Multivariate Data Analysis Project Report

Data Analysis 2022/2023

Alexandre Sobreira nº 59451, André Dias nº 59452, Martim Silva nº 51304,
Miłosz Włoch nº 59437, Tiago Rodrigues nº 49593

Introduction

As a result of an experimental situation, a dataset has to be analysed. Given that the researcher has rudimentary knowledge of multivariate analysis, he requests the help of statisticians to work on his problem.

This project's primary goal is to perform an exploratory analysis on the dataset provided to find existing relations between individuals or reduce the number of variables considered. A preliminary data analysis will first be performed to obtain descriptive statistics of data and observe the distribution of the variables. Afterwards, principal component analysis will be performed to verify if dimensionality reduction can be applied to simplify the problem's complexity. Relationships between individuals will then be evaluated through clustering, which groups individuals with similar characteristics in the same cluster. In the end, a comparison between the two techniques will be performed, and a scientific evaluation of the results and practical interpretations will be done.

The dataset provided contains values of six haematological variables that were measured in 51 workers of a given company. Those variables are described as follows: "HGLB" refers to the concentration of haemoglobin (an oxygen-carrying protein) in a given volumetric quantity of blood; "VGLOB" refers to the ratio of cells found in a given volumetric quantity of blood; "GBR" refers to white blood cell count; "LINF" stands for lymphocyte count, a white blood cell; "NEUTR" stands for the neutrophil count, another type of white blood cell; "CCSER" stands for Serum lead concentration.

Preliminary analysis of the data

Descriptive statistics is one of the approaches for realising descriptive analytics. It is a collection of tools that quantitatively describes the data in summary and graphical forms. In addition, such tools make computations that can be broken down into measures of central tendency or variability (spread) (Sarmento & Costa, 2017).

Descriptive statistics summarise the studied sample without drawing any inferences based on probability theory. Even if the primary aim of a study involves inferential statistics, descriptive statistics are still used to give a general summary. When the population is described using tools such as frequency distribution tables, percentages, and other measures of central tendency like the mean, we are talking about descriptive statistics (Yellapu, 2018).

The distribution of a variable in a dataset plays an essential role in data analytics. It shows all the possible values of the variable and the frequency of occurrence of each value. It is used to present a quantitative analysis of the data set. In a study, numerous variables are to be measured, and descriptive statistics are used to break down this massive amount of data into the simplest form. The variable's values distribution is depicted using a table or function. Measures of dispersion (e.g., variability) include minimum and maximum values, range, quantiles, and standard deviation/variance. The most commonly used measure of a tendency is the mean value. Conversely, the median is generally used when some values differ significantly from the rest.

Overall, descriptive statistics play a vital role at the time of data analysis as well as providing the foundation for comparing variables. Therefore, it is recommended as a good research practice to report the most suitable descriptive statistics with the help of a systematic approach to reduce the chances of presenting misleading results (Sharma, 2019).

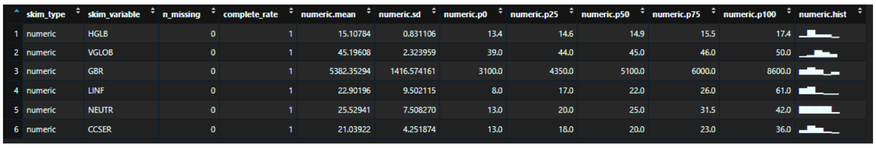
The following analysis and methodologies were performed with the programming language R, using the integrated development environment (IDE) RStudio.

To begin the preliminary data analysis, the CSV file was "called" into R Studio, using the function read_csv from the package "readr".

The data had a dimension of 51 rows (individuals) per 6 columns (variables). Regarding the structure of the data, every variable type was numeric.

Using the skim function from the "skimr" package, a complete description of the data was obtained (Table 1). In this table, it is possible to observe the inexistence of missing values, several descriptive statistics (mean, sd), the quantiles and a histogram for each variable.

Table 1 – Table with various descriptive statistics (mean, sd), the quantiles and a histogram for each variable.



	skim_type	skim_variable	n_missing	complete_rate	numeric.mean	numeric.sd	numeric.p0	numeric.p25	numeric.p50	numeric.p75	numeric.p100	numeric.hist
1	numeric	HGLB	0	1	15.10784	0.831106	13.4	14.6	14.9	15.5	17.4	
2	numeric	VGLOB	0	1	45.19608	2.323959	39.0	44.0	45.0	46.0	50.0	
3	numeric	GBR	0	1	5382.35294	1416.574161	3100.0	4350.0	5100.0	6000.0	8600.0	
4	numeric	LINF	0	1	22.90196	9.502115	8.0	17.0	22.0	26.0	61.0	
5	numeric	NEUTR	0	1	25.52941	7.508270	13.0	20.0	25.0	31.5	42.0	
6	numeric	CCSER	0	1	21.03922	4.251874	13.0	18.0	20.0	23.0	36.0	

When observing this table, it is noticed that there are different metrics with a considerate variation in terms of means, standard deviations, and minimum and max values. This is especially true for the variable "GBR", which stands out from all the others, presenting a much more extensive range. Moreover, in many situations, the variables under study are not measured in the same unit, on the same scale, or have very different variances. Thus, the need to establish a uniformization may arise, which can be achieved through the standardisation of the variables, i.e., by dividing the centred value of each variable by the corresponding standard deviation.

To better observe the distribution of the variables, the function Boxplot from the package "car" was used. This visualisation used scaled variables so that some interpretation could be taken. Otherwise, given the variation in the metrics range, it would be challenging to take insights. Through the boxplots (Figure 1), it is possible to observe the distribution of the variable through statistics such as median (centre of violin), 1st quartile (Q1 - 25%), third quartile (Q3 – 75%), the Inter Quartile Range (IQR) obtain by subtracting the Q3 by the Q1, and minimum and max values. In addition, Boxplots are a useful tool for detecting outliers. An outlier can be considered a data point that differs significantly from other observations. An outlier may be due to variability in the measurement, an indication of novel data, or the result of experimental error; the latter are sometimes excluded from the data set. Boxplots present outliers using the following rule: a data point is an outlier if it is more than 1.5*IQR above the Q3 or below the Q1. Said differently, low outliers are below Q1−1.5*IQR, and high outliers are above Q3+1.5* IQR.In Figure 1, the outliers are the circles with correspondent values, located beyond the depicted box plots. Applying the rule presented for outlier detection, the individuals considered outliers identified for each variable were: "HGLB": 12, 23; "VGLOB": 1, 12, 23, 38; "GBR": 23, "LINF": 31 45 47; "NEUTR": no outliers; "CCSER": 10. In addition all variables present a close to symmetric distribution with no skewness being visually detectable.

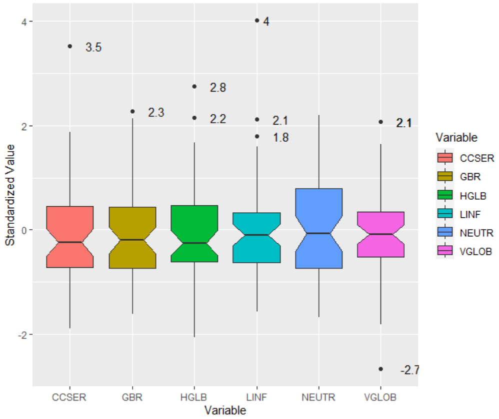


Figure 1 – Boxplots for each variable with outliers

Principal Component Analysis

Principal component analysis (P.C.A.) is a method of analysing multivariate data. It is considered a mathematical algorithm used to reduce the dimensionality of the data while maintaining the majority of its intrinsic variation. It accomplishes this reduction by identifying directions, called principal components (p.c's), along which the variation in the data is maximal (Ringnér, 2008).

Since P.C.A. does not invoke a statistical model, unlike other methods, where the parameters are estimated and are simply a data transformation into new dimensions, no assumptions are required to perform it. Still, certain conditions should be verified before it is applied. For example, the correlation or covariance matrix should contain enough correlation or covariance to make this method worthwhile. If, for instance, the correlation matrix is an identity, performing P.C.A. would be questioned (Shalabh, 2021).

At its core, P.C.A. identifies new variables, the p.c's, which are linear combinations of the original variables. Given that each component should reflect, as much as possible, the characteristics of the data, the one with the maximum variance is chosen from all linear combinations. For example, given two p.c's, the first would represent the direction along which the sample shows the most significant variation. In contrast, the second would be the direction uncorrelated (orthogonal) to the first p.c, along which the samples show the second largest variation (Ringnér, 2008). This process would be repeated if more p.c's existed.

P.C.A. can be performed on either the original raw data or standardised data. Standardisation is obtained by dividing each variable's centred value by the corresponding standard deviation. This procedure is equivalent to analysing the correlation matrix rather than the covariance matrix (Shalabh, 2021). Both matrixes define spread (variance) and orientation (covariance/correlation). From the chosen matrix, a representative vector (eigenvector) pointing in the direction of the largest spread of data and a corresponding number (eigenvalue) equal to the spread (variance) of that direction is obtained. The next step is sorting the eigenvectors in descending order according to their eigenvalues. After this, the first eigenvector will account for the largest spread among data, the second for the second largest and so on, under the already referred condition that all these new directions, which describe a new space, are independent (orthogonal) (Johnson & Wichern, 2013).

The meaning of a p.c. will be interpreted from the variables that most correlate with it. For this, the coefficients of the linear combinations and the correlations between the initial variables and the p.c's (loadings) can be used. If working with the correlation matrix, a new matrix containing the loading can be obtained by doing the correlation between the matrix containing the initial variables and the matrix containing the eigenvectors. An absolute value superior to or equal to .5 is considered acceptable when analysing the loadings (Johnson & Wichern, 2013).

Usually, the main problem in P.C.A.is how many components to keep. This decision is often made with the aid of some statistical tools. For example, it is possible to retain as many p.c's as necessary so that the percentage of variance explained by them is greater than a given value fixed a priori. The scree plot, a graphical device that plots each component's eigenvalues (variances), is another possible way. In this plot, the indicator is a "bend" or "elbow" point, which helps determine the number of components to retain (Shalabh, 2021). Another rule or tool used is Kaiser's Criterion. Kaiser's criterion's logic consists in maintaining components with eigenvalues greater than one, given that any component having an eigenvalue higher than one is responsible for more than the mean total variance, given that the average variance is equal to 1 if referring to the correlation matrix. This is not necessarily the case in a covariance matrix. Nevertheless, the rule of retaining components that exhibit a variance greater than the average can be applied, whatever this average is (Shalabh, 2021). Both the mean and the scree-plot criteria are the most commonly used. The practice has shown that these criteria lead to credible solutions if at least one of the following conditions is met: number of variables less than 30, number of cases (individuals) greater than 250. According to some authors, when the number of variables is greater than 30 (especially if it is greater than 50), a scree-plot must be used in detriment to the mean criterion (Johnson & Wichern, 2013).

Graphical representations are a great auxiliary in interpreting P.C.A. results, with biplots being one the most used methods of visualisation. This method overlays the scores and loadings plots (appropriately scaled) for two p.c's, so that sample-variable relationships are more pronounced. Each axis contains one p.c., and each variable is represented with an arrow pointing to the direction of the correlation. The size of the arrow represents the eigenvalue. If two arrows are perpendicular, these two variables are orthogonal (independent). On the other hand, if they have a small angle between them, they are highly correlated. Although this can be a practical approach for small systems with few samples, variables, and p.c's, it can be very complex for megavariate data sets. Furthermore, the plots are only feasible for two or three dimensions in straightforward cases, so examining multiple pairs may be necessary when various p.c's are considered (Ivosev et al., 2008).

P.C.A. Application

From the previously acquired knowledge, it was decided to conduct a correlation based P.C.A. since the measurement units are not always of the same nature and the mean and standard deviation values observed for each variable vary considerably.

The correlation matrix was obtained using the corrplot function from the "corrplot" package to analyse the correlations between variables (Figure 2). Generally, low correlation but positive, were observed, being the highest correlations for "VGLOB" with "HGLB" and for "LINF" with "GBR. It was observed that "CCSER" had overall very low correlations with the rest of the variables.

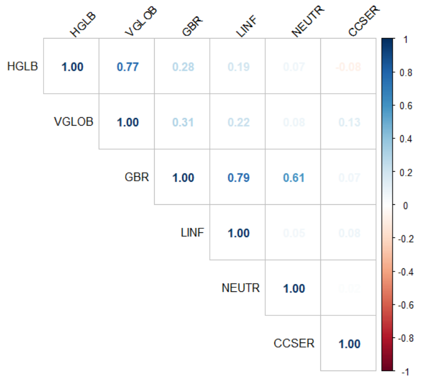
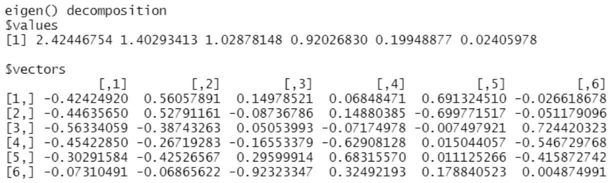


Figure 2 – Correlation matrix

The next step was to obtain the eigenvalues and eigenvectors based on the correlation matrix. These were obtained using the native from R function Eigen and can be seen in (Table 2).



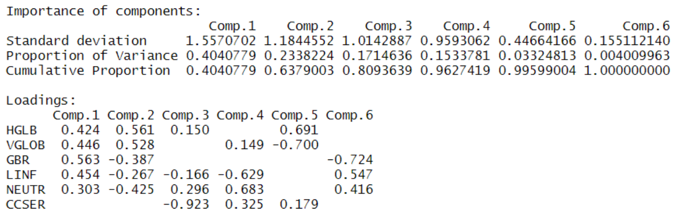
```
eigen() decomposition
$values
[1] 2.42446754 1.40293413 1.02878148 0.92026830 0.19948877 0.02405978

$vectors
            [,1]        [,2]        [,3]        [,4]         [,5]         [,6]
[1,] -0.42424920  0.56057891  0.14978521  0.06848471  0.691324510 -0.026618678
[2,] -0.44635650  0.52791161 -0.08736786  0.14880385 -0.699771517 -0.051179096
[3,] -0.56334059 -0.38743263  0.05053993 -0.07174978 -0.007497921  0.724420323
[4,] -0.45422850 -0.26719283 -0.16553379 -0.62908128  0.015044057 -0.546729768
[5,] -0.30291584 -0.42526567  0.29599914  0.68315570  0.011125266 -0.415872742
[6,] -0.07310491 -0.06865622 -0.92323347  0.32492193  0.178840523  0.004874991
```

Table 2 – Eigenvalues and Eigenvectors for 6 components

Considering the Kaiser's criterion, the three first p.c's were chosen since their eigenvalues are greater than 1 (2.424, 1.403, 1.029). Still, further analysis is required. This criterion alone is not sufficient.

From the previous function, a general notion of our data behaviour was obtained. Therefore, the P.C.A. was applied for which the function pincomp from the package "stats" was used. The output of this analysis is presented in Table 3. From this figure, various observations can be made. In addition to the eigenvalues already known from the previous step, it was possible to observe the proportion of variance that each of the components explained as well as the already known eigenvectors. To inform that the blank spaces in the eigenvectors corresponding to coefficients considered extremely low, therefore omitted. The importance of each p.c. is displayed in descending order according to each component's eigenvalue. Considering the previous application of Kaiser's criterion, where the first three principal components were considered, it was possible to observe that they explained 80.9% of the total variance present in the data, which is regarded as high. The proportion of the (standardised) population standard deviation (sd) due to each p.c. can be obtained by the correspondent eigenvalue by the total variance, which in case, corresponds to the number of variables given that the correlation matrix is being used. The proportions of the (standardised) population sd in were as follows: p.c.1(60%), p.c.2(48%), p.c. 3(41%). Since the correlation matrix was used, this value was obtained by dividing the square root of the eigenvalue by the number of dimensions.



```
Importance of components:
                          Comp.1    Comp.2    Comp.3    Comp.4     Comp.5      Comp.6
Standard deviation     1.5570702 1.1844552 1.0142887 0.9593062 0.44664166 0.155112140
Proportion of Variance 0.4040779 0.2338224 0.1714636 0.1533781 0.03324813 0.004009963
Cumulative Proportion  0.4040779 0.6379003 0.8093639 0.9627419 0.99599004 1.000000000

Loadings:
      Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6
HGLB   0.424  0.561  0.150         0.691       
VGLOB  0.446  0.528         0.149 -0.700       
GBR    0.563 -0.387                      -0.724
LINF   0.454 -0.267 -0.166 -0.629         0.547
NEUTR  0.303 -0.425  0.296  0.683         0.416
CCSER               -0.923  0.325  0.179       
```

Table 3 – P.C.A. output

To support the Kaiser's criterion, a scree plot was performed through the function fviz_screeplot from the package "factoextra" (Figure 3). The scree plot is a graphic that shows the explained variance per newly defined p.c. The measure of the plot is the percentage of the explained variance. When observing the scree plot, it is not clear how many p.c's to maintain. An elbow point is visible where the number of dimensions equals two. Despite this, since the first two p.c's only explain 60% of the variance and the Kaiser's criterion indicates that three p.c's should be retained, increasing the variance explained to 80%, three p.c's were considered as the adequate option.

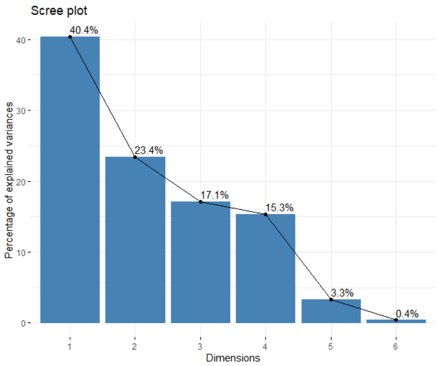
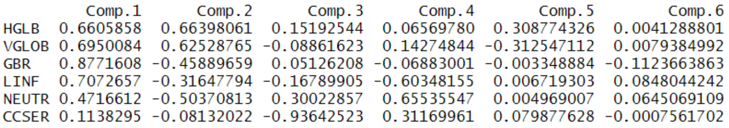


Figure 3 – Scree Plot

In order to interpret each p.c, the loading matrix was obtained (Table 4). As mentioned in the theoretical introduction to this methodology, this matrix can be obtained by doing the correlation between the matrix containing the initial variables and the matrix containing the eigenvectors. An absolute value superior to or equal to .5 is considered acceptable when analysing the loadings.



```
         Comp.1      Comp.2      Comp.3      Comp.4       Comp.5        Comp.6
HGLB  0.6605858  0.66398061  0.15192544  0.06569780  0.308774326  0.0041288801
VGLOB 0.6950084  0.62528765 -0.08861623  0.14274844 -0.312547112  0.0079384992
GBR   0.8771608 -0.45889659  0.05126208 -0.06883001 -0.003348884 -0.1123663863
LINF  0.7072657 -0.31647794 -0.16789905 -0.60348155  0.006719303  0.0848044242
NEUTR 0.4716612 -0.50370813  0.30022857  0.65535547  0.004969007  0.0645069109
CCSER 0.1138295 -0.08132022 -0.93642523  0.31169961  0.079877628 -0.0007561702
```

Table 4 – Loading's Matrix

From the loading matrix, the variables that contributed the most to each component and their explanation can be defined. For the first p.c., these variables were "GBR”, “LINF", and "VGLOB". The contribution of the important variables for each p.c, obtained by the formula present in Equation 1, was as follows: 0.317 ("GBR"), 0.206 ("LINF") and 0.199 ("VGLOB"). The second p.c. was explained by "HGLB" (0.314) and "NEUTR" (0.180). Lastly, the third component was explained by "CCSER" (0.852).

$$a_{ij}^2 = \left(\frac{l_{ij}}{\sqrt{\lambda_j}}\right)^2$$

Equation 1 – Loading divided by the square root of eigenvalue to the power of 2.

The p.c's behaviour was visualised with biplots (Figure 4) aid to consolidate the presented analysis further. As explained in the introduction to this method, given that three p.c's were considered, more than one Biplot was necessary. Below, the graphical representations of all the combinations of the p.c's are presented, which were obtained using Biplots through the function ggbiplot from the package "ggplot2":

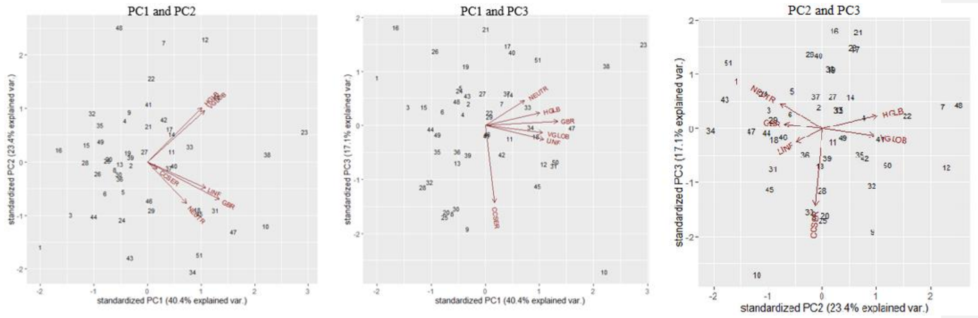


Figure 4 – Biplots for p.c’s combinations. From left to right: p.c.1- p.c.2; p.c.1- p.c.3; p.c.2- p.c.3

When analysing the biplots, first of all, the division of the variables by each p.c. could be confirmed. As explained in the theoretical part of this methodology, the lengths of each of the arrows are proportional to the variability associated with each p.c. and show how well the attributes of the data are explained by the p.c's. In this study, it's possible to observe that CCSER was the variable that stood out the most and was well illustrated by the third p.c. It can also be inferred that "LINF" is strongly correlated to "GBR2 (0.789) and not well correlated with other variables. Moreover, "HGLOB" and "VGLOB" are also strongly correlated, as was previously seen in the correlation matrix (0.77). On the other hand, variables whose vectors have angles with degrees closer to 90º, such as "HGLB" or "VGLOB" with "NEUTR" or "CCSER", signifying that these variables are not correlated (orthogonal). These observations are aligned with the correlation matrix present in Figure 2.

Another valuable observation is that all variables positively correlate with p.c. 1, with "CCSER" having a minimal correlation. For p.c. 2, "HGLB" and "VGLOB" are positively correlated. "NEUTR", "GBR", and "LINF" are negatively correlated, with "NEUTR" being possibly the strongest correlation, "LINF" the smallest and "CCSERT" showing little correlation. For p.c. 3, "CCSER" is strongly and negatively associated but barely correlated with all the others.

P.C.A. is also useful for detecting outliers. Observing the biplots makes it possible to visualise some in each of them. In the Biplot of p.c. 1 with p.c. 2, the following observations were identified: 7, 12, 23, 38 and 48, which can have high levels of "VGLOB" or "HGLB" and observation 1 with probably low levels. In the Biplot of p.c. 2 with p.c. 3, observation 10 stands out with high "CCSER". Finally, in the Biplot of p.c. 1 with p.c. 3, observation 16 is likely to hold shallow values in all variables and somewhat high in regard to "CCSER", while observations 23 and 38 may have very high values for "NEUTR". From crossing the outlier information obtained in the boxplots with the biplots, it is possible to confirm that for "HGLB", 12 (16.9) and 23 (17.4) have high values. For “VGLOB”, 1(39) has the lowest level, 12 (50), 23(50) and 38(50) with the maximum observed level. And for CCSER, observation 10 (36) has the maximum value.

From this P.C.A. process, it was possible to go from 6 variables to characterise this problem to half of its original dimension, with a considerably high value of the total explained variance of 80.9%. This can be valuable to further studies or better interpret the data. Outliers were detected, which should be considered and analysed by the experts. Possible profiles associated with each component could also be inferred. The three new linear combinations (p.c’s) that resulted from the transformation performed on the variables under study is observed in Figure 5. From these new values can be obtain for each individual.

$$P.C.1 = -0.424HGLB - 0.446VGLOB - 0.263GBR - 0.454LINF - 0.303NEUTR - 0.073CCSER$$
$$P.C.2 = 0.561HGLB + 0.528VGLOB - 0.387GBR - 0.267LINF - 0.425NEUTR - 0.069CCSER$$
$$P.C.3 = 0.149HGLB - 0.087VGLOB + 0.051GBR - 0.165LINF + 0.295NEUTR - 0.923CCSER$$

Figure 5 – Three linear combinations of the original variables (p.c.1, p.c.2, p.c.3).

Cluster Analysis

Cluster Analysis (C.A.) is a multivariate statistical technique that identifies groups or clusters within the data (Madhulatha, 2012). Each individual inside a given cluster is more "similar" to others in another cluster (Domany, 2003), making it possible to create patterns within the data. Clustering is a method of exploratory analysis and reduction of the dimensionality of the data (Xia et al., 2021) in which the groups can replace the actual individuals of the data.

Because of these aspects, similarity measures are fundamental components in most clustering algorithms (Jain et al., 1999). Similarities are rules that serve as criteria for grouping or separating items (Saraçli et al., 2013). These distances (similarities) can be based on a single dimension or multiple dimensions, with each dimension representing a rule or condition for grouping objects (Hill et al., 2006).

The most popular way to evaluate a similarity measure is distance measures. The most widely used distance measure is the Euclidean distance which is the distance between points in a straight line (Suwanda et al., 2020). This distance method uses the Pythagorean theorem and is the distance calculation most often used in machine learning (Hamerly & Elkan, 2002). Euclidean Distance formula results from the square root of the distance between each variable, summing the squares and finding the square root of that sum (Madhulatha, 2012; Omran et al., 2007).

$$d_{\text{euclidean}}(\mathrm{x},\mathrm{y}) = \sqrt{\sum_{\mathrm{i}=1}^{\mathrm{n}} (\mathrm{x}_{\mathrm{i}}\text{-}\mathrm{y}_{\mathrm{i}})^2}$$

Figure 6 – Equation of the Euclidean distance

Where dij corresponds to the similarity calculation distance, n to the length of vectors (?) and x, y to the vectors (?).

Since clustering essentially groups closer elements and distant ones separately, if a scaling factor is not applied to the features, more weight will be applied to some of the features compared to others. Because of this factor, it is deemed good practice to scale the data provided (Madhulatha, 2012). In certain situations, it is advisable not to consider all features that show high correlations when applying the Euclidean distance, which can help with the weight each feature will have in the final result (Omran et al., 2007). Highly correlated variables will have double the weight when computing the distance between two points (since all the variables are normalised, the effect will usually double). In short, the variable's strength to influence the cluster formation increases if it has a high correlation with any other variable (Everitt, 2011).

There are several data clustering algorithms, but for this project, only Hierarchical clustering algorithms will be discussed. Briefly, Hierarchical clustering finds successive clusters using previously established ones (Madhulatha, 2012). The main objective of hierarchical classification is to build a tree, whose two-dimensional graphic representation is called a dendrogram, tree diagram or hierarchical tree, where data elements stored at the leaves will be merged two by two to the "closest" sub-sets (stored at nodes) until the root of the tree is reached, which contains all the elements of the dataset (Nielsen, 2016). Hierarchical algorithms can be divided into two approaches: agglomerative (bottom-up) or divisive (top-down) (Sasirekha & Baby, 2013). Agglomerative algorithms begin with each element as a separate cluster and merge them into successively larger clusters. Divisive algorithms begin with the whole set and divide it into successively smaller clusters. In this project, only agglomerative algorithms will be tested.

The process behind the agglomerative approach is the following: one starts with each object forming a separate group. Then, the objects that are close to one another are joined in a cluster. It keeps on doing so until all the groups are merged into one cluster or until a termination condition holds (Reddy et al., 2017).

There are several agglomerative hierarchical algorithms like single-linkage, complete linkage, ward linkage and average linkage.

Single-linkage clustering groups *n* objects into a system of sets (clusters) of similar objects such that, for any given level Δ of clustering intensity, objects I and j are placed in the same set if and only if the resulting set satisfies the single-link criterion (which considers the dissimilarity between a pair of objects). Two objects i and j belong to the same single-link cluster at a clustering level Δ if a chain of size is less than or equal to Δ connecting them. Summarising, it is the shortest distance between a pair of observations in two clusters. As a result, it can sometimes produce clusters where observations in different clusters are closer together than observations within their own clusters (Everitt, 2011). The advantages of the single-linkage method are its speed and its ability to perform well on non-globular data. It can also differentiate between non-elliptical shapes as long as the gap between the two clusters is not small (Przewozniczek & Komarnicki, 2020). The disadvantage of this method is its inability to separate clusters properly if there is noise between the clusters (Przewozniczek & Komarnicki, 2020).

Complete linkage clustering (Gordon, 1999) is the distance measured between two clusters' farthest pair of observations. This method usually produces tighter clusters than single-linkage, but these can end up very close together. Complete linkage also tends to favour larger clusters, which leads to a poor relationship between the clusters and the data structure in some cases (Murtagh, 1985). A disadvantage is that it tends to break large clusters (Murtagh, 1985).

In ward linkage, the dissimilarity between the clusters is calculated based on the distances between the clusters' centroids. A larger distance magnitude represents a higher dissimilarity value between the clusters (Han et al., 2006). When spherical multivariate normal distributions are used, Ward's method is excellent (Kuiper and Fisher, 1975) because this method is based on a sum of squares criterion. However, it must be noted that Ward's method only performs well if an equal number of objects is drawn from each population (Kaufman & Rousseeuw, 1990).

It also has difficulties with clusters of unequal diameters. Moreover, Everitt (1977) demonstrated that Ward's method often leads to misclassifications when the clusters are distinctly ellipsoidal rather than spherical, that is when the variables are correlated within a cluster.

The average linkage is the distance between each pair of observations in each cluster that are added up and divided by the number of pairs to get an average inter-cluster distance (Eisen et al., 1998). One advantage of the average Linkage method is that it works well in separating clusters if there is any noise between the clusters (Zelig & Kaplan, 2020). As for the disadvantage, the method is biased towards globular clusters (Zelig & Kaplan, 2020).

Cluster Application

The first step in performing a hierarchical clustering analysis is calculating the distance matrix, which can be done using various distances. Since the data had different measures and some of the variables' mean or standard deviation had several degrees of deviation, data scaling was applied to the data before calculating the distance matrix. This was achieved through the scale function inherent to R.

After scaling the data, the distance matrix could be calculated. Despite the existence of a high correlation between some of the variables in the data, which is not recommended for some of the distances, the Euclidean distance was still chosen since the data was previously scaled. To calculate the distance matrix, the dist function of R was used with method set to euclidean.

Once the distance matrix was calculated, clustering could be performed through the use of the agnes function from the “cluster” library. As previously mentioned, several methods can be utilised (single linkage, complete linkage, average linkage, ward's method). All these methods were tested, by changing the method parameter of the function accordingly, and then compared.

After clustering with the different methods, $merge can be used to obtain the sequential clustering steps applied by the techniques. The output of the methods can be seen in Figure 7.

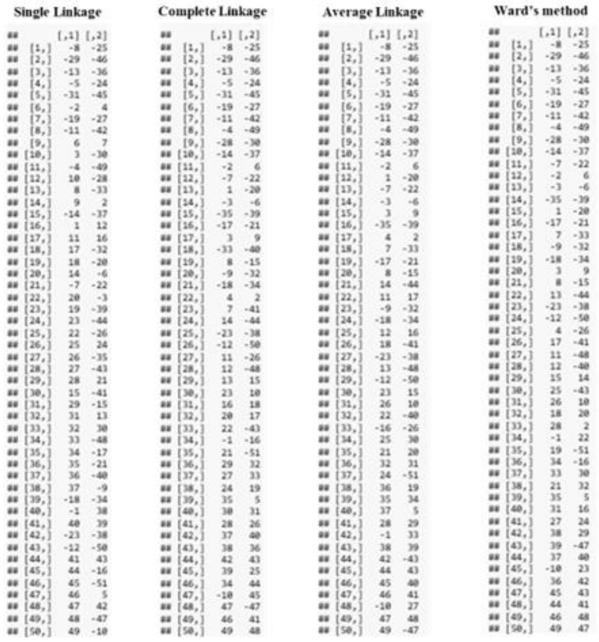
It's important to note that negative numbers indicate individuals, while positive numbers indicate clusters. As such, when two individuals are bundled together, they form a new cluster, while when an individual and a cluster are joined, it means the individual joins an already-formed cluster. This process occurs sequentially until a single global cluster is obtained in the end, as agglomerative methods are being applied.

For example, regarding single linkage, in the first iteration, individuals 8 and 25 were bundled in cluster 1. In the second iteration, individuals 29 and 46 were joined in cluster 2.

A comparison between the sequential formation of the clusters using the single linkage and complete linkage reveals that the first five steps are identical between the two methods, after which clear differences can be seen in the forthcoming steps until the end.

Regarding the average linkage, differences in cluster formation, compared to single linkage, start to appear in the 6th step. On the other hand, differences between average linkage and complete linkage are only seen in the 12th step.

Lastly, for the ward's method, differences in the sequential formation of clusters, when compared to the single linkage, are seen in the 6th step. Compared to the other two methods (complete and average linkage), the first difference occurs at step 11.



Single Linkage

```
##       [,1] [,2]
##  [1,]   -8  -25
##  [2,]  -29  -46
##  [3,]  -13  -36
##  [4,]   -5  -24
##  [5,]  -31  -45
##  [6,]   -2    4
##  [7,]  -19  -27
##  [8,]  -11  -42
##  [9,]    6    7
## [10,]    3  -30
## [11,]   -4  -49
## [12,]   10  -28
## [13,]    8  -33
## [14,]    9    2
## [15,]  -14  -37
## [16,]    1   12
## [17,]   11   16
## [18,]   17  -32
## [19,]   18  -20
## [20,]   14   -6
## [21,]   -7  -22
## [22,]   20   -3
## [23,]   19  -39
## [24,]   23  -44
## [25,]   22  -26
## [26,]   25   24
## [27,]   26  -35
## [28,]   27  -43
## [29,]   28   21
## [30,]   15  -41
## [31,]   29  -15
## [32,]   31   13
## [33,]   32   30
## [34,]   33  -48
## [35,]   34  -17
## [36,]   35  -21
## [37,]   36  -40
## [38,]   37   -9
## [39,]  -18  -34
## [40,]   -1   38
## [41,]   40   39
## [42,]  -23  -38
## [43,]  -12  -50
## [44,]   41   43
## [45,]   44  -16
## [46,]   45  -51
## [47,]   46    5
## [48,]   47   42
## [49,]   48  -47
## [50,]   49  -10
```

Complete Linkage

```
##       [,1] [,2]
##  [1,]   -8  -25
##  [2,]  -29  -46
##  [3,]  -13  -36
##  [4,]   -5  -24
##  [5,]  -31  -45
##  [6,]  -19  -27
##  [7,]  -11  -42
##  [8,]   -4  -49
##  [9,]  -28  -30
## [10,]  -14  -37
## [11,]   -2    6
## [12,]   -7  -22
## [13,]    1  -20
## [14,]   -3   -6
## [15,]  -35  -39
## [16,]  -17  -21
## [17,]    3    9
## [18,]  -33  -40
## [19,]    8  -15
## [20,]   -9  -32
## [21,]  -18  -34
## [22,]    4    2
## [23,]    7  -41
## [24,]   14  -44
## [25,]  -23  -38
## [26,]  -12  -50
## [27,]   11  -26
## [28,]   12  -48
## [29,]   13   15
## [30,]   23   10
## [31,]   16   18
## [32,]   20   17
## [33,]   22  -43
## [34,]   -1  -16
## [35,]   21  -51
## [36,]   29   32
## [37,]   27   33
## [38,]   24   19
## [39,]   35    5
## [40,]   30   31
## [41,]   28   26
## [42,]   37   40
## [43,]   38   36
## [44,]   42   43
## [45,]   39   25
## [46,]   34   44
## [47,]  -10   45
## [48,]   47  -47
## [49,]   46   41
## [50,]   49   48
```

Average Linkage

```
##       [,1] [,2]
##  [1,]   -8  -25
##  [2,]  -29  -46
##  [3,]  -13  -36
##  [4,]   -5  -24
##  [5,]  -31  -45
##  [6,]  -19  -27
##  [7,]  -11  -42
##  [8,]   -4  -49
##  [9,]  -28  -30
## [10,]  -14  -37
## [11,]   -2    6
## [12,]    1  -20
## [13,]   -7  -22
## [14,]   -3   -6
## [15,]    3    9
## [16,]  -35  -39
## [17,]    4    2
## [18,]    7  -33
## [19,]  -17  -21
## [20,]    8  -15
## [21,]   14  -44
## [22,]   11   17
## [23,]   -9  -32
## [24,]  -18  -34
## [25,]   12   16
## [26,]   18  -41
## [27,]  -23  -38
## [28,]   13  -48
## [29,]  -12  -50
## [30,]   23   15
## [31,]   26   10
## [32,]   22  -40
## [33,]  -16  -26
## [34,]   25   30
## [35,]   21   20
## [36,]   32   31
## [37,]   24  -51
## [38,]   36   19
## [39,]   35   34
## [40,]   37    5
## [41,]   28   29
## [42,]   -1   33
## [43,]   38   39
## [44,]   42  -43
## [45,]   44   43
## [46,]   45   40
## [47,]   46   41
## [48,]  -10   27
## [49,]   47   48
## [50,]   49  -47
```

Ward's method

```
##       [,1] [,2]
##  [1,]   -8  -25
##  [2,]  -29  -46
##  [3,]  -13  -36
##  [4,]   -5  -24
##  [5,]  -31  -45
##  [6,]  -19  -27
##  [7,]  -11  -42
##  [8,]   -4  -49
##  [9,]  -28  -30
## [10,]  -14  -37
## [11,]   -7  -22
## [12,]   -2    6
## [13,]   -3   -6
## [14,]  -35  -39
## [15,]    1  -20
## [16,]  -17  -21
## [17,]    7  -33
## [18,]   -9  -32
## [19,]  -18  -34
## [20,]    3    9
## [21,]    8  -15
## [22,]   13  -44
## [23,]  -23  -38
## [24,]  -12  -50
## [25,]    4  -26
## [26,]   17  -41
## [27,]   11  -48
## [28,]   12  -40
## [29,]   15   14
## [30,]   25  -43
## [31,]   26   10
## [32,]   18   20
## [33,]   28    2
## [34,]   -1   22
## [35,]   19  -51
## [36,]   34  -16
## [37,]   33   30
## [38,]   21   32
## [39,]   35    5
## [40,]   31   16
## [41,]   27   24
## [42,]   38   29
## [43,]   39  -47
## [44,]   37   40
## [45,]  -10   23
## [46,]   36   42
## [47,]   45   43
## [48,]   44   41
## [49,]   46   48
## [50,]   49   47
```

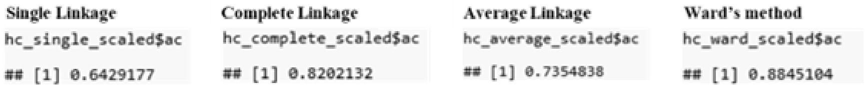
Figure 7 – Sequential cluster formation steps for each employed method.

From this, it's possible to assess that the single linkage method is the furthest from all the others since they all diverge from it at the 6th step. Between the remaining techniques, differences occur around the same step (11/12). Considering the ways clustering is performed in these methods, this could mean that when using the smallest distance between instances (single linkage), greater differences in the cluster formation are seen when compared to all the other metrics (larger distance between instances for complete linkage, average distance between individuals of a cluster for average linkage and distance between cluster's centroids for ward's method).

Further information that can be extracted is the order of the objects on the clustering tree by using $order. This is not very informative and, as such, was only performed once for single linkage. From this command, it's possible to assess that the elements of the dendrogram, from left to right, should follow the mentioned order (1, 2, 5, 24, etc.).

```
[1]  1  2  5 24 19 27 29 46  6  3 26  4 49  8 25 13 36 30 28 32 20 39 44 35 43 7 22 15 11 42 33
14 37 41 48 17 21 40  9 18 34 12 50 16 51 31 45 23 38 47 10
```

An important metric to evaluate the quality of the resulting clustering is the Agglomerative Coefficient (AC). The AC describes the strength of the clustering structure that has been obtained by the employed method. The coefficient takes values from 0 to 1, and it is the mean of the normalised lengths at which the clusters are formed. The closest to one the coefficient is, the better. This coefficient can be obtained using $ac for all the methods.



Single Linkage

```
hc_single_scaled$ac

## [1] 0.6429177
```

Complete Linkage

```
hc_complete_scaled$ac

## [1] 0.8202132
```

Average Linkage

```
hc_average_scaled$ac

## [1] 0.7354838
```

Ward's method

```
hc_ward_scaled$ac

## [1] 0.8845104
```

Figure 8 – Agglomerative Coefficient of each of the tested hierarchical clustering methods.

For single linkage, the AC is not very high, which means that the obtained clustering structure is not very good. However, the AC of the clustering structure saw significant improvement when complete linkage was applied. For average linkage, the AC lies in the middle of the single and complete linkage scores. Lastly, the AC is the highest of any tested methods for the ward's method. This is indicative that the ward method may be the best for this data set. Nevertheless, further exploration of each of the methods may be beneficial to analyse the formed dendrogram.

The clustering structure obtained can be represented as a dendrogram using the functions as.dendrogram, from the "stats" library, and plot, inherent to R. The dendrogram obtained for the single linkage method can be seen below:

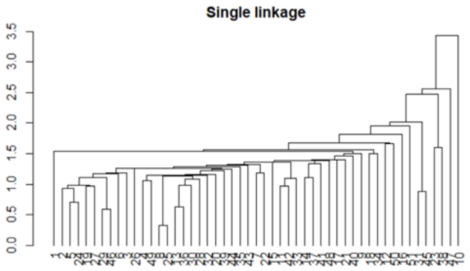


Figure 9 – Obtained dendrogram plot using the single linkage hierarchical method.

Even though three principal components were obtained using PCA, which could be a good starting point for the number of clusters, it is possible to see that a division into four clusters makes sense due to the sudden increase in the dendrogram distance between the last four individuals and the other elements. Therefore, these 4 clusters can be represented as rectangles in the dendrogram plot through the use of the rect.hclust function of the "stats" library, with k set to 4 and border set to 2:4.

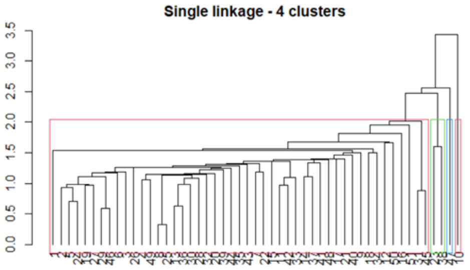


Figure 10 – Dendrogram plot of the single linkage hierarchical method with four clusters.

Two separate representations were performed for each of the remaining methods: (1) using a total of 4 clusters, to compare with the single linkage dendrogram; (2) using a deemed appropriate number of clusters for each method.

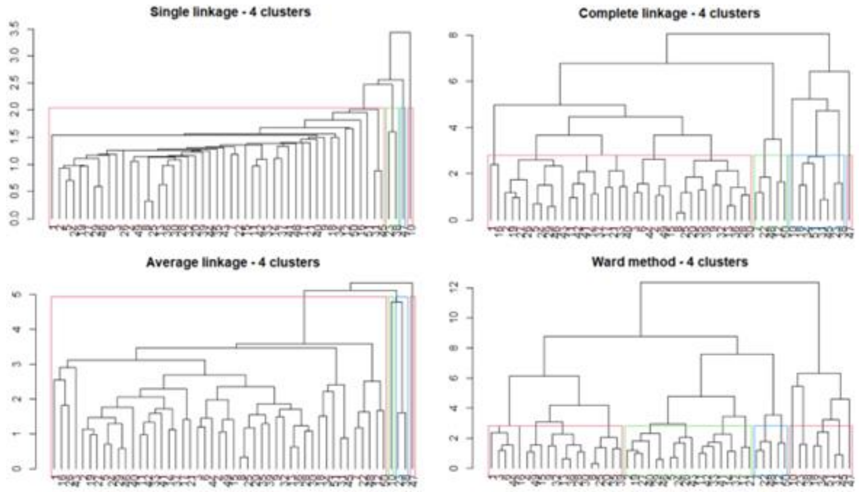


Figure 11 – Obtained dendrogram plots for each of the tested hierarchical methods considering four clusters.

The dendrogram plot shows that the single linkage method appears to form unbalanced clusters, with one cluster being formed by 47 individuals, while four elements only form the remaining 3 clusters.

Regarding complete linkage, when comparing the clusters obtained using this method vs the single linkage method, it is possible to see that despite one of the clusters still having most of the objects, the remaining 3 clusters now hold 14 objects instead of 4. This is indicative that the clusters are more balanced, which is generally agreed as better. Furthermore, better separation between clusters is seen using complete linkage, evidenced by the larger distance between clusters, which can also be seen in the y-scale, which goes from a maximum of 3.5 to a maximum of 8.

For average linkage, it is possible to visualise that the resulting clusters are similar to those obtained using the single linkage method, which indicates that the clusters are not as balanced as the ones seen in the complete linkage method. Furthermore, the worse separation between clusters is also seen, as the height differences between clusters are smaller, and the maximum of the y-scale is 5.

Finally, for the ward's method, it is possible to see that they are the most balanced clusters obtained so far, even better than the ones obtained through complete linkage. Instead of having a single cluster that contains the majority of the objects, two extremely similar clusters are formed with two slightly smaller ones. On top of this, the separation between clusters is better than any of the other methods, also supported by the maximum of the y-scale (12).

Despite this useful comparison, from the dendrogram plots, it is clear that a distinct number of clusters could be considered for methods other than single linkage. This is supported by the distance between groups of individuals, which allows clear separation into more than 4 clusters. For complete linkage, average linkage and ward's method, a total of 8, 7 and 10 clusters were considered, respectively. The dendrogram plots were obtained by changing the parameter k of the function rect.hclust.

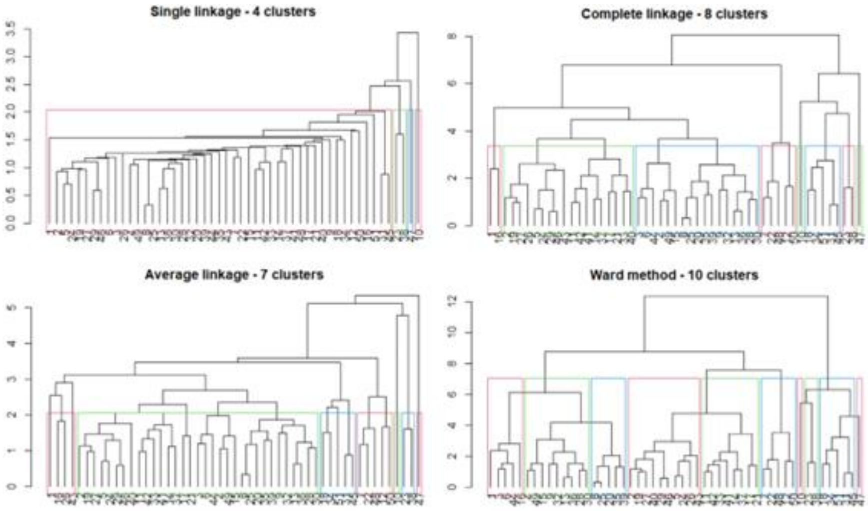


Figure 12 – Obtained dendrogram plots for each of the tested hierarchical methods considering four, eight, seven and ten clusters for single linkage, complete linkage, average linkage, and Ward's method respectively.

These dendrogram plots show clear differences between each method. Generally, the separation between clusters in crescent order is single linkage < average linkage < complete linkage < ward method. Better separation is directly correlated with the number of different clusters that can clearly be identified. As previously mentioned, for smaller numbers of clusters, their balance also varies in the same order, with single linkage producing the most unbalanced clusters, while Ward's method produces the most balanced.

When coupled with the previously obtained AC results, it can be concluded that the ward's method is the better hierarchical clustering method for the provided dataset. Not only does it provide the best AC score of all the methods, but it also gives better cluster separation and allows the distinction of more clusters in the dataset. Without further knowledge of the underlying groups in the data, the mentioned reasons support that this method is the best.

In order to better understand the possible meaning of each cluster, a plot was created using the ggplot function of the "ggplot2" library. For this, the ward method with 4 clusters was initially chosen. When looking at the plot, it is possible to observe four bar graphs, one for each cluster. Given the considerable variation between metrics, a normalised plot was considered for interpretation purposes. Each graph contains six bars representing each variable. Because of the standardisation, each bar behaves almost as a box plot showing its variance around a midpoint, in this case, the mean. With this visualisation, it is possible to understand the behaviour of the variables inside each cluster.

Cluster 1 tends to have lower values for all variables besides "CCSER". Cluster 2 presents a tendency for higher values of "NEUTR" and lower values of "CCSER" and "LINF". Cluster 2 also presents a high variation, as "VGLOB" and "HGBL" are observed with high and low values. Cluster 3 presents values around the mean for "CCER", "GBR", and "LINF". Cluster 3 also shows a tendency for high values of "VGLOB and "HGLB". Cluster 4 has a tendency for overall high values, with only "CSSER" presenting low values. Furthermore, since cluster 1 and 4 have similar high variation of the variable “CCSER” it is safe to assume that this variable does not contribute to the formation of these clusters, since the discerning factor lies in the values of the remaining variables.

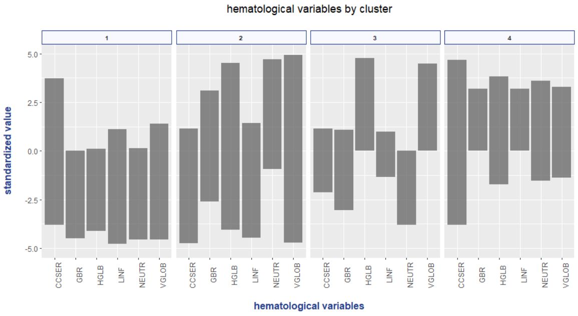


Figure 13 – Analysis of each variable values for every cluster, considering a total of four clusters.

To finalize, the analysis plot was also performed taking into account the separation achieved by the Ward’s method, through the consideration of 10 clusters.

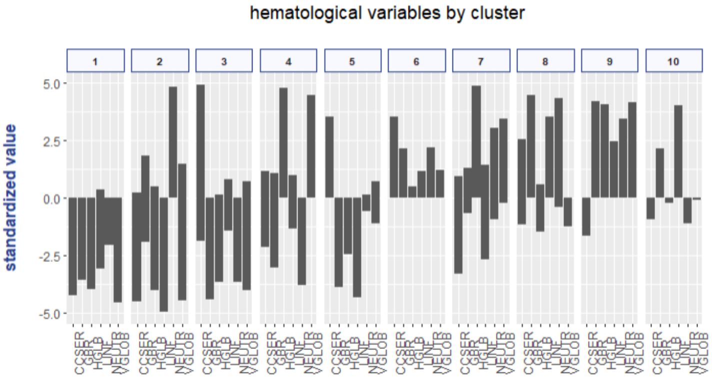


Figure 14 – Analysis of each variable values for every cluster, considering a total of 10 clusters.

From the presented plot, it is possible to see that a greater difference between each cluster can now be identified. From cluster 1 to 5, the majority of the variables have a tendency for low values, while from cluster 6 to 10, the majority of the variables have a propensity for high values.

Cluster 1 tends to have low values for every variable. Cluster 2 leans for low values on “CCSER”, “HGLB”, “LINF” and “VGLOB”, slightly high and low values on “GBR” and values around the mean on “NEUTR”. For cluster 3, high values are seen for “CCSER”, values around the mean for “HGLB” and low values for the remaining variables.

In cluster 4, high values are registered for “HGLB” and “VGLOB”, low values for “CCSER”, “GBR” and “NEUTR” and values around the mean for “LINF”. Regarding cluster 5, only “CCSER” tends to have high values, “GBR”, “HGLB” and “LINF”. Cluster 6 has a tendency for high values of “CCSER”, slightly less high values of “GBR” and “NEUTR” and values around the mean for the remaining variables. For cluster 7, high values are seen for “HGLB”, “NEUTR” and “VGLOB”, low values are seen for “CCSER” and “LINF” and values around the mean are seen for “GBR”. In cluster 8, high values are registered for all variables except “HGLB” and “VGLOB”, which tend to have values around the mean. Moving onto cluster 9, all variables tend to have high values, with the exception of “CCSER”, which tend to have values around the mean. Lastly, in cluster 10 a tendency for high values of “LINF” is seen, with “GBR” having a tendency for slightly less high values. The remaining variables tend to have values around the mean.

The main conclusion from these analysis plots is that an increase in the number of clusters can help increase the differentiation between clusters. Nevertheless, even when considering 4 clusters, there were identifiable differences between the behaviour of the variables.

Comparison of P.C.A. and C.A.

When comparing both methodologies and results, from a visual point of view no resemblances can be salient between the three p.c’s and the ten clusters. Regarding three p.c’s vs four clusters, some comparisons can be made.

Two comparisons can be drawn from analysing the results of P.C.A. and C.A. methodologies. Firstly, the first p.c. has a positive correlation with and is generally explained by all the variables, with the exception of “CCSER”. This behaviour is extremely similar to cluster 4, where, if “CCSER” is excluded (it was already mentioned that this variable does not seem to be a deciding factor for this cluster formation), the variables tend to have high values. Secondly, the second p.c. is mainly explained by “HGLB” and “VGLOB”, which have a positive correlation with it. A comparable behaviour to this p.c. can be seen in cluster 3, whose variables with high values are exactly the same.

Regarding cluster 3, which was mainly explained by “CCSER”, no direct comparable cluster could be identified, since, as previously mentioned, for the cluster 1 and 4, which have high variation of “CCSER”, the similar behaviour of this variable seems to exclude it as one of the major contributors of these clusters’ formation.

It is worth remembering that P.C.A. and C.A. are two completely different methodologies for exploratory analysis. As such, it is not guaranteed that their results will have many similarities.

Interpretation of P.C.A.

From the P.C.A., three practical aspects are considered relevant.

First, by applying this methodology, it was possible to go from 6 variables to characterise this problem to half of its original dimension. With this reduction, three variables instead of six may be used to approach different issues related to the actual six haematological variables in a less complex manner.

Second, three possible profiles of haematological values can be referred to. One possible profile with all but serum lead concentration is positively associated. Another with the concentration of haemoglobin, the ratio of cells positively associated and neutrophil count, white blood cell count and lymphocyte count negatively associated, the neutrophil count having the most substantial negative relation. The final suggested profile has a strong negative association with serum lead concentration and no association with all the other variables.

The third and final point derived from the P.C.A. analysis is connected with outliers. Various outliers were found, meaning that several individuals in the sample supplied for the analysis have statistically distant values from the median point. For Haemoglobin concentration, individuals 12 (16.9) and 23 **(**17.4) were found to have high values, with the value of individual 23 being the max for this variable. For the ratio of cells, individuals 12(50), 23(50) and 38 (50) registered the max value observed for the variable. Still in this variable, individual 1 (39) presented the lowest value. Regarding white blood cell count, individual 23 (8600) showed the highest variable value. In terms of lymphocyte count, individuals 31(43), 45(40) and 47(61) presented high values, with individual 47 being the one with the highest value for this variable. In terms of Neutrophil count, no extreme values were observed. Finally, for serum lead concentration, individual 10 (36) presented the highest variable value. Given this, individual 10, with high levels of serum lead concentration and individual 1, with low values ratio of a cell for the volumetric quantity of blood, are suggested as deserving attention.

Interpretation of Clustering

From the C.A., two main conclusions can be made.

Similarly to P.C.A., if four clusters are considered, the presented problem can be simplified from the six provided variables to four aggregates of individuals.

Furthermore, the behaviour of the variables inside these clusters can be assessed to identify what relationship between individuals is occurring. As such, if four clusters are considered, four distinct groups of individuals can be theorized.

The first group is characterized by low haemoglobin concentration, a low cell ratio in blood and low numbers of white blood cells, lymphocytes and neutrophiles. As such, the individuals of this group will share these characteristics. The second group englobes individuals with low serum lead concentration and low number of lymphocytes, high counts of neutrophils and a wide range of the remaining variables. The individuals of group three tend to have high concentrations of haemoglobin and high cell ratios in blood, while having a low concentration of lead in their serum, as well as low numbers of white blood cells, neutrophiles and sometimes lymphocytes. Lastly, the workers present in group four are characterized by a wide range of lead concentration in serum, and high number in all the other variables.

Succinctly, there are clear differences between each group of individuals, with the second group being the toughest to identify, as it can contain a wide range of values of most variables. The identification of these groups of individuals can help in the administration of correct treatments, or even the search for possible correlations between workers that can lead to these variable levels, such as if their eating habits are similar, do they work on the same position in the company, among others.

Bibliography

Commented [AS1]: Missing: CLuster and prelimianary analysis

Sarmento, R., & Costa, V. (2017). Comparative approaches to using R and Python for statistical data analysis. Descriptive Analysis. DOI: 10.4018/978-1-68318-016-6.ch004.

Yellapu, V. (2018). Descriptive statistics. International Journal of Academic Medicine, DOI:

Sharma, S. (2019). Descriptive Statistics and Factorial Design. Australian Journal of Crop Science 13 (05), 3-5.

Johnson, R. A., & Wichern, D. W. (2013). *Applied Multivariate Statistical Analysis*. Pearson Education Limited. https://books.google.pt/books?id=1XMnngEACAAJ

Ringnér, M. (2008). What is principal component analysis? *Nature Biotechnology*, *26*(3), 303–304. https://doi.org/10.1038/nbt0308-303

Shalabh, S. (2021). Univariate, bivariate and multivariate statistics using R: Quantitative tools for data analysis and data scienceDaniel J.Denis2020, Hobokon, NJ, Wiley, ISBN 9781119549932, pp. xvii + 366. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *185*. https://doi.org/10.1111/rssa.12791