Sensors and Actuators B 155 (2011) 183–190

Contents lists available at ScienceDirect

Sensors and Actuators B: Chemical

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

On data analysis in PTR-TOF-MS: From raw spectra to data mining

Luca Cappellin[a,b], Franco Biasioli[a,*], Pablo M. Granitto[c], Erna Schuhfried[b], Christos Soukoulis[a], Fabrizio Costa[a], Tilmann D. Märk[b], Flavia Gasperi[a]

[a] IASMA Research and Innovation Centre, Fondazione Edmund Mach, Food Quality and Nutrition Area, Via E. Mach, 1, 38010, S. Michele a/A, Italy
[b] Institut für Ionenphysik und Angewandte Physik, Leopold Franzens Universität Innsbruck, Technikerstr. 25, A-6020 Innsbruck, Austria
[c] CIFASIS, French Argentina International Center for Information and Systems Sciences, UPCAM (France)/UNR-CONICET (Argentina), Bv 27 de Febrero 210 Bis, 2000 Rosario, Argentina

ARTICLE INFO

Article history:
Received 20 August 2010
Received in revised form
12 November 2010
Accepted 22 November 2010
Available online 1 December 2010

Keywords:
Proton transfer reaction-mass spectrometry
Time of flight
Data analysis
Data mining
Volatile organic compounds

ABSTRACT

Recently the coupling of proton transfer reaction ionization with a time-of-flight mass analyser (PTR-TOF-MS) has been proposed to realise a volatile organic compound (VOC) detector that overcomes the limitations in terms of time and mass resolution of the previous instrument based on a quadrupole mass analysers (PTR-Quad-MS). This opens new horizons for research and allows for new applications in fields where the rapid and sensitive monitoring and quantification of volatile organic compounds (VOCs) is crucial as, for instance, environmental sciences, food sciences and medicine. In particular, if coupled with appropriate data mining methods, it can provide a fast MS-nose system with rich analytical information. The main, perhaps even the only, drawback of this new technique in comparison to its precursor is related to the increased size and complexity of the data sets obtained. It appears that this is the main limitation to its full use and widespread application. Here we present and discuss a complete computer-based strategy for the data analysis of PTR-TOF-MS data from basic mass spectra handling, to the application of up-to date data mining methods. As a case study we apply the whole procedure to the classification of apple cultivars and clones, which was based on the distinctive profiles of volatile organic compound emissions.

1. Introduction

In recent years increasing attention has been paid to the development of proton transfer reaction-mass spectrometry (PTR-MS), a rapidly developing technique which is proving, thanks to its short response time and low detection limits, quite successful for the on-line monitoring of volatile organic compounds (VOCs) [1] and for the rapid fingerprinting of the head space of complex samples, in particular food samples [2]. In fact, PTR-MS allows for the real-time detection of VOC concentrations at pptv levels [1]. It is based on proton transfer from the H_3O^+ ion, which undergoes non-dissociative proton transfer reactions with most VOCs, while it does not react with the constituents of clean air. PTR-MS, in its quadrupole based version (PTR-Quad-MS) has been applied to a large variety of fields, ranging from environmental research [3] to medical applications [4,5] and food analysis [6,7]. The PTR-Quad-MS instrument is characterized by unit mass resolution and simple spectral fingerprints consisting in fact of histograms with 100–200 bins. Several studies showed that PTR-Quad-MS fingerprints can be efficiently coupled with data mining or multivariate methods to develop useful models for sample classification [2] or for correlation with other characteristics of the samples, such as sensory analysis [8] or genetic information [9]. Recently, a new instrument based on a time of flight mass analyser has been proposed [10] and successfully commercialized as PTR-TOF-MS [11]. The main advantages of this new instrument are the higher mass range, the faster measuring time (a complete mass spectrum in a split second) and the higher mass resolution, which multiply the analytical information contained in the spectra. However, such advantages come at the expense of having to deal with larger and more complex spectra. Hence arises the necessity to develop new procedures to extract manageable datasets, which can be employed in preliminary data visualization and analysis or as inputs for data mining procedures. From our experience with PTR-TOF-MS [12,13] we know that handling the data produced by PTR-TOF-MS is a difficult task which can hinder the widespread use of this interesting technique and the exploitation of its full potentiality.

In the present work we describe in detail a full methodology, summarised in Fig. 1, that can be applied directly to standard PTR-TOF-MS spectra generated from commercial instruments, i.e., from spectra to data visualization/exploration and application of up-to-date data mining methods. It starts with the internal calibration of PTR-TOF-MS spectral data, in accordance with a recently proposed method [14], followed by data pre-processing, such as denoising and baseline removal. For the spectra treated in this manner we then developed a semi-automatic method for peak

* Corresponding author. Tel.: +39 0461 615187; fax: +39 0461 650956.
E-mail address: franco.biasioli@iasma.it (F. Biasioli).

0925-4005/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.snb.2010.11.044

identification and peak area extraction, which can be used to produce data matrices for preliminary data exploration or to feed data mining algorithms which will also be further discussed. Finally, as a case study, we apply the whole procedure to a practical problem: the classification of apple cultivars and clones. Here we do not concentrate solely on preliminary data analysis and PTR-TOF-MS characterisation that have been, at least partly, addressed elsewhere [11,15]. Nor do we describe the fundamentals of PTR-Quad-MS or PTR-TOF-MS which have been thoroughly described and reviewed elsewhere [1,11,16]. We rather present the development and test of a complete procedure for data analysis and data processing for PTR-TOF-MS that can be applied "as is" to fundamental and applied issues as a very fast and highly sensitive MS-nose with rich analytical information. The functions to implement the proposed procedure in MATLAB (R2007a) will be freely provided upon request. This paper is organized as follows: in the next section we describe the proposed procedure for data analysis and data mining of PTR-TOF-MS data; in Section 3 we discuss a case study using different apple cultivars and clones, closing the work in Section 4 with our conclusions.

2. PTR-TOF-MS data analysis

2.1. Internal calibration

Fundamental issues in mass spectrometry are mass calibration and mass accuracy. Indeed, through them the observed spectrometric peaks can be identified with the help of mass to charge ratios, isotope ratios and fragmentation patterns. For the small (up to 300 Da) volatile organic compounds commonly monitored by PTR-MS analysis, a mass accuracy of 5 ppm is usually sufficient for the exact determination of the elemental composition. In the case of PTR-TOF-MS spectra the advantage of attaining a good mass accuracy is twofold. Besides compound identification, it also reduces the peak shifts between different spectra thus allowing for considerable signal to noise ratio improvements by averaging many spectra from the same sample. Mass accuracy of PTR-TOF-MS raw data is limited to external calibration, which refers to fixing a proper set of calibration coefficients which are employed during the entire duration of mass spectral acquisition. However, our experience shows that, due to a lack of stability in instrumental parameters, external calibration in the commercial PTR-TOF-MS instrument does not guarantee mass accuracy for a sufficiently long time. A common solution to this problem is the use of an internal calibration based on the known exact mass of selected ions. In the case of PTR-MS useful choices are, for instance, NO^+, $O_2{}^+$ and protonated acetone at nominal masses 30, 32 and 59, which are always present at reasonable concentrations [17]. Other ions can be used if, for example, GC analysis can identify the nature of some of the more intense peaks. External calibration is useful for the automatic identification of such peaks within the PTR-TOF-MS spectrum. Internal calibration then proceeds straightforwardly using the formula [14]

$$At^2 + Bt + C = \frac{m}{z},$$

where A, B, C are fitting constants, to fit theoretical mass/charge (m/z) values versus measured time of flights of the selected calibration peaks. Time of flights are estimated by determined peak centroids using for example a Gaussian function to fit the peak shape. When selecting calibration peaks it is important to avoid saturated peaks and low intensity peaks, for which centroids cannot be properly determined. In practice we consider only peaks of which the maximum height lies in the 10–1000 cps range. In practise this ensures that the peak centroid can be properly estimated, in particular that the peak height is large enough for the peak to be well shaped and small enough to avoid saturation effects [18]. For further details see Ref. [14]. MATLAB data processing capabilities have been used directly (as the interface to HDF5 files) to implement specific functions for calibration.

2.2. Noise reduction and baseline removal

PTR-TOF-MS spectra are affected by two main sources of error: electronic random noise and saturation effects.

Various solutions exist to overcome the former problem. If many spectra of the same sample are available, random noise is reduced and the quality of the signal is improved by simply averaging over the spectra. However, there is an important caveat. The averaging procedure is appropriate only if the spectra are properly aligned in terms of mass scale. Here the discrete nature of TOF-MS sig-

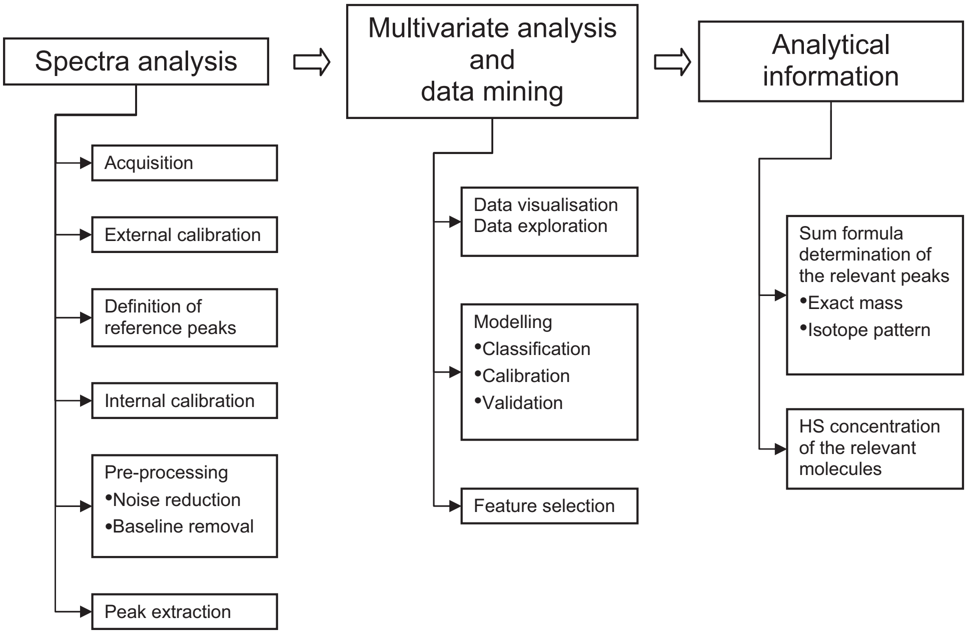

Fig. 1. Schematics of the data analysis methodology presented.

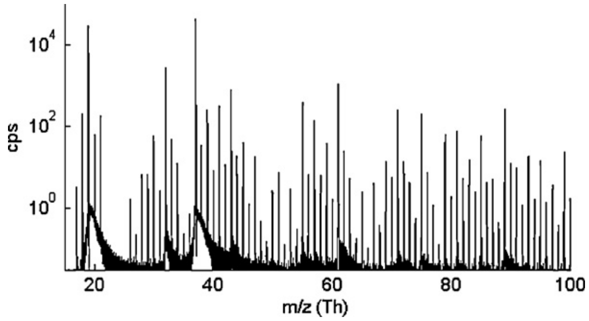

Fig. 2. Example of PTR-TOF-MS spectrum (average of 25 spectra belonging to the same sample). For the sake of clarity, only the low mass region is displayed.

nals plays an important role, becoming a limitation and a possible source of error. A spectrum can be composed of about 400.000 (as in the data presented below) or more non-negative numbers, each referring to an equally separated range of ion time of flights. Internal calibration assigns a mass over charge value corresponding to each bin. But bin positions may be in principle different for different spectra. Therefore, averaging over spectra first requires setting a m/z axis common to all spectra and then evaluating the spectral signals at selected m/z values. Proper alignment would require to know the signal shape *a priori,* so that no error will arise from this step. This not being the case, one choice would be to take the mean or the distance-weighted mean of two signal values at the closest m/z. However, if the purpose is not only to obtain a good calibration but also to properly estimate compound concentrations, computing a mean is not completely appropriate since it may cause a distortion in the ion counts. Thus we propose to not modify ion counts, and to simply shift the spectrum to the closest among the selected m/z values. We propose this method as a reasonable trade off between optimum calibration and ion count uncertainties, the error induced to the calibration being less than 1 ppm.

Further noise reduction may be achieved by using standard denoising methods [19,20]. In particular, wavelet denoising [21] has proved useful to smooth the spectrum without causing large distortions. In the spectral data processing and analysis procedure proposed in the present work, wavelet denoising is only employed as an intermediate step to slightly improve the performance of the automatic peak finder even if this is often not necessary for PTR-TOF-MS data.

Fig. 2 exemplifies that PTR-TOF-MS spectra are characterized by a baseline that especially affects those peaks that are close to saturated peaks. It is safe to say that standard applications of PTR-TOF-MS do not lead to saturated peaks at masses above approximately 250 a.m.u. In the case of masses below this value, there exist no compounds that have m/z values at semi-integer m/z. It is therefore straightforward to build the baseline for each nominal mass by using the spectral signal at m/z values at about the following and the preceding semi-integer values, and then applying a polynomial fit to approximate the baseline at the m/z value under consideration. The baseline is then subtracted. For high nominal masses the baseline is close to zero and therefore there is no need to correct against it. Specific MATLAB functions have been implemented for these data pre-processing procedures.

2.3. Peak detection

The PTR-TOF-MS technique is designed to measure intensities of ions corresponding to the protonated form of volatile organic compounds (VOCs) [11,1]. The peak position is determined by the m/z value of the protonated VOC while the peak area represents the number of ions that reach the MS detector during the set acquisition time. Peaks referring to isobaric compounds and more generally to compounds having close enough m/z values, are superimposed within the spectrum, resulting in highly complex peak structures. Disentangling such spectral signals and reconstructing signal peaks corresponding to different VOCs is a difficult, and sometimes impossible task [15]. However, this step is compulsory when the purpose is to estimate VOC concentrations in the measured sample. But, on the other hand, it is not needed if the goal is to provide a fingerprint of the sample. For each spectrum, we propose to construct a standardized data matrix based on a partial disentanglement of spectral peak structures. We start with the investigation of PTR-TOF-MS peak shapes and in particular of peak widths, which is related to the instrument's mass resolution, defined as

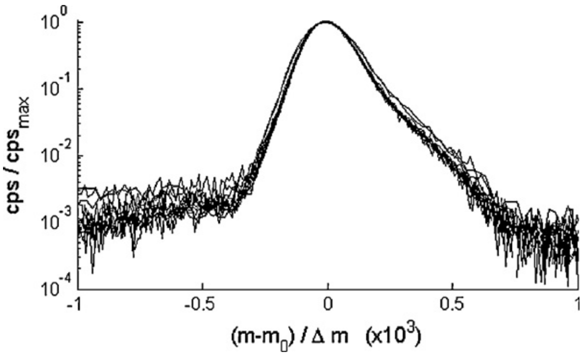

Fig. 3. Shape of PTR-TOF-MS spectral peaks. If isolated peaks in the same spectrum are rescaled by the maximum height and by the centroid m/z value, they approximately converge onto the same curve. Twelve peaks at nominal masses in the range 18–205 a.m.u. are represented in this figure with the following m/z values: 18, 21, 59, 89, 103, 117, 131, 145, 159, 173, 187 and 205. The peaks displaying slightly larger FWHM are those at lower nominal masses (18, 21 and 59).

$$R = \frac{m}{\Delta m},$$

where Δm is the peak full width at half maximum (FWHM).

We observed a roughly constant mass resolution over a wide range of masses. However, our analysis shows that there is some evidence for deviation: the resolution is generally found to be slightly increasing with mass, variations usually being less than 20% in the 1–400 Da mass range. This is in agreement with suggestions of Coles and Guilhaus [22], who also provided a proper modelling of such non-linearities. The implementation of this model in our strategy via the fitting of the proposed function to the peak widths in the averaged spectrum is straightforward and has been used recently [15]. In the following we will use the same approach. It is however useful to notice that, in practical cases, the use of a constant resolution will avoid the preliminary calibration and yield basically the same results. A further effect influencing peak widths and thus the estimation of instrumental resolution is the detector dead time [23]. Commercially available PTR-TOF-MS instruments employ Poisson statistics to correct against dead time effects. A well known effect of dead time is to distort the peak shape, therefore if the applied correction is not completely appropriate, a change in the estimated resolution may be caused. This is likely to happen at high ion intensities, typically a few ions/pulse, where the standard corrections based on Poisson statistics is not appropriate [18]. In order to overcome this problem we do not consider peaks displaying an intensity larger than 5 ions/pulse.

The overall peak shape is another issue for which, to our knowledge, there is no definitive solution in the literature. PTR-TOF-MS peaks are approximately Gaussian, especially in the top part, while asymmetries often appear as right-side (higher m/z) tails. An important observation is that peak shapes are roughly independent on ion intensities and ion m/z values. Fig. 3 exemplifies this: when isolated peaks in the same spectrum are rescaled by the maximum height

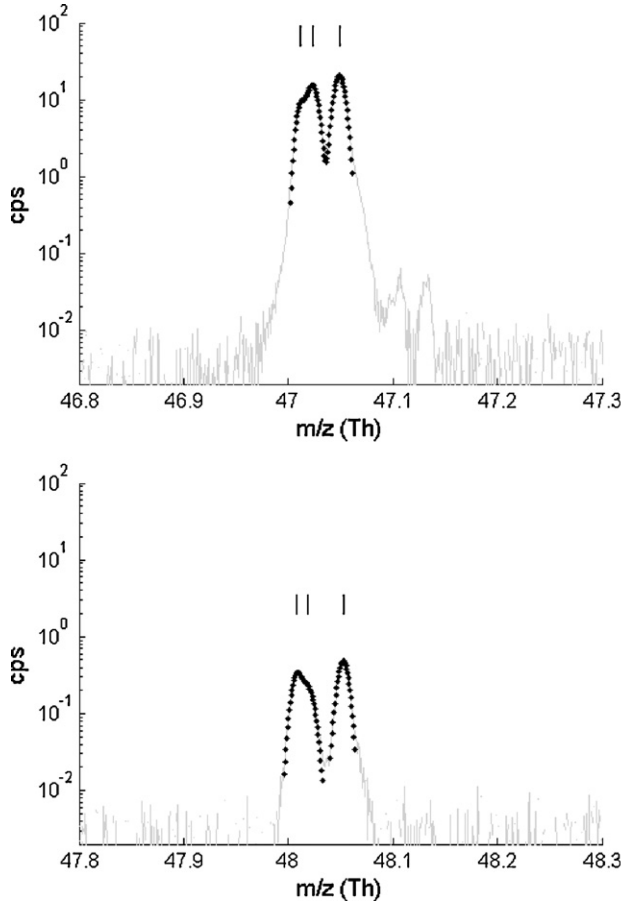

Fig. 4. Example of results from the peak detection algorithm described in the present work. At nominal mass 47 (upper panel) three peaks are detected and a linear sum of three Gaussian functions is fitted. Vertical lines highlight the estimated positions of the corresponding peak centroids. Results for nominal mass 48 corresponding to the mono substituted isotopologues (lower panel) are also reported. Note that in this case peak detection would not be possible using a single spectrum instead of the average of 25 spectra, the signal intensity always being below 1 cps which is close to the noise level in single spectra.

and by the centroid m/z value, they approximately overlap, which indicates that proper modelling of the peak shape is possible. In the present work, we employ Gaussian functions to fit mass spectrometric peaks, since this provides a satisfactory shape matching while reducing computational effort. Other possible choices are for instance to employ modified Gaussian functions [15] or functions consisting of a Gaussian part and an extreme value distribution [24].

Let us assume that the peak shape and the resolution are known for a given nominal mass within the spectrum. A strategy to partially disentangle a complex spectrometric peak structure, as for example the one shown in Fig. 4, could be the following. We choose the number N of peaks that are likely superimposed in the structure. Peak widths are given by the resolution. Since peak shapes are known, only two parameters per peak have to be calculated (centroid and height), amounting to $2N$ parameters. These parameters are obtained by fitting the peak structure with a linear sum of peaks of the known shape. On the basis of the goodness of the fit we can optimize the number of peaks. Given that we are using a Gaussian function instead of the exact peak curve and that the resolution can be affected by an error of up to about 20%, the real application of this method is more complex than in the given idealized example. The achievable disentanglement is in general not complete, and depends on the signal to noise ratio and on the errors in the determination of peak shape and width. The error affecting the resolution is accounted for by allowing the peak width parameters to vary within a range corresponding to an error of 20% in the resolution. The problem concerning the approximation of peak shapes with Gaussian functions cannot be solved in absence of a reliable theory on peak shapes and such a refinement could also prevent, at the present time, the complete automation of the procedure. We propose therefore to employ the automatic procedure as a preliminary but rather efficient screening, which then requires a further refinement or validation, for instance taking advantage of the graphical facilities provided by MATLAB (see below). Specific MATLAB functions have been developed for peak finding and disentanglement.

2.4. Multivariate analysis and data mining

The data matrix (rows = number of samples, columns = number of identified peaks) constructed via the described procedure is relatively large (usually more than 1000 columns) and characterized by strong collinearity induced by diverse causes as, for example, fragmentation. Thus, multivariate analysis and data mining methods are needed for data exploration and visualization, in order to set classifications or calibration models and for feature extraction.

A preliminary analysis can be carried out by employing principal component analysis (PCA) [25] which is often suitable for deriving and visualizing unsupervised classification models. However, we expect that supervised classification methods should, in general, give a better performance. Among them we mention Random Forest (RF) [26], Penalized Discriminant Analysis (PDA) [27], Discriminant Partial Least Squares (PLS) [27] and Support Vector Machines [28], all of which have already been successfully applied to PTR-Quad-MS data [2,29]. Validation of supervised methods is necessary and, in the case of relatively small samples sets, leave-one-out (LOO) validation is often the best choice: at each step a sample is selected as test set and excluded from the data used to fit the models; these are then used to classify the independent test samples. Classification results are analysed by means of confusion matrices, in which the true classes correspond to rows and the predicted ones correspond to columns [30]. Results are given as the number of samples belonging to the class or subclass identified by the row title that is assigned by the classifier to the class or subclass identified by the column title, so that the diagonal entries of the confusion matrix correspond to correct classifications.

An advantage of RF as compared with other classification tools, is the possibility of an unbiased graphical investigation of the classification efficiency, in analogy to the well known PCA. As pointed out by Granitto et al. [31], RF complements PCA analysis since it employs information about real classes. At the same time, visualizations from RF is less biased than that of PDA and PLS, because RF bases its output plots on unseen samples.

Other than knowing if samples can be discriminated, it is always important to know which are the peaks that contribute most to the separation of the different classes. An efficient way to do this is to use an appropriate feature selection method, as for example Random Forest-Recursive Feature Elimination (RF-RFE), introduced by Granitto et al. [29]. This is a recursive method, which at each step, selects (and eliminates) the less relevant peaks in the input data using information extracted from an RF model fitted to the same data. Technical details of the method can be found in Granitto et al. [29], where the authors also showed that RF-RFE can identify the most relevant peaks in a multivariate and collinear data matrix, even in situations when the number of samples is much lower than the number of measured peaks. It is worth mentioning that the peaks obtained by RF-RFE are only the most relevant ones to the problem and not the ones that lead to the smallest error in the classification models. Other versions of RFE, for example the original SVM-RFE [32] can also be used at this step. Here, multivariate analysis and data mining were carried out here using R packages [33].

3. Case study: apple cultivars

As an example we discuss in detail the entire procedure as applied to a PTR-TOF-MS dataset obtained by the VOCs profiling of apple fruits.

We consider three genetically distinct varieties indicated as 1, 2 and 3. For the first one we have two different clones indicated by 1-A and 1-B. Since clones often generate from point mutagenic events, they share almost the same genetic constitution and expected differences are smaller than the ones among cultivars. This results in four different classes and for each class we measured 8 single apple fruits.

3.1. PTR-TOF-MS measurements

Measurements were performed two months after harvest as described in [14], with a commercial PTR-TOF-MS 8000 apparatus supplied by Ionicon Analytik GmbH, Innsbruck (Austria) in its standard configuration [11]. The sampling time per channel in the TOF is 0.1 ns, amounting to about 350,000 channels for a mass spectrum up to 400 Th, and the proton transfer ionization conditions are controlled by drift voltage (600 V), drift temperature (110 °C) and drift pressure (2.25 mbar).

For VOCs profiling fruits were taken after two months of cold storage and kept at room temperature overnight before the actual measurement. Each single fruit was placed in glass jars (1000 ml, 30 °C) supplied with two teflon/silicone septa on opposite sides. After 30 min a Peek tube (110 °C, 0.055″ diameter) was connected between the inlet of the PTR-MS and the headspace collecting glass jars via the septum and 100 standard cubic centimetres of headspace air per minute were continuously extracted for 25 s via one of the septa, whereas the other septum was connected to clean air, allowing to obtain 25 TOF spectra from virtually 0 up to 400 Th within this time. Every single spectrum is the sum of 28,600 acquisitions lasting for 35 μs each: the complete dataset consists of 25 spectra for each apple sample, amounting to 800 complete PTR-TOF-MS spectra each consisting of 350,000 points.

3.2. Calibration and pre-processing

Mass calibration was carried out and all spectra were aligned on a common m/z axis according to the above described procedures. The identification of spectrometric peaks to be used in the calibration step in the case of apple fruits has already been described elsewhere [14]. Noise was reduced by averaging over the 25 spectra belonging to the same sample (element), that is to the same apple fruit, and the baseline was then removed. In the following we will always refer to these average baseline corrected spectra. A mean spectrum of all three classes was also computed and an average resolution was calculated by fitting Gaussian functions to estimate FWHM of calibration peaks. Our algorithm for automatic peak finding was then applied to such a mean spectrum. Fig. 4 exemplifies the results for nominal mass 47 and 48. For instance, for mass 47 we found three peaks at m/z = 47.013, 47.023, 47.049 Th, that, in agreement with the findings of Herbig et al. [17], are compatible with $CH_3O_2^+$, $H_3N_2O^+$, $C_2H_7O^+$, respectively. See below for further details on the assignment of the sum formula to detected peaks. Our interactive graphical tool was used for the refinement steps in order to include peaks that are clearly visible in single element spectra, while being averaged out in the mean spectrum, and to eliminate spurious peaks which are proposed by the peak finding program because of the already mentioned discrepancies between a Gaussian function and the proper peak curve. This is the only non-automated and time consuming step that can require up to 3 h of work for a trained operator. However, this step is needed only once for a given sample typology (apples in this case).

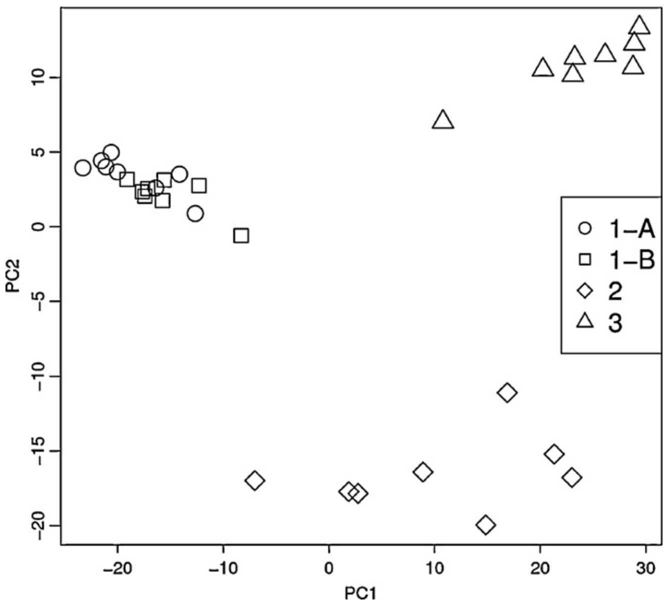

Fig. 5. First and second component of the PCA analysis of the PTR-TOF-MS spectral data of all samples.

3.3. Peak extraction

Finally, automatic peak extraction is performed on the spectra corresponding to each of the elements, but this time the number of peaks and their positions are fixed from the results obtained after the refinement step on the mean spectrum: for each nominal mass a linear sum of Gaussian functions with fixed mean parameters is fitted and the peak heights are estimated as the maximum height of the corresponding Gaussians. Peak widths are then determined from an average resolution estimated as in the case of the mean spectrum (see Section 2.3). This practice gives more reliable estimates than using the peak widths resulting from the fits. Therefore at the end of the whole procedure a data matrix of peak areas is constructed, each row corresponding to different elements (32 apples in this case) and each column corresponding to a peak with a defined m/z value (951 peaks were detected in this case study). Such a matrix is the starting point for further data analysis and data mining.

3.4. Multivariate analysis and data mining

A first analysis of the 32 samples is carried out via PCA, which clearly highlights the separation into three different classes (see Fig. 5). However, the two subclasses composing class 1 are superposed and cannot be distinguished in this analysis. Hence the need arises for supervised classification methods.

Fig. 6 shows the first two components of the RF visualizations of these data sets. The figure clearly confirms that two classes cannot be easily separated, as suggested by the PCA analysis. However, in this supervised visualization the two subclasses of class 1 seem to be more separated than in the PCA analysis, which is in agreement with the results of the four discriminant methods and indicated by the confusion matrices in Table 1. The classification errors are 9%, 6%, 12% and 9% for RF, PDA, PLS and SVM, respectively. The three classes are perfectly separated, while small errors arise in the

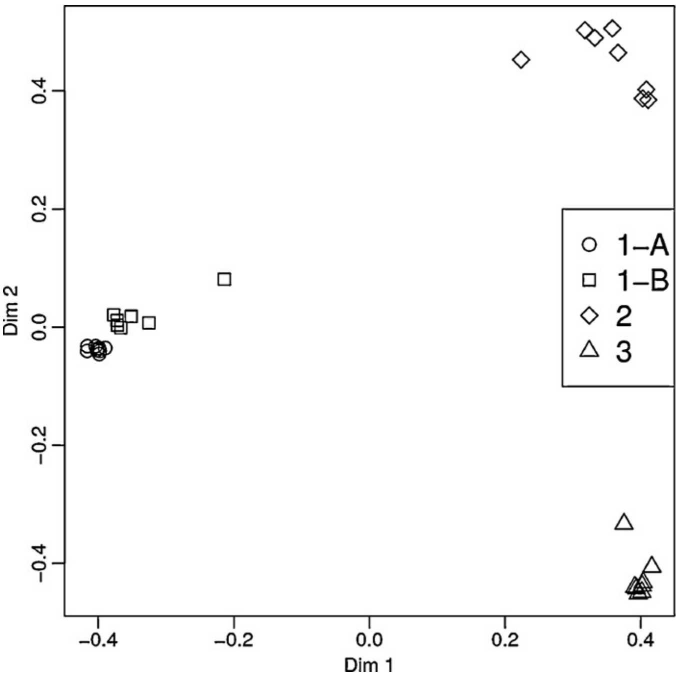

Fig. 6. Random Forest graphical output for the discriminant analysis of the PTR-TOF-MS spectral data of all samples.

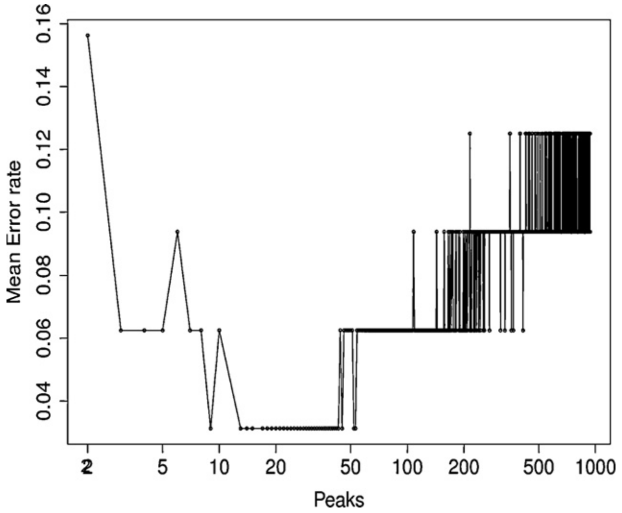

Fig. 7. Mean classification error rate as a function of the number of peaks selected by the RF-RFE method for the 3 classes (32 samples) problem. Note the log scale on the number of peaks.

classification of the samples belonging to the two subclasses (in agreement with the visual analysis in Fig. 6).

The selection of discriminant peaks is done in two steps. First, an analysis of the mean discrimination error of the RF-RFE method as a function of the number of masses in the problem is shown in Fig. 7. This figure is used to select a number of masses for the discrimination models that should be as low as possible, but also with as little loss in discriminant capabilities as possible. In our case study the selection is easy, as the minimum average error rate is obtained with only 9 masses. In the second step, we analyse which are the most important peaks for models constructed by RF-RFE using 9 masses. The results are reported in Fig. 8. The figure indicates how often (over the 32 LOO replicated RF-RFE experiments) masses listed in the figure are selected among the 9 most relevant by RF-RFE. In order to simplify the subsequent analysis, we only considered the 12 masses that appear in more than 20% of the replicated experiments, which can be considered as the more relevant for the separation of the four classes. Of course, this limit can be changed in order to broaden or narrow the number of peaks considered for further analysis.

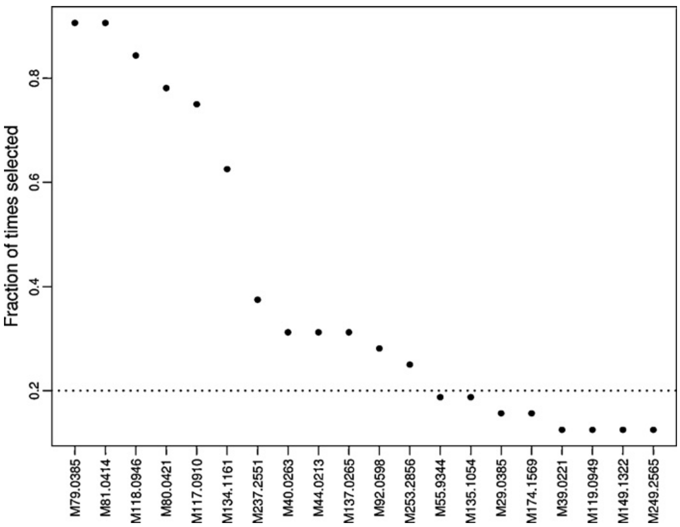

Fig. 8. Fraction of times that each peak was selected among the 9 more discriminant features over the 32 LOO replicated experiments, for the 3 classes problem. The dotted line shows the 20% lower limit considered in this work.

We also investigate which peaks are responsible for the discrimination of the two subclasses 1-A and 1-B. For addressing this issue we applied RF-RFE to the 16 samples composing class 1. We repeated the two-step procedure described before for this reduced problem (figures not shown). In this case we considered subsets with only 5 peaks. The second step highlighted masses 79.0385, 81.0414, 40.0263, 80.0421 and 107.0519 as the most relevant for the problem. Some of the peaks correspond to isotopes of the same component, and have, correspondingly, a very high correlation. This is a limitation of this analysis based on RFE rankings, that in some cases is not able to keep only one of a pair of correlated but highly discriminant masses.

An interesting case for further analysis is mass 107.0519. The previous results showed that it is one of the five more relevant peaks for the discrimination of the two subclasses. In Fig. 9, we show the signals at nominal mass 107 for the samples belonging to subclasses 1-A and 1-B. The discriminant peak $m/z = 107.0519$ indicated by the RF-RFE method appears superimposed to a more intense peak, the superimposed signals of subclass 1-A being higher

Table 1
Confusion matrices for the classification by RF, PDA, PLS and SVM of all apple samples. 1-A and 1-B indicate the clones of the same cultivar while 2 and 3 are the other two cultivars.

RF	1-A	1-B	2	3	PDA	1-A	1-B	2	3	PLS	1-A	1-B	2	3	SVM	1-A	1-B	2	3
1-A	6	2	0	0	1-A	7	1	0	0	1-A	6	2	0	0	1-A	6	2	0	0
1-B	1	7	0	0	1-B	1	7	0	0	1-B	2	6	0	0	1-B	1	7	0	0
2	0	0	8	0	2	0	0	8	0	2	0	0	8	0	2	0	0	8	0
3	0	0	0	8	3	0	0	0	8	3	0	0	0	8	3	0	0	0	8

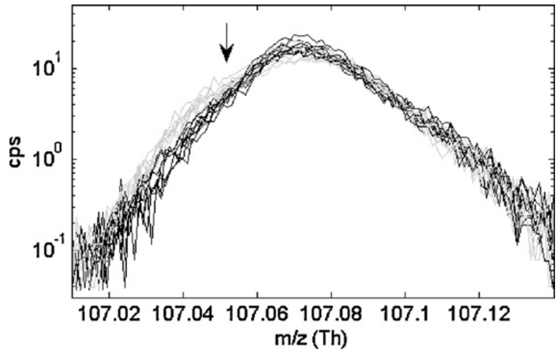

Fig. 9. PTR-MS-TOF signals at nominal mass 107 for class 1. Gray and black spectral extracts refer to elements belonging to subclass 1-A and 1-B, respectively. The arrow indicates the position of the discriminant peak identified by our methodology coupled with RF-RFE.

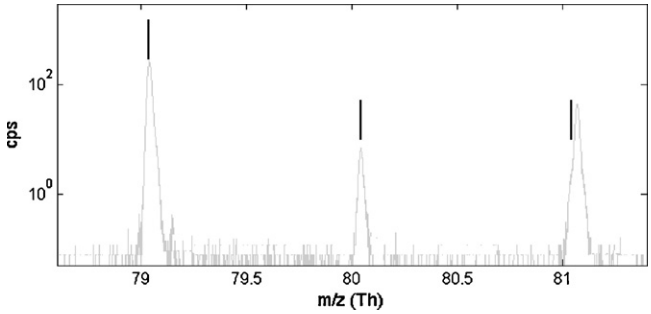

Fig. 10. Example of PTR-TOF-MS signal at nominal masses $m/z = 79$, 80, 81 Th. The black vertical lines mark the theoretical peak positions for $C_2H_7O_3^+$ ($m/z = 79.0390$ Th) and its isotopologues. Note that the isotope at nominal $m/z = 81$ Th appears as a (left) shoulder of a more intense peak.

than those of subclass 1-B. Nevertheless, a marked difference between the spectra of the two subclasses is now evident. Our variable selection methodology proves able to highlight the role of relevant features that would be difficult to detect with more naïve approaches. The presence of superimposed peaks is indeed very common in PTR-TOF-MS spectra and it is therefore important to have a tool that allows the extraction of relevant information from complex peak structures.

As an example of the analytical information provided by PTR-TOF-MS we can consider three top masses identified by RF-RFE in Fig. 8: 79.0385, 80.0421 and 81.0414. An illustrative PTR-TOF-MS signal at those masses for one sample is depicted in Fig. 10. With the given mass accuracy (better then 0.001 Th for $m/z = 79$ Th [14]) the chemical formulas encompassing $C_{0\text{-}10}H_{0\text{-}100}O_{0\text{-}10}S_{0\text{-}10}N_{0\text{-}10}$, compatible with 79.0385 are $C_2H_7O_3^+$ ($m/z = 79.0390$ Th) and $H_5N_3O_2^+$ ($m/z = 79.0376$ Th). The isotopic abundances calculated for M+1 (2.41% of the base peak) and M+2 peak (0.70%) then allow to rule out $H_5N_3O_2^+$. We could have reached the same conclusion by noticing that mass 80.0421 is compatible with the mass of the isotope of $C_2H_7O_3^+$ but not with that of the isotope of $H_5N_3O_2^+$.

4. Conclusions

PTR-TOF-MS is a powerful method that can be used for the very rapid classification and characterisation of samples on the basis of their fingerprint obtained by measuring their head space volatile organic compound profiles. The main obstacle for its wider use is probably the size and complexity of the data sets generated and the lack of clear receipts for data analysis. Here we described a complete methodology ranging from data pre-processing and peak detection to data mining and feature selection. The key aspect of our procedure is the analysis as a whole since it straightforwardly allows for extracting highly relevant features from large amounts of complex spectral data. Our methodology is not restricted to, but has proved very successful in many food science and technology studies, ranging from cheese samples [13] to yoghurt fermentation studies [12]. Future developments of the presented method should include the use of a proper peak shape, not yet available, instead of the nonetheless well established Gaussian approximations. The investigation and correction of the effects caused by the dead time of the detector are another issue that should be addressed in order to improve and extend the capabilities of our methodology. Many relevant aspects affecting the analysis, such as, e.g., the employed primary ion/ions and the proper compound concentration determination, have not been discussed here since they are not typical of TOF analysers and have already been addressed elsewhere [1,34,35].

Acknowledgments

This work is partly supported by Autonomous Province of Trento (APR 2009/2010). PMG thanks partial support from ANPCyT (Grant PICT 643).

References

[1] W. Lindinger, A. Hansel, A. Jordan, On-line monitoring of volatile organic compounds at pptv levels by means of proton-transfer-reaction mass spectrometry (PTR-MS) – medical applications, food control and environmental research, Int. J. Mass Spectrom. 173 (1998) 191–241.

[2] P. Granitto, F. Biasioli, E. Aprea, D. Mott, C. Furlanello, T. Mark, et al., Rapid and non-destructive identification of strawberry cultivars by direct PTR-MS headspace analysis and data mining techniques, Sens. Actuators B: Chem. 121 (2007) 379–385.

[3] J. De Gouw, C. Warneke, Measurements of volatile organic compounds in the earth's atmosphere using proton-transfer-reaction mass spectrometry, Mass Spectrom. Rev. 26 (2007) 223–257.

[4] A. Critchley, T. Elliott, G. Harrison, C. Mayhew, J. Thompson, T. Worthington, The proton transfer reaction mass spectrometer and its use in medical science: applications to drug assays and the monitoring of bacteria, Int. J. Mass Spectrom. 239 (2004) 235–241.

[5] U. Riess, U. Tegtbur, C. Fauck, F. Fuhrmann, D. Markewitz, T. Salthammer, Experimental setup and analytical methods for the non-invasive determination of volatile organic compounds, formaldehyde and NO_x in exhaled human breath, Anal. Chim. Acta 669 (2010) 53–62.

[6] E. Aprea, F. Biasioli, F. Gasperi, D. Mott, F. Marini, T. Mark, Assessment of Trentingrana cheese ageing by proton transfer reaction-mass spectrometry and chemometrics, Int. Diary J. 17 (2007) 226–234.

[7] M. Mateus, C. Lindinger, J. Gumy, R. Liardon, Release kinetics of volatile organic compounds from roasted and ground coffee: online measurements by PTR-MS and mathematical modeling, J. Agric. Food Chem. 55 (2007) 10117–10128.

[8] F. Biasioli, F. Gasperi, E. Aprea, I. Endrizzi, V. Framondino, F. Marini, et al., Correlation of PTR-MS spectral fingerprints with sensory characterisation of flavour and odour profile of "Trentingrana" cheese, Food. Qual. Pref. 17 (2006) 63–75.

[9] E. Zini, F. Biasioli, F. Gasperi, D. Mott, E. Aprea, T.D. Märk, et al., QTL mapping of volatile compounds in ripe apples detected by proton transfer reaction-mass spectrometry, Euphytica 145 (2005) 269–279.

[10] C. Ennis, J. Reynolds, B. Keely, L. Carpenter, A hollow cathode proton transfer reaction time of flight mass spectrometer, Int. J. Mass Spectrom. 247 (2005) 72–80.

[11] A. Jordan, S. Haidacher, G. Hanel, E. Hartungen, L. Mark, H. Seehauser, et al., A high resolution and high sensitivity proton-transfer-reaction time-of-flight mass spectrometer (PTR-TOF-MS), Int. J. Mass Spectrom. 286 (2009) 122–128.

[12] C. Soukoulis, E. Aprea, F. Biasioli, L. Cappellin, E. Schuhfried, T.D. Märk, et al., Proton transfer reaction time-of-flight mass spectrometry monitoring of the evolution of volatile compounds during lactic acid fermentation of milk, Rapid Commun. Mass Spectrom. 24 (2010) 2127–3134.

[13] A. Fabris, F. Biasioli, P. Granitto, E. Aprea, L. Cappellin, E. Schuhfried, et al., PTR-TOF-MS and data-mining methods for rapid characterisation of agro-industrial samples: influence of milk storage conditions on the volatile compounds profile of Trentingrana cheese, J. Mass. Spectrom. 45 (2010) 1065–1074.

[14] L. Cappellin, F. Biasioli, A. Fabris, E. Schuhfried, C. Soukoulis, T.D. Märk, et al., Improved mass accuracy in PTR-TOF-MS: another step towards better compound identification in PTR-MS, Int. J. Mass Spectrom. 290 (2010) 60–63.

[15] M. Graus, M. Müller, A. Hansel, High resolution PTR-TOF: quantification and formula confirmation of VOC in real time, J. Am. Soc. Mass Spectrom. 21 (2010) 1037–1044.

[16] R. Blake, P. Monks, A. Ellis, Proton-transfer reaction mass spectrometry, Chem. Rev. 109 (2009) 861–896.

[17] J. Herbig, M. Müller, S. Schallhart, T. Titzmann, M. Graus, A. Hansel, On-line breath analysis with PTR-TOF, J. Breath Res. 3 (2009) 027004.

[18] L. Cappellin, et al., Extending the dynamic range of proton transfer reaction time-of-flight mass spectrometers by a novel dead time correction, Rapid Commun. Mass Spectrom. 24 (2010) 1–5.

[19] J. Hamilton, Time Series Analysis, Princeton University Press, Princeton, NJ, 1994.

[20] H. Kantz, Nonlinear Time Series Analysis, 2nd ed., Cambridge University Press, Cambridge UK, New York, 2004.

[21] S. Mallat, A Wavelet Tour of Signal Processing, 2nd ed., Academic Press, San Diego, 1999.

[22] J.N. Coles, M. Guilhaus, Resolution limitations from detector pulse-width and jitter in a linear orthogonal-acceleration time-of-flight mass-spectrometer, J. Am. Soc. Mass Spectrom. 5 (1994) 772–778.

[23] P.B. Coates, Analytical corrections for dead time effects in the measurement of time-interval distributions, Rev. Sci. Instrum. 63 (1992) 2084.

[24] T. Titzmann, M. Graus, M. Müller, A. Hansel, A. Ostermann, Improved peak analysis of signals based on counting systems: illustrated for proton-transfer-reaction time-of-flight mass spectrometry, Int. J. Mass Spectrom. 295 (2010) 72–77.

[25] I. Jolliffe, Principal Component Analysis, 2nd ed., Springer, New York, 2002.

[26] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.

[27] S. Wold, M. Sjostrom, L. Eriksson, PLS-regression: a basic tool of chemometrics, Chemometr. Intell. Lab. Syst. 58 (2001) 109–130.

[28] V. Vapnik, The Nature of Statistical Learning Theory, Springer, New York, 1995.

[29] P. Granitto, C. Furlanello, F. Biasioli, F. Gasperi, Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products, Chemometr. Intell. Lab. Syst. 83 (2006) 83–90.

[30] I. Witten, Data Mining: Practical Machine Learning Tools and Techniques, 2nd ed., Morgan Kaufman, Amsterdam, Boston, MA, 2005.

[31] P. Granitto, F. Gasperi, F. Biasioli, E. Trainotti, C. Furlanello, Modern data mining tools in descriptive sensory analysis: a case study with a random forest approach, Food Qual. Pref. 18 (2007) 681–689.

[32] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Gene selection for cancer classification using support vector machines, Mach. Learn. 46 (2002) 389–422.

[33] Development Core Team, A Language and Environment for Statistical Computing, Vienna, Austria, 2009.

[34] L. Cappellin, M. Probst, J. Limtrakul, F. Biasioli, E. Schuhfried, C. Soukoulis, et al., Proton transfer reaction rate coefficients between H_3O^+ and some sulphur compounds, Int. J. Mass Spectrom. 295 (2010) 43–48.

[35] A. Jordan, S. Haidacher, G. Hanel, E. Hartungen, J. Herbig, L. Märk, et al., An online ultra-high sensitivity Proton-transfer-reaction mass-spectrometer combined with switchable reagent ion capability (PTR + SRI - MS), Int. J. Mass Spectrom. 286 (2009) 32–38.

Biographies

Luca Cappellin received his MS degree in physics from University of Padua in 2009, integrating his studies at the Galilean School of Higher Education. He is a PhD candidate at University of Innsbruck, sponsored by the Edmund Mach Foundation (FEM-IASMA, Italy). His current research interests include PTR-MS fundamentals and data analysis.

Franco Biasioli studied physics at the University of Trento (MSc) and at the University of Innsbruck (PhD). Since 2000 he has been a researcher at the Innovation and Research Centre of the Fondazione Edmund Mach (former Istituto Agrario di S. Michele a/A) where he set up a PTR-MS based facility to investigate the applications of direct injection mass spectrometry in food science and technology: rapid product characterisation, on-line process monitoring, correlation of food volatile compounds profile with sensory analysis and genomics, and nose-space measurements. His activity further comprises multivariate analysis and data mining methods applied to spectrometric and sensory data.

Pablo M. Granitto obtained a MSc and PhD in Physics in 1997 and 2003, respectively, both from Universidad Nacional de Rosario, Argentina. He was a Post-Doc at Istituto Agrario San Michele all'Adige, Trento, Italy. He's currently at CIFASIS and UNR, Rosario, Argentina. His research interests include the application of modern machine learning techniques to agroindustrial problems and the development of clustering and feature selection methods.

Erna Schuhfried is a PhD candidate at the Institute of Ion Physics and Applied Physics, University of Innsbruck, Innsbruck, Austria.

Christos Soukoulis completed his MSc in 2004 at the School of Chemical Engineering of National Technical University of Athens. In 2009 received his PhD from NTUA having worked at the Laboratory of Food Chemistry and Technology. From 2009 is a post-doctoral fellow at IASMA – Fondazione Edmund Mach. His interest is on the area of food science and engineering.

Fabrizio Costa received his PhD in Fruit Crop Science at the University of Bologna, Italy. Currently is a researcher at the Innovation and Research Centre of Edmund Mach Foundation focused on molecular genetic of fruit ripening.

Tilmann Märk, Professor of Physics in the Department of Ion Physics and Applied Physics of the University of Innsbruck, is currently Vicerector for Research at the University of Innsbruck. He received his PhD degree from this university in 1968 and was a guest professor at universities in Konstanz, Beograd, Bratislava, Boulder and State College. His current research focuses on electron and ion physics, clusters, biomolecules, mass spectrometry and chemical analysis. He has published more than 650 papers in refereed journals, given more than 250 invited lectures at international conferences and received a Dr.h.c. degree from the University of Lyon and University of Bratislava.

Flavia Gasperi leads the Food Quality and Consumer Choice program at FEM. She investigates, by sensory and instrumental techniques, the effect of the transformation processes on the final quality of food, the development of new products and the correlation between classical sensory analysis and innovative instrumental analysis of textural properties and volatile compounds.