



AJUSTE DE UM MODELO DE MISTURA BI-GAUSSIANO PELO ALGORITMO EM COMBINADO AO MÉTODO *BOOTSTRAP* PARA A EVAPOTRANSPIRAÇÃO POTENCIAL NA CIDADE E MARINGÁ – PR ENTRE OS ANOS DE 2003 E 2014

Marcos Jardel Henriques (PIBIC/Uem), Oilson Alberto Gonzatto Junior (PIBIC/Uem), Hellen Geremias dos Santos (PIC/Uem), Rosângela Getirana Santana (Orientadora), Isolde Terezinha Santos Previdelli (Coorientadora)
e-mail: rgsantana@uem.com

Universidade Estadual de Maringá/Centro de Ciências Exatas/Maringá, PR.

Ciências Exatas e da Terra – Probabilidade e Estatística

Palavras-chave: distribuição de probabilidade, métodos de estimação, distribuição amostral via bootstrap.

Resumo:

Os métodos estatísticos, de modo geral, pressupõem que as amostras coletadas em um experimento seguem uma mesma distribuição de probabilidade, no entanto, existem experimentos cujas informações amostrais são insuficientes para que alguma distribuição de probabilidade teórica conhecida se ajuste aos dados, de modo que não é possível identificar um único padrão para todas as observações. Situações como estas podem ser resolvidas com o uso do que é conhecido como modelo de mistura finita. Tais modelos usam uma mistura entre distribuições paramétricas conhecidas a fim de obter um modelo adequado aos dados observados. O presente trabalho tem como objetivo a obtenção das estimativas pontuais e intervalares para um modelo como esse utilizando o algoritmo de Maximização da Esperança (*Expectation Maximization* – EM) combinado ao método *bootstrap*. Ambos os métodos foram implementados por meio da PROC IML disponível no *software* SAS. Como uma aplicação dessa teoria, a distribuição de frequências da evapotranspiração potencial na cidade de Maringá – PR foi analisada com base em observações de alguns meses entre os anos de 2003 a 2014, a fim de propor um modelo bi-gaussiano que descreva seu comportamento.

Introdução

Um modelo de mistura finita é uma alternativa paramétrica que descreve uma distribuição desconhecida em termos de misturas de distribuições



conhecidas, cujas viabilidades analíticas são comprovadas. Essa distribuição é expressa como uma soma ponderada de distribuições que representam, individualmente, subgrupos de observações (MCLACHLAN; PEEL, 2000).

Na modelagem estatística, uma mistura finita é capaz de prover suficiente flexibilidade para descrever uma grande variedade de fenômenos aleatórios e, devido a essa utilitária característica os métodos de modelagem que a envolve têm recebido atenção por anos.

O algoritmo de Maximização da Esperança (EM) é um algoritmo de otimização com forte motivação estatística e é responsável pelo cálculo iterativo dos estimadores de máxima verossimilhança (PAWITAN, 2001), ele simplifica o ajuste dos modelos de mistura finita e tem um número considerável de propriedades, como a estabilidade numérica, implementação simples e convergência global confiável (MCLACHLAN; NG, 2009). Uma desvantagem é o fato de não fornecer, de modo imediato, o erro padrão das estimativas, essa desvantagem pode ser contornada pelo uso alternativo de uma simulação *bootstrap* (MILLAR, 2011).

Para aplicar essa teoria, esse trabalho tem como objetivo ajustar um modelo bi-gaussiano à distribuição de frequências da evapotranspiração (transferência de água para a atmosfera via evaporação dos ambientes e transpiração das plantas) potencial na cidade e Maringá – PR com base em observações coletadas no site do Instituto Nacional de Meteorologia (INMET) em alguns meses no período de 2003 a 2014. Os dados foram aferidos em milímetros pela estação climatológica OMM: 83767 com a seguinte localização: latitude: -23.4°, longitude: -51.91° e altitude: 542.00m.

Materiais e métodos

A função densidade de probabilidade de uma variável aleatória Y cuja distribuição é uma mistura de duas densidades gaussianas é expressa por $f(y; \mu, \sigma, \nu, \tau) = \frac{p}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right\} + \frac{1-p}{\sqrt{2\pi}\tau} \exp\left\{-\frac{1}{2}\left(\frac{y-\nu}{\tau}\right)^2\right\}$, nesta função densidade de probabilidade os parâmetros podem ser estimados pela maximização da função de verossimilhança. O algoritmo EM será utilizado nesse estudo como descrito por McLachlan (2009) e Chang (2007). O processo de estimação pontual realizado pelo algoritmo EM foi combinado ao método *bootstrap*, como sugerido por Millar (2011), para a obtenção dos erros padrão e das estimativas intervalares para os parâmetros. Para o método *bootstrap*, a estimação pontual por meio do algoritmo EM foi repetida 10000 vezes. Ambos os processos foram implementados na PROC IML do software SAS.



Resultados e Discussão

O histograma apresentado na Figura 1 permite entender que esses dados podem ser modelados por uma distribuição de frequência bimodal, isso sugere que um modelo de mistura pode se adequar a esse problema.

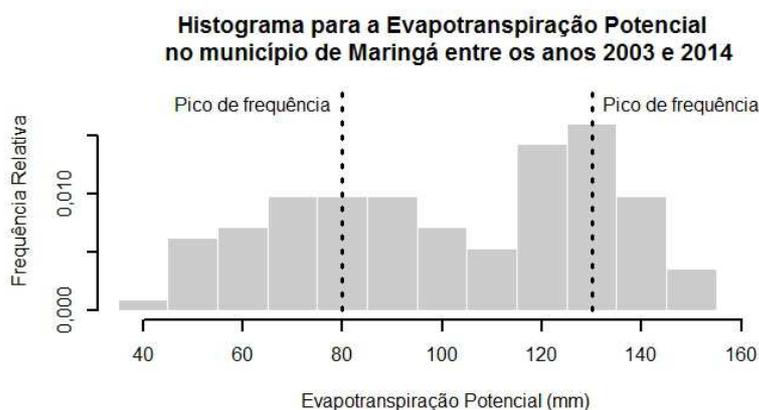


Figura 1 – Histograma para a Evapotranspiração Potencial no município de Maringá entre os anos de 2003 e 2014.

O processo *bootstrap* forneceu um vetor com 10000 estimativas para cada um dos parâmetros p , μ , σ , ν e τ , de onde foram obtidos, a estimativa, o erro-padrão e os percentis das distribuições amostrais empíricas para a obtenção de um intervalo com 95% de confiança, Tabela 1.

Tabela 1 – Estimativas dos parâmetros, erros padrões e intervalos de confiança obtidos via *bootstrap*.

<i>Parâmetro</i>	<i>Estimativa</i>	<i>Erro Padrão</i>	<i>Intervalo com 95% de confiança</i>
p	0,5440094	0,0652849	(0,4076974 0,6684229)
μ	77,977158	3,8909431	(70,136123 85,548466)
σ	18,031071	2,2503439	(13,420912 22,446244)
ν	129,12552	2,5168029	(123,34466 133,29597)
τ	10,506852	1,7194453	(7,7740614 14,446122)

Note que, expostos os resultados, é possível afirmar que a distribuição de frequência da evapotranspiração potencial do município de Maringá – PR pode ser descrita pela soma de duas distribuições gaussianas com proporções iguais, ou muito próximas ($p \approx 0,544$ e $1 - p \approx 0,456$), mas com médias e variâncias expressivamente distintas. As implicações de um comportamento desses devem ser estudadas com atenção, a fim de estabelecer critérios adequados de cuidado e manejo dos recursos hídricos.



Conclusões

O modelo de mistura bi-gaussiana se mostrou adequado para descrever o fenômeno da evapotranspiração potencial na cidade de Maringá – PR e pôde ser utilizado para explicar a sua distribuição frequências. Foi possível identificar a variação latente inerente ao fenômeno, e observada na amostra, bem como avaliar as distintas variações e comportamentos esperados.

Agradecimentos

Os autores agradecem a bolsa de fomento cedida pela UEM.

Referências

CHANG, S. C.; KIM, H. J. **EM Algorithm**. 2007. Disponível em: <www.stat.uiowa.edu/~kcowles/s166_2007/chang166.pdf>. Acesso em: 19 jun. 2015.

MCLACHLAN, G.; PEEL, D. **Finite Mixture Models**. New York: John Wiley & Sons, Inc., 2000. 419 p.

MCLACHLAN, G. J.; NG, S. K. EM algorithm. In: WU, X.; KUMAR, V. (Org.). **The top ten algorithm in Data Mining**. New York: Taylor & Francis Group, 2009. p. 96-113.

MILLAR, R. B. **Maximum Likelihood Estimation and Inference: with examples in R, SAS, and ADMB**. New York: John Wiley & Sons, Inc., 2011. 357 p.

PAWITAN, Y. In **All Likelihood: Statistical Modelling and Inference Using Likelihood**. New York: Oxford University Press Inc., 2001. 528 p.