Estudo comparativo dos algoritmos de normalização por combinação de histogramas e de dois passos

André Luiz Nogueira
Instituto Federal de Sergipe, IFS. Rod. Lourival Batista, s/n, CEP 49400-000, Aracaju, SE.
andreln27@yahoo.com

Casimiro S. Munita
Instituto de Pesquisas Energéticas e Nucleares, IPEN-CNEN/SP.
Av. Prof. Lineu Prestes, 2242, CEP 05508-000, São Paulo, SP.
camunita@ipen.br

Resumo

O uso de técnicas multivariadas em resultados experimentais representa uma crescente responsabilidade sobre o pesquisador, no sentido de entender, avaliar e interpretar os resultados, em especial os mais complexos. Um auxílio nessas tarefas é a compreensão das características básicas dos dados e de suas relações. O ponto de partida na interpretação dos resultados é caracterizar a forma de sua distribuição. Neste trabalho foram estudados dois algoritmos estatísticos: combinação de histogramas e de dois passos. Os estudos foram realizados usando uma base de dados de 161 amostras nas quais foram determinadas as frações de massa de K, La, Lu, Na, U, Yb, CE, Cr, Cs, Eu, Fe, Hf, SC, Tb e Th usando o método de análise por ativação com nêutrons. Os resultados foram comparados usando gráficos de histogramas, de probabilidade normal e o Teste de Shapiro-Wilks. Os resultados mostraram que o algoritmo de dois passos é superior.

Introdução

Tendo em vista o crescente avanço das técnicas físico-químicas a quantidade de dados gerados (resultados) tem aumentado significativamente. Para a interpretação desses resultados, faz-se necessário o uso de métodos estatísticos cada vez mais sofisticados, tais como as técnicas multivariadas. Em geral, os métodos estatísticos multivariados permitem avaliar um conjunto de amostras, levando em consideração as correlações existentes entre as variáveis. Estas técnicas, de uma forma geral, consideram que cada amostra analisada pode ser representada como um ponto no espaço multidimensional, onde cada dimensão do hiper-espaço corresponde a eixos determinados pela composição físico-química das amostras. Com o objetivo de agrupar as amostras, conforme sua similaridade/dissimilaridade, se devem formar grupos de amostras de acordo com alguns critérios estatísticos. Os resultados podem ser organizados dentro de uma matriz de dados X_{np}, sendo n o número de amostras variando de 1 até n e p o número de variáveis (elementos químicos) variando de 1 até p (Mucha & Bartel, 2015). Inicialmente é necessário verificar se os resultados atendem as condições de uma distribuição normal. Em casos em que todas as variáveis exibem a normalidade univariada ajudam a obter normalidade multivariada. Para normalizar esses dados aplicam-se algumas transformações matemáticas como*: log, x^2, 1/x*, e outras (Hair et al.).

Nesse trabalho foram estudados dois métodos de normalização: combinação de histogramas e de dois passos.

Metodologia

Normalização baseada em combinação de histogramas

A combinação de histogramas tem por objetivo definir uma transformação *HM()* que transforme o score *s* em *HM(s)*, de forma que o histograma de *HM(s)* tenha a forma desejada (Fu & Qiu, 2016). Supondo que se quer transformar um score $\{s_1,s_2,...,s_n\}$ com distribuição uniforme em um score $\{s_1',s_2',...,s_n'\}$, com distribuição normal. Seja $H=\{h_v, v=1,2,...,M\}$ o histograma com distribuição uniforme e $H'=\{h_v', v'=1,2,...,M\}$, o histograma com distribuição normal. A função distribuição acumulativa normalizada desses dois histogramas são $F=\{f_v,v=1,2,...,M\}$ e $F'=\{f_v',v'=1,2,...,M\}$, sendo

$$f_v=\sum_{i=1}^{c} h_i/\sum_{i=1}^{c} h_i \quad \text{e} \quad f_v'=\sum_{i=1}^{c} h_i'/\sum_{i=1}^{c} h_i' \quad (1)$$

onde a descrição do algoritmo é dada por:
a) para cada score *s* obtém-se o índice *v* correspondente;
b) posteriormente, obtém-se o f_v e v' tal que $f_v=f_v'$;
c) por meio de v', obtém-se s' e $HM(s)=s'$.

Normalização baseada no algoritmo de dois passos

O Método de dois passos consiste em transformar uma distribuição não-normal contínua das variáveis em uma distribuição normal (Templeton, 2011). A abordagem é realizada em dois passos:
Passo 1: envolve a transformação da variável original em uma variável estatisticamente uniforme, por meio da função "rank" percentil:

$$RP(X_i)=1- Rank(X_i)/n, \quad (2)$$

onde *n* é o número de amostras e $Rank(X_i)$ é o valor do "rank" de X_i.

Passo 2: consiste em transformar a variável do passo anterior em uma variável com distribuição normal, usando a função distribuição normal inversa:

$$p = \mu + \sqrt{2}\,\sigma\, erf^{-1}(-1+2Pr) \quad (3)$$

onde *p* é o score do passo 2, μ é a média de *p* (=0), σ desvio padrão de *p* (=1), erf^{-1} é função erro inversa, *Pr* probabilidade do passo 1.

Resultados

Os dois métodos foram comparados utilizando uma base de 161 amostras, nas quais foram determinadas as frações de massa de K, La, Lu, Na, U, Yb, CE, Cr, Cs, Eu, Fe, Hf, SC, Tb e Th. As variáveis (elementos) não apresentaram uma distribuição significativamente normal $(p<0.05)$, como mostrado nos gráficos de Lu e Eu nas Figuras 1 e 2.

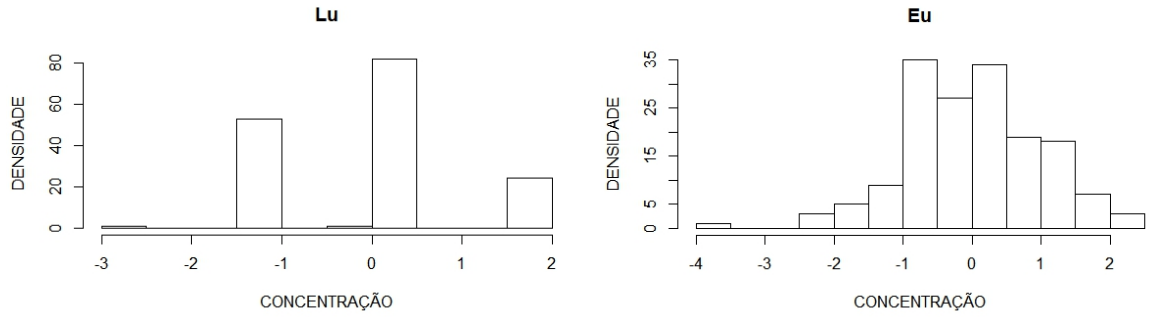

Figura 1. Histogramas para Lu e Eu dos dados brutos

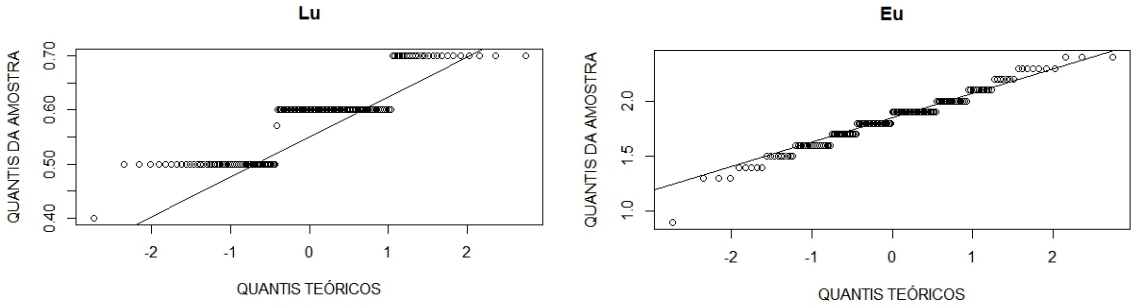

Figura 2. Gráficos de probabilidade normal para Lu e Eu dos dados brutos

A normalidade univariada foi verificada por meio de histogramas, gráficos de probabilidade normal e do Teste de Shapiro-Wilks (Royston, 1986). Após a aplicação do método baseado na combinação de histogramas, observou-se que os elementos Na, Lu, Eu e Tb apresentam um desvio da normalidade, conforme pode-se notar nos histogramas da Figura 3 e nos gráficos de probabilidade normal da Figura 4 dos elementos Lu e Eu, sendo confirmado com o teste de Shapiro-Wilks, já que as variáveis não apresentaram uma distribuição significativamente normal ($p<0.05$). Por outro lado, ao aplicar o método baseado no algoritmo de dois passos, todos os elementos não apresentaram desvios significativos nem assimetrias, o que foi confirmado pelo teste de Shapiro-Wilks, onde o grau de significância foi maior que 0.1 ($p>0.1$), como pode ser obsevado, por exemplo, nos histogramas da Figura 5 e nos gráficos de probabilidade normal da Figura 6 para Lu e Eu.

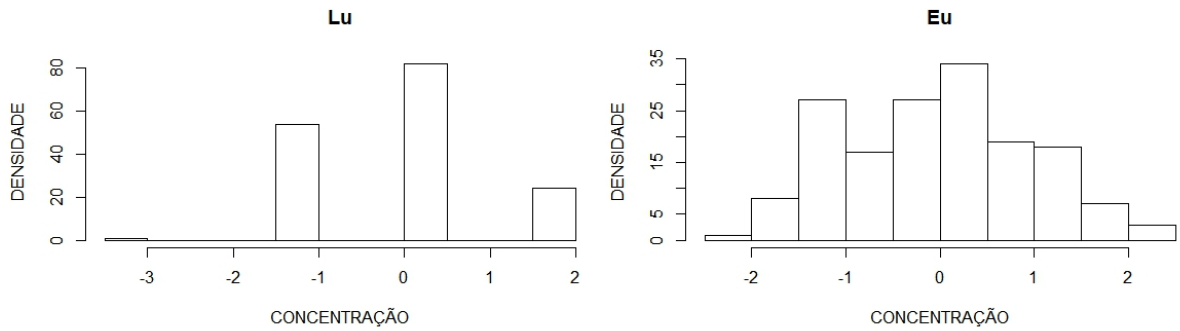

Figura 3. Histogramas para Lu e Eu após aplicação do algoritmo combinação de histogramas

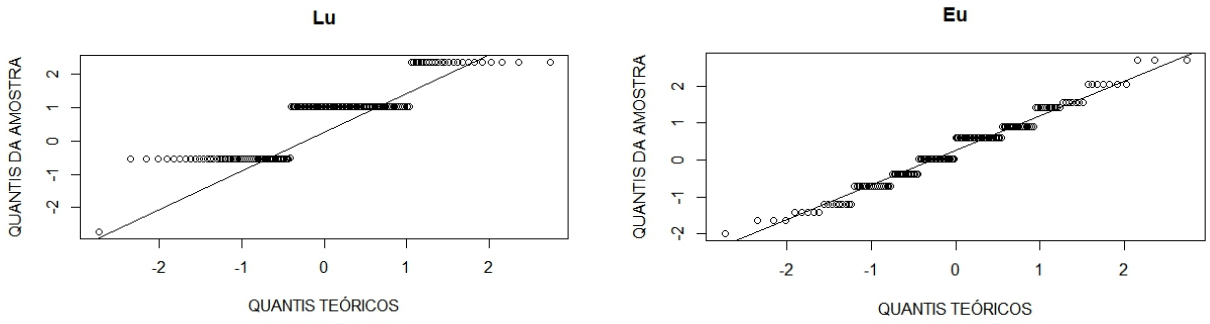

Figura 4. . Gráficos de probabilidade normal para Lu e Eu após aplicação do algoritmo combinação de histogramas

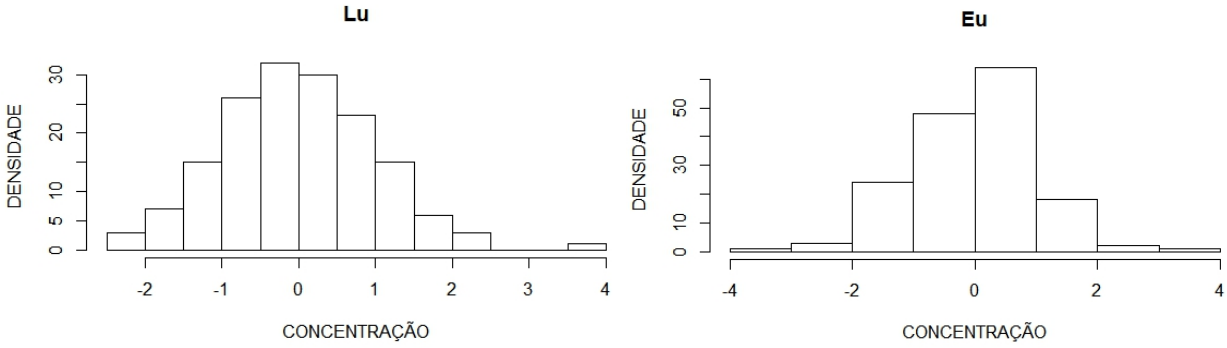

Figura 5. histograma dos elementos Lu e Eu após aplicação do algoritmo de dois passos

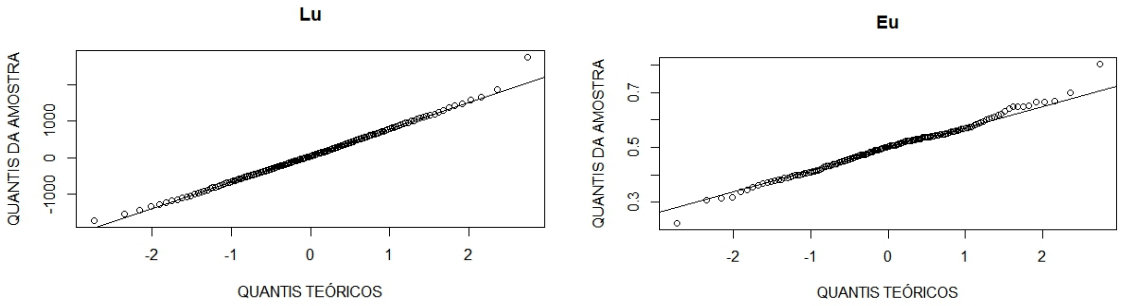

Figura 6. Gráficos de probabilidade normal para Lu e Eu após aplicação do algoritmo combinação de histogramas

Conclusão

Os métodos de normalização por histograma e dois passos usados na base de dados de 161 amostras, mostrou que o algoritmo de dois passos teve um melhor desempenho já que após sua aplicação nas 15 variáveis não ocorreram desvios significativos de normalidade nem assimetrias, o que foi verificado pelos histogramas, pelos gráficos de probabilidade normal e o teste de Shapiro-Wilks.

Referências

Fu, H.; Qiu, G. A new normalization approach for combining similarities. 8th International Conference on Intelligent Human-Machine Systems and Cybernetics, p. 404-407, 2016.

Hair, J. et al. Análise multivariada de dados. 6 ed. Bookman, 2009.

Mucha, H. J.; Bartel, H. G. Resampling techniques in cluster analysis: is subsampling better than bootstraping? In: *Data science, learning by latent structures, and knowledge discovery*. B. Lausen, S. Krolok-Schwerdt, M. Bohmer, (eds.), p. 113-122, 2015.

Royston, P. An Extension o Shapiro and Wilk's W test for normality to large samples. Applied Statistics, v. 31, p. 115-124, 1986.

Templeton, G. F. A Two-Step Approach for Transforming Continuous Variables to Normal: Implications and Recommendations for IS Research. Communications of the Association for Information Systems: v. 28, Article 4, p. 41-58, 2011.