Talanta 85 (2011) 264–275

Contents lists available at ScienceDirect

Talanta

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

Multivariate curve resolution modeling of liquid chromatography–mass spectrometry data in a comparative study of the different endogenous metabolites behavior in two tomato cultivars treated with carbofuran pesticide

Gabriel G. Siano[a], Isidro Sánchez Pérez[c], María D. Gil García[c], María Martínez Galera[c,*], Héctor C. Goicoechea[a,b,**]

[a] Laboratorio de Desarrollo Analítico y Quimiometría (LADAQ), Cátedra de Química Analítica I, Facultad de Bioquímica y Ciencias Biológicas, Universidad Nacional del Litoral, Ciudad Universitaria, Santa Fe (S3000ZAA), Argentina
[b] Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Avda. Rivadavia 1917, CP C1033AAJ, Buenos Aires, Argentina
[c] Departamento de Hidrogeología y Química Analítica, Facultad de Ciencias Experimentales, Universidad de Almería, La Cañada de San Urbano, 04120 Almería, Spain

ARTICLE INFO

Article history:
Received 18 February 2011
Received in revised form 21 March 2011
Accepted 24 March 2011
Available online 5 April 2011

Keywords:
Multivariate curve resolution
Partial least squares discriminant analysis
Tomato metabolites
Metabonomic
Evolutionary time profiles

ABSTRACT

A metabonomic study based on the application of multivariate curve resolution and alternating least squares (MCR-ALS) to three-way data sets obtained by liquid chromatography coupled to mass spectrometry detection (LC–MS) was carried out for Rambo and Raf tomato cultivars treated with carbofuran pesticide. Samples were picked up during a 21 days period after treatment and analyzed by LC–MS in scan mode, along with the corresponding blank samples. Then, MCR-ALS was applied to the three-way data sets using column wise augmented matrices, and the evolutionary profiles as a function of the time after treatment were estimated for the metabolites present in both cultivars, as well as their corresponding pure spectra estimations. A comparative study using those estimations showed that some of these metabolites followed different behavior for the different cultivars after treatment. Since all treated and untreated Rambo and Raf samples were picked up according to the same sampling protocol and in a similar state of maturation, any difference in the behavior between profiles can be interpreted as an effect due to the presence of pesticide and to the kind of cultivar. Based on this hypothesis, several PLS-DA approaches were tested to check if it would be possible to classify samples by using the metabolites MCR estimations. Results showed that PLS-DA models for classification of treated or non-treated (blank) samples were the best ones obtained (98.44% of correct classifications for the validation set), which supports the stress effects related to carbofuran treatment. In addition, excellent discrimination among the four groups could be attained (89.06% of correct classifications for the validation set).

1. Introduction

Tomato (*Lycopersicum esculentum*) is a warm season crop with origins in elevated regions of Peru and Ecuador. A member of the *Solanaceae* family, tomato is the most widespread vegetable in the world, as well as that showing the higher economic value. Tomatoes have traditionally been grown in the field, but the use of protected culture has allowed expand their cycle and availability through the year.

Metabolites are compounds, which are responsible of physical and chemical characteristics establishing differences between cultivars. The term metabolite refers to all small molecular weight compounds including organic acids, sugars, amino acids, vitamins, as well as small peptides. These compounds are involved in metabolic processes, whether they are the products or being necessary to the mechanism. As detection of all metabolites by an universal method is impossible, a number of techniques have been applied with this aim, such as NMR, Fourier transform infrared spectroscopy (FT-IR), pyrolysis/electron impact-mass spectrometry (pyrolisis/EI-MS), electron impact-mass spectrometry (EI/MS), electrospray mass spectrometry (ESI-MS) the most widely used being NMR and techniques based on MS detection [1].

Most MS applications in metabonomics use a separation method before mass detection, typically gas chromatography (GC), liquid chromatography (LC) or capillary electrophoresis (CE). GC–MS and

* Corresponding author. Tel.: +34 950 015313.
** Corresponding author at: Universidad Nacional del Litoral, Laboratorio de Desarrollo Analitico y Quimiometria,Facultad de Bioquimica y Ciencias Biologicas, Catedra de Quimica Analitica I, S300ZAA Santa Fe, Argentina. Tel.: +54 342 4575205..
E-mail addresses: mmartine@ual.es (M.M. Galera), hgoico@fbcb.unl.edu.ar (H.C. Goicoechea).

0039-9140/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.talanta.2011.03.064

LC–MS are frequently used techniques that can detect a wide variety of compounds at lower concentrations compared, for example, to NMR. LC is probably the most versatile separation technique as it allows separation of compounds of a wide range of polarity and thermal stability, with little effort in sample preparation, compared with GC.

However, metabonomic studies using data collected from hyphenated techniques such as GC–MS or LC–MS in full SCAN mode involve some drawbacks due to the high background noise, the presence of artefacts and also the occurrence of redundancy (i.e. different *m/z* ratios that are related to the same metabolite). Moreover, the background noise can vary with the retention time and, additionally the peak shape and retention times can vary from one run to another for complex matrices. At all events, complete chromatographic resolution of all peaks is impossible to obtain for such a large number of analytes in a single separation dimension. These shortcomings are usually overcome by chemometric approaches for denoising and compression of the data matrices without losing relevant chemical information and models using the second order advantage that are able to handle problems related to peaks shifts. Curve resolution or deconvolution methods are mainly applied to reach these aims [2–7].

In a previous paper [7] we developed a chemometric strategy based on multivariate curve resolution and alternating least squares (MCR-ALS) applied to LC–MS three-way data arrays to perform a metabonomic study in tomato (*Lycopersicon esculentum*) fruits (cultivar Rambo) following treatment with carbofuran. MCR-ALS was performed on augmented matrices built with the LC–MS three-way data obtained from treated and non-treated samples through the sampling time. The strategy allowed us to obtain the concentration and spectra profiles of the main components (previously estimated with the SVD algorithm) from samples treated with pesticide as well as from blank samples, showing how they vary with time after plants treatment with the pesticide. In addition, a simple resolved mass spectrum was obtained corresponding to the peaks of a particular component in all matrices, thus avoiding ambiguity in the compound identity assignment. Different time profiles were found for some metabolites in treated and non-treated samples, which demonstrate that the presence of pesticide causes changes through the time in the behavior of certain endogenous tomato metabolites as the result of a physiological stressful situation [7].

In this paper, we applied for the first time the above mentioned strategy in combination with some PLS-DA approaches to check the possibility of classifying tomato samples belonging to two cultivars (Rambo and Raf), treated and untreated with Carbofuran, by using their MCR-ALS resolved concentration profiles.

2. Theory

2.1. Partial least squares-discriminant analysis (PLS-DA)

Pure pattern recognition techniques are oriented to discriminate among different groups of samples and operate by dividing the hyperspace in as many regions as the number of groups. Thus, if a sample is represented in the region of the space corresponding to a particular category, it is classified as belonging to that category. In this case, each sample is always assigned to one and only one group [8]. A discriminant separating two classes can be developed by linear regression. On the other hand, when classes are described by multivariate objects and there are more variables than objects, a conventional PLS regression can be used to model a linear discriminant for classification [9,10]. PLS [11] was initially built for quantitative analysis, but it can be also used for pattern recognition. This supervised analysis is based on the relation between spectral intensity and sample characteristics. Actually, PLS latent variables are built to find a proper compromise between two purposes: describing the set of explanatory variables and predicting the response ones [8]. Binary classification is done by encoding class membership in the property vector **y**, and PLS-1 regression is used to model the single discriminant and a threshold value [10]. It involves a calibration step in which the relation between the spectra and sample codes is estimated from a set of standards, and a prediction step in which the calibration results are used to estimate the codes for unknown samples [12]. When more than two classes are present, class information can be treated as a binary classification and PLS-1 used to develop a set of discriminant boundaries between each target class and all other classes. It is also possible to encode the set of class identities in a matrix, **Y**, where each column represents the binary class membership of each sample, and to use a PLS-2 regression to develop a set of discriminant functions separating each class from all others [10].

2.2. Multivariate curve resolution and alternating least squares (MCR-ALS)

Second order instruments can provide bilinear and non-bilinear data, which determines the type of algorithms that can be used for data treatment. If data are bilinear, the **D** ($n \times m$) matrix representing the data output of the second-order instrument can be decomposed into the product of other two matrices **C** and $\mathbf{S}^T$ as follows:

$$\mathbf{D} = \mathbf{C}\mathbf{S}^T + \mathbf{E} \quad (1)$$

where the columns vectors of matrix **C** ($n \times k$) correspond to the profiles of the k pure components that are present in matrix **D**, whereas the row vectors of matrix $\mathbf{S}^T$ ($k \times m$) correspond to the spectra of the k pure components and **E** is the matrix of the residuals.

For a mixture of k compounds in high-pressure liquid chromatography (HPLC) with mass (MS) detection, **D** is the two-dimensional spectrochromatogram of MS spectra (horizontal) as a function of elution time (vertical). The matrix **C** contains the k elution profiles (column-wise) and the matrix $\mathbf{S}^T$ contains the k MS spectra (row-wise) of the k pure compounds.

Basically, the aim of multivariate-analysis algorithms applied to bilinear data from second order instruments is to solve Eq. (1), removing the noise **E** of the signal and obtaining both **C** and $\mathbf{S}^T$ [13]. These two matrices contain all the individual signals of each pure compound in the two orders of measurement.

Second order data can be arranged giving rise to a column wise augmented data matrix. In this way, several matrices that belong to different processes are appended one on top of each other so that the spectral direction is common and the data matrix length is augmented in the process direction, so the resolved pure mass spectra are common to all experiments and the concentration profiles can be different from experiment to experiment. Then, MCR-ALS can be applied in order to obtain the resolved concentration and mass spectral profiles corresponding to the components or endogenous metabolites, through the sampling period.

3. Experimental

3.1. Chemicals and solvents

Magnesium sulphate anhydrous ($MgSO_4$) and sodium acetate 3-hydrate ($AcONa{\cdot}3H_2O$) were obtained from Merck (Germany). Acetonitrile (ACN) of HPLC grade was obtained from J.T. Baker (Holland) and acetic acid glacial (AcOH, 99.7%) from Panreac (Spain). Ultra pure water was provided by a Milli-Q water purification system from Millipore (Bedford, MA, USA). Mobile phases were filtered

through a 0.45 μm cellulose acetate (water) or polytetrafluoroethylene (PTFE) (organic solvents) filters, and degassed with helium prior to and during use. All the extracts were filtered through a Millipore membrane of cellulose acetate (0.45 μm particle size) before pumping them into the chromatographic system. Finally a concentrated suspension of Botrán 20 (carbofuran 20%, w/v) was obtained from Tragusa (Sevilla, Spain).

3.2. Instrumentation and software

LC separation was carried out with a Hewlett Packard (H-P) series 1100 system (Hewlett Packard, Wilmington, DE, USA) provided with an H-P Chem Station for MS control and spectral processing. The HPLC system consisted of a model G 1311 gradient pump and a Rheodyne six-port injection valve (model 7725i) with a 20 μL loop. The analytical separation was performed with a 150 mm × 4.6 mm i.d. Agilent Zorbax EclipseXDB C_8 column (5 μm particle size). An H-P G 1948 A Platform benchtop single quadrupole mass spectrometer with an ESI interface was used to detect the target compounds in the LC column effluent.

A 230 V–50 Hz crusher (0.5 Kw maximum potency) from Sammic S.L. (Azpeitia, Spain) and a polytron PT1035 from Kinematica AG (Switzerland) were also used. A rotary evaporator (R-114) with a B-480 thermostated water bath was purchased from Buchi (Flawil, Switzerland). A Sigma 4–15 centrifugal provided with a Sigma 11150, 143/F, 5100/min rotor was used during the extraction step. Finally crushed and homogenized samples were stored in a −86 °C ultralow freezer.

MatLab 7.6.0 R2008a (The MathWorks, Natick, MA, USA) was used as the development platform. A graphical interface was used to apply MCR-ALS, which additionally provides detailed information about the implementation of this algorithm [14] Statistics Toolbox™ for MatLab, PLS Toolbox 3.52 [15] and home-made routines were used in this work to do the maths.

3.3. Field trial design, pesticide treatment and sampling

Three different tomato cultivars (Rambo, Raf and Zayno) were grown in a 1 ha greenhouse located at the Experimental Farm UAL-ANECOOP (Almería, Spain). Plants were arranged according to a design following the criteria previously published in a previous paper [7], which is depicted in Fig. 1.

Plants, receiving routine horticultural practices, were treated with Botrán 20 (carbofuran 20%, w/w), which was incorporated in the irrigation water at the doses recommended by the supplier (4 L/ha), following the scheme depicted in Fig. 1. Finally, Rambo and Raf tomatoes were sampled following the protocol described in a previous work [7].

3.4. Extraction step and LC–ESI-MS analysis

Rambo and Raf tomato extracts were obtained by using the original QuEChERS method [16,17] including some modifications [7]. The steps in the extraction process are as follows: (1) weigh 15 g of thoroughly homogenized sample into a 50 mL Teflon centrifuge tube; (2) add 15 mL of ACN acidified with 1% AcOH; (3) add 6 g of anhydrous $MgSO_4$ and 2.5 g of $AcONa{\cdot}3H_2O$; (4) shake vigorously for 3 min by hand; (5) centrifuge the tube at 3700 rpm for 5 min. On the other hand, a preconcentration step was carried out by evaporating to dryness aliquots of 10 ml of supernatant in a rotatory evaporator, which were reconstituted with 1 ml of ACN. Finally, the extracts were filtered through Millipore membrane Teflon filters (0.45 μm particle size) before injection into the chromatographic system. The chromatographic separation was carried out in an identical way for both classes of samples, following the procedure previously described [7].

3.5. Data sets and multivariate analysis

Eight treated and non-treated Rambo and Raf tomato samples from each sector (A, B and C) of the greenhouse were used in this work, in such a way that the total number of available samples for this study was 96.

After LC–MS analysis in full scan mode and zero values elimination, every sample was represented by a 507 × 710 matrix, where the first value corresponds to the number of retention times (30 min of chromatographic run with 507 points) and the second one to the number of mass (50–750 amu with 710 *m/z* values). These data files were provided in *cdf* format by the HP Chem Station Software and then they were converted to ASCII format to be processed with MatLab.

Matrices were undergone to a reduction in its dimensionality by using a Discrete Wavelet Transform (DWT) with Haar wavelet, which compressed them to a quarter of its size, without losing relevant chemical information. After this procedure, all matrices had a size of 127 × 178. More details about this pretreatment can be found in a previous paper [7].

Because of the complexity of the data set and hardware limitations, every wavelet reduced matrix was divided in four chromatographic time regions (data points 1–35, 36–70, 71–100 and 101–127, respectively). In this way, equivalent regions of all samples were used to compose column-wise augmented matrices (one per region). Each of these augmented matrices was constituted by second order data corresponding to the eight Rambo and Raf treated samples, which were picked through the eight sampling days from the sectors A, B and C, and by the eight non-treated samples picked on the same days, from these same sectors.

Before MCR-ALS, SVD was applied to the augmented matrices, in order to estimate the number of significant components responsible for variance generation. Finally, MCR-ALS resolved the column-wise augmented matrices into individual concentration and spectral profiles using non-negativity (spectra and concentrations) and unimodality (concentration) constraints. Concentration profiles in combination with sampling time provided the evolutionary profiles of the endogenous metabolites present in both treated and non-treated Rambo and Raf samples.

The evolutionary profiles were compared by using a home-made routine which allowed us calculate the Pearson's correlation coefficient between profiles. These profiles were also partially compared through relative movements of one over another to check if metabolite pathways were retarded or accelerated in terms of sampling time.

On the other hand, concentration profiles of the MCR-ALS resolved components were used to build PLS-DA classification models. To do that, concentration profiles from the four obtained MCR-ALS resolutions (one per region) were compiled in a single matrix which had a size of NSamp × Ncomp, being NSamp the total number of available samples (96) and Ncomp the sum of the components estimated by SVD on the four column-wise augmented matrices and resolved by the MCR-ALS procedure. In this sense, every second order data matrix was represented by a single first order vector of resolved components areas.

It should be taken into account that classifying groups of samples based on their metabolic profile might not be an easy task due to the low number of objects compared to the large number of variables (in our case, concentration profiles of MCR-ALS resolved components). The large number of peaks in these samples that are all potential biomarkers involve modeling and validation challenges. The number of samples needed to accurately describe such a classification problem increases exponentially with the number

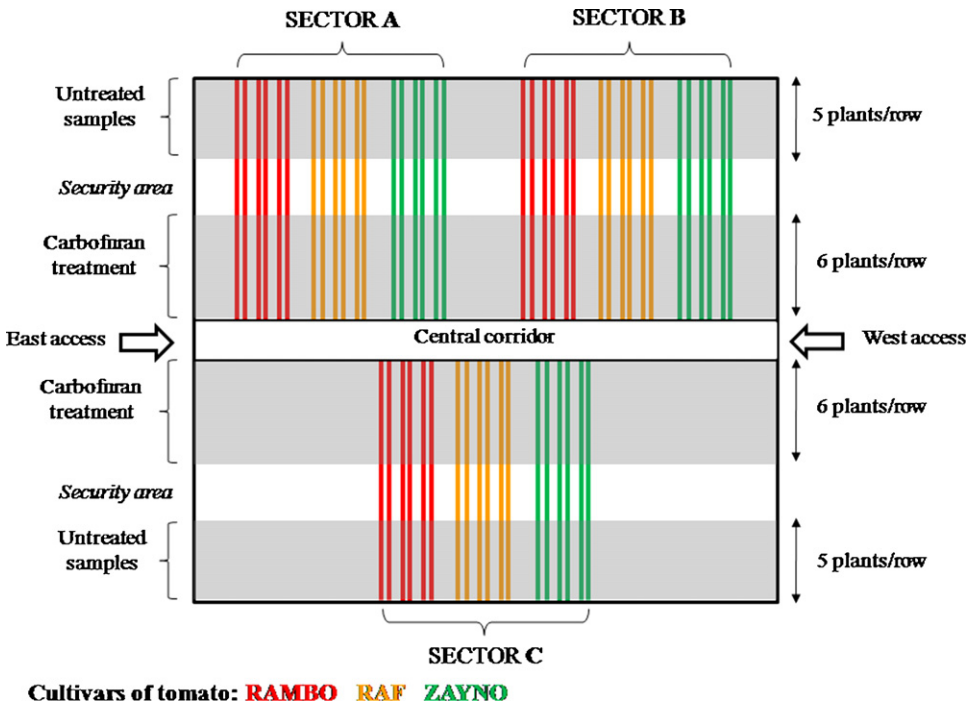

Fig. 1. Scheme of the planting design developed in the greenhouse and spatial distribution of the plants untreated and treated with carbofuran pesticide.

of variables measured. However, the number of samples used in these applications is usually much smaller than the number of variables. Although PLS-DA is one of the data analysis methods used in these kind of cases, unfortunately this procedure eagerly overfits the data [18] So, with that in mind, a simple procedure was tested to select components in an attempt of reducing the number of active variables. Given the areas of all the MCR-ALS resolved components in a set of known samples (the calibration set) and the respective classes of them, it was evaluated, component by component, if there was statistical equality of mean areas between samples of different classes. With this aim, the Levene's test was applied to check if variances in both groups could be considered statistically equal or not, and then a *t*-test was performed taking that into account. Finally, all components for which the hypothesis of equality in mean areas could not be rejected were discarded, being the rest of the components the selected ones. When more than two classes of samples were involved, the components selection proceeded evaluating "one vs one". Then all selected components were joined, and finally a copy of that selection was obtained discarding all the repeated components.

In summary, PLS-DA models were built with and without components selection. In both cases, matrices of both predictor and predicted variables were mean-centered and this was the only preprocessing applied to the data. Besides the fact that some problems could arise due to the use of cross validation (CV) in models optimization [18,19] and taking into account that the main objectives of this work were not related to an exhaustive study of these issues, the number of latent variables selected for each model was obtained by leave-one-out cross validation (LOOCV) on the calibration sets. In general, this number of latent variables was the one represented by the first minimum in Root Mean Squares Error of Cross Validation (RMSECV) vs number of latent variables plot. It should be noted that LOOCV of models which were not built for binary classifications (models using the PLS2 algorithm, in which a unique number of latent variables must be used for all modeled classes) had some disagreements in terms of what number of latent variables could be considered as the optimal one, since the minimum CV error for each class could be obtained for different number of factors. In these cases, the number of factors selected was the one coincident with the minimum error for the worst predicted or generally worse predicted class, even when for the rest of the classes this number would not have been the best choice, in a try to keep all the classes more or less contemplated by the models.

Finally, since PLS-DA models were built with PLS Toolbox 3.52, each sample was assigned to the modeled class for which the sample prediction was bigger than the threshold of that class. These thresholds were estimated using the Bayes Theorem and the available data in order to minimize total errors. When predictions were superior to more than one threshold, samples were assigned to the class with the highest prediction probability. More details about the calculus of these probabilities and the thresholds can be found elsewhere [15]. Fig. 2 shows a scheme of the steps involved in this study.

4. Results and discussion

4.1. Sampling and samples pre-treatment

The pre-harvest interval of a pesticide can be defined as the minimum time that must elapse between its application and the crop harvesting. After that time, the concentration of pesticide in fruits is expected to be lower than its MRL (maximum residues limit). In tomato crops carbofuran is used at a dose of 4 L/ha with a pre-harvest interval of 45 days [20].

Usual horticultural practices for this type of crop were followed and then, tomato fruits were sampled according to a protocol proposed by the European Union [21] as indicated in Section 3. After sampling, Rambo and Raf tomato samples were extracted according to the QuEChERS method [16,17]. In this step, the dispersive-solid phase extraction (SPE) clean-up using PSA (primary-secondary amine), included in the original QuEChERS method, was not performed after ACN extraction as for profiling an exhaustive extraction method is the best

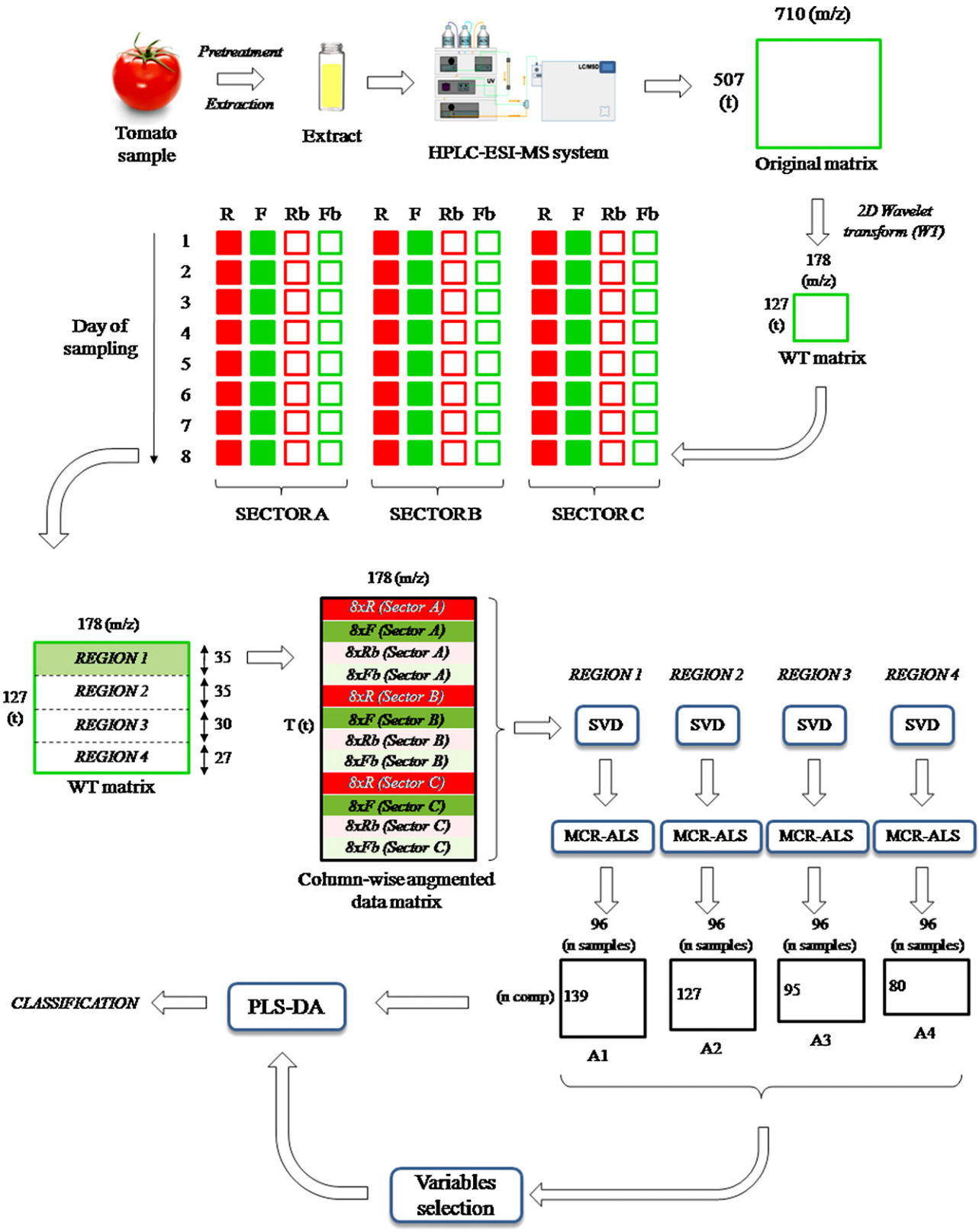

Fig. 2. Scheme of the steps involved in this study. Rambo (R), Raf (F), Rambo Blank (Rb) and Raf Blank (Fb).

choice [7]. On the other hand, the preconcentration step carried out before LC–MS analysis allowed us to increase the sensitivity in the signals in compensation with the loss of sensitivity when the acquisition of the data is performed in scan mode.

4.2. Reduction of the dimensionality

Often, LC–MS data sets obtained in full scan mode are of large size and as a result their processing becomes difficult. In this work, the Discrete Wavelet Transform

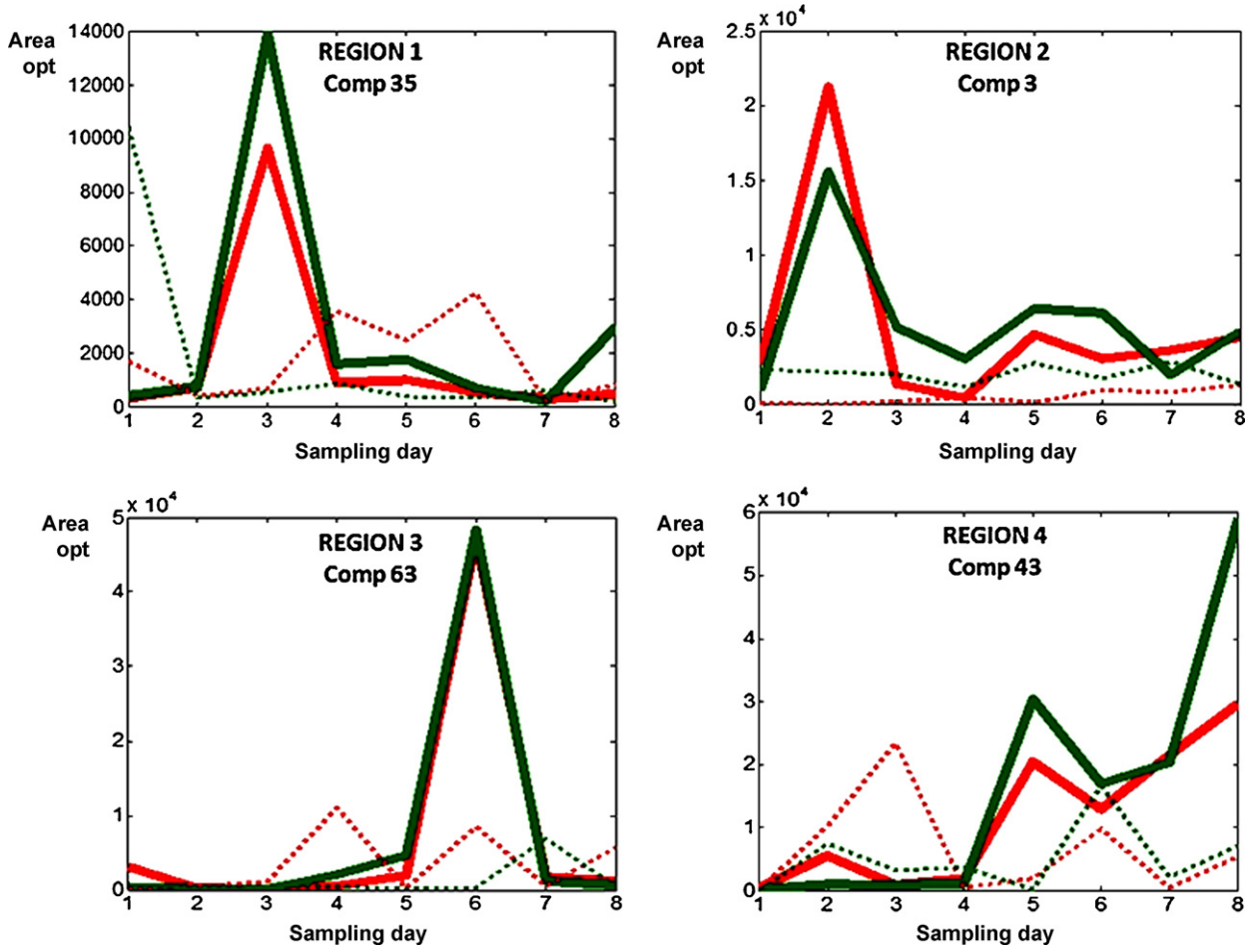

Fig. 3. Evolutionary profiles of four components (one per region) through the eight sampling days in Rambo treated (____), Raf treated (____) and their respective non treated samples, Rambo (- - - -) and Raf (- - - -). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

(DWT) technique [22] was used to perform data compression with the aim of facilitating further chemometric data treatments, without losing important chemical information.

Among the different DWT filters, the Haar wavelet (the simplest and first member of the Daubechies family of orthogonal wavelets) was chosen because it is the only wavelet which keeps the non-negativity property in the approximations (low frequency), allowing the application of ordinary multivariate curve resolution methods with non-negativity constraints [23]. In this way, DWT was applied to both dimensions of each individual matrix by using the standard approach. In this procedure, matrices were decomposed to level 2 in the wavelet coefficients domain, as a compromise between compression and resolution, in such a way that the computer worked fast enough without losses of important MS spectral information. Finally, the wavelet approximations coefficients corresponding to the optimal decomposition level were used to reconstruct the final reduced matrices in its own signal domain. Compression to the above indicated level reduced the MS spectra from 710–178 *m/z* values in each MS spectrum, whereas in the other dimension the number of rows (retention times) was reduced from 507 to 127. This procedure was applied to each individual data matrix. Although the size of the data matrices was reduced to 25% of their original size in both dimensions, limitations still related to computational capabilities led us to apply a further strategy. Thus, by establishing 4 regions in the time dimension, each 127×178 WT matrix was subdivided into 4 submatrices of dimensions 35×178 for the first and second regions, 30×178 for third region and 27×178 for fourth region (see Fig. 2).

4.3. MCR-ALS of augmented data

The first step consisted in building up augmented column-wise matrices **D** from the individual data matrices, by setting one on top of the other and keeping the column vector space in common. This procedure was carried out for the four regions from the original WT matrices, obtaining four augmented data matrices. The dimensions of these four augmented data matrices depended on the dimensions of each region. In this way, augmented data matrices built from regions 1 and 2 had dimensions of 3360×178, whereas augmented data matrices corresponding to regions 3 and 4 had dimensions of 2880×178 and 2592×178, respectively.

A previous step before MCR-ALS analysis was to determine the number of components explaining an acceptable value of the total variance of each augmented data matrix. In this work, a 90% of total variance explained was chosen as a compromise between number of component and computational time.

The number of components in each augmented matrix **D**, explaining a 90% of total variance, was estimated throughout SVD algorithm. This estimation gave a total of 139 principal components in region 1, 127 in region 2, 95 in region 3 and 80 in region 4.

The matrices **C** and $\mathbf{S}^T$ were estimated from each augmented data matrix by an ALS procedure, starting with the implementation of initial estimates to spectra profiles, which were calculated by applying the SIMPLISMA algorithm, fixing the noise level in 0.1.

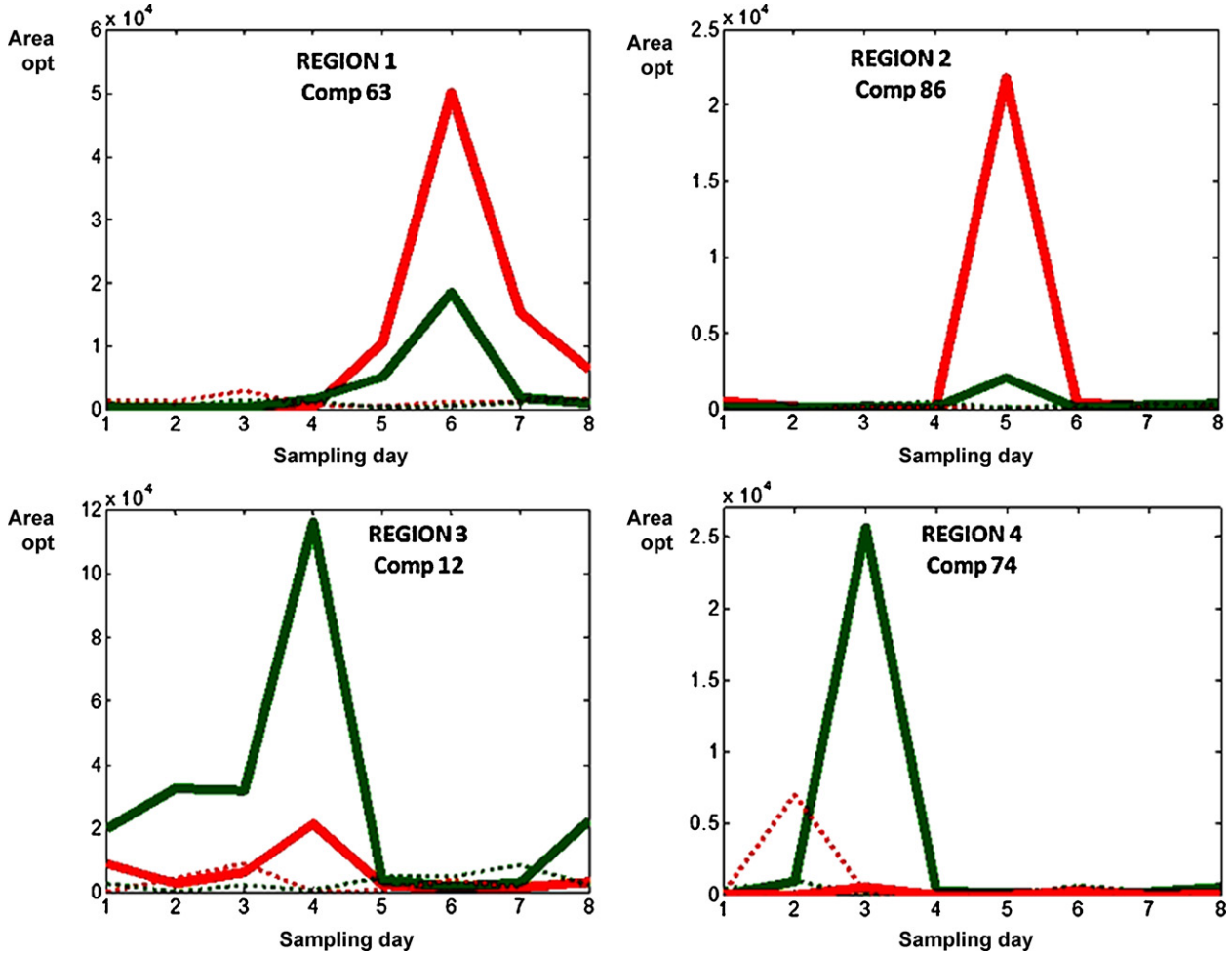

Fig. 4. Evolutionary profiles of some components through the eight sampling days in Rambo treated (_____) and Raf (_____) samples, extracted from the four regions. In all these examples the kinetic curves are similar but the concentration level of the components are different. Dashed lines correspond to blank samples, Rambo (- - - -) and Raf (- - - -). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

Also, during the iterative optimization non-negativity (applied to concentrations and spectra) and unimodality (applied to concentrations) constraints were applied to obtain chemically meaningful solutions.

Analytical figures of merit, obtained after MCR-ALS optimization, are summarized in Table 1. As can be observed, all the MCR-ALS analysis resulted in a percent of variance explained (r^2) higher than 90%.

4.4. Comparison of evolution profiles in different samples

A matrix **A** containing the areas under the concentration profiles for each component at a sampling date was also obtained, where each value a_{ij} of this matrix corresponds to the area of the ith component in jth sampling day. Every row of **A** corresponds to a MCR-ALS resolved component, in such a way that the matrix **A** contains a total of 139, 127, 95 or 80 rows, depending on the region being considered. Thus, each row of **A** represents the evolutionary profile of each component through the time, which shows how vary the concentration of each component through the eight sampling days in the four kinds of samples (treated and untreated Rambo and Raf cultivars).

Table 1
Figures of merit of MCR-ALS analysis for each region.

Region	Lack of fit (%)	r^2 (%)
1	22.07	95.12
2	19.77	96.09
3	19.46	96.21
4	13.84	98.08

An exhaustive comparison of the individual evolutionary profiles of all the components which are present in treated and non-treated samples of Rambo vs the corresponding profiles in Raf cultivars was carried out for each region. A Pearson's coefficient (r) for each pair of evolutionary profiles was also obtained.

Thus, r values close to 1 indicate a high degree of similarity in the pair of evolutionary profiles being compared, that is, a similar behavior over time for those components in both kinds of samples.

This behavior is shown in Fig. 3 for a component belonging to each region. In this case the components are in similar level of concentration for both Rambo and Raf samples.

However, in some cases, despite of obtaining satisfactory correlations in the evolutionary profiles in both kinds of samples, that is, similar kinetics of evolution, the profiles appears at different levels of concentration. In this way, good correlations were obtained for the evolutionary profiles of some components, for which the evolution curve of a Raf treated sample is below the curve of the Rambo treated sample or vice versa. This can be interpreted assuming that, although the pesticide altered the behavior of the components in comparison with the blank samples, did not modify the kinetic

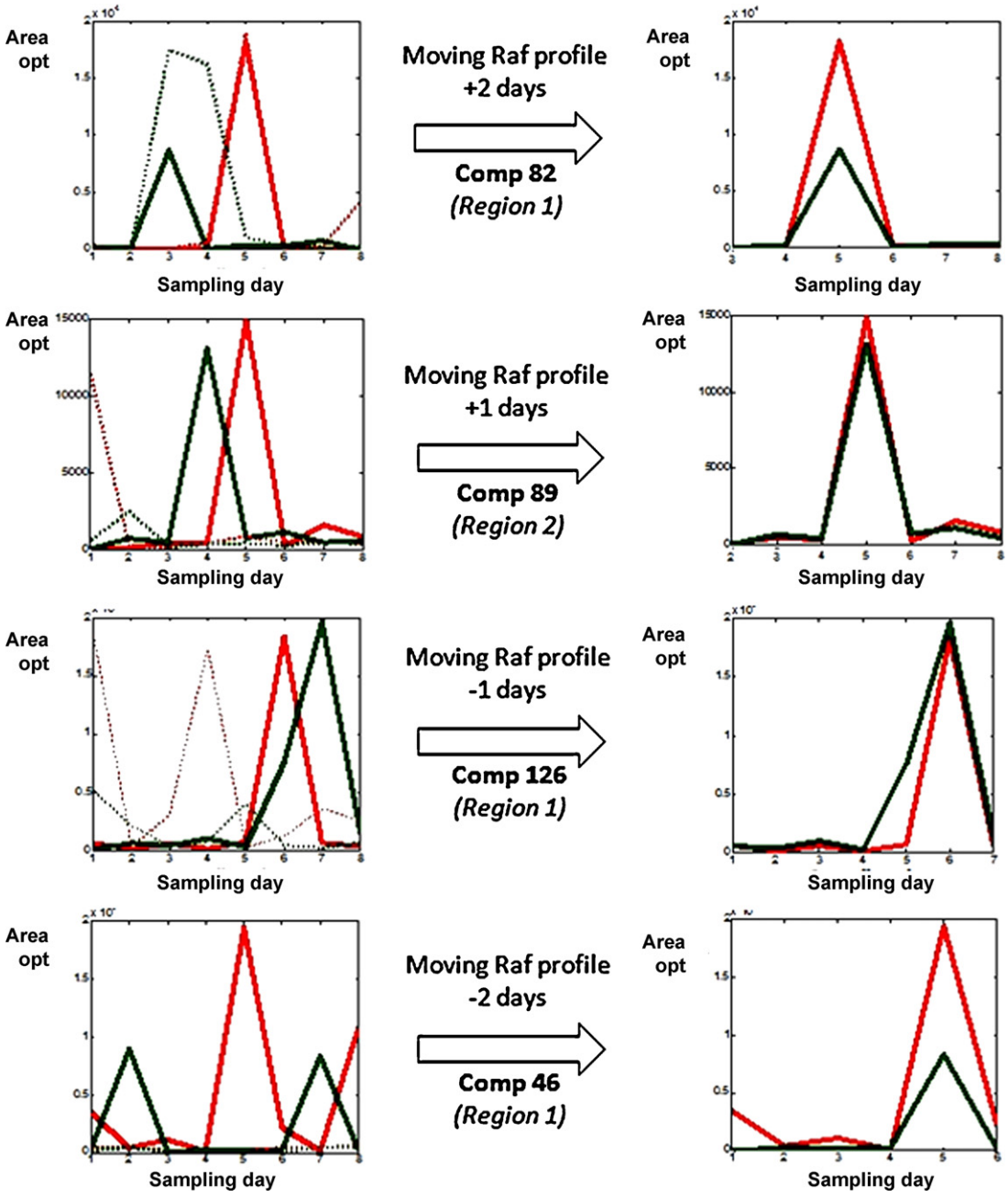

Fig. 5. Moved evolutionary profiles of some components through the eight sampling days in Rambo treated samples (——) and in Raf treated samples (——). Dashed lines correspond to blank samples, Rambo (– – –) and Raf (– – –). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

evolution of the component in both cultivars of tomato but did alter its concentration level in a pattern depending on the cultivar.

Some examples of these behaviors are depicted in Fig. 4. The two upper examples show evolution profiles where the level of concentration in a Rambo sample is higher than in a Raf samples, and the two lower examples show an inverse behavior. In all cases, the evolution profiles of the components are different in treated and non treated samples, for each cultivar. Note how the kinetic profiles of the blank samples appear at very low concentration levels in comparison with their respective treated samples.

Fig. 5 shows some examples where the metabolism of the component has been delayed in a Rambo treated sample in comparison with the metabolism in the Raf treated sample and vice versa.

Finally, some evolutionary profiles showed very low r values, which usually indicate a very poor correlation between the profiles, that is, a different behavior in the kinetic profiles depending on the cultivar of tomato.

4.5. PLS-DA classification

After MCR-ALS resolution, areas from all the resolved components (matrices A1, A2, A3 and A4, obtained from the MCR-ALS analysis of the 4 regions) were compiled in a single matrix which size was 96×441 (samples by components). This matrix was used to test different approaches of classification.

Since all treated and untreated Rambo and Raf samples were picked up according to the same sampling protocol and in a similar

Table 2
Results obtained from different PLS-DA models (R = Rambo Treated, RF = Raf Treated, Rb = Rambo Blank, RFb = Raf Blank, letters in brackets means that these classes are used as a unique class).

PLS-DA models results	Without components selection			With components selection		
Model for:	Rambo/Raf	Blank/Treated	4 classes	Rambo/Raf	Blank/Treated	4 classes
Classes	(R+Rb) (RF+RFb)	(R+RF) (Rb+RFb)	(R) (Rb) (RF) (RFb)	(R+Rb) (RF+RFb)	(R+RF) (Rb+RFb)	(R) (Rb) (RF) (RFb)
LV (latent variables)	3	2	6	3	2	5
Number of used components/variables	441	441	441	41	63	127
Correct classifications in calibration set	32/32 (100%)	32/32 (100%)	32/32 (100%)	31/32 (96.88%)	32/32 (100%)	31/32 (96.88%)
Correct classifications in validation set	60/64 (93.75%)	63/64 (98.44%)	52/64 (81.25%)	58/64 (90.63%)	63/64 (98.44%)	57/64 (89.06%)

state of maturation, any difference in the behavior between profiles can be interpreted as an effect due to the presence of pesticide, to the kind of cultivar or to both.

For the application of the PLS-DA methodologies, a calibration set was made taking one of every three samples (represented by the areas of the MCR-ALS resolved components) from the 96 available samples (starting from the first of them), being the rest of samples assigned to the validation set to test the screening capabilities. In summary, calibration and validation sets were built with 32 and 64 samples, respectively. It can be noted that each set contained a similar proportion of the four kinds of samples (Rambo, Raf, Blank Rambo and Blank Raf, or R, RF, Rb and RFb, respectively) from the 3 sectors (A, B and C), and also that the same calibration and validation sets of samples were used both for binary ("Blank/Treated" and "Rambo/Raf") and quaternary ("Rambo/Raf/Rambo Blank/Raf Blank") models. Depending on the model, coded values were assigned to each calibration sample according to its category. In the case of "Blank vs Treated" models as well as in the case of "Raf vs Rambo" models, classes were represented in a vector by zeros and ones, respectively, and the PLS1 algorithm was applied to obtain binary decision models. On the other hand, for "R/RF/Rb/RFb" (four different classes) models, the codification was done by using a dummy matrix **Y** composed of binary digits with as many rows as training samples were used and as many columns as categories were present, in such a way that the presence of a unique logic one per row indicated the class of the sample. Also, when it was needed, classes R, RF, Rb and RFb were coded with the integer numbers 1, 2, 3 and 4, respectively. In the four class models, the PLS2 algorithm for multivariate predicted variables was used when modeling.

Table 2 shows a summary of the results obtained from different PLS-DA models. All models presented similar results when calibration samples were predicted and in general terms, the whole set is correctly predicted without problems, although this fact could be due to an overfitting situation respect to the calibration set. Looking at binary models, especially "Blank/Treated" ones, it can be seen that they only need a small number of latent variables to deal with the differences between classes, and taking the good validation results into account, one could assume that no overfitting occurred in the training step. Maybe the same hypothesis would not be valid in the case of the quaternary model without components selection, since it had a perfect prediction for the training set but in terms of validation results the performance was not at the same or similar level than the rest of the models, in such a way an overfitting in the modeling stage could have occurred. Another reason for this lower performance could be that the number of samples representing each of the four classes was too small to show differences among classes, especially when all the original variables were used. Moreover, modeling with the PLS2 algorithm has the extra difficult of selecting only one optimal number of latent variables for all modeled classes. In the case of this quaternary model without variables selection, this pseudo optimal number of factors was the one for which the CV error for one class (Raf treated) was almost three times higher than for the rest of classes (data not shown). In fact, all wrong predictions in the validation set were for treated samples. Nevertheless, the PLS2 algorithm takes advantage of possible correlations between **Y** variables [8], but if the classes were independent, this advantage would not be obvious. In our case, classes had some degree of relationship, since for example Rambo treated and untreated samples; although they were considered as two different classes, it is supposed that they preserved their common Rambo nature.

Whenever the components selection procedure was applied, it was performed at a 95% of confidence level. As can be seen, it reduced the number of active components in the classifications, leaving near 10%, 15% and 30% of the originally available variables for "Rambo/Raf", "Blank/Treated" and "4 classes" models, respectively, which may lead to more parsimonious models. Seeing at validation results, the performance of models which classify "Rambo"/"Raf" or "Treated/Blank" samples is basically the same, independently of components selection. On the other hand, in the case of four different classes, components selection seems to improve considerably the classification both in terms of correct predictions and also in the reduction of the needed latent variables. Related to the last item, CV errors were reduced for all classes (data not shown).

Fig. 6 shows the predicted vs nominal coded values when PLS-DA was applied to discriminate between the four classes (Rambo Treated: 1, Raf Treated: 2, Rambo Blanks: 3 and Raf Blanks: 4) and using variable selection. The confidence interval for each category was estimated as the product of the calculated standard deviations of the results for the training samples and the Student *t*-value with $(n-1)$ degrees of freedom. As was commented before, in this figure it can clearly be seen that calibration samples were correctly predicted without problems. On the other hand, when validation

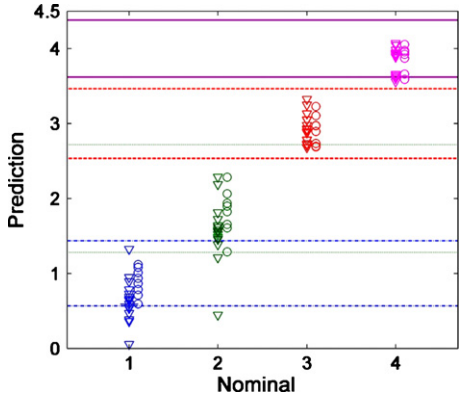

Fig. 6. Plot of the PLS-DA predicted vs nominal coded values for validation samples (triangles) in the quaternary models with variable selection. Rambo Treated (1-blue), Raf Treated (2-green), Rambo Blanks (3-red) and Raf Blanks (4-magenta). Circles to the right of each group indicate the values of the corresponding training samples and were shifted for clarity. Regions between lines of the same color indicate the confidence interval for each class (see text). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

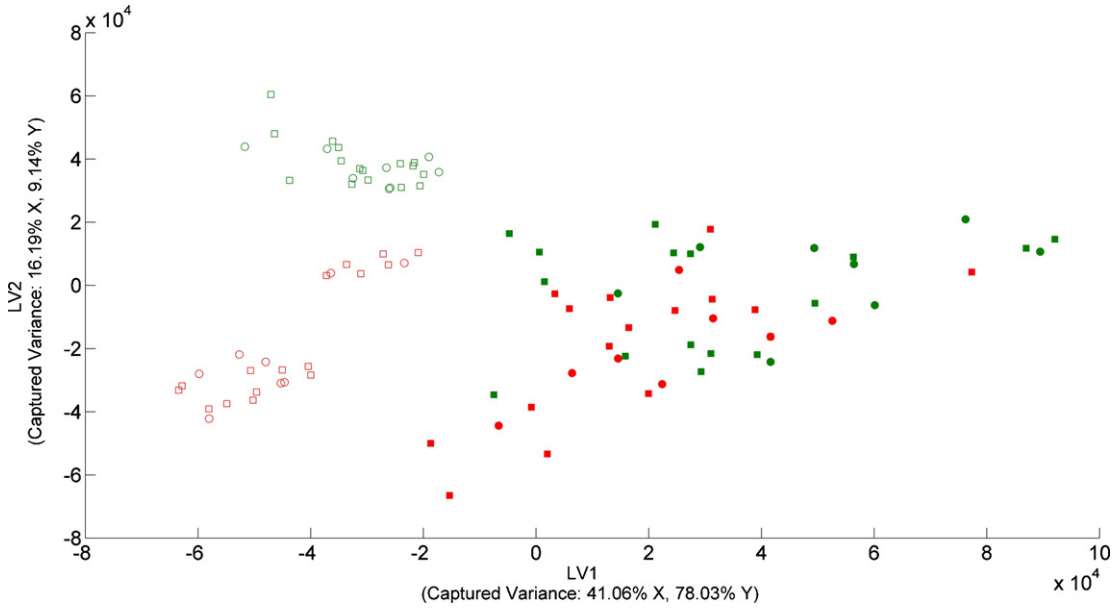

Fig. 7. Distribution of calibration (circles) and validation (squares) samples in classification models for "Treated/Blank" samples with components selection (green = Rambo, red = Raf, filled = Treated, empty = Blank, LV = latent variable). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

samples were predicted, a few samples layed out of the confidence limits.

Besides the fact that components selection did not change the performance of all models in a radical way, some aspects should be taken into account. For example, Figs. 7 and 8 show the distribution of calibration/validation samples in classification models for "Treated/Blank" samples with and without components selection, respectively. As can be seen, in both models blank samples are on the left part of the plane (negative scores in LV1) while treated samples are mainly on the right part. On the other hand, if LV2 is analyzed, one can see that components selection leads to a separation between Rambo and Raf untreated samples even when no information about it was given when the model was done, since in the modeling stage, calibration untreated samples represent only one group besides the fact of their Rambo or Raf nature. Also, the components selection process itself did not have any information about these natures. Thus, under this point of view, components selection seems to preserve natural and non-modeled characteristics of the samples. Similar conclusions were obtained in other work [18], in which it was stated that although the score plot should not be used to infer class separation, it might reveal structures (e.g. subgroups) within a class, since if the model is not forced to show this difference (as it was in our case), this is not a result of overfit, and thus, such information could be inferred from the score plot. Fig. 7, which corresponds to the "Treated/Blank" model with components selection, shows that these facts are also valid for the validation set.

Fig. 9 shows the regression vector for Treated/Blank samples with components selection and also a loading plot in which it

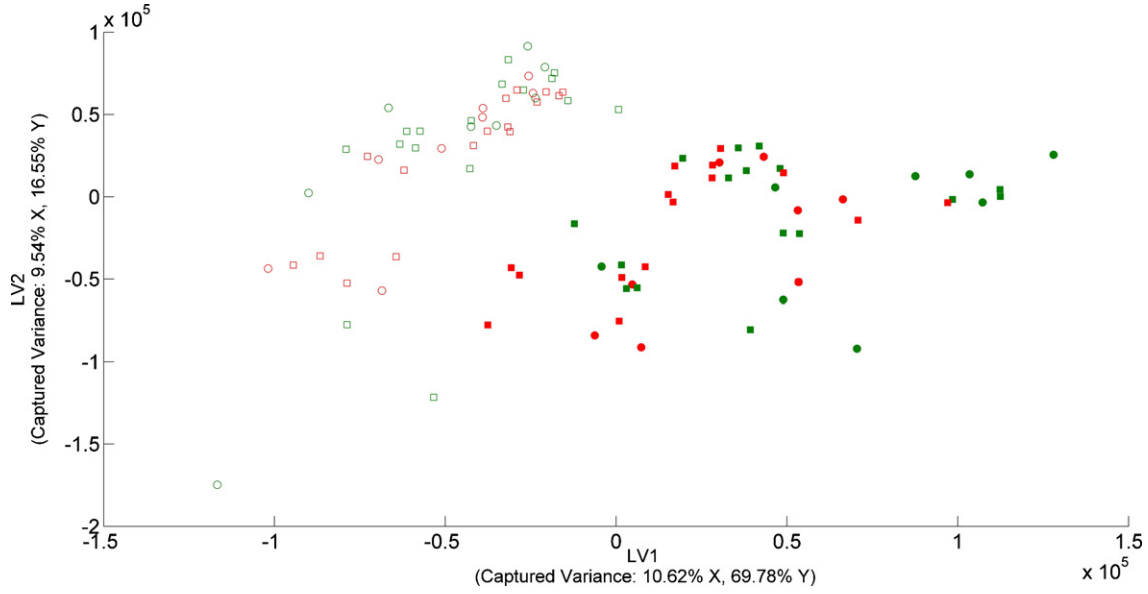

Fig. 8. Distribution of calibration (circles) and validation (squares) samples in classification models for "Treated/Blank" samples without components selection (green = Rambo, red = Raf, filled = Treated, empty = Blank, LV = latent variable). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

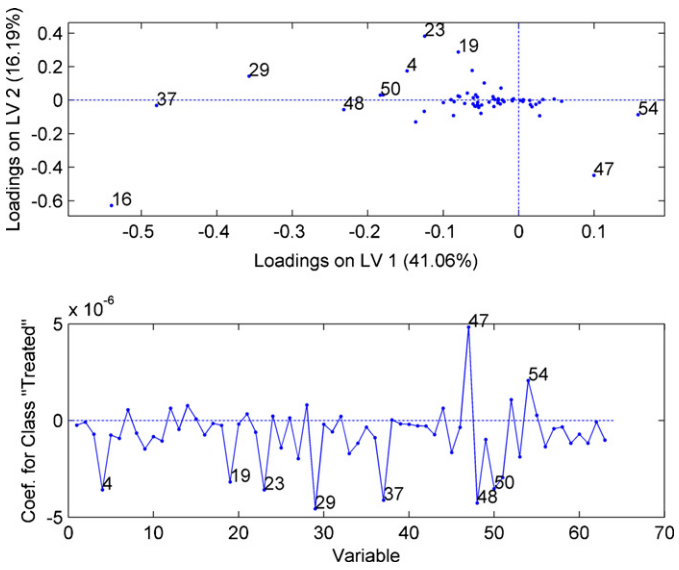

Fig. 9. Loading plot (above) and regression vector (below) for Treated/Blank model with component selection.

can be seen the relationship between variables. In both graphics numbered points corresponds to variables with the high absolute values (relative to the values present in each plot) so that these selected components can be considered as responsible markers for the differentiation between classes. Fig. 10 shows resolved component areas for the 96 available samples, evaluated at some of these potential markers. It should be noted that samples 1–32, 33–64 and 65–96 correspond to different sectors of the greenhouse (A, B and C, respectively). As can be seen, results are sometimes different among sectors, so it could be questionable the fact that all these samples have been grouped in a unique set from which calibration samples were selected to represent all modeled phenomena. Nevertheless, some tendencies are more or less clear inside each sector. Generally speaking, treated samples show similar levels for these components. Sometimes blank samples of only one class are near to those levels, so these blanks could be confounded with treated samples. These are the cases for Raf blanks with components 4 and 19, and for Rambo blanks with component 48. On the other hand, component 50 shows a clear difference between treated and blank samples across all the sectors, so this component could be inferred as the most powerful marker. It could also be noted that for a given sector and component, in all cases the blanks are nearly separated.

It should be noted that the separation between Rambo and Raf classes obtained for untreated samples is not so clear for treated ones. This may be because the treatment itself leads to a partial homogenization, in terms of present or major metabolites, of the treated samples. As a result, treated samples seems to be non differentiable, at visual level in score plots, in terms of Rambo/Raf classes. The hypothesis of partial homogenization in terms of metabolites due to treatment agrees with the fact that in the 4 classes models the validation samples that were wrong predicted were from treated classes, since some Raf treated samples were predicted as Rambo treated ones and vice versa, but no confusions were present for blank samples of both kinds of cultivars. Also, when components selection was performed for Rambo/Raf models, the number of variables that was conserved (41) was smaller than for the others models (63 for Blank/Treated, 121 for the quaternary model), even when the same parameters were used in all these selections, which suggests that less components could give concentration profiles that could be considered statically different or, in other words, that the treatment leads to a situation in which most of the metabolites behave similarly.

Rambo/Raf classification models are a bit more complex, since they need one more factor than Treated/Blank models. The necessity of an additional factor could be due to the fact that in the calibration step, Rambo and Raf classes are not only represented by blank samples, which represent the real nature of both kind of cultivars, but also by Rambo and Raf treated samples. As stated

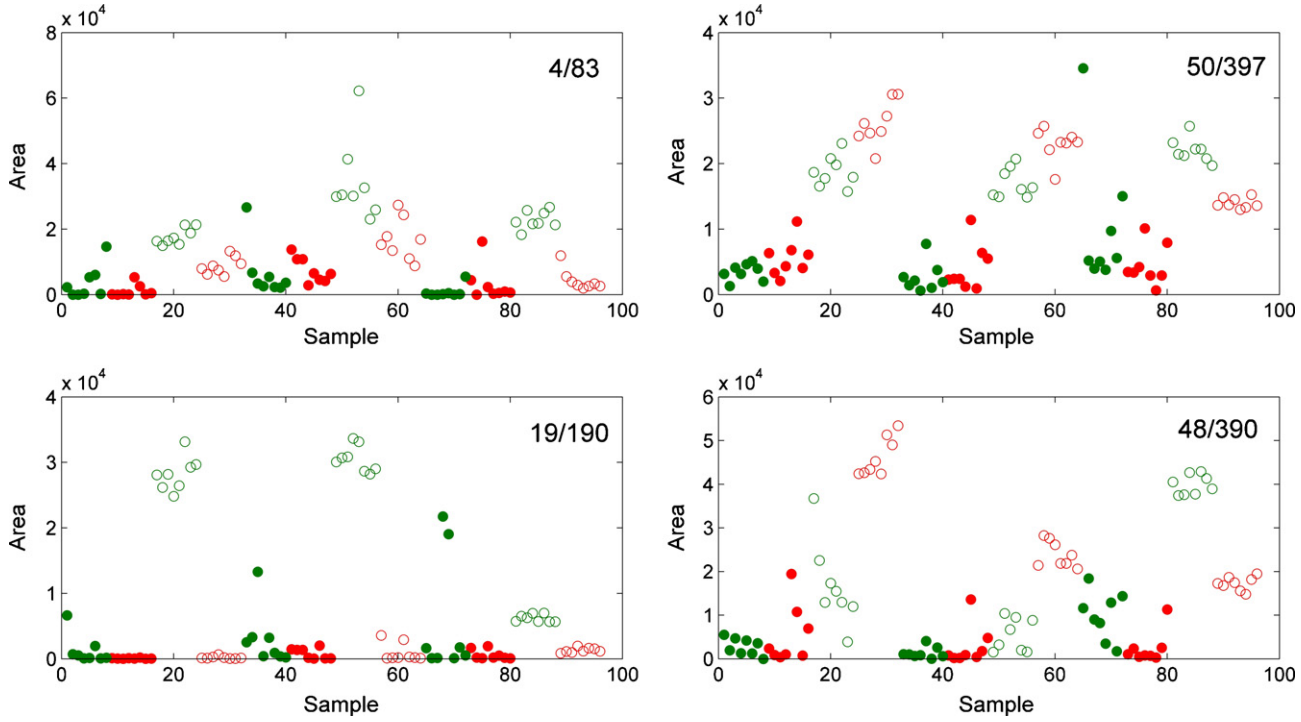

Fig. 10. Resolved component areas in the 96 available samples for some components obtained from the Treated/Blank model with components selection. Numbers such as X/Y inside each plot mean the order in the component selected subset (X) of all the 441 MCR/ALS resolved components (Y) (green = Rambo, red = Raf, filled = Treated, empty = Blank). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

before, there is a potential homogenization of those samples due to the treatment, thus it is possible that those samples were not so different at a signal level (as apparently could be the blank samples of both cultivars). In other words, treated samples of both classes would be present calibrating in different groups or representing different cultivar natures, although showing similar information. Then, in the calibration step, it could result in a sort of confusion that could be avoided by using one more factor for the classification task. Score plots for these models showed Rambo and Raf samples spatially separated. No extra separations were clearly seen as in the case of Blank/Treated models, although some samples were close to be clustered, especially with components selection and for Rambo Blank/Treated samples (data not shown).

As in the case of Treated/Blank models, a comparison between regression vectors with and without components selection (data not shown) would show that the most of the most relevant components are present on both kinds of vectors, suggesting again that the principles of selection performed by the proposed method and by PLS-DA were compatible. It was also seen that some of the relevant components for Rambo/Raf classifications were also relevant for Treated/Blank models. When the resolved component areas were evaluated for the 96 available samples, it was observed that the best separations between Rambo and Raf samples were obtained on the basis of components that were relevant only for Rambo/Raf classes and not for Treated/Blank ones. In that case, it suggests that these markers could be useful for Rambo/Raf differentiation, independently of the treatment and intrinsically related to the nature of each kind of tomato.

5. Conclusion

MCR-ALS has been applied to three-way data sets in column wise augmented matrices in a metabonomic study carried out on two different tomato cultivars (Rambo and Raf) after treatment with carbofuran pesticide. The methodology demonstrated to be appropriate to capture individual trajectories of endogenous metabolites through the time. The evolutionary profiles were obtained over the time for those metabolites that were present simultaneously in both cultivars with and without carbofuran treatment, which released useful information to make comparisons and to obtain general conclusions.

It was observed that the presence of the pesticide in tomato plants involves a physiological stress situation, which is cultivar dependent for some metabolites but on the other hand, the treatment itself could lead to a partial homogenization of samples in terms of present metabolites.

PLS-DA models could be useful to quickly classify samples as corresponding to different classes. Treated or non-treated (Blank) PLS-DA models were the best ones obtained herein, which is important for this work, since it intended to detect stress effects related to carbofuran treatment. We also noticed that a simple variable selection approach could improve the understandability of classification models and it could reveal hidden information in the data. Since classifications results were good in general, we believe that the MCR-ALS resolution obtained was representative of the main metabolic situations and that the method itself preserved the differential information contained in the distinct kinds of samples. Although components identification was not possible, relatively easy analysis of regression vectors and interactions between variables showed how to find potential markers to detect carbofuran treatment and nature of tomatoes.

Acknowledgements

The authors are grateful to Junta de Andalucia (Project AGR-973), Universidad Nacional del Litoral (Project CAI+D No. 12-65) and CONICET (Consejo Nacional de Investigaciones Científicas y Técnicas, Project PIP 2988) for financial support. G.G.S. thanks CONICET for his fellowship.

References

[1] C. Ducruix, D. Vailhen, E. Werner, J.B. Fievet, J. Bourguignon, J.C. Tabet, E.E. Ezan, C. Junot, Chemomtr. Intell. Lab. Syst. 91 (2008) 67.
[2] P. Jonsson, J. Gullberg, J. Nordström, M. Kowalczyk, M. Sjöström, T. Moritz, Anal. Chem. 76 (2004) 1378.
[3] P. Jonsson, S.J. Bruce, T. Moritz, J. Trygg, M. Sjöström, R. Plumb, J. Granger, E. Maibaum, J.K. Nicholson, E. Holmes, H. Antti, Analyst 130 (2005) 701.
[4] J. Halket, A. Przyborowska, S.E. Stein, W.G. Mallard, S. Down, R.A. Chalmers, Rapid Commun. Mass Spectrom. 13 (1999) 279.
[5] H.L. Shen, B. Grung, O.M. Kvalheim, I. Eide, Anal. Chim. Acta 446 (2000) 313.
[6] P. Jonsson, E. Sjövik Johansson, A. Woulikainen, J. Lindberg, I. Schuppe-Kiostinen, M. Kusano, M. Sjöström, J. Trugg, T. Moritz, H. Antti, J. Proteome Res. 5 (2006) 1407.
[7] I. Sánchez Pérez, M.J. Culzoni, G.G. Siano, M.D. Gil García, H.C. Goicoechea, M. Martínez Galera, Anal. Chem. 81 (2009) 8335.
[8] O. Galtier, O. Abbas, Y. Le Dréau, C. Rebufa, J. Kister, J. Artaud, N. Dupuy, Vib. Spectrosc. 55 (2011) 132.
[9] M. Barker, W. Rayens, J. Chemometrics 17 (2003) 166.
[10] W. Ni, S.D. Brown, R. Man, Anal. Chem. 81 (2009) 8962.
[11] D.M. Haaland, E.V. Thomas, Anal. Chem. 60 (1988) 1193.
[12] J.A. Arancibia, C.E. Boschetti, A.C. Olivieri, G.M. Escandar, Anal. Chem. 80 (2008) 2789.
[13] A.K. Smilde, R. Tauler, J.M. Henshaw, L.W. Burgess, B.R. Kowalski, Anal. Chem. 66 (1994) 3345.
[14] J. Jaumot, R. Gargallo, A. de Juan, R. Tauler, Chemometr. Intell. Lab. Syst. 6 (2005) 101.
[15] B.M. Wise, N.B. Gallagher, R. Bro, J.M. Shaver, W. Windig, R.S. Kock, PLS_Toolbox 3.52, Eigenvector Research, Manson, WA, 2005.
[16] M. Anastassiades, S.J. Lehotay, J. AOAC Int. 86 (2003) 412.
[17] S.J. Lehotay, K. Mastovská, A.R. Lightfield, J. AOAC Int. 88 (2005) 615.
[18] J. Westerhuis, H. Hoefsloot, S. Smit, D. Vis, A. Smilde, E. van Velzen, J. van Duijnhoven, F. van Dorsten, Metabolomics 4 (2008) 81.
[19] E. Anderssen, K. Dyrstad, F. Westad, H. Martens, Chemometr. Intell. Lab. Syst. 84 (2006) 69.
[20] Registro de Productos fitosanitarios, No° Registro: 23.092, Nombre comercial: Botran 20 SC.
[21] Dirección General de Agricultura (Comisión de la Comunidad Europea) Anexo 1.
[22] S. Mallat, IEEE Trans. Pattern Anal. Machine Intell. 11 (1989) 674.
[23] E. Peré-Trepart, R. Tauler, J. Chromatogr. A 1131 (2006) 85.