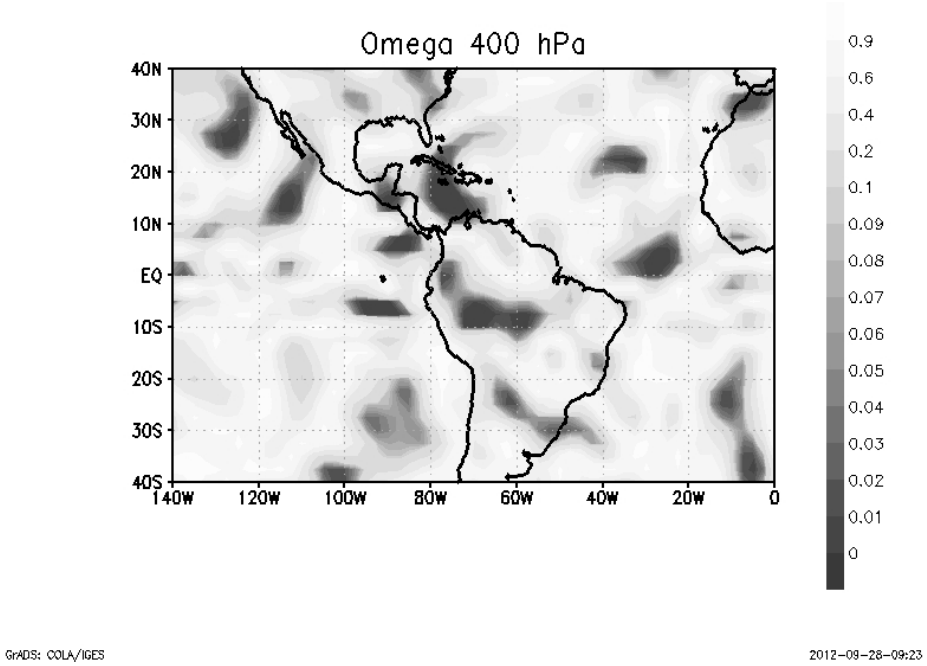


(a)

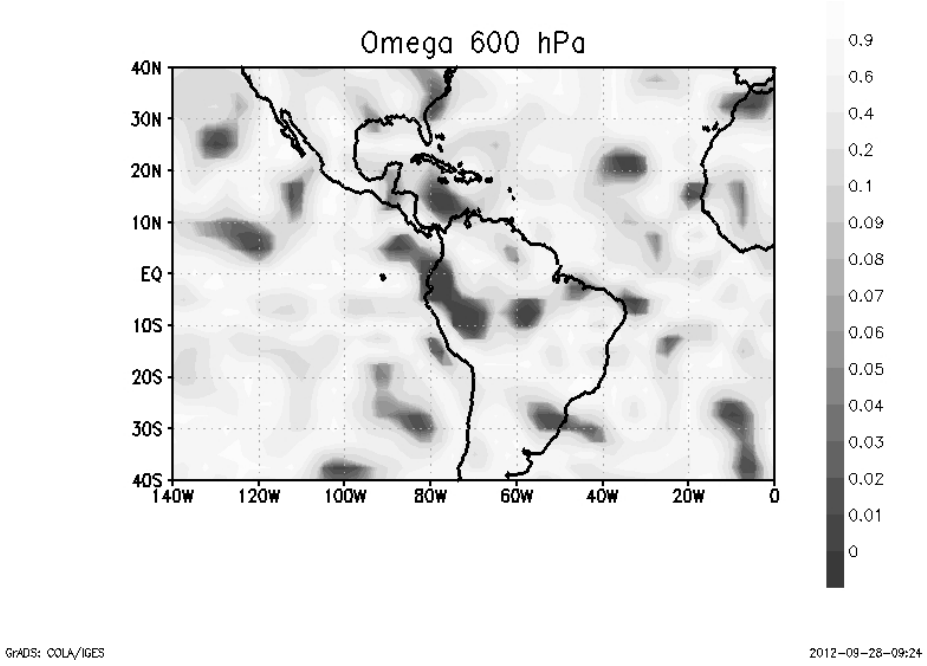


(b)

Figura .3 - Representação em p-valores da influência das variáveis climatológicas na seca da Amazonia - sec XXI.

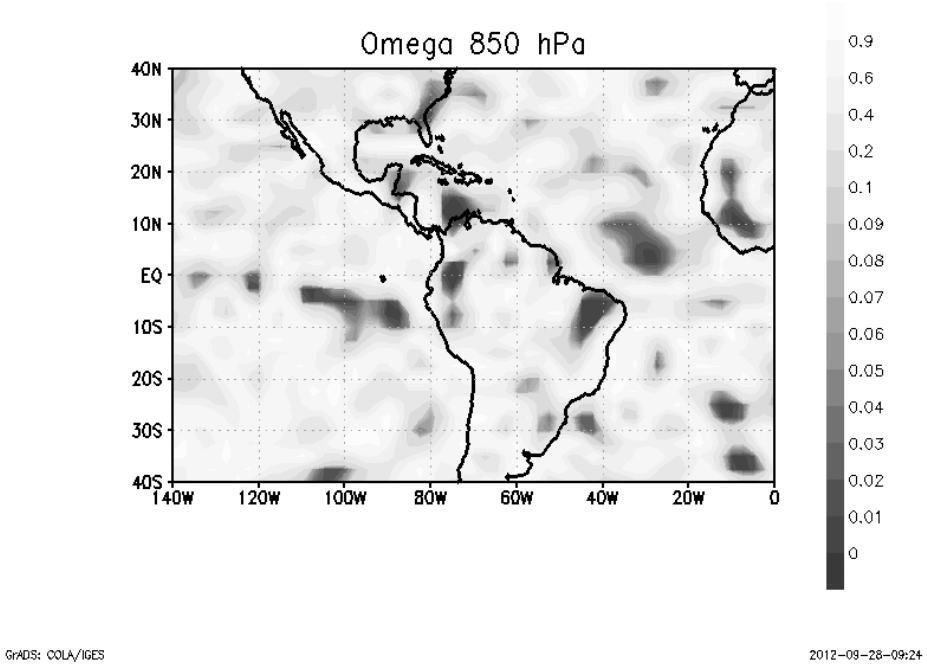


(a)

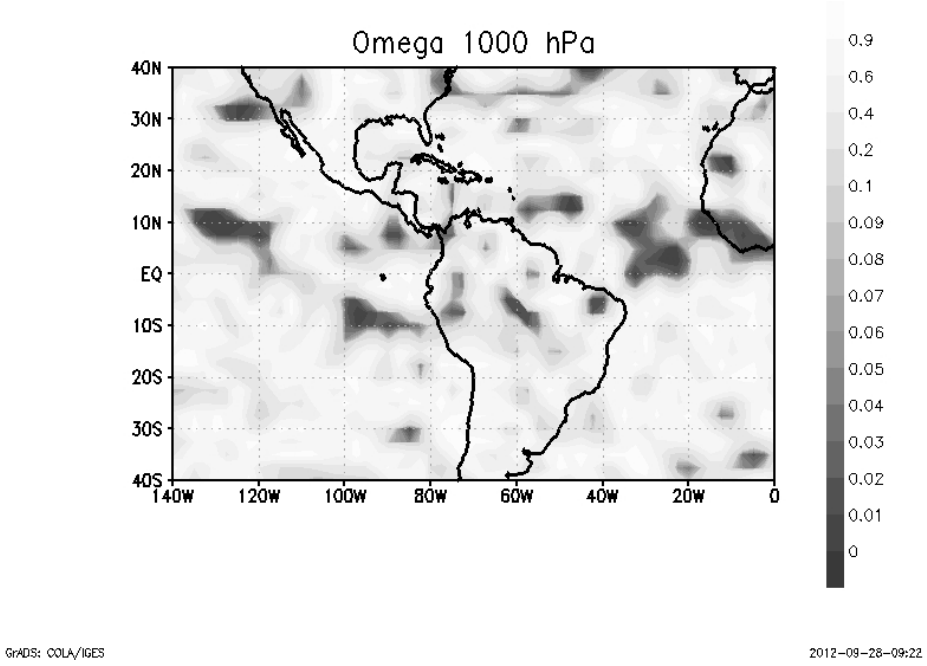


(b)

Figura .4 - Representação em p-valores da influência das variáveis climatológicas na seca da Amazonia - sec XXI.

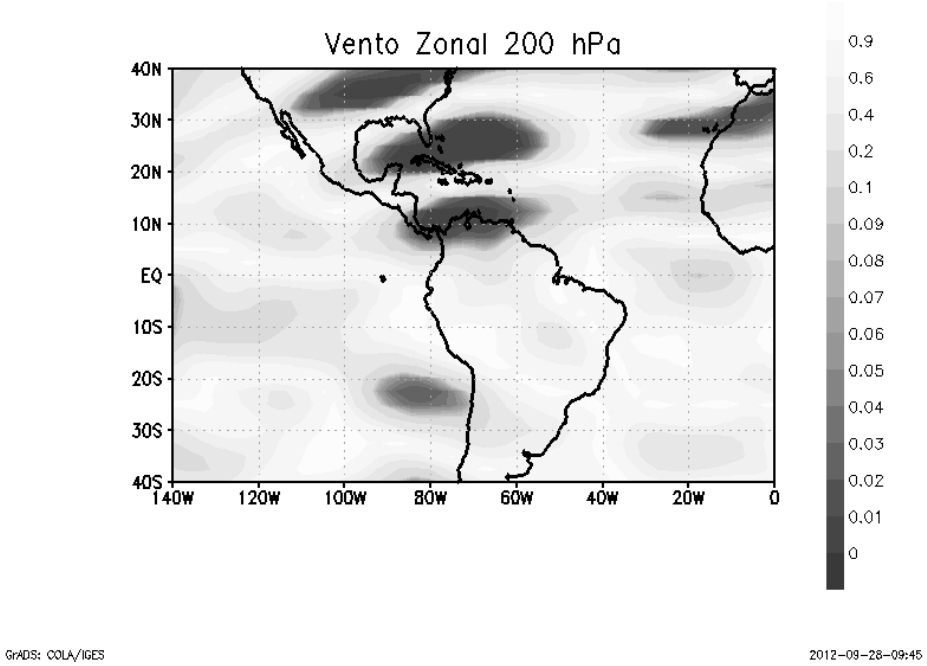


(a)

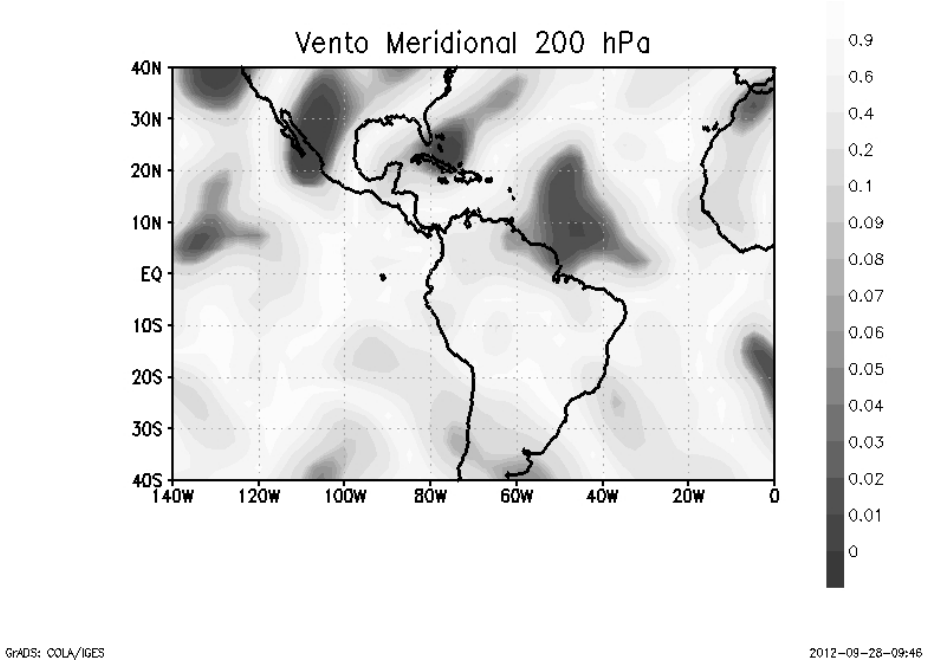


(b)

Figura .5 - Representação em p-valores da influência das variáveis climatológicas na seca da Amazonia - sec XXI.

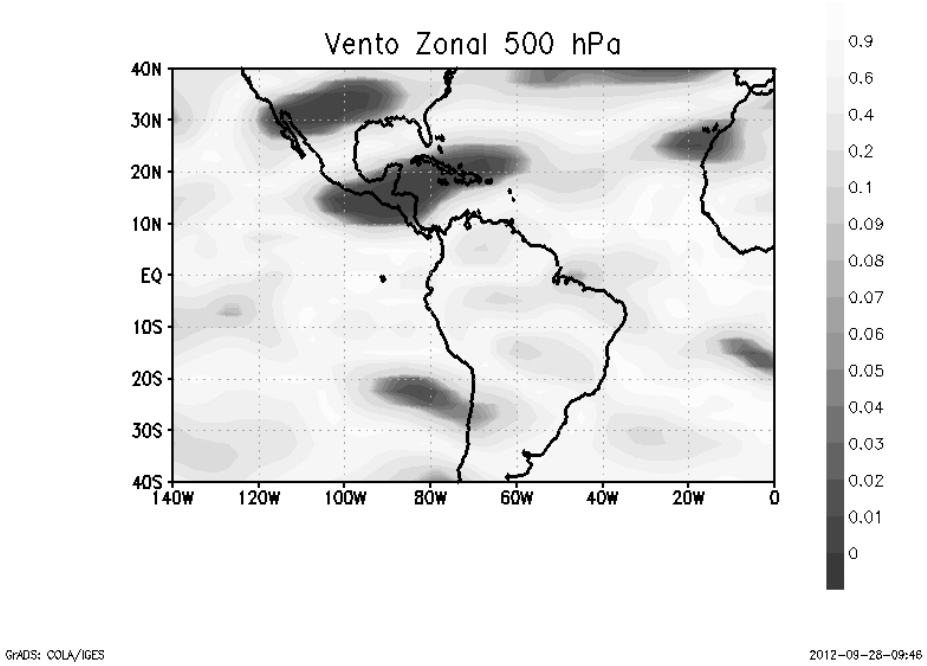


(a)

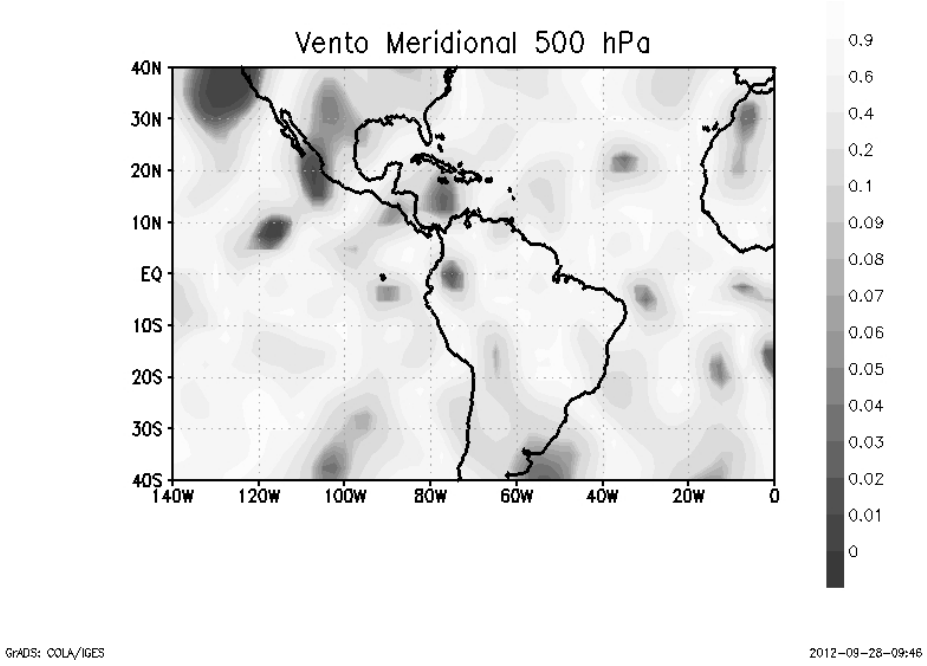


(b)

Figura .6 - Representação em p-valores da influência das variáveis climatológicas na seca da Amazonia - sec XXI.

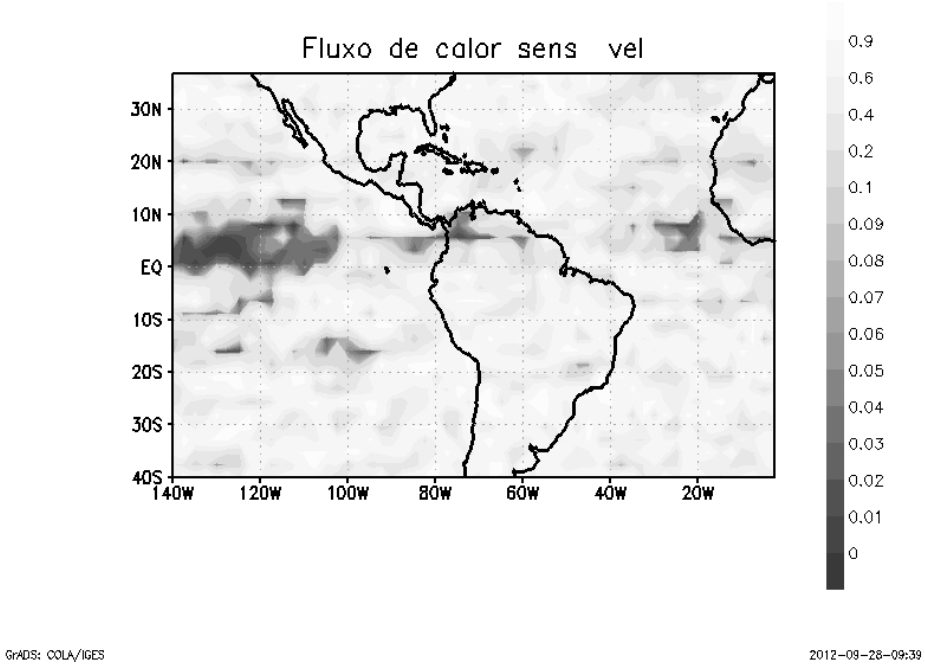


(a)

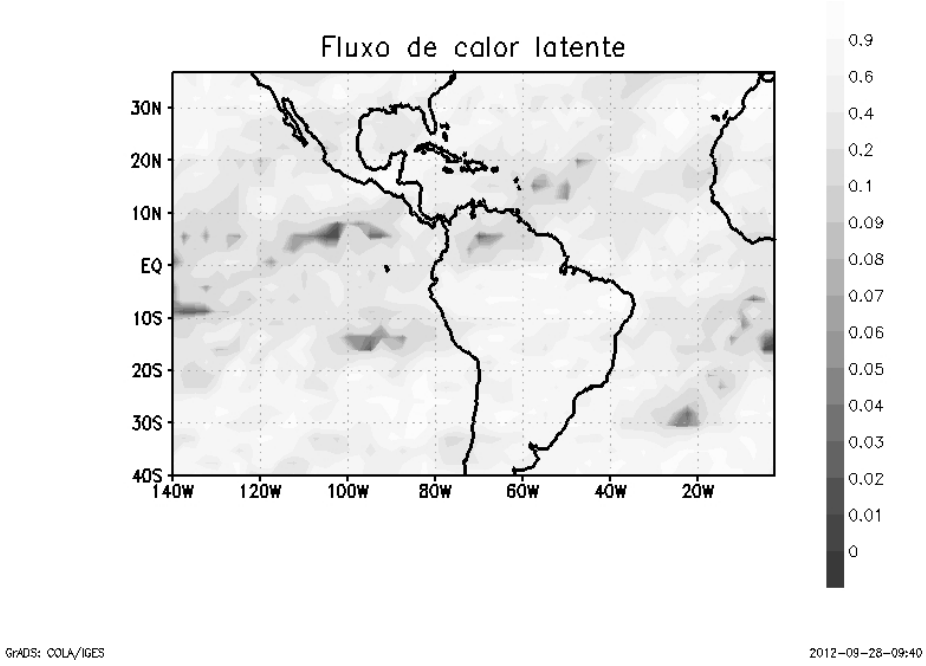


(b)

Figura .7 - Representação em p-valores da influência das variáveis climatológicas na seca da Amazonia - sec XXI.



(a)

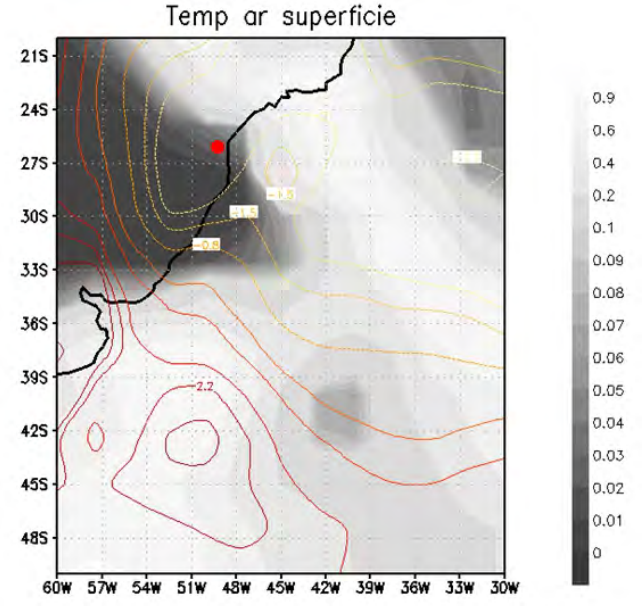


(b)

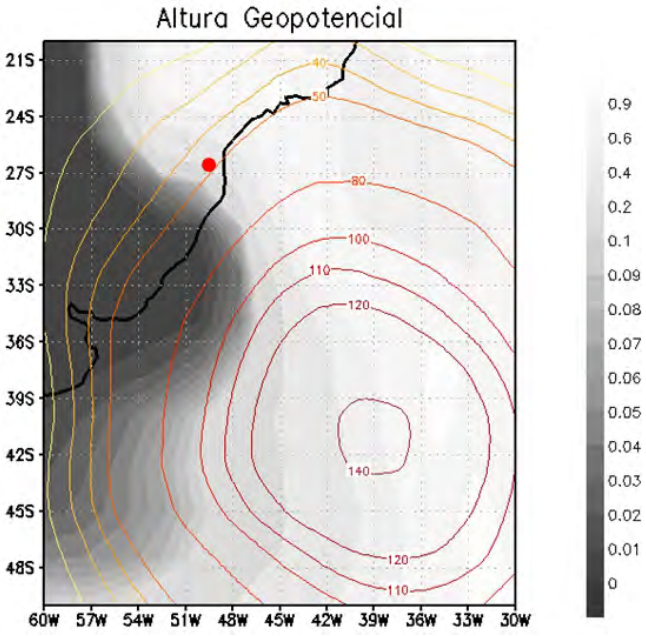
Figura .8 - Representação em p-valores da influência das variáveis climatológicas na seca da Amazonia - sec XXI.

ANEXO B - RESULTADOS ADICIONAIS - SANTA CATARINA

Os resultados apresentados das Figuras .1 e 5.12 referem-se a análise da precipitação extrema em Santa Catarina. Correspondem a comparação entre chuva abundante x chuva moderada e são apresentados em campos de p-valores.



(a)



(b)

Figura .1 - Representação em p-valores da influência das variáveis climatológicas na cheia de Santa Catarina - 2008