20 e 21 de outubro
Instituto Nacional de Pesquisas Espaciais - INPE
São José dos Campos - SP

Metodologias de Mineração de Dados em Análise Climática

Heloisa Musetti Ruivo[1], Haroldo Fraga de Campos Velho[2], Fernando Manuel Ramos[2]

[1]Programa de Doutorado em Computação Aplicada – CAP
Instituto Nacional de Pesquisas Espaciais – INPE

[2]Laboratório Associado de Computação e Matemática Aplicada – LAC
Instituto Nacional de Pesquisas Espaciais – INPE

{heloisa.ruivo, haroldo, fernando}@lac.inpe.br

Abstract. *Data mining methodologies are used to analyse extremes meteorological phenomena. Three methodologies were evaluated: statistical classification of DNA microarray (MA) analysis, decision trees (DT), and the technique of empirical orthogonal functions (EOF). The MA analysis is implemented in the computational tool BRB-ArrayTools Software, an bioinformatic software that has been adapting and applying to the environmental problems. The decision trees algorithms are implemented in WEKA Software. For the EOF scheme, a software package installed in GRADS computer environment was employed. We used these techniques for analyzing the drought of Amazon in 2005. The main goal is to have different techniques to evaluate extreme events. This research is one task of the the National Institute of Science and Technology for Climate Change*

Resumo. *Este trabalho utilizará metodologias de mineração de dados para analisar fenômenos metereológicos extremos. Três metodologias serão avaliadas: classificação estatística utilizada na análise de micro-arranjos (MA) de DNA; árvores de decisão (AD); e a técnica de funções ortogonais empíricas (FOE). O método estatístico de MA está implementado na ferramenta computacional BRB-ArrayTools, software desenvolvido na área de bioinformática, aqui adaptado e aplicada à área ambiental. Algorítmos de árvores de decisão, estão implementados no software WEKA. Para a técnica de FOE, utilizar-se-á ferramentas computacionais já implementadas, como no pacote de pós-processamento GRADS. Estas técnicas serão testadas para análise da grande seca do Amazonas ocorrida em 2005, onde se pode apontar alguns parâmetros climatológicos responsáveis pelo evento. O objetivo aqui é analisar vários eventos extremos e contribuir com um conjunto de ferramentas de avaliação com o Instituto Nacional de Ciência e Tecnologia para Mudanças Climáticas (INCT-MC).*

Palavras-chave: *Mineração de dados, árvore de decisão, análise estatística, climatologia.*

1. Introdução

Há evidências experimentais do aquecimento do sistema climático global, através do monitoramento das temperaturas médias globais do ar e dos oceanos. Em um planeta mais aquecido, os fenômenos climáticos e meteorológicos extremos como secas, inundações, tempestades severas, ventanias e incêndios florestais se tornam mais freqüentes [IPCC 2007]. O progresso tecnológico ocorrido nas últimas décadas possibilitou que a grande quantidade de dados gerada aumentasse rapidamente. Com esse crescimento explosivo de dados armazenados, 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. Este trabalho fará uso desta quantidade de dados armazenada e de ferramentas computacionais, apontando as variáveis climatológicas responsáveis por eventos climáticos extremos.

A mudança climática envolve um dinamismo mais complexo do que a simples elevação da média térmica, mesmo porque o clima não se define só pela temperatura. Contudo, a reação em cadeia que se estabelece a partir do aquecimento deve ser avaliada em profundidade [Conti 2005]. Neste contexto, é de grande importância a interpretação dos dados quantitativos e qualitativos relacionados às perdas e prejuízos no contexto ambiental e sócio econômico. Esta interpretação requer técnicas de mineração dos dados para posterior avaliação das análises. Estas técnicas buscam transformar os dados armazenados em conhecimento.

O INPE é a instituição lider do Instituto Nacional de Ciência e Tecnologia para Mudanças Climáticas (INCT-MC), que é uma das maiores redes de pesquisa em mudancas 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 [INPE]. Neste sentido, este trabalho analisará inicialmente a grande seca da Amazônia ocorrida em 2005 que causou grande prejuízo a população de boa parte da bacia Amazônica [Marengo et al. 2008].

A floresta Amazônica tem importante papel no clima do planeta, por várias razões, em particular devido ao grande estoque de carbono, tanto na biomassa como no solo, bem como o transporte de energia devido a intensa convecção registrada na região. As variações climáticas na região Amazônica podem estar relacionadas às mudanças climáticas globais decorrentes de causas naturais ou de fatores antropogênicos como desflorestamento e queimadas, que provocam mudanças climáticas nesta região.

1.1. Mineração e classificação de dados

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. 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 de dados utilizadas em bioinformática que permitem reduzir o banco de dados identificando os genes mais significantes.

Neste trabalho serão empregadas técnicas computacionais de mineração e classificação de dados para analisar dados climáticos. As aplicações iniciais realizadas investigaram, quais foram os fatores climáticos associados a grande seca de 2005 no Amazonas. As metodologias de mineração de dados que serão empregadas no presente estudo são: algoritmos de árvore de decisão disponíveis no pacote WEKA [Witten 2000]; classificação de dados através da ferramenta computacional BRB-ArrayTools [Simon and Lam 2006] (desenvolvida para aplicações em bioinformática); e funções ortogonais empíricas utilizadas em análises de dados metereológicos.

2. Mineração de dados - abordagem teórica

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. Técnicas de mineração de dados têm sido crescentemente desenvolvidas. O conceito de Mineração de dados está se tornando cada vez mais popular como uma ferramenta de descoberta de informações, podendo revelar estruturas de conhecimento.

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, freqüentemente 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 1996].

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.

A tarefa mais significativa que será abordada neste trabalho é a classificação de dados.

2.1. Classificação

Classificação é a tarefa de mineração de dados que tem sido mais estudada ao longo do tempo. Essa tarefa consiste em classificar um item de dado como pertencente a uma determinada classe dentre várias classes previamente definidas. Cada classe corresponde a um padrão único de valores dos atributos previsores (demais atributos que caracterizam o exemplo). 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.

O principal objetivo da construção de um classificador é descobrir algum tipo de relação entre os atributos previsores e as classes. O procedimento de construção deste classificador é baseado em particionamentos recursivos do espaço de dados. O espaço é dividido em áreas e a cada estágio é avaliado se cada área deve ser dividida em subáreas, a fim de obter uma separação das classes.

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 1984]. Esse conhecimento é geralmente representado na forma de regras "se"..(condições).. "então"..(classe).., cuja interpretação é: "se" os valores dos atributos satisfazem as condições da regra, "então" o exemplo pertence à classe prevista pela regra. Um importante conceito da tarefa de classificação é a divisão dos dados entre dados de treinamento e dados de teste. Inicialmente, um conjunto de dados de treinamento é disponibilizado e analisado, e um modelo de classificação é construído baseado nesses dados. Então o modelo construído é utilizado para classificar outros dados, chamados dados de teste, os quais não foram contemplados pelo algoritmo durante a fase de treinamento [Carvalho 2004].

2.2. Classificação - Árvores de decisão

Métodos de **árvore de decisão** representam um tipo de algoritmo de aprendizado de máquina que utilizam 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 1), 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 atributos (parâmetros) e os resultados deste teste formam os ramos das árvores [Witten 2005]. Cada nó folha, nas extremidades da árvore, representa um valor de predição para o atributo meta [Meira 2008].

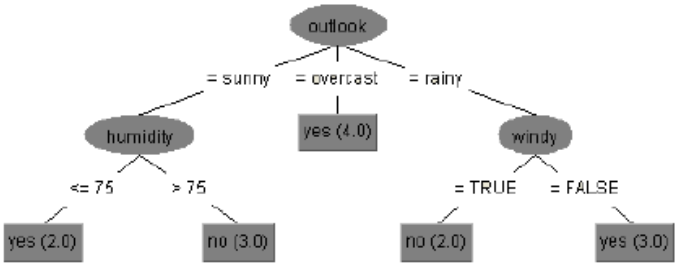

Figura 1. Exemplo de árvore de decisão simples baseada no problema modelo weather [Quinlan 1993], gerada com a ferramenta Weka.

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 em um nó folha, que representa a classe de predição exemplo [Han 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 2001].

Existem várias implementações utilizando **algoritmos** de indução (construção) baseados em árvores de decisão conhecidos na literatura. Neste trabalho será utilizado o J48 que emprega 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 que é definido por um conjunto de atributos.

Para a escolha dos atributos a serem testados, o J48 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.

Uma das ferramentas utilizadas neste trabalho para aplicar técnicas de mineração de dados foi o pacote **WEKA** desenvolvido na Universidade de Waikato na Nova Zelândia [Witten 2000], que tem disponível o algoritmo J48. Este pacote 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 que torna o sistema portável.

2.3. Classificação - BRB ArrayTools

Outra ferramenta de classificação de dados utilizada neste trabalho provém da bioinformática. Trata-se do pacote **BRB-ArrayTools** versão 3.7.0 desenvolvido pelo *Biometric Research Branch of the Division of Cancer Treatment and Diagnosis of the National Cancer Institute*, sob a direção do Dr. Richard Simon. É um software livre, voltado para análise de dados de MA de DNA.

O software foi desenvolvido por estatísticos experientes em análise de dados de microarranjos, mas possui uma interface gráfica que facilita a utilização por biólogos. A tecnologia de microarranjos (MA) de DNA consiste em medir o nível de expressão de milhares de genes simultaneamente. A idéia fundamental é comparar níveis de expressão do gene entre duas amostras de tecidos, uma normal e outra com tumor [Hautaniemi 2003]. A tecnologia de MA é um processo baseado em hibridização que possibilita observar a concentração de mRNA de uma amostra de células analisando a luminosidade de sinais

fluorescentes. Hibridização é o processo bioquímico onde duas fitas de ácido nucléico com seqüências complementares se combinam.

A Figura 2 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. Em seguida, a lâmina é digitalizada [Krutovskii and Neale 2001].

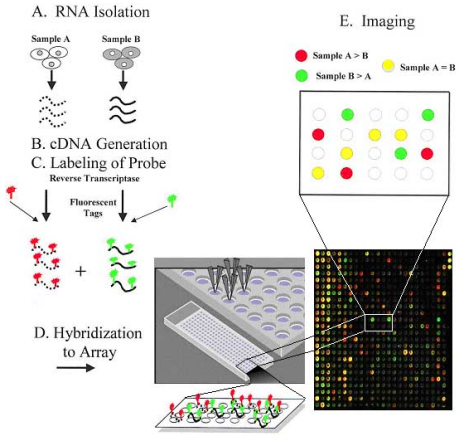

Figura 2. Construção de Microarray.

Neste trabalho, a comparação de classes objetiva determinar se o perfil de expressão do gene difere entre amostras selecionadas de classes pré-definidas e identificar qual gene é diferentemente expresso entre as classes. Uma característica desta classificação é que as classes são pré-definidas independente do perfil de expressão. A comparação de classes entre grupos de amostras calcula o t-teste, ou t-estatístico (quando se tem mais de duas classes) separadamente para cada gene. O t-estatístico é então 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 and Cabrera 2004].

2.4. Funções Ortogonais Empíricas

O conceito de Funções Ortogonais Empíricas (Empirical Orthogonal Functions - EOF) foi introduzido por [Lorenz 1956] em estudos meteorológicos, com a finalidade de encontrar uma maneira eficaz de extrair uma representação simplificada ou compacta de um conjunto de dados.

A decomposição em EOF é uma técnica estatística multivariada usada tanto para se conhecer as dependências existentes entre um conjunto de dados como também para estruturar tal conjunto a fim de se reduzir o número de variáveis inter-relacionadas para um conjunto menor de componentes, que são combinações lineares das variáveis originais.

É um método bastante usado em meteorologia pois permite que a descrição de um campo seja feita por um número relativamente pequeno de funções e coeficientes temporais associados, e também permite investigar processos geofísicos complexos, tais como variações oceânicas ou alterações climáticas a curto prazo [Andreoli 2003].

Os resultados obtidos através do uso de EOF aparecem sob a forma de auto-valores e auto-vetores. Esses auto-vetores formam um conjunto ortogonal completo de vetores através dos quais as observações originais podem ser representadas.

3. Resultados

Este trabalho tem por objetivo comparar metodologias de mineração de dados na análise de dados ambientais, a fim de apontar as variáveis que influenciaram eventos climáticos extremos. Como um estudo preliminar, foi analisado a grande seca ocorrida no Amazonas em 2005.

Foram utilizados conjuntos de dados em grade fornecidos pelo CPTEC/INPE. São dados globais de reanálise (isto é, assimilados e integrados), tomados ao nível de superfície, com resolução espacial de $2.5°$ x $2.5°$, com exceção da temperatura da superfície do mar, cuja resolução é de $2°$ x $2°$.

Este sistema de previsão e análise para executar assimilação de dados usa dados de 1979 até o presente [Kanamitsu et al. 2002]. Todos os dados são mensais e cobrem o período de janeiro de 2000 a dezembro de 2006 (84 meses). Da grade global selecionou-se uma subregião com coordenadas $140W$ a 0W, e 40N a 40S, conforme ilustrado na Figura 3. Dentro desta subregião, foram calculados valores médios mensais de cada grandeza, em quadriláteros de $20°$ de longitude por $20°$ de latitude.

Figura 3. Região analisada: 140W à 0W, 40N à 40S.

3.1. Árvore de decisão

Os algoritmos classificadores de árvores de decisão objetivam criar regras. Neste trabalho, os testes foram executados utilizando o software *WEKA* que tem disponível vários algoritmos de árvore de decisão. Primeiramente foram feitos testes utilizando o algoritmo J48, mas optou-se por testar os outros algoritmos para comparar a robustez dos mesmos.

Neste conjunto de testes, as instâncias representam as médias mensais, e os atributos, as variáveis climatológicas. A classificação baseou-se na vazão do Rio Amazonas em Óbidos (que por estar a jusante, contem "informação" de vários afluentes inclusive o Madeira). Foram consideradas duas classes: alta - valores acima da mediana, e baixa - valores abaixo da mediana. As opções de testes de algoritmos foram:

- *Use training set*: faz a predição (regras) e testa com o próprio conjunto de treinamento submetido ao classificador;

- *Cross-validation*: o classificador é avaliado por validação cruzada. O conjunto de teste é dividido em partes iguais e a predição é aplicada a cada um separadamente.

A Figura 4 representa a árvore obtida utilizando o conjunto de treinamento. A avaliação do conjunto de treinamento classificou corretamente 100% de instancias.

```
20S0S60W40W - Comp Meridional Vento <= -0.06049
|   20S0S120W100W - Temperatura do ar <= 24.5987: baixo (39.0)
|   20S0S120W100W - Temperatura do ar > 24.5987: alto (7.0)
20S0S60W40W - Comp Meridional Vento > -0.06049
|   sst - Atl N (5-20N, 60-30W) <= 28.27: alto (35.0)
|   sst - Atl N (5-20N, 60-30W) > 28.27: baixo (3.0)
```

Figura 4. Árvore de decisão utilizando o algoritmo J48 com opção *Use training set*.

Com os mesmos parâmetros, o algoritmo com a opção *Cross-validation* obteve o melhor resultado para 28 validações cruzadas e com 98.8095% de instancias corretas. A matriz de confusão está representada na Figura 5.

```
 a  b   <-- classified as
41  1 |  a = baixo
 0 42 |  b = alto
```

Figura 5. Matriz de confusão da opção *Cross-validation*.

Foram feitos testes em outros algoritmos de árvore de decisão disponíveis no WEKA, e as porcentagens de acertos estão indicadas na Tabela 1.

Tabela 1. Desempenho de algoritmos de árvore de decisão disponíveis no WEKA

Algoritmo	Opção de teste	Folds	% acertos
ADTree	Use training set		100
BFTree	Use training set		96.4286
RandomTree	Use training set		100
REPTree	Use training set		83.3333
SimpleCart	Use training set		100
J48	Use training set		100
ADTree	Cross-validation	20	98.8095
BFTree	Cross-validation	20	95.2381
RandomTree	Cross-validation	20	92.8571
REPTree	Cross-validation	15	86.9048
SimpleCart	Cross-validation	15	95.2381
J48	Cross-validation	28	98.8095

Observa-se que os algoritmos que melhor classificam os dados foram o J48 e o ADTree. O número de validações cruzadas (Folds) apontados na tabela foi aquele que apontou o maior número de acertos nos testes.

3.2. Classificação - BRB ArrayTools

Esta análise utiliza uma metodologia destinada à análise de dados biológicos. Dentro do espírito da analogia com a análise de MA, cada mês de dados representa um "paciente", e uma coluna na base de dados. Já cada grandeza climatológica (temperatura, velocidade do vento, vazão de rio, etc.), corresponde a um "gene", e uma linha na base de dados. Estendendo a analogia, pode-se imaginar a seca de 2005 como uma "doença", e os fatores climáticos causadores do fenômeno, os genes reguladores ainda desconhecidos. Todos os dados utilizados foram tabulados em termos de anomalias (isto é, o valor corrente menos a média do mês para os 7 anos considerados), normalizadas para variarem no intervalo [-1, 1]. Este procedimento garante que todos os dados terão, em princípio, o mesmo peso na análise.

A extração do conhecimento do banco de dados é realizada através de "projetos". Um projeto necessita a definição das "classes" que nortearão as operações de classificação e agrupamento. Cada projeto busca responder a uma pergunta padronizada do tipo "Quais são as variáveis climáticas, dentre as consideradas no banco de dados, responsáveis pela propriedade *X* (por exemplo, vazão média mensal do rio Amazonas em Óbidos) pertencer a uma determinada classe?".Nesta aplicação, foi considerada a propriedade *X*: vazão do rio Amazonas em Óbidos.

Para esta análise foram consideradas três classes, definidas da seguinte maneira: classe 1 = valores entre -1 e (mediana-0.03), classe 2 = valores entre (mediana-0.03) e (mediana$+0.03$), e classe 3 = valores entre (mediana$+0.03$) e 1. A análise foi feita utilizando-se p-valor de 0.01.

Como o objetivo desta aplicação é analisar o período de seca (mais precisamente, de vazões decrescentes), foram considerados apenas os meses de julho de 2000 a novembro de 2006. Mais precisamente, foram considerados apenas os meses de julho a novembro de 2000 (estiagem moderada), e julho a novembro de 2006 (estiagem extrema).

A Figura 6 apresenta os resultados do agrupamento dos dados onde observa-se o nítido agrupamento baseado em 25 parâmetros, apresentando uma forte diferenciação entre os períodos de seca e chuva. Estes parâmetros podem ser melhor interpretados pela Figura 7.

Resultados desta análise mostram que parâmetros como a temperatura da superfície do mar na região do Atlântico Norte entre as coordenadas 20W e 60W e a temperatura do ar na mesma região aparecem como variáveis chave para explicar a seca de 2005. Neste período, observa-se que a baixa umidade relativa na região do Atlântico Norte, com valores próximos do mínimo, e o fraco movimento vertical sobre a região do Amazonas são também relevantes.

[Marengo et al. 2008] atribuem o aumento da SST no Atlântico Norte tropical como o principal responsável pela seca de 2005, na ausência do fenômeno El Niño. Os autores apontam, também, o fraco movimento vertical sobre a região do Amazonas como um dos possíveis causadores da seca. Estes resultados corroboram as nossas análises, sobretudo no que se refere ao papel fundamental da SST na seca de 2005. Outra análise, publicada em [Trenberth and Shea 2006], também destaca o papel da SST na região do Atlântico Norte ($10°$N - $20°$N) na seca da Amazônia.

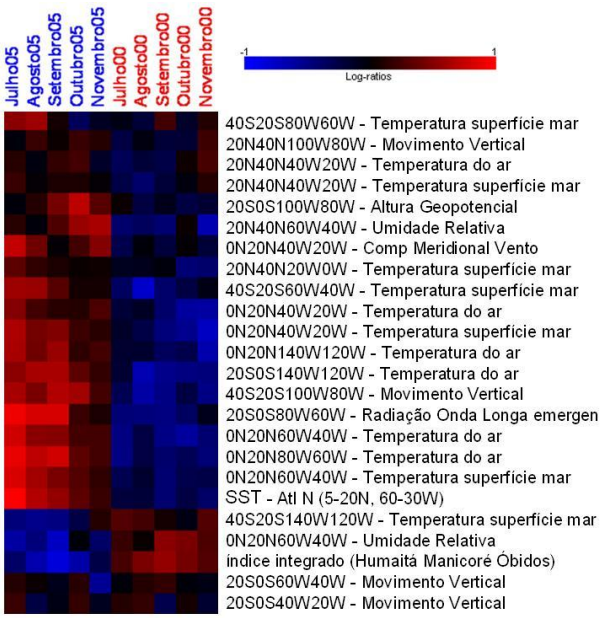

Figura 6. Agrupamento do índice de vazão do Rio Amazonas em Óbidos utilizando-se 3 classes - períodos de Jun à Nov em 2000 e 2005.

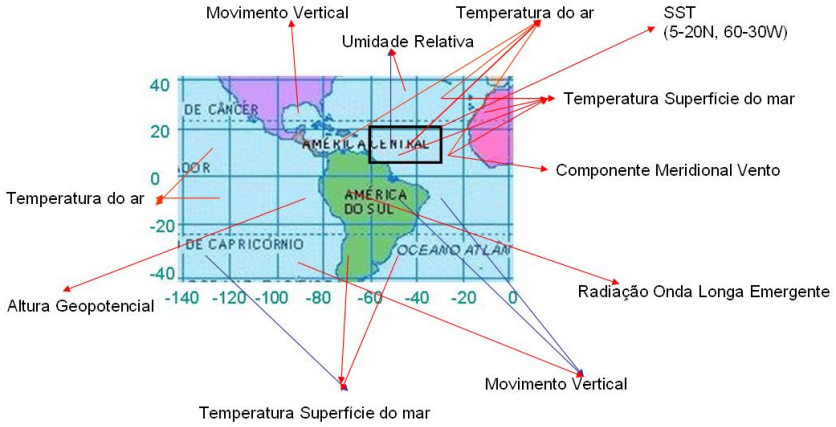

Figura 7. Localização geográfica dos parâmetros mais relevantes encontrados na Figura 6

3.3. Funções Ortogonais Empíricas

Nesta seção serão apresentados os padrões de anomalia da temperatura do ar identificados por meio da análise do padrão espacial da primeira e da segunda EOF. Para o cálculo da anomalia, foi utilizado o padrão de temperatura do periodo da seca para realizar a diferença da média. A interpretação dos resultados da análise é baseada na análise das imagens dos componentes (Figuras 8(a) e 9(a)) e nos gráficos de autovetores (Figuras 8(b) e 9(b)), que representam a série temporal. Estes dois primeiros modos explicam 80.98% e 11.42% respectivamente da variância total.

Observa-se que as séries temporais definem muito bem o ciclo anual onde as componentes mostram o padrão de verão de cada hemisfério. Mais especificamente, o modo 1 é definido por variações sinópticas, enquanto o modo 2, por variações entre continente e oceano. Estes valores podem ser considerados como anomalias positivas ou negativas

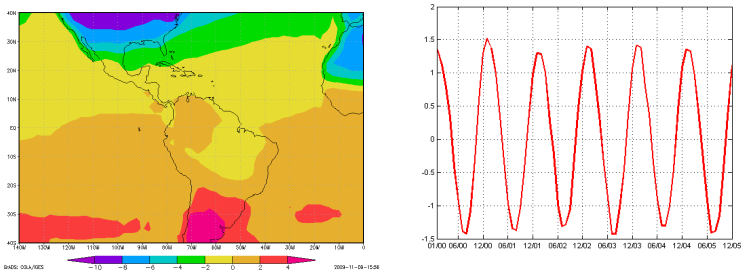

Figura 8. (a) Padrão espacial do primeiro modo; (b) componente principal do primeiro modo.

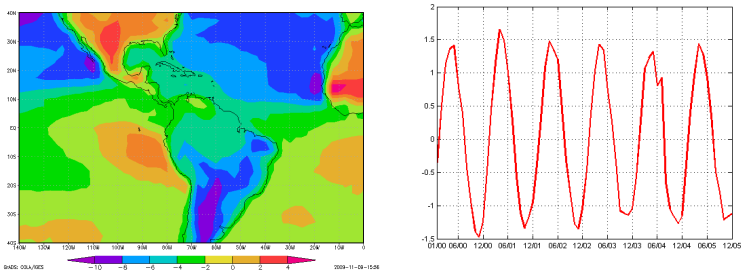

Figura 9. (a) Padrão espacial do segundo modo; (b) componente principal do segundo modo.

em relação à média mensal de temperatura do ar. Se um mês apresenta um alto valor de autovetor indica que ele contém um padrão espacial muito semelhante ao representado pela imagem componente. Deve-se ressaltar, que quando o autovetor for negativo indica que este mês tem padrão inverso do demonstrado na imagem componente, ou seja, se o valor de temperatura do ar está aparecendo alto na imagem componente, na realidade, naquele mês ele foi baixo.

4. Conclusão

Os resultados mostram que metodologias de mineração de dados podem apontar fatores climatológicos que influenciam fenômenos metereológicos. O método estatístico que é utilizado para análise de dados de MA, foi adaptado e validado na análise de dados climatológicos. Observa-se que os algoritmos de árvore de decisão apresentam uma grande porcentagem de acerto na classificação dos dados. Já as funções ortogonais empíricas apresentam padrões bem definidos quando se analisa parâmetros climatológicos responáveis pelo evento analisado.

Referências

Amaratunga, D. and Cabrera, J. (2004). *Exploration and analysis of DNA microarray and protein array data*. Wiley Interscience, New Jersey.

Andreoli, R. V. (2003). Variabilidade e previsibilidade da temperatura da superfície do mar no allântico tropical.

BREIMAN, L.; FRIEDMAN, J. O. R. S. C. (1984). *Classification and Regression Trees*. Wadsworth and Brooks, Monterey, Ca.

Carvalho, D. R. (2004). Árvore de decisão / algoritmo genético para tratar o problema de pequenos disjuntos em classificação de dados.

Conti, J. B. (2005). Considerações sobre as mudanças climáticas globais. *Revista do Departamento de Geografia USP*, 16:70–75.

Fayyad, U.; Piatesky-Shapiro, G. S. P. U. R. (1996). *Advances in Knowledge Discovery and Data Mining*. The MIT Press, California.

Han, J.; Kamner, M. (2001). *Data mining: concepts and techniques*. Morgan Kaufmann Publishers, San Francisco.

Hautaniemi, S. (2003). Studies of microarray data analysis with applications for human cancers.

INPE. Ministério da ciência e tecnologia - instituto nacional de pesquisas espaciais.

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

Kanamitsu, M., Ebisuzaki, W., Woollen, J., Yang, S., Hnilo, J. J., Fiorino, M., and Potter, G. L. (2002). Ncep-doe amip-ii reanalysis (r-2). *American Meteorological Society*, 83:1631–1643.

Krutovskii, K. V. and Neale, D. B. (2001). Forest genomics for conserving adaptive genetic diversity.

Lorenz, E. N. (1956). Empirical orthogonal functions and statistical weather predictio. In *Statistical Forecasting Project - Sci. Rep. No. 1*, page 48, Cambridge.

Marengo, J. A., Nobre, C. A., Tomasella, J., Oyama, M. D., Oliveira, G. S., Oliveira, R., Camargo, H., Alves, L. M., and Brown, I. F. (2008). The drought of amazonia in 2005. *Journal of Climate*, 21(3).

Meira, C. A. A. (2008). 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.

Quinlan, J. R. (1993). *C4.5: programs for machine learning.* Morgan Kaufmann Publishers, San Francisco.

Simon, R. and Lam, A. P. (2006). *BRB-ArrayTools - version 3.4 - User's manual*. National Cancer Institute.

Trenberth, K. E. and Shea, D. J. (2006). Atlantic hurricane and natural variability in 2005. *Geophysical Research Letters*, 33.

Witten, I. H, . F. E. S. (2000). *Data mining: Practical machine learning tools and techniques with java implementation.* Morgan Kaufmann Publishers, California.

Witten, I. H.; Frank, E. (2005). *Data mining: practical machine learning tools and techniques.* Morgan Kaufmann Publishers, San Francisco.