FOODINTEGRITY

Ensuring the Integrity of the European food chain

613688: Collaborative Project
Seventh Framework Programme
KBBE.2013.2.4-01: Assuring quality and authenticity in the food chain

Deliverable D18.5

White paper ("Good practices and methodological guidelines for the validation and application of the untargeted analysis for food authenticity and traceability") draft

Author(s): Monica Locatelli, Cristiano Garino, Luigi Portinale, Giorgio Leonardi, Maurizio Rinaldi, Vito Gallo, Linda Monaci, Vincenzo Lippolis, Emiliano De Dominicis, Marialuisa Piva, Elisa Gritti, Michal Godula, Joana Amaral, Isabel Mafra and Marco Arlorio

Beneficiary(s): P56

Date of preparation: 21/12/2018
Period covered: 01/08/2017 - 31/08/2018
Status: version 1

<table>
<thead>
<tr>
<th>Dissemination level</th>
<th>PU</th>
<th>Public</th>
<th>PP</th>
<th>Restricted to other participants</th>
<th>RE</th>
<th>Restricted to a group specified by the consortium</th>
<th>CO</th>
<th>Confidential, only members of the consortium</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The project has received funding from the European Union’s Seventh Framework Programme for research, technological development and demonstration under grant agreement No. 613688.

Disclaimer: The information expressed in this deliverable reflects the authors’ views; the European Commission is not liable for the information contained therein.
TABLE OF CONTENTS

1. Description of Deliverable p 1
2. Guideline draft p 2
3. Achievement of the Deliverable p 75

Deliverable 18.5

White paper (“Good practices and methodological guidelines for the validation and application of the untargeted analysis for food authenticity and traceability”) draft

1. Description of Deliverable

The main objective of this deliverable (developed by the Authors of the WP18 in collaboration with other Partners within the Food Integrity Project) has been to prepare a draft for a “consensus paper” focused on the best practices for an effective, functional and validated application of non-targeted analytical methods, aimed at the determination of food authenticity, origin and integrity. The document is primarily addressed to all the interested technical Stakeholders, and presents a selected method for the statistical validation of a “non-targeted” analysis. The report includes a concise description of the validation procedure, focusing on the selection of the criteria for a successful validation. Common misleading errors and problems are presented, and consequent re-directing actions are proposed.

The content of this Deliverable will be edited as public “Guidance on the best practices for the development and validation of non-targeted methods for food integrity assessment”, directly downloadable from the Food integrity web site.
2. Guideline draft

SECTION 1

This document has been prepared by the working Group of WP18: Università degli Studi del Piemonte Orientale (IT), Meriéux Nutrisciences (IT), Thermo Scientific (D), ICETA (PT), ISPA-CNR (IT), in collaboration with Innovative Solutions (IT) and COOP (IT), in the framework of Food Integrity Project.

PURPOSE

Considering the gap already identified by the Food Integrity Consortium, this document provides good practices and methodological guidelines on the development and validation of non-targeted analytical methods for the classification of food commodities. Non-targeted methods consist of analytical measurements coupled with a suitable statistical elaboration, which enables recognizing trends and deviations from a given pattern of signals assigned to the reference model. Non-targeted methods can be applied to tackle a variety of issues related to food classification topics, like the assessment of the food integrity, and in general the compliance with all the information accompanying the commodity (its integrity, origin, composition, freshness, etc.).

OVERVIEW

In spite of the widespread tendency to consider any analytical problem as quantitative, a considerable number of practical cases that require an analytical solution within the food sciences have to solve qualitative questions.

Typical cases concern identification of raw materials and ingredients, monitoring of ripening, investigations on evolution during storage/shelf life, verification of authenticity of finished products, and assessments of quality and origin. In all of these cases, the answer to the problem of interest can be provided by applying appropriate classification strategies, usually based on multivariate data treatment of the analytical results.1

Beside classic targeted methods, aimed at identifying and quantifying one or multiple primary and secondary analytical markers, in the last years several advanced analytical approaches coupled to chemometric tools have been proposed to characterize and classify foods based on their non-targeted fingerprint. However, most of them are not subjected to a well-defined validation protocol, mainly because there are neither formally defined parameters for performance evaluation available, nor procedures for standardised validation. The main

---

The purpose of this document is thus to tackle this gap and provide a methodological guidance, based on which harmonizing the validation protocols for non-targeted methods.

Despite there are no universally accepted criteria on how to approach and carry out non-targeted analyses, several papers proposing solutions and protocols have been published over the years. As following reported, all the efforts were directed to discuss and by-pass the principal gaps in this field. A brief summary of the main publications issued over the last 10 years, that tried to propose specific solutions and suggestions to build a validation strategy is reported below:

- 2007, Goodacre et al., Proposed minimum reporting standards for data analysis in metabolomics. It proposes the minimum level of reporting required in order to account accurately for any chemometric/statistical methods supporting the conclusions of metabolomics. The Authors suggest structuring the report into 3 distinct sections, such as the capture of the raw data and the formation of the initial data matrix, the pre-processing and the preparation of the data for the main modelling process, and the analysis of the clean (transformed) data matrix.2

- 2012, Dunn et al., The importance of experimental design and QC samples in large-scale and MS-driven untargeted metabolomics studies of humans. This review describes the importance of the experimental design and of the incorporation of quality control samples in metabolomic studies, to perform signal correction, reduce the analytical variation and quantitatively determine the analytical precision, whether single or multiple analytical experiments are applied. The Authors describe how to apply this in quality assurance processes.3

- 2013, Naz et al., Multiplatform Analytical Methodology for Metabolic Fingerprinting of Lung Tissue. It presents the analytical method optimization (sample treatment and analysis), characterization and validation of a non-targeted global fingerprinting approach using multiple complementary analytical techniques (LC, GC and CE–MS) to analyse lung tissue.4

- 2013, Tengstrand et al., A concept study on non-targeted screening for chemical contaminants in food using liquid chromatography–mass spectrometry in combination with a metabolomics approach. It proposes a generic method to screen for unexpected contaminants in food using UHPLC-TOF-MS. Orange juice was chosen as example.5

2 Dunn, W. B., Wilson, I. D., Nicholls, A. W., & Broadhurst, D. (2012). The importance of experimental design and QC samples in large-scale and MS-driven untargeted metabolomic studies of humans. Bioanalysis, 4(18), 2249-2264
- 2014, Kiss et al., Data-handling strategies for metabonomic studies: example of the UHPLC-ESI/ToF urinary signature of tetrahydrocannabinol in humans. The paper presents a detailed metabonomic study on human urine and the corresponding validation steps, with a special emphasis on the importance of assessing the data quality. It also describes, step by step, the statistical tools currently used and offers a critical view on some of their limits.

- 2014, López et al., Validation of multivariate screening methodology. Case study: Detection of food fraud. It proposes a strategy to validate non-targeted analysis based on 2 steps: 1) establish the multivariate classification model and use receiver operating characteristic (ROC) curves to optimize the significance level for setting the model boundaries; 2) evaluate the performance parameter from the contingency table results and performance characteristic curves. The adulteration of hazelnut paste with almond paste and chickpea flour was the case study.6

- 2014, Naz et al., Method validation strategies involved in non-targeted metabolomics. The Authors recommend some steps to consider during the development of a non-targeted metabolomics analytical method. Recommendations involve how to select the samples, how to set analytical conditions, how to obtain dependable, consistent and accurate data by running quality control samples, how to cross-validate statistical models.

- 2014, Franceschi et al., MetaDB a data processing workflow in untargeted MS-based metabolomics experiments. It present a bioinformatic pipeline to process MS-based metabolomics data, suggesting ways to pre-process and annotate them, create sample lists, organize the storage and generate survey plots for quality assessment.7

- 2015, Egert et al., A peaklet-based generic strategy for the untargeted analysis of comprehensive two-dimensional gas chromatography mass spectrometry data sets. It presents an automated workflow for the analysis of large GC × GC-qMS metabolomics data sets.8

- 2015, Riedl et al., Review of validation and reporting of non-targeted fingerprinting approaches for food authentication. It develops a generic scheme for multivariate model validation that shall guide users through validation and reporting of non-targeted fingerprinting results.9

7 Franceschi, P., Mylonas, R., Shahaf, N., Scholz, M., Arapitsas, P., Masuero, D., ... & Wehrens, R. (2014). MetaDB a data processing workflow in untargeted MS-based metabolomics experiments. Frontiers in bioengineering and biotechnology, 2, 72
- 2015, Mastrangelo et al., *From sample treatment to biomarker discovery: A tutorial for untargeted metabolomics based on GC-(EI)-Q-MS*. This tutorial provides a comprehensive description of the GC-MS-based untargeted metabolomics workflow, including ethical approval requirement, sample collection and storage, equipment maintenance, and setup, sample treatment, monitoring of analytical variability, data pre-processing, data processing, statistical analysis and validation, detection of outliers and biological interpretation of the results.10

- 2016, Di Guida et al., *Non-targeted UHPLC-MS metabolomic data processing methods: a comparative investigation of normalisation, missing value imputation, transformation and scaling*. The paper deals with data processing applied to UHPLC-MS datasets, trying to define the most appropriate workflow. The Authors suggest different normalization procedures, pointing out that the choice of method is essential to maximise the biological derivations.11

- 2016, Alewijn et al., *Validation of multivariate classification methods using analytical fingerprints – concept and case study on organic feed for laying hens*. It proposes a new approach for a full validation of multivariate classification methods applied to food and feed analysis in practice. The Authors suggest using the kernel density estimate (KDE) function to interpolate the probabilistic results obtained for training and validation sets, creating an expanded probability distribution for the prediction of the classes. Such distribution, based on the observed cross validation and external validation distribution, offers an easy way to judge the future performance of the method.12

- 2016, Bijttebier et al., *Bridging the gap between comprehensive extraction protocols in plant metabolomics studies and method validation*. It compares comprehensive extraction protocols for plant metabolomics with LC-amMS, in order to select the most performing one and validate it. The Authors point out that none of the procedures tested outperforms the others, and compromises have to be made during method selection.13

- 2018, Dudzik et al., *Quality assurance procedures for mass spectrometry untargeted metabolomics. a review*. It considers the occurrence of unwanted variation regarding the pre-analytical, analytical and post-analytical phase of mass spectrometry-based metabolomics

---


experiments, analysing possible sources. A workflow aimed at minimizing these variations is proposed.14

- 2018, McGrath et al., What are the scientific challenges in moving from targeted to non-targeted methods for food fraud testing and how can they be addressed? – Spectroscopy case study. In this opinion paper, many aspects relating to the role of non-targeted spectroscopy based methods for food fraud detection are considered: (i) a review of the current non-targeted spectroscopic methods to include the general differences with targeted techniques; (ii) overview of in-house validation procedures including samples, data processing and chemometric techniques with a view to recommending a harmonized procedure; (iii) quality assessments including QC samples, ring trials and reference materials; (iv) use of “big data” including recording, validation, sharing and joint usage of databases.15

- 2018, Cavanna et al., The scientific challenges in moving from targeted to non-targeted mass spectrometric methods for food fraud analysis: A proposed validation workflow to bring about a harmonized approach. It summarizes the scientific activities published from 2001 to 2017 on the non-targeted mass spectrometry methods to food fraud detection. The Authors proposed a possible approach for the development and validation of these types of methods and also providing global considerations on their applicability for legal purposes. Additionally, the Authors concluded that both public and private institutions will have to increase their efforts to finalize a shared approach aimed to guide the development and validation of robust and reliable non-targeted method for food fraud detection.16

- 2018 (in press), Ballin and Laursen, To target or not to target? Definitions and nomenclature for targeted versus non-targeted analytical food authentication. In this commentary paper the authors propose definitions and nomenclature for targeted and non-targeted methods as a preliminary step for the harmonization process. A focus on different analytical techniques (analytical chemistry, microscopy, molecular biology) and their use in food authentication, and the identification of possibilities and limitations of targeted and non-targeted analytical approaches, are further reported.17

An example of non-targeted method validation: ‘Appendix XVIII: Guidance On Developing and Validating Non-Targeted Methods For Adulteration Detection’ (US Pharmacopeia)

In November 2016, a “Guidance on developing and validating non-targeted methods for adulteration detection” has been released by the US Pharmacopeia. The document is intended to assist users in supply chain management by providing information that can generally be applied to testing and authentication of raw materials with a variety of analytical techniques. It provides guidance on how to develop and implement one-class, non-targeted classification methods for the detection of economically motivated adulteration (EMA)-related adulterants in food, independent of the analytical technology used, but it is not intended to cover the use of multi-class classification methods.

The Guidance is structured as follows:

- a general OVERVIEW on non-targeted methods applied to EMAs;
- an OUTLINE AND SCOPE that includes Glossary of Terms and examples of in-scope and out-of-scope methods;
- a description of THE GENERIC THOUGHT PROCESS for the validation. The process includes:
  - establishing an Applicability Statement;
  - determining common ranges and levels of adulterants based on a thorough risk assessment;
  - selecting an appropriate analytical approach;
  - defining Reference and Test sets;
  - establishing Performance Criteria, such as Sensitivity and Specificity rate;
  - validating the Optimized Method using Typical and Atypical samples;
- a section on how to USE/MAINTAIN/MONITOR/REVALIDATE the developed method;
- a series of Follow-up Actions, including confirmatory analyses and actions to prevent further adulteration

This version of the USP Guidelines is still a draft, and it has already been updated.

END USERS

This guidance for harmonization in validation of non-targeted methods is intended to be useful to analysts working for public and private Research Centres, public Authorities, public Bodies dedicated to official control of food, as well as to private laboratories providing services of food analysis. The person - or the team - responsible for setting up the method should be well trained both in analytical chemistry and in advanced chemometrics. Collaboration with mathematicians and experts in computer science (particularly on data mining) is strongly recommended.
GLOSSARY OF TERMS (alphabetical order)

Analytical marker: or more simply “marker”, is an individual, specific, and pre-determined analytical target, directly or indirectly related to the authenticity issue.\(^{18}\)

Balanced dataset: a dataset that is not unbalanced.

Class: a class is defined as a group of objects/individuals that share some common properties. In the present context, such properties are described by numerical variables, and objects in a given class are characterized by close in range values of such variables.

Classification: a procedure receiving as input a sample (i.e. a dataset row) and producing a class as output. The output can also be a probabilistic distribution over the possible classes.

Classifier: an algorithm implementing a classification.

Cross validation: a procedure where a given dataset is split into different folds, using some of them as training set and some as test set.

Dataset: a data matrix where each row represents a sample (i.e., a set of possibly pre-processed measurements) and columns represent features.

Data dimensionality: the number of columns (features) in the considered dataset.

Data size: the number of rows (samples or instances) in the considered dataset.

Dimensionality reduction: the process of reducing the number of features in a given dataset; it is already part of the data analysis procedure, but it can be considered a data pre-processing step for classification.

False negative: the number of positive samples that are mislabelled as negative.

False positive: the number of negative samples that are incorrectly labelled as positive.

Feature: the attribute or property of an object/individual belonging to a particular class; in the present context a feature refers to a numeric variable resulting from a measurement process, usually followed by an analysis pre-processing treatment.

Fingerprint: refers to the display of multiple non-targeted parameters comprising information about the analytical method used (most frequently, NMR and MS fingerprints).\(^{18}\)

Fold: a subset of samples from a given dataset; to be used in cross-validation (every time some folds are part of the training set, one fold represents the test set).

**Leave-One-Out (LOO) cross validation**: a cross validation where each time one different example is used as test and the remaining ones as training.

**Method development**: the development of a method is the process of optimizing the experimental conditions for sample collection, preparation, and measurement that produce accurate and repeatable results. In the development of a new procedure, the choice of analytical instrumentation and methodology should be based on the intended purpose and scope of the analytical method.\(^{19}\)

**Method validation**: validation is the process used to demonstrate that the developed procedure is suitable for its intended use.

**Model**: the specific function or predictor used to perform the classification: it is a synonym of classifier.

**Model Bias**: a set of erroneous assumptions in the learning algorithm. High bias can cause an algorithm to miss the relevant relations between features and output class (underfitting).

**Model Learning**: the procedure with which a model is obtained (i.e., learnt) from a given dataset.

**Model Variance**: the sensitivity to small fluctuations in the training set. High variance can cause an algorithm to model the random noise in the training data, rather than the intended outputs (overfitting).

**N-fold cross validation**: a cross validation using N different folds. N different tests are performed, each one by using N-1 folds as training set and the remaining fold as test set. Each time the fold for test is different.

**Non-targeted analysis**: it simultaneously detects numerous unspecified targets or data points. In literature the terms non-targeted, nontargeted, untargeted and not targeted are used as synonyms; here the term non-targeted was used being it more frequently reported in food authentication literature.\(^{18}\) Concerning this topic, a broad discussion about the need of harmonization also of terms regarding the non-targeted methods is worldwide taking place. The non-targeted analysis is often referred to as “fingerprinting”.

**Positive and negative samples**: samples belonging to one or more classes of interest are defined as positive; all the other samples are defined as negative.

---

**Pre-processing:** a set of data manipulation procedures resulting in the generation of a feature for a raw instrumental measurement.

**Primary marker:** it provides results directly linked to a specific authenticity issue; it is generally a chemical compound, often subjected to specific legal limits that can be used as an adulterant or can derive from blending and/or undeclared processes.\(^\text{18}\)

**Profiling:** analysis aimed at targeting multiple secondary markers. The number of markers constituting a profile is neither harmonized nor described in the scientific literature; however, it is advisable to have at least 3 markers to obtain a “profiling” analysis.\(^\text{18}\)

**Robust Model:** a model insensitive to small departures from underlying probabilistic assumptions. It provides a good trade-off between bias and variance.

**Instance:** a sample.

**Secondary marker:** it provides information about the authenticity issue only through an indirect way.\(^\text{18}\)

**Signature:** term frequently used by the mass spectrometry community for data from stable isotope or multi-element analysis; can be considered a synonym of profile.\(^\text{18}\)

**Targeted analysis:** analysis aimed at detecting and/or quantifying one or more predetermined analytical target(s). The analytical targets can be either primary or secondary markers.

**Test set:** a dataset used to test the performances of a model.

**Training set:** a dataset used to perform model learning.

**True negative:** the number of negative samples that are correctly labelled by the classifier.

**True positive:** the number of the positive samples that are correctly labelled by the classifier.

**Unbalanced dataset:** a dataset for which the number of samples of one or more classes is very small with respect to other classes (i.e., the number of samples is far from being uniformly distributed among the classes).

**Validation set:** a dataset used to validate a given model, by choosing the suitable model parameters.

**APPLICABILITY OF THE GUIDANCE**

In recent years there has been a significant increase in the number of publications concerning food authenticity and food frauds. The growing interest of the scientific community and the industry was triggered by the willing to prevent both unfair competition and economic fraud,
to comply with legal regulations in force and to address food safety issues. Often, conventional chemical or microbiological tests are not sufficient for such purposes, thus coupling of food fingerprinting and chemometrics represents a valuable tool to unravel the question. For this reason, this guideline is focused not only on analytical aspects, but also on statistical and chemometric approaches.

This guidance is applicable to different fields of analysis and to different analytical techniques coupled to different chemometric tools.

Non-targeted approaches can be applied not only to the classical analytical chemistry techniques (specifically chromatography, spectrometry, spectroscopy and related hyphenated methods), but also to microscopy and molecular biology based techniques\textsuperscript{18,17}

However, due to intrinsic differences of these techniques, in order to define a generalizable and harmonized method development and validation approach, we referred particularly to the analytical chemistry ones, even if the general concepts of the proposed approach can be generally extended.

In the following sections an overview on the possible application fields, the main analytical techniques to which the present guideline is addressed, and the principal approaches for data analysis and the related chemometric tools are presented.

**Examples of application fields**

Non-targeted methods are mainly used to solve classification problems related to food authenticity issues; consequently, most of the application fields of non-targeted methods are related to food frauds prevention and/or discovery. Food frauds were widely described in literature.\textsuperscript{20,21} As summarized by Esslinger and co-authors (2014),\textsuperscript{22} and successively by Ballin and Laursen (2018, in press),\textsuperscript{17} food frauds can be categorized in: i) substitution frauds, in which similar, but generally cheaper, ingredients are used to partially or completely substitute high value ingredients (e.g. substitution of for natural vanillin with synthetically produced vanillin); ii) extension frauds, in which non-authentic substance are added in order to mask low quality ingredients (e.g. addition of Sudan Red dyes to enhance to the colour of poor-quality paprika) or blending and/or undeclared processes (e.g. irradiation); iii) origin frauds (identification of origin), intended in the widest meaning of the term, for example geographic origin, botanical, varietal. Within the frauds related to the origin, the evaluation of the “method of production” can be further included, for example organic vs non organic cultivations and/or productions, wild vs farmed, aging, and so on.

Considered Analytical Techniques

Typical analytical techniques, to which the present guideline is addressed, can be divided into:

1) Mass Spectrometric techniques, 2) Spectroscopic techniques, 3) Miscellaneous

1) Mass spectrometry is a powerful analytical technique in the recent years also exploited in food fingerprinting analysis. Both low resolution and High Resolution Mass Spectrometry have been recently applied to the field of food authentication also coupled with different ionization sources proving to be promising for this application field. In particular, high-resolution mass spectrometry (HRMS) allows high sensitivity and selectivity generating highly informative spectra that can be downstream processed via sophisticated software capable of handling such huge output data file. A wide range of instrumental and technical variants are available for MS spectrometry that can be divided into two major categories:

a) Direct mass spectrometry: Desorption ElectroSpray Ionization (DESI-MS) and Direct Analysis in Time (DART-MS) represent innovative approaches mainly used for food authentication studies. DART-MS showed to be a suitable tool for fast analysis with minimal sample preparation in the evaluation of authenticity of olive oil, meat, beer, vegetables and salmons. Proton Transfer Reaction MS (PTR-MS) and headspace/solid phase micro-extraction MS (HS/SPME-MS), which are also classified as MS-based electronic noses, can be considered as useful approaches for fingerprinting of volatile compounds without sample preparation procedure. These techniques have been mainly used for the rapid detection of adulteration and discrimination of geographical and botanical origin of olive oils, cheeses, wines, honeys and apples.

b) Mass spectrometry coupled to chromatographic techniques: The coupling between gas chromatography and mass spectrometry (GC-MS) has proved to be a successful combination for the analysis of volatile and semi-volatile compounds. Various studies investigated flavour fingerprinting of several foodstuffs (e.g. juices, edible oils, alcoholic beverages, tea and coffee, meat and vegetable and fruits. Liquid chromatography mass spectrometry (LC-MS) has been used for non-targeted analysis of several

---

compounds such as proteins, amino acids, carbohydrates, vitamins, phenolic compounds, triglycerides, pigments, and many others. In the field of food authenticity LC-MS has been applied to different matrices, such as alcoholic beverages, coffee and tea, fish, vegetable and fruit juices.\textsuperscript{22, 25, 26} The use of these techniques increase the number of acquired information but also the mathematical complexity needs to be handled (i.e. three dimensional data).

2) Spectroscopic techniques:

a) Nuclear Magnetic Resonance: NMR spectroscopy results particularly advantageous as fingerprinting technique as it provides information on metabolome by analysing relatively small amounts of sample in a non-destructive, easy and quick way, with a convenient sample preparation. Many studies dealing with NMR based food analysis have been reported emphasizing the power of NMR methods in the discrimination of varieties, crop years, agricultural practices and geographical origin.\textsuperscript{27, 28, 29} Particular attention has been also paid to authentication and fraud detection of a huge list of food products such as olive oil, alcoholic beverages, juices, honey, coffee, cocoa, spices, milk, meat, fish, cereals, etc.

b) Vibrational spectroscopy: Vibrational spectroscopy includes non-invasive fingerprinting techniques that permit rapid and inexpensive analysis. Vibrational spectroscopy techniques applied to food authenticity include Fourier Transform Infrared Spectroscopy (FT-IR), FT-Raman spectroscopy and near infrared (NIR) spectroscopy. The use of FT, for both IR and Raman spectroscopy, allows fast spectra acquisitions and improved signal-to-noise ratio. The applicability of these techniques to food authenticity is widely demonstrated by the huge number of publications reporting their feasibility for several commodities, including edible oils, meat, dairy, alcoholic beverages, juices, tea and coffee, honey, herbs and spices (McGrath et al., 2018).\textsuperscript{15} In some specific cases, the advantages of applying data fusion strategies using complementary instrumental information for these techniques have been demonstrated.


3) Miscellaneous:

Sensor-based techniques and others: other systems based on instrumental techniques that emulate human sensory responses, including mainly electronic nose (e-nose) and electronic tongue (e-tongue) have been proposed for food authentication purposes. E-noses permit to analyse volatile compounds in gaseous phase, while e-tongues allow analysis of medium/low volatility components in liquid phase. These techniques, used alone or in combination, can generate a global fingerprint of foodstuffs with benefits in terms of simplicity, rapidity and cheapness. Different methods have been developed for the authentication of edible oil, cheese, honey, meat and tea.\textsuperscript{30} Ion Mobility Spectrometry is also a technique that is expanding its application fields demonstrating to be suitable for food authentication. The implementation of ion mobility in combination with GC has demonstrated to be successful for untargeted profiling of volatile compounds for the authentication of the botanical origins of honey combined with optimized chemometric techniques.\textsuperscript{31} In particular, ion mobility combined with mass spectrometry has great potential in metabolite and lipid profiling with applications to food authentication.

**Chemometrics tools**

Data analytical techniques should exploit state-of-the-art approaches, and in particular those developed inside the Machine Learning community, usually underexploited in the context addressed by the present guidelines.

Classification problems can be categorized into one of the following approaches\textsuperscript{1}:

One-class classification\textsuperscript{32,33}: the goal is to verify whether a sample is compatible or not with the characteristics of a single class of interest. One-class classifiers are predictor models that are learnt from a training set containing only samples of the class of interest (target class). A representative sampling is not strictly required for non-target classes. The approach allows one to label as “anomalies” those samples that are not classified into the target class. One-class approach should be adopted when the interest is focused on a single target class and the aim is to verify compliance of samples with the features of that class. One-class methods build an enclosed class space around the class samples. The shape of the class space depends on


\textsuperscript{33} Khan, S.S., Madden, M.G. (2014) One-class classification: taxonomy of study and review of techniques, The Knowledge Engineering Review 29, pp. 345-374
the particular method applied, while its size is a function of the confidence level that is selected a-priori by the user for the specific case. Samples that fall outside model boundaries are considered as non-compatible with the class of interest. Some of the most popular techniques for this task include: SIMCA (largely used in chemometric), density estimation techniques based on multivariate Gaussian, lazy learning based on k-Nearest-Neighbour, one-class support vector machines, as well as various clustering approaches.34

Binary classification: the goal is to verify whether a particular sample belongs to one of two classes of interest, usually called the positive and the negative class. Even if the one-class approach could be framed into the present one, the main difference stands in the fact that, in order to learn a binary classifier, we usually need a relevant number of examples of both classes. Several classifiers are binary “by nature”, since they are models able to discriminate only between two classes. Most popular binary classifiers include: logistic regression, support vector machines, binary Bayesian networks, Multi-Layer Perceptron with one output unit35.

Multiclass classification: the goal is to verify whether a particular sample belongs to one of three or more classes. Some classifiers can naturally deal with multiclass problems such as Bayesian networks, Multi-layer Perceptron with multiple output units, Decision trees, lazy learners based on k-Nearest-Neighbour. However, binary classifiers can be adapted to deal with multiclass problems as well, by considering one of the possible approaches36:

1) one-vs-all: if there are N>2 classes to be discriminated, a set of N binary classifiers is learnt, each one able to discriminate one class from all the others. The approach requires each classifier to return a confidence score for its decision; the final predicted class is the one for which the corresponding binary classifier outputs the highest score.

2) one-vs-one: if there are N>2 classes to be discriminated, a set of N(N-1)/2 binary classifiers is learnt, each one trained on one possible pair of classes; at prediction time every classifier is applied to the test sample and a majority voting decides the class.

Binary and multiclass approaches assign samples to one among a number of predefined classes (at least two) that are available during training (i.e. the training set contains samples representative of every class to be discriminated). In such approaches is fundamental that all of the classes are not only meaningfully defined but also sampled in a fully representative way.

Most conventional classifiers assume more or less equally balanced data classes and do not work well when any class is severely under-sampled or is completely absent. This may have consequences at several levels:\(^{37}\):

1) the performance measures should be suitably considered, since in case of unbalanced datasets simple accuracy may not be significant (e.g., in a dataset of 1000 samples having 998 positive samples and only 2 negative samples, a binary classifier always predicting the positive class would provide a 99.8% accuracy); in this case measures like precision and recall (and strictly connected measures like F-score, AUC or area under PR curve) should be taken into account for the evaluation;

2) the over-sampled classes could bias the classifier towards them, and consequently it may be really difficult to recognize test samples of the under-sampled classes. To avoid this, a cost-based learning approach can be adopted. It means that under-sampled classes are weighted more than over-sampled ones, as to simulate a certain number of replica for the formers, in order to get the dataset balanced.

SECTION 2
GENERAL PROPOSED APPROACH

The process of developing and validating a non-targeted method can be divided into four main phases (Figure 1):

1. Definition of the scope
2. Feasibility study: performing a pilot study to test the operative conditions of the chosen analytical protocol
3. Method development: collection of the analytical data and selection of the best performing classification model

After having defined the problem to be solved and having set technique and classification approach to be used (Scope definition), the study has to be opportunely planned by defining samples, sampling procedures and sample preparation. The best operative conditions can be identified through a Feasibility study, in which an exploratory statistical analysis should give first indications about the fitness of the chosen method. The method, initially tested on a small subset of representative samples, should thus be developed (Method development) by
analysing the entire sample set (model set initially defined by the applicability limits). Obtained raw non-targeted data need then to be opportunely pre-treated (data pre-processing), in order to obtain a new tidy data matrix useful to develop the best classification model. Finally, the model needs to be firstly internally validated through statistical tools and subsequently challenged with a new independent validation set (Method Validation).

A workflow schematizing the key steps aiming at assisting analysts on how to correctly undertake a non-targeted analysis for food authenticity studies is proposed in figure 2. After each validation phase, performance evaluation has to be provided, in order to decide whether the method is fit-for-purpose or it has to be reviewed, either for the internal validation process (failure of external validation) or for an entire re-development (failure of internal validation).

![Validation workflow](image)

*Figure 2. Proposed workflow.*

*Full arrows define the workflow suggested in case of positive evaluation of the performances; dashed arrows indicate alternative workflow in case performances are not acceptable and the method needs re-validation (starting over from the method development) or re-development (starting over from the feasibility study).*
In the following sections, the steps of this general approach are specifically described and commented.

**SCOPE DEFINITION**

Defining the scope is fundamental and has a strong impact on the protocol that follows. Main questions to be answered are:

- **Why?** What is the problem? Studies can be focused on several aspects, such as the evaluation of the food integrity, the assessment of its origin (geographical, botanical, etc.), the evaluation of its “production” procedure (e.g. biological vs non-biological, fished vs farmed, aging protocol, etc.), and more.

  What is the purpose? The final purpose of the method to be validated can vary, it could be used as a proof-of-principle of the classification strategy, or as decision tool on real-life screening samples, where the result of the classification needs to be used to take legal action. The purpose has a strong impact on the level of acceptability of the classification method.

- **What?** What is the target (type of commodity)? Foods can be solid, liquid, fresh, processed, preserved, consisting of a single ingredient, an admixture, etc. Analysts should be well aware (or ask for the support of an expert in the field) of all the variables affecting the composition of the food commodity, and of their influence on the classification to be made.

- **How?** How will the samples be analysed? Decision on which analytical technique is the most promising to achieve the expected classification results should be based as much as possible on the existing literature, as well as on the skills of the analyst and of the instrument availability.

  Which classification approach should be followed? One-class vs multiclass. The classification approach influences directly the sampling. If the interest is characterizing a single class (e.g. only the pure food, only one geographical region, only one agronomical practice), then all samples not belonging to that particular class should be considered equally in the study. However, if more than one class is under investigation, all classes should be extensively and equally represented in the sampling, while samples not belonging to these classes should not be considered.

- **Where?** What are the limits of applicability? They depend on the variability of the samples collected for the development of the method. All the classes, groups and categories present within the food commodity are variables that can affect its composition and chemical profile (e.g. origin, seasonality, number of varieties, taxonomy, year of production, agronomical practices, post-harvest conditions, industrial processing techniques, etc.), and should therefore be carefully taken into account in the experimental design. The validated method will be applicable only to those future samples that will fall into the classes, groups and categories represented in the samples set.

- **How much?** What are the Accepted performances? They depend on the purpose and on the level of certainty required. In a classification type of problem, the performances of the classifiers refer to their ability to correctly classify an object, related to the probability of
making a mistake in its classification. If the validated method needs to be used to take legal action based on its results, the predictive performance should be higher than if the method is used just for screening. The level of acceptance should be initially set by the analyst based on the existing knowledge or on his/her experience. Moreover, the weight of possible errors in the classification should be carefully evaluated and set in this phase: different types of error (false positive vs. false negative) might in fact have a different impact on the final outcome. Finally, the decision on whether to accept the classification results should be also based on how the probability is distributed within a given class (typically in a decision tool-type of method). Different levels of acceptance (accuracy, specificity, sensibility, probability, other) should be therefore set in advance.

FEASIBILITY STUDY

A feasibility study addresses the likelihood that a method will be able to achieve its intended purpose. It is a preliminary phase to the method development useful to identify the best operative conditions and, based on the problem to be solved (scope of the method), to investigate if the chosen conditions are promising for the future development of the method itself.

**Definition of the sampling procedure.** Depending on the food product submitted to investigation, instructions on collection operations and storage conditions (from the collection site to the laboratory) over the period before the analysis must be defined.

**Sampling of the small subset.** For the collection of a small subset (i.e. at least 10% of the sample set used for the method development) refer to the following section “Samples collection and storage”: as a general rule, the subset should be a fair representation of the sample chosen for the model development.

**Sample pre-treatment.** Sample preparation should reflect the type of analysis that the food material will undergo. In general, for non-targeted methods sample pre-treatment should be kept minimal, therefore extraction protocols should aim at extracting the majority of compounds to maximize the number of signals and to be deemed representative of the food matrix under analysis. It is important that the amount of food to be extracted is compatible with the type of downstream analysis. Sample preparation should be of broad coverage (metabolites, lipids or proteins) to be suitable for non-targeted analysis. It will typically include a first step of grinding or milling (for solid samples) to increase homogeneity. In case of mass spectrometry or chromatographic methods, extraction of either the polar or the apolar

---

fraction will guide towards the choice of the best extractive solution. In case of vibrational spectroscopy methods, samples do not usually need to undergo an extraction step. In contrast, NMR methods might include a quick preparation step, usually by adjusting the pH of the sample (e.g. wines) or, in case of signal suppression caused by water, by replacing common solvents with more compatible ones.

**Development of the analytical protocol.** The analytical procedure (fit for the purpose, considering available knowledge of the problem, instrumental availability and costs) and its best operative conditions (extractions protocols, settings of the instrument) should be defined in this phase. Once the operative conditions are set, a defined number of replicates must be produced from a selection of samples, in order to study the repeatability and the reproducibility of the analytical protocol. This will allow gaining insights into the variability of the analytical protocol. Indicatively, the protocol should introduce less variability than the natural one occurring between the independent samples. A strategy to estimate the variability should be defined according to the structure of the output data. The variability (standard deviation) calculated for the chosen number of replicates must be lower than 30%.

**Exploratory data analysis and protocol evaluation.** An exploratory data analysis that would provide an indication whether the analytical protocol is sufficiently promising for the purpose should be performed. In this phase also classical unsupervised chemometric techniques as Principal Component Analysis (PCA) can be applied. A specific example is reported in the Annex 1.

**METHOD DEVELOPMENT**

Development of a non-targeted method is typically underpinned by the employment of an analytical technique that provides a detailed profile of the sample under analysis. Although the profile can vary depending on the type of analysis, the implementation of advanced statistical tools allows comparing the profile of samples to a library of profiles gathered over the time and representing the authentic food. The application of such model allows classifying the unknown samples with a certain degree of uncertainty according to the acceptance criteria set during method development.

**Samples collection and storage.** When choosing a sample within a total population, we agree that we will get only an estimate of the indicator of interest, and not the true population value. One of the crucial points when developing a non-targeted method is to have a sufficient representative number of unique samples to cover all variations associated with the commodity of interest. All variables possibly affecting the chemical fingerprint of the food commodity (size, seasonality, cultivar/breed, area of production, cultivation/feeding process, ripening degree, age at the slaughtering, harvesting procedure, post-harvest conditions), should be taken into account and included in underlying groups of the class(es) of interest. This variability within each class needs to be maximized, in order to avoid that, in the future,
a new sample belonging to that class will not be recognized by the trained algorithm. Considered the myriad of possible cases, it is hence impossible to suggest an average or even a minimum number of samples to be included in the study. According to the recent scientific opinion issued by Mc Grath et al., 2018, the minimum number of representative samples recommended for a non-targeted method under development is of approximately 200 unique samples collected for each class to build up a robust model. However, the Authors of this document warn the reader against trusting a fixed general number. Several publications on the theory of sampling and on how to correctly calculate sample size are available in the literature, the review of which is outside the scope of this document. Generally, two very important estimate’s characteristics need to be bear in mind:

- **the accuracy**: it refers to how close the estimate is likely to be to the true value. It is a measure of the validity of the estimate, and a lack of it is related to **bias**, which reflects the difference between the sample estimate and the true population value due to error in measurement, selection of a non-representative sample or factors other than sample size;
- **the precision**: it refers to how similar will be the results if several measurements are conducted with the same methodology. It is a measure of the consistency of the estimate, and it is related to the **sampling error**, which is the difference between the sample estimate and the population value due to measuring a sample instead of the whole population (uncertainty).

Sampling error can be controlled by acting on the sample size. However, a larger sample size increases the precision but does not guarantee the absence of bias! A derived parameter, often included in the proposed equations for calculating the sample size, is the confidence interval, which depends on the chosen level of confidence. Other parameters, such as the expected size of the population, its standard deviation, or the proportion between the classes of interest, could be retrieved from available prevalence studies or import/export statistics. Once these values are known (or guessed), it is possible to have an estimate of the sample size according to the desired level of confidence. If such estimate is too far larger than the acceptable one (according to practical reasons, such as time and costs per analysis), it is possible to narrow the confidence interval to obtain a much lower estimate, provided that such action is compatible with the degree of precision expected at the end. It should be kept in mind in fact that the precision of the results will drop, increasing the chance that a new sample unknown to the system will not be properly classified.

Moreover, the classification approach has also an impact on the definition of the sample. If more than one class is considered, samples representing the classes should be balanced in number, as much as possible, and all classes should be internally balanced according to the

---

relevant underlying groups. If it is a one-class problem, test samples not belonging to that class should be collected wisely, bearing in mind what is the real life situation. If it is an adulteration problem, artificially contaminated samples should be included, again bearing in mind what is the real market situation, both in terms of the nature and in terms of the amount of the adulterating agent(s).

The choice of sampling method is also important, and should be directly related to the objectives of the study. Moreover, in order to obtain robust and reliable methods, it is mandatory to obtain authentic samples with full traceability for model development. Once the samples have arrived to the laboratory, they should be immediately labelled and catalogued, keeping records of each individual sample and annotating their physical conditions. Each defect or lack of compliance with the expected, likewise variations in the temperature of the samples, must be noted, and those samples eventually removed. A traceability system should be put in place to keep records of the entering samples, including the name of the operator handling the sample. Storage of the samples should be properly done according to their kind. In case of perishable food materials (e.g. fresh meat or fish), samples should be aliquoted as quickly as possible and stored under refrigerated conditions (typically -20°C or -80°C), depending on the assessed stability of the food or on the estimated storage period before analysis. When dealing with more thermostable foods (e.g. dried food materials like spices, coffee or honey), they should be preferably stored at their usual storage conditions and, in any case, under conditions preserving their quality and characteristics at best (for example, room temperature, assuring the right conditions of humidity and light exposition).

**Analytical phase.** This section refers to the steps following sample pre-treatment (previously described in the Feasibility Study section) and relates to the sample itself, the instrument and method parameters that will affect the final outcome of the experiment. Sources of variations in this phase include: the apparatus, system stability, temperature, eventual sources of contaminations during the analysis, and also batch effect, matrix effect and carry over depending on the specific technique being considered. Since every analytical technique has its own characteristics, this guidance will not enter into the description of the details of each and every one. Analysts approaching this type of study should be aware of how to operate their instrument, and analyses should be carried out based on the indications obtained during method development. However, the incorporation of quality control samples (QCs) is highly recommended to i) ensure that reproducible data are acquired along the analysis, ii) provide data to calculate technical precision within each analytical block (considering appropriate a maximum CV of 20% for HPLC-MS analysis), iii) allow signal correction within and between analytical blocks. QCs should be representative of the qualitative and quantitative composition of the study samples. Different types of QCs can be enrolled in the study: pooled samples, surrogates, commercially available standards. Generally, the use of pooled QC
samples deriving from all the subjects of the set is highly recommended in case a huge number of metabolic features are analysed. The choice of the type of QCs should be appropriately done during feasibility study. The implementation of QCs allows controlling the stability of the analytical measurements. It enables the quick detection of abnormalities and thus allows the fast correction to obtain reliable measurements.

Multiple analysis of the same QC sample should be carried out and typically ≥ 3 multiple and consecutive analysis from the same batch are recommended. Likewise, each sample should be analysed in ≥ 3 technical replicates to be able to statistically process the final dataset. The results obtained from the QCs should be monitored and checked for systematic and random errors. In general, the use of QCs turns useful especially in long term metabolomics studies, where a periodic analysis of a biological QC sample together with the subject samples is required. At the end of the experimental analysis, each metabolic feature is normalised to each QC sample set. A correction curve for the whole analysis is then interpolated, to which the total data set for that feature is normalized. The plot of the QCs should be inspected in real time to check the quality of the analytical measurements. When a significant and constant variation of the signal is detected over time, the analysis can be adjusted by applying a corrective factor, to avoid losing the acquired data. As a general rule, once a consistent deviation is noticed in the QCs plot, the analysis should be stopped, and all the apparatus and analytical conditions should be carefully checked, eventually re-calibrating the instrument if necessary before starting a new analysis.

QCs are commonly utilised in LC and GC-MS metabolomics studies enabling signal drift correction that may occur within and between analytical batches. This will allow increasing the accuracy of the measurement. A representative number of QCs should be typically injected at the start, the end and at regular intervals throughout the analysis every approximately 5 up to 10 runs (total QCs about 20-30% of the total number of investigated samples).²

**Data pre-processing.** Once raw data are produced, several steps of data pre-processing (e.g. centroiding, binning, noise reduction, light scatter correction, baseline correction, alignment, feature detection, normalization, scaling, transformation, smoothing) should be considered, the order and the number of which depend on the analytical technique and on the nature of the data. Raw data might be treated either inside or outside the vendor software. Generally speaking, the chosen pre-processing procedures will impact the classification outcomes. During data pre-processing, two main aspects should be carefully monitored by using exploratory analyses: the performances of the quality controls, expected to cluster close to each other, and the percentage of relative standard deviation of each feature across the QC samples. In particular, with reference to large scale studies, when data external to the main data set are used as QCs, specific strategies such as data filtering and signal correction based on QC measurements should be properly designed. For example, it could be advised to filter
the data choosing the signals present in 80% of a sample class, and showing a relative standard deviation (RSD) lower than 20 or 30% in the respective QCs (whether based on HPLC-MS or GC-MS analysis). Also, for all data that pass filters, it should be clearly specified how the missing values are calculated to avoid altering the final outcomes of the statistical analysis.  

Finally, a tidy data matrix, that will ease the statistical analysis, should be built. In a tidy data matrix samples are reported in each row, while single features are reported in each column. As a result of data pre-processing, the dimension of the data matrix should be reported as model input information.  

**Data analysis.** Different statistical tools should be tested in parallel, in order to be able to compare and criticize the outcomes. For major details, see the sections below.

### METHOD VALIDATION

After having defined the best analytical approaches (feasibility study and method development), the method must be validated in order to confirm its effectiveness for its intended use (*fit-for-purpose*).

In this guideline we suggest that the validation process should include two distinctive phases. The first step, or **internal validation**, must be performed on known representative samples, using data sets created as previously described. Independently from the selected classification approach (one- or multi-class), the main aim of the internal validation is to generate and fix suitable models to be adopted for the external validation, through which the developed method will be challenged. In the second step of the process (**external validation**), a new set of “blind” samples (i.e. samples unknown to the classifiers, but whose origin is known to the analyst), potentially bringing more variability respect to the previous set, must be collected and analysed, and the prediction ability of the developed method have to be further tested and verified.

Following each validation step, the performances of the method must be evaluated through opportune validation parameters (see sections below for more details), in order to verify that the acceptance thresholds set at the beginning of the study are satisfied, and to confirm that the method is *fit-for-purpose*.

In the scheme of **Figure 3** the decision tree concerning the validation process is represented.

---

Internal validation and critical evaluation of the results.
An internal statistical validation (e.g. cross-validation) must be performed. Cross-Validation (CV) has been chosen as a key point for validation, since it can provide an indication about how well the model could perform on unseen data, by basically using a part of the dataset for training the classifier and the remaining part for validating it. Cross-validation can be performed through different techniques, with the possibility (or not) to iterate on different combinations of training/validation splits, in order to reuse the same data both for training and for validating the classifier. The main cross-validation techniques are listed in the following:
- **Holdout**: this strategy aims to split the dataset into two different and independent datasets, labelled as training and validation datasets. This split can be done selecting a percentage of 60(training)/40(validation) or 70/30 or 80/20 split, for example, and ensuring that the distribution of the data in the different classes is respected both in the training and in the validation datasets (stratification). The training set is then used for training the classifier, while the validation set is tested, to provide an independent measure of the performance of the trained classifier. The holdout measure is computationally efficient, since the dataset is used only once for training and validation, but the obtained classifier performance strongly depends on the way the samples have been chosen during the split operations. Different choices lead to different validation results.

- **K-Fold CV**: the dataset is divided in K different subsets and the hold-out strategy is repeated K times, ensuring that, in each iteration, a different subset is used as validation and the remaining part of dataset as training. Stratification can also be applied to ensure that the same distribution in classes is also reflected in each of the K subsets (stratified K-Fold CV). The advantage of this method is that we use all the dataset both for training and for validation, and the classifier performance is the mean average of each iteration, reducing the bias caused by the “one-shot” choice of the samples of the Holdout technique. On the other hand, the K-Fold CV is computationally less efficient than Holdout.

- **Random subsamples**: similar to K-fold strategy, it selects, in each iteration, some samples for validating and some samples for training, in a randomized way. Advantages over the K-Fold CV are that one can freely decide the number of iterations and can also decide the length of each training and validation sets. On the other hand, the sampling is done without replacement, therefore there is a chance that some samples could never be selected in the validation set, while others may be selected more than once (this strategy is a non-exhaustive).

- **Leave-One-Out CV**: it can be seen as a particular case of K-Fold CV where, for each iteration, one sample is selected for validation and all the remaining N-1 (being N the number of samples in the dataset) samples are used for training. This process is then repeated for N iterations, selecting each time a different sample for validation. This technique has the advantage of using all the dataset both for training and for validation and of being an exhaustive strategy, because it trains and tests all possible ways, but is it also computationally less efficient than the two techniques above.

The choice of the CV strategy can be guided by balancing the numerosity of the dataset and the advantages of the CV techniques, with respect to their computational impact. In general, the K-Fold CV can be considered as a feasible compromise between advantages and computational efficiency, and as a starting point to decide if it is better to move to more or to less efficient techniques, gaining or losing some of the relative advantages.

The statistical analyses are performed on the dataset prepared for the internal validation, after a first step of feature selection. Feature selection is a supervised data analysis technique,
with the aim to automatically select the most relevant sub-set of features (columns in the data matrix), for training the classifiers. Feature selection is an important step, since it reduces the computational complexity while, at the same time, improving the reliability of the classifier by removing noise and redundant information and tackling the problem of overfitting, which generally arises when the number of features is much higher than the number of samples. Feature selection can be performed, for example, with correlation-based algorithms\textsuperscript{42}, which evaluate the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them and their correlation with respect to the class, or searching for the best feature subset using genetic algorithms\textsuperscript{43}. The importance of feature selection as a pre-processing step to classification is much more important in applications like the one described here, where the number of instances is often limited, in contrast to the potential number of features that may be potentially very large. To avoid overfitting problems, the number of feature must be reduced in a principled way; since we are in a supervised learning context, this can be done by exploiting information about the correct class of each instance available in the dataset.

The dataset, after being reduced through feature selection, is then used to train and to validate a set of classifiers among, for example:

- Soft independent modelling by class analogy (SIMCA);
- Partial least squares and Discriminant analysis (PLS, DA);
- Linear classifiers, e.g. logistic regression\textsuperscript{44} or Naïve Bayes Classifier\textsuperscript{45};
- Bayesian Networks with different learning algorithm, e.g. K2\textsuperscript{46} or TAN\textsuperscript{47};
- Lazy algorithms, e.g. K Nearest Neighbor\textsuperscript{48};
- Decision trees, e.g. C4.5 trees\textsuperscript{49};
- Random Forest\textsuperscript{50};

Mathematical functions, such as Multi Layer Perceptrons\textsuperscript{51} or Support Vector Machines\textsuperscript{52}. All the classifiers are trained and validated on this dataset, searching for the classifier setting, for each classifier (e.g. the number of parents per node in a Bayesian Network, or the number of hidden layers in a Multilayer Perceptron), allowing it to obtain its best validation

performance. Therefore, at the end of this series of statistical analyses, a set of classifiers are obtained, each of them performing at its best on the provided dataset. Once the statistical analysis is validated, analyst must accept/reject the method based on the comparison between the performances of the used classifiers, in terms of the validation parameters outcomes provided (see next section) and the starting level of acceptance set. The choice regarding both the kind of validation parameters and of their relative acceptance values should be the result of the cooperation between all specialists involved in the study (food commodity specialists, laboratory analysts, mathematicians), who will share their knowledge and respective expertise, focusing on the tackled problem and on the goals they aim to achieve. All classifiers performing as or better than the predetermined levels of acceptance should be considered for the next step of validation. If the method is rejected, counteractions are suggested in the proposed decision tree. Firstly, a revision of the pre-processing step is advised: a newly generated data matrix (newly sized) might lead to better classification results without the need of adding new samples. Finally, if the existing data matrix cannot be modified or statistically processed in a different way to improve the classifiers performances, adding new samples and more analytical data might improve the classification outcomes and allow reaching the minimum levels of acceptance. However, this might not always be the case; for example, when the analytical method is unable to generate features allowing for samples separation, or when the analytical method generates too much noise, and feature selection is unable to mitigate the problem, or when the method creates a dataset generating overfitting, the addition of new samples will most likely worsen the validation performance instead of improving it.

**External validation.** A new set of samples must be collected and analysed, and the prediction ability of the developed method have to be further tested for the final validation.

The goal of the external validation is twofold: (1) to further validate each classifier using an external dataset, in order to verify that the performances obtained during the internal validation phase are maintained when new and previously unseen samples are provided, and (2) to consider validated and reliable the only classifier models passing this further test.

To perform the external validation, the developed prediction model (classifiers selected during the internal validation step) must be challenged with a new set of test samples, which will be analysed following the protocol previously set up (analytical + chemometrics). For the collection of the external validation set, new known samples not belonging to the previous set, but sharing the same or, if possible, higher representativeness, should be chosen. The size of the validation set will depend upon that of the previous set. Even if it could be theoretically slightly smaller, ideally it should be equal or larger than the previous set.

The performances of the classifiers must be equal or higher than the levels of acceptance previously set for the external validation acceptance. All the classifiers presenting performances equal or higher than the acceptance thresholds can be considered validated. If
the performances are lower, the method should be declared failed. However, as a corrective action, merging the validation set into the model set and recalculating a new prediction model is suggested, in order to improve the representativeness of the model set. In this case, if the newly developed model is accepted, a new sample set for a further external validation must be collected. If the performances of the classifiers are still lower than the defined levels of acceptance, the proposed method should be definitively declared failed. In this latter case, an alternative method could be considered and tested for the validation.

Validation parameters

This section presents the main validation parameters and measures suggested for assessing the overall quality of the developed method (intended as the analytical procedure coupled to the classification approaches), in terms of their ability in recognizing the correct class labels assigned to the analyzed food samples.

Considering that untargeted method are mainly used to solve classification problems related to food authenticity issues (identification of origin, intended in the widest meaning of the term; adulteration), we have drawn from specific literature the most used parameters able to measure the performances of classification approaches, which were also adaptable to different chemometric techniques (regression models, algorithms, neural networks, ...) and classification approaches (one-class and multi-class).

These parameters rely on the definition of: positive samples ($P$, samples belonging to one or more classes of interest) and negative samples ($N$, all the other samples). The attempt of one classifier to identify them, by assigning labels starting from the samples’ features, generates 4 different results:

- **True Positives (TP):** the number of the positive samples that were correctly labeled by the classifier
- **True Negatives (TN):** the number of negative samples that were correctly labeled by the classifier
- **False Positives (FP):** the number of negative samples that were incorrectly labeled as positive
- **False Negatives (FN):** the number of positive samples that were mislabeled as negative

These terms are summarized in a confusion matrix, shown in Figure 4.

![Confusion matrix, with totals for positive and negative samples.](image)
The confusion matrix is an important information about the validation, since it can show the picture about how many samples have been classified correctly ($TP$ out of $P$, and $TN$ out of $N$) and how many have been misclassified (the remaining $FN$ and $FP$), for each class. The table may have additional rows or columns to provide totals. For example, in the confusion matrix of Figure 1, $P$ and $N$ are shown. In addition, $P'$ is the number of samples that were labeled as positive ($TP + FP$) and $N'$ is the number of samples that were labeled as negative ($TN + FN$). The total number of samples is $TP + TN + FP + TN$, or $P + N$, or $P' + N'$. Note that, although the confusion matrix shown is for a binary classification problem, confusion matrices can be easily drawn for multiple classes in a similar manner.

On the basis of this matrix, a number of evaluation measures are obtained, and used as validation parameters. As first,

- **Accuracy rate** $= \frac{TP + TN}{P + N}$: reflects how well the classifier recognizes samples of the various classes;

- **Kappa statistic**: $K = \frac{Pr(a) - Pr(e)}{1 - Pr(e)}$, $Pr(a) = \frac{TP + TN}{P + N}$, $Pr(e) = \frac{P' \cdot P + N' \cdot N}{(P + N)^2}$ $^{53}$ measures inter-rater agreement for qualitative (categorical) items, and can be seen as a measure for assessing accuracy and reliability of a statistical classification. When there is no agreement other than that which would be expected by chance $K$ is zero. When there is total agreement, $K$ is one;

In case of class unbalance, which occurs when the classes of interest in a dataset are not equally represented, appropriate measures should be taken. In fraud detection, for example, the class, generally speaking, “fraud” occurs much less frequently than the generally defined “non-fraudulent” class. In these cases, if a multi-class classification approach is used, it is important to concentrate on validation parameters able to manage the unbalanced distribution of the samples, assessing how well the classifier can recognize the positive samples, and how well it can recognize the negative samples. Sensitivity and specificity per class, beside the area under the ROC curve, are parameters useful for overcoming these limitations and more appropriately measure the performances of the classification model.

- **Precision** $= \frac{TP}{TP + FP}$: measures the “exactness” of the prediction, in terms of what percentage of samples labeled as positive are actually such.

---

• **Sensitivity or Recall or Non-error rate** \( TP \): also defined as true positive rate, measures the “completeness” of the prediction, which is the proportion of the positive samples correctly identified;

• **Specificity** \( TN \): also defined as true negative rate, is the proportion of the negative samples correctly identified;

• **Fall-out** \( FP \): also defined as false positive rate, is the proportion of negative examples incorrectly classified. It is equal to 1-specificity;

• **ROC (Receiving Operating Characteristic) Curve**: is the plot of the sensitivity of a classifier as a function of the fall-out (1-specificity); in other terms, the curve obtained by plotting and joining a set of points \((fpr,tpr)\) where \(fpr\) is the false positive rate and \(tpr\) is the true positive rate of a classifier. Different curves can be obtained by changing the classification parameters of the model (e.g., the threshold for classifying the positive class);

• **Area under ROC curve**\(^{54}\): it is the area of the region under the ROC curve; it can be interpreted as the probability that the model ranks a random positive example more highly than a random negative example. An area of 1 represents a perfect test; an area of .5 represents a worthless test, while the traditional academic point system defines the interpretation of this parameter through the following thresholds:
  o .90-1 = excellent (A)
  o .80-.90 = good (B)
  o .70-.80 = fair (C)
  o .60-.70 = poor (D)
  o .50-.60 = fail (F).

**Accuracy rate, Precision, Sensitivity** and **Specificity** can be expressed in the range 0-1 or as percentage values (0-100%). For reference about these parameters, see\(^{55}\).

The following are aggregate measures, summarizing different validation parameters for a further inspection of the classifiers performances:

- **F-measure:** \( F_1 = 2 \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \): represents the harmonic mean of precision and recall. The \( F_1 \) score reaches its best value at 1 (perfect precision and recall) and worst at 0.

- **Matthews correlation coefficient:** \( \text{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \): to be considered in particular for measuring the performance of binary classifications. MCC measures the correlation between the observed and predicted classifications. MCC has a range of -1 to 1, where -1 indicates a completely wrong binary classifier, while 1 indicates a completely correct binary classifier.

Additionally, cost-based validation measures should be considered. In fact, in some cases there is the necessity to individuate different weight for the errors, not only because the datasets can be unbalanced, but also because different types of error (e.g., false positive vs false negative) might have a different impact on the final outcome, depending on the specific purpose of the study. In these cases, cost-based validation measures should be considered. The costs (or weights) can be assigned on the basis of the balance of each class or based on the specific knowledge of the application field (assigning a higher cost to misclassification of the negative samples, for example), or combining both of these aspects.

Cost-based evaluation relies on a cost matrix, assigning, for each class \( i \), the cost of predicting a sample belonging to \( i \), to the same class \( i \), or to a different class \( j \neq i \). A cost matrix \( C \) has the following entries \( C(i,j) = \text{cost of assigning an element of class } i \text{ to class } j \) (with positive cost for misclassification, i.e. non diagonal entries)

<table>
<thead>
<tr>
<th>ACTUAL CLASS</th>
<th>PREDICTED CLASS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Class=Yes</td>
</tr>
<tr>
<td>Class=Yes</td>
<td>( C(\text{Yes</td>
</tr>
<tr>
<td>Class=No</td>
<td>( C(\text{Yes</td>
</tr>
</tbody>
</table>

*Figure 5. A cost matrix for binary classification.*

---


As previously defined, the costs can be assigned on the basis of the cardinality of each class or exploiting knowledge about the application domain or combining both of these aspects.

Additionally, a-priori probabilities $P_g$ can be set for each class $g$, in a uniform way ($P_g = 1/G$, for each class $g$, with $G = \text{no. of classes}$) or using the cardinalities of each class (e.g. for a binary classifier $P_{\text{yes}} = P / (P+N); P_{\text{no}} = N / (P+N)$).

After assigning costs and probabilities, the cost-based misclassification risk for each class $g$ can be calculated as:

$$MR_g = \frac{\sum g, CM_{gg'}CS_{gg'}}{N_g} \cdot P_g$$

In this formula, $CM$ represents the confusion matrix (see Figure 4), $CS$ represents the cost matrix (see Figure 5) and $g$ is the considered class. Given $g$, $CM_{gg'}$ is the number of samples of class $g$ classified as $g'$. For example, in a binary classification, $CM_{\text{yes,yes}}$ is the number of samples belonging to the actual class $\text{yes}$, correctly classified as $\text{yes}$ (TP), while $CM_{\text{yes,no}}$ is the number of samples belonging to the actual class $\text{yes}$ incorrectly classified as $\text{no}$ (FN). $CS_{gg'}$ is the cost value in the cost matrix, giving the cost of assigning an element of class $g$ to class $g'$. For example, in binary classification, $CS_{\text{yes,no}}$ is the value in the cost matrix for assigning a sample belonging to the actual class $\text{yes}$ the incorrect class $\text{no}$. Finally, $P_g$ is the a-priori probability for class $g$ and $N_g$ is the number of samples in the class $g$. The total misclassification risk is the sum of all the risks for each class.

As final aspect, because of the probabilistic nature of the classification approaches (a classifier assigns a class to each unknown sample on the basis of a probability value), we suggest to include, among the other validation parameters, the probability assigned to the samples to belong to a specific class; this parameter is a measure of the confidence through which the correct classifications and/or misclassifications are performed. However, due to the high number of samples expected to be tested, it can be unpractical to report the probability assigned for all the samples classified; thus, in order to overcome this limitation, we suggest at least calculating and report the mean and variance of the probabilities determined for each sample typology ($TP, TN, FP$ and $FN$).

All of these validation parameters must be carefully chosen for both the internal and external validation, and have to be produced for each classification model tested.

**Acceptance criteria for validation**

After having defined the most appropriated validation parameters, the main question to be answered is “what values are accepted to consider the method validated?” Acceptance values for each parameter (or at least for those parameters considered as the most critical based on
the specifics aims of the study designers) should be clearly defined and declared a priori, taking into account the purpose of the study and the level of uncertainty eventually admitted. However, it is difficult to generalize and indicate specific thresholds to reach. The level of acceptance should be initially set by the analysts based on the existing knowledge or on their experience, depending on the specific purpose of the study. For example, if the validated method needs to be used to classify unknown samples (designation of the origin, identification of adulterated samples,...), the predictive performance should be higher than if the method is used just for preliminary screening (for example in the case of identification of “suspected samples” to be further analyzed by targeted methods for confirmations).

Accuracy rate, precision and sensitivity rate are usually considered important parameters, and high values are generally associated to good performances. However, in case for instance high specificity is sought (e.g. in an adulteration type of study one might be interested in minimizing the number of false positive), the consequence might be a decrease for these three parameters associated values. This is nicely represented by the receiver operating characteristic (ROC) curve.

![Figure 6. ROC curve](image)

Furthermore, it is known from Machine Learning that, in general, classifiers perform better during the internal validation with respect to the external validation. For this reason, it can be suggested that for the internal validation (generally performed through the cross-validation technique) the selected parameters could have more restrictive thresholds in respect to those set for the external validation step (performed on blind samples, generally expressing more variance respect to that expressed by the internal validation sample set). A specific example is presented in Annex 1; considering that the purpose of the study was to develop a preliminary non targeted LC-LRMS method able to discriminate the origin of Acacia honey, but

---

not allowing the identification of unknown samples, we have set Accuracy and Sensitivity values at least of 95% in order to accept the internal validation, and values >75% to accept the external validation. These thresholds appeared satisfactory to validate a preliminary method, the aim of which was limited to confirm its adequateness for future implementation (considering other variables as seasonality, manufacturing process, packaging, storage and so on). On the other hand, as a general rule, we can additionally suggest that methods aimed to recognize unknown samples should present definitively higher accuracy even after the external validation phase, for example setting the level at least at 95%.

Report from validation process

Another fundamental aspect that should be carefully considered is how to report the validation performances and the results. A validation report should be always accompanied by additional information concerning method specifications, including its specific purpose, the main characteristics of the samples and the sampling procedures, the detail of the analytical method, the description of the pre-processing procedures and of the resultant dataset(s), the detailed description of the classification approaches. All these aspects define the application limits of the method, and the validation parameters, when the acceptance levels are reached, are verified within these limits of acceptability.

The method should be reproducible in its principles, although not necessarily reaching the same performances.

The validation report should not only detail the results of the validation process, but also information identifying the specific application field of the validated method.59

Typical information in a validation report should include:

1. The main purpose of the study and the applicability limits of the method;

2. The description of the samples, identifying the main variables considered during the collection of the samples;

3. The description of the analytical method, defining sample preparation, instruments, operative conditions, preparation of QC samples, etc.;

4. The description of pre-processing methods applied to raw data and the description of the resulting datasets;

5. The validation procedure, including type of classifiers employed, of tests run and the description of the acceptance criteria;

6. The results and their interpretation.

HOW TO KEEP THE METHOD UPDATED

Once the validation process is completed, the method is valid only within the pre-set limits of applicability. However, to maintain it useful and eventually broaden its limits, it needs to be constantly monitored and regularly updated. While monitoring does not involve a change in the structure of the data matrix used to build the method, updating the system requires revalidation.

Monitoring

Over time different sources of variations could reduce the reliability of the validated method. Such variations could be represented by internal factors, like the natural mechanical deterioration of the instrument, or the change of the consumable supplier. For all these cases, standards are needed to verify that the performances of the validated method remain acceptable. Moreover, other external sources of variation might occur. When dealing with classification problems, it is possible that new individual samples, somehow different from those employed to build the method (either due to natural environmental changes or to human intervention in the food processing), though still belonging to the class of interest, show up on the market. In this case there is a chance that the validated method will struggle in recognizing these new samples, which chemical profile presents features unknown to the algorithm, classifying them as (false) negative. This type of error has nothing to do with the performances of the method, it is related to its accuracy, namely its likelihood to the true value. The method is biased by the difference between what was represented in the sample and the new reality, which changed. Because the appearance of such new challenging samples cannot be predicted, validated methods need to be constantly monitored.

Monitoring is practiced by testing known samples (that are blind for the classifier) at a chosen frequency and analysing the output for possible misclassifications. A number of new representative samples should be regularly collected, stored, processed and measured at the chosen frequency.

During monitoring, two main types of known samples should be analysed:

1. old samples belonging to the model set and QCs: the stability of the method is assessed by re-measuring the samples already analysed during a defined time. Samples
employed for this purpose need to be stable over time, maintaining their chemical fingerprint intact;

2. new samples (with a known origin) that fall into the applicability limits of the method: these samples should be correctly classified. In case they are not, all incorrect classifications must be registered and possible explanations for the misclassification should be investigated. The samples correctly classified should be used for following method updates.

Moreover, it is also advisable to collect and test new samples that do not fall into the applicability limits of the method: examples of this kind of sample would be individuals belonging to the class of interest but coming from a subgroup (e.g. a different region, harvest practice, feeding program) that was not considered in the primary sampling plan. These samples might or might not be correctly classified, but all classification results should be registered. These samples, collected during the monitoring, might be useful during the method updating: indications coming from the classification outcomes of these samples will be used to modify the limits of applicability, or to suggest the exclusion of a specific category.

All data must be processed and analysed following the validated method. Incorrect classifications should be recorded and statistically analysed, to verify the current state of the method and to monitor systematic trends in the method performance. In case old samples are misclassified, statistically significant shifts compared to the historical measures will trigger an alert that there has been a change in the overall system. In case of misclassification of new samples, a deep investigation on the possible reasons should take place.

The ‘Appendix XVIII: Guidance On Developing and Validating Non-Targeted Methods For Adulteration Detection’, released by The USP Expert Panel on Non-Targeted Methods for Milk Ingredients in 2016, proposes an internal control plan for the monitoring a validated non-targeted method. We refer to this document for a list of specific actions.

Method Updates/Revalidation

While monitoring can be done regularly and in parallel to the current usage of the method, updating events needs to be scheduled in advance. During updates, the method needs to be revalidated, because the size and the composition of the reference data matrix are changed. A new validated method, possibly characterized by new limits of applicability and performances, is released after the update.

As already mentioned, updates can follow a set schedule, or they can be specially required when there has been a significant change in the new samples being analysed during the routine monitoring. In these cases, updating events might be anticipated respect to the schedule.
During the update, all data coming from the previous external validation set and all data collected during the monitoring must be merged to the original data matrix of the model set. If necessary, more samples (both inside and outside the limits of applicability) are collected specifically for the update event. Eventually, data of samples belonging to the model set that are identified as “old” (for example because they belong to a subgroup that is no longer represented by the new unknown samples to be tested) can be subtracted from the data matrix: this might help the classification power of the new prediction model, especially if the profiles of the old samples differ from those of the new samples representative of that class. Once the new data matrix has passed the internal validation step, a new validation set must be collected, according to the previously set criteria, in order to complete the validation procedure.

**COMMUNICATION**

The complete procedure, from the study planning to the validation and the monitoring/updating plan (see the proposed workflow) should be correctly described, and the results should be efficiently expressed (see Report from validation process section).

**CONCLUSIONS**

The present document is the results of the activities of the WP18 “INTELLItrace” within the Food Integrity Project (Seventh Framework Programme for research, technological development and demonstration; under grant agreement no 613688), which brought together multidisciplinary expertise (food chemistry; analytical chemistry; chemometrics; data mining and machine learning). The principal purpose of the Authors was to provide good practices and methodological guidelines on the development and validation of non-targeted analytical methods useful to tackle food integrity, authenticity and traceability issues. This guidance originates both from the evaluation of experimental cases-studies, opportunely designed to represent different food fraud typologies (among them, as examples, geographic origin of honey; wild vs farmed salmons; partial substitution of high value saffron powder with turmeric), and a wide revision of the existing literature. A long and intense discussion within the WP 18 Consortium has derived, and several experts in the field were also invited to share their experience and provide valuable comments and suggestions. As final result, this guideline would like to represent an attempt of harmonization of the main key steps required to develop and validate non-targeted methods, particularly addressed to analytical chemistry techniques. Nevertheless, we must point out that the proposed approach is restricted to an *in house* (intra-laboratory) validation protocol. Surely, a more complete methodological approach of validation should be implemented, including an inter-laboratory phase. Some analytical techniques here considered, however, do not easily allow obtaining equivalent
datasets that can be integrated in or evaluated by a unique developed method. Specific trials should be planned to overcome these limitations. However, even if currently not applicable in a general way to all the analytical techniques, a specific protocol for inter-laboratory comparisons based on non-targeted NMR analysis is proposed in the Annex 2. Based on this protocol, useful to validate unbiased and multi-user NMR based classification tool, other studies can be planned, in order to identify the best operative procedures for different analytical techniques. Finally, the Authors hope that this document will stimulate further discussions aimed at improving the approach here suggested.

Acknowledgments

The Authors of the Deliverables are grateful to: Paul Brereton (Queen's University Belfast, UK); James Donarski (FERA, York, UK); Saskia van Ruth (WUR, Rikilt, The Netherlands); Martin Alewijn, WUR, Rikilt, The Netherlands); Laszlo Hollosi (Thermo Scientific Company); Carsten Fauhl-Hassek (Bfr, Berlin, Germany).
Case study: assessment of geographic origin of Acacia honey

1. Definition of the scope

The general purpose of the present study was to develop a preliminary non-targeted method able to discriminate Italian Acacia honey from Eastern European ones. More particularly, the specific aim is to test the potential application of LC-LRMS technique to this specific problem, thus confirming, as future perspective, the possibility to implement the method, making it a routine method for the quality control at industrial level.

2. Feasibility study

The feasibility study was performed on 18 individual samples (9 for each class of interest, i.e. Italy and Eastern Europe), corresponding to about 20% of the total samples selected for the method development and the internal validation (see the following section). The aim of this phase was to identify, in a preliminary way, if the selected method could be potentially able to distinguish Italian Acacia honey from Eastern European ones.

The feasibility study was conducted to evaluate the variability within and between the sample sets belonging to the different classes and was based on three main points:

a) sample preparation;

b) analysis;

c) exploratory data analysis.

a) Sample preparation protocols

The 18 selected honey samples were extracted to obtain extracts generating m/z signals useful to differentiate the geographic origin of the samples. As a general rule, the extraction solvent was selected to extract medium-polar compounds (log P in the range from -1.5 to 7) amenable of ionization in both positive and negative polarity. The selected extraction protocol was carried out as follows. One gram of honey was weighted in a plastic ultracentrifuge tube and added to 3 ml of water. After 8 min shaking, 6 ml of 1% CH₃COOH in CH₃CN were added, and the whole mixture was stirred for 10 min. Phases separation was achieved by addition of 4 g of MgSO₄ and 1 g of sodium acetate. Samples were stirred for 10 min, then centrifuged and left in refrigerator for almost 2 hours. The acetonitrile phase of each sample was divided in
two aliquots: one to be analyzed in positive mode (POS) and one to be analyzed in negative mode (NEG).

b) Analytical protocol

Analysis of the extracts were performed by using a SCIEX QTRAP 6500+ LC-MS/MS system coupled with a 3000 Ultimate Thermo UPLC.

The column was used and maintained according to the supplier’s indications, and was checked twice before analysing the first sample.

Prior to injection, the POS aliquots were purified with 300 mg of MgSO4 and 75 mg of PSA sorbent bulk, centrifuged and diluted 1:1 with ammonium formate 10mM pH 4 containing internal standard Chlorpyrifos Ethyl D10. The NEG aliquots were purified with 200 mg of MgSO4 and 35 mg of C18 sorbent bulk, centrifuged and diluted 1:1 with ammonium formate 10mM pH 4 containing internal standard Nicarbazine.

Operating conditions:

- **LR-MS**
  - Scan Mode: Full scan; scan range: 50-1000 Da
  - Polarity: positive/negative

  **POS**
  - Curtain gas (CUR): 30 psi
  - Collision Gas (CAD): Medium
  - Ion Spray Voltage (IS): +5500 V
  - Temperature (TEM): 450°C
  - Ion Source Gas 1 (GS1): 45 psi
  - Ion Source Gas 2 (GS2): 40 psi
  - Declustering Potential (DP): 80
  - Entrance Potential (EP): 10

  **NEG**
  - Curtain gas (CUR): 30 psi
  - Collision Gas (CAD): Medium
  - Ion Spray Voltage (IS): -4500 V
  - Temperature (TEM): 450°C
  - Ion Source Gas 1 (GS1): 45 psi
  - Ion Source Gas 2 (GS2): 40 psi
  - Declustering Potential (DP): -70
  - Entrance Potential (EP): -10
• **LC**

1. Column: *Acquity UPLC BEH C18 130Å (50 mm x 2.1 mm i.d. x 1.7 µm)*
2. Flow: 0.2 ml/min
3. Mobil phase: A) Ammonium formate 10 mM pH4  
   B) MetOH

<table>
<thead>
<tr>
<th>Time (min)</th>
<th>%B</th>
</tr>
</thead>
<tbody>
<tr>
<td>0,0</td>
<td>0</td>
</tr>
<tr>
<td>0,5</td>
<td>0</td>
</tr>
<tr>
<td>14,0</td>
<td>100</td>
</tr>
<tr>
<td>16,0</td>
<td>100</td>
</tr>
<tr>
<td>16,5</td>
<td>0</td>
</tr>
<tr>
<td>19,0</td>
<td>0</td>
</tr>
</tbody>
</table>

c) **Exploratory data analysis**

The MarkerView software (AB Sciex) was used to detect mass features (i.e. signals corresponding to individual metabolites characterized by a set retention time, accurate mass and intensity) and then to select the major ones able to differentiate the geographic origin of the honey samples. Based on the Volcano plot reporting P values against fold changes, 8 mass features for POS analyses and 5 for NEG analysis were selected for the data analysis.

An unsupervised PCA (Principal Component Analysis) was selected as exploratory chemometric analysis. Prior to the PCA, the data were mean centered and subjected to Pareto scaling. The PCA score plot (representing the first two principal components, cumulative $R^2 = 0.96$ and cumulative $Q^2 = 0.88$) highlights a good separation of the samples based on their geographic origin on the first principal component. This result suggests that the selected method is promising to achieve its intended purpose.
3. Method development

Samples

Honey samples for both internal and external validation sets were provided by an Italian dealer that guaranteed their origin. All the samples were provided in glass jars (closed with screw caps) and stored in the dark at room temperature until the sampling for the extraction and the analysis.

- **Internal validation set**

A first set of 69 Acacia honey samples was obtained for developing the analytical method and the first step of the validation process (internal validation). More particularly, 30 honey samples were from Italy and 39 from East Europe. Eastern European honey samples included 15 honeys from Hungary, 3 from Romania, 3 from Serbia, 1 from Moldavia; the origin of the 17 remaining honey samples was declared as not quantitatively defined mixtures of the
previously cited East European countries.

- **External validation set**

A second set of 82 Acacia honey samples (20 from Italy, 62 from East Europe (including Hungary, Romania, Serbia, Moldavia, Croatia and Ukraine) was obtained to fulfil the second step of the validation process (external validation with blind samples). The specific composition of Eastern European samples was as follows: 2 from Croatia, 10 from Romania, 24 from Hungary; the remaining 26 samples were indicated as different mixtures from the previously specified countries.

**Analytical method**

Extraction and analytical protocols were as previously described in the Feasibility study section. Each sample was extracted in single and analyzed in triplicate.

**Pre-processing method**

Background subtraction, alignment and mass feature detection were all performed in MarkerView software (AB Sciex).

Mass feature detection was obtained based on the following criteria:

1) m/z extraction tolerance: 1 Da;
2) Retention time extraction window: 3-15 min;
3) Retention time tolerance: 0.5 min;
4) Intensity Threshold Feature > $10^5$;
5) Intensity Noise Threshold: 1000.

Only mass features detected in at least two of the three replicates for each sample were considered; null values were assigned to the features having Intensity Threshold < $10^5$.

In order to avoid systematic bias due to analytical variation, all samples were injected under a randomized sequence and QC samples, obtained by mixing an equal amount of 15 Italian honey samples and 15 East European honey samples, were analyzed at regular intervals through the analysis (about every other 20 samples, corresponding to the start, the middle and the end of each analytical batch). The robustness of the analytical procedure was demonstrated by the tight clustering of QC samples. Relative standard deviation (RSD%) of QC samples was < 30%.
The detected mass features were then sequentially subjected to:

1. Normalization based on the internal standard intensity;

2. Normalization based on the sum of the intensity values (the sum of all features of each observation was set to 1).

Pre-processed data were organized in a tidy matrix of dimensions $i \times j$, where $i$ is the number of the instances (corresponding to the number of samples multiplied for three analytical replications) and $j$ represents the number of the mass features.

From the analysis of the internal validation sample set, two different data sets were produced, one obtained from the positive mode analysis (POS) and the other from the negative mode (NEG). The first dataset was composed by 207 instances (90 for Italian honey samples; 117 for Eastern European ones) and 2900 features, the second one included the same number of instances, but only 578 features. For the external validation set, the total instances were 247, while the number of the features was 2900 and 578, for the positive and negative ionization data sets, respectively.

**Data analysis and classification approaches**

As suggested by this guideline, the data analysis and classification testing/validation phase should start with a feature selection step, followed by the choice of an appropriate cross-validation strategy. To this purpose it is important to choose a set of statistical approaches to be trained and validated in parallel, in order to obtain a comprehensive analysis of the data from different point of view and to foster the identification of a proper classification strategy, among the classifiers trained and validated for the proposed goal.

As first step, feature selection was performed through a correlation-based algorithm, using two different search methods: (1) a sequential forward-backward selection on the original feature space, resulting in the selection of 23 analytical attributes out of 2900 for POS dataset and 14 out of 578 for the NEG one; and (2) a genetic algorithm-based search, resulting in the selection of 839 features and 134 for POS and NEG datasets, respectively.

10-Fold cross-validation was selected as cross-validation strategy for all the tests, and subsequently the complete and the reduced datasets have been analyzed with the following set of mathematical classifiers, all set up for binary classification:

- Bayes Net BN with Cooper/Herskovits algorithm with automatic selection of the best number of parents per node;
- K Nearest Neighbor KNN with automatic selection of the best value of K, cross-validation and KDTree search strategy;
- Decision tree J48 with automatic selection of the best confidence factor value;
- Multi Layer Perceptron (MLP) with 1 hidden layer, with automatic selection of the best number of hidden units;
- Support Vector Machine SMO (Sequential Minimal Optimization), with: (1) Pearson Kernel and Platt’s scaling, or (2) polynomial kernel, for output class probability estimation;

In addition, the dataset has been tested with a Soft Independent Modelling of Class Analogy (SIMCA) for two classes, matched with Partial Least Squares - Discriminant Analysis (PLS-DA); in this case, prior to the analysis, the datasets have been subjected to the Pareto scaling.

3. Method validation

Validation parameters and acceptance criteria

The performances of the method (both internal and external validation phases) were evaluated and described by calculating the following parameters:

Confusion matrix, showing how many samples were classified correctly and how many were misclassified, for each class;

Accuracy;
Kappa statistic;
Sensitivity;
Specificity;
Precision;
Recall;
F-Measure;
Matthews correlation coefficient (MCC);

Misclassification Risk (MR);

ROC Area (receiver operating characteristic, area under the curve) (not determined for SIMCA/PLS-DA approach).

Mean and variance of classification/misclassification probabilities for each classifier (excluding SIMCA/PLS-DA approach) were also considered.
Concerning the internal validation, according to what stated in the guidelines the method is considered acceptable for **Accuracy**, **Sensitivity** and **Specificity** if showing values better than/equal to 95%, above which the following step of external validation can be carried out.

In order to accept the overall method for future implementations (considering other variables as seasonality, raw materials, manufacturing process, packaging, transport, storage and so on), the external validation will not be considered acceptable for **Accuracy** and **Sensitivity** values lower than 75%.

### 3.4 Internal validation

The internal validation procedure aims at obtaining a classifier (or a set of classifiers) which sub-set of validation parameters, chosen by the experts, matches or overcomes the acceptability thresholds, that have been set as minimum performance requirements. To assess the geographic origin of honey, the internal validation has been performed using a binary classification approach (Italian honey vs East Europe), applying a 10-fold cross validation to each combination of the classifiers and the model samples in each of the datasets (both POS and NEG), considering each of them as rough data without feature selection, or with feature selection through linear search, or with feature selection using the genetic algorithm. After analyzing the results of each classifier, trained with each of these datasets, we obtained the best classification results for data from positive ionization mode analysis (POS), followed by correlation-based feature selection (no genetic algorithm) and sequential selection on the original feature space.

The following results are, thus, referred to the POS dataset reduced to 207x23 (instances x features) dimensions.

Concerning the thresholds set for the acceptability of the internal validation classification models, we concentrated, on one hand, on the general accuracy, setting a threshold of more than 95% for this parameter and, on the other hand, on parameters to handle the problem of class imbalance affecting the datasets used for this study. For this purpose, we set sensitivity and specificity ratio higher than 0.95 for each class.

The **Misclassification Risk (MR)** was determined by defining a cost matrix to measure only the impact of the misclassifications; therefore, we placed the cost 1 in the positions corresponding to misclassifications in the cost matrix (C(no|yes) and C(yes|no) ), and the cost 0 elsewhere. Misclassifications of Italian honey samples were considered equivalently to misclassifications of East European ones. A-priori probabilities have been set considering the cardinality of the samples for each class. Therefore, the a-priori probabilities for the misclassification risk in the internal validation dataset are set as: \( P_{\text{Italian}} = 90/207 \) and \( P_{\text{East Europe}} = 117/207 \), while the a-priori probabilities for the external validation dataset are set as: \( P_{\text{Italian}} = 60/246 \) and \( P_{\text{East Europe}} = 186/246 \).
The classification parameters for each classifier, providing the best classification results with the selected dataset are summarized as follows:

- Bayes Net BN: K2 algorithm and max. of 2 parents per node;
- K Nearest Neighbor: K=1 and Euclidean distance;
- Decision tree J48: C4.5 Decision Tree with confidence factor of 0.23;
- Multi-layer Perceptron: no. of hidden units = (attrs + classes) / 2;
- Support vector machine: polynomial kernel;
- Soft Independent Modeling of Class Analogy: 4 component for both classes; Partial Least Squares - Discriminant Analysis: 5 component

These classifiers produced the validation results listed in the following tables:

**BN:**

<table>
<thead>
<tr>
<th>Accuracy</th>
<th>205</th>
<th>99.0338 %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kappa statistic</td>
<td>0.9804</td>
<td></td>
</tr>
<tr>
<td>Total Number of Instances</td>
<td>207</td>
<td></td>
</tr>
</tbody>
</table>

***Detailed Accuracy By Class***

<table>
<thead>
<tr>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
<th>MCC</th>
<th>ROC Area</th>
<th>MR</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,000</td>
<td>0,983</td>
<td>0,978</td>
<td>1,000</td>
<td>0,989</td>
<td>0.981</td>
<td>1,000</td>
<td>0.00% Italian</td>
<td></td>
</tr>
<tr>
<td>0,983</td>
<td>1,000</td>
<td>1,000</td>
<td>0,983</td>
<td>0,991</td>
<td>0.981</td>
<td>1,000</td>
<td>0.85% East Europe</td>
<td></td>
</tr>
<tr>
<td>0,990</td>
<td>0,993</td>
<td>0,991</td>
<td>0,990</td>
<td>0,990</td>
<td>0.981</td>
<td>1,000</td>
<td>0.85% Overall</td>
<td></td>
</tr>
</tbody>
</table>

***Confusion Matrix***

<table>
<thead>
<tr>
<th>a</th>
<th>b</th>
<th>classified as</th>
</tr>
</thead>
<tbody>
<tr>
<td>90</td>
<td>0</td>
<td>a = Italian</td>
</tr>
<tr>
<td>2</td>
<td>115</td>
<td>b = East Europe</td>
</tr>
</tbody>
</table>

***Mean and variance of classification/misclassification probabilities***

<table>
<thead>
<tr>
<th>Mean:</th>
<th>Variance:</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>0,995</td>
<td>0,000</td>
</tr>
<tr>
<td>0,620</td>
<td>0,999</td>
</tr>
<tr>
<td></td>
<td>a</td>
</tr>
<tr>
<td>0,029</td>
<td>0,006</td>
</tr>
<tr>
<td>0,152</td>
<td>0,006</td>
</tr>
</tbody>
</table>

**KNN:**

<table>
<thead>
<tr>
<th>Accuracy</th>
<th>207</th>
<th>100 %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kappa statistic</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>
Total Number of Instances              207

=== Detailed Accuracy By Class ===

<table>
<thead>
<tr>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
<th>MCC</th>
<th>ROC Area</th>
<th>MR</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>0,956</td>
<td>0,983</td>
<td>0,977</td>
<td>0,956</td>
<td>0,966</td>
<td>0,941</td>
<td>0,987</td>
<td>2,22%</td>
<td>Italian</td>
</tr>
<tr>
<td>0,983</td>
<td>0,954</td>
<td>0,966</td>
<td>0,983</td>
<td>0,975</td>
<td>0,941</td>
<td>0,987</td>
<td>0,85%</td>
<td>East Europe</td>
</tr>
<tr>
<td>0,971</td>
<td>0,967</td>
<td>0,971</td>
<td>0,971</td>
<td>0,971</td>
<td>0,941</td>
<td>0,987</td>
<td>3,07%</td>
<td>Overall</td>
</tr>
</tbody>
</table>

=== Confusion Matrix ===

a b <-- classified as

86 4 | a = Italian
2 115 | b = East Europe

=== Mean and variance of classification/misclassification probabilities ===

Mean:

<table>
<thead>
<tr>
<th>a</th>
<th>b</th>
<th>&lt;- classified as</th>
</tr>
</thead>
<tbody>
<tr>
<td>0,995</td>
<td>0,000</td>
<td>a = Italian</td>
</tr>
<tr>
<td>0,000</td>
<td>0,995</td>
<td>b = East Europe</td>
</tr>
</tbody>
</table>

Variance:

<table>
<thead>
<tr>
<th>a</th>
<th>b</th>
<th>&lt;- classified as</th>
</tr>
</thead>
<tbody>
<tr>
<td>0,000</td>
<td>0,000</td>
<td>a = Italian</td>
</tr>
<tr>
<td>0,000</td>
<td>0,000</td>
<td>b = East Europe</td>
</tr>
</tbody>
</table>

J48:

Accuracy              201               97.1014 %
Kappa statistic                          0.9409
Total Number of Instances              207

=== Detailed Accuracy By Class ===

<table>
<thead>
<tr>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
<th>MCC</th>
<th>ROC Area</th>
<th>MR</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>0,956</td>
<td>0,983</td>
<td>0,977</td>
<td>0,956</td>
<td>0,966</td>
<td>0,941</td>
<td>0,987</td>
<td>2,22%</td>
<td>Italian</td>
</tr>
<tr>
<td>0,983</td>
<td>0,954</td>
<td>0,966</td>
<td>0,983</td>
<td>0,975</td>
<td>0,941</td>
<td>0,987</td>
<td>0,85%</td>
<td>East Europe</td>
</tr>
<tr>
<td>0,971</td>
<td>0,967</td>
<td>0,971</td>
<td>0,971</td>
<td>0,971</td>
<td>0,941</td>
<td>0,987</td>
<td>3,07%</td>
<td>Overall</td>
</tr>
</tbody>
</table>

=== Confusion Matrix ===

a b <-- classified as

86 4 | a = Italian
2 115 | b = East Europe

=== Mean and variance of classification/misclassification probabilities ===

Mean:

<table>
<thead>
<tr>
<th>a</th>
<th>b</th>
<th>&lt;- classified as</th>
</tr>
</thead>
<tbody>
<tr>
<td>0,995</td>
<td>0,000</td>
<td>a = Italian</td>
</tr>
<tr>
<td>0,000</td>
<td>0,995</td>
<td>b = East Europe</td>
</tr>
</tbody>
</table>

Variance:

<table>
<thead>
<tr>
<th>a</th>
<th>b</th>
<th>&lt;- classified as</th>
</tr>
</thead>
<tbody>
<tr>
<td>0,000</td>
<td>0,000</td>
<td>a = Italian</td>
</tr>
<tr>
<td>0,000</td>
<td>0,000</td>
<td>b = East Europe</td>
</tr>
</tbody>
</table>
MLP:

**Accuracy**
0.999 0.986 | a = Italian
0.895 0.991 | b = East Europe

**Kappa statistic**
0.012 0.010 | a = Italian
0.008 0.027 | b = East Europe

**Total Number of Instances**
207

--- Detailed Accuracy By Class ---

<table>
<thead>
<tr>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
<th>MCC</th>
<th>ROC Area</th>
<th>MR</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>0.00%</td>
<td>Italian</td>
</tr>
<tr>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>0.00%</td>
<td>East Europe</td>
</tr>
<tr>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>0.00%</td>
<td>Overall</td>
</tr>
</tbody>
</table>

--- Confusion Matrix ---

<table>
<thead>
<tr>
<th>a</th>
<th>b</th>
</tr>
</thead>
<tbody>
<tr>
<td>90</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>117</td>
</tr>
</tbody>
</table>

--- Mean and variance of classification/misclassification probabilities ---

**Mean:**
0.993 0.000 | a = Italian
0.000 0.994 | b = East Europe

**Variance:**
0.020 0.000 | a = Italian
0.000 0.016 | b = East Europe

SMO:

**Accuracy**
0.999 0.986 | a = Italian
0.895 0.991 | b = East Europe

**Kappa statistic**
0.012 0.010 | a = Italian
0.008 0.027 | b = East Europe

**Total Number of Instances**
207

--- Detailed Accuracy By Class ---

<table>
<thead>
<tr>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
<th>MCC</th>
<th>ROC Area</th>
<th>MR</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.967</td>
<td>1.000</td>
<td>1.000</td>
<td>0.967</td>
<td>0.983</td>
<td>0.971</td>
<td>0.983</td>
<td>1.67%</td>
<td>Italian</td>
</tr>
<tr>
<td>1.000</td>
<td>0.967</td>
<td>0.975</td>
<td>1.000</td>
<td>0.987</td>
<td>0.971</td>
<td>0.983</td>
<td>0.00%</td>
<td>East Europe</td>
</tr>
<tr>
<td>0.986</td>
<td>0.981</td>
<td>0.986</td>
<td>0.986</td>
<td>0.985</td>
<td>0.971</td>
<td>0.983</td>
<td>1.67%</td>
<td>Overall</td>
</tr>
</tbody>
</table>
All the classifiers produced very good results on the POS database, reduced with correlation-based feature selection. In particular, comparing these results with the validation parameters thresholds we have previously set, we can see that all the classifiers trained on this dataset matched all the requirements and can be considered as a candidate for further testing through external validation.

In the next section, we tested the validation set analysed in positive mode, considering the same features selected for the model dataset, on each of the classifier models trained for this internal validation step.
3.5 External validation

The second validation step (external validation) was carried out by applying the classifiers trained for the internal validation to the external validation dataset, after selecting by hand only the features selected by the automatic correlation-based feature selection and used to train the classification models in the internal validation step (the dimensions of the final data matrix was of 246 x 23).

The method was thus subjected to the validation process for its original aim; the performances of the method are defined in the following tables.

<p>| | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>201</td>
<td>81.7073 %</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kappa statistic</td>
<td>0.5123</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total Number of Instances</td>
<td>246</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

== Detailed Accuracy By Class ==

<table>
<thead>
<tr>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
<th>MCC</th>
<th>ROC Area</th>
<th>MR</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.650</td>
<td>0.129</td>
<td>0.619</td>
<td>0.650</td>
<td>0.634</td>
<td>0.513</td>
<td>0.859</td>
<td>8,54% Italian</td>
</tr>
<tr>
<td>0.871</td>
<td>0.350</td>
<td>0.885</td>
<td>0.871</td>
<td>0.878</td>
<td>0.513</td>
<td>0.859</td>
<td>9,76% East Europe</td>
</tr>
<tr>
<td>0.817</td>
<td>0.296</td>
<td>0.820</td>
<td>0.817</td>
<td>0.819</td>
<td>0.513</td>
<td>0.859</td>
<td>18,30% Overall</td>
</tr>
</tbody>
</table>

== Confusion Matrix ==

<table>
<thead>
<tr>
<th>a</th>
<th>b</th>
</tr>
</thead>
<tbody>
<tr>
<td>39</td>
<td>21</td>
</tr>
<tr>
<td>24</td>
<td>162</td>
</tr>
</tbody>
</table>

== Mean and variance of classification/misclassification probabilities ==

<table>
<thead>
<tr>
<th>Mean:</th>
<th>Variance:</th>
</tr>
</thead>
<tbody>
<tr>
<td>a b &lt;-- classified as</td>
<td>a b &lt;-- classified as</td>
</tr>
<tr>
<td>0.971 0.977</td>
<td>a = Italian</td>
</tr>
<tr>
<td>0.996 0.999</td>
<td>b = East Europe</td>
</tr>
<tr>
<td>0.106 0.078</td>
<td>a = Italian</td>
</tr>
<tr>
<td>0.011 0.009</td>
<td>b = East Europe</td>
</tr>
</tbody>
</table>

KNN:

<p>| | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>197</td>
<td>80.0813 %</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kappa statistic</td>
<td>0.463</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total Number of Instances</td>
<td>246</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

== Detailed Accuracy By Class ==
<table>
<thead>
<tr>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
<th>MCC</th>
<th>ROC Area</th>
<th>MR</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.600</td>
<td>0.134</td>
<td>0.590</td>
<td>0.600</td>
<td>0.595</td>
<td>0.463</td>
<td>0.733</td>
<td>9.76%</td>
<td>Italian</td>
</tr>
<tr>
<td>0.866</td>
<td>0.400</td>
<td>0.870</td>
<td>0.866</td>
<td>0.868</td>
<td>0.463</td>
<td>0.733</td>
<td>10.16%</td>
<td>East Europe</td>
</tr>
<tr>
<td>0.801</td>
<td>0.335</td>
<td>0.802</td>
<td>0.801</td>
<td>0.801</td>
<td>0.463</td>
<td>0.733</td>
<td>19.92%</td>
<td>Overall</td>
</tr>
</tbody>
</table>

=== Confusion Matrix ===

<table>
<thead>
<tr>
<th>a</th>
<th>b</th>
<th>classifier as</th>
</tr>
</thead>
<tbody>
<tr>
<td>36</td>
<td>24</td>
<td>a = Italian</td>
</tr>
<tr>
<td>25</td>
<td>161</td>
<td>b = East Europe</td>
</tr>
</tbody>
</table>

=== Mean and variance of classification/misclassification probabilities ===

<table>
<thead>
<tr>
<th>Mean:</th>
<th>Variance:</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>0.995</td>
<td>0.995</td>
</tr>
<tr>
<td>0.995</td>
<td>0.995</td>
</tr>
</tbody>
</table>

J48:

Accuracy: 191 77.6423 %
Kappa statistic: 0.4688
Total Number of Instances: 246

=== Detailed Accuracy By Class ===

<table>
<thead>
<tr>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
<th>MCC</th>
<th>ROC Area</th>
<th>MR</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.750</td>
<td>0.215</td>
<td>0.529</td>
<td>0.750</td>
<td>0.621</td>
<td>0.483</td>
<td>0.767</td>
<td>6.10%</td>
<td>Italian</td>
</tr>
<tr>
<td>0.785</td>
<td>0.250</td>
<td>0.907</td>
<td>0.785</td>
<td>0.841</td>
<td>0.483</td>
<td>0.767</td>
<td>16.26%</td>
<td>East Europe</td>
</tr>
<tr>
<td>0.776</td>
<td>0.241</td>
<td>0.815</td>
<td>0.776</td>
<td>0.788</td>
<td>0.483</td>
<td>0.767</td>
<td>22.36%</td>
<td>Overall</td>
</tr>
</tbody>
</table>

=== Confusion Matrix ===

<table>
<thead>
<tr>
<th>a</th>
<th>b</th>
<th>classifier as</th>
</tr>
</thead>
<tbody>
<tr>
<td>45</td>
<td>15</td>
<td>a = Italian</td>
</tr>
<tr>
<td>40</td>
<td>146</td>
<td>b = East Europe</td>
</tr>
</tbody>
</table>

=== Mean and variance of classification/misclassification probabilities ===

<table>
<thead>
<tr>
<th>Mean:</th>
<th>Variance:</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>1,000</td>
<td>1,000</td>
</tr>
<tr>
<td>1,000</td>
<td>1,000</td>
</tr>
</tbody>
</table>
MLP:

Accuracy 203 82.5203 %
Kappa statistic 0.5676
Total Number of Instances 246

=== Detailed Accuracy By Class ===

<table>
<thead>
<tr>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
<th>MCC</th>
<th>ROC Area</th>
<th>MR</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.783</td>
<td>0.161</td>
<td>0.610</td>
<td>0.783</td>
<td>0.686</td>
<td>0.576</td>
<td>0.892</td>
<td>5.28%</td>
<td>Italian</td>
</tr>
<tr>
<td>0.839</td>
<td>0.217</td>
<td>0.923</td>
<td>0.839</td>
<td>0.879</td>
<td>0.576</td>
<td>0.892</td>
<td>12.20%</td>
<td>East Europe</td>
</tr>
<tr>
<td>0.825</td>
<td>0.203</td>
<td>0.847</td>
<td>0.825</td>
<td>0.832</td>
<td>0.576</td>
<td>0.892</td>
<td>17.48%</td>
<td>Overall</td>
</tr>
</tbody>
</table>

=== Confusion Matrix ===

\[
\begin{array}{cc}
    a & b \\
    47 & 13 | a = Italian \\
    30 & 156 | b = East Europe \\
\end{array}
\]

=== Mean and variance of classification/misclassification probabilities ===

<table>
<thead>
<tr>
<th>Mean:</th>
<th>Variance:</th>
</tr>
</thead>
<tbody>
<tr>
<td>a      b</td>
<td>a      b</td>
</tr>
<tr>
<td>0.994 0.832</td>
<td>0.029 0.184</td>
</tr>
<tr>
<td>0.982 0.993</td>
<td>0.029 0.025</td>
</tr>
</tbody>
</table>

SMO:

Accuracy 208 84.5528 %
Kappa statistic 0.6078
Total Number of Instances 246

=== Detailed Accuracy By Class ===

<table>
<thead>
<tr>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
<th>MCC</th>
<th>ROC Area</th>
<th>MR</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.783</td>
<td>0.134</td>
<td>0.653</td>
<td>0.783</td>
<td>0.712</td>
<td>0.612</td>
<td>0.824</td>
<td>5.28%</td>
<td>Italian</td>
</tr>
<tr>
<td>0.866</td>
<td>0.217</td>
<td>0.925</td>
<td>0.866</td>
<td>0.894</td>
<td>0.612</td>
<td>0.824</td>
<td>10.16%</td>
<td>East Europe</td>
</tr>
<tr>
<td>0.846</td>
<td>0.197</td>
<td>0.859</td>
<td>0.846</td>
<td>0.850</td>
<td>0.612</td>
<td>0.824</td>
<td>15.44%</td>
<td>Overall</td>
</tr>
</tbody>
</table>

=== Confusion Matrix ===
a  b  <-- classified as
47  13 |  a = Italian
25 161 |  b = East Europe

=== Mean and variance of classification/misclassification probabilities ===

<table>
<thead>
<tr>
<th>Mean:</th>
<th>Variance:</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>1,000</td>
<td>1,000</td>
</tr>
<tr>
<td>1,000</td>
<td>1,000</td>
</tr>
</tbody>
</table>

SIMCA/PLS-DA:

<table>
<thead>
<tr>
<th>Accuracy</th>
<th>85.3659 %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kappa statistic</td>
<td>0.6032</td>
</tr>
<tr>
<td>Total Number of Instances</td>
<td>246</td>
</tr>
</tbody>
</table>

=== Detailed Accuracy By Class ===

<table>
<thead>
<tr>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
<th>MCC</th>
<th>MR</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.700</td>
<td>0.903</td>
<td>0.700</td>
<td>0.700</td>
<td>0.700</td>
<td>0.603</td>
<td>15.00%</td>
<td>Italian</td>
</tr>
<tr>
<td>0.903</td>
<td>0.700</td>
<td>0.903</td>
<td>0.903</td>
<td>0.903</td>
<td>0.603</td>
<td>4.84%</td>
<td>East Europe</td>
</tr>
<tr>
<td>0.815</td>
<td>0.788</td>
<td>0.815</td>
<td>0.815</td>
<td>0.815</td>
<td>0.603</td>
<td>19.84%</td>
<td>Overall</td>
</tr>
</tbody>
</table>

=== Confusion Matrix ===

<table>
<thead>
<tr>
<th>a  b  &lt;-- classified as</th>
</tr>
</thead>
<tbody>
<tr>
<td>42 18</td>
</tr>
<tr>
<td>18 168</td>
</tr>
</tbody>
</table>

Except for BN, KNN and SIMCA/PLS-DA classifiers, at least the 75% accuracy value and, for each honey class (Italy and East Europe), a sensitivity of at least 0.75 was reached, satisfying the thresholds initially requested for these validation parameters. Consequently, the method can be considered promising for future implementations (considering other variables as seasonality, raw materials, manufacturing process, packaging, transport, storage and so on).

**General considerations from the results.** The results of the external validation reveal that the classifiers trained for the internal validation cannot recognize the external validation samples with the same extremely high performances. Furthermore, from the global evaluation of the results, we have to additionally observe that sensitivity and specificity highlight differences in
the recognition of the Italian samples, with respect to the East Europe ones. For example, the BN has sensitivity of 0.650 for the Italian class and of 0.871 for East Europe, while specificity has values of 0.129 and 0.350 for the same classes, respectively. This situation applies, with more or less impact, also to all the other classifiers, meaning that, in general, the classifiers struggle to identify the origin of the Italian honey samples, while the situation improves when they have to deal with the East Europe ones. The reasons for such a behaviour should be investigated, together with awareness and attention to the representativeness of sampling planned from the beginning in relation to the purpose and the variables involved, which, at a first glance, could be the main problem leading to the great difference in performance between the internal and the external validation.

Overfitting should also be considered as one of the reasons for this behaviour, pointing the focus on the results of the internal validation: they seem, indeed, very optimistic, suggesting that the overfitting problem should be seriously taken into account. One of the aspects the experts could investigate would be the possibility that the dataset for internal validation contains too little variability, therefore the classifiers learned a very “specialized” version of the two categories. This problem can be investigated and tackled by putting all the samples from the internal and external validation together and use them as a new internal validation set to train and validate the classifiers. A new set of unseen data can then be used to verify, as external validation, if the added variability provided by the fusion of the former datasets is able to reduce the problem, obtaining more robust and reliable classifiers.

Furthermore, even if in the present example we have not defined specific thresholds for all these parameters, this possibility should be carefully considered if the identification (or misclassification) of one class respect to the other(s) can have a different impact based on the initial purpose of the method, specifically in the case of fraud detection as in the case of the protection of origin, the detection of adulterants and so on. In these cases, it should be recommended to identify also specific threshold for other parameters, for example Specificity, Precision, ROC Area and Misclassification Risk.
ANNEX 2.

Case study: validation of non-targeted NMR analysis by inter-laboratory comparisons

1. INTRODUCTION

In addition to the in-house validation step, a complete validation approach should include inter-laboratory trials. Considering non-targeted methods, a critical point is to obtain equivalent datasets that can be integrated in a unique developed method; for this reason inter-laboratory comparisons are strongly limited, because data sets generated by different laboratories are generally dependent on instrumental features and environmental conditions, even if analytical protocols are identically performed.

In this annex, even if currently not applicable in a general way to all the analytical techniques, a specific protocol for inter-laboratory comparisons based on non-targeted NMR analysis is proposed; the case study here presented considers the NMR analysis of durum wheat and wheat flour aqueous extracts.

The protocol is integral part of an unbiased and multi-users classification tool for identification of the “origin” of an unknown food sample (Scheme 1). The tool is based on non-targeted NMR analysis. In the following, “origin” is intended in a broad sense and, depending on the specific context, may refer to geographical origin, varietal origin, process step and any other kind of feature of interest.
In Scheme 1, Class 1 indicates samples having a common origin that must be identified by the classifier. The same apply for Class 2 samples. The NMR spectra of the samples belonging to classes 1 and 2 may derive from the analyses carried out by using either a single spectrometer or many spectrometers which are different in term of hardware configuration, magnetic field, manufacturer, age, etc. The NMR spectra of the samples belonging to classes 1 and 2 are collected in a suitably designed database and are used to train \((n + m)\) samples and cross-validate \((n' + m')\) samples the classification models (internal validation). In case of satisfactory results of the internal validation tests, the NMR spectra of the Blind samples are recorded and used for an external validation. Even in this case, the \(x\) Blind samples may derive from different spectrometers. Once ascertained the satisfactory performance of the classifier in the identification of the blind samples, the whole system can be considered suitable for identification of the \(y\) unknown samples. For the sake of clarity, in this annex, a blind sample is intended as a sample whose identity is known to the user, is unknown to the classifier and has not been exploited in the training step. The origin of the blind samples must be included among those of classes 1 and 2. On the other hand, an unknown sample is intended as a sample whose identity is unknown to both user and classifier. The identity of the \(y\) samples must be accepted as correct.
2. THE VALIDATION SCHEME

The validation steps making operative the classification tool shown in scheme 1 are mentioned in the guidelines at section 2. Here, some additional steps are considered to take into proper account the specificity of NMR spectroscopy. In particular, the validation procedure is redefined as summarized in the following scheme 2.

![Scheme 2. Validation procedure for a NMR based classification tool.](image)

Typically, for a wide range of analytical methods, step 1 is performed according to ISO/IEC 17025:2017 [1]. If non-targeted method development is concerned, the absence of internationally agreed procedures paves the way to biased applications of the norm. In this respect, it is important to note that the critical point of the classification tool reported in scheme 1 regard the statistical equivalence of the NMR spectra (strictly speaking, statistical equivalence should be evaluated for “scaled NMR spectra”, as explained below, but in this annex the qualifier “scaled” is omitted for the sake of brevity). In practice, two different samples belonging to the same class must be differentiated each other, independently on the spectrometer used for the measurement. For instance, if 2 of the n samples belonging to Class 1 are subdivided into z aliquots, each aliquot being analysed by a different NMR spectrometer (thus z spectrometers are used), then 2 different groups of z statistical equivalent NMR spectra must be obtained. In this conditions, the variance attributable to the instrumental features (noise) can be considered negligible with respect to the intra-class variance, and then also with respect to inter-class variance. In other words, possible unsatisfactory performance of the
classifier must derive from the specific analytical problem and not from possible instrumental interferences.

Evaluation of the statistical equivalence of the NMR spectra (step 2, scheme 2) requires a suitable scaling procedure. At this stage, it is worthwhile considering the basic equation of quantitative NMR, namely \( I = k \cdot N \) (Equation 1) \[2\]. The equation provides the direct proportionality between the number of moles \( N \) of nuclei generating a signal and the intensity \( I \) of the same signal with a proportionality constant \( k \) being the spectrometer constant which remains the same for all resonances in a NMR spectrum. Thus, in a given spectrum, considering the signal \( a \) having intensity \( I_a \) generated by specific protons belonging to the analyte of interest and the signal \( r \) having intensity \( I_r \) generated by specific protons in a reference compound, the ratio \( \frac{I_a}{I_r} = \frac{N_a}{N_r} \) (Equation 2) is independent from the proportionality constant \( k \) and, as a consequence, it does not depend on the spectrometer. Thus, submitting the NMR spectrum to a bucketing (or binning) procedure and dividing each bucket integral by the integral of a selected signal of a reference compound, results in the generation of a “scaled NMR spectrum” which is independent on spectrometer features.

The statistical equivalence of NMR spectra was evaluated for durum wheat and wheat flour aqueous extracts by an interlaboratory comparison involving 39 different NMR spectrometers \[3\]. In table 1, The 39 NMR spectrometers are differentiated on the bases of magnetic field and manufacturer.

<table>
<thead>
<tr>
<th>Magnetic Field (Larmor frequency for 1H)</th>
<th>Number of spectrometers</th>
</tr>
</thead>
<tbody>
<tr>
<td>9.4 T (400 MHz)</td>
<td>16</td>
</tr>
<tr>
<td>11.7 T (500 MHz)</td>
<td>7</td>
</tr>
<tr>
<td>14.1 T (600 MHz)</td>
<td>14</td>
</tr>
<tr>
<td>16.4 T (700 MHz)</td>
<td>2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Spectrometer manufacturers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bruker</td>
</tr>
<tr>
<td>Varian</td>
</tr>
</tbody>
</table>

In the following, a part of the guidelines for NMR spectra acquisition and processing of the NMR spectra is reported.
General Information

Each pack contains 5 sealed NMR tube (Norell 509-UP 7) labelled with A, B, C, D, E.

NMR tube must be handled by qualified technical staff.

Tube A contains 0.7 mL of deuterated methanol 99.80 %D (CD3OD, CAS number: 811-98-3).

Tubes B and C contain two different wheat extracts in H2O/D2O (90/10, acidic solution at pH = 2 by hydrochloric acid). Tubes contain also sodium azide as biocide (CAS number: 26628-22-8; concentration ca. 15mg/100mL) and 3-(Trimethylsilyl)propionic-2,2,3,3-d4 acid sodium salt (TSP, CAS Number 24493-21-8; concentration ca. 0.2 mg/mL).

Tubes D and E contain two different flour extracts in H2O/D2O (80/10, acidic solution at pH = 2 by hydrochloric acid). Tubes contain also sodium azide as biocide (CAS number: 26628-22-8; concentration ca. 15mg/100mL) and 3-(Trimethylsilyl)propionic-2,2,3,3-d4 acid sodium salt (CAS Number 24493-21-8; concentration ca. 0.2 mg/mL).

The exclusive use of the five tubes is for the inter-laboratory comparison IS-NMR-ILC_001_2014. No different use is allowed without written permission of Innovative Solutions S.r.l.

After use, NMR tube disposal must be carried out according to local legislation. Innovative Solutions S.r.l. is not responsible for uncorrect procedures of NMR tube disposal. No costs can be charged to Innovative Solutions S.r.l. for NMR tube disposal.
NMR measurements

All spectra must be acquired, under manual or automatic procedures, as follows:
- sample loading;
- temperature stabilization at 298.0 ± 0.1 K for at least 5 min;
- lock on solvent signal \( \text{[CD]OD for tube A; H2O/D2O (90/10) for tubes B-E]} \);
- tuning and matching;
- shimming;
- 90° hard pulse calibration (for 1D \(^1\)H-NMR experiments, selective pulse for pre-saturation step must be calculated taking into account the solvent signal width of 25 Hz);
- receiver gain optimization (automatic);
- spectrum acquisition.


Use tube "A" as nuclear magnetic resonance thermometer.

On the temperature control unit of your spectrometer, set the sample temperature to 298 K.

Record a routine \(^1\)H spectrum by 90° single excitation pulse sequence ("s2pu" for Agilent, "tz" for Bruker). Peak pick the two signals at ca. 4.8 and 3.3 ppm and calculate the difference \(\Delta\delta\). Use \(\Delta\delta\) for calculation of the sample temperature \(T\) [K] according to the following equation:

\[
T = -16.7467 \cdot (\Delta\delta)^2 - 52.5130 \cdot \Delta\delta + 419.1381
\]

Adjust the temperature control unit so to obtain a \(\Delta\delta\) value allowing for calculated \(T = 298.0 \pm 0.1\) K.

Once optimized, use the temperature control unit setup for all of the other measurements described in the following. Do not optimize temperature for each session.
Step 2: Spectra registration (tubes A-E)

For each NMR tube, 5 spectra must be recorded to comply with conditions for intermediate precision, i.e. same NMR tube, same spectrometer, same user, at least 24 h delay between runs, removal of the NMR tube from the magnet from run to run. Thus, schematically, 5 registration sessions are required:

1st session:

n. 1 spectrum for tube “A” (routine $^1$H spectrum by single excitation pulse (“12pul” for Agilent, “2g” for Bruker)); filename: TUBE_LABEL_1;

n. 1 spectrum for tube “B” (1D $^1$H-NOESY experiment (“NOESY” for Agilent, “noesypr1d” for Bruker)); filename: TUBE_LABEL_1;

n. 1 spectrum for tube “C” (1D $^1$H-NOESY experiment (“NOESY” for Agilent, “noesypr1d” for Bruker)); filename: TUBE_LABEL_1;

n. 1 spectrum for tube “D” (1D $^1$H-NOESY experiment (“NOESY” for Agilent, “noesypr1d” for Bruker)); filename: TUBE_LABEL_1;

n. 1 spectrum for tube “E” (1D $^1$H-NOESY experiment (“NOESY” for Agilent, “noesypr1d” for Bruker)); filename: TUBE_LABEL_1;

2nd session (at least 24 h later the previous session; pulse programs as in the 1st session):

n. 1 spectrum for tube “A” (routine $^1$H spectrum by single excitation pulse sequence); filename: TUBE_LABEL_2;

n. 1 spectrum for tube “B” (1D $^1$H-NOESY experiment); filename: TUBE_LABEL_2;

n. 1 spectrum for tube “C” (1D $^1$H-NOESY experiment); filename: TUBE_LABEL_2;

n. 1 spectrum for tube “D” (1D $^1$H-NOESY experiment); filename: TUBE_LABEL_2;

n. 1 spectrum for tube “E” (1D $^1$H-NOESY experiment); filename: TUBE_LABEL_2;

3rd session (at least 24 h later the previous session; pulse programs as in the 1st session):

n. 1 spectrum for tube “A” (routine $^1$H spectrum by single excitation pulse sequence); filename: TUBE_LABEL_3;

n. 1 spectrum for tube “B” (1D $^1$H-NOESY experiment); filename: TUBE_LABEL_3;

n. 1 spectrum for tube “C” (1D $^1$H-NOESY experiment); filename: TUBE_LABEL_3;

n. 1 spectrum for tube “D” (1D $^1$H-NOESY experiment); filename: TUBE_LABEL_3;
n. 1 spectrum for tube “E” (1D ²H-NOESY experiment); filename: TUBE LABEL_3;
4th session (at least 24 h later the previous session; pulse programs as in the 1st session):
  n. 1 spectrum for tube “A” (routine ²H spectrum by single excitation pulse sequence); filename:
      TUBE LABEL_4;
  n. 1 spectrum for tube “B” (1D ²H-NOESY experiment); filename: TUBE LABEL_4;
  n. 1 spectrum for tube “C” (1D ²H-NOESY experiment); filename: TUBE LABEL_4;
  n. 1 spectrum for tube “D” (1D ²H-NOESY experiment); filename: TUBE LABEL_4;
  n. 1 spectrum for tube “E” (1D ²H-NOESY experiment); filename: TUBE LABEL_4;
5th session (at least 24 h later the previous session; pulse programs as in the 1st session):
  n. 1 spectrum for tube “A” (routine ²H spectrum by single excitation pulse sequence); filename:
      TUBE LABEL_5;
  n. 1 spectrum for tube “B” (1D ²H-NOESY experiment); filename: TUBE LABEL_5;
  n. 1 spectrum for tube “C” (1D ²H-NOESY experiment); filename: TUBE LABEL_5;
  n. 1 spectrum for tube “D” (1D ²H-NOESY experiment); filename: TUBE LABEL_5;
  n. 1 spectrum for tube “E” (1D ²H-NOESY experiment); filename: TUBE LABEL_5;

Summarizing, each participant will record 25 NMR spectra (5 replicates for each of the 5 NMR tubes)

Acquisition parameters for 1D ²H-NOESY experiments (tubes B-E)

- Agilent spectrometers:
  - Pulse program: NOESY
  - size of f1d (n): 128 K;
  - spectral width (sw): 20 ppm;
  - transmitter offset (tof): ca. 4.70 ppm (set the chemical shift value of the residual water signal);
  - 90° hard pulse (pw): to be optimized by manual or automatic procedures;
  - Steady state (ss): 4;
  - number of transients (nt): 16;

www.innovative-solutions.it
- mixing time (mixN): 0.01 s;
- recycle delay (d1): 30 s;
- no sspu (sspu=’n’)
- no ZQ filter (zqfil=’n’)
- no homo spoil during mixing time (gt1=0, gssl=0 and gstab=0)
- presaturation during the whole length of d1, centered at the HDO residual signal with a yB2 power of about 25 Hz (satmode=’yn’, satsly=d1, satfreq=tof, set satpower in such a way that the output of gatepower(satwtr,tn)=yB2 yB2? is about 25)

- Bruker spectrometers:
  - Pulse program: noesypr1d
  - size of fid (TD): 128 K;
  - spectral width (SW): 20 ppm;
  - transmitter offset: ca. 4.70 ppm (set the chemical shift value of the residual water signal);
  - 90° hard pulse (p1): to be optimized by manual or automatic procedures;
  - power level for presaturation (p8): once optimized p1, calculate p8 by command “pulse 25Hz”;
  - dummy scans (ds): 4;
  - number of scans (ns): 16;
  - mixing time (d8): 0.01 s;
  - recycle delay (d1): 30 s;
Step 3. Processing parameters for 1D $^1$H-NOESY experiments (tubes B-E)

Fourier transform the spectra applying exponential multiplication function with a line broadening of 0.1 Hz. Correct phase and baseline.

Tube A: calculate the $\Delta \delta$ for each session as described above (Step 1).

Tubes B-E: set TSP singlet to 0.00 ppm. Integrate signals using the ranges listed below:

<table>
<thead>
<tr>
<th>Signal label</th>
<th>Integration range</th>
</tr>
</thead>
<tbody>
<tr>
<td>TSP</td>
<td>0.50</td>
</tr>
<tr>
<td>$S_1$</td>
<td>9.50</td>
</tr>
<tr>
<td>$S_2$</td>
<td>7.56</td>
</tr>
<tr>
<td>$S_3$</td>
<td>5.46</td>
</tr>
<tr>
<td>$S_4$</td>
<td>5.28</td>
</tr>
<tr>
<td>$S_5$</td>
<td>4.68</td>
</tr>
<tr>
<td>$S_6$</td>
<td>4.12</td>
</tr>
<tr>
<td>$S_7$</td>
<td>3.33</td>
</tr>
</tbody>
</table>

For each spectrum, calculate the integral ratios $\frac{I_{S_2}}{I_{TSP}}$, $\frac{I_{S_3}}{I_{TSP}}$, $\frac{I_{S_4}}{I_{TSP}}$, $\frac{I_{S_5}}{I_{TSP}}$, $\frac{I_{S_6}}{I_{TSP}}$, $\frac{I_{S_7}}{I_{TSP}}$.

On website http://nmr.mxcs.it/index.php, login with your username and password, click on “Data input” and select “International NMR Inter-Laboratory Comparison IS-NMR-ILC 001_2014”.

In the “Data input” page (Figure 1), upload a zip file containing all of the recorded NMR spectra (25 spectra).

![Figure 1](image)

Then, fill in the form with the values of the integral ratios \( \frac{I_{152}}{I_{153}} \), \( \frac{I_{162}}{I_{163}} \), \( \frac{I_{172}}{I_{173}} \), \( \frac{I_{182}}{I_{183}} \), \( \frac{I_{192}}{I_{193}} \), and \( \frac{I_{202}}{I_{203}} \). Please, consider that:

“Measurement 1” stands for “Session 1”;
“Measurement 2” stands for “Session 2”;
“Measurement 3” stands for “Session 3”;
“Measurement 4” stands for “Session 4”;
“Measurement 5” stands for “Session 5”.

In the “Check” field insert again the value of the corresponding Measurement.
2.1 Statistical elaboration

The selected NMR signal integrals were scaled to the TSP integral and the corresponding \( \frac{I_{\text{signal}}}{I_{\text{TSP}}} \) values were uploaded on the website http://nmr.mxcs.it/index.php, specifically designed and validated for data elaboration in agreement with internationally accepted requirements [4]. \( \frac{I_{\text{signal}}}{I_{\text{TSP}}} \) values were uploaded reporting at least four decimal places. The five \( \frac{I_{\text{signal}}}{I_{\text{TSP}}} \) replicates collected for each signal and for each NMR tube were submitted to the Shapiro-Wilk test to ascertain their normal distribution and to Huber, Dixon, and Grubbs tests for identification of possible outliers. Grubbs tests refer to application of both the classical Grubbs test identifying one outlier and the double Grubbs test which enables the identification of two outliers. Data identified as outliers by all of the four tests were not considered in successive steps. After removing outliers, \( \frac{I_{\text{signal}}}{I_{\text{TSP}}} \) values were used to determine their mean value and the corresponding standard deviation which were considered as intra-laboratory uncertainties of the method. Then, results from all participants were submitted to data elaboration for proficiency test and for determination of the assigned \( \frac{I_{\text{signal}}}{I_{\text{TSP}}} \) values. The lack of official reference data for this case study prompted us to determine assigned values as consensus values from participants [4]. Thus, for each \( \frac{I_{\text{signal}}}{I_{\text{TSP}}} \) ratio, according to the flowchart suggested by Horwitz [5], the 39 standard deviation values were submitted to the Cochran test (provided that all of the 5 replicates successfully passed the abovementioned tests for outliers) with the aim to identify and remove outliers for successive calculations. In turn, mean \( \frac{I_{\text{signal}}}{I_{\text{TSP}}} \) values from data sets which passed successfully the Cochran test were submitted to Huber test with the aim to further refine the quality of the results. All sets of data successfully passing the abovementioned outlier tests were submitted to the Shapiro-Wilk test to ascertain the normal distribution of the population (data were always normal distributed after refinement by the Cochran and Huber tests) and were used to calculate, for each signal, the assigned \( \frac{I_{\text{signal}}}{I_{\text{TSP}}} \) value (Average) [6], the inter-laboratory standard deviation (\( \sigma \)), the coefficient of variation (CV\%), the repeatability variance, the reproducibility limits and other statistical parameters.

2.2 Performance assessment

In the case study, laboratory performance was assessed by calculation of z-scores according to the following equation

\[
\begin{align*}
    z &= \frac{(x_i - \bar{x})}{\sigma} \quad \text{(Equation 3)}
\end{align*}
\]

where \( x_i \) is the mean \( \frac{I_{\text{signal}}}{I_{\text{TSP}}} \) value determined by using the \( i \)th data set, \( \bar{x} \) is the assigned \( \frac{I_{\text{signal}}}{I_{\text{TSP}}} \) value (average), and \( \sigma \) is the inter-laboratory standard deviation, all referred to a single NMR signal.

For z scores, the following limits were considered:
• \(|z| \leq 2.0\) indicates “satisfactory” performance;
• \(2.0 < |z| < 3.0\) indicates “questionable” performance;
• \(|z| \geq 3.0\) indicates “unsatisfactory” performance.

The spectra with at least one \(I_{\text{signal}}/I_{\text{TSP}}\) value falling in the category “unsatisfactory performance” must not be included in the sets to train and test the algorithms.

An alternative index, \(Q_p\)-score, can be used to assess the laboratory performance provided that calibration lines are developed. \(Q_p\)-score is calculated by equation 4

\[
Q_p = \frac{(a_i - \bar{a})}{\sigma_{\text{slope}}} \quad \text{(Equation 4)}
\]

where \(a_i\) is the slope of the calibration line determined by the \(i\)th data set, \(\bar{a}\) is the consensus slope value, and \(\sigma_{\text{slope}}\) is the interlaboratory standard deviation on slopes, all referred to a single NMR signal. The values \(\bar{a}\) and \(\sigma_{\text{slope}}\) are determined using \(a_i\) successfully passing the outlier test. By an analogous reasoning followed for the \(z\)-score, performance by the \(Q_p\)-score are assessed as follow:

• \(|Q_p| \leq 2\) indicates “satisfactory” performance;
• \(2 < |Q_p| < 3\) indicates “questionable” performance;
• \(|Q_p| \geq 3\) indicates “unsatisfactory” performance.

It is demonstrated that \(Q_p\)-score does not depend on instrumental features, but it depends only on molar masses of the analyte and of the reference compound, and on the number of protons generating the NMR signals, as reported in reference [7].

### 2.3 Results

The results are summarized in the scores plot (figures 2 and 3) obtained by PLS (partial least squares) regression and in the confusion matrixes (tables 2 and 3). Figure 2 and table 2 refer to the wheat extracts, while figure 3 and table 3 refer to flour extracts.

PLS was performed by using the NMR spectra selected after performance assessment. The unsatisfactory performances were not considered for classifier based on PLS. Spectra were submitted to rectangular bucketing with regular 0.05 ppm intervals and each bucket was referred to TSP signal. The training set was made of 222 NMR spectra and the test set was made of 146 spectra.
Figure 2. Scores plot deriving from PLS regression applied to NMR spectra of tubes B and C (B and C refers to two different aliquots of the same wheat sample). Variables were scaled to unit variance.

Table 2. Confusion matrix obtained by PLS analysis applied to samples B and C.

<table>
<thead>
<tr>
<th>NMR Spectra</th>
<th>B_true</th>
<th>C_true</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bpredicted</td>
<td>73</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Cpredicted</td>
<td>0</td>
<td>73</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>73</td>
<td>73</td>
<td>100%</td>
</tr>
</tbody>
</table>

Accuracy is calculated according to equation 5.

\[
\text{Accuracy} = \frac{(B_{\text{correctly predicted}} + C_{\text{correctly predicted}})}{\text{Total population}} \times 100
\]  
(Equation 5)
Figure 3. Scores plot deriving from PLS regression applied to NMR spectra of tubes D and E (D and E refers to two different aliquots of the same flour sample). Variables were scaled to unit variance.

Table 3. Confusion matrix obtained by PLS analysis applied to samples D and E.

<table>
<thead>
<tr>
<th>NMR Spectra</th>
<th>D_true</th>
<th>E_true</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>D predicted</td>
<td>73</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>E predicted</td>
<td>0</td>
<td>73</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>146</td>
<td>73</td>
<td>73</td>
</tr>
</tbody>
</table>

Also in this case, Accuracy is calculated according to equation 5, where B and C are replaced by D and E, respectively.
3. CONCLUSIONS

In light of the comments and the results reported above, the procedure reported in scheme 2 should be adopted to validate an unbiased and multi-user NMR based classification tool designed to identify the origin of food products.

Steps 2 and 3 represent backbone of the tool as they allow for multi-user data input and multi-user enquiry. Indeed, experimental evidence of the statistical equivalence of the NMR spectra plays a multiple function:

1. it demonstrates the suitability of the analytical method;
2. it quantifies the instrumental error and compares it with the natural variance of the sample population;
3. it allows for the selection of the laboratory participating to database feed (scheme 3);
4. it enables laboratories to classifier enquiry (schema 4).

Thus, a general validation procedure should provide for the organization of an interlaboratory comparison as described above.

Scheme 3. Concept map for selection of the laboratory participating to database feed.
Concerning step 4 of scheme 2, once the suitable statistical approach (PLS, LDA, SIMCA, ANN, etc.) is defined, the internal validation of the classifier should be considered satisfactory on the bases two factors:

1. the ratio between training samples and total (training + test) available samples;
2. the accuracy of the classifier.

As to point 1, a ratio lower than 80% should be considered, while accuracy higher than 95% should be accepted for satisfactory performance of the classifier.

Finally, step 5 should be considered only when step 4 is successfully passed. In such a case, classifier should be trained by using all the samples available in step 4 and its accuracy should be determined considering the correct classification of the blind samples. Also for this step, accuracy higher than 95% should be accepted for satisfactory performance of the whole classification tool. Only when steps 1-5 are fulfilled, the tool can be used for identification of unknown samples.
4. REFERENCES


3. Achievement of the Deliverable

The issues reported in this Deliverable are related to the drafting of a guidance for the validation of non-targeted food analysis methods. The work, to be understood as a "flow chart" of "best practices", is the main outcome of the research carried out by the WP18 within the Food Integrity project, in relation to the gap identified previously in this field. The WP18 working group aimed at harmonizing the information already published and considered at the time of writing, as well as implementing the procedure, considering both the analytical and the post-analytical aspects. The work was finalized considering the most critical gaps related to some key points in the “in house” validation of an analytical method.

The setting up of the non-targeted analytical method, the best practices regarding sampling, the analytical validation, the selection of the most suitable approach, the validation of the whole analytical process (also considering the acceptability of the outcomes obtainable from statistical data processing, passing through the pre-processing of data, the use of quality control samples and the criteria for the maintenance of the method) were the main targets of discussion, shared with other Food Integrity Partners (WUR, Bfr, FERA).

The present draft of the Guidance is finally the result of the interdisciplinary approach selected by all the Units belonging to the WP18 Consortium.