Journal of Statistics on Health Decision

The dimension reduction power of ClustOfVar: application of the variable cluster analysis technique in a mixed data health database

Natacha Oliveira[1], Milton Severo[2,3]

[1]Department of Mathematics, University of Aveiro, 3810-193 Aveiro, Portugal
[2]EPIUnit - Institute of Public Health, University of Porto, Rua das Taipas, n° 135, 4050-600 Porto, Portugal
[3]Laboratory for Integrative and Translational Research in Population Health (ITR), University of Porto, Rua das Taipas, n° 135, 4050-600 Porto, Portugal

ABSTRACT

Background/Objective: Technological evolution is increasingly making real the elements necessary for the daily practice of personalized medicine, an improved vision of health care whose decisions regarding prognosis, diagnosis and therapeutic strategies depend on the patient's various characteristics. This approach leads to the collection and use of information that is broad in extension and complexity, for which dimensionality reduction techniques are imperative, in order to simplify and understand it. This paper aims to show the value of the ClustOfVar technique, a variable clustering technique capable of dealing with mixed data, resulting in data reduction. Through its hierarchical and non-hierarchical approaches, it replaces sample variables with representative synthetic variables. This dimensional reduction can be extended to individuals by applying Ward's method.

Methods: The cleaning process of anthropometric, obstetric, vital signs and pubertal status data from 700 participants of the Generation XXI cohort and/or their mothers led to variables being removed (181 down to 105 variables, 82 quantitative and 23 qualitative). Then, the hierarchical technique of the ClustOfVar package was applied, which started by building a hierarchy of variables. The optimal number of clusters was then determined, considering the aggregation level plot and the bootstrap methodology, and each cluster was characterized. The partition into clusters was then tried with the non-hierarchical process. Once the partition was defined, Ward's method was applied, dividing the participants into clusters. We finished with their description according to the synthetic variables.

Results: The partition in 8 clusters of variables suggested by the hierarchical technique was chosen, with the first and third cluster being filled mainly by maternal characteristics (relating mainly to menstruation and physical measurements, respectively). While cluster 2 mixes maternal and individual characteristics, cluster 4 contains only patient variables at birth. Cluster 5 is the most diverse, with anthropometric and related measurements of vital signs and blood macromolecules. Cluster 6 has total mass and fat measurements. Finally, cluster 7 is related to pubertal status variables, and cluster 8 includes cholesterol variables. The clustering of individuals results in the creation of specific profiles for each of the 8 clusters of individuals.

Conclusions: The ClustOfVar technique accomplishes a data transformation relevant to the dispersion of personalized medicine. However, it lacks the ability to deal with high proportions of missing data and its bootstrap process is very time-consuming.

Introduction:

Personalized medicine involves an optimized vision of healthcare that has at its core the various differences inherent in patients (genetic, epigenetic, molecular, physiological, physiological, environmental, behavioral, etc.) and relies on them to investigate possible prognoses, determine diagnoses, and choose the therapeutic strategy to be followed [1]. The constant development of procedures and platforms for collecting biological data, such as the Internet of Things, mhealth and omics science techniques (e.g. next-generation sequencing) has led to the creation of big data, thus opening the door to solving complex health enigmas. This type of heterogeneous dataset, by containing the variability of individuals, provides a basis for implementing this type of healthcare practices [2,3]. Considering the varied nature of the information collected, data processing aimed at reducing its dimensionality is crucial to increase its interpretability [4]. In the health sector, this heterogeneity can usually be translated into the presence of mixed data samples (containing both qualitative and quantitative variables). Clustering is a useful technique for exploratory analysis or reduction of mixed data size by forming groups of individuals or groups of variables [5,6]. The techniques of the ClustOfVar package appear as an interesting alternative, capable of applying cluster ana-

Keywords:
ClustOfVar, Clustering, Dimension Reduction, Generation XXI Cohort, Mixed Data

Corresponding author:
Natacha Oliveira
natachasantosoliveira@outlook.com

Conflict of interest:
The authors declare no conflict of interests.

First published: 01JUN2023

lysis of mixed data variables, and being able to adapt this transformation in a subsequent cluster analysis of individuals [7,8].

This paper intends to demonstrate the hierarchical and non-hierarchical variable clustering techniques of the ClustOfVar package, as well as the method of clustering individuals from the results of these techniques, highlighting the relevance of this methodology in the implementation of personalized medicine.

The ClustOfVar package was developed with the objective of creating variable clustering techniques compatible with mixed data. This technique was inspired by the PCAmix method, a PCAMIX method with varimax rotation and addition of the singular value decomposition method, and will be applied to each cluster of variables, acquiring a synthetic non-orthogonal variable for each one [7].

The package in question comprises a hierarchical and a non-hierarchical methodology (similar to k-means), capable of achieving variable clustering. A homogeneity criterion was defined as the core of both techniques, calculated from the relations between the synthetic variable generated for the cluster in question and the squared Pearson correlations (quantitative variables) or squared correlation ratios (qualitative variables) of the variables belonging to the same cluster. The purpose is to maximize this criterion, and this value will be higher the higher the relationship between the sample variables and the synthetic variable. The homogeneity will be translated into the measure of cohesion gain, which corresponds to the proportion of homogeneity obtained from the partition in the defined number of clusters, relative to the homogeneity generated by the division of each variable by different clusters [7-9]. The ClustOfVar package is also equipped with a bootstrap function, capable of indicating the optimal number of clusters by calculating the stability of the partitions of the sample variables into k clusters [7,8].

Methods:

Data from 700 individuals participating in the Generation XXI cohort were used for the present study by the clustering technique of the ClustOfVar package. Out of the initial 181 variables, we considered 105 variables related to anthropometric measures, vital signs, and pubertal status of the subjects included, at various stages of their growth, as well as obstetric information and measures of physical dimensions of their mothers. R software (version 4.1.2) and the RStudio integrated development environment (version 2022.12.0+353) were used to develop this project.

Prior to the cluster construction process, the database was cleaned by removing variables that were not in line with the goal in mind: variables which contained no information about any individual or provided the same information to all individuals and therefore did not add valuable information; variables whose information was already represented in other variable(s), in a different form; variables unable to present values for correlation with any other variable present in the database (given that the computation of the correlation of each pair of variables only takes into account the individuals who do not exhibit missing values for both variables). After completing this step, the initial database was divided into two: one containing all 82 quantitative variables and another containing the remaining 23 qualitative variables. Both were used to establish the hierarchy of association between the sample variables by the described package (function hclustvar), obtaining a representative dendrogram and a graph of aggregation levels (function plot.hclustvar), useful for weighting the number of clusters to be adopted. We proceeded with the application of the bootstrap technique (stability function), in order to verify the stability of the possible partitions and to reflect again on the number of clusters to assume. Finally, the partition was carried out in the number of clusters previously defined (cutreevar function), observed in the hierarchy previously designed (plot.hclustvar and rect.hclust functions). Once the hierarchical approach was finalized, the division into clusters suggested by the non-hierarchical approach (function kmeansvar) was assessed, and the partition associated with the greatest cohesion gain was chosen.

In order to achieve the division of individuals, a matrix containing the synthetic variables previously obtained was defined, used to obtain the matrix of Euclidean distances between individuals (dist function, package stats). Ward's technique was then performed (function hclust, package stats), and the division into clusters of individuals (function cutree), observed in the same way as applied to the grouping of variables. The decision on the number of clusters of individuals to assume was based on the observation of the structural organization of the dendrogram of individuals, having chosen the most prominent partition. In order to compare the clusters of individuals according to the scores of the synthetic variables, boxplots were created.

Results:

Clustering of variables

The dendrogram created by the hierarchical approach of the ClustOfVar package reveals the variables related to testicular volumes at 13 years as the first to cluster, with a height of 5.30×10^{-6} (Figure 3). The aggregation levels plot presented in Figure 1 provides us some optimal partitions choices. Whilst the partition in 3 clusters only gathers groups of variables that individually contribute to a decreasing height, it ignores some clusters that could also reduce this measure considerably. Partitions in 6 and 8 clusters include most of and all the clusters of interest, respectfully. When investigating the stability of the 103 possible partitions, the bootstrap method indicates the creation of 75 clusters as the most stable (mean corrected Rand index of 0.960) (Figure 2). Taking into account the objective of using this methodology in this context, although the presence of 75 clusters is reflected in a reduction of the sample size, this is not sufficient for a synthesis capable of expressing a clear and concise interpretation. Therefore, the partition in 8 clusters was chosen, given that it was the previous option with the highest mean adjusted Rand index. Cluster 1, with 6 variables, is mainly formed by characteristics of the mothers of the individuals, such as their age, education, and developed pregnancies. Cluster 2, although presenting some maternal characteristics, concentrated the variables related to the individuals' ages and some of their laboratory values (liver enzymes, uric acid, etc.), encompassing 16 variables. While cluster 3 revolves around anthropometric measurements of the parents, cluster 4 retains similar variables concerning the individuals at birth, both being composed of 4 variables. Cluster 5 stands out from the rest, as the cluster with the greatest diversity and size of variables, comprising 39 variables. These can be divided into variables related to anthropometric measurements of individuals at the ages of assessment not previously mentioned (waist and hip circumferences, weight, etc.), variables related to glucose and fat (glucose, insulin, triglycerides, etc.), and variables alluding to blood pressures. Cluster 6, of 17 variables, represents measures of total mass and fat, as well as measures of height, while cluster 7 focuses on characteristics related to the pubertal stage and sex of individuals, grouping 15 variables representative of this. Finally, cluster 8 includes 4 variables associated with cholesterol (LDL and cholesterol levels).

The non-hierarchical clustering technique failed in these data, an effect of the presence of missing data in more than one qualitative variable. The function responsible for this process fails by assuming that all variables with missing data have only one category of "NAs", rather than each having a category of "NAs". Since this approach proved infeasible, the partitioning presented by the hierarchical method was retained.

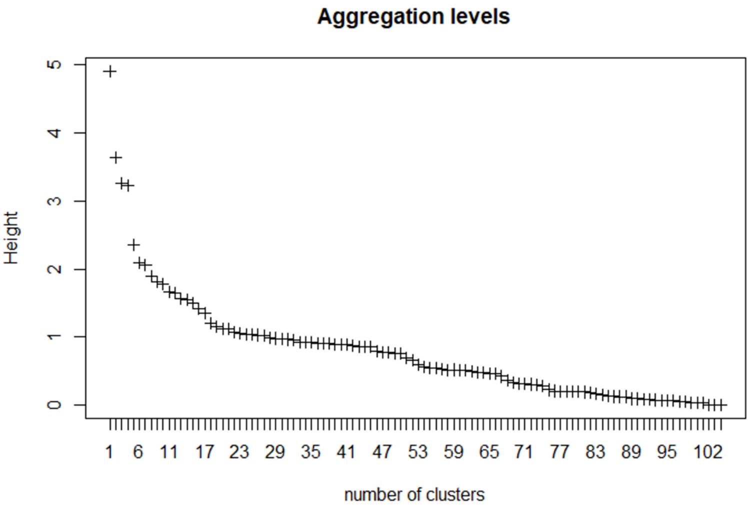


Figure 1 - Aggregation levels plot.

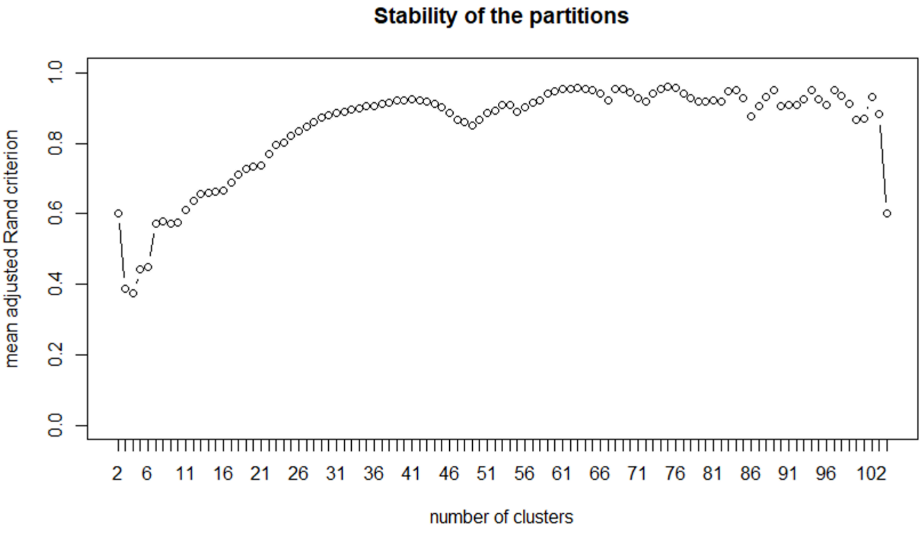


Figure 2 - Stability plot of the partition of the sample variables from the database.

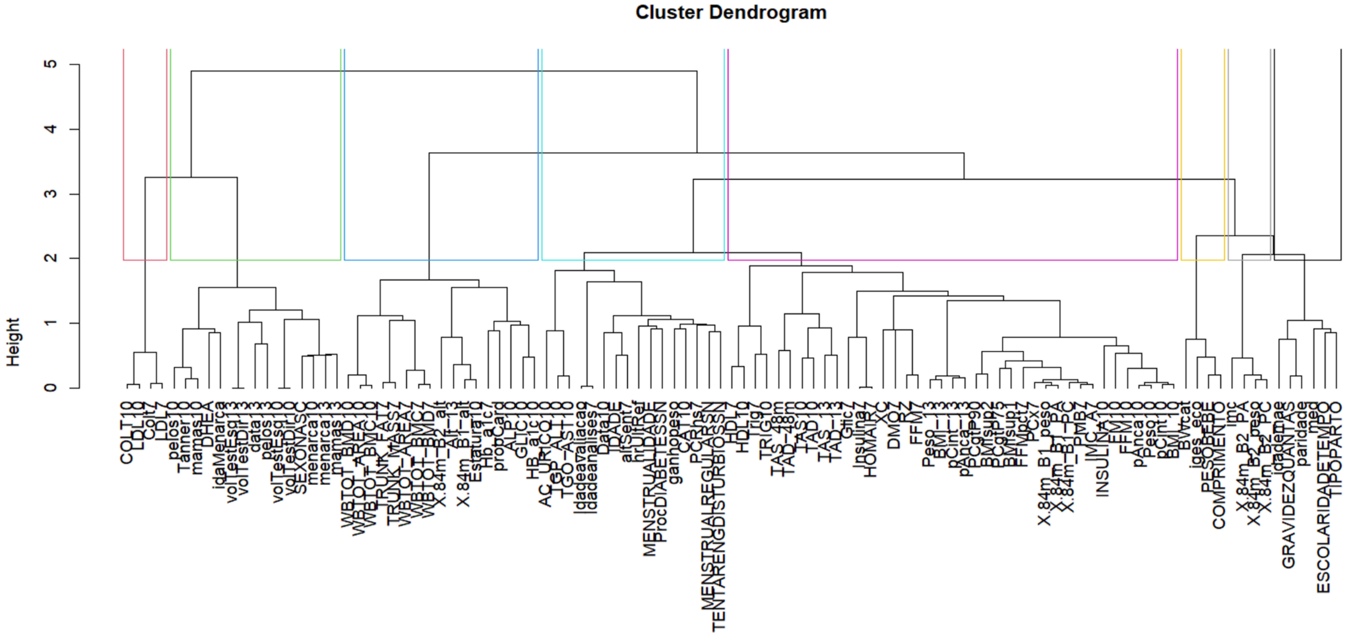


Figure 3 - Dendrogram with the application of partition by the number of clusters established.

Clustering of individuals

The application of Ward's method to the calculated synthetic variables resulted in 8 dimensionally disparate clusters of individuals, with 131 individuals present in the largest (cluster 4) and 11 individuals in the smallest (cluster 8). In Figure 4, the boxplots referring to the first synthetic variable reveal that the 8 clusters of individuals formed are similar. With the exception of cluster 7 (whose median score is negative), these boxplots show similar interquartile ranges and positive median score values for the variable. Again, there is a prominent cluster in the boxplots of synthetic variable 2, with cluster 8 showing negative score values, inferior to the approximately null median scores of the other clusters. While the boxplots of synthetic variable 3 concentrate all their median scores near zero, with slightly similar distributions, not all boxplots of synthetic variable 4 are concordant. In this variable, clusters 6 and 8 are the only ones with median score values considerably far from zero, and these measures are negative and positive, respectively. The synthetic variable 5, on the other hand, shows scattered boxplots, with the medians of clusters 1, 4, 6, and 8 being approximately zero, clusters 2, 3, and 7 being considerably negative, and cluster 3 being notably positive. Although with dissimilar interquartile ranges, most boxplots associated with synthetic variable 6 exhibit median score values close to zero, with the exception of clusters 5 and 6, with their central tendency measures noticeably negative and positive, respectively. The synthetic variable 7 is dichotomized into boxplots of positive median scores (clusters of individuals 1, 2, 3, and 5) and boxplots of negative median scores (remaining clusters). Cluster 3 also stands out, with a higher dispersion than all other clusters. Finally, the synthetic variable 8 exhibits boxplots with similar interquartile intervals, and the graph concerning cluster 6 has a remarkably negative median score value.

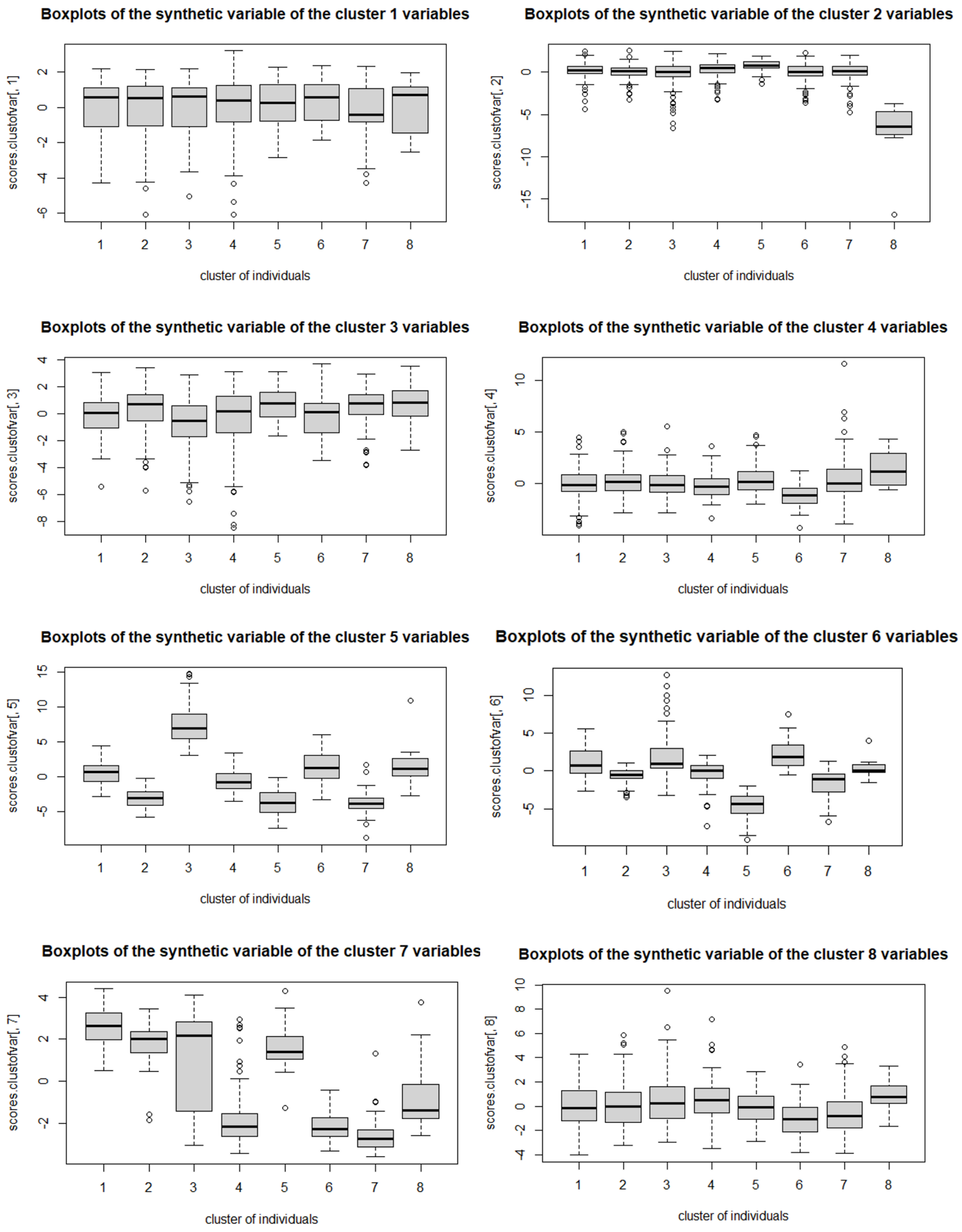
Given the differences characterized above, the subpopulations defined become distinguishable. This is the case, for instance, of cluster 8, which while it is characterized mainly by positive score values for synthetic variables 4 and 7 and negative scores for synthetic variable 2, it ends up differing from cluster 3 (defined by positive scores for synthetic variables 5 and 7) and the others.

Discussion:

The ClustOfVar technique allowed the transformation of an extensive database into a reduced matrix representative of the original data. It was thus possible to concisely identify several profiles reflecting the differentiating nuances of the individuals taken into account. Such stratifying capacity makes it possible to assign a diagnosis or to carry out the most appropriate treatment, given certain characteristics.

It is important to mention that the technique is not completely robust to the presence of missing data, as we have seen in the demonstration previously performed, and data with high proportions of this type of data require the application of more complex external imputation methodologies. On the negative side, one can also point out that the bootstrap process is time-consuming for large sample sizes [7]. For future work, it would be interesting to explore the potential benefit of adopting more objective strategies in the definition of the partition of clusters of individuals.

Given the use that can be given to the ClustOfVar technique, it stands out as a promising way of propelling personalized medicine and making its usual practice a reality.



Boxplots of the synthetic variable of the cluster 1 variables

Boxplots of the synthetic variable of the cluster 2 variables

Boxplots of the synthetic variable of the cluster 3 variables

Boxplots of the synthetic variable of the cluster 4 variables

Boxplots of the synthetic variable of the cluster 5 variables

Boxplots of the synthetic variable of the cluster 6 variables

Boxplots of the synthetic variable of the cluster 7 variables

Boxplots of the synthetic variable of the cluster 8 variables

Figure 4 - Boxplots of the values of the 8 synthetic variables, for the 8 clusters of individuals.

Acknowledgements:

I would like to thank Adelaide Freitas for all the availability, expertise and assistance throughout all aspects of this study. This work is funded by national funds through FCT - Foundation for Science and Technology, I.P., under projects UIDB/04750/2020 and LA/P/0064/2020

References:

1. Goetz LH, Schork NJ. Personalized medicine: motivation, challenges, and progress. Fertil Steril. 2018;109(6). https://doi.org/10.1016/j.fertnstert.2018.05.006
2. Dash S, Shakyawar SK, Sharma M, Kaushik S. Big data in healthcare: management, analysis and future prospects. J Big Data. 2019;6(1):54. https://doi.org/10.1186/s40537-019-0217-0
3. Bibault J, Xing L. The Role of Big Data in Personalized Medicine. In: Precision Medicine in Oncology. Wiley; 2020:229-247. https://doi.org/10.1002/9781119432487.ch8
4. Hassan M, Awan FM, Naz A, et al. Innovations in Genomics and Big Data Analytics for Personalized Medicine and Health Care: A Review. Int J Mol Sci. 2022;23(9):4645. https://doi.org/10.3390/ijms23094645
5. Ahmad A, Khan SS. Survey of State-of-the-Art Mixed Data Clustering Algorithms. IEEE Access. 2019;7:31883-31902. https://doi.org/10.1109/ACCESS.2019.2903568
6. Bry X, Cucala L. A von Mises–Fisher mixture model for clustering numerical and categorical variables. Adv Data Anal Classif. 2022;16(2):429-455. https://doi.org/10.1007/s11634-021-00449-4
7. Chavent M, Kuentz-Simonet V, Liquet B, Saracco J. ClustOfVar : An R Package for the Clustering of Variables. J Stat Softw. 2012;50(13). https://doi.org/10.18637/jss.v050.i13
8. Saracco J, Chavent M. Clustering of Variables for Mixed Data. EAS Publications Series. 2016;77:121-169. https://doi.org/10.1051/eas/1677007
9. Lomax NJ, Scheib SG. Quantifying the degree of conformity in radiosurgery treatment planning. International Journal of Radiation Oncology*Biology*Physics. 2003;55(5):1409-1419. https://doi.org/10.1016/S0360-3016(02)04599-6
10. Kiers HAL. Simple structure in component analysis techniques for mixtures of qualitative and quantitative variables. Psychometrika. 1991;56(2):197-212. https://doi.org/10.1007/BF02294458