Electronic Supplementary Material (ESI) for Analytical Methods.

SUPPLEMENTARY INFORMATION

COMPARISON OF DIFFERENT ANALYTICAL CLASSIFICATION SCENARIOS: APPLICATION FOR GEOGRAPHICAL ORIGIN OF EDIBLE PALM OIL BY STEROLIC (NP)HPLC FINGERPRINTING

Estefanía PÉREZ-CASTAÑO [a,*], Cristina RUIZ-SAMBLÁS [a], Santiago MEDINA-RODRÍGUEZ [a,b], Verónica QUIRÓS-RODRÍGUEZ [a], Ana M. JIMÉNEZ-CARVELO [a], Lucía VALVERDE-SOM [a], Antonio GONZÁLEZ-CASADO [a], Luis CUADROS-RODRÍGUEZ [a]

[a] Department of Analytical Chemistry, University of Granada, c/ Fuentenueva, s.n. E-18071 Granada, Spain.

[b] Department of Signal Theory, Networking and Communications, CITIC-UGR, University of Granada, c/ Periodista Rafael Gómez, E-18071 Granada, Spain.

* Corresponding author: phone: +34 958240797; fax: +34 958243328; email: stefani@ugr.es.

1. Description of the MEDINA function for preprocessing of chromatographic data

In this work, an efficient and user-friendly algorithm for pre-processing of chromatographic data is reported, which can be used to improve the quality of complex chromatographic signals. It provides more reliable information on the content of the chromatograms and facilitates better performing classification/prediction models for the analysis of such complex samples.

Preprocessing of chromatographic data matrices is carried out using a home-made MATLAB function, named "MEDINA" (version 07), programmed by our research group. This function makes use of some of the functions contained in the MATLAB's Bioinformatics Toolbox™ software to improve the quality of raw chromatographic data, and also the "icoshift" algorithm (version 1.2) for solving signal alignment problems in chromatographic data.

The algorithm developed implements the following preprocessing options: *(1)* Selection of the interval of interest from chromatographic data matrices, discarding the analysis of variables outside this range; *(2)* Decimation of the raw chromatographic data with the MATLAB "resample" function. This function makes possible to resample the signal into a more manageable chromatographic data vector, preserving the information contained in the chromatogram; *(3)* De-noising and smoothing the chromatographic signal with peaks using a least-squares digital polynomial filter (i.e., a Savitzky-Golay filter). For it, the MATLAB "mssgolay" function is used. This filter smooths the raw noisy signal data preserving the sharpness (or high-frequency components) of the peaks in the chromatogram; *(4)* Normalization of the chromatographic data. The method scales the intensity values of each chromatogram to a specific intensity value (i.e., the intensity value for a given reference peak); *(5)* Baseline correction of the chromatographic data. The MATLAB "msbackadj" function is used to estimate a low-frequency baseline, which is hidden among high-frequency noise and signal peaks. It then subtracts the baseline from the chromatograms; *(6)* Alignment of the chromatographic profiles using the "icoshift" algorithm. This alignment algorithm is based on correlation shifting of spectral intervals, and employs a fast Fourier transform (FFT) engine that aligns all chromatograms simultaneously. Additionally, the algorithm developed also allows the alignment of the chromatographic data using the MATLAB "msalign" function. This function aligns the set of chromatograms to a set of reference peaks given. These reference peaks can be chosen manually or automatically (based on an intensity threshold pre-selected or calculated automatically). Once the chromatographic data preprocessing is

carried out, it is then possible to use classification and statistical learning tools to create classifiers.

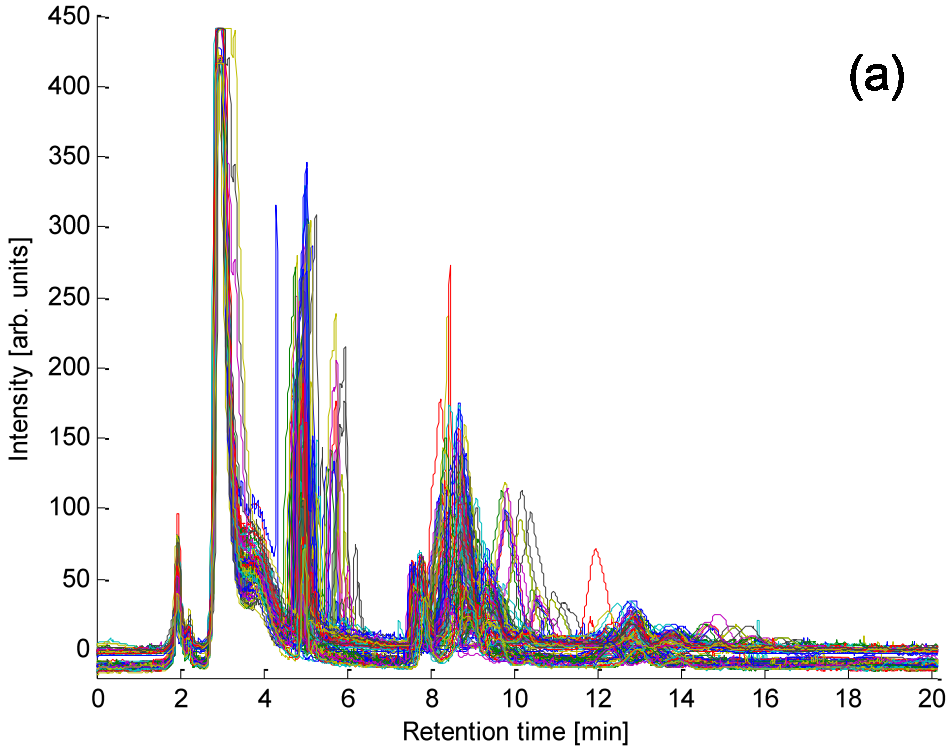
The efficacy of the algorithm has been demonstrated on a set of 102 chromatographic fingerprints of edible palm oil samples obtained from two different normal-phase HPLC systems. Figure SI-1 summarizes the preprocessing steps of the “MEDINA” function described above.

Figure SI-1

Figure illustrates the effect of chromatographic data processing on a set of 102 chromatographic fingerprints of edible palm oil obtained from the CAD detector. (a) superposition of raw chromatograms; (b) raw chromatograms after selecting the data set of interest (retention times between 6.7 and 16 min); (c) chromatograms after decimation process (a decimation factor equal to 2 was used in this example) and smoothing with a Savitzky-Golay filter; (d) chromatograms after normalization (the method scales the intensity values of each chromatogram to a specific intensity value (in this example, the maximum intensity value of the reference peak located between 7.3 and 9.0 min); (e) chromatograms after baseline shifting removal (for this, the MATLAB "msbackadj" function was used); (f) chromatograms after subsequent “icoshift” alignment with customized intervals (in this example, (0-8.11], (8.11-9.95], (9.95-11.53], (11.53-15.86] min); (g)-(h) heat map to observe the alignment of the spectra before and after applying the alignment algorithm.

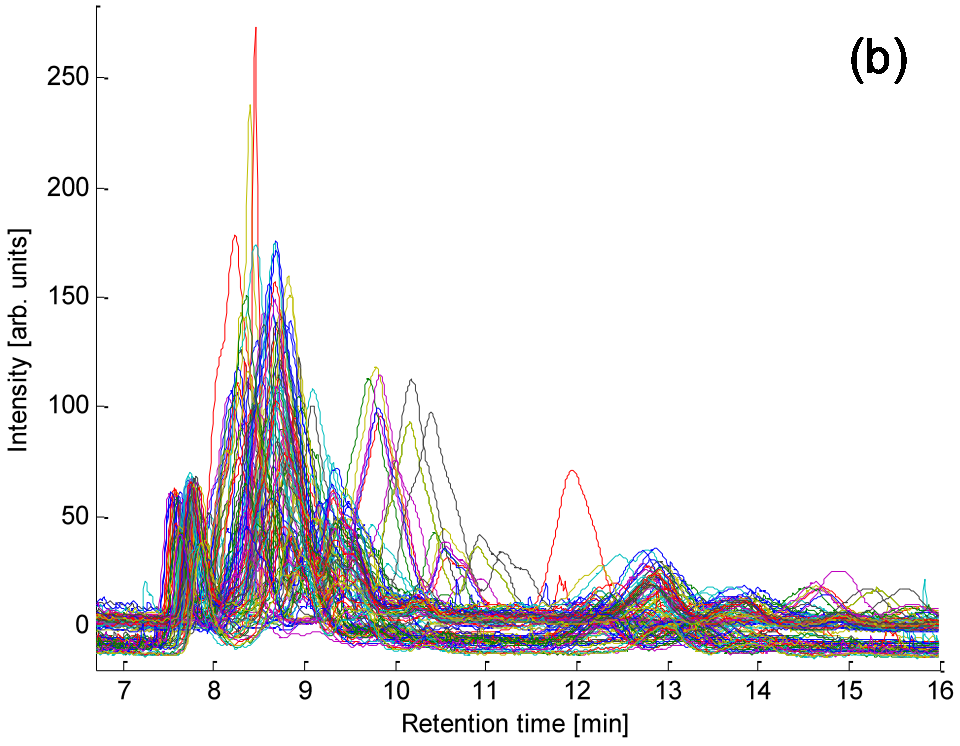
For further information, please contact Dr. Santiago Medina (smedina@ugr.es).

93


(a)
Intensity [arb. units]
Retention time [min]
0
50
100
150
200
250
300
350
400
450
0
2
4
6
8
10
12
14
16
18
20

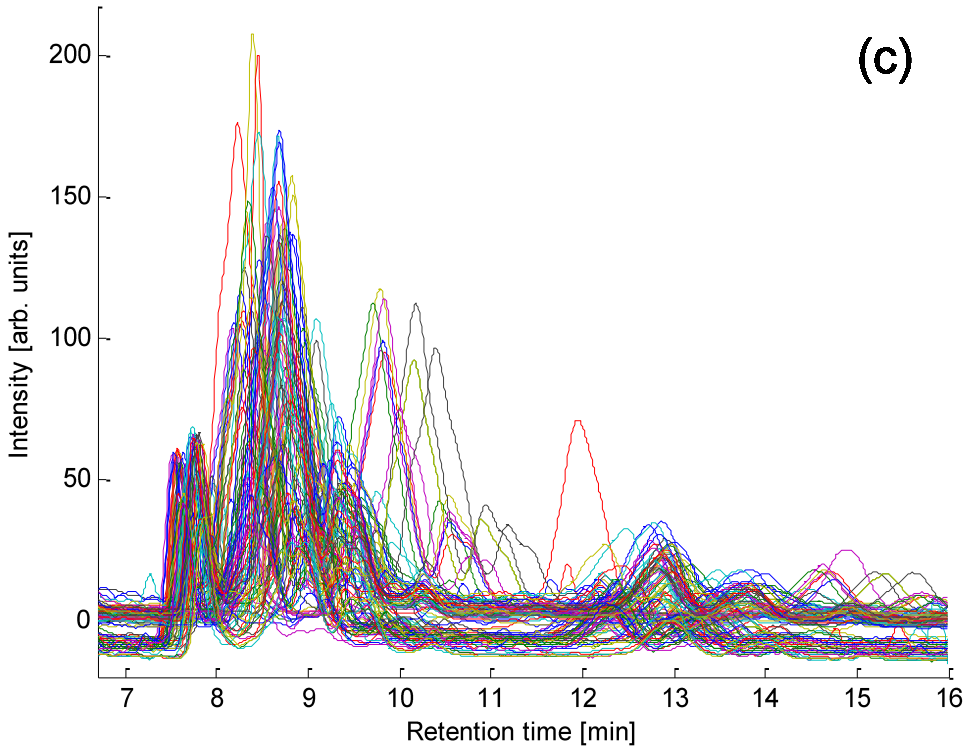
94

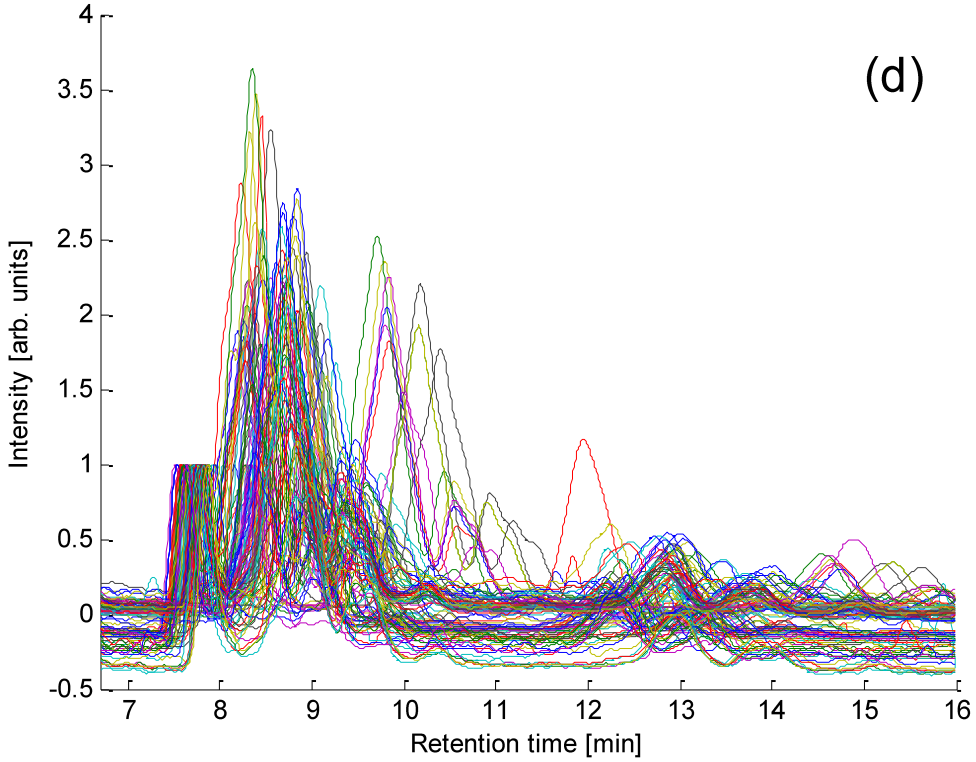
95


(b)
Intensity [arb. units]
Retention time [min]
0
50
100
150
200
250
7
8
9
10
11
12
13
14
15
16

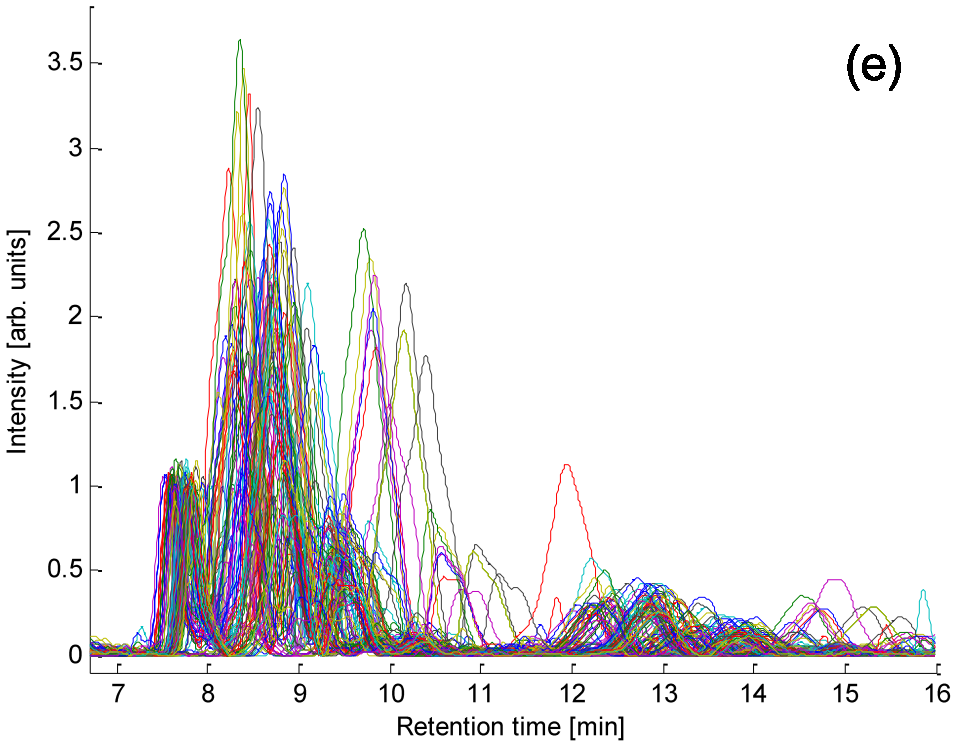
96

97


(c)
Intensity [arb. units]
200
150
100
50
0
7
8
9
10
11
12
13
14
15
16
Retention time [min]

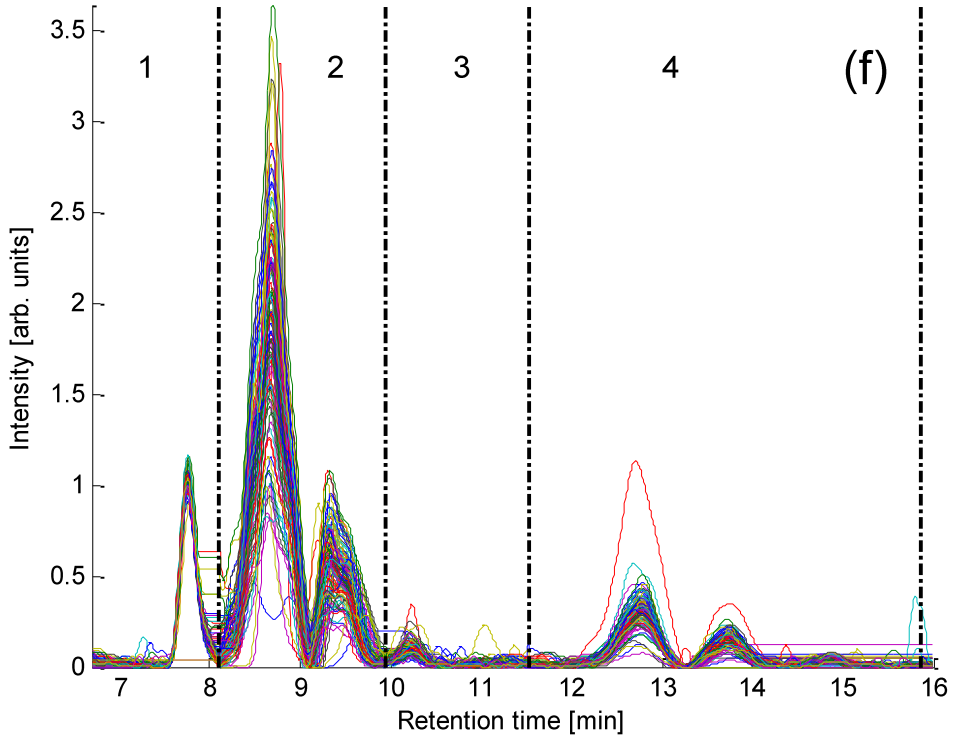

(d)
Intensity [arb. units]
4
3.5
3
2.5
2
1.5
1
0.5
0
-0.5
7
8
9
10
11
12
13
14
15
16
Retention time [min]

103



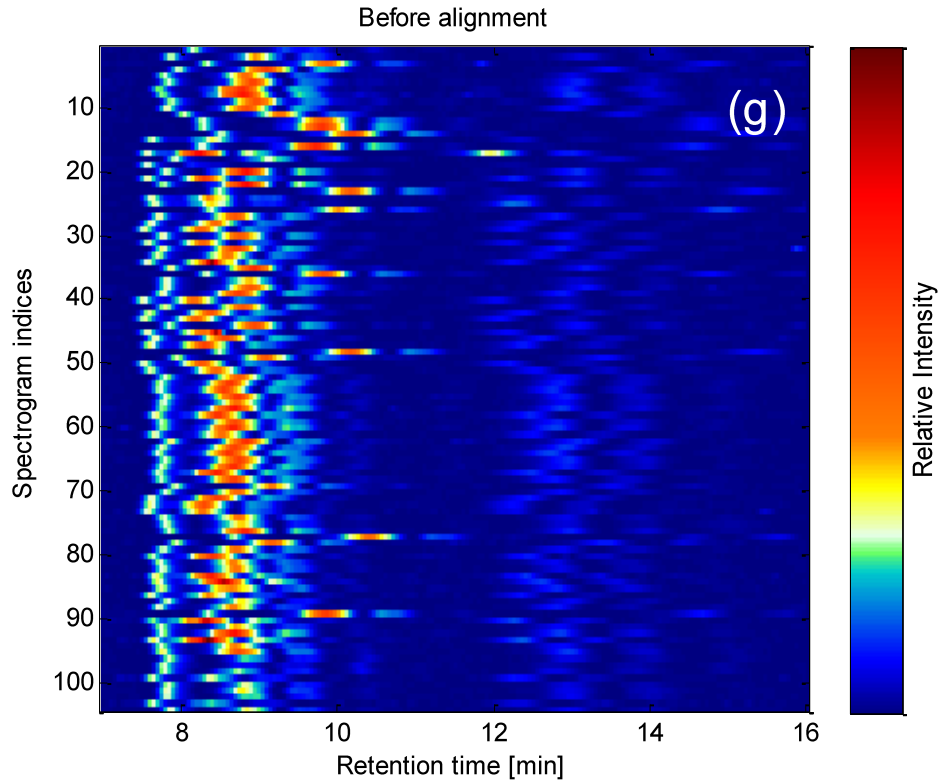
104

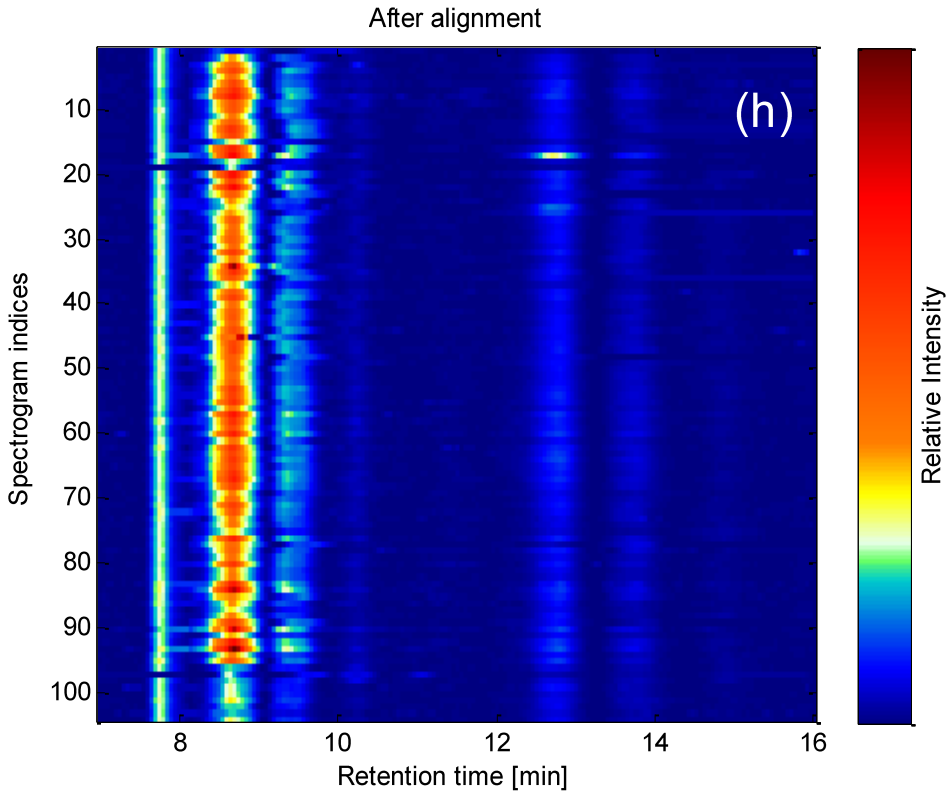
105



106

107


Before alignment
(g)
Spectrogram indices
10
20
30
40
50
60
70
80
90
100
8
10
12
14
16
Retention time [min]
Relative Intensity


After alignment
(h)
Spectrogram indices
10
20
30
40
50
60
70
80
90
100
8
10
12
14
16
Retention time [min]
Relative Intensity