Archetypal Analysis as an imputation method for missing data

Pórtya Piscitelli Cavalcanti [1], **Matthew Frederick Parry** [2], **Bryan Frederick John Manly** [3] **Carlos Tadeu dos Santos Dias** [4]

1 Introduction

In real applications, specially in multivariate dataset, missing data are common, which results in incomplete data. Three types of missing data can be cited: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). In MCAR the causes of missing values are unrelated to the data; in MAR the missing values probability is related to the values for some other observed variables; and in MNAR the probability of missing an observation is related to its value (EPIFANIO; IBÁÑEZ; SIMÓ, 2018).

Regardless the type of missing data, most statistical techniques assume the completeness of the data, and so the incomplete observations are often discarded. Thus, sometimes an entire row of observations is discarded due to a missing value in only one column. This way, one may loose valuable information (EPIFANIO; IBÁÑEZ; SIMÓ, 2018).

A common approach to deal with it is the imputation method, which means to replace the missing data with an estimate (DONDERS et al., 2006). Once the missing values are filled, the full dataset can be analyzed normally. Among the techniques to impute missing data, one is the Archetypal Analysis (AA).

AA is a multivariate technique, proposed by Cutler and Breiman (1994), whose objective is to reduce the space of observations of a dataset by means of convex combinations of the data. In this way, archetypes are convex combinations of data points and data points are explained as convex combinations of archetypes (BAUCKHAGE et al, 2015). In addition, a set of archetypes consist on extreme elements that lies on the boundary of the data convex hull (CUTLER; BREIMAN, 1994).

Bauckhage et al. (2015) presented another approach to AA where they used subgradient algorithms for optimization over the simplex to determine archetypes and reconstruction coefficients. They also applied AA as an autoencoder which inspired the use of AA to deal with missing data.

The aim of this study is to evaluate the use of archetypes to impute values in a dataset with simulated missing data.

2 Material and methods

This study was performed on a dataset about the endemic New Zealand Hector's dolphin, seeking differences between the South Island and the North Island (Maui's dolphin)

[1]Department of Mathematics & Statistics - University of Otago/ Departamento de Ciências Exatas - ESALQ/USP. e-mail: *portya@usp.br*

[2]Department of Mathematics & Statistics - University of Otago. e-mail: *mparry@maths.otago.ac.nz*

[3]Manly-Biostatistics Limited. e-mail: *BryanManly@xtra.co.nz*

[4]Departamento de Ciências Exatas - ESALQ/USP/ Departamento de Ciências do Solo - CCA/UFC. e-mail: *ctsdias@usp.br*

populations (BAKER; SMITH; PICHLER, 2002). The following six measurements (variables) were taken from dolphin skeletons in museums: condylobasal length (CBL), rostrum width at midlength (RWM), rostrum length (RL), zygomatic width (ZW), mandible length (ML) and rostrum width at base (RWB). From a total of 59 specimens, 13 were from the North Island and 46 from the South Island.

In this type of data, there may be missing values due to damaged or missing bones in the collection, which could occur completely at random. So, in order to simulate MCAR data, nine observations were randomly removed from the original dataset, and then some variables values were randomly removed from each one. This procedure was done considering that the number of removing values r in each observation follows a zero-truncated Binomial distribution $ZTBin(6, p)$ with probability p (Figure 1).

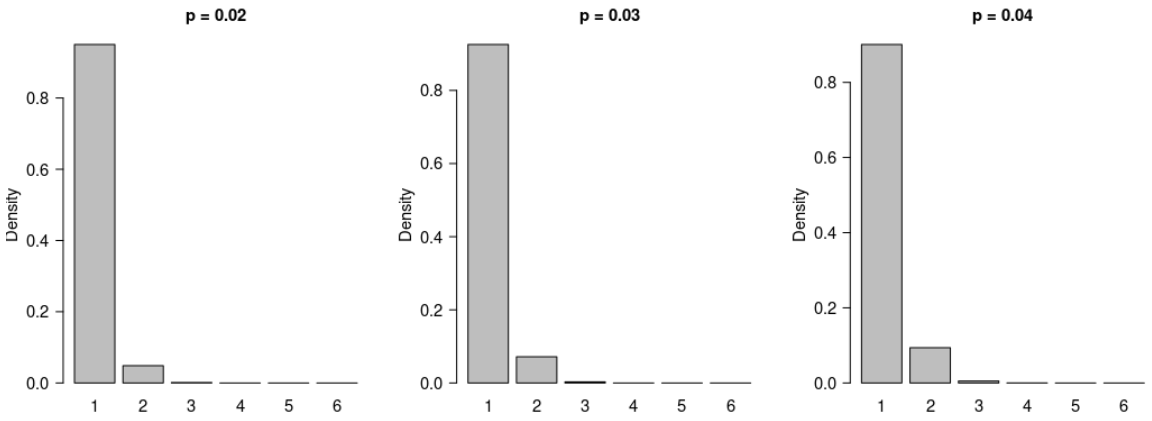

Figure 1: Probabilty density of a zero-truncated Binomial distribution $ZTBin(6, p)$. Source: author's own elaboration.

As nine observations were removed, the probability of removing one value ($r = 1$) is $p = 0.0272$. Hence, three values of the probability p were evaluated in different scenarios: $p = 0.02$, 0.03 and 0.04 (Figure 1).

The AA was performed on the scaled complete data (50 observations), without the nine incomplete values. This technique consists on determine stochastic matrices $\boldsymbol{B}$ and $\boldsymbol{A}$ for a given data matrix $\boldsymbol{X}$ and for a number K of archetypes, such that $\boldsymbol{X} \approx \boldsymbol{XBA} = \boldsymbol{ZA}$, where $\boldsymbol{Z}$ is the archetypes matrix. The analysis was performed according to the Bauckhage et al. (2015) approach, based on the Frank-Wolfe algorithm. The stochastic matrices were determined in an alternating manner, assuming $\boldsymbol{Z} = \boldsymbol{XB}$ to update the current estimate of $\boldsymbol{A}$, and then fixing $\boldsymbol{A}$ to update the current estimate of $\boldsymbol{B}$. From there, three archetypes were chosen.

Thus, to impute the missing data by means of its archetypes, we estimated $\boldsymbol{A}*$ just for the missing values, that were also scaled in accordance with the complete data, and then multiplied by the archetypes, such that $\tilde{\boldsymbol{X}} = \boldsymbol{ZA}*$, where $\tilde{\boldsymbol{X}}$ is the matrix of imputed values. This procedure was repeated 100 times for each p evaluated and the results obtained were compared with the original data by calculating the Mean Squared Error (MSE).

Also, the Principal Components Analysis (PCA) was performed in the scaled complete data in order to characterize and visualize the dataset, their archetypes and the imputed values. All computing was performed in R software (R CORE TEAM, 2019).

3 Results and discussion

From the removed observations, only one was from the North Island (Table 1), which was expected as the number of North Island's specimens was smaller.

Table 1: Removed observations from the original dataset.

Island	CBL	RWM	RL	ZW	ML	RWB
North	317	63	165	167	257	94
South	276	49	135	142	220	77
South	284	48	138	139	222	78
South	276	50	130	147	222	80
South	271	51	133	138	219	77
South	302	55	147	150	243	80
South	282	51	139	140	226	76
South	278	54	133	146	221	79
South	285	46	140	141	223	75

Source: author's own elaboration.

The MSE values of the 100 scenarios of each p evaluated can be seen in Figure 2, where all the values were much lower than 1, as expected.

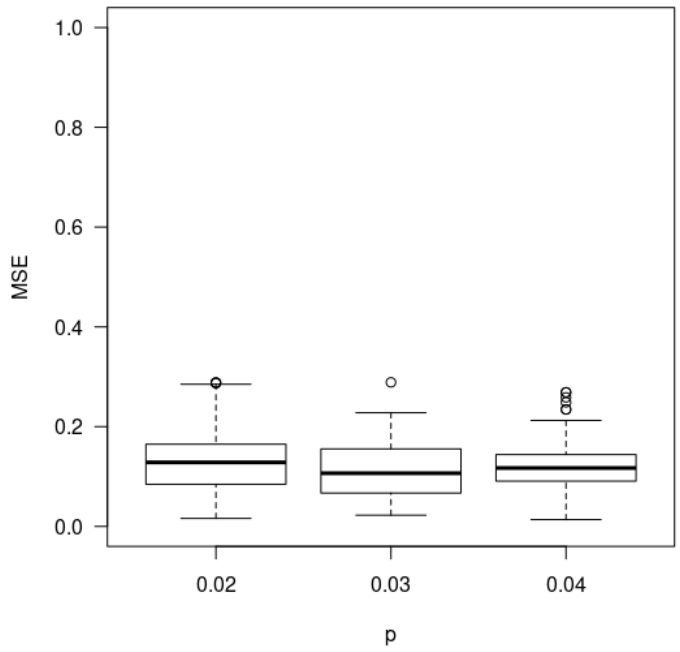

Figure 2: Boxplot of the MSE in all 100 scenarios of each p. Source: author's own elaboration.

The scenario that presented the lowest MSE was obtained by $p = 0.04$ probability of removing values and $r = 1$ variables removed in each observation. Thus, it was choosen

to ilustrate the imputation via archetypes (Table 2). The imputed values are highlighted in bold type.

Table 2: Ilustration of the removed observations with imputed values (bold).

Island	CBL	RWM	RL	ZW	ML	RWB
North	**314.51**	63.00	165.00	167.00	257.00	94.00
South	276.00	**50.49**	135.00	142.00	220.00	77.00
South	284.00	48.00	138.00	139.00	222.00	**73.90**
South	276.00	50.00	130.00	**140.85**	222.00	80.00
South	271.00	51.00	133.00	138.00	219.00	**77.89**
South	**298.64**	55.00	147.00	150.00	243.00	80.00
South	282.00	51.00	139.00	140.00	**224.46**	76.00
South	278.00	54.00	133.00	**142.58**	221.00	79.00
South	**281.63**	46.00	140.00	141.00	223.00	75.00

Source: author's own elaboration.

The results are noteworthy as the imputed data were similar to the original data. Next, in Figure 3, one can see the biplot with the results of PCA for the complete data (red and blue dots), their archetypes (purple dots) and the example of imputed values (red and blue triangles).

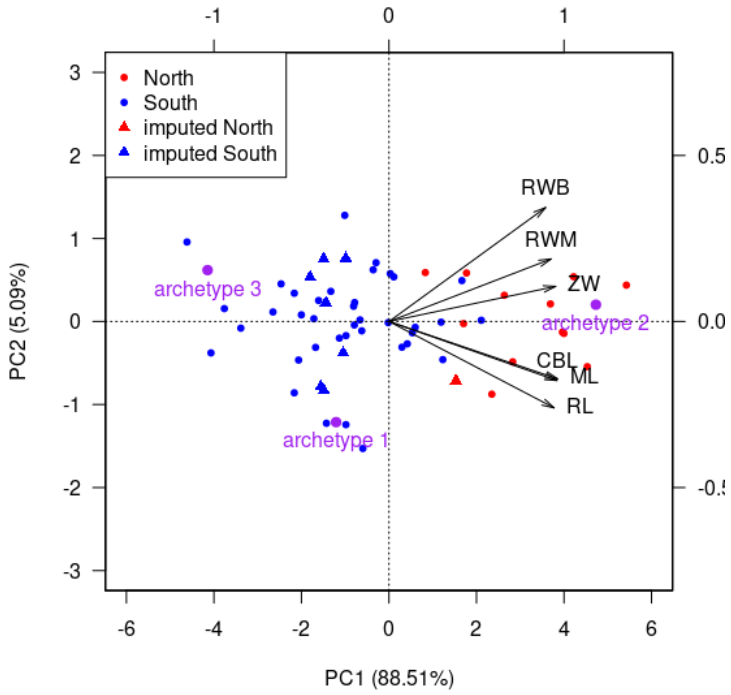

Figure 3: Biplot of the first two principal components PC1 and PC2. Source: author's own elaboration.

The first two principal components explained together 93.6% of the total variance, then, they are enough to characterize dolphins' measurement. One more time, it can be

noted that the imputed data were similar to the original ones and corroborated with the interpretation of the dataset (Figure 3).

The North Island dolphins (red dots) are located in the first and fourth quadrants of the graph as well as the variables vectors, which means that these specimens presented higher values of the response variables. The majority of the South Island dolphins (blue dots) are located in the second and third quadrants and in the opposite direction of the variables vector, this way, they presented lower values and can be considered smaller than the North Island dolphins (Figure 3).

In Figure 3 one can also see the three archetypes. Archetype 3 had the lowest values of all measurements, representing the smallest dolphins, archetype 2 had the higher values of all variables, representing the largest dolphins, and archetype 1 represents the medium-size dolphins.

4 Conclusion

Based on the results, the use of AA as an imputation method in a MCAR situation was promising and the body measurement of dolphins were successfully imputed via archetypes.

In future work we aim to address the AA method on other types of missing data, like MAR and MNAR.

Acknowledgements

We are grateful to CAPES and CNPq for the financial support.

References

BAKER, A. N.; SMITH, A. N.; PICHLER, F. B. Geographical variation in Hector's dolphin: Recognition of new subspecies of *Cephalorhynchus hectori*. *Journal of the Royal Society of New Zealand*, v. 32, n.4, p. 713-727, 2002.

BAUCKHAGE, C.; KERSTING, K.; HOPPE, F.; THURAU, C. Archetypal analysis as an autoencoder. In: WORKSHOP NEW CHALLENGES IN NEURAL COMPUTATION, 2015, Aachen. *Machine learning reports*, p. 8-15, 2015.

CUTLER, A.; BREIMAN, L. Archetypal analysis. *Technometrics*. v. 36, p. 338-347, 1994.

DONDERS, A. R. T.; VAN DER HEIJDEN, G. J. M. G.; STIJNEN, T.; MOONS, K. G. M. A gentle introduction to imputation of missing values. *Journal of clinical epidemiology*. v. 59, n. 10, p. 1087-1091, 2006.

EPIFANIO, I.; IBÁÑEZ, M. V.; SIMÓ, A. Archetypal analysis with missing data: see all samples by looking at a few based on extreme profiles. *The American Statistician* (just-accepted), p. 1-39, 2018.

R CORE TEAM. *R*: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2019. ISBN 3-900051-07-0, URL `http://www.R-project.org/`.