Anais 2º Simpósio de Geotecnologias no Pantanal, Corumbá, 7-11 novembro 2009,
Embrapa Informática Agropecuária/INPE, p.323-332.

Análise de padrões seqüenciais em série histórica do rio Paraguai

Laurimar Gonçalves Vendrusculo[1]
Stanley Robson de Medeiros Oliveira[1]
Júlio César Dalla Mora Esquerdo[1]
João Francisco Gonçalves Antunes[1]

[1] Embrapa Informática Agropecuária - CNPTIA
Av. André Tosello, 209 - Caixa Postal 6041
13083-886 - Campinas - SP, Brasil
(laurimar, stanley, julio, joaof}@cnptia.embrapa.br

Resumo. O crescente armazenamentos de dados com características temporais desafia os pesquisadores a elaborar algoritmos eficientes para a descoberta de conhecimento. A mineração de dados, por meio da descoberta de padrões seqüenciais, contribui para o entendimento de ocorrência de fenômenos que possuam um ciclo não conhecido. Neste trabalho, foram analisados dados referentes à altura do Rio Paraguai de uma série histórica centenária. Utilizou-se a técnica SAX para redução da dimensionalidade e representação simbólica dos dados. Por meio do algoritmo GeneralizedSequentialPatterns, as seqüências mais freqüentes encontradas explicaram a periodicidades transitórias e permanentes do ciclo hidrológico do Rio Paraguai. Como validação da técnica SAX, utilizou-se índices de reflectância advindos da banda 3 A (infra-vermelho próximo) de imagens AVHRR-NOOA. Os resultados nas duas metodologias identificaram similarmente períodos de cheia e estiagem no período de 2004 a 2006.

Palavras-Chave: mineração de dados, padrões seqüenciais, séries temporais

Abstract. The increase of data storage capacity with spatial characteristics poses new challenges to researchers to design efficient algorithms for knowledge discovery. Data mining, through the discovery of sequential patterns, contributes to the understanding of natural phenomena that have an unknown cycle. In this article, it was analyzed data related to height of the River Paraguay from a historical series. The SAX technique was applied to dimensionality reduction and symbolic representation of the data. By using the algorithm GeneralizedSequentialPatterns, the most frequent sequences explained the transitory and constant periods of the hydrological cycle of the River Paraguay. As a validation of the technique SAX it was used reflectance indices arising from the band 3 A (near infrared) of AVHRR-NOOA images. The results presented by the two methods identified similar periods of flood and drought in the period from 2004 to 2006.

Key-words: data mining, sequential patterns, temporal series.

1. Introdução

A mineração de padrões seqüenciais abrange o desafio de localizar subseqüências freqüentes, as quais representam padrões em grandes bases de informação. As técnicas para localização destes padrões podem ser usadas em uma ampla gama de aplicações tais como: análise de mercado varejista, weblogging, tratamento de doenças, análise de seqüências de DNA entre outras. Pode-se perceber que todas essas aplicações incorporam aspectos relacionados ao tempo.

O problema de mineração de padrões seqüenciais foi inicialmente discutido por Agrawal & Srikant (1995). Dada uma base de dados de seqüências, onde cada seqüência é uma lista de transações ordenadas pelo tempo e cada transação uma lista de itens, o problema relacionado à mineração de padrões seqüenciais resume-se em encontrar subseqüências freqüentes, que satisfaçam o suporte mínimo especificado pelo usuário, isto é, localizar aquelas subseqüências cuja freqüência de ocorrência no conjunto de seqüências não seja menor que o suporte mínimo.

Uma série temporal $T = t_1,t_2 \ldots ,t_m$, pode ser definida como um conjunto escalar ou multivariado de observações medidas no tempo, em intervalos iguais. Para um valor de m elevado, extrair informação útil passa a ser um desafio.

No contexto das tarefas de mineração de dados, o importante é obter propriedades locais e não globais da série. Para tanto, a abordagem mais direta é o particionamento da série temporal em subseqüências $C = t_p, \ldots , t_p+n-1$ o de T, com tamanho $n<<m$, onde p é uma posição aleatória e para $1 \leq p \leq m - n + 1$. Por meio das subseqüências pode-se analisar os eventos próximos, e.g. período de valores elevados (cheia) de um rio, ou todas as possíveis ocorrências da subseqüência extraídas na série como um todo, e.g. conjunto de baixos valores de vazão hidrológica em uma série centenária.

A estatística clássica oferece um conjunto de técnicas para análise das séries temporais. Para a análise de tendências, pode-se ajustar modelos de regressão polinomial na série como um todo ou em na vizinhança de algum ponto de interesse. Outra classe de modelos matemáticos aplicados em séries temporais, onde não se detecte tendência ou sazonalidade, são os modelos auto-regressivos (Latorre & Cardoso, 2001).

Outra abordagem para analisar a série temporal seria utilizar as tarefas da mineração de dados. Segundo Han & Kamber (2006) é grande o interesse em descobrir conhecimento implícito em séries temporais. Estudos que retratam o interesse da comunidade científica em tarefas de mineração relevantes como indexação, agrupamento, classificação e segmentação, são respectivamente, descritos por Yi & Faloutsos (2000), Keogh & Pazzani (1998) e Ge & Smyth (2000).

O foco deste estudo é a descoberta de padrões seqüenciais, a qual tem por objetivo identificar a seqüência que ocorre mais freqüentemente na série temporal. Para tanto, será utilizada uma representação simbólica dos dados denominada SAX (*Symbolic Aggregate approXimation*), proposta por Lin et al. (2007). SAX permite a redução da dimensionalidade

dos dados e apresenta uma medida de distância eficaz para cálculo da similaridade entre a série discretizada e a série dos dados originais *(Lower Bounding*).

O restante deste artigo será organizado da seguinte forma: Seção 2 discute brevemente as técnicas de representação simbólica voltada à mineração de dados, enfatizando o método SAX. Seção 3 contém uma avaliação experimental da abordagem simbólica utilizando dados de uma série histórica estática de cento e seis anos das cotas diárias (altura) do Rio Paraguai. A validação do método SAX é realizada por meio de índices de refletância da banda 3 A de imagens AVHRR-NOAA. Finalmente, a seção 4 discute alguns dos resultados obtidos e a seção 5 oferece algumas conclusões e sugestões para trabalhos futuros.

2. Material e Métodos

2.1. Representação simbólica das séries temporais

O grande gargalo para análise computacional de uma série temporal com grande quantidade de dados tem sido processá-la totalmente na memória principal.

Uma abordagem genérica para processar dados históricos procura criar uma aproximação dos dados que se ajuste a memória principal, como primeira etapa. A aproximação dos dados, com os aspectos de interesse é submetida ao algoritmo da mineração de dados. A fase final consiste em realizar alguns acessos à série original, armazenada em disco, de forma a confirmar a solução obtida na fase anterior.

Porém de nada vale esta abordagem se a aproximação dos dados não for significativa em relação à série original. Os acessos ao disco na última fase podem ficar numerosos e inviabilizar uma solução eficiente.

Para facilitar a aproximação dos dados ou sua representação simbólica, vários autores propuseram soluções tais como: Transformada discreta de Fourier - DFT (Faloutsos et al., 1994), Transformada discreta de ondaletas - DWT (Chan & Fu, 1999), Piecewise linear e Piecewise constant models - PAA (Keogh et al., 2001). A **Figura 1** ilustra as representações mais utilizadas no processo de discretização dos dados.

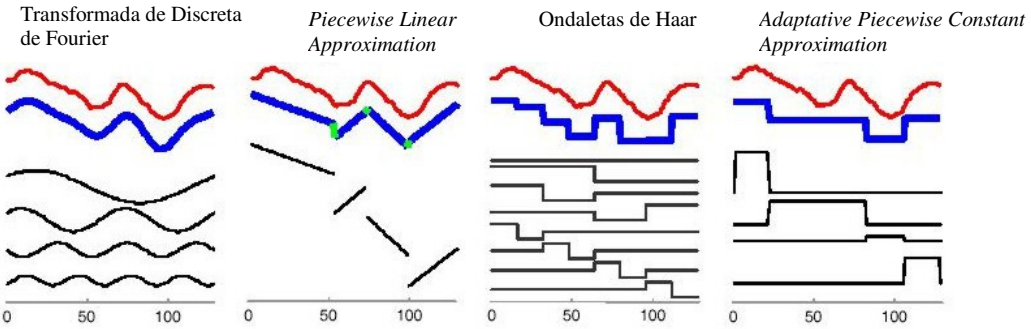


Figura 1. Representações mais usuais para séries temporais utilizadas em tarefas de mineração de dados (Figura extraída de Li et al.. (2007).)

A técnica SAX permite que uma série de tamanho n seja reduzida para uma palavra (*string*) de tamanho w, onde normalmente $w << n$. O tamanho do alfabeto é um inteiro representado por a, onde $a > 2$.

A primeira fase de SAX transforma os dados na representação *Piecewise Aggregate Approximate* (PAA). Este algoritmo divide uma série em N segmentos de tamanho fixo (*frames*), adjacentes um ao outro, porém disjuntos, por meio do cálculo da média dos valores dos pontos no interior de cada segmento. Desta forma obtém-se uma série intermediária formada pelos valores médios de cada um dos segmentos.

De imediato, a vantagem no uso de SAX é a redução da dimensionalidade, onde uma série temporal *C* de tamanho *n* pode ser representada no espaço *w*-dimensional pelo vetor Č = $\check{c}_1$, ..., $\check{c}_w$ O *n*-enésimo elemento de Č pode ser calculado pela **equação 1**:

$$\overline{c}_i = \frac{w}{n} \sum_{j=\frac{n}{w}(i-1)+1}^{\frac{n}{w}i} c_j \quad \textbf{(1)}$$

Visualmente, a representação SAX pode ser vista como uma tentativa de aproximar os dados originais utilizando uma combinação linear de segmentos, conforme mostra a **Figura 2**. Para simplificar, SAX considera que *n* é divisível por *w*. Antes de se converter para a representação PAA, a série é normalizada para ter média zero e desvio padrão igual a um.

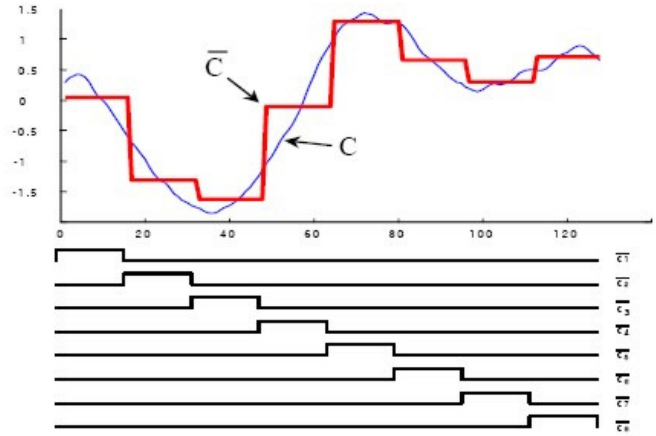


Figura 2. Representação em PAA de uma série temporal de tamanho original 128, reduzida para 8 dimensões (Figura extraída de Li et al.. (2007)).

O passo seguinte a transformação em PAA é obter uma representação discreta. Para tanto, parte-se do princípio que a técnica de discretização produza símbolos com probabilidade iguais (eqüiprobabilidade) de ocorrência. Considera-se que as séries temporais normalizadas têm distribuição gaussiana ou normal.

Após a normalização dos dados, determinam-se quantos *breakpoints* a curva gaussiana produzirá. Os *breakpoints* produzem *m* áreas de igual tamanho sobre a curva. Na técnica SAX o número de *breakpoints* varia de três a dez.

Após obter a série PAA, o algoritmo verifica os coeficientes que estão abaixo do menor *breakpoint*, neste caso a este segmento é associado à letra “a”. Todos os coeficientes maiores que o menor e menores que o próximo breakpoint são mapeados com o símbolo “b” e assim sucessivamente, conforme ilustra a **Figura 3**. A palavra encontrada na série é : **baabccbc**

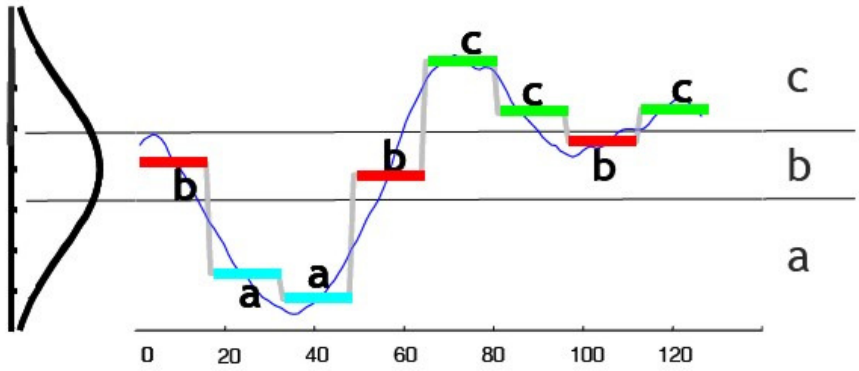


Figura 3. Utilização de *breakpoints* predeterminados para mapear os coeficientes PAA. O exemplo mostra uma série com $n = 128$, $w = 8$ e $a = 3$. (Figura extraída de Li et al. (2007)).

2.2. Avaliação experimental da abordagem simbólica

Os dados utilizados neste estudo referem-se aos valores diários obtidos por uma régua de medição, do nível do Rio Paraguai, localizada no 6° distrito Naval da Marinha em Ladário, Mato Grosso do Sul (Lat: -19° 0' 06'' e Long -57° 35' 39"), conforme ilustra a **Figura 4.** Os dados foram extraídos no sistema de Informações Hidrológicas – HidroWeb[1], desenvolvido pela Agência Nacional de Águas (Ana), relativos ao período de 1900 a 2006.

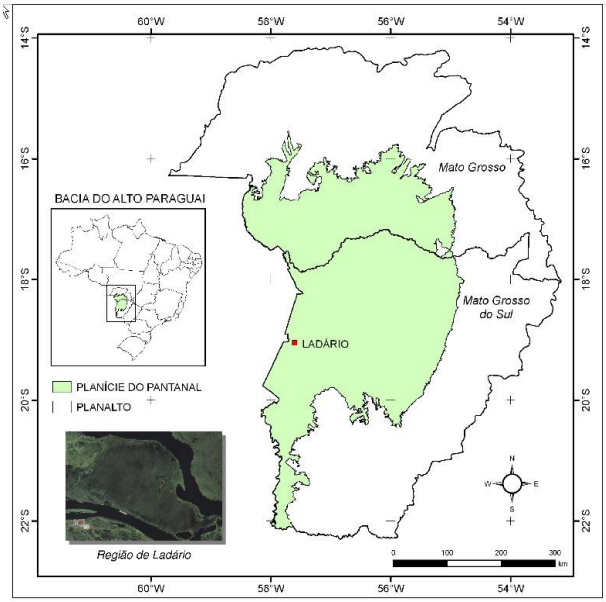


Figura 4. Bacia do Alto Paraguai e a localização de Ladário, Mato Grosso do Sul.

A importância do Rio Paraguai é inegável, pois constitui-se num dos principais tributários da Bacia do Prata, que é a segunda maior bacia da América do Sul. Daqueles rios que compõem a Bacia do Alto Paraguai, é o que mais se direciona para o centro do continente e abrange regiões distintas como o Planalto e o Pantanal.

O Planalto é uma região com alta altitude, chegando a atingir valores acima de 200 m até 1.400 m. A drenagem nesta região é bem definida e convergente. Ao contrário, o Pantanal é uma região com baixas cotas de altitude.

Os dados obtidos do sistema Hidroweb são obtidos na forma de uma planilha eletrônica. Cada linha da planilha original apresenta dados mensais e os valores diários, em centímetros,

[1] Disponível na URL: http://hidroweb.ana.gov.br/

representados pelos atributos Cota *n*, onde *n* representa o dia do mês. O esforço inicial foi transpor manualmente as linhas mensais em valores diários na forma de colunas.

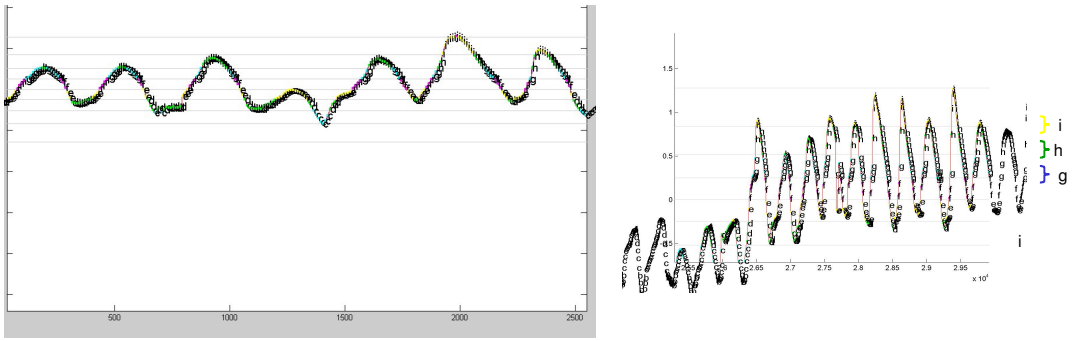
No período de 1977 a 1989 os dados da planilha Excel apresentaram-se redundantes. Para cada mês havia três registros idênticos. A exclusão das linhas se deu por inspeção visual e de forma manual.

Na fase seguinte, foi aplicada a conversão da série em PPA e posteriormente a atribuição de valores simbólicos para cada segmento da curva. Utilizou-se os algoritmos MatLab disponíveis na Home Page (http://www.cs.ucr.edu/~jessica/sax.htm) do Projeto SAX. O ambiente MATLAB Student 6.5, foi utilizado, customizando os seguintes programas para a série em estudo:

- Timeseries2symbol.m - Converte a série em uma String SAX
- sax_demo.m - Normaliza os dados e apresenta graficamente a série original, em PAA e SAX.

A série original possuía 38.652 registros equivalentes aos 106 anos da série. Como o valor de *w* deve ser divisível pelo tamanho original, retirou-se os últimos dois valores, arredondando a série para 38.650 registros. Optou-se pelo valor de 10 para o *breakpoint*, para tanto o valor de *w* passou a ser 3.865. O alfabeto para *w* = 10 foi representado pelos seguintes símbolos: a,b,c,d,e,f,g,h,i.

A curva de saída gerada pelos algoritmos é mostrada pela **Figura 5.**



(a) (b)

Figura 5. Curva PPA com representação simbólica (a) Visão ampliada de parte da representação simbólica.

Os algoritmos disponibilizaram uma palavra de 3.865 caracteres, onde cada caracter associava o valor representativo de um (1) decêndio (dez dias) do mês.

Como cada letra representava o período de um decêndio, houve esforço manual para criar subseqüências que representassem o período anual (36,6 strings). Neste caso, optou-se por considerar 37 letras. A trigésima sétima letra sempre recebia o mesmo valor da trigésima sexta posição.

Posteriormente escolheu-se reduzir a janela temporal para um mês, gerando então um novo arquivo com subseqüências de três letras (Ex: a2006,"f,f,g,1m"). Para o décimo segundo mês, optou-se por acrescentar mais uma letra, com valor igual ao último decêndio (Ex: a2006,"f,f,g,g,12m").

Para análise de padrões utilizou-se um dos mais populares ambientes para descoberta do conhecimento de domínio público, a *ferramenta Waikato Environment for Knowledge Analysis* – WEKA (Witten & Frank, 2005). O WEKA fornece uma API para suporte ao aprendizado de máquina, implementado em Java, que incorpora vários algoritmos para:

seleção de atributos, seleção de instâncias, algoritmos de aprendizagem supervisionada e outros.

O algoritmo utilizado foi o *GeneralizedSequentialPatterns* no Ambiente Weka 3.5.6. Para um suporte de 0.01, os padrões mais encontrados, para seqüências de tamanho um, e o respectivo número de vezes, foram:

{h,h,h} - (99)	**{e,e,e} - (111)**	**{c,c,c} - (110)**	**{d,d,d} - (165)**

Para analisar o ciclo hidrológico do Rio Paraguai em toda a extensão da série temporal, construiu-se numa planilha com a presença dos padrões mais encontrados associados a cada mês. Uma janela temporal de janeiro de 2004 a dezembro de 2006 dessa planilha é ilustrada pela **Figura 6**.

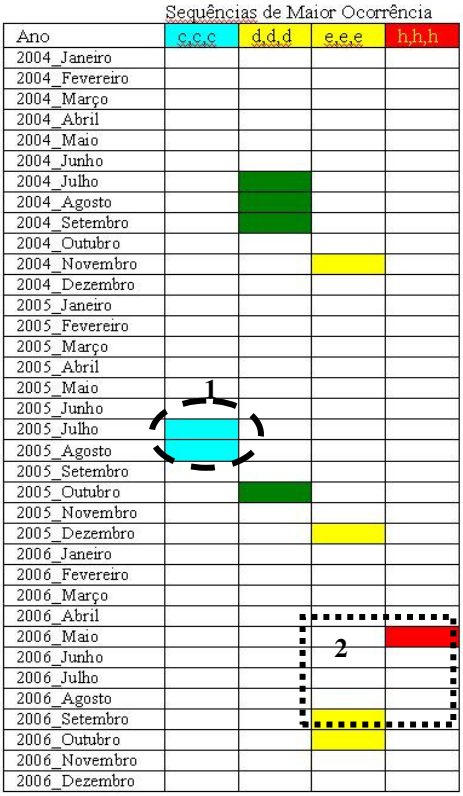


Figura 6. Destaque das subseqüências que mais ocorreram na série temporal hidrológica no período de 2004 a 2006.

Com o objetivo de validar a metodologia SAX utilizada nesse trabalho, em cotas do rio Paraguai, optou-se por utilizar a técnica proposta por Antunes & Esquerdo (2007) e comparar os resultados obtidos. Foram utilizadas imagens AVHRR-NOAA obtidas do acervo do Centro de Pesquisas Meteorológicas e Climáticas Aplicadas à Agricultura (CEPAGRI/UNICAMP), em seu estado bruto. Foram processadas ao todo 1628 imagens das passagens diurnas do satélite NOAA-17, entre janeiro de 2004 e dezembro de 2006. Foi analisada somente a banda 3A cuja faixa espectral varia de 1,58 a 1,64 μm se situando então na região do Infravermelho médio (IVM). Esta banda, segundo Antunes & Esquerdo (2007), enfatiza significativamente a presença de água nos locais amostrados. Utilizou-se um algoritmo baseado em análise harmônica para suavizar a curva temporal do IVM, mostrada na **Figura 7**. Ressalta-se que a refletância no IVM, mostrada na **Figura 7**, é inversamente proporcional à presença da água, ou seja, quanto menor a refletância da superfície, maior a quantidade de água.

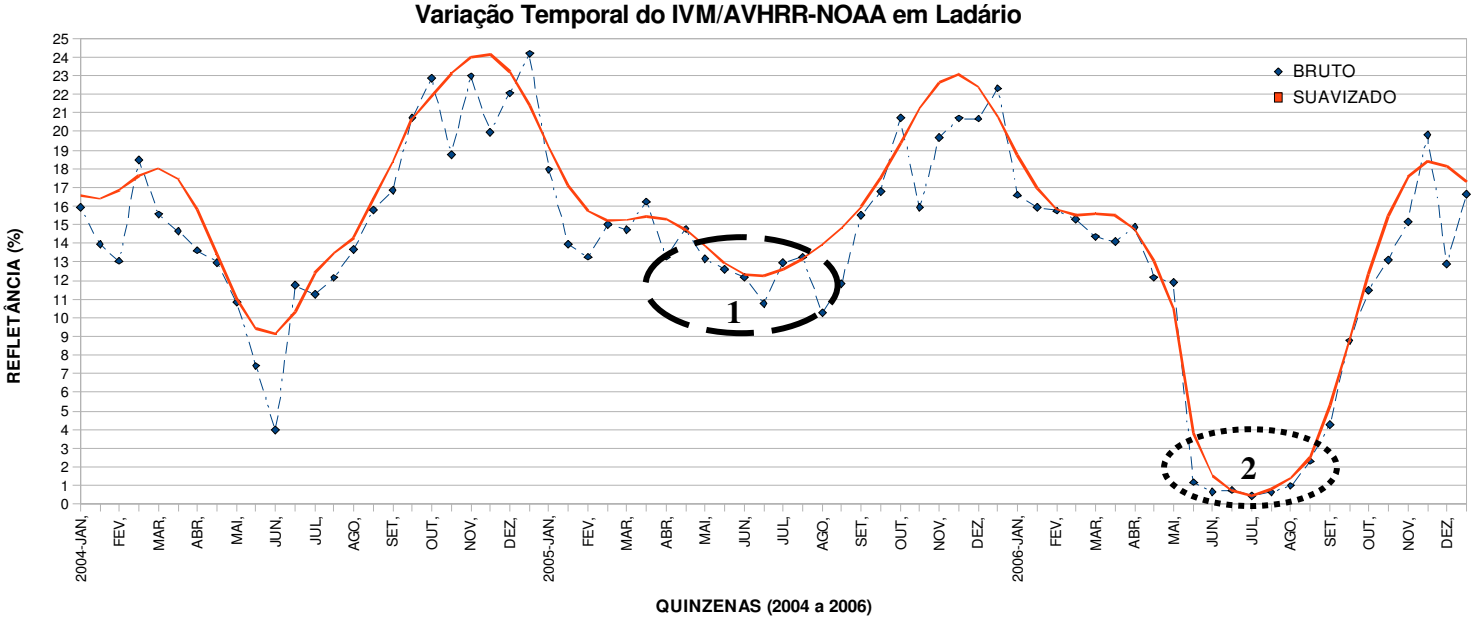


Figura 7. Variação temporal do IVM (banda 3A) no município de Ladário, MS.

3. Resultados e Discussão

A opção de se estudar o ciclo hidrológico do Rio Paraguai em um período mensal mostrou-se promissora, pois para períodos maiores (semestrais ou anuais), houve poucas ocorrências de padrões similares. A opção mensal também coincidiu com o período das estações do ano e os resultados permitiram confirmar que, posterior a uma cheia (período chuvoso) verifica-se uma estação de baixa (seca)

Analisando-se toda a série histórica, encontrou-se uma periodicidade das cotas médias (**d**,**e**); ocorrendo alternadamente, em um período de três em três anos, desde o início do século até 1936. Sá et al. (1998), por meio de ondaletas, encontrou o período de quatro anos.

De 1936 a 1950 verifica-se uma periodicidade quase anual de alternância entre as letras **c** e **d**, nos meses de novembro a março (período chuvoso)

Constatou-se que até 1960 a periodicidade de altas cotas (**h**) era de quatro a cinco anos, com duração de três meses. A partir da década de 70 a freqüência muda para um ano, porém com menor duração (um mês).

Uma melhoria deste estudo considera que os períodos trimestrais devam ser associados aos seus anos de ocorrência e não mais considerados como trimestres independentes. Desta forma espera-se encontrar padrões de tamanho maior que um.

Embora os resultados do algoritmo *GeneralizedSequentialPatterns* mostraram apenas o somatório das maiores ocorrências, uma análise destes padrões mensais permitiu encontrar um média de cinco a seis das subseqüências mais encontrados no período de um ano. Portanto, estas são subseqüências representativas do ciclo.

Os produtos resultantes do processamento da banca 3 A do sensor AVHRR-NOAA são similares aos resultados encontrados pela metodologia SAX, para o período de Janeiro de 2004 e dezembro de 2006, nas duas situações.

Observa-se que o círculo 1, da **Figura 7**, corresponde ainda a um período de estiagem, atípico para esta época no pantanal onde a cheia já deveria ter se iniciado a partir do mês de

abril, conforme informa a Embrapa Pantanal (referência). O círculo 1, da **Figura 6**, representa este período de seca com cotas baixas, por meio do padrão "c,c,c".

Ilustrando um período bem definido de Cheia na **Figura 7**, o círculo 2, no período de maio a agosto de 2006, também é representado pelos padrões por cotas altas ("e,e,e" e "h,h,h").

4. Conclusões

A abordagem de redução de dimensionalidade e a representação simbólica dos dados, implementadas pela técnica SAX, mostraram-se eficazes e rápidas, no que tange ao processamento computacional de uma série temporal centenária.

A técnica também prescinde o conhecimento do comportamento dos dados. Tal característica é interessante uma vez, que nem sempre o especialista em mineração de dados conhece profundamente a temática da base de dados.

A validação metodológica da técnica SAX com imagens orbitais da faixa do infravermelho médio mostrou-se coerente, onde os períodos de cheia e seca foram detectados claramente.

Como trabalhos futuros estão previstas a aplicação de algoritmos para predição da cota do Rio Paraguai, utilizando classificadores. Deve-se aplicar a técnica em outras bases de dados, tais como séries agroclimatológicas que contemplem os atributos de temperatura e precipitação.

Pretende-se automatizar os algoritmos, de forma que se integrem com sistemas legados que já armazenem dados com características temporais. Sugere-se também como trabalhos futuros que alguma técnica de visualização deve ser adotada para que os padrões sejam visualizados na própria série temporal.

5. Agradecimentos

Os autores agradecem o CEPAGRI/UNICAMP pela cessão das imagens NOAA para este trabalho.

6. Referências

Agrawal, R.; Srikant,R.. Mining Sequential Patterns. **In Proc. of the 11th Int'l on Data Engineering**, Taipei, Taiwan:March 1995.

Antunes, J. F. G. ; Esquerdo, J. C. D. M. . Geração automática de produtos derivados de imagens AVHRR-NOAA para monitoramento de áreas inundáveis do Pantanal. RBC. **Revista Brasileira de Cartografia,** v. 59, p. 115-122, 2007.

Chan, K.; Fu, A.W. Efficient Time Series Matching by Wavelets. In **proceedings of the 15 th IEEE Int. Conference on Data Engineering.** Sydney:Australia. p. 126-133, Marc. 1999.

Faloutsos, C.; Ranganathan, M.; Manopopulos, Y. Fast Subsequence Matching in Time-Series Databases. **SIGMOD Record**. v. 23. p. 419-429. 1994.

Ge, X.; Smyth, P. Deformable markov model templates for time-series pattern matching. In **proceedings of the 6 th A CM SIGKDD Int Conference on Knowledge Discovery and Data Mining**. Boston:MA,. p 81-90,. Aug 2000

Han, J. Kamber, M. **Data mining concepts and techniques**. San Diego, CA: Academic Press, 2001. 550 p.

Keogh, E. ; Pazzani, M. (1998). An enhanced representation of time series which allows fast and accurate classification, clustering artd relevance feedback. In **proceedings of the 4 th Int. Conference on Knowledge Discovery and Data Mining.** New York: NY, p 239-241, Aug 1998.

Latorre, M. Do R. D. De O; Cardoso, M. R. A. Time series analysis in epidemiology: an introduction to methodological aspects. **Rev. Bras. Epidemiol**. , São Paulo, v. 4, n. 3, 2001 . p. 145-152. Disponível em:

<http://www.scielosp.org/scielo.php?script=sci_arttext&pid=S1415-790X2001000300002&lng=en&nrm=iso>. Acesso em: 17 Dec 2007.

Lin, Jessica; Keogh, Eamonn; Wei, Li; Lonardi, Stefano Experiencing SAX: a novel symbolic representation of time series**. Journal Data Mining and Knowledge Discovery.** v. 15, n. 2 ,. p. 107-144, Out. 2007

Sá, L. A. A; Sambatti, S. B. M. ; Galvão, G. P Ondeleta de Mortlet Aplicada ao Estudo da Variabilidade do Nível do Rio Paraguai em Ladário, MS. **Pesq. Agropec. Bras.**, Brasília v. 33, Número Especial, p. 1775-1785, 1998

Yi, B.; Faloutsos, C.. Fast time sequence indexing for arbitrary lp norms. In **proceedings of the 26 th Intl Conference onVery Large Databases**. Cairo:Egypt,. p 385-394, Sept 2000.

Witten, I.H.; Frank, E. **Data Mining: Practical machine learning tools and techniques**, Morgan Kaufmann, San Francisco, 525 p.2005.