Hierarchical classification and matching of mid-infrared spectra of paint samples for forensic applications.

Raffaele Vitale, Giulia Spinaci, F. Marini, Philippe Marion, Martine Delcroix, Arnaud Vieillard, François Coudon, Olivier Devos, Cyril Ruckebusch

Version of Record: https://www.sciencedirect.com/science/article/pii/S0039914022001564
Manuscript_ef4574464d5c59a9d5acf8605a286d97

Hierarchical classification and matching of mid-infrared spectra of paint samples for forensic applications

Raffaele Vitale[a,*], Giulia Spinaci[a,b], Federico Marini[b], Philippe Marion[c], Martine Delcroix[c], Arnaud Vieillard[c], François Coudon[c], Olivier Devos[a], Cyril Ruckebusch[a]

[a] *Univ. Lille, CNRS, LASIRE (UMR 8516), Laboratoire Avancé de Spectroscopie pour les Interactions, la Réactivité et l'Environnement, F-59000 Lille, France*
[b] *Department of Chemistry, Università degli Studi di Roma "La Sapienza", Piazzale Aldo Moro 5, 00185 Roma, Italy*
[c] *Laboratoire de Police Scientifique, Service National de Police Scientifique, 7 Boulevard Vauban CS80007, 59041 Lille Cedex, France*

Abstract

A novel fast and automatic methodology for the hierarchical classification and similarity matching of mid-infrared spectra of paint samples based on the principles of Soft Independent Modelling of Class Analogy (SIMCA) and on the definition and properties of the Mahalanobis distance is here proposed. This approach was tested in a so-called market study (*i.e.*, targeting products largely accessible to the general public and conceived for a considerably wide range of usages) conducted across the surroundings of the city of Lille, in France, and has permitted not only to successfully achieve the chemical characterisation of most of the analysed samples but also to discover specific commonality patterns among specimens sharing the same chemical features.

Keywords: classification, matching, hierarchical methods, Soft Independent Modelling of Class Analogy (SIMCA), Hotelling's T^2, paints

1. Introduction

Crime scene investigation often involves the assessment and analysis of paint traces (*e.g.*, graffiti), which are generally carried out by means of spectroscopic and/or microscopic techniques [1, 2]. Among these, Fourier Transform Infrared (FTIR) spectroscopy has proven to be one of the most reliable and performant approaches, especially for paint discrimination and characterisation [3–6]. Paints, in fact, can be regarded as mixtures of two primary constituents: i) an organic polymeric resin enabling its binding to the target surface and ii) organic/inorganic pigments dispersed in this polymeric resin (for instance, titanium dioxide, talc, kaolin or calcium carbonate), regularly at sufficient concentration levels to allow FTIR determination. In such scenarios, two of the basic tasks forensic agents and examiners usually need to address are i) the identification

*Corresponding author:
Telephone number: +33320434748
Email address: raffaele.vitale@univ-lille.fr (Raffaele Vitale)

Submitted to Talanta *March 3, 2022*

of the chemical composition of the unknown samples collected on-field through the interpretation of the spectral profiles registered and ii) the comparison between these spectroscopic profiles and those of databased specimens for similarity recognition. The former can provide essential information on the primary source of a paint trace: commonly, in fact, paints of different nature have diverse end-uses and the evaluation of their formulation might aid the identification of the items or supports from which they come (motor vehicles, maintenance tools, *etc.*). The latter, instead, may permit the direct association of the recovered evidence with objects available on the market or found during the inspection of other crime scenes [7].
In order to get insights into the chemical composition of paint traces by FTIR analysis, practitioners typically resort to *ad hoc* flow-diagrams (like the so-called "automotive paint binder infrared classification flow-chart") guiding the users throughout the visual interpretation of the individual peaks detected [8]. On the other hand, database searching/matching is ordinarily performed either manually through pairwise comparisons of spectra or by commercial software suites (like KnowItAll Spectroscopy Edition - John Wiley & Sons, Inc., Hoboken, United States of America - and OMNIC Spectra Software - Thermo Fisher Scientific, Inc., Waltham, United States of America) that rely on proximity measures directly estimated from the spectroscopic data at hand. Although both these strategies constitute the state of the art in paint forensics, they suffer from severe drawbacks which hamper their utilisation particularly when large amounts of spectral profiles are to be handled and processed. Indeed, being mostly based on visual/manual procedures, such methodologies are extremely time consuming and prone to errors principally induced by the mood/fatigue of the operators and/or by the complexity and the pronounced entanglement of the instrumental response recorded [9]. Furthermore, even when computer algorithms are employed for this purpose, it is well-known that chance matches are frequent and the risk of false positives/negatives remains high, mainly when sample differentiation strictly depends on minor chemical components [10]. For overcoming these limitations while guaranteeing a rapid, automatic and robust characterisation of paint samples, multivariate (chemometric) methods can be alternatively applied. Recent studies have already demonstrated that they can represent feasible solutions for the discrimination of synthetic resins [11] and proteinaceous binders [12] contained in varnishes for art purposes, architectural finishes (household paints [13]), automotive clear coats [14], and spray paint specimens of distinct colours [9] as well as for developing spectral pre-filters to facilitate library or database comparisons [15]. For this reason, the present article describes a novel hierarchical approach based on Soft Independent Modelling of Class Analogy (SIMCA [16, 17]) and on the definition and properties of the Mahalanobis distance [18, 19] to achieve the characterisation of commercial spray paint samples, largely accessible to the general public and conceived for a considerably wide range of usages - what is also known as a *market study* [9, 20–25] - and, consecutively, discover similarity patterns among specimens sharing the same chemical features. Tackling these two objectives sequentially is key in contingencies like this: seeking such similarity patterns only for paints exhibiting the same composition, in fact, can significantly reduce the impact of the aforementioned chance or false matches. To the best of the authors' knowledge, works targeting these two goals in a similar fashion and by the combination of FTIR spectroscopy and multivariate statistics have never been reported in literature.

2. Materials and methods

The proposed workflow for paint identification and matching encompasses two sequential steps, whose detailed description will be given in the following subsections:

1. the SIMCA-based classification of paint specimens according to their FTIR spectral signature;
2. the recognition of the databased samples most spectroscopically similar to such specimens. Notice that this assessment is conducted through the estimation of a pairwise Mahalanobis distance metric representative of how resemblant the individual spectral response of these latter is to the spectral profiles of paints with the same chemical composition.

2.1. Paint sample classification by Soft Independent Modelling of Class Analogy (SIMCA)

Let $\mathbf{X}$ be a N samples $\times$ J variables (wavenumber channels, in this case) dataset, made up of Z blocks, $\mathbf{X}_z$ ($N_z \times J$), each one containing spectral profiles of (paint) specimens of the same unique category (*i.e.*, a particular chemical composition). SIMCA separately decomposes every (centred) $\mathbf{X}_z$ array according to a Principal Component Analysis (PCA [26, 27]) model of appropriate dimensionality (say A_z) as:

$$\mathbf{X}_z = \mathbf{T}_z \mathbf{P}_z^{\mathrm{T}} + \mathbf{E}_z \tag{1}$$

where $\mathbf{T}_z$ ($N_z \times A_z$), $\mathbf{P}_z$ ($J \times A_z$) and $\mathbf{E}_z$ ($N_z \times J$) denote the scores, loadings and residuals matrices resulting from the factorisation of $\mathbf{X}_z$. After having defined the individual class subspaces as in Equation 1, the degree of *outlyingness* of any new generic measurement observation, $\mathbf{x}_{\mathrm{new}}^{\mathrm{T}}$ ($1 \times J$), with respect to them can be assessed in terms of the so-called reduced distance [28] which is calculated as:

$$d_{\mathrm{new},z} = \sqrt{\left(\frac{T^2_{\mathrm{new},z}}{T^2_{\mathrm{lim},z}}\right)^2 + \left(\frac{Q_{\mathrm{new},z}}{Q_{\mathrm{lim},z}}\right)^2} \tag{2}$$

with $T^2_{\mathrm{new},z}$ reflecting the (Mahalanobis) distance between the origin of the z-th model hyperplane and the projection of $\mathbf{x}_{\mathrm{new}}^{\mathrm{T}}$ onto it, and $Q_{\mathrm{new},z}$ reflecting the perpendicular (orthogonal) distance between $\mathbf{x}_{\mathrm{new}}^{\mathrm{T}}$ and the z-th model hyperplane. $T^2_{\mathrm{lim},z}$ and $Q_{\mathrm{lim},z}$ connote empirical thresholds for the former and the latter statistical indices, respectively, usually corresponding to a significance level of 95% and estimated based on the elements in $\mathbf{X}_z$. In this article, the observation $\mathbf{x}_{\mathrm{new}}^{\mathrm{T}}$ is considered an outlier for the model of the z-th class and, thus, not recognised as its member if $d_{\mathrm{new},z}$ is found to be larger than $\sqrt{2}$. Conversely, if $d_{\mathrm{new},z} \leq \sqrt{2}$, the corresponding sample is assigned to the z-th category [29–31].

In SIMCA, the classification performance is commonly evaluated according to the following figures of merit:

$$\mathrm{sensitivity}_z = \frac{\mathrm{TP}_z}{\mathrm{TP}_z + \mathrm{FN}_z} \times 100 \qquad \forall z \in [1, \ldots, 3] \tag{3}$$

$$\mathrm{specificity}_z = \frac{\mathrm{TN}_z}{\mathrm{TN}_z + \mathrm{FP}_z} \times 100 \qquad \forall z \in [1, \ldots, 3] \tag{4}$$

$$\mathrm{efficiency}_z = \sqrt{\mathrm{sensitivity}_z \times \mathrm{specificity}_z} \qquad \forall z \in [1, \ldots, 3] \tag{5}$$

with TP_z, TN_z, FP_z and FN_z standing for the amount of true positives (objects correctly identified as belonging to the z-th category), true negatives (objects correctly identified as not belonging to the z-th category), false positives (objects mistakenly identified as belonging to the z-th category) and false negatives (objects mistakenly identified as not belonging to the z-th category) returned by the classification procedure. The optimisation of the complexity (number of principal components per class) of SIMCA class models, instead, can be carried out in various ways. Here, a resampling strategy encompassing 300 repetitions of random subset cross-validation and aiming at maximising the resulting efficiency values was resorted to [32–35].
Contrarily to standard discriminant approaches - such as Partial Least Squares Discriminant Analysis (PLSDA [36, 37]) - that strictly partition the multivariate space of the registered variables into as many subregions as the number of categories of objects at hand, SIMCA (as well as other *class modelling* techniques) independently defines a multivariate frontier for each individual category under study, delimiting a specific region of the aforementioned multivariate space where specimens belonging to it are more likely to be found (see also Figure 1 for a schematic representation of this difference) [38]. In other words, discriminant strategies always assign each one of the data

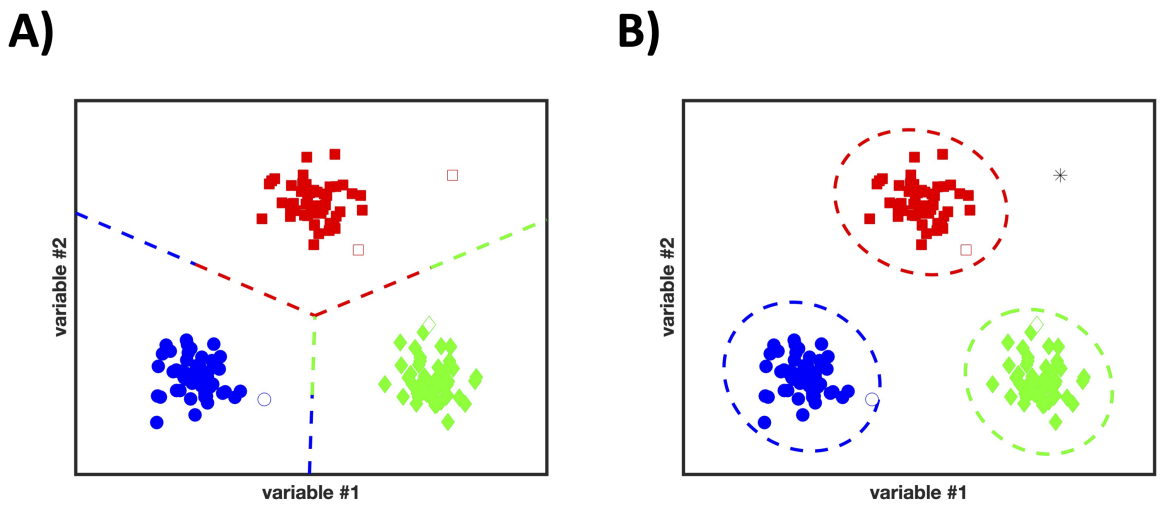

Figure 1 – Schematic representation of the basic principle of A) a discriminant and B) a class-modelling techniques. The former strictly partitions the multivariate space of the registered variables into as many subregions as the number of classes of objects at hand and always assigns each one of such objects to a certain category. The latter independently defines a multivariate frontier for each individual class under study, delimiting a specific region where specimens belonging to it are more likely to be found. In this specific case, the observation lying on the upper right area of the two plots would be recognised as member of the red square category by a discriminant approach, but would be rejected by all the three independent class models one could possibly construct. Notice that empty symbols (as well as the black star) denote hypothetical test samples, *i.e.* samples not taken into account when defining the classification boundaries/rules.

items to a certain class (the one and only one within whose boundaries the point corresponding to its measurement vector falls), while, in SIMCA, samples can be recognised as members of none, one or multiple modelled categories, which renders the application of this methodology perfectly

suited when such categories are expected to constitute only a reduced part of those that could be potentially encountered and explored - a rather common case in paint forensics. Moreover, once samples are identified as outliers by all the class models available at a given moment, their respective spectral profiles can be further examined and investigated in order to get insights into their chemical composition. If a new particular chemical composition gets sufficiently represented, given the total independence among class models, an additional one can be at any time built without necessarily having to modify those constructed in the initial training step. In principle, this cannot be achieved through classical discriminant tools: first of all, as outlined previously, they cannot easily spot outlying observations unless they are utilised in a so-called *soft* classification framework; second, if new classes of objects become available, discriminant models always need to be entirely recalibrated, which, in theory, also applies to these more recently developed soft classification methods [39, 40].

2.2. Spectral similarity assessment based on pairwise Mahalanobis distance estimation

Once a sample (say again the observation $\mathbf{x}_{\text{new}}^{\text{T}}$) is successfully assigned to one or more of the classes under study, its similarity with other specimens belonging to those (and only those) categories can be assessed in terms of their pairwise Mahalanobis distance [41] as:

$$d_{\text{M},z}\left(\mathbf{x}_{\text{new}}^{\text{T}}, \mathbf{x}_{n_z}^{\text{T}}\right) = \sqrt{\left(\mathbf{t}_{\text{new},z}^{\text{T}} - \mathbf{t}_{n_z,z}^{\text{T}}\right)\mathbf{S}_z^{-1}\left(\mathbf{t}_{\text{new},z} - \mathbf{t}_{n_z,z}\right)} \qquad \forall n_z \in [1, \ldots, N_z] \tag{6}$$

where $\mathbf{t}_{\text{new},z}^{\text{T}}$ and $\mathbf{t}_{n_z,z}^{\text{T}}$ (both of dimensions $1 \times A_z$) are the row vectors containing the scores resulting from the projection of $\mathbf{x}_{\text{new}}^{\text{T}}$ and of each one of the N_z observations of $\mathbf{X}_z$ ($\mathbf{x}_{n_z}^{\text{T}}$ - $1 \times J$) onto the z-th class model subspace, while $\mathbf{S}_z = \frac{\mathbf{T}_z^{\text{T}}\mathbf{T}_z}{N_z - 1}$ connotes the $\mathbf{T}_z$ scores covariance matrix. Clearly, the higher $d_{\text{M},z}\left(\mathbf{x}_{\text{new}}^{\text{T}}, \mathbf{x}_{n_z}^{\text{T}}\right)$, the more dissimilar $\mathbf{x}_{\text{new}}^{\text{T}}$ with respect to $\mathbf{x}_{n_z}^{\text{T}}$.
Three aspects are worth to be mentioned here. First of all, this procedure neglects the distance of the compared samples from the z-th principal component hyperplane since such a distance is assumed to exhibit *in-control* values (*i.e.*, lower than its respective threshold) if these samples are not rejected by the corresponding class model. Secondly, it reduces the chance of false matches as the pairwise Mahalanobis distance estimation is uniquely carried out for specimens sharing the same spectroscopic features. Finally, it is not limited to the set of N_z measurements exploited for training purposes, but can be directly extended to any object having been previously recognised as member of the z-th category.

2.3. Dataset

The dataset analysed to test the proposed methodology consists of 246 spectra of spray paints (see Figure 2A) recorded in transmission mode by a Nicolet Continuμm infrared microscope coupled to a NEXUS 670 FTIR platform for spectroscopic measurements (Thermo Fisher Scientific, Inc., Waltham, United States of America) within the range 4000-650 cm^{-1} and at a resolution of 4 cm^{-1} (128 scans per sample). All specimens proceed from as many commercial products purchased - for the purpose of the study itself, *i.e.*, to estimate the variety of their composition - in a set of shops of the city of Lille, in France, and its surroundings, and manufactured by over 30 multinational companies active in the same area. Brands/family brands and colours (see the

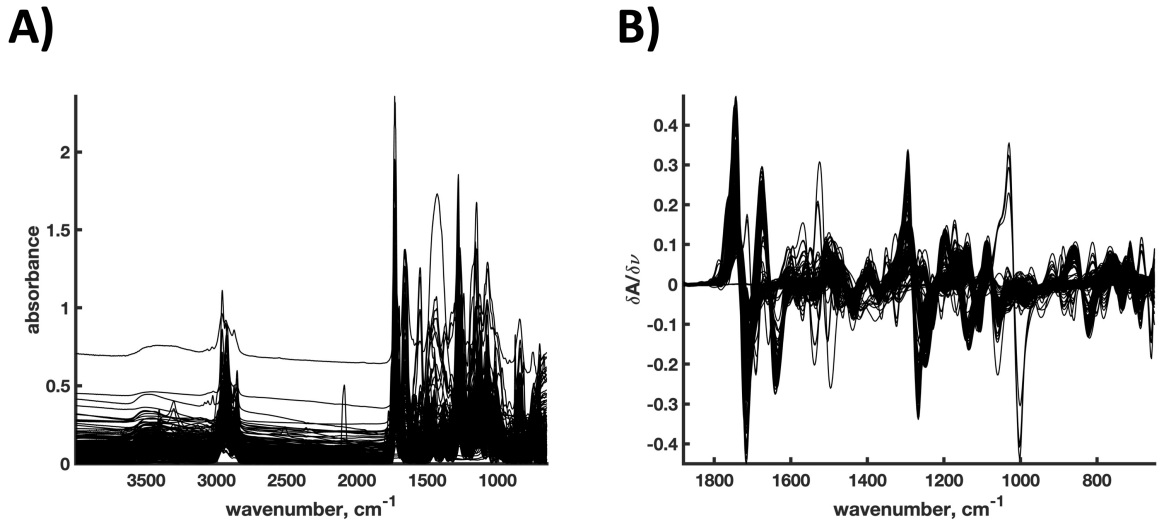

Figure 2 – A) Raw and B) reduced-preprocessed FTIR spectral data analysed in the present study.

supplementary material for a comprehensive list) were selected so as to span most of the spray paints available on the market in the aforementioned region and, more specifically, those most commonly sold by local distributors (as per the information received directly from sales representatives). Sample preparation encompassed the following steps: every spray can was initially shaken for 3 minutes in order to homogenise its content. The paint was then vaporised for 15 seconds onto two glass slides (previously cleansed with anhydrous ethanol and placed at a distance of 30 cm from the actuator) which were let to dry horizontally for 72 hours. A dry paint sample was afterwards collected by means of a scalpel and a hand lens, flattened by using a 10-ton hydraulic press and deposited onto a potassium bromide (KBr) spectroscopic window.

105 of the investigated paint specimens were chemically characterised (by validating preliminary information supplied by the providers through the direct interpretation of the registered profiles) and identified as belonging to three different categories: 35 orthophtalic alkyd-based paints (chemical composition #1), 52 orthophtalic alkyd-, nitrocellulose- and styrene-based paints (chemical composition #2) and 18 poly-methyl-methacrylate-based paints (chemical composition #3). No details on the constituents of the remaining 141 samples were instead available. The spectral data related to the first 105 specimens were split into a training and a test set (constituted by approximately 70% and 30% of the total number of observations available, respectively - see also Table 1) by means of the Duplex algorithm executed category-wise [42]. The former was resorted to for the calibration of three one-class SIMCA models (for chemical composition #1, #2 and #3), while the latter was exploited in order to address an external validation of their performance in terms of the figures of merit defined in Section 2.1.

For the sake of data processing and modelling, only the wavenumber interval between 1900 and 650 cm^{-1} (including the so-called *fingerprint* range that contains selective information for individual chemical constituents and, therefore, is usually exploited for the comprehensive characterisation of complex and heterogeneous chemical systems) was taken into account. Standard Normal

Table 1 – Composition of the training and test set analysed in the present study.

	training set	test set
composition #1 class model	25 samples	10 samples
composition #2 class model	37 samples	15 samples
composition #3 class model	13 samples	5 samples

Variate [43] combined with Savitzky-Golay derivation [44] (spectral window size: 19 points; polynomial function degree: second; derivative order: first) were utilised for spectral preprocessing (see Figure 2B).

3. Results and discussion

The outcomes obtained from the application of SIMCA to the spectral data collected for the samples of known composition are summarised in Table 2 and Figure 3. They clearly highlight the accomplishment of a satisfactory differentiation among the paints exhibiting distinct chemical characteristics. Moreover, the so-called Coomans plots [45] in Figure 4 (combined bivariate representations of the graphs in Figure 3 which enable the visualisation of the degree of confusion between every possible couple of classes) indicate that no sample was assigned to multiple categories at the same time, *i.e.*, no symbol falls in their bottom-left region, delimited by the reduced distance thresholds estimated for the two compared class models. Thereafter, for all the specimens recognised as member of at least one of the categories at hand, the assessment outlined in Section

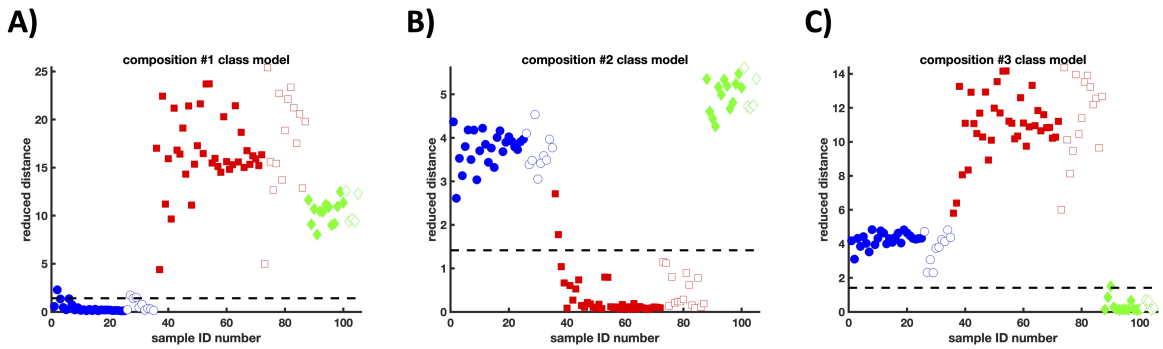

Figure 3 – Reduced distance plots for the training and test paint samples of known chemical composition. A) Class #1 model; B) class #2 model; C) class #3 model. Legend: filled blue dots - training specimens of chemical composition #1; empty blue dots - test specimens of chemical composition #1; filled red squares - training specimens of chemical composition #2; empty red squares - test specimens of chemical composition #2; filled green diamonds - training specimens of chemical composition #3; empty green diamonds - test specimens of chemical composition #3. The black dotted line denotes the classification threshold set at $\sqrt{2}$.

Table 2 – Classification sensitivity, specificity and efficiency yielded by the three SIMCA class models in training, cross-validation (CV) and external validation (test), respectively.

	number of PCs	sensitivity (training)	specificity (training)	efficiency (training)	sensitivity (CV)	specificity (CV)	efficiency (CV)	sensitivity (test)	specificity (test)	efficiency (test)
composition #1 class model	1	96.0%	100.0%	98.0%	87.8%	100.0%	93.7%	80.0%	100.0%	89.4%
composition #2 class model	1	94.6%	100.0%	97.3%	94.1%	100.0%	97.0%	100.0%	100.0%	100.0%
composition #3 class model	1	92.3%	100.0%	96.1%	84.6%	100.0%	92.0%	100.0%	100.0%	100.0%

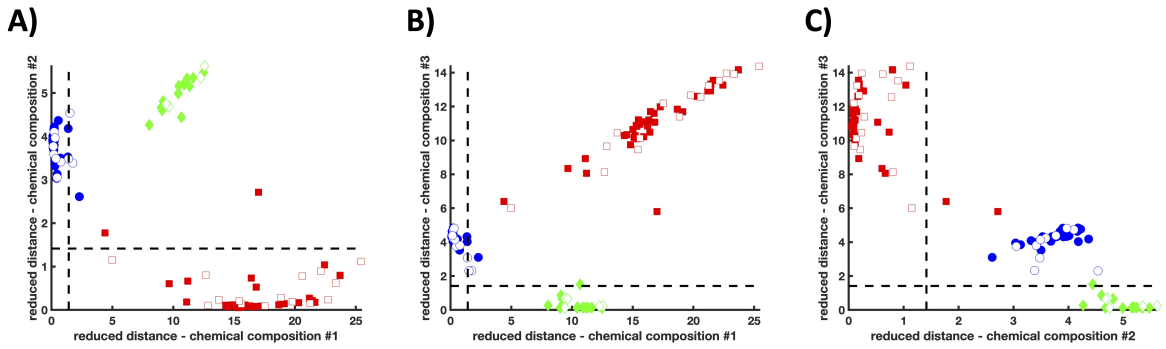

Figure 4 – Coomans plots for A) class model #1 *vs* class model #2, B) class model #1 *vs* class model #3 and C) class model #2 *vs* class model #3. Legend: filled blue dots - training specimens of chemical composition #1; empty blue dots - test specimens of chemical composition #1; filled red squares - training specimens of chemical composition #2; empty red squares - test specimens of chemical composition #2; filled green diamonds - training specimens of chemical composition #3; empty green diamonds - test specimens of chemical composition #3. The black dotted lines denote the classification thresholds set at $\sqrt{2}$ for all categories.

2.2 was conducted[i]. Figures 5A, 5B and 5C contain the waterfall representation of the derivative spectra of three test samples assigned to the first, the second and the third class, respectively, and of those of their 10 most similar training objects, coloured according to their corresponding pairwise Mahalanobis distance values. Specific commonality patterns can be easily discerned [46–48]:

1. in Figure 5A, the signals at around 1450, 1270, 1130, 1070, 740 and 700 cm^{-1}, generally attributed to the vibration modes of the orthophtalic alkyd molecular groups;
2. in Figure 5B, together with those listed before, the signals at around 1650, 1280 and 840 cm^{-1}, generally attributed to the vibration modes of the nitrocellulose molecular groups, and the signals at around 1490, 1450 and 760 cm^{-1}, generally attributed to the vibration modes of the styrene molecular groups;
3. in Figure 5C, the signals at around 1450, 1380, 1270, 1240, 1150 and 970 cm^{-1}, generally attributed to the vibration modes of the poly-methyl-methacrylate molecular groups.

It has to be mentioned here that the developed methodology is capable of providing investigators and forensic scientists with insights about the spectroscopic similarity between databased samples (*e.g.* historical bodies of evidence) and newly collected ones. This would directly support them in significantly narrowing the amount of target objects with respect to which additional comparisons (in terms of other attributes like colour or production site) might be addressed. For example, in this particular case, the two training paints exhibiting the highest commonality with the test ones to which Figures 5A and 5B refer were found to share with them the same colour shade (brilliant blue and night blue, respectively). The first can even be traced back to the same distribution chain.

[i]Notice that, as for all the class models the estimated number of components equals 1, Equation 6 reduces to $d_{\mathrm{M},z}\left(\mathbf{x}_{\mathrm{new}}^{\mathrm{T}}, \mathbf{x}_{n_z}^{\mathrm{T}}\right) = \sqrt{\frac{\left(t_{\mathrm{new},z} - t_{n_z,z}\right)^2}{s_z^2}}$, with $t_{\mathrm{new},z}$ and $t_{n_z,z}$ being two scalars and s_z^2 denoting the variance of the PCA scores of the training samples.

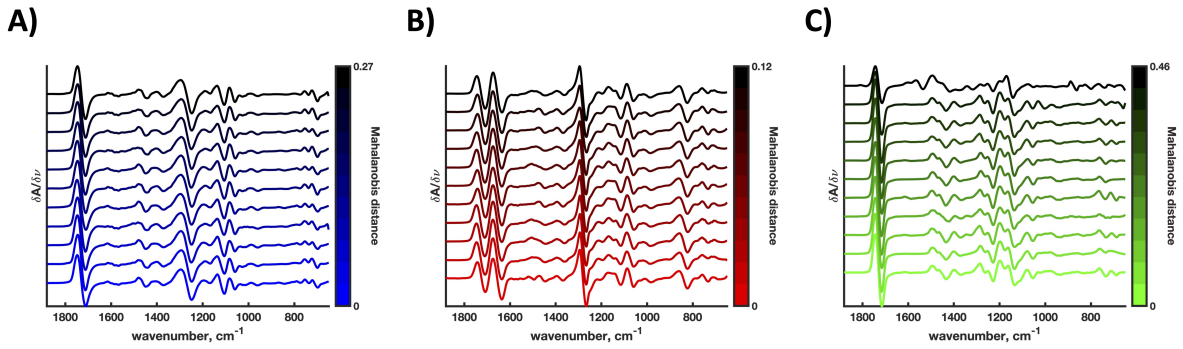

Figure 5 – A) Waterfall representation of the derivative spectral profiles of a test sample of known chemical composition assigned to the first class of paints (blue solid line corresponding to a Mahalanobis distance equal to 0) and of its 10 most similar training specimens belonging to the same category; B) waterfall representation of the derivative spectral profiles of a test sample of known chemical composition assigned to the second class of paints (red solid line corresponding to a Mahalanobis distance equal to 0) and of its 10 most similar training specimens belonging to the same category; C) waterfall representation of the derivative spectral profiles of a test sample of known chemical composition assigned to the third class of paints (green solid line corresponding to a Mahalanobis distance equal to 0) and of its 10 most similar training specimens belonging to the same category. The colour coding reflects the variation of the pairwise Mahalanobis distance values across training objects.

Furthermore, it is important to stress that the Mahalanobis distance between two spectral profiles (as calculated in this work) is not necessarily linked to their correlation coefficient (see Table SM.2 for an indicative comparison). The latter, in fact, being estimated from raw data, does not directly benefit from all the intrinsic properties of PCA modelling (which guarantees, for instance, dimensionality reduction, white noise filtering, *etc*.).
The same hierarchical approach was finally applied to the FTIR data measured for the paints with unknown chemical composition. Figures 6, 7 and 8 display the results yielded by its two

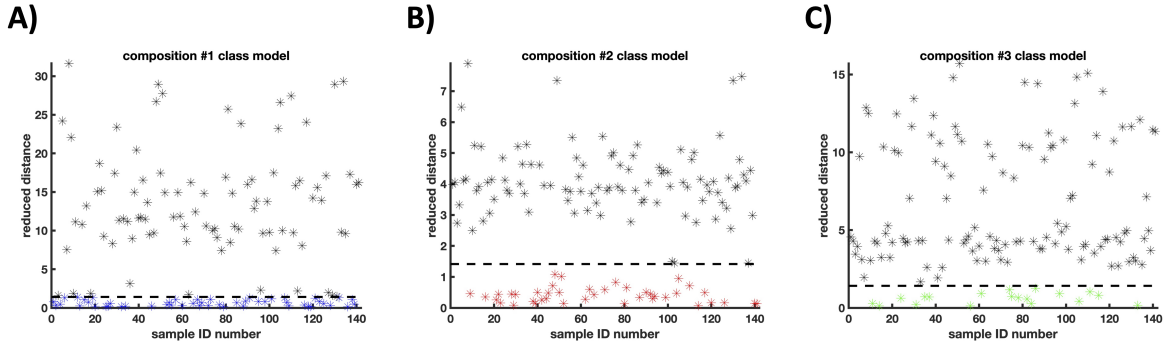

Figure 6 – Reduced distance plots for the external paint samples of unknown chemical composition. A) class #1 model; B) class #2 model; C) class #3 model. The black dotted line denotes the classification threshold set at $\sqrt{2}$. When a specimen is assigned to a specific category (*i.e*., its estimated reduced distance value is lower than $\sqrt{2}$), its respective symbol is coloured in accordance with Figure 1.

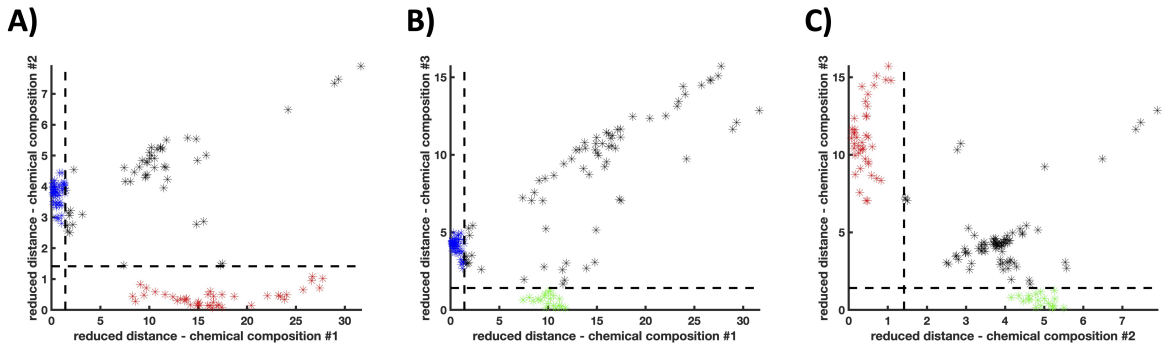

Figure 7 – Coomans plots for A) class model #1 *vs* class model #2, B) class model #1 *vs* class model #3 and C) class model #2 *vs* class model #3. The black dotted lines denote the classification thresholds set at $\sqrt{2}$ for all categories. When a specimen is assigned to a specific category (*i.e.*, its estimated reduced distance value is lower than $\sqrt{2}$), its respective symbol is coloured in accordance with Figure 1. No multiple simultaneous assignation was observed.

sequential computational steps: out of a total number of 141 specimens, 50 were recognised as characterised by chemical composition #1, 41 as characterised by chemical composition #2, 19 as characterised by chemical composition #3, while 31 were rejected as outliers by all the trained class models (indeed, their spectroscopic fingerprint presents considerable differences from that of the paints belonging to three categories under study - see Figure 9 for an exemplifying illustration). No multiple simultaneous assignation was observed. Also in the light of what stated for Figures 3 and 5, it is evident how the proposed multivariate analysis pipeline was capable not only of successfully achieving the chemical identification of such specimens[ii] but also of unveiling their most spectroscopically resemblant ones among those sharing the same chemical features. The consistency of this conclusion was also verified for most of the other paint samples not explicitly taken into account for the generation of Figure 8.

4. Conclusions

The chemical characterisation of paint samples as well as the discovery of their spectroscopic similarities are tasks that forensic operators usually need to address manually, being potentially biased by subjectivity and human errors. In this article, a hierarchical chemometric approach based on the principles of SIMCA modelling and on the definition and properties of the Mahalanobis distance was proposed for sequentially tackling both of them. Such a method was tested on FTIR data collected during a so-called market study conducted across the French city of Lille and its surroundings and permitted not only to successfully achieve the aforementioned characterisation for most of the analysed specimens but also to satisfactorily unveil common spectral signal patterns shared by those exhibiting the same chemical features. Altogether, the performance it guaranteed,

[ii]Paints of unknown composition were, in fact, assigned to unique classes of samples which share with them distinctive spectral bands, typical of specific chemical constituents underlying one of the three types of formulations considered here and not observed for any of the other two.

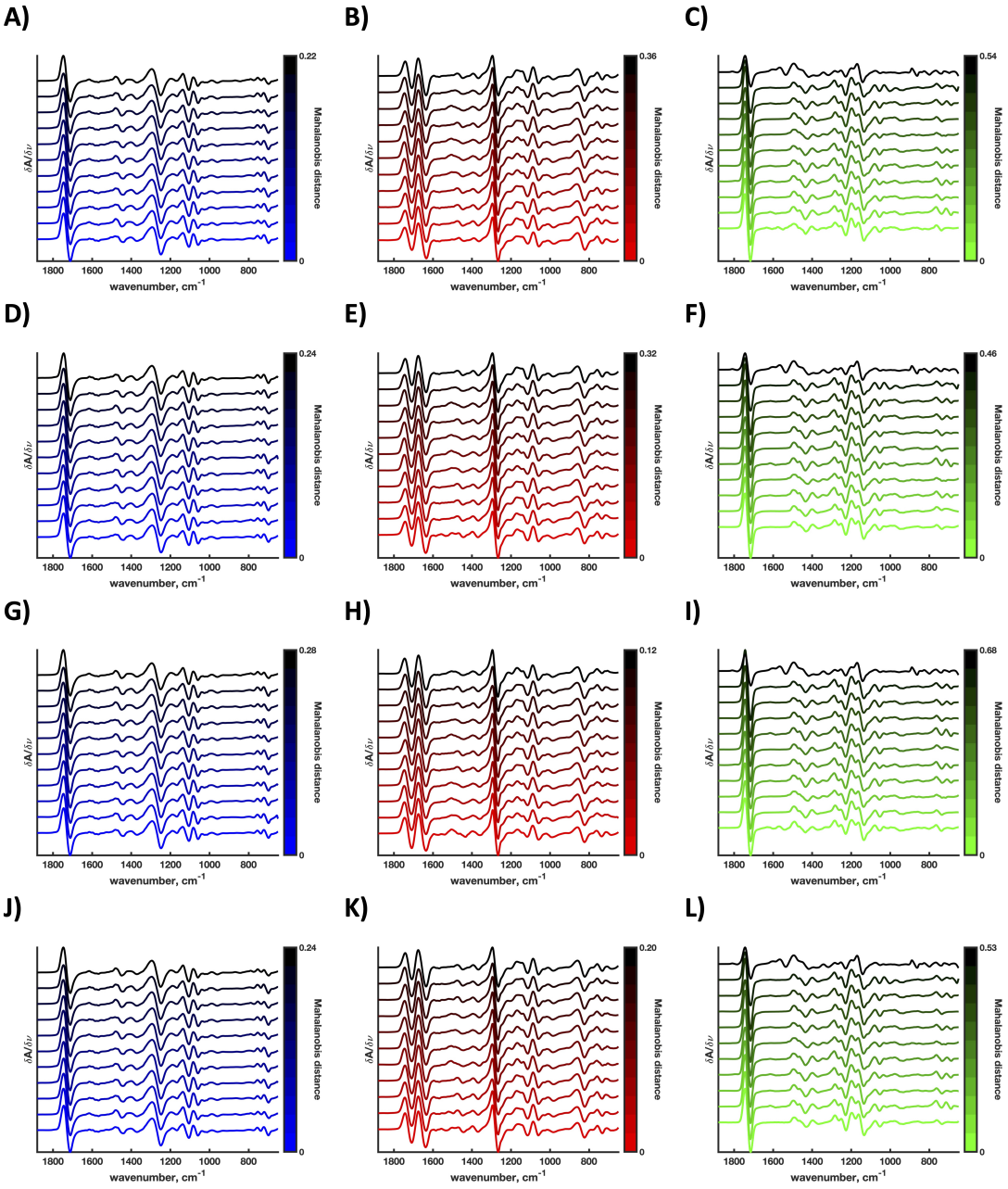

Figure 8 – A)-D)-G)-J) Waterfall representation of the derivative spectral profiles of four external samples of unknown chemical composition assigned to the first class of paints (blue solid lines corresponding to a Mahalanobis distance equal to 0) and of their 10 most similar training specimens belonging to the same category; B)-E)-H)-K) waterfall representation of the derivative spectral profiles of four external samples of unknown chemical composition assigned to the second class of paints (red solid lines corresponding to a Mahalanobis distance equal to 0) and of their 10 most similar training specimens belonging to the same category; C)-F)-I)-L) waterfall representation of the derivative spectral profiles of four external samples of unknown chemical composition assigned to the third class of paints (green solid lines corresponding to a Mahalanobis distance equal to 0) and of their 10 most similar training specimens belonging to the same category. The colour coding reflects the variation of the pairwise Mahalanobis distance values across training samples.

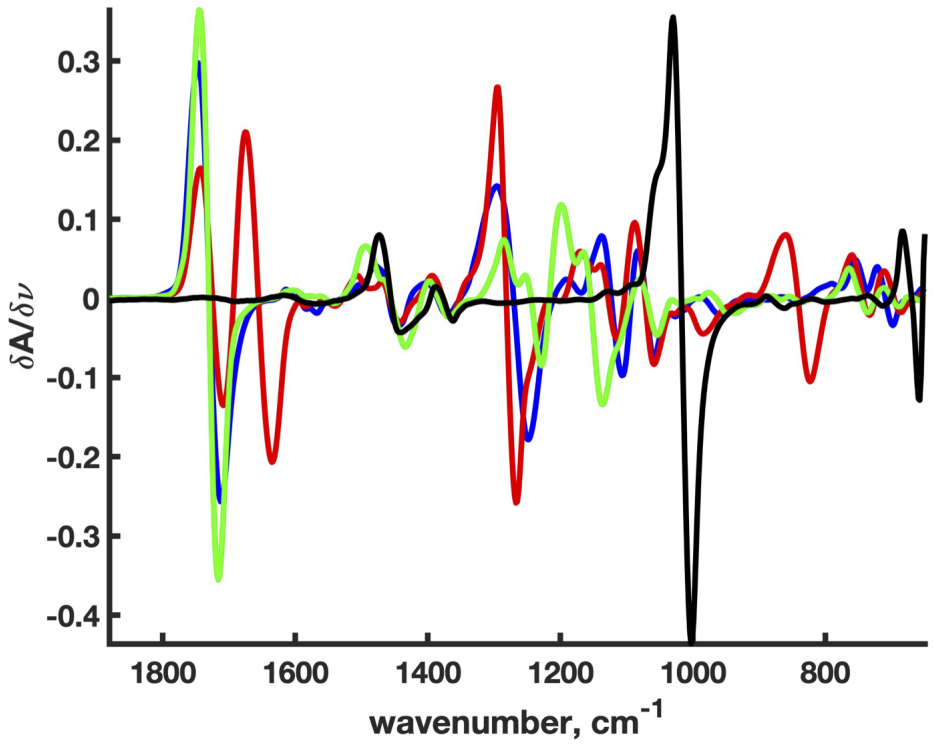

Figure 9 – Average derivative FTIR profiles of the training samples of chemical composition #1, #2 and #3 (blue, red and green solid line, respectively). Derivative FTIR profile of one of the paint specimens rejected as outlier by all the three SIMCA classification models built in this study (black solid line).

its simplicity and its rapid, automatic and objective nature (*i.e.*, mathematically grounded) constitute promising aspects in the light of its potential implementation in more extensive investigation campaigns and, possibly, for on-field applications. In complex scenarios like these, where paints may exhibit more chemical compositions and more resembling instrumental responses, building SIMCA models encompassing a larger number of categories and/or tuning differently their parameters [49] could in principle represent feasible strategies to accomplish a satisfactory resolution of the identification/matching problems at hand. In the former case, in practice, one might get insights into the chemical composition of a new sample by evaluating in parallel the responses output for it by all the class models that have been trained.

5. References

References

[1] R. Saferstein, Criminalistic: An Introduction To Forensic Science, Pearson Education, Inc., Upper Saddle River, United States of America, 2015.

[2] R. Saferstein, Forensic Science Handbook, Volume I, CRC Press, LLC, Boca Raton, United States of America, 2020.

[3] J. Wilkinson, J. Locke, D. Laing, The examination of paints as thin sections using visible microspectrophotometry and Fourier Transform infrared microscopy, Forensic Sci. Int. 38 (1988) 43–52.
[4] G. Ellis, M. Claybourn, S. Richards, The application of Fourier Transform Raman spectroscopy to the study of paint systems, Spectrochim. Acta A-M 46 (1990) 227–241.
[5] T. Allen, Paint sample presentation for Fourier Transform infrared microscopy, Vib. Spectrosc. 3 (1992) 217–237.
[6] G. Massonnet, W. Stoecklein, Identification of organic pigments in coatings: applications to red automotive topcoats. Part II: infrared spectroscopy, Sci. Justice 39 (1999) 135–140.
[7] H. Humecki, Practical Guide to Infrared Microspectroscopy, Marcel Dekker, Inc., New York, United States of America, 1995.
[8] R. Saferstein, Forensic Science Handbook, Volume III, Prentice Hall, Hoboken, United States of America, 2009.
[9] C. Muehlethaler, G. Massonnet, P. Esseiva, Discrimination and classification of FTIR spectra of red, blue and green spray paints using a multivariate statistical approach, Forensic Sci. Int. 244 (2014) 170–178.
[10] S. Ellison, S. Gregory, Predicting chance infrared spectroscopic matching frequencies, Anal. Chim. Acta 370 (1998) 181–190.
[11] J. Peris-Vicente, M. Lerma-García, E. Simó-Alfonso, J. Gimeno-Adelantado, M. Doménech-Carbó, Use of linear discriminant analysis applied to vibrational spectroscopy data to characterize commercial varnishes employed for art purposes, Anal. Chim. Acta 589 (2007) 208–215.
[12] R. Checa-Moreno, E. Manzano, G. Mirón, L. Capitan-Vallvey, Comparison between traditional strategies and classification technique (SIMCA) in the identification of old proteinaceous binders, Talanta 75 (2008) 697–704.
[13] S. Bell, L. Fido, S. Speers, W. Armstrong, Rapid forensic analysis and identification of "Lilac" architectural finishes using Raman spectroscopy, Appl. Spectrosc. 59 (2005) 100–108.
[14] E. Liszewski, S. Lewis, J. Siegel, J. Goodpaster, Characterization of automotive paint clear coats by ultraviolet absorption microspectrophotometry with subsequent chemometric analysis, Appl. Spectrosc. 64 (2010) 1122–1125.
[15] B. Lavine, A. Fasasi, N. Mirjankar, C. White, J. Mehta, Search prefilters for library matching of infrared spectra in PDQ database using the autocorrelation transformation, Microchem. J. 113 (2014) 30–35.
[16] S. Wold, Pattern recognition by means of disjoint principal components models, Pattern Recogn. 8 (1976) 127–139.
[17] S. Wold, M. Sjöström, SIMCA: a method for analyzing chemical data in terms of similarity and analogy, in: B. Kowalski (Ed.), Chemometrics: Theory and Application, American Chemical Society, Washington DC, United States of America, 1977, pp. 243–282.
[18] P. Mahalanobis, On the generalized distance in statistics, P. Natl. Ins. Sci. India 2 (1936) 49–55.
[19] R. De Maesschalck, D. Jouan-Rimbaud, D. Massart, The Mahalanobis distance, Chemometr. Intell. Lab. 50 (2000) 1–18.
[20] F. Govaert, G. de Roy, B. Decruyenaere, Analysis of black spray paints by Fourier Transform infrared spectrometry, X-ray fluorescence and visible microscopy, Probl. Forensic Sci. 47 (2001) 333–339.
[21] P. Buzzini, G. Massonnet, A market study of green spray paints by Fourier Transform infrared (FTIR) and Raman spectroscopy, Sci. Justice 44 (2004) 123–131.
[22] F. Govaert, M. Bernard, Discriminating red spray paints by optical microscopy, Fourier Transform infrared spectroscopy and X-ray fluorescence, Forensic Sci. Int. 140 (2004) 61–70.
[23] R. Gosse, S. Milet, B. Espanet, Discrimination of black spray paints, in: Proceedings of the 11th ENFSI (European Network of Forensic Science Institutes) European Paint & Glass Working Group Meeting, Berlin, Germany, 2005.
[24] S. Ryland, Discrimination of retail black spray paints, J. Am. Soc. Trace Evid. Examiners 1 (2010) 109–126.
[25] M. Falardeau, V. Moran, C. Muehlethaler, A random object-oriented population study of household paints measured by infrared spectroscopy, Forensic Sci. Int. 297 (2019) 72–80.
[26] K. Pearson, On lines and planes of closest fit to systems of points in space, Lond. Edinb. Dubl. Philos. Mag. J. Sci. 2 (1901) 559–572.
[27] H. Hotelling, Analysis of a complex of statistical variables into principal components, J. Educ. Psychol. 24 (1933) 417–441.

[28] H. Yue, S. Qin, Reconstruction-based fault identification using a combined index, Ind. Eng. Chem. Res. 40 (2001) 4403–4414.
[29] F. Marini, Classification methods in chemometrics, Curr. Anal. Chem. 6 (2010) 72–79.
[30] C. Durante, R. Bro, M. Cocchi, A classification tool for N-way array based on SIMCA methodology, Chemometr. Intell. Lab. 106 (2011) 73–85.
[31] E. Salvatore, M. Bevilacqua, R. Bro, F. Marini, M. Cocchi, Chapter 14 - Classification methods of multiway arrays as a basic tool for food PDO authentication, in: M. de la Guardia, A. Gonzálvez (Eds.), Food Protected Designation of Origin - Methodologies and Applications, Elsevier, B.V., Oxford, United Kingdom, 2013, pp. 339–382.
[32] M. Bevilacqua, R. Bucci, A. Magrì, A. Magrì, F. Marini, Tracing the origin of extra virgin olive oils by infrared spectroscopy and chemometrics: a case study, Anal. Chim. Acta 717 (2012) 39–51.
[33] R. Vitale, M. Bevilacqua, R. Bucci, A. Magrì, A. Magrì, F. Marini, A rapid and non-invasive method for authenticating the origin of pistachio samples by NIR spectroscopy and chemometrics, Chemometr. Intell. Lab 121 (2013) 90–99.
[34] A. Biancolillo, R. Bucci, A. Magrì, A. Magrì, F. Marini, Data-fusion for multiplatform characterization of an Italian craft beer aimed at its authentication, Anal. Chim. Acta 820 (2014) 23–31.
[35] R. Vitale, F. Marini, C. Ruckebusch, SIMCA modeling for overlapping classes: fixed or optimized decision threshold?, Anal. Chem. 90 (2018) 10738–10747.
[36] S. Wold, C. Albano, W. Dunn III, K. Esbensen, E. Hellberg, E. Johansson, M. Sjöström, Pattern recognition: finding and using regularities in multivariate data, in: H. Martens, H. Russwurm Jr. (Eds.), Food Research and Data Analysis, Applied Science Publishers, Ltd., London, United Kingdom, 1983, pp. 147–188.
[37] M. Barker, W. Rayens, Partial least squares for discrimination, J. Chemometr. 17 (2003) 166–173.
[38] A. Biancolillo, F. Marini, C. Ruckebusch, R. Vitale, Chemometric strategies for spectroscopy-based food authentication, Appl. Sci. - Basel 10 (2020) article number 6544.
[39] A. Pomerantsev, O. Rodionova, Multiclass partial least squares discriminant analysis: taking the right way - A critical tutorial, J. Chemometr. 32 (2018) article number e3030.
[40] R. Calvini, G. Orlandi, G. Foca, A. Ulrici, Development of a classification algorithm for efficient handling of multiple classes in sorting systems based on hyperspectral imaging, J. Spectr. Imaging 7 (2018) article number a13.
[41] M. Derde, L. Kaufman, D. Massart, A non-parametric class modelling technique, J. Chemometr. 3 (1989) 375–395.
[42] R. Snee, Validation of regression models: methods and examples, Technometrics 19 (1977) 415–428.
[43] R. Barnes, M. Dhanoa, S. Lister, Standard Normal Variate transformation and de-trending of near-infrared diffuse reflectance spectra, Appl. Spectrosc. 43 (1989) 772–777.
[44] A. Savitzky, M. Golay, Smoothing and differentiation of data by simplified least squares procedures, Anal. Chem. 36 (1964) 1627–1639.
[45] D. Coomans, I. Broeckaert, M. Derde, A. Tassin, D. Massart, S. Wold, Use of a microcomputer for the definition of multivariate confidence regions in medical diagnosis based on clinical laboratory profiles, Comput. Biomed. Res. 17 (1984) 1–14.
[46] J. Workman Jr., L. Weyer, Practical Guide to Interpretive Near-Infrared Spectroscopy, CRC Press, LLC, Boca Raton, United States of America, 2008.
[47] J. Zięba-Palus, A. Michalska, A. Wesełucha-Birczyńska, Characterisation of paint samples by infrared and Raman spectroscopy for criminalistic purposes, J. Mol. Struct. 993 (2011) 134–141.
[48] R. Wiesinger, L. Pagnin, M. Anghelone, L. Moretto, E. Orsega, M. Schreiner, Pigment and binder concentrations in modern paint samples determined by IR and Raman spectroscopy, Angew. Chem. Int. Edit. 57 (2018) 7401–7407.
[49] S. Małyjurek, R. Vitale, B. Walczak, Different strategies for class model optimization. A comparative study, Talanta 215 (2020) article number 120912.

- A hierarchical approach for paint sample classification and matching
- A market study conducted across the city of Lille, in France, and its surrounding
- Automatic identification of the chemical composition of spray paint samples
- Objective discovery of similarity patterns shared by paints of equal composition
- Analysis speed compatible with on-field applications

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: