2021 International Nuclear Atlantic Conference – INAC 2021
Virtual meeting, Brazil, November 29 – December 2, 2021

ISBN: 978-65-594-1256-3
9 786559 412563

Comparison of data imputation using the methods of Fuzzy logic, mean and autoencoder neural network

A. L. Nogueira[1], C. S. Munita[2]

[1]*andre.nogueira@ifs.edu.br, Instituto Federal*
de Sergipe (IFS), 49400-000, Lagarto, SE
[2]*camunita@ipen.br, Instituto de Pesquisas*
Energéticas e Nucleares (IPEN/CNEN-SP),
05508-000, São Paulo, SP

1. Introduction

The predicted future advancement of physicochemical techniques means that the quantity of results generated will increase significantly. For results analysis, it is necessary to use more sophisticated methods, such as multivariate techniques. In general, multivariate statistical methods allow one to evaluate a set of samples in terms of the correlations between variables. These techniques consider that each sample can be represented as a point in multidimensional space, where each dimension of hyperspace corresponds to an axis determined by the physicochemical composition of the samples. Missing values make data analysis difficult. The problems associated with missing values are: loss of efficiency, complications in treatment and data analysis, and bias resulting from differences between missing and complete data [1]. This article aims to study three imputation techniques, namely: autoencoder neural network [2], mean [3] , and fuzzy c-means algorithm [4].

The study was performed using one database of 122 samples, in which the massfractions of Na, K, La, Yb, Lu, U, Sc, Cr, Fe, Cs, Eu, Tb, Hf, Tb, respectively, were determined by instrumental neutron activation analysis, INAA. Discriminant analysis and mean distance were used to evaluate the methods.

2. Imputation techniques

2.1 Autoencoder neural network

The autoencoder is an artificial neural network trained to copy its input into the output and consists of two parts: the encoder and the decoder. The encoder is a function $h = f(x) = s_f(W_x + b_h)$, where s_f is an activation function, the encoder is parameterized by a matrix weight W_x , of order $d_h \times d_x$ and a vector bias $b_h \in R^{d_h}$. On the other hand, the decoder is a function g that maps the representation h in the reconstruction $y = g(h) = s_g\big(W'_h + b_y\big)$, where s_g is the activation function of the decoder. The training of the autoencoder network consists of determining the parameters $W_x, W'_h, b_h$and b_y that minimize the error of reconstruction in the examples of the training set D_m that acts to minimize the following objective function [2] :

$$J_{AE} = \sum\nolimits_{z \in D_n} L\big(x, g\big(f(x)\big)\big) \tag{1}$$

L is the reconstruction error, a typical choice is $L(x,y) = \big||x-y|\big|^2$.

2.2 Mean

The most commonly used approach is mean imputation [3]. In this approach the missing value from the variable is replaced by the average of the values of the data from that variable [5], i.e.,

$$\tilde{x_j} = \frac{1}{n-1} \sum_{m=1, m \neq i}^{n} x_{mj} \tag{2}$$

2.3 Fuzzy c-means algorithm

The c-mean algorithm is nothing more than a fuzzy version of the k-mean algorithm, in which the data may belong to more than one class. The following is a simple imputation version of the algorithm proposed by [4] and the imputed values are updated with the expression:

$$v_{ij} = \sum_{g=1}^{k} \mu_{ig} c_{gj}, \forall\ (i,j) \in M \tag{3}$$

where M is the set of coordinates of the missing values, μ_{ig} is the data membership function i in the class g and c_{gj} is the value of the j-th variable in the class g.

2.4 Dataset

In the tests, one database was used, consisting of 122 ceramic samples excavated from three archaeological sites that are located superficially on the slope of a hill with a water course running below it [6]. The ceramics located in these sites are associated with food preparation, funerary urns and decorative use.

2.5 Sample preparation and description of the Method

The samples were obtained by cleaning the outer surface of the ceramic artefact and extracting powder from its interior using a tungsten carbide rotary file attached to the end of a variable speed drill with a flexible shaft. After that, the powder was dried in an oven at 105°C for 24 h and stored in a desiccator.

Constituent Elements in Coal Fly Ash, NIST-SRM-1633b, were used as standards, and IAEA-Soil-7, Trace Elements in Soil, were used to check samples in every analysis. These materials were dried in an oven at 105°C for 4h [7].

About 100 mg of different ceramic samples, NIST-SRM-1633b, and IAEA-Soil-7 were weighed in polyethylene bags and wrapped in aluminium foil. Groups of 8 samples, and one of each reference material were packed and irradiated in the research reactor pool (IEA-R1) at IPEN-CNEN/SP, at a thermal neutron flux of about 5×10^{12} $\text{cm}^{-2} \times \text{s}^{-1}$ for 8h.

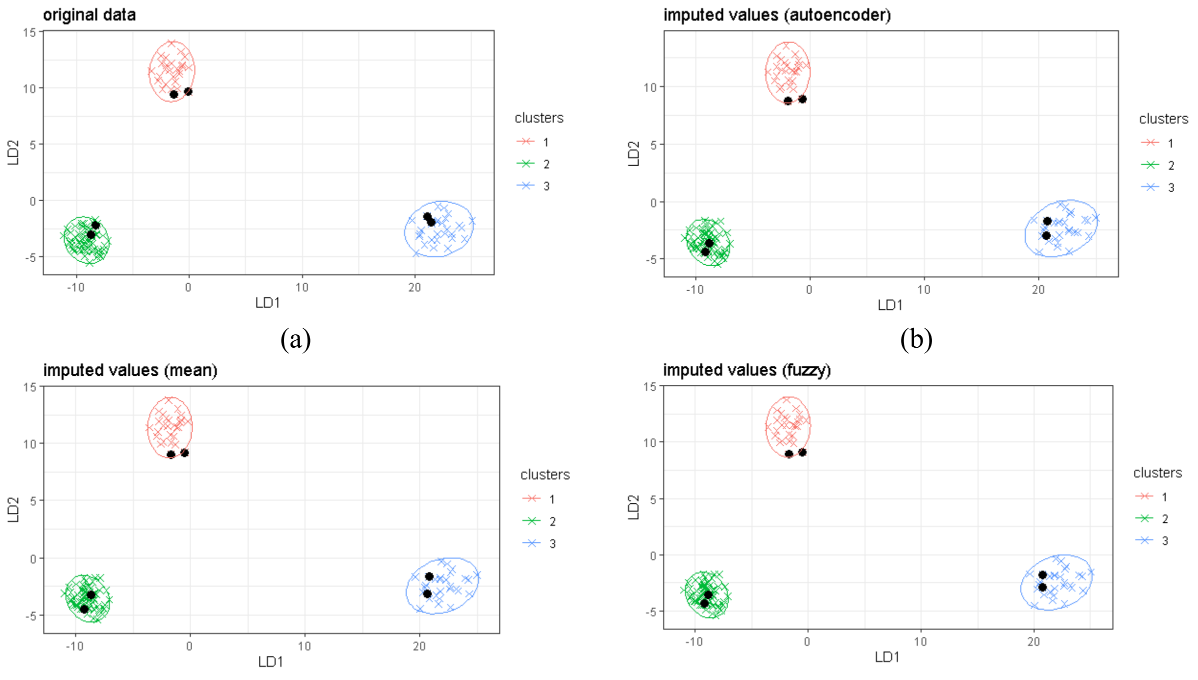
Two measurements series were carried out using a Ge (hyperpure) detector, model GX 1925 from Canberra, which has a resolution of 1.90 keV at the 1332.49 keV gamma peak of ^{60}Co, with S-100 MCA of Canberra with 8192 channels. Gamma ray spectra analysis and the concentration measurements were carried out using the Genie-2000 Neutron Activation Analysis Processing Procedure from Canberra. A detailed description of the method, the samples, the standard sample preparation, and the archaeological sites were published elsewhere [6-8].

3. Results and Discussion

The tests were performed using one database containing 122 samples (consisting of ceramic artefacts from the three sites). The elements determined for the dataset were Na, K, La, Yb, Lu, U, Sc, Cr, Fe, Cs, Eu, Tb, Hf and Tb . The mass fraction of the each samples were obtained by INAA [6, 7]. The tests were performed by selecting the two lowest concentrations of element La in each of the three groups that make up the base. Discriminant analysis was used to evaluate the results because it is a statistical technique that identifies the most relevant variables for the classification of the data into groups.

Fig. 1 shows the discriminant scores of the base with original and imputed values. Comparing graphs (a) and (b) of Fig. 1, there is a small difference between the location of the original values and those imputed by the method based on the autoencoder neural network. Nevertheless, the imputed values remain within the ellipses with a 95% confidence level. The same occurs when comparing graph (a), with graphs (c) and (d) of Fig. 1. It can therefore be noted that the imputed values do not cause harm in the classification of the data.

As indicated in Fig. 1, graphically there is no significant difference in the performance of the imputation methods presented. To assist in choosing the imputation method with the best performance, the mean distance (MD) between the original and imputed values was calculated, as shown in Table I. The lower the MD value, the closer the original and imputed values are, thereby improving the performance of the method. Table I shows that the imputed values closest to the originals were obtained with the method based on the autoencoder neural network, with MD=15.25. Mean imputation had the worst performance with MD=20.95 and the method based on the Fuzzy c-mean algorithm scored in-between with MD=16.23.



(a)

(b)

(c)

(d)

Figure 1: Graph of discriminant function 1 vs discriminant function 2. The ellipse represents a 95% confidence level.

Table I: Mean distance between original and imputed values.

Methods	Mean distance
autoencoder	15.25
mean	20.95
fuzzy	16.23

4. Conclusions

The three imputation techniques (autoencoder, media and c-mean algorithm) in the determination of missing values were applied in a database with 122 samples. Fig. 1 showed that graphically there is no significant difference in the performance of the methods presented and that the imputed values did not cause harm in the classification of the data. On the other hand, Table I shows that the imputed values closest to the originals were obtained with the method based on the autoencoder neural network, with MD=15.25. Mean imputation had the worst performance with MD=20.95 and the method based on the Fuzzy c-mean algorithm scored in-between with MD=16.23.

References

[1] G. Hawthorne, G. Hawthorne, and P. Elliott, "Imputing cross-sectional missing data: comparison of common techniques," *Australian & New Zealand Journal of Psychiatry*, vol. 39, n. 7, pp. 583-590 (2005).

[2] S. Rifai, P. Vincent, X. Muller, X. Glorot, and Y. Bengio "Contractive auto-encoders: Explicit invariance during feature extraction," *Proceedings of the 28th International Conference on International Conference on Machine Learning*, June, pp. 833-840 (2011).

[3] R. Malarvizhi, and A. Thanamani, "K-nearest neighbor in missing data imputation," *International Journal of Engineering Research and Development*, vol. 5, n.1, pp. 5-7 (2012).

[4] S. Nikfalazar, C. Yeh, S. Bedingfield, and H. Khorshidi, "A new iterative fuzzy clustering algorithm for multiple imputation of missing data," *IEEE International Conference on Fuzzy Systems*, July, pp. 1-6 (2017).

[5] R. Little, and D. Rubin, *Statistical analysis with missing data*, John Wiley & Sons, Hoboken & USA (2019).

[6] C. Munita, R. Paiva, M. Alves, P. De Oliveira, and E. Momose, "Provenance study of archaeological ceramic," *Journal of Trace and Microprobe Techniques*, vol. 21, n.4, pp. 697-706 (2003).

[7] J. Santos, M. Reis, C. Munita, and J. Silva, "Box-Cox transformation on dataset from compositional studies of archaeological potteries," *Journal of Radioanalytical and Nuclear Chemistry*, vol. 311, n. 2, pp. 1427-1433 (2017).

[8] C. Munita, R. Paiva, M. Alves, P. De Oliveira, and E. Momose, "Major and trace element characterization of prehistoric ceramic from Rezende archaeological site,"*Journal of Radioanalytical and Nuclear Chemistry*, vol. 248, n. 1, pp. 93-96 (2001).