Correlação cruzada em janela deslizante ajustável aplicada a séries temporais de descargas elétricas atmosféricas e radar meteorológico

João Victor Cal Garcia[1], **Dr. Stephan Stephany**[2], **Dr. Augusto Brandão d'Oliveira**[3]

[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

[3]Centro de Previsão de Tempo e Estudos Climáticos – CPTEC
Instituto Nacional de Pesquisas Espaciais – INPE

sawamano@gmail.com,stephan@lac.inpe.br,augusto.oliveira@cptec.inpe.br

Abstract. *One of the techniques used to study data series is the cross-correlation, which results can be affected by the presence of gaps in the data. One way to tackle this problem is the use of sliding windows. This paper presents a variation of the cross-correlation technique based on sliding windows with variable size based on data integrity in order to better evaluate the series studied. The case of study employs meteorological data.*

Resumo. *Uma das técnicas para estudar séries de dados é correlação cruzada, cujo resultado pode ser alterado pela presença de falhas nos dados. Uma forma de contornar o problema é o uso de janelas deslizantes. O presente trabalho apresenta uma variação da técnica de correlação cruzada por janelas deslizantes que, baseada na integridade dos dados, altera o tamanho da janela de forma a melhor avaliar as séries estudadas. O estudo é aplicado então a dados meteorológicos.*

Palavras-chave: *Radar Meteorológico, Séries Temporais, Descargas Elétricas Atmosféricas, Correlação Cruzada*

1. Introdução

Técnicas para estudar o comportamento conjunto de duas ou mais séries temporais, em específico a correlação cruzada, são severamente prejudicadas por lacunas nos dados [Box et al. 1976, Brockwell and Davis 2009, Fuller 1976]. O uso de uma janela deslizante para limitar a quantidade de dados que será analisada cada vez pode amenizar esse problema, além de permitir uma descrição detalhada da influência das lacunas no resultado [Yang et al. 2009]. Esse estudo é aqui aplicado na análise de séries de dados meteorológicos relacionados com atividade convectiva severa, especificamente a densidade de ocorrência de descargas elétricas atmosféricas, utilizando dados provindos do sistema RINDAT, e a refletividade medida pelo radar meteorológico da UNESP de Bauru do IPMet.

2. Análise por janela deslizante adaptativa

Sejam $X = x_1, x_2, ..., x_n$ e $Y = y_1, y_2, ..., y_n$ duas séries temporais de n amostras, definidas por $x_t = f(t)$ e $y_t = g(t)$, $t \in \Re$. Seja $X_{a,a+b} = x_a, x_{a+1}, ..., x_{a+b}$, e similarmente, $Y_{a,a+b} = y_a, y_{a+1}, ..., y_{a+b}$, um subintervalo das séries X e Y. Então o intervalo $[a, a+b]$ dos índices de x e y formam uma janela $W_{a,a+b}$, com $a \in [1, n-b]$ e $b \in [0, n-1]$. Pode-se dizer que $b+1$ é o comprimento dessa janela, o subconjunto dos dados sequenciais que se estende de x_a e y_a até x_{a+b}, y_{a+b}. O deslocamento é concluído quando a janela atinge o final da série.

Seja L a extensão da maior lacuna contínua e L' a soma das extensões de todas as lacunas. Para não haver janelas compostas apenas de dados de lacunas, $b + 1 > L$. O tamanho máximo da janela é aquele em que a maior parte das janelas possui uma fração menor de lacunas que a série em seu total. Se só houvesse uma lacuna, $b < n - 2 \cdot L'$, com n o comprimento da série, e haveria s janelas influenciadas pela lacuna, tal que $L' < s < 2 \times L'$. O tamanho da janela ideal estará no intervalo $b \in [L, n - 2 \cdot L')$. Esse tamanho ótimo será o ponto no dado intervalo cuja quantidade média de lacunas em cada janela seja mínima.

O algoritmo da correlação cruzada é de complexidade de ordem $O(n \cdot \log(n))$ no tempo, já que n se refere ao número de amostras de uma série temporal. Com o uso das janelas faz necessário que o cálculo seja realizado $n-1$ vezes, uma para cada possível tamanho de janela b, e o processo total passa a ter ordem $O(n^2 \cdot \log(n))$. Calcular os resultados para todas as janelas possíveis em cada tamanho de janela seria inviável na maioria dos casos já que seria necessário se calcular a correlação cruzada para as $n - b$ posições de cada janela, o que acarretaria em uma complexidade $O(n^3 \cdot \log(n))$. A solução é encontrar um tamanho de janela ótimo e calcular a correlação apenas para esse tamanho.

3. Aplicação

Os testes foram realizados a partir de um conjunto de séries de dados meteorológicos relacionados com atividade convectiva severa, no período de março a dezembro de 2009, num circulo de raio 60 km a partir da cidade de Bauru, São Paulo. As séries temporais correspondem à média da densidade de ocorrência de descargas elétricas atmosféricas geradas a partir dos dados provindos do sistema RINDAT, utilizando o Software EDDA [Strauss et al. 2010], e à refletividade média do radar meteorológico da UNESP de Bauru do IPMet. A correlação entre dados de descargas elétricas atmosféricas e dados de radar pode atingir valores próximos a 0.9 [Carey and Rutledge 2000]. Supondo a hipótese que esses resultados são válidos para os dados estudados, é esperado que a correlação seja alta.

Não havia leituras de radar para cerca de 8.5% do período estudado. As falhas da série de radar são apresentadas na Figura 1. Todos as falhas foram supostas com o valor 0, que índica a não-detecção do fenômeno estudado. As séries possuem $n = 22400$ pontos, uniformemente amostrados a cada 15 minutos.

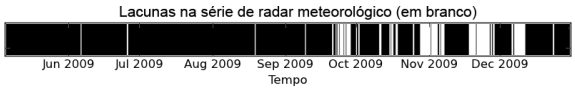

Figura 1: Lacunas em branco na série temporal de radar meteorológico.

O máximo valor da correlação cruzada entre as duas séries foi inferior a 0.15, com um atraso de 5 unidades de tempo, conforme mostra a Figura 2, indicando pouca correlação entre as duas séries.

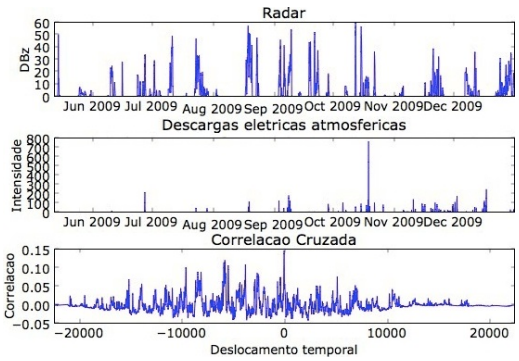

Figura 2: Correlação cruzada entre as séries estudadas.

No uso das janelas deslizantes, encontrou-se $L = 1287$, $L' = 1906$, e $b \in (1287, 18588)$. A janela de tamanho mínimo é apresentada na Figura 3, com o eixo das ordenadas sendo a posição do início da janela e o das abscissas, a correlação cruzada daquela janela entre as séries em valor absoluto; a tabela de cores se refere a correlação. Os resultados de pico são em torno de 0.8 na região inicial da série, onde ocorrem poucas lacunas; porém a janela relativamente pequena faz com que as lacunas mais longas alterem mais severamente os resultados posteriores.

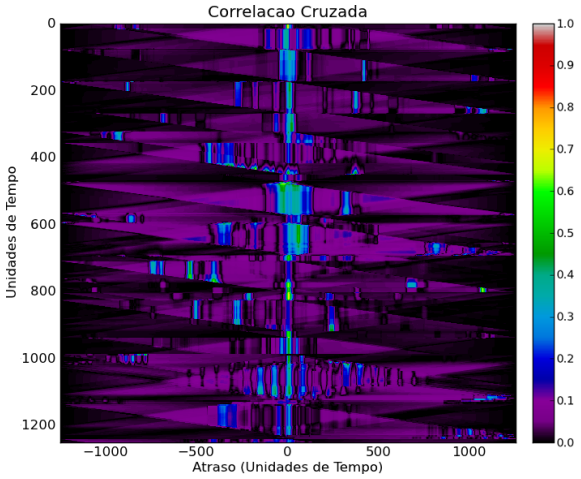

Figura 3: Resultado da correlação cruzada por janela deslizante com b=1288.

Empiricamente foi obtido $b = 11004$ como tamanho ótimo para janela, mostrado na Figura 4. A máxima correlação tende para um atraso pequeno, próximo de 0. A partir da janela de posição 4200, há uma notável alteração nos valores médios da correlação, que seguiam um padrão similar até então. Esse comportamento, detectado pelo uso das janelas delizantes, se deve à maior ocorrência de lacunas por janela próximas ao final da série.

No começo da análise as séries tem um deslocamento temporal próximo a 3, com correlação de 0.3 a 0.8. No entanto, assim que a janela começa a ter forte presença das

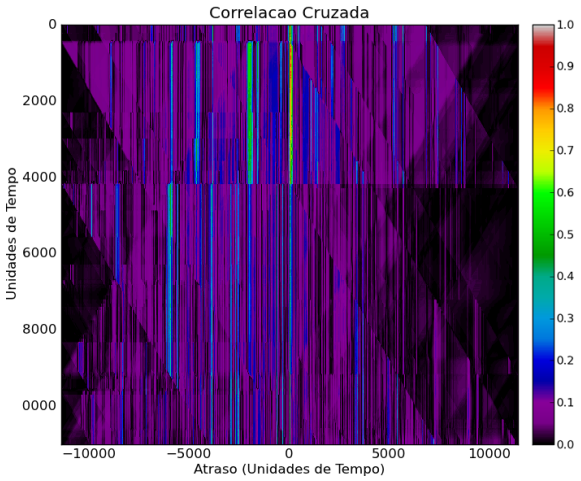

Figura 4: Resultado da correlação cruzada por janela deslizante com b=11004.

lacunas os resultados se modificam, apontando correlação inferior a 0.3 e deslocamento temporal superior a 400 unidades de tempo.

4. Conclusões

Foi descrito o problema do cálculo da correlação cruzada entre duas séries temporais quando ocorrem falhas em parte dos dados. O uso de janelas deslizantes foi apresentado como uma maneira de tratar o problema. Ele melhora os resultados de cálculo da correlação cruzada, e dá uma ideia melhor de como as lacunas nos dados afetam a correlação da série. No entanto, os resultados obtidos dizem respeito apenas a esse conjunto de dados, e ainda é necessário mais estudo a fim de validar a hipótese proposta para outras classes de dados. Encontrar o tamanho ideal para a janela é computacionalmente caro. Trabalhos futuros poderiam ser realizados a fim de encontrar uma maneira mais eficiente de determinar o tamanho ótimo de janela.

Referências

Box, G., Jenkins, G., Reinsel, G., et al. (1976). *Time series analysis: forecasting and control*, volume 16. Holden-day San Francisco.

Brockwell, P. and Davis, R. (2009). *Time series: theory and methods*. Springer Verlag.

Carey, L. D. and Rutledge, S. A. (2000). The Relationship between Precipitation and Lightning in Tropical Island Convection: A C-Band Polarimetric Radar Study. *Monthly Weather Review*, 128(8):2687–2710.

Fuller, W. A. (1976). *Introduction to statistical time series*. Wiley, New York.

Strauss, C., Stephany, S., and Caetano, M. (2010). A ferramenta EDDA de geração de campos de densidade de descargas atmosféricas para mineração de dados meteorológicos. In *33º Congresso Nacional de Matemática Aplicada e Computacional (CNMAC-2010)*, Água de Lindóia, SP.

Yang, H., Zhu, L., and Chu, R. (2009). Fault-plane determination of the 18 april 2008 Mount Carmel, Illinois, earthquake by detecting and relocating aftershocks. *Bulletin of the Seismological Society of America*, 99(6):3413.