# Discrimination of geographical origin of oranges (*Citrus sinensis L. Osbeck*) by mass spectrometry-based electronic nose and characterization of volatile compounds

Valentina Centonze[a], Vincenzo Lippolis[b]*, Salvatore Cervellieri[b], Anna Damascelli[b], Grazia Casiello[a], Michelangelo Pascale[b], Antonio Francesco Logrieco[b], Francesco Longobardi[a,b]

[a]*Dipartimento di Chimica, Università di Bari "Aldo Moro", Via Orabona 4, 70126 Bari, Italy.*

[b]*Institute of Sciences of Food Production (ISPA), National Research Council of Italy (CNR), Via G. Amendola 122/O, 70126 Bari, Italy*

* Corresponding author: Tel: +39-080-5929457; fax +39-080-5929374;

e-mail address: vincenzo.lippolis@ispa.cnr.it

## Abstract

An untargeted method using headspace solid-phase microextraction coupled to electronic nose based on mass spectrometry (HS-SPME/MS-eNose) in combination with chemometrics was developed for the discrimination of oranges of three geographical origins (Italy, South Africa and Spain). Three multivariate statistical models, i.e. PCA/LDA, SELECT/LDA and PLS-DA, were built and relevant performances were compared. Among the tested models, SELECT/LDA provided the highest prediction abilities in cross-validation and external validation with mean values of 97.8% and 95.7%, respectively. Moreover, HS-SPME/GC-MS analysis was used to identify potential markers to distinguish the geographical origin of oranges. Although 28 out of 65 identified VOCs showed a different content in samples belonging to different classes, a pattern of analytes able to discriminate simultaneously samples of three origins was not found. These results indicate that the proposed MS-eNose method in combination with multivariate statistical analysis provided an effective and rapid tool for authentication of the orange's geographical origin.

**Key-words:** oranges, MS-based electronic nose, geographical origin, volatile compounds, chemometrics.

## 1. Introduction

Sweet orange (*Citrus sinensis L. Osbeck*) is one of the most popular fruits all over the world because it is very well-accepted by consumers for its nutritional, nutraceutical and sensorial attributes. Sweet oranges are usually classified into three main groups (i.e. common, navel, blood), with a diversification in terms of agronomical features within each group. Sweet orange accounting for 70% of worldwide citrus production is widely consumed both as fresh fruit as well as fruit juice. An annual global production was estimated at 73 million tons for oranges in 2016, and the main producers were Brazil, China, India and the United States (FAO, 2016).

Orange production at EU level was higher than 6 million tons for the 2016/2017 harvest and it is mainly concentrated in the Mediterranean basin with Spain and Italy representing about 80% of the total yield, followed by Greece and Portugal (Citrus Annual, 2017).

For both the high productive vocation of its geographical area and the quality of products, Italy occupies a prominent position in the production of oranges that amounted to 1,2 milion tons in the 2016/2017 season with geographic origin brands recognized by the European Commission (DOOR, 2018). Sicily and Calabria are the predominant production regions covering 80% of total Italian production (ISTAT, 2016). Although Italian oranges are considered of a premium quality, in the last few years the country has lost its leading role in the Mediterranean basin, due to the high costs of production and considerable loss of production due to the recent epidemic of the Citrus tristeza virus (CTV). In this contest, Italy commonly imports oranges from Spain and South-Africa which are gradually increasing their productions (Citrus Annual, 2017). Imports are mainly requested to cover the lack of availability of Italian products in some periods of the year (i.e. summer months), providing year-round availability for consumers. However, there are overlapping periods between Italian

and foreign productions, with an increased possibility for the consumer to buy mislabelled products with a lower product quality. Indeed, the geographical origin identification becomes more important for food with an origin label (e.g. "Made in Italy") since these products having acquired an “added-value” are more likely to become a target for frauds.

Considering that the illegal food trade is increasing around the world, traceability certifying the food authenticity including a correct labelling of origin is of great importance for traders, producers and consumers. Determination of food origin is commonly applied to control products with labelled or undeclared geographical provenances, for customs control and for self-control programs in the food industry.

For this reason, the development of rapid and reliable analytical methods to assess the geographical origin of food is highly demanded. For food authentication, non-targeted analysis (fingerprint) in combination with multivariate statistical analysis is a promising approach that allowing the detection of many metabolites as possible and permits to classify samples based on pattern of metabolites. A variety of non-targeted analytical techniques, mainly based on vibrational spectroscopy, mass spectrometry and NMR, have been applied to discriminate geographical origin of several food matrices and recently reviewed (Cubero-Leon et al. 2014, Essingler et al. 2014, Danezis et al. 2016), including fresh oranges (Diaz et al. 2014, Jandric and Cannavan 2017). Among these techniques, non-chromatographic mass spectrometry (MS) is an emerging approach for food authentication studies due to its several advantages in terms of rapidity, sensitivity, selectivity and high-throughput analysis (Danezis et al. 2016). In particular, the mass spectrometry-based electronic nose (MS-eNose) technique based on the use of headspace solid-phase microextraction (HS-SPME) directly coupled to MS is one of the most innovative approach that can be used to analyse volatile organic compounds (VOCs) of complex matrices. This technique provides a global mass

spectrometric fingerprint of VOCs of a sample, analyzed without chromatographic separation, in which each m/z ratio acts as a “sensor” whose intensity derives from the contribution of each compound producing that fragment. The main advantages of the proposed methodology are the minimum sample preparation required and the speed of analysis. The MS-eNose technique has been successfully applied to characterize several food matrices (i.e. coffee, raw spirits, milk and honey) for different purposes including the discrimination of geographical origin (Perez Pavòn et al. 2006, Jeleń et al. 2010, Liberto et al. 2013, Smyth and Cozzolino 2013).

Indeed, the aroma is one of the most important factors to discriminate food products and determine their quality. In particular, the aroma of fresh squeezed orange juice is a complex mixture of VOCs that is related to several factors, including orange cultivar, environment, geographical origin, degree of ripeness and storage conditions (Perez-Cacho and Rouseff 2008a, Cuevas et al. 2017). The VOCs of orange consist of esters, alcohols, aldehydes, ketones, terpenes and furans (Perez-Cacho and Rouseff 2008a, Cuevas et al. 2017). VOCs of fresh oranges and orange juices have been successfully characterised for different purposes by several studies using GC-MS analysis (Cuevas et al. 2017, Cerdàn-Calero et al. 2013, Reinhard et al. 2008, Cerdàn-Calero et al. 2012, Reid 2003, Zierler et al. 2004). However, to date the MS-eNose technique has not been applied for the discrimination of geographical origin of oranges.

For this reason, the aim of this study was to demonstrate the feasibility of MS-eNose technique applied to the VOCs analysis for the discrimination of the geographic origin of oranges. In particular, a robust and suitable non-targeted MS-based electronic nose method in combination with multivariate statistical analysis was developed and validated for the discrimination of oranges of three different geographical origins, i.e. Italy, South Africa and

Spain. Moreover, an HS-SPME/GC-MS method was used to characterize the possible pattern of volatile compounds having a role in this discrimination.

## 2. Materials and Methods

### *2.1 Chemicals and reagents*

Methanol (HPLC grade) and (E)-3-hexen-1-ol (≥98%) was purchased from Sigma Aldrich (Milan, Italy). Ten milliliters headspace vials with magnetic screw cap containing a pierceable PTFE/silicon septa were purchased from Agilent Technologies (Palo Alto, CA, USA). Helium at a purity of 99.9995% was obtained by Sapio s.r.l. (Bari, Italy). The automatic solid-phase microextraction (SPME) fiber holder was obtained from Gerstel (Mulheim an der Ruhr, Germany).

SPME-Fast Fit Fiber Assembly (FFA) divinylbenzene/carboxen/polydimethylsiloxane (DVB/CAR/PDMS, 50/30 µm, 1 cm fiber length), SPME Fiber Assembly for manual use divinylbenzene/carboxen/polydimethylsiloxane (DVB/CAR/PDMS, 50/30 µm, 1 cm fiber length), polydimethylsiloxane/divinylbenzene (PDMS/DVB, 65 µm, 1 cm fiber length), carboxen/polydimethylsiloxane (CAR/PDMS, 85 µm, 1 cm fiber length) and the manual SPME holder were purchased from Supelco (Bellafonte, PA, USA).

### *2.2 Samples collection and sample preparation*

Orange samples of the 2014/2015 crop season were collected from producers. A total of 137 samples of different cultivars of three different geographical origins, i.e. Italy, South Africa and Spain, was collected. Table 1 reports the number and the cultivars of samples collected for each geographic area.

The collected samples (five oranges for each sample) were squeezed and the juice was frozen at -20 ° C until the analyses. The stored juice after thawing was centrifuged for 20 min at 13000 rpm. Aliquots (2g) of the supernatant were placed in 10 mL headspace vials, adding as internal standard (E)-3-hexen-1-ol in methanol to obtain a concentration of 2 µg/g. Then, vials were sealed for the analysis with both analytical methods (i.e. HS-SPME/MS-eNose and HS-SPME/GC-MS). The extraction, desorbtion and sample introduction of the samples were performed automatically in HS-SPME/MS-eNose and manually in HS-SPME/GC-MS analysis.

*2.3 HS-SPME/MS-eNose analysis*

The squeezed orange juice samples were analysed by the mass spectrometry-based electronic nose (MS-eNose) GERSTEL Headspace ChemSensor System (GERSTEL, Mülheim, Germany) consisted of a headspace multi-purpose sampler MPS 2 (Gerstel, Mulheim an der Ruhr, Germany) and the Agilent 7890A GC System (Agilent Technologies, Palo Alto, CA, USA), modified for non-separative analysis with a deactivated fused-silica tubing (transfer column, 10 m x 0.18 mm i.d., 0 µm film thickness, Agilent Technologies), coupled to the Agilent 5975C inert MSD mass spectrometer. Moreover, the MS-eNose was online integrated with a multi-purpose sampler MPS 2 (Gerstel, Mulheim an der Ruhr, Germany), which was equipped with headspace incubation chamber and SPME sampling unit. An HS-SPME/MS-eNose protocol of analysis was *in-house* optimized according the procedure reported by Cefola et al. 2018. In particular, the headspace vial was kept at temperature of 40 °C for 10 min in the incubator-agitator of the MPS 2 autosampler to generate the headspace. The extraction from the headspace was performed by exposing a divinylbenzene/carboxen/polydimethylsiloxane (DVB/CAR/PDMS) fiber at 40 °C for 30 min. After extraction, compounds were thermally desorbed exposing the fiber in the CIS-4

programmed temperature vaporization (PTV) injector (Gerstel) of the MS-eNose at 250 °C for 5 min. Then, the MS-eNose analyses were carried out for 5 min using the following experimental conditions: the injection port fitted with a 1 mm i.d. liner was maintained at 250 °C in splitless mode; the oven, transfer line, ion source and quadrupole temperatures were 180, 280, 230 and 150 °C, respectively; the helium flow rate was held constant at 1 mL/min; Electron impact Ionization (EI+) mode with an electron energy of 70 eV was used; the mass spectrometer acquired data in full scan mode (scan range: 40–300 amu).

For each analysis mass spectral fingerprint was obtained by the software Chemsensor 6.912 (Gerstel, Mülheim and der Ruhr, Germany) corresponding to the sum of mass spectra obtained in the time range 0.22-2.0 min. Mass intensities of mass spectral fingerprint were estimated as relative abundances by comparing the mass intensity of each ion with the intensity of ion at 43 amu of the internal standard (i.e. (E)-3-hexen-1-ol).

*2.4 HS-SPME/GC-MS analysis*

A subset of 27 orange samples of different geographical origin (9 samples for each origin) were randomly selected from the entire set of 137 samples and analysed in duplicate by HS-SPME/GC–MS. The extraction and desorption steps of volatile compounds were performed following the same experimental parameters optimized for the HS-SPME/MS-eNose method, while GC-MS analysis was carried out according the procedure reported by Cefola et al. 2018 with some modifications. In particular, the GC-MS analyses were carried out by an Agilent 6890 Series GC system (Agilent Technologies, Palo Alto, CA, USA) equipped with a VF-WAXms (60 m x 0.25 mm i.d., 0.25 µm film tickness, Agilent Technologies) fused-silica capillary column and coupled to an Agilent 5973 Network Mass Selective Detector mass spectrometer. The injection port fitted with a 0.75 mm i.d. liner was maintained at 250 °C in

splitless mode. The analyses were performed with programmed temperature: initial temperature 40 °C maintained for 6 min, from 40 to 120 °C at 2 °C/min, 120 to 230 °C at 10 °C/min, the final temperature being maintained for 10 min. The helium flow rate was held constant at 1 mL/min. The transfer line, ion source and quadrupole temperatures were 280, 290 and 150 °C, respectively. Electron impact Ionization (EI+) mode with an electron energy of 70 eV was used. The mass spectrometer acquired data in full scan mode (scan range: 40–300 amu). The compounds were identified by comparison of experimental mass spectra with ones present in the NIST v2.0 and Wiley 138 libraries using a match quality higher than 70. The identification of volatile compounds was also verified by comparison of their linear retention indices (LRI) determined in relation to the retention times of C5–C14 and C8-C40 n-alkanes series, with those reported in literature [Zellner et al., 2008]. Quantification of compounds was performed by the same method of internal standardization used for HS-SPME/MS-eNose analysis. The amount of each identified compound was estimated by comparing the total ion current (TIC) peak area with (E)-3-hexen-1-ol peak area and expressed as area ratio. All mass spectrum fingerprints were combined to obtain a data matrix containing 137 objects and 260 variables that was submitted to statistical analyses.

*2.5 Statistical data analysis*

Before chemometric analysis, data obtained by the Chemsensor 6.912 software were pre-treated by baseline correction, through noise subtraction, and by internal normalization of the signal from each sample (Perez Pavon et al. 2006). The internal normalization allowed to remove the efficiency loss during the extraction process of volatile components due to the variability of the SPME fiber performance and MS signal instability. Subsequently, data were pre-processed by Pareto scaling and then submitted to multivariate statistical analyses.

The presence of outliers was evaluated observing the influence plot obtained by applying PCA (Principal Component Analysis) for each single class of different geographical origin. Samples identified as extreme outliers will be excluded. For this reason, NIPALS (Non-linear Iterative Partial Least Squares) algorithm was applied, considering V-fold equal to 10 in the cross validation process (CV=10), establish the exact number of PCs to use to build PCA models. PCA was also applied as exploratory technique with the aim to visualize if sample clustering was present as a function of the geographical origin of the samples (Jolliffe 2002).

Then, supervised pattern recognition techniques, i.e. Linear Discriminant Analysis (LDA) and Partial Least Squares Discriminant Analysis (PLS-DA) (Oliveri and Downey 2012), were used in order to classify orange samples on the basis of their geographical origin. For this purpose, the data matrix was divided in two subsets: a modeling set (containing 90 samples) and a test set (containing 47 samples). In particular, the modeling set, built using only Navel orange samples, was represented by 30 samples for each class (different geographical origin) and was used to build the statistical models, while the test set, consisting of 19 Italian, 22 African and 6 Spanish samples, was used to their validation.

In the case of LDA, to prevent model overfitting two different strategies PCA (unsupervised approach) and SELECT (a supervised feature selection algorithm) were used to reduce the number of variables that exceeded the number of objects (Berrueta et al. 2007, Casale et al. 2010, Vandeginste et al 1998). In particular, the number of variables should not exceed (n-g)/3, where n is the number of objects and g is the number of categories, i.e. 29, considering 90 objects (number of samples) and 3 categories (number of geographical origins).

On the other hand, PLS-DA was used as an alternative approach to avoid variables reduction being it frequently used in the case of large number of variables (Massart et al. 1997, Oliveri , 2017).

231 The PCA/LDA and PLS-DA models were built evaluating the proper number of principal
232 components and latent variables, respectively, which returned the lowest root mean square
233 error of cross validation (RMSECV). This parameter can guarantee that feature variables are
234 collected as much as possible and they are not overfitted. Therefore, performances of the
235 PCA/LDA, SELECT/LDA and PLS-DA models were compared in terms of recognition ability,
236 i.e. its ability to correctly classify the samples used for the building of the model, prediction
237 ability in cross-validation (CV), i.e. its ability to correctly classify samples of a test set
238 generated in a V-fold cross validation (with V equal to 10) and prediction ability in external
239 validation calculated using the test set.
240 Univariate statistical analysis, i.e. one-way analysis of variance (ANOVA) followed by a post
241 hoc Tukey's honestly significant difference (HSD) test ($p < 0.05$), was performed to assess
242 the differences between mean peak area ratios of identified volatile molecules of orange
243 samples of three different geographical origins obtained by HS-SPME/GC-MS analysis.
244 Data analyses were performed by using Pirouette software ver. 4.0 (Infometrix, Inc., Bothell,
245 WA, USA), V-Parvus release 2010 (http://www.parvus.unige.it, Genova, Italy), Classification
246 Toolbox in Matlab (Mathworks Inc., Natick, Massachusetts, USA) and Statistica 6.0 (StatSoft,
247 Tulsa).
248
249
250 **3. Results and Discussion**
251 *3.1 Geographical origin discrimination using HS-SPME/MS-eNose*
252 In order to find anomalous samples (outliers), data were processed in specific PCA models
253 for each geographical origin (class) showing that 9, 11, and 9 PCs explained 97.0, 97.0, and
254 95.0% of the total variance, for the Italian, South African, and Spanish origin, respectively.

Influence plots obtained by plotting the Mahalanobis distance versus sample residual showed that all the samples coming from a specific class fit in the respective model then excluding the presence of outliers.

Subsequently, to get a general overview of the data distributions an explorative PCA was applied on all data and by plotting the PC1 vs. PC2 sample scores (Figure 1) a poor visual clustering of the objects based on their geographical origin was showed (PC1 and PC2 explained respectively 42.5% and 14.5% of the total variance). Moreover, no significant separation was evidenced when the score plots of the remaining PCs were observed. This aspect was also confirmed evaluating the PC Fisher weights (FW) values, i.e. the measure of the between-class variance/within-class variance ratio, that resulted to be considerably lower than 1 (data not shown), meaning that no single PC was sufficiently suitable to distinguish samples for their geographical origin (Harper et al. 1977). Therefore, these results highlighted the necessity to use supervised techniques, i.e. discriminant techniques such as LDA and PLS-DA. These classification techniques were applied to data matrix divided into the two data subsets: modeling set and test set. The overall results of these classification models are reported and compared in Table 2.

In the case of LDA, two variable reduction strategies, i.e. PCA and SELECT were adopted to avoid model overfitting. In particular, PCA was applied to compress the information and the number of PCs chosen to get the lowest error in prediction cross validation and then used to build the PCA/LDA model was of 13 (CV procedure, V=10). The PCA/LDA model provided mean values of the recognition (classification) ability and CV prediction ability of 82.2% and 78.9%, respectively (Table 2). In particular, the model correctly predicted 21/30 Italian samples, 22/30 South African samples and 28/30 Spanish samples. Despite these low performances, the applicability of the model was also evaluated by an external test obtaining

279 a similar mean prediction ability of 80.9%. In the case of SELECT, 29 variables out of 260
280 were selected and then used to build the SELECT/LDA model. Mean percentages of
281 recognition ability and CV prediction ability obtained using SELECT/LDA model, were 100.0%
282 and 97.8%, respectively (Table 2). In particular, this model permitted to correctly classify all
283 South African and Spanish samples and 28 samples out of 30 Italian samples giving a value
284 of specific prediction percentage of 93.3% for this class.
285 Therefore, the supervised selection approach SELECT allowed significantly improvement of
286 results in terms of recognition and prediction abilities than those obtained using the
287 unsupervised PCA compression method. These results can be justified by considering that
288 the direction of maximum variability of data, used in the PCA approach, of the data could not
289 correspond to the direction of maximum discrimination among defined classes. Indeed, if the
290 variability associated with geographical origin is small with respect to the total variability, the
291 use of PCA variable reduction method can partially hide its specific feature contribution. On
292 the other hand, SELECT, by choosing the variables that contain the best information for the
293 under study classification, provides decorrelated variables avoiding redundant information. In
294 order to confirm these results, the SELECT/LDA model was also validated using the external
295 set. An external prediction ability of 95.7% was obtained with all South African and Spanish
296 samples correctly recognized (specific prediction rates of 100.0%) while only 2 samples out of
297 19 Italian samples were not correctly assigned, with a specific prediction rates of 89.5%.
298 Furthermore, PLS-DA was applied to test an alternative multivariate statistical approach of
299 classification and avoiding the process of variables reduction. By implementing a 10-fold
300 cross-validation, 12 latent variables guaranteed the optimal model complexity, leading to a
301 97.8% average recognition rate. In particular, the totality of the Italian and Spanish samples
302 were correctly classified, and only two out of 30 African samples was not correctly assigned.

The average CV prediction rate was 85.6% with CV prediction abilities for the Italian, African and Spanish categories of 83.3%, 76.7% and 96.7%, respectively. Moreover, the external validation procedure provided prediction abilities of 84.2% for Italy, 81.8% for South Africa and 100.0% for Spain orange samples, corresponding to an average prediction rate of 85.1%. These results showed that although PLS-DA model permitted acceptable prediction abilities, they were significantly lower than those obtained by the SELECT/LDA model.

These results demonstrated that HS-SPME/MS-eNose experimental data contain enough information to allow the construction of appropriate models for the discrimination of orange samples on the basis of their geographical origin.

*3.2 Characterization of the pattern of volatile compounds by HS-SPME/GC–MS*

In order to identify the most important volatile organic compounds (VOCs) to be used as markers in distinguishing oranges according to the country of origin, 27 orange samples of the three different geographical origins (9 for each class) were analysed by HS-SPME/GC-MS technique under the optimized experimental conditions. A total of 65 VOCs have been identified belonging to a wide range of chemical classes including aldehydes (5), ketones (4), esters (16), acids (3), alcohols (8), terpenes (20), heterocyclic compounds (4), saturated, unsaturated and aromatic hydrocarbons (5) (Table 3). In particular, aldehydes as hexanal and terpenes as D-limonene and β-linalool are responsible for characteristic orange juice flavour. Conversely, fruity notes are mainly due to ethyl butanoate with minor contributions from ethyl 2-methylpropanonate and ethyl 2-methylbutanoate (Perez-Cacho et al., 2008b). Moreover, high content of ester compounds have been shown to discriminate oranges obtained under conventional procedures from oranges cultivated under organic ones, characterized by high content of some terpenes and neryl acetate and geranyl acetate (Cuevas et al, 2017).

Accordingly, the composition (calculated as ratio of the analyte peak area relative to (E)-3-hexen-1-ol peak area) of fresh squeezed orange juice headspaces was investigated by one-way ANOVA analysis followed by a *post hoc* Tukey's HSD test in order to detect molecules discriminating samples in relation to their geographical origins. Among the identified compounds, the content of 28 molecules was significantly different among samples belonging to the three classes ($p < 0.05$). As reported in Table 3, 13 analytes allowed to dicriminate the geographical differences between Italian oranges and samples of the other two classes while only 2 volatile compounds discriminated South African oranges from the others. Moreover, 8 and 3 molecules differentiated Italian samples from Spanish and South African samples, respectively. Moreover, Spanish oranges were distinguished from the samples of the other two classes by the high content of methyl butanoate and only from South African samples by the presence of 1-terpinen-4-ol.

Among the selected analytes, (E)-2-hexen-1-ol, (Z)-3-hexen-1-ol, (E)-β-ionone and 6-methyl-5-hepten-2-one in the volatile fraction of Italian oranges showed the highest increase of contents from 7 to 19 times with respect to those measured for South African and Spanish oranges. (E)-2-hexen-1-ol and (Z)-3-hexen-1-ol have been already reported as volatile compounds of fresh prepared orange juice (Perez-Cacho et al., 2008b). Moreover, 4-methyl-heptane and 1-penten-3-one were not detected in the headspace of Italian samples while ethyl octanoate, 1-octen-3-ol and β-caryophyllene were absent in volatile fraction of Spanish oranges. Most molecules in the selected pattern have already been associated to fresh orange juice (Cuevas et al, 2017; Bai et al., 2014; Perez-Cacho et al., 2008b; Sádecká et al., 2014) with the exception of 2-methyl-1-pentene, 4-methyl-heptane, o-cymene and 4-methyl-2-heptanone, which were related for the first time to orange fresh squeezed juices.

However, although the HS-SPME/GC–MS analysis clearly highlighted difference in the VOCs profile of the oranges from the three geographical origins, a pattern of analytes able to discriminate simultaneously samples of the three different origins was not found. Consequently, the application of the multivariate analysis to the whole dataset obtained by MS-eNose analysis was confirmed to be the most appropriate approach to permit the rapid prediction of geographical origin of oranges.

## 4. Conclusion

In this study, a rapid and inexpensive method based on MS-eNose analysis in combination with chemometrics was successfully used to classify orange samples of three different geographical origins, i.e. Italy, South Africa and Spain.

In particular, three multivariate statistical approaches, i.e. PCA-LDA, SELECT-LDA and PLS-DA, were tested. Although, all tested statistical models permitted acceptable recognition and prediction abilities, the SELECT/LDA model showed the highest percentages in terms of prediction ability in cross-validation and external validation, with average values of 97.8% and 95.7%, respectively. The performances of the proposed method makes it suitable as powerful tool to assess the authenticity of oranges.

Although, HS-SPME/GC–MS analysis showed the absence of specific markers, differences in the pattern and content of VOCs of orange samples of the three different geographical origins were observed confirming the validity of the multivariate statistical approach used in this study.

## References


Bai, J., Baldwin, E., Hearn, J., Driggers, R., & Stover, E. (2014). Volatile profile comparison of usda sweet orange-like hybrids versus ‘Hamlin’ and ‘Ambersweet’. *Hortscience, 49*, 1262–1267.

Berrueta, L. A., Alonso-Salces, R. M., & Héberger, K. (2007). Supervised pattern recognition in food analysis. *Journal of Chromatography A, 1158*, 196–214.

Casale, M., Casolino, C., Oliveri, P., & Forina, M. (2010). The potential of coupling information using three analytical techniques for identifying the geographical origin of Liguria extra virgin olive oil. Food Chemistry, 118(1), 163-170.

Cefola, M., Damascelli, A., Lippolis, V., Cervellieri, S., Linsalata, V., Logrieco, A., & Pace, B. (2018). Relationships among volatile metabolites, quality and sensory parameters of 'Italia' table grapes assessed during cold storage in low or high CO2 modified atmospheres. Postharvest Biology and Technology, 142, 124-134.

Cerdán-Calero M., María Sendra J., & Sentandreu E., (2012). Gas chromatography coupled to mass spectrometry analysis of volatiles, sugars, organic acids and aminoacids in Valencia Late orange juice and reliability of the Automated Mass Spectral Deconvolution and Identification System for their automatic identification and quantification. *Journal of Chromatography A, 1241*, 84-95.

Cerdán-Calero M., Izquierdo L., & Sentandreu E., (2013). Valencia Late orange juice preserved by pulp reduction and high pressure homogenization: Sensory quality and gas chromatography–mass spectrometry analysis of volatiles. *LWT - Food Science and Technology, 51,* 476–483.

Citrus Annual, (2017). USDA Foreign Agricultural Service – Global Agricultural Information Network. 31.12.2017. https://www.fas.usda.gov/data/eu-28-citrus-annual-2. Accessed 17.07.2018.

Cubero-Leon, E., Peñalver, R., & Maquet, A., (2014). Review on metabolomics for food authentication. *Food Research International, 60*, 95–107.

Cuevas F. J., Moreno-Rojas J. M., & Ruiz-Moreno M. J., (2017). Assessing a traceability technique in fresh oranges (Citrus sinensisL. Osbeck) with an HS-SPME-GC-MS method. Towards a volatile characterisation of organic oranges. *Food Chemistry, 221*,1930–1938.

Danezis G. P.,Tsagkaris A. S., Brusic V., & Georgiou C. A., (2016). Food authentication: state of the art and prospects. *Current Opinion in Food Science, 10*, 22–31.

Díaz R., Pozo O. J., Sancho J. V., & Hernández F., (2014). Metabolomic approaches for orange origin discrimination by ultra-high performance liquid chromatography coupled to quadrupole time-of-flight mass spectrometry. *Food Chemistry, 157*, 84–93.

DOOR (Database of Origin and Registration), (2018). The Database of Origin and Registration (DOOR). European Commission, Agricultural and Rural Development, online database. URL http://ec.europa.eu/agriculture/quality/door/list.html. Accessed 17.07.2018

Esslinger, S., Riedl J., & Fauhl-Hassek C., (2014). Potential and limitations of non-targeted fingerprinting for authentication of food in official control. *Food Research International, 60*, 189–204.

FAOSTAT (Food and Agriculture Organization of the United Nations). Crops data, 2016. URL 1http://www.fao.org/faostat/en/#data/QC. Accessed 17.07.2018

Harper, A. M., Duewer, D. L., Kowalski, B. R., & Fashing J. L. (1977). ARTHUR and experimental data analysis: The Heuristic use of a polyalgorithm. In: B. R. Kowalski (Ed.),

ACS Symposium series 52. Chemometrics: Theory and Application (pp. 14–52). Washington: American Chemical Society.

ISTAT (Istituto Nazionale di Statistica). Tavola C24 - Superficie (ettari) e produzione (quintali): arancio, mandarino, clementina, limone, 2016. URL http://agri.istat.it/jsp/dawinci.jsp?q=plC240000010000012000&an=2016&ig=1&ct=506&id=15A|21A|31A. Accessed 17.07.2018

Jandric, Z., & Cannavan, A., (2017). An investigative study on differentiation of citrus fruit/fruit juices by UPLC-QToF MS and chemometrics. *Food Control, 72*, 173-180.

Jeleń, H. H.; Ziolkowska, A.; & Kaczmarek, A., (2010). Identification of the botanical origin of raw spirits produced from rye, potato, and corn based on volatile compounds analysis using a SPME-MS method. *Journal of Agricultural and Food Chemistry, 58*, 12585−12591.

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

Kim, H. K., & Verpoorte, R. (2010). Sample preparation for plant metabolomics. *Phytochemical Analysis, 21*, 4–13.

Liberto E., Ruosi M. R, Cordero C., Rubiolo P., Bicchi C. & Sgorbini B., (2013). Non-separative Headspace Solid Phase Microextraction−Mass Spectrometry Profile as a Marker To Monitor Coffee Roasting Degree. *Journal of Agricultural and Food Chemistry,* 61, 1652−1660.

Massart, D. L., Vandeginste, B. G. M., Buydens, L. M. C., De Jong S., Lewi P. J. & Smeyers-Verbeke J., (1997). Handbook of Chemometrics and Qualimetrics: Part A. Amsterdam: Elsevier.

Oliveri, P. & Downey, G. (2012). Multivariate class modelling for the verification of food-authenticity claims. *Trends in Analytical Chemistry, 35*, 74–86.

441 Oliveri, P., (2017). Class-modelling in food analytical chemistry: Development, sampling,
442 optimisation and validation issues – A tutorial. *Analytica Chimica Acta, 982*, 9-19.
443 Perez-Cacho P. R. & Rouseff R. (2008a). Processing and storage effects on orange juice
444 aroma: A review. *Journal of Agricultural and Food Chemistry, 56*, 9785–9796.
445 Perez-Cacho, P. R. & Rouseff, R. L. (2008b). Fresh Squeezed Orange Juice Odor: A Review.
446 *Critical Reviews in Food Science and Nutrition, 48*, 681–695.
447 Pérez Pavón, J. L. P., del Nogal Sanchez, M., Pinto, C. G., Laespada, E. F., Cordero, B. M.,
448 & Peña, A. G., (2006). Strategies for qualitative and quantitative analyses with mass
449 spectrometry-based electronic noses. *Trends in Analytical Chemistry, 25*, 257−266.
450 Reid W.J., (2003). Instrumental methods in detecting taints and off-flavours. In B. Baigrie
451 (Eds.) *Taints and Off-Flavours in Foods, A volume in Woodhead Publishing Series in Food*
452 *Science, Technology and Nutrition* (pp 31–63), Cambridge : Woodhead Publishing.
453 Reinhard H., Sager F., & Zoller O., (2008). Citrus juice classification by SPME-GC-MS and
454 electronic nose measurements. *LWT - Food Science and Technology, 41*, 1906–1912.
455 Sádecká, J., Polovka, M., Kolek, E., Belajová, E., Tobolková, B., Daško, L. & Durec, J.
456 (2014). Orange juice with pulp: impact of pasteurization and storage on flavour,
457 polyphenols, ascorbic acid and antioxidant activity. *Journal of Food & Nutrition Research,*
458 *53,* 371-388.
459 Smyth, H., & Cozzolino, D., (2013). Instrumental Methods (Spectroscopy, Electronic Nose,
460 and Tongue) As Tools To Predict Taste and Aroma in Beverages: Advantages and
461 Limitations. *Chemical Reviews, 113*, 1429−1440.
462 Vandeginste, B.G.M., Massart, D.L., Buydens, L.M.C., De Jong, S., Lewi, P.J., & Smeyers-
463 Verbeke, J. (1998). Handbook of Chemometrics and Qualimetrics: Part. B. In B.G.M.

Vandeginste, & S.C. Rutan (Eds.), Supervised Pattern Recognition (pp. 207-241). Amsterdam: Elsevier.

Zellner, B. D. A., Bicchi, C., Dugo, P., Rubiolo, P., Dugo, G., & Mondello, L. (2008). Linear retention indices in gas chromatographic analysis: a review. *Flavour and fragrance journal, 23*, 297–314.

Zierler, B., Siegmund, B., Pfannhauser, W., (2004). Determination of off-flavour compounds in apple juice caused by microorganisms using headspace solid phase microextraction–gas chromatography–mass spectrometry. *Analytica Chimica Acta, 520*, 3–11.

**Figure captions**

**Figure 1.** *PC1 vs PC2 scatter plot for orange samples. Geographical origins: Italy (o), Africa (□), Spain (+).*

## Table 1

Orange samples for each geographic area

| Geographical origin | Italy | South Africa | Spain |
|---|---|---|---|
| N. Samples | 49 | 52 | 36 |
| Cultivars | Washington Navel, Newhall Navel, Navel Foglia, Navelina<br>Bionda IGP, Duretta IGP, Ovale<br>Valencia, Lane Late | Navel, Navelina navel | Lane Late, Navel Pawel Navel |

## Table 2

Recognition, CV prediction abilities and external prediction for all models built classifying oranges samples according to their geographical origin.

| | Model performance (%) | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Recognition ability (Modelling)** | | | | **Prediction ability (CV[d] 10)** | | | | **External Prediction** | | | |
| | ITA[a] | S.A.[b] | SPA[c] | Mean | ITA | S.A. | SPA | Mean | ITA | S.A. | SPA | Mean |
| **PCA/LDA** (13 Principal Components) | 76.7 (23/30) | 80.0 (24/30) | 90.0 (27/30) | 82.2 | 70.0 (21/30) | 73.3 (22/30) | 93.3 (28/30) | 78.9 | 68.4 (13/19) | 86.4 (19/22) | 100.0 (6/6) | 80.9 |
| **SELECT/LDA** (29 variables) | 100.0 (30/30) | 100.0 (30/30) | 100.0 (30/30) | 100.0 | 93.3 (28/30) | 100.0 (30/30) | 100.0 (30/30) | 97.8 | 89.5 (17/19) | 100.0 (22/22) | 100.0 (6/6) | 95.7 |
| **PLS-DA** (12 Latent Variables) | 100.0 (30/30) | 93.3 (28/30) | 100.0 (30/30) | 97.8 | 83.3 (25/30) | 76.7 (23/30) | 96.7 (29/30) | 85.6 | 84.2 (16/19) | 81.8 (18/22) | 100.0 (6/6) | 85.1 |

[a]: Italy; [b]: South Africa; [c]: Spain; [d]: Cross Validation.

## Table 3

Volatile compounds (n = 65) identified by HS-SPME/GC–MS analysis of Italian, South African and Spanish oranges.

| Volatile Compound | Code | $LRI_{lt}/LRI_{sp}$[d] | Volatile Compound | Code | $LRI_{lt}/LRI_{sp}$ |
|---|---|---|---|---|---|
| ***Hydrocarbons*** | | | ***Alcohols*** | | |
| 2-methyl-1-pentene[b] | 1 | -[e]/644 | 2-methyl-2-propanol | 4 | 900/904 |
| 4-methyl-heptane[b] | 2 | 790/765 | ethanol | 5 | 937/937 |
| undecane | 15 | 1100/1100 | (Z)-2-penten-1-ol | 30 | 1329/1330 |
| 1,3-bis(1,1-dimethylethyl)-benzene | 36 | 1423/1433 | (Z)-3-hexen-1-ol[a,b] | 34 | 1393/1394 |
| 4-acetyl-1-methylcyclohexene | 43 | 1568/1566 | (E)-2-hexen-1-ol[a,b] | 35 | 1417/1416 |
| ***Terpenes*** | | | 1-octen-3-ol[b] | 38 | 1459/1459 |
| 1R-α-pinene[b] | 9 | 1022/1022 | 1-octanol | 44 | 1575/1577 |
| α-thujene | 11 | 1030/1027 | 2,4-bis(1,1-dimethylethyl)-phenol | 62 | 2321/2321 |
| β-phellandrene | 17 | 1183/1118 | ***Acids*** | | |
| 3-carene | 19 | 1146/1146 | acetic acid | 39 | 1468/1467 |
| D-limonene[a,b] | 23 | 1200/1203 | nonanoic acid[b] | 61 | 2184/2184 |
| γ-terpinen | 27 | 1249/1249 | dodecanoic acid | 63 | 2503/2495 |
| o-cymene[b] | 28 | 1276/1273 | ***Esters*** | | |
| α-terpinolene[a,b] | 29 | 1285/1286 | ethyl acetate | 3 | 895/896 |
| β-linalool | 42 | 1563/1562 | ethyl propanoate | 6 | 961/961 |
| β-caryophyllene[a,b] | 46 | 1603/1602 | ethyl 2-methyl-propanoate | 7 | 966/969 |
| 1-terpinen-4-ol[c] | 47 | 1616/1611 | methyl butanoate[b,c] | 8 | 993/992 |
| 4,11-selinadiene[a,b] | 52 | 1656/1703 | ethyl butanoate | 12 | 1041/1041 |
| α-terpineol | 53 | 1707/1708 | ethyl 2-methyl-butanoate | 13 | 1057/1056 |
| valencene[a,b] | 54 | 1728/1736 | diethyl carbonate | 16 | 1102/1114 |
| (S)-(+)-carvone[a,b] | 55 | 1748/1752 | ethyl (E)-2-butenoate[a,b] | 20 | 1164/1168 |
| nerol[a,c] | 56 | 1817/1819 | methyl hexanoate | 22 | 1192/1194 |
| (Z)-carveol[b] | 57 | 1856/1857 | ethyl hexanoate | 26 | 1239/1239 |
| (E)-carveol | 58 | 1882/1888 | 2-hexenyl acetate | 31 | 1323/1340 |
| β-ionone[a,b] | 59 | 1967/1969 | ethyl 2-hexenoate | 33 | 1357/1351 |
| nootkatone[b] | 65 | 2563/2590 | ethyl octanoate[a,b] | 37 | 1442/1442 |
| ***Aldehydes*** | | | methyl benzoate | 48 | 1636/1636 |
| hexanal[a,b] | 14 | 1086/1086 | butyrolactone | 49 | 1643/1647 |
| heptanal[a,b] | 21 | 1191/1191 | ethyl 3-hydroxy-hexanoate | 51 | 1673/1693 |
| (E)-2-hexenal | 25 | 1222/1222 | ***Heterocyclic compounds*** | | |
| benzaldehyde[a,c] | 41 | 1535/1535 | furfural | 40 | 1475/1475 |
| 3-methyl-benzaldehyde | 50 | 1624/1663 | 5-methyl-2-furfural | 45 | 1587/1586 |

| | | | | | |
|---|---|---|---|---|---|
| ***Ketones*** | | | 2,5-furandicarboxaldehyde | 60 | 2022/2017 |
| | | | 5-(hydroxymethyl)-2-furfural | 64 | 2530/2543 |
| 1-penten-3-one[a] | 10 | 1025/1025 | | | |
| 4-heptanone | 18 | 1162/1128 | | | |
| 4-methyl-2-heptanone[a] | 24 | 1206/1211 | | | |
| 6-methyl-5-hepten-2-one[a] | 32 | 1343/1343 | | | |

[a]: compound selected by *post hoc* Tukey's HSD test and discriminating between Italian and South African oranges.
[b]: compound selected by *post hoc* Tukey's HSD test and discriminating between Italian and Spanish oranges.
[c]: compound selected by *post hoc* Tukey's HSD test and discriminating between South African and Spanish oranges.
[d]: $LRI_{lt}$: Linear Retention Index reported in literature by www.pherobase.com, www.flavornet.org, www.chemspider.com and www.nist.gov; $LRI_{sp}$: Linear Retention Index calculated against n-alkanes ($C_5$–$C_{14}$ and $C_8$-$C_{40}$) on VF-WAXms column .
[e] not available.

**Figure 1**

**Click here to download Figure(s): Figure 1_Centonze et al.docx**

Figure 1