IAHR **CIC**

XXV CONGRESSO LATINO-AMERICANO DE HIDRAULICA
SAN JOSÉ, COSTA RICA, 9 AO 12 DE SETEMBRO DE 2012

PREVISÃO DAS VAZÕES MÉDIAS MENSAIS NO RIO TOCANTINS ARAGUAIA USANDO POLINÔMIOS PONDERADOS

Julián David Rojo H.[1], Nelson J. Ferreira[2], Luis Fernando Carvajal[1],Daniel A. Rodriguez[3]
[1]Escola de Geociências e Meio Ambiente. Universidade Nacional da Colômbia, sede Medellín - Colômbia
[2]Centro de Previsão de Tempo e Estudos Climáticos - CPTEC/INPE - Cachoeira Paulista, SP-Brasil
[3]Centro de Ciência do Sistema Terrestre - CCST/INPE - Cachoeira Paulista, SP-Brasil
jdrojoh@unal.edu.co, lfcarvaj@unal.edu.co, nelson.ferreira@cptec.inpe.br, daniel.andres@inpe.br

RESUMO:

O presente trabalho tem por objeto disponibilizar uma nova ferramenta para a previsão de vazões médias mensais utilizando o método não paramétrico de regressão conhecido como polinômios localmente ponderados ou mínimos quadrados móveis. O esquema de prognóstico proposto tem por conceito básico ponderar com maior valor aquelas observações mais próximas no momento de efetuar um prognóstico; apresentam-se as equações básicas do método e o procedimento para encontrar os parâmetros das regressões locais. A metodologia proposta aplicada à previsão das vazões médias mensais do rio Tocantins-Araguaia no Brasil conseguiu representar adequadamente as principais características das series de tempo para as diferentes estações fluviométricas.

ABSTRACT:

This paper aims to introduce a new tool for prediction of monthly mean river flows using the nonparametric regression method known as locally weighted polynomial. Localization of the regression is achieved by using k nearest neighbors of the point of estimate and a monotonic distance based weight function, we present the basic equations of the method and procedure to find the parameters of the local regressions. The proposed methodology is applied in the monthly mean river flows predictions for the Tocantins-Araguaia River in Brazil and the results show an improvement in river flow prediction.

PALAVRAS CHAVES:

Previsão de vazões, regressão não paramétrica, polinômios ponderados.

INTRODUÇÃO

Nos últimos anos, a frequente aplicação de analises estatística em inúmeros problemas da hidrologia motivou a procura de soluções não habituais que se adaptassem aos requerimentos e às circunstâncias atuais de previsão não linear e não estacionaria. O campo não paramétrico é um dos mais populares, uma vez que oferece uma alternativa mais sofisticada em relação aos modelos paramétricos tradicionais, na exploração de dados univariados ou multivariados, sem pressupor nenhuma distribuição específica dos dados. O presente trabalho tem por objeto desenvolver uma ferramenta para a previsão de vazões médias mensais no rio Tocantins-Araguaia usando uma aproximação não paramétrica provida pelos polinômios ponderados de regressão.

POLINÔMIOS PONDERADOS

A estimativa não paramétrica da distribuição de probabilidade constitui-se em um importante objeto de pesquisa estatística, embora as primeiras tentativas de estimação não paramétrica das distribuições de probabilidade começaram na década de 1930. A preocupação com o desenvolvimento desse tema só surgiu nos anos oitenta, com a publicações de inúmeros trabalhos sobre aspectos teóricos sobre essa estimação. A esse tipo de técnicas não paramétricas de regressão correspondem as funções Kernel (Priestley & Chao, 1972), os polinômios localmente ponderados (mínimos quadrados locais, polinômios móveis) (Cleveland 1979,1988) e as funções de influência radial (Powell, 1987). Estas metodologias são conhecidas como técnicas de interpolação e suavização, porque aperfeiçoam o ajuste atribuindo diferentes pesos aos dados que coexistem na vizinhança. A ideia consiste em ponderar com maior valor aquelas observações mais próximas no momento de elaborar o prognóstico e com menor valor aquelas mais distantes.

Os polinômios localmente ponderados (também chamados regressão polinomial móvel) ou LWP (Locally Weighted Polynomials) foram desenvolvidos por Cleveland (1979) como método de interpolação e logo melhorado pelo autor (Cleveland, 1988) mediante a implementação das aproximações polinomiais locais. Alguns exemplos de aplicação são apresentados por Schmerling e Peil (1985). A aproximação LWP elabora-se mediante um ajuste pontual de polinômios de baixa ordem em subconjuntos localizados de dados. A ideia básica é usar um polinômio ordinário para a regressão local, selecionando intervalos $[x-b, x+b]$ com largura de banda b, onde para cada intervalo estima-se uma função $\hat{y}(x)$. Os coeficientes do polinômio ajustado são determinados mediante mínimos quadrados ponderados, dando um peso maior na ponderação aos pontos de dados mais próximos e um menor peso aos mais distantes dentro do intervalo $[x-b, x+b]$.

Dado um modelo localizado, por exemplo, um polinômio de primeiro grau:

$$\hat{y}(x) = \beta_0 + \beta_1 x \qquad [1]$$

Os coeficientes β frequentemente são estimados minimizando a expressão:

$$\beta = \min \sum_{k=1}^{N} (y_k - \hat{y}(x_k))^2 = \min \sum_{k=1}^{N} (y_k - \beta_0 + \beta_1 x_k)^2 \qquad [2]$$

Para o ajuste de uma regressão linear localmente ponderada deve subministrar-se um ponto de referência x_{query} com base ao qual construísse um novo ajuste linear que está mais influenciado pelos pontos na vizinhança de x_{query}, segundo a distância euclidiana estre os pontos. A ponderação local é obtida mediante a ponderação de cada ponto na vizinhança de x_{query} em função da sua distância euclidiana, assim um ponto que esteja distante de x_{query} terá uma ponderação nula e um ponto próximo terá uma ponderação alta. Os parâmetros da regressão podem ser estimados como:

$$\beta = \min \sum_{k=1}^{N} w(x_{query}, x_k)(y_k - \hat{y}(x_k))^2 = \min \sum_{k=1}^{N} w(x_{query}, x_k)(y_k - \beta_0 + \beta_1 x_k)^2 \qquad [3]$$

No lado direito da Figura 1 destaca-se o efeito da ponderação, próximo do ponto de referência (marcado com um X) os resíduos maiores são fortemente penalizados, enquanto aqueles distantes a ponderação é desprezível. Se o ponto de referência muda de posição, os pesos sobre os dados podem mudar e gerar um novo ajuste linear, por isso o método também recebe o nome de mínimos quadrados móveis.

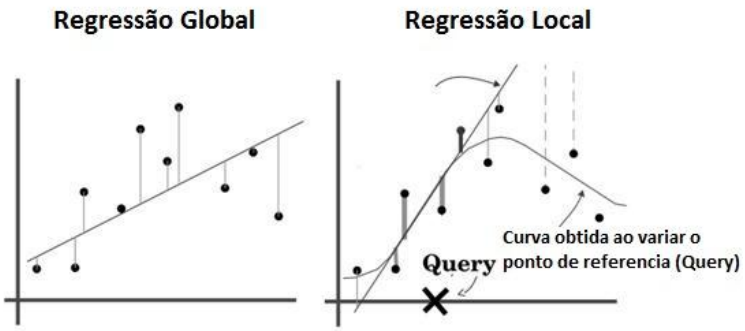

Figura 1.- Comparação entre uma regressão global e uma regressão local.

A função de ponderação w depende da distância euclidiana entre o ponto de referência x_{query} e as observações na vizinhança x, uma das funções de ponderação mais utilizadas é a função Gaussiana, dada por:

$$w(x_{query}, x_i) = \exp(-\alpha\mu_i^2) \qquad [4]$$

Sendo

$$\mu = \|x_{query}, x_{farthest}\| \qquad [5]$$

Onde $\|\ \|$ é o operador da distância euclidiana, $x_{farthest}$ é o ponto mais distante da vizinhança; cada aproximação local é controlada mudando o valor do coeficiente α e o melhor valor de dito parâmetro obtém-se automaticamente pela validação cruzada. Em termos gerais uma regressão polinomial local, construída na vizinhança de um ponto de referência x_{query}, é expressa por:

$$\hat{y}(x) = \beta_1 t_1(x) + \beta_2 t_2(x) + ... + \beta_M t_M(x) \qquad [6]$$

Onde $t_j(x)$ é uma função que gera o j-ésimo termo do polinômio de regressão, por exemplo, para um polinômio biquadrático com dados de entrada (x_1, x_2) se tem: $t_1(x) = 1$, $t_2(x) = x_1$, $t_3(x) = x_2$, $t_4(x) = x_1^2$, $t_5(x) = x_2^2$, $t_6(x) = x_1 x_2$. As equações 4-5 podem ser reescritas de uma forma compacta como:

$$\hat{y}(x) = \beta^T t(x) \qquad [7]$$

Onde $t(x)$ é o vetor dos termos do polinômio $t(x) = [t_1(x), t_2(x), ..., t_M(x)]$. A ponderação para o k-ésimo dado é computada como uma função decrescente da distância euclidiana entre x_k e x_{query}. Os valores dos coeficientes β são estimados minimizando a expressão:

$$\sum_{k=1}^{N} w_k \left(y_k - \beta^T t_k(x)\right)^2 \qquad [8]$$

Com $w_k = w(x_{query}, x_k)$, usando mínimos quadrados:

$$\beta = \left(X^T X\right)^{-1} X^T y \qquad [9]$$

Sendo $\left(X^T X\right)$ uma matriz de longitude $M \times M$ e $X^T y$ uma matriz de $M \times 1$, onde:

$$\left(X^T X\right)_{i,j} = \sum_{k=1}^{N} w_k t_i(x_k) t_j(x_k) \qquad [10]$$

$$\left(X^T y\right)_i = \sum_{k=1}^{N} w_k t_i(x_k) y_i \qquad [11]$$

CALIBRAÇÃO E VALIDAÇÃO DO MODELO

Pré-processamento dos dados

Para a operação eficiente do modelo proposto é necessário um tratamento prévio da informação que inclui um processo de normalização e escalonamento dos dados. Um primeiro passo consiste em remover da serie original o coeficiente de assimetria com o objeto de obter uma distribuição de probabilidades que esteja centrada na média, para tal efeito aplica-se a seguinte transformação:

$$X_{v\tau} = \log\left(Q_{v\tau} - c_\tau \overline{Q}_\tau\right) \qquad [12]$$

Onde

$$c_\tau = a \Big/ g_\tau^2 \qquad [13]$$

Sendo $Q_{v\tau}$ a vazão média para o mês τ ($\tau = 1,...,12$) e o ano v ($v = 1,..., N_a$) com N_a o número de anos da serie; $\overline{Q}_\tau$ a vazão média mensal do mês τ e a é um parâmetro adimensional cujo valor é de 0.35, o qual resulta de uma analise de regressão entre g_τ e c_τ, g_τ é o coeficiente de assimetria para o conjunto $Q_{1\tau}, Q_{s\tau},..., Q_{Na\tau}$; e $X_{v\tau}$ é a serie normalizada das vazões para o ano v e o mês τ. Para eliminar a periodicidade, a serie de vazões é padronizada mediante a seguinte equação:

$$Y_{v\tau} = \frac{Q_{v\tau} - \overline{Q}_\tau}{s_{Q\tau}} \qquad [14]$$

Onde $\overline{Q}_\tau$ e $s_{Q\tau}$ são a média e o desvio padrão da serie de vazões para o mês τ, e $Y_{v\tau}$ são os valores padronizados para o ano v e o mês τ. Finalmente, uma transformação adicional é aplicada à serie $Q_{v\tau}$ para levar os dados a uma escala adequada que permita o processamento com algum dos métodos de ajuste propostos. O intervalo será reduzido a valores entre 0 e 1 usando a seguinte transformação:

$$Z_t = \frac{Q_t - Q_m}{Q_M - Q_m} \qquad [15]$$

Sendo Q_t o valor da vazão com $t = 12(v-1) + \tau$; Q_M é o máximo valo da serie $Q_{v\tau}$; Q_m é o valor mínimo da serie.

Ajuste dos parâmetros no espaço de componentes principais

A análise clássica de ACP é usada com múltiplas series de tempo a fim de obter as direções principais de uma sequencia de vetores M-dimensionais $\left(X_i, 1 \le i \le N\right)$, a expandi-los com respeito a uma base ortogonal $\left(E^k, 1 \le k \le M\right)$.

$$X_{i,j} = \sum_{k=1}^{cp} a^k{}_i E_j^{\,k} \quad 1 \le j \le M \qquad [15]$$

Os coeficientes de projeção, $a^k{}_i$, são chamados de Componentes Principais (CP) e os vetores, $E^k{}_i$, são as funções ortogonais empíricas (FOE). Para a implementação da análise de componentes principais se tem uma expansão da forma:

$$X_{i+j} = \sum_{k=1} a_i^{\,k} E_j^{\,k} \quad 1 \le i \le n;\ 0 \le j \le M-1 \qquad [16]$$

As FOE são os autovetores da matriz Toepletiz, T_x, que contem os coeficientes de covariação cruzada dos diferentes vetores para retardos de 0 a *M-1*. As anteriores equações resultam da aplicação da expansão biortogonal de Karhunen-Loeve, muito usada no processamento de sinais

digitais (Ghil et al., 2002). A ortogonalidade no tempo (covariação cruzada igual a zero para dois CP no retardo zero) e espaço (ortogonalidade das FOE) implica que λ_k (autovalor k da matriz da matriz de Toeplitz) representa a variância da k -ésima CP.

Função periódica de regressão

Propõe-se construir funções de regressão especializadas no prognóstico de cada mês com o objeto de capturar com maior fidelidade a variabilidade das vazões, expressa da seguinte forma:

$$Q_{t,\tau} = \hat{y}(X,\tau) = \begin{cases} \hat{y}_1(X_{1,1}, X_{2,1}, \cdots, X_{n,1}) \Rightarrow \tau = 1 \\ \hat{y}_2(X_{1,2}, X_{2,2}, \cdots, X_{n,2}) \Rightarrow \tau = 2 \\ \vdots \\ \hat{y}_\tau(X_{1,\tau}, X_{2,\tau}, \cdots, X_{n,\tau}) \\ \vdots \\ \hat{y}_{12}(X_{1,12}, X_{2,12}, \cdots, X_{n,12}) \Rightarrow \tau = 12 \end{cases} \quad [17]$$

Onde $\hat{y}_\tau$ corresponde ao polinômio ponderado ajustado para o mês τ, e os $X_{n,\tau}$ são as n variáveis explicativas definidas mediante análise de correlação cruzada entre as series de vazões e os valores antecedentes (retardos) das variáveis explicativas (vazões, chuva e variáveis macro-climáticas). Propõe-se utilizar a Análise espacial de Componentes Principais (ACP) para reduzir a dimensionalidade e ajustar o modelo periódico às sete estações mediante a aplicação de funções ortogonais empíricas (FOE). Com o intuito de selecionar aquelas variáveis explicativas que maximizam a possibilidade de ajuste extraem-se dois conjuntos de dados da informação histórica, um para calibrar os parâmetros e outro para validar os resultados.

Validação dos resultados

Se os indicadores estatísticos de validação satisfazem as expectativas de quem aplica o modelo as previsões são aceitas, caso contrario devem-se mudar os parâmetros do modelo selecionado ou as variáveis explicativas. A precisão dos prognósticos para o período de validação pode ser estimada mediante diferentes indicadores de erro, apresentados na literatura de series de tempo, cujo cálculo basea-se na seguinte notação: y_t denota a observação histórica no tempo t e $\hat{y}_t$ corresponde ao valor esperado da predição para o período t, onde $t = 1,2,...,T$. Define-se então o erro (ou resíduo) da predição como $e_t = y_t - \hat{y}_t$, e o erro percentual como $p_t = 100 \times e_t / y_t$. Com base na notação anterior definem-se as seguintes medidas do erro:

Somatória de erros ao quadrado: $SSE = \sum_{t=1}^{T} e_t^2$

Erro quadrático médio ou variância do predictor: $MSE = \frac{SSE}{T}$

A raiz do erro quadrático médio: $RMSE = \sqrt{MSE}$

Se y corresponde ao valor médio da serie de dados, então o erro quadrático médio expressado como porcentagem do valor médio dos dados é expresso por:

$$\%RMSE = \frac{RMSE}{y} \times 100 \quad [18]$$

E o valor médio do erro percentual absoluto é dado por:

$$MAPE = \frac{1}{T}\sum_{i=1}^{T}\left|p_t\right| \qquad [19]$$

Outro método utilizado para estimar a possibilidade de ajuste de um modelo consiste em estimar o coeficiente de determinação ou de correlação de Pearson entre os valores previstos e os valores históricos durante o período de validação. O procedimento geral para a previsão de vazões, usando o método proposto é esquematizado na Figura 2.

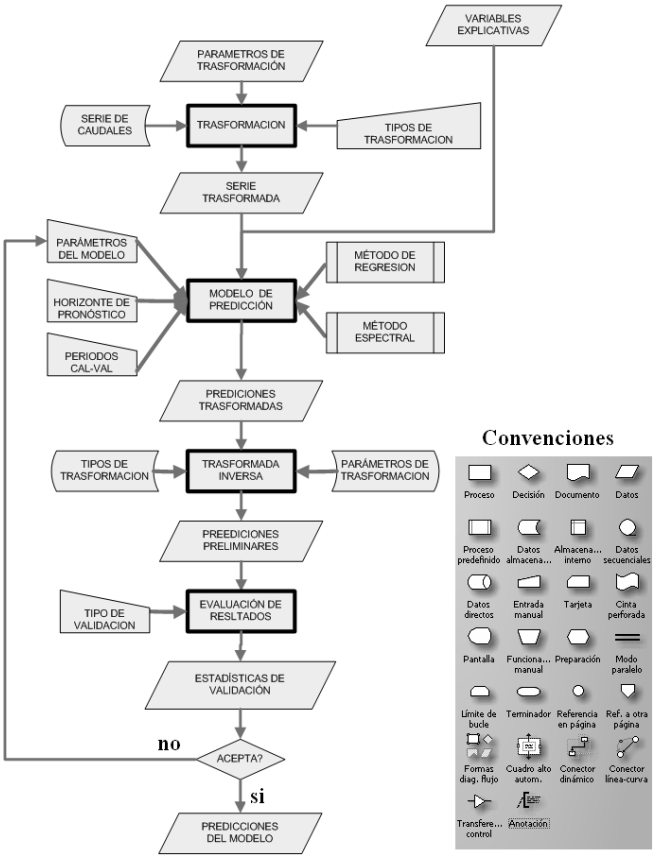

Figura 2.- Metodologia geral para a previsão de vazões.

APLICAÇÃO AO RIO TOCANTINS

A região hidrográfica do rio Tocantins – Araguaia possui um área de drenagem de 918.822 km², cobre aproximadamente o 11% da área total do Brasil e é a segunda bacia em importância do país, despois da bacia do rio Amazonas (3.869.953 km^2); nesta estão localizados alguns dos projetos hidroelétricos mais estratégicos do Brasil, entre eles estão a usina hidroelétrica de Serra da Mesa em Goiás (1.275 Mw), a usina hidroelétrica de Luiz Eduardo Magalhães no Tocantins (902,5 Mw) e a usina hidroelétrica de Tucuruí no Pará (8.370 MW). A Figura 3 apresenta a localização das estações de vazões utilizadas no presente trabalho com seu respectivo código. A informação hidrológica dessas estações foi subministrada pela Agência Nacional de Águas (ANA) do Brasil. As series de vazões mensais contam com dados desde janeiro de 1980 até dezembro de 2005 (312 dados). Além disso, conta-se com os campos de precipitação a escala mensal gerado pelo Centro de Previsão do Tempo e Estudos Climáticos (CPTEC) do INPE e foi recopilada uma quantidade significativa de informação macro-climática proveniente da Reanalise NCEP/NCAR e a Reanalise II de NCEP-DOE.

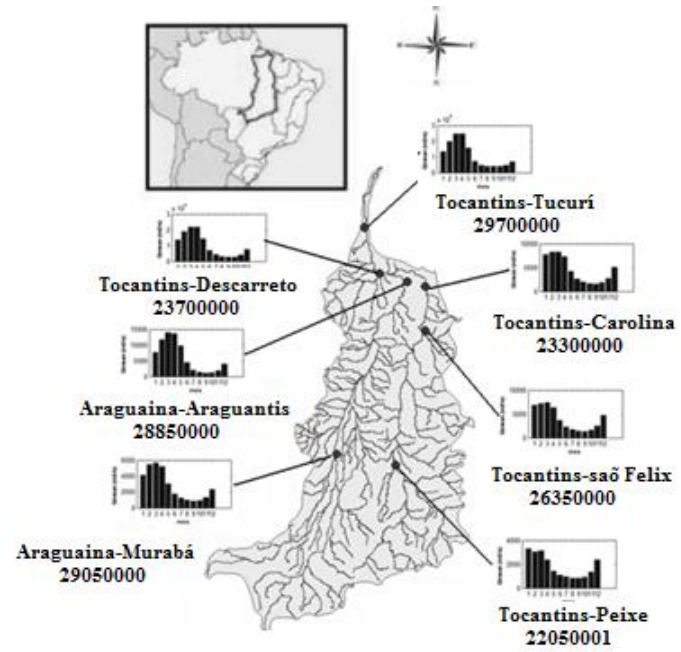

Figura 3.- Estações de vazões utilizadas no presente trabalho.

O rio Tocantins e seus tributários encontram-se localizados na região do nordeste do Brasil, cuja variabilidade está relacionada com as mudanças das temperaturas superficiais do mar nas regiões tropicais do oceano Pacífico (ENSO) e Atlântico (Atlântico tropical norte e Atlântico tropical Sul). Alguns autores consideram que a relação entre o sistema ENSO e o clima do nordeste brasileiro não é direta, pois se processa via oceano Atlântico Tropical, em particular no setor sul (Hastenrath, 1978). Note-se na Figura 4 que as máximas correlações (superiores a 0.4) se apresentam no oceano Atlântico tropical sul. Estes autores afirmam que, em parte, as anomalias climáticas do nordeste brasileiro, podem ser relacionadas com as variações inversas da pressão ao nível do mar entre o Pacífico tropical leste e o Atlântico tropical, e tais variações fazem parte do acomodamento das massas de grande escala associadas ao sistema ENSO. De modo consistente com essa hipótese Saravanan e Chang (2000) propuseram que as teleconexões do ENSO tem um papel importante na variabilidade climática do Atlântico tropical, afetando o clima do nordeste brasileiro e em consequência as vazões dos rios Tocantins-Araguaia. Em 1981 Moura e Shukla propuseram um mecanismo dinâmico para explicar as secas no nordeste brasileiro em função das anomalias de temperaturas sobre a superfície do oceano Atlântico tropical, a ocorrência simultânea de anomalias positivas (ou negativas) no ATN e negativas (ou positivas) no ATS, é o que tem sido denominado como “modo dipolo”. Assim, as TSM da região do Pacífico Tropical leste (de frente ao litoral colombiano) e as TSM no Atlântico Tropical (o gradiente de temperaturas nessa região) são indicadores da variabilidade das vazões na região Nordeste do Brasil.

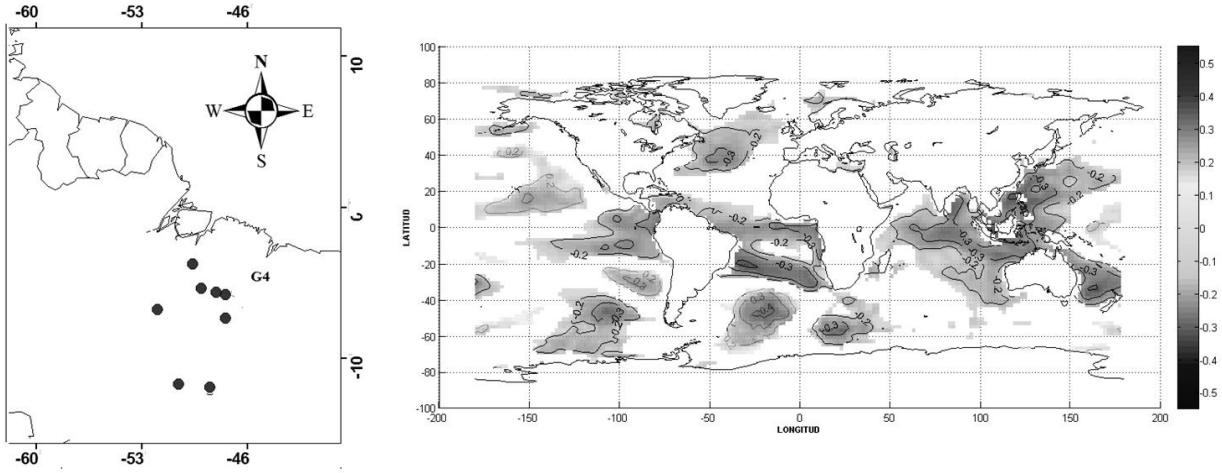

Figura 4.- Análises de correlação entre a primeira componente principal das vazões meias mensais e as temperaturas da superfície do mar

Os dados de vazões e chuva antecedente (com retardo de até três meses), e as temperaturas da superfície do mar (SST) no Pacífico Tropical e o Atlântico Tropical (ATN, ATS) (com retardos de um mês) foram selecionados como variáveis explicativas mediante uma analise de correlação (Figura 5). Os modelos de prognóstico são calibrados e validados para a previsão de vazões médias mensais nas diferentes estações do rio Tocantins-Araguaia para horizontes de prognóstico de 1, 3 e 6 meses. O período de calibração compreende janeiro de 1980 até dezembro de 1999 e o de validação desde janeiro de 2000 até dezembro de 2005. Foi utilizada uma analise de componentes principais para reduzir a dimensionalidade e predizer de forma simultânea todas as series de tempo. Os resultados para o prognóstico da estação Tucurí, com horizonte de prognóstico de um mês é apresentado na Figura 5 e na Figura 6 a distribuição do erro MAPE para todas as estações nos diferentes horizontes de prognóstico.

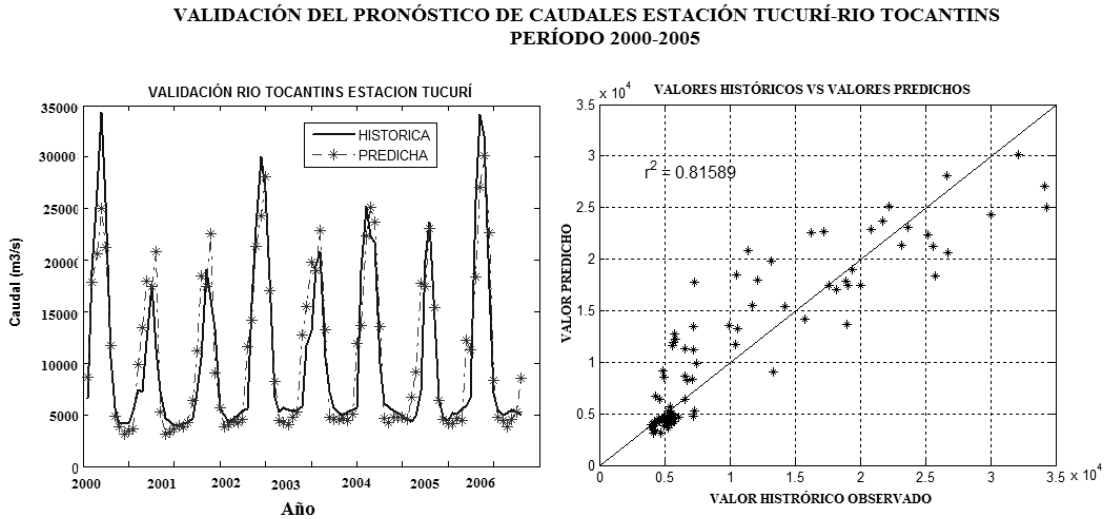

Figura 5.- Resultados da previsão para o rio Tocantins estação Tucurí para um horizonte de um mês

Os erros de validação para todas as estações com diferentes horizontes de prognóstico não excedem o 35% a exceção da estação Peixe, onde são superiores ao 47%, grande parte do erro pode ser explicado pela subestimação das vazões mínimas durante a época seca.

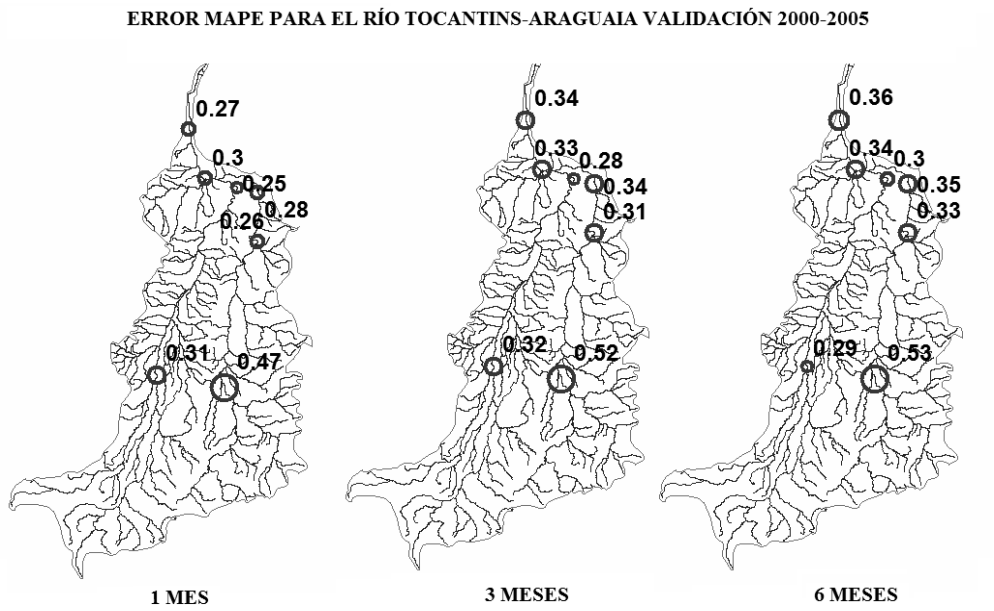

Figura 6.- Erro médio percentual (MAPE) para as diferentes estações fluviométricas no rio Tocantins.

CONCLUSÃO

A metodologia aqui exposta tem a vantagem de ser não-paramétrica tal que o cálculo dos parâmetros de calibração não dependem da distribuição probabilística dos dados, o que converte aos polinômios localmente ponderados em excelentes candidatos para simular processos não lineares e não estacionários como as vazões. Além disso, considerou-se o efeito das mudanças climáticas nas series de vazões em longo prazo. Embora a aproximação não linear e não paramétrica, fornecida pelos mínimos quadrados móveis não seja perfeita, consegue-se representar as principais características das series de vazões para as diferentes estações fluviométricas. Os prognósticos obtidos tiveram uma boa aproximação aos valores observados de vazão na bacia do rio Tocantins-Araguaia com o uso das temperaturas da superfície do mar, a precipitação e os valores antecedentes de vazões como variáveis explicativas.

AGRADECIMENTOS

Os autores agradecem ao projeto: "E*stágios de pesquisas entre o Instituto Interamericano para a Pesquisa em Mudanças Globais (IAI) e o Centro de Previsão do Tempo e Estudos Climáticos (CPTEC) do Instituto Nacional de Pesquisas Espaciais (INPE) MCT"* pela formação recebida e o apoio necessário para o desenvolvimento do presente trabalho.

REFERÊNCIAS

Cleveland, W. S. (1979). "Robust locally weighted regression and smoothing scatterplots". *Journal of American Statistical Association*, 74 (368), 829-836.

Cleveland, W. S. and Devlin, S. J. (1988). "Locally weighted regression: an approach to regression analysis by local fitting". *Journal of American Statistical Association*., 83 (403), 596-610.

Ghil M., Allen M., Dettinger M., Ide M., Kondrashov D., Mann M., Robertson A., Saunders A., Tian Y. y Varadi F. y Yiou P. (2002). "Advanced Spectral Methods for Climatic Time Series". *Review of Geophysics. American Geophysical Union,* Vol. 40, N° 1. pp. 1-41.

Hastenrath, S. (1978). "On modes of tropical circulation and climate Anomalies". *J. Atmos.* Sci., 35, 2222-2231.

Priestley, M. B. y Chao, M. T. (1972): "Non-parametric function fitting". *J. Royal Stat. Soc*, B, 34, 385-392.

Poveda G. (2006). "Aplicación de los métodos MARS, Holt-Winters y ARIMA generalizado en el pronóstico de caudales medios mensuales en ríos de Antioquia". *Rv Colombia Meteorologia Colombiana*, ISSN: 0124-6984 ed: Gente Nueva v.10 fasc.1 p.36 - 46,

Powell M.J.D. (1987). " Radial basis functions for multivariable interpolation: a review, Algorithms for Approximation", Mason J.C., Cox M.G. (eds.), London, Oxford University Press.

Queipo NV, Haftka RT, Shyy W, Goel T, Vaidyanathan R, Tucker PK (2005). "Surrogate-based analysis and optimization". *Prog Aerosp Sci* 41:1–28

Moura, A. D., and J. Shukla, (1981): On the dynamics of droughts in northeast Brazil: Observations, theory and numerical experiments with a general circulation model. *J. Atmos*. Sci., 38, 2653-2675.

Saravanan, R., and P. Chang, (2000). "Interaction between tropical Atlantic variability and El Niño–Southern Oscillation. *J. Climate*,13, 2177–2194.

Schmerling, S. & Peil, J. (1986): "Improvement of the method of kernel estimation by local polynomial approximation of the empirical distribution function and this application to empirical regression". *egenbaurs morphologisches Jahrbuch*, 132,29-35.