A different way to look at pharmaceutical Quality Control data: multivariate instead of univariate.

1. INTRODUCTION

In the pharmaceutical industry, Quality Control (QC) data are typically arranged in datasets that contain the results from different types of measurements (chemical and microbiological) on different samples each representative of a production lot.

For a given active chemical entity (API, Active Pharmaceutical Ingredient), or dosage form, it therefore exists a data table (or *data matrix*) each row of which contains the results of different measurements (*e.g.*, pH value, assay, *etc.*) carried out on a specific lot. The first column of the data table contains the lot numbers.

In practice, each row of such data table contains the information typically listed in the certificate of analysis issued for that lot. From a QC perspective, this information represents the "analytical profile" of that specific lot.

As for each active chemical entity, or dosage form, there is a specific dataset and since all lots listed therein are manufactured using the same approved process, the dataset contains the "analytical fingerprint" of that manufacturing process.

As required by regulations, QC data must be reviewed, evaluated and trended for knowledge and insight [1]. This task is usually carried out in a *univariate mode*, *i.e.*, each type of data is individually analyzed using statistical tools such as control charts, box plots, *etc.* The dataset is therefore studied "by columns".

In this post, it is proposed a different way to analyze QC data, *i.e.*, by using a *multivariate approach* instead of a *univariate* one. Multivariate statistical analysis is the simultaneous analysis of a collection of variables and it improves upon separate univariate analyses of each variable by using information about the relationships between the variables [2]. Moreover, the combination of multivariate methods with the power of the programming language R and its unsurpassed graphic tools, allows analyzing data mainly relying on graphics and, as stated by Chambers *et al*, "there is no statistical tool that is as powerful as a well-chosen graph" [4].

This post shows how using R for combined *multivariate data analysis* and visualization, the information contained in QC chemical dataset can be easily extracted and converted into "knowledge ready to use".

2. EXPERIMENTAL SECTION

As case study, it has been considered a hypothetical QC chemical dataset containing the analytical results obtained for thirty-one (31) samples (or individuals) of a drug substance, each representative of a different manufacturing lot. Obviously, all lots are assumed manufactured using the same production method.

Even if no missing data are expected in QC datasets, as this it would prevent the lot approval, if a test leads to a result below the quantitation limit, it is common practice that of reporting: <LOQ. In these cases, instead of removing data or variables, the result "<LOQ" has been replaced with the numerical value of the corresponding limit of detection (LOD). In fact, if the removal of rows or columns to eliminate missing data affects the dataset, the use of other methods (*e.g.*, iterative methods) to impute missing data often leads to data values with no meaning.

Each lot of the hypothetical dataset considered is characterized by an array of analytical parameters (or variables) that are listed in Table 1 together with the requirements (allowed range of variability or specifications) to whom they have to comply. Table 1 is completed by the abbreviations that it will be used, from now on, to identify the analytical parameters in graphs, *etc*.

Table 1

Analytical parameter (or variable)	Units	Allowed Range of Variability	Analytical Technique	Abbreviation
pH	pH units	5.0 – 8.0	pH-metry	ph
Residual water content	%	1.0 – 5.0	Karl-Fisher titration	h2o
Assay	%	80 - 92	HPLC	assay
Starting material residual content	%	≤ 0.20	HPLC	sm
Largest known impurity	%	≤ 0.20	HPLC	known
Largest unknown impurity	%	≤ 0.20	HPLC	unk
Total impurities content	%	≤ 1.0	HPLC	total
Residual solvent 1 content	%	≤ 5.0%	Gas-chromatography	solv1
Residual solvent 2 content	%	≤ 5.0%	Gas-chromatography	solv2
Residual solvent 3 content	%	≤ 1.0%	Gas-chromatography	solv3

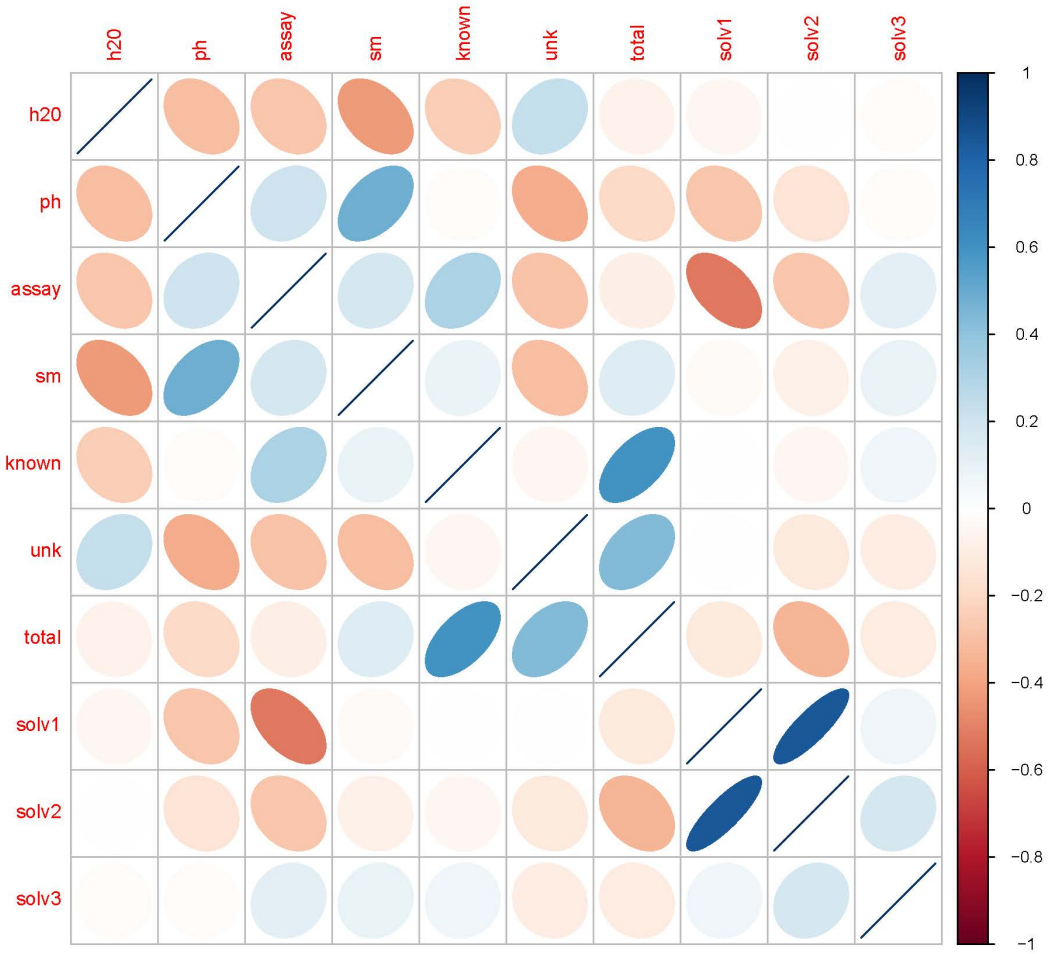
Data analysis and visualization have been performed using RStudio version 1.0.153 and R version 3.4.1 (The R Foundation for Statistical Computing). The following specific R packages have been used:

- *tidyverse* (H. Wickham, RStudio Inc., Boston, USA)[5]
- *FactoMineR* (F. Husson, Agrocampus Ouest, Rennes University, France) [6, 7]
- *factoextra* (A. Kassambara, HalioDx, Marseille, France) [8, 9,10]
- *corrplot* (T. Wei, Fujian Agriculture and Forestry University, China) [11, 12]
- *scatterplot3D* (U. Ligges, TU Dortmund, Germany) [13]
- *cluster* (M. Mächler, ETH Zürich, Switzerland) [14]

3. RESULTS AND DISCUSSION

In Figure 1 is displayed the correlation plot obtained on the initial data autoscaled.

Figure1



Each element of this diagram is a geometrical figure that becomes more and more elliptical and colored as the two initial variables gets more related each other. On the main diagonal, where the correlation is maximum (in fact the correlation of each element with itself is equal to one) the ellipses become a segment. Ellipses are right-oriented and blue colored if the two variables are positively correlated each other, while they are left oriented and red/brown colored if negatively correlated.

The lack of many elongated and deeply colored ellipses in Figure 1 indicate at glance that, in general, the variables are not highly correlated each other.

Figure 1 in fact shows strong correlations only between:

- the amount of the largest known impurity (known) and the total impurities amount (total) - positive
- the residual amount of solvents 2 and 1 - positive
- the residual amount of solvent1 (solv1) and the assay value (assay) – negative

For the rest, Figure 1 indicates weaker correlations such as those between:

- residual starting material (sm) and pH value (ph) - positive
- largest unknown impurity (unk) and impurities total amount (total) - positive
- residual amount of water (h2o) and largest unknown impurity content - positive
- residual starting material (sm) and residual water content (h2o) – negative.

These correlations, in particular the stronger ones, indicate some aspects of this manufacturing process worthy of further investigation such as the influence of the largest known impurity on the total impurities content or that to an increase in solvent 2 corresponds an increase in solvent 1 and a decrease in the assay value. Obviously, the correlations pattern becomes more and more informative as the manufacturing process in known in detail.

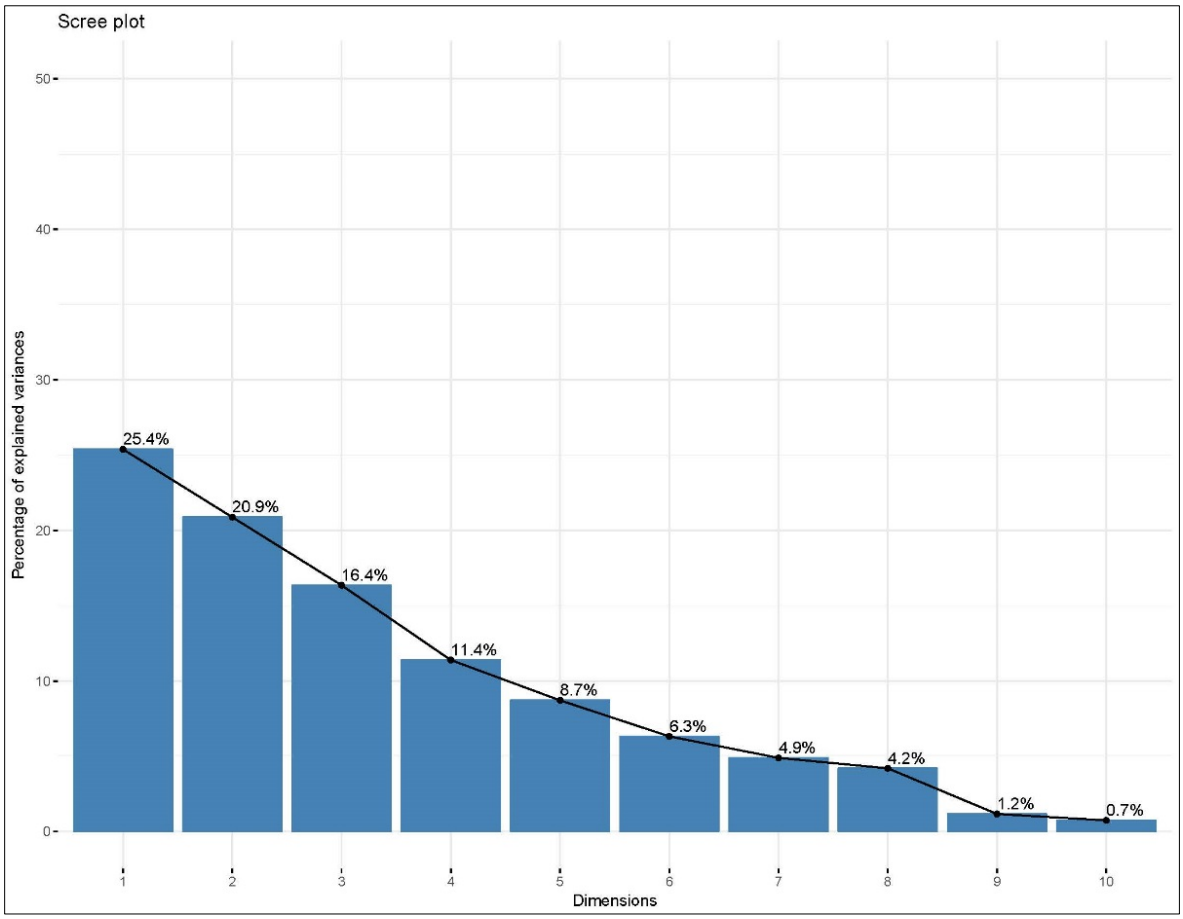
The correlation matrix, visualized in Figure 1 using the function *corrplot()* of R *corrplot* package, can be calculated using the function *cor()* of R *stats* package.

Principal Component analysis (PCA) [15], the oldest and most widely used multivariate method, is a powerful tool to summarize and visualize the information contained in a dataset described by multiple inter-correlated quantitative variables, which is a QC chemical dataset.

As, it has been said many times, the human eye, is the best pattern recognizer. However, this is true only when objects to be classified can be represented in two (or sometimes three) dimensions, that is, when they are characterized by only two or three variables. [16]. PCA with its capability of reducing the dimensionality of the initial data by removing noise and redundancy, is a useful tool for data display.

In Figure 2 is shown the scree plot, initially proposed by Cattell (1966) [17], which shows the eigenvalues of each component are plotted in successive order from the largest to the smallest. The scree plot has been obtained using the function *fviz_eig ()* of R *factoextra* package.

Figure 2



In light of the lack of strong correlations between variables already observed examining Figure 1, the diagram in Figure 2 shoes a clear *elbow* only after the eighth component (*i.e.*, at about 98% of explained variance). Nonetheless, to gain insight into data structure, the first two components (that account for about 46% of the total variation in the data) are enough.

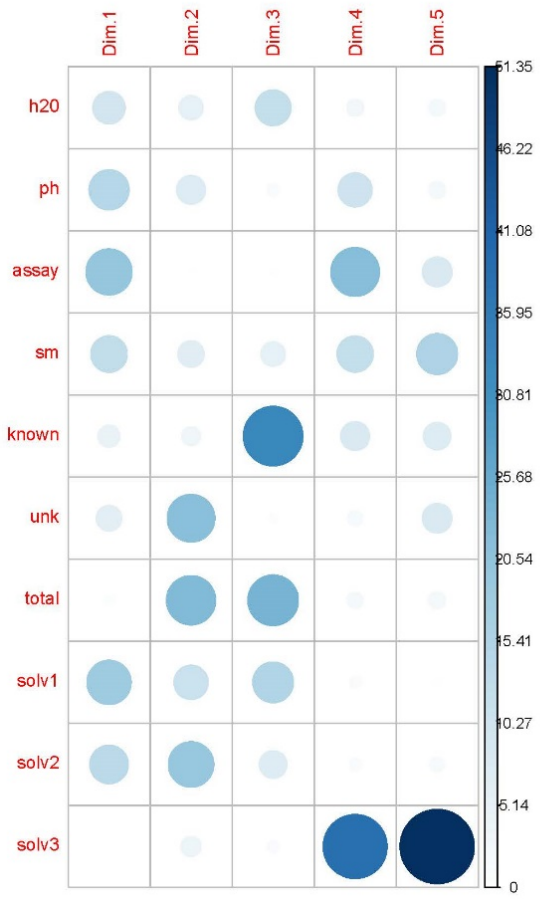
This looks clear examining Table 2 that summarizes the numerical compositions of the first five principal components (or dimensions).

Table 2

Variable	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
h20	9.76	5.45	11.89	2.63	2.50
ph	14.88	7.64	1.34	10.74	2.55
assay	19.73	0.00	0.38	22.08	8.20
sm	12.34	6.40	5.49	12.18	15.44
known	4.44	3.20	32.92	7.76	6.96
unk	6.03	21.10	0.72	1.95	8.13
total	0.88	22.48	23.80	2.17	2.52
solv1	18.17	10.89	15.32	1.19	0.51
solv2	13.77	19.20	7.00	1.22	1.83
solv3	0.01	3.64	1.15	38.09	51.35

To the first two principal components (*i.e.*, Dim. 1 and Dim. 2), in fact, contribute the majority of variables (eight out of ten) and each occurs with an important coefficient in the linear combination. An exception is represented by the contents of known impurity (known) and solvent 3 (solv3) that occur, first, in the third and fourth components. The observations deductible from data Table 2 are clearly visualized by the diagram in Figure 3.

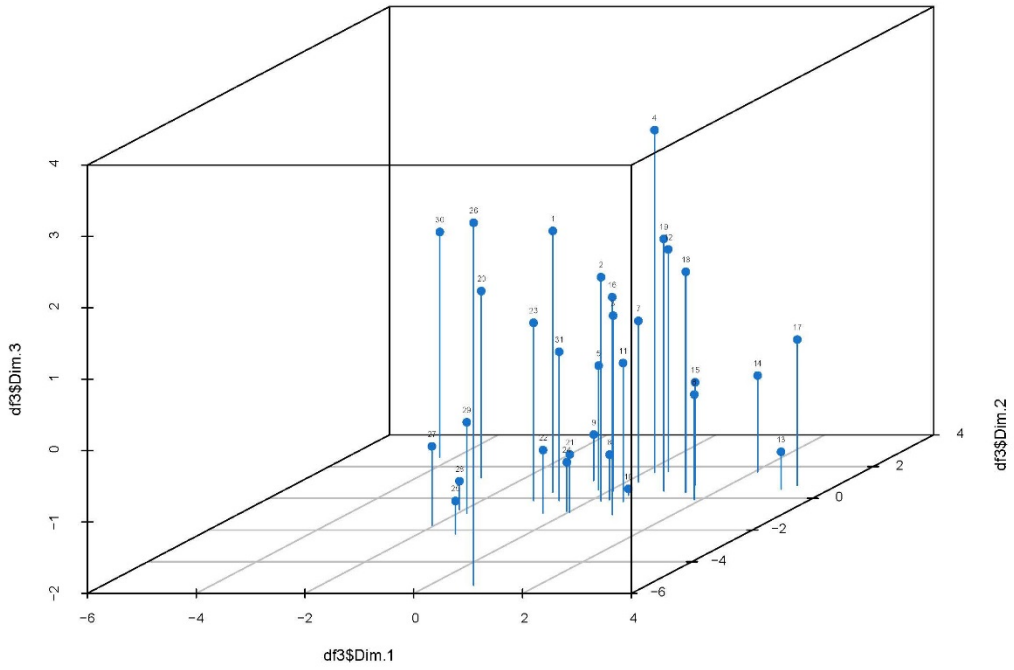
Figure 3



The plot in Figure 3 displays the contribution of each variable to each principal component or dimension. The larger is the contribution correlation, the darker blue is the spot. This plot has been obtained using the function *corrplot ()* of R *corrplot* package [11, 12].

Scatterplot is the oldest and widely used static graphical technique to begin exploring data. Considering the first three components (*i.e.*, about 63% of the total variation in the data), Figure 4 shows a 3D-scatterplot of the individuals obtained using the *scatterplot3d ()* function of the *scatterplot3d* R package for visualizing multivariate data [13]. In this scatterplot, each point represents a single lot.

Figure 4



In spite of the high percentage of total variation in the data considered, the diagram of Figure 4 does not visualize much about data distribution. For a more informative view it should be used a scatterplot matrix enhanced with contours of a 2d-density estimate such as those shown in Figures 5-7.

All these diagrams, obtained using in combination the *ggscatter ()* and the *geom_density2d ()* functions of the *ggplot2* R package [18], correspond to projections of the data points on two-dimensional sections of the scatterplot shown in Figure 4. Each section is cut along a plane defined by two axis each corresponding to a principal component, or dimension.

Figure 5

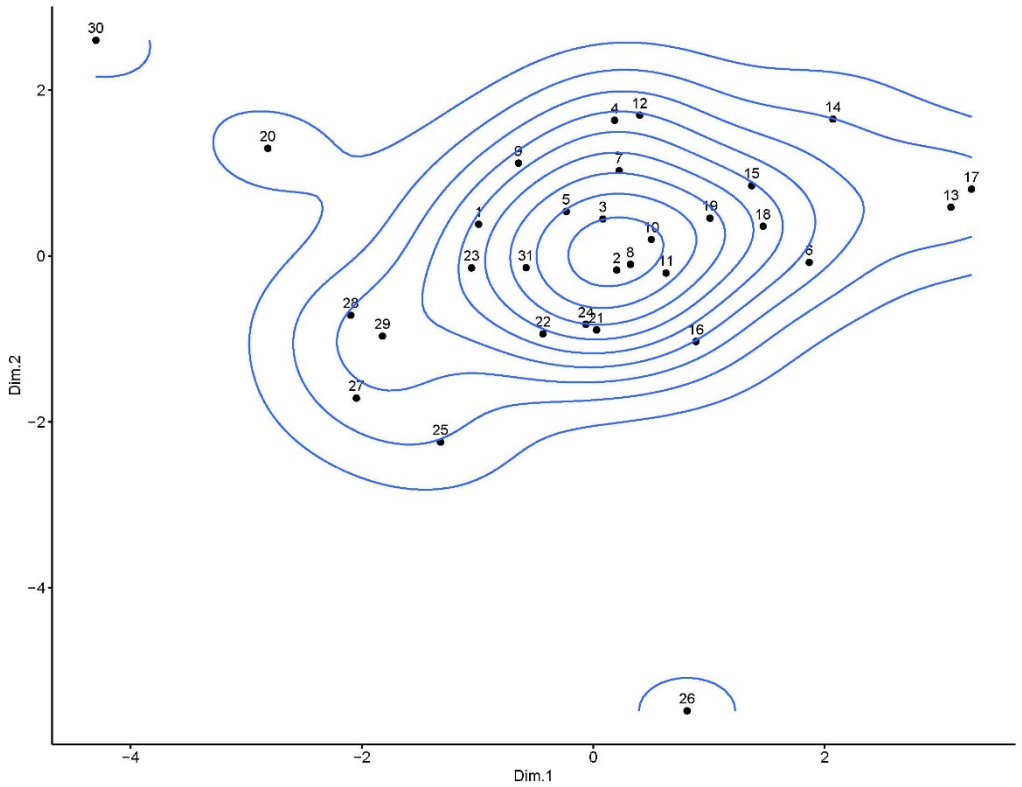


Figure 6

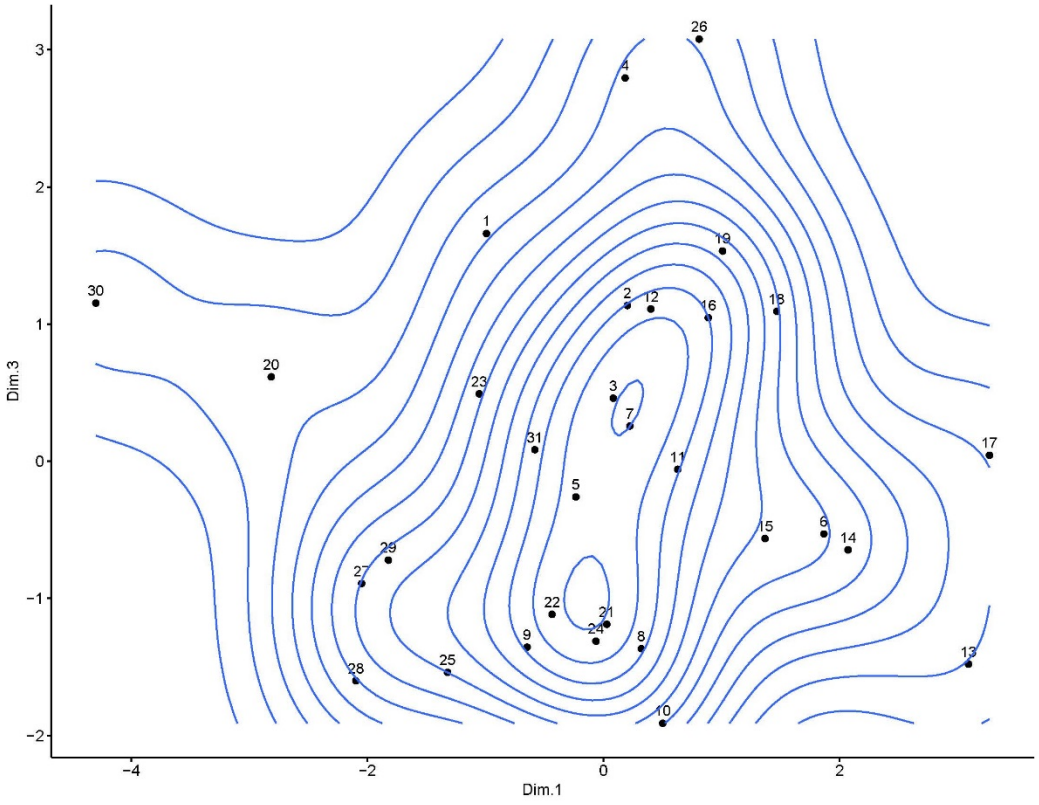
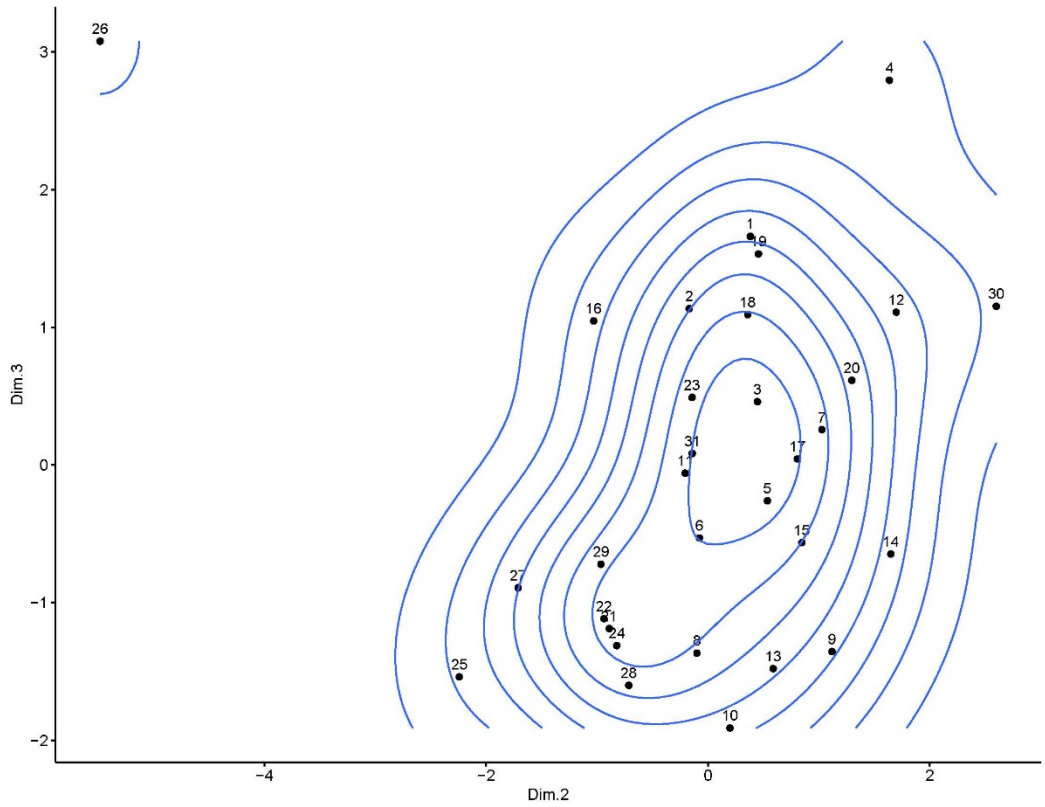


Figure 7



Among the three 2d-contour plots shown in Figures 5-7 that in Figure 5 is the one that better displays the cloud of data points. In the plane defined by the two first principal components, in fact, the cloud is projected in such a manner that the distortion of the cloud of points is minimized and, at the same time, it is captured the maximum variability.

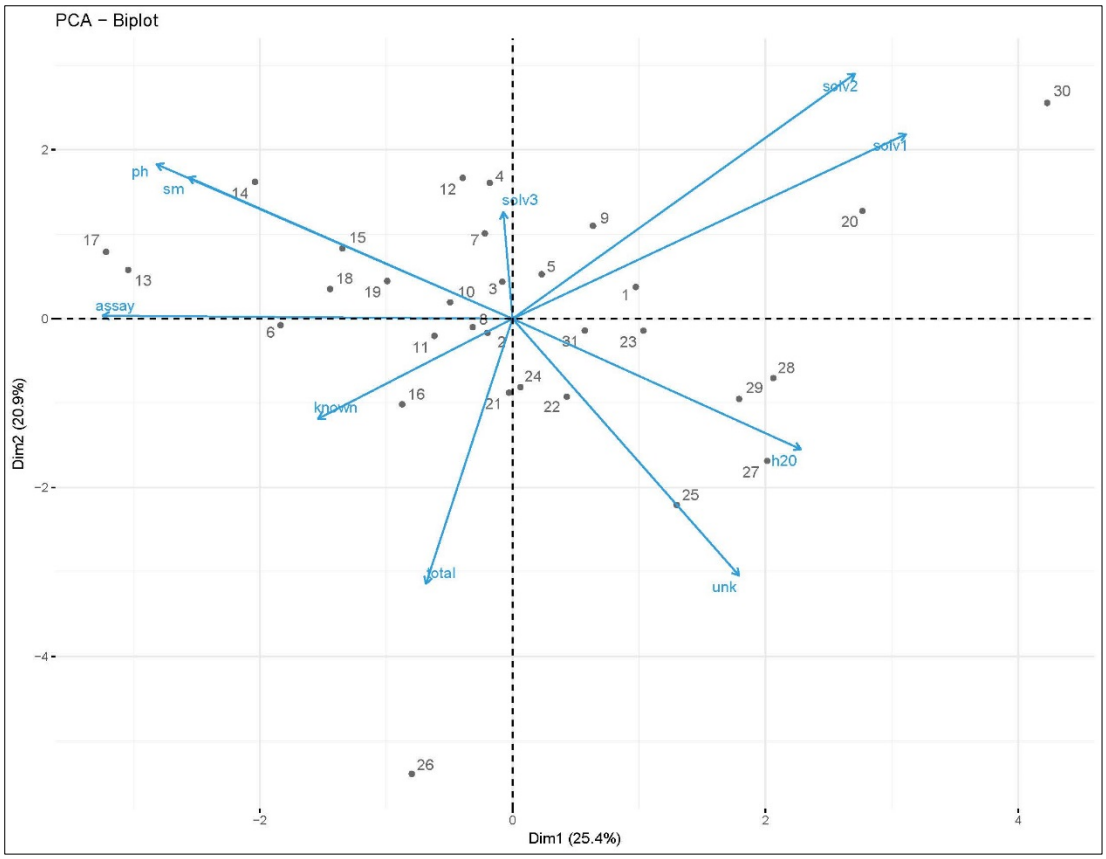
The exam of Figure 5 shows in fact that:

- the majority of lots appear centered around a central nucleus (see data points 2, 8 and 10)
- a few lots form a separate group on the left of the diagram (25, 27, 28, 29) and
- three lots are evidently unrelated with the rest (30, 26 and 20).

Figures 6 and 7 also capture the anomaly represented by these three last lots and suggest a possible data point's disposition aggregated around two centers.

In Figure 8 is displayed a PCA-Biplot obtained using the function *fviz_pca_biplot ()* of R *factoextra* package

Figure 8



This type of graphs display simultaneously individuals (*i.e.*, lots) and variables (*i.e.*, analytical parameters). The Biplot in Figure 8 is drawn using the first and second dimensions and it shows that:

- positively correlated variables (*e.g.*, solvent 1 and 2, water content and single largest unknown impurity, single known impurity and total impurities, pH and residual starting material) are grouped together,
- variables negatively related are on opposite quadrants,
- the intensity of each vector measures the quality of the variable represented on the map. For instance, the intensity of the vector associated to solvent3 is small as this variable is well represented on starting from the fourth dimension,
- the angles between vectors indicate the size of their correlations, small angles correspond to high correlations (*e.g.*, $\theta_{\text{pH, sm}} = 0°$) while wide angles indicate low correlations (*e.g.*, $\theta_{\text{solv1, known}} = 180°$).

Besides showing relationships between variables, Figure 8 also displays the relationships existing between the manufactured lots, each being represented as a point.

Figure 8, in fact shows:

- points look spread on the plane even if more concentrated near the origin. This finding is in line with what observed discussing Figure 5
- a few points look isolated and at the borders of the quadrants (*e.g.*, 30, 20 and 26). This suggests that the corresponding lots display characteristics different among them and with those of the rest of the lots. Being the variable *assay* coincident with the *x*-axis (the assay variable "dominates" the first component as to it correspond the higher coefficient), is reasonable to expect differences in this parameter among these lots. Moreover, being, for instance, lot 30 in the first quadrant and 26 in the opposite quadrant, one will display an assay value greater than the other.
- lots whose corresponding data points are very close on the map (*e.g.*, 21 and 24) display similar characteristics while those corresponding to data points a little bit more separated (*e.g.*, 17 and 13 or 28 and 29), slightly differ in their analytical profiles.

In light of the findings arising from the exam of Figures 5 – 8, it is reasonable considering *clustering* algorithms to investigate if, among the lots constituting the data set, it can be identified groups of similar individuals.

In particular, using the *HCPC ()* function of *FactoMineR* package to compute hierarchical clustering on principal components and the *fviz_cluster ()* function of *factoextra* R package to visualize individual clusters, it can be obtained the factor maps shown in Figures 9 and 10.

Figure 9

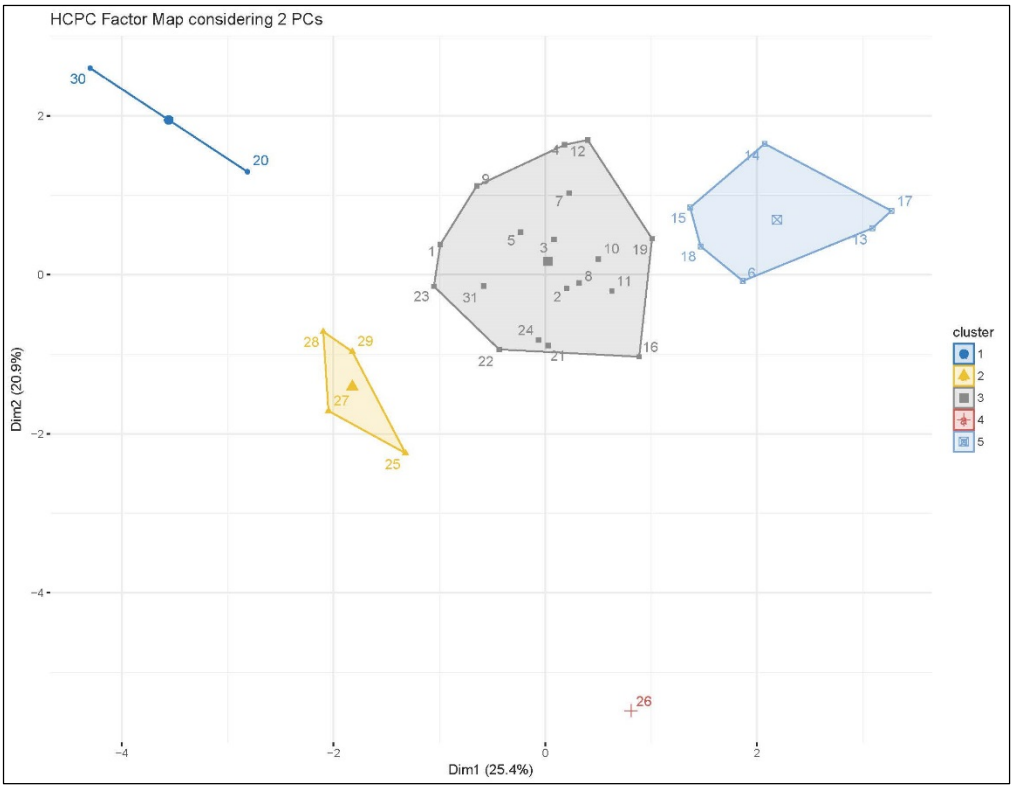
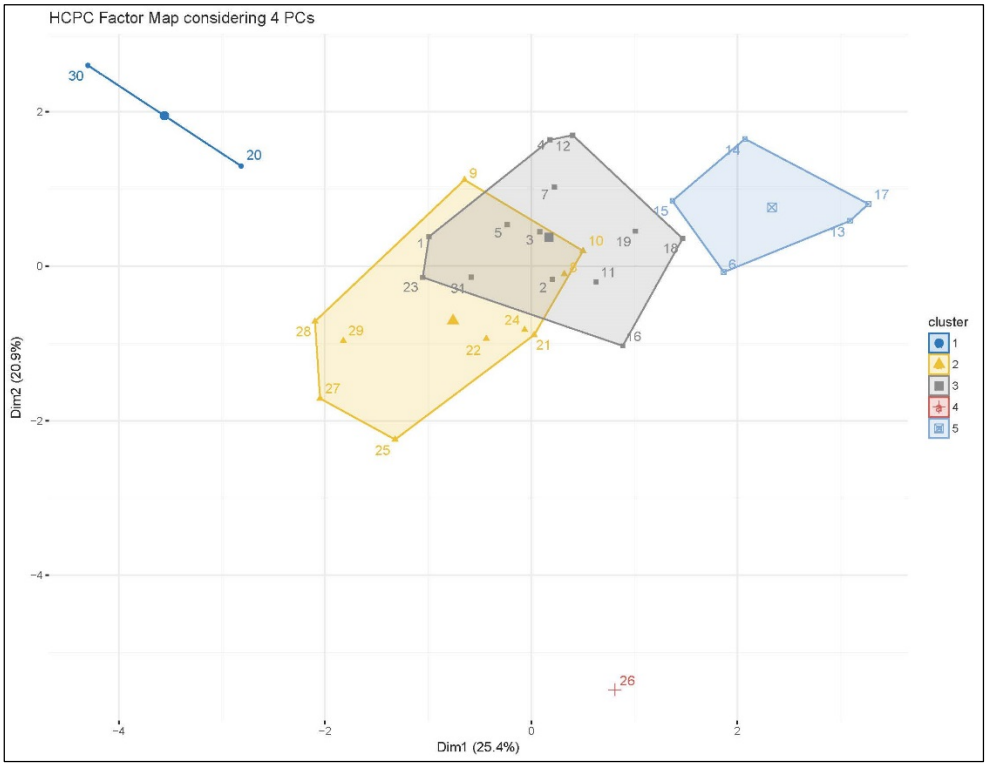


Figure 10



Both factor maps shown in Figures 9 and 10, even if obtained considering 2 and 4 principal components (that account, respectively, for 46% and 74% about of the total variation in the initial data), display five clusters, two of which show the same structure in both cases. In both factor maps, there are in fact two clusters (*i.e.*, 1 and 4) clearly separated from the remaining data points. Cluster 1 consists of two data points (*i.e.*, 20 and 30) while Cluster 4 of just one (*i.e.*, 26). This net separation between them and from the rest of data points is a clear indication of a different nature of the corresponding lots with respect to all others.

This fact was already evident examining of Figure 4.

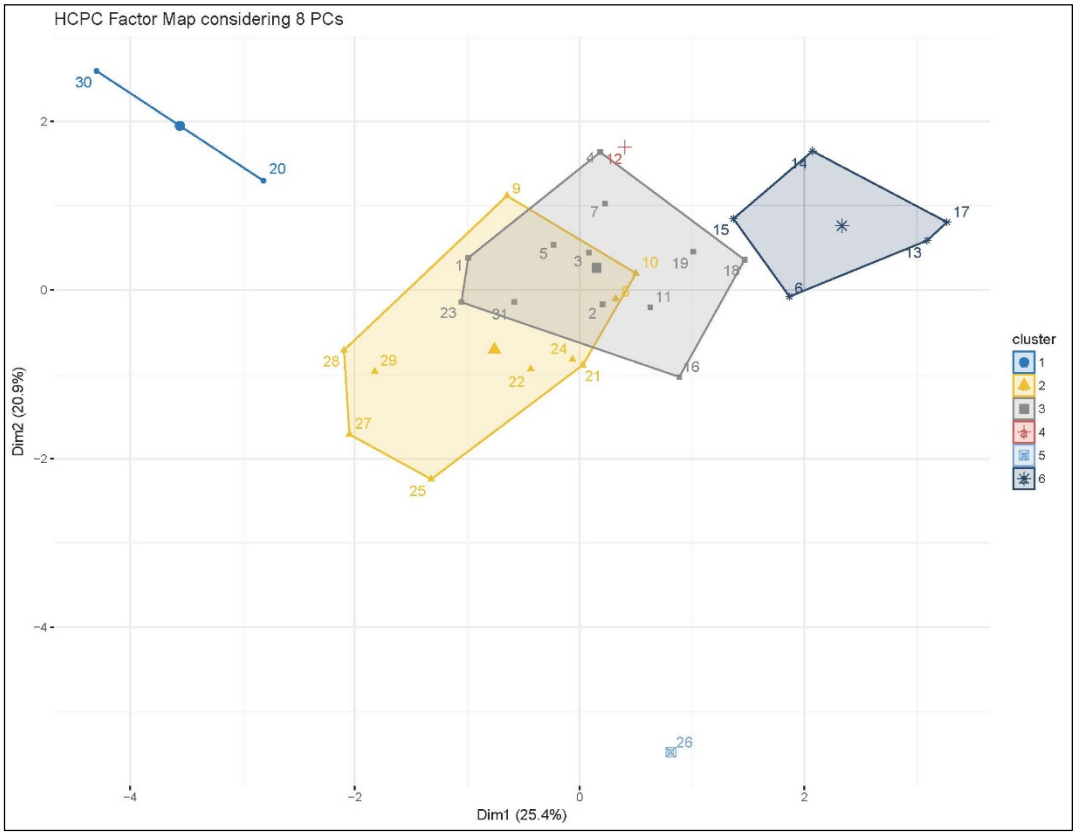
Interestingly, increasing the number of principal components from two to four, just leads to a change in the shape of clusters 2, 3 and 5 while the position of all centroids remains the same.

The presence of multiple clusters may indicate that the production method used is characterized by such a wide variability that groups of similar lots form populations different from each other. At the extreme, in case of clusters widely separated, one can even hypothesize that lots were manufactured using different methods instead of just one as expected.

In the example under analysis it applies the first possibility as the initial data have been intentionally built up to allow pattern recognition.

A further increase in the number of principal components considered (*e.g.*, eight, that correspond to about 98% of the total variation in the initial data) does not dramatically change the clusters structure as shown in Figure 11 that looks rather similar to Figure 10. In this case, duplicating the number of principal components considered for clustering does not duplicate the information returned by cluster analysis.

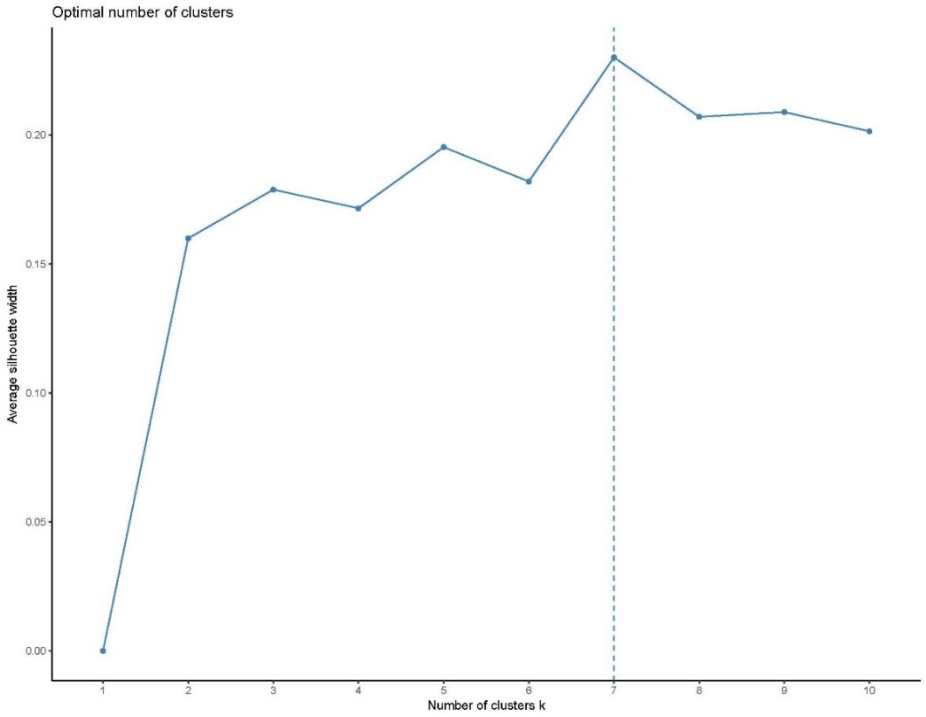
Figure 11



Besides hierarchical clustering other classification techniques are available (*e.g.*, partitioning and mixture models) and, obviously, different classification methods can lead to different patterns. As hierarchical clustering techniques impose a hierarchical structure on data [19], by way of example, it has also been considered a different clustering approach. In particular, it has been selected the PAM algorithm (*Partitioning Around Medoids*) as it represents, among the partitioning clustering methods, that less sensitive to outliers. As partitioning clustering algorithms require to specify the number of clusters to be generated, this value has been calculated using the *pam ()* function of the *cluster* R package [14] and visualized using the *fviz_nbclust ()* function of *factoextra* R package.

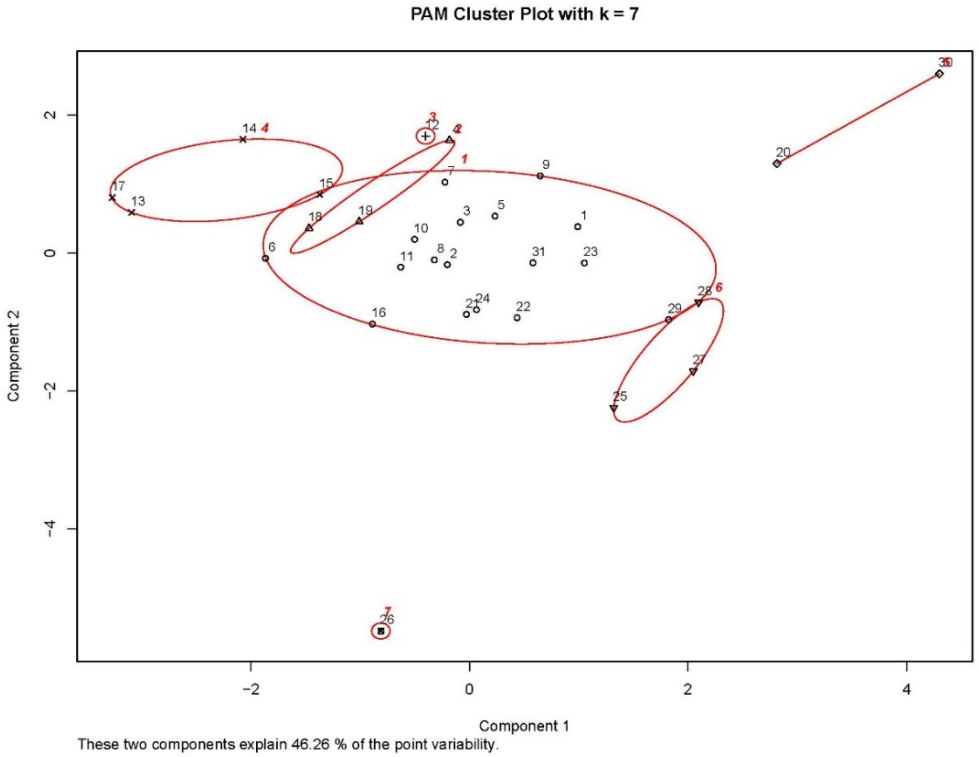
The resulting diagram is displayed in Figure 12.

Figure 12



Combining the *pam ()* and the *fviz_cluster ()* functions, respectively from *cluster* and *factoextra* R packages, in Figure 3 are displayed the clusters.

Figure 13



The comparison between Figure 13 and Figures 9 – 11 is very interesting. In fact, even if they show the results of two different types of cluster analysis, based on different approaches (hierarchical *vs.* partitioning) and on different data (principal components *vs.* initial data scaled) they display similar patterns. In both cases in fact data points 20 and 30 form a single cluster (cluster 6 in Figure 13) as well as data point 27 (cluster 7 in Figure 13). Moreover, the central part of the factor maps show similar patterns.
This very short cluster analysis, even if just sketched, clearly shows the presence of underlying patterns in data and suggests the need of a more in-depth study.

4. CONCLUSIONS

Multivariate analysis is a tremendously powerful tool for data analysis in general and it is extremely useful even for pharmaceutical Quality Control data as it provides an immediate data overview not otherwise possible with a univariate approach. The combined use of many excellent statistical and graphical packages available for R makes easy the data analysis and the interpretation of results. The multivariate approach does not just allow to reveal patterns or outliers at a glance as well as fake data, but, more important, it provides in depth insight of the manufacturing process and of the relationships existing between analytical parameters.

5. ACKNOWLEDGMENTS

I wish to express my deepest gratitude and appreciation to all Authors of the R packages that I have used in this post.

6. BIBLIOGRAPHY

1. Code of Federal Regulation part 21 §§ 211.22 and 212.70
2. B. Everitt, T. Hothorn, *An Introduction to Multivariate analysis with R*, Springer, 2011
3. S. Wold *et al.*, *Multivariate Data Analysis in Chemistry*, Proceedings of the NATO Advanced Study Institute on Chemometrics – Mathematics and Statistics in Chemistry, Cosenza, Italy, September 12-223, 1983, edited by B.R. Kowalski, Springer, 1984
4. Chambers, J. M., Cleveland, W. S., Kleiner, B., and Tukey, P. A. (1983). *Graphical Methods for Data Analysis*. New York: Chapman and Hall.
5. H. Wickham, G. Grolemund, *R for Data Science*, 2017, O'Reilly
6. F. Husson, S. Lê, J. Pagès, *Exploratory Multivariate Analysis by Example using R*, 2011, CRC Press.
7. 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.
8. A. Kassambara, *R Graphics Essentials for Great Data Visualization*, STHDA, 2017
9. A. Kassambara, *Practical Guide to Principal Component Methods in R*, STHDA, 2017
10. A. Kassambara, *Practical Guide to Cluster Analysis in R*, STHDA, 2017
11. M. Friendly, *Corrgrams: Exploratory displays for correlation matrices*, The American Statistician, 56 (2002) 316-324
12. D.J. Murdoch, E.D. Chow, *A graphical display of large correlation matrices*, The American Statistician, 50 (1996) 178-180

13. U. Ligges, M. Mächler, *Scatterplot3d – an R Package for Visualizing Multivariate Data*, J. of Statistical Software, 8 (11), 2003, 1-20
14. L. Kaufman, P.J. Rosseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*, Wiley, New York, 1990
15. I.T. Jolliffe, *Principal Component Analysis*, 2nd Edition, 2002, Springer
16. D.L. Massart, L. Kaufman, *The Interpretation of Analytical Chemical Data by the use of Cluster Analysis*, J. Wiley & Sons, New York, 1983
17. R.B. Cattell, *The screen test for the number of factors*, Multivariate Behavioral Research, 1(1966), 140-161
18. H. Wickham, *ggplot2: Elegant Graphics for Data Analysis*, Second Edition, Springer, 2016
19. B.S. Everitt, G. Dunn, *Applied Multivariate Data Analysis, Second Edition*, Wiley, 2001