Chemometrics applied to the comparison of synthetic fibers by MSP in Forensic Sciences

Hamza GHEDJATI and Abdesselam SEBAÏ

***Abstract*—In this article we discuss the advantages of utilizing chemometric approaches in the field of forensic science to resolve and compare synthetic textile fibers. It is well known that micro-spectro-photometry (MPS) is a rapid, accurate and reproducible method used in comparing colored fibers; however, these comparisons are often hampered by raw spectra showing little details or fewer points to allow more precise distinctions. The use of chemometric methods applied to spectroscopic data using multivariate statistics can provide valuable information especially when considering complex data sets. Such efforts could save time and resources in forensic examinations and help minimize variability between individual examiners. Differentiating a group of samples using chemometric analysis increases the probative value in comparing fibers by decreasing the probability of erroneous combinations. In this study, the visible spectra subtracted and normalized sixteen dyed synthetic samples with different dye compositions ranging from green, blue and red. Each of these samples was differentiated by three methods Hierarchical Ascending Classification (HAC), Principal Component Analysis (PCA) and Linear Discriminate Analysis (LDA). The results of the HAC and the PCA were consistent, showing similar spectra gathering near each other. Analysis of the LDA revealed a total classification accuracy of 100%, which corroborates the results of Hotelling test (T2).**

***Index Terms*--Textile fibers, Fiber comparison, Microspectro-photometry, chemometric.**

I. INTRODUCTION

Chemometrics is the discipline that uses mathematical and statistical methods to evaluate chemical data in order to select optimal measurement procedures. Thus, it has the ability to identify patterns and clusters of large, complex data sets with greater accuracy than visual examination alone. It can also investigate the dependence between variables, make predictions and be used to test hypotheses.

The visual comparison of marks can be quite subjective as there is no statistical basis for the findings of the examiner. Such bias poses a concern as to the reliability and relevance of scientific evidence, as raised in the criminal case of Daubert vs. Merrell Dow Pharmaceuticals. The chemometric analysis of multivariate data, often found in traces, could meet the requirements to counter effect similar issues encountered in the routinely work of forensic laboratories.

H.Ghedjati is with the National Institute of Criminalistics and Criminology of the National Gendarmerie, BP 194 Cheraga 16014 / Algiers /Algeria (e-mail: incc.cgn@mdn.dz).

A. Sebaï is with *University of Science and Technology Houari Boumediene.*

II.MATERIALS AND METHODS

A. Sampling

A total of sixteen (16) samples of synthetic fibers shown in Figure 1, were used in this study. Samples number (3), (4), (7) and (8) are reference fibers in the collection named: Forensic Fiber reference Collection "FFRC". The latter was provided by the department of trace evidence LLC (USA). Three different types of materials were analyzed, comprised of four (04) acrylic samples, eight (08) samples of polyester, and four (04) samples of nylon.

Fig.1: Photograph showing the microscopic fiber preparations used in the study, numbered 01 to 16

Samples (3) and (4) are respectively from a military sweater and a blanket, whereas samples (7) and (8) were selected from the consultation of various cases handled by the lab of textile fibers analysis of the National Institute of Criminalistics and Criminology of the National Gendarmerie (INCC/GN). The selected samples had different diameters and unknown dye concentrations.

The fibers were positioned on quartz slides using micro-tweezers. Each fiber has been mounted using the spectral

quality glycerin and quartz cover slips. The fibers were then identified and each group coded (as shown in Table I) on the basis of their observed color, preparing them for UV-Vis MSP analysis.

In our study, the spectra collection were captured on 20 test sites for each fiber, a total of eight (08) spectra were randomly selected for each example, only four replicated spectra are represented in the data sets, the others are used as data cross-validation.

B. Spectral collection and pretreatment

A Micro-Spectrophotometer (J & M), model TIDAS800, with two light sources; the ultraviolet (UV) deuterium lamp and the visible (VIS) tungsten lamp, both were used in transmitted light mode. The microscope was calibrated by Köhler illumination, and the spectrophotometer was calibrated to NIST traceable standards before each use of the instrument [2,3].

The Spectrophotometer is coupled to a Leica DM 4500P microscope and CCD detector. Spectra were obtained by averaging 100 scans over a spectral range of 400 to 850 nm. A 40 × collection lens was used to focus the light source onto an area within the diameter of the fiber samples, and the replicated spectra were taken along the same fiber to reflect the intra-fiber variation.

For discrimination by multivariate analysis, wavelength ranges of all spectra were bounded between wavelength ranges of 400 to 760 nm.

Conventionally, each set of collected data consisted of a matrix with n rows and p columns. Data were saved as variables files separated by commas (CSV).

Before using statistical evaluation tools, data were pretreated to remove systematic and random noise, using the correcting baseline and normalization. There are many methods to correct the baselines in spectroscopy, we opted to resize each spectrum assuming the lowest non- zero intensity on a spectrum lies in a region where the signal is zero [4]. This intensity is then subtracted from all other points of this spectrum. This eliminated the effects of scattered light and reduced the baseline to zero.

Then, each spectrum was normalized to the length of the unit vector by applying the auto scaling.

After this conversion, data are labeled "reduced-centric data". Principal Component Analysis (PCA) applied to the transformed data is called standard PCA.

C.Principal Component Analysis (PCA) and Hierarchical Ascending Classification (HAC):

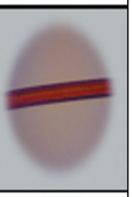
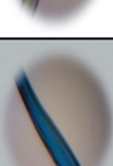
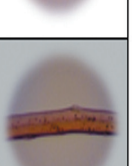
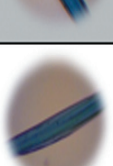
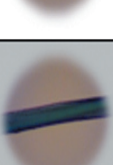
After pretreatment, all sets of spectra were subjected to PCA. The number of relevant principal components (PC) to use in the models was chosen via Scree Plot displaying the percentage of variance captured by each PC. So, we set a minimum percentage of total variance of 95%.

Then, each spectrum was normalized to the length of the unit vector by applying the auto scaling. We then applied the HAC to automatically classify the observations in each color group. Generally, the greater interclass distance between the distances of the groups, the better separation obtained.

Finally, a new sample from same each fiber is introduced at each previous classification result as to test its strength.

The proximity measure used was the Euclidean distance and the aggregation method used to group the samples was the Ward method.

Table I
List of synthetic fibers studied, as coded and **numbered.**

Code FFRC	item	Type	Image	Code FFRC	item	Type	Image
38	1	Acrylique		157	9	Polyester	
38	2	Acrylique		157	10	Polyester	
/	3	Acrylique		158	11	Polyester	
/	4	Acrylique		155	12	Polyester	
166	5	Polyester		103	13	Nylon	
166	6	Polyester		103	14	Nylon	
/	7	Polyester		103	15	Nylon	
/	8	Polyester		87	16	Nylon	

D. Cross Validation and Hotelling Test

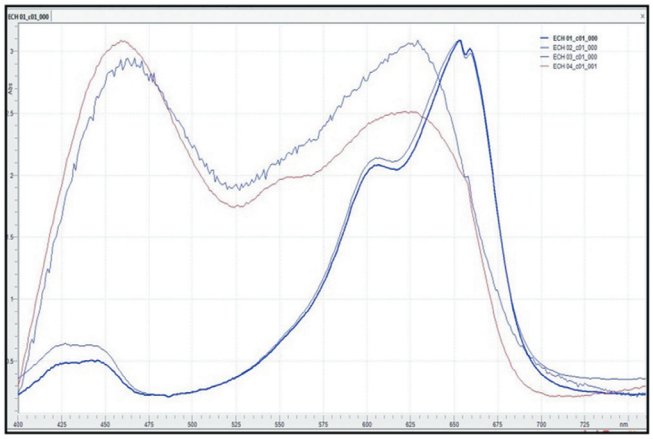
It is accepted to rely on a classification of cases when it is based on the same data used to calculate the discriminate functions, but in order to classify cases predicatively, it is necessary to collect new data to verify the validity of discriminate functions and to interpret the confusion matrix generated [5].

Once the appropriate number of CPs was selected, the linear discrimination analysis (LDA) was used by projecting the data in the space of canonical variables to predict the class membership of the cross-validation samples. The classification accuracies are obtained by assigning each spectrum to the group with the shortest Mahalanobis distance. Finally, the pairs

of samples underwent the Hotelling test (T2) for comparison. All these chemometric techniques were performed using R.

III.RESULTS AND DISCUSSION

A.Green Acrylic samples:



.Fig.2: Representative spectra of green acrylic fibres

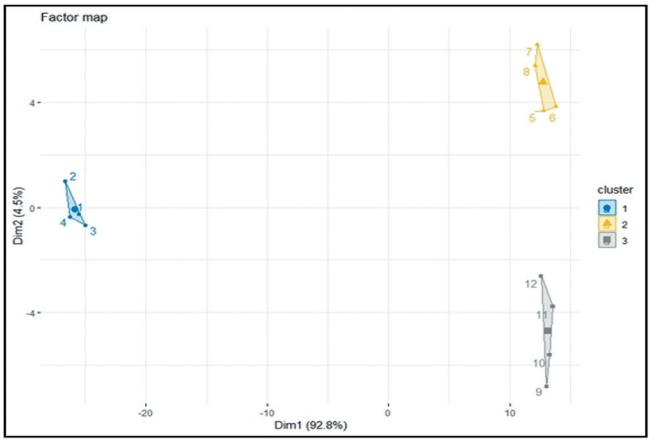


Fig.3: Classification of the four acrylic fibres in Group **A**

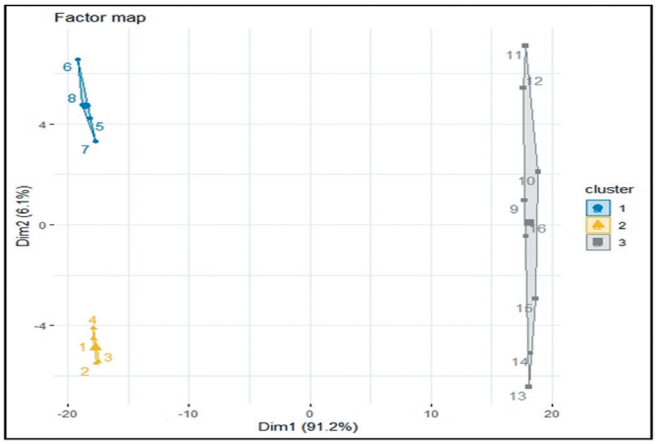


Fig.4: Classification of the four acrylic fibres in Group **B**

Table II
Confusion matrix: green acrylic fibers (Group A).

LDA Classes	2	3	4
2	4	0	0
3	0	4	0
4	0	0	4

Table III
Results of Hotelling test T2.

item	1	2	3	4
1	1	0.943	$3.021e^{-10}$	$2.359e^{-10}$
2	0.943	1	$3.017e^{-10}$	$2.365e^{-10}$
3	$3.021e^{-10}$	$3.017e^{-10}$	1	$2.113e^{-06}$
4	$2.359e^{-10}$	$2.365e^{-10}$	$2.113e^{-06}$	1

The result of the Group “A” classification of green acrylic fibers is shown in Figure 3. Three distinct widely dispersed classes have been indicated.

The classes were easily distinguished from each other. Classes 1, 2, and 3 represent the replicates of samples (2), (3), and (4), respectively.

It is noted, that the fiber (2) gave very similar and very close spectra compared to the two other fibers.

The smallest separation between two classes of fibers was obtained with the fibers (3) and (4).

For Group “B” where data from sample (1) is entered, the new HAC classification is shown in Figure 4.

LDA result (Table II) showed that the four replicas of each fiber were classified correctly with an accuracy of 100%. These results are consistent with those of the HAC, and those of the PCA.

Finally, table III lists P-values of the Hotelling test between each pair of samples.

The calculations were performed with R to 5 degrees of freedom on the table of the law of Fisher-Snedecor.

Except for samples (1) and (2), the rest of the samples gave P-values < 5%.

We conclude that we can reject the H_0 hypothesis and that the barycenters of these groups (with P-values less than 0.05) are quite distinct. These results are consistent with every previous technique.

B. Blue Samples Polyester

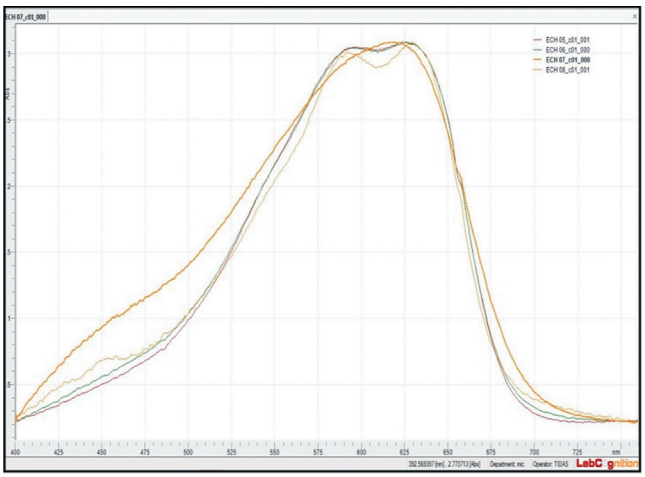


Fig.5: Representative spectra of blue polyester fibres

The result of Group “A” classification of blue polyester fibers is shown in Figure 6.

Three classes were easily distinguishable from each other, dispersed over great distances.

Classes 1, 2, and 3 represent the replicas of fibers (8), (6) and (7), respectively.

For Group "B" where data from sample (5) is entered into the dataset, the new HCA classification is shown in Figure 7.

Samples (5) and (6) were construed in the same class which is a good indicator that the reclassification was done successfully for the simple reason that the two samples (5) and (6) are the two halves of the same polyester fiber.

No noteworthy changes in the classification and distribution of other fibers.

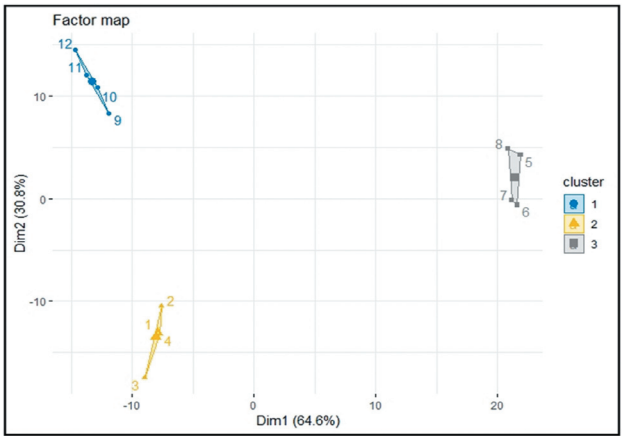


Fig.6: Classification of the three polyester fibres Group **A**

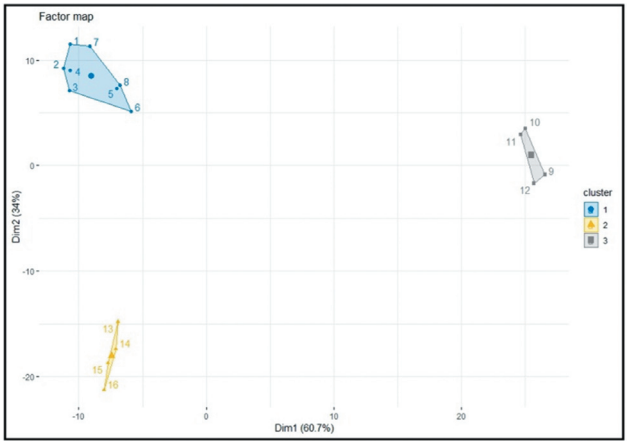


Fig.7: Classification of the four polyester fibres Group B

The LDA result (Table IV) shows that all four replicates of each fiber are classified correctly with 100% accuracy.

Table IV
Confusion matrix: blue polyester fibers (Group A).

LDA Classes	6	7	8
6	4	0	0
7	0	4	0
8	0	0	4

The Hotelling test (Table V) shows that except for samples (5) and (6) from the same fiber, we obtain P-values < 5% for the rest of the samples.

We conclude that the barycenters of the groups with P-values less than 0.05 are quite distinct, and these results are in corroboration with each result from previous techniques.

Table V
Results of the Hotelling test T^2 .

LDA Classes	5	6	7	8
5	1	0.902	$2.360e^{-08}$	$2.062e^{-10}$
6	0.902	1	$2.377e^{-08}$	$2.110e^{-10}$
7	$2.360e^{-08}$	$2.377e^{-08}$	1	$3.192e^{-05}$
8	$2.062e^{-10}$	$2.110e^{-10}$	$3.192e^{-05}$	1

C. Samples Reds

Five distinct classes scattered at different distances were indicated. Classes 1, 2, 3, 4 and 5 represent the replicates of samples (15), (10), (16) (11) and (12), respectively.

It is noted that the fiber (16) gave very similar and very close spectra to the barycenter compared to the other fibers.

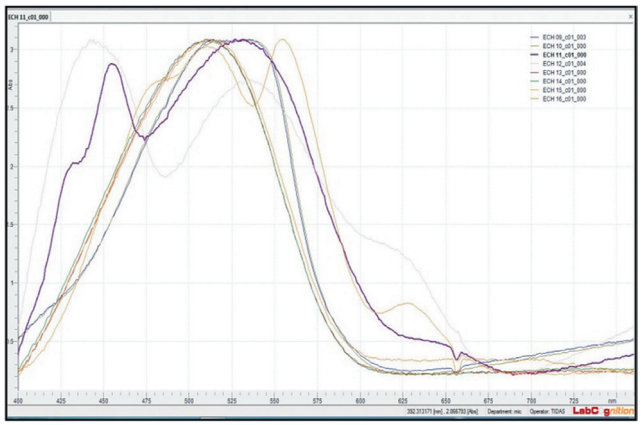


Fig.8: Representative spectra of the six red fibres

The smallest separation between two classes of fibers was obtained between fibers (10) and (15), due to the resemblance of the characteristics of their spectra as shown in Figure 9.

For Group "B" where data from sample (9) is entered into the dataset, the new HCA classification is shown in Figure 10. As expected, samples (9) and (10) have been grouped precisely in the same class (class 1), and no change is noticed for the other classes.

The LDA result (Table VI) showed that the four replicates of each fiber were classified correctly.

The Hotelling test (Table VII) shows calculations performed for 3 degrees of freedom on the Fisher-Snedecor law table.

Except for samples (9) and (10), we obtain P-values < 5% for the rest of the samples.

We therefore conclude that the barycenters of these groups are quite distinct, and these results are consistent with previous techniques.

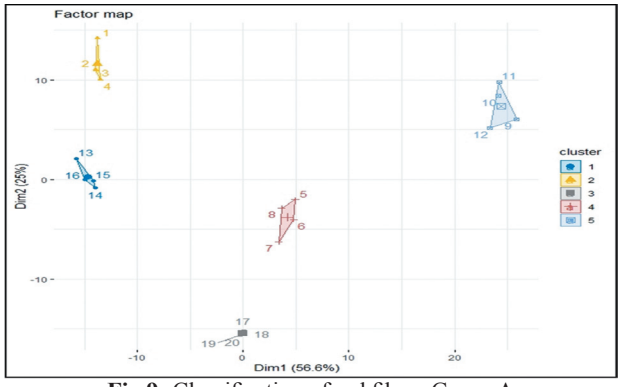


Fig.9: Classification of red fibres Group A

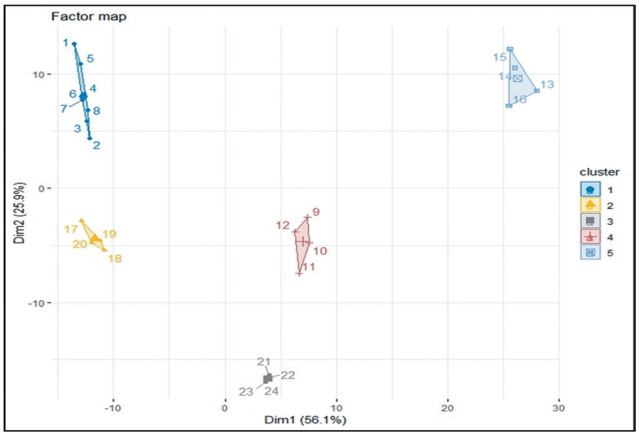


Fig.10: Classification of the six red fibres Group B

Table VI
Confusion matrix (Group A)

LDA Classes	10	11	12	15	16
10	4	0	0	0	0
11	0	4	0	0	0
12	0	0	4	0	0
15	0	0	0	4	0
16	0	0	0	0	4

Table VII
Results of the Hotelling test T^2

item	9	10	11	12	15	16
9	1	0.714	$2.512e^{-07}$	$2.698e^{-06}$	$2.756e^{-06}$	$1.359e^{-05}$
10	0.714	1	$1.775e^{-07}$	$2.745e^{-06}$	$2.534e^{-06}$	$1.289e^{-05}$
11	$2.512e^{-07}$	$1.775e^{-07}$	1	$1.883e^{-05}$	$3.927e^{-07}$	$3.648e^{-06}$
12	$2.698e^{-06}$	$2.745e^{-06}$	$1.883e^{-05}$	1	$2.967e^{-06}$	$1.754e^{-09}$
15	$2.756e^{-06}$	$2.534e^{-06}$	$3.927e^{-07}$	$2.967e^{-06}$	1	$3.443e^{-09}$
16	$1.359e^{-05}$	$1.289e^{-05}$	$3.648e^{-06}$	$1.754e^{-09}$	$3.443e^{-09}$	1

IV. CONCLUSION

The chemometric treatment of visible spectra of fibers with different dye load has proven to be a reliable and efficient way to discriminate fibers. Higher classification clarification was observed when the fibers were treated using the HAC, PCA and LDA, which led to several conclusions.

The PCA has provided an initial understanding of the spectra of clusters. The spectra of clusters replicated within each class are grouped, which can indicate an additional structure (subclasses) in the data set.

The PCA results were consistent with those of the HAC. The results of ADL created a model that provides error rate to all classified spectra.

Overall, the performances of the model provided were excellent, with an overall classification accuracy of 100% using both cross-validation and the Hotelling test.

We recommend, the participation in inter laboratory studies to verify the reproducibility of the procedure with chemometric MSP method. Additionally, it should be noted that a practical software application with many pre-processing options, that can significantly improve the structure of the spectra, is strongly recommended to determine if calculations augur.

We also suggest that further studies should include a set of data containing fibers with the same diameter, for more representative conclusions, as well as datasets comprising damaged fibers, altered or washed in order to evaluate the effects of these variables on the chemometric analysis. Furthermore, data sets from other characterizations such as FTIR spectroscopy, Raman and PLM, could be used with chemometric procedure to classify and compare different types and subtypes of fiber.

REFERENCES

[1] National resource Council, Strengthening Forensic Science in the United States: A Path Forward. National Academies Press: Washington, dC, 2009.

[2] Standard Guide for Microscopical, Examination of Textile Fibers, ASTM E2228– 18. https://www.nist.gov/publications/multiple-fiber-technique-single-fiber- fragmentation-test.

[3] Martin Y. Sohn et al, Koehler Illumination for High-resolution Optical Metrology, March 01, 2006.

[4] S. Golotvin, A. Williams, Improved Baseline Correction of FT NMR Spectra. Advanced Chemistry development, NMR Newsletter Advanced Chemistry development, 1999.

[5] StatSoft, Inc. Techniques Exploratoires, Multivariées : Analyse discriminante. https://www .stats oft.fr /concepts -statistiques/analyse-discriminante/analyse- discriminante.php, 2019.