2024 International Nuclear Atlantic Conference – INAC 2024
Escola de Guerra Naval, Rio de Janeiro, RJ, May 6 - 10, 2024

ISBN: 978-65-594-1256-3

Imputation in experimental results

A. L. Nogueira[1], C. S. Munita[2], and P. R. Carvalho[3]

[1] *andreln27@yahoo.com, Federal institute of Sergipe, IFS-SE, Rod. Lourival Batista, s/n, Povoado Carro Quebrado, Lagarto, SE, CEP 49400-000, Brazil*
[2] *camunita@yahoo.com, Nuclear and Energy Research Institute, IPEN-CNEN/SP, Av. Professor Lineu Prestes, 2242, São Paulo, SP, CEP 05508-000, Brazil,*
[3] *e-mail da Patrícia, Nuclear and Energy Research Institute, IPEN-CNEN/SP, Av. Professor Lineu Prestes, 2242, São Paulo, SP, CEP 05508-000, Brazil,*

1. Introduction

In experimental results, the occurrence of missing (absent) values is frequent. The problems associated with absent values are: loss of efficiency, complications in data analysis, bias resulting from differences between absent and complete results, reduction of statistical performance, among others [1]. These difficulties emerge because statistical methods consider that the same variables were determined in all samples and included in the sample matrix.

Frequently, variables or samples with missing values are excluded from the data matrix. This is the most common procedure used by statistical programs, since such programs are unable to process samples with missing values [2]. In this sense, researchers have studied strategies to substitute missing values for plausible values, a process generally referred to as data imputation [3].

Imputation methods are classified into two types: simple and multiple [4]. Simple imputation refers to the substitution of a missing value by a plausible value of a variable within a set of samples. In simple imputation methods, one assumes that the imputed value is unique and correct. Instead of substituting a unique value for each absent sample or variable, multiple imputation substitutes several plausible values and then determines the uncertainty of the values being imputed. However, if the proportion of absent values is small, less than 5%, simple imputation can be very precise [5]. The aim of this article is to evaluate the performance of four methods of data imputation: mean [6], autoencoder [7], clustering [8], and c-means [9].

The study was carried out using a database composed of 146 ceramic samples from 3 archaeological sites in Brazil, named Água Limpa, Prado and Rezende, the former located in Monte Castelo city, São Paulo state, and the latter located in Predizes and Centralina city, both in Minas Gerais state. The INAA technique was used to determine As, Ce, Cr, Eu, Fe, Hf, La, Na, Nd, Sc, Sm, Th, and U. Table 1 shows the mean and standard deviation of the elements in each of the sites [10, 11].

Table 1: Means and standard deviations for ceramic samples from Água Limpa, Prado and Rezende archaeological sites, in $\mu g g^{-1}$, unless indicated.

Element	Água Limpa n=76	Prado n=34	Resende n=30
As	2.17 ± 0.97	1.57 ± 0.34	1.84 ± 0.48
Ce	122.29 ± 20.90	115.22 ± 9.92	83.95 ± 34.57
Cr	160.85 ± 30.31	138.20 ± 20.61	218.13 ± 28.43
Eu	2.46 ± 0.34	1.40 ± 0.16	3.17 ± 0.41
Fe, %	3.27 ± 0.72	2.85 ± 0.56	1.09 ± 0.24
Hf	8,40 ± 1.02	8.87 ± 0.69	11.44 ± 0.71
La	70.58 ± 10.18	33.23 ± 3.97	37.27 ± 6.16
Na, %	0.19 ± 0.06	0.06 ± 0.01	0.016 ± 0.004
Nd	57.57 ±9.81	38.23 ± 7.59	52.26 ± 9.16
Sc	15.55 ± 2.33	29.65 ± 2.03	43.87 ± 3.02
Sm	9.57 ± 1.35	7.45 ± 0.63	10.37 ± 1.50
Th	12.81 ± 1.94	17.47 ± 0.96	6.37 ± 0.76
U	1.38 ± 0.29	4.24 ± 0.87	1.36 ± 0.23

2. Methodology

To evaluate the estimates obtained by the imputation methods, the normalized root mean squared error (NRMSE) was used, which calculates the error between the real value y_j, and the estimated (imputed) value $\widehat{y_j}$, in the following way:

$$NRMSE = \frac{1}{\sigma_y}\sqrt{\frac{\sum_{j=1}^{N}\left(y_j - \widehat{y_j}\right)^2}{N}}, \quad (1)$$

where σ_y is the standard deviation of N real values corresponding to all of the missing values [12]. The NRMSE takes into account the scale of the values.

3. Results and Discussion

The study was carried out using two samples from each site (Água Limpa, Prado and Rezende) and the elements La, Na and Sm, since upon irradiation with thermic neutrons, these give origin to Na24, La140 and Sm153 whose mean half-lives are 15.0, 40.3 and 47.3 h, respectively. With this type of analysis, its precision can be affected depending on the concentration of the elements, and whether or not the samples are measured at the correct decay time. Furthermore, the energy peak of the radioisotope may not appear on the spectrum as a consequence of its decay time and half-life. In these cases, it is frequent for the analyst to eliminate the element of the base sample, which can harm the interpretation of the results. One way of getting around this problem is through data imputation. To study the imputation methods, the results of the selected samples (corresponding to the lowest values in each site) were excluded from the database. Table 2 presents the samples from the Água Limpa, Prado and Rezende sites, and the mass fractions of variables La, Na and Sm which were excluded from the base. To analyze the performance of each of the imputation methods, we used:the normalized root mean squared error (NRMSE), equation 1.

Table 2: Site, samples and results, in $\mu g g^{-1}$, of the mass fractions that were excluded

	Element					
	La		Na		Sm	
Site	Sample	Excluded value	Sample	Excluded value	Sample	Excluded value
A	13	27.26	1	302.12	13	6.35
	33	28.12	6	328.49	25	6.51
B	68	54.95	95	841.09	68	7.03
	85	54.26	109	700.01	85	7.34
C	121	28.12	123	92.91	121	7.45
	124	26.21	136	93.12	124	8.46

Table 3 presents the NRMSE values of La for each of the imputation methods, using equation 1. The imputed values were calculated using samples from the 3 sites simultaneously. The method based on clustering (0.93) was superior to imputation by the mean (1.57), autoencoder (1.59) and c-means (1.57).

Table 3: Method and value of the NRMSE for La, $n = 140$

Method	NRMSE
Mean	1.57
Autoencoder	1.59
Clustering	0.93
C-Means	1.57

Table 4 shows the NRMSE values of Na for the four methods, calculated using the mass fractions for all three sites. The method based on clustering (1.02) performed better than those based on the mean (2.80), autoencoder (2.83) and c-means (2.71).

Table 4: Method and value of the NRMSE for Na, n=140

Method	NRMSE
Mean	2.80
Autoencoder	2.83
Clustering	1.02
C-Means	2.71

Table 5 shows the NRMSE values for Sm, calculated using the mass fractions for all three sites. Differently to what occurred in Tables 3 and 4, in which the best performance was reached by the clustering method, Table 5 shows that the methods: mean (2.78), autoencoder (2.86) and c-means (2.72), performed better than the method based on clustering (3.93).

Table 5: Method and value of the NRMSE for Sm, $n = 140$

Method	NRMSE
Mean	2.78
Autoencoder	2.86
Clustering	3.93
C-Means	2.72

We also evaluated the impact of imputation of La, Na and Sm upon determining groups by the clustering methods, through the error rate, after the application of the imputation methods. Before the clustering analysis, logarithmic transformations were applied on the complete base, as well as on the imputed values. Comparing the results of the clustering methods with the imputed values and the complete base, there was no impact on the determination of the groups, since alteration of the error rates (0%) did not occur.

4. Conclusions

The tests performed with a base of 146 samples showed that simple imputation methods based on the mean, autoencoder network, clustering and c-means, presented different NRMSE values. Despite that, data imputation had no impact on the determination of groups by either hierarchical (simple linkage, mean, complete and Ward) and partitional (k-means, k-medoids and hybrid) clustering methods. This is because the calculated error of the complete base and of the bases with imputed values was the same.

References

[1] G. Hawthorne, P. Elliott, "Imputing cross-sectional missing data: comparison of common techniques", *Aust NZ J Psychiatry*, vol. 39(7), pp. 583-590 (2005). https://doi.org/10.1080/j.1440-1614.2005.01630.x

[2] M. Misztal, "Comparison of selected multiple imputation methods for continuous variables-preliminary simulation study results" *Acta Univ Lodz Folia Oecon*, vol. 6, pp. 73-98 (2018). https://doi.org/10.18778/0208-6018.339.05

[3] T. D. Little, T. D. Jorgensen, K. M. Lang, E. W. G. Moore, "On the joys of missing data", *J Pediatr Psychol* vol. 39(2), pp. 151-162 (2014). https://doi.org/10.1093/jpepsy/jst048

[4] S. Sinharay, H. S. Stern , D. Russell, "The use of multiple imputation for the analysis of missing data", *Psychol Methods*, vol. 6(3), pp. 317-329 (2001). https://doi.org/10.1037/1082-989X.6.4.317

[5] J. L. Schafer, "Multiple imputation: a primer", *Stat Methods in Med Res*, vol. 8(1), pp. 3-15 (1999). https://doi.org/10.1177/096228029900800102

[6] R. Malarvizhi R, A. Thanamani, "K-nearest neighbor in missing data imputation", *Int. J of Eng Res Dev*, vol, 5(1), pp. 5-7 (2012).

[7]. P. Vincent, H. Larochelle,I. Lajoie, Y. Bengio, P. A. Manzagol, "Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion", *J Mach Learn Res*, vol. 11(12), pp. 3371-3308 (2010).

[8]. H. Shi, P. Wang, X. Yang, H. Yu, "An improved mean imputation clustering algorithm for incomplete data", *Neural Process Lett*, vol. 54, pp. 3537-3550 (2020). https://doi.org/10.1007/s11063-020-10298-5

[9]. S. Nikfalazar, C. H. Yeh, S. Bedingfield, H. A. Khorshidi, "A new iterative fuzzy clustering algorithm for multiple imputation of missing data", *In FUZZ-IEEE*, Nápolis/Italy, 09-12 jul., pp. 1-6 (2017). https://doi.org/10.1109/FUZZ-IEEE.2017.8015560

[10] C. S. Munita, R. P. Paiva, M. A. Alves, P. M. S. De Oliveira, E. F. Momose,"Major and trace element characterization of prehistoric ceramic from Rezende archaeological site", *J Radioanal Nucl Chem*, vol. 248(1), pp. 93-96 (2001) https://doi.org/10.1023/a:1010682209370

[11] C. S. Munita, R. P. Paiva, M. A. Alves, P. M. S. De Oliveira, E. F. Momose "Provenance study of archaeological ceramic", *J Trace Microprobe Techn*, vol. 21(4), pp. 697-706 (2003). https://doi.org/10.1081/TMA-120025819

[12] L. P. Brás, J. C. Menezes, "Improving cluster-based missing value estimation of DNA microarray data", *Biomol Eng*, vol. 24(2), pp. 273-282 (2007). https://doi.org/10.1016/j.bioeng.2007.04.003