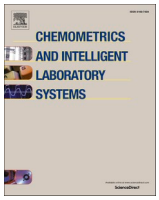
Chemometrics and Intelligent Laboratory Systems 250 (2024) 105150

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemometrics



Combining PLS-DA and SIMCA on NIR data for classifying raw materials for tyre industry: A hierarchical classification model

Riccardo Voccio [a,b], Cristina Malegori [a], Paolo Oliveri [a,*], Federica Branduani [c], Marco Arimondi [c], Andrea Bernardi [c], Giorgio Luciano [d], Mattia Cettolin [c]

[a] University of Genova, Department of Pharmacy (DIFAR), Viale Cembrano, 4, Genova, 16148, Italy
[b] University of Genova, Department of Chemistry and Industrial Chemistry (DCCI), Via Dodecaneso, 31, Genova, 16146, Italy
[c] Pirelli Tyre SpA, Viale Piero e Alberto Pirelli, 25, Milano, 20126, Italy
[d] Istituto di Scienze e Tecnologie Chimiche "Giulio Natta", Consiglio Nazionale delle Ricerche, SCITEC-CNR, Via De Marini 6, 16149, Genova, Italy

ARTICLE INFO

Keywords:
Tyre
Near infrared spectroscopy
Chemometrics
Raw material identification
Compliance verification

ABSTRACT

Tyre materials are complex products, as they are prepared using a number of raw materials, each of them with its specific chemical composition and functionality in the final product. It is, therefore, of crucial importance to avoid mislabeling errors and even to verify the compliance of raw materials entering the factory.

The present study proposes a strategy that makes use of near infrared (NIR) spectroscopy combined with chemometrics for raw material identification (RMID) and compliance verification of the most common raw materials used in the tyre industry. In particular, the chemometric model developed consists of a global hierarchical classification model, which combines nested PLS-DA nodes for RMID and SIMCA nodes for compliance verification, in a two-step approach.

The global model showed satisfactory results, as a 100 % of total correct predictions and a sensitivity higher than 90 % in the test set were obtained for most of the classes of interest.

The strategy obtained has the final goal of being directly applied on the raw materials at their receiving stage in factory, with the double advantage of minimizing the risk of mislabeling and, at the same time, decreasing the number of suspicious samples that need to be analyzed in the laboratory, by means of traditional methods, for verifying their compliance.

1. Introduction

Tyre materials are particularly complex under the chemical and structural point of view, and their chemical-physical properties are strictly dependent on the composition of the formulation from which they are prepared. Indeed, few differences in the compound composition might affect the properties of the final product [1]. In particular, for compound preparation, different raw materials are used, each one with its specific function in the final product.

Among the tyre ingredients, the most important is the raw elastomer, which constitutes the polymeric skeleton of the tyre. Raw elastomers are generally of two types: natural rubber and synthetic rubber. The most common synthetic rubbers are isoprene rubber (IR), styrene-butadiene rubber (SBR), polybutadiene rubber (BR), butyl rubber (IIR), chlorobutyl rubber (CIIR), bromobutyl rubber (BIIR) [1] but, generally, these elastomers show lack of ultimate strength and, therefore, they cannot be used alone as they are insufficient to meet the demanding requirements for tyres [2].

Therefore, for compound preparation, other ingredients are needed. The most commonly used are: curing agents, fillers, curing accelerators, activators, vulcanization retardants, organic acids, antioxidants, and various additives [3].

Curing agents – generally, sulphurs, peroxides, and metal oxides – are used to generate the vulcanization process, by means of which, the network junctures between the polymer chains are produced [4]. The formation of these crosslinking bonds is necessary to impart tyre its characteristic features.

Fillers are defined as materials that are added to a polymer formulation to lower the compounding cost or to improve properties. The most common fillers are carbon black and silica. All the fillers can be classified into two groups – namely, reinforcing fillers and non-reinforcing fillers. Reinforcing fillers impart good mechanical properties to the

* Corresponding author.
E-mail address: paolo.oliveri@unige.it (P. Oliveri).

https://doi.org/10.1016/j.chemolab.2024.105150
Received 18 April 2024; Received in revised form 15 May 2024; Accepted 18 May 2024
Available online 19 May 2024

vulcanizates, while non-reinforcing fillers or inert fillers are generally used to reduce the cost of the compound [1].

Accelerators are used to control the rate of the vulcanization process and activators are used to promote the action of accelerators. Vulcanization retardants interact with the system curing agent/accelerator/activator, to avoid vulcanization process for a controlled time. Organic acids react with the activators promoting the reaction between accelerators and activators. The use of antioxidants protects tyres from oxidation processes that can modify the mechanical properties of the tyres themselves [3]. In addition to the above-mentioned raw materials, various additives are also used, in order to improve the performance of the final product.

Due to the number and to the specific functionality of ingredients used for preparing tyre, it is crucial to perform a reliable raw material identification (RMID) directly at the receiving stage.

Moreover, it is not only important to recognize the raw material but even to verify its compliance in respect to the industrial standards. It is, therefore, of fundamental importance to have an accurate method to verify that the materials received and addressed to the compound preparation are compliant with the target specifications.

Up to date, methods implemented in the tyre industry for raw material identification and compliance verification often rely on International Standards (i.e. ISO or ASTM), and need long procedures, sophisticated instrumentation and a qualified personnel able to use them. Such instrumentation may include GC-MS, HPLC, DSC, and similar analytical technique.

In light of the considerations presented above, the application of non-destructive spectroscopic techniques for both RMID and compliance verification in the tyre industry seems to be advantageous. Near infrared (NIR) spectroscopy, among non-destructive spectroscopic techniques, is particularly valuable, as it is fast, non-invasive and user-friendly [5] and, for all these reasons, it is suitable for automation of RMID and compliance verification in raw material acceptance phases.

Indeed, NIR spectroscopy is more and more widely used for these objectives in the industries. In particular, it finds application in the pharmaceutical industry, where NIR spectroscopy was used, for example, for obtaining classification models able to classify between the most common excipients and active ingredients [6,7] Moreover, NIR spectroscopy was also applied for classifying between the most common raw materials used in the cosmetic industry [8].

For what is currently known by the authors, NIR spectroscopy has not been applied yet for RMID and compliance verification in the tyre industry, but its application is worthy of being studied.

When dealing with RMID in the tyre industry, we are facing the so-called multi-class problem, as there are many classes to be predicted. In the case of a multi-class problem, conventional classification models generally do not perform properly, as they are mostly efficient to solve binary/ternary classification cases [9]. Despite of the numerosity of the classes involved, the raw materials can be grouped according to their functionality in the final product, leading to a so-called nested structure. From an analytical perspective, similar functionality is often reflected on similar chemical composition of the raw materials, which is mirrored, in turn, on the corresponding NIR spectra.

The creation of a hierarchical model is an efficient strategy for handling a multi-class problem in the presence of nested structures. In hierarchical models, in fact, the classes are modelled into a tree in which each node is a simple classifier, usually a binary classifier, which makes the discrimination between the different child clusters, until reaching the leaf nodes, which contain only a single class [9].

This approach, used for classification purposes, proved to be advantageous, as documented in different works. For example, the hierarchical approach was applied on hyperspectral imaging data in the NIR region, for classifying different polymers in mixed plastic waste [10], and it was also applied on bioacoustics signals for classifying different anuran species [11].

Moreover, the hierarchical approach was used for addressing regression purposes, for example, in determining water solubility and the Setschenow coefficient of ~300 compounds [12] and, also, for estimating the human age from face images [13].

In the present work, NIR spectroscopic data are used to create a hierarchical classification tree that combines a nested discriminant classification strategy, for RMID purposes, and a class-modeling approach, for compliance verification. In more detail, Partial Least Squares – Discriminant Analysis (PLS-DA) [14] implemented by the Automated Hierarchical Model Builder (AHIMBU) [15] is combined with Soft Independent Modeling of Class Analogy (SIMCA) [16] in a global hierarchical structure. This is the novelty rational of the presented work, because generally, in the literature, discriminant classification methods and class-modeling strategies are used separately. In fact, the discriminant approach is used to classify between two or more classes when they are meaningfully represented, while the class-modeling approach focuses on a specific target class with the aim of verifying the compliance of samples with the features of that class [17].

A multivariate classification strategy that combines discriminant and class-modeling approaches is already present in literature in the works published by Małyjurek et al. for the authenticity verification of different indigenous herbal tea. But, differently from the strategy proposed in the present study, Małyjurek et al. apply, at first, a class-modeling method (Potential Function Method – PFM) and, subsequently, a Random Forest discriminant algorithm, in a so-called two-step approach [18]The same authors also published the comparison between their two-step approach and other classification methods [19].

2. Materials and methods

For developing the classification models, a starting database containing historical NIR data acquired between 2021 and 2022 have been used.

This starting database contains NIR spectra that have been acquired in two different production plants – one is referred to as Plant 1 and the other to as Plant 2 – for taking into account the variability between two different laboratories.

Moreover, new experimental data have been acquired in January 2023, after a rational selection of the most significant missing samples, in order to span chemical variability of the raw materials themselves. In particular, the new acquisitions were performed with the aim of having available at least 20 samples for one of the two production plants, minimizing the experimental effort. To do this, the raw materials with a number of samples between 12 and 20 in one of the two production plant and less than 12 in the other one were selected. The spectrometers used in both the production plants are Frontier™ devices (PerkinElmer, Inc., Shelton, CT, US), operating in the range between 10,000 cm^{-1} (1000 nm) and 4000 cm^{-1} (2500 nm). The spectral resolution is equal to 16 cm^{-1} and the number of scans is equal to 32. All the samples were acquired as solids, within a Petri dish, in the reflection acquisition mode. Spectra were stored in the database either with the reflectance percentage scale (%R) or with the pseudo-absorbance (pseudo-Abs) scale, depending on the plant and on the operator. For each sample, 3 or 4 spectral replicates were acquired.

From the whole database, a dataset containing 1916 samples belonging to 36 raw materials was extracted and submitted to data processing. In particular, the raw materials studied in this work were selected according to the following constraints.

i) availability of the information about the chemical composition of the raw material;
ii) availability of at least 20 samples acquired in one of the two production plants under study.

All the raw materials are coded using an alphanumeric code containing three characters. The first character is a letter and is indicated as the “family” of the raw material. The second and third characters are

numbers and indicate the raw material itself.

The subdivision of the raw materials into families was made so that raw materials with similar functionality in the final product and/or similar chemical composition belong to the same family. The number of families is equal to 11 and the number of raw materials inside each family is between 1 and 9. One of the families, moreover, contain miscellaneous raw materials that cannot be categorized into the other 10 families, and it is the family coded with the letter 'K'.

The real codes of the raw materials are encrypted, being considered as sensitive information protected by industrial secrecy, but the belonging family for different raw materials is maintained, to allow the reader to delve into the chemical differences between the various classes to be predicted.

In Fig. 1, a scheme of the structure of the dataset is presented, taking as an example the specification code A01, where the first character, A, corresponds to the family (functionality in the final product) and 01 corresponds to the raw material.

In Table S1 (Supplementary Material), the numerosity of the dataset is reported detailing the number of samples acquired in Plant 1 and Plant 2 and their separation in training set and test set.

3. Data processing

3.1. Automatization of data import and exploratory analysis

For data import purposes, a code under the MATLAB® environment (Version 2022b, The MathWorks, Inc., Natick, MA, US) was created. The most important steps performed by the code are the calculation of averages on replicates of the same samples and the checking of the spectra measurement scale, with the consecutive transform from %R to pseudo-Abs [20]for all the spectra that were stored with the %R scale.

An exploratory analysis by means of principal component analysis (PCA) [21] was then performed on the spectroscopic data, in order to understand the variance pattern of the dataset and to visually evaluate the presence of evident groupings within the raw materials.

Moreover, this exploratory phase, was used to understand the effect of different spectral pre-treatments and to verify how their application affects the data structure. The final aim of this data pre-processing was to minimize the systematic differences between the two plant laboratories and to maximize the differences due to the chemical structure of the raw materials themselves.

3.2. Classification strategy

For properly developing, optimizing and validating the chemometric strategy, the dataset was subdivided into a training set and a test set, according to a class-wise Venetian blind scheme [14] in which, for each class, 2/3 of the samples are employed as the training set and 1/3 as the test set, for a total of 1293 samples in the training set and 623 samples in the test set (see Table S1 for observing the numerosity for each class).

The first step for the development of the hierarchical global classification tree was RMID, performed by means of PLS-DA models. The structure can be schematized as follows.

a) nested PLS-DA nodes, in which the family (first character of the specification code – Fig. 1) is classified;
b) within the family, further PLS-DA nodes, in which the raw material (the numeric part of the specification code – Fig. 1) is classified.

The model for family recognition – identified as *a)* – is itself a hierarchical model, which was created automatically using the Automated HIerarchical Model BUilder (AHIMBU) [15]. AHIMBU, in fact, makes use of a discriminant classification method to create the hierarchical model automatically, based on the similarities between classes. In particular, AHIMBU performs PLS-DA on all the combinations between pairs of classes. The minimum of nonerror rate in cross-validation is used as a parameter to evaluate the quality of these binary classifications. The worst discriminated classes are merged together and constitute one PLS-DA node of the hierarchical model. This process is iterated until all the classes are well separated and, afterwards, the hierarchical structure is released as an output. AHIMBU is advantageous for developing the hierarchical skeleton of the model *a)*, because the present dataset is characterized by a multi-class problem, in which the spectroscopic similarities between classes are not evident.

PLS-DA nodes identified as *b)* were added to the family-classification AHIMBU model, with the aim of discriminating the raw materials inside the single family, —, and leading to the global hierarchical model for RMID.

Afterwards, the models for compliance verification, based on the SIMCA method were added to the leaves of the RMID tree, with the final goal of verifying the compliance of each new sample belonging to a specific raw material. This further step allows not only to verify compliance but also to make the model for RMID more robust. Indeed, a sample misclassified by the RMID model is very likely to be considered non-compliant by the SIMCA of the class that was predicted. A further verification of the mislabeling presence is, therefore, added.

The global model obtained is schematized in Fig. 2, where the model is applied, as an example, on a test sample of class A03.

3.3. Model optimization and sample diagnostics

All the nodes of the nested PLS-DA model for RMID were optimized, in terms of pre-processing and complexity, in cross-validation (CV) applied only on samples of the training set (see Paragraph 3.2), using the Model Optimizer tool in the PLS_Toolbox software, version 9.2

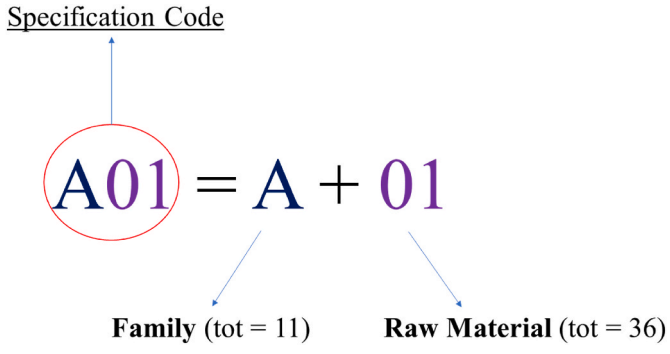


Fig. 1. Schematization of the raw material identification code: A01, A indicates the family and 01 indicates the raw material. In total we have 36 raw materials that belong to 11 families.

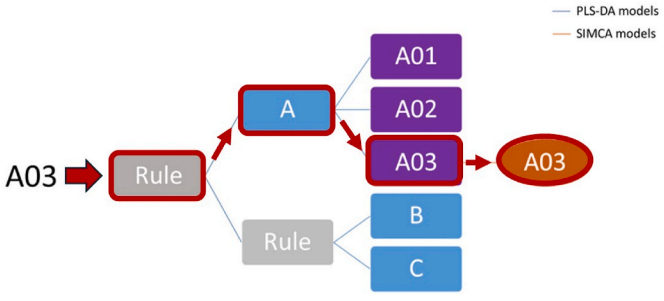


Fig. 2. Schematization of application of the global hierarchical model for a test sample of raw material A03, in which the PLS-DA models for family recognition are represented in blue and the PLS-DA models for raw material identification in violet. On the terminal leave, the SIMCA model for compliance verification is shown in orange. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

(Eigenvector Research, Inc., Manson, WA, US).

The cross-validation scheme used was Venetian-blind with number of data splits equal to 6.

Model complexity was optimized studying the number of latent variables in a range between 2 and 10.

The pre-processing studied are: i) Standard Normal Variate (SNV) transform [22]; ii) Multiplicative Scatter Correction (MSC) [23]; iii) 1st and 2nd Savitzky–Golay derivatives, with polynomial order equal to 2 and 3 and datapoint window size equal to 7 [24]; iv) detrending [22].

Parameters used for evaluating the models were the classification error rate in CV [25] and the root mean square error in CV, RMSECV [26]. In particular, a joint interpretation of the two parameters was carried out, trying to minimize at first the complexity (number of latent variables) of the model.

In particular for the nodes in which the less satisfactory values were observed for the figures of merit in CV, an accurate model diagnostic was performed, focusing on the chemical reasons of class separation. When the model diagnostic phase showed that two classes were chemically very similar, variable selection was evaluated and implemented.

The same CV scheme was applied for building the SIMCA models. Moreover, also the SIMCA models were optimized in terms of pre-processing and complexity (in a range between 1 and 10 principal components). In particular, the number of principal components to be included in the model were chosen following a rigorous approach [27] i.e., observing the inflection point in the scree plots (percentage of explained variance associated to each PC) of the training set for each raw material.

After having optimized the PLS-DA nodes and SIMCA nodes, the global hierarchical model was applied to the test set for validating the robustness and calculating the error in prediction. In particular, the % of Total Correct Prediction for the PLS-DA nodes and the sensitivity for the SIMCA nodes have been used. For the SIMCA nodes, in particular, only sensitivity was evaluated according to a rigorous approach [27]. In this approach, in fact, only the target class is taken into account and used to evaluate the performances of the SIMCA models. The evaluation of parameters that take into account also the non-target class can, indeed, determine a bias in the model evaluation [27].

For 5 of the 26 raw materials less than 5 samples are present in the test set. For these less-represented raw materials the sensitivity results on the test set are to be considered as a proof-of-concept of the applicability of the developed model.

Finally, a chemical interpretation of the misclassifications obtained was performed.

4. Results and discussion

4.1. Exploratory analysis

After applying PCA on the raw data with column mean-centering, it is observed that the greatest source of variability between samples is given by the production plant in which the samples have been analyzed.

This behavior, observed in PCA, is due to systematic differences of the NIR signals acquired in the two production plants, ascribable to global intensity effects and baseline vertical shifts; these differences are imputable to unavoidable and expected instrumental differences, which might be minimized applying a proper row pre-treatment to the data.

Standard Normal Variate (SNV) transform is one of the most common pre-treatments used to remove these unwanted signal variations [28] As an example of the influence of SNV on the present dataset, raw signals and the SNV-transformed signals of sample B02 are reported in Fig. 3a and d, respectively. Moreover, the PCA score plot defined by the two lowest-order components for the same spectra is reported in Fig. 3b and e.

In Fig. 3a, a vertical shift is observed, in the whole spectral range, between the spectra acquired in the two production plants. This behavior is reflected in the corresponding score plot (Fig. 3b), where samples tend to group, according to the plant laboratory, along the direction of the lowest-order component. Moreover, PC1, explaining 99.8 % of variance, indicates that most of the variability in raw data is related to systematic differences between instruments of the plants in which the spectra have been recorded. These results are, moreover, confirmed observing the corresponding loading line plot of PC1 (Fig. 3c), in which

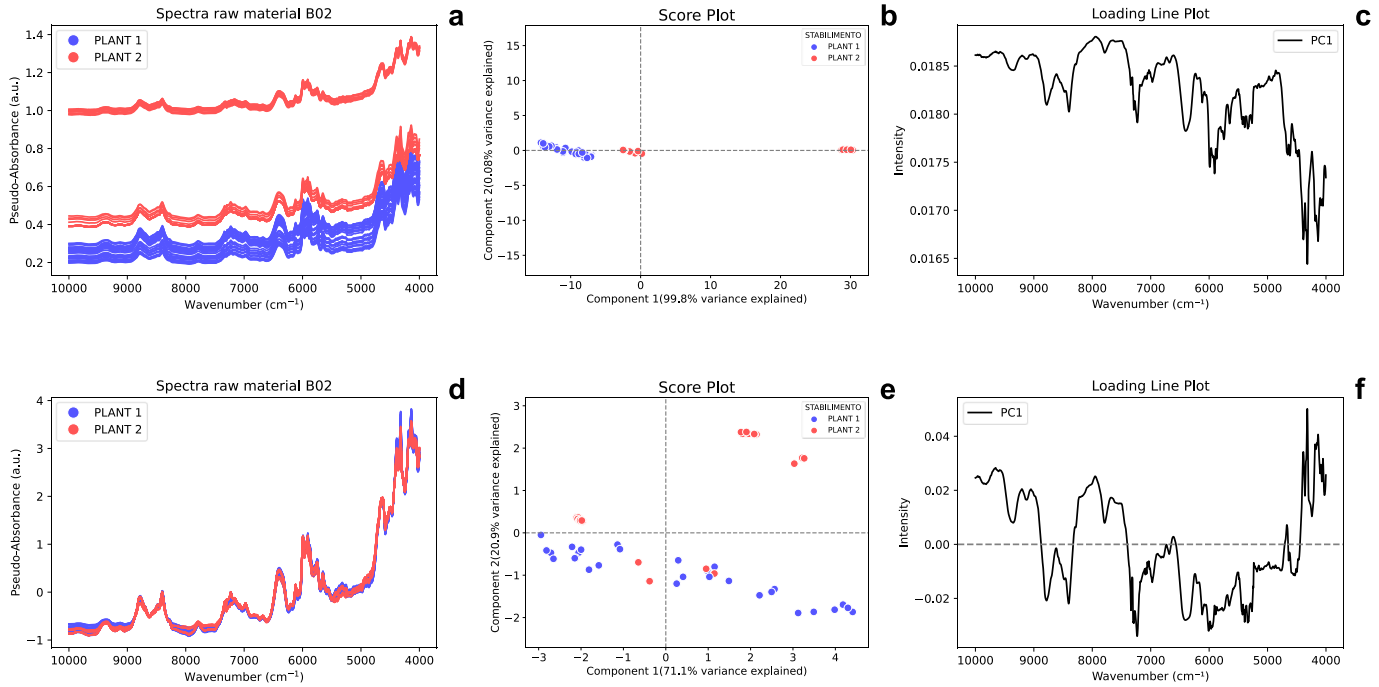


Fig. 3. Raw spectra of samples of raw material B02 (a), the corresponding PC1 vs. PC2 score plot (b) and loading line plot of the PC1 (c). SNV-transformed spectra (d), the corresponding PC1 vs. PC2 score plot (d) and loading line plot of the PC1 (f). Blue: spectra measured in Plant 1. Red: spectra measured in Plant 2. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

a global intensity effect is observable.

This information, retained in the raw data, is minimized after applying the SNV transform on the spectra, as represented in Fig. 3d, in which the vertical shift present in the raw spectra (Fig. 3a) is almost completely removed, leading to the overlapping of the two groups of spectra. This change in the data profiles is reflected in the corresponding score plot (Fig. 3e), in which samples are no longer grouped according to the production plant laboratory. Indeed, samples analyzed in different production plants are in the same region of the score plot (Fig. 3e) and the residual presence of less defined groupings is not due to the systematic difference highlighted.

Observing, finally, the corresponding loading line plot of PC1 (Fig. 3f), the global intensity effect is not present anymore.

The trend observed for raw material B02 is generally found for all the raw materials involved in this case study.

To observe the general behavior of samples after applying one of the most common pre-treatments, the SNV transform, PCA was applied on the SNV-transformed dataset. In more detail, the score plot defined by the two lowest-order components can be observed in Fig. 4a (samples colored according to the raw material) and in Fig. 4b (samples colored according to the family).

Observing Fig. 4, it is possible to see how samples tends to group according both to the family and to the raw material. This is a clue to how it is possible to minimize unwanted systematic differences due to the production plant and to maximize the spectral differences due to the chemical composition, by performing a suitable pre-processing. In lights of these results, it is, therefore, worth creating a classification model for the recognition of the raw materials.

4.2. Nested PLS-DA for RMID

Considering the RMID model, the first step for recognizing the family was developed automatically using the AHIMBU algorithm [15], obtaining the classification tree schematized in Fig. 5.

To each leaf of the AHIMBU tree obtained, PLS-DA nodes for recognizing the raw materials were added obtaining the global hierarchical model for RMID (Fig. 6).

In Table 1, for each node of the nested PLS-DA hierarchical model, the classes considered, the pre-processing performed, the number of latent variables and the spectral range included are listed.

For almost all nodes, in optimization phase the lowest value of error rate in CV and RMSECV has been obtained applying SNV as pre-treatment. This pre-treatment has indeed been chosen for almost all nodes. The only exception is the PLS-DA node that discriminates raw materials belonging to family I (Table 1 – Node 2.I). The minimum of RMSECV and Error Rate in CV has been obtained applying first derivative to this node, instead of SNV, due to a higher efficiency in minimizing the unwanted systematic effects that characterize these signals.

Focusing on the description of the AHIMBU structure (Fig. 5), families B, C, D and E appear as the simplest to be discriminated, thanks to

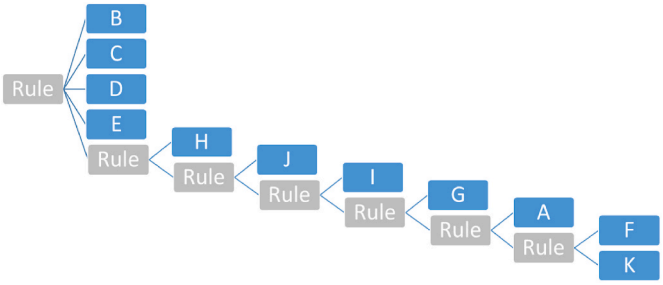


Fig. 5. AHIMBU model created for family classification.

the chemical similarities of the materials reflected in the characteristic spectral profiles. As an example, family D is composed by 5 silica with a very peculiar profile and with reduced differences among them (Fig. 7). So, samples belonging to family D are easy to be distinguished from the other families, but samples of this family are difficult to be discriminated as single raw materials. This is confirmed by the number of latent variables used in the corresponding node (Table 1 – Node 2.D), which is equal to 7 – the highest one, if compared with the other nodes. Except for this node, it is generally observed that the number of latent variables included is relatively low, indicating that the PLS-DA models have not a great complexity and, therefore, they are characterized by a high reliability and robustness.

In the optimization phase, wavelength range selection was also evaluated. In most of the nodes, minimum values of Error Rate in CV and RMSECV were obtained including the whole spectral range. However, there are two exceptions: the node that discriminates between family H and the group of families J, I, G, A, F, K (Table 1 – Node 1.2) and the node that discriminates between family F and family K (Table 1 – Node 1.7). For these two nodes, the best performances in CV were observed including in the model only the spectral range between 6000 and 4500 cm^{-1}. Regarding the node 1.2, this behavior was observed as some raw materials inside the H family have a spectral behavior similar to the raw material J01. Therefore, selecting, for this specific node, the spectral range in which major spectral differences are observed, the model performances are improved. Regarding the node 1.7, it is also the last node of the AHIMBU tree. This means that the AHIMBU algorithm recognize families F and K as the most difficult to be distinguished. The raw materials belonging to family K are, indeed, miscellaneous raw materials. Differently from the other families, within family K raw materials have more variability in the chemical composition. Some of the raw materials inside the K family have, indeed, more similar chemical composition to raw materials of other families respect to the ones of the K family. The spectral range selection between 6000 and 4500 cm^{-1} (Table 1 – Node 1.7) determines, therefore, an improvement in the classification performances.

After having optimized all the nodes, the global RMID model was applied to the test set, obtaining the results reported in Table 2, which

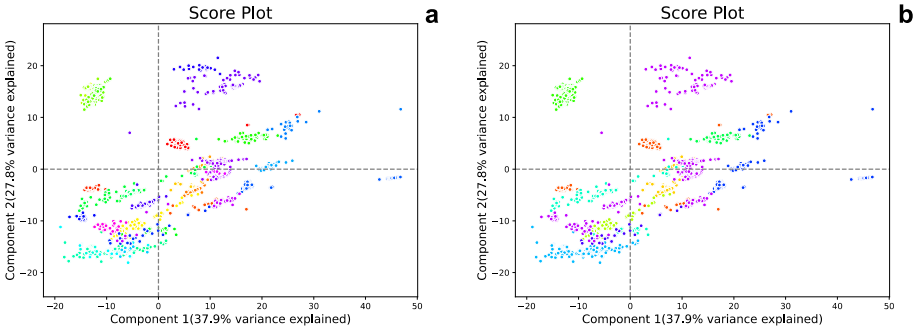


Fig. 4. PC1 vs. PC2Score plots obtained on the whole dataset after SNV transform. Samples are colored according to the raw material (a) and according to the family (b).

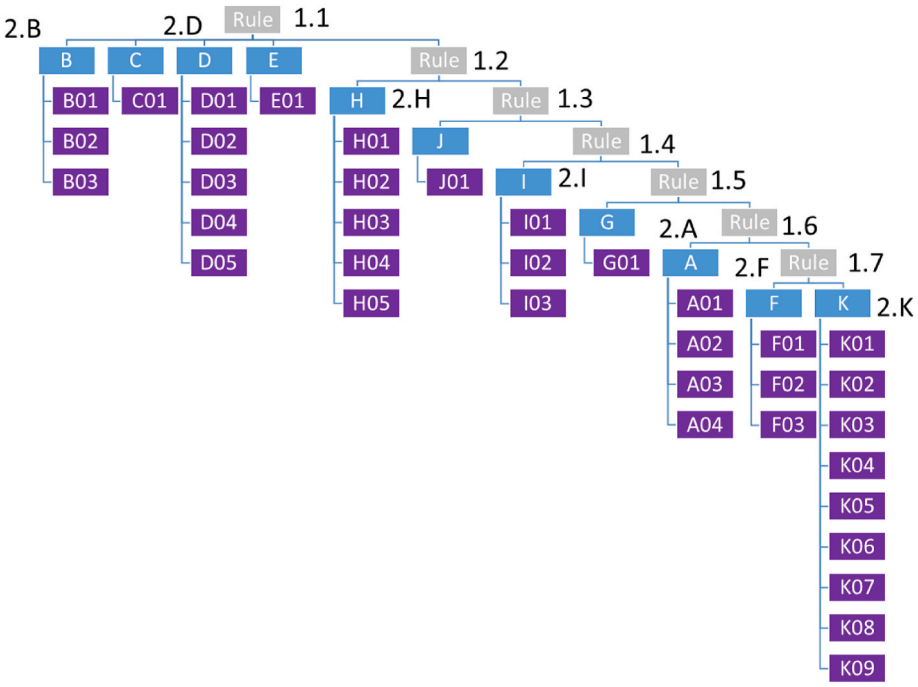


Fig. 6. Hierarchical nested PLS-DA model built for RMID.

Table 1
Classes, preprocessing, number of latent variables and spectral range for each node of the nested PLS-DA model.

Node	Classes Considered	Preprocessing selected	Number of latent variables	Spectral range included
1.1	[B, C, D, E, OTHERS]	SNV	6	10,000–4000 cm^{-1}
1.2	[H, OTHERS]	SNV	5	6000–4500 cm^{-1}
1.3	[J, OTHERS]	SNV	5	10,000–4000 cm^{-1}
1.4	[I, OTHERS]	SNV	5	10,000–4000 cm^{-1}
1.5	[G, OTHERS]	SNV	4	10,000–4000 cm^{-1}
1.6	[A, OTHERS]	SNV	5	10,000–4000 cm^{-1}
1.7	[F, K]	SNV	5	6000–4500 cm^{-1}
2.B	[B01, B02, B03]	SNV	2	10,000–4000 cm^{-1}
2.D	[D01, D02, D03, D04, D05]	SNV	7	10,000–4000 cm^{-1}
2.H	[H01, H02, H03, H04, H05]	SNV	5	10,000–4000 cm^{-1}
2.I	[I01, I02, I03]	First Derivative	5	10,000–4000 cm^{-1}
2.A	[A01, A02, A03, A04]	SNV	3	10,000–4000 cm^{-1}
2.F	[F01, F02, F03]	SNV	3	10,000–4000 cm^{-1}
2.K	[K01, K02, K03, K04, K05, K06, K07, K08, K09]	SNV	5	10,000–4000 cm^{-1}

lists, for each raw material: the results in prediction, expressed as % of total correct predictions of class membership, the number of samples in the test set, and number of samples that were misclassified.

As reported in Table 2, only 13 out of the 623 samples of the test set were misclassified. It is worth noticing that, for 27 out of the 36 raw materials under study, a 100 % of total correct predictions was achieved.

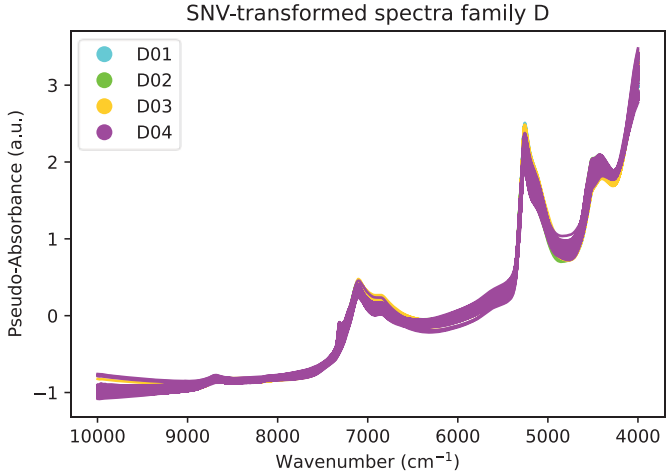


Fig. 7. NIR spectra SNV-transformed of D raw materials.

Moreover, for 8 of the remaining raw materials, a percentage of total correct predictions higher than 90 % of was obtained.

The class with the highest number of misclassifications is D04, in which 4 prediction errors are observed. These 4 samples are all classified as D02, probably because the raw materials belonging to the D families are very similar from a structural and spectroscopic perspective. But, although the spectra are quite similar, as it is observable in Fig. 7, a percentage of total correct predictions higher than 90 % was obtained.

Only K02 had 0 % of total correct predictions, and this may be ascribable to the small number of samples (=20) in the training set (Table 1). For this reason, the model was not enough robust for recognizing K02 samples.

An interesting case of misclassification regards a sample labeled as K05 that has been predicted by the nested PLS-DA model as a B02. Observing the NIR spectra of these two classes, and comparing them to the spectrum of this sample (Fig. 8), it can be observed that this sample has a spectral profile totally resembling the one of the class predicted (B02), instead of the one of the class labeled. Therefore, it is most

Table 2
Prediction results obtained on the test set applying the nested PLS-DA model for each raw material.

Raw Material	% of Total Correct Predictions	N. of samples in Test Set	N. of samples wrongly predicted
A01	100	19	0
A02	92.3	13	1
A03	100	17	0
A04	91.0	11	1
B01	100	19	0
B02	100	24	0
B03	100	12	0
C01	100	30	0
D01	95.5	22	1
D02	100	9	0
D03	94.7	19	1
D04	91.0	44	4
D05	100	33	0
E01	100	18	0
F01	100	14	0
F02	100	20	0
F03	100	25	0
G01	100	19	0
H01	100	24	0
H02	100	15	0
H03	100	13	0
H04	100	5	0
H05	100	10	0
I01	100	15	0
I02	100	5	0
I03	100	24	0
J01	100	9	0
K01	100	8	0
K02	0	1	1
K03	95.5	22	1
K04	91.3	23	2
K05	97.0	34	1
K06	100	17	0
K07	100	3	0
K08	100	23	0
K09	100	4	0

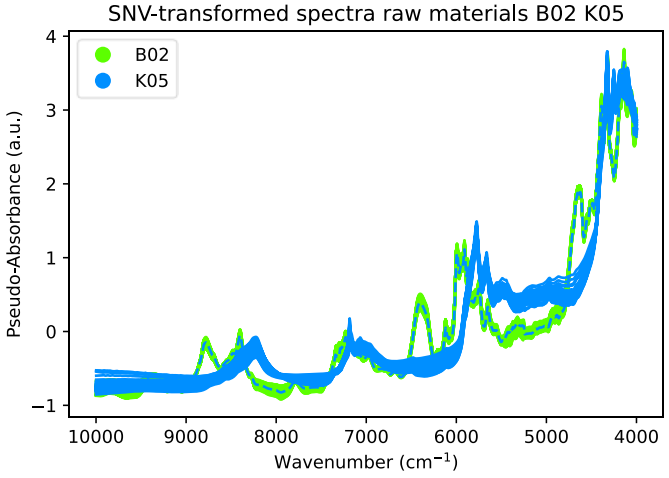


Fig. 8. Spectra of classes B02 (green) and K05 (blue). The mislabeled K05 sample is represented as a dashed line. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

probably a case of mislabeling that was properly recognized by the model.

4.3. Combining nested PLS-DA and SIMCA

For each class predicted by the nested PLS-DA model, SIMCA nodes were added for compliance verification.

The hierarchical model obtained combining nested PLS-DA and SIMCA is schematized in Fig. 9, where the global model is applied for the classification of four samples – two correctly identified as A01 and A02 (true-positive) and two incorrectly classified: a false-positive (a sample of type A02 classified as A01) and a false-negative (a sample of type A01 classified as A02).

Results obtained applying the SIMCA models to the test set are schematized in Table 3 where, for each raw material, the number of principal components included in the model and the evaluation parameter sensitivity is reported [17].

In Table 3, it is possible to observe that the number of PCs used for developing the SIMCA classification ranges between 1 and 5, demonstrating a generally low complexity of the models. The sensitivities obtained are satisfactory, as the sensitivity is higher than 90 % for 27 of the 36 raw materials under study, highlighting the ability of SIMCA models to accept samples belonging to the target class and, therefore, to avoid false-negative errors. Moreover, it was observed that the SIMCA models rejected the misclassified samples of the RMID model, avoiding the false-positive error. This is another evidence of the robustness of the SIMCA models and how they are able not only to verify the compliance of raw materials but to make the entire classification strategy applied robust.

5. Conclusions

From the results presented and discussed above, it can be concluded that the use of NIR spectroscopy combined with a chemometric model based on a two-step approach (nested PLS-DA and SIMCA) represents a valid strategy to deal with both the crucial issues of RMID and compliance verification in manufacturing industries.

Particularly noteworthy are the results obtained from the hierarchical tree combining nested PLS-DA models for RMID, in which it was observed that the hierarchical approach is particularly suitable for solving cases of multi-class classification [9]. The advantage of using AHIMBU to generate a classification structure for the families of raw materials was demonstrated. Moreover, PLS-DA provided satisfactory results when dealing with raw material classification, once the family was recognized.

Finally, thanks to the addition of the SIMCA nodes downstream the PLS-DA models, it was possible to carry out both RMID and compliance verification by means of a single global hierarchical model. Particularly valuable is the fact that the SIMCA nodes can reject the misclassified samples of the previous nested PLS-DA model, making the overall global model more robust in identifying mislabeled samples.

Results obtained applying the global model to the test set are satisfactory, as 100 % of total correct predictions was obtained for most of the raw materials under study, and the SIMCA models showed good results in terms of sensitivity and specificity.

However, robustness of the model can be still improved, especially for those raw materials, such as K02, whose dataset numerosity can be increased.

Moreover, for some of the raw materials selected, less than 20 samples for one of the production plants were available. Due to this non-representative sampling, a lower reliability of the class membership prediction for that plant could be expected. This limitation needs to be considered in an industrial implementation of the developed model.

In conclusion, even if for some raw materials, few samples were incorrectly recognized as non-compliant, the model can be considered efficient, because its final aim is to perform a screening of the raw materials entering the factory, and to select only few samples, exactly those predicted as non-compliant, on which further analyses in the chemical laboratory can be carried out.

CRediT authorship contribution statement

Riccardo Voccio: Writing – original draft, Software, Investigation,

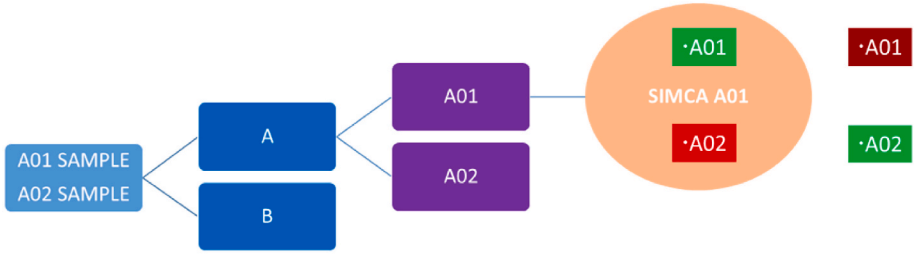


Fig. 9. Global hierarchical model composed by a nested PLS-DA model for RMID and SIMCA models for compliance verification.

Table 3
Number of principal components and sensitivity are reported for each raw material classified.

Raw Material	N. of PCs	Sensitivity (%)
A01	1	100
A02	2	100
A03	3	94.0
A04	3	80.0
B01	4	95.0
B02	4	96.0
B03	2	83.0
C01	2	100
D01	3	100
D02	5	100
D03	4	94.0
D04	4	90.0
D05	4	91.0
E01	2	100
F01	2	71.0
F02	3	95.0
F03	2	92.0
G01	2	100
H01	2	96.0
H02	3	93.0
H03	3	92.0
H04	3	80.0
H05	3	90.0
I01	3	93.0
I02	3	80.0
I03	1	92.0
J01	2	100
K01	4	88.0
K02	3	100
K03	3	86.0
K04	3	100
K05	4	91.0
K06	2	81.0
K07	3	100
K08	4	87.0
K09	3	100

Data curation. **Cristina Malegori:** Writing – review & editing, Supervision, Methodology, Formal analysis, Conceptualization. **Paolo Oliveri:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Federica Branduani:** Investigation, Data curation. **Marco Arimondi:** Resources, Project administration, Conceptualization. **Andrea Bernardi:** Investigation, Data curation. **Giorgio Luciano:** Software, Methodology, Investigation. **Mattia Cettolin:** Writing – review & editing, Supervision, Resources, Project administration, Conceptualization.

Declaration of competing interest

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.

Data availability

The data that has been used is confidential.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.chemolab.2024.105150.

References

[1] K.S. Sisanth, M.G. Thomas, J. Abraham, S. Thomas, General introduction to rubber compounding, in: Progress in Rubber Nanocomposites, 2017, https://doi.org/10.1016/B978-0-08-100409-8.00001-2.
[2] M.S. Evans, Tyre Compounding for Improved Performance, iSmithers Rapra publishing, 2002.
[3] G.T. Viola, Bacchelli fabio, fabbri augusto, elastomeri, in: Enciclopedia Degli Idrocarburi, 2006, pp. 789–835. Treccani.
[4] A.Y. Coran, Vulcanization, in: Science and Technology of Rubber, 2005, https://doi.org/10.1016/B978-012464786-2/50010-9.
[5] C. Pasquini, Near infrared spectroscopy: fundamentals, practical aspects and analytical applications, J. Braz. Chem. Soc. 14 (2003), https://doi.org/10.1590/S0103-50532003000200006.
[6] L. Sun, C. Hsiung, C.G. Pederson, P. Zou, V. Smith, M. Von Gunten, N.A. O'Brien, Pharmaceutical raw material identification using miniature near-infrared (MicroNIR) spectroscopy and supervised pattern recognition using support vector machine, Appl. Spectrosc. 70 (2016), https://doi.org/10.1177/0003702816638281.
[7] W.B. Mroczyk, K.M. Michalski, Application of modern computer methods for recognition of chemical compounds in NIRS, Comput. Chem. 22 (1998), https://doi.org/10.1016/S0097-8485(97)00042-9.
[8] H. Grunewald, C. Kurowski, D. Timm, U. Grummisch, U. Meyhack, Rapid non-destructive raw material identification in the cosmetic industry with near-infrared spectroscopy, J. Near Infrared Spectrosc. 6 (1998), https://doi.org/10.1255/jnirs.197.
[9] M. Aly, Survey on multiclass classification methods, Neural Netw. 19 (2005) 1–9.
[10] G. Bonifazi, G. Capobianco, S. Serranti, A hierarchical classification approach for recognition of low-density (LDPE) and high-density polyethylene (HDPE) in mixed plastic waste based on short-wave infrared (SWIR) hyperspectral imaging, Spectrochim. Acta Mol. Biomol. Spectrosc. 198 (2018), https://doi.org/10.1016/j.saa.2018.03.006.
[11] J.G. Colonna, J. Gama, E.F. Nakamura, Recognizing family, genus, and species of anuran using a hierarchical classification approach, in: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2016, https://doi.org/10.1007/978-3-319-46307-0_13.
[12] C. De Stefano, G. Lando, C. Malegori, P. Oliveri, S. Sammartano, Prediction of water solubility and Setschenow coefficients by tree-based regression strategies, J. Mol. Liq. 282 (2019), https://doi.org/10.1016/j.molliq.2019.03.029.
[13] P. Thukral, K. Mitra, R. Chellappa, A hierarchical approach for human age estimation, in: ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 2012, https://doi.org/10.1109/ICASSP.2012.6288182.
[14] D. Ballabio, V. Consonni, Classification tools in chemistry. Part 1: linear models. PLS-DA, Anal. Methods 5 (2013), https://doi.org/10.1039/c3ay40582f.
[15] L. Marchi, I. Krylov, R.T. Roginski, B. Wise, F. Di Donato, S. Nieto-Ortega, J.F. Q. Pereira, R. Bro, Automatic hierarchical model builder, J. Chemom. 36 (2022), https://doi.org/10.1002/cem.3455.
[16] S. Wold, Pattern recognition by means of disjoint principal components models, Pattern Recogn. 8 (1976), https://doi.org/10.1016/0031-3203(76)90014-5.
[17] P. Oliveri, Class-modelling in food analytical chemistry: development, sampling, optimisation and validation issues – a tutorial, Anal. Chim. Acta 982 (2017), https://doi.org/10.1016/j.aca.2017.05.013.
[18] Z. Małyjurek, D. de Beer, H. van Schoor, J. Colling, E. Joubert, B. Walczak, Class-modelling of overlapping classes. A two-step authentication approach, Anal. Chim. Acta 1191 (2022), https://doi.org/10.1016/j.aca.2021.339284.

[19] Z. Małyjurek, D. de Beer, E. Joubert, B. Walczak, Combining class-modelling and discriminant methods for improvement of products authentication, Chemometr. Intell. Lab. Syst. 228 (2022), https://doi.org/10.1016/j.chemolab.2022.104620.
[20] H.K. Hughes, Beer's law and the optimum transmittance in absorption measurements, Appl. Opt. 2 (1963) 937–945.
[21] S. Wold, K. Esbensen, P. Geladi, Principal component analysis, Chemometr. Intell. Lab. Syst. 2 (1987) 37–52, https://doi.org/10.1016/0169-7439(87)80084-9.
[22] R.J. Barnes, M.S. Dhanoa, S.J. Lister, Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra, Appl. Spectrosc. 43 (1989), https://doi.org/10.1366/0003702894202201.
[23] A. Kohler, M. Zimonja, V. Segtnan, H. Martens, Standard normal variate, multiplicative signal correction and extended multiplicative signal correction preprocessing in biospectroscopy, in: Comprehensive Chemometrics, 2009, https://doi.org/10.1016/B978-044452701-1.00102-2.
[24] A. Savitzky, M.J.E. Golay, Smoothing and differentiation of data by simplified least squares procedures, Anal. Chem. 36 (1964), https://doi.org/10.1021/ac60214a047.
[25] J.-H. Kim, Estimating classification error rate: repeated cross-validation, repeated hold-out and bootstrap, Comput. Stat. Data Anal. 53 (2009) 3735–3745, https://doi.org/10.1016/j.csda.2009.04.009.
[26] O.M. Kvalheim, R. Arneberg, B. Grung, T. Rajalahti, Determination of optimum number of components in partial least squares regression from distributions of the root-mean-squared error obtained by Monte Carlo resampling, J. Chemom. 32 (2018), https://doi.org/10.1002/cem.2993.
[27] O.Ye Rodionova, P. Oliveri, A.L. Pomerantsev, Rigorous and compliant approaches to one-class classification, Chemometr. Intell. Lab. Syst. 159 (2016) 89–96, https://doi.org/10.1016/j.chemolab.2016.10.002.
[28] P. Oliveri, C. Malegori, R. Simonetti, M. Casale, The impact of signal pre-processing on the final interpretation of analytical outcomes – a tutorial, Anal. Chim. Acta 1058 (2019), https://doi.org/10.1016/j.aca.2018.10.055.

Table S1: Numerosity of the dataset for each raw material and each production plant

		Plant 1	Plant 2	Plant 1	Plant 2
Family	Raw Material	N° Sample Train		N° Sample Test	
A	01	25	14	14	5
	02	19	7	9	4
	03	11	22	4	13
	04	17	4	10	1
B	01	20	16	12	7
	02	23	22	13	11
	03	3	22	0	12
C	01	23	35	12	18
D	01	33	13	16	6
	02	18	2	8	1
	03	34	4	16	3
	04	40	46	20	24
	05	28	36	14	19
E	01	13	22	7	11
F	01	1	26	2	12
	02	28	8	13	7
	03	2	46	2	23
G	01	0	36	0	19
H	01	0	48	0	24
	02	0	28	0	15
	03	0	26	0	13
	04	0	20	0	5
	05	0	21	0	10
I	01	27	1	15	0
	02	16	4	4	1
	03	24	25	10	14
J	01	12	8	7	2
	01	19	2	7	1
	02	20	0	1	0

K	03	29	14	16	6
	04	27	18	13	10
	05	38	29	21	13
	06	20	13	10	7
	07	0	20	0	3
	08	7	38	2	21
	09	20	0	2	2