A geographical origin assessment of Italian hazelnuts: Gas chromatography-ion mobility spectrometry coupled with multivariate statistical analysis and data fusion approach

Food Research International

# A Geographical Origin assessment of Italian Hazelnuts: Gas Chromatography-Ion mobility spectrometry coupled with Multivariate Statistical Analysis and Data Fusion approach

--Manuscript Draft--

| | |
|---|---|
| Manuscript Number: | FOODRES-D-23-01114R2 |
| Article Type: | Research Paper |
| Keywords: | Hazelnut (Corylus avellana); Gas Chromatography-Ion Mobility System; Food authenticity; multivariate statistical analysis; sensory analysis; Data Fusion. |
| Corresponding Author: | Michele Suman<br><br>Parma, ITALY |
| First Author: | Giuseppe Sammarco |
| Order of Authors: | Giuseppe Sammarco |
| | Daniele Bardin |
| | Federica Quaini |
| | Chiara Dall'Asta |
| | Joscha Christmann |
| | Philipp Weller |
| | Michele Suman |
| Abstract: | Hazelnut is a commodity that has gained interest in the food science community concerning its authenticity. The quality of the Italian hazelnuts is guaranteed by Protected Designation of Origin and Protected Geographical Indication certificates. However, due to their modest availability and the high price, fraudulent producers/suppliers blend, or even substitute, Italian hazelnuts with others from different countries, having a lower price, and often a lower quality. To contrast or prevent these illegal activities, the present work investigated the application of the Gas Chromatography-Ion mobility spectrometry (GC-IMS) technique on the hazelnut chain (fresh, roasted, and paste of hazelnuts). The raw data obtained were handled and elaborated using two different ways, software for statistical analysis, and a programming language. In both cases, Principal Component Analysis and Partial Least Squares-Discriminant Analysis models were exploited, to study how the Volatile Organic Profiles of Italian, Turkish, Georgian, and Azerbaijani products differ. A prediction set was extrapolated from the training set, for a preliminary models' evaluation, then an external validation set, containing blended samples, was analysed. Both approaches highlighted an interesting class separation and good model parameters (accuracy, precision, sensitivity, specificity, F1-score). Moreover, a data fusion approach with a complementary methodology, sensory analysis, was achieved, to estimate the performance enhancement of the statistical models, considering more discriminant variables and integrating at the same time further information correlated to quality aspects. GC-IMS could be a key player as a rapid, direct, cost-effective strategy to face authenticity issues regarding the hazelnut chain. |
| Suggested Reviewers: | Carlo Bicchi<br>University of Torino<br>carlo.bicchi@unito.it |
| | Lourdes Arce<br>Universidad de Cordoba<br>lourdes.arce@uco.es |

# *JOURNAL SUBMISSION COVER LETTER*

Title: *A Geographical Origin assessment of Italian Hazelnuts: Gas Chromatography-Ion mobility spectrometry coupled with Multivariate Statistical Analysis and Data Fusion approach*

*Food Research International*

Submission date: *15/02/2023*

Dear Editor-In-Chief,

I am writing to submit our manuscript entitled "A Geographical Origin assessment of Italian Hazelnuts: Gas Chromatography-Ion mobility spectrometry coupled with Multivariate Statistical Analysis and Data Fusion approach" for consideration as a Food Research International research paper.

The study mainly focuses on how gas chromatography-ion mobility spectrometry (GC-IMS) technology could be a valuable tool for geographical origin discrimination, with particular attention to Italian hazelnuts. Fresh, roasted, and paste hazelnuts from Italy, Turkey, Georgia, and Azerbaijan were analysed, to gain information about their volatile organic compounds (VOCs), which could be related to geographical location. In all these matrices, chemical differences among Italian and not Italian samples were always confirmed, according to multivariate statistical analysis carried out with both software and coding. Thus, this approach results in a potential key analytical strategy for hazelnut traceability.

Furthermore, GC-IMS data were positively merged with data from sensory analysis as a complementary technique, aimed at improving the models' performance. Indeed, this data fusion approach allowed us to build multivariate statistical models, from unsupervised to supervised, underlining the separation between Italian and non-Italian clusters.

Considering the growing request for assessing the geographical location of hazelnuts, because of the high quality of the Italian "Tonda Gentile Trilobata", "Tonda Gentile Romana", and "Nocciola di Giffoni" (PGI and PDO certified), we assume that the outcomes presented in our paper will appeal to the food scientist community who subscribe to Food Research International. Our findings will allow your readers to get updated about a potential and innovative analytical solution for authenticity studies.

Each of the authors confirms that this manuscript has not been previously published and is not currently under consideration by any other journal. Additionally, all of the authors have approved the contents of this paper and have agreed to Food Research International's submission policies.

Each named author has substantially contributed to conducting the underlying research and drafting this manuscript. Additionally, to the best of our knowledge, the named authors have no conflict of interest, financial or otherwise.

Sincerely,

Giuseppe Sammarco

Advanced Laboratory Research, Barilla G. e R. Fratelli S.p.A., Parma, Italy

Via Mantova 166 - 43100 Parma (Italy)

Department of Food and Drug, University of Parma, Parma, Italy

Parco Area delle Scienze, 27 a, 43100 Parma PR

Giuseppe.sammarco@barilla.com

Giuseppe.sammarco1@unipr.it

Tel.: +39 0521 263662

Additional Contact

Prof. Dr. Michele Suman, Corresponding Author

Advanced Laboratory Research, Barilla G. e R. Fratelli S.p.A., Parma, Italy

Via Mantova 166 - 43100 Parma (Italy)

Department for Sustainable Food Process, Catholic University Sacred Heart, Piacenza, Italy

Via Emilia Parmense, 84, 29122 Piacenza (Italy)

Michele.suman@barilla.com

Reviewer #1: This study investigated the application of the GC-IMS coupled with multivariate statistical analysis could effectively identify the authenticity of hazelnut chain origin. While the study is interesting, there were several shortcomings in this manuscript as the follows:

1. The introduction part is too long, with too much content, and the language is not concise enough.

Response: the chapter has been shortened, and the language is now more concise.

2. Line 123-124, "extensively" only cites 2 documents, which cannot explain the extensive research.

Response: the sentence has been corrected.

3. Line 125-127, please provide theoretical basis and reference.

Response: the reference about this work is reported later in the sentence, line 129 (Ciarmiello, et al., 2014).

4. Compared to fresh hazelnuts and hazelnut pastes, the accuracy of the model prediction for roasted hazelnuts is very high, even up to 100%. I am curious about the reasons for this difference. what was authors' explanation?

Response: the number of samples considered for the roasted hazelnut analyses was bigger than the other hazelnut-based matrices. Both peeled and unpeeled samples were included. This bigger number of samples lowered the data dispersion, enhancing the model's performance. Further, the roasted samples presented a richer chromatogram than the other matrices. This is probably due to the industrial process that favoured the presence of diverse compounds in the volatile fraction. A higher number of samples and variables could so explain excellent values of accuracy for the model prediction.

5. Raw data on sample GC-IMS and sensory analysis should be presented in the text, including sensory characteristics of hazelnuts from different producing regions.

Response: raw data obtained from GC-IMS analyses are intensities related to the spots of the chromatogram. They do not contain relevant/interpretable information as they are, for this reason we did not insert them into the paper. As regard the sensory data and characteristic, the table including all the data were placed into the supplementary material in order to not overload the text with three big tables.

6. Line 9-14 "permitted valuable discrimination among the groups (Manfredi, et al., 2018). polymerase chain reaction (PCR)". Is this one sentence or two?

Response: these are two sentences, a capital letter was used for the word 'Polymerase'.

7. Line 142, "0.2 µg/m3 to 2 mg/m3". Is the unit format correct?

Response: yes, according to the source considered: Ruszkiewicz, D., Myers, R., Henderson, B., Yusof, H., Meister, A., Moreno, S., . . . Thomas, C. (2022). Peppermint protocol: first results for gas chromatography-ion mobility spectrometry. Journal of Breath Research, 036004.

8. Line 224, "(Vera, Companioni, Meacham, & Gygax, 2016)". Which sentence does this reference belong to?

Response: the reference belongs to this part: “The output of a GC-IMS analysis was a 3D graph, with the y-axis as GC retention time, the x-axis as the IM drift time, and the z-axis showing the detector response or signal intensity. To favour the graphical visualisation, the heat map 2D fingerprint representation was ideal, the 3D-2D conversion was achieved by transforming the z-axis into a colour signature for each spot. Thus, the more intense the signal, the more coloured the spot” (lines 219-224). The position of the dot has been changed, after the reference.

9. The writing style is too verbose and needs to be streamlined.

Response: some less concise parts have been modified, in order to make the paper easier to read.

Reviewer #2: The manuscript FOODRES-D-23-01114 (Geographical Origin assessment of Italian Hazelnuts: Gas Chromatography-Ion mobility spectrometry coupled with Multivariate Statistical Analysis and Data Fusion approach) by G. Sammarco et al. deals with the application of a recent technique (GC-IMS) combined with statistical methods to the evaluation of the geographical origin of Italian hazelnuts. In my opinion the manuscript is interesting and can be accepted for publication in Food Research International after minor revision.

Two important integrations are required:

1) The first point concerns the application of the method to routine quality control. The authors correctly write in the Conclusions (lines 538-540 page 27) "Simultaneously, the speed, sensitivity, and cost effectiveness of the tool are also exploitable for routinary analysis in the companies quality control department." However, this statement must be confirmed by an experimental comparison with the technique normally available in the quality control laboratory, i.e. mostly gas chromatography-quadrupole mass spectrometry (GC-qMS). This consideration is even more useful in this case where sampling by static-HS does not involve a concentration of the analytes and the contribution of GC in terms of separation is less decisive because the analysis is performed isothermally. The interest in the proposed method would be increased by introducing a brief discussion of this aspect in the text.

Response: a brief discussion about the advantages of using GC-IMS in the quality control laboratory, instead of the standard technique GC-qMS, has been written in the Conclusions.

2) The second point is indirectly related to the previous one. The authors report an explained variance between 20 and 30%, which is objectively rather low or in some cases insufficient (Line 319 page 12). Why did the authors not optimise the PCA procedure by selecting significant loadings and eliminating those that are redundant or do not contribute to sample discrimination? A reduction in statistical noise would probably have been useful to increase the explained variance and "smooth out" some of the classification inconsistencies.

Response: a feature selection was carried out by selecting only the variables having a VIP score major than 1. This led to a higher explained variance (>40%) with a discrete separation being preserved. The score plots were added in the text, together with the selection criteria of the variables.

Minor points

Line 52, page 2: not "class" but "species" in agreement with the official botanical nomenclature

Response: corrected

Line 60, page 2: not "bigger" but "higher"

Response: corrected

Line 539 page 27: NOT "companies" but "company" or "companies' "

Response: corrected

# *A Geographical Origin assessment of Italian Hazelnuts: Gas Chromatography-Ion mobility spectrometry coupled with Multivariate Statistical Analysis and Data Fusion approach*

*Giuseppe Sammarco*[1,3], *Daniele Bardin*[1], *Federica Quaini*[1], *Chiara Dall'Asta*[3], *Joscha Christmann*[4], *Philipp Weller*[4], *Michele Suman*[1,2]

[1] Sensory and Analytical Food Science, Barilla G. e R. Fratelli S.p.A., Parma, Italy
[2] Department for Sustainable Food Process, Catholic University Sacred Heart, Piacenza, Italy
[3] Department of Food and Drug, University of Parma, Parma, Italy
[4] Institute of Analytics and Bioanalytics, Faculty of Biotechnology, Mannheim University of Applied Sciences, Mannheim, Germany

**Highlights**

- Geographical origin assessment is a growing requirement for both food quality and safety
- GC-IMS is proposed as an effective technique for the geographical assessment of Italian hazelnuts
- VOCs analysed differ in samples, according to their provenience
- GC-IMS data are successfully merged with sensory analysis ones, providing better statistical models performance

Revised Manuscript Click here to view linked References

# *A Geographical Origin assessment of Italian Hazelnuts: Gas Chromatography-Ion mobility spectrometry coupled with Multivariate Statistical Analysis and Data Fusion approach*

*Giuseppe Sammarco*[1,3], *Daniele Bardin*[1], *Federica Quaini*[1], *Chiara Dall'Asta*[3], *Joscha Christmann*[4], *Philipp Weller*[4], *Michele Suman*[1,2]

[1] Sensory and Analytical Food Science, Barilla G. e R. Fratelli S.p.A., Parma, Italy
[2] Department for Sustainable Food Process, Catholic University Sacred Heart, Piacenza, Italy
[3] Department of Food and Drug, University of Parma, Parma, Italy
[4] Institute of Analytics and Bioanalytics, Faculty of Biotechnology, Mannheim University of Applied Sciences, Mannheim, Germany

**Abstract**

Hazelnut is a commodity that has gained interest in the food science community concerning its authenticity. The quality of the Italian hazelnuts is guaranteed by Protected Designation of Origin and Protected Geographical Indication certificates. However, due to their modest availability and the high price, fraudulent producers/suppliers blend, or even substitute, Italian hazelnuts with others from different countries, having a lower price, and often a lower quality. To contrast or prevent these illegal activities, the present work investigated the application of the Gas Chromatography-Ion mobility spectrometry (GC-IMS) technique on the hazelnut chain (fresh, roasted, and paste of hazelnuts). The raw data obtained were handled and elaborated using two different ways, software for statistical analysis, and a programming language. In both cases, Principal Component Analysis and Partial Least Squares-Discriminant Analysis models were exploited, to study how the Volatile Organic Profiles of Italian, Turkish, Georgian, and Azerbaijani products differ. A prediction set was extrapolated from the training set, for a preliminary models' evaluation, then an external validation set, containing blended samples, was analysed. Both approaches highlighted an interesting class separation and good model parameters (accuracy, precision, sensitivity, specificity, F1-score). Moreover, a data fusion approach with a complementary methodology, sensory analysis, was achieved, to estimate the performance enhancement of the statistical models, considering more discriminant variables and integrating at the same time further information correlated to quality aspects. GC-IMS could be a key player as a rapid, direct, cost-effective strategy to face authenticity issues regarding the hazelnut chain.

**Keywords**

Hazelnut (*Corylus avellana*), Gas Chromatography-Ion Mobility System, Food Authenticity, Multivariate Statistical Analysis, Sensory Analysis, Data Fusion.

## ***1. INTRODUCTION***

Authentic food can be defined as "a product where there is a match between its actual characteristics and the corresponding food product claims" (Food Integrity Handbook, 2018). According to the Food and Agriculture Organization of the United Nations (FAO) and the World Health Organization (WHO), "food authenticity is the quality of a food to be genuine and undisputed in its nature, origin, identity, and claims, and to meet expected properties" (Food and Agriculture Organization of United Nations & World Health Organization, 2018). Italy, as well as other Southern Europe countries (Greece, Spain, Portugal…), is increasingly interested in food authenticity research since it produces a relevant number of foods with Protected Designation of Origin (PDO), Protected Geographical Indication (PGI), and Traditional Speciality Guaranteed (TSG) certifications. In spite of that, nations such as China and the United States are significantly increasing their commitment to the food authenticity field in the last decade (Danezis, Tsagkaris, Camin, Brusic, & Georgiou, 2016). Hazelnuts are nuts belonging to the species~~class~~ *Corylus avellana* and *C. maxima*, the main global country producer is Turkey (ca. 70 % of the global production), followed by Italy, which is one of the principal importers (Maestri, Imperiale, & Marmiroli, 2020). Hazelnuts are commercialized with or without shells and consumed fresh and roasted, or they can be employed as an ingredient since hazelnut paste is a popular commodity in confectionery. The nation of origin and the indication of the species could be one of the most important characteristics for tree nuts concerning food fraud, as this information is typically required on the label, and frequently these features are the basement for the development of protected denominations. In particular, Italian hazelnuts represent a high-quality food, certified by the abovementioned PDO and PGI brands. This aspect could have determined a higher~~bigger~~ price than the other producing/importing countries. The vast availability of hazelnuts from Turkey, and the lower selling price of other big producers, such as Azerbaijan and Georgia, favoured illegal activities of false origin declaration by fraudulent suppliers, "contaminating" Italian batches with cheaper and/or more abundant products or even entirely substituting them (Lang, et al., 2021). According to the Food Integrity Handbook (Morin & Lees, 2018), several analytical approaches were exploited to solve hazelnut-related food authenticity issues. Ruiz del Castillo *et al.* evaluated the potential of enantiomeric analysis of (*E*)-5-methyl-hept-2-en-4-one (filbertone) for authenticity control of hazelnut and hazelnut oil ~~. A pre-separation of the hazelnut extract with high-pressure liquid chromatography (HPLC) was followed by the enantioselective gas chromatography (GC) analysis of the resulting fraction~~ (Ruiz del Castillo, Gomez Caballero, Blanch, & Herraiz, 2002). Filbertone was also analysed for the authenticity assessment of hazelnut-based products. The compound was determined through headspace solid phase micro extraction-GC-mass spectrometry (HS-SPME-GC-MS), and it was considered a proper marker to estimate the quality of commercial hazelnut spreads ~~. According to these analyses, the spreads were classified into three groups, samples with minimal hazelnuts content (less than 1 % - filbertone concentration lower than 4 μg/kg), with middle hazelnuts content (1-10 % - filbertone 4-45 μg/kg), and with high hazelnuts content (> 10 %, filbertone > 45 μg/kg)~~ (Cizkova, Rajchl, Snebergrova, & Voldrich, 2013). Fourier-transform infrared (FTIR) spectroscopy was also employed for the discrimination of hazelnuts from different origins/cultivars, according to their IR signal intensities. The multivariate statistical approach comprehended the unsupervised principal component analysis (PCA), followed by linear discriminant analysis (LDA) and partial least square-discriminant analysis (PLS-DA), with or without variable selection, and it permitted valuable discrimination among the groups (Manfredi, et al., 2018). P~~p~~olymerase chain reaction (PCR), and immunological methods, like enzyme-linked immunosorbent assay (ELISA) are often performed for hazelnut species/cultivars/origins characterization (Rohman, et al., 2021). ~~However, in the evaluation of the geographical origin, DNA-based markers are less effective than other chemical compounds (Esteki, et al., 2017).~~ Regarding other spectroscopic techniques exploited for hazelnuts' authenticity issues, near infrared (NIR) and nuclear magnetic resonance (NMR) were the most relevant in the scientific literature. NIR analyses were performed by Biancolillo *et al.*, to authenticate the Italian PDO hazelnut "Nocciola Romana". Two different classification approaches were taken into account, PLS-DA and soft independent modelling of class analogies (SIMCA), and both of them showed a high predictive ability, being applied in external validation on a test set (Biancolillo, et al., 2018). The same technology was tested also by Moscetti *et al.*, for the assessment of the same PDO hazelnut. In this study, ~~an algorithm for the selection of the best pre-treatment was carried out, and, besides SIMCA and PLS-DA, also *k*-Nearest Neighbour (KNN) and Support Vector Machine (SVM) were considered as discriminant routines.~~ support vector machine (SVM) has been revealed to be the optimal approach for the classification, with the best ~~classification~~ performance rate (Moscetti, Radicetti, Monarca, Cecchini, & Massantini, 2014). $^{1}$H-NMR was used for geographical origin determination by Bachmann *et al.*, considering 262 authentic samples from five different countries~~, over four harvesting years Both non-polar and polar extraction protocols were achieved, but the polar fraction was the most suitable for the work. It was used for the data analysis, monitoring of the sample preparation, and measurement through PCA~~. LDA was then found to be the best machine learning algorithm for the model classification, providing a 91 % cross-validation accuracy on the training set, and 96 % on the test set (Bachmann, Klockmann, Haerdter, Fischer, & Hackl, 2018). Despite an interesting application of spectroscopic strategies for the authentication of hazelnuts, the researchers' focus was mainly on spectrometric techniques, especially for the geographical origin assessment. Rosso *et al.* focused on the volatile metabolome of high-quality hazelnuts, "Ordu" from Turkey and "Tonda Romana" from Italy, to evaluate their evolution along the production chain, from the harvest to the storage and the roasting phase, employing HS-SPME-GCxGC-MS. Selected pattern recognition was used to mine the outputs, whereas PCA, Fisher ratio, HCA, and analysis of variance (ANOVA) permitted to find 'decisional markers' among the most important features (Rosso, et al., 2018). Comprehensive bi-dimensional GC hyphenated with MS technology was applied for an advanced fingerprinting approach aimed at the comparative analysis of the volatile fraction of roasted hazelnuts from different origins. A HS-SPME-GCxGC-qMS solution was exploited, ~~gaining patterns processed with a 'chromatographic fingerprinting' and a 'comprehensive template matching',~~ and the resulting markers had a distribution that can be linked to the sensory properties, the geographical origin, and the effect of thermal treatment on the compound classes (Cordero, et al., 2010). Mono-dimensional GC, coupled with MS was approached by Han *et al.* for an automatic metabolic profiling analysis with chemometrics (AuMPAC), for the geographical origin estimation of hazelnuts from six Chinese regions. ~~. AuMPAC consented to monitor the Total Ion Chromatographic (TIC) peaks showing evident differences among the abovementioned regions. Afterward, a chemometric peak resolution method was applied for the screened peaks, and the recovered features were then analysed by ANOVA.~~ The peaks with significant differences were employed for an origin discrimination model, based on two-way encoding PLS, obtaining an accuracy of up to 98 % (Han, et al., 2018). Hazelnut cultivars from different areas were ~~extensively~~ studied as well, to differentiate according to their geographical origins. Agronomical and pomological characterisations were initially achieved, then a polyphenolic extract was analysed through HPLC/UV analysis, and the phenolic fraction from the HPLC separation was also analysed by electrospray ionization-multistage ion trap MS (ESI-ITMS$^n$). The protein fraction was also extracted with a different protocol, and it was analysed via matrix assisted laser desorption ionization-time of flight-MS (MALDI-ToF-MS) (Ciarmiello, et al., 2014). Klockmann, in two diverse papers, perpetrated both food fingerprinting, by discerning hazelnuts from different locations with an untargeted metabolomics approach, using ultra high performance LC-quadrupole ToF (UPLC-QToF) and food targeting, determining the hazelnuts geographical origin through a targeted metabolomics approach, exploiting LC-triple quadrupole-MS/MS (LC-QqQ-MS/MS) for the quantitation of the proper 'origin markers', identified by the previous strategy (Klockmann, Reiner, Bachmann, Hackl, & Fischer, 2016) (Klockmann, Reiner, Cain, & Fischer, 2017). An innovative technique that could be potentially helpful in origin investigation is gas chromatography-Ion mobility spectrometry (GC-IMS). This analytical solution interfaces the comprehensive separation by gas chromatography and ion mobility systems allowing fingerprinting of the volatile fraction from solid and liquid samples, with a limited or even inexistent sample preparation. In addition, this technology enhances the analysis dimensionality, by combining the analytical power of the high-res chromatographic separation with the analytical selectivity of IMS, which has a Limit of Detection (LOD) range from 0.2 µg/m$^3$ to 2 mg/m$^3$ (Ruszkiewicz, et al., 2022). IMS is a technology for the detection of separated gaseous compounds in a mixture of analytes. It includes an ionisation source and a drift tube, which is then constituted of an ionisation zone and a migration zone (Yin, et al., 2021). The beta-emitting tritium ($^3H$) source is very solid and does not require an additional power supply. The Reactant Ion Peak (RIP), in the case of the radioactive source employed, represents $H_3O^+(H_2O)^n$ ions (hydronium ions), that are fundamental for the ionisation of the Volatile Organic Compounds (VOCs), through charge-transfer reaction (Jurado-Campos, Martin-Gomez, Saavedra, & Arce, 2021). Subsequently, the ions pass across the drift tube, a fixed distance tube, under an electric field applied. They are divided passing into the drift tube, at a time that depends on their shape and charge distribution. Therefore, the IMS tool permits the separation of isomeric compounds. When hyphenated with a GC column, everything that is eluted is ionised and subsequently separated in the IMS cell, the electrical signal generated is recorded at the end of the tube through a Faraday plate, and the ions discrimination is favoured by a counter-flowing inert gas in the tube. A 3D graph is obtained as result, each compound is characterised by a retention time, a drift time, and an intensity value (Eiceman, Karpas, & Hill, 2016) (Garrido-Delgado, Dobao-Prieto, Arce L., & Valcarcel, 2015). This technology was also performed for food authenticity issues, particularly on the geographical origin assessment. Honey samples with different botanical origins were studied, merging HS-GC-IMS with multivariate statistical models, PCA, LDA, and kNN. It was demonstrated, by comparing the PCA-LDA models, the complementarity of the technique with the NMR-based profiling of honey samples (Gerhardt, Birkenmeier, Schwolow, Rohn, & Weller, 2018). Another vulnerable commodity whose geographical origin was assessed by GC-IMS is olive oil. Gerhardt *et al.* compared the technology to the conventional isothermal capillary column (CC)-IMS system, in order to differentiate between extra virgin olive oil (EVOO) from Italy and Spain. GC-IMS highlighted a valuable resolving power for the untargeted profiling of VOCs in a complex matrix such as EVOO (Gerhardt, Birkenmeier, Sanders, Rohn, & Weller, 2017). Afterwards, GC-IMS data were also merged with fourier-transform mid-infrared (FT-MIR) data, for the authentication of olive oil and honey samples. Datasets were combined with a low-level data fusion approach, and a multivariate classification was performed, by PCA-LDA or PLS-DA. Data fusion is an effective tool for better classification performance.

-The present study focuses on the geographical assessment of the Italian hazelnut chain, from the fresh matrix to the roasted and the pasted one. This chain does not present robust scientific literature concerning authenticity frauds, and GC-IMS could represent an innovative, functional, and cost-effective strategy to deal with them. Furthermore, to evaluate how and whether the data fusion approach could improve the statistical analysis, the GC-IMS data matrix was merged with a matrix derived from sensory analysis, a complementary methodology that can provide more discriminant variables to the model for the geographical origin assessment.

## *2. MATERIALS AND METHODS*

### *2.1 Sampling*

Fresh, roasted, and paste hazelnut batches were sampled from different Italian and non-Italian regions, considering both the 2020 and the 2021 harvesting campaigns, and different storage shelf life, short and long (roughly 6 months after harvest) for 2020. The Italian samples include PGI "Tonda Gentile delle Langhe" from Piedmont, PDO "Nocciola Romana" from Lazio, and "Mortarella" from Campania. For each matrix, these three varieties were equally blended to have 'Italian samples' (N=36, 9 raw, 9 roasted, 9 peeled roasted, and 9 paste of hazelnuts) The same number of lots, considering the same sampling factors, were from Turkey, Azerbaijan, and Georgia, for a total of 108 non-Italian samples. All the samples were stored in a cold room, with a controlled temperature of 4-6 °C. For the validation set, these samples were used to create mixed ones; for each product, the Italian hazelnuts were mixed with one of the non-Italian, at different percentages of adulteration (10, 20, 50, 70, 90 %). Fresh samples were constituted of Italian and Georgian hazelnuts, the roasted ones had Italian and Azerbaijani matrices inside, whereas the pastes were made by blending Italian and Turkish samples. These mixes were employed as a test set and analysed as real samples. Tables S1 and S2 (Supplementary materials) show the training and the test sets, respectively.

### *2.2 Sample preparation*

Ca. 10 g of both fresh and roasted hazelnuts were initially minced with the knife mill Grindomix GM 200 (Retsch, Haan-Gruiten, Germany). 0.5 g were weighed into a 20 mL headspace vial, incubated then 60 °C for 5 mins, under agitation (500 rpm). Hazelnut pastes were directly weighed, incubated, and analysed as they were, without any mincing step. To evaluate the method repeatability, each sample was double-prepared and injected.

### *2.3 Instrumental parameters*

The GC-IMS instrument (FlavourSpec®, G.A.S. Dortmund, Dortmund, Germany) was equipped with a syringe and the autosampler PAL3-RSI Series II (CTC Analytics AG, Zwingen, Switzerland) for the headspace injection mode. The injection volume was set to 0.5 mL, and both the syringe and the injector port temperatures were at 80 °C. The chromatographic separation step was carried out with an FS-SE-54-CB-0.5 GC column (30 m length, internal diameter 0.32 mm, film thickness 0.5 μm), at the constant temperature of 40 °C. Nitrogen was the carrier gas. The separation was done without a thermal ramp, only a flow ramp was exploited: the program started at 2 mL/min for 5 mins, then the flow was brought to 31 mL/min in 4 mins, then to 100 mL/min in 20 sec, keeping it at this value for the last 2 mins, for a total GC runtime of 11 mins. The elute was subjected to the drift tube, for the ion mobility separation step. Both drift tube flow and temperature were kept constant, 150 mL/min and 45 °C, respectively. The carrier gas was nitrogen, the tube length was 9.8 cm, and the drift voltage and time were 5 kV and 30 ms, operated in positive ionisation mode.

### *2.4 Data elaboration*

Fresh, roasted, and paste hazelnut samples were separately considered, so data analyses were individually performed on each matrix sample set. This was due to the strong impact that the product processing could have on the volatile profiles of the different hazelnut-based commodities. The output of a GC-IMS analysis was a 3D graph, with the y-axis as GC retention time, the x-axis as the IM drift time, and the z-axis showing the detector response or signal intensity. To favour the graphical visualisation, the heat map 2D fingerprint representation was ideal, the 3D-2D conversion was achieved by transforming the z-axis into a colour signature for each spot. Thus, the more intense the signal, the more coloured the spot~~.~~ (Vera, Companioni, Meacham, & Gygax, 2016). An example of a fresh hazelnut sample GC-IMS 2D heat map is shown in Figure 1.

GC-IMS raw data were elaborated in two different ways: by manually selecting the most intense spots, or by working on the whole spectrum. The former is achieved by creating a manual area set, which was done through the VOCal software (version 0.1.3 – G.A.S. Dortmund, Dortmund, Germany), selecting all the visible spots on the heat map of the entire samples set, considering both training and test sets. The list of spots/areas can be better visualised using a VOCal software module, called "Galerie". The module permits observation of all the areas picked for the project, showing them in all the samples analysed. From this plot, the signal intensities were exported in an Excel spreadsheet, and this obtained matrix was processed by SIMCA software (Version 16.0.1, Umetrics, Umea, Sweden). In this case, also to assess the method repeatability, both replicates were included in the processing and modelling steps. For the statistical analysis and the model setup, the spectra were normalised according to the reactant ion peak (RIP) position, the red line in Figure 1 at 1.0 ms drift time. The data elaboration of the full spectrum was achieved using the programming language Python (Python, version 3.8.12) via Visual Studio Code (Microsoft & Electron Framework, version 1.74.1), through the *gc-ims-tools* package (version 0.1.2) for GC-IMS data handling and analysis (Christmann, Rohn, & Weller, 2022). In this other case, the package allowed us to create a mean value of the two replicates, returning score plots with a minor data dispersion. Starting from the initial spectra, all of them were normalised according to RIP, set at 1.0 ms drift time. Subsequently, it was cut, in order to better visualise the graph, and to define the area where all the spots were, that will be considered for the elaboration. Data were auto-scaled, so the mean centering was combined with weighting by the inverse standard deviation for each variable, prior to being processed with multivariate statistical models (Christmann, Rohn, & Weller, 2022). These two different approaches were evaluated to compare the data elaboration performed only on the spots of interest, using dedicated software, with defined, or eventually limited, statistical models to be applied, and the elaboration was done via coding with Python, which consents to work on the whole graph, including the risk of picking background signals, or others due to instrumental variation. However, this last approach is more flexible, as every statistical model can be applied with the proper code.

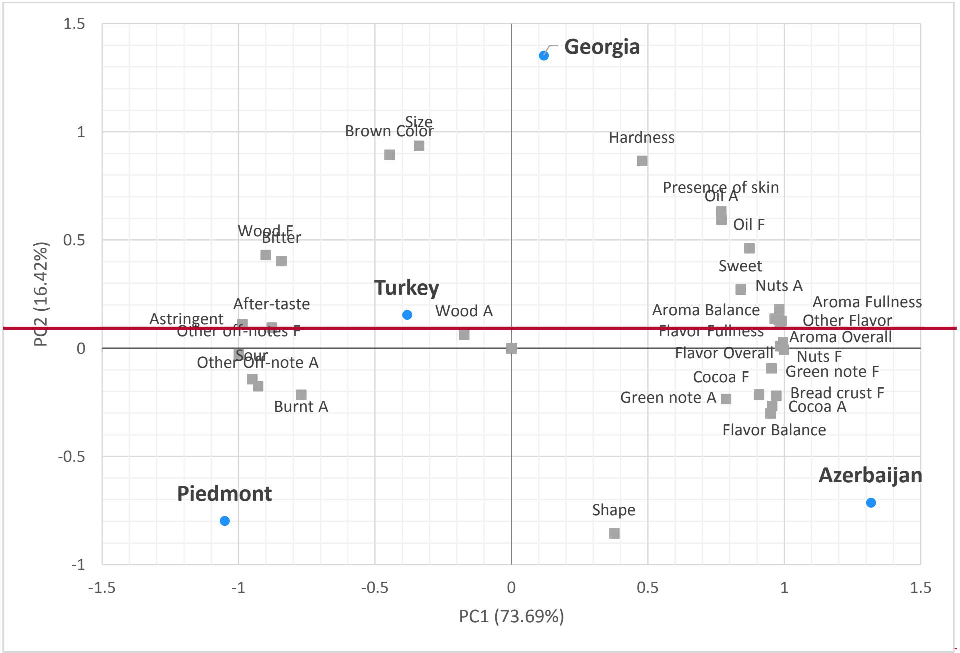
### *2.5 Sensory analysis*

Sensory evaluation is defined as the discipline used to evoke, measure, analyze and interpret reactions to those characteristics of foods and materials as they are perceived by the sense of sight, smell taste, touch and hearing (Stone & Sidel, 1985).

The methodology used for this study is the profile attribute analysis (PAA) that is a sensory analysis methodology carried by a panel composed by expert tasters to objectively describe and measure the sensory characteristics of a product. Data obtained with the PAA can be entered efficiently in automated processing systems and are perfectly suitable for statistical analysis and aggregations.

Tasting booths following ISO 8589:2007 (*General guidance for the design of test rooms Sensory evaluations*) were used to perform the sensory evaluation. The panel was composed of eight (8) experienced assessors. The sample set tasted by the assessors was composed by one batch for each sample, for a total of 48 samples (12 fresh hazelnut lots, 24 roasted ones, unpeeled and peeled, 12 paste ones). A first tasting session was run to describe the sensory characteristics of all the samples to select the attributes to be used for the quantitative evaluation. In the subsequent session, the intensity of the sensory attributes was measured for each hazelnut sample, using a 12 points scale.

The results were then analysed by running the ANOVA test and the PCA. The main aim of ANOVA is to identify and quantify which factors are responsible for the variability of the response, while PCA maps out how products perform in terms of sensory characteristics and provide a visual snapshot of the sensory space that the samples sit in. Both ANOVA and PCA allow to analyse sensory data for detecting and quantifying differences between products. Figure 2 displays an example of PCA score plot of the fresh hazelnut sample set.




~~Fig. 2) Example of a PCA score plot of the fresh hazelnut sample set.~~

### *2.6 Data Fusion*

Data from GC-IMS and sensory analyses were merged by low-level data fusion with the original data. The GC-IMS data matrix obtained from the manual area set presented a suitable dimensionality for the fusion with the sensory analysis matrix. Thus, the original data were merged without pre-treatment with a SIMCA software functionality, and then the fused block was auto-scaled before the processing. On the other hand, the entire spectrum handled with Python code, returned a huge data matrix. This was due to the unfolding of three-way GC-IMS data, that generated a matrix with a lower number of rows (number of samples) and a larger number of columns (retention x drift times). After this unfolding, the matrix could be merged with the sensory one by concatenation (Schwolow, Rohn, Gerhardt, & Weller, 2019). However, the enormous difference in features number between the two datasets (ca. 3 million from GC-IMS vs 40 from sensory), was difficult to handle, even with a pre-treatment, which included "within-" and "between-block" scaling. The dominance of the GC-IMS dataset strongly impacted the model, and the data fusion did not significantly improve the models. Therefore, this specific approach was employed only using the matrix from the manual area set, concatenating it with the sensory analysis matrix by using SIMCA software, to evaluate how the merging could potentially work on the models.

# *3. RESULTS AND DISCUSSION*

## *3.1 SIMCA software data elaboration*

The data matrix obtained from the GC-IMS analysis by manually selecting the visible spots had between 70 and 100 variables, ~~a manageable dimension by SIMCA software,~~ and the multivariate statistical analysis was fundamental to deal with all of these variables at the same time, finding a correlation among them. A variables selection was carried out to lower the data dispersion and increase the explained variance of the models. This was achieved by filtering the most important variables considering the variable importance in projection (VIP) score. Only the features having a VIP score > 1.0 were kept for the statistical analysis. The workflow adopted started with PCA for a preliminary visualisation of the class clustering. Figures 3A)-B)-C) show the PCA score plots of fresh, roasted, and hazelnut paste, respectively.

The plots highlight a discrete separation between Italian and non-Italian samples.~~, t~~The overall explained variance was higher than 40~~between 20 and 30~~ %. Despite this value is acceptable,~~,~~ this pointed out a data dispersion, mainly due to the relevant number of qualitative variables / DoE factors (storage shelf-life, harvesting year, presence of peel) related to the number of samples. The main target of the PCA is the data dimensionality reduction, useful ~~in order to be able~~ to visualise a 2D plot and extract the features. It is an unsupervised approach since it does not label the classes, hence the observations are positioned into the plot only depending on the variables that drive them (Ghojogh & Crowley, 2022). Supervised analysis, such as PLS-DA, by labelling the groups, permits to reach of a better clustering, decreasing the data dispersion. Figures 4A)-B)-C) show the PLS-DA score plots of fresh, roasted, and paste hazelnut sample sets, underlining a better class separation, with a clear division between Italian and non-Italian products~~, even with a discrete grouping inside the non-Italian cluster, in some cases~~.

A misclassification table was created ~~as well. It is a tabular way~~ to estimate the performance of the prediction model. Each item in the misclassification table designates the number of predictions by the model, where the classification is done correctly or incorrectly (Mohajon, 2020). In the present study, 26 samples for fresh hazelnuts, 47 for roasted hazelnuts, and 34 for hazelnut pastes were picked from the training set and used as a prediction set. An accuracy score of 100 % for all the supervised models was obtained, confirming their statistical robustness (Tables 1A-B-C).

~~A)~~

| | Members | Correct | Georgia | Azerbaijan | Turkey | Italy | No class (YPred <= 0) |
|---|---|---|---|---|---|---|---|
| Georgia | 4 | 100% | 4 | 0 | 0 | 0 | 0 |
| Azerbaijan | 4 | 100% | 0 | 4 | 0 | 0 | 0 |
| Turkey | 6 | 100% | 0 | 0 | 6 | 0 | 0 |
| Italy | 12 | 100% | 0 | 0 | 0 | 12 | 0 |
| No class | 0 | - | 0 | 0 | 0 | 0 | 0 |
| Total | 26 | 100% | 4 | 4 | 6 | 12 | 0 |

~~B)~~

| | Members | Correct | Georgia | Azerbaijan | Turkey | Italy | No class (YPred <= 0) |
|---|---|---|---|---|---|---|---|
| Georgia | 9 | 100% | 9 | 0 | 0 | 0 | 0 |
| Azerbaijan | 11 | 100% | 0 | 11 | 0 | 0 | 0 |
| Turkey | 11 | 100% | 0 | 0 | 11 | 0 | 0 |
| Italy | 16 | 100% | 0 | 0 | 0 | 16 | 0 |
| No class | 0 | - | 0 | 0 | 0 | 0 | 0 |
| Total | 47 | 100% | 9 | 11 | 11 | 16 | 0 |

~~C)~~

| | Members | Correct | Georgia | Azerbaijan | Turkey | Italy | No class (YPred <= 0) |
|---|---|---|---|---|---|---|---|
| Georgia | 7 | 100% | 7 | 0 | 0 | 0 | 0 |
| Azerbaijan | 8 | 100% | 0 | 8 | 0 | 0 | 0 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| ~~**Turkey**~~ | ~~7~~ | ~~100%~~ | ~~0~~ | ~~0~~ | ~~7~~ | ~~0~~ | ~~0~~ |
| ~~**Italy**~~ | ~~12~~ | ~~100%~~ | ~~0~~ | ~~0~~ | ~~0~~ | ~~12~~ | ~~0~~ |
| ~~**No class**~~ | ~~0~~ | - | ~~0~~ | ~~0~~ | ~~0~~ | ~~0~~ | ~~0~~ |
| ~~Total~~ | ~~34~~ | ~~100%~~ | ~~7~~ | ~~8~~ | ~~7~~ | ~~12~~ | ~~0~~ |

~~Table 1A) *Misclassification table of the selected fresh hazelnut prediction set (n=26). 1B) Misclassification table of the selected roasted hazelnut prediction set (n=47). 1C) Misclassification table of the selected hazelnut paste prediction set (n=34). (Green cell: samples correctly classified, yellow cell: samples misclassified)*~~

To further assess the robustness of the predictive models, and also their feasibility in a real environment, the test set samples were~~as~~ analysed. The selected workflow was the same one used for the training set, as well as the manual area set and the matrix export. Thus, the test set was used as a validation~~prediction~~ set, considering only the Italian class, and the other one was used for the mixed samples. This allowed~~was suitable to simulate~~ an industrial approach simulation, where~~a sample is analysed, and~~ if the sample's~~its~~ volatile profile matches the Italian ones~~samples,~~ it is considered authentic, otherwise, it is discarded or, eventually, analysed with a confirmatory methodology.

A classification list was performed on the test set: it displayed the observation (sample IDs), the original dummy variables as YVarPS, which can range from 1 to 0, and the predicted dummy variables as YPredPS. From this last value, it is possible to define the sample class:

- <0.35 the samples do not belong to the class
- Between 0.35 and 0.65 the samples are borderline
- >0.65 the samples belong to the class

(MKS Umetrics, 2015). In this work, only samples with a YPredPS bigger than 0.65 were considered Italian. Tables S3A)-B)-C) report the classification list outcomes from the fresh, roasted, paste of hazelnut test sets.

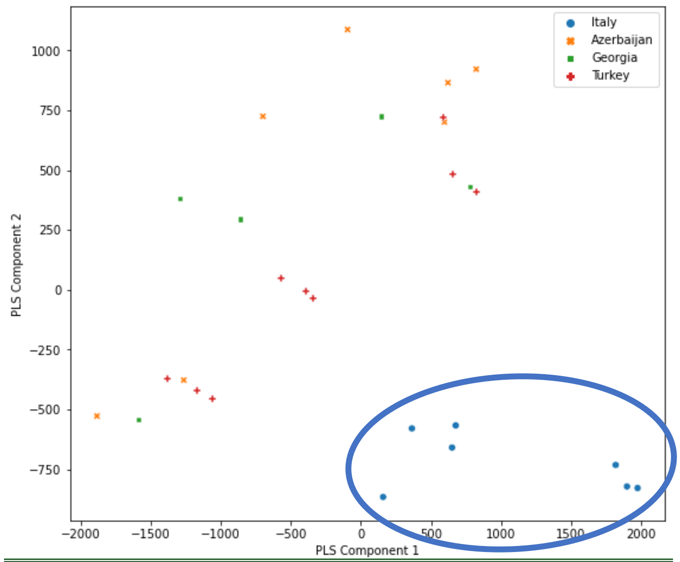
The Classification list related to the fresh hazelnut samples gave valuable outcomes, as ~~since~~ none of the mixes were classified as Italian. This confirmed the good sensitivity and the robustness of the technique on this matrix, as even the samples having only the 10 % of the non-Italian products were ~~not wrongly~~ well-classified. Concerning the roasted hazelnut samples, the performance was worse than the one of~~referred to~~ the fresh ones. Indeed, all the samples, both peeled and unpeeled roasted hazelnuts, with 90 % Italian products were classified as Italian; on the other hand~~also~~, many samples having 80 and 50 % Italian hazelnuts were misclassified. Thus, with the roasted matrix, the analytical tool is less robust, with~~and~~ only samples having more than 50 % of non-Italian hazelnuts ~~were~~ grouped as anomalous. Hazelnut paste samples were all classified as non-Italian, also with high values of YPredPS. This points out a good performance of the methodology to tackle the authenticity issues of Italian materials. However, samples with a big percentage of Italian products presented relevant YPredPS values for the Turkish class; this does not denote an ideal performance by the model. Likely, a higher number of samples should be needed to better understand the robustness of the methodology on the paste matrix.

### *3.2 Python data elaboration with gc-ims-tools package*

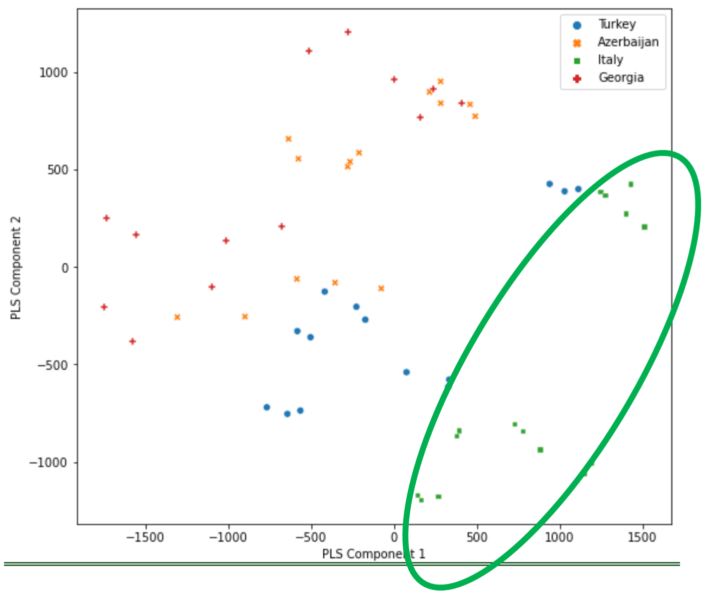
The workflow selected for the data elaboration with the “gc-ims-tools” package in Python coding was analogue to the one adopted with the SIMCA software. ~~, so PCA as an unsupervised model, then PLS-DA as supervised, and other tests, on the prediction set, taken from the training set, and on the test set, to assess the model’s robustness.~~ Figures 5A-B-C display the PCA score plots of fresh, roasted, and paste hazelnut samples, respectively.

As expected, a preliminary geographical origin-based separation was achieved by considering PCs different from the first one, since, on the whole spectrum, the harvesting year and the storage shelf-life factors were more relevant than in the model made from the manual area set. It was evident in the score plot of the roasted hazelnut sample set ~~PCA~~, where six Italian samples were out of the Italian cluster~~, out of the Italian cluster, there were six Italian samples, from the 2020 harvesting campaign, with short storage shelf-life~~. Nevertheless, this trend was not so evident also in the models related to the fresh and paste of hazelnuts, where the Italian clusters were quite well-defined~~, considering the unsupervised method~~. Moving to the supervised models, figures 6A-B-C show PLS-DA score plots of fresh, roasted, and paste hazelnut sample sets, respectively.

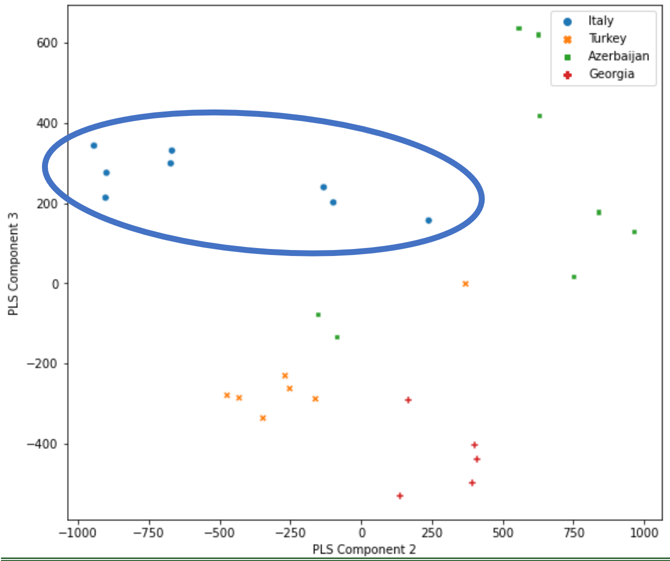
~~A)~~


Italy
Azerbaijan
Georgia
Turkey
PLS Component 1
PLS Component 2

~~B)~~


Turkey
Azerbaijan
Italy
Georgia
PLS Component 1
PLS Component 2

~~C)~~




~~Fig. 6A) *PLS-DA score plot of the fresh hazelnut sample set. (Blue dots: Italy, orange crosses: Azerbaijan, green squares: Georgia, red crosses: Turkey, blue ellipse: Italian cluster)* 6B) *PLS-DA score plot of the roasted hazelnut sample set. (Blue dots: Turkey, orange crosses: Azerbaijan, green squares: Italy, red crosses: Georgia, green ellipse: Italian cluster)* 6C) *PLS-DA score plot of the hazelnut paste sample set. (Blue dots: Italy, orange crosses: Turkey, green squares: Azerbaijan, red crosses: Georgia, blue ellipse: Italian cluster)*~~

Predictably, the clustering improved, and it was possible to define a precise Italian group. In the roasted hazelnut score plot~~,~~ it is still evident a clear division between samples having a different harvesting year and storage shelf-life. It is interesting how the Italian samples were well-separated from the other classes.~~;~~ T~~t~~his could be explained by taking into account the areas of cultivation in the non-Italian countries in the present study. Turkey, Azerbaijan, and Georgia are all around the "Black Sea Region", hence the climatic and soil conditions are similar. This could be reflected in products having analogue chemical characteristics, and so the VOCs. To initially estimate the robustness of the supervised model, the train-test split was done, extracting 20 % of samples from the training set to generate the prediction~~test~~ set. On this latter set, the PLS-DA algorithm was applied, in order to obtain the model accuracy score. Fresh, roasted, and paste of hazelnut models reported an accuracy score of 86 %, 100 %, and 88 %, respectively. These values underlined a good model performance.~~;~~ H~~h~~owever, to better assess it, and to simulate the application of the approach in a real situation, the validation set, that included both blended samples and others picked from the training set, was employed for a Random Forest Classification. ~~It is a so-called "ensemble learning" machine learning~~

~~method, used to create prediction models.~~ Random Forests are a collection of classification and regression trees, elementary models using binary split on the variable to define predictions. Decision trees provide a perceptive method for predicting output, that divides “high” vs. “low” values of a variable linked to it. Nonetheless, this methodology often returns low accuracy in the case of intricate datasets. Random Forest builds classification and regression trees exploiting random training sets and subsets variables for modelling outputs. Outcomes from each tree are cumulated, returning a prediction for each observation. This leads to higher accuracy than a single decision tree model (Breiman, Random Forests, 2001) (Breiman, Friedman, Stone, & Olshen, 1984) (Speiser, Durkalski, & Lee, Random forest classification of etiologies for an orphan disease, 2015) (Speiser, Miller, Tooze, & Ip, 2019). Random Forest Classifier was then applied on the validation set after the model was trained using the training set. Prediction variables were generated and compared with the ones from the validation set, and, from this comparison, the following parameters were evaluated:

- Accuracy: the ratio of the correctly classified samples to the whole pool of samples.
- Precision: the ratio of the correctly positive classified by the model to all positive samples (true and false positive).
- Sensitivity: the ratio of the correctly positive classified by the model to all that are positive (true positive and false negative).
- F1 score: the harmonic mean of the precision and sensitivity.
- Specificity: the correct negative classified by the model to all that are negative (true negative and false positive).

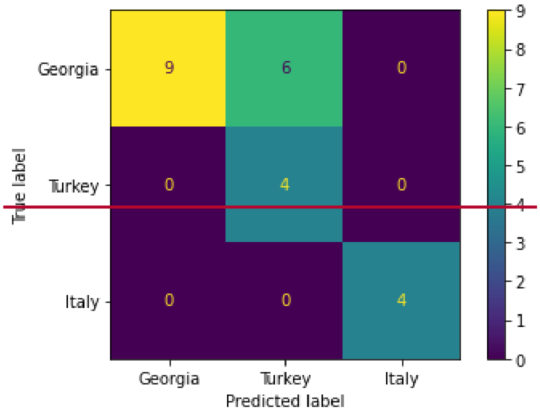
(Ghoneim, 2019). Table 2 reports the abovementioned parameters after the Random Forest Classification of fresh, roasted, and paste hazelnut validation sets.

| ~~MATRIX~~ | ~~ACCURACY~~ | ~~PRECISION~~ | ~~SENSITIVITY~~ | ~~F1-SCORE~~ | ~~SPECIFICITY~~ |
|---|---|---|---|---|---|
| ~~FRESH HAZELNUTS~~ | ~~74 %~~ | ~~90 %~~ | ~~74 %~~ | ~~76 %~~ | ~~74 %~~ |
| ~~ROASTED HAZELNUTS~~ | ~~100 %~~ | ~~100 %~~ | ~~100 %~~ | ~~100 %~~ | ~~100 %~~ |
| ~~HAZELNUT PASTES~~ | ~~92 %~~ | ~~92 %~~ | ~~92 %~~ | ~~90 %~~ | ~~92 %~~ |

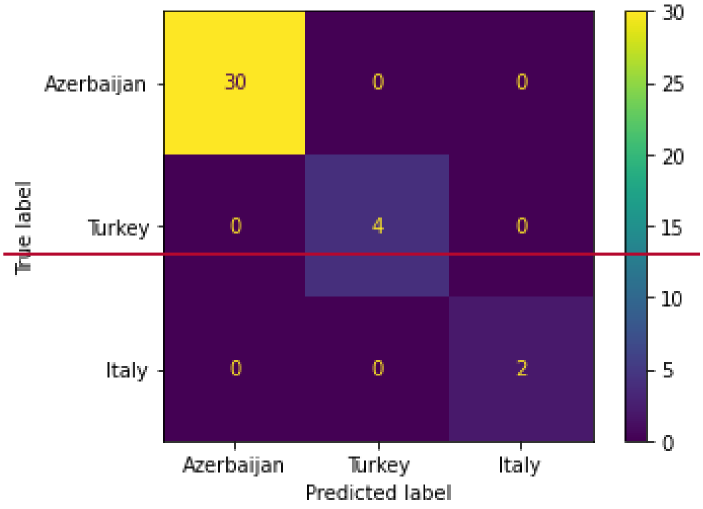
~~Table 2 *Accuracy, Precision, Sensitivity, F1-Score, and Specificity values of fresh, roasted, paste hazelnut validation sets.*~~

To conclude the assessment and to better visualise the model performance on the validation set, a confusion matrix was built. It is a table where each row represents the number of samples in the actual class, whereas each column is about the number of samples in the predicted class (Powers, 2011). Figures 3~~7~~A-B-C represent the confusion matrix related to, respectively, fresh, roasted, and paste hazelnut validation sets.

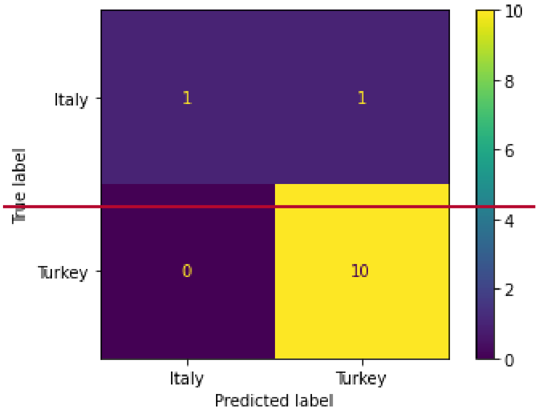
~~A)~~




~~B)~~




~~C)~~




~~Table 7A) *Confusion matrix of the fresh hazelnut validation set.* 7B) *Confusion matrix of the roasted hazelnut validation set.* 7C) *Confusion matrix of hazelnut paste validation set.*~~

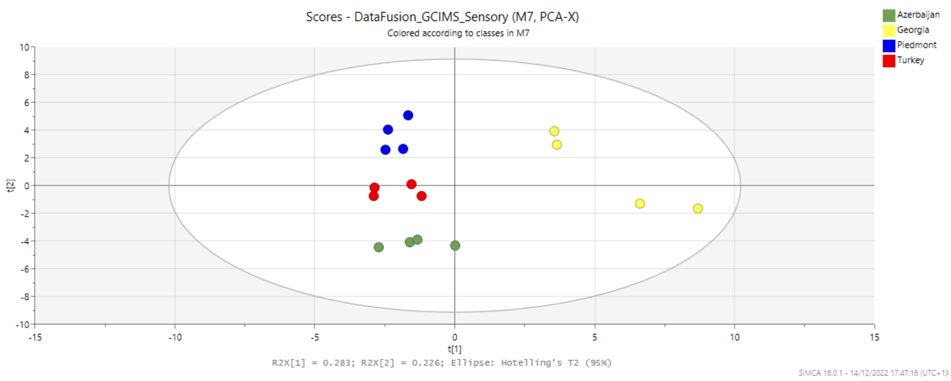
All the mixed samples (Italian – non-Italian at different percentages) were placed into the non-Italian class before the statistical analysis, and they were correctly classified as non-Italian. ~~This approach also explains the non-ideal parameter values for the fresh hazelnut model.~~ Regarding the fresh samples, ~~S~~some of the mixed samples, ~~initially~~ labelled as Georgian, were classified as Turkish~~, the class related to the other samples picked from the training set,~~ since the mix between Italian and Georgian products was evaluated, in some cases, ~~by the model~~ as Turkish class by the model. This led to lower accuracy, precision, sensitivity, specificity, and F1-score. However, the model results were effective, because all the not 100 % authentic Italian samples were classified as non-Italian, making the detection of anomalous lots~~it~~ possible~~, also on external sets, to detect anomalous lots~~, even with a percentage of Italian products.

### *3.3 Data fusion*

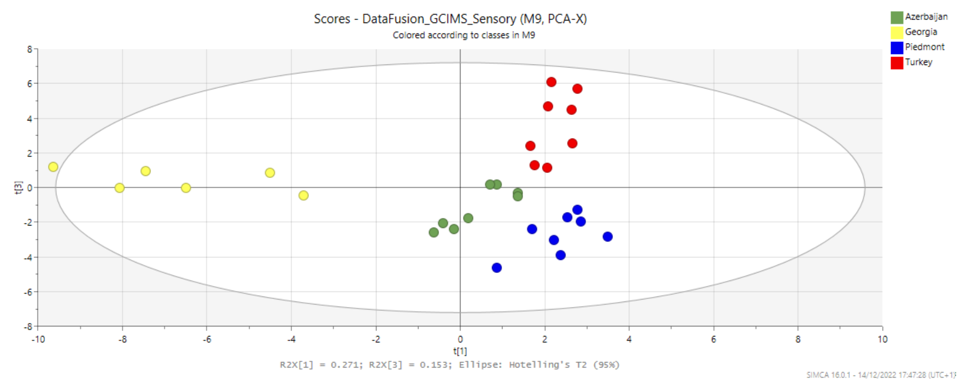
The sensory analysis encompasses several sciences to better comprehend the sensory properties of commodities and the consumer's response to these properties. (Chambers & McGuire, 2003) Some parameters, according to the matrix analysed, are initially set. To assign values to these parameters, category scales are employed, for objective assessment and subjective response. (McEwan & Lyon, 2003) Tables S4A-B-C display all the sensory parameters of the hazelnuts analysis, and their relative values, obtained by doing the mean of the assessors' evaluation of each sample.

These features were concatenated with the ones from the GC-IMS analysis, ~~and extracted with the manual area set,~~ employing the SIMCA software functionality. Since the number of features from the two data sets was similar, no pre-treatment was achieved prior to merging them. Low-level data fusion was then performed, and, afterward, the entire data matrix obtained was auto-scaled~~,~~ to make all the variables comparable. One batch, in duplicate, per type of sample (considering origin, harvesting year, and storage shelf-life) was used for the sensory analysis. Thus,~~, hence~~ only the matching data from the GC-IMS analysis were merged, to better investigate the efficacy of the approach, avoiding the dominance of one technique on the other. C~~Thus, c~~onsidering the limited number of samples employed, this could be defined as a 'proof-of-concept' of the analytical-sensory data fusion approach. The same protocol for the statistical analysis of the GC-IMS outputs was carried out, starting from the unsupervised PCA to the supervised PLS-DA. Figures 7A-B-C show the PCA score plots of the fresh, roasted, and paste hazelnut samples after the data fusion.

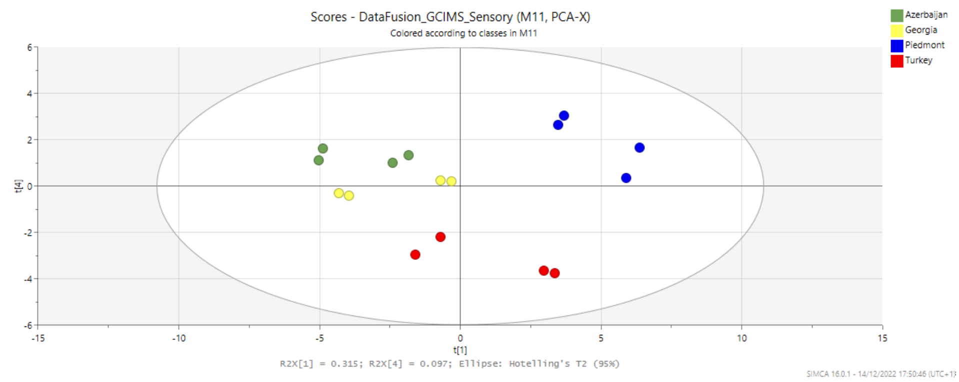
~~A)~~




~~B)~~



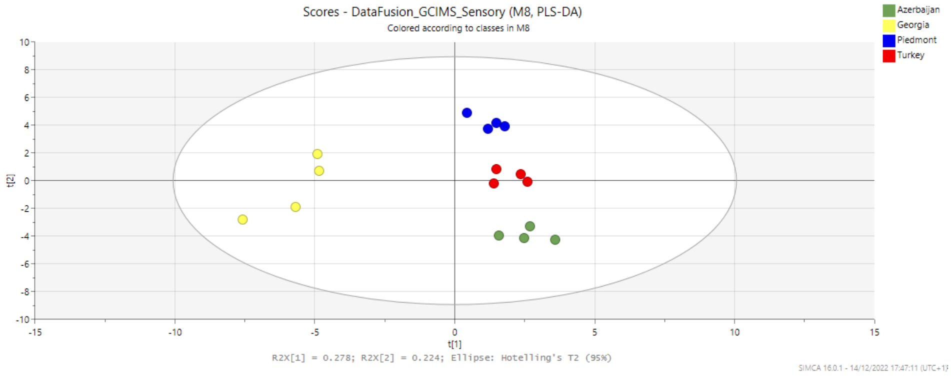

~~C)~~



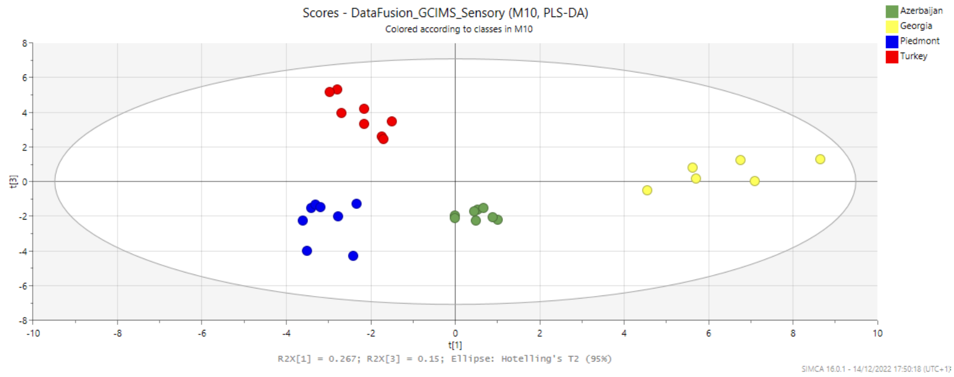

~~Figure 7A) *PCA score plot of fresh hazelnut samples after data fusion. 7B) PCA score plot of roasted hazelnut samples after data fusion. 7C) PCA score plot of hazelnut paste samples after data fusion. (Green dots: Azerbaijan, yellow dots: Georgia, blue dots: Piedmont, red dots: Turkey)*~~

The Italian mix could not be reproduced for the sensory analysis, hence only samples from one Italian region, Piedmont, were considered (PDO "Tonda Gentile Trilobata"). As it is possible to see, the group clustering seems to be improved, the classes were clearly divided. In addition, the first PC already can efficiently separate the samples according to their geographical origin, so the explained variance was increased, with a lower data dispersion and harvesting year effect. It is curious to notice how the fusion of the data sets brought a distinct division of the Georgian cluster from the others in the fresh and roasted score plots, and this grouping is then inverted in the hazelnut paste score plot, where Piedmont is the class finely separated, similarly to the GC-IMS outcome. Figures 8A-B-C indicate the PLS-DA score plots of the fresh, roasted, and paste hazelnut samples after the data fusion.

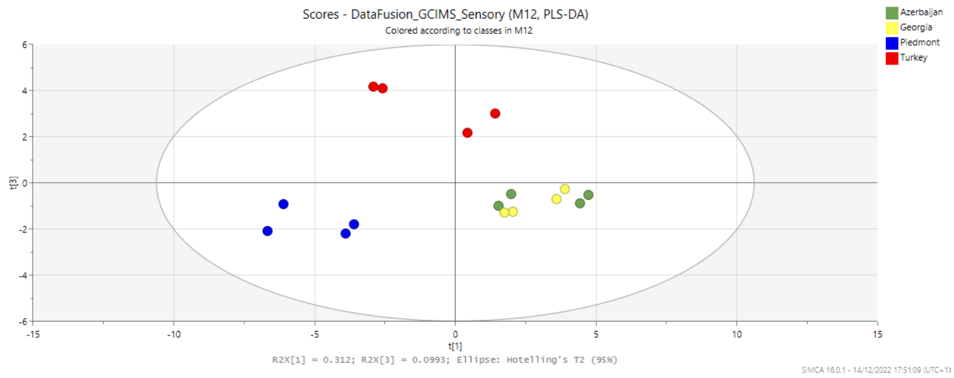
~~A)~~


Scores - DataFusion_GCIMS_Sensory (M8, PLS-DA)
Colored according to classes in M8
Azerbaijan
Georgia
Piedmont
Turkey
t[2]
t[1]
R2X[1] = 0.278; R2X[2] = 0.224; Ellipse: Hotelling's T2 (95%)
SIMCA 16.0.1 - 14/12/2022 17:47:11 (UTC+1)

~~B)~~


Scores - DataFusion_GCIMS_Sensory (M10, PLS-DA)
Colored according to classes in M10
Azerbaijan
Georgia
Piedmont
Turkey
t[3]
t[1]
R2X[1] = 0.267; R2X[3] = 0.15; Ellipse: Hotelling's T2 (95%)
SIMCA 16.0.1 - 14/12/2022 17:50:18 (UTC+1)

~~C)~~


Scores - DataFusion_GCIMS_Sensory (M12, PLS-DA)
Colored according to classes in M12
Azerbaijan
Georgia
Piedmont
Turkey
t[3]
t[1]
R2X[1] = 0.312; R2X[3] = 0.0993; Ellipse: Hotelling's T2 (95%)
SIMCA 16.0.1 - 14/12/2022 17:51:09 (UTC+1)

~~Figure 8A) *PLS-DA score plot of fresh hazelnut samples after data fusion.* 8B) *PLS-DA score plot of roasted hazelnut samples after data fusion.* 8C) *PLS-DA score plot of hazelnut paste samples after data fusion. (Green dots: Azerbaijan, yellow dots: Georgia, blue dots: Piedmont, red dots: Turkey)*~~

As expected, the supervised models improved the preliminary grouping, highlighting the inter-class differences. Therefore, this ‘proof-of-concept’ study reported interesting outcomes about the data fusion strategy, that revealed to be a tool for statistical model improvement when complementary/orthogonal techniques were considered. However, considering the limited number of samples per class, and the also the moderate number of variables included, further studies are necessary to robustly define the effectiveness of the approach.

## ***4. CONCLUSIONS***

Hazelnuts and their processed products are very common in bakery companies, for their versatility in different preparations. The growing demand for high-quality ingredients has led producers/suppliers to search for the best varieties, that, in the case of the hazelnut commodity, are strictly related to the geographical area of cultivation. Therefore, the authentication of this matrix is a relevant aspect from both a quality and safety point of view. The GC-IMS technology could be a functional candidate for facing authenticity issues linked to the hazelnut chain. The instrumental parameters can be set and optimised for gaining the best sample fingerprinting, as well as the sample preparation for the ideal extraction of the VOCs. This opens up the possibility for research on this technology, which can be performed at both academic and industrial levels. Simultaneously, the speed, sensitivity, and cost-effectiveness of the tool are also exploitable for routinary analysis in the companies’ quality control department. In comparison with the GC-quadrupole MS (GC-qMS), a technique usually available in the quality control laboratory for the VOCs analysis, GC-IMS present diverse advantages. In terms of costs, the routine maintenance can be conducted by the end-user, the technology requires only high-purity nitrogen as carrier and drift gas without any expensive vacuum pump to be maintained. Moreover, it is as sensitive as the standard GC-MS system, and the two-dimensional separation permits faster throughput of complex samples. Finally, no particular pre-treatment is required before the analysis, that makes this approach more user-friendly and green. T~~In addition, t~~he coupling with multivariate statistical analysis permits to deal~~ing~~ with the big number of features of the 3D data matrix~~ obtained~~, extrapolating useful chemical info. The statistical models created~~ employing a relevant number of samples~~ can be applied to~~on~~ real lots in an automatiz~~s~~ed way. Furthermore, this study~~it~~ also demonstrated preliminary efficacy in merging data from this technique with a complementary one, such as sensory analysis, concerning the statistical modelling.

## ***5. REFERENCES***


Bachmann, R., Klockmann, S., Haerdter, J., Fischer, M., & Hackl, T. (2018). 1H NMR Spectroscopy for Determination of the Geographical Origin of Hazelnuts. *Journal of Agricultural and Food Chemistry*, 11873-11879.

Biancolillo, A., De Luca, S., Bassi, S., Roudier, L., Bucci, R., Magrì, A. D., & Marini, F. (2018). Authentication of an Italian PDO hazelnut ("Nocciola Romana") by NIR spectroscopy. *Environmental Science and Pollution Research*, 28780–28786.

Breiman, L. (2001). *Random Forests.* Berkeley, CA.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and Regression Trees.* New York: Taylor & Francis.

Chambers, E., & McGuire, B. (2003). SENSORY EVALUATION | Practical Considerations. In B. Caballero, *Encyclopedia of Food Sciences and Nutrition (Second Edition)* (p. 5136-5141). Elsevier Science Ltd.

Christmann, J., Rohn, S., & Weller, P. (2022). gc-ims-tools – A new Python package for chemometric analysis of GC-IMS data. *Food Chemistry*, 133476.

Ciarmiello, L. F., Mazzeo, M. F., Minasi, P., Peluso, A., De Luca, A., Piccirillo, P., . . . Carbone, V. (2014). Analysis of Different European Hazelnut (Corylus avellana L.) Cultivars: Authentication, Phenotypic Features, and Phenolic Profiles. *Journal of Agricultural and Food Chemistry*, 6236-6246.

Cizkova, H., Rajchl, A., Snebergrova, J., & Voldrich, M. (2013). Filbertone as a marker for the assessment of hazelnut spread quality. *Czech Journal of Food Science*, 81-87.

Cordero, C., Liberto, E., Bicchi, C., Rubiolo, P., Schieberle, P., Reichenbach, S. E., & Tao, Q. (2010). Profiling food volatiles by comprehensive two-dimensional gas chromatography coupled with mass spectrometry: Advanced fingerprinting approaches for comparative analysis of the volatile fraction of roasted hazelnuts from different origins. *Journal of Chromatography A*, 5848-5858.

Danezis, G. P., Tsagkaris, A. S., Camin, F., Brusic, V., & Georgiou, C. A. (2016). Food authentication: Techniques, trends & emerging approaches. *Trends in Analytical Chemistry, 85*, 123-132. doi:https://doi.org/10.1016/j.trac.2016.02.026

Eiceman, G., Karpas, Z., & Hill, H. J. (2016). *Ion Mobility Spectrometry, Third Edition.* Boca Raton: Taylor & Francis Group.

Esteki, B., F., S., A., R., B., Z., A., E., D., . . . Y., V. H. (2017). Classification and authentication of Iranian walnuts according to their geographical origin based on gas chromatographic fatty acid fingerprint analysis using pattern recognition methods. *Chemom. Intell. Lab. Syst.*, 251-258. doi:10.1016/j.chemolab.2017.10.014

Food and Agriculture Organization of United Nations & World Health Organization. (2018). CODEX COMMITTEE ON FOOD IMPORT AND EXPORT INSPECTION aAND CERTIFICATION SYSTEMS. *DISCUSSION PAPER ON FOOD INTEGRITY AND FOOD AUTHENTICITY*, (p. 1-7). Brisbane, Australia.

Food Integrity Handbook. (2018). doi:https://doi.org/10.32741/fihb

Garrido-Delgado, R., Dobao-Prieto, M., Arce L., & Valcarcel, M. (2015). Determination of volatile compounds by GC-IMS to assign the quality of virgin olive oil. *Food Chemistry*, 572-579.

Gerhardt, N., Birkenmeier, M., Schwolow, S., Rohn, S., & Weller, P. (2018). Volatile-Compound Fingerprinting by Headspace-Gas-Chromatography Ion-Mobility Spectrometry (HS-GC-IMS) as a Benchtop Alternative to 1H NMR Profiling for Assessment of the Authenticity of Honey. *Analytical Chemistry*, 1777–1785.

Ghojogh, B., & Crowley, M. (2022). Unsupervised and Supervised Principal Component Analysis: Tutorial. *arXiv*.

Ghoneim, S. (2019, April 2). *Accuracy, Recall, Precision, F-Score & Specificity, which to optimize on?* Tratto da Towards Data Science: https://towardsdatascience.com/accuracy-recall-precision-f-score-specificity-which-to-optimize-on-867d3f11124

Han, L., Zhang, Y., Song, J., Fan, M., Yu, Y., Liu, P., . . . She, Y. (2018). Automatic untargeted metabolic profiling analysis coupled withChemometrics for improving metabolite identification quality toenhance geographical origin discrimination capability. *Journal of Chromatography A*, 12-20.

Jurado-Campos, N., Martin-Gomez, A., Saavedra, D., & Arce, L. (2021). Usage considerations for headspace-gas chromatography-ion mobility spectrometry as a suitable technique for qualitative analysis in a routine lab. *Journal of Chromatography A*, 461937.

Klockmann, S., Reiner, E., Bachmann, R., Hackl, T., & Fischer, M. (2016). Food Fingerprinting: Metabolomic Approaches for Geographical Origin Discrimination of Hazelnuts (Corylus avellana) by UPLC-QTOF-MS. *Journal of Agricultural and Food Chemistry*, 9253-9262.

Klockmann, S., Reiner, E., Cain, N., & Fischer, M. (2017). Food Targeting: Geographical Origin Determination of Hazelnuts (Corylus avellana) by LC-QqQ-MS/MS-Based Targeted Metabolomics Application. *Journal of Agricultural and Food Chemistry*, 1456-1465.

Lang, C., Weber, N., Moller, M., Schramm, L., Schelm, S., Kohlbacher, O., & Fischer, M. (2021). Genetic authentication: Differentiation of hazelnut cultivars using polymorphic sites of the chloroplast genome. *Food Control*, 108344.

Maestri, E., Imperiale, D., & Marmiroli, N. (2020). Nuts, nut products and other seeds. In *Food Integrity Handbook.* doi:https://doi.org/10.32741/fihb

Manfredi, M., Robotti, E., Quasso, F., Mazzucco, E., Calabrese, G., & Marengo, E. (2018). Fast classification of hazelnut cultivars through portable infrared spectroscopy and chemometrics. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 427-435.

McEwan, J. A., & Lyon, D. H. (2003). SENSORY EVALUATION | Sensory Rating and Scoring Methods. In B. Caballero, *Encyclopedia of Food Sciences and Nutrition (Second Edition)* (p. 5148-5152). Elsevier Science Ltd.

MKS Umetrics. (2015, January 26). *Dynacentrix.* Tratto da https://www.dynacentrix.com/telecharg/SimcaP/SIMCA14_User_Guide.pdf

Mohajon, J. (2020, May 9). *Confusion Matrix for Your Multi-Class Machine Learning Model.* Tratto da Towards Data Science: https://towardsdatascience.com/confusion-matrix-for-your-multi-class-machine-learning-model-ff9aa3bf7826

Morin, J., & Lees, M. (2018). *FoodIntegrity Handbook.* Nantes: Eurofins Analytics France.

Moscetti, R., Radicetti, E., Monarca, D., Cecchini, M., & Massantini, R. (2014). Near infrared spectroscopy is suitable for the classification of hazelnuts according to Protected Designation of Origin. *Journal of the Science of Food Agriculture*, 2619-2625.

Powers, D. M. (2011). Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation. *Journal of Machine Learning Technologies*, 37-63.

Rohman, A., Erwanto, Y., M., H. M., M., R., Aldawoud, M., T., & Galanakis, C. M. (2021). The application of DNA-based methods for authentication analysis: Examples in halal. *Food Authentication and Traceability*, 195-213.

Rosso, M. C., Liberto, E., Spigolon, N., Fontana, M., Somenzi, M., Bicchi, C., & Cordero, C. (2018). Evolution of potent odorants within the volatile metabolome of high-quality hazelnuts (Corylus avellana L.): evaluation by comprehensive two-dimensional gas chromatography coupled with mass spectrometry. *Analytical and Bioanalytical Chemistry*, 3491-3506.

Ruiz del Castillo, M. L., Gomez Caballero, E., Blanch, G. P., & Herraiz, M. (2002). Enantiomeric composition of filbertone in hazelnuts and hazelnut oils from different geographical origins. *Journal of the American Oil Chemists' Society*, 589-592.

Ruszkiewicz, D., Myers, R., Henderson, B., Yusof, H., Meister, A., Moreno, S., . . . Thomas, C. (2022). Peppermint protocol: first results for gas chromatography-ion mobility spectrometry. *Journal of Breath Research*, 036004.

Schwolow, S., Rohn, S., Gerhardt, N., & Weller, P. (2019). Data fusion of GC-IMS data and FT-MIR spectra for the authentication of olive oils and honeys—is it worth to go the extra mile? *Analytical and Bioanalytical Chemistry*, 6005-6019.

Speiser, J. L., Durkalski, V. L., & Lee, W. M. (2015). Random forest classification of etiologies for an orphan disease. *Statistics in Medicine*, 887-899.

Speiser, J. L., Miller, M. E., Tooze, J., & Ip, E. (2019). A comparison of random forest variable selection methods for classification prediction modeling. *Expert Systems with Applications*, 93-101.

Stone, H., & Sidel, J. L. (1985). *Sensory Evaluation Practices.* London, UK: Academic Press.

Vera, L., Companioni, E., Meacham, A., & Gygax, H. (2016). Real Time Monitoring of VOC and Odours Based on GC-IMS at Wastewater Treatment Plants. *Chemical Engineering Transactions*, 79-84.

Yin, J., Wu, M., Lin, R., Li, X., Ding, H., Han, L., . . . Li, Z. (2021). Application and development trends of gas chromatography–ion mobility spectrometry for traditional Chinese medicine, clinical, food and environmental analysis. *Microchemical Journal*, 3745.

## *FIGURES*

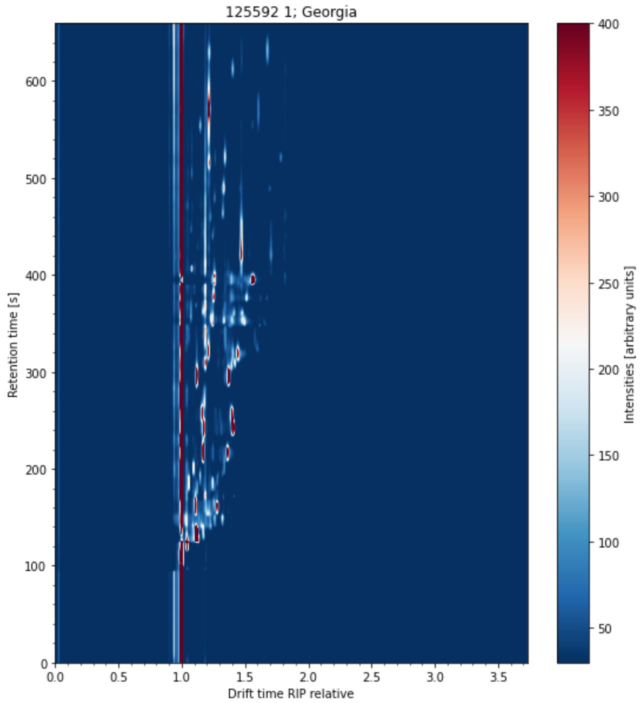



Fig. 1 *2D GC-IMS chromatogram of a fresh hazelnut sample*

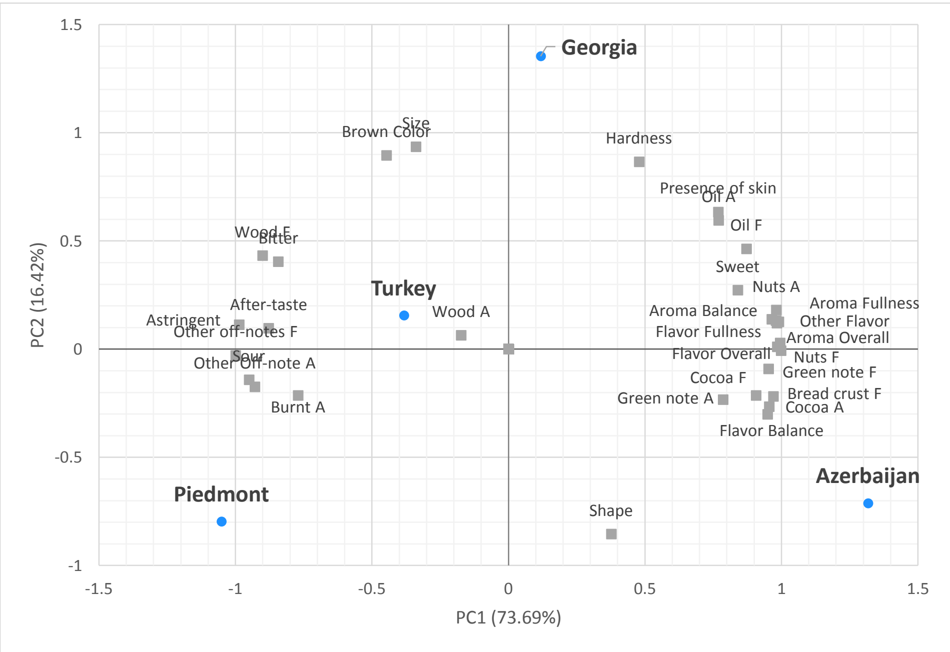
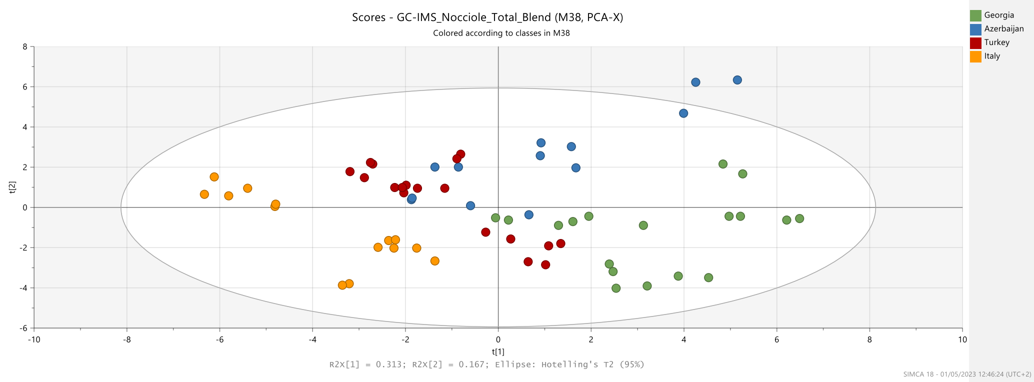
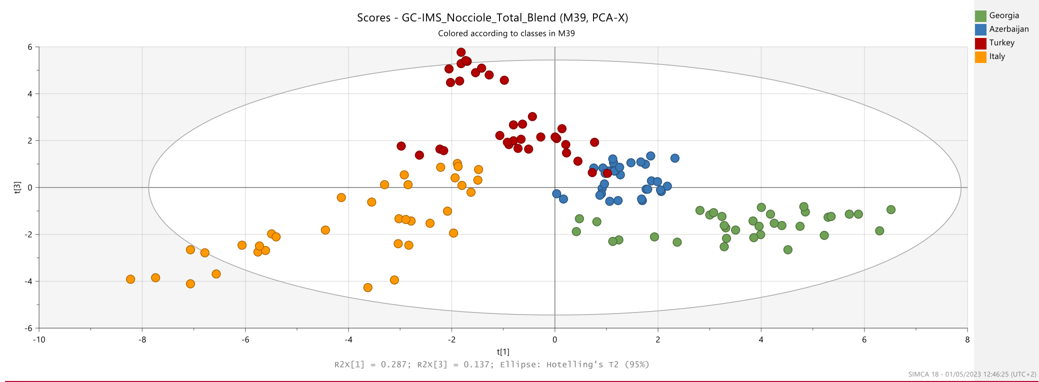



Fig. 2 *Example of a PCA score plot of the fresh hazelnut sample set.*

A)




B)




C)

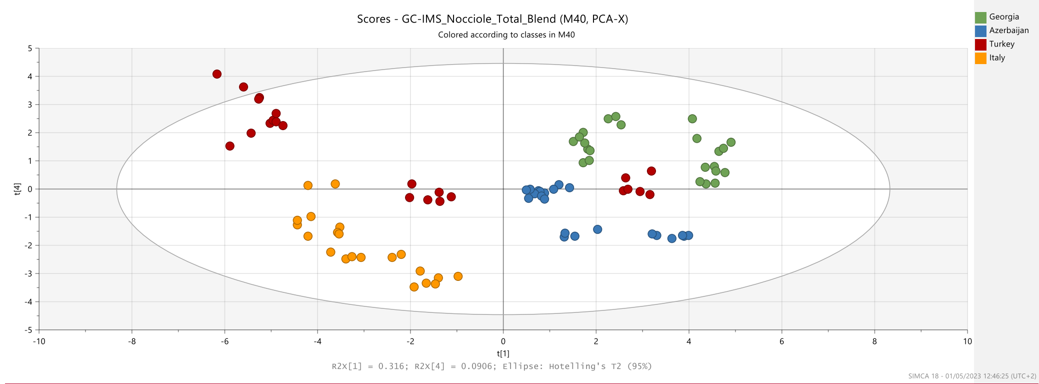
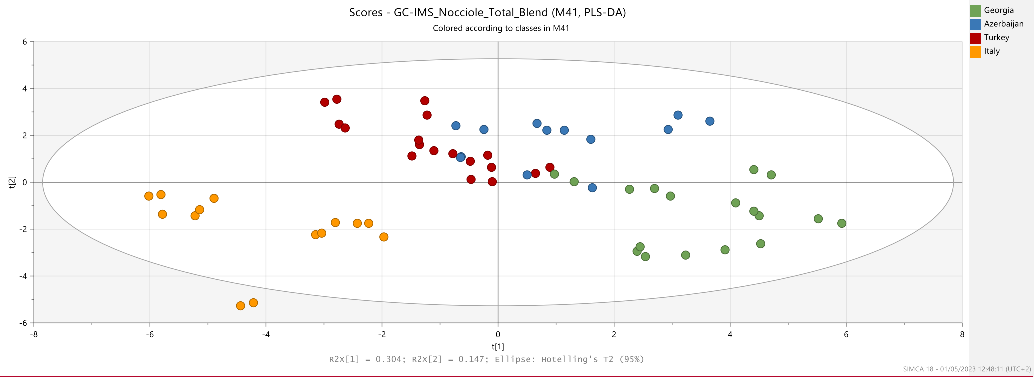
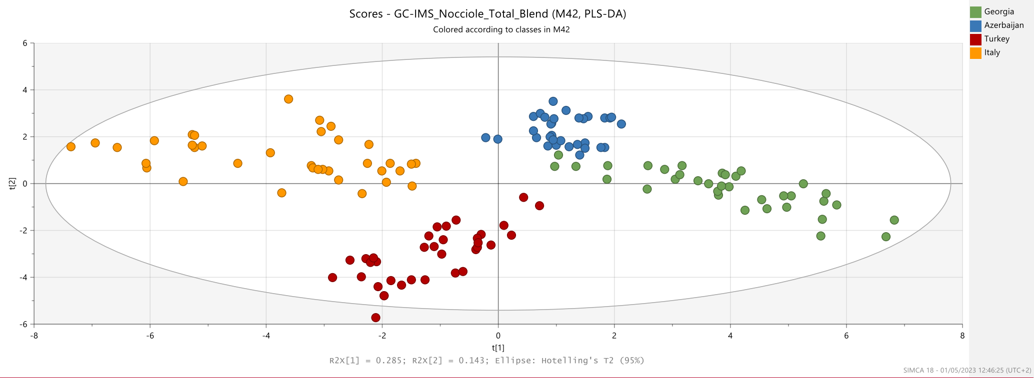



*Fig. 3A) PCA score plot of the fresh hazelnut sample set. 3B) PCA score plot of the roasted hazelnut sample set. 3C) PCA score plot of the hazelnut paste sample set. (Green dots: Azerbaijan, yellow dots: Georgia, blue dots: Italy, red dots: Turkey)*

A)


Scores - GC-IMS_Nocciole_Total_Blend (M41, PLS-DA)
Colored according to classes in M41
Georgia
Azerbaijan
Turkey
Italy
t[1]
t[2]
R2X[1] = 0.304; R2X[2] = 0.147; Ellipse: Hotelling's T2 (95%)
SIMCA 18 - 01/05/2023 12:48:11 (UTC+2)

B)


Scores - GC-IMS_Nocciole_Total_Blend (M42, PLS-DA)
Colored according to classes in M42
Georgia
Azerbaijan
Turkey
Italy
t[1]
t[2]
R2X[1] = 0.285; R2X[2] = 0.143; Ellipse: Hotelling's T2 (95%)
SIMCA 18 - 01/05/2023 12:46:25 (UTC+2)

C)

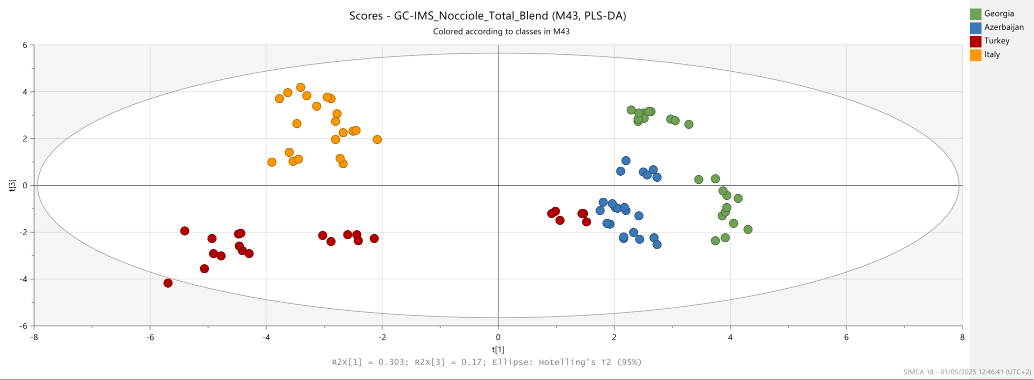
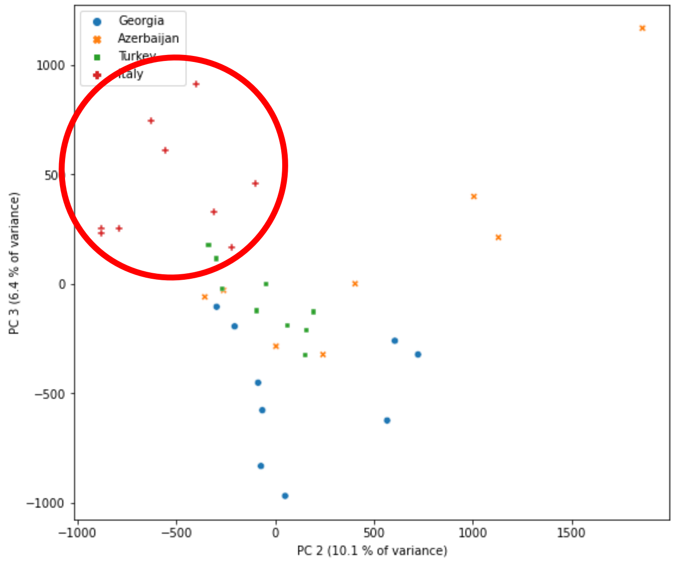
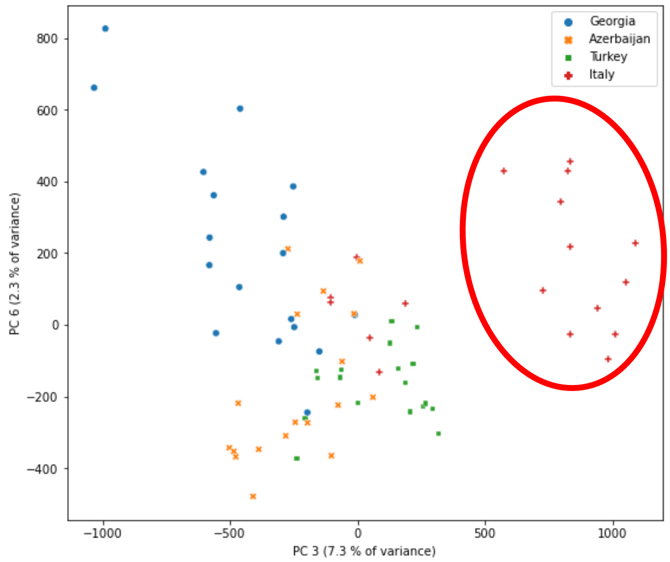

Scores - GC-IMS_Nocciole_Total_Blend (M43, PLS-DA)
Colored according to classes in M43
Georgia
Azerbaijan
Turkey
Italy
t[1]
t[3]
R2X[1] = 0.303; R2X[3] = 0.17; Ellipse: Hotelling's T2 (95%)
SIMCA 18 - 01/05/2023 12:46:41 (UTC+2)

Fig. 4A) *PLS-DA score plot of the fresh hazelnut sample set.* 4B) *PLS-DA score plot of the roasted hazelnut sample set.* 4C) *PLS-DA score plot of the hazelnut paste sample set. (Green dots: Azerbaijan, yellow dots: Georgia, blue dots: Italy, red dots: Turkey)*

A)




B)




C)

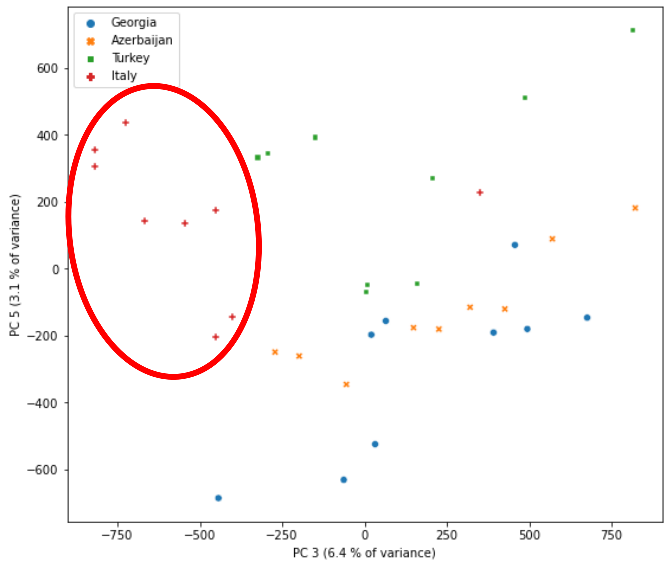
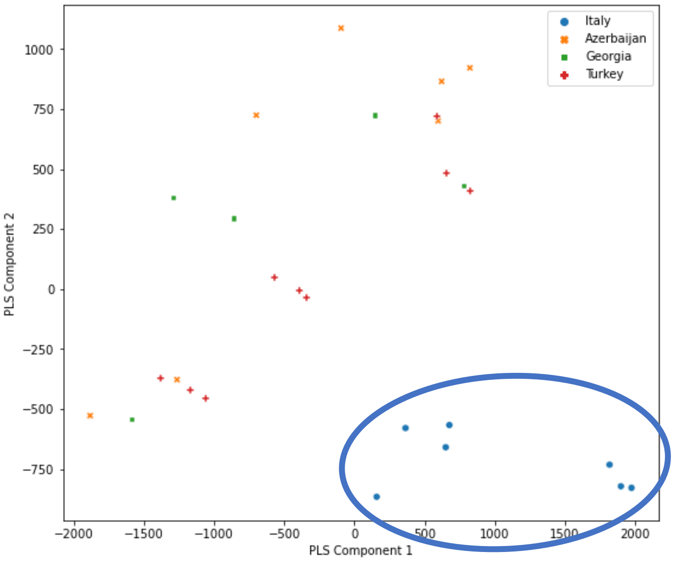
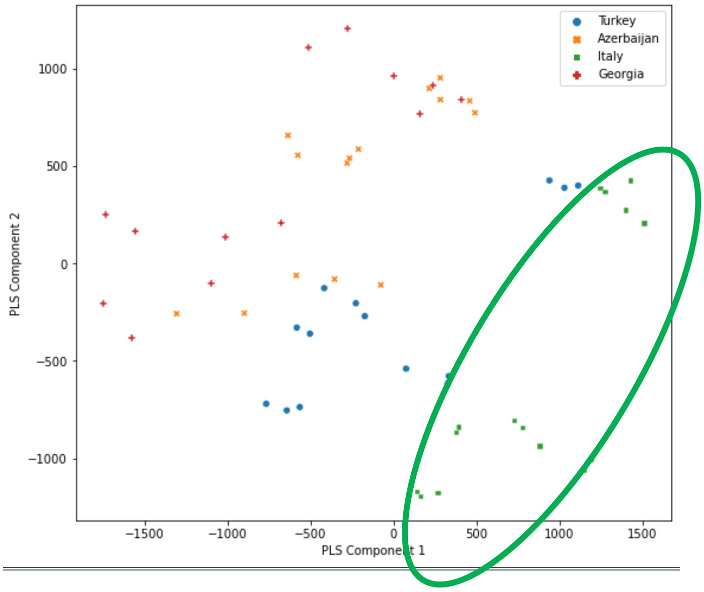



*Fig. 5A) PCA score plot of the fresh hazelnut sample set. 5B) PCA score plot of the roasted hazelnut sample set. 5C) PCA score plot of the hazelnut paste sample set. (Blue dots: Georgia, orange crosses: Azerbaijan, green squares: Turkey, red crosses: Italy, red ellipse: Italian cluster)*

A)


Italy
Azerbaijan
Georgia
Turkey
PLS Component 2
PLS Component 1
1000
750
500
250
0
−250
−500
−750
−2000
−1500
−1000
−500
0
500
1000
1500
2000

B)


Turkey
Azerbaijan
Italy
Georgia
PLS Component 2
PLS Component 1
1000
500
0
−500
−1000
−1500
−1000
−500
0
500
1000
1500

C)

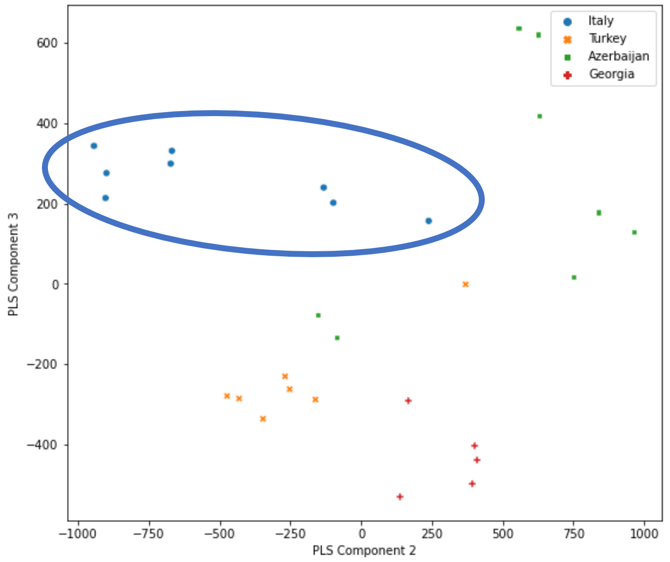
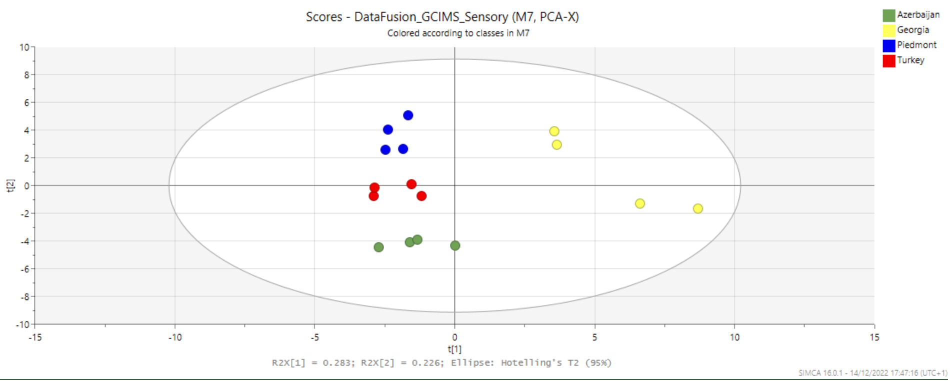
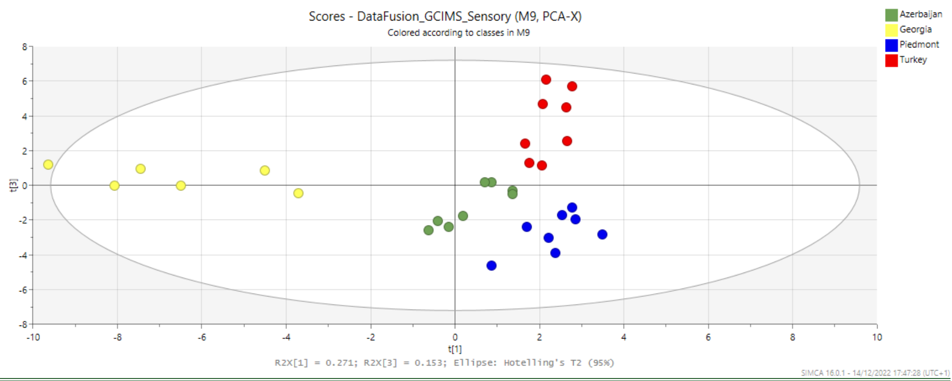



Fig. 6A) *PLS-DA score plot of the fresh hazelnut sample set. (Blue dots: Italy, orange crosses: Azerbaijan, green squares: Georgia, red crosses: Turkey, blue ellipse: Italian cluster) 6B) PLS-DA score plot of the roasted hazelnut sample set. (Blue dots: Turkey, orange crosses: Azerbaijan, green squares: Italy, red crosses: Georgia, green ellipse: Italian cluster)* 6C) *PLS-DA score plot of the hazelnut paste sample set. (Blue dots: Italy, orange crosses: Turkey, green squares: Azerbaijan, red crosses: Georgia, blue ellipse: Italian cluster)*

A)



B)




C)

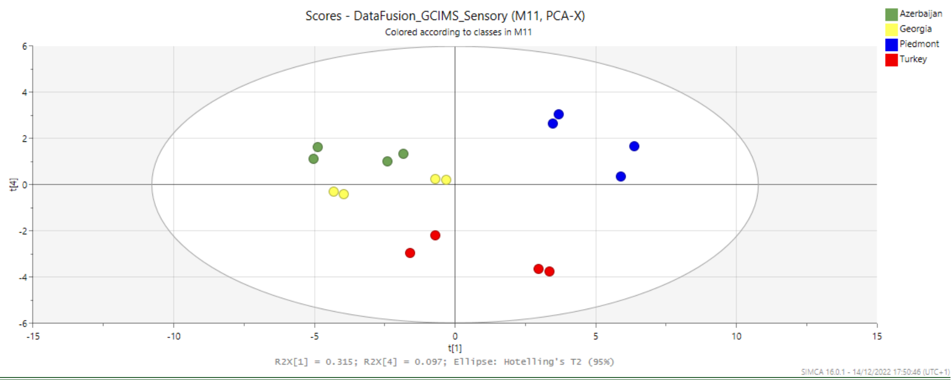
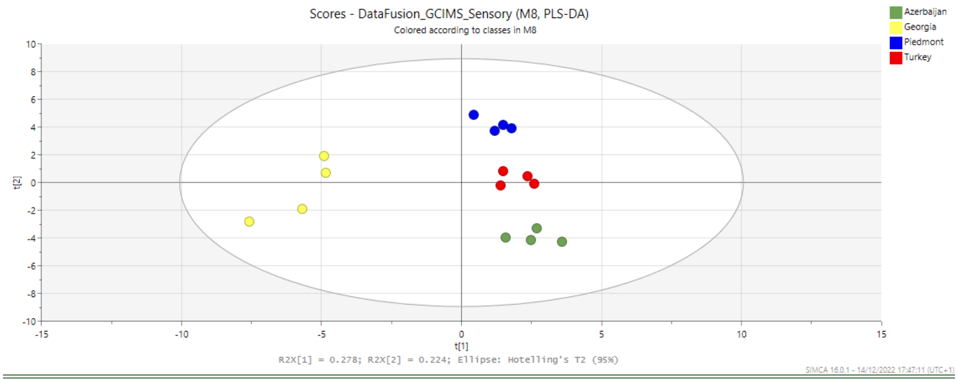
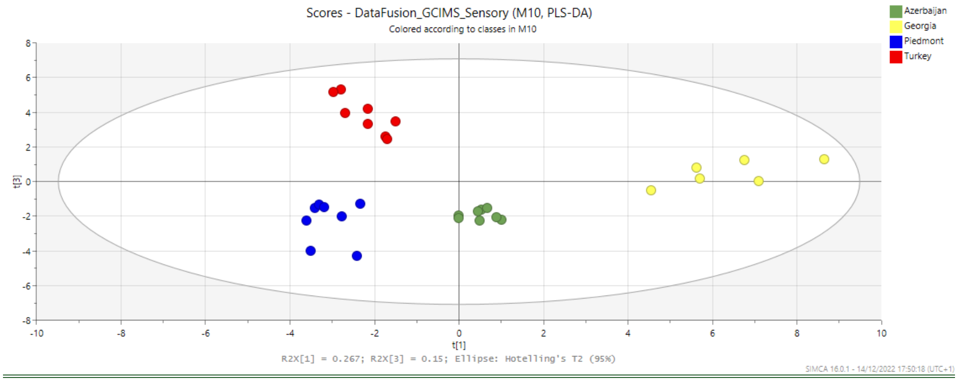



Figure 7A) *PCA score plot of fresh hazelnut samples after data fusion.* 7B) *PCA score plot of roasted hazelnut samples after data fusion. 7C) PCA score plot of hazelnut paste samples after data fusion. (Green dots: Azerbaijan, yellow dots: Georgia, blue dots: Piedmont, red dots: Turkey)*

A)


Scores - DataFusion_GCIMS_Sensory (M8, PLS-DA)
Colored according to classes in M8
Azerbaijan
Georgia
Piedmont
Turkey
t[2]
t[1]
R2X[1] = 0.278; R2X[2] = 0.224; Ellipse: Hotelling's T2 (95%)
SIMCA 16.0.1 - 14/12/2022 17:47:11 (UTC+1)

B)


Scores - DataFusion_GCIMS_Sensory (M10, PLS-DA)
Colored according to classes in M10
Azerbaijan
Georgia
Piedmont
Turkey
t[3]
t[1]
R2X[1] = 0.267; R2X[3] = 0.15; Ellipse: Hotelling's T2 (95%)
SIMCA 16.0.1 - 14/12/2022 17:50:18 (UTC+1)

C)

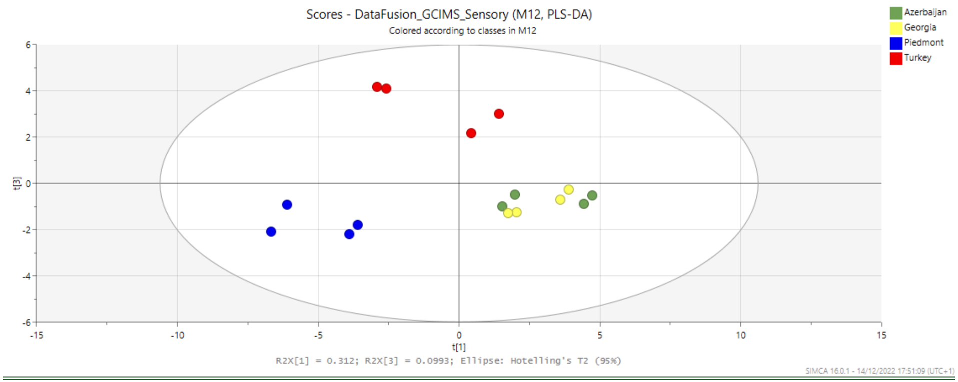

Scores - DataFusion_GCIMS_Sensory (M12, PLS-DA)
Colored according to classes in M12
Azerbaijan
Georgia
Piedmont
Turkey
t[3]
t[1]
R2X[1] = 0.312; R2X[3] = 0.0993; Ellipse: Hotelling's T2 (95%)
SIMCA 16.0.1 - 14/12/2022 17:51:09 (UTC+1)

Figure 8A) *PLS-DA score plot of fresh hazelnut samples after data fusion.* 8B) *PLS-DA score plot of roasted hazelnut samples after data fusion.* 8C) *PLS-DA score plot of hazelnut paste samples after data fusion. (Green dots: Azerbaijan, yellow dots: Georgia, blue dots: Piedmont, red dots: Turkey)*

## *TABLES*

A)

| | **Members** | **Correct** | **Georgia** | **Azerbaijan** | **Turkey** | **Italy** | **No class (YPred <= 0)** |
|---|---|---|---|---|---|---|---|
| **Georgia** | 4 | 100% | 4 | 0 | 0 | 0 | 0 |
| **Azerbaijan** | 4 | 100% | 0 | 4 | 0 | 0 | 0 |
| **Turkey** | 6 | 100% | 0 | 0 | 6 | 0 | 0 |
| **Italy** | 12 | 100% | 0 | 0 | 0 | 12 | 0 |
| **No class** | 0 | | 0 | 0 | 0 | 0 | 0 |
| Total | 26 | 100% | 4 | 4 | 6 | 12 | 0 |

B)

| | **Members** | **Correct** | **Georgia** | **Azerbaijan** | **Turkey** | **Italy** | **No class (YPred <= 0)** |
|---|---|---|---|---|---|---|---|
| **Georgia** | 9 | 100% | 9 | 0 | 0 | 0 | 0 |
| **Azerbaijan** | 11 | 100% | 0 | 11 | 0 | 0 | 0 |
| **Turkey** | 11 | 100% | 0 | 0 | 11 | 0 | 0 |
| **Italy** | 16 | 100% | 0 | 0 | 0 | 16 | 0 |
| **No class** | 0 | | 0 | 0 | 0 | 0 | 0 |
| Total | 47 | 100% | 9 | 11 | 11 | 16 | 0 |

C)

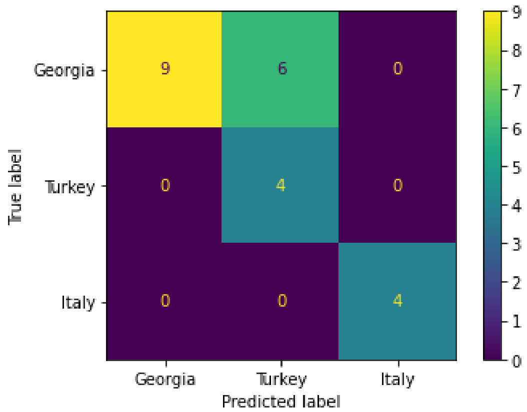
| | **Members** | **Correct** | **Georgia** | **Azerbaijan** | **Turkey** | **Italy** | **No class (YPred <= 0)** |
|---|---|---|---|---|---|---|---|
| **Georgia** | 7 | 100% | 7 | 0 | 0 | 0 | 0 |
| **Azerbaijan** | 8 | 100% | 0 | 8 | 0 | 0 | 0 |
| **Turkey** | 7 | 100% | 0 | 0 | 7 | 0 | 0 |
| **Italy** | 12 | 100% | 0 | 0 | 0 | 12 | 0 |
| **No class** | 0 | | 0 | 0 | 0 | 0 | 0 |
| Total | 34 | 100% | 7 | 8 | 7 | 12 | 0 |

Table 1A) *Misclassification table of the selected fresh hazelnut prediction set (n=26).* 1B) *Misclassification table of the selected roasted hazelnut prediction set (n=47).* 1C) *Misclassification table of the selected hazelnut paste prediction set (n=34). (Green cell: samples correctly classified, yellow cell: samples misclassified)*

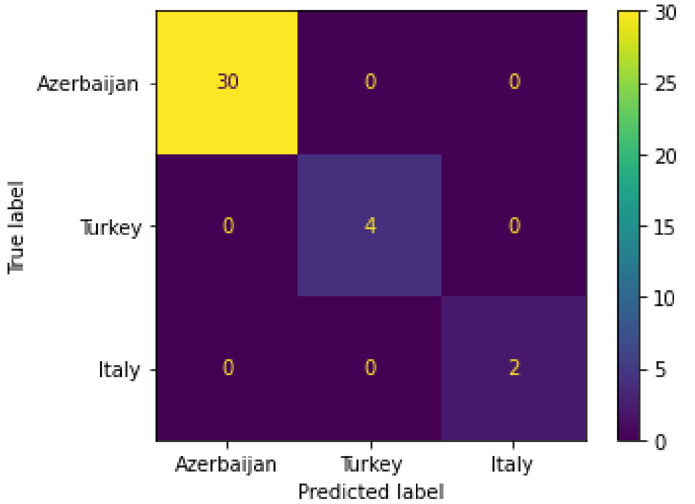
| MATRIX | ACCURACY | PRECISION | SENSITIVITY | F1-SCORE | SPECIFICITY |
|---|---|---|---|---|---|
| FRESH HAZELNUTS | 74 % | 90 % | 74 % | 76 % | 74 % |
| ROASTED HAZELNUTS | 100 % | 100 % | 100 % | 100 % | 100 % |
| HAZELNUT PASTES | 92 % | 92 % | 92 % | 90 % | 92 % |

Table 2 *Accuracy, Precision, Sensitivity, F1-Score, and Specificity values of fresh, roasted, paste hazelnut validation sets.*

A)




B)




C)

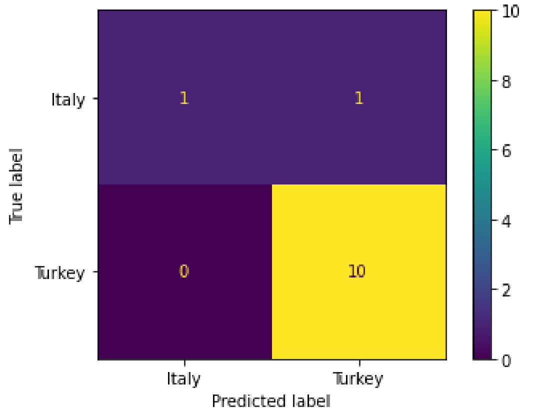



Table 3A) *Confusion matrix of the fresh hazelnut validation set.* 3B) *Confusion matrix of the roasted hazelnut validation set.* 3C) *Confusion matrix of hazelnut paste validation set.*

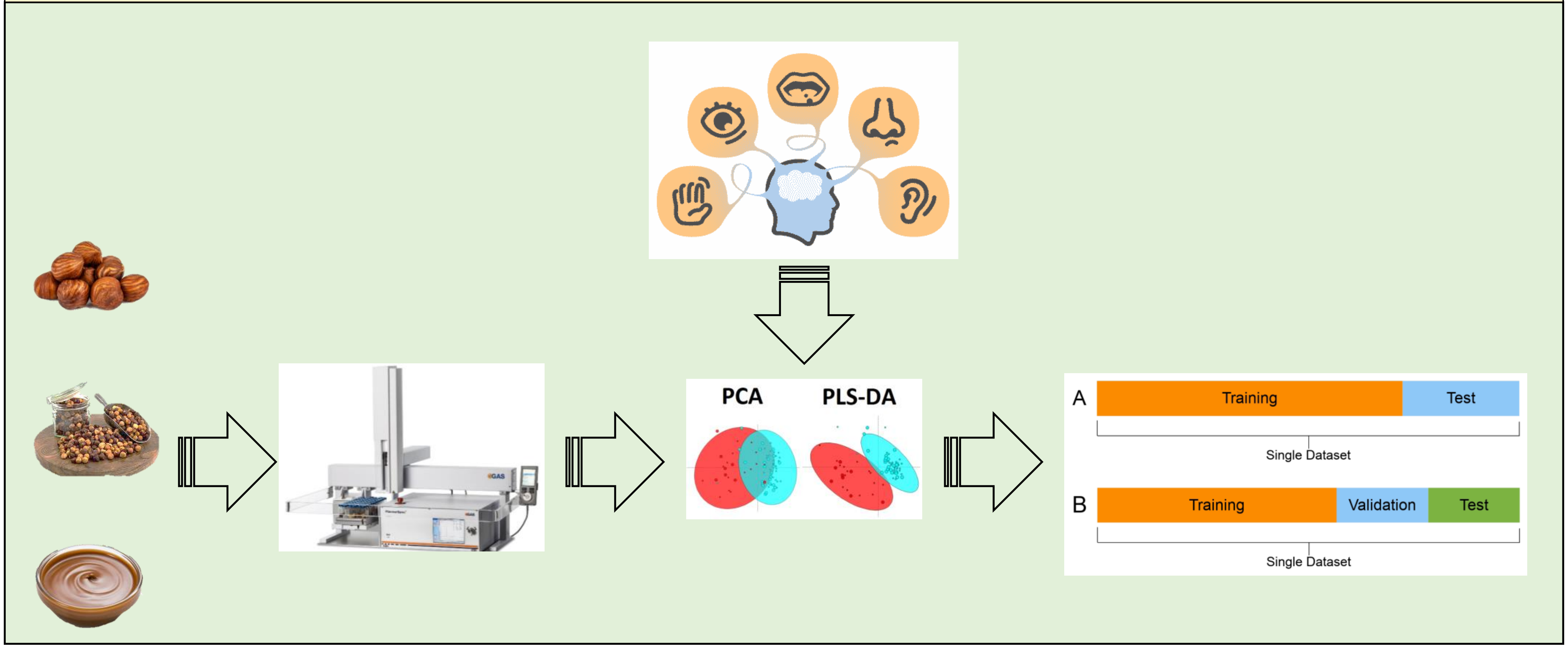
**Formatted:** Left

**Formatted:** Font: Not Italic

Click here to access/download
**Supplementary material for online publication only**
Supplementary materials.docx

Graphical Abstract

# A Geographical Origin assessment of Italian Hazelnuts: Gas Chromatography-Ion mobility spectrometry coupled with Multivariate Statistical Analysis and Data Fusion approach



# A Geographical Origin assessment of Italian Hazelnuts: Gas Chromatography-Ion mobility spectrometry coupled with Multivariate Statistical Analysis and Data Fusion approach

*Giuseppe Sammarco[1,3], Daniele Bardin[1], Federica Quaini[1], Chiara Dall'Asta[3], Joscha Christmann[4], Philipp Weller[4], Michele Suman[1,2]*

[1] Sensory and Analytical Food Science, Barilla G. e R. Fratelli S.p.A., Parma, Italy
[2] Department for Sustainable Food Process, Catholic University Sacred Heart, Piacenza, Italy
[3] Department of Food and Drug, University of Parma, Parma, Italy
[4] Institute of Analytics and Bioanalytics, Faculty of Biotechnology, Mannheim University of Applied Sciences, Mannheim, Germany

***CRediT author statement***

Giuseppe Sammarco: methodology, data curation, formal analysis, investigation, validation, visualization, writing – original draft, writing – review & editing; Daniele Bardin: data curation, formal analysis, investigation, visualization, writing – originl draft; Federica Quaini: data curation, investigation, methodology, supervision, visualization, writing – review & editing; Chiara Dall'Asta: conceptualization, methodology, supervision, visualization, writing – original draft, writing – review & editing; Joscha Christmann: data curation, investigation, validation, visualization, writing – original draft; Philipp Weller: methodology, data curation, supervision, visualization, writing – original draft; Michele Suman: project administration, conceptualization, methodology, supervision, visualization, writing – original draft, writing – review & editing.

**Declaration of interests**

☒The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: