Principal Component Analysis and Cluster Analysis as statistical tools for a multivariate characterization of pharmaceutical raw materials

1. INTRODUCTION

As known, the pharmaceutical industry processes are complex and influenced by numerous parameters that determine their variability. One of these is certainly represented by the *raw materials* which, together with everything that happens during the production process, determine the quality of the finished product. Variations, sometimes even limited, in the quality of raw materials can in fact cause problems in production. This variability arises from the different origin of raw materials, for which pharmaceutical companies usually have more than one supplier, and from the intrinsic variability in the production processes of the raw materials themselves.

The examination of raw materials characteristics is, for instance, a key step in all investigations resulting from the presence of anomalous data. The FDA guideline on OOS, for example, is extremely precise in this regard:

" *OOS results may indicate a flaw in product or process design. For example, a lack of robustness in product formulation,* ***inadequate raw material characterization or control****, substantial variation introduced by one or more unit operations of the manufacturing process, or a combination of these factors can be the cause of inconsistent product quality. In such cases, it is essential that redesign of the product or process be undertaken to ensure reproducible product quality* "

The same guideline also states that:

"…*Current good manufacturing practice for APIs includes the* ***performance of scientifically sound raw material testing***…"

This concept of "scientifically sound raw material testing" is also emphasized in the ICH Q7 guideline which, in paragraph 11, states that:

" *All specifications, sampling plans, and test procedures should be* ***scientifically sound*** *and appropriate to ensure that raw materials, intermediates, APIs, and labels and packaging materials conform to established standards of quality and/or purity.* "

In common practice, the characterization of raw materials is carried out in a *univariate* way, *i.e.* by comparing the average values, relative to several batches, of each parameter that characterizes a given raw material (*e.g.*, assay, pH, *etc.*) using statistical tools such as *2-sample t-test* and ANOVA. This approach often does not allow to adequately characterize a given raw material because it does not capture, in its entirety, the singularity of each batch and therefore does not contextualize it fully with respect to the others coming from the same supplier. Therefore, only the adoption of a multivariate approach can respond to such *global* needs and therefore provide an ***adequate raw material characterization***.

In this post, we will illustrate the application to multivariate characterization of raw materials of two of the main techniques of *Multivariate Statistical Data Analysis* (*MVDA*) [1,2,3,4,5] and precisely the *Principal Component Analysis* (*PCA*) [6] and *Cluster Analysis* (*CA*) [1,2,3].

To illustrate the details, a raw material (synthetic intermediate) was chosen, supplied by three different suppliers and characterized by a limited number of parameters (four) but rather different from each other by nature: assay, largest known impurity, color and residual water content according to Karl Fischer. These, in fact, were the only continuous (or numerical) variables present and common to the various suppliers.

The *multivariate characterization of the raw materials* proposed here, in addition to the case described, can also be directly extended to other situations such as, for example, the *comparison between batches of the same finished product*. In this case, in fact, there are only more parameters to consider, but, for the rest, nothing changes.

2. EXPERIMENTAL SECTION

Table 1 below shows the database used. These are the values of assay, largest known impurity, color (percentage transmittance at predetermined λ) and Water Content according to Karl-Fischer including their units of measurement and specifications, relating to a total of 41 lots from three different suppliers (14 from S1, 13 from S2 and 14 from S3).

Table1

Supplier	Supplier_Batch No.	Assay (%)	Largest known Imp. (%)	Color (%)	Water Content (%)
Supplier 1	1	99,9	0,03	98,3	0,07
	2	99,9	0,02	98,5	0,11
	3	99,9	0,03	97,7	0,09
	4	99,9	0,04	98,1	0,09
	5	99,9	0,03	96,6	0,09
	6	99,9	0,02	98,3	0,07
	7	99,9	0,02	98,0	0,13
	8	99,9	0,03	98,4	0,11
	9	99,9	0,02	98,3	0,13
	10	99,9	0,02	98,2	0,09
	11	99,9	0,03	97,3	0,11
	12	99,9	0,04	98,4	0,09
	13	99,9	0,04	98,0	0,11
	14	99,9	0,02	97,9	0,11
Supplier 2	15	99,8	0,05	94,6	0,20
	16	99,8	0,05	94,9	0,10
	17	99,8	0,04	94,4	0,10
	18	99,9	0,03	95,1	0,10
	19	99,8	0,02	95,2	0,10
	20	99,8	0,03	95,6	0,10
	21	99,8	0,03	95,1	0,10
	22	99,9	0,02	96,4	0,10
	23	99,4	0,05	95,3	0,20
	24	99,8	0,03	94,6	0,10
	25	99,9	0,02	95,1	0,10
	26	99,8	0,04	96,3	0,10
	27	99,8	0,03	95,3	0,10
Supplier 3	28	99,8	0,07	91,0	0,16
	29	99,8	0,06	91,0	0,14
	30	99,9	0,07	91,0	0,16
	31	99,7	0,08	91,0	0,16
	32	99,8	0,06	91,0	0,15
	33	99,8	0,07	91,0	0,16
	34	99,7	0,08	91,0	0,15
	35	99,9	0,08	92,0	0,18
	36	99,9	0,07	94,0	0,20
	37	99,9	0,07	93,0	0,30
	38	99,9	0,08	91,0	0,20
	39	99,9	0,06	94,0	0,20
	40	99,9	0,08	95,0	0,20
	41	99,9	0,07	93,0	0,20
Specifications		NLT 95%	NMT 0,10%	NLT 90,0%	NMT 0,5%

The dataset relevant to the forty-one batches considered consists of a table in which each row contains the data relating to a given batch while each column refers to a specific measured analytical parameter, or *variable*.

This data table is usually known in statistical jargon as the *data matrix*. Data analysis and their visualization were conducted using Minitab 19 (GMSL S.r.l. - Via Giovanni XXIII, 21 - 20014 Nerviano (Milan), Italy) and RStudio version 1.3.1093 and R version 4.0.3 (The R Foundation for Statistical Computing). The following specific R packages have been used:

- *FactoMineR* (F. Husson, Agrocampus Ouest, Rennes University, France) [7, 8]
- *factoextra* (A. Kassambara, HalioDx, Marseille, France) [9, 10,11]
- *NbClust* (M. Charrad, N. Ghazzali, V. Boiteau, A. Niknafs, Laval University, Canada) [12
- *corrplot* (T. Wei, Fujian Agriculture and Forestry University, China) [13, 14]
- *fpc* (C. Hennig, Bologna University, Italy)

3. RESULTS AND DISCUSSION

Principal Component Analysis, PCA, is one of the most powerful and widely used multivariate methods for data exploration. It is used when a simpler representation of a set of related variables is desired. In practice, the starting variables that describe the data are transformed into new variables (*i.e.*, the *main components*, PC) which are linear combinations of the initial variables and are, by construction, orthogonal to each other. The PCA assumes that the directions where there is greater variability are the most "important" or, indeed, the "principal" ones. PCA does not work if the original variables are not related to each other. The existence of correlation indicates redundancy in the data.

The first step is therefore to verify the degree of linear correlation existing between the four variables involved. Table 2 shows the *correlation matrix* while Figure 1 shows the so-called *matrix plot*, *i.e.*, a graphical representation of the relationships between pairs of variables.

Figure 1

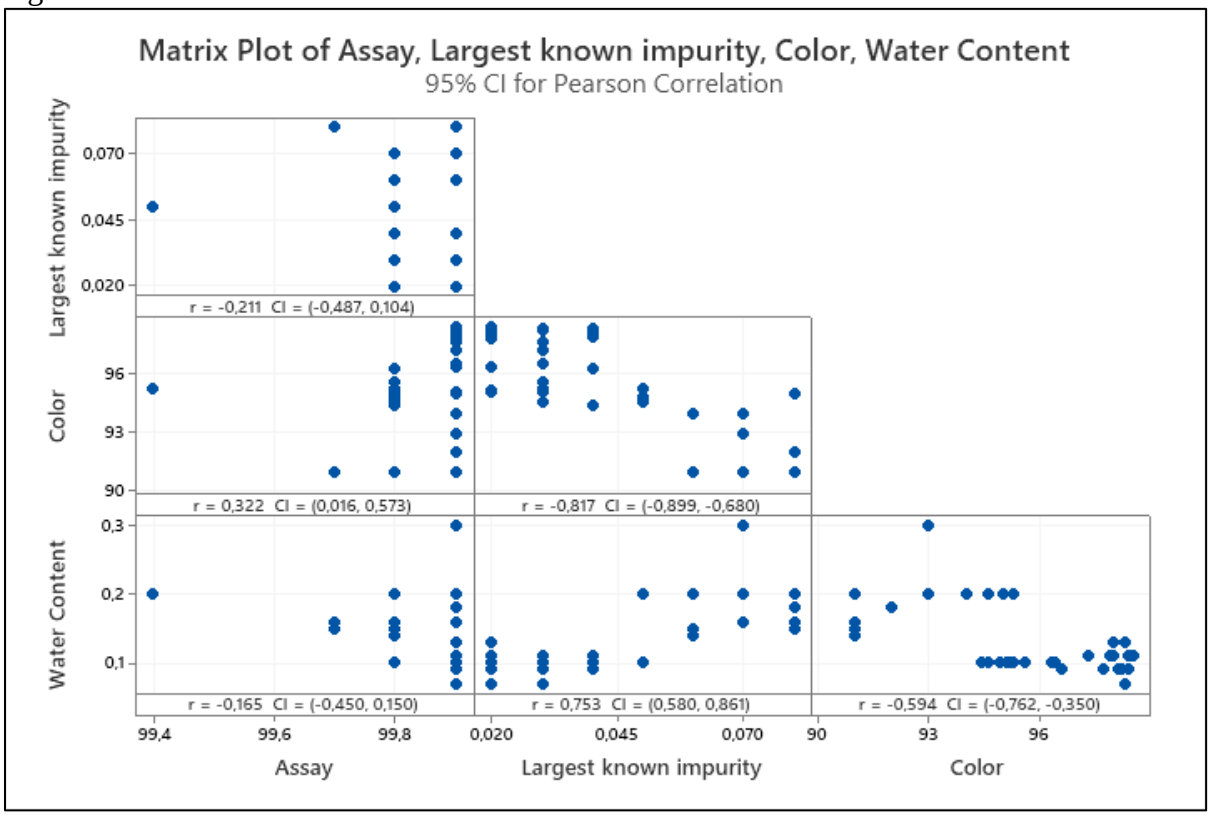
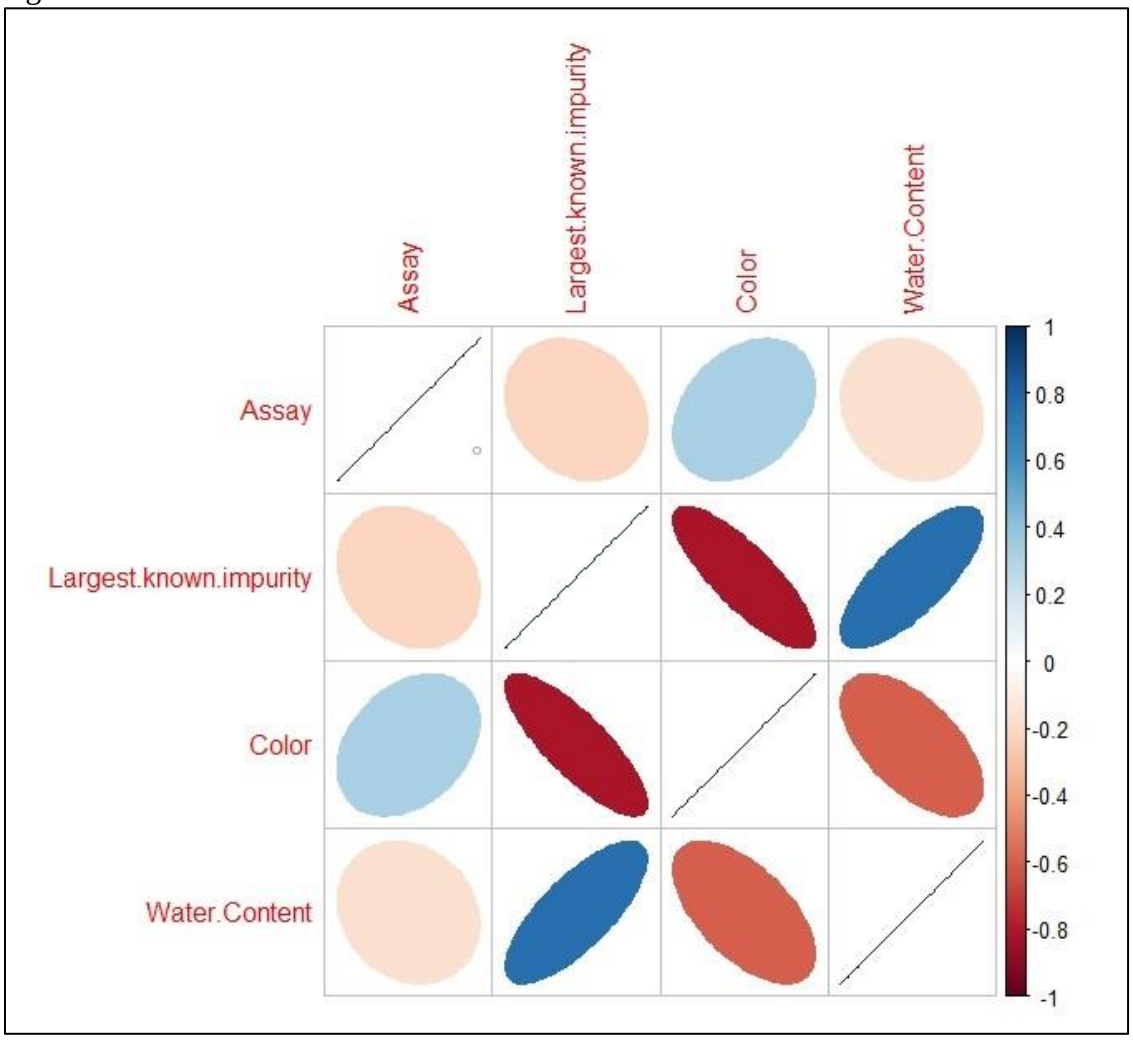


Table 2

	Assay	Largest known imp.	Color
Largest known imp.	-0,211		
Color	0,322	**-0,817**	
Water Content	-0,165	**0,753**	-0,594

The examination of the values in Table 2 shows how some variables (*e.g.*, amount of largest known impurity and color, *etc.*) are highly correlated with each other and therefore there are the conditions for applying the PCA. A graphic representation that allows to better grasp the relationships existing between the different variables is that provided by the *correlogram* shown in Figure 2.

Figure 2



Each element of this graph is a geometric figure that becomes more and more elliptical and intensely colored the more the two variables are linearly related. On the main diagonal, where the correlation is maximum (in fact the correlation of a variable with itself is equal to 1) the ellipses become a segment.

The ellipses are oriented to the right and colored in blue if the two variables are positively correlated to each other, while they are oriented to the left and colored in red / brown if they are negatively correlated.

The examination of the correlogram shows a strong correlation between the couples:

- Largest known impurity - Color: negative
- Largest known impurity - Water Content: positive
- Water Content - Color: negative

Beyond the more statistical evaluations that follow, it is interesting to observe how already from these first findings it is possible to derive correlations of potential chemical interest, *e.g.*, the higher the residual Water Content, the higher the content of Largest known impurity, *etc*.

To better understand the degree of linear correlation existing between the variables mentioned above, the relationships between them can be analyzed using Simple Linear Regression. As an example, this analysis is shown below only for those pairs of variables that show the highest linear correlation coefficients.

Regression Analysis: Color *vs.* Largest known impurity

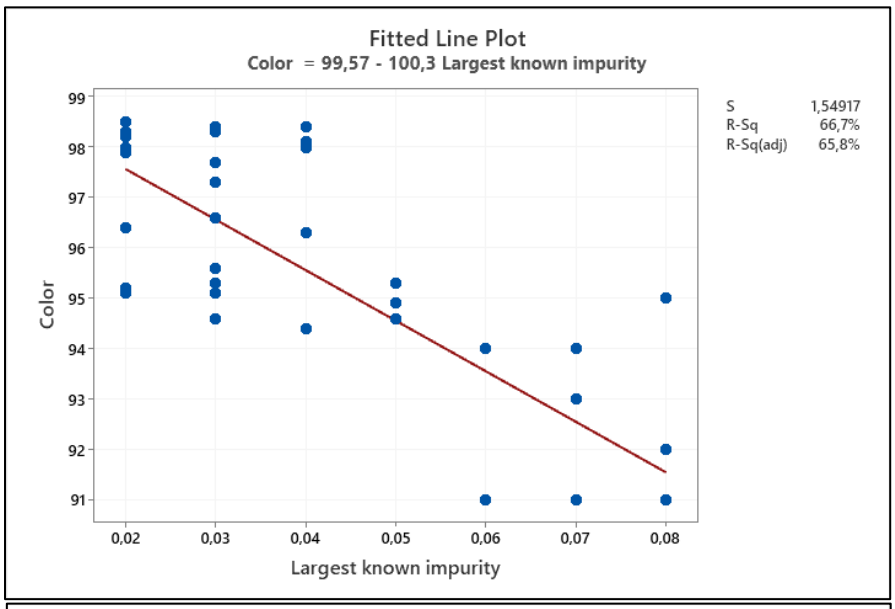
The regression equation is **Color = 99,57 - 100,3 Largest known impurity**

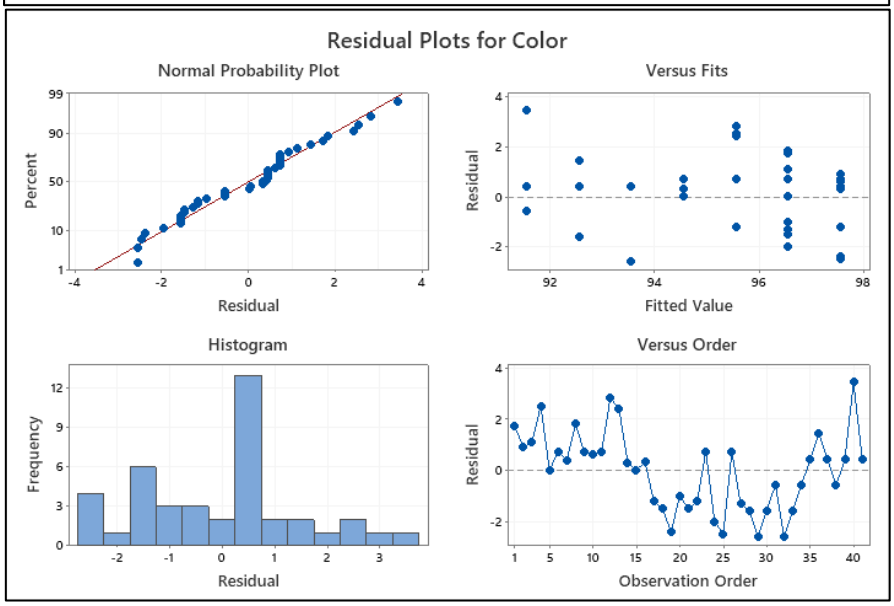
Model Summary

S	R-sq	R-sq(adj)
1,54917	66,69%	65,84%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	187,422	187,422	78,09	0,000
Error	39	93,597	2,400		
Total	40	281,019			





Regression Analysis: Water Content *vs.* Largest known impurity

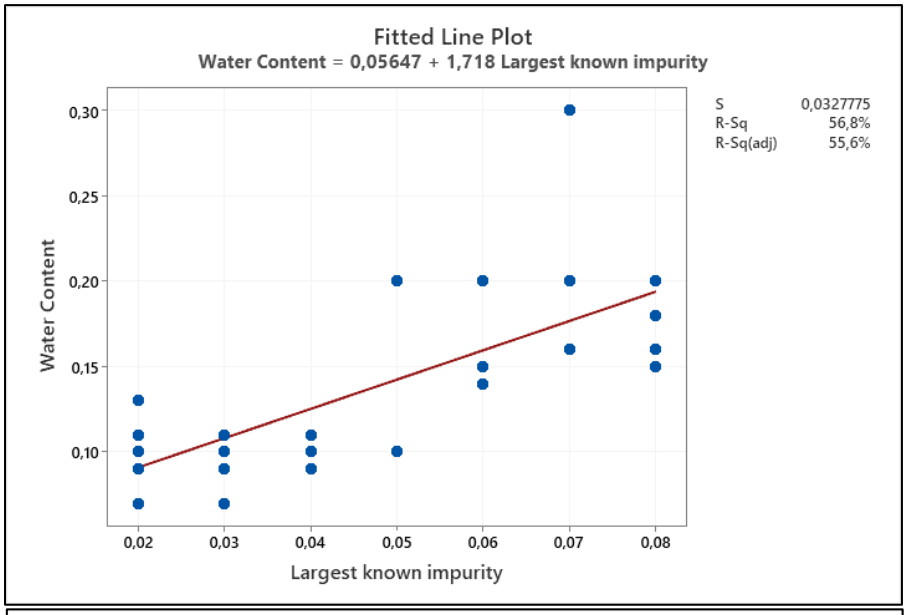
The regression equation is **Water Content = 0,05647 + 1,718 Largest known impurity**

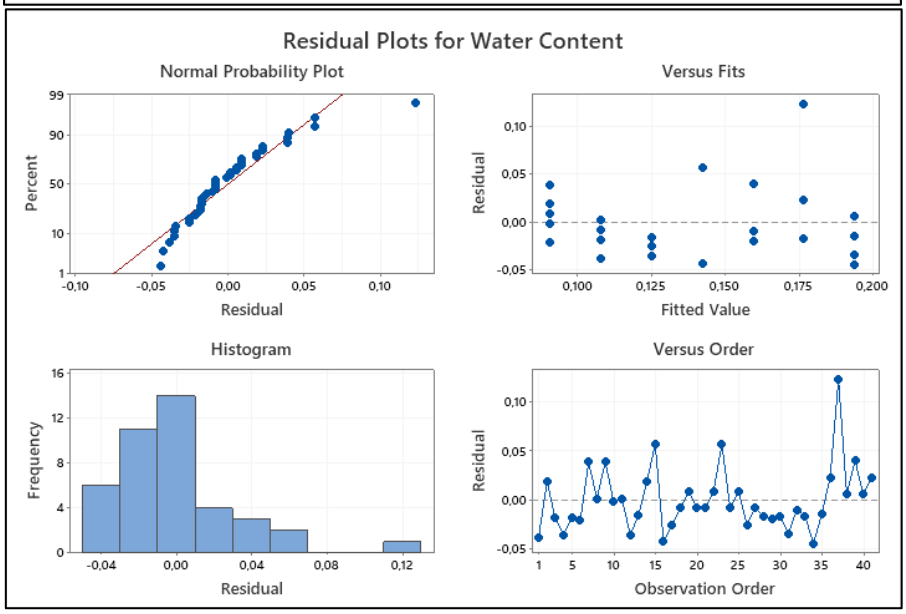
Model Summary

S	R-sq	R-sq(adj)
0,0327775	56,75%	55,65%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	0,0549877	0,0549877	51,18	0,000
Error	39	0,0419001	0,0010744		
Total	40	0,0968878			





Given the evident correlations that exist, it is therefore possible to apply the PCA to the starting database (or *data matrix*) reported in Table 1. Leaving aside the more mathematical and calculation aspects, it is interesting to focus on the output resulting from PCA which is summarized below:

Principal Component Analysis:

Table 3: Eigenanalysis of the Correlation Matrix

Eigenvalue	2,5514	0,9174	0,3926	0,1386
Proportion	0,638	0,229	0,098	0,035
Cumulative	0,638	0,867	0,965	1,000

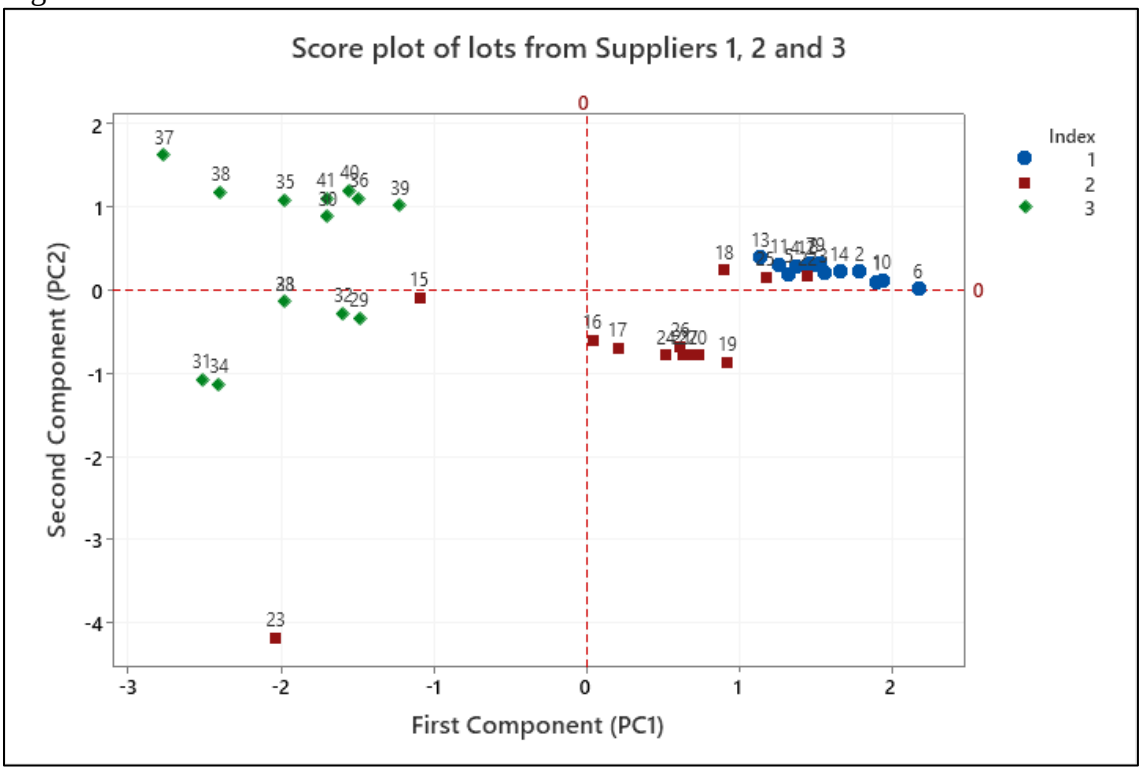
Table 4: Eigenvectors

Variable	PC1	PC2	PC3	PC4
Assay	0,252	0,948	-0,180	0,075
Largest known impurity	-0,586	0,189	-0,150	-0,774
Color	0,562	0,012	0,623	-0,543
Water Content	-0,526	0,257	0,746	0,317

Table 3 shows the so-called *eigenvalues* which, beyond their complex mathematical meaning, in this context represent the variance associated with each *eigenvector*, that is, with each *principal component* (PC). The *eigenvalues* are ordered in descending order and that is to say that, passing from one *principal component* to the next, the variability that each of them intercepts gradually decreases, so much so that the smaller *eigenvalues* are associated with information that is generally not relevant. The first *principal component* (PC1 - Table 4) is therefore the most important to represent the variation in the measurements of the 41 batches of raw material considered here. In fact, it explains 63.8% of the variability in the data, while the second component only 22.9% and so on (Proportion – Table 3). Considering only the first and second components, it is therefore possible to explain 86.7% of the variability in the starting data. In practice, from the four initial variables we are reduced to only two. Although this involves an overall loss of information of 13.3% (100% - 86.7%), however, there is the enormous advantage of being able to work with only two coordinates and therefore represent each single lot as a point in a Cartesian plane identified by the axes PC1 and PC2.

At this point the *score plot* in Figure 3 below is highly explanatory.

Figure 3



Each point of the graph identifies a lot (or *individual* or *observation* as we would say in statistical jargon) of those shown in Table 1.

The points are of three colors, one for each of the three suppliers from which the individual lots come.

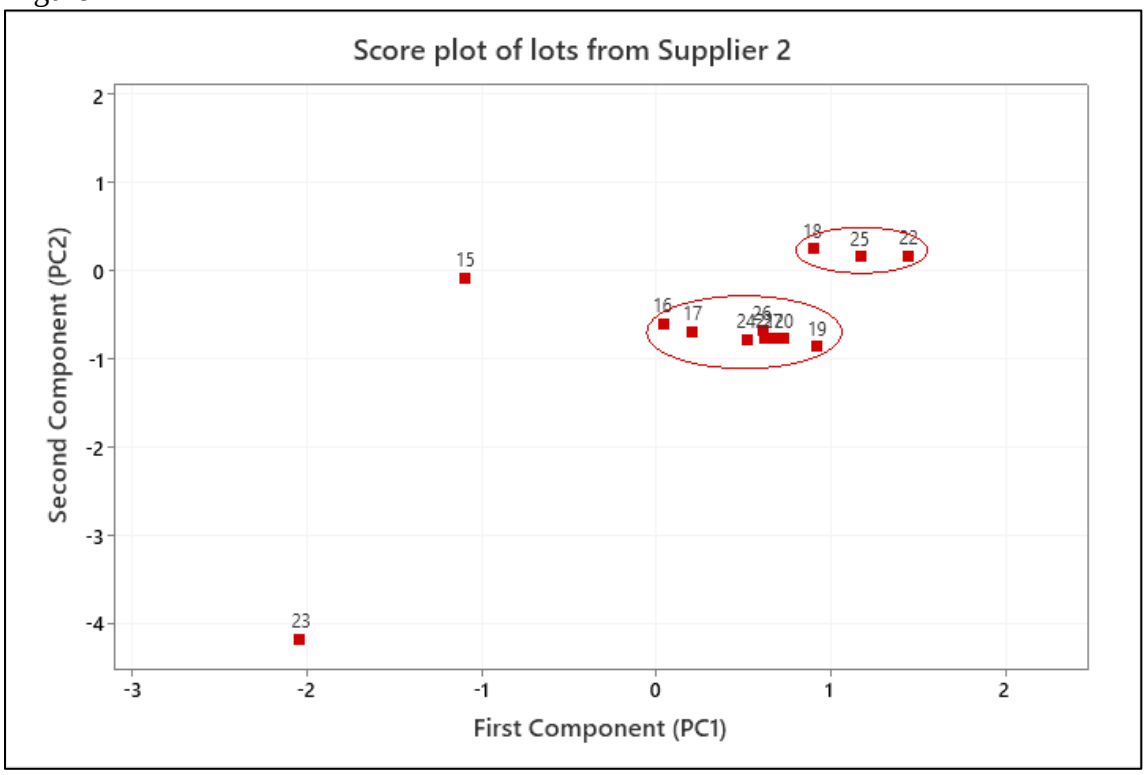
At first glance some important aspects are evident:

- the points (and therefore the lots) are distributed mainly horizontally, *i.e.* in the direction of the first principal component and this is obvious since, as mentioned above, it intercepts as much as 63.8% of the entire variability of the data,
- the datapoint corresponding to lot 23 is completely separate from all the others to indicate that it clearly differs from them,
- the datapoints corresponding to the lots of supplier 1, and marked with blue dots, are all close to each other, while those of supplier 3, marked with green diamonds, are a little more dispersed in the plan, but still contained in a fairly limited "cloud".

  The datapoints associated with the lots of supplier 2, marked with red squares, are instead scattered on the plane with two fairly distinct central nuclei and two points distant from the others. In particular, the point, and therefore the lot, 23, appears so far from the others coming from the same supplier as to suggest that it does not have much to do with them.

A better perception of the dispersion of the lots of Supplier 2 and of the degree of separation that distinguishes lot 23 from the others provided by the same supplier is given by Figure 4 which is nothing more than the *score plot* of Figure 3 in which only the datapoints relating to the lots of supplier 2 appear.

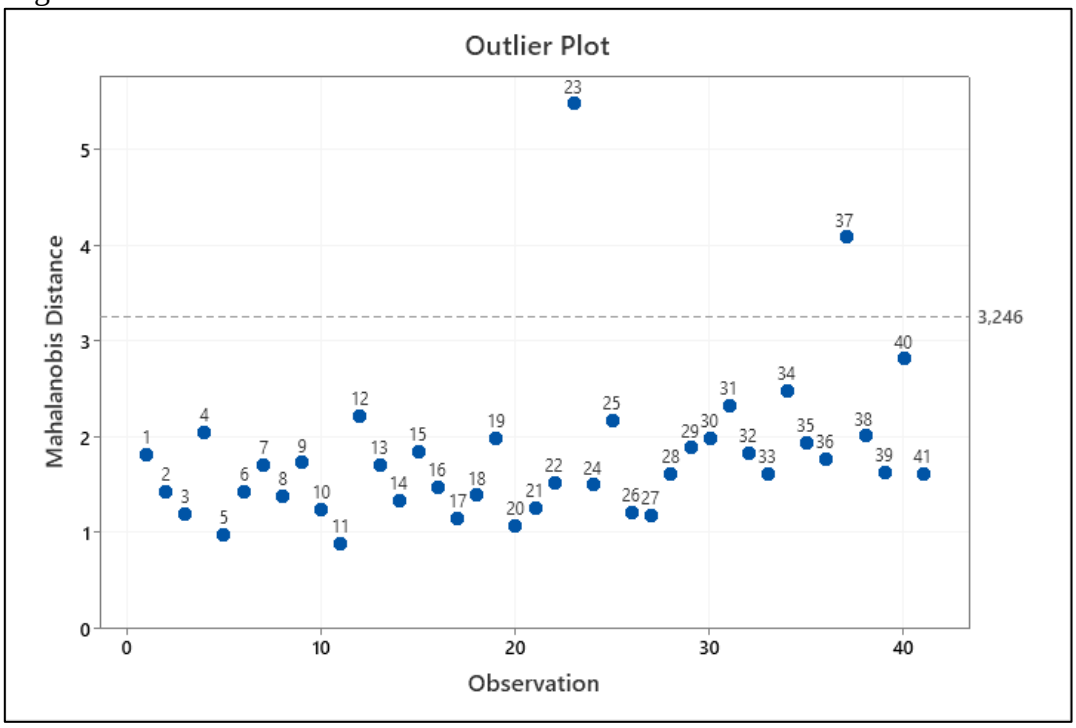
Figure 4



In Figure 4 the two distinct aggregation nuclei of the lots of the supplier 3 mentioned above are evident. To these are added lots 15 and 23 which, in this coordinate system (PC1, PC2), are clearly separated from the others.

The anomalous nature of lot 23 is also well highlighted by the *outlier plot* in Figure 5 which shows the Mahalanobis distances of each point (or lot) with respect to the *centroid* of each group of points.

Figure 5



The graph of Figure 5 also identifies lot 37 as certainly anomalous and the point corresponding to lot 40 appears close to the reference line too. Examining the *score plot* in Figure 3, it can be seen that point 37 is only the most marginal one among those of supplier 3 (green diamonds) while point 40 mixes with the others.

A better perception of the anomaly of lots 37 and 40 is obtained by examining the score plots obtained using PC3 (Figure 6) and PC4 (Figure 7) in ordinate instead of PC2.

Figure 6

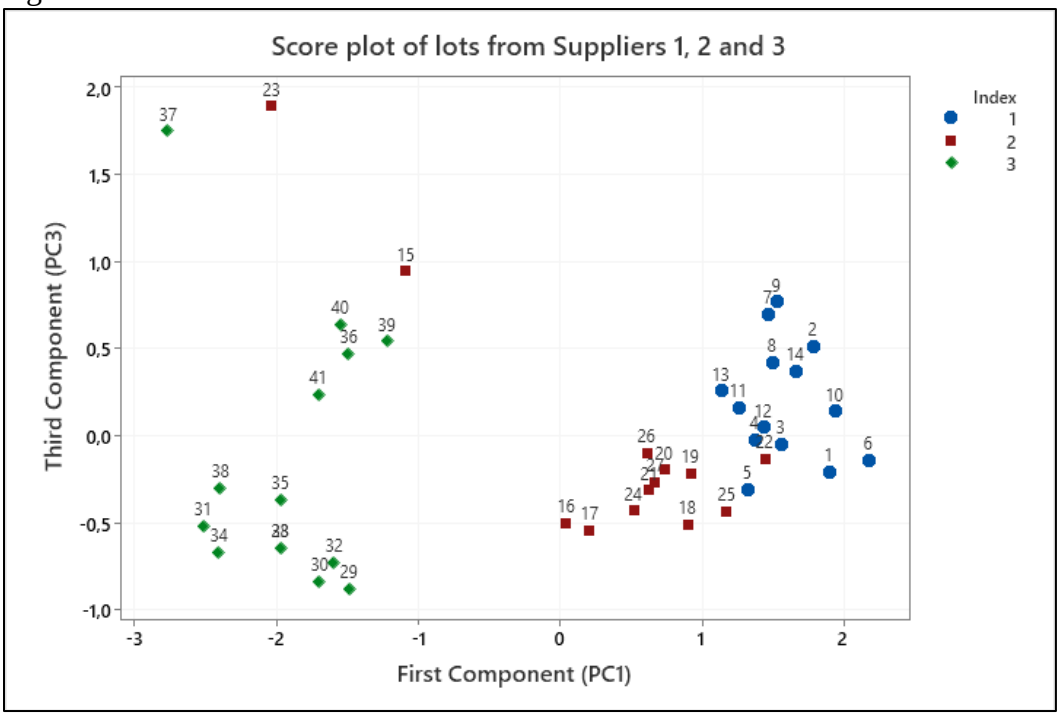
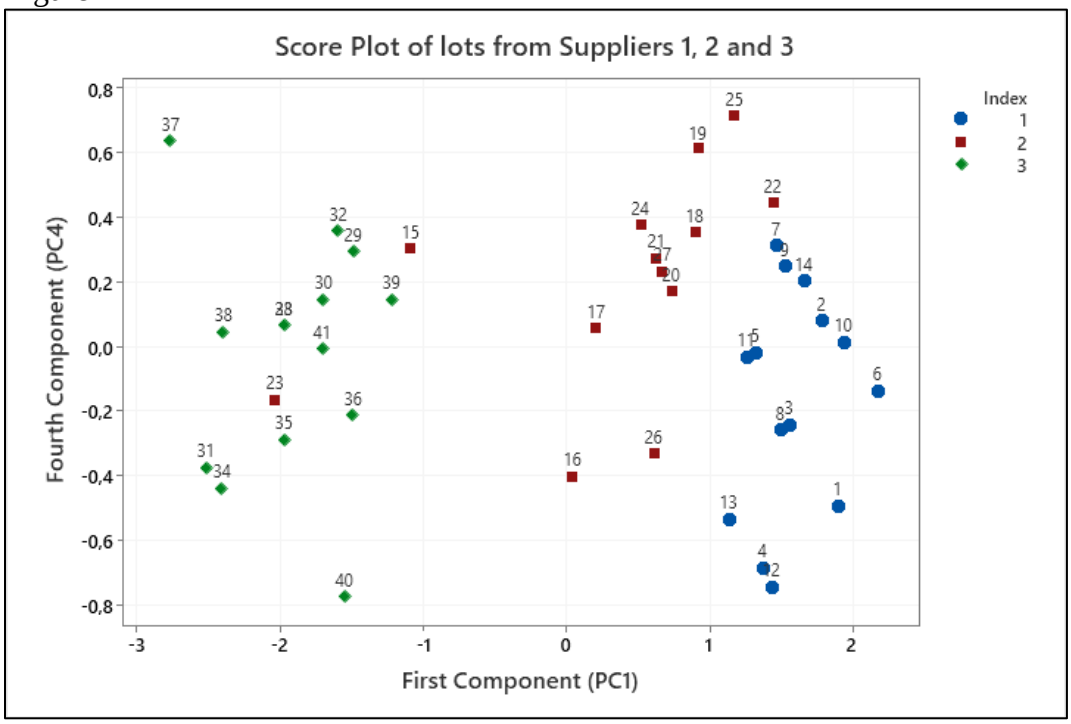


Figure 7



Point 37 is far from other datapoints in both the score plots of Figures 6 and 7, while point 40 is well separated only in Figure 7.

This different "resolving power" of the *score plots*, and which derives from the choice of a different ordinate, follows precisely from the composition of the single coordinates or *principal components*. This composition is summarized in Table 4 and can easily be understood thinking that, for example, the *first principal component* is, in fact, described by the following mathematical relationship:

Z_1 = 0.252 *assay* – 0.586 *largest known impurity* + 0.562*color* -0.526 *water content*

and, similarly, it applies to the other components.

Examining the structure of this equation it is evident that the first component is approximately an average of three of the four variables that characterize each batch, namely: *largest known impurity*, *color* and *water content*.

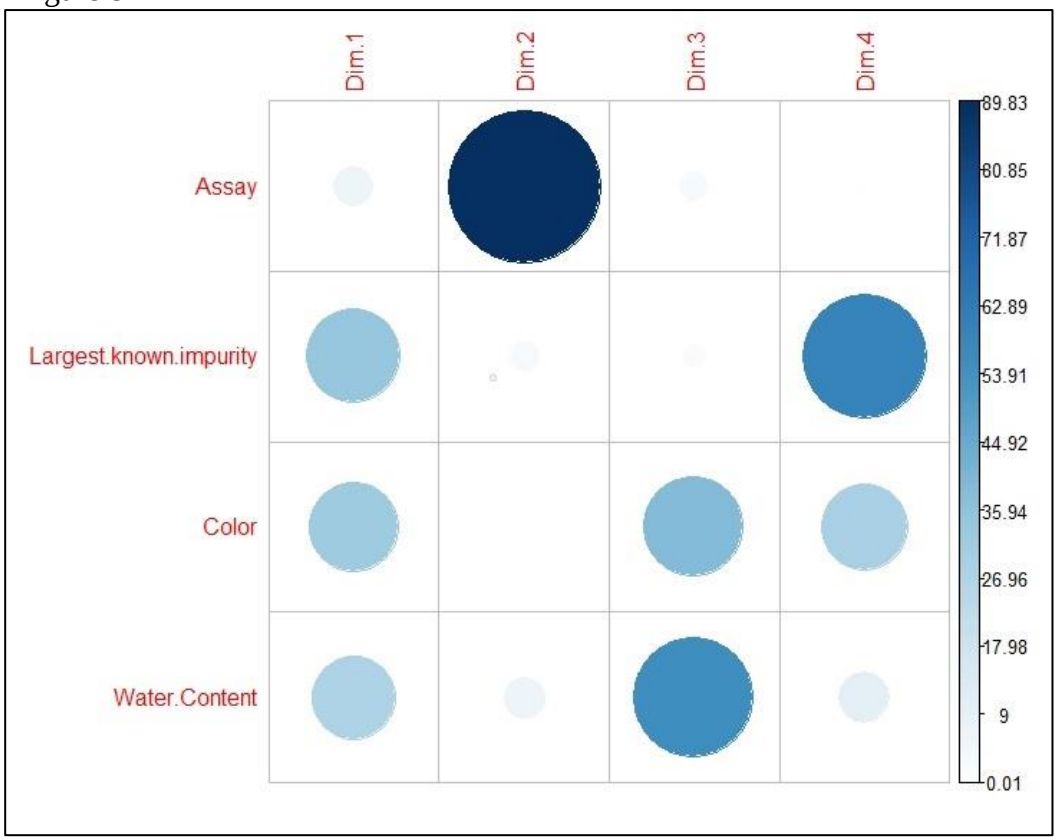
The *second component*, instead, has the following structure:

Z_2 = 0.948 *assay* + 0.189 *largest known impurity* + 0.012*color* +0.257 *water content*

In this case the *assay* variable is the one that weighs the most of all while *color*-type variables contribute in a practically insignificant way. In this regard, consider that the point corresponding to lot 23 is so well separated in the score plot of Figure 3 (PC1, PC2) because it is the one with the lowest assay value (*i.e.*, 99.4% - Table 1).

A graphic representation of the *principal components* that allows you to immediately grasp their different compositions at a glance is that provided by the *correlogram* in Figure 8 below.

Figure 8



The graph in Figure 8 is self-explanatory. From it, for example, it is confirmed that the second main component is practically determined by the *assay* variable alone, while the third component is determined by *water content* and *color*.

For practical purposes, the score plots in figures 3, 4, 6 and 7 indicate that, in general, the quality of the product supplied by supplier 1 is the most constant and reproducible. This is followed by that of supplier 3, while that of supplier 2 is affected by high variability that could make it a source of problems.

It is useless to say how useful such a discriminating characterization is. The *univariate* approach to the problem, namely the comparison of averages and dispersions of the individual variables in the three cases, does not allow the information to be viewed as well, which, using PCA, becomes intuitive even to non-experts.

Finally, it is important to consider that the approach illustrated here, in addition to the case of the same raw material supplied by multiple suppliers discussed here, can also be applied to other very common situations in the pharmaceutical industry such as, for example:

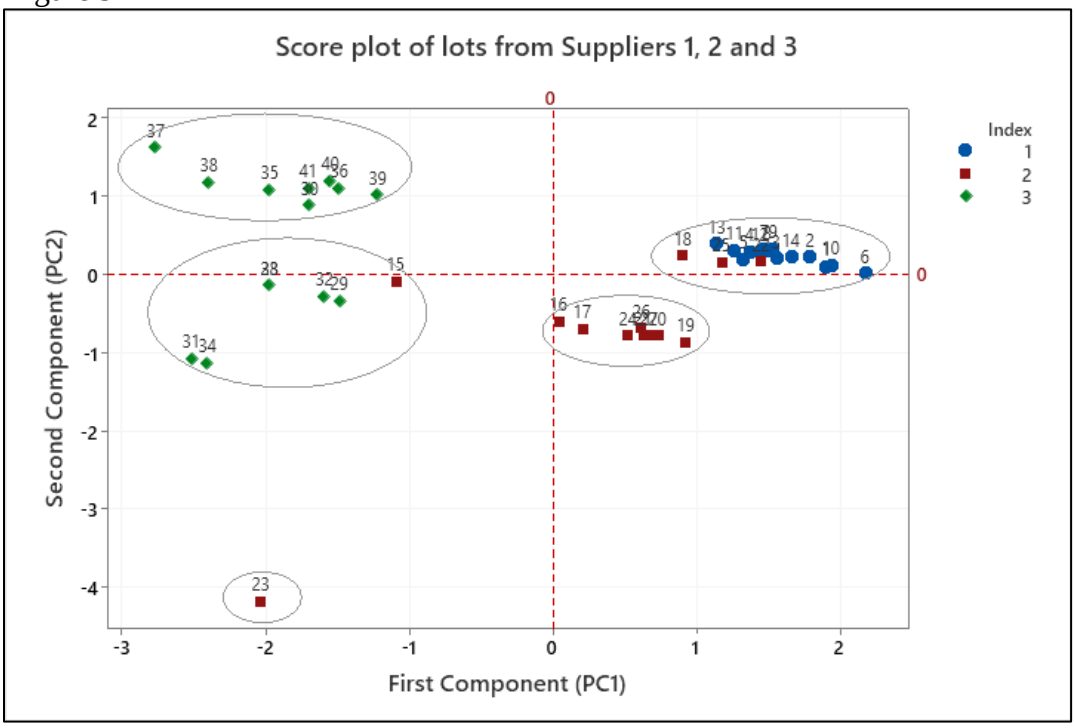
- comparative evaluation of finished product lots, for example for the purposes of Annual Product Quality Review (APQR)
- comparative evaluation of series of measurements performed by different operators, *etc.*

In all these cases, the information hidden in the folds of the numerous numerical variables that describe the analytical profile of a batch of finished product or in the measurements performed by different operators, is extracted and made immediately available in a ready-to-use format.

The examination of the *score plots* of Figures 3,4, 6 and 7, if separated from the supplier's groups to which they belong, returns an overall photograph of the arrangement of the raw material batches in the plan identified by two main components.

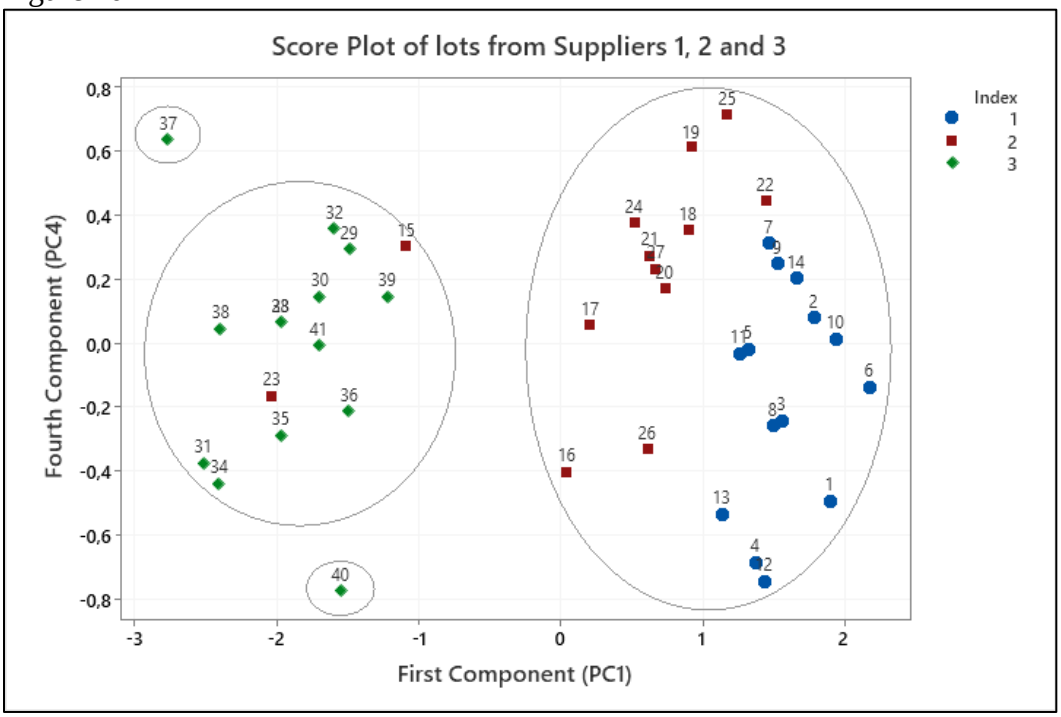
The examination of the score plot in Figure 3, for example, shows in the plan identified by the two main components PC1 and PC2 (which, together, explain 86.7% of the variability in the data) the presence of four nuclei dense of points and one group only consisting of a single element (23) as illustrated in Figure 9 below.

Figure 9



The score plot in Figure 7, on the other hand, shows in the plan identified by PC1 and PC4 (which, together, explain the 63.8 + 3.5 = 67.3% of the data variability) two main groups of points and two isolated points from the rest as illustrated in Figure 10 below.

Figure 10

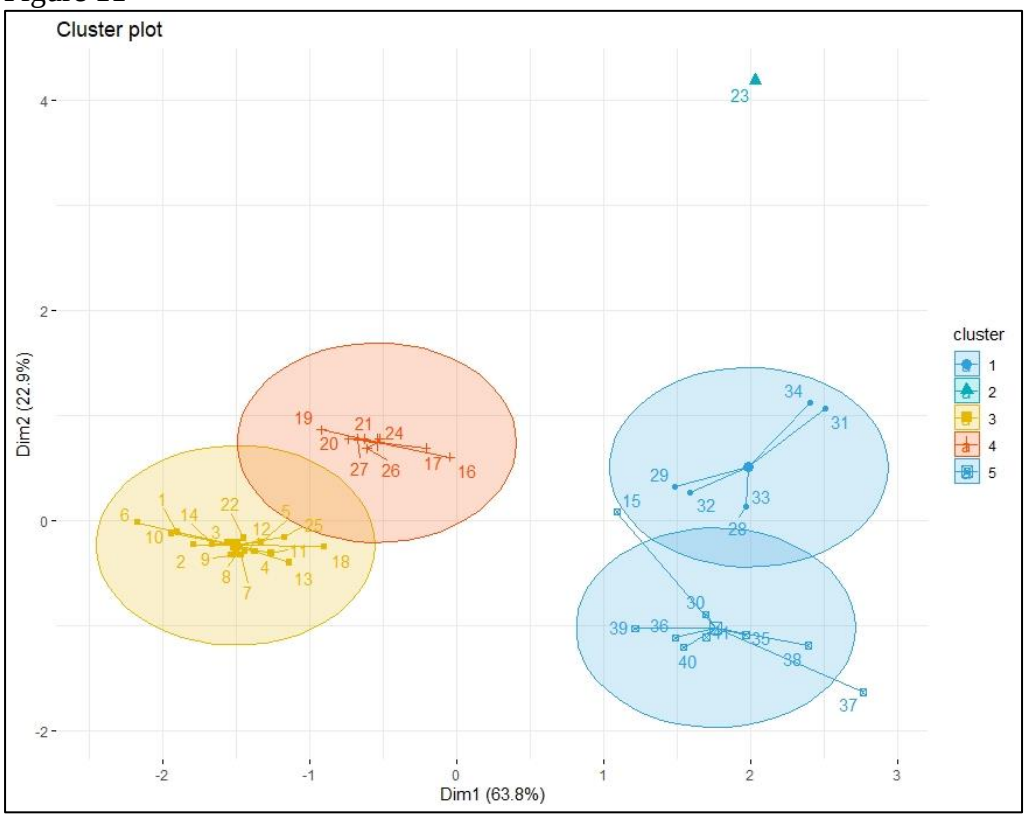


Therefore, depending on the perspective from which the data is looked at, a different number of homogeneous groups, or *clusters*, can be identified. In practice, these are "sub-populations" present within the "population" made up of all lots considered.

To define the number of homogeneous groups on a non-individual basis, such as the visual one, multivariate methodologies are used that are particularly useful for the purpose and known, in general, as *Cluster Analysis*. This term means all those multivariate methodologies that solve the problem of classification, that is, the aggregation of statistical units to form groups (or *clusters*) of elements as homogeneous as possible and as isolated from each other as possible. Within these methodologies there are indices that allow you to establish with a good level of certainty the real number of homogeneous groups, or *clusters*, present. Furthermore, these methods contain indices, or *statistics*, such as Hopkin's that allow to establish a priori whether a certain dataset has the characteristics to be divided into *clusters* or not.

In the case chosen here, for example, by applying the *Cluster Analysis* to the data and choosing *K-means* as grouping method (a so-called *partitioning clustering method*), the diagram shown in Figure 11 is obtained.

Figure 11

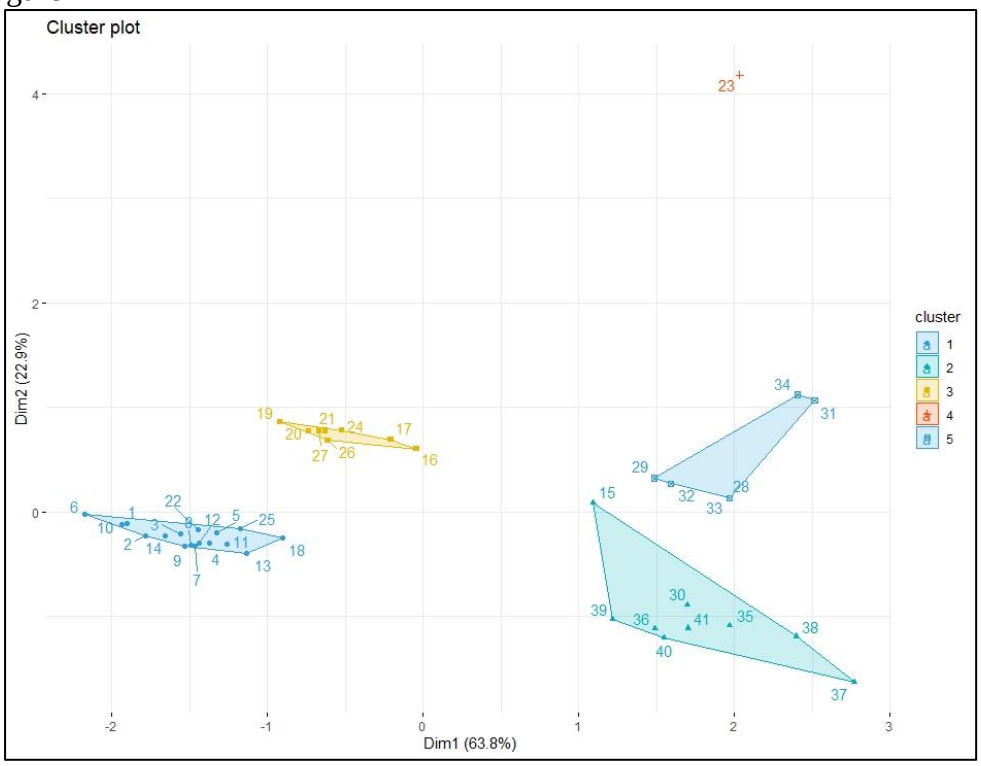


As can be seen, the diagram in Figure 11 highlights four main *clusters* and two separate datapoints: 23 completely isolated from everyone and 37 also however outside the closest homogeneous group.

Each group is built around a *centroid* well indicated in the graph.

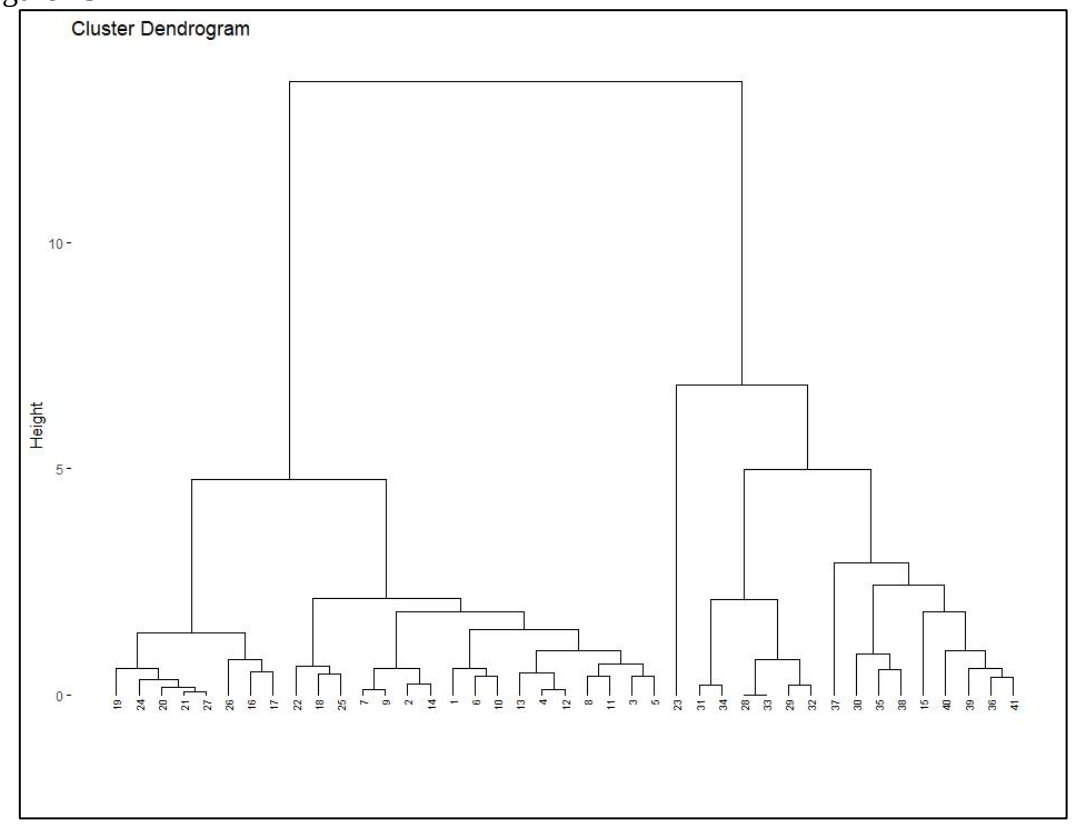
A similar distribution (Figure 12) is also achieved using the so-called *Hierarchical Clustering methods*.

Figure 12



This second type of methods is based on the construction of hierarchical trees or *dendrograms*, such as the one represented in Figure 13.

Figure 13



From the above findings it emerges that, regardless of the technique used, in the dataset it can be identified four main *clusters* consisting of multiple datapoints and one consisting of a single datapoint (point 23) completely separated from all the others. Two datapoints (points 37 and 15) occur in marginal positions.

4. CONCLUSIONS

Numerous factors contribute to the variability of the pharmaceutical industry processes and among these the raw materials play a primary role as they often come from different sources that use different production processes.

The characterization of raw materials therefore plays a fundamental role in terms of Quality which, by its nature, is "the enemy of variability".

Multivariate Statistical Analysis of Data (MVDA), beyond of its complex mathematical structure, is presented here as a powerful and practical tool for the study and classification of raw materials. In fact, thanks to the use of multivariate techniques such as *Principal Component Analysis* (PCA) or *Cluster Analysis* (CA), it is possible to graphically represent each lot, defined by the values of the different analytical parameters that characterize it, as a point in a Cartesian diagram whose coordinates are the *principal components*. Since these components are built to intercept the variability in the data, these graphs reveal characteristics which would escape other types of surveys and therefore allow to catalog the lots based on the degree of intrinsic homogeneity that defines them and identify any anomalous behavior. This approach can therefore be used both initially, to characterize the incoming raw materials, and subsequently, in the case of any anomalies, to see how the raw materials of the batches under investigation were located compared to those that had not given problems.

The techniques that have been detailed here can also be extended to other typical situations in the pharmaceutical industry such as, for instance:

- comparative evaluation of finished product lots, for example for the purposes of Annual Product Quality Review (APQR),
- comparative evaluation of series of measurements performed by different operators, *etc.*

Once again, statistical methods show how it is possible to "simplify complexity" and extract practical and "ready-to-use" knowledge from data sets by capturing their information content.

5. BIBLIOGRAPHY

1. K.V. Mardia, J.T. Kent, J.M. Bibby, *Multivariate Analysis*, Academic Press, 1989
2. W.R. Dillon, M. Goldstein, *Multivariate Analysis*, J. Wiley &Sons, 1984
3. B.S. Everitt, G. Dunn, *Applied Multivariate Data Analysis*, 2nd Ed., Wiley, 2001
4. W. Härdle, L. Simar, *Applied Multivariate Statistical Analysis*, 2nd Ed., Springer, 2007
5. B.F.J. Manly, J.A. Navarro Alberto, *Multivariate Statistical Methods, A Primer*, CRC Press, 4th Ed., 2017
6. I.T. Jolliffe, *Principal Component Analysis*, 2nd Ed., Springer, 2002
7. F. Husson, S. Lê, J. Pagès, *Exploratory Multivariate Analysis by Example using R*, 2011, CRC Press.
8. F. Husson, J. Josse, J. Pagès, *Principal component methods – hierarchical clustering – partitional clustering: why would we need to choose for visualizing data?*, September 2010, Technical Report - Agrocampus.
9. A. Kassambara, *R Graphics Essentials for Great Data Visualization*, STHDA, 2017
10. A. Kassambara, *Practical Guide to Principal Component Methods in R*, STHDA, 2017
11. A. Kassambara, *Practical Guide to Cluster Analysis in R*, STHDA, 2017
12. Malika Charrad, Nadia Ghazzali, Veronique Boiteau, Azam Niknafs (2014). *NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set*, Journal of Statistical Software, 61(6), 1-36. URL http://www.jstatsoft.org/v61/i06/.
13. M. Friendly, *Corrgrams: Exploratory displays for correlation matrices*, The American Statistician, 56 (2002) 316-324
14. D.J. Murdoch, E.D. Chow, *A graphical display of large correlation matrices*, The American Statistician, 50 (1996) 178-180