Class-modeling approach to PTR-TOFMS data: a peppers case study

Cosimo Taiti[1], Corrado Costa[2*], Paolo Menesatti[2], Diego Comparini[1], Nadia Bazihizina[1], Elisa Azzarello[1], Elisa Masi[1], Stefano Mancuso[1]

[1] Università degli Studi di Firenze, Dipartimento di Scienze delle Produzioni Agroalimentari e dell'Ambiente – Viale delle Idee, 30, 50019 – Sesto Fiorentino (FI), Italy.

[2] Consiglio per la Ricerca e la sperimentazione in Agricoltura, Unità di ricerca per l'ingegneria agraria – Via della Pascolare 16, 00015 Monterotondo scalo (Rome), Italy.

* Corresponding author: Corrado Costa - Consiglio per la Ricerca e la sperimentazione in Agricoltura, Unità di ricerca per l'ingegneria agraria – Via della Pascolare 16, 00015 Monterotondo scalo (Rome), Italy - Phone +39-0690675214 - Fax +39-0690625591 - E-mail corrado.costa@entecra.it

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/smj.6761

ABSTRACT

BACKGROUND: Proton Transfer Reaction-Mass Spectrometry (PTR-MS), in its recently developed implementation based on a time-of-flight mass spectrometer (PTR-TOFMS), was used to rapidly determine the volatile compounds present in fruits of *Capsicum spp*. RESULTS: We analyzed the volatile organic compounds emission profile of freshly cut chili peppers belonging to 3 species and 33 different cultivars. PTR-TOFMS data, analyzed with appropriate and advanced multivariate class-modeling approaches, perfectly discriminated among the three species (100% of correct classification in validation set). Variable Importance in Projection scores was used to select the 15 most important volatile compounds in discriminating the species. Particularly the best candidates for *Capsicum* species were compounds with measured m/z of 63.027, 101.096 m/z and 107.050, which were, respectively, tentatively identified as dimethylsulfide, hexanal and benzaldehyde. CONCLUSIONS: Based on the promising results, the possibility of introducing multivariate class-modeling techniques, differently from the classification approaches, in the field of volatile compounds analyses is discussed.

KEYWORDS: Chili pepper (*Capsicum spp.*); Proton transfer reaction-mass spectrometry; Volatile organic compounds (VOC); PLS-DA; Class-modeling; VIP scores.

INTRODUCTION

The genus *Capsicum* of the Solanaceae family is cultivated worldwide as spices, food and medicine and includes five main species: *Capsicum annuum*, *C. chinense*, *C. frutescens*, *C. baccatum*, *C. pubescens*. Two main factors contribute to the flavor perception of the fruits of Capsicum: pungency and aroma, and these are associated to the fruit volatile compounds. Over the last decades, consumers have become more demanding in term of experiencing new aromas and flavors, and therefore the profile of fruit volatile compounds is now considered an important factor in

determining fruit quality [1]. In this context, the characterization and evaluation of Capsicum's species is of particular importance for gene bank curators, since these species present a wide variability yet to be fully known and exploited [2,3]. The *Capsicum* species are traditionally classified by morphological descriptors or related traits [3], with the following traits being the main taxonomic descriptors: *i.* flowers morphology, including the flower color; *ii.* calix constriction; and *iii.* number of floral for axil [2,3]. Although fruit volatile composition is often not considered an important factor for fruit classification, it has recently emerged that the different compositions of fruit volatiles compounds could be used to separate different *Capsicum* species [4].

Plants emit a multitude of volatile organic compounds (VOCs) in various tissues, and these are fundamental for the characterization of agro-industrial products, including fruits, and for consumer choice [5,6]. Most importantly, fingerprinting of VOCs present in fruits can be used for non-destructive characterization and identification of the different cultivars (e.g. strawberry fruits) [5]. Although currently gas chromatography–mass spectrometry (GC-MS) is the technique commonly used to identify VOCs, as it is a highly sensitive tool for VOCs detection (with suitable pre-treatment and pre-concentration stages, GC-MS systems can reach detection limits as low as 0.1 pptv), this technique is expensive and time consuming [6, 7]. Furthermore, GC-MS suffers from a relatively low time resolution and has an elevated risk of artefacts. Therefore an alternative physical tool, the PTR-MS was initially developed to study VOCs in air [8, 9] and was later on extended to food chemistry [10]. More recently, the evaluation of organic volatile compounds emitted by food has improved thanks to a new version of the PTR-MS that has been coupled with a time of flight mass analyser (PTR-TOFMS) with a PTR ion source and drift tube reaction chamber, which enables a precise, highly sensitive and real-time monitoring of many volatile compounds [11]. One of the advantages of the PTR-TOFMS is the enhanced analytical information provided, which allow a high mass resolution power ($m/\Delta m \sim 4000$) that enables to separate between many isobaric compounds [7, 12]. Since its development, the PTR-TOFMS has been used in a wide range of

fields, including environmental sciences, food monitoring [13] and VOC emissions from plants during various stress conditions [14, 15].

As stated by Forina et al [16], multivariate class-modeling techniques answer to the general question of whether an object *O*, stated of class *A*, really belong to class *A*. This is a typical question that is addressed in the traceability of Protected Denomination of Origin (PDO) foods or in multivariate quality control. On the contrary, the classification techniques assign objects to one of the classes in the problem. For example, linear Discriminant Analysis assigns an object to the class with the maximum posterior probability [5]. However, these classification techniques are not very useful in the control of quality, variety, origin or genuineness of a sample when considering their VOC profiles [17, 18]. Nevertheless, almost all research papers on food control use classification techniques; furthermore, also when a class-modeling technique is applied, the attention is focused on its classification performance rather than on its modeling characteristics. Class-modeling techniques calculate the "prediction probability" with a classification threshold for each modeled class. Using a class-modeling approach, it is possible to attribute objects not only into one or more classes but also to none (*i.e.*, in this case, the object is an outlier) [19].

The aim of the present work is the application of a Partial Least Squares Discriminant Analysis (PLS-DA) class-modeling approach based on volatiles data collected with the PTR-TOFMS to correctly classify different cultivars of pepper according to their species. In particular, we studied the VOC emission profile of 33 cultivars belonging to 3 different *Capsicum* species, as an an exercise to highlight the high potential of the PTR-TOFMS, when coupled with appropriate and advanced multivariate class-modeling approaches, in the field of volatile compounds analyses.

EXPERIMENTAL

Plant Material

A collection of chili peppers belonging to 3 species and 33 cultivars was analyzed (Table 1). For each cultivar 5 plants were grown. The seeds were placed in a substrate of peat and compost in separated jars and were maintained in darkness at 26°C until germination. Seedlings were moved to a greenhouse (25/18°C day/night) and were transplanted into larger pots 30 days after germination. Each variety was divided into blocks inside the greenhouse to limit the occurrence of cross-pollination; furthermore, given that the main effects of cross-pollination are generally present in the subsequent generations, its influence has not been considered.

For the VOC emission analyses, 3 uniform plants were selected and subsequently the uniform sized fruits were collected when the optimal ripening stage, based on the color changes (100% of coloration), was reached. All analyses were conducted at room temperature (25°C ±1).

PTR-TOFMS and VOC Emissions

VOCs emitted from chili peppers were analyzed with a PTR-TOFMS 8000 (Ionicon Analytik GmbH, Innsbruck, Austria). For a detailed explanation of the system see Lindinger et al. [8] and Brilli et al. [12]. Briefly, 20 g of freshly cut chilli pepper (including the seeds) were inserted in a glass jar (500 mL at 25°C, with a dynamic headspace flushing flow rate of 200 ml/min) equipped with two Teflon inlet and outlet tubes on opposite side, which were, respectively, connected to a zero-air generator (Peak Scientific) and the PTR-TOFMS. VOCs were then measured by direct injection of the head space mixture into the PTR-TOFMS drift tube via a heated (60°C) peek inlet tube with a flow rate of 100 sccm for 5 min. Measurements were carried out as previously described in Cappellin et al. [18] using a PTR-TOFMS in its standard configuration. The sampling time for each channel of TOF acquisition was 0.1 ns, amounting to 350,000 channels for a mass spectrum ranging up to m/z=250. The conditions in the drift tube were: drift voltage 600 V, temperature 110°C, pressure 2.25 mbar, extraction voltage at the end of the tube (Udx) 32V. Compounds with an exactly known mass such as 1,4 dichlorobenzene (m/z = 146.976) and 1,2,3 trichlorobenzene (m/z = 180.937) were continuously, and together with other known low mass ions, used for a

precise conversion of ‘‘time-of-flight’’ into ‘‘mass-to-charge’’ ratio (m/z) in order to assign the exact mass scale and the sum formula of all ions during VOC analysis [12].

For the data analysis, the average of the signal intensity was used. The external calibration automatically done by the acquisition program achieved a mass accuracy of 0.001 Th for the considered mass range, which was in most cases sufficient for sum formula identification. Spectra raw data (averaged count rate of the analytes recorded expressed in number of counts for second, cps) were acquired with TofDaq software (Tofwerk AG, Switzerland). For each sample, the average data, resulting from the last 30 consecutive seconds of the measurement, were extracted after 3 minutes from the beginning of the measurements. All spectra were corrected for count losses due to the detector dead time, applying Poisson correction in the DAQ settings of TofDAQ configuration options. Background measurements were run before every set of experiments by sampling the empty glass jar and were always subtracted before VOCs emission rates calculation.

Class-Modeling Approach

A Partial Least Squares Discriminant Analysis (PLS-DA) approach was used in order to predict the species identity of each sample. PLS-DA consists of a classical PLS regression analysis where the response variable is categorical (y-block; replaced by a set of dummy variables describing the categories, i.e. species identity), thus expressing the class membership of the statistical units [20-22]. The model included a calibration phase and a cross-validation phase. The prediction ability of PLS-DA also depends on the number of latent vectors (LV) used in the model. The x-block were pre-processed using a autoscale algorithm (i.e. centers columns to zero mean and scales to unit variance). For each sample a prediction probability to belong to each of the modeled y-block categories (*i.e.*, species identity) was calculated (modeling characteristics *sensu* [16]). Threshold values of the prediction probability were calculated for each y-block category. The threshold values are calculated using the observed distribution of predicted values and Bayesian statistics. By this way, an object could belong to none (outlier), one or more than one category, if the prediction

probabilities for each category exceed the threshold values. This analysis also expresses the statistical parameters indicating the modeling efficiency in terms of sensitivity and specificity of the parameters. The sensitivity is the percentage of the samples of a category accepted by the class model. The specificity is the percentage of the samples of the categories different from the modeled one, which is rejected by the class model. Generally, the trend of the residual errors is decreasing in the calibration phase (root mean square error of calibration—RMSEC) and increasing in the cross-validation phase (root mean square error of cross-validation RMSECV). The entire dataset was subdivided into two groups: (1) 90% of specimens for the class modeling and cross-validation, and (2) 10% of specimens for the independent test (*i.e.*, validation), optimally chosen with the Euclidean distances based on the algorithm of Kennard and Stone [23] that selects objects without the a priori knowledge of a regression model (i.e., the hypothesis is that a flat distribution of the data is preferable for a regression model). For each LV models with the mean higher performance value were considered to be more robust (robustness *sensu* [24]). Moreover, the Variable Importance in Projection (VIP) scores was calculated [25]. VIP scores estimate the importance of each variable, for each species, in the PLS-DA model. Variables with VIP scores significantly higher than 1 (one) are of great importance and might be good candidates for indicators for species selection. The models were developed using a procedure written in the MATLAB 7.1 R14 environment.

RESULTS

Peak extraction allowed the detection of more than 700 peaks in the range of measured masses (*m/z*=30-250), derived from the protonation of various VOCs, with an estimated headspace concentration higher than 1 ppbv. The average mass spectra obtained for each variety (in the range of *m/z*=30-250) were classified according to the different species, as displayed in Figure 1.

Table 2 shows the performance indicators of the model with 5 LVs selected as the more robust. It is possible to observe how the model could perfectly discriminate among all the samples of the three species, both in the model/validation dataset and in the independent test set (percentages of correct

classification). Also the specificity and sensitivity were both 100%, with the mean classification error equal to zero and a very low RMSEC (0.35). Figure 2 shows the plot of the scores of the 99 samples (3 replicates for each cultivar), grouped following their species identity, on the first three LV was represented. Considering that the whole model is composed by 6 LVs, the first three ones (x-block 63.3%; y-block 47.1% of cumulated variance) could still return a partial separation among the groups. For each species identity, the threshold values, calculated using the observed distribution of predicted values and Bayesian statistics. are equal to 0.08 (*C. annuum*), 0.19 (*C. baccatum*) and 0.05 (*C. chinense*). Considering the prediction probabilities, all the samples, both in the training/validation dataset and in the test dataset, exceed the threshold values only of its species identity. That means that no sample belonged to more than one species category, or was an outlier.

Table 3 reports these protonated masses together with the molecular formula, the tentative of identifications and the VIP scores for each species. Bibliographic citations listed in Table 3 refer to volatile compounds identified in other studies on pepper by means of gas chromatography and showing the same molecular mass.

DISCUSSION AND CONCLUSIONS

PTR-TOFMS is a tool, with a great potential in a wide range of fields, including food monitoring [13]. Nevertheless, the huge amount of data produced by the instrument can be underexploited and with the present paper we wants to introduce the concept of multivariate modeling, in particular using a class-modeling approach, in the field of volatile compounds analyses to further improve PTR-TOFMS data mining. On the basis of literature data, the detected peaks with VIP scores higher than 1 among the different species were identified (Table 3). In order to have more reliable results, we considered only VOCs cited in the literature data on peppers (see Table 3 for major details on the references), taking into account of the available fragmentation patterns of pure standards [26]. The only exception was the compound with measured protonated *m/z*=71.049, tentatively identified

as methyl vinyl ketone or methacrolein, as it is produced by plant leaves through the oxidation of isoprene [12, 27].

Volatile compounds with higher VIP value might be good candidates as indicators for species selection. In particular the chemical species with the higher significance both for *C. baccatum* and *C. annuum* were C_2H_6S (measured *m/z* = 63.027), $C_6H_{12}O$ (measured *m/z* = 101.096) and C_7H_6O (measured *m/z* = 107.050), while for *C. chinense* are C_2H_6S (measured *m/z* = 63.027), $C_{11}H_{16}$ (measured *m/z* = 149.133) and $C_9H_{14}N_2O$ (measured *m/z* = 167.219). The five volatile compounds listed above possibly refer to the following molecules: dimethylsulfide [28], hexanal [29, 30], benzaldehyde [31, 32], ectocarpene [28], 2-isobutyl-3-methoxypyrazine [24, 27]. Dymethilsulfide showed the higher VIP score for the classification of *C. annuum* and *C.baccatum*. This compound has been previously measured by PTR-TOFMS [32] and reported, by using GC-MS, as a typical volatile emitted by *C. baccatum* [28], as clearly shown in figure 3. Carbonyl compounds, in particular those with measured *m/z*= 87.045, 101.096, and 107.050 were also found to have a high VIP score (Tab. 3); these compounds are likely to correspond respectively to 2,3-butanedione, benzaldehyde and hexanal, which are chemical compounds typically produced by enzymatic action upon tissue destruction [30, 31]. Furthermore the analysis highlighted other two important carbonyl compounds with measured *m/z*= 99.081 and 117.093, tentatively attributed to cis-3-hexenal and hexanoic acid, which are among others possible carbonyl compounds produce by *Capsicum* species [28, 33].

Terpenoids (measured *m/z*= 137.133 and 181.250), sesquiterpenoids (*m/z*= 205.195) and the compounds with *m/z*= 167.219 also emerged as good candidates as indicators for species selection (Tab. 3). Indeed, other studies conducted with gas chromatography-mass spectrometry showed the presences of different type of terpenoids and sesquiterpenoids in *Capsicum* species [4, 24, 26], and cubebene, copaene and β-caryophyllene have been shown to be common sesquiterpenes in the

Capsicum genus whereas the most common terpenes are limonene, pinene and δ³-carene [4, 28, 30, 31]. In particular the chemometric analysis identified the ion with a measured m/z=181.250 as an important mass for the discrimination between *Capsicum* species. This ion was tentatively identified as the dihydroactinidiolide, a volatile terpene already reported in peppers [34, 35]. Finally, the chemometric analysis highlighted several nitrogen compounds, but only the ion with measured *m/z*= 167.219, identified as 2-isobutyl-3-methoxypyrazine [28] showed a high VIP score value. Pyrazine and other alkyl-methoxypirazine have been found to be important compounds in the *Capsicum* genus [30]; in particular, it has been shown that 2-isobutyl-3-methoxypyrazine is a chemical compounds produced in *C. annuum* which is associated with typical fresh green pepper flavor, and this compound has been used as an aroma descriptor of Jalapeno, Anaheim and Fresno cultivar [31].

VOCs fingerprinting with the PTR-TOFMS is a tool with a great potential, but due to the huge amount of data produced by the instrument there is the risk to underexploited them, limiting the potential of the instrument. There is a limited number of works that have explored the data provided by the PTR-TOFMS in a multivariate way [18]. As reported by Costa et al. [19], it is important to distinguish two main analytical approaches for multivariate supervised techniques: modeling and classification. For the modeling approach, it is possible to attribute objects not only into one or more classes but also to none (i.e. in this case, the object is an outlier). Moreover, modeling techniques calculate the "prediction probability" with a classification threshold for each modeled class. In the present work for the first time we applied such approach, which enabled us to further exploit the whole potentiality of the PTR-TOFMS and to have perfect classification results in both training and validation phases. PTR-TOFMS analysis together with the class-modeling approach presented here, within the food sector could aim to improve the *i.* food quality control [11] through aromatic profile, *ii.* food safety (such as biological contamination) and *iii.* food traceability integrated with infotracing systems [31] for the PDO. Thus from an applicative point of view, in

food science and technology, PTR-TOFMS offers the possibility to use VOCs spectra as fingerprints to rapidly identify food samples; therefore, this tool could be used as an MS-e-nose, as an implementation of an electronic nose based on a mass spectrometer [7].

In the present work, fingerprinting of the volatile compounds emitted by chili pepper fruits enabled us to perfectly discriminate among the three specie; these results suggests that combining the dataset provided by the PTR-TOFMS with multivariate class-modeling techniques can be used for the rapid and non-destructive classification of the fruits. By using PTR-TOFMS analysis we were been able to detect in the volatile fraction more than 200 peaks, identifying significant differences both quantitative and qualitative among the three different species. In conclusion, volatile compounds involved in the creation of aroma and flavor typical of the chili species *C. annuum*, *C. baccatum* and *C. chinense* were fingerprinted using PTR-TOFMS in order to discriminate among the three species. The species analyzed were perfectly discriminated using a PLS-DA modeling approach; interestingly it was shown that only 15 volatile compounds were sufficient to characterize the different aromatic profile of the three species. By generalizing the results obtained with chili peppers in the present work, we hope to encourage the introduction of multivariate modeling techniques in the field of volatile compounds analyses.

ACKNOWLEDGEMENTS

This work was partially funded by the Italian Ministry of Agriculture, grant PEPIC.

REFERENCES

[1] Garruti D, Pinto F, Alves V C, Penha M F, Tobaruela E & Araujo I M Volatile profile and sensory quality of new varieties of *Capsicum chinense* pepper. Ciência e Tecnologia de Alimentos, Campinas, 33(supl. 1), 102–108 (2012).

[2] Ince A G, Karaca M & Onus A N Development and utilization of diagnostic DAMD-PCR markers for *Capsicum* accessions. Genetic Resources and Crop Evolution, 56, 211–221 (2009).

[3] Sudré C P, Gonçalves L S A, Rodrigues R, do Amaral Júnior A T, Riva-Souza E M & Bento S C Genetic variability in domesticated *Capsicum spp* as assessed by morphological and agronomic data in mixed statistical analysis. Genetics and Molecular Research, 9(1), 283–294 (2010).

[4] Bogusz J F, Marchi T A, Teixeira F J, Alcaraz Z C & Teixeira G H Analysis of the volatile compounds of Brazilian chilli peppers (*Capsicum spp.*) at two stages of maturity by solid phase micro-extraction and gas chromatography-mass spectrometry. Food Research International, 48, 98–107 (2012).

[5] Biasioli F, Gasperi F, Aprea E, Mott D, Boscaini E, Mayr D & Märk T D Coupling proton transfer reaction-mass spectrometry with linear discriminant analysis: a case study. Journal of Agricultural and Food Chemistry, 51(25), 7227–7233 (2003).

[6] Aparicio R & Harwood J Handbook of Olive Oil: analysis and properties. Second Edition, Springer (2013).

[7] Cappellin L, Loreto F, Aprea E, Romano A, Sánchez del Pulgar J, Gasperi F & Biasioli F PTR-MS in Italy: a Multipurpose Sensor with Applications in Environmental. Sensors, 13, 11923–11955 (2013).

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

[9] Lindinger W, Hansel A & Jordan A Review: On-line monitoring of volatile organic compounds at pptv levels by means of Proton-Transfer-Reaction Mass Spectrom-etry (PTR-MS) medical applications, food control and environmental research. International Journal of Mass Spectrometry and Ion Processes, 173, 191–241 (1998b).

[10] Hansel A, Jordan R, Holzinger P, Prazeller W, Vogel W & Lindinger Proton-transfer-reaction mass spectrometry: on-line trace gas analisys at ppb levels. International Journal of Mass Spectrometry and Ion Processes, 149–150, 609–619 (1995).

[11] Sulzer P, Edtbauer A, Hartungen E, Jürschik S, Jordan A, Hanel G, Feil S, Jaksch S, Märk L & Märk T D From conventional Proton-Transfer-Reaction Mass Spectrometry (PTR-MS) to universal trace gas analysis. International Journal of Mass Spectrometry, 321–322, 66–70 (2012).

[12] Brilli F, Ruuskanen T M, Schnitzhofer R, Muller M, Breitenlechner M, Bittner V, Wohlfahrt G,Loreto F & Hansel A Detection of plant volatiles after leaf wounding and darkening by proton transfer reaction 'time-of-flight' mass spectrometry (PTR-TOF). Plos One, 6, e20419. doi:10.1371/journal.pone.0020419 (2011).

[13] Fabris A, Biasioli F, Granitto P M, Aprea E, Cappellin L, Schuhfried E, Soukoulis C, Märk T D, Gasperi F & Endrizzi I 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. Journal of Mass Spectrometry, 45, 1065–1074 (2010).

[14] Brilli F, Hörtnagl L, Bamberger I, Schnitzhofer R, Ruuskanen T M, Hansel A, Loreto F & Wohlfahrt G Qualitative and quantitative characterization of volatile organic compound emissions from cut grass. Environmental Science & Technology, 46 (7), 3859–3865 (2012).

[15] Ruuskanen T M, Müller M, Schnitzhofer R, Karl T, Graus M, Bamberger I, Hörtnagl L, Brilli F, Wohlfahrt G & Hansel A Eddy covariance VOC emission and deposition fluxes above grassland using PTR-TOF. Atmospheric Chemistry and Physics, 11, 611–625 (2011).

[16] Forina M, Oliveri P, Lanteri S & Casale M Class-modeling techniques, classic and new, for old and new problems. Chemometrics and Intelligent Laboratory Systems, 93(2), 132–148 (2008).

[17] Granitto P M, Biasioli F, Aprea E, Mott D, Furlanello C, Mark T D & Gasperi F Rapid and non-destructive identification of strawberry cultivars by direct PTR-MS headspace analysis and data mining techniques. Sensors and Actuators B—Chemical, 121, 379–385 (2007).

[18] Cappellin L, Biasioli F, Granitto P B, Schuhfried E, Soukoulis C, Costa F, Märk T D & Gasperi F On data analysis in PTR-TOF-MS: from raw spectra to data mining. Sensors and Actuators B: Chemical, 155, 183–190 (2011).

[19] Costa C, Antonucci F, Pallottino F, Aguzzi J, Sun D W & Menesatti P Shape analysis of agricultural products: a review of recent research advances and potential application to computer vision. Food and Bioprocess Technology, 4, 673–692 (2011).

[20] Sjöström M, Wold S & Söderström B PLS Discrimination plots. In E. S. Gelsema & L. N. Kanals (Eds.), Pattern recognition in practice II. Amsterdam: Elsevier (1986).

[21] Sabatier R, Vivein M Amenta P Two approaches for discriminant partial least square. In: M. Schader, W. Gaul, & M. Vichi (Eds.), Between data science and applied data analysis. Berlin: Springer (2003).

[22] Menesatti P, Antonucci F, Pallottino F, Giorgi S, Matere A, Nocente F, Pasquini M, D'Egidio M G & Costa C Laboratory vs. in-field spectral proximal sensing for early detection of Fusarium head blight infection in durum wheat. Biosystems Engineering, 114, 289–293 (2013).

[23] Kennard R W & Stone A Computer aided design of experiments. Technometrics, 11, 137–148 (1969).

[24] Swierenga H, de Groot P J, de Weijer A P, Derksen M W J & Buydens L M C Improvement of PLS model transferability by robust wavelength selection. Chemometrics and Intelligent Laboratory Systems, 41, 237–248 (1998).

[25] Chong I G & Jun C H Performance of some variable selection methods when multicollinearity is present. Chemometrics and Intelligent Laboratory Systems, 78, 103–112 (2005).

[26] Soukoulis C, Cappellin L, Aprea E, Costa F, Viola R, 393 Märk TD, Gasperi F & Biasioli F PTR-ToF-MS, a novel, rapid, high sensitivity and non-invasive tool to monitor volatile compound release during fruit post-harvest storage: the case study of apple ripening. Food and Bioprocess Technology, 6, 2831–2843 (2013).

[27] Liu Y J, Herdlinger-Blatt I, McKinney K A & Martin S T Production of methyl vinyl ketone and methacrolein via the hydroperoxyl pathway of isoprene oxidation. Atmospheric Chemistry and Physics, 13, 5715–5730 (2013).

[28] Kollmannsberger H, Rodriguez-Barruezo A, Nitz S & Nuez F Volatile and capsaicinoid composition of aji (*Capsicum baccatum*) and rocoto (*Capsicum pubescens*), two Andean species of chili peppers. Journal of the Science of Food and Agriculture, 91, 1598–1611 (2011).

[29] Pino J, Sauri-Duch E, & Marbot R Changes in volatile compounds of Habanero chile pepper (*Capsicum chinense* Jack. cv. Habanero) at two ripening stages. Journal of Food Chemistry, 94, 394–398 (2006).

[30] Ziino M, Condurso C, Romeo V, Tripodi G & Verzera A Volatile compounds and capsaicinoid content of fresh hot peppers (*Capsicum annuum* L.) of different Calabrian varieties. Journal of the Science of Food and Agriculture, 89, 774–780 (2009).

[31] Mazida M M, Salleh M M & Osman H Analysis of volatile aroma compounds of fresh chilli (*Capsicum annuum*) during stages of maturity using solid phase microextraction (SPME). Journal of Food Composition and Analysis, 18, 427–437 (2005).

[32] Sánchez del Pulgar J S, Soukoulis C, Biasioli F, Cappellin L, García C, Gasperi F, Granitto P, Märk T D, Piasentier E & Schuhfried E Rapid characterization of dry cured ham produced following different PDOs by proton transfer reaction time of flight mass spectrometry (PTR-ToF-MS). Talanta, *85*, 386–393 (2011).

[33] McGaw D R, Holder R, Commissiong E & Maxwell A Extraction of volatile and fixed oil products from hot peppers. In: Proceedings of the 6th International Symposium on Supercritical Fluids, International Society for Advancement of Supercritical Fluids, Versailles, France, 28–30 (2003).

[34] Gahungu A, Ruganintwali E, Karangwa E, Xiaoming Z & Mukuzi D Volatile Compounds and Capsaicinoid content of fresh hot peppers (*Capsicum chinense*) Scotch Bonnet variety at red stage. Journal of Food Science and Technology, 3, 211–218 (2011).

[35] Grover N, Patni V 2013 Phytochemical characterization using various solvent extracts and GC-MS analysis of methanolic extract of Woodfordia fructicosa (L.) Kurz. leaves. International Journal of Pharmacy and Pharmaceutical Sciences 5, 291-295 (2013)

[36] Papetti P, Costa C, Antonucci F, Figorilli S, Solaini S & Menesatti P. A RFID web-based infotracing system for the artisanal Italian cheese quality traceability. Food Control, 27, 234-241 (2012).

TABLES

TABLE 1: List of the genotypes studied and their species membership

Code Number	**Species**	**Variety or common name**
1	*C. annuum*	Adorno
2	*C. annuum*	Amoremio
3	*C. annuum*	Akrata
4	*C. annuum*	Arlecchino
5	*C. annuum*	Black Pearl
6	*C. annuum*	Bolivian Rainbow
7	*C. annuum*	Cancun
8	*C. annuum*	Cascabel
9	*C. annuum*	Cayambé
10	*C. annuum*	Ciliegino
11	*C. annuum*	Cayenna Red
12	*C. annuum*	Dolcevita
13	*C. annuum*	El Fuego
14	*C. annuum*	Explosive ember
15	*C. annuum*	Fuego caliente
16	*C. annuum*	Grappolino
17	*C. annuum*	Passerotto
18	*C. annuum*	Pyramid
19	*C. baccatum*	Bird Aji
20	*C. baccatum*	Bishon Crown
21	*C. baccatum*	Brasileiro
22	*C. baccatum*	Campana
23	*C. baccatum*	Hot Lemon
24	*C. baccatum*	Jamy
25	*C. baccatum*	Rocotillo
26	*C. chinense*	Carioca
27	*C. chinense*	Cheiro
28	*C. chinense*	Fatalii
29	*C. chinense*	Habanero chocolate
30	*C. chinense*	Habanero Red Caribbean
31	*C. chinense*	Naga Morich
32	*C. chinense*	Peruvian Orange
33	*C. chinense*	Scotch Bonnet Red

TABLE 2: Characteristics and principal results of the PLS-DA model. N is the number of samples. n° units (Y-Block) is the number of species to be discriminated by the PLSDA. n° LV is the number of latent vectors for each model. Random Probability (%) is the probability of random assignment of an individual into a unit. *Ca* = *C. annuum*, *Cb* = *C. baccatum*; *Cc* = *C. chinense*.

N	99
n° units (Y-block)	3
n° LV	6
% Cumulated Variance X-block	63.3
% Cumulated Variance Y-block	47.1
Mean Specificity (%)	100.00
Mean Sensitivity (%)	100.0
Random Probability (%)	33.33
Mean Class. Err. (%)	0.00
Mean RMSEC	0.35 (0.56 *Ca*; 0.26 *Cb*; 0.23 *Cc*)
Mean RMSEP	0.47 (0.80 *Ca*; 0.36 *Cb*; 0.26 *Cc*)
% Corr. Class. Model	100
% Corr. Class. Independent Test	100

TABLE 3: Protonated masses having VIP scores greater than 1, molecular formula and references which already evidenced the importance of each volatile compound in chilli peppers.

Protonated measured *m/z*	**Protonated chemical formula**	**Protonated theoretical *m/z***	**Tentative identification**	**VIP scores**			**References**
				C. annum	***C. chinense***	***C. baccatum***	
59.048	$C_3H_7O^+$	59.049	2-propanone	1.56	1.4	1.88	[31]
63.027	$C_2H_7S^+$	63.026	Dimethylsulfide	2.23	1.52	3.02	[29]
71.049	$C_4H_7O^+$	71.049	Methyl vinyl ketone, methacrolein	1.33	0.85	1.64	[20, 28]
81.070	$C_6H_9^+$	81.069	Alkyl fragment	1.77	1.21	1.91	[27]
87.045	$C_4H_7O_2^+$	87.044	2,3-butanedione	1.36	1.03	1.52	[31, 32]
95.050	$C_6H_7O^+$	95.049	Phenol	1.46	0.65	1.98	[31]
99.081	$C_6H_{11}O^+$	99.081	cis-3-Hexenal	1.25	0.77	1.65	[30]
101.096	$C_6H_{13}O^+$	101.096	Hexanal	1.34	0.87	1.99	[31, 36]
107.050	$C_7H_7O^+$	107.049	Benzaldehyde	1.47	1.23	2.08	[32]
117.093	$C_6H_{13}O_2^+$	117.091	Hexanoic acid/hexanoates	1.62	1.07	1.98	[31, 34]
137.133	$C_{10}H_{17}^+$	137.132	Monoterpenes	1.32	0.53	1.49	[29-31]
149.133	$C_{11}H_{17}^+$	149.132	Ectocarpene	1.12	1.52	0.88	[29]
167.219	$C_9H_{15}N_2O^+$	167.228	2-isobutyl-3-methoxypyrazine	0.87	1.51	0.82	[29, 31]
181.250	$C_{11}H_{17}O_2^+$	181.254	Dihydroactinidiolide (terpene)	0.85	1.22	0.39	[35, 36]
205.195	$C_{15}H_{25}^+$	205.195	Sesquiterpenes	1.3	1.38	0.84	[1, 29, 30]

FIGURE CAPTIONS

Figure 1. Low mass region of the average PTR-TOFMS spectra in 3 different varieties belonging to three different species (Cayenna, *Capsicum annuum*; Hot lemon, *C. baccatum*; Naga Morich, *C. chinense*).

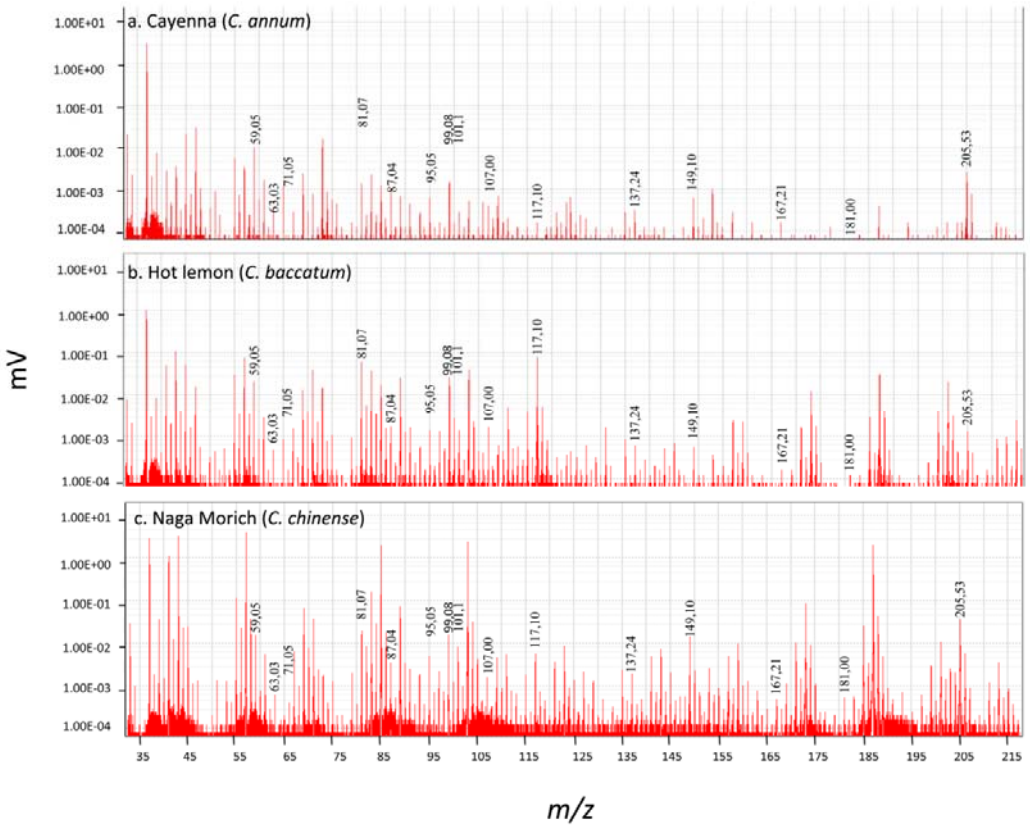

Figure 2: Representation of the peppers' samples on the first three axes of the PLS-DA (LV) model composed by 6 LVs.

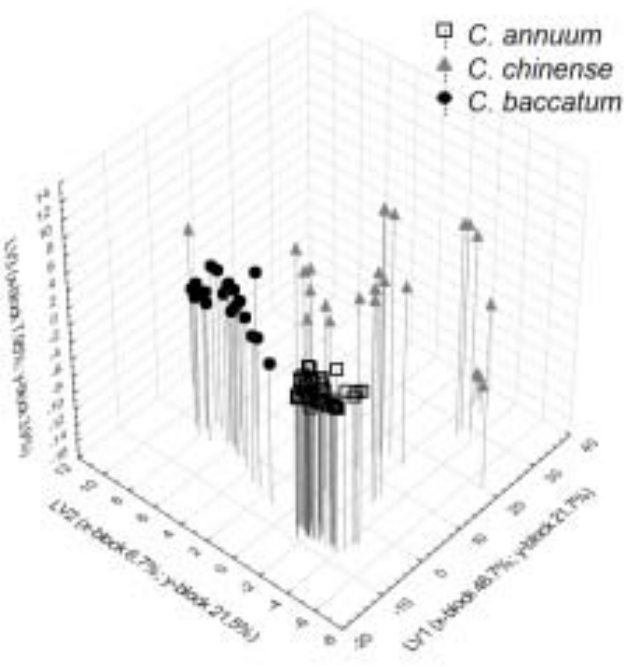

Figure 3: Example of the different VOC emission rates in the different species: emission rate of dimethylsulfide (m/z= 63.027) in freshly cut chilli pepper fruits from 3 different varieties belonging to three different species (Cayenna, *Capsicum annuum*; Nagamorich, *C. chinense*; Hot lemon, *C. baccatum*). Different colours indicate different species.