3

Principal Component Analysis: A Powerful Interpretative Tool at the Service of Analytical Methodology

Maria Monfreda
Italian Customs Agency
Central Directorate for Chemical Analysis and Development of Laboratories, Rome, Italy

1. Introduction

PCA is one of the most widely employed and useful tools in the field of exploratory analysis. It offers a general overview of the subject in question, showing the relationship that exists among objects as well as between objects and variables.

An important application of PCA consists of the characterization and subsequent differentiation of products in relation to their origin (known as traceability). PCA is often applied in order to characterize some products obtained via a manufacturing process and the transformation of some raw materials. In this case, there are two kinds of elements linkable to the differentiation of products in relation to their origin: the variability associated to the raw material and the differences in various production techniques used around the world. In this study, two examples of PCA application to some products obtained via a manufacturing process are presented. These products, belonging to completely different fields (foodstuffs and petroleum based fuel) show one element in common: their traceability is correlated to the raw material and the production process.

The strength of PCA is that it provides the opportunity to visualize data in reference to objects described by more than 3 variables. Indeed, PCA allows us to study and understand such systems, helping the human eye to see in two or three dimension systems that otherwise would necessarily have to be seen in more than three dimensions in order to be studied. PCA allows data to maintain their original structure, making only an orthogonal rotation of variables, which helps to simplify the visualization of all the information already contained in the data. Consequently, PCA can be considered the best technique to begin to approach any qualitative multivariate problem, be it unsupervised or supervised. Needless to say, supervised problems - following a primary study by PCA - require the application of either a classification or a class modeling method. In this study, three cases regarding supervised problems which involved the preliminary application of PCA are put forward. Results from PCA have been compared to those obtained from classification or class modeling tools.

2. PCA and traceability

PCA is widely used to characterize foodstuffs according to their geographical origin (Alonso-Salces et al., 2010; Diaz et al., 2005; Gonzalvez et al. 2009; Marini et al., 2006). Such a requirement is becoming prominent in the control field, especially in the marketing of products with PDO (Protected Denomination or Origin) or PGI (Protected Geographical Indication) markings. The PDO marking is awarded to products linked strictly to a typical area. Both the production of raw materials and their transformation into the final product must be carried out in the region that lends its name to the product. As a consequence, some analytical methods, whose results could be directly linked to the sample origin, would be extremely useful in the legal battle against the fraudulent use of PDO or PGI marking.

The local nature of a food product, strongly associated with its geographical location, can be correlated to the quality of the raw material used and its production techniques. Environmental conditions in a specific geographical area also provide the raw material with set characteristics, becoming a factor of primary importance in determining the final product "typicality". The production technique is of primary importance for both agricultural products and so-called transformed products, where culture, the instruments used, the ability and experience of the operator and the addition of particular ingredients create a unique product. Brescia et al. (2005) characterized buffalo milk mozzarella samples with reference to their geographical origin (two provinces, namely Foggia, in Apulia and Caserta, in Campania, were considered), by comparing several analytical and spectroscopic techniques. Some analyses were also performed on the raw milk (from which mozzarella had been obtained) with the purpose of evaluating how the differences among milk samples had transferred to the final product. In this study, a further PCA was applied only to those analytical variables measured on both milk and mozzarella samples: fat, ash, Li, Na, K, Mg, Ca, $\delta^{15}N/^{14}N$ e $\delta^{13}C/^{12}C$, disregarding all the analyses carried out only on mozzarella samples for which any comparison with milk samples could not be performed and vice versa. The biplots relative to PCA carried out on milk and mozzarella samples are reported in figures 1 and 2 respectively. It is easy to see that the milk samples are completely separated, according to their origin, on the PC1 (figure 1), whilst mozzarella samples lose such a strong separation, even though they maintain a good trend in their differentiation.

As already stated by Brescia et al., milk samples from Campania have a higher ^{13}C content, whilst samples from Apulia have a greater Li, Na and K content. If PCA results relative to mozzarella samples are compared to those from milk samples, it can be deduced that geographical differences, very clearly defined in the raw material, tend to drop slightly in the final product. There is a factor (K content) whose distribution is inverted between the raw material and the final product (positive loading on PC1 for milk samples and negative loading on PC1 for mozzarella samples). Another factor (Na content) was a discriminator for the raw milk (high positive loading on PC1) but its loading in mozzarella samples rises on the PC2 (the direction perpendicular to the geographical separation) and becomes negative on PC1. As Na content is known to be linked to the salting process of a cheese, the production technique is thought to reduce some differences originating from the raw materials. In other words, the differences that exist between buffalo mozzarella from Campania and Apulia are mainly determined by the differences between the two types of raw milk, rather than between manufacturing processes.

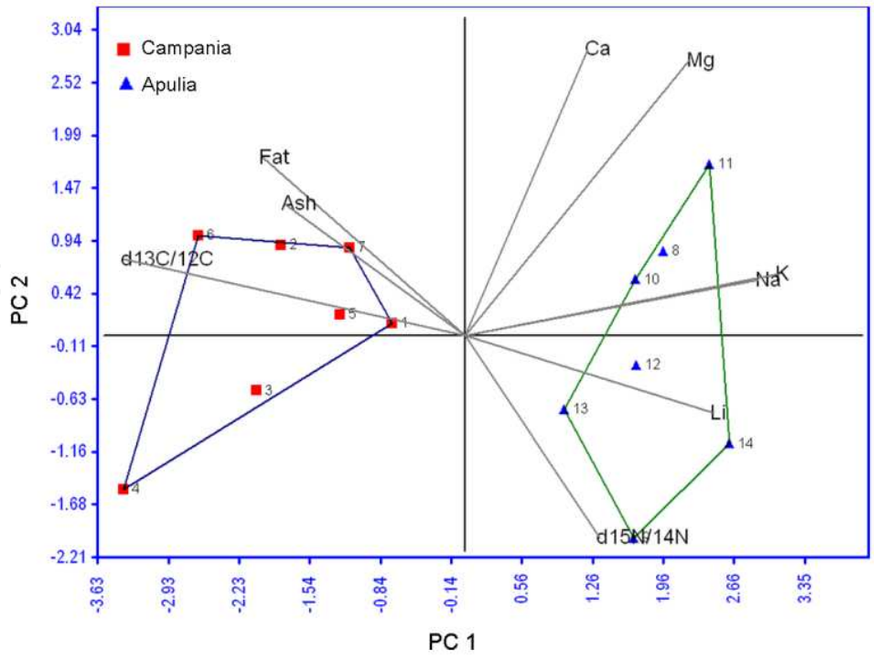

Fig. 1. Score plot of PC2 versus PC1 for milk samples.

Tables 1 and 2 show variances and cumulative variances associated to the principal components with eigenvalues greater than 1 for milk and mozzarella samples respectively. 4 PCs were extracted for both data set, which explain 86% of the variance for milk samples and 83% of variance for mozzarella samples.

PC	Variance %	Cumulative %
1	40.23	40.23
2	20.93	61.16
3	13.57	74.73
4	11.33	86.06

Table 1. PCs with eigenvalues greater than 1, extracted applying PCA to milk samples.

PC	Variance %	Cumulative %
1	35.95	35.95
2	20.45	56.40
3	15.10	71.50
4	11.83	83.33

Table 2. PCs with eigenvalues greater than 1, extracted applying PCA to mozzarella samples.

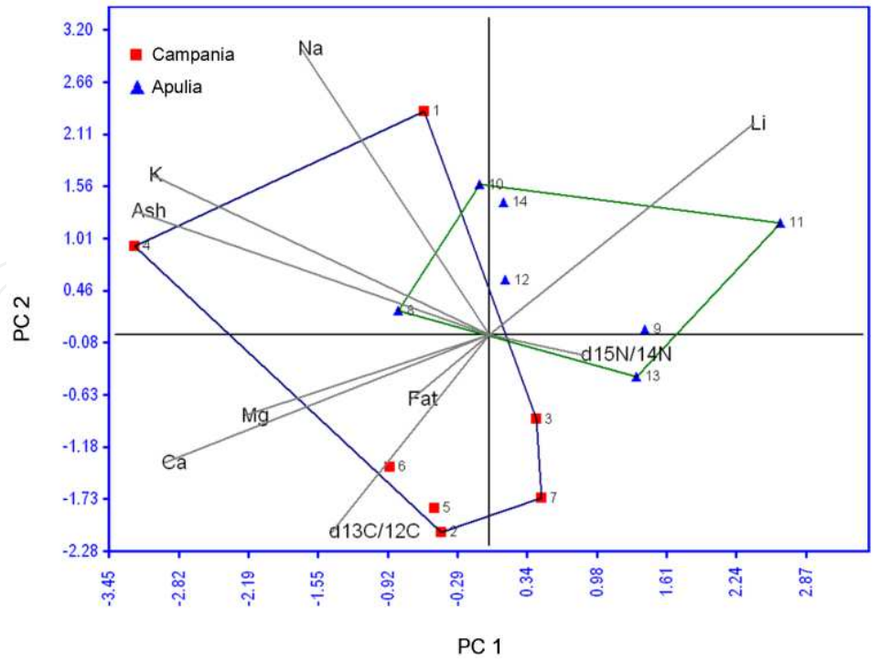

Fig. 2. Score plot of PC2 versus PC1 for mozzarella samples.

From this example, it can be deduced that the application of PCA to results obtained from chemical analyses of the raw material from which a transformed product has been obtained allows a characterization of the raw material in relation to its geographical origin. Secondly, the transformed product characterization allows to see how geographical differences among the raw materials have been spread out in the final product. In particular, it can be seen whether production techniques amplified or, indeed, reduced the pre-existing differences among the varying classes of the raw material. In other words, the application of PCA to the chemical analyses of a food product – as well as the raw material from which it has been made - allows to understand what the main elements are that provide a product characterization in relation to its origin: i.e. the quality of the raw material, the production techniques, or in fact a combination of both.

The characterization of products in relation to their origin is, however, not only important for food products. In forensic investigations, for example, it is becoming increasingly essential to identify associations among accelerants according to their source. Petroleum-based fuels (such as gasoline, kerosene, and diesel), which are often used as accelerants as they increase the rate and spread of fire, are also in fact transformed products from raw material (petroleum). Differentiation of such products in relation to their source (brand or refinery) depends both on the origin of the petroleum and the specific production techniques used during the refining process. Monfreda and Gregori (2011) differentiated 50 gasoline samples belonging to 5 brands (indicated respectively with the letters A, B, C, D and E) according to their refinery. Samples were analyzed by solid-phase microextraction (SPME) and gas chromatography-mass spectrometry (GC-MS). Some information on the origin of the crude oil was available but only for two of the brands: A samples were obtained from crude oil coming from only one country, whilst D samples were produced from crude oil coming from several countries. In addition A samples were

tightly clustered in the score plots while D samples were fairly well spread out in the same score plots. This evidence was explained by considering that crude oil coming from only one place might have consistent chemical properties, compared to crude oils coming from several countries. Therefore differences existing between the raw materials had been transferred to the final products, determining very clustered samples with consistent chemical properties (for A brand) and samples with a greater variability within the class (for D brand). The score plot of PC2 versus PC1, shown in figure 3, was obtained by Monfreda and Gregori (2011).

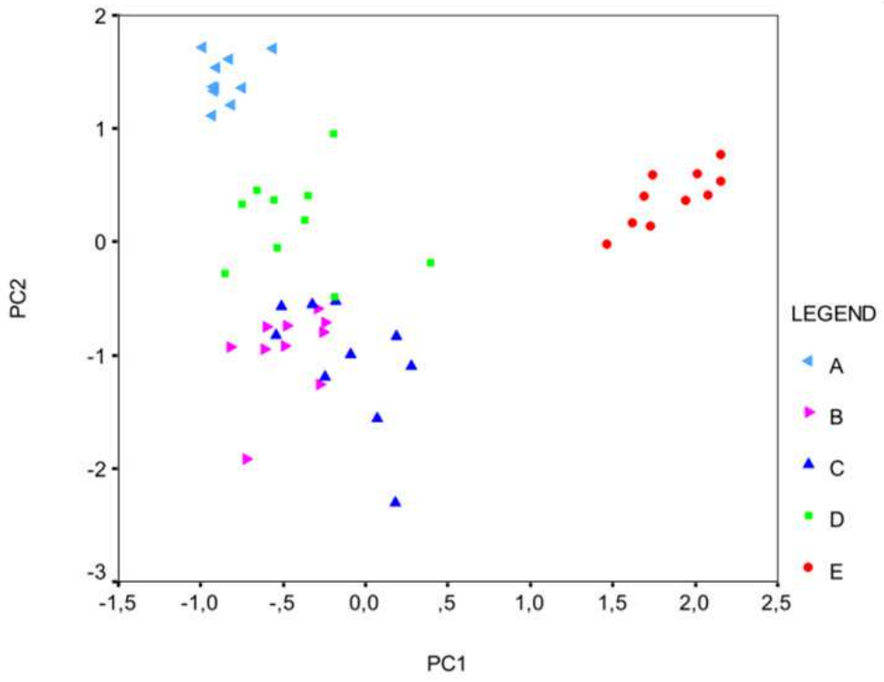

Fig. 3. Score plot of PC2 versus PC1 for gasoline samples (obtained by Monfreda & Gregori, 2011).

In the study presented here, 25 diesel samples belonging to the same 5 brands studied by Monfreda and Gregori were analysed using the same analytical procedure, SPME-GC-MS. As in the previous work, chromatograms were examined using the TCC approach (Keto & Wineman, 1991, 1994; Lennard at al., 1995). Peak areas were normalized to the area of the base peak (set to 10000), which was either tridecane, tetradecane or pentadecane, depending on the sample. Three independent portions for each sample of diesel were analyzed and peak areas were averaged. Analysis of variance was carried out before the multivariate statistical analysis, in order to eliminate same variables whose variance between classes was not significantly higher than the variance within class. Tetradecane, heptadecane, octadecane and hexadecane tetramethyl were then excluded from multivariate statistical analysis. PCA was finally applied to a data set of 25 samples and 33 variables, as listed in table 3.

Variable	COMPOUND
1	Nonane
2	Octane, 2,6-dimethyl
3	Benzene, 1-ethyl, 2-methyl
4	Decane
5	Benzene, 1,2,3-trimethyl
6	Benzene, 1-methyilpropyl
7	Nonane, 2,6-dimethyl
8	Benzene, 1-methyl-2-(1-methylethyl)
9	Benzene, 1,2,3-trimethyl
10	Cyclohexane, butyl
11	Benzene, 1-methyl-3-propyl
12	benzene, 4-ethyl-1,2-dimethyl
13	benzene, 1-methyl-2-propyl
14	Benzene, 1-methyl-4-(1-methylethyl)
15	Benzene, 4-ethyl-1,2-dimethyl
16	Undecane
17	Benzene, 1-ethyl-2,3-dimethyl
18	Benzene, 1,2,3,5-tetramethyl
19	Benzene, 1,2,3,4-tetramethyl
20	Cyclohexane, pentyl
21	Dodecane
22	Undecane 3,6-dimethyl
23	Cycloexane, hexyl
24	Tridecane
25	Naphthalene, 2-methyl
26	Naphthalene, 1-methyl
27	Pentadecane
28	Hexadecane
29	Pentadecane tetramethyl
30	Nonadecane
31	Eicosane
32	Heneicosane
33	Docosane

Table 3. Target compounds used as variables in multivariate statistical analysis of diesel samples.

Three PCs were extracted, with eigenvalues greater than 1, accounting for 92.16% of the total variance, as shown in table 4. From the score plot of PC2 versus PC1 (figure 4), it can be seen that a separation of samples according to the refinery was achieved, because each group stands in a definite area in the plane of PC1 and PC2. A samples are more clustered than D samples, according to the results obtained for gasoline samples.

PC	Variance %	Cumulative %
1	59.48	59.48
2	20.70	80.18
3	11.98	92.16

Table 4. PCs with eigenvalues greater than 1, extracted applying PCA to diesel samples.

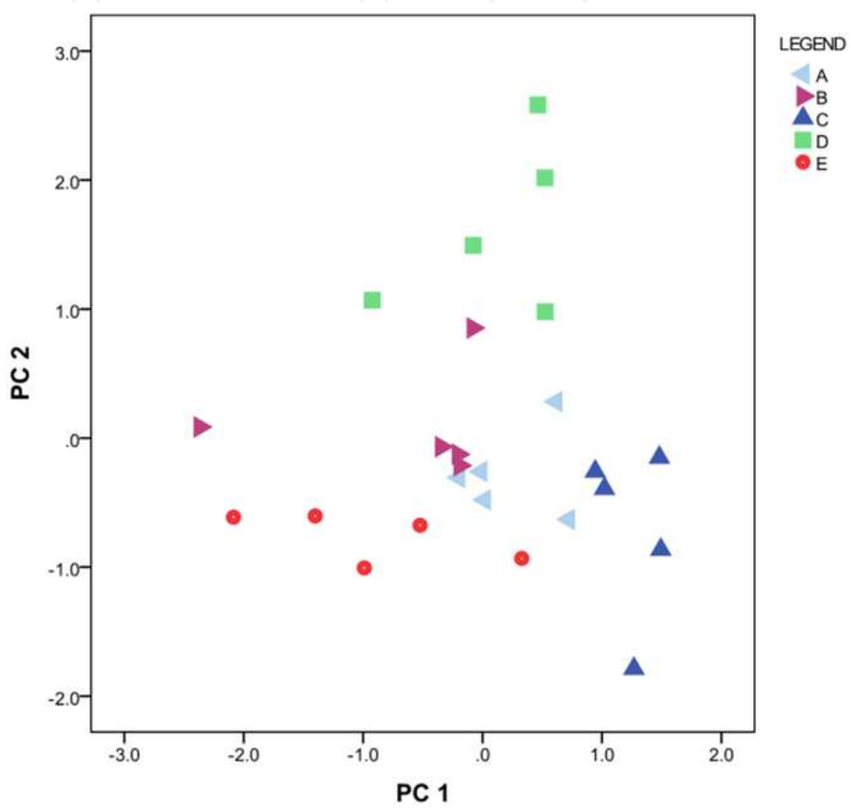

Fig. 4. Score plot of PC2 versus PC1 for diesel samples.

Results of both studies, carried out respectively on gasoline and diesel samples coming from the same five refineries, allow to achieve a traceability of these products according to their brands, that is to say that production techniques give well-defined features to these products. Properties of crude oil, otherwise, show a strong influence on the homogeneity of samples distribution within their class, based on information availability (only for two of five refineries).

3. The PCA role in classification studies

3.1 Case 1

The gasoline data matrix has been used in real cases of arson to link a sample of unevaporated gasoline, found at a fire scene in an unburned can, to its brand or refinery. This helped to answer, for example, questions posed by a military body about the origin of an unevaporated gasoline sample taken from a suspected arsonist. The gasoline sample

under investigation was analyzed with the same procedure adopted by Monfreda and Gregory (2011) and using the same devices. Analyses were carried out almost in the same period in which the 50 samples of the previous work had been analyzed. Three independent portions of the sample were analyzed and from the Total Ion Chromatogram (TIC) of each analysis, a semi-quantitative report of peak areas of the same target compounds (TCs) used by Monfreda and Gregori was obtained. Areas were normalized to the area of the base peak (benzene, 1,2,3-trimethyl), set to 10000, as in the previous study. The average areas (of the three portions analyzed) corresponding to the aromatic compounds were appended to the data matrix of 50 gasoline samples analyzed by Monfreda and Gregori. A PCA was then applied to a data set of 51 samples and 16 variables. Results are shown in the scatter plots of figures 5, 6 and 7. From these scatter plots it can be seen that the sample under investigation is significantly different from those of the A and E brands. As a consequence, these two refineries could be excluded from further investigations by the relevant authorities because the membership of the unknown sample to A or E brands was less likely than it belonging to other classes. The score plot of PC2 versus PC1 (figure 5) shows the unknown sample among the classes B, C and D. From the score plot of PC3 versus PC1 (figure 6), it can be seen that the unknown sample is very close to those of class B, and quite distant from class C. The unknown sample, however, falls into an area where some samples of D brand are also present. Finally, from the scatter plot of PC3 versus PC2 and PC1 (figure 7), the sample under investigation would appear to fall between the B and D classes.

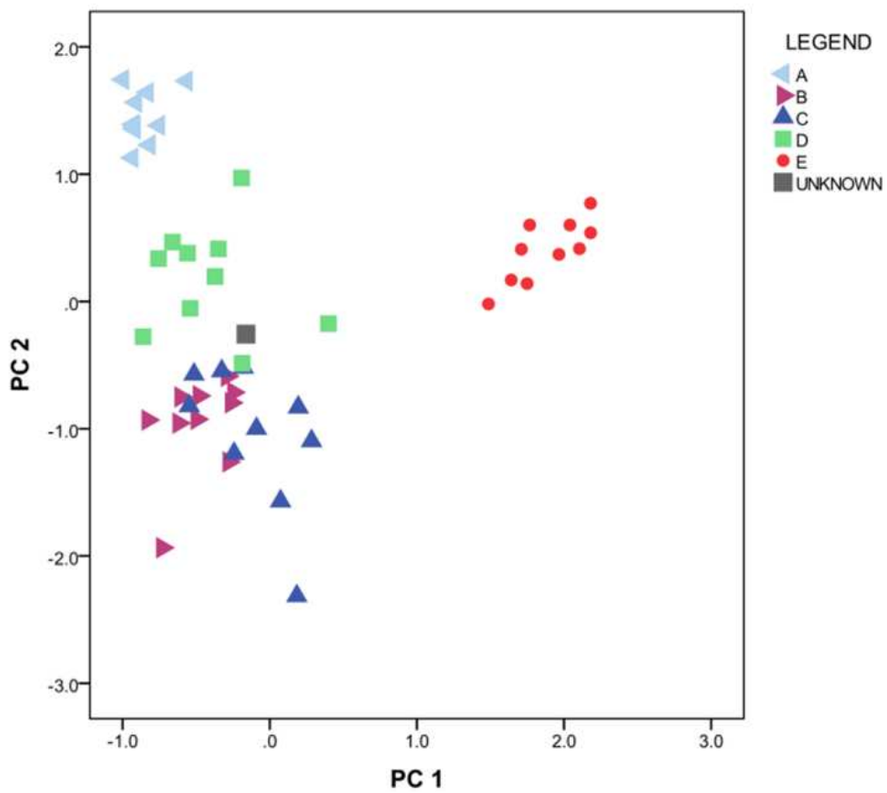

Fig. 5. Score plot of PC2 versus PC1 for 51 gasoline samples.

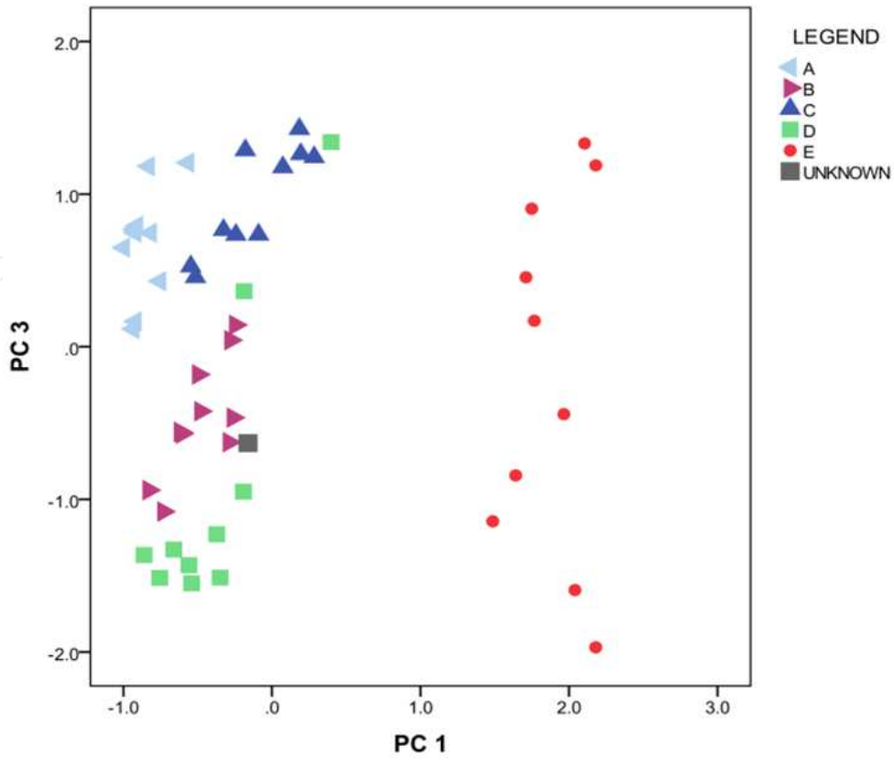

Fig. 6. Score plot of PC3 versus PC1 for 51 gasoline samples.

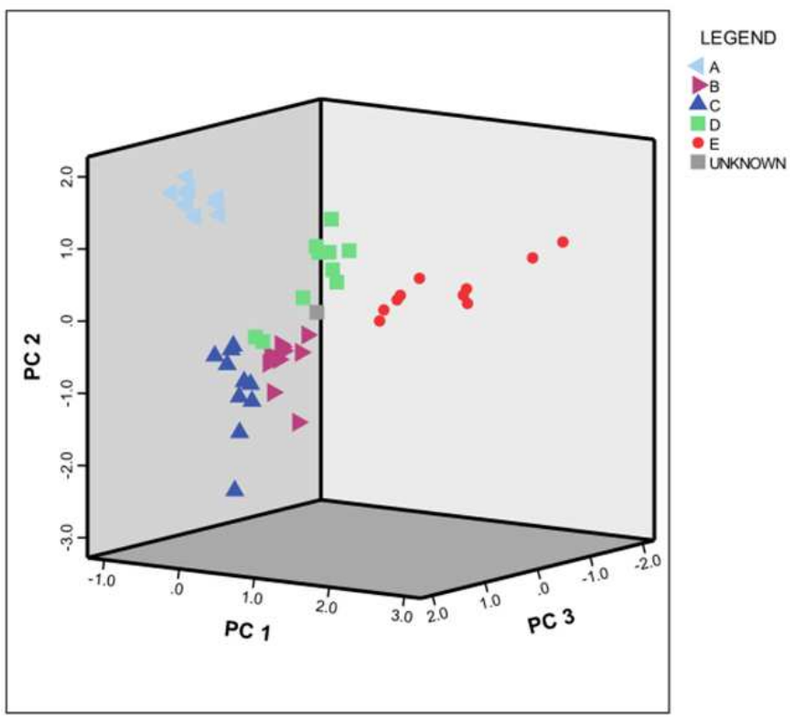

Fig. 7. Score plot of PC3 versus PC2 versus PC1 for 51 gasoline samples.

The application of PCA was especially useful for an initial visualization of data, however the question posed by the military body also needed to be handled with some supervised methods; in other words, discriminant analysis or class modeling tools. In such a way, the system is forced to create a boundary between classes and eventually the unknown sample is processed. For this kind of problem, class modeling tools are clearly preferable to discriminant analysis, in that they first create a model for each category as opposed to creating a simple delimiter between classes. The modeling rule discriminates between the studied category and the rest of the universe. As a consequence, each sample can be assigned to a single category, or to more than one category (if more than one class is modeled) or, alternatively, considered as an outlier if it falls outside the model. Discriminant analysis tends, however, to classify in any case the unknown sample in one of the studied categories even though it may not actually belong to any of them. In this case, the class modeling technique known as SIMCA (Soft Independent Models of Class Analogy) was applied to the data set under investigation. SIMCA builds a mathematical model of the category with its principal components and a sample is accepted by the specific category if its distance to the model is not significantly different from the class residual standard deviation. This chemometric tool was applied considering a 95% confidence level to define the class space and the unweighted augmented distance (Wold & Sjostrom, 1977). A cross validation with 10 cancellation groups was then carried out and 8 components were used to build the mathematical model of each class. The boundaries were forced to include all the objects of the training set in each class, which provided a sensitivity (the percentage of objects belonging to the category which are correctly identified by the mathematical model) of 100%. Results are shown in the Cooman's plots (figures 8, 9 and 10), where classes are labeled with the numbers 1 to 5 instead of the letters A to E respectively. The specificity (the percentage of objects from other categories which are classified as foreign) was also 100%.

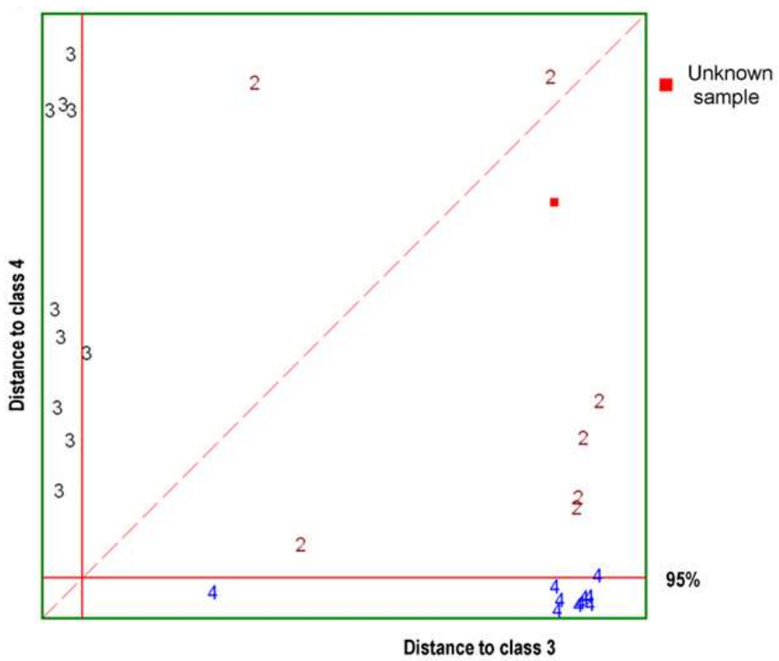

Fig. 8. Cooman's plot for the classes 3 (C) and 4 (D).

From the Cooman's plot of classes 3 (C) and 4 (D) (figure 8), the unknown sample (red square) results in an outlier but is closer to class 4 than to class 3. In figure 9, the distances from classes 2 (B) and 4 (D) are displayed and the sample under investigation remains an outlier, but its distance from class 2 is shorter than the equivalent from class 4. In figure 10, where the distances from classes 1 (A) and 5 (E) are plotted, the unknown sample is missing as it is too far from both classes.

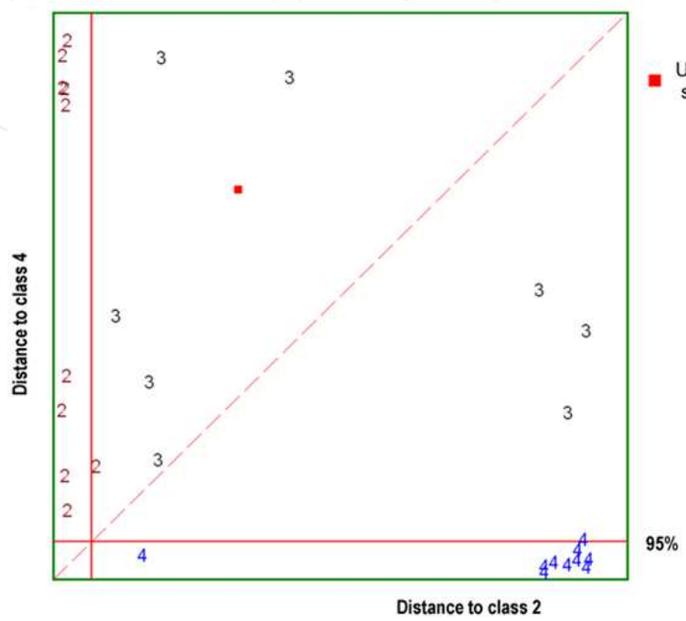

Fig. 9. Cooman's plot for the classes 2 (B) and 4 (D).

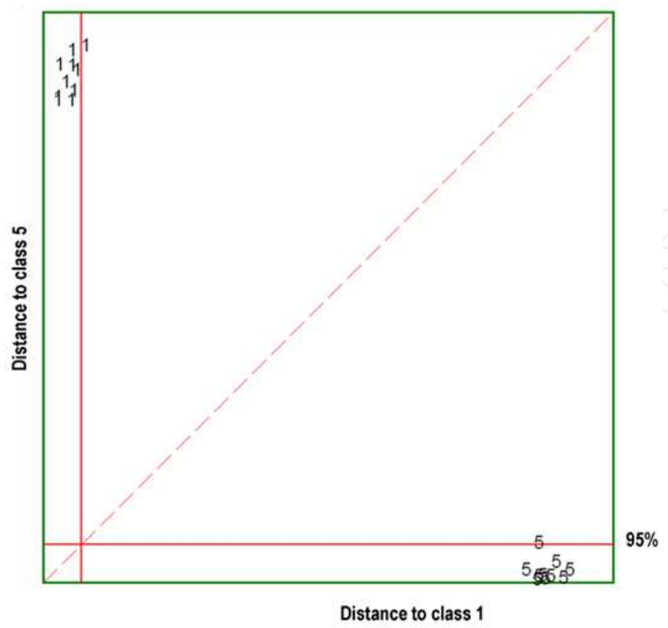

Fig. 10. Cooman's plot for the classes 1 (A) and 5 (E).

SIMCA confirms, therefore, the results obtained with PCA in so far as the unknown sample was significantly different from the A and E samples. Regarding classes B, C and D, SIMCA allows to conclude that the unknown sample, outlier for all classes, is nevertheless closer to class B (figure 9) than to the others. Finally, it can be concluded that the sample under investigation does not belong to any of the classes studied (for example, it comes from another refinery, not included in the data matrix); otherwise the sample could belong to one of the classes studied (the most probable class is number 2, followed by class 4) but the variability within each class might not have been sufficiently represented in the data matrix used.

3.2 Case 2

There are other examples that show the importance of PCA for data visualization. In forensic investigations, there is often the need to compare very similar samples. These comparisons invariably require the use of specific devices. One example was a specific request to compare three paint samples (A, B & C) in order to discover whether sample C was more similar to A or B. At first glance, all three samples appeared to be very alike. The analytical method that might have allowed to answer this question is pyrolysis, followed by GC, but the laboratory in question wasn't equipped with the necessary devices. Therefore, FT-IR (Fourier Transform Infrared Spectroscopy) analyses were carried out in transmission on 10 portions for each sample (these are both quick and relatively cheap) in order to characterize each sample variability: in other words, each sample was treated as if it were a class with 10 samples. PCA was applied to a data set relative to 30 samples and variables obtained from a data spacing of 64 cm^{-1} (with a smooth of 11, corresponding to 21.213 cm^{-1}) of FT-IR transmittances, in order to obtain a first data visualization. From the score plot of PC1 vs PC2 vs PC3, shown in figure 11, a trend can be seen in the separation between samples of classes A and B, while C samples are more frequently close to A samples than to B samples. Therefore, the similarity between C and A classes is assumed to be bigger than the one between C and B classes.

As the analytical problem required the classification of C sample to one of two classes, A or B, a discriminant analysis tool was then applied, with discriminant functions calculated only for A and B classes, while C samples were considered as unknown. The aim of this analysis was to verify in which of the two classes (A or B) samples C were more frequently classified. Discriminant analysis always classifies an unknown sample in one class (even if it is an outlier or it belongs to a different class from those implemented), because it calculates only a delimiter between the known classes. For the purpose of this case study, this tool was therefore preferable to a class modeling tool, which builds, on the other hand, a defined mathematical model for each class. Discriminant analysis was performed calculating canonical discriminant functions and using the leave-one-out method; this method is an extension of Linear Discriminant Analysis (LDA), which finds a number of variables that reflect as much as possible the difference between the groups.

The results of discriminant analysis, apart from indicating a classification ability of 100% for both classes A and B and a prediction ability of 70% and 80% respectively, show that seven C samples were classified in class A against three samples classified in class B. To conclude, the results obtained perfectly reflected those achieved in a laboratory equipped with pyrolysis devices.

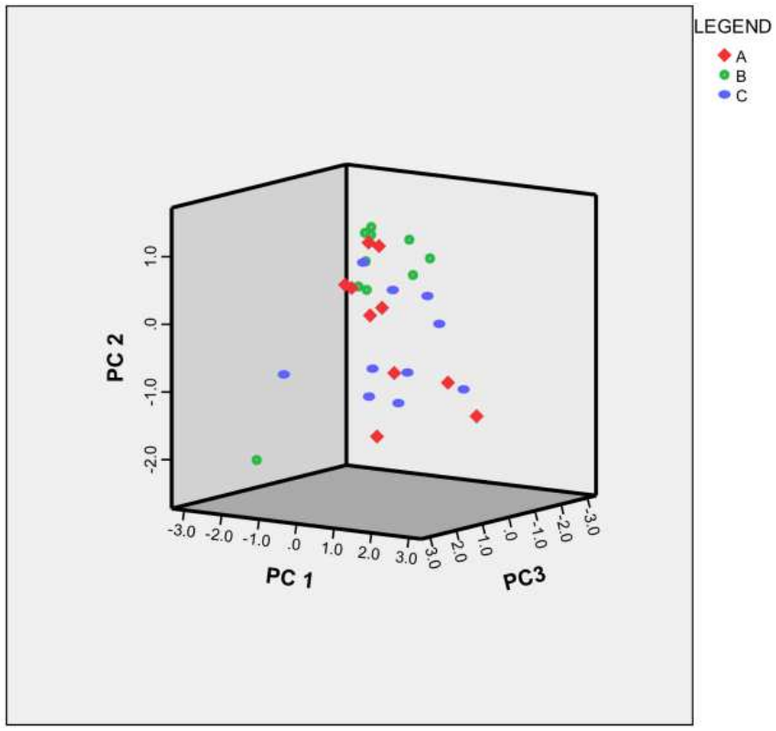

Fig. 11. Score plot of PC3 versus PC2 versus PC1 relative to FT-IR data for 30 paint samples.

A further observation about how discriminant analysis was applied in this case needs to be made. Indeed, this chemometric tool was applied to 5 principal components which account for 97,48% of the total variance, instead of the original variables. Such a procedure was adopted because the original variables were more than the samples used for building the classification rule between A and B classes (20). PCA is, therefore, imperative in classification problems where the number of variables is greater than that of the samples. In these cases, the application of discriminant analysis to the original variables would cause some overfitting problems; in other words, a sound and specific model would be obtained only for the training set used for its construction. The application of this model in real cases (like this one) would not prove very reliable. In reality, with overfitting, classification ability tends to increase while prediction ability tends to decrease. The best approach to take in these cases is to apply discriminant analysis to the PCs, by using a number of PCs (obviously less than the number of original variables) that explain a fair quantity of the variance contained in the original data. DA provides reliable results if the ratio between the number of samples and variables is more than 3.

3.3 Case 3

Another (forensic) case which involved a comparison between very similar samples was the comparison between a piece of a packing tape used on a case containing drugs with a roll of packing tape found during a house search, in order to establish whether the packing tape could have been ripped from the roll. Finding such evidence would have been of utmost importance in building a strong case against the suspect. Both exhibits, analyzed by FTIR in transmission, revealed an adhesive part of polybutylacrylate and a support of

polypropylene. Both supports and adhesive parts showed significant similarity in IR absorptions. This similarity, though necessary, was not sufficient in itself to establish whether the packing tape had been ripped from the exact same roll seized at the suspect's home. The compatibility between the two exhibits was studied through a multivariate approach, analyzing, via FTIR, 10 independent portions of the adhesive part for each exhibit. 10 portions of the adhesive part (in polybutylacrylate) of two other rolls of packing tape (not linked to the case) were also analyzed. PCA was then applied to a data set relative to 40 samples and variables obtained from a data spacing of 16 cm^{-1} (with a smooth of 11, corresponding to 21.213 cm^{-1}) of FT-IR transmittances. Six PCs were extracted, with eigenvalues greater than 1, explaining 98,15% of the total variance.

The score plot of the first three principal components is shown in figure 12, where samples taken from the seized roll are indicated as class 1, the other two rolls are indicated respectively as classes 2 and 3, while the piece of packing tape is indicated as class 4. From the score plot it can be seen that points of class 4 are fairly close to those of class 1, indicating a decent similarity between the two classes of interest. However, points of class 4 are also rather close to points of class 3, suggesting a similarity also between classes 4 and 3, while points of class 2 appear more distant, showing a lower similarity between classes 2 and 4. In this case, PCA gave a first display of data, but could not be used as definitive proof to establish the compatibility between classes 1 and 4 because class 4 appears also to be consistent with class 3.

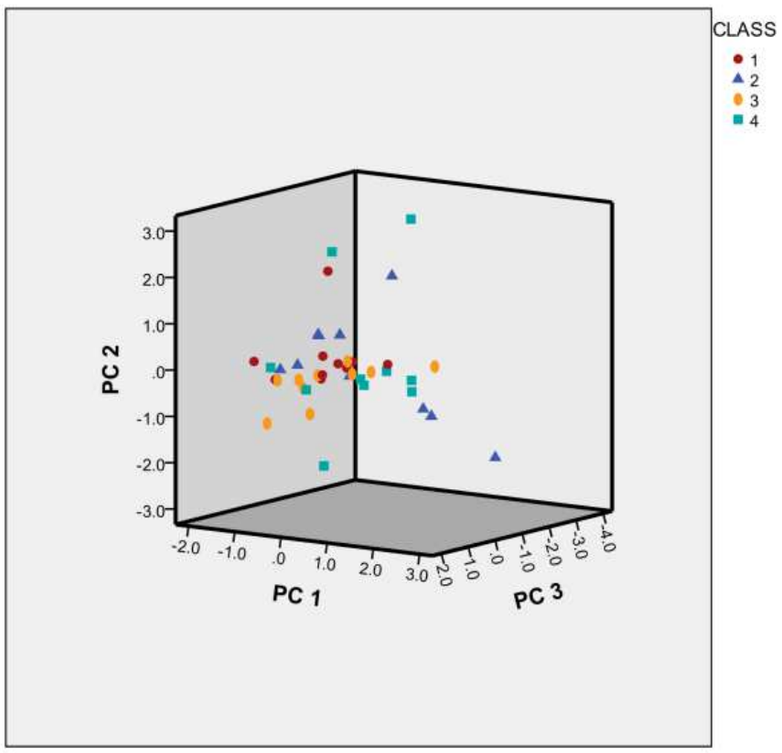

Fig. 12. Score plot of PC3 versus PC2 versus PC1 relative to FT-IR data for 40 packing tape samples.

SIMCA was then applied, considering a 95% confidence level to define class space and the unweighted augmented distance (Wold & Sjostrom, 1977). A cross validation with 10 cancellation groups was carried out and 8 components were used to build the mathematical model of each class. The boundaries were forced to include all the objects of the training set in each class, which provided a sensitivity of 100% for each class. With regard to the specificity, class 4 showed a specificity of 90% towards class 2, 80% for class 3 and 10% towards class 1. Such results can be visualized in the Cooman's plots.

For classes 1 and 4, the Cooman's plot is shown in figure 13. It can be seen that 9 samples of class 1 fall in the common area between classes 1 and 4 (the specificity of class 4 towards class 1 was in fact 10%). This kind of result indicates a significant similarity between classes 1 and 4, that is between the roll of packing tape found in the suspect's house and the piece of packing tape stuck on the case containing drugs.

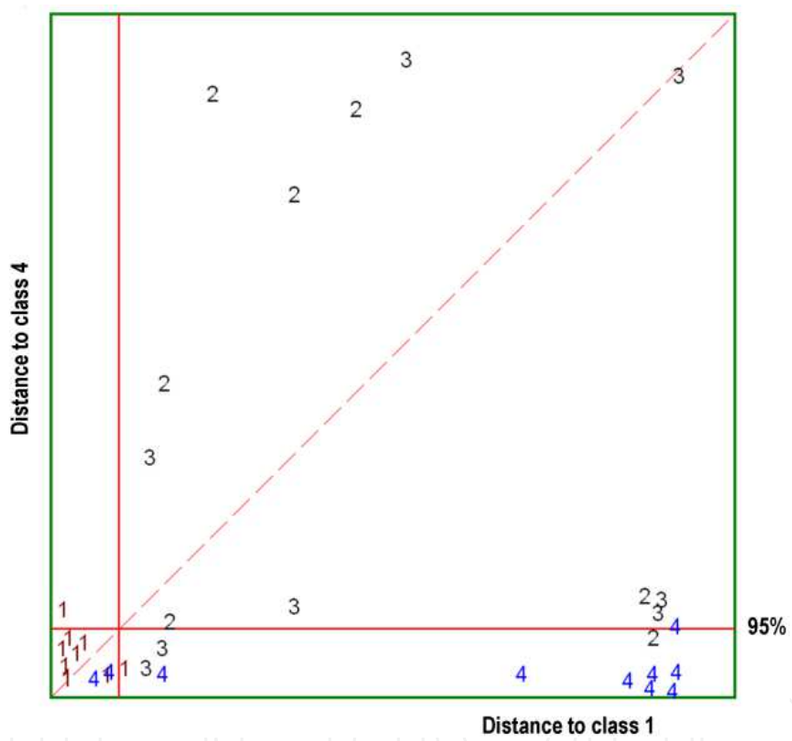

Fig. 13. Cooman's plot for the classes 1 and 4.

From the Cooman's plot relative to classes 2 and 4 (figure 14), it can be deduced that only one sample from class 2 is classified in the common area between classes 2 and 4 (specificity of class 4 towards class 2 equal to 90%), while no samples from class 4 are classified also in class 2. The similarity between classes 2 and 4 can therefore be considered insignificant.

Finally, from the Cooman's plot relative to classes 3 and 4 (figure 15), it is clearly visible that only 2 samples of class 3 fall in the overlapping area with class 4 (the specificity of class 4 towards class 3 was in fact 80%), whilst there are no samples from class 4 that fall in the overlapping area with class 3. From this last figure it can be deduced that the similarity between classes 1 and 4 is significantly higher than the similarity between classes 3 and 4.

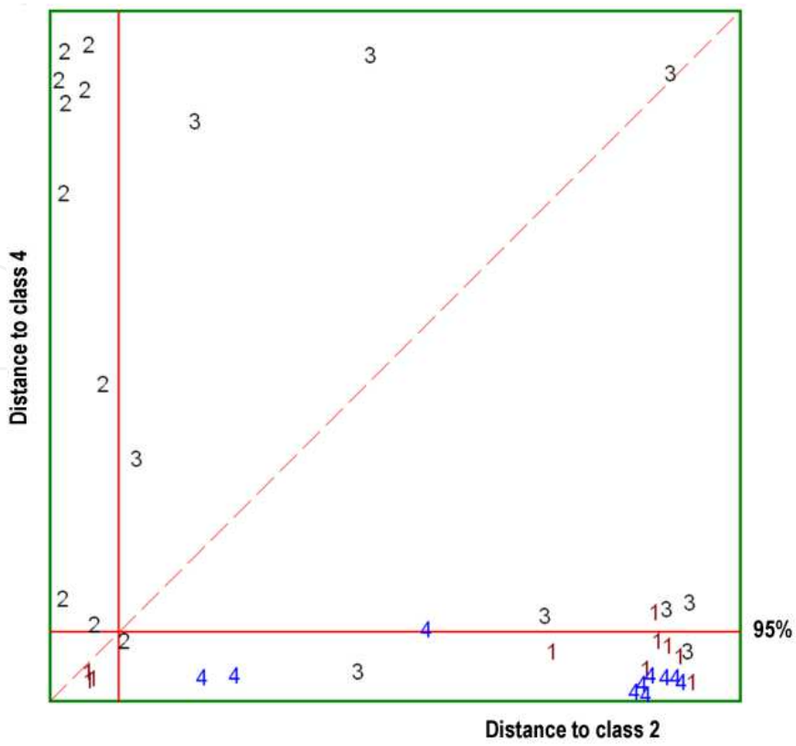

Fig. 14. Cooman's plot for the classes 2 and 4.

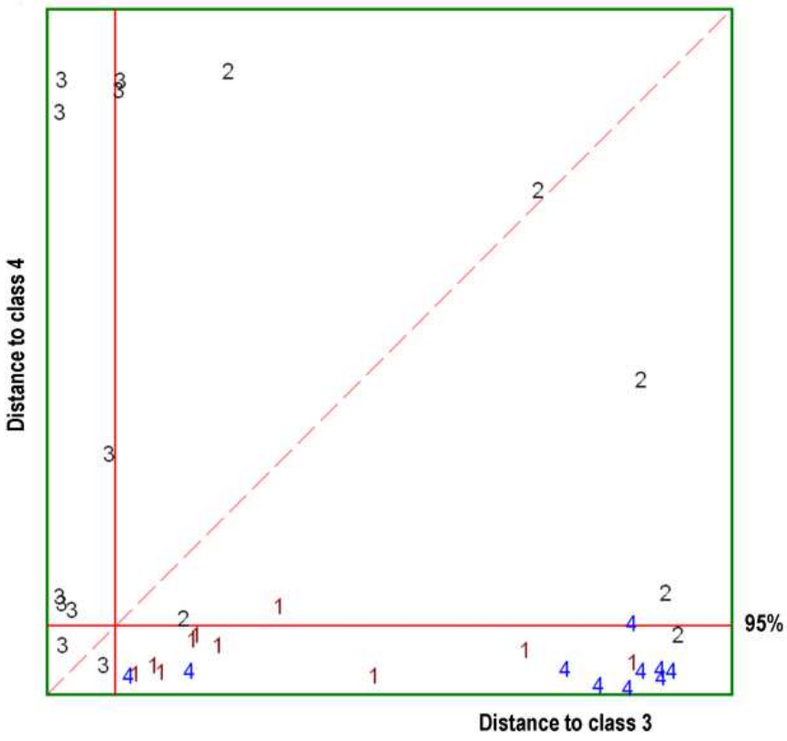

Fig. 15. Cooman's plot for the classes 3 and 4.

In conclusion, SIMCA analysis allowed the comparison between a piece of a packing tape and three rolls of packing tape that had the same chemical composition, finding the most significant similarity with the seized roll. Such a degree of similarity was measured in terms of specificity of the tape class (4) with the roll classes (1, 2 and 3): the lower the specificity is, the higher the similarity between the two classes under study is. SIMCA results are fairly consistent with PCA results, which gave a simple visualization of data. Both techniques found that class 4 had the lowest similarity with class 2. In addition, SIMCA, as a class modeling tool, gave better results than PCA with regards classes 1, 3 and 4.

4. Conclusions

This study shows the importance of PCA in traceability studies which can be carried out on different kind of matrices. As the majority of products come about from a transformation of some raw material, traceability has components deriving from both the fingerprint geographical characteristics transfer to the raw material and the production techniques developed in a specific context.

Moreover, PCA is a very useful tool for dealing with some supervised problems, due to its capability of describe objects without altering their native structure. However, it must be noted that, especially in forensics, results originating from a multivariate statistical analysis need to be presented and considered in a court of law with great care. For these kinds of results, the probabilistic approach is different from the one generally adopted for analytical results. In fact, in univariate analytical chemistry, the result of a measurement is an estimate of its true value, with its uncertainty set at a stated level of confidence. On the other hand, the use of multivariate statistical analysis in a court of law would imply a comparison between an unknown sample and a data set of known samples belonging to a certain number of classes. However, there remains the real possibility that the unknown sample might belong to yet another class, different from those of the known samples. In case 1, for example, the unknown sample might have been produced in a refinery that had not been included in the data matrix used for the comparison, or in case 3, the piece of packing tape, might not have belonged to any of the rolls analyzed. (Case 2 appears to be different, because sample C was specifically required to be classified in class A or B).

In these cases, an initial approach to the analytical problem by using PCA is fundamental because it allows the characteristics of the unknown sample to be compared with those of samples of which the origin is known. Depending on the results obtained at this step, a potential similarity between the unknown sample and samples from some specific classes may be excluded, or the class presenting the best similarity with the unknown sample might be found.

Results derived from PCA present a real picture of the situation - without any data manipulation or system forcing - and as such can form the basis for further deduction and the application of any other multivariate statistical analysis. A second step might be the application of some discriminant analysis or class modeling tool and an attempt to classify the sample in one of the classes included in the data matrix. A good result is achieved when PCA results fit those of supervised analysis. However, in a court of law these results would only become compelling alongside other strong evidence from the investigation, because, as already stated, the sample would have been compared with samples belonging to some distinct classes (and not all existing ones) and the data matrix might not adequately show the variability within each class.

5. References

Alonso-Salces, R.M. Héberger, K. Holland M.V., Moreno-Rojas J.M., C. Mariani, G. Bellan, F. Reniero, C. Guillou. (2010). Multivariate analysis of NMR fingerprint of the unsaponifiable fraction of virgin olive oils for authentication purposes. *Food Chemistry*, Vol. 118 pp. 956-965.

Brescia, M.A. Monfreda, M. Buccolieri, A. & Carrino, C. (2005). Characterisation of the geographical origin of buffalo milk and mozzarella cheese by means of analytical and spectroscopic determinations. *Food Chemistry*, Vol. 89, pp. 139-147.

Diaz, T.G. Merás, I.D. Casas, J.S. & Franco, M.F.A. (2005). Characterization of virgin olive oils according to its triglycerides and sterols composition by chemometric methods. *Food Control,* Vol. 16 pp. 339-347.

Gonzalvez, A. Armenta, S. De la Guardia, M. (2009). Trace-element composition and stable - isotope ratio for discrimination of foods with Protected Designation of Origin. *Trends in Analytical Chemistry*, Vol. 28 No.11, 2009.

Keto, R.O. & Wineman, PL. (1991). Detection of petroleum-based accelerants in fire debris by target compound gas chromatography/mass spectrometry. *Analitycal Chemistry*, Vol. 63 pp. 1964-71.

Keto, R.O. & Wineman, PL. (1994). Target-compound method for the analysis of accelerant residues in fire debris. *Analytica Chimica Acta,* Vol.288 pp.97-110.

Lennard, C.J. Tristan Rochaix, V. Margot, P. & Huber, K. (1995). A GC–MS Database of target compound chromatograms for the identification of arson accelerants. *Science & Justice*; Vol. 35 No.1 pp.19–30.

Marini, F. Magrì, A.L. Bucci, R. Balestrieri, F. & Marini, D. (2006). Class –modeling techniques in the authentication of Italian oils from Sicily with a Protected Denomination of Origin (PDO). *Chemometrics and Intelligent Laboratory Systems*, Vol. 80, pp. 140-149.

Monfreda, M. & Gregori, A. (2011). Differentiation of Unevaporated Gasoline Samples According to Their Brands, by SPME-GC-MS and Multivariate Statistical Analysis. *Journal of Forensic Sciences*, Vol. 56 (No. 2), pp. 372-380, March 2011.

Wold, S. & Sjostrom, M. (1977). SIMCA: A Method for Analyzing Chemical Data in Terms of Similarity and Analogy In: *Chemometrics, Theory and Application*, Kowalsky B.R. pp. 243-282, American Chemical Society Symposium Series No. 52 Washington.

Principal Component Analysis
Edited by Dr. Parinya Sanguansat

ISBN 978-953-51-0195-6
Hard cover, 300 pages
Publisher InTech
Published online 02, March, 2012
Published in print edition March, 2012

This book is aimed at raising awareness of researchers, scientists and engineers on the benefits of Principal Component Analysis (PCA) in data analysis. In this book, the reader will find the applications of PCA in fields such as image processing, biometric, face recognition and speech processing. It also includes the core concepts and the state-of-the-art methods in data analysis and feature extraction.

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen