ANÁLISE E DISCRIMINAÇÃO DE DISTRIBUIÇÕES USADAS NA ANÁLISE DE VARIÁVEIS CLIMATOLÓGICAS

Larissa Bueno Fernandes (PIBIC/CNPq/FA/Uem), Josmar Mazucheli (Orientador), e-mail: lbf.estatistica@gmail.br.

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

Ciências Exatas e da Terra/ Probabilidade e Estatística.

Palavras-chave: climatologia, discriminação, PBCM

Resumo:

A escolha de uma distribuição de probabilidade para descrever observações aleatórias é uma etapa crucial na análise de dados. Usualmente, seleciona-se o modelo que apresenta uma melhor medida de qualidade de ajuste. Entretanto, tais medidas não levam em conta a complexidade dos modelos em sua forma funcional. O *parametric bootstrap cross-fitting method* (PBCM), proposto por Wagenmakers et al. (2004) quantifica o viés devido ao mimetismo dos modelos candidatos, ajudando a evitar tal problema. O método foi aplicado aos dados mensais de precipitação da estação meteorológica de Maringá-PR, objetivando a discriminação entre as distribuições Gama e Nakagami. Em geral, o modelo Gama apresentou maior capacidade de mimetismo. Por outro lado, a distribuição Nakagami se mostrou mais adequada para a maioria dos conjuntos de dados considerados, de acordo com os resultados do PBCM.

Introdução

Na análise de dados climatológicos, a escolha de um modelo para descrever um fenômeno em estudo é de fundamental importância, uma vez que esta decisão determina a estrutura do estudo e a ferramenta a ser usada: o modelo de probabilidade (Meylan et al., 2011).
Usualmente, seleciona-se o modelo que apresentar o ajuste mais próximo, indicado por alguma medida de qualidade de ajuste. Tais medidas, em geral, não levam em conta a complexidade dos modelos. Este método pode levar a

uma escolha enganosa, tendendo a selecionar modelos mais complexos, mesmo se um modelo mais simples gerou os dados (Pitt e Myung, 2002).
O *parametric bootstrap cross-fitting method* (PBCM), proposto por Wagenmakers et al. (2004) quantifica o viés devido ao mimetismo dos modelos candidatos, ajudando a evitar tal problema.

Materiais e métodos

De acordo com Wagenmakers et al. (2004), mimetismo é a capacidade de um modelo explicar dados que foram gerados por um modelo concorrente. Para quantificar essa característica, o procedimento PBCM gera uma distribuição de diferenças de GOF sob cada um dos modelos concorrentes. No artigo em que introduz o PBCM, Wagenmakers et al. (2004) propõem duas variações do método.
A primeira versão, denotada por DIPBCM, consiste nos seguintes passos:

1. Retirar uma amostra com repetição x^* dos dados observados x;
2. Ajustar ambos os modelos A e B a amostra x^*, obtendo os estimadores de máxima verossimilhança dos vetores de parâmetros $\hat{\theta}_A^*$ e $\hat{\theta}_B^*$;
3. Aplicar o *bootstrap* paramétrico para ambos os modelos, gerando dados simulados a partir do modelo A ($D(\hat{\theta}_A^*)$) e do modelo B ($D(\hat{\theta}_B^*)$);
4. Ajustar os modelos A e B a $D(\hat{\theta}_A^*)$, obtendo as estimativas dos parâmetros e a diferença de GOF, representada por $\Delta GOF_{AB}^*|A = GOF_A^* - GOF_B^*$. Do mesmo modo, ajustando ambos os modelos a $D(\hat{\theta}_B^*)$ e calcular a diferença: $\Delta GOF_{AB}^*|B = GOF_A^* - GOF_B^*$;
5. Repetir os passos anteriores M vezes.

Já a segunda variação, DUPBCM, difere da primeira na geração dos conjuntos de dados simulados. De acordo com Schultheis e Naidu (2014), para gerar parâmetros no DUPBCM, primeiro fixa-se um intervalo de valores possíveis para cada um dos parâmetros do modelo e uma distribuição de probabilidade entre cada um destes intervalos. Os valores para a geração de dados são então amostrados a partir dos intervalos de acordo com as distribuições associados.
Em resumo, como apontado por Wagenmakers et al. (2004), o DIPBCM é útil para avaliar a adequação e o mimetismo do modelo para um conjunto de dados específico. Quando for necessário fazer afirmações mais gerais sobre a plausibilidade dos tipos de modelo, o DUPBCM é mais apropriado.
Ambas versões geram duas distribuições de diferenças de GOF, uma distribuição sob o modelo A e outra sob o B. A probabilidade relativa de que

os dados são originários do modelo A, pode ser quantificado pela divisão das alturas estimadas das distribuições de diferença no valor observado δ_{AB}: $P(\delta_{AB} \mid A)/P(\delta_{AB} \mid B) = P_A(x)/P_B(x)$. O critério de decisão ótimo é dado por $P_A(x)/P_B(x) = 1$ (Wagenmakers et al., 2004). A diferença entre o critério nominal ($\Delta GOF_{AB} = 0$) e o critério ótimo é o viés do mimetismo β_m.

Para a aplicação da versão de dados informados do PBCM, o número M de iterações foi fixado em 1000 e foram comparadas as distribuições Gama e Nakagami. Os dados utilizados foram coletados do Banco de Dados Meteorológicos para Ensino e Pesquisa (BDMEP) e referem-se aos volumes mensais de precipitação (mm) da estação meteorológica de Maringá – Paraná no período de 1983 a 2016, divididos em 12 séries mensais. A medida de GOF utilizada para calcular as diferenças foi a de *Kolmogorov-Smirnov* (KS). As análises foram realizadas com o auxílio do ambiente estatístico R.

Resultados e Discussão

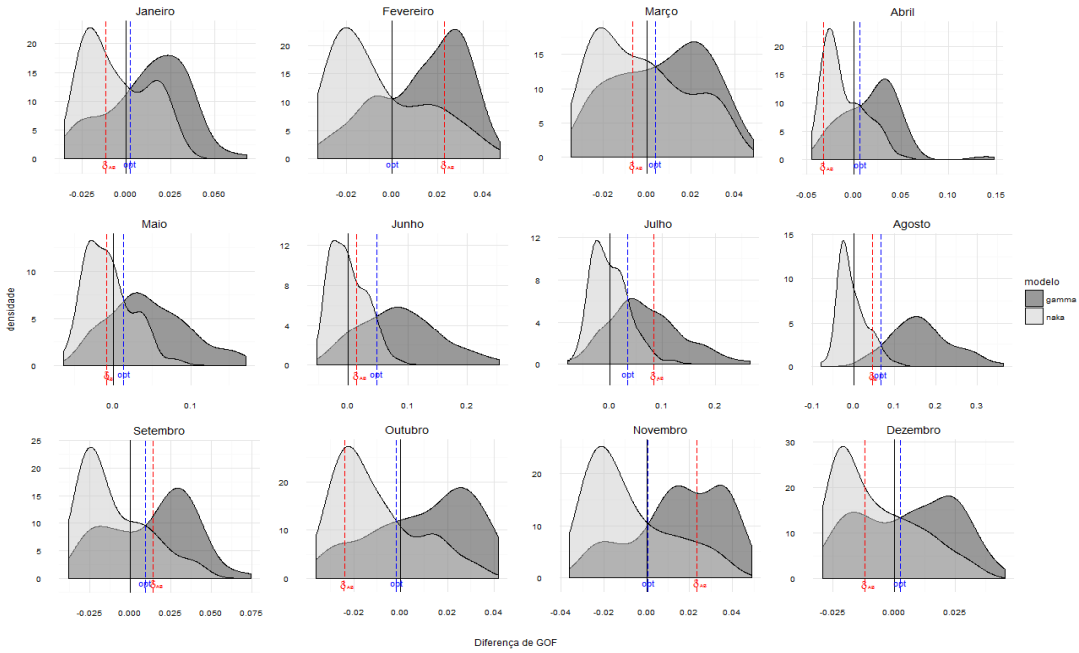

Figura 1 - Distribuições das diferenças de GOF obtidas pela aplicação do DIPBCM aos dados mensais de precipitação na estação de Maringá.

A Figura 1 apresenta os resultados da aplicação do DIPBCM para os dados de precipitação, divididos por mês. A área sobreposta entre as duas

distribuições de diferença de GOF, separadas pelo critério ótimo, indicam o mimetismo de cada modelo. Foi observado que para 9 dos 12 conjuntos de dados observados, o modelo Gama se mostrou mais flexível no ajuste de dados gerados pela distribuição Nakagami, com maior capacidade de mimetismo. Ainda, nota-se que para a maioria dos conjuntos de dados (8 dos 12 meses), o PBCM indicou que a distribuição Nakagami caracteriza-se como a distribuição mais provável de geração dos dados, sendo que para os meses de Junho e Julho, a conclusão foi contrária a indicada pela análise da estatística KS isoladamente.

Conclusões

Apesar da distribuição Nakagami ser popular em algumas áreas, ainda são poucas as aplicações envolvendo variáveis climatológicas. Neste trabalho, os resultados obtidos neste trabalho, utilizando dados reais de precipitação apontaram um melhor ajuste da distribuição Nakagami em relação a Gama. Além disso, foi observado que a distribuição Gama se mostrou mais complexa que a Nakagami, para os conjuntos de dados específicos, apresentando na maioria dos casos maior flexibilidade para explicar os dados gerados sob o modelo concorrente. Em alguns casos, o viés devido ao mimetismo influenciou a seleção do modelo, indicando a distribuição contrária a apontada pelo critério de GOF utilizado, a estatística KS.

Agradecimentos

Agradecemos ao CNPq-FA-UEM, por ter possibilitado e financiado este projeto.

Referências

Meylan, P. and Favre, A.C. and Musy, A. **Predictive Hydrology:** A Frequency Analysis Approach. 1.ed. CRC Press, 2011.
Pitt, M. A.; Myung, J. When a good fit can be bad. **Trends in Cognitive Sciences**, 6, 421-425, 2002.
Schultheis, H.; Naidu, P. Multi-Model Comparison Using the Cross-Fitting Method. **COGSCI**, 1389-1394, 2014.
Wagenmakers, E.; Ratcliff, R.; Gomez, P.; Iversonc, G. Assessing model mimicry using the parametric bootstrap. **Journal of Mathematical Psychology**, 48, 28-50, 2004.