Multiclass methods in the analysis of metabolomic datasets: The example of raspberry cultivar volatile compounds detected by GC–MS and PTR-MS

Luca Cappellin [a], Eugenio Aprea [a], Pablo Granitto [b], Andrea Romano [a], Flavia Gasperi [a], Franco Biasioli [a,*]

[a] Research and Innovation Centre, Fondazione Edmund Mach, Via E. Mach (FEM), 1, 38010, S. Michele a/A, Italy
[b] 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

ABSTRACT

Multiclass sample classification and marker selection are cutting-edge problems in metabolomics. In the present study we address the classification of 14 raspberry cultivars having different levels of gray mold (*Botrytis cinerea*) susceptibility. We characterized raspberry cultivars by two headspace analysis methods, namely solid-phase microextraction/gas chromatography–mass spectrometry (SPME/GC–MS) and proton transfer reaction-mass spectrometry (PTR-MS). Given the high number of classes, advanced data mining methods are necessary. Random Forest (RF), Penalized Discriminant Analysis (PDA), Discriminant Partial Least Squares (dPLS) and Support Vector Machine (SVM) have been employed for cultivar classification and Random Forest-Recursive Feature Elimination (RF-RFE) has been used to perform feature selection. In particular the most important GC–MS and PTR-MS variables related to gray mold susceptibility of the selected raspberry cultivars have been investigated. Moving from GC–MS profiling to the more rapid and less invasive PTR-MS fingerprinting leads to a cultivar characterization which is still related to the corresponding *Botrytis* susceptibility level and therefore marker identification is still possible.

© 2013 Elsevier Ltd. All rights reserved.

1. Introduction

Multiclass sample classification and "explanatory" variable selection are cutting-edge problems in metabolomics. Often metabolomic studies concentrate on two-class problems, considering for instance treated and non-treated samples. In the vision of integration with genomic and transcriptomic studies for the investigation of gene functions, metabolomic approaches should deal with metabolite differences within a specie by considering populations of individuals. A possible first approach could be the evaluation of different cultivars of a single species. It is therefore important to develop and test methodologies for addressing multiclass problems in this context. Typical datasets from metabolomic experiments have a number of variables considerably exceeding the number of measured samples. This poses a serious limitation to the strategy that may be used for sample classification or response prediction. For example, classical approaches such as Fisher's Linear Discriminant Analysis or General Linear Models are not suitable. A standard method for addressing high dimensional data in food metabolomics is unsupervised Principal Component Analysis (PCA) (Jolliffe, 2002). However, the performance of PCA may be limited, especially when irrelevant factors dominate the variance or when the number of sample classes is very high (Boccard et al., 2010; Jolliffe, 2002). Moreover, PCA alone does not provide a quantification of class separation. Addressing multiclass separation in metabolomics requires more sophisticated

tools, such as machine learning methods (Boccard et al., 2010; Pers, Albrechtsen, Holst, Sorensen, & Gerds, 2009; Scott et al., 2010). Here we discuss the use of supervised classification methods to actually assess the separability of classes. Random Forest (RF) (Breiman, 2001), Penalized Discriminant Analysis (PDA) (Wold, Sjöström, & Eriksson, 2001), Discriminant Partial Least Squares (dPLS) (Wold et al., 2001) and Support Vector Machine (SVM) (Vapnik, 1995) were applied, as a working example, to the identification of raspberry cultivars based on the corresponding volatile profiles.

Volatile organic compounds (VOCs) are important secondary metabolites that can be measured with non-invasive and non-destructive techniques. Their presence is ubiquitous. They are highly studied in plant biology, atmospheric chemistry, and breath analysis (Atkinson, 2000; Buszewski, Kęsy, Ligor, & Amann, 2007; Tholl et al., 2006). In food industry, volatile compounds are a key aspect of the quality of food products and particularly with reference to acceptance by consumers (Klee, 2010). GC–MS profiling triggered the arise of metabolomics and it is still unsurpassed in compound identification capabilities. However, it suffers from relatively time-consuming measurements that render very large studies unpractical. Alternatives are fingerprinting techniques that privilege rapidity over analytical information, and have little sample preparation and no chromatography (Han, Datla, Chan, & Borchers, 2009). Such techniques, on one hand, allow screening a broader number of samples and, on the other hand, minimize the potential artifacts due to the extraction and concentration procedures (Han et al., 2009). Moreover, machine learning methods on such datasets are often more robust given the larger number of measured samples (Han et al., 2009). Here we

employ two different headspace techniques: on the one hand the well-established GC–MS, on the other hand an innovative, rapid and non-invasive (no sample treatment) methodology that allows VOC fingerprinting and measurement repetitions during product shelf life. Proton transfer reaction-mass spectrometry (PTR-MS) is a hyphenated technique that was developed about two decades ago by Lindinger and co-workers (Lindinger, Hansel, & Jordan, 1998). Already in its original versions it was equipped with a quadrupole mass analyzer, having about unit mass resolution. It couples high sensitivity (ppbt) with a large dynamic range and a fast response time (Lindinger et al., 1998). A complete spectrum (from 0 to 200 Th) is acquired in a few seconds. Recently PTR-MS has also been coupled with time of flight (Jordan et al., 2009) and ion trap (Mielke et al., 2008; Prazeller, Palmer, Boscaini, Jobson, & Alexander, 2003) detectors, reaching much higher mass resolutions. Compared to GC–MS, PTR-MS reduces measurement time of about 100 times. Moreover, it provides simple output data, requiring no pre-processing before statistical analysis. Therefore, a sample can be fingerprinted in a few seconds in total. In this sense, PTR-MS is an ideal tool for metabolomic investigations. A pioneer feasibility study in this field has been carried out a few years ago by Granitto et al.(2007). After improving the data analysis methodology (Cappellin, Biasioli, et al., 2010; Cappellin et al., 2011), recently, the first application followed (Cappellin, Soukoulis, et al., 2012), showing that PTR-MS coupled to a time-of-flight spectrometer and to suitable data mining methods is a powerful tool for separating apple cultivars and clones on the basis of their VOC emission profiles. Another high-throughput approach has been attempted by several other research groups through direct infusion of non-volatile compounds (mainly from liquid samples) into mass spectrometers (Favé et al., 2011; Højer-Pedersen, Smedsgaard, & Nielsen, 2008; Mattoli et al., 2010; McDougall, Martinussen, & Stewart, 2008) but with scarce results mainly due to ion suppression (Annesley, 2003; Sterner, Johnston, Nicol, & Ridge, 2000) and very complex spectra.

Raspberry (*Rubus idaeus* L.) is a member of the Rosaceae family, grown primarily for its edible berries. Raspberry fruits are important dietary sources of antioxidant compounds, in particular, polyphenols (Kähkönen, Hopia, & Heinonen, 2001), which are renowned for their health benefits (Larrosa, González-Sarrías, García-Conesa, Tomás-Barberán, & Espín, 2006). Their typical flavor makes these fruits easily recognizable and appreciated not only for their health impact (Aprea, Carlin, Giongo, Grisenti, & Gasperi, 2010).

Literature studies about the volatile emission from raspberry are scarce and mainly concentrate on gas chromatographic techniques (Aprea et al., 2010; Guichard & Issanchou, 1983; Malowicki, Martin, & Qian, 2008). We already pointed out the viability of the PTR-MS approach for in vivo characterization of raspberry products in a recent study involving two raspberry cultivars (Aprea, Biasioli, Carlin, Endrizzi, & Gasperi, 2009). In the present work we employ GC–MS profiles and PTR-MS fingerprints for the discrimination of 14 raspberry cultivars, which are listed in Table 1, via the already mentioned classification methods. Moreover, we employ a multivariate method for the identification of the features (peaks/volatile compounds) that are most relevant for the raspberry classification problem.

Another area where modern multiclass methods can replace traditional approaches (as linear regression or PLS) is grading problems, i.e. problems in which samples are divided into numbered classes that are ordered according to a given measure of similarity. Simple examples are qualities or tolerance to stress factors. In our case, differences in volatile emission between raspberry cultivars may be related to the diverse level of susceptibility to certain infections. Here we concentrate on gray mold caused by *Botrytis cinerea*(Elad et al., 2004; Jarvis, 1962). In a precedent study on GC–MS data (Aprea et al., 2010) we identified nine compounds which were negatively correlated to *Botrytis* susceptibility. These compounds were mainly monoterpenes, such as α-pinene, β-phellandrene, *p*-cymene, and 4-terpineol, and sesquiterpenes, namely trans-caryophyllene and caryophylleneoxide. Moreover, 2-heptanol, β-damascenone and dehydro-β-ionone were found. PTR-MS is sensible

Table 1
List of considered raspberry cultivars. Gray mold susceptibility values from Aprea et al.(2010) are also reported.

	Variety	*Botrytis* susceptibility
1	Anne	4
2	Autumn Bliss	3
3	Caroline	0
4	Heritage	2
5	Himbo-top	1
6	Josephine	0
7	Opal	3
8	Pokusa	4
9	Polana	4
10	Polesie	5
11	Polka VV3-536	4
12	Polka VV5-657	4
13	Popiel	5
14	Tulameen	2

to monoterpenes but cannot separate them, since isobaric ions generate superposing signals. Similar remarks hold for sesquiterpenes. A key question is in fact whether the accuracy of cultivar grading related to *Botrytis* susceptibility diminishes due to this loss of information. It is therefore interesting to assess the relationship between rapid PTR-MS fingerprints and gray mold susceptibility grading for the considered raspberry cultivars. We also use the four machine learning methods named before to grad GC–MS profiles and PTR-MS fingerprints into 6 levels of *Botrytis* susceptibility (also listed in Table 1). The application of diverse machine learning techniques could potentially highlight new information about the problem at hand (grading based on GC–MS profiles), as was demonstrated in previous works (Cappellin, Soukoulis, et al., 2012; Granitto et al., 2007). Again, we apply a multivariate method for the identification of the features (peaks/compounds) that are most relevant for this grading. We recall from Aprea et al.(2010), that the degree of *Botrytis* susceptibility was assessed for each genotype by reporting an index on a scale from 0 to 5, 0 meaning that no fruits showed any damage caused by this infection.

In summary, the scope of the present work is twofold. On the one hand we show that supervised multivariate techniques, such as machine learning ones, may successfully address different multiclass analysis problems in metabolomics: the classification of raspberry cultivars or the grading of the same cultivars into levels of *Botrytis* susceptibility. On the other hand, we discuss the advantages of rapid PTR-MS fingerprint of volatiles compared with traditional GC–MS profiles.

2. Materials and methods

2.1. GC–MS and PTR-MS analyses of raspberry cultivars

This study is a derivative work following our recent study on GC–MS profiling of raspberry cultivars (Aprea et al., 2010). We therefore refer to Aprea et al.(2010) for a detailed description of samples and give only a brief summary here.

Fruits were produced under standard conditions (Aprea et al., 2010) and collected from the Edmund Mach Foundation experimental orchard located in Vigolo Vattaro (Trentino, Italy). In order to take into account possible variability two different seasons (2006 and 2007) were considered, and three batches for each of the 14 cultivars were collected on three different days in each year. The actual number of measured samples depends on the analytical technique. The GC–MS dataset obtained for the 14 cultivars has extensively been described in Aprea et al.(2010). Briefly, ripe berries (a batch of about 250 g per each variety) were harvested manually, placed in plastic container and immediately transported to the laboratory, in ice packs, where samples were stored at 4 °C for 1 day before analyses. From 4 to 5 fruits (about 20 g) per each variety where grouped

together and used for GC analysis, obtaining six data points for GC dataset (one data point × three different sampling days × two years) per each variety. When enough material was available more data points were added. From the same batch, three to six fruits (according to availability) per each variety were sampled and individually measured by PTR-MS, obtaining at least 20 data points for PTR-MS dataset (data points × three different sampling days × two years) per each variety. Only for Josephine variety we collected 15 points (over the two years) being the available material scarcer.

The SPME/GC–MS analysis procedure has already been described elsewhere (Aprea et al., 2009).

PTR-MS measurements were conducted using a high sensitivity PTR-MS manufactured by Ionicon (IONICON Analytik GmbH, Innsbruck, Austria). The conditions in the drift tube were 2.04 mbar pressure, 520 V drift tube voltage and 50 °C temperature, corresponding to a E/N of about 120 Td. The dwell time was set to 0.2 s and the m/z range to 20–240 Th. We chose to measure mashed berries and not intact fruits for this study because of superior level of VOC emission of the former. A comparison between intact fruit and mashed fruit regarding their PTR-MS fingerprint can be found in Aprea et al.(2009). The measuring procedure was set according to our previous work (Aprea et al., 2009). A day after harvest, single berry fruits were taken from the 4 °C storage space, left at room temperature for 90 min, then gently mashed and placed into a sealed glass vessel (323 mL) equipped with silicon septa on two opposite sides. After equilibrating for 60 min at room temperature, the inlet of the PTR-MS was connected by a 1/16″ PTFE tube kept at 70 °C, terminating with a stainless-steel needle to be introduced into one of the glass vessel septa. The opposite septum was connected to a 1/4″ PTFE tube through a second stainless-steel needle to allow outdoor air to enter the vessel, thus replacing the headspace air that was continuously extracted for 4 min (corresponding to the acquisition of five complete spectra) at 10 $cm^3\ min^{-1}$. Special care was devoted to avoid systematic memory effects: replicate order was randomized, different glass vessels were used for each sample, and the apparatus was flushed with outdoor air for 6 min between consecutive measurements. Spectral data were normalized by the primary ion as described in Lindinger et al.(1998) and employing a constant reaction rate coefficient of $2 \cdot 10^{-9}\ cm^3$ s. The systematic error that is introduced in the concentration determination for each compound is in most cases below 30% and can be accounted for if the actual rate coefficient is available (Cappellin, Karl, et al., 2012; Cappellin, Probst, et al., 2010).

2.2. Statistical analysis

We analyzed two datasets. The GC–MS dataset consisted of 94 rows, corresponding to the measured samples, and 45 columns, each corresponding to an identified compound. The PTR-MS dataset consisted of 358 rows (samples) and 141 columns, each corresponding to the normalized intensity of a PTR peak. The datasets were built to take into account the intra-seasoning (three different days of sampling spanned over six weeks, see Aprea et al.(2009) for more details) and inter-seasoning (two years, Aprea et al., 2010) variabilities of the fruits, in order to build more robust classification models.

Supervised classification models were built using RF, PDA, dPLS and SVM on both GC–MS and PTR-MS datasets. All methods were described in previous works on PTR-MS fingerprint analysis (Granitto et al., 2007). In all cases, we used implementations available as free packages for the R statistical environment software (R Development Core Team, 2009). To evaluate the results of the classification methods we used a leave-one-out (LOO) procedure: we iterated the process of leaving a sample out as test set and using the remaining of the dataset to fit the models. The free parameters of each classifier, such as the C constant of SVM or the number of dimensions considered in dPLS, were selected at this step by internal cross validation using only the training dataset. After that, those models were used to individually classify the sample in the independent test batch. We analyzed the classification results using confusion matrices, in which rows correspond to the true classes and columns to the predicted ones. The diagonal entries of the confusion matrix correspond to correct classifications. The results are given in number of samples of each cultivar that the classifier assigns to the cultivar given by the column title.

Relevant Feature Identification was done using Random Forest-Recursive Feature Elimination (RF-RFE), introduced by Granitto, Furlanello, Biasioli, & Gasperi, (2006). The procedure has two steps. First, RF-RFE is applied separately to each one of the several partitions in training and test sets produced by the LOO procedure described previously. The method produces an average error curve relating the classification error with the number of compounds/peaks used in the model. We use that curve to select a number p of peaks that is as low as possible but still yields good discriminant models. In the second step, we select the top p compounds/peaks from each run of the RF-RFE. We compute the average number of times that each peak is selected in these reduced lists of p discriminant inputs, and keep only the compounds/peaks that were selected more often. It is important to note that the output of the process is a list of compounds/peaks that are highly relevant to the problem, not the subset that produces the lowest classification error.

3. Results and discussion

3.1. Classification of raspberry cultivars

In the present section we aim at presenting the raspberry cultivar multiclass problems addressed by the selected data analysis methods on both datasets.

Table 2
GC–MS profiles. Confusion matrix for the classification by RF of 14 raspberry cultivars considered in this work.

		11	12	14	10	13	9	7	5	8	3	4	1	2	6
		Polka VV3-536	Polka VV5-657	Tulameen	Polesie	Popiel	Polana	Opal	Himbotop	Pokusa	Caroline	Heritage	Anne	Autumn Bliss	Josephine
11	Polka VV3-536	1	6	0	0	0	0	0	0	0	0	0	0	0	0
12	Polka VV5-657	6	0	0	0	0	0	0	0	0	0	0	0	0	0
14	Tulameen	0	0	4	1	1	0	0	0	0	0	1	0	0	0
10	Polesie	0	0	0	7	1	0	0	0	0	0	0	0	0	0
13	Popiel	0	0	0	0	8	0	0	0	0	0	0	0	0	0
9	Polana	0	0	0	0	0	6	0	0	0	0	0	0	0	0
7	Opal	0	0	0	0	0	0	5	0	1	0	0	0	0	0
5	Himbotop	0	0	0	0	0	0	0	5	1	0	0	0	0	0
8	Pokusa	0	0	0	0	0	0	0	1	5	0	0	0	0	0
3	Caroline	0	0	0	0	0	0	0	0	0	4	2	0	0	0
4	Heritage	0	0	0	0	0	0	0	0	0	1	10	0	0	0
1	Anne	0	0	0	0	0	0	0	0	0	0	0	6	0	0
2	Autumn Bliss	0	0	0	0	0	0	0	0	0	0	0	0	5	1
6	Josephine	0	0	0	1	0	0	1	1	0	0	0	0	2	0

Table 3
PTR-MS fingerprint. Confusion matrix for the classification by RF of 14 raspberry cultivars considered in this work.

		14	11	12	8	6	9	7	3	4	1	13	10	2	5
		Tulameen	Polka VV3-536	Polka VV5-657	Pokusa	Josephine	Polana	Opal	Caroline	Heritage	Anne	Popiel	Polesie	Autumn Bliss	Himbotop
14	Tulameen	18	0	0	0	0	0	1	0	1	0	0	0	0	0
11	Polka VV3-536	1	14	8	0	0	1	0	0	0	0	2	0	1	0
12	Polka VV5-657	0	11	9	0	0	1	0	0	0	0	0	0	0	0
8	Pokusa	0	0	1	22	0	1	0	0	0	0	0	0	0	1
6	Josephine	0	0	0	0	12	1	0	0	1	0	0	0	1	0
9	Polana	0	1	0	0	0	23	0	0	0	1	0	0	0	0
7	Opal	0	0	0	0	0	0	22	1	1	0	0	0	0	0
3	Caroline	0	0	0	0	0	0	1	16	7	0	0	0	0	0
4	Heritage	0	0	0	0	0	0	1	1	42	0	0	0	0	0
1	Anne	1	0	0	0	0	0	0	0	0	20	1	0	0	0
13	Popiel	0	1	1	0	0	0	0	0	1	1	28	0	0	0
10	Polesie	0	1	1	0	0	0	0	0	2	0	0	22	0	1
2	Autumn Bliss	1	2	1	0	1	0	0	0	0	0	0	4	9	4
5	Himbotop	0	0	0	0	0	0	0	0	5	0	0	1	1	23

The confusion matrix reported in Table 2 provides an insight in the classification performance provided by RF on the GC–MS dataset. The corresponding average classification error is 0.298, meaning that on average about 70% of the samples are assigned to the correct class by the model. The generally good prediction performance on the analyzed cultivars entangles a significant and robust difference in their VOC profiles. A close look at Table 2 suggests the existence of groups of cultivars which are generally confused. This is particularly true in the case of Polka VV3-536 and Polka VV5-657. In fact, these two classes correspond to the same cultivar (Polka), confirmed by genetic analysis (data not shown) and have been considered separately because at the beginning of the harvesting campaign their identity was doubtful. Thus, it does not come as a surprise that the model is not able to distinguish between them and can be seen as a sort of validation of the method. Caroline and Heritage are confused in about 17% of the cases. In fact, Caroline is genetically close to Heritage, being a crossing between Heritage and Geo-1 (Aprea et al., 2010). The discriminate model is not able to correctly classify the Josephine samples. Moreover, a marked confusion with Autumn Bliss is evident. Autumn Bliss has a complex parentage including *Rubus strigosus*, *Rubus arcticus*, and *Rubus occidentalis*, and 6 red raspberry varieties (US Plant Patent 6597).

The analogous confusion matrix built on PTR-MS data is reported in Table 3. The corresponding average classification error (0.218) is rather lower than in the case of GC–MS, almost 80% samples being correctly classified. The expected confusion between Polka VV3-536 and Polka VV5-657, even if less marked, is confirmed. The confusion between Heritage and Caroline (in 8 out of 68 cases, 12%) is found in analogy to the outcome of the analyses on the GC–MS data. Contrary to the model built on GC–MS data, RF on PTR-MS data is able to correctly assign Josephine sample in 80% of the cases. Probably

this superior performance is mainly related to the larger number of Josephine samples, 15 instead of 6, that were analyzed with PTR-MS compared to GC–MS. The same reason is probably more generally at the basis of the lower prediction errors that multivariate models on PTR-MS fingerprints show compared to the corresponding models built on GC–MS data.

Table 4 reports a comparison between the considered classification methods. In the case of PTR-MS data, RF shows the lowest prediction errors, followed by PDA; PLS and SVM gives poorer results. Previous results on multiclass classification on PTR-MS data (Granitto et al., 2007) also showed a good performance of RF on this kind of data. For GC–MS, very similar prediction performances are found for the four methods, SVM giving slightly lower prediction errors than the other methods in this case. Overall, all four methods show a good performance in both problems. Confusion matrices for the other methods are qualitatively similar to those showed in Tables 2 and 3 (not shown for lack of space). Table 4 also reports the prediction performances considering only 13 classes, where the two Polka cultivars are grouped together. The results improve in all cases, showing that these two mixed and isolated classes correspond in fact to a single class. With 13 classes the best prediction errors are reduced to 0.191 for GC–MS and to 0.156 for PTR-MS.

3.2. Feature selection

In this section we address the problem of highlighting the most relevant variables (GC–MS compounds or PTR-MS peaks) for

Table 4
Average classification errors for the selected multivariate methods. Results are reported for the 14 raspberry cultivar multiclass problem along with those of reduces number of classes after merging. See text.

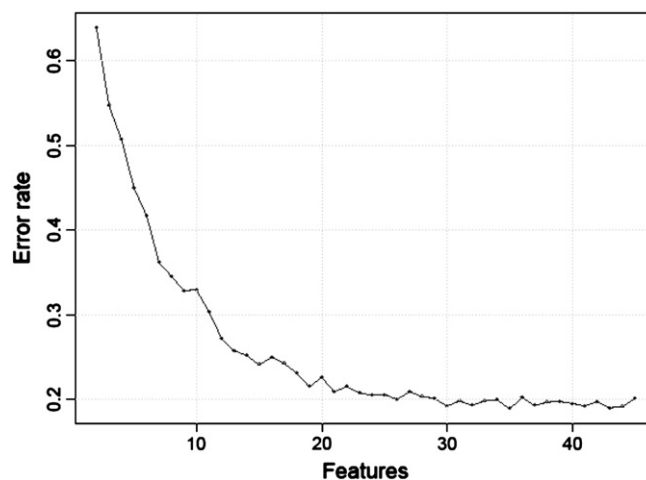|	14 classes	13 classes
PTR-MS		
RF	0.218	0.156
PDA	0.271	0.218
PLS	0.310	0.243
SVM	0.291	0.246
mean	0.272	0.216
GC–MS		
RF	0.298	0.191
PDA	0.277	0.202
PLS	0.266	0.245
SVM	0.255	0.170
mean	0.274	0.202

Fig. 1. GC–MS profiles. Mean prediction error of Random Forest over the 98 LOO replications as a function of the number of variables used in the models during the feature selection process.

Table 5
GC–MS profiles. Fraction of times that each compound was selected among the 20 more discriminant features on RF over LOO replicated experiments, for the 14 class problem.

Compounds	Fraction of times selected
2-Heptanone	1.00
2-Heptanol	1.00
Trans-caryophyllene	1.00
Dehydro-β-ionone	1.00
α-Phellandrene	0.96
Benzyl alcohol	0.94
Trans-3-methyl-1,3,5-hexatriene	0.92
Ethyl acetate	0.84
Theaspirane B	0.84
Limonene	0.82
β-Phellandrene	0.82
Linalool	0.74
Geraniol	0.74
α-Pinene	0.70
p-Cymene	0.70
Caryophyllene oxide	0.70
3,4-Didehydro-β-ionone (t.i.)	0.68
β-Damascenone	0.62
Acetic acid	0.60
β-Pinene	0.58
Acetoin	0.58
β-Myrcene	0.56
α-Ionol	0.40
β-Ionone	0.38
Hexanal	0.36
Unidentified sesquiterpene	0.36
Acetato di esile	0.28
Hexanoic acid	0.26
5-Ethyl-(3H)-furan-2-one (t.i.)	0.18
4-Terpineol	0.14
γ-Terpinene	0.10

Table 6
GC–MS profile. Confusion matrix for the grading by RF of the raspberry samples into the 6 gray mold susceptibility level. The class number represents the susceptibility level.

	0	1	2	3	4	5
0	3	0	4	4	0	0
1	0	4	0	2	0	0
2	1	0	15	0	0	2
3	0	0	0	12	0	0
4	0	1	0	1	28	1
5	0	0	0	0	4	12

Interestingly, the four variables which are always selected within the 20 used to build the models belong to different classes of VOCs, namely ketones (2-heptanone), alcohols (2-heptanol), sesquiterpenes (trans-caryophyllene), and C13-norisoprenoids (dehydro-β-ionone). This suggests that dissecting the VOC emissions of the considered raspberry cultivars entangles coarse grain differences in the emission of diverse classes of VOCs. Multivariate models provide a useful tool to capture such differences for classification purposes and allow the highlighting of VOCs that are most relevant in the discrimination.

Feature selection in the case of PTR-MS data was carried out in an analogous way. We chose to keep, again, 20 peaks over the initial 141 peaks measured by PTR-MS (figure not shown in this and following experiments). The final results are reported in Fig. 2. In this case it is less straightforward to draw conclusions. Among the more relevant variables, we find peaks 69, 95 and 137, which mainly correspond to monoterpenes and their fragmentation (Steeghs, Crespo, & Harren, 2007; Tani, Hayward, Hansel, & Hewitt, 2004). The peak at m/z 73 is related to aldehyde fragments and 2-butanone and the one at m/z 83 corresponds to a fragment of hexanal (Aprea et al., 2009). m/z 41 is a general fragment. In analogy to the case of GC − MS, VOCs belonging to very different classes are important for differentiating the studied raspberry cultivars. Note that RF-RFE may include isotopic peaks that entangle the same information, such as for instance peak at 137 amu and its isotope at 138 amu.

3.3. Botrytis susceptibility

The basic question addressed in this paragraph is whether moving from GC–MS profiling of the selected raspberry cultivars to their
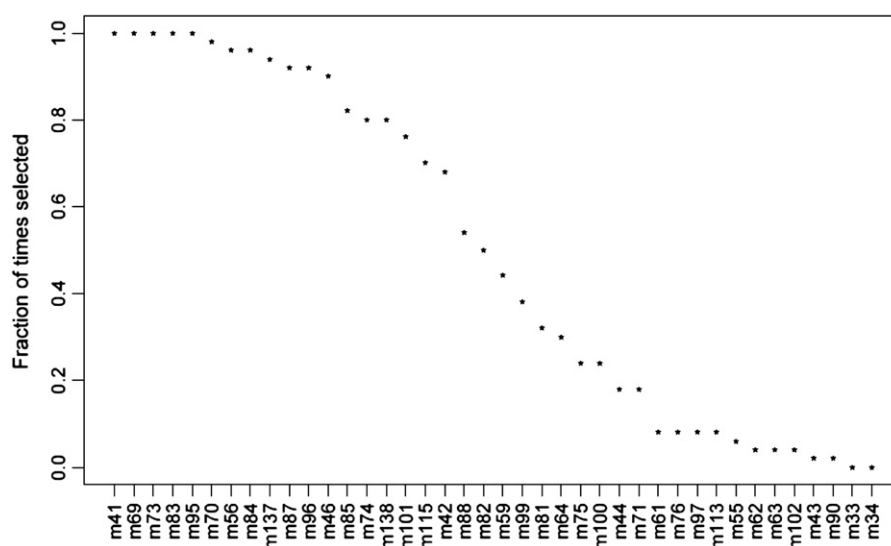
separating the 14 raspberry classes. As we explained in Section 2, feature selection based on RF-RFE proceeds as follows. As a first step we assess the behavior of the mean discrimination error of RF as a function of the number of variables used in the model. In order to clarify the procedure, in the case of the GC–MS dataset the results are showed in Fig. 1. A trade-off between model simplicity and discrimination error is represented by 20 variables in this case, from the original total of 45. In a subsequent step we report how often each variable is selected among the 20 most relevant peaks by RF, over the 94 LOO RF-RFE experiments. The results are reported in Table 5.

Fig. 2. PTR-MS fingerprints. Fraction of times that each compounds was selected among the 20 more discriminant features on RF over LOO replicated experiments, for the 14 class problem. Higher fraction means more relevant variable for the discriminant process.

Table 7

PTR-MS fingerprint. Confusion matrix for the grading by RF of the raspberry samples into the 6 gray mold susceptibility level. The class number represents the susceptibility level.

	0	1	2	3	4	5
0	26	0	8	3	1	1
1	0	22	6	0	2	0
2	1	0	59	2	2	0
3	2	0	4	25	12	3
4	0	0	1	0	116	3
5	0	0	2	2	12	43

PTR-MS fingerprints leads to a cultivar characterization which is still related to the corresponding gray mold susceptibility level; or whether, at the contrary, it leads to a loss of information that disrupts that relationship. We recall that in the case of the PTR-MS dataset a larger number of samples per each raspberry cultivar have been considered, given the highly reduced measurement time required by PTR-MS compared to GC–MS.

We first apply a multiclass analysis approach to both GC–MS and PTR-MS data in order to grade the samples into classes of equal susceptibility level.

Tables 6 and 7 report the confusion matrices for RF models in the case of GC–MS and PTR-MS data, respectively. The corresponding average prediction errors are 0.21 and 0.19, meaning that in both datasets the correct *Botrytis* susceptibility level is assigned by the model to about 80% of the samples. A comparison between the performances of the various classification methods is reported in Table 8. In general RF is found to reliably provide good results. In fact, RF (prediction error 0.19) outperforms all other methods in the case of the PTR-MS, the other methods showing poorer results. For the GC–MS dataset RF (prediction error 0.21), PDA (0.20) and SVM (0.22) display similar performances, while PLS gives a higher error (0.27). As this is a grading problem, it is important to consider also the type of error produced by the diverse classifiers. Beside classification error, Table 8 also reports the fraction of samples that are assigned to a class distant 1 level (denoted as "1 Level") from the correct one or more than 1 level (denoted as ">1 Level"). Interestingly, the results are rather different for the two datasets. In fact PTR-MS not only produces better results in general, i.e. lower prediction errors, but also the confusion is more related to neighbor classes than in the case of GC–MS data. Again, this fact could be probably attributed to the larger number of samples considered for PTR-MS investigations. In conclusion both PTR-MS and GC–MS are suitable to address the presented grading problem and when experiment time is an issue PTR-MS should be preferred.

Variable selection was performed in order to unveil which variables are more important for dissecting gray mold susceptibility in the selected raspberry cultivars.

Table 9 reports variable selection results for RF-RFE applied to the GC–MS dataset. The four most relevant variables, selected in between 80% and 100% of the LOO experiments, are α-phellandrene, *p*-cymene, 4-terpineol, and dehydro-β-ionone. Consistently, such compounds were also found by Aprea et al.(2010) using a regression method and Martens' uncertainty test.

Table 8

Average classification errors for the employed multivariate methods. Results are reported for the *Botrytis* susceptibility multiclass problem for both PTR-MS and GC–MS headspace analyses.

Method	PTR-MS			GC–MS		
	Error	1 Level	>1 Level	Error	1 Level	>1 Level
RF	0.187	0.109	0.078	0.213	0.064	0.149
PDA	0.282	0.14	0.142	0.202	0.085	0.117
PLS	0.299	0.142	0.157	0.266	0.117	0.149
SVM	0.257	0.162	0.095	0.223	0.063	0.160

Table 9

GC–MS profiles. Fraction of times that each compound was selected among the 6 more discriminant features on RF over LOO replicated experiments, for the 6 class grading problem.

Compounds	Fraction of times selected
Dehydro-β-ionone	1.00
4-Terpineol	0.98
p-Cymene	0.91
α-Phellandrene	0.89
Trans-caryophyllene	0.65
Unidentified sesquiterpene	0.59
Theaspirane B	0.52
β-phellandrene	0.22
γ-Terpinene	0.11
Geraniol	0.09
2-Heptanone	0.04

The RF-RFE variable selection in the case of the PTR-MS is reported in Fig. 3. Again, terpenes are among the most relevant features. In fact, the peak at m/z 69 is a common fragment of terpenes and aldehydes; the peak at m/z 137 is related to monoterpenes, m/z 95 is a terpene fragment (Aprea et al., 2009). m/z 115 is probably related to 2-heptanone (Aprea et al., 2009). Such results are consistent with the findings using GC–MS and suggest that rapid PTR-MS fingerprint captures properties of raspberry cultivars that are closely connected to their resistance to gray mold susceptibility. The potential role of monoterpenes and sesquiterpenes in the inhibition of gray mold infections has been pointed out by many other studies (Bouchra, Achouri, Hassani, & Hmamouchi, 2003; Daferera, Ziogas, & Polissiou, 2003; Reddy, Angers, Gosselin, & Arul, 1998; Sekine, Sugano, Majid, & Fujii, 2007). For example, Reddy et al.(1998) highlighted the action of essential oils from *Thymus vulgaris* against *B. cinerea* for strawberry. Sekine et al.(2007) tested the effect of *p*-cymene and cuminaldehyde vapor phase concentrations on the mycelial growth inhibition of phytopathogenic fungi such as *B. cinerea*.

4. Conclusions

In the present study we showed that supervised multivariate techniques, such as machine learning ones, may successfully address different multiclass analysis problems in metabolomics. We employed modern machine learning methods to analyze diverse aspects of two different multiclass problems. First, we classified 14 raspberry cultivars on the basis of their GC–MS profiles and PTR-MS fingerprints. Good results were achieved with both techniques, with slightly lower classification errors for the PTR-MS dataset, probably because of the larger number of analyzed samples. In fact, PTR-MS is a high-throughput technique that allows the reduction of measurement time by about 100 times compared to standard GC–MS analysis. Groups of cultivars with similar volatile emission were consistently identified using confusion matrices. These similarities were related to genetic affinities, varieties sharing common parents are generally grouped together. Among the classification methods considered, Random Forest showed the best classification performance, in particular for PTR-MS data, but the other methods also showed to be effective in both cases.

Feature selection by RF-RFE allowed the identification of the peaks/compounds that are more relevant to these classification problems and suggested that VOCs of very diverse compound classes are needed for the full discrimination of the considered raspberry cultivars.

The same analysis procedure was employed to grade the raspberry cultivars on levels of gray mold susceptibility. Models based on the GC–MS dataset and on the PTR-MS one displayed similar grading errors but marked differences in the confusion matrices. In fact, in the PTR-MS case multivariate model prediction errors were primary based on confusions between raspberries belonging to cultivars with close level of susceptibility, while this did not hold true in the
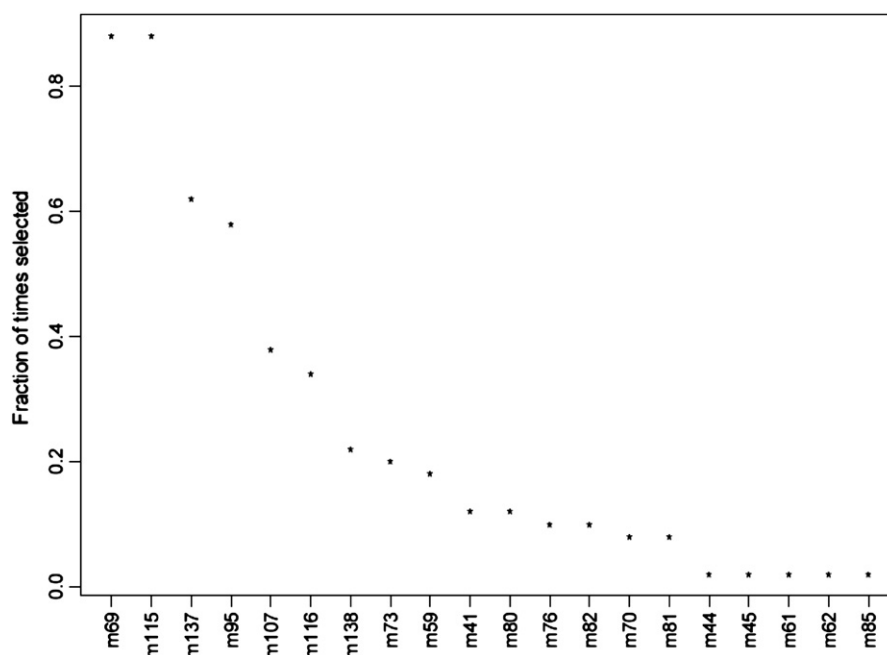
Fig. 3. PTR-MS fingerprints. Fraction of times that each compound was selected among the 6 more discriminant features over LOO replicated experiments for the gray mold susceptibility grading problem. Higher fraction means more relevant variable for the discriminant process.

GC–MS case. Again, RF reliably displayed the best grading capabilities. Several VOCs, in particular terpenes, were found to be related to the gray mold susceptibility level. We showed that moving from GC–MS profiling to PTR-MS fingerprinting leads to a cultivar characterization which is still related to the corresponding *Botrytis* susceptibility level and therefore marker identification is still possible.

Acknowledgments

References

Annesley, T. M. (2003). Ion suppression in mass spectrometry. *Clinical Chemistry*, *49*(7), 1041–1044. http://dx.doi.org/10.1373/49.7.1041.

Aprea, E., Biasioli, F., Carlin, S., Endrizzi, I., & Gasperi, F. (2009). Investigation of volatile compounds in two raspberry cultivars by two headspace techniques: Solid-phase microextraction/gas chromatography–mass spectrometry (SPME/GC–MS) and proton-transfer reaction-mass spectrometry (PTR-MS). *Journal of Agricultural and Food Chemistry*, *57*(10), 4011–4018. http://dx.doi.org/10.1021/jf803998c.

Aprea, E., Carlin, S., Giongo, L., Grisenti, M., & Gasperi, F. (2010). Characterization of 14 raspberry cultivars by solid-phase microextraction and relationship with gray mold susceptibility RID C-1218-2010 RID E-9184-2011 RID B-8104-2011. *Journal of Agricultural and Food Chemistry*, *58*(2), 1100–1105. http://dx.doi.org/10.1021/jf902603f.

Atkinson, R. (2000). Atmospheric chemistry of VOCs and NOx. *Atmospheric Environment*, *34*(12–14), 2063–2101. http://dx.doi.org/10.1016/S1352-2310(99)00460-4.

Boccard, J., Kalousis, A., Hilario, M., Lanteri, P., Hanafi, M., Mazerolles, G., et al. (2010). Standard machine learning algorithms applied to UPLC–TOF/MS metabolic fingerprinting for the discovery of wound biomarkers in *Arabidopsis thaliana* RID A-8870-2011. *Chemometrics and Intelligent Laboratory Systems*, *104*(1), 20–27. http://dx.doi.org/10.1016/j.chemolab.2010.03.003.

Bouchra, C., Achouri, M., Hassani, L., & Hmamouchi, M. (2003). Chemical composition and antifungal activity of essential oils of seven Moroccan Labiatae against *Botrytis cinerea* Pers: Fr. *Journal of Ethnopharmacology*, *89*(1), 165–169. http://dx.doi.org/10.1016/S0378-8741(03)00275-7.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Buszewski, B., Kęsy, M., Ligor, T., & Amann, A. (2007). Human exhaled air analytics: Biomarkers of diseases. *Biomedical Chromatography*, *21*(6), 553–566. http://dx.doi.org/10.1002/bmc.835.

Cappellin, L., Biasioli, F., Fabris, A., Schuhfried, E., Soukoulis, C., Mark, T., et al. (2010). Improved mass accuracy in PTR-TOF-MS: Another step towards better compound identification in PTR-MS. *International Journal of Mass Spectrometry*, *290*(1), 60–63. http://dx.doi.org/10.1016/j.ijms.2009.11.007.

Cappellin, L., Biasioli, F., Granitto, P. M., Schuhfried, E., Soukoulis, C., Costa, F., et al. (2011). On data analysis in PTR-TOF-MS: From raw spectra to data mining. *Sensors and Actuators B-Chemical*, *155*(1), 183–190. http://dx.doi.org/10.1016/j.snb.2010.11.044.

Cappellin, L., Karl, T., Probst, M., Ismailova, O., Winkler, P. M., Soukoulis, C., et al. (2012). On quantitative determination of volatile organic compound concentrations using proton transfer reaction time-of-flight mass spectrometry. *Environmental Science & Technology*, *46*(4), 2283–2290. http://dx.doi.org/10.1021/es203985t.

Cappellin, L., Probst, M., Limtrakul, J., Biasioli, F., Schuhfried, E., Soukoulis, C., et al. (2010). Proton transfer reaction rate coefficients between H3O+ and some sulphur compounds. *International Journal of Mass Spectrometry*, *295*(1–2), 43–48. http://dx.doi.org/10.1016/j.ijms.2010.06.023.

Cappellin, L., Soukoulis, C., Aprea, E., Granitto, P., Dallabetta, N., Costa, F., et al. (2012). PTR-ToF-MS and data mining methods: A new tool for fruit metabolomics. *Metabolomics*, *8*(2). http://dx.doi.org/10.1007/s11306-012-0405-9.

Daferera, D., Ziogas, B., & Polissiou, M. (2003). The effectiveness of plant essential oils on the growth of *Botrytis cinerea*, *Fusarium* sp and *Clavibacter michiganensis* subsp *michiganensis*. *Crop Protection*, *22*(1), 39–44. http://dx.doi.org/10.1016/S0261-2194(02)00095-9.

Elad, Y., Williamson, B., Tudzynski, P., Delen, N., Elad, Y., Williamson, B., et al. (2004). Botrytis: Biology, pathology and control. Downloaded from. http://www.cabdirect.org/abstracts/20053018846.html;jsessionid=954C14134777AEA39E13325E12130E8A

Favé, G., Beckmann, M., Lloyd, A. J., Zhou, S., Harold, G., Lin, W., et al. (2011). Development and validation of a standardized protocol to monitor human dietary exposure by metabolite fingerprinting of urine samples. *Metabolomics*, *7*(4), 469–484. http://dx.doi.org/10.1007/s11306-011-0289-0.

Granitto, P., Biasioli, F., Aprea, E., Mott, D., Furlanello, C., Mark, T., et al. (2007). Rapid and non-destructive identification of strawberry cultivars by direct PTR-MS headspace analysis and data mining techniques. *Sensors and Actuators B: Chemical*, *121*(2), 379–385. http://dx.doi.org/10.1016/j.snb.2006.03.047.

Granitto, P., Furlanello, C., Biasioli, F., & Gasperi, F. (2006). Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products. *Chemometrics and Intelligent Laboratory Systems*, *83*(2), 83–90. http://dx.doi.org/10.1016/j.chemolab.2006.01.007.

Guichard, E., & Issanchou, S. (1983). Extraction of volatile compounds from raspberry by different methods — application of principal component analysis to gas-chromatographic data. *Sciences Des Aliments*, *3*(3), 427–438.

Han, J., Datla, R., Chan, S., & Borchers, C. H. (2009). Mass spectrometry-based technologies for high-throughput metabolomics. *Bioanalysis*, *1*(9), 1665–1684. http://dx.doi.org/10.4155/bio.09.158.

Højer-Pedersen, J., Smedsgaard, J., & Nielsen, J. (2008). The yeast metabolome addressed by electrospray ionization mass spectrometry: Initiation of a mass spectral library and its applications for metabolic footprinting by direct infusion mass spectrometry. *Metabolomics*, *4*(4), 393–405. http://dx.doi.org/10.1007/s11306-008-0132-4.

Jarvis, W. R. (1962). The infection of strawberry and raspberry fruits by *Botrytis cinerea* Fr. *The Annals of Applied Biology*, *50*(3), 569–575. http://dx.doi.org/10.1111/j.1744-7348.1962.tb06049.x.

Jolliffe, I. (2002). *Principal component analysis* (2nd ed.). New York: Springer.

Jordan, A., Haidacher, S., Hanel, G., Hartungen, E., Mark, L, Seehauser, H., et al. (2009). A high resolution and high sensitivity proton-transfer-reaction time-of-flight mass spectrometer (PTR-TOF-MS). *International Journal of Mass Spectrometry*, *286*(2–3), 122–128. http://dx.doi.org/10.1016/j.ijms.2009.07.005.

Kähkönen, M. P., Hopia, A. I., & Heinonen, M. (2001). Berry phenolics and their antioxidant activity. *Journal of Agricultural and Food Chemistry*, *49*(8), 4076–4082. http://dx.doi.org/10.1021/jf010152t.

Klee, H. J. (2010). Improving the flavor of fresh fruits: Genomics, biochemistry, and biotechnology. *The New Phytologist*, 187(1), 44–56. http://dx.doi.org/10.1111/j.1469-8137.2010.03281.x.

Larrosa, M., González-Sarrías, A., García-Conesa, M. T., Tomás-Barberán, F. A., & Espín, J. C. (2006). Urolithins, ellagic acid-derived metabolites produced by human colonic microflora, exhibit estrogenic and antiestrogenic activities. *Journal of Agricultural and Food Chemistry*, 54(5), 1611–1620. http://dx.doi.org/10.1021/jf0527403.

Lindinger, W., Hansel, A., & Jordan, A. (1998). Proton-transfer-reaction mass spectrometry (PTR-MS): On-line monitoring of volatile organic compounds at pptv levels. *Chemical Society Reviews*, 27(5), 347–354.

Malowicki, S. M. A., Martin, R., & Qian, M. C. (2008). Volatile composition in raspberry cultivars grown in the Pacific northwest determined by stir bar sorptive extraction-gas chromatography–mass spectrometry. *Journal of Agricultural and Food Chemistry*, 56(11), 4128–4133. http://dx.doi.org/10.1021/jf073489p.

Mattoli, L., Cangi, F., Ghiara, C., Burico, M., Maidecchi, A., Bianchi, E., et al. (2010). A metabolite fingerprinting for the characterization of commercial botanical dietary supplements. *Metabolomics*, 7(3), 437–445. http://dx.doi.org/10.1007/s11306-010-0268-x.

McDougall, G., Martinussen, I., & Stewart, D. (2008). Towards fruitful metabolomics: High throughput analyses of polyphenol composition in berries using direct infusion mass spectrometry. *Journal of Chromatography B*, 871(2), 362–369. http://dx.doi.org/10.1016/j.jchromb.2008.06.032.

Mielke, L. H., Erickson, D. E., McLuckey, S. A., Müller, M., Wisthaler, A., Hansel, A., et al. (2008). Development of a proton-transfer reaction-linear ion trap mass spectrometer for quantitative determination of volatile organic compounds. *Analytical Chemistry*, 80(21), 8171–8177. http://dx.doi.org/10.1021/ac801328d.

Pers, T. H., Albrechtsen, A., Holst, C., Sorensen, T. I. A., & Gerds, T. A. (2009). The validation and assessment of machine learning: A game of prediction from high-dimensional data. *PLoS One*, 4(8). http://dx.doi.org/10.1371/journal.pone.0006287.

Prazeller, P., Palmer, P. T., Boscaini, E., Jobson, T., & Alexander, M. (2003). Proton transfer reaction ion trap mass spectrometer. *Rapid Communications in Mass Spectrometry*, 17(14), 1593–1599. http://dx.doi.org/10.1002/rcm.1088.

R Development Core Team (2009). *R: A language and environment for statistical computing.* Vienna, Austria: Recuperato da (http://www.R-project.org).

Reddy, M. V. B., Angers, P., Gosselin, A., & Arul, J. (1998). Characterization and use of essential oil from *Thymus vulgaris* against *Botrytis cinerea* and *Rhizopus stolonifer* in strawberry fruits. *Phytochemistry*, 47(8). http://dx.doi.org/10.1016/S0031-9422(97)00795-4.

Scott, I. M., Vermeer, C. P., Liakata, M., Corol, D. I., Ward, J. L., Lin, W., et al. (2010). Enhancement of plant metabolite fingerprinting by machine learning. *Plant Physiology*, 153(4), 1506–1520. http://dx.doi.org/10.1104/pp. 109.150524.

Sekine, T., Sugano, M., Majid, A., & Fujii, Y. (2007). Antifungal effects of volatile compounds from black zira (*Bunium persicum*) and other spices and herbs. *Journal of Chemical Ecology*, 33(11), 2123–2132. http://dx.doi.org/10.1007/s10886-007-9374-2.

Steeghs, M. M. L., Crespo, E., & Harren, F. J. M. (2007). Collision induced dissociation study of 10 monoterpenes for identification in trace gas measurements using the newly developed proton-transfer reaction ion trap mass spectrometer. *International Journal of Mass Spectrometry*, 263(2–3), 204–212. http://dx.doi.org/10.1016/j.ijms.2007.02.011.

Sterner, J. L, Johnston, M. V., Nicol, G. R., & Ridge, D. P. (2000). Signal suppression in electrospray ionization Fourier transform mass spectrometry of multi-component samples. *Journal of Mass Spectrometry*, 35(3), 385–391. http://dx.doi.org/10.1002/(SICI)1096-9888(200003)35:3<385::AID-JMS947>3.0.CO;2-O.

Tani, A., Hayward, S., Hansel, A., & Hewitt, C. N. (2004). Effect of water vapour pressure on monoterpene measurements using proton transfer reaction-mass spectrometry (PTR-MS). *International Journal of Mass Spectrometry*, 239(2–3), 161–169. http://dx.doi.org/10.1016/j.ijms.2004.07.020.

Tholl, D., Boland, W., Hansel, A., Loreto, F., Röse, U. S. R., & Schnitzler, J. (2006). Practical approaches to plant volatile analysis. *The Plant Journal*, 45(4), 540–560. http://dx.doi.org/10.1111/j.1365-313X.2005.02612.x.

Vapnik, V. (1995). *The nature of statistical learning theory.* New York: Springer.

Wold, S., Sjöström, M., & Eriksson, L. (2001). PLS-regression: A basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58, 109–130. http://dx.doi.org/10.1016/S0169-7439(01)00155-1.