



HAL
open science

A Novel Methodology for Determining Effectiveness of Preprocessing Methods in Reducing Undesired Spectral Variability in Near Infrared Spectra

Jhon Buendia Garcia, Julien Gornay, Marion Lacoue-Negre, Sílvia Mas Garcia, Jihane Er-Rmyly, Ryad Bendoula, Jean-Michel Roger

► **To cite this version:**

Jhon Buendia Garcia, Julien Gornay, Marion Lacoue-Negre, Sílvia Mas Garcia, Jihane Er-Rmyly, et al.. A Novel Methodology for Determining Effectiveness of Preprocessing Methods in Reducing Undesired Spectral Variability in Near Infrared Spectra. *Journal of Near Infrared Spectroscopy*, 2022, 30 (2), pp.74-88. 10.1177/09670335211047959 . hal-03685527

HAL Id: hal-03685527

<https://ifp.hal.science/hal-03685527>

Submitted on 2 Jun 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

1 A novel analysis methodology for preprocessing methods 2 effectiveness determination in reducing undesired spectral 3 variability in near-infrared spectra acquisition

4 Jhon Buendia Garcia^{a,c}, Julien Gornay^a, Marion Lacoue-Negre^a, Silvia Mas Garcia^{b,c}, Jihane Er-Rmyly^a,
5 Ryad Bendoula^b, Jean-Michel Roger^{b,c}

6 ^a IFP Energies Nouvelles, Rond Point de l'échangeur de Solaize, France

7 ^b ITAP-INRAE, Institut Agro, University Montpellier, Montpellier, France

8 ^c ChemHouse Research Group, Montpellier, France

9

10 Corresponding Authors:

11 Jhon Buendia (jhon.andersson@gmail.com)

12 Abstract

13 This study uses a novel analysis methodology based on the Hierarchical Clustering Analysis (HCA) to
14 determine the effectiveness of different preprocessing methods in minimizing undesired spectral
15 variability in near-infrared spectroscopy due to both the consecutive and repetitive acquisition of the
16 spectrum and the sample temperature. Nine preprocessing methods and different combinations of
17 them were evaluated in four case studies: reproducibility, repeatability, sample temperature, and
18 combination of the before mentioned cases. Eighty-four spectra acquired on seven different
19 hydrocarbon samples from catalytic conversion processes have been selected as the real case study
20 to illustrate the potential of the mentioned methodology. The approach proposed allows a more
21 detailed discriminatory analysis compared to the classical methods for comparing the between-class
22 and the within-class variances, such as the Wilks' lambda criterion, and hence constitutes a powerful
23 tool to determine adequate spectral preprocessing strategies. This study also proves the potential of
24 the discrimination analysis methodology as a general scheme to identify atypical behaviors either in
25 the spectrum acquisition or in the measured samples.

26 Keywords

27 Spectral variability, Hierarchical Clustering Analysis (HCA), Principal Components Analysis (PCA),
28 Preprocessing effectiveness, outliers, Qresidual, Hotelling's T^2 , reproducibility, repeatability, sample
29 temperature.

30 1 Introduction

31 In the past few decades, the use of Near-Infrared Spectroscopy (NIRS) in the development of non-
32 destructive and rapid measurement applications has been significantly increasing in several
33 industries, such as food^{1, 2}, pharmaceuticals^{3, 4}, and petroleum^{5, 6}. Due to the recent growth boom in
34 using the NIR spectrum for real-time acquisition data^{7, 8}, the need to determine, analyze, and
35 minimize spectral variability that is not associated with the physicochemical characteristics of the
36 sample has notably arisen. This need becomes particularly evident when spectral variability is mainly
37 generated by factors associated with the spectrum acquisition, such as spectrometer system,

38 operator, measurement conditions, and environmental factors such as temperature and humidity,
39 rather than the physicochemical characteristics of the sample. An example of this is the possible
40 generation of spectral variability in the consecutive and repetitive acquisition of NIR spectra on a
41 sample whose physicochemical characteristics remain constant over the spectrum acquisition. A lack
42 of minimization of this type of spectral variability can result in inaccurate analysis and interpretation
43 of spectroscopic information, misleading conclusions and flawed decision making^{9,10}.

44 Among the classical performance parameters needed to validate a measurement methodology,
45 precision is the most affected by the aforementioned factors. Precision is defined as the closeness of
46 agreement between measured values obtained by replicate measurements on the same or similar
47 samples under conditions of repeatability or reproducibility¹¹. Repeatability conditions include the
48 same measurement procedure, the same operator, the same instrument and measurement
49 conditions, the same location, and a short interval between repetitions¹². On the other hand,
50 reproducibility implies successive measurements of the same sample under changing measurement
51 conditions¹³, such as measurement principle, measurement method, operator, measurement
52 instrument, reference standard, location, conditions of use, and time. The NIR spectra acquisition is
53 very sensitive to any change in measurement; even ensuring that both the spectrum acquisition
54 conditions and the physicochemical characteristics of the sample do not change in a repetitive NIR
55 spectrum acquisition, the resulting spectra may have differences that can lead to random errors and
56 deviations, which must be corrected or minimized.

57 Due to its high impact on NIR spectra acquisition accuracy, the temperature is the most studied
58 influencing parameter^{14, 15}. Hansen *et al.*¹⁶ showed that molecular bonds vibration intensity depends
59 on temperature, leading to changes in the spectrum according to temperature variation.
60 Furthermore, some physicochemical properties of samples, such as viscosity and density, are
61 temperature-dependent, and many changes in the sample due to temperature are not permanent
62 and do not reflect the intrinsic nature of the sample¹⁷⁻¹⁹. Nevertheless, these changes can
63 significantly affect spectrum acquisition. As with the spectral variability generated by repetitive
64 spectrum acquisition of a specific sample, the variability caused by sample temperature must be
65 minimized to ensure the reliable description of the sample physicochemical behavior from the
66 spectroscopic information extracted.

67 Data preprocessing is a common step for reducing undesired effects and for minimizing spectral
68 variability. There are different preprocessing algorithms for the correction of the undesired spectral
69 variation; these can be divided into two main categories: scatter-correction methods, employed to
70 correct the additive and multiplicative effects, and spectral derivatives, used to minimize the sources

71 of unwanted and non-informative spectral variations²⁰. Among the most commonly preprocessing
72 methods used in NIR spectroscopy, Savitsky-Golay derivative (Sav-Gol)²¹, Extended Multiple Signal
73 Correction (EMSC)²², Standard Normal Variate (SNV)²³, and recently, the Variable Sorting for
74 Normalization (VSN)²⁴, can be highlighted. However, the effectiveness of the preprocessing methods
75 is highly dependent on the type of spectroscopic information analyzed and the factors that are
76 causing its variability⁹.

77 The preprocessing method effectiveness evaluation is generally based on the performance of
78 prediction models²⁵⁻²⁷. Among the contributions reported in the literature, the work of Gerretzen et
79 al.,²⁸ which presents a novel approach for the selection of the most appropriate preprocessing
80 methods based on the design of experiments, is worth mentioning. Similarly, the studies of Devos et
81 al.,²⁹ and Allegrini et al.,³⁰ which, by means of a parallel workflow approach of preprocessing and
82 variable selection, present an interesting alternative to the optimization of the preprocessing method
83 selection. Nonetheless, the application of these approaches may be limited when the variability of
84 the physicochemical characteristics of the samples is negligible, but significant spectral variability
85 exists as a result of the repetitive spectrum acquisition and the sample temperature. In that case, a
86 different analysis approach may yield more detailed results, helping to improve understanding of the
87 impact of these parameters. Another less common approach to assessing preprocessing methods
88 effectiveness is analyzing the spectral variance³¹. Different statistical tools are available to determine
89 both within-class variance (multiple measurements of the same sample) and between-class variance
90 (measurements of different samples). One of the most common criteria used to evaluate between-
91 class and within-class variances is the Wilks' lambda³².

92 In this study, a novel and general strategy based on the Hierarchical Clustering Analysis (HCA)³³ was
93 proposed for evaluating the effectiveness of preprocessing methods in reducing the spectral
94 variability generated by parameters related to the continuous and dynamic spectrum acquisition. To
95 this aim, the effectiveness of nine preprocessing methods and different combinations of them in
96 minimizing undesired spectral variability due to repeatability, reproducibility, sample temperature
97 and combination of these parameters were evaluated. Eighty-four spectra acquired on seven
98 different hydrocarbon samples from catalytic conversion processes have been selected as the real
99 case study to illustrate the potential of the mentioned methodology.

100 To obtain reliable conclusions and validate the results obtained by the analysis methodology
101 proposed, the Wilks' lambda criterion³² was used as a reference method.

102 2 Material and methods

103 2.1 Samples

104 Twenty-four vacuum gasoil (VGO) samples were processed in the catalytic conversion pilot plant
105 reactors at IFPEN (Solaize, France). From these reactors, ninety-three different hydrocarbon samples,
106 known as total effluent, were obtained (see references^{34, 35} for a detailed description of catalytic
107 conversion processes). From these 93 samples, 7 samples were selected, ensuring their
108 representativeness and physicochemical diversity. Table 1 summarizes four relevant physicochemical
109 properties of the selected samples: the density³⁶ and the simulated initial boiling point and
110 distillation temperatures range to obtain both 5% and 95% of sample distillate (Simulated Distillation
111 IBP, T5 and T95)³⁷. It can be observed that physicochemical variability between the selected samples
112 is guaranteed.

113 **Table 1 Samples physicochemical properties. SimDis IBP, T5 & T95 description: Simulated distillation to determine the**
114 **temperatures to start the sample evaporation and to recover both 5% and 95% of sample distillate**

Sample Physicochemical Properties				
Sample ID	Density (gr/ml)	Distillation Temperatures		
		IBP (°C)	SimDis T5 (°C)	SimDis T95 (°C)
Sample 1	0.8049	84.7	121.6	425.8
Sample 2	0.8186	79.8	111.5	474.5
Sample 3	0.8219	80.3	117.8	516
Sample 4	0.8411	83.8	136.1	503.5
Sample 5	0.847	83.7	129.1	512.6
Sample 6	0.8962	148.9	227.6	504.6
Sample 7	0.9181	159.8	231.6	502.9

115

116 2.2 Spectral acquisition

117 The spectra were recorded with a Fourier Transform Near-Infrared spectrometer (FT-NIR) MATRIX-F
118 (Bruker, Optik GmbH, Ettlingen - Germany) within the range of 9090 - 4600 cm⁻¹ (1100 - 2160 nm) and
119 a resolution of 4 cm⁻¹. 32 scans were used to obtain the final spectrum after each measurement. For
120 acquiring absorbance spectra, the spectrometer system was equipped with an immersion
121 transmittance probe with an optical path fixed at 2 mm withstanding temperatures ranging from -40
122 °C to 200 °C. The software used with the spectrometer was OVP (OPUS Validation Program - Bruker,
123 Optik GmbH, Ettlingen - Germany) which automatically performs a series of analyses of the
124 instrument's performance, evaluates them and ensures that it is operating within specifications.
125 Besides, to ensure the spectrometer operation within specifications and that the spectral variability
126 generated was due to the parameters evaluated and not to the instrument's inadequate functioning,
127 the spectrometer performance was validated once a day using cyclohexane as an external reference

128 sample. Before NIR analysis, the samples were heated in closed flasks at 60°C for 1 hour in a water
129 bath and shaken manually to ensure their liquid state and homogeneity. The initial boiling point (IBP)
130 reported in Table 1 guarantees no loss of volatiles.

131 Ensuring the integrity and stability of both the sample and the NIR spectrum acquisition conditions,
132 spectral variability due to repeatability, reproducibility, and sample temperature was generated. A
133 short description of the spectrum acquisition for the cases evaluated in this study is presented
134 below.

- 135 • **Case 1. Spectral variability due to reproducibility:** Each of the seven samples was analyzed
136 one time per day during five consecutive days at 60 °C. Thirty-five spectra were obtained.
- 137 • **Case 2. Spectral variability due to repeatability:** Each of the seven samples was analyzed
138 three times on the same day at 60 °C. All samples were analyzed in less than 8 hours.
139 Twenty-one spectra were obtained.
- 140 • **Case 3. Spectral variability due to the sample temperature:** Each of the seven samples was
141 analyzed at five different temperatures, ranging from 60 °C to 80 °C with a temperature
142 increment of 5°C. The samples were heated in closed flasks at the desired temperature for 1
143 hour. Evaporation losses of volatiles were null or negligible (see IBP in Table 1). Twenty-eight
144 spectra were obtained.
- 145 • **Case 4. Spectral variability due to the combination of the aforementioned cases:** In this
146 case, all spectra acquired in the above-described cases were used.

147 For each case, a matrix was generated. Each analyzed sample was defined as a class; thus, seven
148 classes were defined in all matrices.

149 **2.3 Analysis methodology**

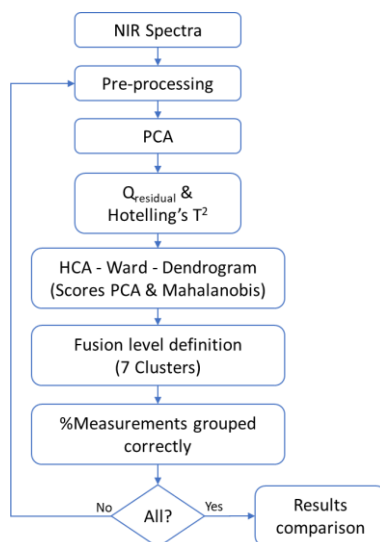
150 The main steps of the data analysis workflow proposed in this study are schematized in Figure 1. A
151 brief explanation of the procedure is given as follows.

152 The first step consisted in preprocessing each of the generated matrices. The nine most commonly
153 preprocessing methods used in NIR data were divided into two categories; the filtering and
154 normalization methods. The preprocessing methods from each category were analyzed individually.
155 If the total reduction or compensation of the studied spectral variability were not achieved, the
156 evaluated preprocessing method were complemented with the methods belonging to the opposite
157 category. This allowed the evaluation of all possible combinations and order of use of the
158 preprocessing methods. The methods evaluated are described in Table 2. It should be emphasized
159 that each preprocessing scenario evaluated includes the data centering by columns.

Table 2 Preprocessing methods description

#	Category	Name	Acronym	Parameters
1	Filtering	Automatic Weighted Least Squares Baseline ²⁰	AWLS-B	
2		Norris-Williams Derivation ³⁸	NW-D	15-point window, gap size = 7, First order derivation
3		Savitsky-Golay Derivative ²¹	SG-D	15-point window, polynomial order = 2 First order derivative
4		Detrend ²³	Dtd	Polynomial order = 1
5		Extended Multiplicative Scatter/Signal Correction ²²	EMSC	Reference spectrum (basis to remove the scatter) = mean of each matrix generated, polynomial order = 2, whole spectral range
6	Normalization	Multiplicative Signal Correction ³⁹	MSC	Reference data = mean of data, whole spectral range
7		Standard Normal Variate ²³	SNV	
8		Probabilistic Quotient Normalization ⁴⁰	PQN	
9		Variable Sorting for Normalization ²⁴	VSN	Automatic calculation

161



162

163

Figure 1 – Methodology flow diagram

164 Afterward, a preliminary inspection and dimension reduction of the corresponding dataset was
 165 performed by principal component analysis (PCA)⁴¹. The number of the chosen principal components
 166 (PCs) captured at least 99 % of the total variance in the dataset. The Q residual and Hotelling's T²
 167 tests were performed to determine the possible presence of anomalous data⁴².

168 The chosen PCs scores were then used to perform the hierarchical clustering analysis (HCA)
 169 employing Ward's algorithm and the Mahalanobis distance. The HCA aims to group clusters to form a

170 new one to either minimize a statistical distance between classes or maximize a measure of similarity
171 between them^{43, 44}. The analysis starts with as many groups as individuals contained in the dataset.
172 From these initial groups, clusters are formed in an ascending manner until all cases treated are
173 included in at least one of them. Ward's algorithm seeks to minimize each group variance by
174 calculating all samples mean in each cluster. The algorithm then calculates each case distance and
175 the cluster mean, adding up the distances between all cases. Finally, the clusters whose sum of
176 distances is minimal are grouped. This procedure creates homogeneous groups of a similar amount
177 of individuals. For achieving the grouping of classes, it is necessary to define a comparison parameter
178 to calculate the variance of each class concerning the others. The most common is the Mahalanobis
179 distance³⁹.

180 A widespread manner of displaying the cluster analysis results is constructing a tree diagram known
181 as a dendrogram. The resulting diagram shows the different groups' clustering order and the
182 association measure's value, also known as the fusion level. The fusion level was defined for
183 obtaining seven clusters corresponding to the seven classes. Finally, the number of correctly grouped
184 sample measurements in each cluster was determined, and the percentage of clustering was
185 calculated as: Number of correctly grouped samples / Total number of samples * 100.

186 These steps were repeated for each preprocessing method scenario, and the results obtained were
187 compared using the percentage of samples correctly grouped as a figure of merit to determine the
188 effectiveness of the preprocessing methods evaluated.

189 All the analyses were conducted with the PLS_Toolbox version 8.8 (Eigenvector Research Inc.,
190 Wenatchee, WA, USA) for MATLAB version R2019b (MathWorks, Natick, MA, USA).

191 **2.4 Results comparison**

192 To validate the results obtained with the methodology proposed in this study, the Wilks' lambda was
193 used as a comparative criterion. This criterion evaluates how well the data set classes are separated
194 by calculating a ratio involving between-class and within-class variances. Several versions of the
195 Wilks' lambda exist. In this article, the ratio of the between-class variance over the total variance was
196 used. This ratio varies between 0 and 1, where 0 means that all the classes are superimposed, and 1
197 means that all the classes are perfectly separated.

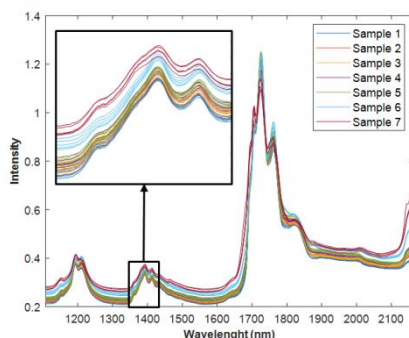
198 **3 Results and discussion**

199 This section shows the results obtained from the application of the proposed methodology. A
200 comprehensive description of its application for the case study of reproducibility (Case 1) is
201 presented. However, only the main results from the case studies of repeatability, temperature effect,

202 and the combination of all cases are showed. Finally, the validation and the advantages of the
203 proposed methodology are offered.

204 3.1 Case 1 - Reproducibility

205 For the reproducibility case, 35 spectra were used (5 spectra for each sample, see section 2.2). Figure
206 2 shows the raw spectra over the entire spectral range used. At first glance it can be observed the
207 variability of spectra due to the physicochemical nature of the sample, showing, with some
208 exceptions, a trend consistent with the properties reported in Table 1; i.e., the spectra acquired on
209 the sample with the lowest density (sample 1 - dark blue color) is at the bottom of the plot, and the
210 spectra acquired on the sample with the highest density (sample 7 - red color) is at the top. This
211 observation becomes more evident in the black square of Figure 2 where the spectra are magnified
212 over a defined range (1340nm - 1430nm). In this same figure it is possible to visualize that the
213 reproducibility measurements made on a single sample generate a variability that has a similar
214 behavior to the variability caused by multiplicative effects, which can impact the final results
215 obtained.

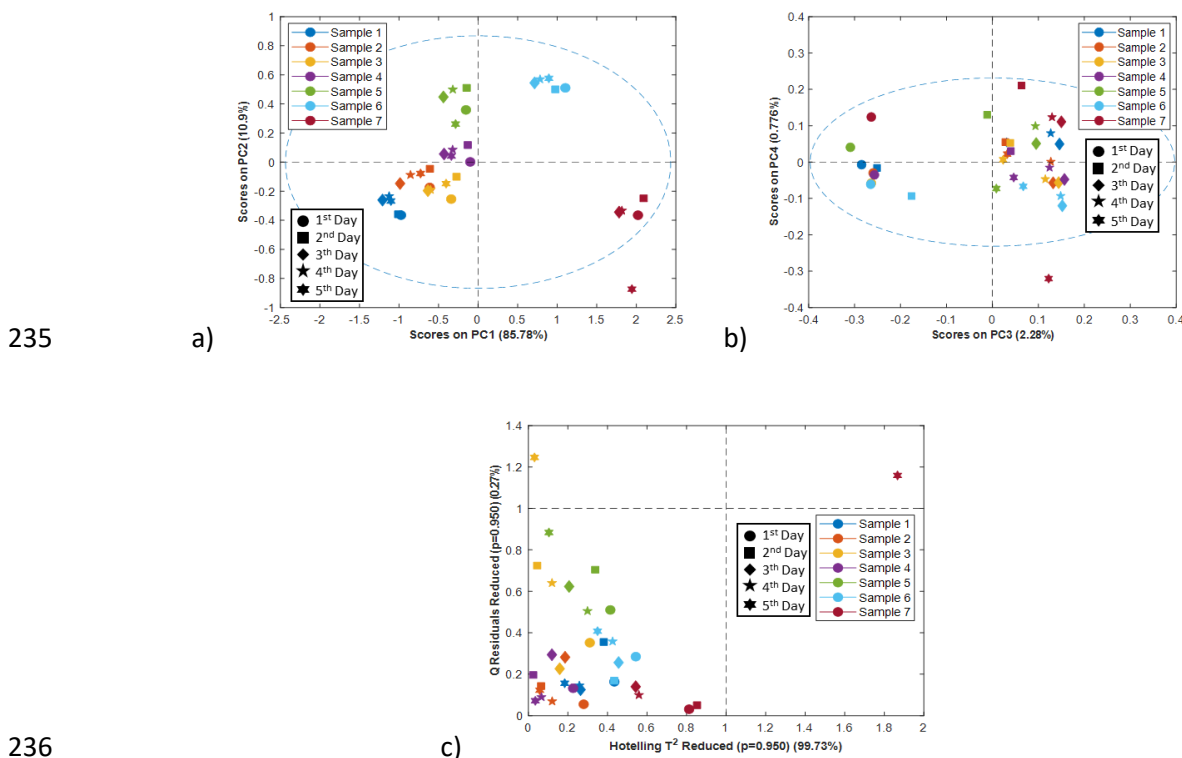


216
217 **Figure 2 – 35 raw spectra used in Case 1 evaluation over the entire spectral range. Legend: Color → samples analyzed.**
218 **Black square → 35 raw spectra used in Case 1 evaluation magnified over the 1340nm - 1430nm spectral range**

219 In this first case, the proposed methodology application shows the impact of the variability in NIR
220 spectra acquisition caused by reproducibility measurements as well as the effectiveness of
221 preprocessing methods in minimizing this spectral variability.

222 Firstly, an exploratory analysis using PCA was performed to visualize the similarity/dissimilarity
223 among the 5 spectra (1 per day) acquired for each of the 7 hydrocarbon samples analyzed. The data
224 set (35 spectra) was merely centered. The first two components (PC1 and PC2) explained 96.7% of
225 the data set total variance. From the score plot of the first two components (see Figure 3a), it can be
226 seen that all 5 measurements of samples 1 to 6 are grouped in a consistent pattern regarding to the
227 variance within classes. However, there are measurements of different samples that intersect with

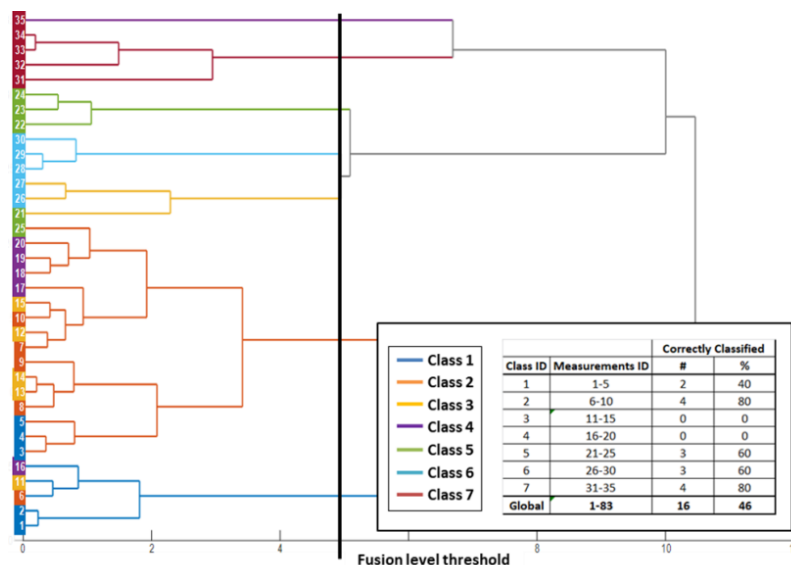
228 each other (between-class variance) (see Figure 3a-b), and hence their clustering can be influenced.
 229 Additionally, from these score plots, it can also be observed that the measurement of sample 7 on
 230 the fifth day is relatively distant from the other measurements of this sample. This could mean the
 231 presence of possible outliers in the dataset. Figure 3c shows the Hotelling's T^2 and Qresidual scores
 232 for the dataset. It can be observed that the same measurement identified previously (Sample 7, 5th
 233 acquisition day) was found to be above the threshold of both tests. Therefore, this measurement can
 234 be confirmed as an outlier.



237 **Figure 3 – a) Score plot of PC1 & PC2 for Case 1 with centered spectra, b) Score plot of PC3 & PC4 for Case 1 with centered**
 238 **spectra, c) Reduced Qresidual & Hotelling's T^2 for Case 1 using 4 PCs with centered spectra. Legend: Shapes → acquisition**
 239 **day. Color → samples analyzed (classes)**

240 The scores of the first 4 principal components yielded by the PCA analysis (99% explained variance)
 241 were used to perform the HCA analysis. Figure 4 shows the dendrogram obtained from the HCA,
 242 where a fusion level (black line) for obtaining seven clusters corresponding to the seven hydrocarbon
 243 samples is defined. This Figure also shows the correct grouping percentage achieved for each class.
 244 From the table embedded in the Figure, it can be observed that no sample achieves an accurate
 245 grouping of all its measurements. Samples 2 and 7 (orange and red classes) are the classes with the
 246 highest correct grouping percentage (4 out of 5 for 80%), while classes 3 and 4 (yellow and purple
 247 classes) do not have any correctly grouped measurements. Furthermore, measurement 35 that
 248 belongs to class 7 (red class) and identified in the previous steps as a potential outlier is the only

249 measurement grouped in class 4 (purple class). Based on these results, it can be considered that this
 250 measurement has no similarity with any of the other 34 measurements, which confirms its outlier
 251 status.

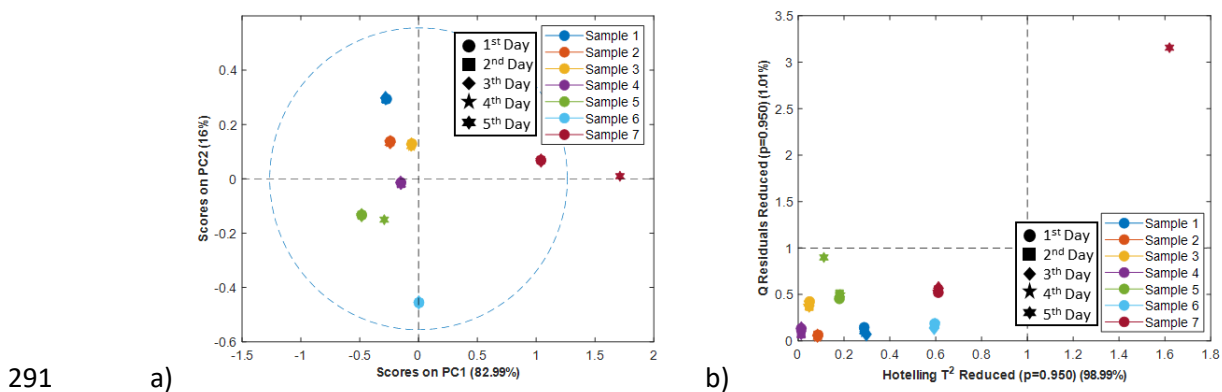


252
 253 **Figure 4 – Dendrogram for Case 1 (all measurements) with centered spectra. Legend: Color samples measured (classes).**
 254 **Table: correct grouping percentage for each class**

255 In order to compensate the spectral variability caused by the lack of reproducibility and hence
 256 achieve a better clustering of all the classes, the use of appropriate preprocessing methods is
 257 needed. Nine preprocessing methods and different combinations of them were evaluated. Figure
 258 App 1, which can be found in the article's supplementary section, summarizes the correct grouping
 259 percentage results for each preprocessing scenario evaluated in the 4 cases. From this Figure, it can
 260 be seen that for Case 1 (diamond shape - blue color), none of the evaluated scenarios reaches a
 261 correct grouping percentage of 100%, implying that none of the scenarios achieved the total
 262 reduction of the spectral variability generated by the reproducibility measurements. With a correct
 263 grouping of 85%, the EMSC was the most effective preprocessing method scenario. Figure 5 shows
 264 for this scenario the score plot of the first two principal components of the PCA analysis and the
 265 Qresidual & Hotelling's T^2 results achieved using 2 PCs. It can be seen that all measurements are
 266 consistently grouped and are within the threshold of the two tests, except for measurement 35
 267 (potential outlier identified). From the HCA dendrogram and the table of its corresponding correct
 268 grouping percentage (see Figure 6), it could be observed that this measurement is still wrongly
 269 grouped as the unique measurement in class 2 (orange class). From these results, it can be assumed
 270 that all other measurements could be correctly grouped without this misgrouped measurement.
 271 Therefore, measurement 35 was removed from the data set, and the preprocessing methods
 272 scenarios were re-evaluated to confirm this assumption.

273 Figure App 1 (diamond shape – red color) shows that by removing measurement 35 from the data
 274 set, correct grouping of all measurements is possible in 6 scenarios (EMSC, AWLS-B+MSC, AWLS-
 275 B+SNV, AWLS-B+VSN, SG-D+SNV, and MSC+AWLS-B). However, only one scenario uses a single
 276 method, which is the EMSC. In order to prevent loss of relevant information, the use of a minimum
 277 number of methods in the data preprocessing is generally recommended. Therefore, in this case,
 278 EMSC could be selected as the most efficient preprocessing method scenario to reduce the spectral
 279 variability due to the lack of reproducibility (see HCA dendrogram in Figure App 2 in the
 280 supplementary section).

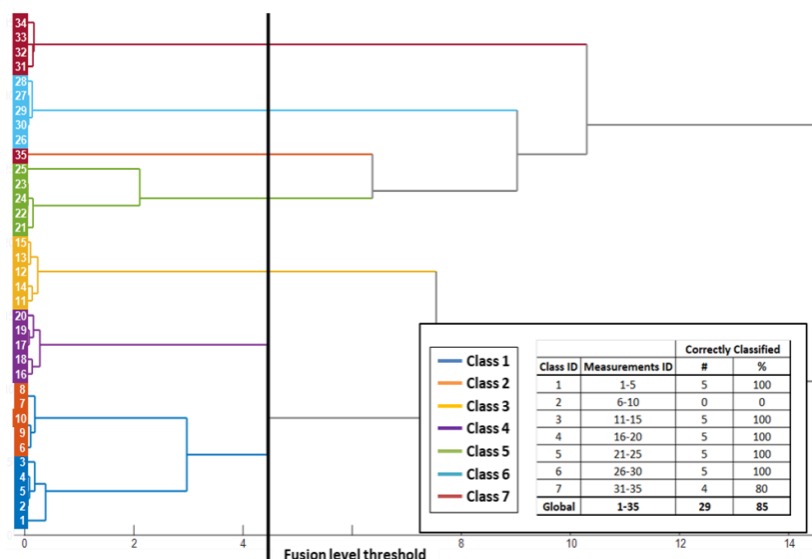
281 Figure App 1 also shows that both the order of use and the combination of preprocessing methods
 282 can influence the final result. An example of this statement can be seen by comparing the scenarios
 283 from Case 1 without the measurement 35, where the NW-D method was combined with the MSC,
 284 PQN, SNV, and VSN methods. When the NW-D method (filtering category) is used before applying
 285 the other preprocessing methods (normalization category), the correct clustering is lower (about
 286 30%) compared to when using the normalization preprocessing methods before the NW-D method. It
 287 can also be found by comparing the same scenarios that using complementary preprocessing
 288 methods does not always yield better correct clustering results than using a single method (NW-
 289 D+PQN → 38% Vs NW-D → 47%). This is an important consideration when using more than one
 290 preprocessing method.



291 a) Score plot of PC1 & PC2 for Case 1 with EMSC preprocessing, b) Reduced Qresidual & Hotelling's T² for Case
 292 1 using 2 PCs with EMSC preprocessing. Legend: Shapes → day of measurement. Color → samples measured (classes)
 293

294 The results obtained in this case show the proposed discrimination methodology's ability to evaluate
 295 the effectiveness of preprocessing methods in minimizing spectral variability in NIR measurements
 296 due to the lack of reproducibility. Moreover, it is worth mentioning that these results also illustrate
 297 the versatility of the proposed methodology for detecting potential anomalous data (outliers) caused
 298 by possible errors in spectrum acquisition.

299 As mentioned before, no detailed description for cases 2, 3, and 4 are presented; only their main
 300 results are shown. The detailed results of these cases are shown in the article's supplementary
 301 section.

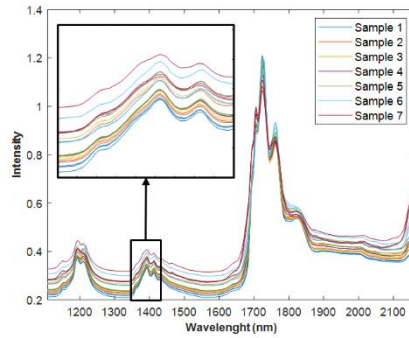


302
 303 **Figure 6 – Dendrogram for Case 1 (all measures) with EMSC preprocessing. Legend: Color samples measured (classes).**
 304 **Table: correct grouping percentage for each class**

305 3.2 Case 2 – Repeatability

306 Figure 7 shows the raw spectra used in the analysis of the variability caused by repeatability
 307 measurements. Analogous to case 1, the spectral variability generated by the physicochemical nature
 308 of the sample can be observed, presenting the same relationship (trend) with the properties
 309 reported in Table 1. However, the black square in Figure 7 shows that the variability caused by
 310 repeatability measurements seems to present a similar behavior to the variability generated by the
 311 combination of two different effects (additive and multiplicative), making the spectral differences of
 312 a single sample more evident in comparison with case 1.

313 In this second case, a total of 21 spectra (7 samples performed in triplicate) were analyzed to
 314 demonstrate the proposed methodology's ability to evaluate the effectiveness of preprocessing
 315 methods in minimizing unwanted spectral variability due to the lack of repeatability. The correct
 316 grouping percentage achieved from all the preprocessing methods scenarios is shown in Figure App 1
 317 (squared shape – orange color).



318

319 **Figure 7 – 21 raw spectra used in Case 2 evaluation over the entire spectral range. Legend: Color → samples analyzed.**

320 **Black square → 21 raw spectra used in Case 2 evaluation magnified over the 1340nm - 1430nm spectral range**

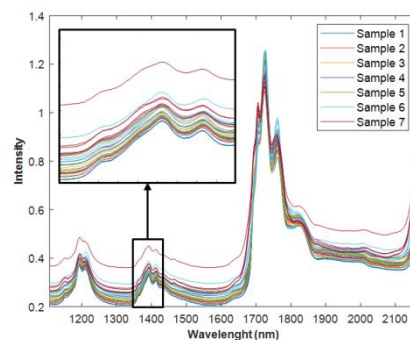
321 In the scenario that is assumed that the variability generated in the repeatability measurements do
 322 not have a significant impact on the grouping of sample measurements, that is, that no additional
 323 preprocessing methods are needed besides the data centering, only 48% of the measurements were
 324 correctly grouped (see Figure App 1 (squared shape – orange color) mean center scenario).
 325 Therefore, it can be preliminarily concluded that preprocessing methods to reduce spectral variations
 326 due to repeatability are needed. From the PCA analysis (data not shown) of this scenario (centered
 327 data), it could be determined that the first repetition of each sample differs significantly from its
 328 second and third repetitions. However, all of them were within the Qresidual and Hotelling's T^2 test
 329 thresholds (data not shown). Therefore, they cannot be considered as anomalous data (outliers).

330 From all the evaluated scenarios, fourteen achieved a correct grouping of 100% (AWLS-B+MSC,
 331 AWLS-B+SNV, Dtd+MSC, Dtd+PQN, Dtd+SNV, EMSC+PQN, EMSC+SNV, MSC+AWLS-B, MSC+Dtd,
 332 PQN+EMSC, SNV+AWLS-B, SNV+Dtd, SNV+EMSC, and VSN). Among these scenarios, only one uses a
 333 single method, which is VSN (see HCA dendrogram in Figure App 4). Therefore, VSN could be selected
 334 as the most effective preprocessing method to minimize the spectral variability due to the lack of
 335 repeatability. As in Case 1, Figure App 1 shows that the order of use of the preprocessing methods
 336 affects the correct grouping result. Comparing the same scenarios as in Case 1, it can be reaffirmed
 337 that the results are more promising when the filtering methods are applied after the normalization
 338 methods.

339 From these results, it can be concluded that spectral variability due to repeatability has a lesser
 340 impact than those generated by reproducibility. Nevertheless, the proposed methodology
 341 demonstrated that the two cases' variability could be entirely compensated using an appropriate
 342 data preprocessing strategy.

343 3.3 Case 3 – Temperature effect

344 As previously mentioned in the introduction of the manuscript, sample temperature is one of the
345 factors having significant impact on the NIR spectra acquisition. Figure 8 shows that the spectral
346 variability generated by the sample temperature presents a behavior similar to the multiplicative
347 effect. However due to the absorbance shift caused by the temperature increase, which prevents
348 having a direct relationship between this parameter and the height of the acquired spectra, the
349 spectral difference presents a nonlinear growth⁴⁵. This can be corroborated in the black square of
350 Figure 8, where it is observed that with a temperature variation greater than 15°C the spectral
351 variability is more evident than when the delta in temperature is less than 15°C. This non-linear
352 impact of the sample temperature on the spectrum acquisition could limit the performance of the
353 different preprocessing methods evaluated.



354

355 **Figure 8 – a) 35 raw spectra used in Case 3 evaluation over the entire spectral range. Legend: Color → samples analyzed.**
356 **Black square → 35 raw spectra used in Case 3 evaluation magnified over the 1340nm - 1430nm spectral range**

357 In this third case, spectra acquired at 5 different sample temperatures (35 spectra) were analyzed to
358 find the most effective preprocessing scenario to reduce undesired spectral variability due to the
359 sample temperature. The correct grouping percentage achieved from all the evaluated preprocessing
360 strategies is shown in Figure App 1 (triangle shape – green color).

361 From Figure App 1, it can be seen that if no other preprocessing method than data centering is
362 applied, the percentage of correctly grouped measurements is 49%. This result reflects, as expected,
363 the need to apply preprocessing methods to reduce variability caused by sample temperature. The
364 clustering results shown in Figure App 1 reveal that no evaluated preprocessing scenario could
365 entirely compensate the spectral variability due to temperature variations, meaning that the entire
366 accurate measurement grouping was not achieved in any scenario. The best performing scenario is
367 the SG-D+SNV with an accurate grouping percentage of 80%. Although Figure App 1 shows that the
368 best performance scenario does not achieve the correct grouping of all 35 measurements, the results
369 shown in Figure App 5 (scenario SG-D+SNV) show that samples 1, 2, 3, 5, and 6 have an accurate

370 grouping in all their measurements (5 out of 5 = 100%). The class affecting the overall measurement
371 clustering is sample 4 (purple), which does not have any measurements grouped correctly. As in Case
372 1, it could be assumed that the whole misgrouping of sample 4 is due to the presence of some
373 atypical data. However, no measurement was found above the Qresidual and Hotelling's T^2 tests
374 thresholds in any scenario (data not shown).

375 The results analyzed in this case show that sample temperature is a very influential parameter on the
376 spectrum acquisition. Therefore, it is recommended to use a strategy that evaluates this variable's
377 impact more thoroughly for a more efficient solution^{9,45}.

378 **3.4 Case 4 – cases 1, 2 & 3 combined**

379 The parameters causing unwanted spectral variability evaluated in the 3 cases previously described
380 are likely to occur simultaneously, mainly when online NIR measurement is used for real-time data
381 analysis. For this reason, a fourth case was evaluated where the spectral variability generated by
382 these three cases was combined.

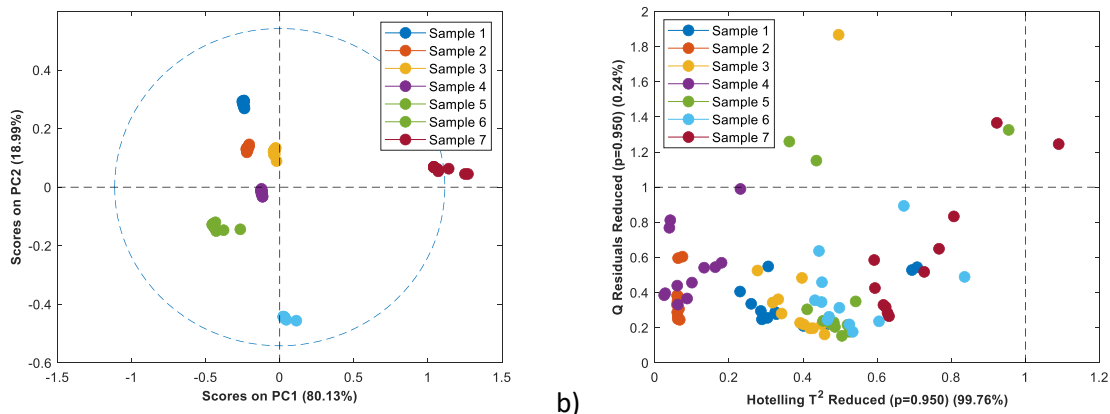
383 The dataset used in this case comprises 84 spectra (35 from reproducibility Case, 21 from
384 repeatability Case, and 28 from sample temperature Case). As in the 3 cases already studied, the
385 Qresidual and Hotelling's T^2 analyses were applied to the centered dataset to determine the possible
386 presence of atypical data (data not shown). The measurement identified as an outlier in Case 1
387 (Sample 7, 5th acquisition day) once again exceeds the thresholds of the two tests. Thus, this
388 measurement was removed, and the methodology proposed in this study was applied to the other
389 83 measurements.

390 In this case, 5 scenarios (EMSC, EMSC+PQN, PQN+EMSC, SNV+EMSC, and VSN+EMSC) had the best
391 performance, where 80 out of 83 measurements were correctly grouped, that is, 96% of correct
392 grouping. All 5 scenarios identified involve the use of EMSC, which, when evaluated individually,
393 yields the same percentage of correct grouping (96%). It could be preliminarily inferred that for this
394 case, the normalization methods do not give any additional improvement, and thus EMSC would be
395 selected as the most effective preprocessing method scheme. Nevertheless, the use of the proposed
396 methodology allows a more detailed analysis to determine at what level each type of variability is
397 minimized in each of the 5 scenarios mentioned. In this way, the most effective preprocessing
398 scheme's determination can be done more conveniently according to the researcher's objective.

399 Figure 9 shows the PCA score plot and the Qresidual & Hotelling's T^2 results for the VSN+EMSC
400 preprocessing method scenario. The three measurements that were not correctly grouped
401 correspond to the measurement at 80°C of samples 3 (highest value of residual Q test) and 4, and the

402 measurement at 75°C of sample 4. From Figure App 1, it is also concluded that the EMSC and VSN
 403 methods have a better performance when grouping the samples measured at different temperatures
 404 in Case 4 (circle shape – purple color, combined cases) than in Case 3 (triangle shape – green color,
 405 sample temperature case).

406 Although the conclusion made previously that no method can entirely compensate the measurement
 407 variability caused by sample temperature is reaffirmed, the reduction of this variability was quite
 408 considerable in delta temperature lower than 20°C.



409 a) b)
 410 **Figure 9 – a) Score plot of PC1 & PC2 for Case 4 with VSN+EMSC preprocessing, b) Reduced Qresidual & Hotelling's T² for**
 411 **Case 4 using 3 PCs with VSN+EMSC preprocessing. Legend: Color → samples measured (classes)**

412 To sum up the results obtained by using the analysis methodology proposed, it could be concluded
 413 that for cases 1 and 2 (reproducibility and repeatability), the variability affecting the measurement
 414 clustering was fully compensated using a single preprocessing method. On the other hand, cases 3
 415 and 4 (sample temperature and combination of cases) needed the combination of two methods, and
 416 still, the correct grouping of all the measurements was not achieved. The sample temperature has a
 417 high impact on the spectrum acquisition. Therefore, it is recommended to use a methodology that
 418 evaluates this variable more thoroughly for a more efficient solution^{9, 45}.

419 **4 Results comparison**

420 In order to validate the consistency and reliability of the proposed approach, the analysis
 421 methodology results were compared with those obtained by Wilks' lambda criterion.

422

423 Table 3 summarizes each case's most relevant results achieved by both the proposed approach and
 424 Wilks' Lambda.

425
426
427

Table 3 % Results comparison using Wilk's Lambda algorithm

Case	Preprocessing method	Methodology (% Grouped/100)	Wilk's Lambda
Case 1 (All Measurements)	Mean Center	0.46	0.95
	EMSC	0.85	0.96
Case 1 (Measurement 35 removed)	Mean Center	0.53	0.68
	EMSC	1.00	0.99
Case 2	Mean Center	0.48	0.69
	VSN	1.00	0.99
Case 3	Mean Center	0.49	0.55
	SavGol + SNV	0.82	0.95
Case 4	Mean Center	0.51	0.66
	VSN+EMSC	0.96	0.99

428

429 The results shown in

430

431 Table 3 validate the approach proposed in this study. It can be seen that the results between the two
432 methodologies are comparable, except for Case 1, including all measurements, when the data have
433 been only mean-centered. In this case, Wilks' Lambda value is close to 1, while the value obtained by
434 the proposed methodology is 0.46. The difference may be attributable to the presence of the outlier
435 identified in Case 1 and how each approach handles this type of data. While the Wilks' lambda
436 criterion assumes that there is no presence of outliers in the analyzed dataset, the methodology used
437 in this study provides a preliminary analysis of the dataset for the identification and removal of
438 possible anomalous data. This premise can be supported by observing that the two approaches'
439 results are comparable when the identified outlier is removed from the dataset (see

440

441 Table 3 - Case 1 (Measurement removed)).

442 Moreover, the proposed methodology provides a more detailed discrimination analysis in
443 comparison with Wilks' lambda criterion. As a way of example, both the proposed approach and the
444 Wilks' lambda results obtained from the individual evaluation of the nine preprocessing methods in
445 Case 1 were compared. From Table 4, it can be seen that Wilks' lambda criterion presents no
446 significant differences in 4 preprocessing methods (SG-D, EMSC, MSC, SNV), which could lead to the
447 conclusion that the 4 methods have equal effectiveness in minimizing the unwanted spectral

448 variability. On the contrary, the proposed methodology allows to identify that out of these 4
 449 methods, two are equally effective in minimizing the unwanted spectral variability (MSC, SNV), the
 450 least effective is the SG-D, and the most effective preprocessing method, which achieves the
 451 maximum compensation of the unwanted spectral variability, is the EMSC. Therefore, the
 452 methodology used in this work could help to select the preprocessing method in a more precise and
 453 reliable way. Finally, the proposed methodology allows identifying measurements and samples that
 454 have been properly or poorly discriminated, information that the Wilks' lambda does not provide as
 455 it is a global measurement.

456

Table 4 Case 1 detailed results comparison

Preprocessing Method	Explored Methodology (% Grouped/100)	Wilk's Lambda
Mean Center	0.53	0.68
AWLS-B	0.65	0.88
NW-D	0.56	0.73
SG-D	0.65	0.98
Dtd	0.65	0.90
EMSC	1.00	0.99
MSC	0.82	0.99
PQN	0.68	0.87
SNV	0.82	0.99
VSN	0.94	0.85

457 **5 Conclusions**

458 The results obtained in this study shows the capacity of the analysis methodology used to assess the
 459 effectiveness of preprocessing methods in reducing the undesired spectral variability of near-infrared
 460 spectroscopy (NIRS) measurements in a more thoughtfully and detailed manner than other
 461 approaches based on the dataset variance analysis such as the Wilks' lambda criterion.

462 In this study, an original strategy not previously reported in literature was proposed to evaluate and
 463 determine the effectiveness of different preprocessing methods in minimizing the unwanted spectral
 464 variability due to parameters related to the continuous and repetitive NIR spectra acquisition such as
 465 repeatability, reproducibility, sample temperature, and the combination of these three parameters.

466 It is essential to stress the twofold benefit of using the proposed methodology. On the one hand, the
 467 detailed discrimination analysis provides a significant aid in determining the most effective data
 468 preprocessing scheme. On the other hand, the methodology provides a preliminary data analysis

469 step for identifying and removing the potential anomalous data from the dataset, thus improving the
470 reliability of the final results.

471 The results obtained using the proposed analysis methodology suggest that the variability caused by
472 repeatability and reproducibility can be fully corrected when using the adequate preprocessing
473 scheme; however, no preprocessing scenario could entirely compensate the unwanted spectral
474 variability caused by the sample temperature. Similarly, the detailed discriminant analysis employed
475 in this study showed that the EMSC preprocessing method presents interesting and promising results
476 in all cases.

477 The preprocessing scheme's ultimate selection should be conducted in a careful manner considering
478 the researcher's objective. The proposed methodology offers an analysis strategy that could help
479 determine the most effective preprocessing scheme more reliably.

480 The conclusions reached in this work promote further optimization and automation of the proposed
481 methodology to improve its implementation in large datasets.

482 The strategy proposed was shown to work for a case study including seven different hydrocarbon
483 samples but can be generally applicable in any study involving spectroscopic information analysis

484 **Acknowledgments**

485 The authors would like to thank IFP Energie Nouvelles for providing the hydrocarbon samples
486 obtained in their HCK pilot plant reactors and the facilities for spectra acquisition and data analysis.
487 Thanks also go to Axel One for providing the spectrometer for the NIR spectra acquisition.

488 **6 Declaration of conflicting interests**

489 The Author(s) declare(s) that there is no conflict of interest

490

491 **References**

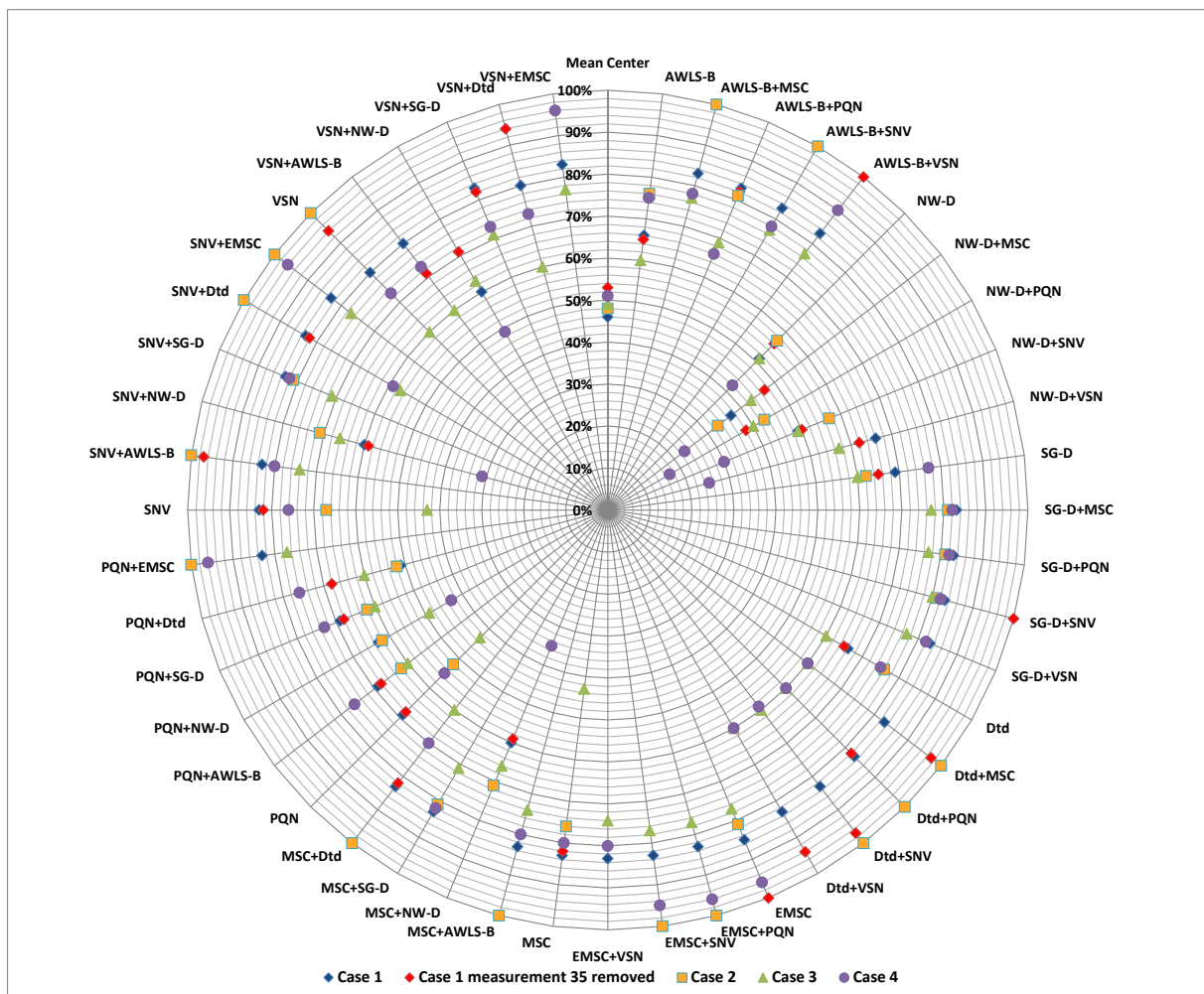
- 492 1. J. Fernández-Navales, M.-I. López, M.-T. Sánchez, J.-A. García and J. Morales, "A feasibility study on
493 the use of a miniature fiber optic NIR spectrometer for the prediction of volumic mass and reducing
494 sugars in white wine fermentations", *Journal of Food Engineering*, **89**, 3 (2008).
- 495 2. W. Saeys, N. Nguyen Do Trong, R. van Beers and B.M. Nicolai, "Multivariate calibration of
496 spectroscopic sensors for postharvest quality evaluation: A review", *Postharvest Biology and
497 Technology*, **158** (2019).
- 498 3. M. Blanco and A. Peguero, "Analysis of pharmaceuticals by NIR spectroscopy without a reference
499 method", *TrAC Trends in Analytical Chemistry*, **29**, 10 (2010).

- 500 4. T. de Beer, A. Burggraeve, M. Fonteyne, L. Saerens, J.P. Remon and C. Vervaet, "Near infrared and
501 Raman spectroscopy for the in-process monitoring of pharmaceutical production processes",
502 *International journal of pharmaceuticals*, **417**, 1-2 (2011).
- 503 5. R.M. Balabin, E.I. Lomakina and R.Z. Safieva, "Neural network (ANN) approach to biodiesel
504 analysis: Analysis of biodiesel density, kinematic viscosity, methanol and water contents using near
505 infrared (NIR) spectroscopy", *Fuel*, **90**, 5 (2011).
- 506 6. Zanier-Szydłowski N., Quignard A., Baco F., Biguerd H., Carpot L., "Control of Refining Processes on
507 Mid-Distillates by Near Infrared Spectroscopy", *Oil & Gas Science and Technology - Rev. IFP* (1999).
- 508 7. P.R. Wahl, I. Pucher, O. Scheibelhofer, M. Kerschhaggl, S. Sacher and J.G. Khinast, "Continuous
509 monitoring of API content, API distribution and crushing strength after tableting via near-infrared
510 chemical imaging", *International journal of pharmaceuticals*, **518**, 1-2 (2017).
- 511 8. R.R. de Oliveira, R.H.P. Pedroza, A.O. Sousa, K.M.G. Lima and A. de Juan, "Process modeling and
512 control applied to real-time monitoring of distillation processes by near-infrared spectroscopy",
513 *Analytica chimica acta*, **985** (2017).
- 514 9. F. Chauchard, J.M. Roger and V. Bellon-Maurel, "Correction of the temperature effect on near
515 infrared calibration—application to soluble solid content prediction", *Journal of Near Infrared
516 Spectroscopy*, **12** (2004).
- 517 10. B. Igne, M.N. Hossain, J.K. Drennen and C.A. Anderson, "Robustness Considerations and Effects of
518 Moisture Variations on near Infrared Method Performance for Solid Dosage Form Assay", *Journal of
519 Near Infrared Spectroscopy*, **22**, 3 (2014).
- 520 11. J.M. Betz, P.N. Brown and M.C. Roman, "Accuracy, precision, and reliability of chemical
521 measurements in natural products research", *Fitoterapia*, **82**, 1 (2011).
- 522 12. A.C. Olivieri and N.M. Faber, "Validation and Error", in: *Comprehensive chemometrics. Chemical
523 and biochemical data analysis*, Ed by S.D. Brown, Elsevier, Amsterdam, pp. 91–120 (2009).
- 524 13. ISO, "Measurement management systems - Requirements for measurement processes and
525 measuring equipment", 10012:2003. UNE.
- 526 14. Florian Wülfert,[†] Wim Th. Kok,[†] and, and Age K. Smilde*,[†], "Influence of Temperature on
527 Vibrational Spectra and Consequences for the Predictive Ability of Multivariate Models", *Analytical
528 Chemistry*, **70** (1998).
- 529 15. Hideyuki Abe, Chie Iyo, and Sumio Kawano, "A Study on the Universality of a Calibration with
530 Sample Temperature Compensation", *Journal of Near Infrared Spectroscopy*, **8** (2000).
- 531 16. W.G. Hansen, S.C.C. Wiedemann, M. Snieder, and V.A.L. Wortel, "Tolerance of near Infrared
532 Calibrations to Temperature Variations; A Practical Evaluation", *Journal of Near Infrared
533 Spectroscopy*, **8** (2000).
- 534 17. Jasem M. Al-Besharah/Saed A. Akashah/Clive J. Mumford, "The effect of temperature and
535 pressure on the viscosities of crude oils and their mixtures", *Industrial & Engineering Chemistry
536* (1989).
- 537 18. P. Luo and Y. Gu, "Effects of asphaltene content on the heavy oil viscosity at different
538 temperatures", *Fuel*, **86**, 7-8 (2007).
- 539 19. R. Payri, F.J. Salvador, J. Gimeno and G. Bracho, "The effect of temperature and pressure on
540 thermodynamic properties of diesel and biodiesel fuels", *Fuel*, **90**, 3 (2011).
- 541 20. Å. Rinnan, F. van den Berg and S.B. Engelsen, "Review of the most common pre-processing
542 techniques for near-infrared spectra", *TrAC Trends in Analytical Chemistry*, **28**, 10 (2009).
- 543 21. Abraham. Savitzky/M. J. E. Golay, "Smoothing and Differentiation of Data by Simplified Least
544 Squares Procedures.", *Analytical Chemistry*, **36** (1964).

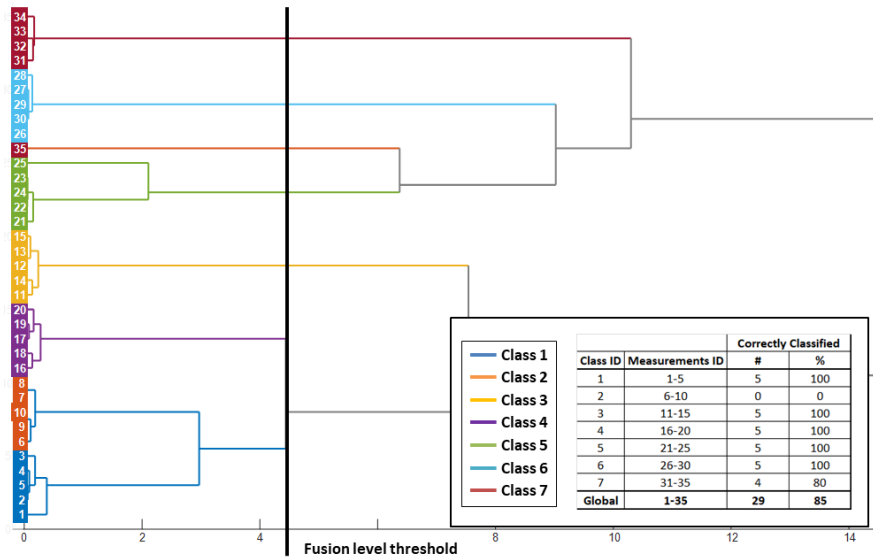
- 545 22. M. Harald and Edward Stark, "Extended multiplicative signal correction and spectral interference
546 subtraction: new preprocessing methods for near infrared spectroscopy", *Journal of Pharmaceutical*
547 *& Biomedical Analysis*, **9** (1991).
- 548 23. Barnes, R. J., Dhanoa, M. S., & Lister, S. J., "Standard Normal Variate Transformation and De-
549 Trending of Near-Infrared Diffuse Reflectance Spectra.", *Applied Spectroscopy*, **43** (1989).
- 550 24. G. Rabatel, F. Marini, B. Walczak and J.-M. Roger, "VSN: Variable sorting for normalization",
551 *Journal of Chemometrics*, **34**, 2 (2020).
- 552 25. A. Gholizadeh, L. Borůvka, M.M. Saberioon, J. Kozák, R. Vašát and K. Němeček, "Comparing
553 different data preprocessing methods for monitoring soil heavy metals based on soil spectral
554 features", *Soil & Water Res.*, **10**, No. 4 (2016).
- 555 26. Y. Jiao, Z. Li, X. Chen and S. Fei, "Preprocessing methods for near-infrared spectrum calibration",
556 *Journal of Chemometrics*, **130**, 1 (2020).
- 557 27. C.D. Brown, L. Vega-Montoto and P.D. Wentzell, "Derivative Preprocessing and Optimal
558 Corrections for Baseline Drift in Multivariate Calibration", *Appl Spectrosc*, **54**, 7 (2000).
- 559 28. J. Gerretzen, E. Szymańska, J.J. Jansen, J. Bart, H.-J. van Manen, E.R. van den Heuvel and L.M.C.
560 Buydens, "Simple and Effective Way for Data Preprocessing Selection Based on Design of
561 Experiments", *Analytical Chemistry*, **87**, 24 (2015).
- 562 29. O. Devos and L. Duponchel, "Parallel genetic algorithm co-optimization of spectral pre-processing
563 and wavelength selection for PLS regression", *Chemometrics and Intelligent Laboratory Systems*, **107**,
564 1 (2011).
- 565 30. F. Allegrini and A.C. Olivieri, "An integrated approach to the simultaneous selection of variables,
566 mathematical pre-processing and calibration samples in partial least-squares multivariate
567 calibration", *Talanta*, **115** (2013).
- 568 31. D.L. Luthria, S. Mukhopadhyay, L.-Z. Lin and J.M. Harnly, "A comparison of analytical and data
569 preprocessing methods for spectral fingerprinting", *Applied Spectroscopy*, **65**, 3 (2011).
- 570 32. S. S Wilks, "The large-sample distribution of the likelihood ratio for testing composite
571 hypotheses", *Mathematical Statistics* (1962).
- 572 33. L. Rokach and O. Maimon, "Clustering Methods". Springer-Verlag.
- 573 34. Maureen Bricker, Vasant Thakkar, John Petri, "Hydrocracking in Petroleum Processing", in:
574 *Handbook of Petroleum Processing 2014*, pp. 1–35.
- 575 35. J.G. Speight, "Hydrocracking", in: *The Refinery of the Future*, Elsevier, pp. 275–313 (2011).
- 576 36. ASTM D1218 - 12, "Standard Test Method for Refractive Index and Refractive Dispersion of
577 Hydrocarbon Liquids", <https://www.astm.org/Standards/D1218.htm>.
- 578 37. ASTM D2887 - 19ae1, "Standard Test Method for Boiling Range Distribution of Petroleum
579 Fractions by Gas Chromatography", <https://www.astm.org/Standards/D2887.htm>.
- 580 38. K.H. NORRIS and P.C. WILLIAMS, "Optimization of mathematical treatments of raw near-infrared
581 signal in the measurement of protein in hard red spring wheat. I. Influence of particle size.", *Cereal*
582 *Chemistry*, **61** (Mar. 1984).
- 583 39. H. Martens and T. Naes, *Multivariate calibration*. Wiley, Chichester (1989).
- 584 40. F. Dieterle, A. Ross, G. Schlotterbeck and H. Senn, "Probabilistic quotient normalization as robust
585 method to account for dilution of complex biological mixtures. Application in 1H NMR
586 metabonomics", *Analytical Chemistry*, **78**, 13 (2006).
- 587 41. S. Wold, K. Esbensen and P. Geladi, "Principal component analysis", *Chemometrics and Intelligent*
588 *Laboratory Systems*, **2**, 1-3 (1987).

589 42. L.S. Chen, D. Paul, R.L. Prentice and P. Wang, "A regularized Hotelling's T2 test for pathway
590 analysis in proteomic studies", *Journal of the American Statistical Association*, **106**, 496 (2011).
591 43. K. Sasirekha and P. Baby, "Agglomerative Hierarchical Clustering Algorithm- A Review",
592 *International Journal of Scientific and Research Publications*, **3** (2013).
593 44. F. Murtagh and P. Legendre, "Ward's Hierarchical Agglomerative Clustering Method: Which
594 Algorithms Implement Ward's Criterion?", *J Classif*, **31**, 3 (2014).
595 45. J.-M. Roger, F. Chauchard and V. Bellon-Maurel, "EPO-PLS external parameter orthogonalisation
596 of PLS application to temperature-independent measurement of sugar content of intact fruits",
597 *Chemometrics and Intelligent Laboratory Systems*, **66**, 2 (2003).
598

599 **Appendix**
600



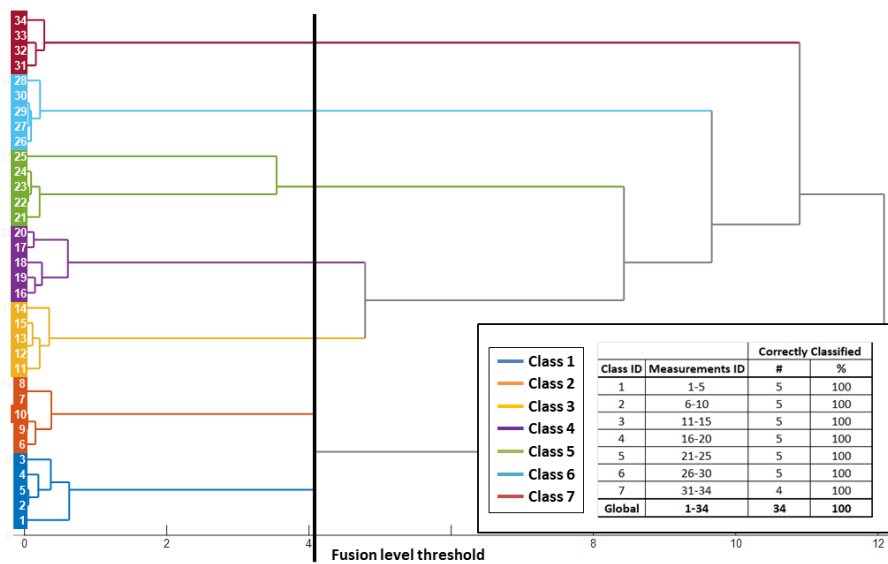
601 **Figure App 1 Comparative summary of the effectiveness of the preprocessing scenarios evaluated in the 4 case studies.**
602 **Legend :** Shapes and color → Case studies. Description : Circumferential lines → Percentage of correct grouping (0%
603 Center - 100% outer line). Radial lines → Evaluated preprocessing scenario, from left to right the order of use of the
604 methods.
605



606
607
608

Figure App 2 Dendrogram for Case 1 (all measurements) with EMSC preprocessing. Legend: Color samples measured (classes). Table: correct grouping percentage for each class

609



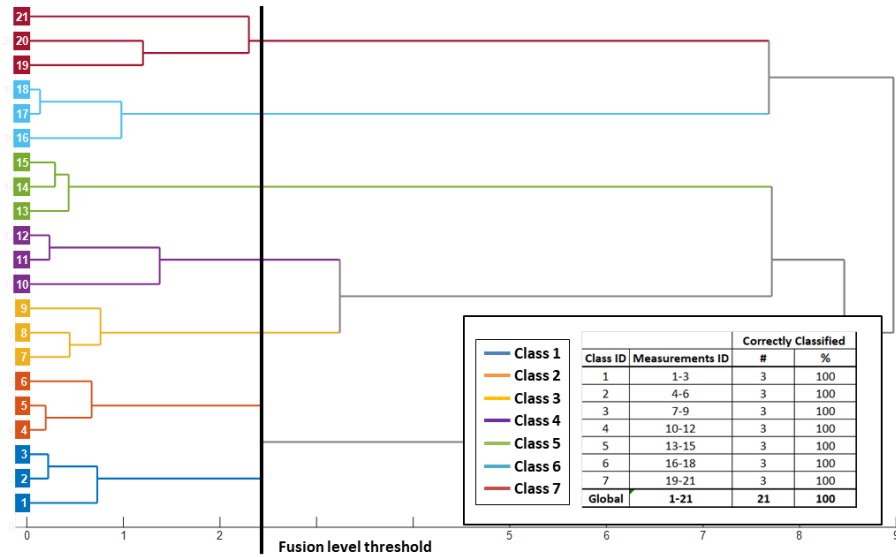
610

611
612

Figure App 3 Dendrogram for Case 1 (removing measurement 35) with EMSC preprocessing. Legend: Color samples measured (classes). Table: correct grouping percentage for each class

613

614



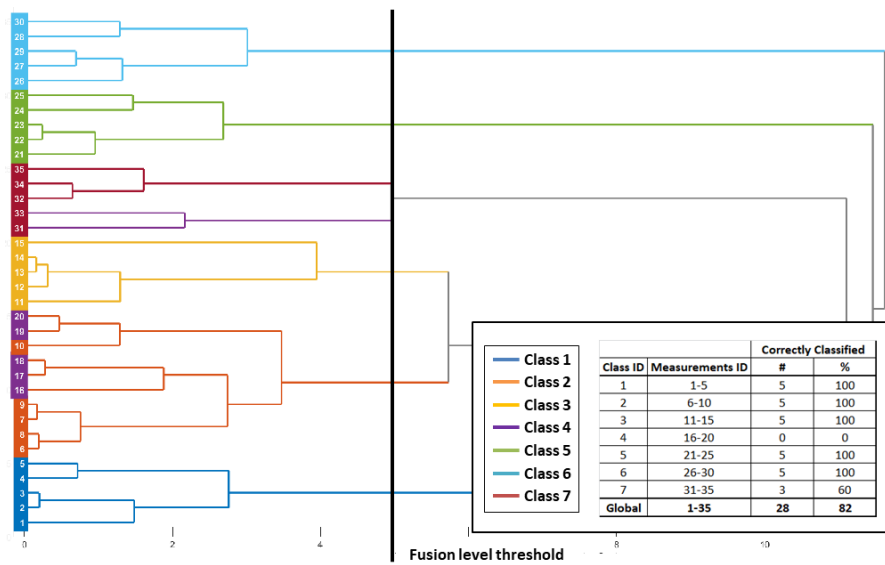
615

616

617

Figure App 4 dendrogram for Case 2 with VSN preprocessing. Legend: Color samples measured (classes). Table: correct grouping percentage for each class

618



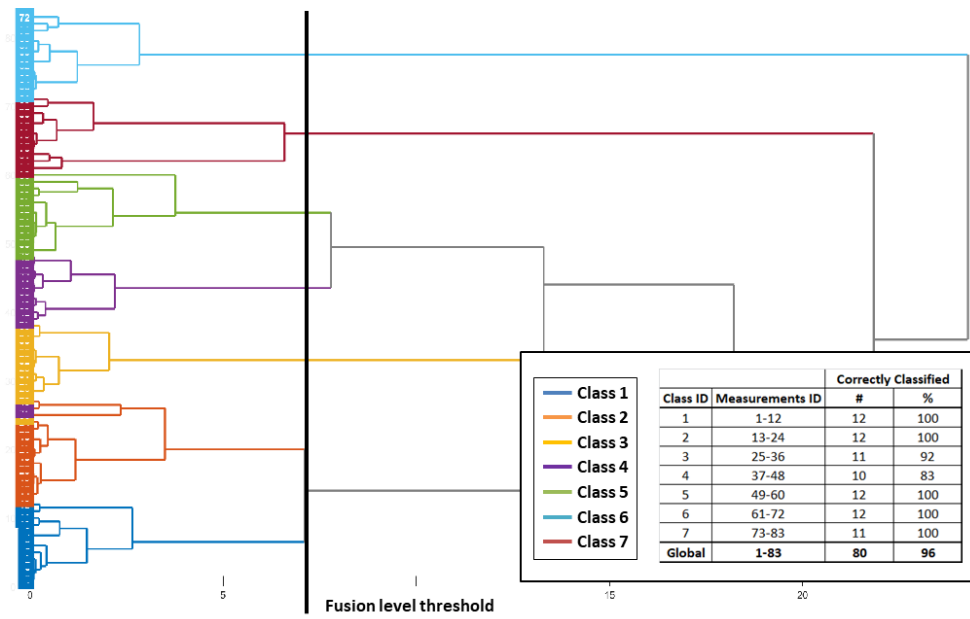
619

620

621

Figure App 5 dendrogram for Case 3 with SG-D+SNV preprocessing. Legend: Color samples measured (classes). Table: correct grouping percentage for each class

622



623

624

625

Figure App 6 dendrogram for Case 4 with VSN+EMSC preprocessing. Legend: Color samples measured (classes). Table: correct grouping percentage for each class