Comunicação oral

Estudo comparativo entre três métodos de imputação

P.T.M.S. Oliveira, C.S. Munita
Instituto de Pesquisas Energéticas Nucleares, IPEN - CNEN / SP, ptoliveira@ipen.br

Palavras–chave: imputação, normal univariada, normal multivariada, decomposição por valor singular

Abstract: Many multivariate statistical methods, such as cluster analysis, principal component analysis and discriminant analysis are used to interpret the results of ceramic provenience studies in archaeometry. The studies aim to find similarities in the elemental composition, which indicate the origin of raw material used in the ceramic manufacture. One requirement to use those multivariate statistical techniques is that the sample matrix must be complete, i.e., it is required the absence of missing values. The purpose of this paper is the use of some techniques for assigning missing values in a elemental concentration matrix, such as univariate normal imputation, multivariate normal imputation and singular value decomposition. The data matrix refers to the elemental concentrations of K, La, Lu, Na, Nd, Sb, Sm, U, Yb, Ce, Co, Cr, Cs, Eu, Fe, Hf, Rb, Sc, Ta, Tb, Th and Zn obtained by Instrumental Neutron Activation Analysis (INAA) of archaeological ceramic samples.

1 Introdução

Tendo em vista o crescente avanço das técnicas físico-químicas em estudos arqueométricos, a quantidade de dados gerados 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. 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, devem ser formados grupos de amostras de acordo com alguns critérios estatísticos. Os resultados podem ser organizados dentro de uma matriz de dados X_{np}, sendo o n o número de amostras variando de 1 até n e p o número de variáveis variando de 1 até p.

2 Desenvolvimento

O estudo das concentrações faltantes consiste na substituição da concentração faltante por uma concentração consistente, com a finalidade de completar a matriz das amostras, para ser utilizada na interpretação dos dados [1]. Existem vários procedimentos para imputar os valores faltantes em uma base de dados [3]. Entretanto, neste trabalho foram estudados 3 procedimentos de imputação:

a) Imputação pelo valor normal univariado. Consiste em substituir os valores das concentrações faltantes de cada variável pelo valor estimado a partir de uma variável normal univariada, das concentrações das outras amostras da mesma variável.

b) Imputação por valor normal multivariado. Neste procedimento estima-se a média μ e matriz de dispersão Σ como um valor das médias amostrais de uma matriz de covariância usando todos os dados. Utilizam-se estas estimativas para calcular a regressão linear das variáveis perdidas em relação ás variáveis existentes em cada amostra. Substituindo os valores das concentrações das variáveis na curva da regressão, obter-se-ão os valores das concentrações das variáveis faltantes [4].

c) Imputação pelo método de decomposição do valor singular (DVS.) O método de decomposição do valor singular (DVS), consiste na substituição do valor faltante pelo valor estimado a partir do método de decomposição do valor singular, que considera a decomposição da matriz dos valores das variáveis dos scores das componentes principais [2].

Todos os estudos foram realizados em resultados de amostras de cerâmica obtidos por meio de análise por ativação com nêutrons instrumental. O propósito deste trabalho foi encontrar a forma mais adequada para substituir os valores faltantes.

3 Discussão

Entre os diferentes métodos utilizados para resolver o problema dos dados faltantes, encontrou-se que o procedimento que teve melhor comportamento foi o de imputação por normal univariada, isto é, padronizando os dados de concentração por variável com mais de 99% de acerto. Enquanto que o de mais baixo desempenho foi o método de normal multivariada com cerca de 15,5% de acerto. Estes resultados podem ser atribuídos ao fato de estarem trabalhando com dados de concentração que são considerados como dados não aleatórios.

4 Conclusão

Entre os métodos utilizados para o estudo de imputação de dados, o procedimento por normal univariada mostrou ser o mais adequado, obtendo mais de 99% de acerto.

Referências

[1] Fitzmaurice, G. (2008). Missing data: implications for analysis. *Nutrition* 24, 200–202.

[2] Krzanowski, W.J. (1987). Cross-validation in principal component analysis. *Biometrics* 43, 578–584.

[3] Barroso, L. (1995). *Inputação de Dados em Painéis para Populações Finitas*. Tese de Doutorado em Estatística, Instituto de Matemática e Estatística, Universidade de São Paulo. São Paulo SP.

[4] Bérgamo, G. (2007). *Inputação múltipla livre de distribuição a decomposição por valor singular em matriz de interação*. Tese de Doutorado em Agronomia, área de concentração: Estatística e Experimentaccão Agronômica, Escola superior de Agronomia Luiz de Queiróz, Universidade de São Paulo. Piracicaba SP.